Traballo Fin de Mestrado

Predición de series temporais. Comparación de modelos ARIMA vs. modelos GAM

Marta Cousido Rocha

Mestrado en Técnicas Estatísticas

Curso 2015-2016

Proposta de Traballo Fin de Mestrado

Título en galego: Predición de series temporais. Comparación de modelos ARIMA vs. modelos GAM

Título en español: Predicción de series temporales. Comparación de modelos ARIMA vs. modelos GAM

English title: Time series forecasting: ARIMA vs. GAM

Modalidade: Modalidade B

Autora: Marta Cousido Rocha, Universidade de Santiago de Compostela

Director: Javier Roca Pardiñas, Universidade de Vigo

Titor: Antonio Vidal Vidal, Optare Solutions

Breve resumo do traballo:

A predición de series temporais é unha ferramenta empregada en diferentes campos co obxectivo de adiantarse aos feitos permitindo tomar decisións anticipadas. O obxectivo principal deste traballo é determinar, de entre unhas cantas metodoloxías, a máis axeitada na predición de valores futuros de tres series temporais que impactan no negocio e nos sistemas dunha operadora de telecomunicacións. O traballo comprende as seguintes tarefas que se corresponden con cada unha das partes da memoria:

- Introdución do problema.
- Revisión das metodoloxías.
- Aplicación das metodoloxías sobre as series temporais.
- Conclusións.

Recomendacións:

Ter cursado as seguintes materias do Mestrado Interuniversitario en Técnicas Estatísticas: Series de Tempo e Estatística non Paramétrica. Ademais destas materias tamén é necesario coñecer a linguaxe estatística de programación R.

Don Javier Roca Pardiñas e don Antonio Vidal Vidal informan que o Traballo Fin de Mestrado titulado

Predición de series temporais. Comparación de modelos ARIMA vs. modelos GAM

foi realizado baixo a súa dirección por dona Marta Cousido Rocha para o Mestrado en Técnicas Estatísticas. Estimando que o traballo está rematado, dan a súa conformidade para a súa presentación e defensa ante un tribunal.

En Vigo, a 1 de xullo de 2016.

O director:

Don Javier Roca Pardiñas

O titor:

Don Antonio Vidal Vidal

A autora:

Dona Marta Cousido Rocha

Agradecementos

En primeiro lugar agradecerlle á empresa Optare Solutions a oportunidade de introducirme no mercado laboral mediante esta experiencia, en particular o meu titor Antonio Vidal Vidal. No aspecto académico a Javier Roca Pardiñas como director deste proxecto.

Por outra parte agradecer a Irene Castro Conde e David Lozano Núñez, compañeiros e amigos, o trato recibido durante a estancia na empresa, así como o apoio brindado durante a escritura desta memoria. En particular gustaríame reiterar o meu agradecemento a Irene Castro Conde xa que as súas indicacións permitíronme adquirir novos coñecementos que sen dúbida foron de gran utilidade na realización deste proxecto.

Índice xeral

Resumo

Resumo en Galego

O obxectivo deste traballo é seleccionar unha metodoloxía axeitada para predicir eventos futuros de tres series temporais que impactan no negocio e nos sistemas dunha operadora de telecomunicacións. O primeiro capítulo do traballo adicámolo a introducir o problema a tratar a través da súa exposición e da mención das metodoloxías seleccionadas para a súa resolución, así como a presentar brevemente a empresa onde se levaron a cabo as prácticas. No segundo capítulo centrámonos na revisión das metodoloxías seleccionadas que son a metodoloxía Box-Jenkins, a metodoloxía GAM (do inglés Generalized Additive Models) e as Árbores de Divisións Recursivas, expondo con maior detalle as dúas primeiras e brevemente a última delas. Ao final deste capítulo tamén expomos os métodos de predición simples (un exemplo destes métodos é predicir os valores futuros como a media da serie) ademais do procedemento que empregamos para cuantificar a exactitude das predicións realizadas con cada unha das metodoloxías. O terceiro capítulo adicámolo á aplicación das metodoloxías revisadas sobre as tres series temporais, tendo presente o fundamento teórico exposto no capítulo anterior. Para rematar achegamos as conclusións xerais do traballo realizado.

English abstract

The aim of this work is to choose a suitable methodology in order to predict future events of three time series that impact in the business and the system of a telecommunications operator. The first chapter begins with a brief introduction about the company that collaborates on this project (where the internship took place) and the proposed problem. After describing the problem, we mention the selected methodologies for the study of the three time series, that are Box-Jenkins methodology, Generalized Additive Models, and Recursive Partitioning Trees. After this quick introduction, we begin the second chapter with a deep review of the Box-Jenkins methodology and the Generalized Additive Models. Next, we introduce the Recursive Partitioning Trees and simple methods, and for ending the second chapter we show the procedure that we use for quantifying the accuracy of the given predictions by each one of the procedures. The third chapter is devoted to the application of the reviewed methodologies to the three time series, and finally we give the general conclusions of the project.

Prefacio

Neste traballo imos abordar o estudo de tres series temporais importantes para unha compañía de telecomunicacións seguindo diferentes metodoloxías co obxectivo de seleccionar finalmente unha delas. Consecuentemente comezamos definindo o concepto de serie temporal, o cal presentamos como o resultado de observar os valores dunha variable ao longo do tempo en intervalos regulares. O primeiro campo de traballo no desenvolvemento da metodoloxía de análise de series temporais ten as súas raíces nos estudos de series astronómicas e climáticas, que deron lugar á teoría de procesos estocásticos estacionarios, desenvolta polos matemáticos Kolmogorov, Wiener e Cramer na primeira metade do século XX. A partires destes inicios ata hoxe en día realizáronse numerosos estudos científicos sobre a análise de series temporais en distintos campos proporcionándonos unha ampla gama de metodoloxías para o seu estudo. Neste traballo seleccionamos tres metodoloxías para o estudo das series temporais. A primeira delas é a metodoloxía Box-Jenkins, froito das investigacións realizadas polos británicos G.E.P. Box e G. Jenkins, que estudaban o problema de predición e control de series temporais industrias nos anos 60. Esta metodoloxía marca un punto e aparte no estudo das series temporais pois é unha metodoloxía unificada para estudar series estacionarias e non estacionarias, algo inexistente ata ese momento. A segunda das metodoloxías que empregamos son os Modelos Aditivos Xeneralizados, coñecidos como modelos GAM, introducidos por Hastie e Tibshirani en 1990 como unha extensión dos modelos tradicionais de regresión lineal incorporando a non linearidade e a flexibilidade dos modelos non paramétricos. Esta metodoloxía non é propia para o estudo de series temporais como é o caso da metodoloxía Box-Jenkins pero a pesar disto actualmente emprégase no estudo de series temporais de diferentes campos proporcionando resultados satisfactorios. A terceira das metodoloxías seleccionada tampouco é exclusiva para o estudo de series temporais pero tamén é empregada neste campo no ámbito da consultoría estatística, estamos a falar das Árbores de Divisións Recursivas. Ademais destas metodoloxías cabe mencionar que usaremos os denominados métodos de predición simples co obxectivo de levar a cabo certas comparativas.

O primeiro capítulo deste traballo adicámolo á presentación do problema a tratar e a xustificar a relevancia da súa resolución. O segundo capítulo iníciase coa revisión da metodoloxía Box-Jenkins tras a cal expomos certos aspectos metodolóxicos relevantes dos modelos GAM, introducindo estes modelos partindo do Modelo Lineal e pasando polo Modelo Lineal Xeneralizado. Logo presentamos brevemente o procedemento das Árbores de Divisións Recursivas, seguidamente amosamos os métodos de predición simples ou métodos inxenuos, e xa para pechar este capítulo presentamos o procedemento deseñado para cuantificar a exactitude das predicións realizadas por cada metodoloxía.

No terceiro capítulo levamos a cabo a aplicación das metodoloxías expostas nas series temporais de interese para a compañía de telecomunicacións. É dicir no estudo de cada unha das series temporais empregamos a metodoloxía Box-Jenkins, os modelos GAM, as Árbores de Divisións Recursivas e os métodos simples, achegando ao final do mesmo as conclusións acadadas. Cabe mencionar que nesta aplicación se emprega como ferramenta informática o *software* estatístico R Core Team (2015). Para rematar expomos nun último capítulo as conclusións xerais do traballo.

En canto á bibliografía empregada, para a sección introdutoria do segundo capítulo, Sección 2.2.1, empregamos dúas referencias clásicas no estudo de series temporais como son Wei (2006) e Peña (2010). Esta última tamén a empregamos xunto a Brockwell e Davis (2002) e Cryer e Chan (2008) na Sección 2.2 a cal aborda a revisión da metodoloxía Box-Jenkins. As referencias empregadas nestas

seccións son referencias habituais para un curso académico básico de series temporais. Por outra banda empregamos Wood (2006) para a revisión dos modelos GAM, pois é unha das referencias usuais no estudo destes modelos. Con respecto ás Árbores de Divisións Recursivas fundamentamos a descrición das mesmas no artigo de Hothorn et al. (2006). Para rematar, cabe mencionar que os métodos simples foron consultados en Hyndman (2011).

Para un correcto seguimento do proxecto é moi recomendable ter presentes coñecementos adquiridos nas seguintes materias do Mestrado en Técnicas Estatísticas: Series Temporais e Estatística non Paramétrica.

Capítulo 1

Introdución

Neste primeiro capítulo contextualizamos e expomos o problema abordado no presente traballo. Para iso presentamos brevemente a empresa colaboradora na realización do mesmo, Optare Solutions, e seguidamente centrámonos en expor o problema a tratar xunto á utilidade da súa resolución. Tras isto mencionamos os motivos polos cales seleccionamos cada unha das metodoloxías que empregaremos na resolución do problema, e para rematar este capítulo introdutorio recalcamos o obxectivo que desexa cumprir a empresa nesta cooperación co Mestrado en Técnicas Estatísticas.

1.1. Optare Solutions

Optare Solutions naceu no ano 2002 co obxectivo de proporcionar consultoría técnica para provisión de servizos complexos a unha operadora emerxente do mercado español. Posteriormente, outras operadoras solicitaron os seus servizos, ampliando deste modo o seu rango de clientes, feito que permitiu a empresa apostar pola investigación e o desenvolvemento creando no ano 2005 un departamento de I+D. A maior parte dos proxectos levados a cabo neste departamento pertencen ao ámbito da analítica predictiva, dentro do cal se enmarca o problema abordado neste traballo. Cabe mencionar que segundo o informe Ardán da Zona Franca, Optare Solutions é agora a terceira empresa de Galiza en inversión en proxectos de I+D (informes do 2015 e 2016).

A sede central de Optare Solutions encóntrase no parque Tecnolóxico e Loxístico de Vigo, e amais desta sede contan con outras dúas, unha delas situada en Madrid e outra en México. Para máis información acerca desta empresa pódese consultar a súa páxina web http://optaresolutions.com.

1.2. Exposición do problema

Como xa mencionamos con anterioridade o problema consiste en seleccionar unha familia de modelos axeitada para predicir eventos futuros de tres series temporais que impactan no negocio e nos sistemas dunha operadora de telecomunicacións. Dispor dunha familia de modelos apropiada á hora de predicir valores futuros permite axudar á operadora na toma de decisións co obxectivo de mellorar os resultados de negocio. No mundo dos servizos é vital coñecer as necesidades actuais do cliente e poder prever as súas necesidades futuras para realizar así accións adiantadas que permitan maximizar os beneficios da compañía. A predición de series temporais permite levar a cabo esta tarefa, por este motivo a empresa está interesada na predición das seguintes series temporais:

- "Número de altas diarias de teléfono móbil".
- "Número de baixas diarias de teléfono móbil".
- "Número de usuarios diarios do servizo de vídeo baixo demanda".

O modelado e posterior predición de cada unha destas series temporais ten un interese para a operadora. A predición das dúas primeiras series de tempo emprégase na mercadotecnia (do termo inglés "Marketing") de servizos, xa que ao dispor da previsión do número de altas e baixas esperadas a empresa pode adiantarse aos feitos, por exemplo, realizando accións de retención e fidelización de clientes en risco de baixa, se as predicións indican un número de baixas elevado, ou ofertando promocións a clientes potenciais, se o número de altas predito é inferior ao esperado. Por outra banda, o interese de predición da terceira serie mencionada está relacionado co ámbito industrial. Coñecer por adiantado o número de usuarios que se van conectar cada día ao servizo de vídeo baixo demanda permite anticiparse a posibles anomalías no seu uso axustando os medios técnicos ás necesidades esixidas.

1.3. Métodos de resolución

O estudo e análise de series temporais experimentou un desenvolvemento rápido desde o seu nacemento, por este motivo neste momento dispoñemos dunha ampla variedade de técnicas para o estudo de series temporais de entre as cales seleccionamos a metodoloxía Box-Jenkins, a metodoloxía GAM e as Árbores de Divisións Recursivas para o presente estudo.

Como xa mencionamos a Metodoloxía Box-Jenkins é unha técnica clásica de estudo de series temporais, a cal se fundamenta na consideración exclusiva dos valores pasados da serie temporal para explicar a súa evolución presente e futura. Os motivos que nos levan a considerar está técnica son os seguintes:

- Metodoloxía amplamente estudada e referenciada.
- Existencia dunha guía clara e consolidada da súa aplicación empírica.
- Utilidade na predición a corto prazo de series temporais reais.

Por outra banda seleccionamos a metodoloxía GAM porque a pesar de non ser unha técnica deseñada exclusivamente para o estudo de series temporais, como é o caso da metodoloxía Box-Jenkins, proporciona resultados satisfactorios na predición de series temporais reais. Concretamente goza dunha importante vantaxe que é a gran flexibilidade que este modelo permite nos efectos das covariables sobre a resposta.

Amais da metodoloxía Box-Jenkins e GAM empregamos as Árbores de Divisións Recursivas, pois no mundo da consultoría son moi empregadas sobre todo na clasificación de individuos debido a súa sinxela interpretación, e tamén como modelos de regresión no estudo de series temporais.

Os métodos de predición simples empregámolos co obxectivo de comparar os resultados proporcionados polas outras metodoloxías cos obtidos empregando este métodos, pois a consultoría debe coñecer se a empresa de telecomunicacións pode obter resultados similares aos que lle proporcionan empregando métodos de predición triviais, pois de ser así isto resta valor ao traballo realizado pola consultoría.

1.4. Obxectivo

O obxectivo principal deste traballo, como xa mencionamos, consiste na selección dunha familia de modelos axeitada para a predición das series temporais de relevancia para a empresa. Para iso levaremos a cabo o modelado e predición das series temporais empregando cada unha das metodoloxías mencionadas, para finalmente concluír, tras unha comparativa dos resultados obtidos, cal delas é a máis indicada para a análise das series temporais de interese.

Para cumprir o obxectivo exposto é necesario levar a cabo unha revisión das metodoloxías que imos empregar no estudo das series temporais. No seguinte capítulo centrámonos nesta tarefa.

Capítulo 2

Aspectos metodolóxicos

Neste capítulo afróntase a revisión dos fundamentos das técnicas estatísticas que serán empregadas na análise das series temporais no Capítulo 3. Na primeira parte deste capítulo expomos a definición formal de serie temporal xunto con algúns conceptos básicos relacionados con dito termo. A continuación abordamos a exposición da metodoloxía Box-Jenkins abarcando, dentro desta metodoloxía tan ampla, simplemente os coñecementos necesarios para o posterior estudo das series temporais. Na Sección 2.3 centrámonos en proporcionarlle ao lector coñecementos acerca da metodoloxía GAM que lle permitan comprender a aplicación da mesma na análise das serie temporais de interese. Tras isto expomos brevemente o procedemento básico das Árbores de Divisións Recursivas, e presentamos os métodos de predición simples. Xa para rematar o capítulo, explicamos o procedemento que empregamos para cuantificar a exactitude das predicións realizadas con cada unha das metodoloxías

2.1. Conceptos previos

Unha serie temporal é o resultado de observar os valores dunha variable ao longo do tempo en intervalos regulares. Por exemplo, a serie "número de altas diarias de teléfono móbil" é o resultado de observar os valores da variable "número de altas de teléfono móbil" todos os días. Agora ben, unha serie temporal tamén se pode definir doutros xeitos. Deseguido expomos unha definición máis formal.

Un proceso estocástico é unha familia de variables aleatorias indexada polo tempo, o cal pode ser representado polo conxunto

$$y = \{Y(w,t), w \in W, t \in I\}, \tag{2.1}$$

onde W denota o espazo mostral e I o correspondente conxunto de índices. A partir da expresión (2.1), e considerando $I = \mathbb{Z}$, temos que tomando un valor fixo do parámetro asociado ao tempo, $t^* \in \mathbb{Z}$, o resultado obtido é unha variable aleatoria sobre W, a cal representamos polo conxunto $\{Y(w,t^*), w \in W\}$. Por outra banda se fixamos $w^* \in W$, o conxunto (2.1) convértese no conxunto $\{y(w^*,t), t \in \mathbb{Z}\}$ que non é máis que unha observación do proceso estocástico que se denomina realización ou traxectoria do proceso con respecto a W. Unha serie temporal non é máis que unha realización ou traxectoria parcial dun proceso estocástico, polo que unha serie temporal de tamaño T pódese representar como $y = \{y(t), t = 1, \cdots, T\} = \{y_t, t = 1, \cdots, T\}$, obviando o parámetro w fixado. Para máis información pódese consultar Wei (2006), Sección 2.1, Páxina 6.

Dada unha serie temporal, a tarefa principal é construír un proceso estocástico que de xeito razoable puidera ter xerado unha serie temporal coas características da serie dada[1]. Por este motivo é relevante coñecer como caracterizamos un proceso estocástico. Un proceso queda caracterizado se definimos a distribución de probabilidade conxunta das variables aleatorias Y_t con $t \in \{1, \cdots, T\}$, para calquera valor de T; estas distribucións denomínanse distribucións finito-dimensionais do proceso. Polo tanto,

[1]Nótese que para as series de datos reais non existe un proceso estocástico que as xerase, polo cal o que procuramos é un proceso susceptible de xerar unha serie temporal coas características presentes na serie temporal de interese.

coñecemos a estrutura probabilística dun proceso estocástico cando se coñecen estas distribucións, que determinan a distribución de calquera subconxunto de variables.

A continuación definiremos varias medidas características e propiedades de interese dun proceso estocástico. A posesión dalgunhas destas propiedades facilitará e posibilitará a tarefa principal exposta.

Dado un proceso de valores reais, o cal denotamos de forma abreviada omitindo o parámetro w como $y = \{Y_t, t \in \mathbb{Z}\}$, podemos definir a función de medias, medida de posición de carácter central de Y_t, como

$$\mu_t = \mathbb{E}(Y_t),$$

a función de varianzas, medida do grao de variabilidade de Y_t, como

$$\sigma_t^2 = \mathrm{Var}(Y_t) = \mathbb{E}\left((Y_t - \mu_t)^2\right),$$

e a función de autocovarianzas, medida do grao de dependencia lineal existente entre Y_t e Y_{t+k}, como

$$\gamma(t, t+k) = \mathrm{Cov}(Y_t, Y_{t+k}) = \mathbb{E}\left((Y_t - \mu_t)\left(Y_{t+k} - \mu_{t+k}\right)\right).$$

O número enteiro k denomínase retardo da función de autocovarianzas.

Como é coñecido, a autocovarianza ten unidades, concretamente o cadrado da unidade na que está medida a serie temporal. Por este motivo esta medida non é axeitada para comparar series temporais con diferente unidade de medición. Polo tanto, deseguido presentamos dúas medidas adimensionais da dependencia lineal entre calquera dúas variables do proceso.

Definición 2.1.1 *Dado un proceso estocástico* $\{Y_t, t \in \mathbb{Z}\}$ *defínese a función de autocorrelacións simples (fas) entre* Y_t *e* Y_{t+k} *como* $\rho(t, t+k) = \dfrac{\gamma(t, t+k)}{\sigma_t \sigma_{t+k}}$. *Como se pode observar a función de autocorrelacións simples non é máis que un medida do grao de dependencia lineal existente entre* Y_t *e* Y_{t+k}, *con maior interpretabilidade que a función de autocovarianzas pois toma valores no intervalo* $[-1, 1]$.

Definición 2.1.2 *Dado un proceso estocástico* $\{Y_t, t \in \mathbb{Z}\}$ *defínese a función de autocorrelacións parcias (fap) entre* Y_t *e* Y_{t+k} *como*

$$\alpha(t, t+k) = \frac{\mathrm{Cov}\left(Y_t - \widehat{Y}_t^{(t,t+k)}, Y_{t+k} - \widehat{Y}_{t+k}^{(t,t+k)}\right)}{\sqrt{\mathrm{Var}\left(Y_t - \widehat{Y}_t^{(t,t+k)}\right)\mathrm{Var}\left(Y_{t+k} - \widehat{Y}_{t+k}^{(t,t+k)}\right)}},$$

onde $\widehat{Y}_t^{(t,t+k)}$ *denota ao mellor predictor lineal de* Y_t *construído a partir das variables medidas nos instantes de tempo comprendidos entre o instante* t *e o* $t+k$. *Polo tanto* $\alpha(t, t+k)$ *é unha medida do grao de dependencia lineal existente entre* Y_t *e* Y_{t+k} *unha vez que se lles extraeu o efecto lineal que sobre cada unha delas exercen as variables medidas nos instantes comprendidos entre* t *e* $t+k$.

Para máis información das medidas expostas pódese consultar Peña (2010), Subsección 3.2.1, Páxinas 87-90.

A obtención das distribucións de probabilidade do proceso é posible en certas situacións, agora ben, en xeral só podemos observar unha realización do proceso. Por isto o proceso existe conceptualmente, pero non é posible obter mostras sucesivas ou realizacións independentes do mesmo. Polo tanto, para poder estimar características do proceso, como por exemplo a función de medias ou de varianzas, debemos supor certa estabilidade ao longo do tempo, o que conduce ao concepto de estacionariedade, que definimos de contado.

Definición 2.1.3 *Un proceso estocástico* Y_t *con* $t \in \mathbb{Z}$ *dise estacionario en sentido débil* [2] *se verifica*

[2] Neste documento só presentamos o concepto de estacionariedade débil pois a estacionariedade estricta é unha condición difícil de contrastar na práctica. Este concepto pódese consultar en Peña (2010), Subsección 3.3.1, Páxina 92.

- $\mu_t = \mu$, $\forall t \in \mathbb{Z}$.
- $\sigma_t^2 = \sigma^2$, $\forall t \in \mathbb{Z}$.
- $\gamma(t, t+k) = \gamma_k$, $\forall t \in \mathbb{Z}, k \in \mathbb{N}$.

Cando o proceso estocástico é estacionario denotamos por ρ_k e α_k as autocorrelacións simples e parcias, respectivamente, existentes entre dúas variables separadas k instantes de tempo.

Como mencionamos anteriormente a estacionariedade dota ao proceso estocástico de propiedades de estabilidade na media, na varianza e nas autocovarianzas. Estas propiedades permítenos estimar distintas características do proceso a partires dunha realización parcial do mesmo $\{y_t, t = 1, ...T\}$, é dicir, permítenos estimar características do proceso a partires dunha única observación do mesmo, tal e como amosamos de contado. Supoñamos entón un proceso estacionario con media $\mu = \mathbb{E}(Y_t)$, varianza $\sigma^2 = \mathrm{Var}(Y_t)$ e covarianzas $\gamma_k = \mathrm{Cov}(Y_t, Y_{t+k})$ do que observamos unha realización $\{y_t, t = 1, ...T\}$.

Un estimador centrado da media poboacional é a media mostral. En efecto, chamando $\overline{y}$ a media mostral, temos que

$$\overline{y} = \frac{\sum_{t=1}^{T} y_t}{T}$$

é un estimador centrado da media poboacional, pois

$$\mathbb{E}(\overline{y}) = \mathbb{E}\left(\frac{\sum_{t=1}^{T} y_t}{T}\right) = \frac{\sum_{t=1}^{T} \mathbb{E}(y_t)}{T} = \frac{T\mu}{T} = \mu.$$

Exposto o xeito de estimar a media proseguimos expondo o estimador das autocovarianzas e autocorrelacións. Se a media do proceso é coñecida, o estimador das autocovarianzas de orde k é

$$\widetilde{\gamma}_k = \frac{1}{T-k} \sum_{t=1}^{T-k} (y_t - \mu)(y_{t+k} - \mu), \tag{2.2}$$

o cal é centrado para estimar $\gamma_k = \mathbb{E}((Y_t - \mu)(Y_{t+k} - \mu))$. Pola contra, se μ é descoñecida e substituímos esta polo seu estimador, $\overline{y}$, na expresión (2.2) temos que o estimador resultante non é centrado. Un estimador alternativo, que ten mellores propiedades cando μ é descoñecida, é

$$\widehat{\gamma}_k = \frac{1}{T} \sum_{t=1}^{T-k} (y_t - \overline{y})(y_{t+k} - \overline{y}). \tag{2.3}$$

O estimador (2.3) é tamén un estimador sesgado da función de autocovarianzas, a pesar disto ten mellores propiedades que o estimador (2.2) pois o seu erro cadrático de estimación é menor. Nótese que en particular a varianza do proceso se estima como $\widehat{\gamma}_0$.

A partires do estimador (2.3) podemos construír un estimador para a función de autocorrelacións simples

$$\widehat{\rho}_k = \frac{\widehat{\gamma}_k}{\widehat{\gamma}_0}.$$

Para rematar imos proporcionar un estimador da función de autocorrelacións parcias. O estimador da autocorrelación parcial de orde k, α_k, é o estimador mínimo cadrático do coeficiente α_{kk} no modelo de regresión lineal múltiple

$$y_t = \alpha_{k0} + \alpha_{k1} y_{t-1} + \cdots + \alpha_{kk} y_{t-k} + \varepsilon_t,$$

sendo $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ e independentes.

2.2. Modelos Box-Jenkins

Nesta sección abordamos a presentación dos modelos ou procesos Box-Jenkins. Comezamos coa definición dos procesos Box-Jenkins estacionarios e non estacionarios. Feito isto, continuamos coa explicación das técnicas necesarias para identificar un modelo ou proceso estocástico dos expostos como posible xerador dunha serie temporal. Seguido disto centrámonos na construción de estimadores para os parámetros do modelo identificado previamente. A continuación, presentamos as técnicas necesarias para a comprobación das hipóteses realizadas sobre o modelo, é dicir, expomos o procedemento de diagnose do modelo. Deseguido, amosamos o sistema de predición de valores futuros da serie en base ao modelo estimado. E xa para rematar proporcionamos unhas directrices básicas sobre o tratamento de datos atípicos.

2.2.1. Definición

A metodoloxía Box-Jenkins fundaméntase na procura dun proceso ou modelo estocástico que, de forma razoable, puidera ter xerado a serie temporal que desexamos estudar. Para unha correcta comprensión da metodoloxía Box-Jenkins empezamos esta presentación amosando o proceso autorregresivo, coñecido como AR (Autoregressive), o proceso de medias móbiles, coñecido como MA (Moving Average) e a súa fusión, proceso ARMA (Autoregressive Integrated Moving Average). Estes modelos son relevantes no estudo de procesos estacionarios, que son os máis simples cos que podemos tratar.

Definición 2.2.1 *Un proceso estacionario* $\{Y_t, t \in \mathbb{Z}\}$ *que admite a representación*

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + a_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + a_t,$$

para todo $t \in \mathbb{Z}$*, onde* c*,* ϕ_1*,* $\cdots$*,* ϕ_p *son constantes e* $\{a_t, t \in \mathbb{Z}\}$ *é ruído branco, é dicir, unha secuencia de variables aleatorias incorreladas, con media nula e varianza finita* σ_a^2*, coñécese como un proceso autorregresivo de orden* p *(AR(*p*)).*

Un proceso autorregresivo de orde p non é mais que un proceso que explica o valor actual, Y_t, como unha función lineal de p valores anteriores ao actual, $Y_{t-1}, Y_{t-2}, \cdots, Y_{t-p}$.

Cabe mencionar que de agora en diante indicaremos que un proceso a_t é ruído branco con media cero e varianza σ_a^2 coa notación $a_t \sim WN(0, \sigma_a^2)$. Tamén nos referiremos a estas variables co nome de innovacións.

Definición 2.2.2 *Un proceso estacionario* $\{Y_t, t \in \mathbb{Z}\}$ *que admite a representación*

$$Y_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q} = c + \sum_{i=1}^{q} \theta_i a_{t-i} + a_t,$$

para todo $t \in \mathbb{Z}$*, onde* c*,* θ_1*,* $\cdots$*,* θ_q *son constantes e* $\{a_t\} \sim WN(0, \sigma_a^2)$*, coñécese como un proceso de medias móbiles de orden* q *(MA(*q*)).*

Un proceso de medias móbiles de orde q explica o valor actual, Y_t, como unha función lineal dos q valores anteriores dun proceso de ruído branco $a_{t-1}, a_{t-2}, \cdots, a_{t-q}$.

A partir dos modelos AR e MA que acabamos de definir podemos construír un modelo lixeiramente máis complexo que é o modelo autorregresivo de medias móbiles.

Definición 2.2.3 *Un proceso estacionario* $\{Y_t, t \in \mathbb{Z}\}$ *que admite a representación*

$$\begin{aligned} Y_t &= c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q} \\ &= c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i a_{t-i} + a_t, \end{aligned} \tag{2.4}$$

para todo $t \in \mathbb{Z}$, onde c, ϕ_1, $\cdots$, ϕ_p, θ_1, $\cdots$, θ_q son constantes e $\{a_t\} \sim WN(0, \sigma_a^2)$, coñécese como un proceso autorregresivo de medias móbiles de ordes p e q (ARMA(p,q)).

En termos do operador retardo B, definido por $BY_t = Y_{t-1}$, podemos escribir de forma máis compacta a expresión (2.4) como

$$\phi(B)Y_t = c + \theta(B)a_t, \tag{2.5}$$

onde $\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$ e $\theta(B) = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)$.

O modelo ARMA que vimos de definir é unha familia moi flexible de procesos estacionarios, pois está incluído na clase de procesos lineais [3]. Esta clase de procesos forman o marco xeral para o estudo dos procesos estacionarios, xa que calquera proceso estacionario ou ben é lineal ou ben pode ser transformado para que o sexa ("Descomposición de Wold"). Agora ben, non abundan series reais xeradas por procesos estacionarios, pois soen presentar compoñentes determinísticas como tendencia e/ou estacionalidade [4]. Polo tanto necesitamos ampliar a clase de procesos ARMA, de modo que a nova clase de modelos nos permita incorporar estas características. Dende outro punto de vista, necesitamos detraer a compoñente determinística do proceso para convertelo en estacionario e modelalo seguindo un proceso ARMA.

Algúns procesos non estacionarios debido á presenza de tendencia pasan a selo simplemente coa aplicación dunha ou varias diferenzas regulares, é dicir, tras restar a cada observación Y_t a observación anterior, é dicir, o proceso diferenciado regularmente é $Y_t - Y_{t-1}$. Esta cuestión é a base da clase de modelos que deseguido definimos.

Definición 2.2.4 *Se d é un número enteiro non negativo, un proceso autorregresivo integrado de medias móbiles de ordes p, q, e d, proceso ARIMA(p,d,q) (Autoregressive Integrated Moving Average), é aquel proceso que tras a aplicación de d diferencias regulares, se converte nun proceso ARMA(p,q). En notación matemática,*

$$\{Y_t, t \in \mathbb{Z}\} \text{ é un } ARIMA(p,d,q) \Leftrightarrow (1-B)^d Y_t \text{ é un } ARMA(p,q).$$

Outra escritura da definición dun proceso ARIMA(p, d, q) é a seguinte.

Definición 2.2.5 *Un proceso $\{Y_t, t \in \mathbb{Z}\}$ é un proceso ARIMA(p,d,q) se admite unha representación do tipo*

$$\phi(B)(1-B)^d Y_t = c + \theta(B)a_t,$$

onde $\phi(B)$ e $\theta(B)$ se definen coas mesmas expresións que na ecuación (2.5), o polinomio $\phi(B)$ non ten raíces de módulo 1 e $\{a_t\} \sim WN(0, \sigma_a^2)$.

Os procesos ARIMA que acabamos de definir permiten captar non estacionariedades provocadas pola presenza de tendencia pero non capturan non estacionariedades provocadas pola presenza de compoñente estacional. Por este motivo seguidamente procedemos á ampliación da clase de procesos ARIMA presentada co obxectivo de que a nova clase capture tanto non estacionariedades provocadas pola presenza de tendencia como de compoñente estacional.

Os modelos ARMA expostos ata o momento só modelizan dependencia entre observacións consecutivas acontecidas no pasado inmediato (dependencia regular), polo cal é necesario presentar os modelos ARMA estacionais os cales permiten modelizar dependencia entre observacións acontecidas en instantes separados por múltiplos do período estacional s, é dicir, dependencia estacional.

Definición 2.2.6 *Un proceso estacionario $\{Y_t, t \in \mathbb{Z}\}$ que admite a representación*

$$\begin{aligned} Y_t &= c + \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \cdots + \Phi_P Y_{t-Ps} + a_t + \Theta_1 a_{t-s} + \Theta_2 a_{t-2s} + \cdots + \Theta_Q a_{t-Qs} \\ &= c + \sum_{i=1}^{P} \Phi_i Y_{t-is} + \sum_{i=1}^{Q} \Theta_i a_{t-is} + a_t, \end{aligned}$$

[3] A definición formal de proceso lineal pódese consultar en Brockwell e Davis (2002), Sección 2.2, Páxina 51.

[4] Con tendencia dunha serie temporal referímonos ao comportamento a longo prazo da mesma e con estacionalidade ao comportamento periódico da serie temporal.

onde $c, \Phi_1, \cdots, \Phi_P, \Theta_1, \cdots, \Theta_Q$ *son constantes,* s *é un número enteiro non negativo e* $\{a_t\} \sim WN(0, \sigma_a^2)$*, coñécese como un proceso ARMA estacional de período* s*, ARMA(P,Q)*$_s$.

Nótese que un proceso ARMA$(P, Q)_s$ se pode escribir de forma compacta como

$$\Phi(B^s)Y_t = c + \theta(B^s)a_t, \tag{2.6}$$

onde $\Phi(B^s) = (1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps})$, $\Theta(B^s) = (1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs})$, e B^s denota ao operador retardo estacional, definido por $B^s Y_t = Y_{t-s}$.

Consecuentemente, se combinamos as dúas clases de procesos ARMA expostos obtemos o seguinte proceso que permite modelizar conxuntamente dependencia regular e estacional

$$\phi(B)\Phi(B^s)Y_t = c + \theta(B)\Theta(B^s)a_t,$$

o cal se coñece como proceso ARMA estacional multiplicativo (ARMA$(p, q) \times (P, Q)_s$).

O proceso anterior modeliza conxuntamente dependencia regular e estacional pero non capta non estacionariedades provocadas por presenza de tendencia e/ou estacionalidade. Xa mencionamos anteriormente que a non estacionariedade provocada pola presenza de tendencia se podía solventar cunha ou varias diferencias regulares tal é como presentamos no modelo ARIMA, pois ben, empregando diferencias estacionais podemos solventar a non estacionariedade provocada pola presenza de compoñente estacional, sendo o proceso diferenciado estacionalmente $Y_t - Y_{t-s}$. Deseguido presentamos unha extensión do proceso ARMA estacional multiplicativo que ten en conta estas cuestións.

Definición 2.2.7 *Un proceso estocástico* $\{Y_t, t \in \mathbb{Z}\}$ *é un proceso ARIMA(p,d,q)×(P,D,Q)*$_s$ *(ou ARIMA estacional multiplicativo) se admite unha representación do tipo,*

$$\phi(B)\Phi(B^s)(1 - B)^d(1 - B^s)^D Y_t = c + \theta(B)\Theta(B^s)a_t,$$

onde $\phi(B)$ *e* $\theta(B)$ *se definen como na expresión (2.5),* $\Phi(B^s)$ *e* $\Theta(B^s)$ *se definen como na expresión (2.6),* $\{a_t\} \sim WN(0, \sigma_a^2)$*,* s, d*, e* D *son números enteiros non negativos e* $\phi(z)\Theta(z^s)$ *non ten raíces de módulo 1.*

Polo cal un ARIMA estacional multiplicativo non é máis que un proceso que tras aplicarlle d diferencias regulares e D diferencias estacionais de período s se converte nun proceso ARMA$(p, q) \times (P, Q)_s$. Estes procesos permiten modelar series temporais que presentan tendencia e compoñente estacional.

Agora ben, non todas as series non estacionarias poden ser transformadas nunha serie estacionaria empregando diferencias regulares e/ou estacionais. Moitas series temporais son estacionarias na media pero non estacionarias en termos de varianza debido a unha asociación da variabilidade da mesma co tempo, presenza pois de heterocedasticidade. Na literatura existen moitas transformacións para eliminar este suceso. No seguinte apartado presentamos unha delas, a transformación Box-Cox.

Heterocedasticidade. Transformación Box-Cox

En moitas series temporais a variabilidade das mesmas cambia a medida que cambia o nivel, é dicir,

$$\sigma_t = cf(\mu_t),$$

onde c é unha constante e f unha función da media. Se consideramos como función $f(\mu_t) = \mu_t^{1-\lambda}$, situación moi usual na práctica, unha familia de transformacións moi útiles é a familia de transformacións Box-Cox introducida por George E.P. Box e David Cox no ano 1964. Esta transformación defínese como aquela que transforma y_t en

$$y_t^{(\lambda)} = \frac{y_t^\lambda - 1}{\lambda},$$

sendo λ un número real distinto de cero. Nótese que se $\lambda = 1$ a transformación devolve a serie orixinal desprazada. Cando λ é cero a transformación Box-Cox correspóndese coa transformación logarítmica.

$$\lim_{\lambda \to 0} y_t^{(\lambda)} = \lim_{\lambda \to 0} \frac{y_t^{\lambda} - 1}{\lambda} = \log(y_t).$$

Sinalar que esta transformación só é válida para series temporais cuxos valores sexan positivos para todo instante de tempo, debido á presenza da función log. A pesar disto este aspecto pódese solventar coa selección dunha constante fixa $\varepsilon \in \mathbb{R}^+$ que verifique $y_t + \varepsilon > 0$, para todo t.

2.2.2. Identificación

Na subsección 2.2.1 mostramos dous tipos de procesos estocásticos claramente diferenciados, os procesos estacionarios dentro dos cales destacamos a familia de procesos ARMA, e os procesos non estacionarios onde definimos a familia de procesos ARIMA. O seguinte obxectivo é dotar ao lector dos coñecementos necesarios para identificar a partires dunha serie real un destes procesos como posible xerador da mesma.

Consideremos primeiramente os procesos estacionarios. Dada unha serie temporal debemos asesorarnos acerca de se dita serie puido ou non ser xerada por un proceso estacionario. Un xeito moi sinxelo e rápido de realizar esta tarefa é a través do estudo do gráfico secuencial da serie temporal e da función de autocorrelación simple mostral (*fas* mostral). Se o gráfico secuencial amosa nivel e variabilidade constantes e a *fas* mostral converxe rapidamente a cero a medida que o retardo crece, podemos asumir polo momento que a serie foi xerada por un proceso estacionario seguindo a Definición 2.1.3. Na Figura 2.1 pódese ver a modo de exemplo o gráfico secuencial dunha serie de tempo estacionaria xunto a súa *fas* mostral. Unha vez que asumimos que a serie puido ser xerada por un proceso estacionario debemos identificar os parámetros dun modelo ARMA. Deseguido expomos brevemente como levar a cabo esta labor.

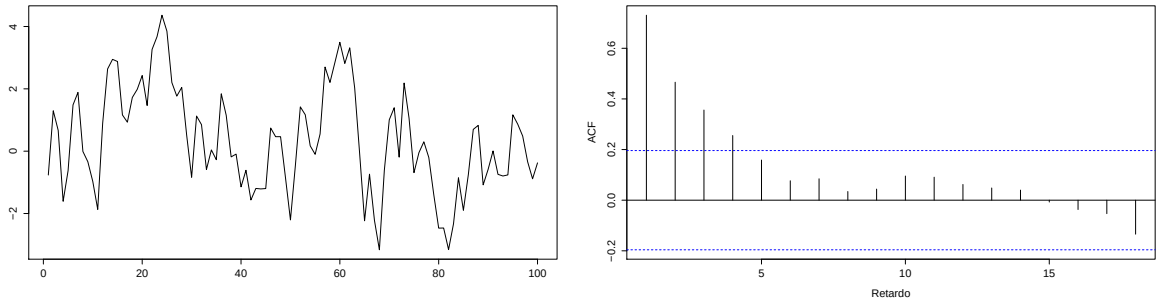


Figura 2.1: Identificación de procesos estacionarios. Á esquerda podemos ver o gráfico secuencial dunha serie de tempo xerada por un proceso estacionario. Á dereita podemos observar a súa *fas* mostral. As liñas descontinuas presentes na *fas* mostral representan o límite inferior e superior do correspondente contraste de significación realizado sobre a *fas*. Consecuentemente, as correlacións que se encontran dentro destes límites pódense asumir nulas.

Comecemos por expor o procedemento de identificación do proceso estocástico estacionario máis simple, o proceso de ruído branco. Un proceso de ruído branco caracterízase porque tanto as autocorrelacións simples como as parcias son nulas. Agora ben no ámbito práctico non dispoñemos de función teóricas como é o caso das autocorrelacións simples e parcias, polo cal é necesario extraer conclusións das súas estimacións, isto é o que nos permite o seguinte resultado. Se unha serie temporal de tamaño

T, suficientemente grande, foi xerada por un proceso de ruído branco verifícase que

$$\widehat{\rho}_k \approx N\left(0, \frac{1}{\sqrt{T}}\right) \text{ e } \widehat{\alpha}_k \approx N\left(0, \frac{1}{\sqrt{T}}\right) \ \forall k.$$

Polo cal unha serie temporal de tamaño T foi xerada por un proceso de ruído branco se para cada retardo k se ten que

$$\widehat{\rho}_k \in \left(\frac{-z_{\alpha/2}}{\sqrt{T}}, \frac{z_{\alpha/2}}{\sqrt{T}}\right) \text{ e } \widehat{\alpha}_k \in \left(\frac{-z_{\alpha/2}}{\sqrt{T}}, \frac{z_{\alpha/2}}{\sqrt{T}}\right),$$

sendo α o nivel de significación e $z_{\alpha/2}$ o cuantil de orde $1-\alpha/2$ dunha normal estándar.

