Traballo Fin de Máster

Desenvolvemento dun indicador de alta frecuencia para o seguimento da economía española

Lucía Gil Rial

Máster en Técnicas Estatísticas

Curso 2020-2021

Proposta de Traballo Fin de Máster

Título en galego: Desenvolvemento dun indicador de alta frecuencia para o seguimento da economía española

Título en castelán: Desarrollo de un indicador de alta frecuencia para el seguimiento de la economía española

English title: Development of a high-frequency index to track spanish economy

Modalidade: Modalidade B

Autor/a: Lucía Gil Rial, Universidade de Santiago de Compostela

Director/a: Salvador Naya Fernández, Universidade da Coruña; Javier Tarrío Saavedra, Universidade da Coruña

Titor/a: Belén María Fernández de Castro, ABANCA

Breve resumo do traballo:

Na Dirección de Planificación Estratéxica e PMO de ABANCA están interesados no desenvolvemento dun indicador da economía española baseado en métricas de alta frecuencia de monitorización, isto é, métricas onde o intervalo temporal entre a toma de observacións é moi curto, por exemplo, días ou semanas. Este índice permitiría anticipar o comportamento da actividade económica nun menor prazo de tempo que os indicadores tradicionais, como o PIB, que adoitan ter unha frecuencia trimestral. Ademais, no contexto no que nos atopamos, este índice resulta de máximo interese. Neste traballo partiremos da demanda de electricidade total diaria no país para a construción do dito indicador.

Don Salvador Naya Fernández, catedrático da Universidade da Coruña, don Javier Tarrío Saavedra, profesor contratado e doutor da Universidade da Coruña, dona Belén María Fernández de Castro, xerente en ABANCA, informan que o Traballo Fin de Máster titulado

Desenvolvemento dun indicador de alta frecuencia para o seguimento da economía española

foi realizado baixo a súa dirección por dona Lucía Gil Rial para o Máster en Técnicas Estatísticas. Estimando que o traballo está rematado, dan a súa conformidade para a súa presentación e defensa ante un tribunal.

En A Coruña, a 22 de xuño de 2021.

O director:

Don Salvador Naya Fernández

O director:

Don Javier Tarrío Saavedra

A titora:

Dona Belén María Fernández de Castro

A autora:

Dona Lucía Gil Rial

Agradecementos

Comezo agradecendo á entidade ABANCA por ofrecerme a posibilidade de desenvolver este traballo na súa institución. A experiencia aportoume un primeiro contacto co mundo laboral e novos coñecementos na área da estatística. Cabe facer especial mención a Belén Fernández de Castro, Teresa Veiga Rodríguez e Sergio Díaz Canosa.

Doutra banda, tamén me gustaría agradecer aos titores académicos, Salvador Naya Fernández e Javier Tarrío Saavedra, por acompañarme neste proxecto.

Índice xeral

Resumo

Resumo

Os índices macroeconómicos tradicionais, como o Produto Interior Bruto (PIB), adoitan ter unha frecuencia trimestral e, ademais, existen atrasos na súa publicación, polo que resulta difícil analizar a evolución do estado económico do país cando se dan situacións tan bruscas como a pandemia provocada pola COVID-19. Polo tanto, xorde a necesidade de indicadores de alta frecuencia, por exemplo semanais ou diarios, que permitan rastrexar a economía en tempo real, no senso de que non hai un atraso significativo na súa difusión. Na actualidade existen diversas variables que se toman en intervalos de tempo moi pequeno, como o fluxo de mobilidade, os pagos con tarxetas e a demanda de enerxía, que poden ser empregadas para avaliar o comportamento da produción industrial, sector servizos, etc.

Ao longo deste traballo preséntase o comezo do desenvolvemento dun índice macroeconómico de alta frecuencia, considerando como variable de partida a demanda de electricidade diaria no país. É interesante coñecer o comportamento da dita variable, polo que primeiro móstrase unha análise anual da mesma. Logo, preséntanse diversas metodoloxías para limpar a demanda eléctrica, ben de estacionalidades ou ben da influencia de variables esóxenas. Unha vez que se obtén a variable corrixida, constrúese un indicador mensual e de frecuencias inferiores e estúdase a súa dinámica, así como a relación que garda con indicadores usuais, como o Índice de Produción Industrial (IPI) e o PIB.

Resumen

Los índices macroeconómicos tradicionales, como el PIB, suelen tener una frecuencia trimestral y, además, existen retrasos en su publicación, por lo que resulta difícil analizar el progreso del estado económico del país cuando se producen situaciones tan bruscas como la pandemia provocada por la COVID-19. En consecuenia, surge la necesidad de indicadores de alta frecuencia, por ejemplo semanales o diarios, que permitan rastrear la economía en tiempo real, en el sentido de que no hay un retraso significativo en su difusión. En la actualidad existen diversas variables que se toman en intervalos de tiempo muy pequeño, como el flujo de movilidad, los pagos con tarjetas y la demanda de energía, que pueden usarse para evaluar el comportamiento de la producción industrial, sector servicios, etc.

A lo largo de este trabajo se presenta el inicio del desarrollo de un índice macroeconómico de alta frecuencia, considerando como variable de partida la demanda de electricidad diaria en el país. Resulta interesante conocer el comportamiento de esta variable, por lo que primero se muestra un análisis anual de la misma. Después, se presentan diversas metodologías para limpiar la demanda eléctrica, bien de estacionalidades o bien de la influencia de variables exógenas. Una vez que disponemos de la variable corregida, se construye un indicador mensual y de frecuencias inferiores y se estudia su dinámica, así como la relación con indicadores tradicionales, como el IPI y el PIB.

Abstract

Traditional macroeconomic indicators, such as GDP, often have a quarterly frequency and, moreover, there are delays in their publication. This makes it difficult to analyse the progress of the economy of the country when situations as abrupt as the pandemic caused by COVID-19 occur. Therefore, there is a need for high frequency indicators which either weekly or daily let us track the state of economy in real time avoiding meaningful delays. Currently, there are several variables that are taken at very short time intervals, as mobility, card payments and energy demand, which can be used to assess the dynamics of industrial production, service sector, etc.

Throughout this report, the beginning of the development of a high-frequency macroeconomic index is introduced, considering as a starting variable the daily electricity demand in Spain. Firtsly, we will show an annual analysis of this variable in order to understand its behaviour. Next, we will present several methodologies to correct electricity demand either from seasonality or from the effects of exogenous variables. Once we have the clean variable, we will calculate a monthly and lower frequency index and we will study its dynamics as well as its relation with traditional indicators, such as IPI and GDP.

Prefacio

No Departamento de Planificación Estratéxica e PMO de ABANCA están interesados na construción dun indicador de alta frecuencia para medir a actividade económica no país, entendéndose por alta frecuencia un intervalo de tempo entre a toma de observacións moi curto, por exemplo, días. A inesperada pandemia provocada pola COVID-19 deixou clara a necesidade do desenvolvemento do dito indicador, pois os máis tradicionais, como o PIB, adoitan ter unha frecuencia trimestral e, ademais, contan con certos días de atraso na súa publicación, polo que dificulta a posibilidade de anticipar o comportamento económico nun período curto de tempo.

Así, este proxecto ten por obxectivo encamiñar a construción dun indicador de alta frecuencia de rastrexo da economía española. Os indicadores macroeconómicos son o resultado da agregación de múltiples variables, pero como punto de partida para a creación do noso índice, consideraremos só a demanda de electricidade diaria, pois esta variable ten unha grande importancia no progreso económico, dado que a produción de bens e servizos precisan de enerxía.

O traballo está configurado en dúas partes. A primeira está adicada á presentación da metodoloxía que se precisou para tratar e dar solución ao problema que nos compete e a segunda mostra os resultados prácticos. A continuación detállase a distribución das ditas partes.

Metodoloxía

O Capítulo 1 comeza introducindo os conceptos básicos no ámbito de series temporais, para logo expoñer a metodoloxía Box-Jenkins, que busca modelar un proceso estocástico que puidera xerar a serie e facer predicións. Por último, exponse a metodoloxía clásica de descomposición de series temporais e o método STL, que é un procedemento máis robusto.

O Capítulo 2 está adicado aos modelos de regresión. Primeiro revísanse brevemente os xa coñecidos modelos de regresión lineares, seguidos polos modelos de regresión polinómicos locais, que se enmarcan no contexto non paramétrico. Así, estes últimos non precisan de hipóteses tan estritas, que poden non cumprirse, e aportan unha maior flexibilidade. Pero no caso de dispoñer de múltiples variables explicativas, pode darse o problema da maldición da dimensión. Para resolver este inconveniente xorden os modelos de regresión semiparamétrica, dos cales presentamos os modelos aditivos.

O Capítulo 3 achega un tema máis novo con respecto ás materias que se imparten no Máster en Técnicas Estatísticas. Descríbense os números índices, amplamente empregados nas ciencias sociais, especialmente na economía.

Práctica

Para o desenvolvemento da parte práctica fíxose uso do software estatístico libre R, para o que pode consultarse R Core Team (2020). As gráficas que se mostran ao longo do traballo están feitas mediante o paquete *ggplot2* (Wickham 2016) e para o manexo de *data frames* empregouse o paquete *dplyr* (Wickham et al 2021), dotada dunha linguaxe fácil de entender. Para que o formato das datas e horas estivese en galego especificamos `Sys.setlocale("LC_TIME","galician")`.

Para poñer en contexto a situación na que nos atopamos, o Capítulo 4 céntrase na presentación dunha pequena introdución ao problema que nos atangue.

É fundamental coñecer a dinámica da serie de demanda de electricidade diaria, que consideramos co fin último de construír un índice. O capítulo 5 mostra a análise da variable ao longo de 2017, un ano usual en termos de evolución económica. Este estudo compleméntase co Apéndice A, onde se detalla o comportamento da demanda eléctrica en 2018 e 2019, co obxectivo de coñecer se a evolución da serie anualmente coincide. Ademais, estudouse brevemente a tendencia da serie entre 2015 e 2019, deixando a un lado o 2020 polo seu carácter atípico, co fin de coñecer se o nivel da serie subiu, baixou ou se mantivo constante ao longo destes anos. Logo, estúdase a dinámica de demanda eléctrica no ano 2020, que resulta ter un carácter atípico polo impacto da pandemia de COVID-19. Como se mostra neste capítulo, a demanda eléctrica presenta estacionalidades e influencias de variables esóxenas e, dado que o interese radica na información da demanda eléctrica relacionada coa actividade económica, cómpre corrixir estes efectos. O Capítulo 6 céntrase nesta tarefa. Comézase corrixindo as observacións atípicas da demanda eléctrica, para logo eliminar a estacionalidade semanal, obtendo unha serie máis homoxénea. A corrección do efecto mensual resulta máis complicado de solucionar, polo que se desenvolven diferentes metodoloxías. Os dous primeiros procedementos, baseados no axuste de modelos de regresión da demanda eléctrica sobre a temperatura ou en variacións foron descartados por non proporcionar resultados axeitados. O último procedemento, baseado na variación da mediana da demanda de electricidade mensual respecto da mediana desta variable ao longo dos anos considerados, resultou ser unha boa opción.

Unha vez que se dispón da demanda eléctrica limpa, pártese dela para a construción dun índice, cuxos resultados se mostran no Capítulo 7, así como a comparación con indicadores tradicionais, como o IPI, PIB e o IRE. Cabe resaltar especialmente a dinámica ao longo do 2020, onde o indicador que calculamos capta acertadamente o gran descenso producido neste ano, como consecuencia do confinamento.

Ao longo do Capítulo 8 preséntase unha aplicación que se fixo mediante a libraría *Shiny* (Chang et al 2020). Segundo o rango de datas que o usuario especifica, descárganse automaticamente as observacións da demanda eléctrica diaria no país e móstranse diferentes gráficos desta variable durante ese período de tempo. Tamén se presentan os índices das diferentes frecuencias consideradas, xunto coa súa comparación cos agregados macroeconómicos, IPI e PIB, e co IRE. Ademais, pódense descargar os datos dos indicadores nun arquivo Excel.

Por último, no Capítulo 9 preséntase un pequeno resumo coas conclusións do proxecto e as posibles liñas de investigación de cara ao futuro.

Parte I

Metodoloxía

Capítulo 1

Introdución ás series de tempo

Este capítulo está adicado á presentación das series de tempo e da metodoloxía Box-Jenkins. Para a súa realización empregáronse como referencias principais Cryer e Chan (2010) e Peña (2005). Na Sección 1.1 defínense os conceptos elementais para a análise das series temporais. Unha vez visto isto, na Sección 1.2 preséntase a metodoloxía Box-Jenkins, onde se definen os modelos ARIMA e as etapas na procura e axuste dun modelo estocástico que, razoablemente, puidera xerar a serie de tempo observada.

1.1. Conceptos previos

Nesta sección introducimos os conceptos básicos no marco das series de tempo. Como referencia básica pode consultarse Cryer e Chan (2010), páxina 11. Comezamos definindo un proceso estocástico, que serve como modelo para unha serie de tempo observada.

Definición 1.1. Un proceso estocástico é un conxunto de variables aleatorias, que denotaremos por $\{Y_t\}_{t\in C}$, definidas sobre o mesmo espazo de probabilidade.

No noso contexto, encadrámonos naqueles procesos estocásticos definidos no conxunto de números enteiros, $\{Y_t : t \in \mathbb{Z}\}$. Denotaremos con letras minúsculas, $y_t\,, t \in \mathbb{Z}$, a unha realización deste proceso.

Definición 1.2. Defínese unha serie temporal como unha realización parcial dun proceso estocástico, isto é, $\{y_1, \dots, y_T\}$.

As observacións da serie temporal foron tomadas en intervalos regulares de tempo, por exemplo, cada día. Os subíndices destas observacións indican o instante no que foron observadas. Así, seguindo co exemplo anterior, a observación y_1 é o dato correspondente ao primeiro día, y_2 ao segundo día e así sucesivamente ata o T-ésimo día.

Supoñemos que a serie de tempo foi xerada por un proceso estocástico. Polo tanto, á hora de analizar a serie, interésanos coñecer este proceso que a puido orixinar. Pero isto non é sinxelo, dado que a serie $\{y_t\}_{t\in\{1,\dots,T\}}$ só nos proporciona unha observación das variables Y_t, $t = 1, \dots, T$, do proceso estocástico $\{Y_t\}$, onde $t \in \mathbb{Z}$. En consecuencia, para facilitar esta tarefa, introducimos a continuación algunhas funcións características dos procesos estocásticos.

Definición 1.3. Sexa $\{Y_t\}_{t\in\mathbb{Z}}$ un proceso estocástico.

(I) A función de medias dun proceso defínese como

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

isto é, o valor esperado do proceso na variable t.

(II) A función de varianzas dun proceso é

$$\sigma_t^2 = Var(Y_t) = \mathbb{E}\left[(Y_t - \mu_t)^2\right].$$

Se ben a función de medias é unha medida de posición central da variable Y_t, a función de varianzas mide o grao de variabilidade desta variable.

(III) A función de autocovarianzas dun proceso defínese como

$$\gamma(t,s) = Cov(Y_t, Y_s) = \mathbb{E}\left[(Y_t - \mu_t)(Y_s - \mu_s)\right] = \mathbb{E}(Y_t Y_s) - \mu_t \mu_s.$$

Esta permite medir o grao de dependencia linear entre as variables Y_t e Y_s, $t, s \in \mathbb{Z}$. As autocovarianzas teñen dimensión (a dimensión da serie ao cadrado), polo que non son axeitadas para comparar series medidas en unidades diferentes. No seguinte apartado definimos tamén unha medida da dependencia linear, pero adimensional.

(IV) A función de autocorrelacións simples dun proceso é unha medida de dependencia linear entre Y_t e Y_s, pero toma valores en $[-1, 1]$.

$$\rho(t,s) = \frac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t)}\sqrt{Var(Y_s)}} = \frac{\gamma(t,s)}{\sigma_t \sigma_s}.$$

Aqueles valores de $\rho(t,s)$ próximos a ± 1 indican unha relación de dependencia linear forte entre as variables e aqueles valores de $\rho(t,s)$ próximos a cero indican pouca relación de dependencia linear. No caso de que $\rho(t,s) = 0$, as variables Y_t e Y_s son incorreladas.

(V) A función de autocorrelacións parciais dun proceso estocástico defínese como

$$\alpha(t,s) = \frac{Cov\left(Y_t - \widehat{Y}_t^{(t,s)}, Y_s - \widehat{Y}_s^{(t,s)}\right)}{\sqrt{Var\left(Y_t - \widehat{Y}_t^{(t,s)}\right) Var\left(Y_s - \widehat{Y}_s^{(t,s)}\right)}}.$$

onde $\widehat{Y}_i^{(t,s)}$ denota o mellor preditor linear de Y_i construído mediante aquelas variables medidas entre o tempo t e s e sen considerar estas. Así, esta función mide a dependencia linear entre dúas variables unha vez que se eliminou o efecto linear que exercen as variables medidas entre os instantes t e s sobre cada unha delas. Tamén está definida en $[-1, 1]$.

As funcións que vimos de definir dependen do proceso estocástico, que non é coñecido, polo que debemos estimalas. Pero, como xa comentamos, a serie temporal só proporciona unha observación das variables Y_t, $t \in \{1, \dots, T\}$. En consecuencia, para poder realizar as mencionadas estimacións, precisamos asumir certas suposicións que simplificarán a tarefa. Neste caso, a suposición será a estacionariedade do proceso, cuxa idea é que a lei de probabilidade que rexe o comportamento do proceso non cambia ao longo do tempo.

Definición 1.4. Un proceso estocástico $\{Y_t\}_{t\in\mathbb{Z}}$ dise estritamente estacionario se a distribución conxunta de $\{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}\}$ é a mesma que a do conxunto $\{Y_{t_1-k}, Y_{t_2-k}, \dots, Y_{t_n-k}\}$ para toda colección de instantes $t_1, t_2, \dots, t_n$ e todo retardo $k \in \mathbb{Z}$.

No caso de que $n = 1$, as variables Y's son marxinalmente identicamente distribuídas. Así, teñen a mesma función de medias e de varianzas constantes ao longo do tempo, isto é,

$$\mathbb{E}(Y_t) = \mathbb{E}(Y_{t-k}), \forall\, t,\, k,$$
$$Var(Y_t) = Var(Y_{t-k}), \forall\, t,\, k.$$

No caso de que $n = 2$, a distribución das variables Y_t e Y_s é a mesma que a de Y_{t-k} e Y_{s-k}, é dicir,

$$Cov(Y_t, Y_s) = Cov(Y_{t-k}, Y_{s-k}), \forall t,\, s,\, k.$$

A condición de estacionariedade que vimos de definir é moi forte, polo que xeralmente se emprega unha versión máis relaxada, que describimos a continuación.

Definición 1.5. Un proceso estocástico $\{Y_t\}_{t\in\mathbb{Z}}$ dise debilmente estacionario se:

1. A función de medias é constante ao longo do tempo, $\mu_t = \mu\,, \forall\, t$.
2. A función de varianzas é constante ao longo do tempo, $\sigma_t^2 = \sigma^2\,, \forall\, t$.
3. A función de autocovarianzas entre dúas variables só depende da distancia entre os instantes que as separan, $\gamma(t, t+k) = \gamma(k)\,, \forall\, t,\ k$.

A propiedade 3 tamén se da nas funcións de autocorrelacións simples e parciais. Así,

$$\rho(t, t+k) = \rho(k)\,, \forall\, t,\ k,$$
$$\alpha(t, t+k) = \alpha(k)\,, \forall\, t,\ k.$$

A partir de aquí, cada vez que mencionemos que un proceso é estacionario referímonos á estacionariedade en sentido débil.

Unha vez presentada a condición de estacionariedade, podemos estimar as funcións anteriormente definidas, dada unha serie $\{y_1, \dots, y_T\}$, como segue:

(I) Media mostral, $\bar{y} = \frac{\sum_{t=1}^{T} y_t}{T}$.

(II) Función de autocovarianzas mostrais, $\widehat{\gamma}(k) = \frac{\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{T}$.

(III) Función de autocorrelacións simples mostrais, $\widehat{\rho}(k) = \frac{\widehat{\gamma}(k)}{\widehat{\gamma}(0)}$.

(IV) Función de autocorrelacións parciais mostrais, $\widehat{\alpha}(k) = \widehat{\alpha}(k,k)$, onde $\widehat{\alpha}(k,k)$ denota o estimador mínimo cadrático de $\alpha(k,k)$ na regresión $y_t = \alpha(k,0) + \alpha(k,1)y_{t-1} + \dots + \alpha(k,k)y_{t-k} + \varepsilon$, onde ε é o erro de regresión.

Un exemplo moi importante de proceso estacionario (neste caso, no sentido estrito) é o **ruído branco**, que se define como un conxunto de variables aleatorias $\{a_t\}_t$ incorreladas, con media nula e varianza finita σ_a^2. Así,

$$\mu_t = 0\,, \forall\, t,$$
$$\sigma_t^2 = \sigma_a^2\,, \forall\, t,$$
$$\gamma(s,t) = \mathbb{E}(a_s a_t) - \mu_s\mu_t = \begin{cases} \sigma_a^2 & \text{se } s = t \\ 0 & \text{se } s \neq t. \end{cases}$$

Chegados a este punto, imos definir algunhas das posibles representacións dos procesos estocásticos.