Proseguimos coa identificación dun proceso autorregresivo de orde p. Este proceso estocástico caracterízase porque as autocorrelacións simples presentan unha estrutura complexa froito dunha mestura de exponenciais e sinusoidais o cal se reflexa en moitos coeficientes non nulos, mentres que as autocorrelacións parcias se anulan para todo retardo maior que o orde p. Na Figura 2.2a amosamos a *fas* e *fap* dun proceso AR(1) como ilustración do comportamento descrito. Dada unha serie temporal descoñecemos estas funcións teóricas polo que de novo necesitamos extraer conclusións das súas estimacións, obxectivo que permite realizar a seguinte afirmación. Unha serie temporal, de tamaño suficientemente grande, xerada por un proceso AR(p) verifica que

$$\widehat{\alpha}_k \approx N\left(0, \frac{1}{\sqrt{T}}\right), \ \forall k > p.$$

É dicir, unha serie temporal de tamaño T foi xerada por un proceso autorregresivo de orde p se para cada retardo $k > p$ se ten que

$$\widehat{\alpha}_k \in \left(\frac{-z_{\alpha/2}}{\sqrt{T}}, \frac{z_{\alpha/2}}{\sqrt{T}}\right).$$

Para máis información acerca da identificación deste proceso pódese ver Cryer e Chan (2008), Sección 6.2, Páxina 115.

Dun xeito similar podemos identificar un proceso de medias móbiles de orde q, pois existe unha dualidade entre procesos AR e MA, de maneira que a *fap* dun MA(q) ten a estrutura da *fas* dun AR(q) e *fas* dun MA(q) ten a estrutura da *fap* dun AR(q). Na Figura 2.2b preséntanse estas funcións para un proceso MA(1). De novo na práctica estudar a *fas* e *fap* non é posible polo cal deseguido expomos un resultado que permite identificar este proceso seguindo a *fas* e *fap* mostral. Unha serie temporal de tamaño T suficientemente grande xerada por un proceso MA(q) verifica que

$$\widehat{\rho}_k \approx N\left(0, \frac{\sqrt{1+2(\rho_1^2+\cdots+\rho_q^2)}}{\sqrt{T}}\right), \ \forall k > q.$$

É dicir, unha serie temporal de tamaño T foi xerada por un proceso de medias móbiles de orde q se para cada retardo $k > q$ se ten que

$$\widehat{\rho}_k \in \left(-z_{\alpha/2}\frac{\sqrt{1+2(\widehat{\rho}_1^2+\cdots+\widehat{\rho}_q^2)}}{\sqrt{T}}, z_{\alpha/2}\frac{\sqrt{1+2(\widehat{\rho}_1^2+\cdots+\widehat{\rho}_q^2)}}{\sqrt{T}}\right).$$

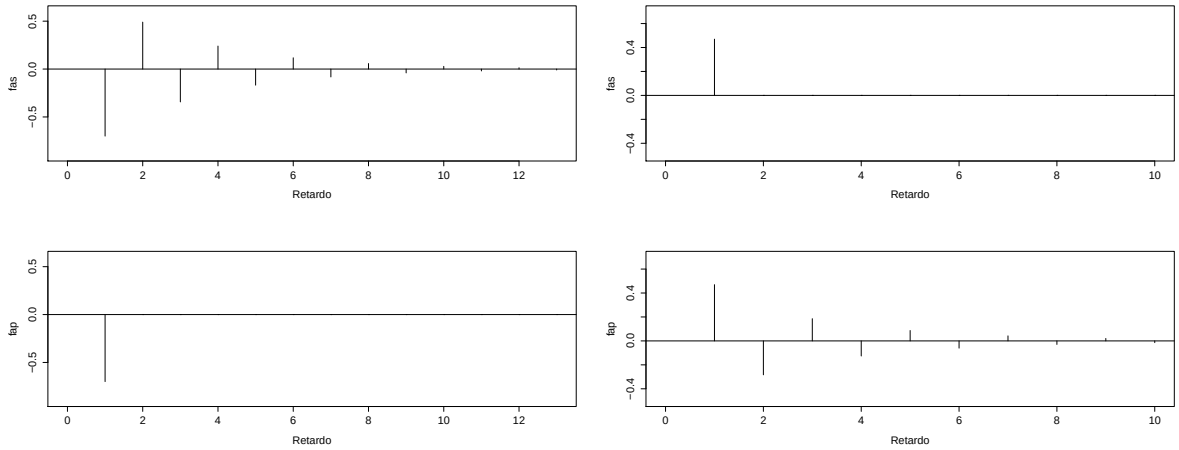
Para máis información acerca da identificación deste proceso pódese consultar Cryer e Chan (2008), Sección 6.1, Páxinas 110-112.

Os modelos AR(p) e MA(q) son casos particulares dun modelo ARMA onde un das ordes é nulo, polo cal debemos abordar arestora a identificación dun modelo ARMA(p,q) con $p \neq 0$ e $q \neq 0$. A *fas* e *fap* destes procesos é o resultado da superposición das súas propiedades AR e MA:

- Na *fas* observamos certos coeficientes iniciais dependentes do orde da parte MA e despois un decrecemento ditado pola parte AR.

- A *fap* presenta valores iniciais dependentes do orde do AR seguidos do decrecemento debido a parte MA.

Esta estrutura complexa fai que o orden dun proceso ARMA sexa difícil de identificar seguindo o estudo da *fas* e *fap*, véxase Figura 2.3.



(a) *Fas* e *fap* dun proceso AR(1)

(b) *Fas* e *fap* fun proceso MA(1)

Figura 2.2: Identificación de procesos AR e MA.

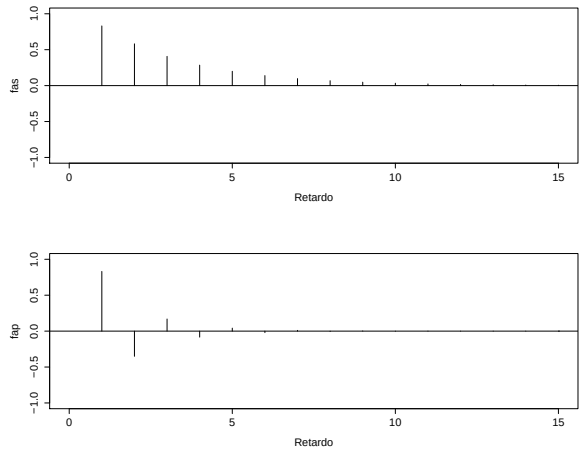


Figura 2.3: Ilustración da dificultade de interpretación da *fas* e *fap* dun proceso ARMA. *Fas* e *fap* correspondentes a un proceso ARMA(1,1)

Debido a isto existen outros métodos para a identificación deste modelo. Un dos métodos máis coñecidos é o proposto por Tsay e Tao no ano 1984 onde suxiren empregar a función de autocorrelación simple extendida (consultar Cryer e Chan 2008, Sección 6.2, Páxinas 116-117), a pesar disto neste documento empregaremos outro procedemento que expomos deseguido. Un método para determinar as ordes dun modelo ARMA é seleccionar o modelo ARMA que minimiza o Criterio de Información Bayesiano definido como

$$BIC = -2\log(L(\widehat{\varphi}_{n+1})) + n\log(T), \tag{2.7}$$

onde $\widehat{\varphi}_{n+1}$ é o vector formado polas estimacións de máxima verosimilitude dos parámetros do modelo e da varianza do proceso de ruído branco σ_a^2, L denota a función de máxima verosimilitude, n é o número de parámetros do modelo e T o tamaño da serie temporal. Nótese que o número de parámetros dun modelo ARMA(p,q) é $n = p + q$ se non consideramos constante na formulación do mesmo ou ben $n = p + q + 1$ se formulamos o modelo seguindo a expresión (2.4).

O que procura este criterio é un modelo que proporcione un bo axuste sen demasiados parámetros. Concretamente, se a serie temporal foi realmente xerada por un proceso ARMA(p,q), entón as ordes especificados na minimización do criterio BIC son consistentes, é dicir, estas aproxímanse as verdadeiras ordes a medida que aumenta o tamaño mostral.

Tamén é usual empregar o Criterio de Información de Akaike (AIC) ou o Criterio de Información de Akaike Correxido (AICC), os cales se poden consultar en Cryer e Chan (2008), Sección 6.5, Páxinas 130-131, onde tamén se pode consultar o criterio BIC exposto.

Cabe destacar que ata o momento só tratamos a identificación dun proceso ARMA que modela dependencia regular, non obstante na Subsección 2.2.1 tamén presentamos os procesos ARMA estacionais. Con respecto a identificación destes modelos mencionar que seguimos en xeral as técnicas desenvolvidas para os modelos ARMA con dependencia regular. Simplemente cabe expor brevemente como se comportan as funcións de autocorrelación simple e parcial dun modelo AR$(P)_s$ e MA$(Q)_s$.

Un proceso autorregresivo estacional de orde P con período estacional s cumpre que tanto a *fas* como a *fap* é nula nos retardos non estacionais, é dicir, nos retardos que non son múltiplos de s xunto con que a *fas* presenta moitos coeficientes non nulos nos múltiplos de s (retardos estacionais) e a *fap* anúlase para todo retardo maior que Ps. Na Figura 2.4a pódese ver unha ilustración deste comportamento.

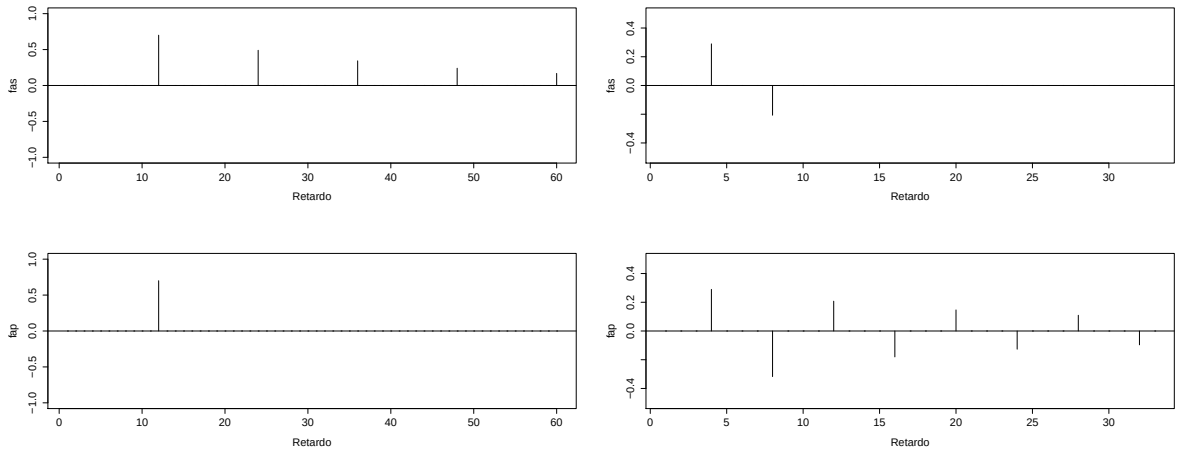
Un proceso de medias móbiles estacional de orde Q con período estacional s cumpre que tanto a *fas* como a *fap* é nula nos retardos non estacionais xunto con que a *fap* presenta moitos coeficientes non nulos nos retardos estacionais e a *fas* se anula para todo retardo maior que Qs. Para maior claridade proporcionase un exemplo na Figura 2.4b. De novo, na práctica é necesario extraer conclusións da *fas* e *fap* mostrais, tarefa que se pode abordar sen máis que empregar un contraste de significación.

Analogamente ao exposto para un ARMA de dependencia regular, cando temos un $ARMA(P,Q)_s$ con $P \neq 0$ e $Q \neq 0$ é moi complexo identificar as ordes seguindo a *fas* e *fap* mostrais, pois estas presentan unha estrutura complexa e consecuentemente moitos coeficientes non nulos nos retardos estacionais. Por este motivo, neste caso empregariamos o criterio BIC exposto para os modelos ARMA de dependencia regular.

Como mencionamos anteriormente, cando combinamos un modelo ARMA que modeliza dependencia regular cun modelo ARMA estacional construímos un modelo ARMA estacional multiplicativo o cal permite modelar conxuntamente dependencia regular e estacional. Seguidamente expomos brevemente como se comportan as autocorrelacións simples e parcias deste modelo rematando así a explicación relativa a identificación dos procesos estacionarios expostos na Subsección 2.2.1.

A *fas* dun proceso ARMA estacional multiplicativo caracterízase porque nos retardos inferiores á metade do período estacional se observa a *fas* da parte regular do modelo ARMA e nos retardos estacionais a *fas* da parte estacional do modelo. Ademais a ambos lados dos retardos estacionais repítese a *fas* da parte regular (se a *fas* no retardo estacional é negativa aparecerá invertida). Con respecto á *fap* sucede o mesmo nos retardos inferiores a metade do período estacional obsérvase a *fap* da parte regular do modelo ARMA e nos retardos estacionais a *fap* da parte estacional do modelo. Amais disto, á dereita de cada retardo estacional aparece a *fap* da parte regular (se a *fap* no retardo estacional é positiva aparecerá invertida), e á esquerda de cada retardo estacional aparece a *fas* da parte regular (se a *fap* no retardo estacional é negativa aparecerá invertida). Na Figura 2.5 proporciónamos dous exemplos para ilustrar o comportamento descrito. Con este coñecemento podemos identificar as ordes dun modelo ARMA$(p,q) \times (P,Q)_s$ con $p = 0$ ou $q = 0$ e con $P = 0$ ou $Q = 0$ simplemente empregando os contastes de significación necesarios para a *fas* e *fap* mostral. Noutra situación será necesario empregar o Criterio

de Información Bayesiano para especificar correctamente o modelo.



(a) *Fas* e *fap* dun modelo $\mathrm{AR}(1)_{12}$ (b) *Fas* e *fap* dun modelo $\mathrm{MA}(2)_4$

Figura 2.4: Identificación procesos ARMA estacionais.

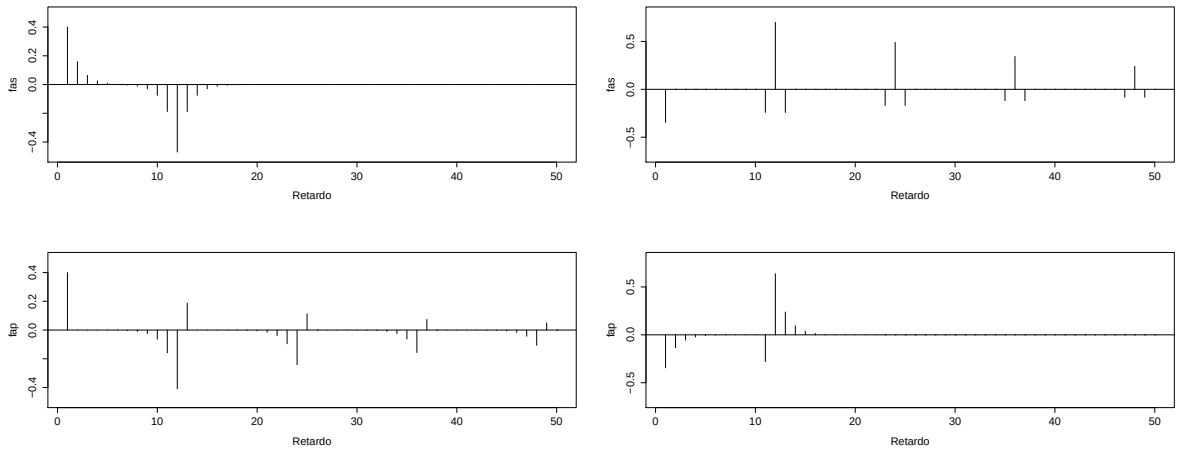


Figura 2.5: Identificación procesos ARMA estacionais multiplicativos. Á esquerda pódese ver a *fas* e *fap* dun modelo $\mathrm{AR}(1)\times\mathrm{MA}(1)_{12}$. Á dereita pódese ver a *fas* e *fap* dun modelo $\mathrm{MA}(1)\times\mathrm{AR}(1)_{12}$.

Abordemos agora a identificación dos procesos non estacionarios. Empecemos por describir o proceso de identificación dun modelo autorregresivo de medias móbiles $\mathrm{ARIMA}(p,d,q)$. Dada unha serie temporal, consideraremos un proceso ARIMA como posible xerador da mesma cando mostre non estacionariedade provocada exclusivamente pola presenza de tendencia. A presenza de tendencia pode ser observada no gráfico secuencial da serie temporal. A pesar disto, en certas situacións pode que o gráfico secuencial non aporte a suficiente información. Por este motivo o estudo do gráfico secuencial compleméntase coa análise da *fas* mostral, pois se esta toma valores positivos, a miúdo próximos a un nos primeiros retardos, e decae lentamente a medida que o retardo crece, isto indica presenza de tendencia na serie temporal. Para maior claridade pódese ver o exemplo exposto na Figura 2.6.

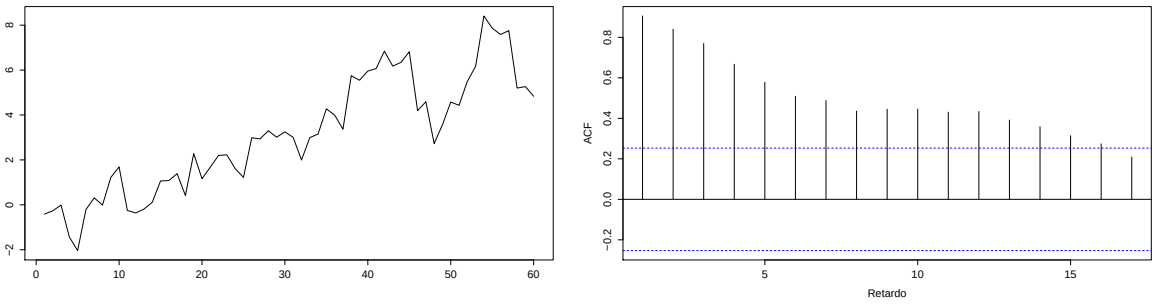


Figura 2.6: Identificación procesos ARIMA. Á esquerda pódese ver o gráfico secuencial dunha serie temporal con tendencia. Á dereita pódese ver a *fas* mostral de dita serie.

Unha vez que comprobamos a presenza de tendencia debemos proseguir transformando a serie temporal co obxectivo de que esta se converta nunha serie xerada por un proceso estacionario. Como xa mencionamos, isto conséguese coa aplicación sucesiva de d diferencias regulares, isto é, se despois dunha diferenza regular persiste a existencia de tendencia, diferenciamos novamente a serie xa diferenciada, e proseguimos con este procedemento ata obter unha serie sen tendencia. Se esta serie temporal sen tendencia é estacionaria podemos modelala a través dun ARMA cuxa identificación é coñecida.

Porén existen moitas series temporais de datos reais que tras a eliminación da tendencia presente na mesma non se poden modelar empregando procesos estacionarios, como é caso do modelo ARMA, debido á non estacionariedade da mesma. Esta falta de estacionariedade vén provocada en moitas ocasións pola presenza dunha compoñente estacional. Neste punto é onde cobra unha relevancia o modelo ARIMA estacional multiplicativo. Por este motivo é importante describir que características delatan a falta de estacionariedade debida á presenza de compoñente estacional. A compoñente estacional moitas veces pode ser detectada no gráfico secuencial da serie temporal, pero outras veces é necesario estudar a *fas* mostral pois se esta presenta forte correlación positiva no retardo estacional, converxe lentamente a cero e presenta periodicidade do mesmo período ca serie podemos afirmar existencia de compoñente estacional, véxase Figura 2.7. Unha vez eliminada a compoñente estacional, mediante diferencias estacionais, podemos obter unha serie estacionaria. Se este é o caso pasaríamos a modelala empregando un modelo ARMA ben con só dependencia regular, con só dependencia estacional ou con ambos tipos de dependencia segundo sexa necesario.

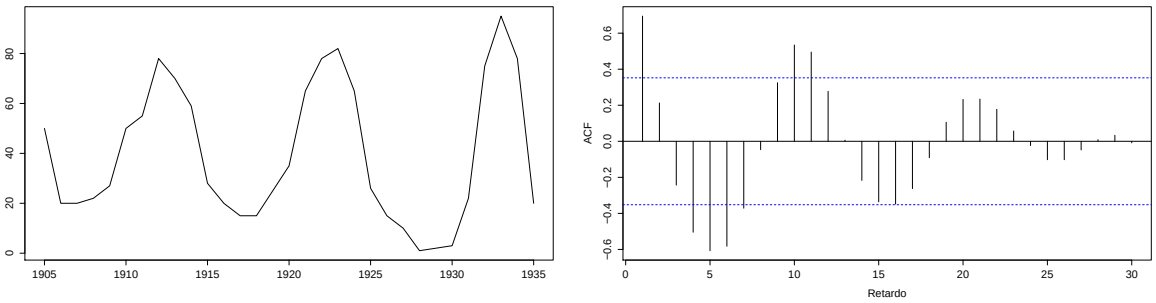


Figura 2.7: Identificación de procesos ARIMA estacionais. Á esquerda pódese ver o gráfico secuencial dunha serie temporal con compoñente estacional de período 5. Á dereita pódese ver a *fas* mostral de dita serie.

Para rematar cabe mencionar que cando temos unha serie temporal non estacionaria é moi importante que antes de proceder á comprobación da existencia de tendencia ou compoñente estacional revisemos a variabilidade da mesma, pois se esta non é estable debemos transformar a serie previamente ao inicio dos pasos de identificación que vimos de expor (transformación Box-Cox exposta na Subsección 2.2.1).

2.2.3. Estimación

Unha vez temos identificado un modelo Box-Jenkins como posible xerador da serie temporal debemos abordar a estimación dos seus parámetros. Como se pode deducir do exposto en subseccións previas, a estimación dun modelo ARIMA redúcese a estimación dun modelo ARMA, pois un modelo ARIMA convértese nun modelo ARMA tras as diferencias necesarias. Neste documento ilustraremos os métodos de estimación para un modelo ARMA de dependencia regular, pois a adaptación a outro dos modelos expostos é inmediata. Isto é, discutiremos a estimación dos parámetros c, $\phi_1, \cdots, \phi_p$, $\theta_1, \cdots, \theta_q$ e σ_a^2 no modelo

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i a_{t-i} + a_t.$$

De entre os diferentes métodos de estimación seleccionamos para tratar neste documento o método de estimación por mínimos cadrados condicionados e o método de máxima verosimilitude (MV).

Mínimos cadrados condicionados

A técnica de mínimos cadrados baséase en minimizar as diferenzas entre os valores observados e os correspondentes valores axustados. Estas diferenzas son coñecidas como residuos asociados a estimación. Concretamente, este método tenta minimizar a suma dos cadrados dos residuos, polo cal proporciona o modelo axustado que mellor se aproxima aos datos segundo o criterio de mínimo erro cadrático.

Como vimos de expor o método de estimación que mostramos fundaméntase na minimización dos residuos, polo cal definimos estes para o modelo ARMA considerado. Os residuos asociados as estimacións $\widetilde{c}$, $\widetilde{\phi}_1, \cdots, \widetilde{\phi}_p$, $\widetilde{\theta}_1, \cdots, \widetilde{\theta}_q$ defínense como

$$\widehat{a}_t = y_t - \left(\widetilde{c} + \widetilde{\phi}_1 Y_{t-1} + \cdots + \widetilde{\phi}_p Y_{t-p} + \widetilde{\theta}_1 \widehat{a}_{t-1} + \cdots + \widetilde{\theta}_q \widehat{a}_{t-q}\right). \tag{2.8}$$

Consecuentemente a estimación dos parámetros $(c, \phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q)$ por medio do método de mínimos cadrados obtense a través dos valores $(\widehat{c}, \widehat{\phi}_1, \cdots, \widehat{\phi}_p, \widehat{\theta}_1, \cdots, \widehat{\theta}_q)$ que minimizan a función $S(\widetilde{c}, \widetilde{\phi}_1, \cdots, \widetilde{\phi}_p, \widetilde{\theta}_1, \cdots, \widetilde{\theta}_q) = \sum_{t=1}^{T} \widehat{a}_t^2$, isto é,

$$(\widehat{c}, \widehat{\phi}_1, \cdots, \widehat{\phi}_p, \widehat{\theta}_1, \cdots, \widehat{\theta}_q) = \arg \min_{\widetilde{c}, \widetilde{\phi}_1, \cdots, \widetilde{\theta}_q} S\left(\widetilde{c}, \widetilde{\phi}_1, \cdots, \widetilde{\phi}_p, \widetilde{\theta}_1, \cdots, \widetilde{\theta}_q\right) = \arg \min \sum_{t=1}^{T} \widehat{a}_t^2, \tag{2.9}$$

sendo $\widehat{a}_t$ os residuos definidos na ecuación (2.8).

Neste método de estimación encontrámonos con dúas dificultades que impiden levar a cabo a mesma. A primeira delas prodúcese cando $p > 0$, pois neste caso non podemos obter os residuos $\widehat{a}_1, \widehat{a}_2, \cdots, \widehat{a}_p$, pois estes dependen de valores no observados de $Y_0, Y_{-1}, \cdots, Y_{1-p}$. Este inconveniente pódese solventar eliminando os residuos descoñecidos $\widehat{a}_1, \widehat{a}_2, \cdots, \widehat{a}_p$, da suma de cadrados definida na función S, é dicir, considerando a seguinte función corrixida

$$S_c\left(\widetilde{c}, \widetilde{\phi}_1, \cdots, \widetilde{\phi}_p, \widetilde{\theta}_1, \cdots, \widetilde{\theta}_q\right) = \sum_{t=p+1}^{T} \widehat{a}_t^2.$$

Solucionado este primeiro atranco no procedemento de estimación seguimos expondo outra dificultade. Como vimos de expor os residuos $\widehat{a}_1, \widehat{a}_2, \cdots, \widehat{a}_p$ son descoñecidos, por este motivo decidimos

eliminar estes da suma de cadrados, agora ben isto non soluciona totalmente o problema pois se $q > 0$ temos que $\widehat{a}_{p+1}$ depende dos valores $\widehat{a}_p, \widehat{a}_{p-1}, \cdots, \widehat{a}_{p+1-q}$ os cales dependen a súa vez dos valores non observados de Y_t. É importante notar que se dada a serie temporal fixamos os valores de

$$\widehat{a}_p, \widehat{a}_{p-1}, \cdots, \widehat{a}_{p+1-q},$$

isto permítenos construír iterativamente os seguintes residuos $\widehat{a}_{p+1}, \widehat{a}_{p+2}, \cdots, \widehat{a}_T$ e por conseguinte realizar o procedemento de estimación. Tendo en conta que $a_t \sim WN(0, \sigma_a^2)$ decidimos fixar a cero os residuos descoñecidos chegando así a definición do método de mínimos cadrados condicionados.

A estimación dos parámetros $(c, \phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q)$ por medio do método de mínimos cadrados condicionados obtense a través dos valores $(\widehat{c}, \widehat{\phi}_1, \cdots, \widehat{\phi}_p, \widehat{\theta}_1, \cdots, \widehat{\theta}_q)$ que minimizan a función

$$S_c\left(\widetilde{c}, \widetilde{\phi}_1, \cdots, \widetilde{\phi}_p, \widetilde{\theta}_1, \cdots, \widetilde{\theta}_q\right) = \sum_{t=p+1}^{T} \widehat{a}_t^2, \tag{2.10}$$

condicionada a $\widehat{a}_p = \widehat{a}_{p-1} = \cdots = \widehat{a}_{p+1-q} = 0$.

Se $q = 0$ a estimación consiste na resolución dun problema de minimización lineal polo que non é necesario unha técnica iterativa para minimizar a función S_c. Pola contra se $q > 0$ este problema convértese nun problema de regresión non lineal cuxa resolución necesita un método de optimización numérica, por exemplo o método de Gauss-Newton. Por último cabe mencionar que cando T é suficientemente grande fixar a cero os residuos $\widehat{a}_p, \widehat{a}_{p-1}, \cdots, \widehat{a}_{p+1-q}$ ten pouca importancia na estimación final dos parámetros mentres que con traxectorias parciais do proceso máis reducidas isto comeza a cobrar relevancia. Para maior detalle sobre este método de estimación consultar Shumway e Stoffer (2011), Sección 3.6, Páxinas 129-130.

Máxima verosimilitude

A estimación de máxima verosimilitude dos parámetros $c, \phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q$ e σ_a^2 obtense a través dos valores que dan maior credibilidade á serie observada $y_1, \cdots, y_T$. A credibilidade que os valores $\widetilde{c}, \widetilde{\phi}_1, \cdots, \widetilde{\phi}_p, \widetilde{\theta}_1, \cdots, \widetilde{\theta}_q, \widetilde{\sigma}_a^2$ dan á serie observada mídese a través da función de máxima verosimilitude,

$$L_{y_1,\cdots,y_T}\left(\widetilde{c}, \widetilde{\phi}_1, \cdots, \widetilde{\phi}_p, \widetilde{\theta}_1, \cdots, \widetilde{\theta}_q, \widetilde{\sigma}_a^2\right) = f_{\widetilde{c},\widetilde{\phi}_1,\cdots,\widetilde{\phi}_p,\widetilde{\theta}_1,\cdots,\widetilde{\theta}_q,\widetilde{\sigma}_a^2}\left(y_1, \cdots, y_T\right),$$

onde $f_{\widetilde{c},\widetilde{\phi}_1,\cdots,\widetilde{\phi}_p,\widetilde{\theta}_1,\cdots,\widetilde{\theta}_q,\widetilde{\sigma}_a^2}$ denota á función de densidade conxunta do vector aleatorio $\left(\widetilde{Y}_1, \cdots, \widetilde{Y}_T\right)^t$ procedente dun proceso ARMA con parámetros $\widetilde{c}, \widetilde{\phi}_1, \cdots, \widetilde{\phi}_p, \widetilde{\theta}_1, \cdots, \widetilde{\theta}_q, \widetilde{\sigma}_a^2$.

Polo tanto, a estimación dos parámetros c, ϕ_1, $\cdots$, ϕ_p, θ_1, $\cdots$, θ_q e σ_a^2 por medio de máxima verosimilitude obtense a través dos valores $\widehat{c}$, $\widehat{\phi}_1$, $\cdots$, $\widehat{\phi}_p$, $\widehat{\theta}_1$, $\cdots$, $\widehat{\theta}_q$ e $\widehat{\sigma}_a^2$ que maximizan á función de verosimilitude L, isto é,

$$\left(\widehat{c}, \widehat{\phi}_1, \cdots, \widehat{\phi}_p, \widehat{\theta}_1, \cdots, \widehat{\theta}_q, \widehat{\sigma}_a^2\right) = \arg \max_{\widetilde{c},\widetilde{\phi}_1,\cdots,\widetilde{\phi}_p,\widetilde{\theta}_1,\cdots,\widetilde{\theta}_q,\widetilde{\sigma}_a^2} L_{y_1,\cdots,y_T}\left(\widetilde{c}, \widetilde{\phi}_1, \cdots, \widetilde{\phi}_p, \widetilde{\theta}_1, \cdots, \widetilde{\theta}_q, \widetilde{\sigma}_a^2\right).$$

Nótese que a maximización da función de verosimilitude se efectúa cun algoritmo de optimización non lineal. É moi usual empregar o método baseado no algoritmo de Gauss-Newton (Peña 2010, Apéndice 10.1).

Unha desvantaxe do método é que previamente ao proceso de maximización debemos traballar na especificación da función de densidade conxunta do proceso, tarefa usualmente complexa.

2.2.4. Diagnose

Tras a estimación dun modelo Box-Jenkins, o seguinte paso é comprobar que as hipóteses básicas realizadas sobre o mesmo son certas. Isto é o que se coñece como diagnose do modelo. A diagnose dun

modelo Box-Jenkins require verificar que as hipóteses básicas realizadas sobre os residuos son certas. Estes deben ter (1) media cero, (2) varianza constante e (3) falta de correlación para calquera retardo. Se o modelo axustado non cumpre calquera destas hipóteses debemos descartar este como posible xerador da serie temporal. Agora ben, a estas tres hipóteses engádese a hipótese de normalidade pois é unha hipótese conveniente xa que nos garante que a incorrelación implica independencia, e que non estamos deixando información por modelar. Nótese que o incumprimento da hipótese de normalidade non inválida o modelo axustado. Con respecto ás tres condicións obrigatorias expostas mencionar que (1) é pouco restritiva mentres que (2) e (3) son condicións máis fortes, concretamente a verificación da condición (3) é fundamental para asegurar que o modelo axustado é correcto.

Contraste de media cero

Para contrastar a hipótese de que os residuos do modelo teñen esperanza nula, supoñendo que temos T residuos, calculamos a súa media

$$\overline{a} = \frac{\sum_{t=1}^{T} \widehat{a}_t}{T},$$

e a súa varianza

$$\widehat{\sigma}_a^2 = \frac{\sum_{t=1}^{T} (\widehat{a}_t - \overline{a})^2}{T},$$

e concluímos que $E[\widehat{a}_t] \neq 0$, se

$$\frac{\overline{a}}{\widehat{\sigma}_a/\sqrt{T}}$$

é significativamente grande con relación a distribución normal estándar.

Contrastes de autocorrelación

O procedemento habitual para verificar a incorrelación dos residuos é debuxar dúas liñas paralelas a distancia $2/\sqrt{T}$ da orixe na súa función de autocorrelación simple estimada, e comprobar se todos os coeficientes $\widehat{\rho}_k$ están dentro destes límites de confianza.

Ademais deste contraste individual cabe presentar un contraste global de que os primeiros h coeficientes son cero, o contraste de Ljung-Box. Se os residuos son realmente ruído branco os coeficientes de correlación estimados $\widehat{\rho}_k$ son asintoticamente normais, con esperanza nula e varianza $\dfrac{(T-k)}{T(T+2)}$ (o razoamento encóntrase en Peña 2010, Capítulo 3). Isto permítenos afirmar que o estatístico

$$Q(h) = T(T+2) \sum_{k=1}^{h} \frac{\widehat{\rho}_k^2}{T-k},$$

se distribúe, asintoticamente, como unha χ^2 con m grados de liberdade, sendo m igual a h menos o número de parámetros estimados no modelo. Consecuentemente concluímos que o modelo é incorrecto segundo a hipótese (3) cando o valor do estatístico presentado é maior que o percentil 0.95 da distribución χ_m^2.

Contrastes de normalidade

A hipótese de que os residuos seguen unha distribución normal pódese comprobar con calquera dos contrastes de normalidade habituais. Un contraste sinxelo é o coñecido como contraste de Jarque-Bera, o cal consiste en calcular o coeficiente de asimetría e o coeficiente de curtosis dos residuos,

$$\alpha_1 = \frac{\sum_{t=1}^{T} (\widehat{a}_t - \overline{a})^3}{\widehat{\sigma}_a^3} \text{ e } \alpha_2 = \frac{\sum_{t=1}^{T} (\widehat{a}_t - \overline{a})^4}{\widehat{\sigma}_a^4},$$

e seguidamente empregar que baixo a hipótese nula de normalidade a variable

$$X = \frac{T\alpha_1^2}{6} + \frac{T(\alpha_2 - 3)^2}{24}$$

segue unha distribución χ_2^2. Polo que seguindo dito contraste rexeitamos a normalidade dos residuos cando o valor de X supera ou iguala ao percentil 0.95 da distribución χ_2^2.

Outro contraste de normalidade excelente é o coñecido como contraste de Shapiro-Wilk. En esencia, este contraste calcula a correlación entre os residuos e os cuantís correspondentes á distribución normal, sendo unha correlación baixa unha evidencia en contra da hipótese nula de normalidade. Para maior detalle consultar Thode (2002), Subsección 2.3.1, Páxina 27.

Finalmente convén estudar sempre o gráfico dos residuos estimados $\widehat{a}_t$ ao longo do tempo, pois a estabilidade da varianza dos residuos compróbase mediante o estudo deste gráfico. Se á vista dos residuos estimados parece existir heterocedasticidade debemos descartar o modelo axustado. Cabe mencionar que en certas situacións o estudo do gráfico dos residuos non proporciona información suficiente sendo necesario aplicar un contraste de homocedasticidade coma o presente en Peña (2010) Subsección 11.3.2, Páxina 327. Outra utilidade do gráfico dos residuos estimados é que situando no mesmo os límites de control $\pm 2\widehat{\sigma}_a$ e $\pm 3\widehat{\sigma}_a$ e estudando os puntos situados fóra dos mesmos, podemos detectar valores atípicos cuxo tratamento abordaremos máis adiante.

2.2.5. Predición

Unha vez comprobamos que o modelo axustado é correcto, o seguinte obxectivo é predicir a partires da serie de tempo observada o valor futuro do proceso dentro de k instantes de tempo, isto é, o valor de Y_{T+k}. Esta predición coñécese como predición con orixe T e horizonte k, e denotarémola por $\widehat{y}_T(k)$.

Como se amosa en Peña (2010), Subsección 8.2.1, Páxinas 224-226, o preditor que minimiza o erro cadrático medio de predición é a esperanza de Y_{T+k} condicionada á información dispoñíbel. É dicir, dada a serie temporal $y_1, \cdots, y_T$ o predictor de Y_{T+k} que minimiza o erro cadrático medio é

$$\widehat{y}_T(k) = \mathbb{E}(Y_{T+k}|Y_1 = y_1, \cdots, Y_T = y_T).$$

Para ilustrar este procedemento poderíase supoñer unha serie xerada por un ARIMA multiplicativo e así mostrar o caso máis xeral de todos, pero por simplicidade suporemos en primeiro lugar que a serie foi xerada por un AR(p) con p coñecido e en segundo lugar por un MA(q) con q coñecido.

Supoñamos que se dispón dunha realización de tamaño T, $y_1, \cdots, y_T$, dun proceso AR(p) cuxos parámetros foron estimados previamente. Se desexamos predicir o seguinte valor da serie temporal, $\widehat{y}_T(1)$, debémonos basear na seguinte relación

$$Y_{T+1} = \widehat{c} + a_{T+1} + \widehat{\phi}_1 Y_T + \widehat{\phi}_2 Y_{T-1} + \cdots + \widehat{\phi}_p Y_{T+1-p},$$

e calcular a esperanza de Y_{T+1} condicionada aos valores observados, tal e como se amosa a continuación.

$$\begin{aligned}
\widehat{y}_T(1) = \mathbb{E}\left(Y_{T+1}|Y_1 = y_1, \cdots, Y_T = y_T\right) &= \\
\mathbb{E}\left(\widehat{c} + a_{T+1} + \widehat{\phi}_1 Y_T + \widehat{\phi}_2 Y_{T-1} + \cdots + \widehat{\phi}_p Y_{T+1-p}|Y_1 = y_1, \cdots, Y_T = y_T\right) &\overset{(1)}{=} \\
\widehat{c} + \mathbb{E}(a_{T+1}|Y_1 = y_1, \cdots, Y_T = y_T) + \widehat{\phi}_1 \mathbb{E}(Y_T|Y_1 = y_1, \cdots, Y_T = y_T) + & \\
+\widehat{\phi}_2 \mathbb{E}(Y_{T-1}|Y_1 = y_1, \cdots, Y_T = y_T) + \cdots + \widehat{\phi}_p \mathbb{E}(Y_{T+1-p}|Y_1 = y_1, \cdots, Y_T = y_T) &\overset{(2)}{=} \\
\widehat{c} + \mathbb{E}(a_{T+1}|Y_1 = y_1, \cdots, Y_T = y_T) + \widehat{\phi}_1 y_T + \widehat{\phi}_2 y_{T-1} + \cdots + \widehat{\phi}_p y_{T+1-p} &\overset{(3)}{=} \\
\widehat{c} + \widehat{\phi}_1 y_T + \widehat{\phi}_2 y_{T-1} + \cdots + \widehat{\phi}_p y_{T+1-p}. &
\end{aligned}$$

Na igualdade (1) empregamos certas propiedades da esperanza condicionada. Concretamente empregamos que dadas as variables aleatorias X, Y e Z tense que $\mathbb{E}(aX|Y = y) = a\mathbb{E}(X|Y = y)$ para

todo $a \in \mathbb{R}$, $\mathbb{E}(X+Y|Z=z) = \mathbb{E}(X|Z=z) + \mathbb{E}(Y|Z=z)$ e $\mathbb{E}(a+X|Y=y) = a + \mathbb{E}(X|Y=y)$. Por outra parte na igualdade (2) utilizamos que os valores das variables $Y_1, \cdots, Y_T$ foron observados e polo tanto a esperanza condicionada coincide co valor observado. Para rematar na igualdade (3) fixemos uso de que a predición a partires da serie temporal do valor de a_{T+1} é a súa media, cero, pois a serie non contén información acerca desta variable, nin en xeral das variables a_t con $t > T$. Deste xeito tan sinxelo obtemos a predición a horizonte 1. O procedemento para a obtención da predición con orixe T e horizonte 2 é o mesmo, simplemente debemos ter en conta que non dispomos do valor da variable Y_{T+1} e polo tanto a esperanza desta variable coincidirá coa predición da mesma $\widehat{y}_T(1)$ realizada con anterioridade. Proseguindo do mesmo xeito podemos obter facilmente a predición para calquera horizonte $k > 0$.

Supoñamos agora que se dispón dunha realización de tamaño T, $y_1, \cdots, y_T$, dun proceso MA(q) cuxos parámetros foron estimados previamente. Se desexamos predicir o seguinte valor da serie temporal, $\widehat{y}_T(1)$, debémonos basear na seguinte relación

$$Y_{T+1} = \widehat{c} + a_{T+1} + \widehat{\theta}_1 a_T + \widehat{\theta}_2 a_{T-1} + \cdots + \widehat{\theta}_q a_{T+1-q},$$

e calcular a esperanza de Y_{T+1} condicionada aos valores observados, tal e como se amosa deseguido.

$$\begin{aligned} \widehat{y}_T(1) = \mathbb{E}\left(Y_{T+1}|Y_1 = y_1, \cdots, Y_T = y_T\right) &= \\ \mathbb{E}\left(\widehat{c} + a_{T+1} + \widehat{\theta}_1 a_T + \widehat{\theta}_2 a_{T-1} + \cdots + \widehat{\theta}_q a_{T+1-q} | Y_1 = y_1, \cdots, Y_T = y_T\right) &\overset{(1)}{=} \\ \widehat{c} + 0 + \widehat{\theta}_1 \mathbb{E}(a_T|Y_1 = y_1, \cdots, Y_T = y_T) + & \\ + \widehat{\theta}_2 \mathbb{E}(a_{T-1}|Y_1 = y_1, \cdots, Y_T = y_T) + \cdots + \widehat{\theta}_q \mathbb{E}(a_{T+1-q}|Y_1 = y_1, \cdots, Y_T = y_T) & \end{aligned} \tag{2.11}$$

No cálculo anterior só empregamos propiedades xa mencionadas con anterioridade no caso do proceso AR(p). Agora ben para obter finalmente a predición necesitamos dispor duns valores axeitados para a esperanza de a_t con $t < T$. A predición deste valores a partires da serie temporal non é inmediata pois esta contén información sobre estas variables. Un procedemento para calcular as predicións destas variables consiste en fixar un valor inicial, en xeral $a_1 = 0$, e calcular recursivamente os valores a_t con $t = 2, \cdots, T$ empregando a identidade $a_t = Y_t - \widehat{c} - \widehat{\theta}_1 a_{t-1} - \widehat{\theta}_2 a_{t-2} - \cdots - \widehat{\theta}_q a_{t-q}$.