Definición 1.6. Un proceso estocástico $\{Y_t\}_t$ dise linear se pode ser representado como unha combinación linear ponderada das observacións de ruído branco, $\{a_t\}$. É dicir, se admite a seguinte representación:

$$Y_t = \dots + \psi_{-1}a_{t+1} + c + \psi_0 a_t + \psi_1 a_{t-1} + \dots,$$

onde $\sum_{i=-\infty}^{\infty} |\psi_i| < \infty$.

Un proceso linear é estacionario, dado que é combinación linear dun proceso estacionario.

Definición 1.7. Un proceso estocástico $\{Y_t\}_t$ dise causal se pode escribirse como

$$Y_t = c + \psi_0 a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots,$$

onde $\sum_{i=0}^{\infty} |\psi_i| < \infty$.

Un proceso causal é un caso particular dun proceso linear onde non hai dependencia cos valores futuros do ruído branco.

Definición 1.8. Un proceso estocástico $\{Y_t\}_t$ dise invertible se pode representarse como

$$Y_t = c + a_t + \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \dots,$$

onde $\sum_{i=1}^{\infty} |\pi_i| < \infty$.

Teorema 1.9 (Descomposición de Wold)**.** *Se o proceso estocástico $\{Y_t\}_t$ é estacionario e, ademais, non contén compoñentes deterministas, entón pode escribirse como*

$$Y_t = \psi_0 a_t + \psi_1 a_{t-1} + \dots,$$

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

Polo tanto, calquera proceso estacionario é linear ou pode transformarse para que o sexa, eliminándolle a compoñente determinista. Os procesos lineares conforman o marco xeral para o estudo dos procesos estacionarios. Na seguinte sección presentaremos a metodoloxía Box-Jenkins, onde moitos dos seus modelos son casos particulares desta representación. Empregaremos a seguinte notación:

- $y_1, \dots, y_T$ denota a serie de tempo.
- $\{Y_t\}$ é o proceso xerador da serie temporal.
- $\{a_t\}$ denota o proceso de ruído branco e suponse que é independente de $Y_{t-1}, Y_{t-2}, \dots$

1.2. Modelos Box-Jenkins

Os modelos Box-Jenkins (tamén coñecidos como modelos ARIMA) son modelos paramétricos que se usan para modelar series de tempo (no noso caso, univariantes) e serven tanto para series estacionarias como non estacionarias, como veremos ao longo do capítulo. Estes modelos demostraron ser útiles como posibles xeradores de series de tempo reais, polo que o seu estudo é importante no noso contexto. Para o desenvolvemento desta sección empregáronse como referencias Cryer e Chan (2010) e Peña (2005).

Interésanos atopar un modelo estocástico que sexa capaz de xerar razoablemente a serie de tempo observada. A continuación presentamos as etapas a seguir no proceso de busca, proposto por Box e Jenkins (1976):

1. Identificación do modelo.
2. Estimación do modelo.
3. Validación do modelo.

No primeiro paso, seleccionase a clase de modelos que poden ser axeitados para a serie de tempo. Á hora de escoller o modelo, este debe ter o menor número de parámetros que representen de forma adecuada a serie (principio de parsimonia). Unha vez elixido o modelo, pasamos á seguinte etapa. A estimación do modelo consiste en buscar as mellores estimacións dos parámetros (descoñecidos) do mesmo. Por último, debemos validar o modelo que vimos de estimar. No caso de que este se axuste adecuadamente aos datos e que non infrinxa as condicións do mesmo, o modelo pode empregarse como xerador da serie. Noutro caso, debemos volver á primeiro etapa. O proceso remata no momento en que atopemos un modelo apropiado.

Esta metodoloxía é a que seguiremos cando precisemos atopar un proceso que poida describir as características dunha serie de tempo. Ademais, esta sección está configurada de tal forma que siga a orde destas etapas.

1.2.1. Modelos para series temporais estacionarias

A primeira etapa consiste na identificación dun modelo estocástico adecuado. Para isto, é interesante coñecer un dos modelos que resultaron ser moi prácticos en situacións reais: os modelos ARMA, cuxa estrutura consta dunha parte autorregresiva e outra parte de medias móbiles. Como complemento a esta subsección pode consultarse Cryer e Chan (2005), páxina 55.

Definición 1.10. Un proceso autorregresivo de orde p, AR(p), pode representarse como:

$$Y_t = c + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + a_t,$$

onde Y_t é un proceso estacionario, c, ϕ_i, $i \in \{1, ..., p\}$ ($\phi_p \neq 0$) son constantes e a_t é ruído branco con varianza σ_a^2.

Sexa μ a media do proceso estacionario Y_t, entón $c = \mu(1 - \phi_1 - ... - \phi_p)$.

Notemos que, por definición, Y_t debe ser un proceso estacionario, senón esta expresión non é un AR(p). A anterior representación resulta nun modelo autorregresivo de orde p (ou nun modelo estacionario) se, e só se, verifica que o seu polinomio característico, $\phi(z)$, non se anula cando z ten módulo 1, isto é, se $\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p \neq 0$, $\forall z : |z| = 1$.

Ademais, por definición, o proceso AR(p) sempre é invertible. Só debemos tomar $\pi_i = \phi_i$ cando $i \in \{1, ..., p\}$ e os restantes parámetros nulos.

Este modelo tamén é causal no caso de que o seu polinomio característico non teña raíces de módulo menor ou igual a 1, é dicir, $\phi(z) \neq 0$, $\forall z : |z| \leq 1$.

Sexa B o operador retardo, que consiste en retardar a variable á que se lle aplica nun instante de tempo, isto é, $BY_t = Y_{t-1}$. Podemos escribir de forma compacta a representación do modelo AR(p) como segue:

$$\phi(B)Y_t = c + a_t,$$

onde $\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p$.

Para identificar este modelo podemos botar man da función de autocorrelacións parciais (FAP), dado que presentan a seguinte estrutura: o último coeficiente de autocorrelación parcial non nulo dase no retardo p, que coincidiría co parámetro ϕ_p, que supuxemos non nulo.

Na seguinte definición presentamos outra subclase dos modelos Box-Jenkins para procesos estacionarios.

Definición 1.11. Un proceso de medias móbiles de orde q, MA(q), pode representarse como:

$$Y_t = c + a_t + \theta_1 a_{t-1} + ... + \theta_q a_{t-q},$$

sendo c, θ_j, $j \in \{1, ..., q\}$ ($\theta_q \neq 0$) constantes e a_t ruído branco con varianza σ_a^2.

Neste caso, a constante c é a media do proceso Y_t.

O proceso de medias móbiles de orde q sempre é linear e causal. Só é preciso tomar $\psi_j = \theta_j$ onde $j \in \{1, ..., q\}$ e os restantes coeficientes iguais a cero. Por ser un proceso linear, os modelos MA(q) sempre son estacionarios. Estes tamén son invertibles se, e só se, o seu polinomio característico verifica que $\theta(z) = 1 + \theta_1 z + ... + \theta_q z^q \neq 0$ para todo $z : |z| \leq 1$.

Tamén podemos empregar a seguinte representación do proceso MA(q):

$$Y_t = c + \theta(B)a_t,$$

con $\theta(B) = 1 + \theta_1 B + ... + \theta_q B^q$, sendo B o operador retardo.

Ao igual que co modelo autorregresivo, podemos tratar de identificar o modelo de medias móbiles de orde q a través da estrutura das súas autocorrelacións. Neste caso, a función de autocorrelacións simples (FAS) ten a seguinte particularidade: o último coeficiente de autocorrelación simple non nulo atópase na posición q.

A continuación, presentamos os procesos ARMA, que combinan a estrutura autorregresiva e de medias móbiles.

Definición 1.12. Un proceso ARMA(p, q) pode representarse como

$$Y_t = c + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + a_t + \theta_1 a_{t-1} + ... + \theta_q a_{t-q},$$

onde Y_t é un proceso estacionario, c, ϕ_i, θ_j, $i \in \{1, ..., p\}$ e $j \in \{1, ..., q\}$ ($\phi_p \neq 0$ e $\theta_q \neq 0$) son constantes e a_t é un proceso de ruído branco con varianza σ_a^2.

Se o proceso Y_t ten media μ, entón $c = \mu(1 - \phi_1 - ... - \phi_p)$.

Pódese observar, a partir da definición, que un ARMA$(p, 0)$ coincide cun AR(p) e que un ARMA$(0, q)$ é un MA(q).

Notemos que se esixe que Y_t sexa estacionario. Para que a simple representación proporcione un modelo ARMA(p, q) (ou un modelo estacionario), o polinomio característico da parte autorregresiva, $\phi(z)$, non pode anularse nos valores z de módulo unidade. Doutra banda, o modelo é causal no caso de que o polinomio característico AR non teña raíces no círculo unidade (raíces de módulo menor ou igual que 1). Ademais, é invertible se o polinomio característico MA non ten raíces no círculo unidade.

Podemos representar de forma compacta a expresión do ARMA tal e como segue:

$$\phi(B)Y_t = c + \theta(B)a_t,$$

onde B é o operador retardo.

Se ben a estrutura das autocorrelacións simples e parciais permitía identificar o modelo de medias móbiles de orde q e o modelo autorregresivo de orde p, respectivamente, isto non é tan sinxelo no caso dos ARMA(p, q). A estrutura da función de autocorrelacións simples e parciais no caso de AR(p) e MA(q), respectivamente, teñen unha estrutura complexa, polo que non é fácil recoñecer os modelos por medio delas. Dado que o ARMA é unha combinación destes dous modelos, a complexa estrutura das autocorrelacións non será de axuda á hora de identificar o modelo.

Os ARMA(p, q) son unha clase de procesos paramétricos estacionarios, que son flexibles á hora de modelar múltiples escenarios onde a serie observada é estacionaria, dado que verifica a seguinte propiedade.

Proposición 1.13. *Sexa Y_t un proceso estacionario. Se $\gamma_Y(h) \to 0$ cando $h \to \infty$, entón existe un proceso ARMA, X_t, tal que $\gamma_X(h) = \gamma_Y(h)\ \forall\, h = 0, 1, ..., k$, para calquera enteiro positivo k.*

O proceso ARMA(p, q) que vimos de definir emprégase para modelar dependencia regular, isto é, a dependencia entre observacións ou ruído branco consecutivos que se deron no pasado inmediato. Pero no caso de que exista unha dependencia estacional, é dicir, a dependencia entre observacións ou ruído branco que se deron en instantes separados por múltiplos do período estacional s, estes procesos contarían cunha gran cantidade de parámetros que, en realidade, non son necesarios. Como solución a este problema, defínense os procesos ARMA estacionais.

Definición 1.14. Un proceso autorregresivo estacional de orde P e período estacional s, AR$(P)_s$, represéntase como

$$Y_t = c + \Phi_1 Y_{t-s} + ... + \Phi_P Y_{t-Ps} + a_t,$$

expresión na que Y_t é un proceso estacionario, c, Φ_i, $i \in \{1, ..., P\}$ ($\Phi_P \neq 0$) son constantes e a_t é ruído branco con varianza σ_a^2.

Se o proceso estacionario Y_t ten media μ, entón $c = \mu(1 - \Phi_1 - ... - \Phi_P)$.

A expresión da definición pode verse como un caso particular dun proceso autorregresivo de orde $p = Ps$ con todos os coeficientes nulos a excepción daqueles que se atopan nos retardos estacionais: $s, 2s, ..., Ps$. Como se engloba nos procesos AR(p), as condicións de estacionariedade da expresión, causalidade e invertibilidade para os AR$(P)_s$ son as mesmas que para este modelo.

Podemos representar este proceso de forma compacta mediante a seguinte expresión:

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

onde $\Phi(B^s) = 1 - \Phi_1 B^s - ... - \Phi_P B^{Ps}$ e B^s denota o operador retardo estacional, que consiste en retardar a variable á que se lle aplica en s instantes de tempo, isto é, $B^s Y_t = Y_{t-s}$.

Tamén cabe mencionar a estrutura teórica da función de autocorrelacións deste modelo, dado que nos podería facilitar a identificación do mesmo. A estrutura característica dáse na función de autocorrelacións parciais, onde se verifica que o último retardo, múltiplo de s, no que non se anula é no Ps. Os coeficientes de autocorrelación parcial en retardos non múltiplos do período estacional son nulos.

Definición 1.15. Un proceso de medias móbiles estacional de orde Q e período estacional s, MA$(Q)_s$, defínese como

$$Y_t = c + a_t + \Theta_1 a_{t-s} + ... + \Theta_Q a_{t-Qs},$$

onde c, Θ_j, $j \in \{1, ..., Q\}$ $(\Theta_Q \neq 0)$ son constantes e a_t ruído branco con varianza σ_a^2.

Se μ denota a media do proceso Y_t, entón $c = \mu$.

Os procesos MA$(Q)_s$ poden verse como un caso particular dos modelos de medias móbiles de orde $q = Qs$ cos únicos coeficientes non nulos aqueles que se atopan nos retardos estacionais: s, $2s, ..., Qs$. En consecuencia, estes procesos son lineares (e, polo tanto, estacionarios) e causais. Ademais, a condición de invertibilidade é a mesma que vimos de ver para os modelos MA(q).

Botando man do operador retardo estacional, B^s, podemos expresar a representación do proceso de forma máis sinxela mediante:

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

sendo $\Theta(B^s) = 1 + \Theta_1 B^s + ... + \Theta_Q B^{Qs}$.

Por último, para a súa posible identificación, exporemos a estrutura da súa función de autocorrelacións. Neste caso, a función de autocorrelacións simples é a que presenta unha dinámica diferenciable: o último retardo, que é múltiplo do período estacional s, no que o coeficiente de autocorrelación simple non se anula dáse no Qs. Os coeficientes de autocorrelación simple que non se atopan en retardos múltiplos de s son nulos.

Da mesma forma que combinábamos os procesos autorregresivos e de medias móbiles para dar lugar aos ARMA, tamén o faremos cos correspondentes procesos estacionais, para obter os procesos ARMA estacionais.

Definición 1.16. Un proceso ARMA$(P, Q)_s$ de ordes P e Q e período estacional s defínese mediante

$$Y_t = c + \Phi_1 Y_{t-s} + ... + \Phi_P Y_{t-Ps} + a_t + \Theta_1 a_{t-s} + ... + \Theta_Q a_{t-Qs},$$

onde Y_t é un proceso estacionario, c, Φ_i, Θ_j, $i \in \{1, ..., P\}$ e $j \in \{1, ..., Q\}$ $(\Phi_P \neq 0$ e $\Theta_Q \neq 0)$ son constantes e a_t un proceso de ruído branco con varianza σ_a^2.

Se denotamos por μ a media do proceso estacionario Y_t, entón $c = \mu(1 - \Phi_1 - ... - \Phi_P)$.

Está claro que un proceso ARMA$(P, 0)_s$ coincide cun AR$(P)_s$ e un proceso ARMA$(0, Q)_s$ cun MA$(Q)_s$, isto é, os modelos autorregresivos e de medias móbiles estacionais son casos particulares dos procesos ARMAs estacionais.

Dado que os procesos ARMA estacionais son, a súa vez, un caso particular dos ARMA non estacionais de ordes sP e sQ con moitos coeficientes nulos, as condicións de estacionariedade, causalidade e invertibilidade destes coinciden coas expostas para os ARMA non estacionais.

É posible representar a expresión da definición dunha forma máis compacta:

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

onde B^s é o operador retardo estacional.

Neste caso, a estrutura teórica da función de autocorrelacións é complexa, polo que non nos permitirá a súa identificación. Esta constitúese por ter moitos coeficientes de autocorrelación simple e parcial non nulos nos retardos estacionais s, $2s, ...$ Os coeficientes das autocorrelacións nos restantes retardos son nulos.

Vimos de ver que os ARMA non estacionais modelan a dependencia regular e que os ARMA estacionais (sendo casos particulares dos primeiros, pero cun menor número de parámetros) modelan a dependencia estacional. Pero é posible modelar ambas dependencias mediante os procesos ARMA multiplicativos, que presentamos a continuación. Pode consultarse Cryer e Chan (2010), páxina 227.

Definición 1.17. Un proceso ARMA estacional multiplicativo de ordes p, q, P, Q e período estacional s, ARMA$(p,q)\times(P,Q)_s$, represéntase mediante

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

onde

$$\begin{aligned}\phi(B) &= 1 - \phi_1 B - ... - \phi_p B^p,\\ \Phi(B^s) &= 1 - \Phi_1 B^s - ... - \Phi_P B^{Ps},\\ \theta(B) &= 1 + \theta_1 B + ... + \theta_q B^q,\\ \Theta(B^s) &= 1 + \Theta_1 B^s + ... + \Theta_Q B^{Qs},\end{aligned}$$

B e B^s denotan o operador retardo e retardo estacional, respectivamente, Y_t é un proceso estacionario e a_t é ruído branco con varianza σ_a^2.

Se o proceso Y_t ten media μ, entón $c = \mu(1 - \phi_1 - ... - \phi_p)(1 - \Phi_1 - ... - \Phi_P)$.

Este proceso é un caso particular dun ARMA$(p + sP, q + sQ)$ con múltiples coeficientes nulos. De feito, o modelo ARMA estacional multiplicativo está definido con só $p+P+q+Q$ coeficientes, facéndoo máis sinxelo. En consecuencia, as condicións de estacionariedade, causalidade e invertibilidade son as definidas no caso do proceso ARMA non estacional.

Para a posible identificación deste modelo, presentamos a estrutura da función de autocorrelacións simples e parciais.

- Estrutura teórica da función de autocorrelacións simples:
  - Nos retardos baixos, 1, 2, ... , $s/2$, captamos os coeficientes de autocorrelación simple da parte regular.
  - Nos retardos estacionais, s, $2s$, ..., captamos os coeficientes de autocorrelación simple da parte estacional.
  - Nos dous lados dos retardos estacionais repítese a estrutura da función de autocorrelación simple da parte regular. No caso de que o coeficiente de autocorrelación simple no retardo estacional sexa negativo, entón a estrutura de autocorrelación simple da parte regular aparecerá invertida.
- Estrutura teórica da función de autocorrelacións parciais:
  - Nos retardos baixos, 1, 2, ... , $s/2$, captamos os coeficientes de autocorrelación parcial da parte regular.
  - Nos retardos estacionais, s, $2s$, ..., captamos os coeficientes de autocorrelación parcial da parte estacional.
  - No lado dereito dos retardos estacionais, móstrase a estrutura da función de autocorrelacións parciais da parte regular. No caso de que o coeficiente de autocorrelación parcial sexa negativo no retardo estacional, a estrutura da función de autocorrelacións parciais da parte regular estará invertida.
  - No lado esquerdo dos retardos estacionais, móstrase a estrutura da función de autocorrelacións simples da parte regular. No caso de que o coeficiente da autocorrelación parcial no retardo estacional sexa negativo, a estrutura da función de autocorrelacións simples da parte regular estará invertida.

Dado que é de gran interese identificar os procesos que vimos de ver, na Táboa 1.1 móstrase o resumo da estrutura da función de autocorrelacións dos mencionados procesos. Aínda que non se menciona na mesma, é importante lembrar que no caso dos procesos estacionais, os coeficientes nos retardos non estacionais, isto é, aqueles distintos de múltiplos de s, son nulos.

Proceso	FAS	FAP
AR(p)	Moitos coeficientes non nulos*	O último coeficiente non nulo dáse no retardo p
MA(q)	O último coeficiente non nulo dáse no retardo q	Moitos coeficientes non nulos*
ARMA(p,q)	Moitos coeficientes non nulos*	Moitos coeficientes non nulos*
AR(P)$_s$	Moitos coeficientes non nulos nos retardos estacionais	O último coeficiente non nulo dáse no retardo Ps
MA(Q)$_s$	O último coeficiente non nulo dáse no retardo Qs	Moitos coeficientes non nulos nos retardos estacionais
ARMA(P,Q)$_s$	Moitos coeficientes non nulos nos retardos estacionais	Moitos coeficientes non nulos nos retardos estacionais

Táboa 1.1: Estrutura teórica das funcións de autocorrelacións dos procesos que se mostran. * indica que a partir dos primeiros retardos, os coeficientes converxen rapidamente a cero, como suma de funcións exponenciais ou sinusoidais.

1.2.2. Modelos para series temporais non estacionarias

Vimos de estudar diferentes procesos que posibilitan a modelaxe de moitas series de tempo estacionarias. Pero na realidade estas situacións non son tan comúns. Polo tanto, é importante estender os procesos ARMA para que poidan modelar series non estacionarias. Como complemento desta subsección pode consultarse Cryer e Chan (2010), páxina 87.

Nesta sección estudaremos como tratar as seguintes causas de non estacionariedade:

1. Heterocedasticidade. A varianza da serie non é constante.

2. Tendencia. A media ou nivel da serie non é constante.

3. Compoñente estacional. Existe un patrón repetitivo na serie.

Xeralmente, no gráfico secuencial[1] da serie pode verse se esta é ou non heterocedástica. A continuación, explicamos como se pode estabilizar a súa varianza no caso de que a serie non sexa homocedástica. Este paso adoita ser a primeira transformación (no caso de que precisara máis) da serie para convertela en estacionaria.