Nótese que outro procedemento para predicir a_t é a través dunha combinación lineal de $y_1, \cdots, y_T$. Os coeficientes da combinación lineal de $y_1, \cdots, y_T$ que mellor predí a_t pódense obter a partires da media e das autocovarianzas de $Y_1, \cdots, Y_T$, e das covarianzas entre cada unha destas variables e a_t.

Se denotamos por $\widehat{a}_1(0), \cdots, \widehat{a}_T(0)$ as predicións obtidas para as variables $a_1, \cdots, a_T$ empregando calquera dos dous procedementos expostos, e retomamos o cálculo presente na ecuación (2.11) temos que

$$\widehat{y}_T(1) = \mathbb{E}\left(Y_{T+1}|Y_1 = y_1, \cdots, Y_T = y_T\right) = \widehat{c} + \widehat{\theta}_1 \widehat{a}_T(0) + \widehat{\theta}_2 \widehat{a}_{T-1}(0) + \cdots + \widehat{\theta}_q \widehat{a}_{T+1-q}(0).$$

Co procedemento que acabamos de expor podemos obter a predición para calquera horizonte $k > 0$.

Este método de predición, que ilustramos nos casos particulares dun proceso AR(p) e MA(q), pódese aplicar a calquera dos modelos expostos na Subsección 2.2.1 sen máis que seguir o procedemento exposto nos exemplos tratados tendo presente certos conceptos mencionados nos mesmos.

2.2.6. Valores atípicos

Con moita frecuencia ocorren nas series reais feitos puntuais que descoñecemos. As observacións afectadas por estas intervencións poden presentar unha estrutura distinta das demais e aparecer como valores atípicos, é dicir, datos que aparentemente non foron xerados igual que os demais. O correcto tratamento de valores atípicos é moi importante pois estes valores poden sesgar a estimación dos parámetros e xerar predicións sesgadas se non se identifican e tratan de maneira axeitada.

Consideramos dous tipos de valores atípicos, atípicos aditivos e innovativos. Diremos que a serie temporal presenta un atípico aditivo (AO) no instante h se o valor da serie temporal nese instante foi xerado de maneira distinta ao resto. O modelo que segue unha serie observada, z_t, afectada por un atípico aditivo no instante h é

$$Z_t = \begin{cases} Y_t & t \neq h \\ Y_t + \omega_A & t = h \end{cases},$$

onde y_t é a serie non contaminada por atípicos que supoñemos segue un modelo ARIMA, o que indícamos como $Y_t = \psi(B)a_t$. Entón, o modelo que segue a serie observada, z_t, é

$$Z_t = \omega_A I_t^{(h)} + \psi(B)a_t,$$

onde $I_t^{(h)} = 0$ para $t \neq h$ e $I_h^{(h)} = 1$.

Por outra parte, diremos que a serie temporal presenta un atípico innovativo (IO) no instante h cando o valor da innovación nese punto está afectado por unha cantidade descoñecida debido a un suceso imprevisto. O modelo que segue unha serie temporal que presenta un atípico innovativo de magnitude ω_I no instante h é

$$Z_t = \psi(B)(\omega_I I_t^{(h)} + a_t).$$

Na práctica, a posición e a natureza dos valores atípicos que poden aparecen na serie temporal é descoñecida. Por este motivo necesitamos un procedemento para identificar e clasificar estes valores. Esquematicamente o modo de proceder consiste en:

1. Detectar o momento de aparición do dato atípico.

2. Identificar o tipo de dato atípico.

3. Estimar a magnitude do mesmo.

Os detalles necesarios para levar a cabo cada unha das etapas enunciadas pódense encontrar en Peña (2010), Sección 13.6.

2.3. Modelo Aditivo Xeneralizado

Nesta sección trataremos os conceptos necesarios para a comprensión do Modelo Aditivo Xeneralizado, modelo GAM, para deste xeito poder aplicalo no modelado das series temporais. Para iso comezaremos pola exposición do modelo de regresión máis simple, o Modelo Lineal, o cal goza dunha ampla aplicación histórica debido a súa sinxela formulación e aplicación, e á intuitiva interpretación dos seus resultados. A pesar disto este modelo presenta unha serie de inconvenientes que limitan a súa correcta aplicación práctica, e que motivaron o nacemento do Modelo Lineal Xeneralizado, coñecido como GLM (do inglés Generalized Linear Models), clara extensión do Modelo Lineal. Seguidamente expomos o Modelo Lineal Xeneralizado e o seu proceso de estimación adaptado ao tipo de variable resposta que nos encontramos no caso práctico que trataremos neste traballo. O modelo GLM é un modelo moi útil e empregado na práctica pero segue a presentar un inconveniente, a excesiva rixidez da súa modelización. Esta é a razón pola cal xurdiu a partires do modelo GLM o modelo GAM, que extende ao modelo GLM proporcionando unha maior flexibilidade ao modelo de regresión. Consecuentemente expomos o modelo GAM adaptado ao caso práctico que nos ocupa.

2.3.1. Modelo Lineal

O obxectivo de calquera estudo de regresión é atopar un modelo matemático que se axuste aos datos, permitindo unha interpretación razoable da relación entre a variable resposta, Y, e o conxunto de p covariables $\mathbf{X} = (X_1, \cdots, X_p)$. Nun modelo de regresión a relación existente entre a variable resposta e as covariables consideradas segue a seguinte formulación.

$$Y = \mu(\mathbf{X}) + \varepsilon,$$

onde $\mu(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$, é dicir, $\mu(\mathbf{X})$ é a media condicional que contabiliza a influencia dos valores das covariables na resposta media, e ε é definida como variable de erro.

A estrutura elixida para o termo $\mu(\mathbf{X})$ determina o modelo de regresión considerado. O modelo de regresión máis simple é o modelo de regresión lineal o cal considera que o efecto das covariables sobre a resposta media é lineal. Máis concretamente, este modelo considera $\mu(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ sendo $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ parámetros descoñecidos. Ademais disto o modelo de regresión lineal parte da suposición de que a variable ε se distribúe de acordo cunha normal de media cero e varianza constante que non depende do valor das covariables $\mathbf{X}$. A pesar de ser un modelo amplamente empregado en diferentes contextos, o uso correcto destes modelos está moi limitado pois só se pode empregar cando a variable resposta é normal e homocedástica. Isto fai que en moitos contextos prácticos estes modelos proporcionen resultados erróneos debido a súa aplicación a datos que non verifican estas asuncións.

Co obxectivo de transformar unha variable que non se encontra nas hipóteses do modelo lineal nunha que si o faga xurdiron diferentes técnicas. Por exemplo para converter unha variable heterocedástica nunha homocedástica pódese empregar a coñecida transformación Box-Cox. Agora ben estas técnicas teñen o inconveniente de que levan os datos a escalas artificiais xunto con que non sempre conseguen que os datos pasen a verificar as hipóteses establecidas no modelo. Por este motivo máis recentemente aparecen os Modelos Lineais Xeneralizados que son unha extensión dos modelos de regresión clásicos que superan algunhas das limitacións presentes nestes.

2.3.2. Modelo Lineal Xeneralizado

Os Modelos Lineais Xeneralizados permiten que a distribución da variable resposta sexa outra distinta a distribución normal. A estrutura xeral dun modelo GLM é

$$\mu(\mathbf{X}) = h(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p),$$

o cal se pode reescribir como

$$h^{-1}(\mu(\mathbf{X})) = g(\mu(\mathbf{X})) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

onde g é unha función monótona suave denominada función "link", e $\beta_0, \beta_1, \cdots, \beta_p$ son os parámetros descoñecidos. Ademais, nun GLM xeralmente realízase a suposición de que a variable Y segue unha distribución pertencente á familia de distribucións exponenciais. A familia de distribucións exponenciais inclúe varias distribucións útiles para a modelaxe práctica como son por exemplo a Poisson, a Binomial, a Normal ou a Gamma. Dependendo da distribución da variable resposta temos diferentes funcións "link" as cales dan lugar a diferentes modelos. Por exemplo, se a resposta segue unha distribución Normal a función link é a identidade e o GLM convértese nun modelo lineal. Agora ben neste traballo imos supoñer que a variable resposta segue unha distribución Poisson, e consecuentemente que a función " link" é o logaritmo natural, pois as series que imos estudar miden variables de conteo. Consecuentemente temos o seguinte modelo

$$\mu(\mathbf{X}) = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p),$$

ou equivalentemente,

$$\ln(\mu(\mathbf{X})) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{2.12}$$

Estimación

Sexa $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ unha mostra, con vector de covariables $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})$, seguindo o modelo (2.12). Unha idea que pode xurdir para levar a cabo a estimación do vector de coeficientes $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)$ é empregar o método de máxima verosimilitude, no cal se trata de obter os coeficientes que maximizan o logaritmo da función de verosimilitude. Agora ben se tentamos estimar os coeficientes mediante o método de máxima verosimilitude encontrámonos co inconveniente de que as ecuacións resultantes do problema de maximización formulado na aplicación deste método son non lineais. A non linearidade das ecuacións imposibilita a resolución exacta das mesmas sendo necesario aplicar algún método iterativo para a súa resolución aproximada. Estes métodos consisten en considerar unha solución inicial $\widehat{\boldsymbol{\beta}}^0$ de $\boldsymbol{\beta}$ e ir dando iterativamente estimacións $\widehat{\boldsymbol{\beta}}^1, \widehat{\boldsymbol{\beta}}^2, \widehat{\boldsymbol{\beta}}^3, \cdots$ que deben converxer a un valor $\widehat{\boldsymbol{\beta}}$ denominado estimador de máxima verosimilitud de $\boldsymbol{\beta}$.

Un dos métodos máis utilizados é o algoritmo "Fisher scoring" (pode verse Wood 2006, Subsección 2.1.2 onde se trata este algoritmo no caso de máis dunha variable resposta) que pode ser visto, en cada iteración, como un problema de mínimos cadrados ponderados. Deseguido enuméranse os pasos deste algoritmo.

- **Inicio.** Calcular os coeficiente iniciais $\widehat{\beta}_0^0 = h^{-1}\left(\overline{Y}\right) = g\left(\overline{Y}\right) = \ln\left(\overline{Y}\right)$, $\widehat{\beta}_1^0 = \cdots = \widehat{\beta}_p^0 = 0$ sendo $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$.
- **Paso 1.** Calcular, para $i = 1, \cdots, n$, as respostas linealizadas

$$Z_i = \eta_i\left(\widehat{\boldsymbol{\beta}}^0\right) + g'\left(\widehat{\mu}_{0i}\right)\left(Y_i - \widehat{\mu}_{0i}\right) = \eta_i\left(\widehat{\boldsymbol{\beta}}^0\right) + \frac{\left(Y_i - \widehat{\mu}_{0i}\right)}{\widehat{\mu}_{0i}}$$

  e os pesos

$$W_i = \frac{1}{g'\left(\widehat{\mu}_{0i}\right)^2 \mathrm{Var}\left(Y_i\right)} = \widehat{\mu}_{0i}$$

  sendo

$$\eta_i\left(\widehat{\boldsymbol{\beta}}^0\right) = \widehat{\beta}_0^0 + \widehat{\beta}_1^0 X_{i1} + \cdots + \widehat{\beta}_p^0 X_{ip}$$

  e

$$\widehat{\mu}_{0i} = \exp\left(\eta_i\left(\widehat{\boldsymbol{\beta}}^0\right)\right).$$

- **Paso 2.** Obter axustando un Modelo Lineal ponderado os coeficientes actualizados

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \cdots, \widehat{\beta}_p\right)^t = \left(\mathbb{X}^t\mathbb{W}\mathbb{X}\right)^{-1}\mathbb{X}^t\mathbb{W}\mathbf{Z}$$

  con

$$\mathbb{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix} \text{ e } \mathbb{W} = diag\left(W_1, \cdots, W_n\right).$$

- **Paso 3.** Repetir o **Paso 1** e **Paso 2** substituíndo as estimacións iniciais, $\widehat{\boldsymbol{\beta}}^0$, polas estimacións $\widehat{\boldsymbol{\beta}}$ obtidas no **Paso 2**, ata acadar a converxencia. Un criterio de converxencia que se pode empregar é parar o algoritmo cando se verifique que

$$\frac{||\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^0||}{||\widehat{\boldsymbol{\beta}}^0||} \leq \varepsilon,$$

  sendo $||\cdot||$ a norma euclídea.

2.3.3. Modelo Aditivo Xeneralizado

Os modelos GLM son unha extensión dos Modelos Lineais que permiten abordar unha ampla gama de problemas prácticos. A pesar disto, como xa mencionamos, en ocasións as modelizacións obtidas mediante técnicas paramétricas resultan demasiado ríxidas, sendo preciso o desenvolvemento e a aplicación de métodos máis flexibles. Como mencionamos no Prefacio, Hastie e Tibshirani propoñen en 1990 o Modelo Aditivo Xeneralizado que evita a suposición de linearidade, proporcionándolle deste xeito maior flexibilidade ao modelo.

O Modelo Aditivo Xeneralizado extende ao Modelo Lineal Xeneralizado substituíndo ao predictor estritamente lineal da forma

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

por un predictor non paramétrico da forma

$$\eta = \beta_0 + f_1(X_1) + \cdots + f_p(X_p), \tag{2.13}$$

onde $f_j(X_j)$ é o efecto parcial suave e descoñecido de X_j no predictor. Para evitar problemas de identificación do modelo debemos engadir á formulación do mesmo a condición $\mathbb{E}\left[f_j\left(X_j\right)\right] = 0$. Este novo predictor permite introducir no modelo todo tipo de efectos e relacións non lineais entre as variables.

Como xa mencionamos estamos interesados no caso particular no que a resposta considerada segue unha distribución Poisson, consecuentemente deseguido amosamos a formulación do modelo GAM con resposta Poisson.

$$\mu\left(\mathbf{X}\right) = \exp(\beta_0 + f_1(X_1) + \cdots + f_p(X_p)),$$

ou equivalentemente,

$$\ln\left(\mu\left(\mathbf{X}\right)\right) = \beta_0 + f_1(X_1) + \cdots + f_p(X_p). \tag{2.14}$$

Estimación

Para levar a cabo a estimación dun modelo GAM é usual empregar unha adaptación do algoritmo Fisher scoring exposto na estimación do modelo GLM, Subsección 2.3.2. A modificación deste método para unha correcta adecuación á estimación dun modelo GAM realízase no **Paso 2** onde axustamos un Modelo Aditivo ponderado en lugar dun Modelo Lineal ponderado.

Sexa $\left(\mathbf{X}_1, Y_1\right), \cdots, \left(\mathbf{X}_n, Y_n\right)$ unha mostra de observacións. A estimación do modelo GAM (2.14) pódese obter mediante a seguinte adaptación do algoritmo iterativo "Fisher scoring".

- **Inicio.** Calcular as estimacións iniciais $\widehat{\beta}_0^0 = h^{-1}\left(\overline{Y}\right) = g\left(\overline{Y}\right) = \ln\left(\overline{Y}\right)$, $\hat{f}_1^0 = \cdots = \hat{f}_p^0 = 0$ sendo $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$.
- **Paso 1.** Calcular, para $i = 1, \cdots, n$, as respostas linealizadas

$$Z_i = \hat{\eta}_i^0 + g'\left(\widehat{\mu}_{0i}\right)\left(Y_i - \widehat{\mu}_{0i}\right) = \hat{\eta}_i^0 + \frac{\left(Y_i - \widehat{\mu}_{0i}\right)}{\widehat{\mu}_{0i}}$$

  e os pesos

$$W_i = \frac{1}{g'\left(\widehat{\mu}_{0i}\right)^2 \operatorname{Var}\left(Y_i\right)} = \widehat{\mu}_{0i}$$

  sendo

$$\hat{\eta}_i^0 = \widehat{\beta}_0^0 + \hat{f}_1^0\left(X_{i1}\right) + \cdots + \hat{f}_p^0\left(X_{ip}\right)$$

  e

$$\widehat{\mu}_{0i} = \exp\left(\hat{\eta}_i^0\right).$$

- **Paso 2.** Axustar un Modelo Aditivo de $\mathbf{Z}$ sobre $\mathbb{X}$, ponderado por $\mathbb{W}$, obtendo para $i = 1, \cdots, n$ as actualizacións

$$\hat{f}_1(X_{i1}), \cdots, \hat{f}_p(X_{ip})$$

e

$$\hat{\eta}_i = \widehat{\beta}_0 + \hat{f}_1(X_{i1}) + \cdots + \hat{f}_p(X_{ip}).$$

- **Paso 3.** Repetir o **Paso 1** e **2** substituíndo $\hat{\eta}_i^0$ por $\hat{\eta}_i$ ata conseguir converxencia. Por exemplo podemos cesar o algoritmo cando se verifique $|\hat{\eta}_i - \hat{\eta}_i^0| \leq \varepsilon$.

O procedemento de estimación do modelo aditivo no **Paso 2** depende do número e da naturaleza das funcións parciais incluídas no modelo. Cando as funcións parciais son paramétricas, por exemplo $f_j(X_j) = \alpha_j X_j$, o modelo de regresión aditivo resolvese usando mínimos cadrados ponderados pois o modelo GAM convertese nun modelo GLM. Agora ben, cando ditas funcións non son paramétricas, necesitamos axustar un Modelo Aditivo ponderado para conseguir as actualizacións $\hat{f}_j$.

Para a estimación do Modelo Aditivo ponderado presente no **Paso 2** existen diferentes técnicas, neste traballo centrarémonos nunha delas que é a metodoloxía de Regresión Spline Penalizada. Esta metodoloxía parte da idea de representar cada unha das funcións $f_j(x_j)$ mediante unha combinación lineal dunha base de funcións, concretamente dunha base de B-splines.

$$f_j(x_j) = \sum_{k=1}^{K_j} b_{jk}(x_j)\beta_{jk} \tag{2.15}$$

onde $b_{jk}(.), k = 1, \cdots, K_j$ son as funcións da base de B-splines considerada, polo cal son funcións coñecidas, K_j é o número de nodos considerado, e $\beta_{jk}(.), k = 1, \cdots, K_j$ son parámetros descoñecidos.

Con esta representación, cada función f_j pasa a ter unha forma paramétrica dependendo unicamente dos parámetros $\beta_{jk}, k = 1, \cdots, K_j$. Así, o modelo aditivo do **Paso 2** pode ser visto como un modelo lineal.

$$\eta_i = \sum_{k=1}^{K_1} b_{1k}(x_{1i})\beta_{1k} + \cdots + \sum_{k=1}^{K_p} b_{pk}(x_{pi})\beta_{pk} = \mathbf{X_i}\boldsymbol{\beta}. \tag{2.16}$$

sendo

$$\mathbf{X}_i = \begin{pmatrix} b_{11}(x_{1i}) & \cdots & b_{1K_1}(x_{1i}) & \cdots & b_{p1}(x_{pi}) & \cdots & b_{pK_p}(x_{pi}) \end{pmatrix}$$

e

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1K_1} & \cdots & \beta_{p1} & \cdots & \beta_{pK_p} \end{pmatrix}^t.$$

Chegados a este punto unha idea que se pode considerar para estimar o modelo é empregar as técnicas de estimación do modelo GLM. Agora ben isto presenta o inconveniente de que previamente a aplicación das técnicas expostas para os modelos GLM teríamos determinar o número e a posición dos nodos para cada unha das covariables.

Un xeito de evitar o problema exposto consiste en seleccionar un número de nodos suficientemente alto, e controlar o grado de suavización incorporando unha penalización no proceso de estimación dos coeficientes do modelo (2.16). Consecuentemente, os coeficientes $\boldsymbol{\beta}$ son estimados minimizando a suma de residuos cadrados penalizados

$$\sum_{i=1}^{n} W_i (Z_i - \mathbf{X}_i\boldsymbol{\beta})^2 + \lambda_1 \int (f''(X_1))^2 dx_1 + \cdots + \lambda_p \int (f''(X_p))^2 dx_p, \tag{2.17}$$

como se pode observar a variabilidade de cada función suave contrólase penalizando o grado de curvatura da función mediante a súa segunda derivada. Os valores λ_j denomínanse parámetros de suavizado, valores de $\lambda_j \to \infty$ levan a estimar f_j como unha recta, mentres que valores próximos a cero dan lugar

á interpolación dos datos. Polo tanto un valor demasiado pequeno ou demasiado grande do parámetro de suavizado implica que a estimación da función suave $\hat{f}_j$ non está próxima á verdadeira función f_j.

O inconveniente de determinar o número e a posición dos nodos para cada covariable transformouse agora no problema de estimación do parámetro de suavización λ_j para $j = 1 \cdots, p$. Como vimos de tratar a determinación destes parámetros é altamente relevante pois ten unha moita influencia na estimación final das funcións suaves consideradas. Existen diferentes criterios e métodos de estimación do parámetro de suavizado, entre os máis empregados encóntranse Validación Cruzada (CV), Validación Cruzada Xeneralizada (GCV), e o Criterio UBRE (Un-Biased Risk Estimator). Este último só se pode empregar en modelos con resposta non gaussiana, pódese consultar Wood (2006), Sección 4.5.

Para abordar brevemente a exposición da expresión do criterio GCV no caso dun modelo GAM é útil unha explicación previa dos criterios CV e GCV nun modelo que considera unha única función suave e unha única covariable, pois o criterio GCV no modelo GAM e unha mera extensión de dito criterio neste modelo simple. Consideremos pois o modelo $y_i = f(x_i) + \varepsilon_i$, sendo y_i a variable resposta, x_i a covariable, e ε_i variables aleatorias i.i.d $N(0, \sigma^2)$, para $i = 1, \cdots, n$.

O ideal sería elixir un λ tal que a estimación $\hat{f}$ este o máis próxima posible a función suave f. Un criterio moi útil para esta tarefa é tomar o λ que minimiza o coeficiente M,

$$M = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}(x_i) - f(x_i) \right)^2,$$

agora ben a función f é descoñecida polo cal non é posible aplicar este criterio no ámbito práctico. Ante esta situación xorde o enfoque de validación cruzada ordinaria que busca o λ que minimiza

$$\zeta_0 = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}^{[-i]}(x_i) - y_i \right)^2,$$

onde $\hat{f}^{[-i]}(x_i)$ é a estimación de $f(x_i)$ realizada con todos os datos a excepción do dato i.

Este enfoque de validación cruzada ordinaria é razoable, pero é ineficiente no cálculo polo que se prefire usar a seguinte expresión.

$$\zeta_0 = \frac{1}{n} \frac{\sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2}{(1 - A_{ii})^2},$$

onde $\hat{f}$ é a estimación da función suave realizada coa totalidade da mostra e $\mathbb{A}$ é unha matriz de influencia cuxa expresión depende da matriz $\mathbb{X}$ e dunha matriz $\mathbb{S}$ cuxos coeficientes son coñecidos e dependen dunha base de splines moi simple (para maior información consúltese Wood 2006, Sección 3.2).

Na práctica os pesos $1 - A_{ii}$ sóense substituír pola media dos pesos, traza$(\mathbb{I} - \mathbb{A})/n$, chegando así o criterio de validación cruzada xeneralizada,

$$\zeta_g = \frac{n \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i) \right)^2}{(\mathrm{tr}(\mathbb{I} - \mathbb{A}))^2}. \tag{2.18}$$

Este criterio presenta unha serie da vantaxes fronte ao criterio CV entre a que se encontra unha mellora computacional.

A extensión do criterio GCV exposto ao modelo GAM é moi simple, o criterio GCV nun modelo GAM busca o valor de λ que minimiza a seguinte expresión

$$\zeta_g = \frac{n || \sqrt{\mathbb{W}} \left(\mathbf{Z} - \mathbb{X}\boldsymbol{\beta} \right) ||^2}{(n - \mathrm{tr}(\mathbb{A}))^2},$$

como se pode observar a suma residual de cadrados presente na expresión (2.18) substitúese pola desviación $|| \sqrt{\mathbb{W}} \left(\mathbf{Z} - \mathbb{X}\boldsymbol{\beta} \right) ||^2 \approx D(\boldsymbol{\beta})$ construída co obxectivo de asimilarse o comportamento da suma

residual de cadrados. Na práctica será necesario empregar unha estimación do vector $\boldsymbol{\beta}$ para levar a cabo a selección do λ óptimo seguindo o criterio GCV, quedando así a expresión a minimizar neste criterio

$$\zeta_g = \frac{nD(\hat{\boldsymbol{\beta}})}{(n - \mathrm{tr}(\mathbb{A}))^2}.$$

2.3.4. Aplicación do GAM en series temporais

A metodoloxía Box-Jenkins exposta na Sección 2.2 está adicada exclusivamente á análise, modelización e predición de series temporais. Pola contra, a metodoloxía GAM revisada na presente sección é unha metodoloxía cunha ampla gama de campos de aplicación, a cal ata o momento expuxemos considerando resposta Poisson pero sen centrarnos no campo que desexamos aplicar este modelo, que é estudo de series temporais. Por este motivo chegados a este punto imos proporcionar unhas cantas directrices relevantes na aplicación deste modelo ás series temporais que desexamos estudar.

A variable resposta considerada no modelo será obviamente a variable medida na serie temporal, que como xa dixemos nas tres series de interese é unha variable de conteo, polo que consecuentemente segue unha distribución Poisson e a función link do modelo GAM é a función logaritmo neperiano. Dito isto, o realmente importante na aplicación dun GAM o estudo destas series é a selección das variables explicativas a considerar. Como expuxemos na Sección 2.2, o modelo considerado debe conseguir captar o comportamento a longo prazo da serie temporal, é dicir, a súa tendencia, xunto co comportamento periódico da serie temporal, é dicir, a súa estacionalidade. Unha variable que se emprega para isto é a variable que neste traballo denominaremos "día de observación". Se y_t con $t = 1, \cdots, T$ é a serie en estudo a variable "día de observación" creada a partires da serie temporal non é máis que a secuencia crecente $1, 2, \cdots, T$. Temos dúas posibilidades á hora de introducir está covariable no GAM, tomar unha función parcial paramétrica para esta covariable ou pola contra unha función parcial non paramétrica. Na revisión da literatura encóntramos numerosos casos de aplicación nos cales se introduce esta variable mediante unha función parcial non paramétrica, pois isto permite captar e controlar a tendencia e a estacionalidade presentes na serie temporal. Xa centrándonos na estacionalidade da serie temporal é frecuente considerar, en caso de existir este tipo de estacionalidade, variables categóricas que indiquen o día da semana, o mes e/ou o ano no que foi observado o valor y_t, pois a introdución destas covariables no modelo permite distinguir entre os distintos días da semana, meses do ano, ou anos proporcionando información que permite modelar estacionalidade semanal, mensual e anual, respectivamente. Estas variables non se introducen mediante función parciais non paramétricas senon mediante funcións paramétricas cuxa forma é a do predictor dun modelo ANOVA, pois deste xeito obtemos a influencia de cada categoría da covariable na variable resposta. Do mesmo xeito tamén se poden introducir outras variables categóricas no modelo se existe algún aspecto da serie temporal que é necesario modelizar mediante este tipo de variables. Ademais destas covariables que creamos a partires da serie temporal e das datas de inicio e remate de medición da variable considerada na serie temporal, tamén se poden introducir como variables explicativas outras series temporais que presenten unha relación coa serie en estudo, aínda que como veremos no capítulo de aplicación no caso que nos ocupa isto non é viable por falta de información.

Algúns artigos onde se atopan modelos GAM que consideran algunha ou varias das variables explicativas mencionadas son Dominici et al. (2002) e Tobías e Saez (2004).

2.4. Outras metodoloxías

Nesta sección realizamos unha breve descrición doutras metodoloxías que empregaremos no modelado e na predición das series temporais.

2.4.1. Árbores de Divisións Recursivas no contexto da regresión

As Árbores de Decisión soen ser moi efectivas e fáciles de interpretar en diversos contextos e por este motivo decidimos explotar a súa utilidade no contexto das series temporais.

Con este fin centrámonos nos modelos de regresión que describen a distribución condicional da variable resposta Y dado un conxunto de p covariables por medio dunha estrutura de Árbore de Divisións Recursivas. Denotamos por $\mathcal{Y}$ o espazo de definición da variable Y e por $\mathbf{X} = (X_1, \cdots, X_p)$ o vector de covariables p-dimensional que toma valores no espazo $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$.

Supoñamos que dispomos dunha mostra $\gamma_n = \{(Y_i, X_{1i}, \cdots, X_{pi}) ; i = 1, \cdots, n\}$, dada esta mostra o algoritmo xenérico para as divisións binarias recursivas pode ser formulado usando pesos non negativos e enteiros $\mathbf{w} = (w_1, \cdots, w_n)$ da seguinte forma: cada nodo da árbore é representado mediante un vector de pesos que non ten elementos nulos cando as correspondentes observacións son elementos do nodo e cero en outro caso. O seguinte algoritmo permite levar a cabo as divisións binarias recursivas:

- **Etapa 1.** Para o vector de pesos $\mathbf{w}$ contrastamos a hipótese global de independencia entre calquera das p covariables e a resposta. Se a hipótese nula non se pode rexeitar debemos parar o algoritmo. Noutro caso seleccionamos a variable X_{j*} que ten unha maior asociación co variable resposta Y.

- **Etapa 2.** Eliximos un conxunto $A^* \subset \mathcal{X}_{j*}$ para dividir deste xeito $\mathcal{X}_{j*}$ en dous conxuntos disxuntos A^* e $\mathcal{X}_{j*} \setminus A^*$. Dando lugar a dous vectores de pesos $\mathbf{w}_{left}$ e $\mathbf{w}_{right}$ onde $w_{left,i} = w_i I(X_{j*i} \in A^*)$ e $w_{right,i} = w_i I(X_{j*i} \notin A^*)$ para todo $i = 1, \cdots, n$. Nótese que $I(\cdot)$ denota a función indicadora.

- **Etapa 3.** Repítense recursivamente as etapas 1 e 2 cos pesos modificados $\mathbf{w}_{left}$ e $\mathbf{w}_{right}$, respectivamente.

Respecto a este algoritmo simplemente cabe expor dúas observacións. Primeiramente dicir que é moi importante ter presente que o algoritmo debe deterse cando a hipótese nula global de independencia entre a variable resposta e calquera das p covariables non se poida rexeitar para un nivel de significación α prefixado, o cal da lugar a un criterio de parada moi intuitivo. En segundo lugar é relevante ter en conta que no contexto no que imos aplicar o algoritmo non desexamos proporcionar unha maior importancia a certas observacións fronte a outras, polo cal os pesos tomarán o valor nulo se a correspondente observación non se encontra no nodo en cuestión, e 1 se a observación esta presente no nodo.

Cabe mencionar que na **Etapa 1** a hipótese nula global de independencia é

$$H_0 = \bigcap_{j=1}^{p} H_0^j,$$

é dicir, a hipótese nula global formúlase en termos de p hipóteses nulas parciais $H_0^j : D(Y|X_j) = D(Y)$, onde $D(Y)$ é a distribución da variable resposta e $D(Y|X_j)$ é a distribución condicional da variable resposta dada a covariable X_j. No artigo Hothorn et al. (2006), Sección 3 pódense consultar tanto os detalles deste contraste como do proceso de selección da covariable con maior asociación coa resposta. Nesta mesma Sección tamén se pode consultar o proceso de elección dos subconxuntos da **Etapa 2** do algoritmo.

Para rematar imos amosar como empregar a árbore construída para predicir os valores da resposta. Unha vez construída a árbore obtemos finalmente r subconxuntos disxuntos $B_1, \cdots, B_r$ que dividen o

espazo de covariables $\mathcal{X} = \bigcup_{k=1}^{r} B_k$, polo cal a hora de predicir é usual estar interesado na estimación da esperanza da resposta $\mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ en cada un dos subconxuntos, cuxa estimación se pode obter mediante a seguinte expresión

$$\widehat{\mathbb{E}}(Y|\mathbf{X} = \mathbf{x}) = \left(\sum_{i=1}^{n} w_i(x)\right)^{-1} \sum_{i=1}^{n} w_i(x) Y_i.$$

Realmente a expresión anterior só indica que para cada subconxunto facemos un promedio das observacións da resposta cuxos valores das variables explicativas se encontran en dito subconxunto. Pódese ver Hothorn et al. (2006), Sección 4.

Consecuentemente, se se desexa predicir un valor Y_i sendo $X_{1i}, \cdots, X_{pi}$ os valores das covariables asociados ao instante que desexo predicir, simplemente se debe ver a que subconxunto $B_1, \cdots, B_r$ pertencen estes valores das covariables e predicir o valor da variable resposta segundo a estimación da esperanza da mesma neste subconxunto.

2.4.2. Métodos de predición simples

Para a compañía Optare Solutions resulta moi relevante que os resultados que proporcionan na súa asesoría á empresa de telecomunicacións sexan mellores, en termos de erro de predición das series temporais, que os resultados que a propia empresa poida obter empregando métodos de predición simples, que poida encontrar implementados en *software* de uso doado e cuxa comprensión e interpretación sexa sinxela. Por este motivo, neste traballo compararemos, no Capítulo 3, os resultados proporcionados pola metodoloxía Box-Jenkins, pola metodoloxía GAM e polas Árbores de Divisións Recursivas cos resultados proporcionados polos métodos de predición simples. Consecuentemente adicamos este apartado a exposición dos métodos simples seleccionados para tal comparativa co obxectivo de levar a cabo a súa aplicación no seguinte capítulo.

O denominado método da Media consiste en predicir calquera valor futuro da serie temporal a través da media dos valores da serie temporal. É dicir, dada a serie temporal $\{y_1, \cdots, y_T\}$ a predición a horizonte k para todo $k \in \mathbb{N}$ é

$$\widehat{y}_T(k) = \overline{y} = \frac{\sum_{i=1}^{T} y_i}{T}.$$

Outro método simple que empregaremos para realizar a comparativa mencionada é o método Naive, o cal tamén se soe denominar método inxenuo. Este método predí calquera valor futuro da serie temporal mediante o último valor da serie temporal. É dicir, dada a serie temporal $\{y_1, \cdots, y_T\}$ a predición a horizonte k para todo $k \in \mathbb{N}$ é

$$\widehat{y}_T(k) = y_T.$$

Unha variante do método Naive é o método Naive estacional, que tal como indica a súa denominación emprega a información do período estacional da serie temporal, predicindo cada valor futuro da serie temporal segundo o último valor observado de igual período.

Por último presentamos o método Drift, este método consiste en predicir os valores da serie temporal como o último valor observado máis unha tasa de cambio. Máis concretamente, dada a serie temporal $\{y_1, \cdots, y_T\}$ a predición a horizonte k para todo $k \in \mathbb{N}$ é

$$\widehat{y}_T(k) = y_T + \frac{k}{T-1} \sum_{t=2}^{T} (y_t - y_{t-1}).$$

Se nos fixamos na expresión previa caemos na conta de que este método simplemente predí os valores futuros seguindo a recta que pasa pola primeira e última observación.

2.5. Medidas de exactitude das predicións

Unha vez revisadas as diferentes metodoloxías que empregaremos na análise, no modelado e na predición das series temporais, é importante expor o procedemento que seguiremos á hora de cuantificar a exactitude das predicións realizadas. Por este motivo nesta sección abordamos a exposición de dito procedemento.

Unha técnica desenrolada co obxectivo de avaliar os resultados dunha análise estatística é o Método de Retención (en inglés Holdout Method). Este método consiste en dividir en dous conxuntos complementarios os datos da mostra, empregar un deles no axuste do modelo (este conxunto denomínase datos de adestramento ou training set) e comprobar o correcto funcionamento do modelo axustado empregando o outro conxunto, coñecido como datos de proba ou test set. Nótese que o modelo axústase empregando só as observacións do training set, e que con este axuste se obteñen as predicións dos valores do test set.

Agora ben esta técnica ten o inconveniente de que os resultados da avaliación poden depender en gran medida dos conxuntos training e test seleccionados. Debido a esta carencia xorde o método de Validación Cruzada (Cross Validation).

O método de Validación Cruzada consiste en repetir a selección dos training e test sets un certo número de veces, e calcular a media das medidas de avaliación correspondentes a cada selección, eliminando así o inconveniente do Método de Retención. Algúns tipos de Validación Cruzada son k-fold, leave-one-out e repeated random sub-sampling validation.

Nesta traballo adoptamos o enfoque do método de Validación Cruzada cunha pequena adaptación ao contexto de series temporais.

Dada unha serie temporal $\{y_1, \cdots, y_T\}$, desexamos comprobar o funcionamento da metodoloxía empregada na predición de valores futuros. Un xeito de comprobar dito funcionamento é comparar as predicións realizadas cos valores reais da serie temporal. Para iso é moi usual dividir a serie temporal en dúas, $\{y_1, \cdots, y_{T_1}\}$ e $\{y_{T_1+1}, \cdots, y_T\}$, axustar un modelo á serie empregando unicamente as observacións da primeira parte $\{y_1, \cdots, y_{T_1}\}$, e logo con este modelo predicir os valores da serie temporal nos instantes $T_1 + 1, \cdots, T$, podendo comparar estas predicións cos valores reais, pois estes encóntranse na segunda parte da serie temporal $\{y_{T_1+1}, \cdots, y_T\}$. Agora ben, isto non é máis que o Método de Retención adaptado ao contexto de series temporais onde o test set se ten que encontrar necesariamente ao final da serie temporal, polo cal ten o inconveniente xa explicado. Por isto decidimos empregar unha adaptación do procedemento de validación cruzada considerando instantes de división da serie temporal, $T_1, T_1 + m, \cdots, T_1 + (m-1)l, T_1 + ml$, onde m é o número de divisións que queremos considerar e l é o número de instantes que desexamos que existan entre cada unha das divisións, dando lugar ás seguintes situacións:[5].

Situación 1: $\text{training}_1 = \{y_1, \cdots, y_{T_1}\}$ e $\text{test}_1 = \{y_{T_1+1}, \cdots, y_{T_1+m_N}\}$.

Situación 2: $\text{training}_2 = \{y_1, \cdots, y_{T_2}\}$ e $\text{test}_2 = \{y_{T_2+1}, \cdots, y_{T_2+m_N}\}$.

$\vdots$ $\vdots$

Situación m-1: $\text{training}_{m-1} = \{y_1, \cdots, y_{T_{m-1}}\}$ e $\text{test}_{m-1} = \{y_{T_{m-1}+1}, \cdots, y_{T_{m-1}+m_N}\}$.

Situación m: $\text{training}_m = \{y_1, \cdots, y_{T_m}\}$ e $\text{test}_m = \{y_{T_m+1}, \cdots, y_{T_m+m_N} = y_T\}$.

En cada unha das situacións axustamos o modelo unicamente coas observacións da primeira parte

[5]Os valores de T_1, m, l e m_N deben ser seleccionados tras realizar certos cálculos para que deste xeito as situacións definidas polos mesmo cumpran certas condicións, como por exemplo, que as mostras training deben ter un número suficiente de datos para o correcto axuste do modelo, e que a mostra test_m debe rematar no último instante da serie temporal.

da serie temporal, e predicimos con este modelo os m_N seguintes valores, predición que comparamos con valores reais contidos na segunda parte da serie temporal.

Para clarificar o procedemento considerado expomos un exemplo. Supoñamos que contamos cunha serie temporal diaria con 800 observacións, e que desexamos que os instantes de divisións sexan 9 e que estean separados 5 instantes, e ademais queremos predicir $m_N = 28$ valores, entón teríamos as seguintes situacións.

$$
\begin{aligned}
&\text{Situación 1:} \quad \text{training}_1 = \{y_1, \cdots, y_{732}\} \text{ e } \text{test}_1 = \{y_{733}, \cdots, y_{760}\}.\\
&\text{Situación 2:} \quad \text{training}_2 = \{y_1, \cdots, y_{737}\} \text{ e } \text{test}_2 = \{y_{738}, \cdots, y_{765}\}.\\
&\text{Situación 3:} \quad \text{training}_3 = \{y_1, \cdots, y_{742}\} \text{ e } \text{test}_3 = \{y_{743}, \cdots, y_{770}\}.\\
&\text{Situación 4:} \quad \text{training}_4 = \{y_1, \cdots, y_{747}\} \text{ e } \text{test}_4 = \{y_{748}, \cdots, y_{775}\}.\\
&\text{Situación 5:} \quad \text{training}_5 = \{y_1, \cdots, y_{752}\} \text{ e } \text{test}_5 = \{y_{753}, \cdots, y_{780}\}.\\
&\text{Situación 6:} \quad \text{training}_6 = \{y_1, \cdots, y_{757}\} \text{ e } \text{test}_6 = \{y_{758}, \cdots, y_{785}\}.\\
&\text{Situación 7:} \quad \text{training}_7 = \{y_1, \cdots, y_{762}\} \text{ e } \text{test}_7 = \{y_{763}, \cdots, y_{790}\}.\\
&\text{Situación 8:} \quad \text{training}_8 = \{y_1, \cdots, y_{767}\} \text{ e } \text{test}_8 = \{y_{768}, \cdots, y_{795}\}.\\
&\text{Situación 9:} \quad \text{training}_9 = \{y_1, \cdots, y_{772}\} \text{ e } \text{test}_9 = \{y_{773}, \cdots, y_{800}\}.
\end{aligned}
$$

En cada unha delas axustaríamos un modelo coas observacións dos datos training, e realizaríamos predicións con este modelo para os seguintes 28 valores, predicións que compararíamos coa realidade (datos test).