Definición 1.18. Sexa y_t unha serie de tempo cuxas observacións son positivas, isto é, $y_t > 0$, $\forall t \in \{1, \dots, T\}$. A familia de transformacións de Box-Cox converte a serie y_t en

$$\begin{cases} \frac{y_t^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0, \\ \log(y_t) & \text{se } \lambda = 0. \end{cases}$$

[1]Representación gráfica das observacións da serie, y_t, fronte ao instante t correspondente, xuntadas mediante segmentos.

Así, no caso de que algunha das observacións da serie observada sexa cero ou negativa, debemos modificar a serie de tal forma que esta sexa positiva antes de aplicar a anterior familia de transformacións. Podemos sumarlle unha constante á serie que faga os valores desta estritamente maiores que cero para todo instante de tempo. Por exemplo, unha opción para esta constante sería $c = |\min(y_t)| + 1$. A selección dun valor exacto de λ non está garantido. Podemos facer uso da función *BoxCox.lambda* do paquete *forecast* (Hyndman et al 2020) do software estatístico libre R para obter unha estimación deste valor mediante o método de Guerrero ou máxima verosimilitude.

Doutra banda, a tendencia e compoñente estacional poden ser eliminadas mediante a aplicación de diferenzas regulares e estacionais de período s, respectivamente. Se a serie só presenta tendencia, aplicaremos d diferenzas regulares para eliminala e obtemos os modelos ARIMA(p, d, q). Doutra banda, se a serie só presenta compoñente estacional de período s, debemos aplicar D diferenzas estacionais de período s para eliminala e obtemos os modelos ARIMA$(P, D, Q)_s$. No caso de que a serie presente tanto tendencia como compoñente estacional, aplícanse d diferenzas regulares e D diferenzas estacionais e obtemos os modelos ARIMA estacionais multiplicativos.

Definición 1.19. Un proceso ARIMA(p, d, q) de ordes p, d e q, onde d é o número de diferenzas regulares que lle aplicamos á serie para eliminar a tendencia, defínese como

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

sendo Y_t un proceso estacionario e a_t un proceso de ruído branco con varianza σ_a^2.

Se μ é a media do proceso Y_t diferenciado regularmente d veces, entón $c = \mu(1 - \phi_1 - ... - \phi_p)$.

Á vista da definición, podemos ver que un proceso autorregresivo integrado de medias móbiles Y_t é un proceso que, logo de aplicarlle d diferenzas regulares, $(1 - B)^d Y_t$, é un proceso ARMA(p, q).

Definición 1.20. Un proceso ARIMA$(P, D, Q)_s$ de ordes P, D, Q e período estacional s, onde D denota o número de diferenzas estacionais de período s necesarias para eliminar a compoñente estacional da serie, pode representarse como

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

onde Y_t é un proceso estacionario e a_t é ruído branco con varianza σ_a^2.

Se o proceso Y_t diferenciado estacionalmente D veces con período s ten media μ, entón $c = \mu(1 - \Phi_1 - ... - \Phi_P)$.

É fácil ver que un proceso ARIMA$(P, D, Q)_s$ non é máis que un proceso Y_t que, logo de ser diferenciado D veces con diferenzas estacionais de período s, $(1 - B^s)^D Y_t$, é un un proceso ARMA$(P, Q)_s$.

Definición 1.21. Un proceso ARIMA$(p, d, q)\times(P, D, Q)_s$ de ordes regulares p, d, q, ordes estacionais P, D, Q e período estacional s defínese como

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

onde Y_t é un proceso estacionario e a_t ruído branco con varianza σ_a^2.

Sexa μ a media do proceso Y_t diferenciado regular e estacionalmente, entón $c = \mu(1 - \phi_1 - ... - \phi_p)(1 - \Phi_1 - ... - \Phi_P)$.

O proceso ARIMA multiplicativo estacional non é máis que un proceso que, logo de aplicarlle d diferenzas regulares e D diferenzas estacionais de período s, é un proceso ARMA$(p, d)\times(P, Q)_s$.

1.2.3. Etapa I. Identificación do modelo

Nas anteriores subseccións definimos os modelos ARIMA, que permiten modelar tanto series estacionarias como non estacionarias. Polo que agora, dada unha serie de tempo, interésanos saber se algún destes procesos puido xerala de forma razoable.

O primeiro paso sería estudar a serie de tempo de forma exploratoria, por exemplo, a través do seu gráfico secuencial e do gráfico de autocorrelacións simples e parciais e decidir se esta é ou non estacionaria. Como xa comentamos, antes de fixarnos se o nivel é constante ou se hai patrón repetitivo, é importante constatar se a varianza da serie é constante. En caso de que non o sexa, debemos aplicar a familia de transformacións Box-Cox. Unha vez que se estabilizou a variabilidade, analizamos se a serie presenta tendencia ou compoñente estacional. Isto pode verse no gráfico secuencial ou no gráfico de autocorrelacións estimadas. No caso de haber tendencia, os coeficientes de autocorrelación simple toman valores positivos altos e estes tenden a cero lentamente. No caso de que exista patrón repetitivo na serie, a compoñente estacional márcase nos coeficientes de autocorrelación de tal forma que os retardos estacionais (múltiplos de s) son positivos altos e van lentamente a cero, onde o período estacional s é o correspondente ao da serie. Como xa vimos, a tendencia elimínase a través da aplicación de d diferenzas regulares e a compoñente estacional mediante D diferenzas estacionais de período s. En xeral, se a serie presenta ambas, primeiro elimínase a tendencia e logo a compoñente estacional. A serie así transformada é estacionaria. No caso de non estar seguros de se esta provén dun proceso estacionario ou non, podemos realizar un contraste de raíces unitarias. Estes indican se se debe aplicar unha diferenza para converter esta serie en estacionaria. O máis coñecido é o contraste de Dickey-Fuller, que pode consultarse en Peña (2005), páxina 257.

O seguinte paso será identificar as ordes p, q, P e Q mediante a estrutura das funcións de autocorrelacións simples e parciais. Notemos que a estrutura que comentamos na Táboa 1.1 é teórica e nós non disporemos das autocorrelacións teóricas, senón de estimacións. Para que poidamos distinguir o modelo a través das autocorrelacións mostrais, introducimos as seguintes proposicións.

Proposición 1.22. *Se as variables do proceso Y_t teñen varianza finita e o tamaño da serie é grande (xeralmente $T \geq 100$), tense que se Y_t é un proceso ARMA$(0,0)$ entón, para todo k:*

$$\begin{cases} \widehat{\rho}(k) \approx N\left(0, 1/\sqrt{T}\right), \\ \widehat{\alpha}(k) \approx N\left(0, 1/\sqrt{T}\right). \end{cases}$$

Proposición 1.23. *Asumindo certas condicións xerais e sendo o tamaño da serie grande, se Y_t é un proceso AR(p) entón, para todo $k > p$:*

$$\widehat{\alpha}(k) \approx N\left(0, 1/\sqrt{T}\right).$$

Proposición 1.24. *Asumindo certas condicións xerais e supoñendo o tamaño da serie grande, se Y_t é un proceso MA(q), entón para todo $k > p$:*

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

Estas proposicións presentan a distribución mostral dos coeficientes de autocorrelación mostral dos procesos ARMA$(0,0)$, AR(p) e MA(q), respectivamente, e poden ser empregadas para contrastar estes modelos. Xeralmente, nos correlogramas debúxanse as bandas de non rexeitamento, dadas por estas proposicións, para o contraste de coeficientes nulos. Se o coeficiente da autocorrelación se atopa dentro das bandas, non rexeitamos a hipótese de que este é cero.

Como dixemos, estas proposicións permiten identificar modelos ARMA$(0,0)$, AR(p) e MA(q). Con este obxectivo, fixámonos nos correlogramas das autocorrelacións simples e parciais mostrais da serie estacionaria e, guiándonos pola Táboa 1.1, decidimos a orde destes modelos no caso de que só presente dependencia regular. Se a estrutura máis sinxela da función de autocorrelacións se da nas simples, fixarémonos no último coeficiente non nulo da mesma e este retardo será a orde q do proceso de medias móbiles. E analogamente, se a función de autocorrelacións parciais da serie ten a estrutura máis simple, fixámonos no último coeficiente non nulo e a súa posición será a orde p do modelo autorregresivo. En

cambio, se non hai estrutura, isto é, se todos os coeficientes de ambas funcións de autocorrelacións son nulos, entón o modelo semella ser un ARMA$(0,0)$. Da mesma forma actuariamos se a serie presenta só dependencia estacional, pero fixándonos nos coeficientes que se atopan nos retardos múltiplos do período estacional e os modelos considerados serían os modelos autorregresivos e de medias móbiles estacionais. Por exemplo, se á vista dos correlogramas a estrutura máis sinxela se presenta na estimación dos coeficientes da función de autocorrelacións parciais, consideraremos o último retardo múltiplo do período estacional no que non se anule como a orde P do modelo AR$(P)_s$. Por último, se a serie presenta tanto dependencia regular como estacional, botaremos man dos procesos multiplicativos xa expostos e lembrando que nos retardos baixos, en xeral ata o $s/2$-ésimo, obsérvase a parte regular e nos retardos múltiplos do período estacional s a parte estacional.

A selección do modelo a partir do estudo das estimacións dos coeficientes de autocorrelacións simples e parciais pode levarnos a identificar máis dun modelo, polo que é importante dispoñer de métodos que poidan propoñer modelos de forma automática e dalgún criterio de modelos óptimos.

Existen diferentes criterios como o criterio de información de Akaike (AIC) e o criterio de información bayesiano BIC, pero, neste caso, traballaremos co criterio de información de Akaike corrixido (AICc). Así, seleccionaríamos o modelo ARMA(p,q) que minimice

$$AICc(p,q) = -2\log(L(\hat{\beta})) + 2\frac{kT+k+2}{T-k-2},$$

onde L é a función de verosimilitude, $\hat{\beta}$ un vector coas estimacións de máxima verosimilitude dos parámetros do modelo (sen contar a varianza de a_t), k o número de parámetros a excepción da varianza do ruído branco e T o tamaño da serie.

Así, o criterio ten en conta tanto a calidade do axuste como o número de parámetros, sendo o óptimo aquel modelo que maximice a función de verosimilitude e sexa máis sinxelo, aínda que na práctica se busca un equilibrio entre ambas características.

1.2.4. Etapa II. Estimación do modelo

Unha vez que especificamos un modelo susceptible de xerar a serie de tempo observada, o seguinte paso é estimar os parámetros do mesmo. Dada a estreita relación entre os modelos ARIMA e ARMA (a serie diferenciada regularmente d veces e estacionalmente D veces de período s é unha serie estacionaria, polo que pode modelarse a través dos procesos ARMA regular, estacional ou multiplicativo), só necesitamos centrarnos na estimación dos parámetros deste último proceso.

Nesta subsección presentamos dous métodos para a estimación dos parámetros do modelo ARMA: mínimos cadrados e máxima verosimilitude. A exposición dunha posible alternativa, o método dos momentos, pode consultarse en Cryer e Chan (2010), páxina 149.

Recordemos que a representación do proceso ARMA(p,q) vén dada por:

$$Y_t = c + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p} + a_t + \theta_1 a_{t-1} + ... + \theta_q a_{t-q},$$

onde Y_t é un proceso estacionario e a_t é ruído branco con varianza σ_a^2. Os parámetros $c, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q, \sigma_a^2$ deben ser estimados a través dalgún dos métodos citados, obtendo $\tilde{c}, \tilde{\phi}_1, ..., \tilde{\phi}_p, \tilde{\theta}_1, ..., \tilde{\theta}_q$.

Método de mínimos cadrados

O primeiro método que exporemos é o de mínimos cadrados. A estimación dos parámetros do proceso ARMA(p,q) mediante este método vén dada pola minimización da suma de residuos ao cadrado, S. É dicir, os estimadores de mínimos cadrados verifican

$$\begin{aligned}(\hat{c}, \hat{\phi}_1, ..., \hat{\phi}_p, \hat{\theta}_1, ..., \hat{\theta}_q) &= \arg \min_{\tilde{c}, \tilde{\phi}_1, ..., \tilde{\phi}_p, \tilde{\theta}_1, ..., \tilde{\theta}_q} S(\tilde{c}, \tilde{\phi}_1, ..., \tilde{\phi}_p, \tilde{\theta}_1, ..., \tilde{\theta}_q) \\ &= \arg \min_{\tilde{c}, \tilde{\phi}_1, ..., \tilde{\phi}_p, \tilde{\theta}_1, ..., \tilde{\theta}_q} \sum_{t=1}^{T} \hat{a}_t^2,\end{aligned}$$

onde $\hat{a}_t = Y_t - (\tilde{c} + \tilde{\phi}_1 Y_{t-1} + ... + \tilde{\phi}_p Y_{t-p} + \tilde{\theta}_1 \hat{a}_{t-1} + ... + \tilde{\theta}_q \hat{a}_{t-q})$, $t \in \{1, ..., T\}$, son os residuos.

Se o proceso ARMA ten parte autorregresiva, é dicir, se $p > 0$, entón atopámonos cun pequeno óbice á hora de obter $\hat{a}_1, ..., \hat{a}_p$, dado que dependen dos valores $Y_0, Y_{-1}, ..., Y_{1-p}$, que non son observados. Unha posible solución é minimizar S a partir do instante $p+1$,

$$S_c(\tilde{c}, \tilde{\phi}_1, ..., \tilde{\phi}_p, \tilde{\theta}_1, ..., \tilde{\theta}_q) = \sum_{t=p+1}^{T} \hat{a}_t^2.$$

Ademais, $\hat{a}_{p+1}$ dependerá dos valores $\hat{a}_p, ..., \hat{a}_{p+1-q}$, que tamén dependen dos valores non observados de Y_t. Pero debemos reparar en que se coñecemos $\hat{a}_p, ..., \hat{a}_{p+1-q}$, poderíamos construír de forma iterativa os valores $\hat{a}_{p+1}, ..., \hat{a}_T$.

Como solución podemos empregar o método de mínimos cadrados condicionados, onde se obteñen as estimacións dos parámetros minimizando a función S_c suxeita a que $\hat{a}_p = ... = \hat{a}_{p+1-q} = 0$.

Método de máxima verosimilitude

Os estimadores de máxima verosimilitude obtéñense maximizando a función de máxima verosimilitude, L,

$$(\hat{c}, \hat{\phi}_1, ..., \hat{\phi}_p, \hat{\theta}_1, ..., \hat{\theta}_q, \hat{\sigma}_a^2) = \arg \max_{\tilde{c}, \tilde{\phi}_1, ..., \tilde{\phi}_p, \tilde{\theta}_1, ..., \tilde{\theta}_q, \tilde{\sigma}_a^2} L_{y_1, ..., y_T}(\tilde{c}, \tilde{\phi}_1, ..., \tilde{\phi}_p, \tilde{\theta}_1, ..., \tilde{\theta}_q, \tilde{\sigma}_a^2),$$

que non é máis que a función de densidade conxunta

$$L_{y_1, ..., y_T}(\tilde{c}, \tilde{\phi}_1, ..., \tilde{\phi}_p, \tilde{\theta}_1, ..., \tilde{\theta}_q, \tilde{\sigma}_a^2) = f_{\tilde{c}, \tilde{\phi}_1, ..., \tilde{\phi}_p, \tilde{\theta}_1, ..., \tilde{\theta}_q, \tilde{\sigma}_a^2}(y_1, ..., y_T)$$

dun vector aleatorio $(\tilde{Y}_1, ..., \tilde{Y}_T)^T$ dun proceso ARMA(p, q) e considerando fixas as observacións $y_1, ..., y_T$.

Baixo certas condicións pode probarse que se o proceso ARMA(p, q) é gaussiano, entón os estimadores de máxima verosimilitude dos parámetros $c, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q$ son asintoticamente óptimos, é dicir, estes son insesgados, eficientes e a súa distribución é gaussiana (o que nos permite construír rexións de confianza para contrastar a significacións dos parámetros) cando T é grande. Ademais, o estimador da varianza do ruído branco, σ_a^2, é consistente. En caso de que o proceso non sexa gaussiano, non se verifica a propiedade de eficiencia. Pode consultarse Cryer e Chan (2010), páxina 160, para afondar nas propiedades dos estimadores.

1.2.5. Etapa III. Validación do modelo

O terceiro paso, logo de estimar o modelo proposto, é a validación do mesmo. Isto consiste en comprobar se o modelo verifica a hipótese de que as súas innovacións a_t son ruído branco. É dicir, debemos contrastar se as innovacións teñen media cero, varianza constante e que sexan incorreladas. No caso de que se incumpra algunha destas condicións, o modelo axustado non se pode considerar como xerador da serie de tempo observada e, en consecuencia, debemos volver á primeira etapa e así sucesivamente ata que o modelo especificado sexa válido.

Sería desexable, aínda que non necesario, que as innovacións a_t fosen tamén gaussianas, pois baixo normalidade a incorrelación equivale a independencia e, ademais, os estimadores de máxima verosimilitude son asintoticamente eficientes.

Dado que as innovacións son descoñecidas, centrámonos nos residuos do modelo, pois no caso de que o modelo estea correctamente especificado e as estimacións dos parámetros do mesmo estean preto dos valores reais, os residuos deberían presentar as características do ruído branco.

Podemos botar man dalgunhas ferramentas gráficas para orientarnos sobre se os residuos cumpren as hipóteses. Por exemplo, no gráfico secuencial dos residuos do modelo podemos ver se a serie presenta tendencia, compoñente estacional ou heterocedasticidade. A simple presenza dunha destas características invalidaría o modelo. Para a hipótese de normalidade podemos empregar un gráfico

qq (cuantil-cuantil), no que se representan os cuantís mostrais fronte aos cuantís dunha distribución normal estándar. Se o gráfico é aproximadamente linear, este suxeriría que os residuos son gaussianos.

As ferramentas gráficas outórgannos unha intuición sobre as características dos residuos, pero a súa interpretación é subxectiva, polo que é recomendable o uso de contrastes de hipóteses. A continuación presentamos algúns contrastes de independencia, de media cero e de normalidade. Xeralmente, a homocedasticidade da serie compróbase a través do gráfico secuencial.

Contraste de independencia

No gráfico de estimacións das funcións de autocorrelacións simples e parciais xunto coas bandas de aceptación, só contrastamos correlación a correlación se estas son nulas. Pero é interesante dispoñer dun contraste máis potente que permita contrastar se H coeficientes de autocorrelación son nulos. Aquí entra en xogo o contraste de Ljung-Box, cuxo estatístico é:

$$Q_H = T(T+2)\sum_{k=1}^{H}\frac{\hat{\rho}(k)^2}{T-k}.$$

Supoñendo que o tamaño mostral, T, é grande e baixo a hipótese nula de que a mostra provén de variables aleatorias independentes e identicamente distribuídas con varianza finita, o estatístico de contraste segue unha distribución χ^2_H. Polo tanto, rexeitaremos a hipótese nula de independencia a un nivel de significación $\alpha = 0.05$ se o estadístico Q_H é maior ou igual que o cuantil 0.95 da distribución χ^2 con H graos de liberdade.

Debemos notar que se axustamos, por exemplo, un ARMA(p, q) sen constante e queremos contrastar se os seus residuos son incorrelados, podemos aplicar este contraste, pero modificando os graos de liberdade do estatístico de contraste para que se teña en conta a estimación dos parámetros. Así, no noso exemplo, os graos de liberdade serían $H - p - q$. Para tratar o tema en máis profundidade, pode consultarse Peña (2005), páxina 315.

Contraste de media nula

Presentamos agora o contraste de media nula, que pode empregarse no caso particular da constatación de se a media dos residuos do modelo que axustamos é cero.

Supoñendo que o tamaño da mostra é grande e baixo a hipótese nula de que a mostra $y_1, \ldots, y_T$ provén de variables aleatorias independentes e identicamente distribuídas con media nula e varianza finita, verifícase que

$$\frac{\bar{y}}{s_y/\sqrt{T}} \approx t_{T-1} \approx N(0,1),$$

onde $\bar{y}$ denota a media da mostra e s_y a desviación típica mostral. Así, rexeitaremos a hipótese nula de media nula a un nivel de significación $\alpha = 0.05$ se $|\bar{y}| \geq 1.96\frac{s_y}{\sqrt{T}}$. Pode consultarse Peña (2005), páxina 318, para afondar neste contraste.

Contraste de normalidade

No caso do contraste de normalidade dunha mostra, podemos empregar os tests de Jarque-Bera e de Shapiro-Wilk, entre outros. O primeiro basease nos coeficientes de curtose e asimetría e o segundo no gráfico cuantil-cuantil. Poden consultarse os artigos Jarque e Bera (1987) e Shapiro e Wilk (1965) para a súa descrición.

1.2.6. Predición

Vimos de describir os tres pasos fundamentais da metodoloxía Box-Jenkins: identificación do modelo, estimación dos seus parámetros e comprobación de que o modelo axustado puido xerar a serie de

tempo da que dispoñemos. Unha vez chegados a este punto, pode ser de gran interese facer predición dos valores futuros da serie.