Para abreviar referirémonos a cada situación simplemente indicando os instantes temporais cuxos valores recolle a segunda parte da serie, que denominaremos ventá temporal, ou ventá de predición. O resto de información é redundante xa que dados os instantes cuxos valores se recollen na parte test xa sabemos que a primeira parte da serie temporal vai do inicio da serie completa ata o instante anterior ao que da comezo a segunda parte da serie temporal. Polo cal no exemplo anterior teríamos 9 ventás temporais que serían:

$$
\begin{aligned}
&\text{Ventá 1:} \quad t \in \{733, 760\}.\\
&\text{Ventá 2:} \quad t \in \{738, 765\}.\\
&\text{Ventá 3:} \quad t \in \{743, 770\}.\\
&\text{Ventá 4:} \quad t \in \{748, 775\}.\\
&\text{Ventá 5:} \quad t \in \{753, 780\}.\\
&\text{Ventá 6:} \quad t \in \{758, 785\}.\\
&\text{Ventá 7:} \quad t \in \{763, 790\}.\\
&\text{Ventá 8:} \quad t \in \{768, 795\}.\\
&\text{Ventá 9:} \quad t \in \{773, 800\}.
\end{aligned}
$$

Para a comparación dos valores reais coas predicións realizadas en cada unha das ventás empregaremos as seguintes medidas de erro de predición:

- O MSE, “Mean Square Error”.
- O MAPE, “Mean Absolute Percentage Error”.
- O MASE, “Mean Absolute Scaled Error”.

Denotamos por $\widehat{y}_t^i$ a estimación de Y_t usando a mostra training$_i$. Baixo esta notación temos que para cada unha das ventás temporais $i \in \{1, \cdots, m\}$ as medidas de erro de predición calcúlanse como seguen.

$$\text{MSE}_i = \frac{1}{m_N} \sum_{j=1}^{m_N} \left(y_{T_i+j} - \widehat{y}_{T_i+j}^i \right)^2 .$$

$$\text{MAPE}_i = 100 \frac{1}{m_N} \sum_{j=1}^{m_N} \frac{\left| y_{T_i+j} - \widehat{y}_{T_i+j}^i \right|}{|y_{T_i+j}|}.$$

$$\text{MASE}_i = 100 \frac{1}{m_N} \sum_{j=1}^{m_N} \frac{\left| y_{T_i+j} - \widehat{y}_{T_i+j}^i \right|}{q}, \text{ sendo } q = \frac{1}{m_N - 1} \sum_{t=T_i+2}^{T_i+m_N} |y_t - y_{t-1}| .$$

Agora ben, como o obxectivo final é comparar as diferentes metodoloxías é interesante resumir os m valores obtidos para cada unha das medidas. Para iso consideramos a media, a mediana e a desviación típica como medidas representativas dos mesmos.

Para rematar esta sección cabe expor algúns comentarios acerca das medidas de erro consideradas. A medida MSE depende da escala da serie temporal polo cal é complexo realizar unha interpretación da mesma, a pesar disto considerámola interesante pois ao elevar ao cadrado as diferenzas entre as predicións e os valores reais penaliza as diferenzas grandes. En contraposición, as medidas MAPE e MASE son independentes da escala, permitindo así unha interpretación das mesmas. O MAPE interprétase como unha porcentaxe de erro, mentres que o MASE ten unha interpretación máis complexa.

Se obtemos valores do MASE menores que 100 significa que a metodoloxía empregada na predición é máis axeitada que predicir cada valor como o valor inmediatamente anterior, pois o denominador desta medida pódese interpretar como a suma das diferenzas, para os valores da ventá temporal, entre o valor real da serie temporal e o valor predito con este procedemento, que é o valor da serie temporal no instante inmediatamente anterior. A cantidade 100-MASE proporciónanos a melloría que obtemos predicindo coa metodoloxía considerada fronte a predición seguindo a filosofía exposta. Por suposto se obtemos que o MASE é igual ao valor 100 a conclusión é que os métodos de predición son equivalentes, e en caso de tomar un valor maior que 100 teríamos que afirmar que procedemento que predí cada observación co valor da serie temporal no instante anterior mellora a predición realizada coa metodoloxía considerada. Por último mencionar que o MAPE non se pode usar cando temos observacións nulas na serie temporal, mentres que o MASE si se pode usar nese caso pois elimina o efecto da escala mediante a cantidade q.

Capítulo 3

Aplicación a datos reais

No capítulo previo levamos a cabo unha revisión metodolóxica dos modelos Box-Jenkins, dos modelos GAM, e das Árbores de Divisións Recursivas no contexto de regresión. O obxectivo final desta revisión era coñecer os fundamentos de ditas metodoloxías para deste xeito realizar unha correcta aplicación das mesmas na resolución do problema que abordamos neste traballo, exposto na Sección 1.2. Recordemos pois que este problema consiste en seleccionar de entre as tres metodoloxías presentadas a máis axeitada para predicir eventos futuros de cada unha das seguintes series temporais.

- "Número de usuarios diarios do servizo de vídeo baixo demanda".
- "Número de baixas diarias de teléfono móbil".
- "Número de altas diarias de teléfono móbil".

Para cumprir este obxectivo modelamos cada unha destas series temporais empregando a metodoloxía Box-Jenkins, a metodoloxía GAM e as Árbores de Divisións Recursivas. Seguidamente comprobamos o funcionamento de predición dos modelos axustados mediante a aplicación do procedemento exposto na Sección 2.5. Este procedemento proporciónanos dúas medidas de posición do erro de predición e unha da variabilidade do mesmo para cada medida de erro considerada, as cales nos permiten levar a cabo un estudo comparativo das tres metodoloxías podendo obter conclusións a cerca de cal delas é a máis eficaz na predición de valores futuros de cada unha das series temporais de interese, chegando así a seleccionar unha metodoloxía para a predición de cada unha das series temporais.

Ademais disto é moi importante para a empresa confrontar os resultados obtidos coas metodoloxías mencionadas cos resultados obtidos se empregamos para predicir valores futuros da serie temporal os denominados métodos simples achegados na Subsección 2.4.2. Por este motivo modelamos cada serie temporal a través destes métodos simples, comprobamos o seu funcionamento de predición seguindo a metodoloxía da Sección 2.5, e para finalizar comparámolos cos resultados obtidos coas outras metodoloxías.

Na primeira sección do presente capítulo realizaremos o estudo exposto na serie temporal "número de usuarios diarios do servizo de vídeo baixo demanda". Tras o estudo desta serie temporal, procederemos o estudo da serie temporal "número de baixas diarias de teléfono móbil", estudo que se encontra na segunda sección deste capítulo. Case para rematar abordamos o modelado da serie de tempo "número de altas diarias de teléfono móbil".

Por último indicar, como xa mencionamos anteriormente, que para o estudo que imos realizar se emprega como ferramenta informática o *software* estadístico R v.3.2.0. En concreto, empregaremos as seguintes librerías: No caso de aplicación da metodoloxía Box-Jenkins empregamos a librería TSA (Chan e Ripley 2012), para a aplicación da metodoloxía GAM fixemos uso da librería mgcv (Wood 2015), e con referencia a aplicación das Árbores de Divisións Recursivas ás series temporais de interese dicir que neste caso utilizamos librería party (Hothorn et al. 2015). Tamén cabe mencionar que para a aplicación dos métodos simples empregamos a librería forecast (Hyndman 2015).

3.1. Estudo da serie número de usuarios diarios do servizo de vídeo baixo demanda

Esta serie temporal comeza o 03/03/2013 e remata o 19/05/2015, e polo tanto dispoñemos de 808 observacións para o presente estudo. O interese de predición desta serie temporal é poder coñecer por anticipado o número de usuarios que van empregar o servizo cada día e así anticiparse a posibles anomalías no seu uso axustando os medios técnicos as necesidades esixidas.

O primeiro paso no estudo dunha serie temporal é a coñecida coma análise descritiva da serie temporal. Dentro desta análise é moi importante a representación do gráfico secuencial da serie temporal, pois este proporciona gran cantidade de información. Por este motivo amosamos na Figura 3.1 o gráfico secuencial da serie "número de usuarios diarios do servizo de vídeo baixo demanda".

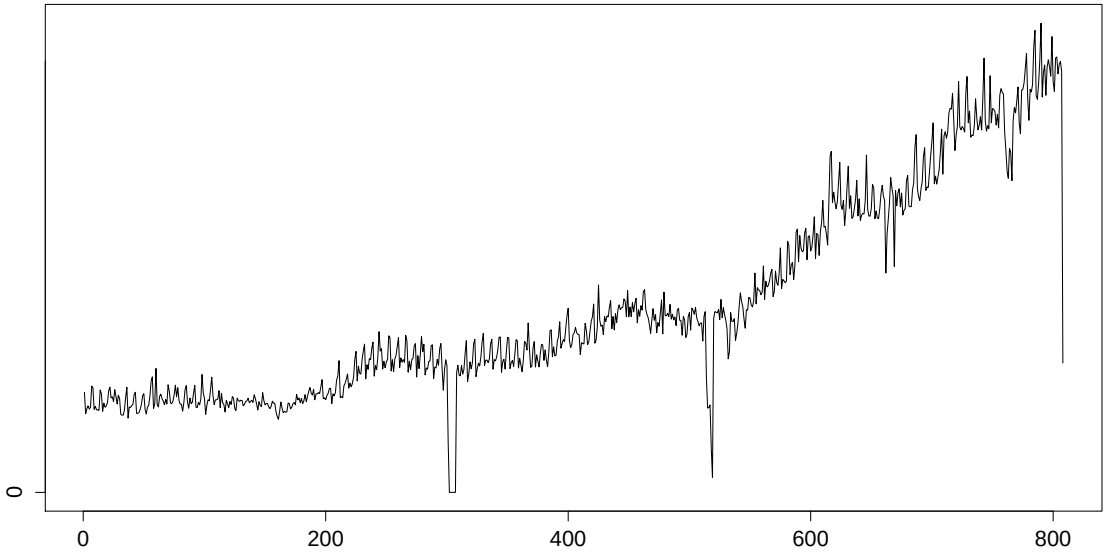


Figura 3.1: Gráfico secuencial da serie temporal "número de usuarios diarios do servizo de vídeo baixo demanda".

Na Figura 3.1 podemos observar que o número de usuarios deste servizo sufriu un gran crecemento durante o período de observación da serie temporal, é dicir, a serie temporal presenta unha tendencia crecente. En canto á variabilidade da serie temporal, observamos que esta presenta maior variabilidade a medida que pasa o tempo, é dicir, ao inicio da mesma a variabilidade é inferior á presente no período final da serie temporal, o que indica pois presenza de heterocedasticidade. Con respecto á compoñente estacional é difícil extraer conclusións mediante a representación gráfica, pero poderíamos pensar pola clase de variable medida nunha posible estacionalidade semanal, pois a demanda deste servizo vai ligada ao tempo de ocio dos clientes, o cal varía ao longo da semana sendo maior no fin de semana. No gráfico secuencial detectamos visualmente valores cuxo comportamento discrepa do comportamento xeral, é dicir, vemos valores atípicos na serie temporal, algúns deles correspóndense con caídas do sistema, é dicir, con días nos que o servizo deixou de estar operativo, mentres que o atípico presente no último valor da serie temporal débese a un fallo na recollida dos datos.

Previamente ao estudo da serie temporal empregando cada un das metodoloxías presentadas, é nece-

sario tratar o inconveniente dos valores atípicos mencionados suavizando o seu comportamento atípico e a posible influencia que estes poidan ter nas modelizacións que imos realizar. Para isto empregamos a función tsclean da librería forecast que nos permite identificar e substituír os valores atípicos e perdidos da serie temporal (Hyndman 2015). Na nosa serie non temos valores perdidos soamente valores atípicos cuxo comportamento se suaviza coa aplicación desta función obtendo a serie temporal que se amosa na Figura 3.2.

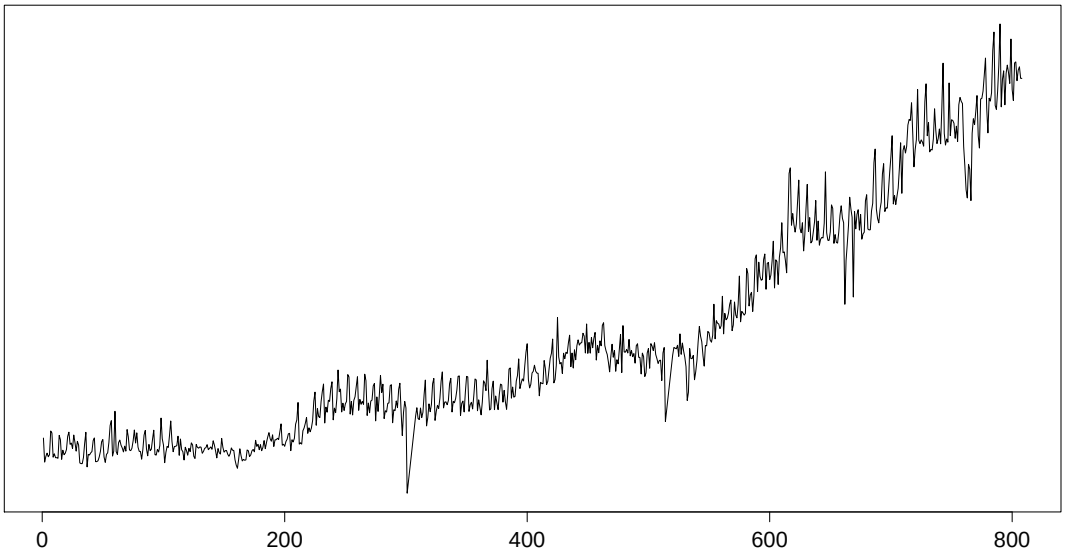


Figura 3.2: Gráfico secuencial da serie temporal "número de usuarios diarios do servizo de vídeo baixo demanda" tras suavizar os valores atípicos.

Outro punto moi importante antes de comezar coa aplicación das metodoloxías tratadas é expor as ventás temporais seleccionadas para levar a cabo o procedemento exposto na Sección 2.5. Consideramos $m = 10$ ventás, cada unha delas con $m_N = 28$ observacións, e tomamos $T_1 = 736$ e $l = 5$, dando así lugar ás seguintes ventás temporais:

- Ventá 1: Dende 08/03/2015 ata 04/04/2015.
- Ventá 2: Dende 13/03/2015 ata 09/04/2015.
- Ventá 3: Dende 18/03/2015 ata 14/04/2015.
- Ventá 4: Dende 23/03/2015 ata 19/04/2015.
- Ventá 5: Dende 28/03/2015 ata 24/04/2015.
- Ventá 6: Dende 02/04/2015 ata 29/04/2015.
- Ventá 7: Dende 07/04/2015 ata 04/05/2015.
- Ventá 8: Dende 12/04/2015 ata 09/05/2015.
- Ventá 9: Dende 17/04/2015 ata 14/05/2015.

- Ventá 10: Dende 22/04/2015 ata 19/05/2015.

Realmente non empregamos todas as ventás temporais na realización do procedemento exposto na Sección 2.5, pois previamente á realización deste proceso parece interesante empregar a última ventá temporal para amosar con detalle o modelo axustado, interpretar o modelo, e comparar as predicións cos valores observados mediante medidas cuantitativas, pero tamén mediante procedementos gráficos. Logo deste estudo considerando a última ventá temporal, o cal permite dispor dunha idea clara do problema tratado, procedemos a realización do procedemento exposto na Sección 2.5 empregando as restantes 9 ventás temporais.

3.1.1. Metodoloxía Box-Jenkins

Como vimos de expor inicialmente consideramos a ventá temporal número 10 e ilustramos o procedemento de análise, modelado e predición nesta ventá. Como xa indicamos empregamos as observacións do período 03/03/2013-21/04/2015 para axustar un modelo Box-Jenkins á serie temporal, mentres que as restantes observacións, período 22/04/2015-19/05/2015, úsanse na tarefa de comprobación das predicións realizadas.

O gráfico secuencial da serie temporal completa (Figura 3.2) parece indicar a presenza de heterocedasticidade, polo que é necesario estabilizar a variabilidade da serie temporal previamente a identificación dun modelo Box-Jenkins. Posto que a variabilidade da serie temporal parece crecer a medida que aumenta o nivel a transformación Box-Cox é a indicada para resolver este problema. Se calculamos o valor do parámetro λ asociado a esta transformación obtemos que este é 0.221. Como este valor é próximo a cero consideramos a posibilidade de asumir que λ é cero, e consecuente realizar unha transformación logarítmica que aporta maior simplicidade ao problema. Se realizamos está transformación a variabilidade estabilizase, polo cal comezamos o proceso de identificación do modelo Box-Jenkins considerando o logarítmo da serie temporal. Na Figura 3.3 podemos ver o gráfico de correlacións simples mostrais desta serie temporal. En dito gráfico observamos que a *fas* mostral toma valores próximos a 1 nos primeiros retardos, e que decae lentamente a cero a medida que o retardo crece o cal indica presenza de tendencia na serie temporal, tal é como explicamos na Subsección 2.2.2.

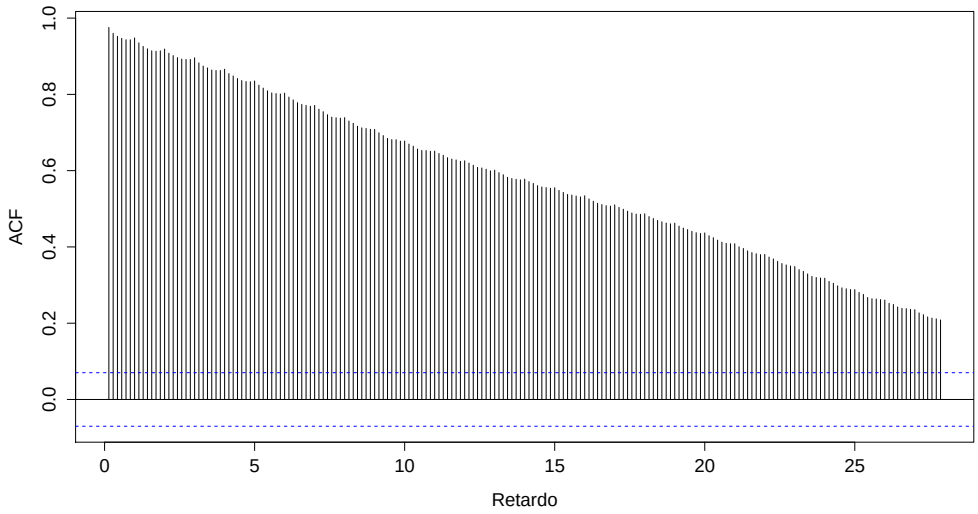


Figura 3.3: Gráfico das correlacións simples mostrais da serie temporal tras a aplicación da transformación logarítmica.

Co obxectivo de eliminar a tendencia aplicamos unha diferenza regular sobre a serie temporal, obtendo como resultado a serie temporal que presentamos na Figura 3.4.

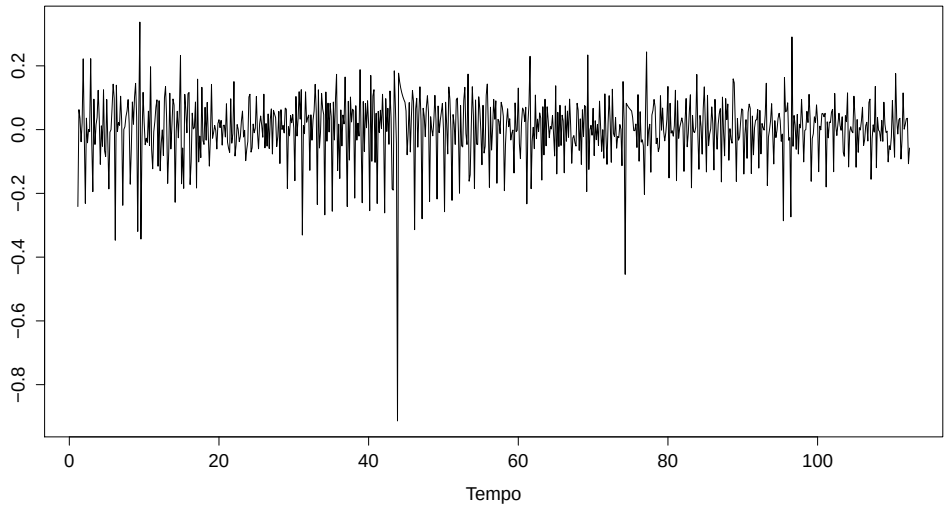


Figura 3.4: Gráfico secuencial da serie temporal diferenciada regularmente.

Na Figura 3.4 observamos que a aplicación dunha diferenza regular consegue eliminar a tendencia, tamén observamos certos valores que poderían ser valores atípicos, sobre os cales traballaremos posteriormente seguindo a metodoloxía exposta na Subsección 2.2.6. Se debuxamos as correlacións simples mostrais da serie temporal diferenciada obtemos o gráfico da Figura 3.5.

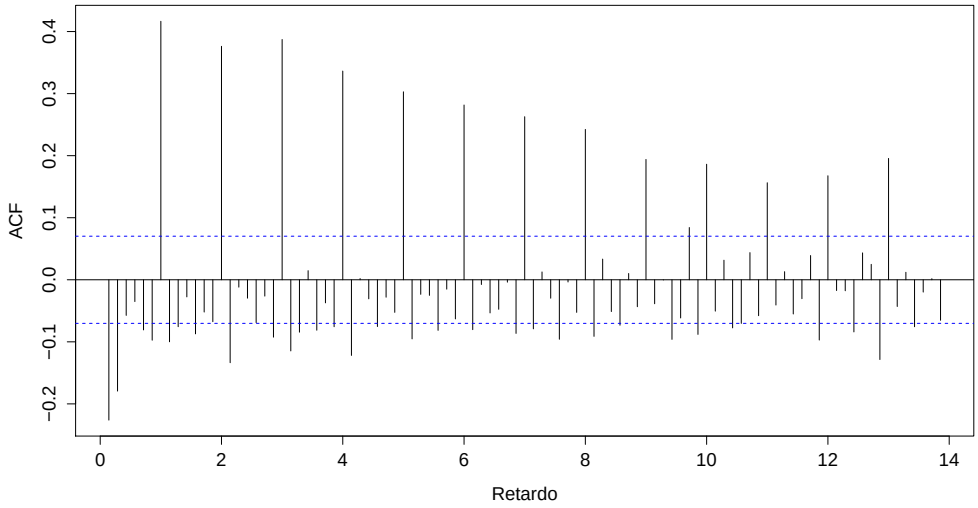
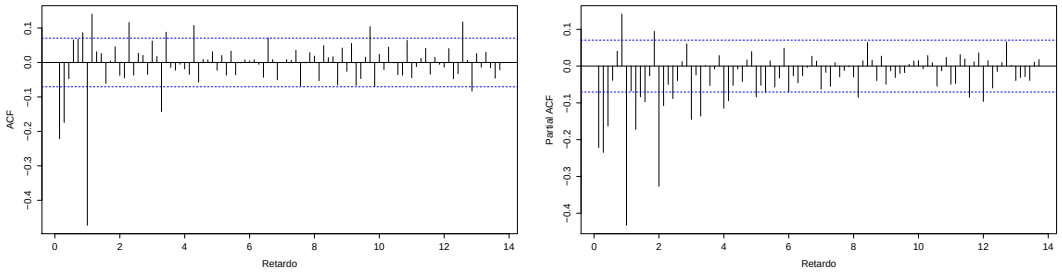


Figura 3.5: Gráfico de correlacións simples mostrais da serie temporal diferenciada regularmente.

Na Figura 3.5 observamos forte correlación no retardo 7 e nos seus múltiplos, ademais de periodicidade de período 7 e converxencia lenta a cero, estas características indican presenza dunha compoñente estacional de período 7 días, é dicir, compoñente estacional semanal. Para poder eliminar esta compoñente estacional e obter unha serie estacionaria na cal poidamos identificar os parámetros dun modelo Box-Jenkins aplicamos unha diferenza estacional obtendo unha serie estacionaria coa *fas* e *fap* mostral presentes na Figura 3.6.



(a) *Fas* mostral (b) *Fap* mostral

Figura 3.6: Gráfico de correlacións simples e parciais mostrais da serie temporal diferenciada regular e estacionalmente.

Mediante a interpretación da *fas* e *fap* mostrais presentes na Figura 3.6 podemos identificar os parámetros dun ARMA estacional multiplicativo pois como se pode observar a *fas* e *fap* mostrais da serie diferenciada regular e estacionalmente presentan dependencia regular e estacional. Se observamos a Figura 3.6 non está claro que parámetros son os axeitados para o modelo ARMA estacional multiplicativo, polo cal decidimos empregar o criterio BIC a través da función best.arima.TSA cuxo código se pode ver no Apéndice B. Está función considera un conxunto de valores para cada un dos parámetros do modelo, axusta todos os modelos que xorden de combinar os posibles valores de cada un dos parámetros, e calcula o BIC asociado a cada modelo para finalmente indicarnos que valores dos parámetros dan lugar ao menor BIC. Observando as correlacións simples e parcias consideramos adecuado indicarlle á función best.arima.TSA que os parámetros poden tomar os seguintes valores $p, q, P, Q \in \{0, 1, 2\}$. Tamén debemos indicarlle o número de diferenzas regulares e estacionais realizadas previamente, que neste caso é unha. Feito isto a función indícanos que debemos axustar un ARIMA$(1,1,1)\times(1,1,2)_7$ sobre a serie orixinal pois o BIC asociado a este modelo é -1737.16 inferior ao BIC asociado a calquera dos outros modelos axustados. A expresión do modelo [1] considerado é

$$(1-\phi_1 B)\left(1-\Phi_1 B^7\right)(1-B)\left(1-B^7\right)Y_t = (1+\theta_1 B)\left(1+\Theta_1 B^7+\Theta_2 B^{14}\right)a_t. \tag{3.1}$$

Os resultados do axuste do modelo (3.1) son os que amosamos no Cadro 3.1. Nel achegamos as estimacións dos coeficientes do modelo obtidas polo método de máxima verosimilitude xunto coa desviación típica asociada a cada estimación, o cal nos permite comprobar mediante un contraste de significación [2] se cada un dos parámetros do modelo é significativamente distinto de cero. Efectivamente se realizamos os contrastes pertinentes obtemos que todos os parámetros son significativamente distintos de cero para un nivel de significación $\alpha = 0.05$.

Tras a estimación do modelo debemos proceder á diagnose do mesmo. Comecemos pola comprobación da hipótese de incorrelación e homocedasticidade. Para a comprobación destas hipóteses

[1] Consideramos o modelo sen a constante c pois incluír constante cando $d+D > 1$ equivale a introducir unha tendencia polinómica de grado $d+D$, que no noso caso é 2, e isto a efectos de predición pode ter unha repercusión negativa.

[2] Se T é suficientemente grande $\widehat{\phi} \approx N(\phi, \sigma)$, polo cal se consideramos o contraste $H_0 := \phi = 0$ fronte a $H_1 := \phi \neq 0$ rexeitaremos H_0 se $|\widehat{\phi}/\widehat{\sigma}| \geq 1.96$.

consideramos os residuos estandarizados en lugar dos residuos orixinais, polas vantaxes que estes presentan. No primeiro gráfico da Figura 3.7 podemos ver os residuos do modelo fronte ao tempo, os cales presentan unha variabilidade estable indicando ausencia de heterocedasticidade. Agora ben vemos que en dito gráfico están presentes dous límites de control [3] co obxectivo de detectar valores atípicos nos residuos do modelo, como se pode observar existen valores dos residuos fóra destes límites de control, por este motivo debemos identificar a súa posición e tipo para a súa modelización.

	$\widehat{\phi}_1$	$\widehat{\theta}_1$	$\widehat{\Phi}_1$	$\widehat{\Theta}_1$	$\widehat{\Theta}_2$
Coeficientes	0.5231	-0.9164	0.7850	-1.5875	0.5876
s.d	0.0445	0.0237	0.0563	0.0727	0.0712

Cadro 3.1: Resultados do axuste do modelo (3.1).

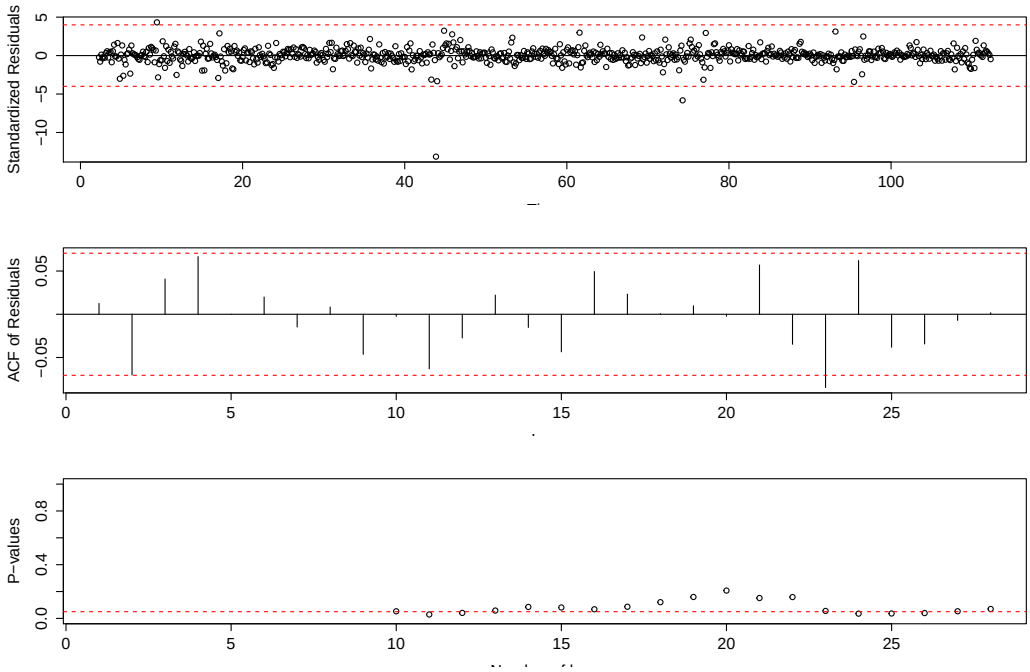


Figura 3.7: Gráficos para a comprobación das hipóteses de incorrelación e homocedasticidade dos residuos para o modelo (3.1). O primeiro gráfico é unha representación dos residuos estandarizados do modelo fronte ao tempo. O segundo gráfico é a función de autocorrelación simple mostral dos residuos. O terceiro gráfico correspóndese co contraste de Ljung-Box, nel debúxanse os p-valores do contraste de Ljung-Box segundo os retardos considerados.

Seguindo o exposto na Subsección 2.2.6 identificamos a posición e o tipo dos atípicos presentes na modelización realizada, obtendo que estes valores atípicos se encontran nas posicións 28,60,113,297,301,

[3] Os límites de control nos residuos estandarizados sitúanse en menos e máis o percentil $0.025/T$ dunha normal estándar.

302,303,304,425,514,646,662 e son todos eles atípicos innovativos. Unha vez identificada a posición e o tipo de atípico debemos abordar a estimación da súa magnitude. Agora ben se introducimos a modelización destes atípicos no modelo (3.1) algúns coeficientes deste modelo que antes eran significativamente distintos de cero pasan agora a ser nulos desaparecendo do mesmo é quedando un modelo ARIMA(1,1,1)$\times$(0,1,1)$_7$ cos atípicos innovativos mencionados, é dicir, temos o seguinte modelo:

$$\begin{aligned} Y_t &= w_{28}\psi(B)I_t^{(28)} + w_{60}\psi(B)I_t^{(60)} + w_{113}\psi(B)I_t^{(113)} + w_{297}\psi(B)I_t^{(297)} + w_{301}\psi(B)I_t^{(301)} \\ &\quad + w_{302}\psi(B)I_t^{(302)} + w_{303}\psi(B)I_t^{(303)} + w_{304}\psi(B)I_t^{(304)} + w_{425}\psi(B)I_t^{(425)} + w_{514}\psi(B)I_t^{(514)} \\ &\quad + w_{646}\psi(B)I_t^{(646)} + w_{662}\psi(B)I_t^{(662)} + \psi(B)a_t, \end{aligned} \tag{3.2}$$

sendo $\psi(B) = (1-B^7)^{-1}(1-B)^{-1}(1-\phi_1 B)^{-1}(1+\theta_1 B)(1+\Theta_1 B^7)$.

Se estimamos os coeficientes presentes no modelo (3.2) obtemos os resultados expostos no Cadro (3.2).

	$\widehat{\phi}_1$	$\widehat{\theta}_1$	$\widehat{\Theta}_1$	w_{301}	w_{514}	w_{302}	w_{303}	w_{60}
Coeficientes	0.4525	-0.8821	-0.7208	-0.9770	-0.3269	-0.8085	-0.4050	0.3439
s.d	0.0485	0.0271	0.0296	0.0534	0.0479	0.0593	0.0596	0.0477
	w_{304}	w_{297}	w_{662}	w_{28}	w_{425}	w_{646}	w_{113}	
Coeficientes	-0.2608	-0.2140	-0.1773	-0.1477	0.2007	0.2173	-0.2120	
s.d	0.0545	0.0483	0.0480	0.0482	0.0478	0.0477	0.0478	

Cadro 3.2: Resultados do axuste do modelo (3.2).

Unha vez que dispomos da estimación do modelo (3.2) no cal emendamos o inconveniente dos valores atípicos mediante a súa modelización retomamos a etapa de diagnose.

Comecemos coa comprobación da hipótese de incorrelación e homocedasticidade. No primeiro gráfico da Figura 3.8 podemos ver os residuos estandarizados do modelo (3.2) fronte ao tempo, os cales presentan unha variabilidade estable indicando ausencia de heterocedasticidade. Ademais disto neste gráfico observamos ausencia dun patrón nos residuos do modelo. No segundo gráfico temos a *fas* mostral dos residuos do modelo. Neste gráfico observamos que a autocorrelación correspondente ao retardo 12 e tamén a correspondente ao retardo 26 exceden os límites de confianza, agora ben consideramos un contraste de incorrelación con $\alpha = 0.05$ polo cal que unha de cada 20 autocorrelacións exceda o límite de confianza non é un motivo para rexeitar a hipótese de incorrelación polo que aceptamos esta hipótese. No terceiro gráfico temos os p-valores do contraste de Ljung-Box, onde observamos que tamén debemos aceptar a incorrelación dos residuos pois todos os p-valores son maiores ou iguais o nivel de significación $\alpha = 0.05$.

Ademais das hipóteses de homocedasticidade e incorrelación, unha hipótese moi importante e cuxo incumprimento pode invalidar o modelo axustado é a hipótese de que os residuos do modelo teñen esperanza nula. Polo cal contrastamos a hipótese nula de que os residuos teñen media cero mediante o contraste exposto na Subsección 2.2.4, obtendo 0.0269 como valor do estatístico de contraste e 0.9786 como p-valor asociado ao contraste, polo que aceptamos a hipótese nula de que os residuos do modelo (3.2) teñen media nula.

Para rematar a diagnose do modelo comprobaremos a hipótese de normalidade dos residuos, cuxo incumprimento non inválida o modelo pero si é unha hipótese axeitada de cara a certos procesos de

inferencia. Se contrastamos esta hipótese mediante o contraste de Shapiro-Wilk obtemos como valor do estatístico de contraste 0.98491 e como p-valor 3.368×10^{-7} o cal nos leva a rexeitar claramente a hipótese de normalidade. A mesma conclusión proporciona o contraste de Jarque-Bera pois o valor do estatístico de contraste é 52.787 cun p-valor asociado igual a 3.44×10^{-12}.

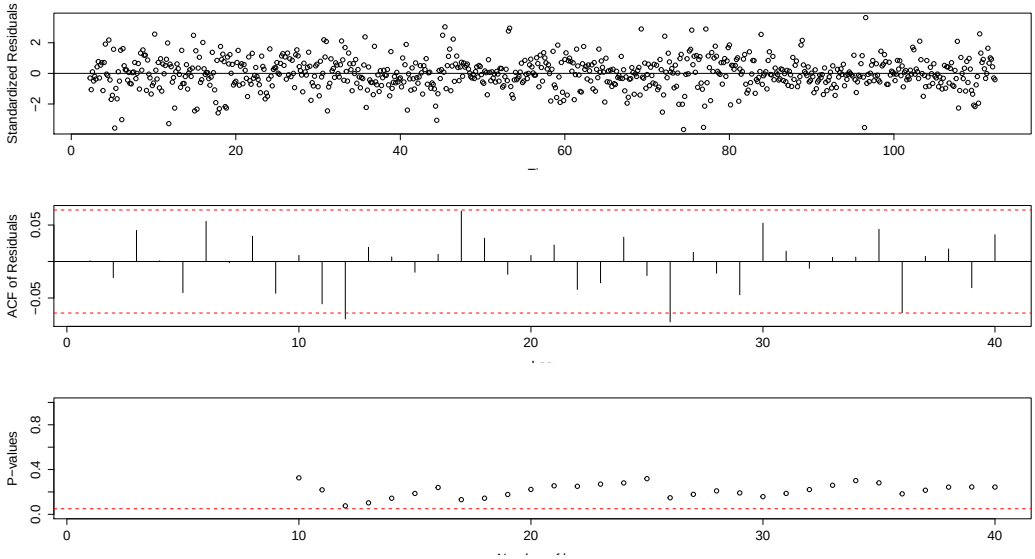


Figura 3.8: Gráficos para a comprobación das hipóteses de incorrelación e homocedasticidade dos residuos para o modelo (3.2). O primeiro gráfico é unha representación dos residuos estandarizados do modelo fronte ao tempo. O segundo gráfico é a función de autocorrelación simple mostral dos residuos. O terceiro gráfico correspóndese co contraste de Ljung-Box, nel debúxanse os p-valores do contraste de Ljung-Box segundo os retardos considerados.

Un vez comprobamos que o modelo é correcto proseguimos abordando a etapa predición en base ao modelo axustado. Como expuxemos con anterioridade empregamos o modelo axustado, recordemos que para o axuste do modelo empregamos unicamente as observacións do período 03/03/2013-21/04/2015, para predicir os valores da ventá temporal 22/04/2015-19/05/2015. Na Figura 3.9 amosamos as predicións para esta ventá temporal xunto cos valores observados. Nesta imaxe observamos que o modelo capta correctamente a tendencia da serie temporal, pero non consegue captar o comportamento semanal real da serie temporal, agora ben isto débese a alta variabilidade e o ruído que a serie presenta no comportamento semanal. Ademais da comparativa gráfica decidimos comparar as predicións cos valores observados seguindo as medidas MSE e MAPE. Os valores obtidos son MSE=980971.9 e MAPE=3.7307 %. O valor de MSE é interesante para realizar comparativas con outros modelos pero non é sinxelo de interpretar. Con respecto ao MAPE observamos un valor de erro moi aceptable a pesares de que graficamente non observemos o comportamento desexado. Agora ben non podemos extraer conclusións calculando medidas de erro nunha única ventá de predición pois pode ser que neste ventá por algún motivo o modelo teña un erro de predición axeitado e noutras ventás este comportamento non se manteña. Polo tanto repetimos o proceso exposto para esta ventá de predición nas 9 ventás mencionadas axustando o modelo unicamente coas observacións previas, e predicindo os valores da ventá considerada. Para cada unha das 9 ventás achamos o valor do MSE e do MAPE, e posteriormente calculamos medidas de erro globais obtendo os resultados que figuran no Cadro 3.3. Os valores das medidas de erro globais son axeitadas polo cal proseguiremos aplicando outras metodoloxías a serie temporal e posteriormente comparemos os resultados obtidos con estes valores.

Media		Mediana		Desviación típica	
MAPE	MSE	MAPE	MSE	MAPE	MSE
8.5221	3844313	7.4363	2569940	2.9247	2758470

Ventá temporal	MAPE	MSE
Ventá 1	6.1669	2170553
Ventá 2	9.1357	4116805
Ventá 3	8.3523	3432046
Ventá 4	6.7897	2187715
Ventá 5	6.7796	2175245
Ventá 6	7.4363	2569940
Ventá 7	15.5854	10744533
Ventá 8	9.8387	4813841
Ventá 9	6.6145	2388144

Cadro 3.3: Resultados das medidas de erro de predición para o modelo (3.2).

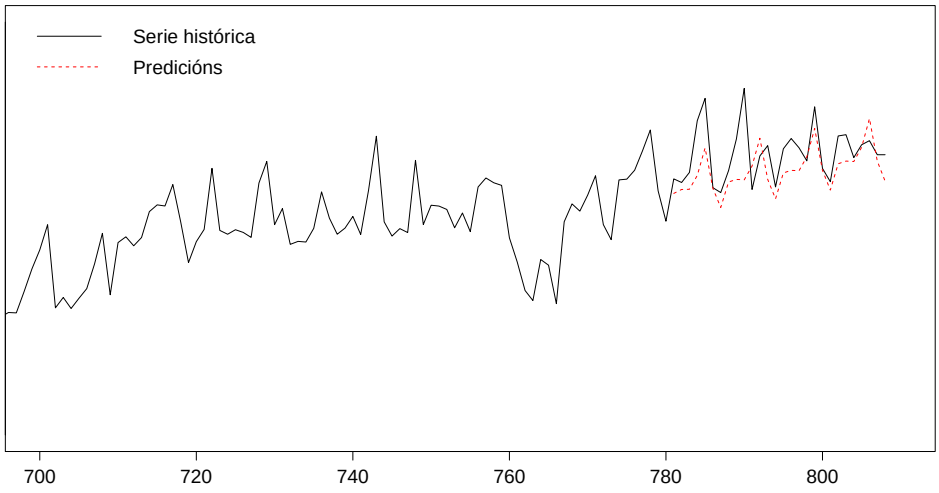


Figura 3.9: Gráfico secuencial dos valores preditos polo modelo (3.2) no período 22/04/2015-19/05/2015 xunto aos valores reais da serie temporal.