O obxectivo é predicir o valor de Y_{T+h}, isto é, coñecer o valor da dita variable que pasará en h instantes de tempo no futuro en relación á serie temporal observada $\{y_1, \dots, y_T\}$. Ao instante h coñécese como horizonte de predición e denotaremos por $\hat{y}_{T+h} = \hat{y}_T(h)$ á predición con orixe T e horizonte h.

Como pode verse en Cryer e Chan (2010), páxina 218, o preditor que minimiza o erro cadrático medio da predición vén dado por

$$\hat{y}_T(h) = \mathbb{E}(Y_T(h)|Y_1, \dots, Y_T).$$

No caso de que o modelo axustado fose un ARMA(p, q), a predición con orixe T e horizonte h é

$$\hat{y}_T(h) = \hat{c} + \phi_1\hat{y}_T(h-1) + \dots + \phi_p\hat{y}_T(h-p) + \theta_1\mathbb{E}(a_{T+h-1}|Y_1, \dots, Y_T) - \dots - \theta_q\mathbb{E}(a_{T+h-q}|Y_1, \dots, Y_T),$$

onde

$$\mathbb{E}(a_{T+i}|Y_1, \dots, Y_T) = \begin{cases} 0 & \text{se } i > 0, \\ a_{T+i} & \text{se } i \leq 0. \end{cases}$$

Notemos que $\hat{y}_T(i) = y_{t+i}$ cando $i \leq 0$. Ademais, precisaremos estimar os valores a_{t+i} con $i \leq 0$. Por exemplo, no caso dun modelo invertible, $Y_t = c + a_t + \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \dots$, a_t pode expresarse mediante unha combinación linear ponderada de pesos π da secuencia Y_t, Y_{t-1}, $Y_{t-2}, \dots$

Convén acompañar as predicións puntuais de intervalos de predición para estudar a precisión das mesmas. No caso de que as innovacións do modelo axustado sexan gaussianas e T sexa grande, tense que o erro de predición a horizonte h cumpre

$$e_T(h) = Y_{T+h} - \bar{Y}_T(h) \approx N(0, \sigma_a^2(1 + \psi_1^2 + \dots + \psi_{h-1}^2)),$$

onde ψ_i son os parámetros da expresión $Y_t = c + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots$ En consecuencia, o intervalo de predición ao 95 % para o valor Y_{T+h} é

$$\left(\bar{Y}_T(h) \pm 1{,}96\sqrt{\sigma_a^2(1 + \psi_1^2 + \dots + \psi_{h-1}^2)}\right).$$

Para afondar máis nos intervalos de predición, pode verse Cryer e Chan (2010), páxina 204.

No caso de que as innovacións do modelo non sexan gaussianas, podemos empregar o seguinte procedemento:

1. Simular os k valores futuros da serie, $y_{T+1}, \dots, y_{T+k}$.
2. Repetir o primeiro paso B veces, obtendo B posibles futuros,

   $$\{y_{T+1}^{(1)}, \dots, y_{T+k}^{(1)}\}, \dots, \{y_{T+1}^{(B)}, \dots, y_{T+k}^{(B)}\}.$$

   Estes obtéñense aplicando *bootstrap* sobre os residuos.
3. Os extremos dos intervalos obtéñense a través dos cuantís de orde α e $1 - \alpha$ de $\{y_{T+h}^{(j)}\}_{j=1}^{B}$ para cada horizonte $1 \leq h \leq k$.

1.3. Descomposición de series de tempo

Xeralmente descomponse unha serie temporal en compoñente de tendencia, estacional e irregular. Describiremos brevemente dous mecanismos: a descomposición clásica de series temporais e a descomposición STL.

1.3.1. Descomposición clásica

Nesta subsección presentamos unha breve descrición da descomposición clásica de series temporais e do modelo de medias móbiles. Para unha explicación máis extensa, pode consultarse Peña (2005), páxina 58.

Sexa $\{y_t,\, t = 1, \dots, T\}$ unha serie de tempo observada. Supoñamos que esta se pode expresar como función de tres compoñentes:

- Compoñente de tendencia, T_t. Esta modela o comportamento a longo prazo da serie temporal.
- Compoñente estacional, S_t. Esta modela o comportamento periódico da serie, é dicir, o patrón repetitivo que se da cada s instantes de tempo. s denomínase período estacional e verifícase que $S_t = S_{t-s} = S_{t+s}$.
- Compoñente irregular, e_t. Esta modela os efectos aleatorios da serie e que as restantes compoñentes non poden explicar.

Neste contexto, os modelos de descomposición de series temporais máis coñecidos son:

- O modelo aditivo, cuxa expresión é $y_t = T_t + S_t + e_t$, onde a compoñente irregular se move ao redor de cero.
- O modelo multiplicativo pode representarse por $y_t = T_t \times S_t \times e_t$, onde a compoñente irregular se move ao redor de un.

Notemos que o modelo multiplicativo coincidiría co modelo aditivo se aplicamos logaritmos.

Escolleremos aquel modelo que poida reunir as principais características da serie temporal adecuadamente. Isto adoita facerse á vista do gráfico secuencial da serie. En xeral, o modelo aditivo está indicado cando a magnitude das oscilacións estacionais non varían ao facelo a tendencia e o modelo multiplicativo cando a magnitude das oscilacións estacionais crece e decrece proporcionalmente cos crecementos e decrecementos da tendencia, respectivamente. Logo de escoller o modelo máis axeitado para a descrición da serie, débense estimar as compoñentes.

Medias móbiles

Á hora de estimar as compoñentes do modelo escollido, poden empregarse dúas vías: métodos paramétricos ou non paramétricos. En xeral, os métodos paramétricos consisten en expresar a relación entre a tendencia e a compoñente estacional co tempo mediante un modelo paramétrico, para logo axustalo á serie. Finalmente, estímanse T_t e S_t polas correspondentes compoñentes axustadas e a compoñente irregular, e_t, despexándoa da expresión do modelo e substituíndo as compoñentes de tendencia e estacional polas súas estimacións. Doutra banda, os métodos non paramétricos supoñen que a relación entre a tendencia e o tempo é suave e aplican algún método de suavización ás observacións da serie para estimar a tendencia e a compoñente estacional. A compoñente irregular estímase da mesma forma que no método paramétrico.

O método de suavización que describiremos é o de medias móbiles, que se encadra nos métodos non paramétricos.

Sexa $k = 2q+1$ un número impar. A aplicación dunha suavización k- MA *(moving average)* á serie $\{y_t\}$ no instante t consiste na transformación do valor y_t na estimación

$$\frac{y_{t-q} + \dots + y_t + \dots + y_{t+q}}{k}.$$

No caso de que $k = 2q$ sexa par, a aplicación dunha suavización $2 \times k$ - MA á serie $\{y_t\}$ no instante t consiste en transformar o valor y_t na estimación

$$\frac{0.5y_{t-q} + y_{t-(q-1)} + \dots + y_t + \dots + y_{t+(q-1)} + 0.5y_{t+q}}{k}.$$

Así, por exemplo, a suavización mediante medias móbiles para unha serie homocedástica con tendencia e compoñente estacional cuxa descomposición é aditiva, consistiría en supoñer que $S_1+...+S_s = 0$ e estimar a tendencia T_t mediante medias móbiles. Obteríamos unha nova serie subtraendo a estimación da tendencia á serie orixinal, $y_t - \widehat{T}_t \approx S_t + e_t$. Logo estimaríamos S_j, onde $j = 1, ..., s$, restándolle á media de $\{y_{j+is} - \widehat{T}_{j+is}\}_i$ a media de $\{y_t - \widehat{T}_t\}_t$.

Empregaranse medias móbiles, aínda que non neste contexto, para a suavización da demanda de electricidade e eliminar, así, a compoñente estacional semanal.

1.3.2. Descomposición STL

STL (*Seasonal-Trend decomposition procedure based on Loess*) é un procedemento de descomposición de series temporais estacionais nas 3 compoñentes que vimos de describir na anterior subsección. Este consiste na aplicación da suavización *loess* (Cleveland 1979), que se presenta na Subsección 2.2.2, e de medias móbiles (ver subsección anterior). Así, as estimacións das compoñentes de tendencia e estacional son robustas, polo que non se ven prexudicadas por observacións atípicas. Pode consultarse Cleveland et al (1990).

Como xa mencionamos, descompoñemos a serie temporal observada, Y_t, en compoñente de tendencia, T_t, estacional, S_t, e irregular, e_t, polo que se ten:

$$Y_t = T_t + S_t + e_t,\ t = 1, ..., T.$$

O método STL conta con dous bucles: un interno e outro externo. No bucle interno calcúlanse as compoñentes de tendencia e estacional. Doutra banda, o bucle externo consiste na execución do bucle interno seguido do cálculo de pesos de 'robustez', que serán empregados na seguinte iteración do bucle interno co obxectivo de reducir o efecto das observacións atípicas nas compoñentes de tendencia e estacional. Na primeira iteración consideraranse os pesos de robustez iguais a 1.

Bucle interno

O bucle interno consta das seguintes etapas, descritas para a k-ésima iteración:

1. Axústase a compoñente de tendencia como segue: $Y_t - T_t^{(k-1)} = S_t^{(k)} + e_t^{(k)}$, onde se considera $T_t^{(0)} = 0$.
2. Cada subserie $Y_t - T_t^{(k-1)}$ de cada ciclo considerado (por exemplo, cada semana para observacións diarias) suavízase mediante a aplicación do *loess*. Obtense, así, unha primeira compoñente estacional $S_{p,t}^{(k)}$.
3. A continuación, aplícase un filtro *low-pass* consistente en medias móbiles e suavización *loess* a $S_{p,t}^{(k)}$, obtendo $L_t^{(k)}$.
4. Obtense a compoñente estacional: $S_t^{(k)} = S_{p,t}^{(k)} - L_t^{(k)}$.
5. Axústase estacionalmente a serie temporal Y_t: $Y_t^{(k)} - S_t^{(k)}$.
6. Obtense a compoñente de tendencia, $T_t^{(k)}$, aplicando un filtro *loess* a $Y_t^{(k)} - S_t^{(k)}$.

Bucle externo

No caso de que sexa necesario pola presenza de atípicos na serie temporal, calcúlanse os pesos de robustez do *loess* do bucle interno para cada observación como función do tamaño da compoñente irregular $e_t = Y_t - T_t - S_t$, que se calcula a partir das estimacións resultantes no bucle interno. Nos pasos 2 e 6, empregaranse como pesos da regresión *loess* os anteriores pesos multiplicados polos pesos de robustez dados pola función *biweight* de Tukey.

Para unha discusión sobre a escolla dos parámetros deste procedemento, como o número de iteracións de cada bucle, pode consultarse en Cleveland et al (1990), así como para unha descrición máis completa do método.

Ao longo do traballo faremos uso da función *tsoutliers* do paquete *forecast* (pode consultarse Hyndman et al 2020) para a detección de atípicos. Esta función emprega para a identificación de atípicos e a estimación de substitutos, no caso de que a serie de tempo sexa estacional, o procedemento STL.

Capítulo 2

Modelos de regresión

Neste capítulo revísanse os coñecidos modelos de regresión lineais como introdución, para logo desenvolver o modelo de regresión local polinómico no ámbito non paramétrico, onde non se fan suposicións sobre a forma da función de regresión, e os modelos aditivos no ámbito semiparamétrico, que combinan os enfoques paramétrico e non paramétrico para solucionar o problema da maldición da dimensión, resultado dunha suavización con múltiples variables explicativas.

Un modelo de regresión (en media) ten como obxectivo explicar a relación funcional entre a media dunha variable resposta Y, que será a variable de interese, e un conxunto de variables explicativas $X_1, \dots, X_k$. Se dispoñemos dunha mostra do vector $(X_1, \dots, X_k, Y)$, queremos determinar a función m, $\mathbb{E}(Y|X_1, \dots, X_k) = m(X_1, \dots, X_k)$. Polo tanto, os modelos de regresión empréganse para coñecer a maneira en que a variable Y depende das variables explicativas e, unha vez estimada m, realizar predicións do valor de Y dados uns valores de $X_1, \dots, X_k$.

2.1. Modelos de regresión paramétricos

Nesta sección revisamos brevemente os modelos de regresión linear cunha e múltiples variables explicativas. Estes sitúanse no marco paramétrico, onde se asume que a función de regresión, m, pertence a unha familia indexada por un vector de parámetros e estes estímanse de tal forma que se axusten ás respostas observadas. Por exemplo, no contexto dos modelos de regresión linear simple asúmese homocedasticidade, normalidade e independencia dos erros e que a función de regresión é linear. Os modelos paramétricos poden axustarse adecuadamente se as hipóteses de partida se cumpren. Pero se se infrinxe algunhas delas, as estimacións poden ser inconsistentes e proporcionarían unha relación entre as variables errónea.

2.1.1. Modelo de regresión lineal simple

Supoñamos que dispoñemos de dúas variables aleatorias distintas, Y e X, características dunha poboación e que nos interesa coñecer a relación entre ambas. Para isto empregaremos un modelo de regresión.

A regresión vén dada pola función

$$m(x) = \mathbb{E}(Y|X = x) \text{ para cada valor } x \text{ de } X,$$

isto é, a media condicionada da variable resposta Y en función do valor x da variable explicativa X. Así, podemos escribir a variable resposta como

$$Y = m(X) + \varepsilon,$$

onde ε é o erro de regresión e verifica $\mathbb{E}(\varepsilon|X = x) = 0$ para todo valor x de X.

No caso do modelo de regresión lineal simple, tense que

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

onde β_0 é a ordenada na orixe e se coñece como intercepto, β_1 é a pendente e ε o erro de regresión.

Supoñamos que dispoñemos dunha mostra $(x_1, Y_1), \dots, (x_n, Y_n)$ baixo deseño fixo. Queremos estimar os parámetros de regresión, β_0 e β_1, en base a esta mostra e denotaremos os ditos estimadores por $\widehat{\beta}_0$ e $\widehat{\beta}_1$. Dado un valor observado x_i de X, obteríamos a predición $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ da variable resposta, denotando por Y_i a observación e $\widehat{Y}_i$ a predición. Coñécense como residuos da regresión aos erros de predición,

$$\widehat{\varepsilon} = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i, \ \forall i \in \{1, \dots, n\}.$$

Estimaremos os ditos parámetros mediante o método de mínimos cadrados, cuxos estimadores terían asociados os residuos mínimos. Así, os estimadores mínimo cadrático defínense como

$$\sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 x_i\right)^2.$$

Unha vez feita a minimización, obtemos os estimadores

$$\widehat{\beta}_0 = \bar{Y} - \frac{S_{xy}}{S_x^2}\bar{x} \in N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}\right)\right) \text{ e } \widehat{\beta}_1 = \frac{S_{xy}}{S_x^2} \in N\left(\beta_1, \frac{\sigma^2}{nS_x^2}\right),$$

denotando con $\bar{Y}$ e $\bar{x}$ a media mostral da variable resposta e explicativa, respectivamente, sendo S_{xy} a covarianza entre X e Y, S_x^2 a varianza de X e σ^2 a varianza do erro.

A estimación dos parámetros que vimos de ver está sustentada nas seguintes hipóteses:

1. Linearidade. A función de regresión é unha recta, isto é, $m(x) = \beta_0 + \beta_1 X$.
2. Homocedasticidade. A varianza do erro é constante, $Var(\varepsilon | X = x) = \sigma^2$ para todo valor x de X.
3. Normalidade. A distribución do erro é gaussiana, $\varepsilon \in N(0, \sigma^2)$.
4. Independencia. Os erros $\varepsilon_1, \dots, \varepsilon_n$ son mutuamente independentes.

Polo tanto, é importante estudar se se cumpren estas suposicións analizando os residuos de regresión.

Mediante o coeficiente de determinación podemos medir o axuste dun modelo de regresión. Para a súa definición, partimos dun modelo de regresión linear simple, $Y = \beta_0 + \beta_1 + \varepsilon$, e dunha mostra baixo deseño fixo. Na Táboa 2.1 móstrase a táboa da análise da varianza. Nesta táboa preséntase a suma total de cadrados da variable Y descomposta en dous sumandos, un representando a variación debida á regresión e outro a variación debida ao erro.

O coeficiente de determinación dun modelo de regresión defínese como a proporción de varianza explicada e plasma a proximidade das observacións ao modelo. Así, este vén dado por

$$R^2 = 1 - \frac{RSS}{TSS},$$

onde $RSS = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$ é a suma residual de cadrados e $TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ é a suma total de cadrados. Un valor próximo a 1 reflicte un modelo de regresión que axusta adecuadamente os datos e que a variable explicativa X está moi relacionada con Y e deixa pouco erro.

No caso de que busquemos comparar modelos de regresión con distinto número de variables explicativas, podemos botar man do coeficiente de determinación axustado, que se define como

$$R^2_{\text{axustado}} = 1 - \frac{RSS/(n-p)}{TSS/(n-1)},$$

denotando con p o número de parámetros de regresión estimados.

Variación	Suma de cadrados	Graos de liberdade
Regresión	$\sum_{i=1}^{n}\left(\widehat{\beta}_0+\widehat{\beta}_1 x_i-\bar{Y}\right)^2$	1
Erro	$\sum_{i=1}^{n}\left(Y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_i\right)^2$	$n-2$
Total	$\sum_{i=1}^{n}\left(Y_i-\bar{Y}_i\right)^2$	$n-1$

Táboa 2.1: Táboa da análise da varianza.

2.1.2. Modelo de regresión lineal xeral

O modelo de regresión linear simple é un modelo moi sinxelo que só conta cunha variable explicativa, pero en moitos escenarios precisaremos de varias variables independentes para poder explicar a variable dependente. Neste contexto emprégase o modelo de regresión linear xeral. Supoñamos que dispoñemos dunha variable resposta Y e un conxunto de $p-1$ variables explicativas, $X_1,\dots,X_{p-1}$. Neste caso podemos escribir

$$Y=\beta_0+\beta_1X_1+\dots+\beta_{p-1}X_{p-1}+\varepsilon,$$

onde $\beta_0,\beta_1,\dots,\beta_{p-1}$ son o intercepto e os parámetros asociados á variable explicativa correspondente e ε o erro.

Se dispoñemos dunha mostra $(x_{1,1},\dots,x_{1,p-1},Y_1),\dots,(x_{n,1},\dots,x_{n,p-1},Y_n)$ e denotamos por $\mathbf{x_i}=(1,x_{i,1},\dots,x_{i,p-1})$ ao vector asociado ás observacións das variables explicativas do individuo i-ésimo e $\beta=(\beta_0,\dots,\beta_{p-1})$ ao vector de coeficientes, podemos expresar a función de regresión como

$$m(x_i)=\mathbf{x_i}^T\beta=\beta_0+\beta_1x_{i,1}+\dots+\beta_{p-1}x_{i,p-1}.$$

Os erros de regresión seguen unha distribución normal $\epsilon_1,\dots,\epsilon_n\in N(0,\sigma^2)$ e son mutuamente independentes. Polo tanto, verifican as hipóteses de normalidade, homocedasticidade e independencia. Seguindo esta notación podemos representar o modelo de regresión lineal múltiple de forma matricial

$$\begin{pmatrix}Y_1\\ \vdots\\ Y_n\end{pmatrix}=\begin{pmatrix}1 & x_{1,1} & \cdots & x_{1,p-1}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n,1} & \cdots & x_{n,p-1}\end{pmatrix}\begin{pmatrix}\beta_0\\ \vdots\\ \beta_{p-1}\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{pmatrix}=\mathbb{X}\beta+\varepsilon,$$

onde $\mathbb{X}$ denota a matriz de deseño do modelo.

A estimación do vector de parámetros β faise mediante o método de mínimos cadrados. Así, o estimador $\widehat{\beta}$ debe verificar:

$$\min_{\beta}(Y-\mathbb{X}\beta)^T(Y-\mathbb{X}\beta).$$

Derivando a función $(Y-\mathbb{X}\beta)^T(Y-\mathbb{X}\beta)$ respecto β e igualando a cero obtemos as ecuacións normais da regresión: $\mathbb{X}^T\mathbb{X}\beta=\mathbb{X}^TY$, cuxa solución é o estimador de β por mínimos cadrados, $\widehat{\beta}=(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^TY$. Para un desenvolvemento máis detallado da tería deste modelo de regresión, pode consultarse Johnson e Wichern (2007).

2.2. Modelos de regresión non paramétricos e semiparamétricos

Na anterior sección presentamos os modelos de regresión paramétrica máis básicos. Estes asumen certas hipóteses que poden non ser certas pola propia índole dos datos ou moi ríxidas á hora de explicar a relación funcional entre as variables, conducíndonos a conclusións erróneas. Entran aquí en xogo os modelos de regresión non paramétricos, que só precisan dalgunhas hipóteses de regularidade razoables e proporcionan unha visión máis flexible ao deixar falar aos datos por si mesmos. Isto é, a forma da función de regresión vén dada a partir da información dos datos mostrais baixo hipóteses xerais de regularidade. Porén, poden ser difíciles de interpretar e proporcionar estimacións inadecuadas se o número de variables explicativas é grande. Este último problema coñécese como a maldición da dimensión e a regresión semiparamétrica xorde para darlle solución, mesturando ambos enfoques: paramétrico e non paramétrico.