3.1.2. Metodoloxía GAM

Analogamente ao realizado na aplicación da metodoloxía Box-Jenkins consideramos primeiramente as observacións que abarcan o período 03/03/2013-21/04/2015, e axustamos un modelo en base a estas

observacións para posteriormente realizar predicións para a ventá temporal 22/04/2015-19/05/2015. Para modelar a serie temporal mediante un modelo GAM debemos ter presente o exposto na Subsección 2.3.4, onde expresamos a importancia da creación das variables explicativas axeitadas para este modelo. Unha destas variables é a que denominamos "día de observación" que neste caso é a secuencia $1, 2, \cdots, 780$. Esta variable introdúcese no modelo mediante unha función parcial suave non paramétrica co obxectivo de captar tendencia e estacionalidade. Ademais desta variable é importante introducir unha variable categórica para distinguir os días da semana, e modelizar así a compoñente estacional semanal. Denotamos por Y a variable resposta, que é o "número de usuarios do servizo de vídeo baixo demanda", por t a variable "día de observación", e por *diasem* a variable categórica referente ao día da semana. Con esta notación o modelo GAM que axustamos a serie temporal formúlase como segue

$$\ln(\mu(\mathbf{X})) = \beta_0 + f_1(t) + \beta_s diasem, \tag{3.3}$$

onde a variable explicativa *diasem* é un factor con posibles niveis $A_0 = "Luns", A_1 = "Martes", \cdots, A_6 = "Domingo"$, polo cal cando escribimos $\beta_s diasem$ estamos facendo un abuso de notación, pois en realidade β_s é un vector $(\beta_{martes}, \beta_{mércores}, \cdots, \beta_{domingo})$, e *diasem* é entendida como a recodificación *dummy* dada por $diasem^* = \{0 \ se \ diasem = A_0, \ 1 \ se \ diasem = A_1, \cdots, \ 6 \ se \ diasem = A_6\}$. Deste xeito cada coeficiente β_j representa o incremento medio da resposta para o nivel A_j en relación ao nivel de referencia A_0. A partires deste momento empregaremos este abuso de notación sempre que introduzamos no modelo unha variable explicativa tipo factor.

Se levamos a cabo o axuste do modelo (3.3) obtemos os resultados presentes no Cadro 3.4. Estes resultados indican que as diferenzas entre as categorías da variable *diasem* son significativas. Ademais, se nos fíxamos nos coeficientes asociados a cada categoría desta variable, podemos ver que o mércores é o día da semana con menor demanda do servizo, mentres que os sábados e domingos a demanda do servizo sofre un aumento con respecto a categoría de referencia que é o luns. Que os días do fin de semana a variable medida sufra un aumento é lóxico pois o uso deste servizo ten unha relación clara co tempo de lecer dos usuarios.

Coeficientes paramétricos

	Estimación	**Sd erro**	**z valor**	**Pr(>|z|)**
β_0	8.8809	0.0011	8306.475	$< 2\times10^{-16}$ ***
β_{martes}	0.0156	0.0015	10.516	$< 2\times10^{-16}$ ***
$\beta_{mércores}$	-0.0040	0.0015	-2.630	0.0085 **
β_{xoves}	0.0073	0.0015	4.864	1.15×10^{-6}***
β_{venres}	0.0210	0.0015	14.139	$< 2\times10^{-16}$ ***
$\beta_{sábado}$	0.0662	0.0015	45.147	$< 2\times10^{-16}$ ***
$\beta_{domingo}$	0.1099	0.0014	75.889	$< 2\times10^{-16}$ ***

Significación dos termos suaves

	edf	**Ref.df**	**Chi.sq**	**p-valor**
$f_1(t)$	8.996	9	1490163	$< 2\times10^{-16}$ ***

Cadro 3.4: Resultados do axuste do modelo (3.3).

Con respecto á función suave non paramétrica observamos que a súa consideración no modelo é significativa. Na Figura 3.10 podemos observar esta función, é dicir, o efecto suave da variable t sobre a resposta. Nótese que esta función está centrada no cero e ademais non se encontra na escala orixinal da serie temporal se non en escala logarítmica. Por outra parte cabe mencionar que na representación da función suave da variable t vese reflexada a súa utilidade á hora de captar a tendencia e a estacionalidade da serie temporal.

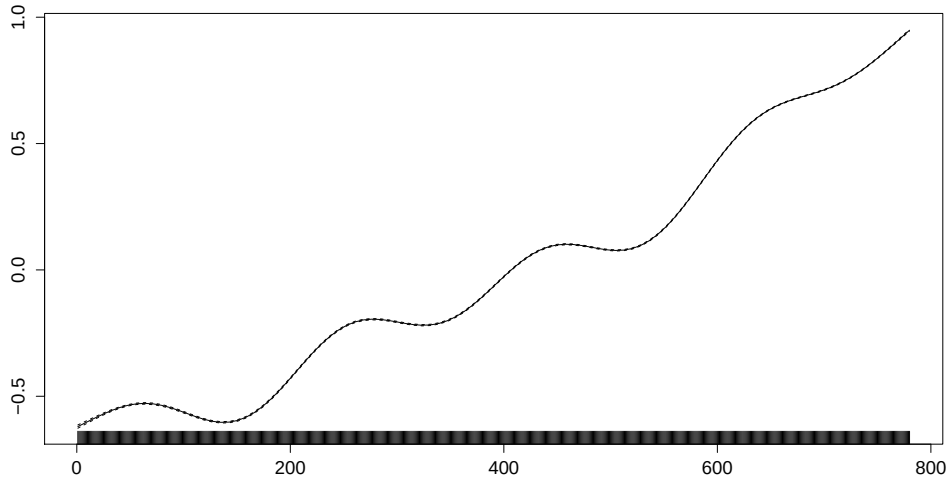


Figura 3.10: Efecto suave da variable t sobre a variable resposta.

Agora que coñecemos os resultados proporcionados polo axuste do modelo proseguimos calculando as predicións realizadas por este modelo na ventá temporal 22/04/2015-19/05/2015. Na Figura 3.11 podemos ver estas predicións xunto cos valores reais da serie temporal. Ademais desta comparativa gráfica cabe mencionar que os valores de MASE e MSE son 5.1574 % e 1668662, respectivamente.

Como se pode observar na Figura 3.11 o modelo (3.3) parece captar unha tendencia máis forte da que realmente presenta a serie temporal en estudo. Como o comportamento do modelo nunha única ventá non é suficiente para xulgar o comportamento xeral do mesmo, aplicamos o procedemento exposto na Sección 2.5 empregando as 9 ventás temporais definidas para este fin obtendo os resultados que se amosan no Cadro 3.5.

Os valores das medidas globais de erro para o modelo (3.3) son superiores aos proporcionados polo modelo Box-Jenkins axustado na sección previa, polo cal imos proseguir este estudo tentando mellorar o modelo exposto. Dada a natureza da variable medida tamén podemos pensar en introducir unha variable categórica que nos permita distinguir entre os diferentes meses do ano, pois nos meses de inverno e outono sería lóxico que a demanda deste servizo sexa máis alta que nos meses de verán e primavera. Denotamos por *mes* a variable categórica mencionada, e axustamos o modelo

$$\ln(\mu(\mathbf{X})) = \beta_0 + f_1(t) + \beta_s diasem + \beta_m mes, \tag{3.4}$$

onde $\beta_s diasem$ ten a mesma estrutura que no modelo (3.3), e $\beta_m mes$ defínese de xeito análogo ao termo $\beta_s diasem$ tomando como categoría de referencia o mes de xaneiro.

Se levamos a cabo o axuste do modelo (3.4) obtemos os resultados presentes no Cadro 3.6. Ao introducir a variable *mes* temos que o coeficiente β_0 agora fai referencia aos luns do mes de xaneiro. Ten do isto presente temos que os resultados do Cadro 3.6 mostran que as diferenzas entre os mércores e os luns de xaneiro non son significativas, o mesmo sucede co mes de outubro, as restantes categorías das

variables *diasem* e *mes* son significativas. Xa anteriormente no modelo (3.3) demos unha interpretación aos coeficientes da semana, resumidamente que os días de menos uso son os luns e mércores e os de maior uso os do fin de semana. Procedamos agora a expor algunhas conclusións acerca dos coeficientes das categorías da variable *mes*. Os resultados indican que o mes de maior uso do servizo é o mes de febreiro, seguido por marzo, novembro e xaneiro, con respecto o mes de menos uso dicir que este é o mes de agosto, seguido por xullo.

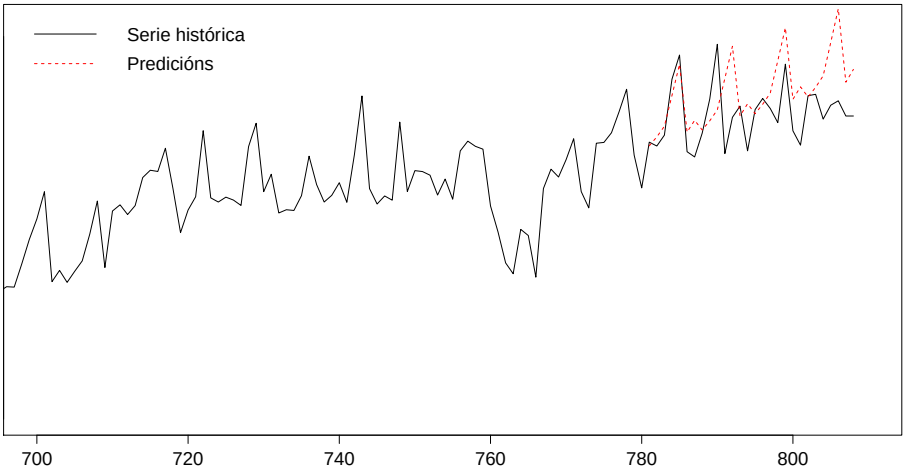


Figura 3.11: Predicións realizadas polo modelo (3.3) na ventá temporal 22/04/2015-19/05/2015 xunto aos valores reais da serie temporal.

Agora que coñecemos os resultados proporcionados polo axuste do modelo proseguimos calculando as predicións realizadas por este modelo na ventá temporal 22/04/2015-19/05/2015. Na Figura 3.12 podemos ver estas predicións xunto cos valores reais da serie temporal. Ademais desta comparativa gráfica cabe mencionar que os valores de MASE e MSE son 6.6337 % e 2548624, respectivamente.

Como se pode observar na Figura 3.12, o modelo (3.4) parece ter o mesmo problema que o modelo (3.3). A pesar disto, como o comportamento do modelo nunha única ventá non é determinante aplicamos o procedemento exposto na Sección 2.5 empregando as 9 ventás temporais definidas para este fin obtendo os resultados que se amosan no Cadro 3.7.

Segundo os resultados do Cadro 3.7 o modelo (3.4) é mellor en termos de erro de predición que o modelo (3.3), pesa a non observar graficamente unha melloría nas predicións da ventá 22/04/2015-19/05/2015, Figura 3.12.

En vez de distinguir os diferentes meses do ano, podemos pensar en introducir unha variable que permita distinguir entre as diferentes estacións do ano, pois como explicamos antes está variable pode estar correlada coa resposta. Denotamos por *estación* a variable mencionada, e axustamos o seguinte modelo

$$\ln(\mu(\mathbf{X})) = \beta_0 + f_1(t) + \beta_s diasem + \beta_e estación, \tag{3.5}$$

onde $\beta_s diasem$ ten a mesma estrutura que no modelo (3.3), e $\beta_e estación$ se define de xeito análogo ao termo $\beta_s diasem$ tomando como categoría de referencia a estación de inverno.

Exposto o modelo (3.5) axustamos este obtendo os resultados presentes no seguinte Cadro 3.8, os cales de contado comentamos e interpretamos. Neste cadro podemos ver que todas as categorías das variables *diasem* e *estación* son significativas, ademais observamos que en primavera e verán o uso do

servizo sofre unha diminución con respecto a categoría de referencia que é a estación de inverno, mentres que no outono o servizo experimenta un aumento con respecto a categoría de referencia mencionada. Estas conclusións extraidas do modelo son concordantes coa idea inicial de que nos meses máis fríos se emprega máis este servizo pois a climatoloxía impide a realización de actividades o aire libre, mentres que nos meses de mellores condicións climatolóxicas se usa menos o servizo.

Media		Mediana		Desviación típica	
MAPE	MSE	MAPE	MSE	MAPE	MSE
10.2268	5037653	11.6011	6042195	4.5667	3015124

Ventá temporal	MAPE	MSE
Ventá 1	7.3389	3002675
Ventá 2	11.6011	6042195
Ventá 3	14.3279	7939382
Ventá 4	13.8427	7535446
Ventá 5	14.7334	7978163
Ventá 6	15.0320	7969264
Ventá 7	6.4642	2307252
Ventá 8	4.5729	1372720
Ventá 9	4.1285	1191775

Cadro 3.5: Resultados das medidas de erro de predición para o modelo (3.3).

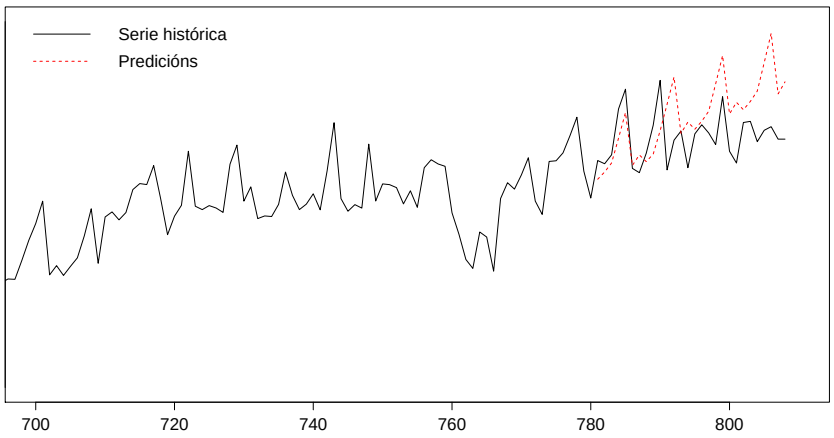


Figura 3.12: Predicións realizadas polo modelo (3.4) na ventá temporal 22/04/2015-19/05/2015 xunto aos valores reais da serie temporal.

Coeficientes paramétricos

	Estimación	**Sd erro**	**z valor**	**Pr(>|z|)**
β_0	8.9126	0.0040	2248.929	$< 2\times10^{-16}$ ***
β_{martes}	0.0162	0.0015	10.972	$< 2\times10^{-16}$ ***
$\beta_{mércores}$	-0.0027	0.0015	-1.794	0.0728
β_{xoves}	0.0074	0.0015	4.934	8.06×10^{-7} ***
β_{venres}	0.0208	0.0015	14.024	$< 2\times10^{-16}$ ***
$\beta_{sábado}$	0.0653	0.0015	44.463	$< 2\times10^{-16}$ ***
$\beta_{domingo}$	0.1089	0.0014	75.134	$< 2\times10^{-16}$ ***
$\beta_{febreiro}$	0.0852	0.0026	33.095	$< 2\times10^{-16}$ ***
β_{marzo}	0.0321	0.0039	8.278	$< 2\times10^{-16}$ ***
β_{abril}	-0.0857	0.0055	-15.531	$< 2\times10^{-16}$ ***
β_{mayo}	-0.0581	0.0067	-8.611	$< 2\times10^{-16}$ ***
$\beta_{xuño}$	-0.0783	0.0072	-10.837	$< 2\times10^{-16}$ ***
β_{xullo}	-0.0952	0.0076	-12.537	$< 2\times10^{-16}$ ***
β_{agosto}	-0.1143	0.0077	-14.817	$< 2\times10^{-16}$ ***
$\beta_{setembro}$	-0.0434	0.0070	-6.121	9.29×10^{-10} ***
$\beta_{outubro}$	-0.0112	0.0061	-1.822	0.0685
$\beta_{novembro}$	0.0307	0.0050	6.197	5.74×10^{-10} ***
$\beta_{decembro}$	-0.0560	0.0031	-18.007	$< 2\times10^{-16}$ ***

Significación dos termos suaves

	edf	**Ref.df**	**Chi.sq**	**p-valor**
$f_1(t)$	8.993	9	1141310	$< 2\times10^{-16}$ ***

Cadro 3.6: Resultados do axuste do modelo (3.4).

Media		Mediana		Desviación típica	
MAPE	MSE	MAPE	MSE	MAPE	MSE
8.7396	3608324	8.3314	3203392	2.9102	1914906

Ventá temporal	MAPE	MSE
Ventá 1	7.4185	2942756
Ventá 2	11.4817	5673891
Ventá 3	13.4148	6750609
Ventá 4	11.0726	4988708
Ventá 5	9.9768	4159619
Ventá 6	8.3314	3203392
Ventá 7	6.4131	1842191
Ventá 8	5.4392	1482985
Ventá 9	5.1079	1430852

Cadro 3.7: Resultados das medidas de erro de predición para o modelo (3.4).

Coeficientes paramétricos

	Estimación	**Sd erro**	**z valor**	**Pr($>\|z\|$)**
β_0	8.8972	0.0018	4950.190	$< 2\times10^{-16}$ ***
β_{martes}	0.0140	0.0015	9.440	$< 2\times10^{-16}$ ***
$\beta_{mércores}$	-0.0059	0.0015	-3.971	7.15×10^{-5} ***
β_{xoves}	0.0054	0.0015	3.609	8.06×10^{-7} ***
β_{venres}	0.0208	0.0015	14.024	0.000308 ***
$\beta_{sábado}$	0.0662	0.0015	45.136	$< 2\times10^{-16}$ ***
$\beta_{domingo}$	0.1108	0.0014	76.498	$< 2\times10^{-16}$ ***
$\beta_{primavera}$	-0.0620	0.0023	-26.710	$< 2\times10^{-16}$ ***
$\beta_{verán}$	-0.0659	0.0026	-25.337	$< 2\times10^{-16}$ ***
β_{outono}	0.0755	0.0025	30.688	$< 2\times10^{-16}$ ***

Significación dos termos suaves

	edf	**Ref.df**	**Chi.sq**	**p-valor**
$f_1(t)$	8.994	9	1260633	$< 2\times10^{-16}$ ***

Cadro 3.8: Resultados do axuste do modelo (3.5).

Continúamos predicindo os valores da ventá temporal 22/04/2015-19/05/2015 seguindo o modelo axustado. Na Figura 3.13 podemos ver estas predicións xunto cos valores reais da serie. Ademais desta comparativa gráfica cabe mencionar MAPE= 4.9345 % e MSE=1511061.

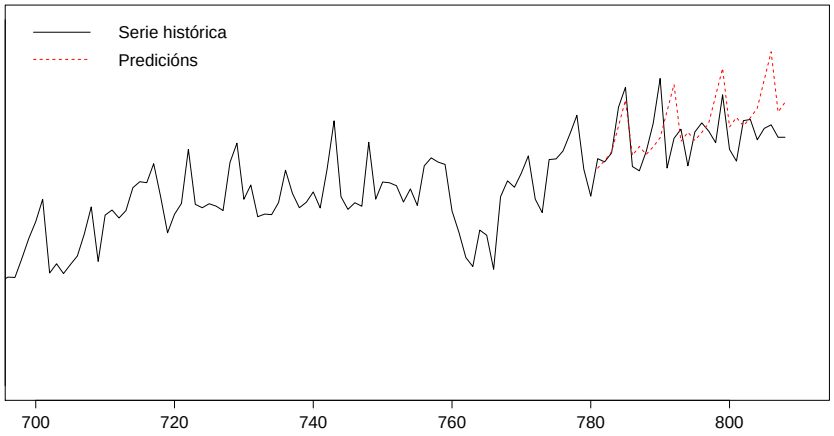


Figura 3.13: Predicións realizadas polo modelo (3.5) na ventá temporal 22/04/2015-19/05/2015 xunto aos valores reais da serie temporal.

Como se pode observar na Figura 3.13 o modelo (3.5) parece ter o mesmo problema que os modelos axustados con anterioridade. A pesar disto como o comportamento do modelo nunha única ventá non é determinante aplicamos o procedemento exposto na Sección 2.5 empregando as 9 ventás temporais definidas para este fin obtendo os resultados que se amosan no Cadro 3.9.

Media		Mediana		Desviación típica	
MAPE	MSE	MAPE	MSE	MAPE	MSE
9.4382	4365422	10.3926	4849213	3.7201	2323927

Ventá temporal	MAPE	MSE
Ventá 1	9.6000	4392798
Ventá 2	12.4037	6654683
Ventá 3	13.5422	7289995
Ventá 4	10.3926	4849213
Ventá 5	11.8390	5518619
Ventá 6	12.8727	6084194
Ventá 7	5.5736	1904431
Ventá 8	4.5430	1370928
Ventá 9	4.1757	1223933

Cadro 3.9: Resultados das medidas de erro de predición modelo (3.5).

As medidas globais de erro poñen de manifesto que en termos de erro de predición o modelo (3.5) é peor que o modelo (3.4). En xeral se revisamos os resultados expostos nos Cadros 3.5, 3.7, 3.9 chegamos a que termos de erro de predición o modelo máis conveniente é o modelo (3.4).

3.1.3. Árbores de Divisións Recursivas

Comezamos por considerar as observacións que abarcan o período 03/03/2013-21/04/2015, e axustamos unha Árbore de Divisións Recursivas empregando unicamente as observacións mencionadas para logo proporcionar predicións para a ventá temporal 22/04/2015-19/05/2015. Na modelización da serie temporal mediante Árbores de Divisións Recursivas é moi importante seleccionar as covariables axeitadas, e é frecuente introducir como variables explicativas o valor da serie temporal en instantes pasados. Concretamente decidimos introducir como variable explicativa o valor da serie temporal 7 días atrás, variable $re7$, pois como razoamos na aplicación da metodoloxía Box-Jenkins existe unha compoñente estacional semanal. Ademais incluímos o valor da serie temporal o día anterior ao que desexamos predicir, variable $re1$. Se axustamos unha Árbore de Divisións Recursivas coas variables mencionadas obtemos o resultado que se amosa de contado.

```
1) re1 <= 10344; criterion = 1, statistic = 751.553
  2) re1 <= 6150; criterion = 1, statistic = 512.436
    3) re1 <= 4987; criterion = 1, statistic = 218.487
      4) re1 <= 3907; criterion = 1, statistic = 60.369
        5) re1 <= 3609; criterion = 1, statistic = 17.989
          6)*  weights = 9
        5) re1 > 3609
          7) re7 <= 4024; criterion = 0.981, statistic = 6.737
            8)*  weights = 20
          7) re7 > 4024
            9)*  weights = 13
      4) re1 > 3907
        10) re7 <= 4429; criterion = 1, statistic = 46.17
          11)*  weights = 115
        10) re7 > 4429
          12) re1 <= 4567; criterion = 0.999, statistic = 11.655
            13)*  weights = 39
          12) re1 > 4567
            14)*  weights = 20
    3) re1 > 4987
      15) re7 <= 5052; criterion = 1, statistic = 24.779
        16)*  weights = 20
      15) re7 > 5052
        17)*  weights = 78
  2) re1 > 6150
    18) re7 <= 7325; criterion = 1, statistic = 190.459
      19) re7 <= 6142; criterion = 1, statistic = 49.495
        20) re1 <= 7165; criterion = 0.967, statistic = 5.732
          21) re7 <= 5443; criterion = 0.999, statistic = 11.594
            22)*  weights = 9
          21) re7 > 5443
            23)*  weights = 21
        20) re1 > 7165
          24)*  weights = 7
      19) re7 > 6142
        25) re1 <= 7448; criterion = 1, statistic = 26.478
          26) re1 <= 6201; criterion = 0.981, statistic = 6.684
            27)*  weights = 7
          26) re1 > 6201
            28)*  weights = 54
        25) re1 > 7448
          29)*  weights = 16
    18) re7 > 7325
      30) re7 <= 8969; criterion = 1, statistic = 74.676
        31) re1 <= 8091; criterion = 1, statistic = 25.256
```

```
        32)*  weights = 52
      31) re1 > 8091
        33) re7 <= 8461; criterion = 0.993, statistic = 8.559
          34)*  weights = 50
        33) re7 > 8461
          35)*  weights = 22
    30) re7 > 8969
      36) re1 <= 8755; criterion = 1, statistic = 18.562
        37)*  weights = 7
      36) re1 > 8755
        38) re7 <= 9792; criterion = 0.999, statistic = 12.541
          39)*  weights = 14
        38) re7 > 9792
          40)*  weights = 9
1) re1 > 10344
  41) re1 <= 15976; criterion = 1, statistic = 156.34
    42) re7 <= 11960; criterion = 1, statistic = 88.815
      43) re7 <= 10167; criterion = 1, statistic = 18.704
        44)*  weights = 8
      43) re7 > 10167
        45)*  weights = 22
    42) re7 > 11960
      46) re7 <= 14150; criterion = 1, statistic = 43.524
        47) re1 <= 13120; criterion = 0.996, statistic = 9.606
          48)*  weights = 27
        47) re1 > 13120
          49)*  weights = 44
      46) re7 > 14150
        50) re1 <= 14311; criterion = 0.997, statistic = 10.298
          51)*  weights = 12
        50) re1 > 14311
          52)*  weights = 17
  41) re1 > 15976
    53) re7 <= 14693; criterion = 1, statistic = 18.018
      54)*  weights = 8
    53) re7 > 14693
      55) re1 <= 17762; criterion = 0.985, statistic = 7.191
        56)*  weights = 36
      55) re1 > 17762
        57) re7 <= 17765; criterion = 0.962, statistic = 5.503
          58)*  weights = 17
        57) re7 > 17765
          59)*  weights = 7
```

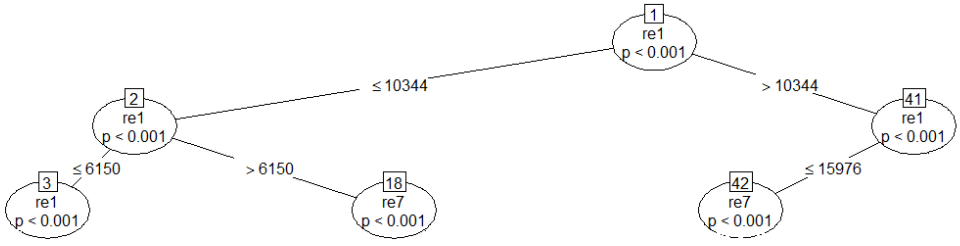


Figura 3.14: Gráfico dos primeiros nodos da Árbore de Divisións Recursivas axustada.

A información que vimos de amosar permite coñecer as divisións realizadas na árbore en base ás covariables consideradas, é dicir, permite coñecer en cada nodo que covariable foi seleccionada xunto os subconxuntos disxuntos que dividiron o espazo de definición da covariable dando lugar a dous novos nodos. Por exemplo, no nodo 1 observamos que seleccionamos a variable $re1$ como a covariable con maior asociación coa resposta, e que dividimos o espazo de definición desta covariable considerando

por unha parte as observacións maiores que 10344, e por outra as menores ou iguais a este valor, dando lugar a dous novos nodos, que seguindo a notación empregada na construción da árbore son o nodo 2 e 41. Proseguimos fixándonos agora no nodo 2, vemos que a variable con maior correlación coa resposta é de novo a variable $re1$, e que dividimos o subconxunto de observacións menores ou iguais a 10344 considerando as observacións menores ou iguais a 6150 por unha parte, e por outra as maiores que este valor, dando lugar aos nodos 3 e 18. Seguindo deste xeito podemos comentar cada un dos pasos levados a cabo polo algoritmo de construción da árbore de divisións recursivas. Unha vez construída a árbore realízamos predicións para valores futuros. O xeito de calcular estas predicións recordemos que consiste en ver a que nodo terminal se corresponden os valores das covariables asociados ao instante que desexamos predicir, e seguidamente predicir segundo a media dos valores da resposta cuxos valores das variables explicativas se encontran en dito nodo. Antes de amosar os valores preditos pola árbore axustada na ventá temporal considerada debemos expor unha matiz en canto ás variables explicativas. Para predicir cada valor da ventá temporal necesito coñecer os valores das covariables para o valor da resposta que desexo predicir, isto non é posible pois a ventá temporal ten 28 observacións entón non podemos coñecer para cada un destes valores o valor anterior da serie temporal, ademais do valor da serie temporal 7 días atrás. Por este motivo facemos o seguinte, predicimos o primeiro valor da ventá temporal, pois para este valor si dispomos dos valores das covariables, e empregamos o valor obtido na predición para actualizar as covariables e así poder proseguir co proceso de predición. Actuamos así deste xeito iterativo ata conseguir as predicións dos 28 valores da ventá, é dicir, predicimos un valor, actualizamos as covariables supoñendo que o valor predito é o valor observado da serie temporal, predicimos o seguinte valor pois xa dispomos da información necesaria, e así sucesivamente. Na Figura 3.15 observamos as predicións realizadas para esta ventá temporal. O resultado non é satisfactorio a pesar de que as medidas de erro son MAPE=11.2905 % e MSE=5876377. Cabe mencionar que non debe sorprender que o modelo realice predicións constantes pois iso simplemente significa que para todos os valores da ventá temporal predicimos segundo o promedio obtido no mesmo nodo terminal.

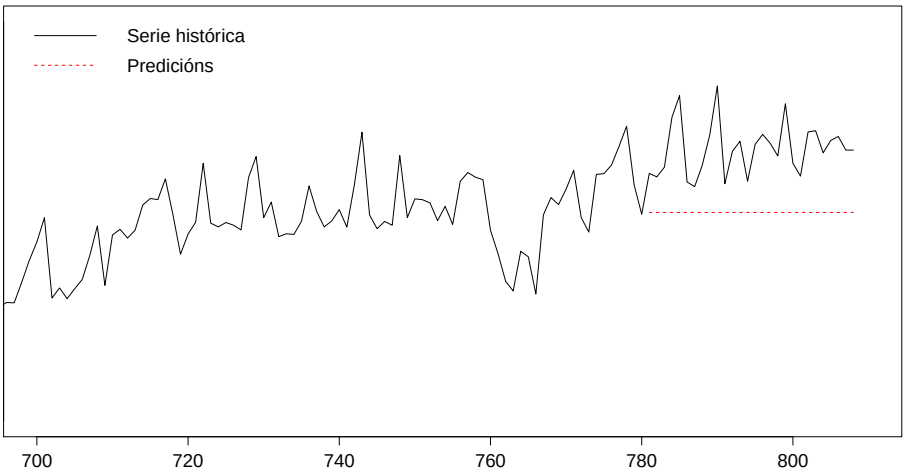


Figura 3.15: Gráfico secuencial dos valores reais da serie temporal xunto os valores preditos pola árbore axustada no período 22/04/2015-19/05/2015.

Como xa mencionamos non debemos extraer conclusións finais mediante os resultados obtidos nunha única ventá, por este motivo seguimos o procedemento exposto na Sección 2.5, e axustamos unha árbore unicamente coas observacións previas a cada unha das 9 ventás temporais, realizamos

predicións para as mesmas, e comparamos os valores preditos cos observados, obtendo así 9 valores para a medida MAPE e 9 valores para medida MSE, os cales resumimos nas medidas globais expostas no Cadro 3.10.

Media		Mediana		Desviación típica	
MAPE	MSE	MAPE	MSE	MAPE	MSE
7.9293	3349083	7.4418	2266968	3.1442	2830240

Ventá temporal	MAPE	MSE
Ventá 1	4.8677	1305991
Ventá 2	6.1075	1787181
Ventá 3	5.6088	1664626
Ventá 4	6.8710	2123073
Ventá 5	7.4418	2266968
Ventá 6	8.1690	3049800
Ventá 7	15.6244	10532600
Ventá 8	8.1792	3635242
Ventá 9	8.4933	3776266

Cadro 3.10: Resultados das medidas de erro de predición para a árbore de divisións recursivas axustada.

Os valores das medidas globais de erro de predición si son axeitadas, isto pon de manifesto que a pesar de que na ventá seleccionada para comentar e ilustrar o modelo o comportamento do mesmo sexa mellorable globalmente predí razoablemente ben.

Podemos pensar en introducir máis variables explicativas ou incluso en substituír as variables $re1$ e $re7$ por outras covariables. Por este motivo consideramos as variables categóricas que indican o día da semana, o mes, e a estación do ano, e realizamos diferentes Árbores de Divisións Recursivas xogando coa introdución das variables mencionadas. Co obxectivo de poder decidir se as probas realizadas melloran os resultados proporcionadas pola árbore amosada con anterioridade, calculamos as medidas de erro de predición globais obtendo sempre valores superiores ou moi semellantes aos expostos no Cadro 3.10, concluíndo pois que as variables $re1$ e $re7$ seleccionadas son as axeitadas.

3.1.4. Métodos de predición simples

Neste apartado imos levar a cabo a aplicación dos métodos de predición simples co obxectivo de comparar os resultados obtidos cos proporcionados polas tres metodoloxías que vimos de aplicar. Consideramos inicialmente a serie temporal composta polas observacións[4] do período 03/03/2013-21/04/2015 mediante as cales axustamos o modelo, que logo empregamos para predicir os valores observados na ventá temporal 22/04/2015-19/05/2015. Na Figura 3.16 podemos ver as predicións

[4]A aplicación destes métodos realízase sobre a serie temporal orixinal, é dicir, sen a eliminación previa dos valores atípicos, pois partimos da suposición de que se a empresa utiliza estes métodos simples na predición de series temporais é porque non dispón de coñecementos sobre o estudo das mesmas, e consecuentemente non realizará un tratamento previo da serie temporal.

realizadas por cada un dos métodos simples expostos na Subsección 2.4.2 na ventá temporal considerada xunto aos valores reais. Ademais desta comparativa gráfica, presentamos no Cadro 3.11 os valores das medidas de erro de predición para cada un dos métodos simples nesta ventá temporal.

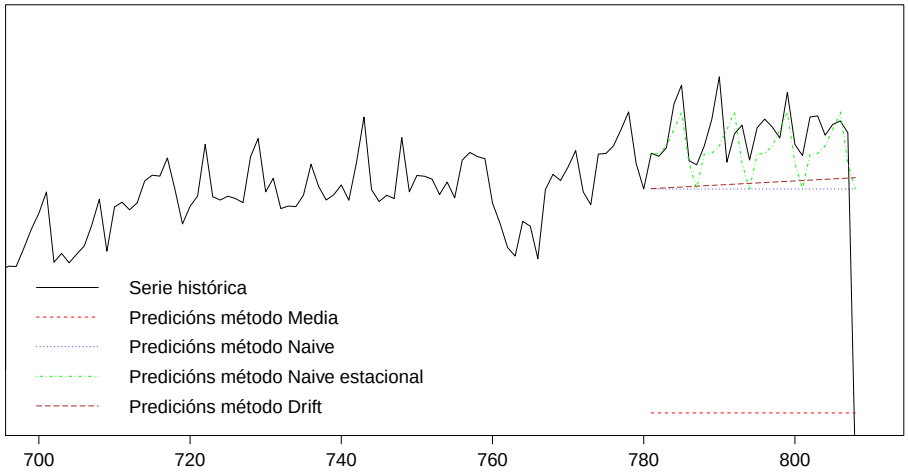


Figura 3.16: Gráfico secuencial dos valores preditos por cada un dos métodos simples no período 22/04/2015-19/05/2015 xunto aos valores reais da serie temporal.

	Método da Media	Método Naive	Método Naive Estacional	Método Drift
MAPE	56.9439	17.9200	11.0872	17.0666
MSE	124079883	10514326	5663992	9902004

Cadro 3.11: Valores de MAPE e MSE para cada un dos métodos simples na ventá temporal 22/04/2015-19/05/2015.

Deseguido proporcionamos algunhas observacións sobre os resultados recollidos na Figura 3.16 e no Cadro 3.11. Comecemos polo método da Media. Este método non capta o nivel da serie temporal dando lugar a predicións moi desatinadas, debido a presenza de tendencia crecente na serie temporal, a cal é imposible captar predicindo mediante a media dos valores observados. Isto tamén se reflexa nos valores do MAPE e MSE pois como se pode observar os valores de erro dispáranse converténdose nos maiores valores de erro rexistrados ata o momento neste estudo. Prosigamos co método Naive, este método está próximo a captar o nivel da serie temporal pero non o consegue porque a serie temporal segue crecendo mentres o método Naive predí constante mediante o último valor observado. A pesar disto as medidas de erro non toman valores descomunais sobre todo se temos en conta que na ventá temporal considerada está presente un valor atípico que produce un aumento das diferenzas entre os valores preditos e os observados. Con respecto ao método Naive estacional observamos que as súas predicións se encontran lixeiramente por debaixo do nivel da serie temporal, polo mesmo motivo exposto para o método Naive, coa diferenza de que o seguimento dun patrón semanal permite reducir

os valores dos erros de predición. A pesar de que non se capte correctamente o comportamento real da serie temporal, pois esta tarefa é realmente complexa polos motivos xa mencionados con anterioridade. Para rematar, o método Drift proporciona un resultado similar ao do método Naive cunha pequena mellora á hora de captar o nivel da serie temporal, a cal se reflexa lixeiramente nas medidas de erro.

Os resultados obtidos nesta ventá temporal serven para comprender que inconvenientes pode ter en xeral a aplicación dos métodos simples na nosa serie temporal, agora ben non é suficiente para realizar unha comparativa coas outras metodoloxías empregadas no estudo desta serie. Para iso necesitamos obter as medidas de erro globais mediante o procedemento exposto na Subsección 2.5 empregando as 9 ventás temporais definidas para este fin. Os resultados obtidos mediante dito procedemento encóntranse no Cadro 3.12. Nótese que os resultados globais obtidos para o método da Media son moi similares os obtidos na ventá temporal 22/04/2015-19/05/2015, mentres que os resultados globais para os restantes métodos indican unha mellora con respecto aos resultados na ventá mencionada.

	Media		Mediana		Desviación típica	
Metodoloxía	MAPE	MSE	MAPE	MSE	MAPE	MSE
Media	55.4696	103005294	56.6456	95743675	1.3422	14256845
Naive	7.7284	3216167	6.2804	2223956	3.3711	2917485
Naive estacional	8.7744	4087925	7.8351	2765712	3.6615	3592455
Drift	7.2887	2865296	6.7149	2155177	3.0810	2486275

	Media		Naive		Naive estacional		Drift	
Ventá temporal	MAPE	MSE	MAPE	MSE	MAPE	MSE	MAPE	MSE
Ventá 1	55.5453	95410444	5.0591	1496417	4.9368	1157484	5.0890	1475363
Ventá 2	54.5768	90734126	6.2135	1795183	5.6951	1630561	6.7149	2155177
Ventá 3	54.1229	88624656	6.2804	1700410	6.7604	2461987	6.0604	1645423
Ventá 4	54.4322	93496755	7.0182	2223956	7.8998	2765712	7.1267	2353754
Ventá 5	54.4953	95743675	8.3348	2795742	7.8360	2438617	7.6040	2350823
Ventá 6	54.6456	99907988	9.8741	4392126	9.2288	3622788	8.8757	3576508
Ventá 7	56.4694	114398863	15.7040	10627119	16.6813	12455790	14.6062	9240638
Ventá 8	57.2551	121544023	6.1262	2260225	12.3927	7262456	5.1472	1715433
Ventá 9	57.6842	127187113	4.9453	1654327	7.5398	2995932	4.3755	1274649

Cadro 3.12: Valores das medidas de erro de predición para os métodos de predición simples.

3.1.5. Conclusións

Tras modelar a serie temporal "número de usuarios diarios do servizo de vídeo baixo demanda" mediante o uso da metodoloxía Box-Jenkins, da metodoloxía GAM, e das Árbores de Divisións Recursivas, ademais dos métodos simples, imos proceder á selección da metodoloxía que a compañía debe empregar na predición de valores futuros desta serie temporal. Para iso no Cadro 3.13 recollemos os valores das medidas globais de erro para cada unha das metodoloxías aplicadas.

	Media		Mediana		Desviación típica	
Metodoloxía	MAPE	MSE	MAPE	MSE	MAPE	MSE
Box-Jenkins	8.5221	3844313	7.4363	2569940	2.9247	2758470
GAM	8.7396	3608324	8.3314	3203392	2.9102	1914906
Árbores	7.9293	3349083	7.4418	2266968	3.1442	2830240
Media	55.4696	103005294	56.6456	95743675	1.3422	14256845
Naive	7.7284	3216167	6.2804	2223956	3.3711	2917485
Naive estacional	8.7744	4087925	7.8351	2765712	3.6615	3592455
Drift	7.2887	2865296	6.7149	2155177	3.0810	2486275

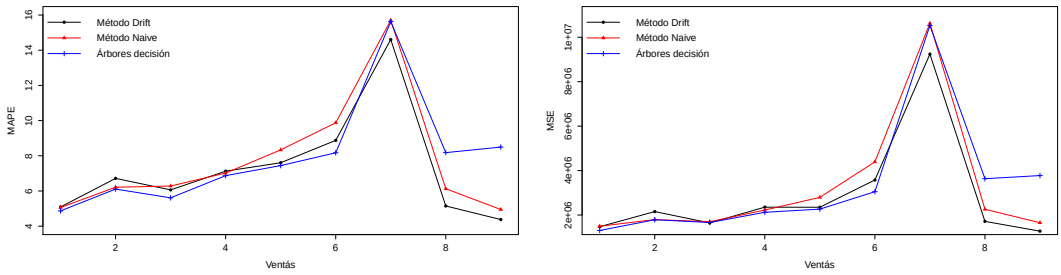
Cadro 3.13: Valores das medidas globais de erro para cada unha das metodoloxías aplicadas.

Se analizamos os resultados presentes no Cadro 3.13 chegamos a conclusión de que a metodoloxía seleccionada debe ser unhas das seguintes tres: método Naive, método Drift, e Árbores de Divisións Recursivas. Pode resultar sorprendente este resultado pois finalmente temos que escoller entre dous métodos simples e as Árbores de Divisións Recursivas que si recordamos en algunha ventá predí constante. Se reflexionamos sobre o tipo de serie temporal que estamos a predicir quizais este feito non sexa tan sorprendente, pois a serie ten unha tendencia clara e unha compoñente estacional semanal, pero o comportamento semanal da mesma é moi variable, polo cal o máis axeitado é captar a tendencia e predicir segundo esta, pois os outros métodos no intento de captar a compoñente estacional cometen un erro maior debido á excesiva variabilidade da serie temporal. Segundo as medidas globais de erro parece que o máis axeitado é o método Drift, a pesar disto decidimos proporcionar máis información para tomar a decisión. Por este motivo no Figura 3.17a e 3.17b mostramos os valores das medidas MAPE e MSE, respectivamente, nas 9 ventás consideradas para cada unha das tres metodoloxías, ademais disto no Cadro 3.14 expomos diferentes características das metodoloxías.

	Método Drift	Método Naive	Árbores de Divisións Recursivas
Tempo de execución	0.02 s	0.04 s	0.08 s
Complexidade	Baixa	Baixa	Media

Cadro 3.14: Características do método Naive, do método Drift, e das Árbores de Divisións Recursivas.

Na Figura 3.17 podemos ver que nas últimas tres ventás o comportamento das Árbores de Divisións Recursivas é peor que o presentado polos métodos simples, agora ben nas outras 6 ventás son as Árbores de Divisións Recursivas as que dan lugar a valores da MAPE e MSE máis baixos. Por outra parte os métodos simples teñen un tempo de computación inferior ás Árbores de Divisións Recursivas, ademais dunha menor complexidade. Consecuentemente se non temos en conta a complexidade e o tempo de execución empregaríamos as Árbores de Divisións Recursivas pois en 6 das 9 ventás proporciona valores menores de MAPE e MSE que o método Drift, mentres que se o que desexamos é obter inmediatamente resultados empregaríamos a metodoloxía Drift.



(a) Valores da medida MAPE.

(b) Valores da medida MSE.

Figura 3.17: Gráficos dos valores de MAPE e MSE para o método Naive, método Drift, e as Árbores de Divisións Recursivas.

3.1.6. Outras aplicacións

A pesar de que a serie temporal ten un comportamento altamente variable, os resultados obtidos en termos de erro de predición son aceptables. Non obstante decidimos cerrar o estudo desta serie temporal aplicando novas ideas buscando unha posible mellora dos resultados. Deseguido expomos brevemente as ideas aplicadas e os resultados obtidos.

Serie Semanal. A primeira idea é construír a partires da serie diaria unha serie semanal, pois quizais deste xeito a serie teña menor variabilidade e isto facilite a súa modelización. É certo que obviamente mediante a modelización desta serie só obteremos predicións semanais, e polo tanto a empresa perderá información, e as predicións non axudarán a anticiparse a posibles anomalías no uso do servizo, pero si permitirán coñecer se as seguintes semanas o número de usuarios está de acordo co crecemento esperado, ou pola contra as predicións indican unha diminución do uso deste servizo.

Na Figura 3.18 podemos ver o gráfico secuencial da serie semanal mencionada. Co obxectivo de modelar a serie semanal aplicamos diferentes metodoloxías, a primeira delas a metodoloxía Box-Jenkins. Tal e como se ve no gráfico da serie esta presenta unha tendencia crecente clara, polo cal na procura dun modelo Box-Jenkins axeitado para a modelización desta serie eliminamos a tendencia mediante unha diferenza regular, obtendo unha serie diferenciada que claramente é ruído branco, é dicir, a aplicación da metodoloxía Box-Jenkins indícanos que unha vez eliminada a tendencia só temos ruído branco, sendo polo tanto o modelo ARIMA(0,1,0) adecuado para a súa modelización. A pesar disto na validación deste modelo observamos que os residuos non teñen variabilidade constante, polo que finalmente aplicamos o modelo mencionado pero a serie transformada mediante a transformación logarítmica. O modelo obtido é concordante co que axustamos na serie diaria, pois ao crear a serie semanal eliminamos a compoñente estacional semanal presente na serie diaria quedando unicamente unha tendencia. Por outra parte modelamos a serie temporal semanal mediante a metodoloxía GAM introducindo mediante unha función non paramétrica parcial a covariable semana de observación, que non é máis que a secuencia crecente $1, 2, \cdots, 115$ que indica a que semana do estudo se corresponde a observación, con esta covariable é suficiente para modelar a serie temporal pois só temos tendencia. De feito estudamos a posibilidade de introducir outras covariables no modelo o cal descartamos tras comprobar que era contraproducente. Por outra banda aplicamos os métodos de predición simples, sendo moi interesante o resultado proporcionado polo método Drift pois ao ter só tendencia o método realiza predicións axeitadas. No Cadro 3.15 podemos ver os valores das medidas globais de erro de predición obtidas mediante o procedemento exposto na Sección 2.5 considerando certas ventás temporais definidas previamente.

Os resultados amosados no Cadro 3.15 indican que tanto a metodoloxía Box-Jenkins como o método Drift son axeitados para predicir a serie temporal semanal dando lugar a un erro de predición satisfactorio. Polo tanto creemos interesante proporcionar a empresa de telecomunicacións as predi-

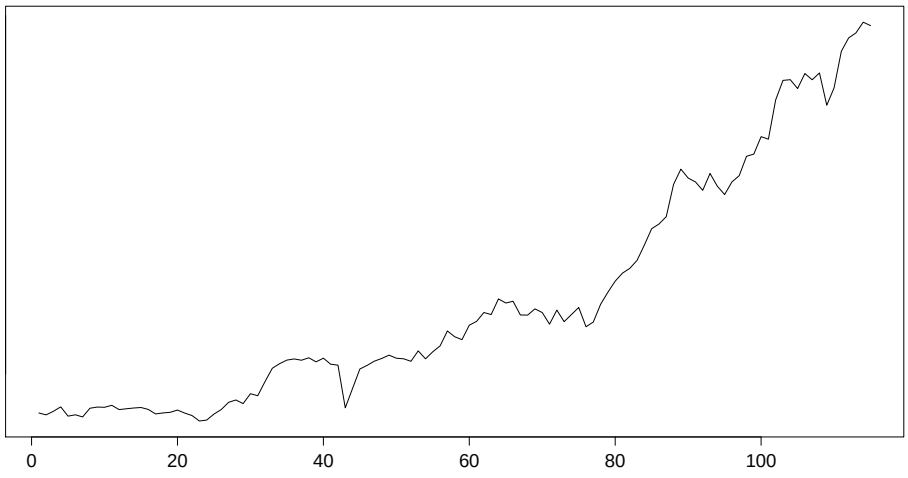


Figura 3.18: Gráfico secuencial da serie semanal construída a partires da serie diaria do "número de usuarios diarios do servizo de vídeo baixo demanda".

	Media		Mediana		Desviación típica	
Metodoloxía	MAPE	MSE	MAPE	MSE	MAPE	MSE
Modelo Box-Jenkins	5.4520	68149019	4.6218	45293978	2.1323	45414973
Modelo GAM	8.2832	152065357	7.1957	121915445	4.8903	124684491
Método Drift	5.3703	71280281	4.9702	45914880	3.0242	71782743

Cadro 3.15: Valores das medidas globais de erro de predición para a serie semanal.

cións semanais como un complemento das predicións diarias, pois as predicións semanais soamente non cobren todas as necesitades da empresa.

Unha serie para cada día da semana. O problema da serie temporal semanal é que obtemos predicións semanais cando en realidade para a empresa o maior interese está en predicións diarias. Por este motivo xorde a idea de considerar 7 series temporais unha para cada día da semana, se lévamos a cabo a construción das mesmas observamos que as novas series só presentan tendencia, a mesma situación ca na serie semanal, polo que os modelos aplicados a serie semanal serían válidos é permitirían realizar predicións diarias. A pesar disto tras certas comprobacións tomamos a decisión de non empregar está idea pois quizais se consiga unha pequena mellora nas predicións pero iso non compensará a necesidade de estudar 7 series temporais, pois para a empresa o factor tempo é relevante.

A temperatura como covariable. Posto que a variable medida na serie temporal en estudo pode ter unha relación coas condicións meteorolóxicas decidimos introducir como covariable a temperatura máxima diaria. Pese a que inicialmente se esperaba que o comportamento do modelo GAM e da árbore de de divisións recursivas mellorase coa introdución desta variable non é así. Pode ser debido a que a influencia desta variable é complexa no seguinte senso: por exemplo un día de 18 grados centígrados en inverno é un día no cal os usuarios poden sentir a necesidade de empregar o tempo de lecer na

realización de actividades ao aire libre, mentres que un día de verán con esta temperatura pode ter o efecto contrario. Como consecuencia desbotamos a idea de introducir esta covariable no modelo.

3.2. Estudo da serie número de baixas diarias de teléfono móbil

Nesta sección abordamos o estudo da serie temporal "número de baixas diarias de teléfono móbil". Esta serie temporal comeza o 01/01/2012 e remata o 31/03/2015, consecuentemente dispoñemos de 1186 observacións para o presente estudo. A predición desta serie temporal permítelle á empresa de telecomunicacións adiantarse aos feitos, por exemplo, realizando accións de retención e fidelización de clientes en risco de baixa se as predicións indican un número elevado de baixas, é dicir, a predición desta serie temporal é útil na mercadotecnia de servizos.

Na Figura 3.19 encóntrase o gráfico secuencial da serie temporal. En dito gráfico observamos que a variabilidade da serie temporal crece a medida que pasa o tempo, sendo considerablemente maior ao final da mesma que ao inicio desta, é dicir, a serie temporal presenta heterocedasticidade. Con respecto á existencia ou non de tendencia é difícil chegar a unha conclusión só co estudo do gráfico secuencial, pois poderíamos pensar nunha posible tendencia crecente pero pode ser que non exista tal tendencia, e que o crecemento observado se deba a variabilidade crecente da serie. Analogamente ao exposto na serie "número de usuarios diarios do servizo de vídeo baixo demanda", é complexo detectar unha compoñente estacional no gráfico secuencial, polo cal máis adiante identificaremos esta compoñente en caso de existir mediante a *fas* mostral da serie temporal. Por outra parte observamos que a variable número de baixas diarias de teléfono móbil toma o valor cero nunha cantidade considerable de días. Agora ben debemos ver a que se deben estes valores nulos, pois está claro que non son valores atípicos, se non que estes valores forman parte do comportamento da serie temporal. Se buscamos a que datas se corresponden estes valores nulos, observamos que a maior parte deles son domingos e luns, días nos que a compañía non leva a cabo os trámites referentes ás baixas. Sen embargo non todos os luns e domingos sucede isto, pois algún luns a empresa contabiliza as baixas, e incluso algún domingo sucede isto. Ademais se seguimos indagando nas datas encóntramos valores nulos noutros días da semana e decatámonos de que algún deles ten relación coa presenza de festivos, pero outros carecen dunha explicación clara, amais se comprobamos todos os festivos vemos que en moitos deles se contabilizan as baixas. Por todo isto chegamos a que o comportamento destes valores nulos é complexo sendo imposible comprender o motivo de cada un dos valores nulos presentes na serie temporal, polo que simplemente nos centraremos en tratar o comportamento maioritario que indica que os domingos e luns a variable medida en xeral non se contabiliza.

Por outra parte, no gráfico secuencial da serie temporal tamén vemos certos días nos que o número da baixas é considerablemente maior que nos restantes días da serie temporal. Se pescudamos acerca de que datas son as correspondentes a estes repuntes chegamos a afirmar que os repuntes observados se corresponden sempre co día 21 de cada mes. Tras isto preguntámonos polo motivo dos mesmos, chegando a que se deben ao sistema de traballo da empresa de telecomunicacións.

Analogamente ao realizado para o estudo da serie temporal "número de usuarios diarios do servizo de vídeo baixo demanda", expomos deseguido as ventás temporais seleccionadas para levar a cabo o procedemento exposto na Sección 2.5. Consideramos $m = 10$ ventás, cada unha delas con $m_N = 28$ observacións, e tomamos $T_1 = 1113$ e $l = 5$, dando así lugar ás seguintes ventás temporais:

- Ventá 1: Dende 18/01/2015 ata 14/02/2015.
- Ventá 2: Dende 23/01/2015 ata 19/02/2015.
- Ventá 3: Dende 28/01/2015 ata 24/02/2015.
- Ventá 4: Dende 02/02/2015 ata 01/03/2015.
- Ventá 5: Dende 07/02/2015 ata 06/03/2015.
- Ventá 6: Dende 12/02/2015 ata 11/03/2015.

- Ventá 7: Dende 17/02/2015 ata 16/03/2015.
- Ventá 8: Dende 22/02/2015 ata 21/03/2015.
- Ventá 9: Dende 27/02/2015 ata 26/03/2015.
- Ventá 10: Dende 04/03/2015 ata 31/03/2015.

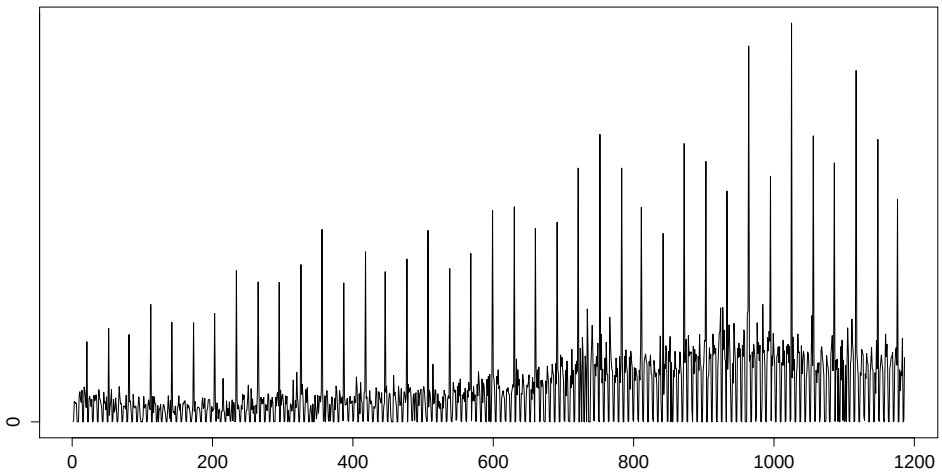


Figura 3.19: Gráfico secuencial da serie temporal número de baixas diarias de teléfono móbil.

Igualmente que no estudo da serie "número de usuarios diarios do servizo de vídeo baixo demanda", non empregaremos todas as ventás temporais para a realización do procedemento exposto na Sección 2.5, se non que empregamos para este procedemento as primeiras 9 ventás temporais e usamos a última delas para ilustrar con detalle a aplicación de cada unhas das metodoloxías.

3.2.1. Metodoloxía Box-Jenkins

Como vimos de expor, para ilustrar con detalle a aplicación de cada metodoloxía consideramos a serie temporal formada polas observacións do período 01/01/2012-03/03/2015 e axustamos un modelo á serie temporal empregando unicamente estas observacións. Unha vez axustado o modelo empregámolo para predicir os valores da ventá temporal 04/03/2015-31/03/2015. Posto que dispomos dos valores reais observados en dita ventá usamos estes para realizar comparativas entre os mesmos e as predicións, comparativa gráfica e cuantitativa.

Na descrición do gráfico secuencial, Figura 3.19, mencionamos que a variabilidade da serie temporal parece ir en aumento a medida que pasa o tempo, polo cal debemos solucionar este inconveniente antes de iniciar o proceso de identificación dun modelo Box-Jenkins. Tal e como expuxemos na Subsección 2.2.1, a transformación Box-Cox permite estabilizar a variabilidade da serie temporal neste tipo de situacións. Polo tanto achamos o valor do parámetro λ da transformación Box-Cox na serie temporal formada pola observacións do período 01/01/2012-03/03/2015 obtendo $\lambda = -0.087$. Ante este valor do parámetro podemos asumir que este é nulo, e consecuentemente aplicar unha transformación logarítmica, a cal é máis sinxela. Agora ben a serie temporal ten valores nulos, polo que non é posible aplicar directamente a transformación logarítmica sendo necesario salvar este inconveniente do seguinte xeito. Seleccionamos unha constante fixa $\varepsilon \in \mathbb{R}^+$, e sumamos dita constante á serie temporal obtendo deste

xeito unha nova serie temporal cuxos valores son todos eles maiores que cero sobre a cal podemos aplicar sen máis a transformación logarítmica. Se levamos a cabo esta tarefa obtemos efectivamente unha serie temporal cuxa variabilidade é constante, podendo proceder a identificación dun modelo Box-Jenkins.

Un primeiro paso para identificar un modelo Box-Jenkins é debuxar e estudar a *fas* mostral da serie temporal. Na Figura 3.20 podemos ver o gráfico da *fas* mostral, nel observamos forte correlación no retardo 7 e nos seus múltiplos, ademais de periodicidade de período 7 e converxencia lenta á cero. Estas características indican a presenza dunha compoñente estacional de período 7 días, é dicir, compoñente estacional semanal, véxase Subsección 2.2.2. Por este motivo aplicamos sobre a serie temporal unha diferenza estacional de período 7 días co obxectivo de eliminar a compoñente estacional detectada e así poder identificar un modelo Box-Jenkins. Na Figura 3.21 podemos ver as correlacións simples e parciais mostrais da serie diferenciada estacionalmente. A diferenciación estacional permitiu eliminar a compoñente estacional, agora ben as correlacións mostran unha forte dependencia estacional estando esta presente en retardos tan altos como o 118, isto supón un problema pois se identificamos un modelo Box-Jenkins baseándonos no comportamento presente na *fas* e *fap* mostral necesariamente o modelo escollido terá unha gran cantidade de parámetros. Esta situación pon de manifesto que a metodoloxía Box-Jenkins non é a máis axeitada para modelar a serie número de baixas diarias de teléfono móbil, a pesar disto decidimos proseguir é comprobar que efectivamente a metodoloxía Box-Jenkins non proporciona un bo modelado da serie temporal.

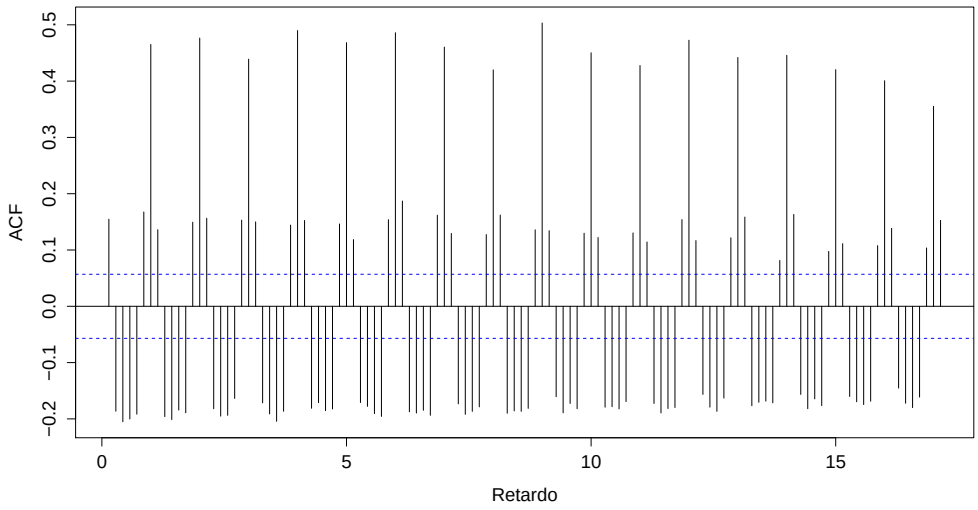
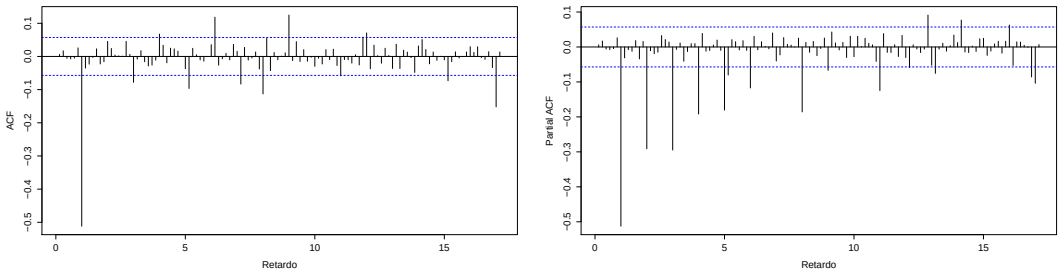


Figura 3.20: Gráfico das correlacións simples mostrais da serie temporal "número baixas diarias de teléfono móbil" tras a aplicación da transformación logarítmica.

Posto que a identificación dun modelo Box-Jenkins seguindo a información amosada pola *fas* e *fap* mostral é moi complexa, decidimos empregar a función auto.arima do paquete forecast a cal identifica un modelo Box-Jenkins para a serie temporal e axusta dito modelo. Se aplicamos esta función e seguidamente calculamos as predicións realizadas polo modelo axustado obtemos o resultado que se pode observar na Figura 3.22.

Como podemos observar, as predicións realizadas polo modelo Box-Jenkins axustado pola función auto.arima non son adecuadas pois o modelo non capta correctamente nin o comportamento semanal presente na serie temporal nin os repuntes referentes o día 21. De entre as medidas expostas na Sección

2.5 decidimos empregar as medidas MASE e MSE para a comparativa dos valores reais coas predicións realizadas, pois a medida MAPE non está definida para esta serie temporal debido aos valores nulos.



(a) *Fas* mostral (b) *Fap* mostral

Figura 3.21: Gráfico de correlacións simples e parciais mostrais da serie temporal diferenciada estacionalmente.

Se calculamos os valores destas medidas obtemos MSE=6453.228 e MASE=72.4945 %, este valor do MASE indica que o modelo Box-Jenkins considerado é un 27.5055 % máis axeitado que o modelo que predí cada valor futuro como o valor observado no instante inmediatamente anterior o que desexamos predicir, pode consultarse a interpretación do MASE na Sección 2.5. Agora ben para posteriormente poder comparar os resultados proporcionados por este modelo cos obtidos empregando outras metodoloxías aplicamos o procedemento exposto na Sección 2.5 empregando as 9 ventás temporais definidas para este fin obtendo os resultados que se amosan no Cadro 3.16.

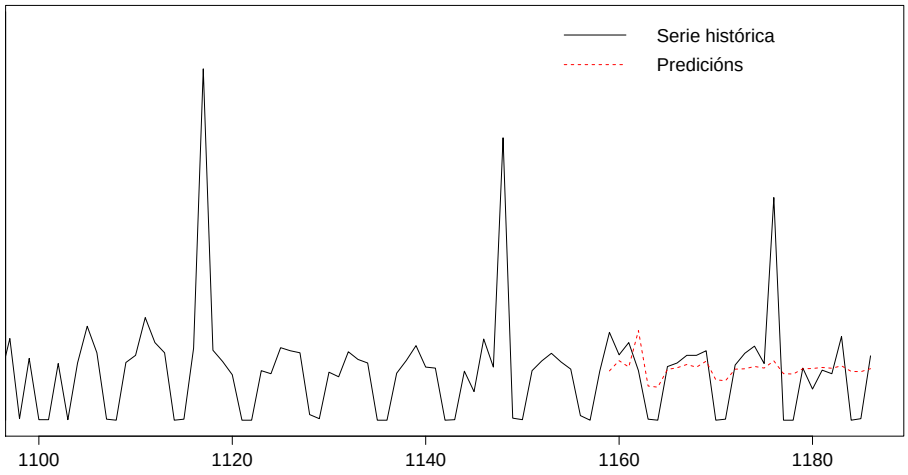


Figura 3.22: Predicións realizadas polo modelo Box-Jenkins na ventá temporal 04/03/2015-31/03/2015 xunto aos valores reais da serie temporal.

Os resultados do Cadro 3.16 indican que o comportamento global do modelo Box-Jenkins axustado pola función auto.arima é similar ao exposto na ventá 04/03/2015-31/03/2015.

Media		Mediana		Desviación típica	
MASE	MSE	MASE	MSE	MASE	MSE
79.1021	9263.605	69.461	9654.952	19.0252	2975.701

Ventá temporal	MASE	MSE
Ventá 1	69.2461	14884.370
Ventá 2	109.6005	4233.784
Ventá 3	78.4726	11073.660
Ventá 4	68.8648	9954.330
Ventá 5	65.4886	9704.923
Ventá 6	63.7216	9616.119
Ventá 7	67.1990	9654.952
Ventá 8	113.5749	7661.216
Ventá 9	75.7508	6589.091

Cadro 3.16: Resultados das medidas de erro de predición para o modelo Box-Jenkins axustado pola función auto.arima.

3.2.2. Metodoloxía GAM

Tras comprobar que a metodoloxía Box-Jenkins non é a apropiada para modelar a serie temporal "número de baixas diarias de teléfono móbil" proseguimos co estudo empregando agora a metodoloxía GAM. Comezamos ilustrando con detalle o axuste dun modelo GAM empregando soamente as observacións do período 01/01/2012-03/03/2015. Unha vez axustado este modelo realizaremos predicións para as observacións da ventá temporal 04/03/2015-31/03/2015 e comparemos entón os resultados proporcionados co acontecido na realidade. Como xa expuxemos na Subsección 3.1.2 é habitual considerar a covariable que denominamos "día de observación", que neste caso é a secuencia $1, 2, \cdots, 1158$, xa que a introdución desta variable no modelo, mediante unha función parcial suave non paramétrica, permite captar tendencia e estacionalidade. Ademais desta variable é importante introducir unha variable categórica para distinguir os días da semana e modelar así a compoñente estacional semanal. Por outra parte é necesario considerar no modelo unha variable que capte o comportamento que a serie temporal presenta os días 21 de cada mes, polo cal consideramos unha variable categórica que diferenza os días 21 dos restantes días.

Denotamos por Y a variable resposta que é o "número de baixas diarias de teléfono móbil", por t a variable "día de observación", por $diasem$ a variable categórica referente ao día da semana, e por $efecto21$ a variable categórica referente aos días 21. Con esta notación o modelo GAM que axustamos a serie temporal considerando as observacións do período 01/01/2012-03/03/2015 formúlase como segue

$$\ln(\mu(\mathbf{X})) = \beta_0 + f_1(t) + \beta_s diasem + \beta_{21} efecto21, \tag{3.6}$$

onde $\beta_{21} efecto21$ se define de xeito análogo ao termo $\beta_s diasem$ tomando como categoría de referencia

os días do mes que non son día 21.

$$\beta_{21} efecto21 = \begin{cases} 0 & se\ efecto21 = \text{“Non é día 21”} \\ \beta_{21,1} & se\ efecto21 = \text{“Día 21”} \end{cases}.$$

Se levamos a cabo o axuste do modelo (3.6), e calculamos os valores axustados obtemos como resultado o gráfico presente na Figura 3.23.

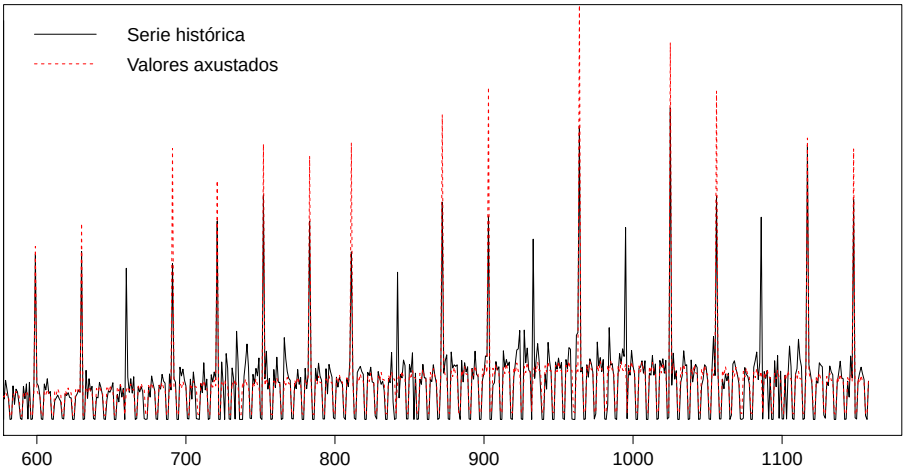


Figura 3.23: Gráfico dos valores axustados polo modelo (3.6) xunto aos valores reais da serie temporal.

Se nos fixamos no comportamento dos valores axustados podemos ver que en xeral se modelizan correctamente os repuntes referentes aos días 21 salvo en algúns casos. Se indagamos neste comportamento chegamos a que eses días nos cales o repunte do día 21 non se modeliza son os días 21 que son luns ou domingo, o motivo disto é o seguinte: os domingos e luns a serie temporal soe tomar o valor nulo polo cal cando estes días da semana coinciden co día 21 do mes o modelo indica que por ser domingo ou luns o valor axustado debe ser practicamente cero, sen ter en conta que o repunte do día 21 se leva a cabo incluso nestes días. Por este motivo e co obxectivo de indicar ao modelo que en que sexa domingo ou luns se estes son día 21 debe modelizar o pertinente repunte, engadimos ao modelo dúas novas variables categóricas que permiten distinguir os domingos día 21 e luns día 21 dos restantes días, $efecto21D$ e $efecto21L$, respectivamente. Polo que o modelo a axustar é

$$\ln(\mu(\mathbf{X})) = \beta_0 + f_1(t) + \beta_s diasem + \beta_{21} efecto21 + \beta_{21d} efecto21D + \beta_{21l} efecto21L, \tag{3.7}$$

onde $\beta_{21d} efecto21D$ e $\beta_{21l} efecto21L$ se definen de xeito análogo ao termo $\beta_s diasem$ tomando como categoría de referencia os días do mes que non son día 21 e domingo, e os días do mes que non son día 21 e luns, respectivamente.

$$\beta_{21d} efecto21D = \begin{cases} 0 & se\ efecto21D = \text{“Non é día 21 e domingo”} \\ \beta_{21,1d} & se\ efecto21D = \text{“Día 21 e domingo”} \end{cases}.$$

$$\beta_{21l} efecto21L = \begin{cases} 0 & se\ efecto21L = \text{“Non é día 21 e luns”} \\ \beta_{21,1l} & se\ efecto21L = \text{“Día 21 e luns”} \end{cases} .$$

Se axustamos o modelo (3.7) obtemos os resultados que se mostran no Cadro 3.17. En dito cadro vemos que todos os coeficientes son significativos incluíndo o termo suave agás o coeficiente β_0 o cal non é significativamente distinto de cero, polo cal os luns que non son día 21 o número de baixas de teléfono móbil é nulo. Con respecto aos domingos que non son día 21 podemos afirmar que o número de baixas indicado polo coeficiente asociado é próximo a cero. Tamén observamos que se é día 21 sen ser domingo ou luns o coeficiente asociado indica unha subida considerable nas baixas, a cal se duplica cando falamos de día 21 coincidente en domingo ou luns pois o coeficiente asociado a estes días debe compensar que o coeficiente asociado aos domingos é próximo a cero, e os luns é nulo.

Coeficientes paramétricos

	Estimación	**Sd erro**	**z valor**	**Pr**$(>\|z\|)$
β_0	-0.0702	0.0769	-0.912	0.3617
β_{martes}	4.3131	0.0774	55.741	$< 2\times10^{-16}$ ***
$\beta_{mércores}$	4.1923	0.0774	54.140	$< 2\times10^{-16}$ ***
β_{xoves}	4.3879	0.0774	56.727	$< 2\times10^{-16}$ ***
β_{venres}	4.2363	0.0774	54.728	$< 2\times10^{-16}$ ***
$\beta_{sábado}$	4.2264	0.0774	54.597	$< 2\times10^{-16}$ ***
$\beta_{domingo}$	0.2821	0.1019	2.769	0.00563**
$\beta_{21,1}$	1.7021	0.0103	165.324	$< 2\times10^{-16}$ ***
$\beta_{21,1d}$	3.8484	0.0715	53.809	$< 2\times10^{-16}$ ***
$\beta_{21,1l}$	3.9829	0.0815	48.876	$< 2\times10^{-16}$ ***

Significación dos termos suaves

	edf	**Ref.df**	**Chi.sq**	**p-valor**
$f_1(t)$	8.858	8.993	15525	$< 2\times10^{-16}$ ***

Cadro 3.17: Resultados do axuste do modelo (3.7).

Tras destacar algúns aspectos dos resultados proporcionados polo axuste do modelo continúamos predicindo os valores da ventá temporal 22/04/2015-19/05/2015 seguindo o modelo axustado. Na Figura 3.24 podemos ver estas predicións xunto cos valores reais da serie temporal. Ademais desta comparativa gráfica podemos calcular as medidas cuantitativas de comparación dos valores reais e das predicións obtidas, obtendo que os valores de MASE e MSE nesta ventá son 27.7483 % e 1062.72, respectivamente.

Na Figura 3.24 observamos que as predicións realizados polo modelo (3.7) asemellan ser adecuadas, ademais o valor do MASE indica que o modelo GAM axustado é un 72.2517 % mellor que o modelo que predí cada observación futura mediante o valor inmediatamente anterior da serie temporal. Posto que o comportamento do modelo nunha única ventá non é determinante aplicamos o procedemento exposto na Sección 2.5 empregando as 9 ventás temporais definidas para este fin obtendo os resultados que se amosan no Cadro 3.18.

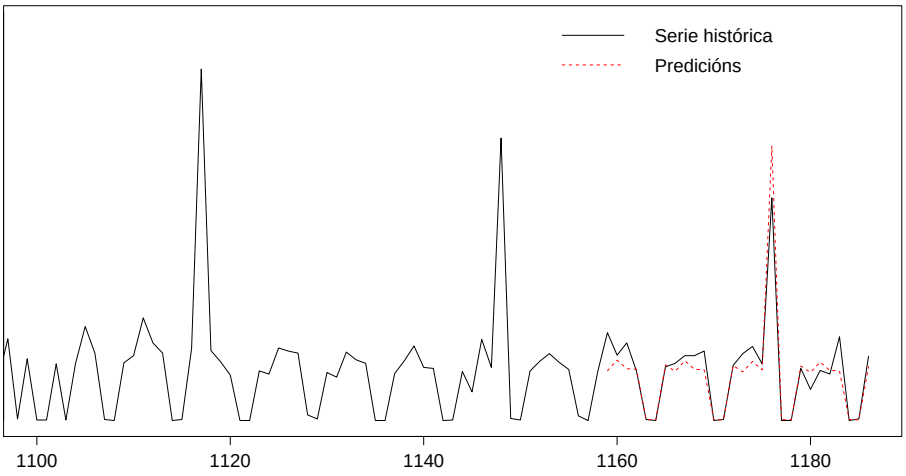


Figura 3.24: Predicións realizadas polo modelo (3.7) na ventá temporal 04/03/2015-31/03/2015 xunto aos valores reais da serie temporal.

Os resultados globais indican que o comportamento xeral do modelo é mellor incluso que na ventá de predición concreta que vimos de expor. A pesar de que os resultados globais son satisfactorios pensamos na posibilidade de introducir algunha covariable extra no modelo (3.7), concretamente pensamos na introdución dunha variable categórica que permita distinguir entre os meses do ano, e comprobamos que os resultados obtidos en termos de medidas globais do erro de predición non apoian a introdución desta covariable, quedándonos pois co modelo (3.7).

3.2.3. Árbores de Divisións Recursivas

Comezaremos por considerar as observacións que abarcan o período 01/01/2012-03/03/2015, e axustaremos unha Árbore de Divisións Recursivas proporcionando predicións para a ventá temporal 04/03/2015-31/03/2015. Como xa mencionamos na modelización da serie temporal mediante Árbores de Divisións Recursivas é frecuente considerar como variables explicativas o valor da serie temporal en instantes pasados. Concretamente decidimos introducir como variable explicativa o valor da serie temporal 7 días atrás, $bm7$ (pois vimos anteriormente que existe unha compoñente estacional semanal), ademais dunha variable que recolle o valor da serie temporal o día anterior ao que desexamos predicir, $bm1$. Por suposto a estas variables debemos engadir a variable categórica que distingue o día 21 de cada mes dos restantes días. Se axustamos unha Árbore de Divisións Recursivas coas variables mencionadas obtemos o resultado que se amosa de contado.

Media		Mediana		Desviación típica	
MASE	MSE	MASE	MSE	MASE	MSE
22.8691	662.1906	19.4784	552.447	6.9158	406.3794

Ventá temporal	MASE	MSE
Ventá 1	26.5781	1588.5247
Ventá 2	31.9604	354.6992
Ventá 3	18.2313	369.7148
Ventá 4	14.9084	263.2605
Ventá 5	18.2996	552.4470
Ventá 6	16.6647	526.9537
Ventá 7	19.4784	569.5676
Ventá 8	33.9116	847.3964
Ventá 9	25.7894	887.1518

Cadro 3.18: Resultados das medidas de erro de predición para o modelo (3.7).

```
Árbore 1
1) efecto21 == {1}; criterion = 1, statistic = 596.807
  2) bm7 <= 77; criterion = 1, statistic = 17.312
    3) bm1 <= 59; criterion = 0.979, statistic = 7.239
      4)*  weights = 19
    3) bm1 > 59
      5)*  weights = 7
  2) bm7 > 77
    6)*  weights = 12
1) efecto21 == {0}
  7) bm7 <= 72; criterion = 1, statistic = 324.348
    8) bm7 <= 11; criterion = 1, statistic = 296.932
      9) bm7 <= 2; criterion = 0.984, statistic = 7.721
        10)*  weights = 304
      9) bm7 > 2
        11)*  weights = 30
    8) bm7 > 11
      12) bm7 <= 37; criterion = 1, statistic = 33.412
        13)*  weights = 166
      12) bm7 > 37
        14)*  weights = 242
  7) bm7 > 72
    15) bm1 <= 115; criterion = 0.972, statistic = 6.731
      16) bm7 <= 190; criterion = 0.962, statistic = 6.171
        17) bm7 <= 85; criterion = 1, statistic = 31.737
          18)*  weights = 33
        17) bm7 > 85
          19) bm7 <= 119; criterion = 0.994, statistic = 9.459
            20) bm1 <= 11; criterion = 0.988, statistic = 8.346
              21)*  weights = 36
            20) bm1 > 11
              22)*  weights = 48
```

```
      19) bm7 > 119
        23)*  weights = 76
    16) bm7 > 190
      24)*  weights = 34
  15) bm1 > 115
    25)*  weights = 151
```

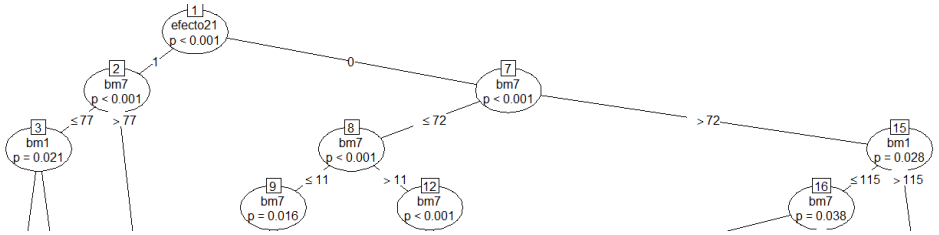


Figura 3.25: Gráfico dos primeiros nodos da Árbore de Divisións Recursivas 1 axustada.

Dito isto, se interpretamos a árbore mostrada podemos coñecer ás divisións realizadas en base as covariables consideradas, é dicir, podemos coñecer en cada nodo que covariable foi seleccionada xunto os subconxuntos disxuntos que dividiron o espazo de definición da covariable dando lugar a dous novos nodos. Por exemplo, no nodo 1 observamos que seleccionamos a variable $efecto21$ como a covariable con maior asociación coa resposta, e que dividimos o espazo de definición desta covariable considerando por unha parte as observacións iguais a 1 (días 21) e por outra as iguais a 0 (días que non son 21) dando lugar a dous novos nodos, que seguindo a notación empregada na construción da árbore son os nodos 2 e 7. Proseguimos fixándonos agora no nodo 2 onde vemos que a variable con maior correlación coa resposta é a variable $bm7$, e que dividimos o seu espazo de definición considerando as observacións menores e iguais a 77 por unha parte, e por outra as maiores que este valor, dando lugar aos nodos 3 e 6. Seguindo deste xeito podemos comentar cada un dos pasos levados a cabo polo algoritmo de construción da árbore de divisións recursivas.

Unha vez construída a árbore realízamos predicións para valores futuros. O xeito de calcular estas predicións recordemos que consiste en ver a que nodo terminal se corresponden os valores das covariables asociados ao instante que desexamos predicir, e seguidamente predicir segundo a media dos valores da resposta cuxos valores das variables explicativas se encontran en dito nodo. Antes de amosar os valores preditos pola árbore axustada na ventá temporal considerada debemos insistir nun matiz en canto ás variables explicativas, xa exposto no estudo da serie "número de usuarios diarios do servizo de vídeo baixo demanda". Para predicir cada valor da ventá temporal necesitamos coñecer os valores das covariables para o valor da resposta que desexo predicir como non dispomos de tales valores solucionamos este inconveniente mediante o procedemento exposto na análise da serie "número de usuarios diarios do servizo de vídeo baixo demanda". Na Figura 3.27 observamos as predicións realizadas para esta ventá temporal, con medidas de erro asociadas MASE=38.4529 % e MSE=1456.89.

Como se pode ver o comportamento da árbore non é o esperado, pois vemos por exemplo que os domingos e luns que non son día 21 a árbore predí por encima do valor real. Tamén observamos que predí constante en algunhas zonas polo cal decidimos substituír a variable $bm7$ pola variable categórica $diasem$ esperando que esta permita captar mellor a compoñente estacional semanal.

```
Árbore 2

1) efecto21 == {1}; criterion = 1, statistic = 596.807
  2) bm1 <= 85; criterion = 0.998, statistic = 11.921
    3)*  weights = 27
  2) bm1 > 85
    4)*  weights = 11
1) efecto21 == {0}
```

```
5) diasem == {2, 3, 4, 5, 6}; criterion = 1, statistic = 483.059
  6) bm1 <= 73; criterion = 1, statistic = 82.802
    7) diasem == {2}; criterion = 1, statistic = 78.053
      8)*  weights = 156
    7) diasem == {3, 4, 5, 6}
      9)*  weights = 345
  6) bm1 > 73
    10) diasem == {4}; criterion = 1, statistic = 37.557
      11)*  weights = 66
    10) diasem == {2, 3, 5, 6}
      12)*  weights = 231
5) diasem == {1, 7}
  13)*  weights = 322
```

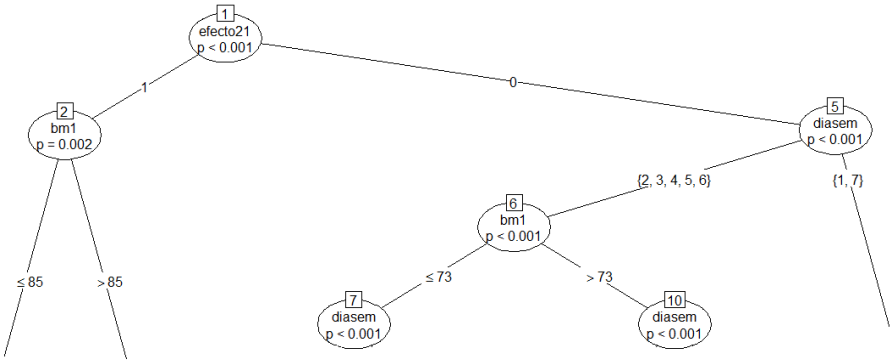


Figura 3.26: Gráfico dos primeiros nodos da Árbore de Divisións Recursivas 2 axustada.