Neste capítulo só presentamos o modelo de regresión polinómico local no ámbito non paramétrico e os modelos aditivos no ámbito semiparamétrico. Poden consultarse en Härdle et al (2004), páxina 85, outros modelos baixo este enfoque, como os *splines* de suavización.

Sexan Y e X dúas variables aleatorias con función de densidade conxunta $f(x,y)$. A función de regresión pode expresarse como

$$m(x) = \mathbb{E}(Y|X=x) = \int yf(y|x)dy = \int y\frac{f(x,y)}{f_X(x)}dy,$$

onde $f(y|x)$ é a función de densidade condicional de Y dado $X=x$ e f_X é a función de densidade marxinal de X. Esta representa a relación das variables en media, polo que o interese radica en estimar m.

2.2.1. Modelo de regresión polinómico local

Dada unha mostra $\{(X_i,Y_i),\, i=1,\dots,n\}$, os procedementos de estimación non paramétrica locais suavizan estes datos mediante a aproximación da variable resposta nun punto x en función das respostas observadas para valores próximos a x.

A regresión polinómica local, como ben indica o seu nome, ciméntase en axustes locais por mínimos cadrados ponderados pero con polinomios de orde d. Notemos que se a función de regresión $m(x)$ ten d derivadas, entón o teorema de Taylor afirma que m se pode aproximar mediante un polinomio de grao d nunha veciñanza de x:

$$m(t) \approx m(x) + m'(x)(t-x) + \dots + \frac{1}{d!}m^{(d)}(x)(t-x)^d,$$

para algún t nunha contorna de x. Sexa $\beta(x) = (\beta_0(x), \beta_1(x), \dots, \beta_d(x))^T$ o vector de parámetros, onde

$$\beta_j(x) = \frac{m^{(j)}(x)}{j!},\ j \in \{0,1,\dots,d\}.$$

Estimaremos o vector de parámetros mediante o método de mínimos cadrados ponderados. Así, os estimadores serán aqueles que minimicen a función

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=0}^{d}\beta_j(x)(X_i-x)^j\right)^2 K_h(X_i-x),$$

onde $K(u)$ é a función *kernel* (xeralmente unha densidade simétrica respecto 0) e adoita escribirse como $K_h(u) = \frac{1}{h}K(\frac{u}{h})$, sendo $h>0$ o parámetro de suavizado, coñecido como xanela, que regula o tamaño da veciñanza que se emprega á hora de axustar o modelo. A selección deste parámetro é fundamental

para un axuste axeitado, pero non tanto a selección da función *kernel*. Pode verse unha discusión sobre a escolla da función tipo núcleo en Härdle et al (2004), páxina 57.

Sexan

$$\mathbf{X}_x = \begin{pmatrix} 1 & (X_1 - x) & \cdots & (X_1 - x)^d \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (X_n - x) & \cdots & (X_n - x)^d \end{pmatrix}, \ \mathbf{W}_x = \begin{pmatrix} K_h(X_1 - x) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K_h(X_n - x) \end{pmatrix}$$

e $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ o vector de respostas. Entón, o estimador de mínimos cadrados ponderados será aquel que verifique

$$\min_{\beta} \left(\mathbf{Y} - \mathbf{X}_x\beta(x)\right)^T \mathbf{W}_x \left(\mathbf{Y} - \mathbf{X}_x\beta(x)\right).$$

A solución á anterior minimización é

$$\widehat{\beta}(x) = (\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x\mathbf{Y}.$$

Se denotamos por $\widehat{\beta}_0(x), \ldots, \widehat{\beta}_d(x)$ ás compoñentes do vector $\widehat{\beta}(x)$, verifícase para todo $j \in \{0, 1, \ldots, d\}$ que

$$m^{(j)}_{PL,d,n}(x) = j!\widehat{\beta}_j(x)$$

é un estimador da j-ésima derivada de m e denomínase estimador da regresión polinómico local ponderado de grao d. En particular, $\widehat{m}^{(0)}_{PL,d,n}(x) = \widehat{\beta}_0(x)$ é o estimador polinómico local da función de regresión m. A estimación depende da xanela h, que controla o grado de suavidade e, polo tanto, a súa selección é de grande importancia.

No caso de que $d = 0$, o polinomio que se axusta é unha constante e $\widehat{m}^{(0)}_{PL,0,n}(x)$ coincide co estimador de Nadaraya-Watson, que pode consultarse en Härdle et al (2004), páxina 90. Se $d = 1$, axústase unha recta localmente e o estimador resultante, que denotaremos por $\widehat{m}_{LL,n}(x)$, coñécese como estimador da regresión lineal local ponderado. Na Táboa 2.2 móstranse o sesgo e a varianza asintóticos destes dous estimadores en deseño aleatorio e baixo certas condicións de regularidade. f denota a función de densidade da variable explicativa, $\mu_2(K) = \int u^2K(u)du$ ao momento de orde 2 da función *kernel* e $R(K) = \int K^2(u)du$ á integral do cadrado de K. O estimador lineal local presenta unhas mellores características en canto a sesgo asintótico e, ademais, ao igual que os estimadores polinómico locais, non sofre o efecto fronteira.

Estimador	Sesgo	Varianza
Nadaraya-Watson	$\frac{h^2}{2}\left(m''(x) + \frac{m'(x)f'(x)}{f(x)}\right)\mu_2(K)$	$\frac{1}{nh}\frac{\sigma^2(x)}{f(x)}R(K)$
Local lineal	$\frac{h^2}{2}m''(x)\mu_2(K)$	$\frac{1}{nh}\frac{\sigma^2(x)}{f(x)}R(K)$

Táboa 2.2: Sesgo e varianza asintóticos dos estimadores.

Selección do grao do polinomio

Para a estimación da función de regresión é necesario especificar tanto o parámetro de suavizado como o grao do polinomio a axustar localmente, así como a función tipo núcleo.

Para a discusión da escolla do grao d do polinomio, presentamos as expresións do sesgo e varianza asintóticas do estimador.

- Se x é un punto do interior do dominio da variable X, baixo certas condicións sobre a función de regresión m e de densidade da variable explicativa X, f, tense a seguinte expresión para o sesgo asintótico condicionado á mostra,

$$\mathbb{E}(\widehat{m}^{(0)}_{PL,d,n}(x) - m(x)|X_1, \dots, X_n) =$$
$$\begin{cases} h^{d+1} m^{(d+1)}(x) \mu_{d+1}(K_{(d)})/(d+1)! + o(h^{d+1}) & \text{se } d \text{ é impar} \\ \left(\frac{m^{(d+1)}(x) f'(x)}{f(x)(d+1)!} + \frac{m^{(d+2)}(x)}{(d+2)!} \right) h^{d+2} \mu_{d+2}(K_{(d)}) + o(h^{d+2}) & \text{se } d \text{ é par,} \end{cases}$$

  onde $\mu_q(K(d)) = \int u^q K_{(d)}(u) du$, sendo $K_{(d)}$ unha función tipo núcleo de orde $d+1$ se d é impar e de orde $d+2$ se d é par.

- A varianza asintótica condicionada á mostra vén dada por

$$Var(\widehat{m}^{(0)}_{PL,d,n}(x)|X_1, \dots, X_n) = \frac{R(K_{(d)})\sigma^2(x)}{nhf(x)} + o\left(\frac{1}{nh}\right).$$

Á vista destas expresións, podemos ver que o grao do polinomio inflúe na orde de converxencia do sesgo a cero, polo que poderíamos botar man deste feito para a escolla do grao d.

No caso de que a función de regresión sexa moi variable, tense que $|m''(x)|$ é grande e a orde h^2 para $d = 1$ pode ser relevante na expresión do sesgo. Así, poderíamos pensar en aumentar a orde de converxencia a h^4 tomando $d = 2$ ou 3.

Tamén podemos ver na expresión do sesgo que o axuste local con polinomios de grao impar d lévanos a un sesgo da mesma orde que o axuste local con polinomios de grao par $d-1$. Por exemplo, os axustes locais constante e lineal, onde $d = 0$ e $d = 1$ respectivamente, teñen sesgo asintótico de orde h^2. Isto, xunto co feito de que o sesgo asintótico para d impar é máis sinxelo e non depende da función de densidade de X, indica que o axuste local con polinomios de grao impar é máis recomendable.

Selección do parámetro de suavizado

O parámetro de suavizado, h, controla a complexidade da estimación. Se o dito parámetro é moi pequeno, só as observacións moi próximas ao punto de estimación son parte do cómputo do estimador, polo que describirá ben os comportamentos locais, pero esta estimación será moi variable con pouco sesgo e moita varianza. En cambio, se a xanela ten un valor grande, no cálculo do estimador téñense en conta observacións afastadas do punto de estimación, polo que non describirá ben os comportamentos locais e leva a moito sesgo e pouca varianza. Así, a selección de h debe ser equilibrada en canto a sesgo e varianza.

Existen múltiples procedementos para a selección do parámetro de suavizado, pero só describiremos o método de validación cruzada. Pode consultarse Härdle et al (2004), páxina 51, para a exposición doutros criterios para a estimación tipo núcleo da función de densidade, que terían unha formulación análoga no ámbito da regresión.

O criterio de validación cruzada trataría de escoller aquela xanela que minimiza a función

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{m}_{PL,d,n,-i}(X_i))^2,$$

onde $\widehat{m}_{PL,d,n,-i}$ é a estimación tipo núcleo da función de regresión en X_i sen ter en conta o par (X_i, Y_i). O parámetro de suavizado escollido por validación cruzada tende, en media, con n á xanela que minimiza o erro cadrático medio (ASE),

$$ASE(\widehat{m}_{PL,d,n}) = \frac{1}{n} \sum_{i=1}^{n} (\widehat{m}_{PL,d,n}(X_i) - m(X_i))^2.$$

2.2.2. Modelo de regresión polinómico local robusto

O modelo de regresión polinómico local robusto é unha versión robusta do axuste polinómico local que vimos de describir e, polo tanto, non se ve tan afectado por observacións atípicas. Este modelo xurdiu no caso bivariante, isto é, unha única variable explicativa e coñécese como un método de suavizado dun diagrama de dispersión.

Imos describir o algoritmo de suavización LOWESS (*locally weighter scatterplot smoothing*) que propuxo Cleveland e pode consultarse con maior detalle en Cleveland (1979).

1. Para cada $i = 1, \dots, n$ calcúlanse as estimacións da función de regresión mediante un axuste polinómico local en X_i, $\widehat{m}^{(1)}_{PL,d,n}(X_i)$. O autor propón o axuste lineal local e a función tipo núcleo *triweight*, que vai suavemente a cero,

$$K(x) = (1 - |x|^3)^3 I_{[-1,1]}(x).$$

   Doutra banda, a xanela escóllese localmente mediante o método dos k veciños máis próximos. Este método emprega os valores de Y que pertencen ás k observacións de X máis próximas ao punto x para estimar $m(x)$. Pode verse Härdle et al, páxina 98, para a descrición do citado método. Neste caso, tómase f tal que $0 < f \leq 1$ e r o menor enteiro máis próximo a fn. A veciñanza de X_k será aquela de ancho h_k, que se calcula como a r-ésima distancia máis pequena de $\{|X_i - X_k|,\ i = 1, \dots, n\}$, para $k = 1, \dots, n$.

2. Consideremos K a función tipo núcleo *bisquare*,

$$K(x) = (1 - x^2)^2 I_{[-1,1]}(x).$$

   E sexan $\widehat{\varepsilon}^{(1)}_i = Y_i - \widehat{m}^{(1)}_{PL,d,n}(X_i)$, $i = 1, \dots, n$, os residuos do axuste. Se denotamos por s á mediana do valor absoluto dos residuos, $\{|\widehat{\varepsilon}^{(1)}_i|,\ i = 1, \dots, n\}$, definimos os pesos de robustez como

$$\delta^{(1)}_i = K\left(\frac{\widehat{\varepsilon}^{(1)}_i}{6s}\right),\ i = 1, \dots, n.$$

3. Realízase un novo axuste local polinómico, obtendo $\widehat{m}^{(2)}_{PL,d,n}(X_i)$, $i = 1, \dots, n$. Neste caso, os pesos de cada observación (X_j, Y_j), $j = 1, \dots, n$ son

$$K\left(\frac{X_j - X_i}{h_j}\right)\delta^{(1)}_j.$$

   E obtéñense os residuos deste axuste, $\widehat{\varepsilon}^{(2)}_i = Y_i - \widehat{m}^{(2)}_{PL,d,n}(X_i)$, $i = 1, \dots, n$.

4. Repítese o procedemento N veces, resultando en que os valores axustados na última etapa constitúen a estimación da función de regresión.

Esta metodoloxía está programada na función *lowess* do paquete base de R (R Core Team 2020). Unha modificación deste para o caso multivariante atópase na función *loess*, tamén do paquete base.

En particular, neste traballo fíxose uso da función *loess*, do paquete base *stats* (pode consultarse R Core Team 2020) con axuste robusto, empregando *re-descending* M-estimadores no lugar de mínimos cadrados ponderados e coa función *biweight* de Tukey. Isto especifícase mediante o argumento *family="symmetric"*. E seleccionamos con anterioridade o parámetro de suavizado h mediante validación cruzada con criterio de erro a mediana dos erros en valor absoluto (MAD), pois as observacións presentan dependencia e o criterio de validación cruzada usual tende a escoller xanelas pequenas, levándonos así a unha infrasuavización da curva de regresión.

Como é ben sabido, o método de mínimos cadrados para a estimación dos parámetros de regresión supón certas condicións. Ademais, supón que o modelo subxacente é válido para cada observación. Na práctica pode que estas hipóteses non se cumpran e existen outros problemas como as observacións atípicas, que incluso cando o seu número é moi pequeno, a dita estimación sofre grandes alteracións. En consecuencia e dado que os atípicos son moi comúns en situacións reais, xorden diferentes métodos no ámbito da Estatística Robusta. A idea da regresión robusta é axustar un modelo que represente á maioría das observacións. Un dos métodos máis empregados neste contexto son os M-estimadores. Estes son unha xeneralización do método de mínimos cadrados, considerando unha función simétrica ρ no lugar da función de perda cadrática,

$$\min_{\widehat{\beta}} \sum_{i=1}^{n} \rho(r_i),$$

sendo $\widehat{\beta}$ un estimador do parámetro β e r_i ao residuo i-ésimo da regresión.

Os *redescending* M-estimadores son aqueles estimadores M onde a derivada da función ρ cumpre que $\lim_{r_i \to \pm\infty} \rho'(r_i) = 0$. Unha das funcións máis coñecidas neste marco é a función *biweight* de Tukey, que se define como segue:

$$\rho'(r) = \begin{cases} r\left(1 - (\frac{r}{c})^2\right)^2 & \text{se } |r| \leq c \\ 0 & \text{se } |r| > c, \end{cases}$$

para certa constante c.

2.2.3. Modelos aditivos

O contido desta subsección, centrada na presentación dos modelos aditivos, baseouse principalmente en Härdle et al (2004).

Os métodos non paramétricos de estimación de curvas de regresión baséanse na aproximación das ditas curvas en cada punto mediante os datos mostrais nunha veciñanza local do punto ata acadar unha curva suave, que describe a estrutura subxacente aos datos. Pero a flexibilidade que ofrece pode resultar nunha perda de precisión se o número de variables explicativas é alto (maldición da dimensión). Ademais, neste caso, a interpretación dos resultados do modelo pode ser moi complexa. Para aliviar este problema xorde a regresión semiparamétrica, do que forman parte os modelos aditivos.

Os modelos aditivos xeneralizan os modelos de regresión lineares e permiten interpretar os efectos marxinais de cada unha das variables independentes sobre a curva de regresión m, cando as demais se manteñen constantes.

Sexa Y a variable resposta e X o vector de variables explicativas de dimensión k. A estrutura aditiva da función de regresión m vén dada por:

$$m(X) = \mathbb{E}(Y|X_1, \ldots, X_k) = \alpha + \sum_{i=1}^{k} m_i(X_i),$$

onde m_i é unha función non paramétrica unidimensional correspondente á variable explicativa i-ésima.

Así, o modelo aditivo é unha xeneralización do modelo linear múltiple deixando que o efecto de cada unha das variables explicativas poida ser explicado de xeito non paramétrico. Ademais, evítase a dificultade da estimación dun modelo non paramétrico de dimensión k mediante k axustes non paramétricos de dimensión 1.

Os modelos aditivos empréganse cando a variable resposta, Y, é continua e, particularmente, cando se pode asumir que a súa distribución é gaussiana. No caso de non cumprirse esta suposición e de coñecer o modelo que segue a variable resposta, empréganse os modelos aditivos xeneralizados. Este

vén dado por:

$$g(\eta) = \sum_{i=1}^{k} m_i(X_i),$$

onde η representa o parámetro da distribución de Y que reflicte a súa media (por exemplo, a probabilidade de éxito en regresión loxística), g é a función de enlace e m_i é a función suave respectiva ao efecto non paramétrico da variable explicativa i.

Tanto os modelos aditivos como os xeneralizados permiten partes paramétricas e non paramétricas e, amais, interaccións entre variables explicativas en ambos campos. Neste último caso, debemos ter en conta que as interaccións non paramétricas de orde grande, maiores que dous, supón o problema da maldición da dimensión.

Para a estimación destes modelos contamos con múltipes algoritmos: *backfitting*, verosimilitude penalizada (P-IRLS), REML, etc. Describiremos brevemente o primeiro, pero pode consultarse Härdle et al (2004), páxina 212, para unha definición detallada deste algoritmo e das súas variantes.

O algoritmo *backfitting* estima en cada un dos pasos unha función suave sobre os residuos do anterior paso. Así, primeiro inicialízase centrando a variable resposta e, con estes valores centrados, estímase m_1 mediante un método de suavización. Cos residuos destes paso, estímase m_2, de novo, por un método de suavización e así sucesivamente ata aproximar as k funcións non paramétricas. Repítense estes dous últimos pasos considerando os residuos dun paso para o outro, ata converxer.

Para o axuste dos modelos modelos aditivos empregamos a función *gam* do paquete *mgcv*, que se pode consultar en Wood (2011).

Capítulo 3

O número índice

A bibliografía sobre o número índice enmárcase maioritariamente no marco económico, pois é amplamente empregado neste ámbito para describir a evolución de variables, como o crecemento económico dun país ao longo do tempo. Neste capítulo centrarémonos na presentación do número índice neste contexto e para a súa realización empregouse principalmente Fernández e Fernández (2004).

Os números índices son unha ferramenta estatística que ten como obxectivo describir e analizar a evolución de variables que se moven no espazo ou no tempo. Estes comparan os valores dunha variable en dúas situacións distintas, ou ben no espazo ou ben no tempo. A situación de referencia (ou inicial) coñécese como período base, mentres que a outra situación, que se compara co período base, coñécese como período actual. A elección do período base do índice é arbitraria e ten como fin representar o momento ou a localización de referencia que se considera como orixe das comparacións.

Os índices poden ser simples ou complexos en función do tipo de magnitude ao que se refiren. Tamén poderían clasificarse segundo a índole da magnitude, por exemplo, os índices de prezo, que se describen en Peña e Romo (1997), páxina 178.

3.1. Número índice simple

O número índice simple considérase cando a magnitude de interese non pode descompoñerse noutras máis simples. Por exemplo, o prezo dun produto. Partimos, pois, dunha variable simple, X, que toma os valores x_0 e x_t no período base e período actual, respectivamente. O número índice simple no período t referente ao período 0 defínese como

$$I_0^t = \frac{x_t}{x_0}$$

e mide a variación en tanto por un que sufriu a variable considerada entre o período actual e base.

No caso de que o índice, I_0^t, se multiplique por cen mídese a variación porcentual que sufriu a variable entre os dous períodos e coñécese como índice en base cen. Así, $100 \cdot I_0^t$ é o valor do índice no período actual cando no período base ten un valor de 100.

Tamén podemos estar interesados nos crecementos relativos respecto ao período base en tanto por cen, que se calculan como

$$100 \cdot \frac{x_t - x_0}{x_0}.$$

Visualicemos este concepto mediante un exemplo. Supoñamos que queremos saber o cambio porcentual do prezo do quilo de patacas en España entre 2018 e 2021. Para isto precisamos coñecer o dito prezo nestes anos e especificar o período base, por exemplo, o ano 2018. Na Táboa 3.1 móstrase o prezo e o índice simple en base 100 do cambio do prezo respecto de 2018. Así, podemos ver que en 2019 houbo un incremento porcentual de -13.04, no 2020 de 11.59 e no 2021 de 17.39. Notemos que no período base o incremento porcentual é 0.

Ano	Prezo (€/Kg)	Índice simple (en base 100)
2018	0.69	100
2019	0.6	86.96
2020	0.77	111.59
2021	0.81	117.39

Táboa 3.1: Prezo do quilogramo de patacas e índice en base 100 do cambio do mesmo respecto do 2018.

3.2. Número índice complexo

No caso de que a variable de interese poida descompoñerse noutras máis sinxelas, emprégase o número índice complexo. Por exemplo, usaríase no caso de que busquemos estudar o prezo dun conxunto de bens ou servizos. Sexa X unha variable descomposta en n variables, X_1, X_2, ..., X_n, que toman os valores $x_{1,0}$, ..., $x_{n,0}$ no período base e $x_{1,t}$, ..., $x_{n,t}$ no período actual. Facendo uso do número índice simple que vimos de definir, podemos saber cal é a variación de cada unha das variables simples X_i:

$$I_{i,0}^t = \frac{x_{i,t}}{x_{i,0}},\ i = 1,\ \ldots,\ n.$$

Coñécese como índice complexo de media aritmética non ponderada (ou índice de Sauerbeck) á media dos anteriores índices simples,

$$\bar{I}_0^t = \frac{1}{n}\sum_{i=1}^{n}\frac{x_{i,t}}{x_{i,0}}.$$

É dicir, por medio deste índice obtemos unha medida da variación do conxunto das n variables X_i.

Para resumir a información dos índices simples das n variables tamén poderían empregarse as medias harmónica e xeométrica, a mediana e incluso a moda. Pode consultarse Dorin et al (2020), páxina 15, para o desenrolo destas técnicas, incluídas as medias ponderadas, e unha discusión sobre que medida é mellor considerar.

Seguindo co índice complexo de media aritmética non ponderada, consideraremos o exemplo que segue. Supoñamos que dispoñemos dos prezos de certos cultivos agrícolas en España nos anos 2020 e 2021. Estes poden consultarse na Táboa 3.2.

Produto / Ano	2020	2021
(1) Pataca	0.77	0.81
(2) Cabaciña	1.12	1.3
(3) Cebola	0.66	0.78

Táboa 3.2: Prezo en euros por quilogramo dos produtos mencionados nos anos 2020 e 2021.

Procedemos agora a calcular a variación media destes produtos no período actual, 2021, respecto do período base, 2020. Para isto, primeiro calculamos os índices simples dos produtos e logo a súa media.

Índices simples: $I^t_{1,0} = \dfrac{0{,}81}{0{,}77} = 1{,}05$, $I^t_{2,0} = \dfrac{1{,}3}{1{,}12} = 1{,}16$, $I^t_{3,0} = \dfrac{0{,}78}{0{,}66} = 1{,}18$.

Índice complexo de media aritmética non ponderada: $\bar{I}^t_0 = \dfrac{1{,}05 + 1{,}16 + 1{,}18}{3} = 1{,}13$.

Polo tanto, a variación media destes produtos no período actual respecto de 2020 é 1.13.

Debemos notar que estamos outorgándolle o mesmo peso ás variables, X_i, $i = 1, \dots, n$, nas que se descompón a variable de interese, X. É dicir, o índice complexo de media aritmética non ponderada dálle a mesma importancia ás variables X_i no seu conxunto. Por exemplo, no anterior exemplo os tres produtos teñen o mesmo peso. No caso de que as variables non teñan o mesmo peso, empregaríase un índice ponderado.

Sexan $\omega_1, \dots, \omega_n$ números reais denominados pesos, que miden a importancia de cada variable X_i no seu conxunto $\{X_i : i = 1, \dots, n\}$. Definimos o índice complexo de media aritmética ponderada como a media ponderada dos n índices simples das variables X_i, $i = 1, \dots, n$,

$$\bar{I}^{*,t}_0 = \frac{\sum_{i=1}^n \omega_i I^t_{i,0}}{\sum_{i=1}^n \omega_i}.$$

Continuando co exemplo anterior, supoñamos que tamén dispoñemos do valor monetario (o produto do prezo pola cantidade consumida) de cada produto no período base, 2020. O valor monetario da pataca neste ano é 50, o do cabaciña 20 e o da cebola 35. Podemos empregar estes valores como pesos, dándolle máis importancia aos produtos máis caros ou que máis se consumen fronte aos máis baratos ou de raro consumo. Así, obteríamos o índice de media ponderada como

$$\bar{I}^{*,t}_0 = \frac{50 \cdot 1{,}05 + 20 \cdot 1{,}16 + 35 \cdot 1{,}18}{50 + 20 + 35} = 1{,}11.$$

Segundo as ponderacións que se empreguen xorden diferentes índices, como o de Laspeyres e o de Paasche no marco dos índices de prezos. Para a consulta destes índices pode verse Dorin et al (2020), páxina 25.

Parte II

Práctica

Capítulo 4

Introdución

A pandemia de COVID-19 provocou unha grande emerxencia sanitaria a nivel global e, en particular, en España. Isto causou que o noso sistema de saúde colapsase e, como consecuencia, o goberno español tivo que tomar medidas moi restritivas para reducir o número de contaxios, nas que se atopaba o confinamento, que dou comezo o 15 de marzo. Polo tanto, moitas actividades económicas víronse interrompidas causando un significativo descenso no estado económico do país. O confinamento durou aproximadamente tres meses, tendo fin no 21 de xuño, pero o estado de alarma seguía vixente. A medición e análise do impacto da COVID-19, en todos os seus ámbitos e extensións, segue sendo un tema de actualidade e entre os numerosos traballos que se publicaron nas últimas datas, podemos destacar os estudos de Chislett (2021) e de la Fuente (2020), este último centrado unicamente no ámbito económico.

Para elaborar unha resposta política axeitada enfocada ao ámbito económico é necesario dispoñer de resumos estatísticos nesta área que conclúan a súa situación. Os principais agregados macroeconómicos[1], como o Produto Interior Bruto, adoitan ter unha frecuencia trimestral e, ademais, contan con atrasos na súa publicación. Isto significa que non se poden tomar medidas inmediatas consecuentes coa dimensión do problema. Polo tanto, este cambio tan brusco pon de manifesto a necesidade de indicadores de alta frecuencia, no senso de métricas onde o intervalo temporal entre a toma de observacións é moi curto como, por exemplo, días. Ao longo do 2020 xurdiron múltiples iniciativas de distintas entidades para o rastrexo da economía en tempo real, por exemplo, o Banco de Portugal construíu un indicador económico diario para o seu país, que pode verse en Lourenço e Rua (2020).

Neste traballo trataremos de desenvolver un indicador de alta frecuencia para o rastrexo da actividade económica do noso país. Para a creación deste indicador poden considerarse múltiples variables que teñan certo impacto sobre a economía, como poden ser a mobilidade, o tráfico de vehículos comerciais, o pago con tarxetas, a demanda de enerxía, etc. A enerxía ten unha grande importancia no crecemento económico, dado que unha boa parte da produción de bens e actividades comerciais (sectores industrial e servizos) precisan dela. Así, nunha primeira instancia, tomaremos a demanda eléctrica total diaria, que inclúe a demanda eléctrica asociado aos fogares, ao sector servizos e a produción industrial, para a construción deste novo índice, que será estudado e avaliado ao longo Capítulo 5. Como resultado, obtemos que a selección desta variable parece ser axeitada para a construción do indicador e pór solución ao problema que se propón neste proxecto.

A construción do indicador de frecuencias diaria, semanal e mensual relátase no Capítulo 7. Resulta de gran interese comparar a dinámica do noso índice coa dos indicadores máis tradicionais, como son o PIB, IPI e IRE, que se definen a continuación.

[1]A macroeconomía é a área de estudo do funcionamento da economía no seu conxunto, en particular céntrase na agregación dos distintos bens e servizos para reducilos a un só ben.

Produto Interior Bruto

O Produto Interior Bruto (PIB) defínese como o valor monetario total de bens e servizos producidos para o mercado durante un ano dentro das fronteiras dun país. Constitúe o agregado macroeconómico máis importante. A serie do PIB ten unha frecuencia trimestral e, xeralmente, para analizar a súa dinámica ao longo do tempo, transfórmanse os seus valores nun índice cun ano base específico.

Índice de Produción Industrial

O Índice de Produción Industrial (IPI) mide a evolución mensual da actividade produtiva das ramas industriais. Este índice preséntase en serie bruta e corrixida de efectos de calendario[2], co obxectivo de eliminar a influencia no número de días laborais e dos festivos e de estacionalidades. Ademais, esta corrección permite facer comparacións homoxéneas entre os meses de diferentes anos, tal e como cita o INE (Instituto Nacional de Estadística), `https://www.ine.es/`.

Índice da Rede Eléctrica

O Índice da Rede Eléctrica (IRE) é un indicador desenvolto pola Rede Eléctrica de España (`https://www.ree.es/es`) co obxectivo de proporcionar información adiantada da evolución da demanda eléctrica do conxunto de empresas que ten unha demanda eléctrica media-alta, tal e como cita a REE. Este indicador publícase tanto en formato bruto como corrixido de efectos de calendario e da evolución das temperaturas, que veremos ao longo deste traballo que ten un efecto sobre a demanda de electricidade. Neste proxecto empregarase o IRE corrixido.

[2]Unha serie desestacionalizada é a serie unha vez que se lle eliminaron a compoñente estacional e os efectos de calendario. A estacionalidade refírese ao patrón periódico e os efectos de calendario abarcan as dinámicas que están relacionados coa distribución do calendario, por exemplo, os anos bisestos, os festivos móbiles e a proporción dos diferentes días da semana ao longo de cada mes.

Capítulo 5

Análise da demanda eléctrica

A base deste traballo fundaméntase na creación dun índice de alta frecuencia baseado na demanda de electricidade total diaria no país. Este capítulo está orientado á análise desta variable. Na Sección 5.1 estudaremos o comportamento da demanda eléctrica en 2017, onde supoñemos que as circunstancias son normais en termos de evolución económica, mentres que na Sección 5.2 facemos unha pequena análise a longo prazo da evolución da variable entre os anos 2015 e 2019, para estudar se existe tendencia ao longo destes anos, deixando a un lado o ano 2020 polo seu comportamento atípico. Por último, na Sección 5.3 estudamos o ano 2020, que presentará unha dinámica distinta á dos anteriores anos como consecuencia da pandemia provocada pola COVID-19. Os datos poden atoparse na páxina web oficial da Rede Eléctrica de España, `https://www.ree.es/es`.

5.1. Análise exploratoria da demanda eléctrica en 2017

Nesta sección presentamos os resultados da análise exploratoria da demanda total de electricidade, que se mide en xiga-watt-hora (XWh), no territorio nacional ao longo do ano 2017. O estudo ten por finalidade coñecer a dinámica da serie durante un ano usual en termos de evolución económica.

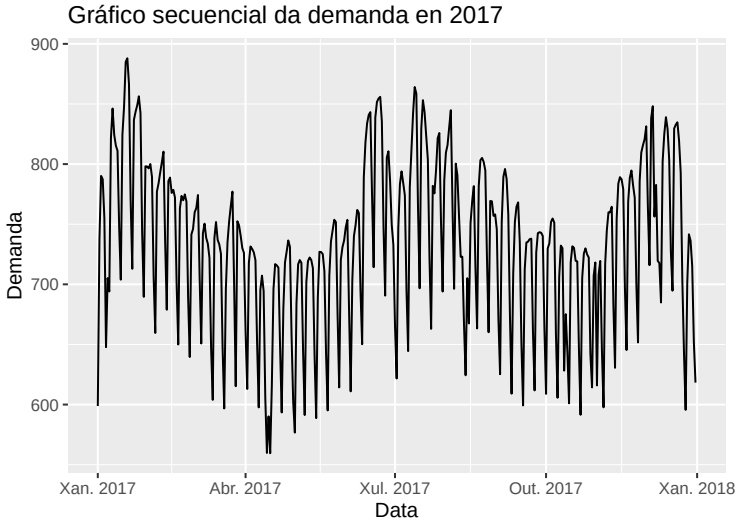


Figura 5.1: Gráfico secuencial da demanda diaria total de electricidade (en XWh) no territorio nacional no ano 2017.

Na Figura 5.1 móstrase o gráfico secuencial da serie de demanda eléctrica en 2017. Nela podemos ver unha posible tendencia e un patrón repetitivo semanal. Esta posible tendencia consiste nunha subida en xaneiro, motivada polo uso de calefacción, para logo ir baixando ata mediados de abril. O nivel comeza a subir levemente ata mediados de xuño, onde hai unha clara subida, debida ao uso do aire acondicionado. Os meses de verán teñen unha demanda maior que os meses veciños, pero tamén se pode ver algunha baixada. A continuación, a demanda comeza a descender ata novembro, onde volve a aumentar, como causa das frías temperaturas. Ao final da serie vese un marcado descenso como resultado dos festivos de Nadal. Respecto do patrón repetitivo semanal, podemos ver unha notable baixada nos domingos conforme aos seguintes días, pois estes non son laborables.

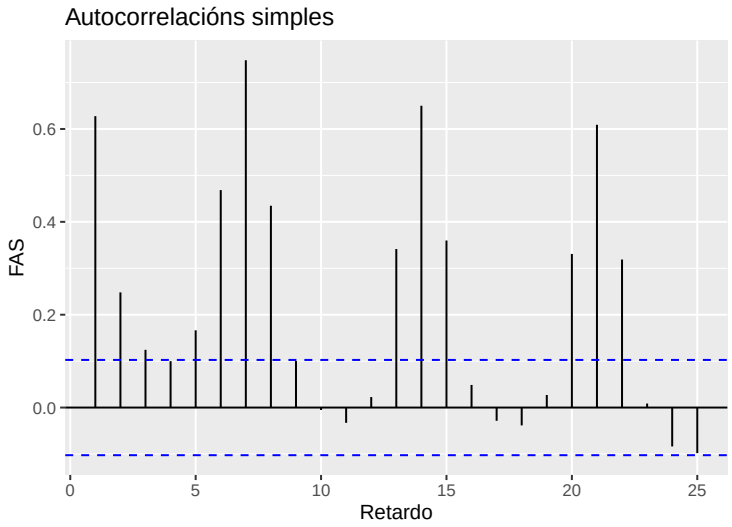


Figura 5.2: Autocorrelacións simples mostrais da serie de demanda eléctrica en 2017.

O gráfico das autocorrelacións simples mostrais da serie, que se mostra na Figura 5.2, indica a presenza de compoñente estacional, dado que no gráfico secuencial se ve un patrón repetitivo e as autocorrelacións simples mostrais mostran repuntes positivos fortes nos múltiplos de 7 e, ademais, estes tardan en diminuír a súa magnitude. Así, o período estacional é $s = 7$. En cambio, as autocorrelacións simples non sinalan a posible tendencia que víamos na Figura 5.1, pois, xeralmente, a tendencia móstrase con autocorrelacións simples mostrais positivas fortes que tardan en baixar.

En conclusión, a serie non é estacionaria dado que presenta estacionalidade. Doutra banda, a súa variabilidade semella ser constante.

Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
559.5	696.2	736.4	733.9	783.2	888

Táboa 5.1: Medidas de posición da demanda eléctrica en 2017 (en XWh).

Na Táboa 5.1 móstrase un resumo das medidas de posición dos datos de demanda eléctrica diaria en 2017. O mínimo presenta un valor aproximado de 559.5 XWh e o máximo de 888 XWh. O primeiro cuartil coincide cun valor de demanda de 696.2 XWh e o terceiro cuartil con 783.2 XWh. Doutra

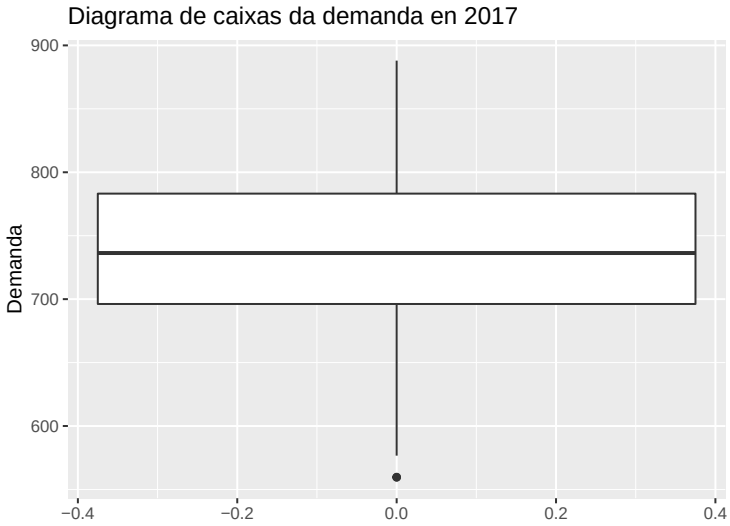


Figura 5.3: Diagrama de caixas da demanda total de electricidade en 2017.

banda, a mediana e a media difiren levemente en 2.5 XWh. Estas medidas tamén se ven reflectidas no diagrama de caixas, que se mostra na Figura 5.3. Este presenta unha lixeira asimetría positiva e un valor atípico.

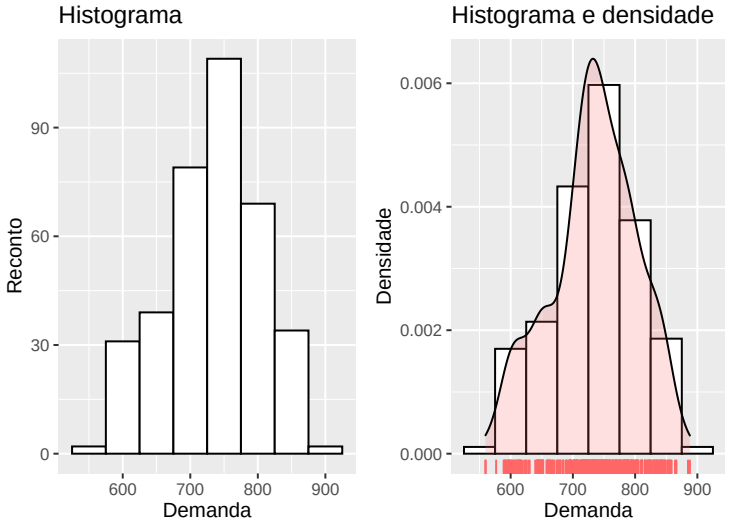


Figura 5.4: Histograma de frecuencias absolutas (esquerda) e histograma e función tipo núcleo da densidade (dereita) da demanda total de electricidade en 2017.

Na Figura 5.4 móstranse, á esquerda, o histograma de frecuencias absolutas e, á dereita, o histograma xunto coa función tipo núcleo da densidade en vermello. No histograma podemos ver que a maior proporción de datos se atopa entre 725 XWh e 775 XWh e tamén a asimetría da que vimos de falar, pois hai un maior volume de datos á dereita. Respecto da densidade, podemos ver que non é simétrica

Mes	Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
Xaneiro	598.8	740.5	798.1	785.1	843	888
Febreiro	639.6	728.6	772.9	751.8	786.6	810.4
Marzo	596.8	704.3	735.2	718.2	751.2	777.1
Abril	559.5	613.8	695	670.5	718.1	736.5
Maio	576.6	660.8	716.7	691.9	726	753.7
Xuño	611	734.7	760.6	767.5	829.8	855.9
Xullo	621.7	734.5	783.2	770.9	822.6	864.1
Agosto	624.4	717.6	766.7	752.3	797.5	844.9
Setembro	599.3	671.5	736.5	715.9	750.2	795.9
Outubro	591.5	645.1	718.4	692.2	730	754.7
Novembro	597.7	698.5	759.9	735.8	785.3	820.6
Decembro	595.6	709	756.4	754.7	827.2	848.2

Táboa 5.2: Medidas de posición da demanda eléctrica por meses en 2017 (en XWh).

(tal é como se intuía) e consta dunha moda.

A continuación presentamos un breve estudo sobre o comportamento mensual da variable. Na Figura 5.5 móstranse os diagramas de caixas por meses e as súas respectivas medidas de posición na Táboa 5.2. Na gráfica podemos ver que hai unha maior demanda en inverno e verán, como consecuencia das temperaturas máis extremas que levan ao uso de calefacción e aire acondicionado, respectivamente. Ademais, nestes meses podemos ver unha maior dispersión dos datos. Isto tamén pode verse en abril, posiblemente motivado polos festivos correspondentes á Semana Santa.

Se nos fixamos no valor da mediana segundo os meses (Táboa 5.2), podemos ver que en xaneiro hai unha maior demanda que nos meses consecutivos. A demanda vai diminuíndo ata maio. De aí tende a subir ata xullo, baixa en agosto e así continúa ata o mes de novembro, onde volve subir. En decembro volve producirse unha leve baixada. Esta dinámica é a forma que víamos no gráfico secuencial da serie (Figura 5.1). Polo tanto, podemos pensar na existencia dun efecto mensual, posiblemente ocasionado pola variación de temperaturas ao longo do ano.

Tamén fixemos unha análise da evolución da demanda eléctrica por días. Na Táboa 5.3 móstrase o resumo das medidas de posición da demanda eléctrica por días e na Figura 5.6 os correspondentes diagramas de caixas. Se nos fixamos na gráfica, podemos ver que, en todos os casos (ao mellor a excepción do sábado), as variables presentan asimetría positiva. Os luns, mércores, xoves e venres presentan algúns valores atípicos, onde aqueles que teñen unha demanda menor da esperada poden

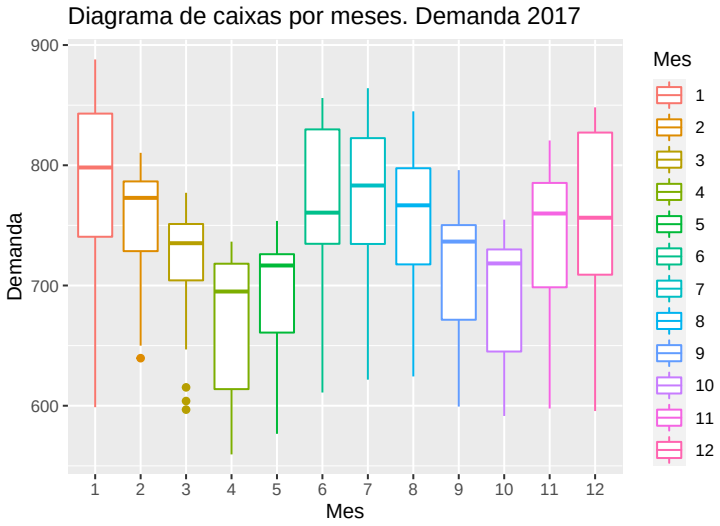


Figura 5.5: Diagrama de caixas por meses da demanda eléctrica en 2017.