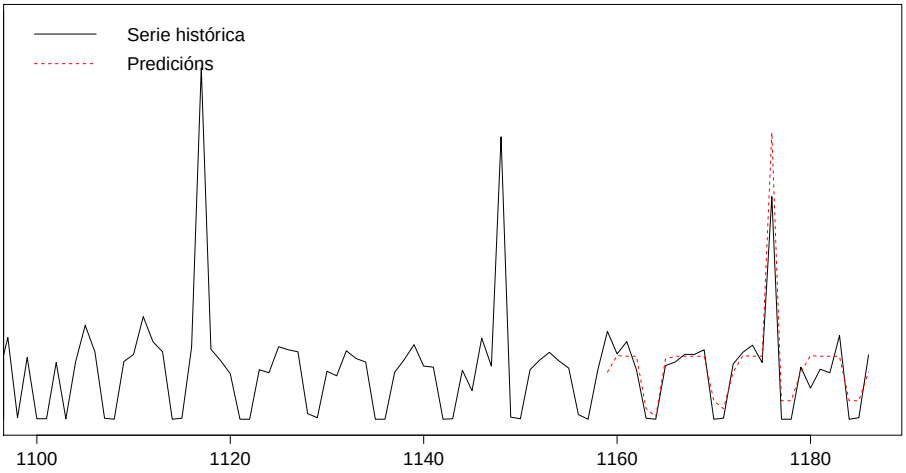


Figura 3.27: Predicións realizadas pola árbore 1 na ventá temporal 04/03/2015-31/03/2015 xunto aos valores reais da serie temporal.

Se empregamos a árbore exposta para realizar predicións no período 04/03/2015-31/03/2015 obtemos os resultados expostos na Figura 3.28 xunto os valores MASE=31.9850 % e MSE=1219.707, os cales son inferiores aos obtidos coa árbore 1.

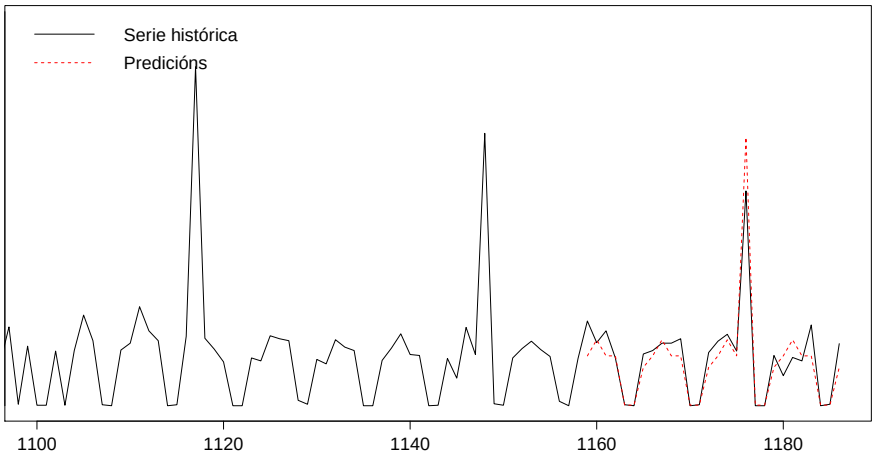


Figura 3.28: Predicións realizadas pola árbore 2 na ventá temporal 04/03/2015-31/03/2015 xunto aos valores reais da serie temporal.

Como se pode ver o cambio da variable *bm*7 pola variable *diasem* parece ter un efecto positivo sobre o modelado arranxando os inconvenientes expostos. Agora ben, como non debemos tomar decisións observando unicamente o comportamento da árbore nunha ventá, calculamos as medidas globais de erro de predición facendo uso das ventás definidas para este fin, no Cadro 3.19 presentamos os resultados.

En base aos resultados obtidos podemos afirmar que a variable *diasem* modela a compoñente estacional mellor que a variable *bm*7 sendo polo tanto a árbore 2 a seleccionada. Chegados a este punto comprobamos se a consideración dunha covariable que distinga os distintos meses pode ter un efecto positivo sobre a predición, e obtemos que a introdución desta covariable non é beneficiosa, sendo isto coherente coa conclusión obtida na aplicación da metodoloxía GAM.

3.2.4. Métodos de predición simples

Nesta sección imos modelar e predicir a serie "número de baixas diarias de teléfono móbil" mediante os métodos de predición simples (véxase Subsección 2.4.2) co obxectivo de comparar os resultados cos proporcionados polas metodoloxías Box-Jenkins, GAM e polas Árbores de Divisións Recursivas.

Analogamente ao realizado para as outras metodoloxías de predición axustamos o modelo empregando unicamente as observacións do período 01/01/2012-03/03/2015, e logo predicimos con este os valores da ventá temporal 04/03/2015-31/03/2015. Na Figura 3.29 podemos ver as predicións obtidas mediante os diferentes métodos simples na ventá considerada xunto aos valores reais da serie temporal. Esta comparativa gráfica compleméntase cos valores das medidas MASE e MSE que se encontran no Cadro 3.20.

Na Figura 3.29 observamos que o método da media capta dun xeito aceptable o nivel da serie temporal grazas a ausencia de tendencia. Con respecto aos métodos Naive e Drift dicir que nesta ventá concretamente non captan correctamente o nivel da serie temporal, ademais polo xeito de predicir en ambos métodos o resultado obtido en outras ventás pode ser considerablemente peor. Por exemplo, se último valor observado é nulo, o método Naive predí constante igual a cero, o cal produce maiores

errores de predición que os producidos nesta ventá concreta. Neste tipo de situacións o método Drift tamén predí dun xeito moi desatinado. Con respecto ao método Naive estacional podemos afirmar que nesta ventá temporal capta adecuadamente a compoñente semanal. O principal inconveniente deste método encóntrase na predición dos repuntes do día 21 que non son modelados. No Cadro 3.20 temos os valores das medidas MASE e MSE nesta ventá, o cal nos permite acompañar as conclusións obtidas a partires do gráfico secuencial coas proporcionadas polos valores destas medidas. Simplemente destacar que o valor do MASE para o método Naive estacional corrobora o que vimos de expor pois indica que este método é un 61.2729 % mellor que o método que predí cada valor da serie temporal mediante o valor da mesma no instante anterior. Para os restantes métodos simples esta melloría é considerablemente menor, concretamente o método da media ten un valor do MASE próximo ao 100 % o que indicaría que os dous métodos son equivalentes. A modelización da serie temporal composta pola observacións do período 01/01/2012-03/03/2015, e a predición dos valores da ventá temporal 04/03/2015-31/03/2015 sérvenos para extraer algunhas conclusións a cerca do comportamento dos métodos aplicados, agora ben non é suficiente pois o comportamento dunha ventá non nos permite extraer conclusións xerais. Por este motivo aplicamos o procedemento exposto na Sección 2.5 empregando as 9 ventás temporais definidas para este fin, e obtemos os resultados que amosamos no Cadro 3.21.

	Media		Mediana		Desviación típica	
Árbore	MASE	MSE	MASE	MSE	MASE	MSE
Árbore 1	35.7304	1035.901	31.4164	756.7353	11.0528	440.9450
Árbore 2	32.7817	1694.811	29.2182	912.9714	23.2102	1748.229

	Árbore 1		Árbore 2	
Ventá temporal	MASE	MSE	MASE	MSE
Ventá 1	31.4164	1208.995	30.9731	4944.7107
Ventá 2	51.9820	756.7353	86.3076	2852.9047
Ventá 3	29.2081	743.9089	49.6949	3887.9234
Ventá 4	29.2555	717.4739	12.2649	199.3743
Ventá 5	31.6066	1261.9604	16.2649	453.3992
Ventá 6	28.0610	719.2013	16.2153	463.0218
Ventá 7	27.0756	688.1990	19.1928	525.3211
Ventá 8	57.1637	2014.4562	34.5740	912.9714
Ventá 9	35.8042	1212.1751	29.2182	1013.6736

Cadro 3.19: Resultados das medidas de erro de predición para as dúas Árbores de Divisións Recursivas axustadas.

Se observamos o Cadro 3.21 cabe destacar a diferenza entre os valores da media e da mediana tanto na medida MASE como na medida MSE nos métodos Naive e Drift, debido á existencia de ventás nas que as medidas de erro toman valores elevados, polo motivo exposto con anterioridade. Esta variabilidade nas medidas de erro nas diferentes ventás tamén se reflexa nos valores da desviación típica. Ademais sobre estes métodos cabe destacar que segundo o valor da media MASE son peores

en termos de erro de predición que o procedemento que predí cada valor da serie temporal co valor da mesma no instante anterior, isto non sucede nin no método Naive estacional nin no método da Media. Para os dous métodos que acabamos de mencionar existe tamén unha diferenza entre a media e mediana o que nos leva a afirmar que existe variabilidade nos valores das medidas de erro, pero inferior á existente para os métodos Naive e Drift.

	Método da Media	Método Naive	Método Naive Estacional	Método Drift
MASE	94.7326	80.2761	38.7271	80.0444
MSE	8641.7370	7593.0710	4629.50	7600.5880

Cadro 3.20: Valores de MASE e MSE para cada un dos métodos simples na ventá temporal 04/03/2015-31/03/2015.

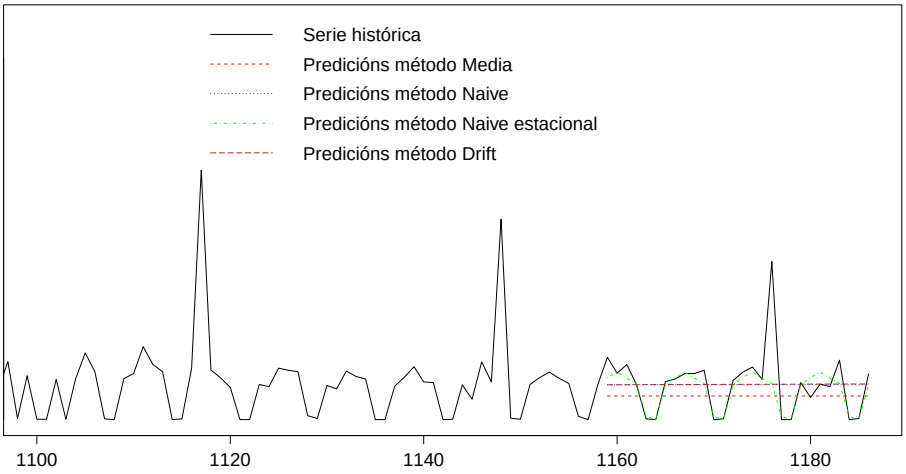


Figura 3.29: Gráfico secuencial dos valores reais da serie temporal xunto aos valores preditos por cada un dos métodos simples na ventá temporal 04/03/2015-31/03/2015.

3.2.5. Conclusións

Tras a aplicación das diferentes metodoloxías imos comparar os resultados obtidos co obxectivo de seleccionar finalmente a metodoloxía que a empresa debe empregar na predición de valores futuros da serie temporal "número de baixas diarias de teléfono móbil". No Cadro 3.22 recollemos os resultados das medidas globais de erro para cada unha das metodoloxías aplicadas.

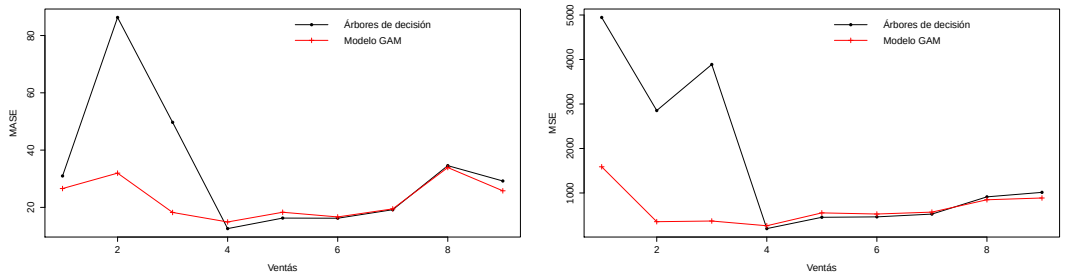
Para a serie temporal "número de baixas diarias de teléfono móbil" non existe dúbida entorno á metodoloxía a seleccionar pois en todas as medidas de erro mostradas no Cadro 3.22 o menor valor é o asociado ao modelo GAM. A pesar disto na Figura 3.30 amosamos dous gráficos nos cales se mostran os valores de MASE e MSE en cada unha das ventás de predición para a metodoloxía GAM

e as Árbores de Divisións Recursivas co obxectivo de comprobar que efectivamente a decisión tomada é acertada. Tamén cabe destacar que os métodos simples están moi lonxe de acadar os valores de erro de predición obtidos na predición empregando a metodoloxía GAM.

	Media		Mediana		Desviación típica	
Metodoloxía	MASE	MSE	MASE	MSE	MASE	MSE
Media	96.2268	10603.73	90.7539	11538.58	15.8764	3784.8
Naive	181.7534	35071.31	90.3511	11144.71	254.0977	67207.26
Naive estacional	87.0803	21880.73	55.1471	12253.11	66.8453	20305.09
Drift	184.0239	35814.25	91.6560	11186.7	258.1873	69232.09

	Media		Naive		Naive estacional		Drift	
Ventá temporal	MASE	MSE	MASE	MSE	MASE	MSE	MASE	MSE
Ventá 1	94.1188	17059.466	90.3511	16344.107	55.1471	12253.107	91.6560	16481.825
Ventá 2	114.1525	3092.037	132.1750	6135.170	226.5519	52248.250	134.9997	6338.645
Ventá 3	84.7461	11538.580	72.5898	10535.321	136.9053	59457.393	72.7889	10524.846
Ventá 4	84.3006	11336.277	112.7424	17425.321	35.5813	6634.964	112.6722	17402.099
Ventá 5	86.4844	11883.264	78.2825	11144.714	37.0812	6881.429	78.9040	11186.701
Ventá 6	85.8567	11845.261	77.0642	11047.821	33.0899	7301.036	77.9830	11118.797
Ventá 7	90.7539	11956.854	125.8198	20494.357	36.4667	7703.571	125.8198	20494.357
Ventá 8	130.6182	8447.680	856.8810	213879.036	133.5320	22600.607	870.0382	220040.9
Ventá 9	95.0099	8274.140	89.8747	8635.964	89.3666	21846.179	91.3539	8740.124

Cadro 3.21: Valores de MASE e MSE para cada un dos métodos simples na ventá temporal 04/03/2015-31/03/2015.



(a) Valores da medida MASE.

(b) Valores da medida MSE.

Figura 3.30: Gráficos dos valores de MASE e MSE para o modelo GAM e a Árbore de Divisións Recursivas.

	Media		Mediana		Desviación típica	
Metodoloxía	MASE	MSE	MASE	MSE	MASE	MSE
Box-Jenkins	79.1021	9263.605	69.461	9654.952	19.0252	2975.701
GAM	22.8691	662.1906	19.4784	552.447	6.9158	406.3794
Árbores	32.7817	1694.811	29.2182	912.9714	23.2102	1748.229
Media	96.2268	10603.73	90.7539	11538.58	15.8764	3784.8
Naive	181.7534	35071.31	90.3511	11144.71	254.0977	67207.26
Naive estacional	87.0803	21880.73	55.1471	12253.11	66.8453	20305.09
Drift	184.0239	35814.25	91.6560	11186.7	258.1873	69232.09

Cadro 3.22: Valores das medidas globais de erro para cada unha das metodoloxías aplicadas.

Efectivamente, os gráficos da Figura 3.30 indican que a metodoloxía GAM é a correcta para realizar predicións da serie temporal pois a árbore ten un comportamento axeitado menos nas primeiras ventás temporais nas que se disparan os valores de erro. Se investigamos que sucede nestas tres ventás chegamos á existencia de problemas á hora de captar a compoñente semanal ademais do comportamento dos días 21.

Consecuentemente, a metodoloxía GAM é a seleccionada para modelar a serie temporal "número de baixas diarias de teléfono móbil". Cabe mencionar que o seu tempo de computación é 0.19 s.

3.2.6. Outras aplicacións

Para rematar o estudo desta serie expomos unha das ideas que xurdiron co obxectivo de tentar mellorar os resultados obtidos coas metodoloxías que vimos de aplicar. Na serie "número de baixas diarias de teléfono móbil" vemos que os días 21 de cada mes teñen un comportamento distinto aos restantes días, polo cal decidimos construír dúas series temporais a partires da serie temporal orixinal, unha serie formada unicamente polas observación dos días 21, e outra serie coas restantes observacións.

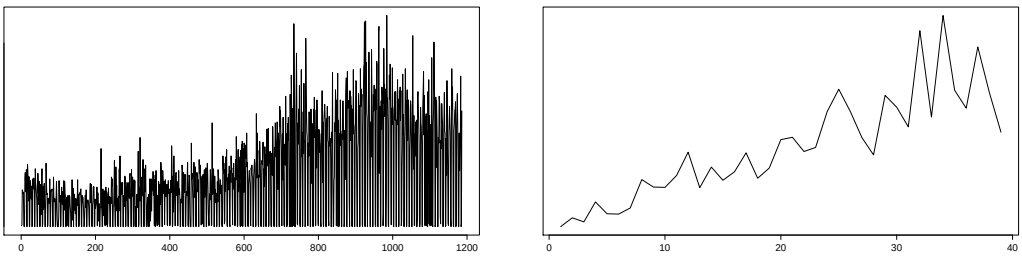


Figura 3.31: Gráficos secuenciais das dúas series temporais construídas a partires da serie orixinal. Á dereita a serie formada unicamente polas observacións dos días 21. Á esquerda serie formada por todas as observacións da serie orixinal agás as correspondentes aos días 21.

O modelado da serie formada unicamente polas observacións dos días 21 é moi simple pois esta

presenta soamente unha tendencia crecente. Pola contra, o modelado da serie formada polas restantes observacións segue a ser complexo, pois ao extraer os días 21 da serie orixinal o que sucede na nova serie é que se distorsiona o comportamento semanal debido a que pasamos do día 20 de cada mes ao día 22 sen pasar polo día 21, e consecuentemente saltámonos un día da semana, por exemplo pasamos dun luns a un mércores, e isto da lugar a problemas de modelado empregando calquera das metodoloxías.

3.3. Estudo da serie número de altas diarias de teléfono móbil

Nesta sección abordamos o estudo da serie temporal "número de altas diarias de teléfono móbil" que comeza o 01/01/2013 e remata o 24/04/2015. Consecuentemente dispoñemos de 844 observacións para o presente estudo. A predición desta serie temporal permítelle á empresa de telecomunicacións adiantarse aos feitos, por exemplo, ofertando promocións a clientes potenciais se o número de altas non é o esperado, é dicir, a predición desta serie temporal é útil na mercadotecnia de servizos.

Na Figura 3.32 encóntrase o gráfico secuencial da serie temporal. En dito gráfico observamos que o comportamento inicial da serie temporal discrepa do comportamento predominante na mesma. Neste período inicial temos maior variabilidade chegando a alcanzar o maior valor da variable medida. Isto pode deberse a que as observacións do inicio da serie temporal se corresponden cun período de adaptación a unha serie de cambios na empresa de telecomunicacións, e polo tanto o seu comportamento non concorda co comportamento xeral da serie temporal. Con respecto á existencia ou non de tendencia dicir que segundo o gráfico secuencial parece que a serie temporal non presenta tendencia. Analogamente ao exposto no estudo das series precedentes, é complexo detectar unha compoñente estacional no gráfico secuencial, polo cal máis adiante identificaremos esta compoñente en caso de existir mediante a *fas* mostral da serie temporal. Por outra parte observamos que a variable "número de altas diarias de teléfono móbil" toma o valor cero nunha cantidade considerable de días. Debemos ver a que se deben estes valores nulos, pois está claro que non son valores atípicos, se non que estes valores forman parte do comportamento da serie temporal, do mesmo xeito que na serie "número de baixas diarias de teléfono móbil". Se buscamos a que datas se corresponden estes valores nulos, observamos que son domingos, días nos que a compañía non leva a cabo os trámites referentes ás altas, e 3 días festivos nos cales tampouco se fixeron estes trámites. Polo cal debemos quedarnos co comportamento maioritario que é que os domingos a empresa de telecomunicacións non traballa na tramitación de altas.

Por outra parte observamos o mesmo comportamento referente aos días 21 de cada mes exposto na serie anterior, coa diferenza de que na serie que agora estudamos non todos os días 21 se produce o repunte, é dicir, se existe un repunte na serie temporal este correspóndese co día 21 dalgún mes pero non todos os días 21 de cada mes se produce un repunte como acontecía na serie de "baixas". Supoñemos que isto se debe á organización da empresa de telecomunicacións.

O comportamento inicial da serie temporal presenta diferenzas con respecto ao comportamento xeral da serie temporal polo motivo exposto con anterioridade, polo cal para limitar o efecto deste período na modelización suavizamos este comportamento mediante a aplicación da función `tsclean`. A serie resultante pode verse na Figura 3.33.

Analogamente ao realizado para o estudo das outras series temporais, expomos deseguido as ventás temporais seleccionadas para levar a cabo o procedemento exposto na Sección 2.5. Consideramos $m = 10$ ventás, cada unha delas con $m_N = 28$ observacións, e tomamos $T_1 = 771$ e $l = 5$, dando así lugar ás seguintes ventás temporais:

- Ventá 1: Dende 11/02/2015 ata 10/03/2015.
- Ventá 2: Dende 16/02/2015 ata 15/03/2015.
- Ventá 3: Dende 21/02/2015 ata 20/03/2015.
- Ventá 4: Dende 26/02/2015 ata 25/03/2015.
- Ventá 5: Dende 03/03/2015 ata 30/03/2015.

- Ventá 6: Dende 08/03/2015 ata 04/04/2015.
- Ventá 7: Dende 13/03/2015 ata 09/04/2015.
- Ventá 8: Dende 18/03/2015 ata 14/04/2015.
- Ventá 9: Dende 23/03/2015 ata 19/04/2015.
- Ventá 10: Dende 28/03/2015 ata 24/04/2015.

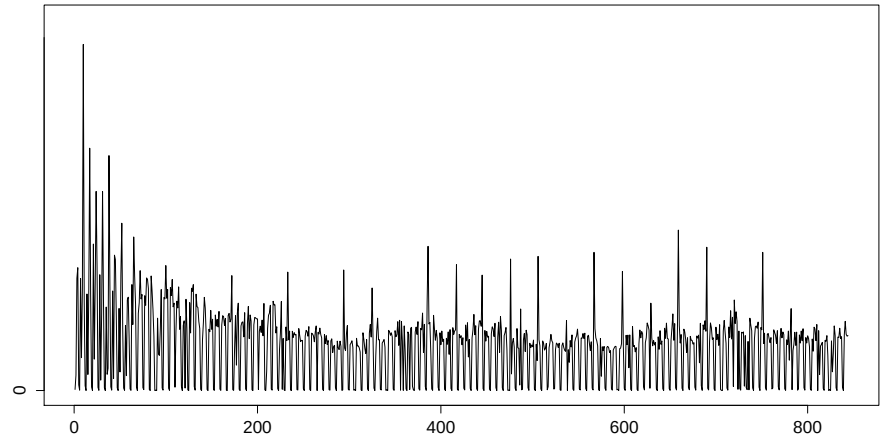


Figura 3.32: Gráfico secuencial da serie temporal "número de altas diarias de teléfono móbil".

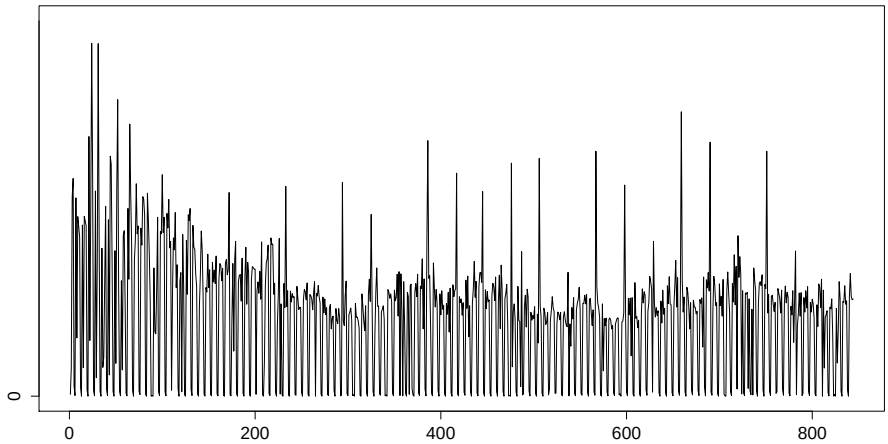


Figura 3.33: Gráfico secuencial da serie temporal "número de altas diarias de teléfono móbil" tras suavizar os atípicos presentes no período de adaptación ao novo funcionamento do sistema de altas.

Recordar unha vez máis que non empregaremos todas as ventás temporais para a realización do procedemento exposto na Sección 2.5, se non que empregamos para este procedemento as primeiras 9 ventás temporais e usamos a última delas para ilustrar con detalle a aplicación de cada unhas das metodoloxías.

3.3.1. Metodoloxía Box-Jenkins

Como vimos de expor, para ilustrar con detalle a aplicación de cada metodoloxía consideramos a serie temporal formada polas observacións do período 01/01/2013-27/03/2015 e axustamos un modelo a serie temporal empregando unicamente estas observacións, unha vez axustado o modelo empregámolo para predicir os valores da ventá temporal 28/03/2015-24/04/2015. Posto que dispomos dos valores reais observados en dita ventá usamos estes para realizar comparativas entre os mesmos e as predicións, comparativa gráfica e cuantitativa.

No gráfico secuencial (Figura 3.33) observamos que a variabilidade ao inicio da serie temporal é considerablemente maior que a presente no resto da serie, a pesar de ter suavizado este comportamento mediante a aplicación da función `tsclean`. Consecuentemente, para poder identificar un modelo Box-Jenkins correcto debemos solucionar previamente este inconveniente. Analogamente ao realizado nas outras dúas series temporais decidimos empregar a transformación Box-Cox co obxectivo de estabilizar dita variabilidade. Achamos o valor do parámetro λ da transformación Box-Cox, na serie temporal formada pola observacións do período 01/01/2013-27/03/2015, obtendo $\lambda = 0.0710$, o cal podemos asumir nulo, e polo tanto aplicar unha transformación logarítmica, a cal é máis sinxela. Agora ben, a serie temporal ten valores nulos, polo que non é posible aplicar directamente a transformación logarítmica, sendo necesario salvar este inconveniente do mesmo xeito que foi salvado na serie "número de baixas diarias de teléfono móbil". Cabe recordar que a solución consistiu en seleccionar unha constante fixa $\varepsilon \in \mathbb{R}^+$, e sumar dita constante a serie temporal obtendo deste xeito unha serie temporal cuxos valores son todos eles maiores que cero, e sobre a cal xa é posible aplicar sen máis unha transformación logarítmica. Efectivamente procedendo deste xeito obtemos unha serie temporal cuxa variabilidade é constante, podendo proceder á identificación dun modelo Box-Jenkins.

Un primeiro paso para identificar un modelo Box-Jenkins é debuxar e estudar a *fas* mostral da serie temporal. Na Figura 3.34 podemos ver o gráfico da *fas* mostral.

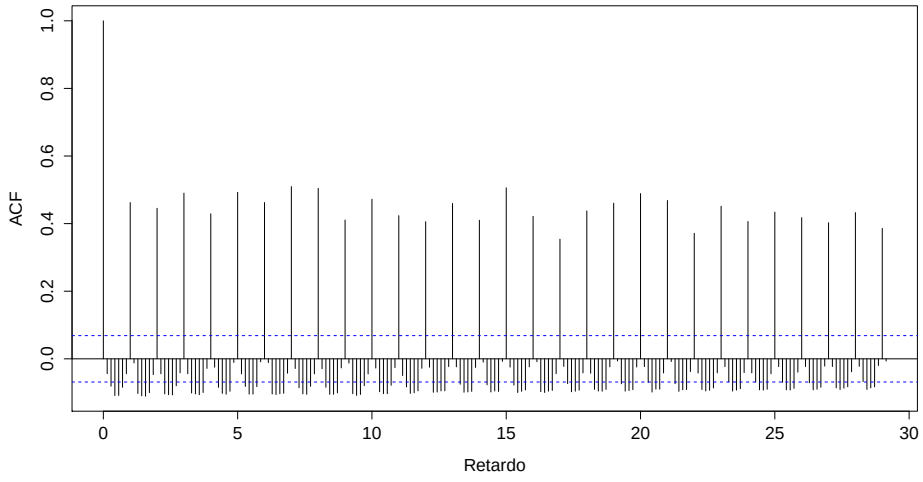
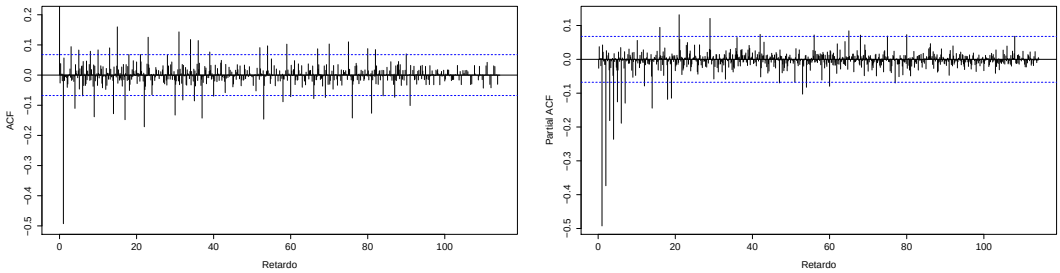


Figura 3.34: Gráfico das correlacións simples mostrais da serie temporal "número altas diarias de teléfono móbil" tras a aplicación da transformación logarítmica.

Na Figura 3.34 observamos forte correlación no retardo 7 e nos seus múltiplos, ademais de periodicidade de período 7 e converxencia lenta á cero, estas características indican a presenza dunha compoñente estacional de período 7 días, é dicir, compoñente estacional semanal, véxase Subsección 2.2.2.

Por este motivo aplicamos sobre á serie temporal unha diferenza estacional de período 7 días co obxectivo de eliminar a compoñente estacional detectada e así poder identificar un modelo Box-Jenkins. Na Figura 3.35 podemos ver as correlacións simples e parciais mostrais da serie diferenciada estacionalmente. A diferenciación estacional permitiu eliminar a compoñente estacional, pero as correlacións mostran unha forte dependencia estacional estando esta presente en retardos moi altos. Isto supón que nos encóntramos co mesmo problema que na serie "número de baixas diarias de teléfono móbil'" polo cal empregamos a función auto.arima do paquete forecast para axustar un modelo Box-Jenkins á serie temporal, e seguidamente calculamos as predicións realizadas polo modelo axustado obtendo o resultado que se pode observar na Figura 3.36.



(a) *Fas* mostral (b) *Fap* mostral

Figura 3.35: Gráfico de correlacións simples e parciais mostrais da serie temporal diferenciada estacionalmente.

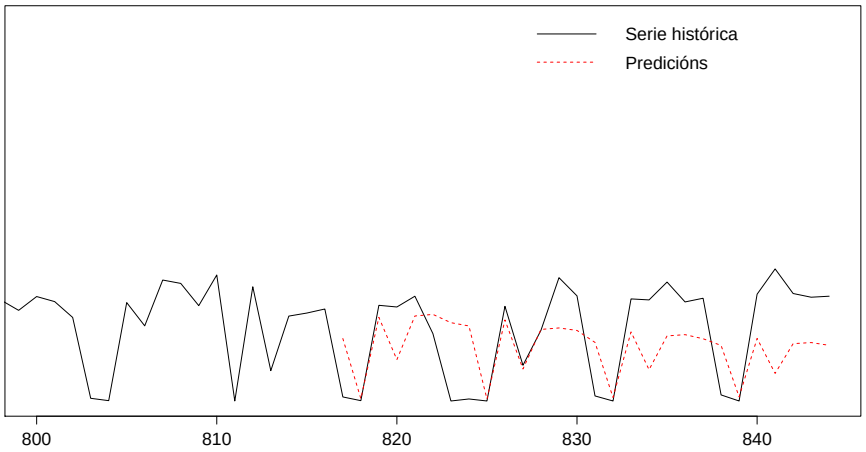


Figura 3.36: Predicións realizadas polo modelo Box-Jenkins axustado pola función auto.arima na ventá temporal 28/03/2015 ata 24/04/2015 xunto aos valores reais da serie temporal.

Como podemos observar as predicións realizadas polo modelo Box-Jenkins axustado pola función auto.arima son mellorables, pois o modelo non capta correctamente o comportamento semanal, xa que predí próximo a cero nos domingos pero non modela correctamente os sábados cuxo número de altas é baixo, isto xunto a que non sempre capta correctamente o nivel de altas nos restantes días da semana, produce que as medidas de erro tomen os valores MSE=18641.22 e MASE=102.3259 %. Nótese que este valor do MASE indica que o modelo Box-Jenkins considerado é menos axeitado que o modelo que predí cada valor futuro como o valor observado no instante inmediatamente anterior ao que desexamos predicir. Agora ben, para posteriormente poder comparar os resultados proporcionados por este modelo cos obtidos empregando outras metodoloxías aplicamos o procedemento exposto na Sección 2.5 empregando as 9 ventás temporais definidas para este fin obtendo os resultados que se amosan no Cadro 3.23.

Media		Mediana		Desviación típica	
MASE	MSE	MASE	MSE	MASE	MSE
66.3326	14763.34	57.8486	11002.87	34.1932	11336.41

Ventá temporal	MASE	MSE
Ventá 1	44.4537	8538.541
Ventá 2	48.2180	9159.873
Ventá 3	62.4178	13474.581
Ventá 4	59.7591	15968.688
Ventá 5	60.1848	9176.706
Ventá 6	52.9477	9533.481
Ventá 7	54.9949	11002.868
Ventá 8	57.8486	11708.544
Ventá 9	156.1688	44306.822

Cadro 3.23: Resultados das medidas de erro de predición para o modelo Box-Jenkins axustado pola función auto.arima.

Os resultados do Cadro 3.23 indican que o comportamento global do modelo é máis axeitado que o exposto na ventá 04/03/2015-31/03/2015, isto débese a que nesta ventá temporal o modelo predí erroneamente nos sábados, tal e como explicamos, mentres que noutras ventás o modelo logra captar correctamente este comportamento. Na ventá 9 vemos uns valores de erro considerablemente altos, isto débese a que o modelo nesta ventá capta mal o nivel da serie temporal.

3.3.2. Metodoloxía GAM

Tras a aplicación da metodoloxía Box-Jenkins proseguimos o estudo centrándonos agora na aplicación da metodoloxía GAM. Comezaremos por ilustrar con detalle o axuste dun modelo GAM empregando soamente as observacións do período 01/01/2013-27/03/2015, seguidamente empregaremos este modelo para predicir os valores da ventá temporal 27/03/2015-24/04/2015 e comparemos entón os resultados proporcionados co acontecido na realidade.

Analogamente ao exposto nas Subseccións 3.1.2 e 3.2.2 consideramos a covariable que denominamos "día de observación" que neste caso é a secuencia $1, 2, \cdots, 816$ xunto a variable categórica que indica os días da semana. Por outra parte é necesario considerar no modelo unha variable que capte o comportamento que a serie temporal presenta os días 21 de cada mes, polo cal consideramos unha variable categórica que diferenza os días 21 dos restantes días.

Denotamos por Y a variable resposta que é o "número de altas diarias de teléfono móbil", por t a variable "día de observación", por $diasem$ a variable categórica referente ao día da semana, e por $efecto21$ a variable categórica referente aos días 21. Con esta notación o modelo GAM que axustamos a serie temporal considerando unicamente as observacións do período 27/03/2015-24/04/2015 formúlase como segue

$$\ln(\mu(\mathbf{X})) = \beta_0 + f_1(t) + \beta_s diasem + \beta_{21} efecto21, \tag{3.8}$$

sendo os termos $\beta_s X_s$ e $\beta_{21} efecto21$ os definidos na Subsección 3.2.2.

Se levamos a cabo o axuste do modelo (3.8), e calculamos os valores axustados obtemos como resultado o gráfico presente na Figura 3.37.

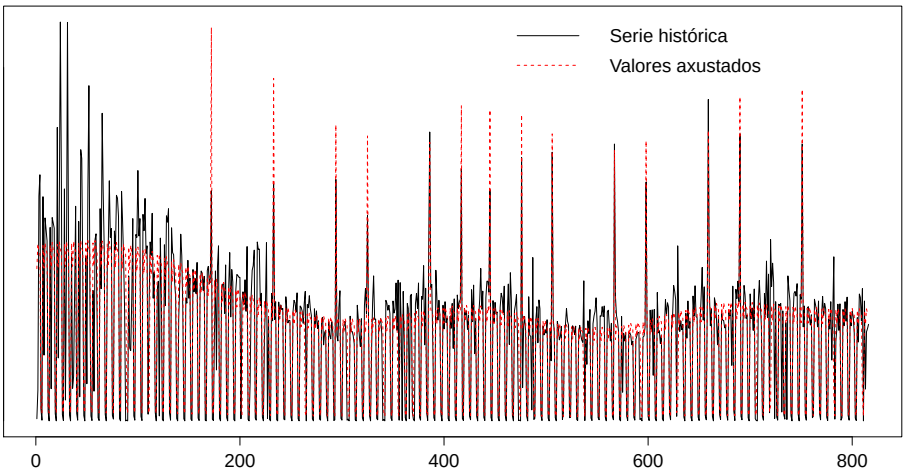


Figura 3.37: Gráfico dos valores axustados polo modelo (3.8) xunto aos valores reais da serie temporal.

Se nos fixamos no comportamento dos valores axustados podemos ver que en xeral se modelizan correctamente os repuntes referentes aos días 21 salvo en algúns casos, por exemplo cara o final da serie temporal vemos que non se modeliza nin o repunte do día 21 de xaneiro nin tampouco o día 21 de decembro, que se corresponden cun sábado e cun domingo, respectivamente. Se pescudamos acerca deste comportamento concluímos que todos os días nos cales o modelo non capta correctamente o repunte son sábados e domingos. O motivo coincide co que producía este mesmo problema no modelado das "baixas", os domingos a serie temporal soe tomar o valor nulo polo cal cando este día da semana coinciden co día 21 do mes o modelo indica que por ser domingo o valor axustado é practicamente cero, sen ter en conta que o repunte do día 21 se leva a cabo incluso en domingo. Con respecto aos sábados chegamos a ver tras un estudo en profundidade dos valores que acada a serie temporal neste días que o número de altas en sábado é moi reducido, máis concretamente a media da variable número de altas de teléfono móbil en sábado é 15.35 e a súa mediana 1, polo cal o modelo ten o mesmo inconveniente que ten nos domingos. Por este motivo é co obxectivo de indicar ao modelo que en que sexa sábado ou domingo se estes son día 21 debe modelizar o pertinente repunte, engadimos ao modelo dúas novas

variables categóricas que permiten distinguir os domingos día 21 e sábados día 21 dos restantes días, $efecto2D$ e $efecto21S$, respectivamente. Polo que o modelo a axustar é

$$\ln(\mu(\mathbf{X})) = \beta_0 + f_1(t) + \beta_s X_s + \beta_{21} efecto21 + \beta_{21s} efecto21S + \beta_{21d} efecto21D, \tag{3.9}$$

onde o termo $\beta_{21s} efecto21S$ se define de xeito análogo ao termo $\beta_{21d} efecto21D$, cuxa definición se encontra na Subsección 3.2.2, considerando como categoría de referencia os días do mes que non son día 21 e sábado.

Se axustamos o modelo (3.9) obtemos os resultados que se mostran no Cadro 3.24. Como se pode observar todos os coeficientes son significativos, polo cal procedemos a súa interpretación. Con respecto aos días da semana observamos que os sábados e domingos se produce unha baixa considerable no número de altas con respecto aos luns, cabe mencionar que esta diminución é maior os domingos que os sábados de acordo ao exposto con anterioridade. Os días 21 o modelo debe modelar o pertinente repunte e isto vese reflexado no modelo cun coeficiente asociado que indica un número de altas maior nestes días. Se ademais dito día coincide con sábado ou domingo vemos que o coeficiente asociado toma un valor considerablemente maior para compensar o feito de que nestes días da semana o modelo encóntrase nun número de altas moi reducido.

Coeficientes paramétricos

	Estimación	**Sd erro**	**z valor**	**Pr($>\|z\|$)**
β_0	5.8298	0.0050	1176.021	$< 2\times10^{-16}$ ***
β_{martes}	-0.1404	0.0072	-19.503	$< 2\times10^{-16}$ ***
$\beta_{mércores}$	0.0143	0.0069	2.071	0.0383 *
β_{xoves}	-0.0335	0.0070	-4.782	$< 1.74\times10^{-6}$ ***
$\beta_{viernes}$	-0.0408	0.0070	-5.847	$< 5.02\times10^{-9}$ ***
$\beta_{sábado}$	-2.9784	0.0229	-130.275	$< 2\times10^{-16}$ ***
$\beta_{domingo}$	-3.8627	0.0350	-110.369	$< 2\times10^{-16}$ ***
$\beta_{21,1}$	0.8727	0.0103	84.792	$< 2\times10^{-16}$ ***
$\beta_{21,1s}$	3.4544	0.0481	71.799	$< 2\times10^{-16}$ ***
$\beta_{21,1d}$	2.4680	0.0383	64.495	$< 2\times10^{-16}$ ***

Significación dos termos suaves

	edf	**Ref.df**	**Chi.sq**	**p-valor**
$f_1(t)$	8.92	8.998	9171	$< 2\times10^{-16}$ ***

Cadro 3.24: Resultados do axuste do modelo (3.9).

Tras destacar algúns aspectos dos resultados proporcionados polo axuste do modelo continúamos predicindo os valores da ventá temporal 22/04/2015-19/05/2015 seguindo o modelo axustado, na Figura 3.38 podemos ver estas predicións xunto cos valores reais da serie temporal. Ademais desta compa-

rativa gráfica calculamos o valor da medida MASE e da medida MSE obtendo 43.5084 % e 6155.391, respectivamente.