Día	Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
Luns	576.6	707.2	744.6	746.1	786	838.6
Martes	667.5	732.1	760.5	768	804.8	853.2
Mércores	615.9	736.9	758.4	769.9	796.1	885
Xoves	605.9	733.6	762.7	767.7	801.6	888
Venres	559.8	722.3	755.8	758.9	793.4	866.3
Sábado	590.2	652.2	686.1	690.1	718.8	775.1
Domingo	559.5	603.9	625.3	638.3	663.3	715.9

Táboa 5.3: Medidas de posición da demanda eléctrica por días en 2017 (en XWh).

ser consecuencia de días festivos. Ademais, hai unha maior demanda nos días laborais que nos fins de semana. Isto tamén se pode ver reflectido na mediana. Así, a demanda é maior, entre os días laborais, os martes e xoves, seguidos moi de preto polos mércores. A demanda decrece considerablemente no sábado e ségueo facendo no domingo. Polo tanto, poderíamos pensar que existe un efecto de fin de semana.

Para contrastar este posible efecto semanal, podemos aplicar un contraste de hipóteses. Pero antes precisamos saber se as mostras por días son ou non independentes. Dado que a variable de demanda eléctrica é continua, podemos empregar o test de independencia da τ de Kendall (Hollander et al 2015, páxina 393) de dous en dous días. En todos os casos, a un nivel $\alpha = 0.05$, rexeitase a hipótese nula de independencia de poboacións. Polo tanto, as mostras son dependentes. Unha vez visto isto, podemos

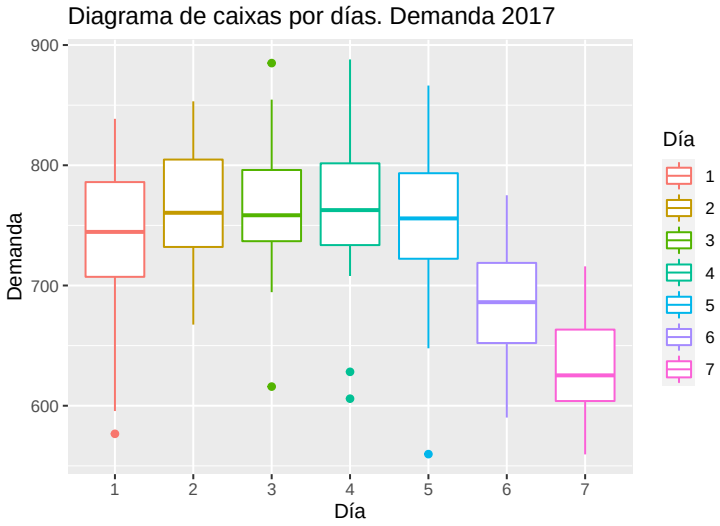


Figura 5.6: Diagrama de caixas por días da demanda eléctrica en 2017.

Día	Luns	Martes	Mércores	Xoves	Venres	Sábado
Martes	0.44					
Mércores	0.27	1				
Xoves	0.42	1	1			
Venres	1	1	1	1		
Sábado	1.8×10^{-6}	2.7×10^{-10}	1.4×10^{-10}	5.9×10^{-10}	6.8×10^{-9}	
Domingo	5.4×10^{-13}	$<2\times10^{-16}$	$<2\times10^{-16}$	9×10^{-16}	3.5×10^{-15}	1.2×10^{-6}

Táboa 5.4: P-valores asociados ao estatístico do test de rangos con signo de Wilcoxon.

aplicar o test de rangos con signo de Wilcoxon (Hollander et al 2015, páxina 115) entre cada par de días para contrastar a igualdade de medianas por días. Na Táboa 5.4 móstranse os p-valores asociados ao estatístico de contraste. Así, a un nivel de significación $\alpha = 0.05$, rexeitamos a hipótese nula en sábados e domingos, polo que os fins de semana son significativamente distintos dos restantes días e concluímos a existencia do efecto semanal.

Tamén fixemos unha análise exploratoria dos datos diarios da demanda de electricidade no territorio nacional nos anos 2018 e 2019, que poden verse no Apéndice A, para estudar se existen diferenzas significativas entre o comportamento da variable ao longo dun ano en condicións normais. Os resultados están na liña dos presentados nesta sección: presenza de compoñente estacional semanal e efecto mensual causado, en principio, pola temperatura (as temperaturas máis frías ou máis quentes provocan unha maior demanda eléctrica).

5.1.1. Suavizado da demanda eléctrica

Vimos de ver que existe un efecto de fin de semana na serie de demanda eléctrica. Isto ocasiona unhas baixadas ao final de cada ciclo semanal, obtendo un aspecto máis rugoso da serie. Para tratar de solucionar este feito e homoxeneizar a demanda eléctrica, suavizamos os seus datos mediante medias móbiles[1] con xanela $h = 7$, onde cada valor da serie temporal se substitúe mediante a media entre as 6 observacións anteriores e ela mesma. Empregando esta xanela obtemos a homoxeneidade que buscabamos, pois cada un dos datos suavizados contén os sete días da semana. Unha vez aplicado o suavizado, eliminamos os seis primeiros datos, pois non están definidos. A partir de agora, denominaremos como demanda eléctrica suavizada á demanda eléctrica suavizada mediante medias móbiles con xanela $h = 7$.

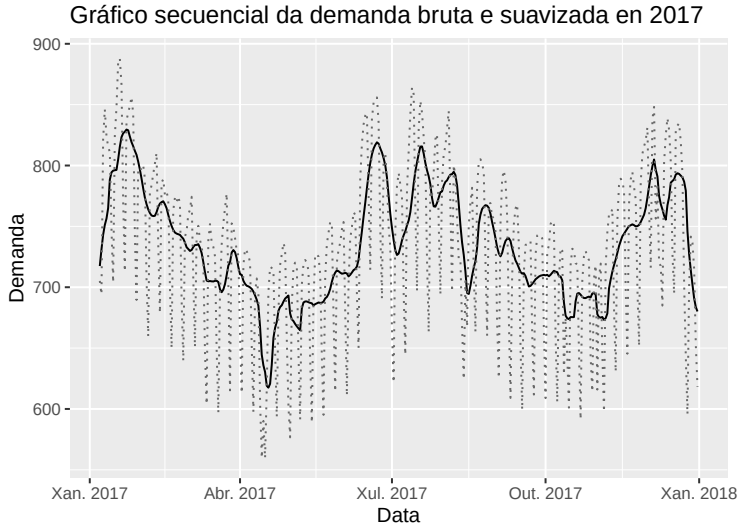


Figura 5.7: Gráficos secuenciais dos datos de demanda eléctrica brutos, en liña punteada, e suavizados mediante medias móbiles con xanela $h = 7$, en negro, no territorio nacional en 2017.

Na Figura 5.7 móstranse os gráficos secuenciais da serie de demanda eléctrica así suavizada en negro e dos datos orixinais en gris e con liña punteada. Nel podemos ver que as observacións suavizadas non teñen este patrón repetitivo semanal, tal e como queríamos, e presentan a mesma tendencia, un maior nivel durante os meses de inverno e verán. Tamén podemos ver unha baixada ao redor de abril, posiblemente motivada pola semana Santa.

Se nos fixamos nas autocorrelacións simples mostrais da serie (Figura 5.8), podemos ver que todas elas son positivas altas e tardan en reducir a súa magnitude a cero. Isto é indicativo da presenza de tendencia.

En resumidas contas, a serie de demanda eléctrica suavizada non é estacionaria, como consecuencia da tendencia. Mentres que a variabilidade da serie é constante.

Para ver se eliminamos o efecto semanal que sofren os datos brutos, fixemos un estudo da demanda eléctrica suavizada por días. O que cabería esperar é un comportamento máis homoxéneo que nos datos

[1] Sexa $\{y_1, \dots, y_T\}$ unha serie de tempo. A aplicación do suavizado de medias móbiles (simples) consiste en calcular, para cada subconxunto $\{y_i, \dots, y_h\}$,

$$\hat{y}_i = \frac{1}{h} \sum_{k=l}^{l+h} y_k,$$

onde $l \leq i$ indica o aliñamento e h é a xanela.

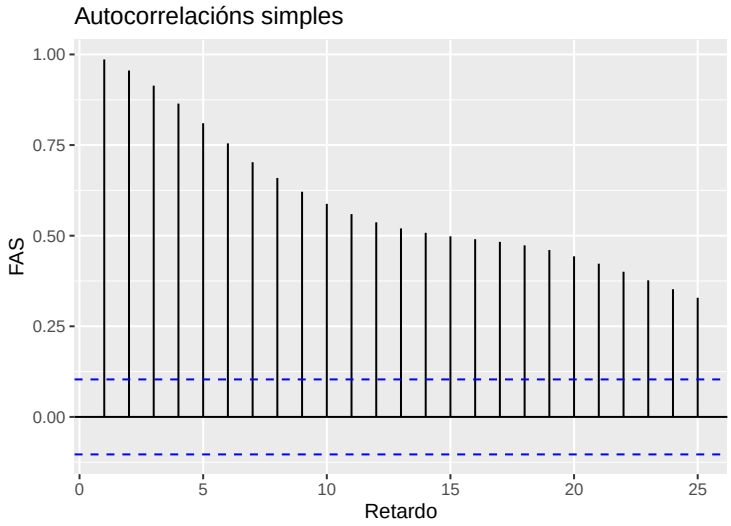


Figura 5.8: Autocorrelacións simples mostrais da serie de demanda eléctrica suavizada en 2017.

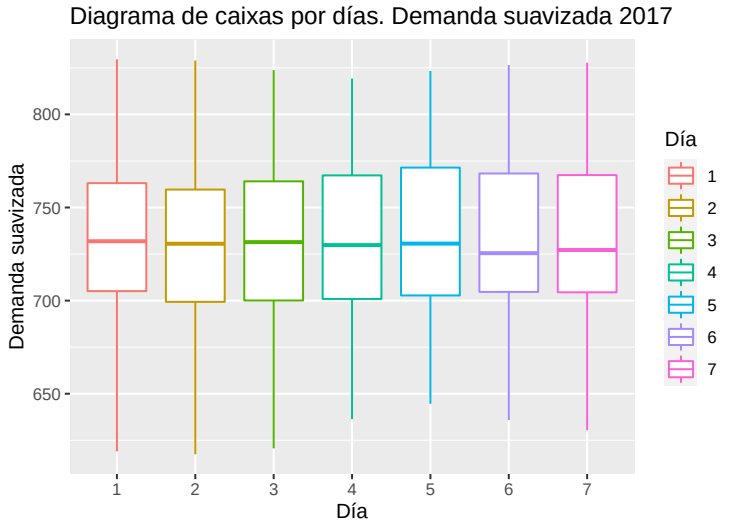


Figura 5.9: Diagrama de caixas por días da demanda eléctrica suavizada en 2017.

orixinais. Na Figura 5.9 móstranse os diagramas de caixas por días. Neles podemos ver asimetría e non hai valores atípicos. Se nos fixamos na mediana diaria, cuxo valor se mostra na Táboa 5.5 (entre outras medidas), podemos ver que son moi semellantes, ao contrario do que pasaba cos datos orixinais. Polo tanto, semella que se eliminou o efecto fin de semana. Para contrastalo, dado que as mostras son dependentes por construción, aplicamos o test de rangos con signo de Wilcoxon. Os p-valores por pares poden verse na Táboa 5.6. Así, a un nivel de significación $\alpha = 0.05$, non rexeitamos a hipótese nula de igualdade de medianas entre días. En consecuencia, mediante a aplicación deste suavizado, eliminamos o efecto semanal da demanda eléctrica.

Día	Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
Luns	619.1	705.1	731.9	734.9	763.1	829.6
Martes	617.5	699.3	730.5	734.7	759.6	828.9
Mércores	620.7	700.1	731.4	734.5	764.1	823.8
Xoves	636.4	700.9	729.8	734.5	767.2	819.3
Venres	644.6	702.8	730.6	734.7	771.4	823.3
Sábado	635.8	704.7	725.5	734.2	768.3	826.5
Domingo	630.4	704.5	727.2	734.3	767.4	827.8

Táboa 5.5: Medidas de posición da demanda eléctrica suavizada por días en 2017 (en XWh).

Día	Luns	Martes	Mércores	Xoves	Venres	Sábado
Martes	1					
Mércores	1	1				
Xoves	1	1	1			
Venres	1	1	1	1		
Sábado	1	1	1	1	1	
Domingo	1	1	1	1	1	1

Táboa 5.6: P-valores asociados ao estatístico do contraste de rangos con signo de Wilcoxon.

Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
617.5	701.6	730.5	734.5	767.4	829.6

Táboa 5.7: Medidas de posición da demanda eléctrica suavizada en 2017 (en XWh).

A continuación, na Táboa 5.7 preséntase un resumo das medidas de posición dos datos suavizados. Estas observacións presentan unha demanda mínima de 617.5 XWh (superior á dos orixinais) e máxima 829.6 XWh (inferior). A mediana é 730.5 XWh (inferior) e a media 734.5 XWh (lixeiramente superior). O primeiro cuartil é 701.6 XWh (superior) e o terceiro 767.4 XWh (inferior). Neste caso o rango intercuartílico é inferior, obtendo así unha menor dispersión dos datos nesta franxa.

Na Figura 5.10 móstrase o diagrama de caixas da demanda suavizada. Nel podemos ver unha leve asimetría positiva.

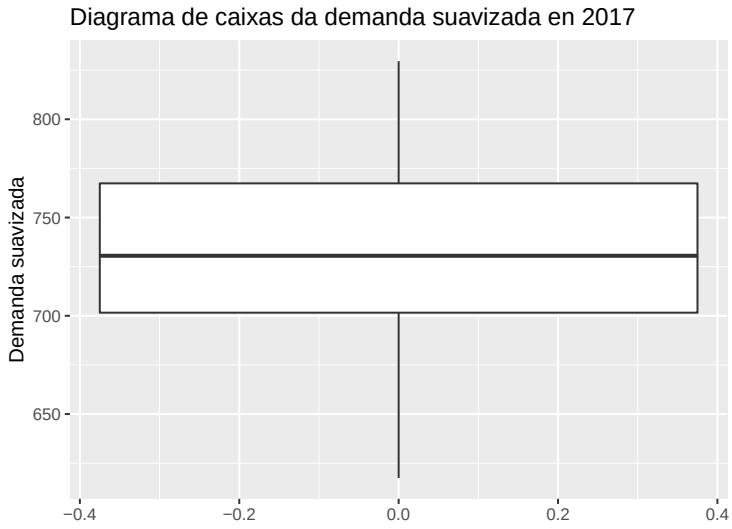


Figura 5.10: Diagrama de caixas da demanda total de electricidade suavizada en 2017.

O histograma de frecuencias absolutas e o histograma xunto coa función tipo núcleo da densidade preséntanse na Figura 5.11. No histograma vemos unha maior proporción de datos entre 675-725 XWh e tamén podemos ver asimetría, pois hai un maior volume de datos á dereita. Respecto da densidade, podemos ver que non é simétrica (tal é como se intuía no estudo anterior) e ten unha moda.

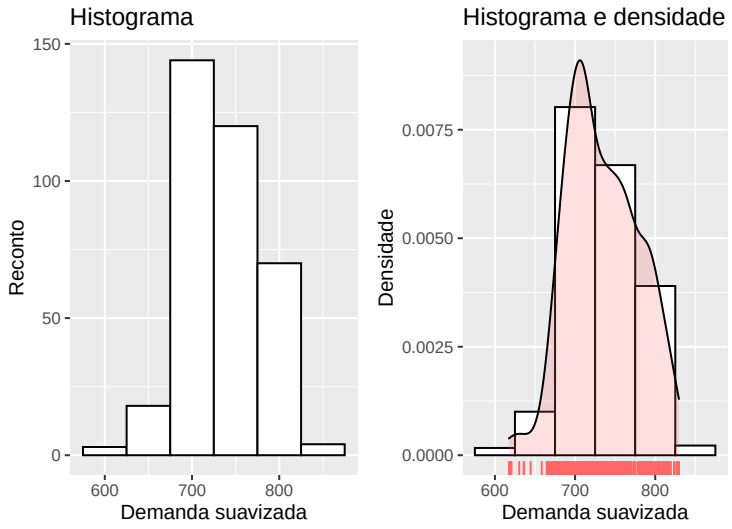


Figura 5.11: Histograma de frecuencias absolutas (esquerda) e histograma e función tipo núcleo da densidade (dereita) da demanda total de electricidade suavizada en 2017.

Mes	Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
Xaneiro	717.4	788.3	803.2	794.8	819.3	829.6
Febreiro	732.8	744.7	759.2	757.9	768.2	789.2
Marzo	695.9	705	717.9	717.2	730.2	735.4
Abril	617.5	660.5	686.7	675.8	698.1	710.6
Maio	664.6	683.8	687.5	688.7	693.3	713.9
Xuño	708.9	714.3	764.9	762.1	805	819.1
Xullo	726.5	747.3	773.1	770.5	793.1	815.7
Agosto	694.5	729.5	760.3	754	783.6	794.9
Setembro	700.6	708.9	714.2	719	731	740.3
Outubro	673.8	689.8	694.1	695.4	709	713.5
Novembro	673.2	700.8	739.9	725.3	750.7	765.1
Decembro	680.1	757.2	779.3	765.7	791.4	805

Táboa 5.8: Medidas de posición da demanda eléctrica suavizada por meses en 2017 (en XWh).

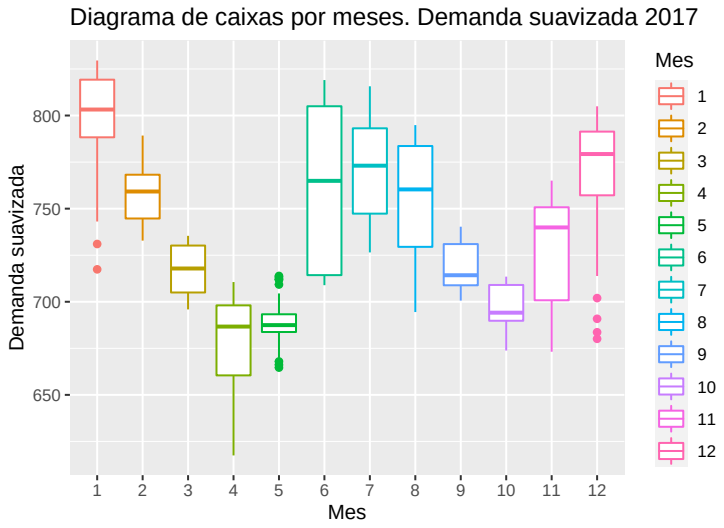


Figura 5.12: Diagrama de caixas por meses da demanda eléctrica suavizada en 2017.

Por último, presentamos unha pequena análise mensual da variable suavizada. O diagrama de caixas por meses móstrase na Figura 5.12 e o seu resumo das medidas de posición na Táboa 5.8. No diagrama de caixas podemos ver que xaneiro, maio e decembro presentan observacións atípicas. Ademais, en todos os meses a demanda é asimétrica. Coa excepción de xuño, a dispersión dos datos é menor en comparación coa demanda orixinal. Se nos fixamos na mediana, podemos observar a tendencia que víamos no gráfico secuencial. A demanda eléctrica tende a subir nos meses de inverno e verán, indicando un posible efecto mensual provocado posiblemente pola temperatura, que tamén presentan os datos orixinais.

En resumo, se suavizamos os datos diarios de demanda eléctrica mediante medias móbiles con xanela $h = 7$, eliminamos o efecto fin de semana que víamos nas observacións brutas, obtendo unha serie máis homoxénea. Os datos suavizados tamén presentan un efecto mensual, que consideramos froito da temperatura, onde as temperaturas máis extremas provocan unha maior demanda.

5.1.2. Estudo do efecto da temperatura sobre a demanda eléctrica

Nas seccións previas xa comentamos o efecto mensual, quizais provocado pola temperatura, que sofren tanto a demanda eléctrica como a demanda eléctrica suavizada mediante medias móbiles con xanela $h = 7$. Este efecto consiste nunha maior demanda de electricidade nos meses de inverno e verán, onde as temperaturas son máis extremas. Nesta subsección centrarémonos no estudo da relación entre a demanda eléctrica e a temperatura media diarias ao longo de 2017.