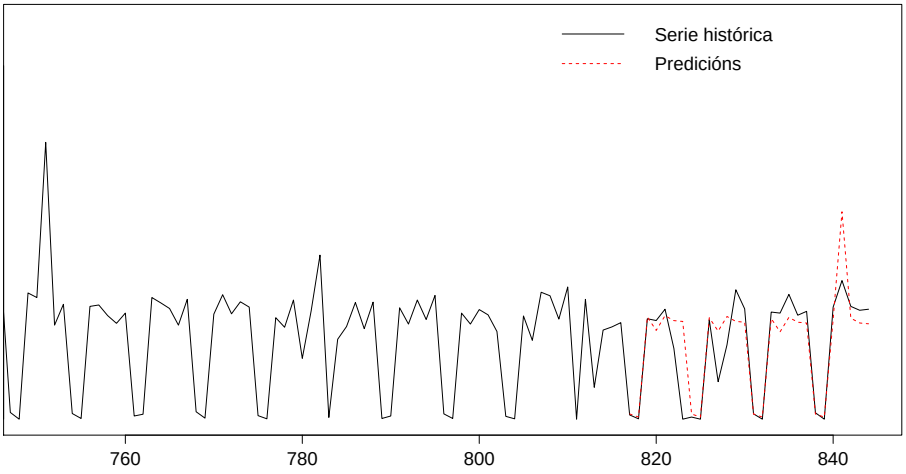


Figura 3.38: Predicións realizadas polo modelo (3.9) na ventá temporal 28/03/2015-24/04/2015 xunto aos valores reais da serie temporal.

Como se pode observar na Figura 3.38 as predicións realizados polo modelo (3.9) asemellan ser adecuadas salvo na predición do repunte do día 21 pois a predición sobrepasa o valor real da serie temporal. Isto débese a que os repuntes nesta serie non se producen todos os días 21 e ademais nalgúns meses prodúcense pero cunha magnitude inferior á doutros. A pesar deste detalle, o valor do MASE indica que o modelo GAM axustado é un 56.4916 % mellor que o modelo que predí cada observación futura mediante o valor inmediatamente anterior da serie temporal (pode consultarse a interpretación do MASE na Sección 2.5). Posto que o comportamento do modelo nunha única ventá non é determinante aplicamos o procedemento exposto na Sección 2.5 empregando as 9 ventás temporais definidas para este fin obtendo os resultados que se amosan no Cadro 3.25.

Os resultados globais indican que o comportamento xeral do modelo é considerablemente mellor que na ventá de predición concreta que vimos de expor, isto débese a que na ventá exposta o modelo se encontraba co inconveniente de modelar o repunte do día 21 e que debido a problemática existente con estes días estimou un efecto maior do mesmo co sucedido na realidade, algo que non sucede en todas as ventás. A pesar de que os resultados globais son satisfactorios pensamos na posibilidade de introducir algunha covariable extra no modelo (3.9), concretamente pensamos na introdución dunha variable categórica que permita distinguir entre os meses do ano, e comprobamos que os resultados obtidos en termos de medidas globais do erro de predición non apoian a introdución desta covariable, quedándonos pois co modelo (3.9), o mesmo que sucedeu no estudo da serie "número de baixas diarias de teléfono móbil".

3.3.3. Árbores de Divisións Recursivas

De novo comezaremos por considerar só as observacións que abarcan o período 01/01/2013-27/03/2015, e axustaremos unha árbore de divisións recursivas, só coas observacións mencionadas, para logo proporcionar predicións para a ventá temporal 28/03/2015-24/04/2015. Para a serie temporal en estudo decidimos introducir como variable explicativa o valor da serie temporal 7 días atrás, $am7$, pois vimos

anteriormente que existe unha compoñente estacional semanal, ademais dunha variable que recolle o valor da serie temporal o día anterior ao que desexamos predicir, $am1$. Por suposto a estas variables debemos engadir a variable categórica que distingue o día 21 de cada mes dos restantes días. Deseguido amosamos a árbore axustada con estas covariables.

Media		Mediana		Desviación típica	
MASE	MSE	MASE	MSE	MASE	MSE
26.9754	3710.477	21.3890	2103.439	9.1427	2464.655

Ventá temporal	MASE	MSE
Ventá 1	18.3476	1608.0956
Ventá 2	21.3890	1720.7501
Ventá 3	16.7712	901.2611
Ventá 4	20.3410	2103.4394
Ventá 5	20.6466	2040.2114
Ventá 6	32.1136	5489.5169
Ventá 7	38.2657	6925.1451
Ventá 8	36.5917	6381.6085
Ventá 9	38.3123	6224.2606

Cadro 3.25: Resultados das medidas de erro de predición para o modelo (3.9).

```
Árbore 1

1) am7 <= 68; criterion = 1, statistic = 409.656
  2)*  weights = 240
1) am7 > 68
  3) efecto21 == {1}; criterion = 1, statistic = 95.315
    4)*  weights = 14
  3) efecto21 == {0}
    5) am7 <= 467; criterion = 1, statistic = 101.055
      6) am7 <= 343; criterion = 0.999, statistic = 12.23
        7)*  weights = 337
      6) am7 > 343
        8)*  weights = 137
    5) am7 > 467
      9) am1 <= 426; criterion = 0.979, statistic = 7.257
        10)*  weights = 53
      9) am1 > 426
        11) am7 <= 651; criterion = 0.977, statistic = 7.105
          12)*  weights = 27
        11) am7 > 651
          13)*  weights = 8
```

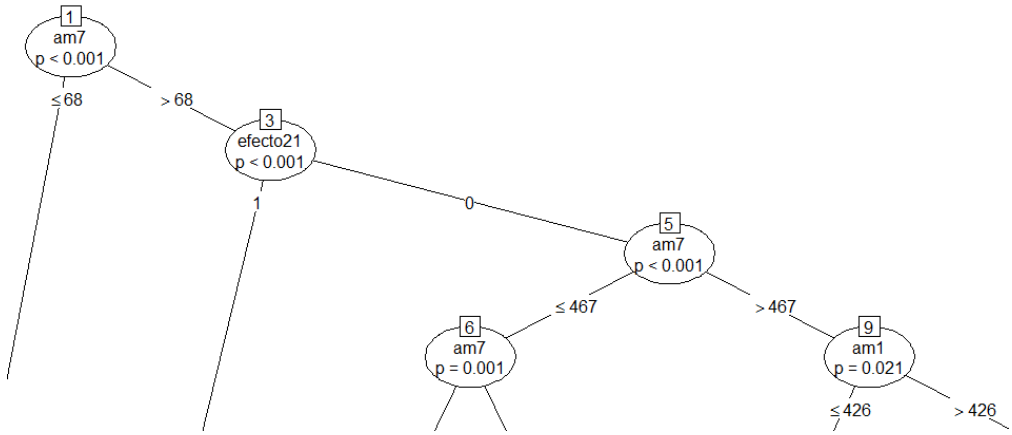


Figura 3.39: Gráfico dos primeiros nodos da Árbore de Divisións Recursivas 1 axustada.

A árbore exposta pódese interpretar tal é como indícamos no estudo das outras dúas series. Se empregamos a árbore exposta para realizar predicións no período 28/03/2015-24/04/2015 obtemos os resultados expostos na Figura 3.41 xunto aos valores MASE=71.285 % e MSE=16167.14. Tanto as medidas de erro como as predicións indican que a modelización pode ser altamente mellorable, pois observamos que o modelo predí un repunte do día 21 de maior magnitude que o acontecido na realidade, non predí correctamente os valores dos días do fin de semana, pois proporciona predicións sempre por encima dos valores reais da serie neste días, e para os restantes días da semana vemos que a veces predí constante durante varios días. A mala predición do días 21 débese a que a magnitude dos repuntes presentes neste días é moi variable a longo da serie, agora ben a modelización da compoñente semanal pódese tentar mellor substituíndo a covariable $am7$ polo covariable $diasem$. Antes de comprobar o funcionamento da árbore co cambio exposto, calculamos as medidas globais de erro de predición para a árbore coas covariables $am1$, $am7$ e $efecto21$, facendo uso das ventás definidas para este fin, co obxectivo de poder realizar unha comparativa entre as dúas árbores, no Cadro 3.26 presentamos os resultados.

Se axustamos a serie temporal composta polas observacións do período 01/01/2013-27/03/2015 unha árbore de divisións recursivas con covariables $diasem$, $am1$ e $efecto21$ obtemos o resultado que se amosa deseguido.

```
Árbore 2
1) diasem == {1, 2, 3, 4, 5}; criterion = 1, statistic = 455.767
  2) efecto21 == {1}; criterion = 1, statistic = 99.448
    3)*  weights = 14
  2) efecto21 == {0}
    4) am1 <= 424; criterion = 0.999, statistic = 13.853
      5) diasem == {1, 3}; criterion = 0.999, statistic = 21.157
        6)*  weights = 212
      5) diasem == {2, 4, 5}
        7)*  weights = 271
    4) am1 > 424
      8)*  weights = 87
1) diasem == {6, 7}
  9)*  weights = 232
```

Se empregamos a árbore exposta para realizar predicións no período 28/03/2015-24/04/2015 obtemos os resultados expostos na Figura 3.42 xunto os valores MASE=51.2445 % e MSE=10203.36. Como se pode observar a introdución da covariable $diasem$ no lugar da covariable $am7$ parece que

proporciona un mellor modelado da compoñente semanal producindo isto unha diminución dos valores das medida de erro, agora ben para comprobar que esta melloría se produce globalmente e non só nesta ventá particular calculamos as medidas globais de erro de predición obtendo os resultados que presentamos no Cadro 3.27.

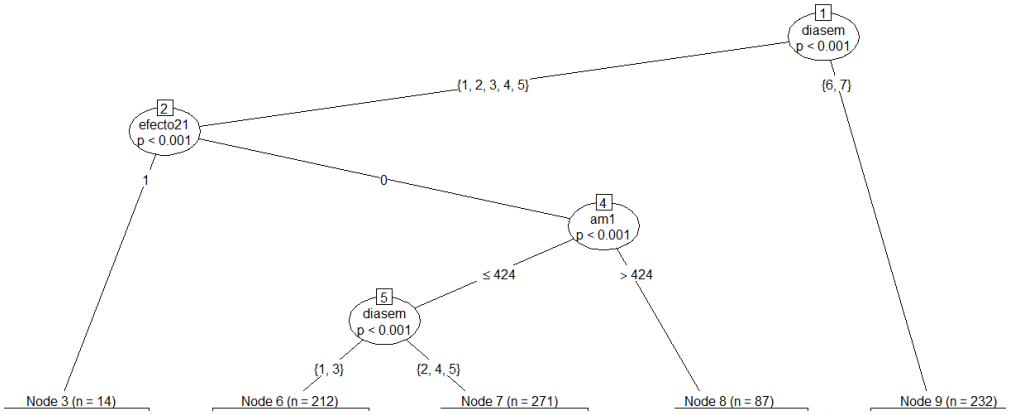


Figura 3.40: Gráfico da Árbore de Divisións Recursivas 2 axustada.

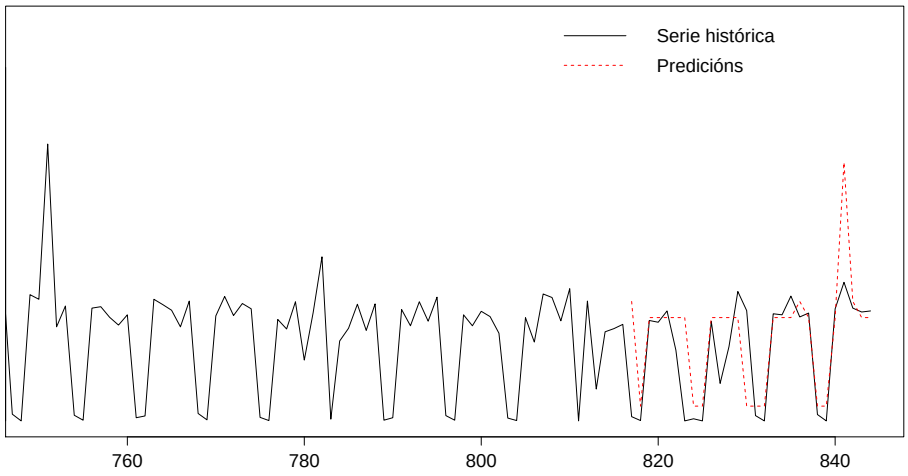


Figura 3.41: Predicións realizadas pola árbore 1 na ventá temporal 28/03/2015-24/04/2015 xunto aos valores reais da serie temporal.

Efectivamente os resultados presentes no Cadro 3.27 indican que a substitución da covariable *am7* polo covariable *diasem* produce unha diminución do valor das medidas globais de erro concluíndo entón que árbore 2 predí dun xeito máis adecuado que a árbore 1. Analogamente ao sucedido na serie temporal "número de baixas diarias de teléfono móbil" comprobamos se a introdución dunha covariable categórica que distinga os distintos meses do ano é útil para a predición, e o resultado é que dita covariable non axuda a mellorar o comportamento da árbore de decisión.

Media		Mediana		Desviación típica	
MASE	MSE	MASE	MSE	MASE	MSE
48.7719	11215.6	48.3277	10117.3	14.3141	4433.051

Ventá temporal	MASE	MSE
Ventá 1	36.3477	8225.681
Ventá 2	39.4221	8238.822
Ventá 3	37.4822	7843.547
Ventá 4	57.1528	17016.231
Ventá 5	36.7450	6694.978
Ventá 6	48.3277	10117.303
Ventá 7	52.4900	11775.887
Ventá 8	49.9645	11301.576
Ventá 9	81.0145	19726.363

Cadro 3.26: Resultados das medidas de erro de predición para a árbore 1.

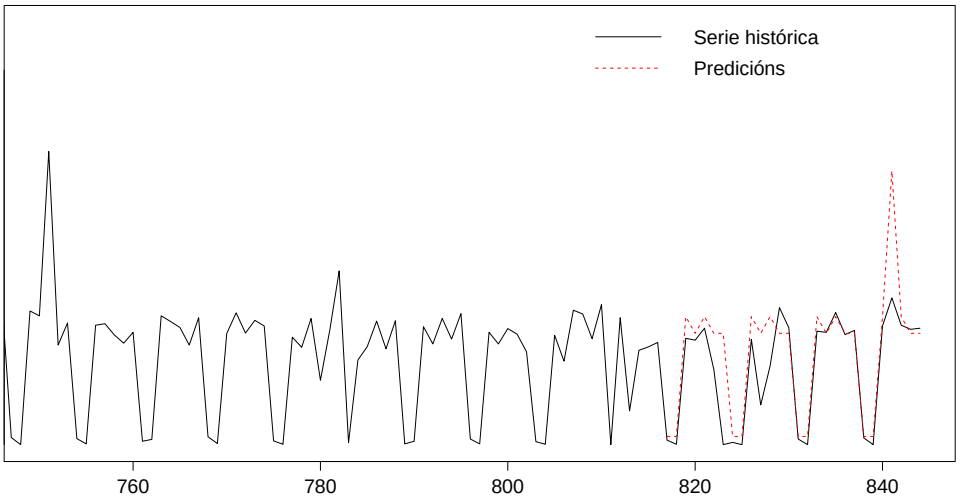


Figura 3.42: Predicións realizadas pola árbore 2 na ventá temporal 28/03/2015-24/04/2015 xunto aos valores reais da serie temporal.

Media		Mediana		Desviación típica	
MASE	MSE	MASE	MSE	MASE	MSE
42.1304	9482.854	37.3512	8927.321	7.8117	2118.922

Ventá temporal	MASE	MSE
Ventá 1	34.4591	8927.321
Ventá 2	37.3512	8942.868
Ventá 3	36.2823	8582.783
Ventá 4	34.3502	7278.683
Ventá 5	36.5690	7345.351
Ventá 6	48.6276	10804.598
Ventá 7	53.7906	12909.285
Ventá 8	51.2972	12550.170
Ventá 9	46.4463	8004.623

Cadro 3.27: Resultados das medidas de erro de predición para a árbore 2.

3.3.4. Métodos de predición simples

Nesta sección imos modelar e predicir a serie "número de altas diarias de teléfono móbil" mediante os métodos de predición simples co obxectivo de comparar os resultados proporcionados por estes métodos cos obtidos na aplicación das outras metodoloxías.

Analogamente ao realizado para as outras metodoloxías de predición axustamos o modelo empregando unicamente as observacións do período 01/01/2013-27/03/2015, e logo predicimos con este os valores da ventá temporal 28/03/2015-24/04/2015. Na Figura 3.43 podemos ver as predicións obtidas polos diferentes métodos simples nesta ventá xunto aos valores reais da serie temporal. Esta comparativa gráfica compleméntase cos valores das medidas MASE e MSE que se encontran no Cadro 3.28.

	Método da Media	Método Naive	Método Naive Estacional	Método Drift
MASE	112.4897	110.3582	105.6496	180.8371
MSE	23693.24	25803.46	29215.46	26243.5

Cadro 3.28: Valores de MASE e MSE para cada un dos métodos simples na ventá temporal 28/03/2015-24/04/2015.

Na Figura 3.43 pódese observar que nesta ventá temporal o método Naive estacional non capta a compoñente estacional semanal adecuadamente, isto débese a que nos últimos 7 días da serie temporal, que é nos que se fundamenta este método para predicir valores futuros, temos presente un día 21

cuxo repunte non ten unha gran magnitude pero a súa presencia produce unha variación no comportamento semanal. Por outra parte o método da media, o método Naive e o método Drift presentan os inconvenientes xa mencionados con anterioridade na serie "número de baixas diarias de teléfono móbil". O método da media ten un mal funcionamento pola presenza dos repuntes asociados aos días 21 que impiden que este capte o nivel da serie temporal, e o comportamento do método Drift e Naive é altamente dependente do último valor da serie temporal, consecuentemente o comportamento dos mesmos pode variar dunha ventá a outra desde aceptable ata realmente desatinado, na ventá concreta que observamos vemos que estes métodos non captan correctamente o nivel da serie temporal.

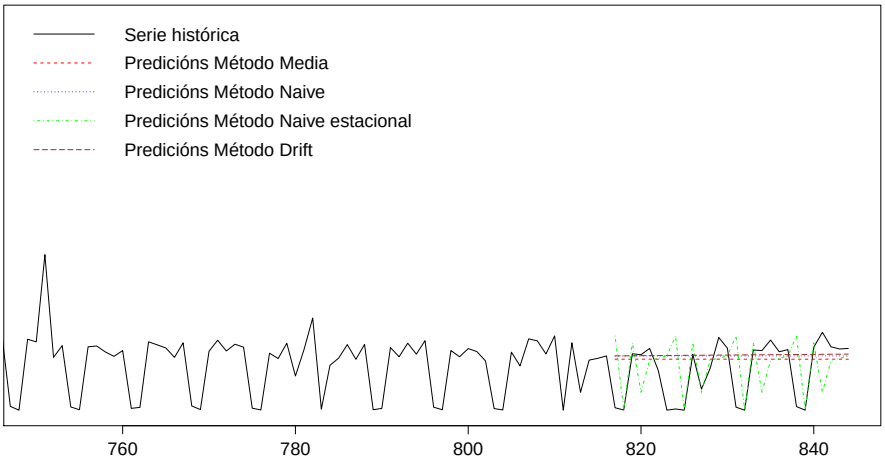


Figura 3.43: Gráfico secuencial dos valores reais da serie temporal xunto aos valores preditos por cada un dos métodos simples na ventá temporal 28/03/2015-24/04/2015.

Se observamos o Cadro 3.28 vemos que todos os valores MASE indican que o método que predí un valor da serie temporal empregando a observación inmediatamente anterior é mellor en termos de erro de predición que calquera dos métodos simples aplicados. No Cadro 3.29 mostramos as medidas globais de predición para cada un dos métodos simples, neste podemos ver que en xeral o comportamento do método da media e do método Naive estacional é máis axeitado que na ventá concreta empregada para ilustrar a gráfica das predicións, pois o MASE é inferior o 100 % o que indica que estes métodos melloran en termos de erro de predición ao método que predí un valor da serie temporal empregando a observación inmediatamente anterior. Isto non sucede para os métodos Naive e Drift pois polos inconvenientes expostos o funcionamento deste métodos é inadecuado en moitas situacións.

3.3.5. Conclusións

Tras a aplicación das diferentes metodoloxías imos comparar os resultados obtidos co obxectivo de seleccionar finalmente a metodoloxía que a empresa debe empregar na predición de valores futuros da serie temporal "número de altas diarias de teléfono móbil". No Cadro 3.30 recollemos os resultados das medidas globais de erro para cada unha das metodoloxías aplicadas.

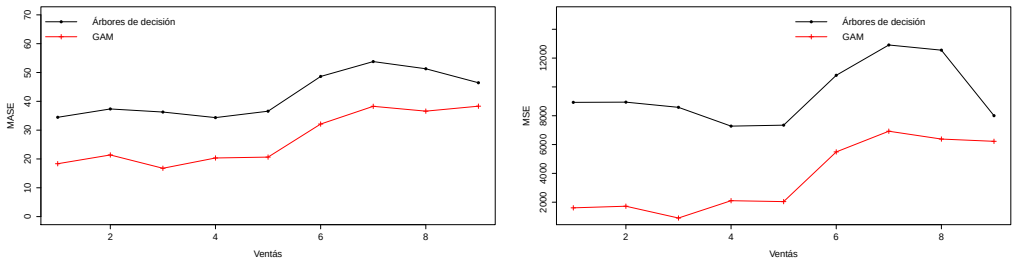
Para a serie temporal "número de altas diarias de teléfono móbil" non existe dúbida entorno a metodoloxía a seleccionar pois en todas as medidas de erro mostradas no 3.30 o menor valor é o asociado ao modelo GAM. A pesar disto na Figura 3.44 amosamos dous gráficos nos cales se mostran os valores de MASE e MSE en cada unha das ventás de predición para a metodoloxía GAM e as Árbores de Divisións Recursivas co obxectivo de ilustrar que efectivamente a decisión tomada é acertada. Tamén

cabe destacar que os métodos simples están moi lonxe de acadar os valores de erro de predición obtidos empregando a metodoloxía GAM, pois o método simple que obtén mellores resultados é o método Naive estacional que case duplica os valores de erro proporcionados pola metodoloxía GAM.

	Media		Mediana		Desviación típica	
Metodoloxía	MASE	MSE	MASE	MSE	MASE	MSE
Media	82.8196	20803.66	80.6842	19536.05	8.5741	1782.29
Naive	113.4571	37505.33	91.7720	30193.21	37.3687	16149.53
Naive estacional	50.0073	13859.59	44.7066	10276	18.2424	7411.565
Drift	114.9546	38258.41	94.0779	31504.9	36.3763	15728.88

	Media		Naive		Naive estacional		Drift	
Ventá temporal	MASE	MSE	MASE	MSE	MASE	MSE	MASE	MSE
Ventá 1	75.0740	19504.71	97.3405	34228.71	33.3501	8914.286	101.8362	35990.27
Ventá 2	79.7699	19284.45	176.4076	67424.14	42.6555	10276.000	176.4819	67461.93
Ventá 3	80.6842	19458.96	83.6223	24616.21	38.0513	10209.071	85.5925	25510.18
Ventá 4	74.6262	19536.05	85.4243	30193.21	72.1167	27582.643	87.4564	31504.90
Ventá 5	76.1775	19188.49	85.5767	27143.36	33.8214	7271.571	88.1887	28420.39
Ventá 6	85.7541	21390.04	152.7922	52676.54	47.6180	11788.750	152.7446	52619.93
Ventá 7	83.9408	22530.86	91.7720	28786.29	50.2810	13188.714	94.0779	29964.82
Ventá 8	87.1892	22510.12	90.5335	19743.86	44.7066	9940.714	90.4760	20078.38
Ventá 9	102.1600	23829.26	157.6453	52735.68	87.4656	25564.536	157.7370	52774.87

Cadro 3.29: Valores de MASE e MSE para cada un dos métodos simples na ventá temporal 28/03/2015-24/04/2015.



(a) Valores da medida MASE. (b) Valores da medida MSE.

Figura 3.44: Gráficos dos valores de MASE e MSE para o modelo GAM e a Árbore de Divisións Recursivas.

	Media		Mediana		Desviación típica	
Metodoloxía	MASE	MSE	MASE	MSE	MASE	MSE
Box-Jenkins	66.3326	14763.34	57.8486	11002.87	34.1932	11336.41
GAM	26.9754	3710.477	21.3890	2103.439	9.1427	2464.655
Árbores	42.1304	9482.854	37.3512	8927.321	7.8117	2118.922
Media	82.8196	20803.66	80.6842	19536.05	8.5741	1782.29
Naive	113.4571	37505.33	91.7720	30193.21	37.3687	16149.53
Naive estacional	50.0073	13859.59	44.7066	10276	18.2424	7411.565
Drift	114.9546	38258.41	94.0779	31504.9	36.3763	15728.88

Cadro 3.30: Valores das medidas globais de erro para cada unha das metodoloxías aplicadas.

Efectivamente os gráficos da Figura 3.44 indican que a metodoloxía GAM é a correcta para realizar predicións da serie temporal pois en todas as ventás consideradas observamos que os valores das medidas MASE e MSE son menores para a metodoloxía GAM.

Consecuentemente a metodoloxía GAM é a seleccionada para a modelización da serie temporal "número de altas diarias de teléfono móbil", cabe mencionar que o seu tempo de computación é 0.28 s.

Capítulo 4

Conclusións

No Capítulo 1 deste proxecto mencionamos que o obxectivo do mesmo era seleccionar unha familia de modelos axeitada para predicir eventos futuros de tres series temporais que impactan no negocio e nos sistemas dunha operadora de telecomunicacións. Ao longo deste traballo realizamos diferentes tarefas para cumprir o obxectivo exposto. A primeira delas consistiu en familiarizar ao lector co concepto de serie temporal, unha vez feito isto realizamos unha revisión metodolóxica dos modelos Box-Jenkins, GAM, e das Árbores de divisións recursivas xunto aos métodos de predición simples co obxectivo de adquirir os coñecementos necesarios para a súa aplicación no estudo das tres series temporais. Ademais disto presentamos tamén un procedemento para verificar a exactitude das predicións realizadas por cada unha das metodoloxías para deste xeito poder comparar ditas metodoloxías en termos de erro de predición e seleccionar así a máis adecuada para a predición de valores futuros, cumprindo así o obxectivo deste traballo. Unha vez dispoñemos desta base metodolóxica aplicamos estes coñecementos no estudo das tres series temporais obtendo as seguintes conclusións: O estudo da serie "número de usuarios do servizo de vídeo baixo demanda" indicounos que ante unha serie similar a esta onde temos tendencia e compoñente estacional xunto a unha alta variabilidade que impide captar correctamente o comportamento estacional a mellor opción é empregar o método simple Drift, salvo que a empresa dispoña do tempo suficiente é desexe empregar as Árbores de Divisións Recursivas que proporcionan uns resultados similares ao método simple. En calquera dos dous casos creemos conveniente proporcionar á empresa de telecomunicacións as predicións semanais como un complemento das predicións diarias. Con respecto ao estudo das series temporais "número de baixas e altas de teléfono móbil" a metodoloxía que a empresa debe empregar para modelar unha serie temporal destas características é sen dúbida a metodoloxía GAM, pois pese a non ser unha metodoloxía deseñada exclusivamente para o estudo de series temporais proporciona, considerando as covariables axeitadas, un modelado máis satisfactorio co obtido coas outras metodoloxías consideradas. Tras a realización destas tarefas e a obtención das conclusións que vimos de expor creemos ter satisfeito o obxectivo do proxecto.

Por outra parte cabe mencionar que este estudo tamén proporcionou conclusións xerais acerca das metodoloxías consideradas neste traballo. Primeiramente, con respecto á metodoloxía Box-Jenkins dicir que é unha metodoloxía útil no modelado de series temporais sempre e cando non sexa necesario considerar covariables externas como a variable que permite modelar o repunte do día 21 nas series temporais "número de baixas e altas". Ademais disto tamén cabe mencionar que o proceso de identificación dun modelo Box-Jenkins pode resultar complexo se non se dispón dos coñecementos expostos no Capítulo 2. Outro punto a destacar, non pola súa complexidade se non polo tempo necesario para a súa realización, é que o proceso de incorporación dos valores atípicos tense que realizar manualmente, pois non existe unha función que o faga automaticamente. Por outra parte, pese a que os modelos GAM teñen unha alta complexidade metodolóxica, a interpretación do mesmo é sinxela, igual que a súa aplicación a cal está automatizada. Isto é unha gran vantaxe pois obter resultados claramente interpretables facilita as tarefas de mercadotecnia. Para rematar dicir que as Árbores de Divisións Recursivas tamén son sinxelas de aplicar e interpretar e que este é o motivo polo cal gozan

de popularidade no mundo empresarial.

Por último, menciónanse algunhas das dificultades afrontadas no estudo destas series temporais. Unha das series temporais máis complexa de modelar foi a serie "números de usuarios do servizo de vídeo baixo demanda" pois a alta variabilidade da mesma xunto cunha cantidade considerable de valores atípicos dificultou o seu modelado. Outra dificultade, neste caso nas series "número de altas e baixas de teléfono móbil", foi a presenza de valores nulos nas series temporais, pois ditos valores non teñen un comportamento totalmente determinado que permita o seu modelado. Por exemplo, na serie "número de baixas" observamos que un alto porcentaxe dos valores nulos se producen en domingo ou luns pero detectamos valores nulos noutros días sen conseguir unha explicación clara deste suceso. Ademais nestas series temporais tamén nos encontramos coa presenza de repuntes no día 21 do mes. Na serie "número de baixas diarias" é sinxelo solventar esta dificultade simplemente introducindo a pertinente covariable, algo que non é tan sinxelo na serie "número de altas" pois os repuntes non se producen todos os meses e tampouco se producen seguindo un patrón determinable.

A pesar destas dificultades cremos ter solventado adecuadamente o problema proposto indicándolle á empresa as metodoloxías que proporcionan menor erro de predición en cada unha das series temporais.

Para rematar cabe mencionar que se poderían utilizar series temporais como as tratadas neste traballo co obxectivo de identificar a taxa de abandono de clientes. Pois no modelado pertinente para levar a acabo a identificación dos clientes que abandonarán a compañía esta presente o problema das clases desbalanceadas, o cal se pode solucionar empregando o método Oversampling (subir o número de rexistros pertencentes a clase minoritaria), Downsampling (reducir o número de rexistros da clase maioritaria), ou un híbrido dos dous que consiste en subir o porcentaxe da clase minoritaria e reducir o da maioritaria. Neste punto é onde podemos usar a serie temporal "número de baixas globais" para coñecer o número de clientes que se darán de baixa nos próximos meses e intentar buscar este número balanceando as clases dun xeito ou doutro.

Apéndice A

Código Subsección 2.2.2

Neste apéndice proporcionamos ao lector o código, en linguaxe R, necesario para realizar os gráficos dos exemplos expostos no proceso de identificación dos modelos Box-Jenkins, Subsección 2.2.2.

Para a realización destes exemplos empregamos basicamente dúas funcións da librería stats do *sowfare* estatístico R:

- A función ARMAacf calcula a función de autocorrelacións teórica simple e parcial dun proceso ARMA.
- A función acf calcula a estimación da función de autocovarianzas e autocorrelacións dunha serie temporal. Por defecto calcula a estimación da función de autocorrelacións.

Deseguido mostramos o código correspondente a cada un dos gráficos da Subsección 2.2.2.

Código correspondente á Figura 2.1:

```
data(arma11.s)
plot(arma11.s, type="l", main="", xlab="", ylab="")
acf(arma11.s, main="", xlab="Retardo", lag.max = 18)
```

Código correspondente á Figura 2.2a:

```
fas <- ARMAacf(ar=-0.7, ma=0, 20)
fas <- fas[-1]
fap <- ARMAacf(ar=-0.7, ma=0, 20, pacf=T)
par(mfrow=c(2,1))
plot(fas, type="h", xlab="Retardo", ylim=c(-0.9,0.6),xlim=c(0,13), main="")
abline(h=0)
plot(fap, type="h", xlab="Retardo", ylim=c(-0.9,0.6), xlim=c(0,13),main="")
abline(h=0)
```

Código correspondente á Figura 2.2b:

```
fas <- ARMAacf(ar=0, ma=0.7, 20)
fas <- fas[-1]
fap <- ARMAacf(ar=0, ma=0.7, 20, pacf=T)
par(mfrow=c(2,1))
plot(fas, type="h", xlab="Retardo", ylim=c(-0.5,0.7), xlim=c(0,10),main="")
abline(h=0)
plot(fap, type="h", xlab="Retardo", ylim=c(-0.5,0.7),xlim=c(0,10),main="")
abline(h=0)
```

Código correspondente á Figura 2.3:

```
fas <- ARMAacf(ar=c(0.7), ma=c(0.5), 20)
fas <- fas[-1]
fap <- ARMAacf(ar=0.7, ma=0.5, 20, pacf=T)
par(mfrow=c(2,1))
plot(fas, type="h", xlab="Retardo", xlim=c(0,15), ylim=c(-1,1), main="")
abline(h=0)
plot(fap, type="h", xlab="Retardo", ylim=c(-1,1), xlim=c(0,15),main="")
abline(h=0)
```

Código correspondente á Figura 2.4a:

```
Phi <- c(rep(0,11), 0.7)
fas <- ARMAacf(ar=Phi, ma=0, 60)
fas <- fas[-1]
fap <- ARMAacf(ar=Phi, ma=0, 60, pacf=T)
par(mfrow=c(2,1))
plot(fas, type="h", xlab="Retardo", ylim=c(-1,1), main="")
abline(h=0)
plot(fap, type="h", xlab="Retardo", ylim=c(-1,1),main="")
abline(h=0)
```

Código correspondente á Figura 2.4b:

```
Theta <- c(rep(0,3), 0.6, rep(0,3), -0.3)
fas <- ARMAacf(ar=0, ma=Theta, 33)
fas <- fas[-1]
fap <- ARMAacf(ar=0, ma=Theta, 33, pacf=T)
par(mfrow=c(2,1))
plot(fas, type="h", xlab="Retardo", ylim=c(-0.5,0.5),main="")
abline(h=0)
plot(fap, type="h", xlab="Retardo", ylim=c(-0.5,0.5),main="")
abline(h=0)
```

Código correspondente á Figura 2.5a:

```
Theta <- c(rep(0,11), -0.7)
fas <- ARMAacf(ar=0.4, ma=Theta, 50)
fas <- fas[-1]
fap <- ARMAacf(ar=0.4, ma=Theta, 50, pacf=T)
par(mfrow=c(2,1))
plot(fas, type="h", xlab="Retardo", ylim=c(-0.5,0.5), main="")
abline(h=0)
plot(fap, type="h", xlab="Retardo", ylim=c(-0.5,0.5),main="")
abline(h=0)
```

Código correspondente á Figura 2.5b:

```
Phi <- c(rep(0,11), 0.7)
fas <- ARMAacf(ar=Phi, ma=-0.4, 50)
fas <- fas[-1]
fap <- ARMAacf(ar=Phi, ma=-0.4, 50, pacf=T)
par(mfrow=c(2,1))
```

```
plot(fas, type="h", xlab="Retardo", ylim=c(-0.7,0.8),main="")
abline(h=0)
plot(fap, type="h", xlab="Retardo", ylim=c(-0.7,0.8),main="")
abline(h=0)
```

Código correspondente á Figura 2.6:

```
data("rwalk")
par(mfcol=c(1,1))
plot(rwalk, type="l", main="", xlab="", ylab="")
acf(rwalk, main="", xlab="Retardo")
```

Código correspondente á Figura 2.7:

```
data(hare)
plot(hare, type="l", main="", xlab="", ylab="")
acf(hare, main="", xlab="Retardo", lag.max = 50)
```

Apéndice B

Código función best.arima.TSA

Neste apéndice proporcionamos unha descrición da función best.arima.TSA empregada na identificación dun modelo Box-Jenkins para o logarítmo da serie "número de usuarios diarios do servizo de vídeo baixo demanda", Sección 3.1.

Como xa indicamos, a función best.arima.TSA proporciona as ordes do modelo Box-Jenkins máis axeitado segundo o criterio que se indique, que no caso da Sección 3.1 foi o criterio BIC.

A liña de código que nos permite facer uso de dita función é

```
best.arima.TSA <- function(x=x, p.max=5, q.max=5, d=0, P.max=5, Q.max=5, D=0, p.min=0,
q.min=0, P.min=0, Q.min=0, period=1, xreg=NULL, include.mean=TRUE, criterio="BIC",
dist.max.crit=2).
```

Como se pode ver a función best.arima.TSA ten unha serie de argumentos, cuxa descrición se aborda de contado.

- x serie de tempo univariante a cal desexamos modelar empregando a metodoloxía Box-Jenkins.
- p.min mínimo orden para tomar na parte autoregresiva regular do modelo. Por defecto 0.
- q.min mínimo orden para tomar na parte de medias móbiles regular do modelo. Por defecto 0.
- p.max máximo orden para tomar na parte autoregresiva regular do modelo. Por defecto 5.
- q.max máximo orden para tomar na parte de medias móbiles regular do modelo. Por defecto 5.
- d número de diferenzas regulares aplicadas para eliminar a tendencia da serie temporal. Por defecto 0.
- P.min mínimo orden para tomar na parte autoregresiva estacional do modelo. Por defecto 0.
- Q.min mínimo orden para tomar na parte de medias móbiles estacional do modelo. Por defecto 0.
- P.max máximo orden para tomar na parte autoregresiva estacional do modelo. Por defecto 5.
- Q.max máximo orden para tomar na parte de medias móbiles estacional do modelo. Por defecto 5.
- D número de diferenzas estacionais aplicadas para eliminar a compoñente estacional da serie temporal. Por defecto 0.
- period valor que reflexa o período de estacionalidade da serie. Por defecto 1.

- xreg opcionalmente, un vector ou matriz de regresores externos, os cales deben ter o mesmo número de filas que x. Por defecto nulo.
- include.mean parámetro lóxico que indica se o modelo debe incluír unha media ou término constante. Por defecto, verdade.
- criterio tipo de criterio aplicado para seleccionar o modelo. Existe a opción de AIC, AICC e BIC (por defecto).
- dist.max.crit distancia máxima entre o mellor modelo e o seguinte modelo, en sentido do criterio, que se quere mostrar por pantalla. Por defecto 2.

Esta función devolve como resultado os seguintes valores.

- p orden da parte autoregresiva regular do modelo.
- q orden da parte de medias móbiles regular do modelo.
- P orden da parte autoregresiva estacional do modelo.
- Q orde da parte de medias móbiles estacional do modelo.
- criterio valor do criterio de información asociado ao modelo.

Xa para rematar de contado proporcionamos o código da función.

```
{
library(TSA)
if (is.ts(x)) period <- frequency(x)
num.x.perdidos <- d + period*D
T <- length(x) - num.x.perdidos
A<-matrix(0,((p.max-p.min)+1)*((q.max-q.min)+1)*((P.max-P.min)+1)*((Q.max-Q.min)+1),5)
if (criterio=="AIC")  factor <- 2
   else if (criterio=="BIC")  factor <- log(T)
fila <- 0
for (p in p.min:p.max)
for (q in q.min:q.max)
for (P in P.min:P.max)
for (Q in Q.min:Q.max)
  {
#optim.control=list(maxit=500)
fila <- fila +1
ajuste<-try(arimax(x=x, order=c(p,d,q), seasonal=list(order=c(P,D,Q), period=period),
xreg=xreg, include.mean=include.mean), silent=TRUE)
if (class(ajuste)=="try-error") {
A[fila, ] <- c(p, q, P, Q, NaN)
  next
}
k <- length(ajuste$coef)
if (criterio=="AICC")  criterio.ajuste <- ajuste$aic +2*(k+1)*(k+2)/(T-k-2)
else criterio.ajuste <- -2*ajuste$loglik + factor*k
A[fila, ] <- c(p, q, P, Q, criterio.ajuste)
}
min.ic <-  min(A[,5], na.rm=TRUE)
A <- A[(A[,5]-min.ic)<=dist.max.crit,]
```

```
if (!is.matrix(A)) A <- t(as.matrix(A))
A <- A[!is.na(A[,1]),]
if (!is.matrix(A)) A <- t(as.matrix(A))
A <- A[order(A[,5]),]
if (!is.matrix(A)) A <- t(as.matrix(A))
A <- as.data.frame(A)
if (criterio=="AIC") names(A) <- c("p", "q", "P", "Q", "AIC")
 else if (criterio=="AICC") names(A) <- c("p", "q", "P", "Q", "AICC")
 else names(A) <- c("p", "q", "P", "Q", "BIC")
return(A[,c(1*(p.max!=0), 2*(q.max!=0), 3*(P.max!=0), 4*(Q.max!=0), 5)])
}
```

Bibliografía

[1] Brockwell PJ, Davis RA (2002) Introduction to Time Series and Forecasting. Springer, New York.

[2] Chan KS, Ripley B (2012) TSA: Time Series Analysis. R package version 1.0.1. http://cran.r-project.org/package=TSA. Accedido 24 de novembro de 2015.

[3] Cryer JD, Chan KS (2008) Time Series Analysis. With Applications in R. Springer, New York.

[4] Dominici F, McDermott A, Zeger SL, Samet JM (2002) On the Use of Generalized Additive Models in Time Series Studies of Air Pollution and Health. American Journal of Epidemiology 156:193-203.

[5] Hothorn T, Hornik K, Strobl C, Zeileis A (2015) party: A Laboratory for Recursive Partytioning. R package version 1.0-25. http://cran.r-project.org/package=party. Accedido 24 de novembro de 2015.

[6] Hothorn T, Hornik K, Zeileis A (2006) Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics 15(3):651-674.

[7] Hyndman R (2011) Forecasting time series using R. http://robjhyndman.com/talks/MelbourneR-UG.pdf. Accedido 02 de xaneiro de 2016.

[8] Hyndman R (2015) forecast: Forecasting Functions for Time Series and Linear Models. R package version 6.2. http://cran.r-project.org/package=forecast. Accedido 27 de novembro de 2015.

[9] Peña D (2010) Análisis de Series Temporales. Alianza Editorial, Madrid.

[10] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[11] Shumway RH, Stoffer DS (2011) Time Series Analysis and Its Applications. With R examples. Springer, New York.

[12] Thode HC (2002) Testing for Normality. Marcel Dekker, New York.

[13] Tobías A, Saez M (2004) Time-series regression models to study the short-term e effects of the environmental factors on health. Working Papers of the Department of Economics, University of Girona. Number 11. https://ideas.repec.org/p/udg/wpeudg/011.html

[14] Wei WW (2006) Time Series Analysis: Univariate and Multivariate Methods. Addison Wesley, Boston.

[15] Wood S (2006) Generalized Additive Models: an introduction with R. Chapman and Hall, London.

[16] Wood S (2015) mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation. R package version 1.8-9. http://cran.r-project.org/package=mgcv. Accedido 24 de novembro de 2015.