Dado que o cálculo da temperatura media nacional non ten sentido dende un punto de vista meteorolóxico, tomaremos os datos da estación de Madrid Retiro, que se atopa na zona centro de España e conta cun gran volume de poboación e industria. No caso de que exista algún dato faltante, tomarase a media do anterior e seguinte día. Os datos de temperatura media diaria, medida en graos Celsius (ºC), poden obterse mediante o paquete *climaemet* (Piñeiro 2020) e mediante unha API Key, que pode conseguirse na páxina web oficial da Axencia Estatal de Meteoroloxía (AEMET), `http://www.aemet.es/es`.

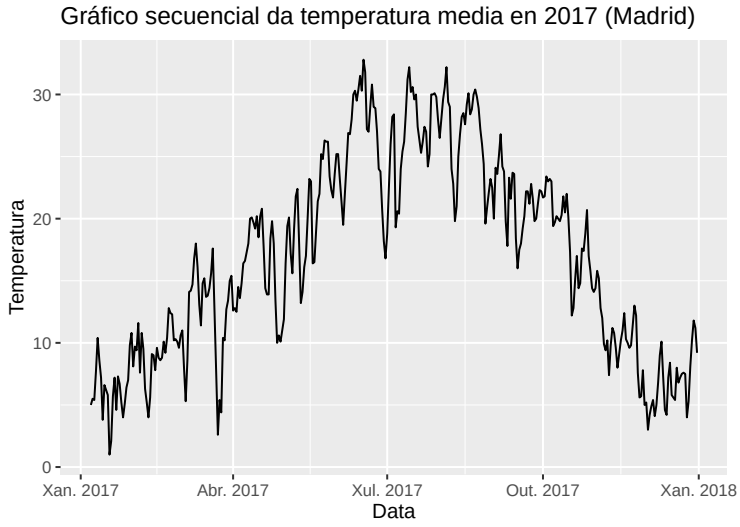


Figura 5.13: Gráfico secuencial da temperatura media diaria na estación Madrid Retiro en 2017.

Na Figura 5.13 móstrase o gráfico secuencial da temperatura media na estación Madrid Retiro no 2017. Podemos ver a posible presenza de tendencia e a variabilidade non semella ser constante, dado

que aumenta co nivel. A dita tendencia consiste nunha subida de temperatura dende principios de ano ata o verán, para logo baixar.

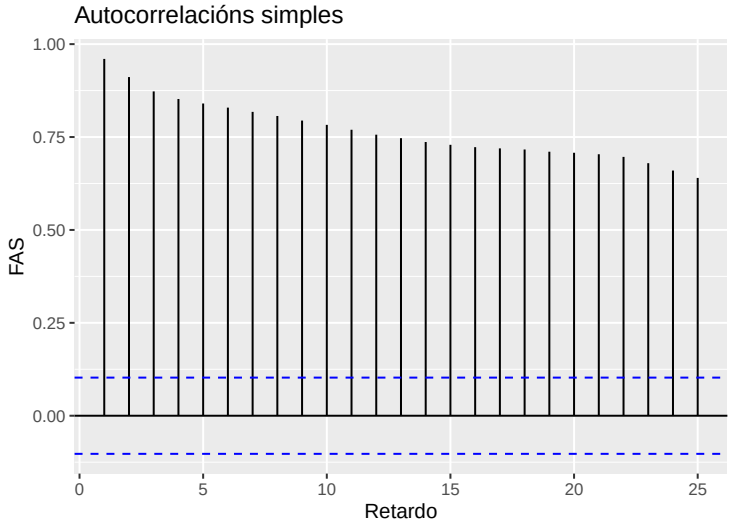


Figura 5.14: Autocorrelacións simples mostrais da serie de temperatura media na estación Madrid Retiro en 2017.

Na Figura 5.14 móstranse as autocorrelacións simples mostrais da serie de temperatura media. Podemos ver que estas correlacións son positivas altas e, ademais, tardan en reducir a súa magnitude, indicando a presenza de tendencia. En consecuencia, a serie non é estacionaria.

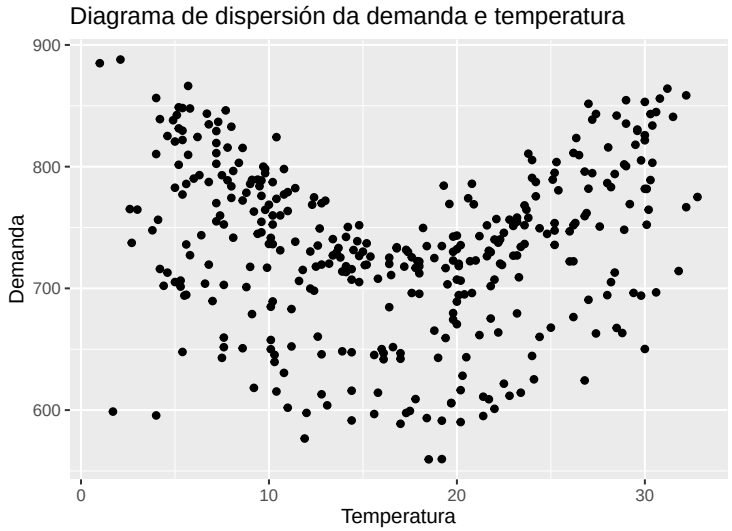


Figura 5.15: Diagrama de dispersión da demanda eléctrica e da temperatura media en 2017.

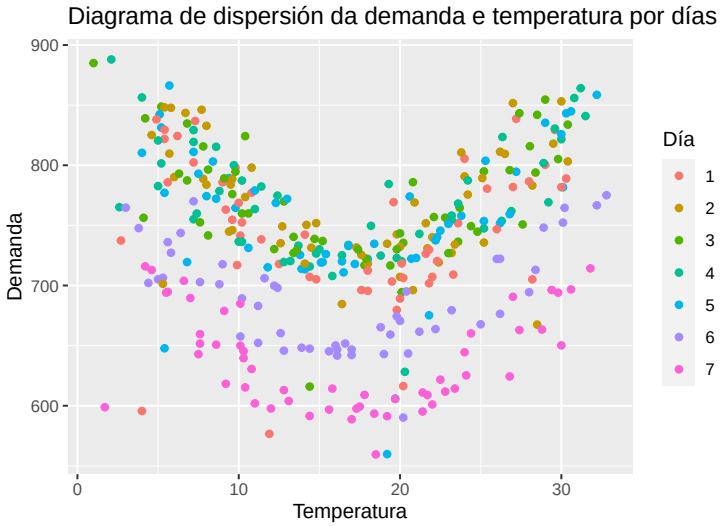


Figura 5.16: Diagrama de dispersión por días da demanda eléctrica e da temperatura media en 2017.

Unha vez visto o comportamento da temperatura, pasamos a estudar a súa relación coa demanda eléctrica. Na Figura 5.15 móstrase o diagrama de dispersión dos datos de demanda eléctrica e temperatura. Podemos ver unha relación en forma de parábola, onde as temperaturas máis extremas provocan unha maior demanda de electricidade. Ademais, vemos certos puntos, seguindo esta tendencia, que presentan unha menor demanda. É posible que estes puntos se dean durante os fins de semana, dado que nestes dous días existe unha demanda eléctrica significativamente inferior, como xa vimos. Para comprobar se este feito é certo, podemos pintar a nube de puntos segundo o día da semana. O resultado pode verse na Figura 5.16. Tal e como esperabamos, os sábados e domingos (especialmente estes últimos) son os que presentan unha menor demanda, provocando unha maior dispersión nos datos.

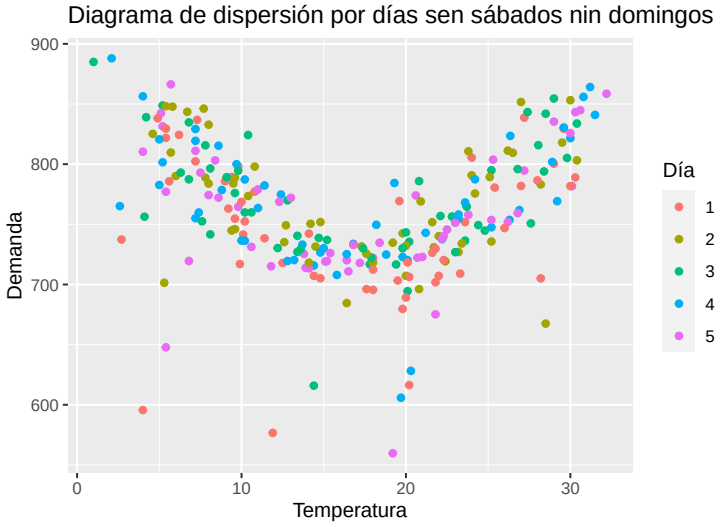


Figura 5.17: Diagrama de dispersión por días da demanda eléctrica e da temperatura media sen sábados nin domingos en 2017.

Se eliminamos as observacións correspondentes aos fins de semana, obtemos unha nube de puntos con menor variabilidade, como podemos ver na Figura 5.17, aínda que existen algunhas observacións atípicas, con menor demanda eléctrica que a esperada. Podemos pensar que estes datos atípicos se dan nos festivos. Para verificalo, sinalamos as festividades no diagrama de dispersión, que se mostra na Figura 5.18. Os festivos que se marcan, que se poden ver na lenda da figura, son nacionais a excepción do 13 e 17 de abril, onde moitas comunidades celebran a dita festividade. O 6 e 8 de decembro semellan presentar un comportamento dentro do esperado, mentres que os restantes festivos presentan unha demanda moito menor do que correspondería ao seu día laboral.

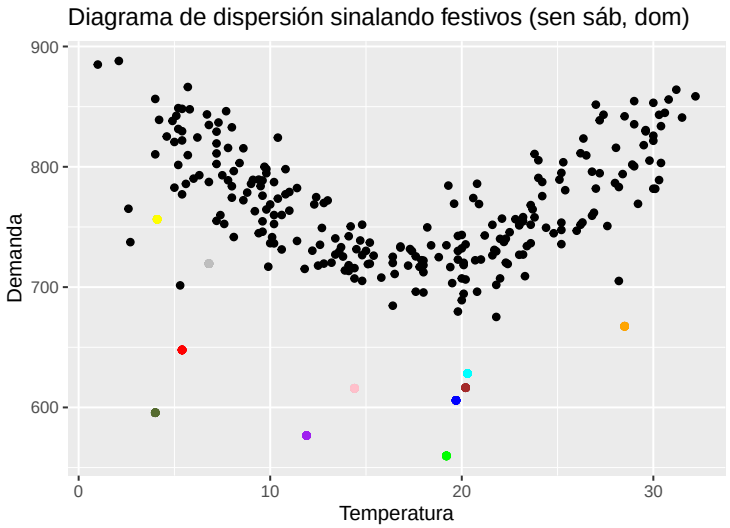


Figura 5.18: Diagrama de dispersión da demanda eléctrica e da temperatura media sen fins de semana en 2017, sinalando os festivos, onde o 6 de xaneiro (venres) se mostra en vermello, 13 de abril (xoves) en azul escuro, 14 de abril (venres) en verde claro, 17 de abril (luns) en marrón, 1 de maio (luns) en morado, 15 de agosto (martes) en laranxa, 12 de outubro (xoves) en ciano, 1 de novembro (mércores) en rosa, 6 de decembro (mércores) en amarelo, 8 de decembro (venres) en gris e 25 de decembro (luns) en verde escuro.

Ademais da análise diaria, tamén é interesante estudar o comportamento por meses. Para isto, pintamos o total das observacións de demanda eléctrica e temperatura por meses. Isto móstrase na Figura 5.19. Como xa vimos, os meses de verán e inverno presentan unha maior demanda de electricidade, posiblemente como consecuencia do uso de aire acondicionado e calefacción, respectivamente.

Logo desta pequena análise exploratoria, aplicamos medias móbiles con xanela $h = 7$ aos datos de temperatura para suavizar o seu comportamento ao longo da semana e eliminamos as seis primeiras observacións por non estar definidas. O seu gráfico secuencial móstrase na Figura 5.20, onde podemos ver que, efectivamente, suavizamos a súa dinámica.

O seguinte paso é analizar a relación entre a temperatura e a demanda eléctrica, ambas así suavizadas. Na Figura 5.21 móstrase o diagrama de dispersión das observacións destas dúas variables por días. Neste caso, ao contrario do que víamos na Figura 5.16, os días presentan unha demanda máis homoxénea (pois xa vimos que ao aplicar medias móbiles con xanela $h = 7$ eliminamos o efecto fin de semana da demanda eléctrica). Tamén podemos notar a presenza de atípicos, que son consecuencia dos

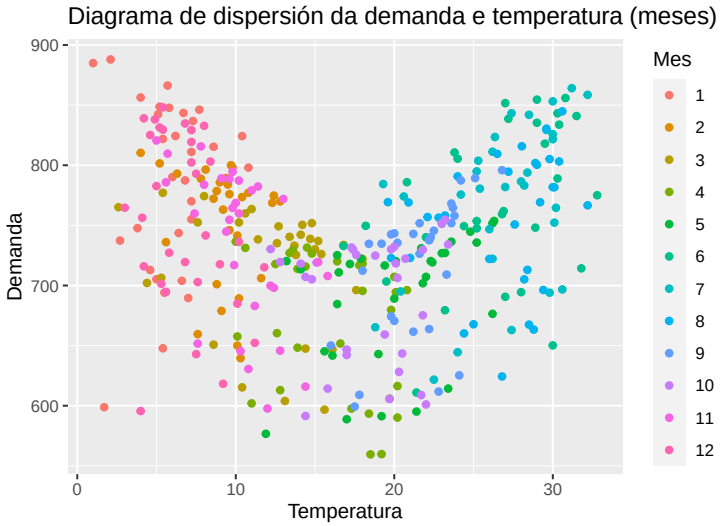


Figura 5.19: Diagrama de dispersión por meses da demanda eléctrica e da temperatura media en 2017.

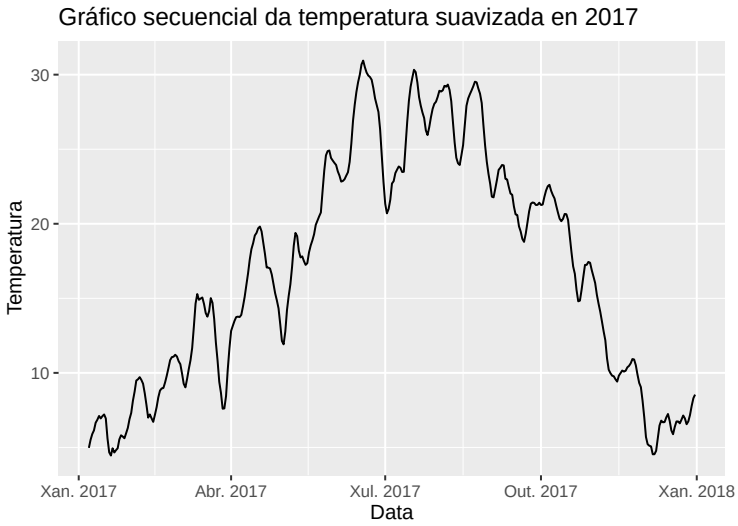


Figura 5.20: Gráfico secuencial dos datos suavizados da temperatura media na estación Madrid Retiro en 2017.

festivos de semana Santa, tal e como se mostra na Figura 5.22. Debemos lembrar que o suavizado de medias móbiles con xanela $h = 7$ consiste na substitución da observación pola media dos seis anteriores días e ela mesma, polo que o efecto da festividade se arrastra en sete días consecutivos.

É interesante estudar o comportamento por meses, dado que xa vimos a existencia dun efecto mensual sobre a demanda eléctrica e consideramos que este viña dado pola temperatura. Para isto, debuxamos o diagrama de dispersión dos datos suavizados por meses. O resultado móstrase na Figura 5.23 e tal e como vimos, todo apunta á existencia dun efecto mensual por mor da temperatura sobre a demanda eléctrica.

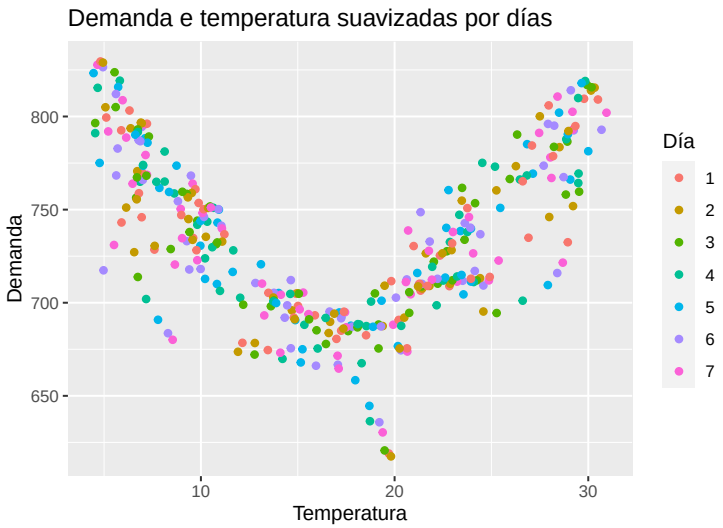


Figura 5.21: Diagrama de dispersión por días da demanda eléctrica e temperatura suavizadas en 2017.

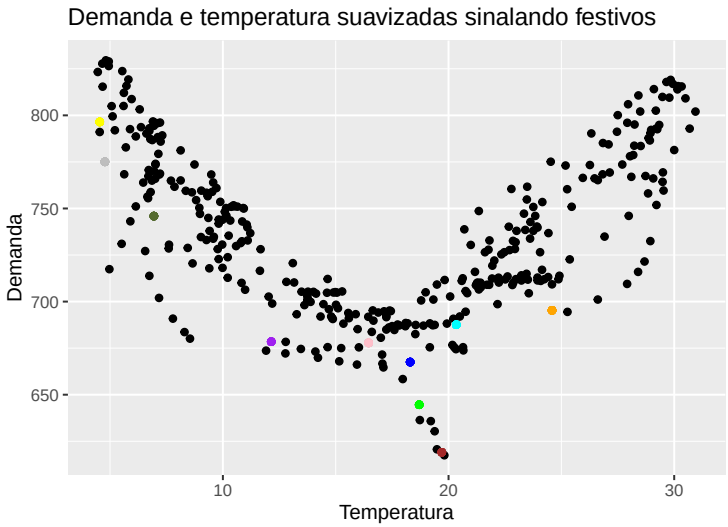


Figura 5.22: Diagrama de dispersión da demanda eléctrica e temperatura suavizadas en 2017 sinalando festivos, onde o 6 de xaneiro (venres) se mostra en vermello, 13 de abril (xoves) en azul escuro, 14 de abril (venres) en verde claro, 17 de abril (luns) en marrón, 1 de maio (luns) en morado, 15 de agosto (martes) en laranxa, 12 de outubro (xoves) en ciano, 1 de novembro (mércores) en rosa, 6 de decembro (mércores) en amarelo, 8 de decembro (venres) en gris e 25 de decembro (luns) en verde escuro.

Unha vez visto que existe un efecto da temperatura sobre a demanda de electricidade, sería importante axustar un modelo de regresión para estudar a relación entre estas variables. Ante estes datos, consideramos que unha boa opción sería o axuste dun modelo de regresión non paramétrico e, dado que conta con algunhas observacións atípicas, sería conveniente que este modelo fose robusto. En consecuencia, axustamos un modelo de regresión polinómico local robusto aos datos suavizados, onde a variable explicativa é a temperatura e a resposta, a demanda eléctrica, mediante a función *loess* do

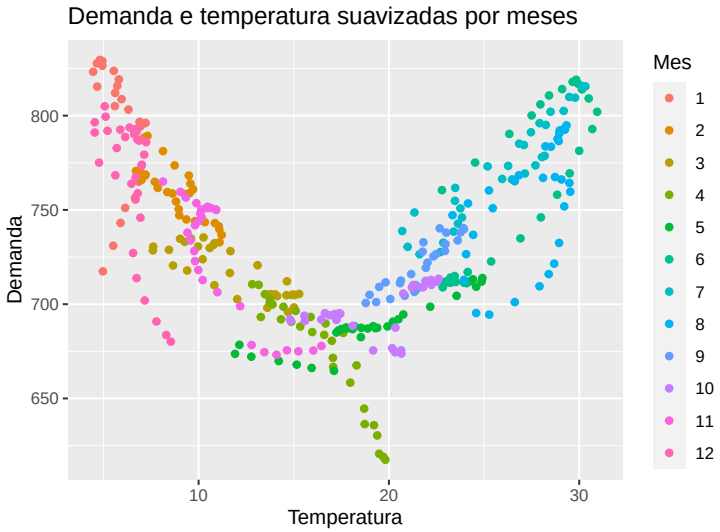


Figura 5.23: Diagrama de dispersión por meses da demanda eléctrica e temperatura suavizadas en 2017.

paquete base (R Core Team 2020). Ademais, seleccionamos o parámetro de suavizado mediante validación cruzada con criterio de erro a mediana dos erros en valor absoluto. A xanela, a proporción de datos no axuste local, é 0.1. Dado que a xanela é pequena, o axuste non será moi suave, tal e como se pode ver na Figura 5.24. Este axuste ten un coeficiente de determinación $R^2 = 0.75$, polo que o modelo explica un 75 % da variabilidade da demanda. Esta proporción parece ser acertada, pois é lóxico que a demanda eléctrica non dependa só da temperatura, senón doutros factores económicos, que son os que interesan medir neste traballo.

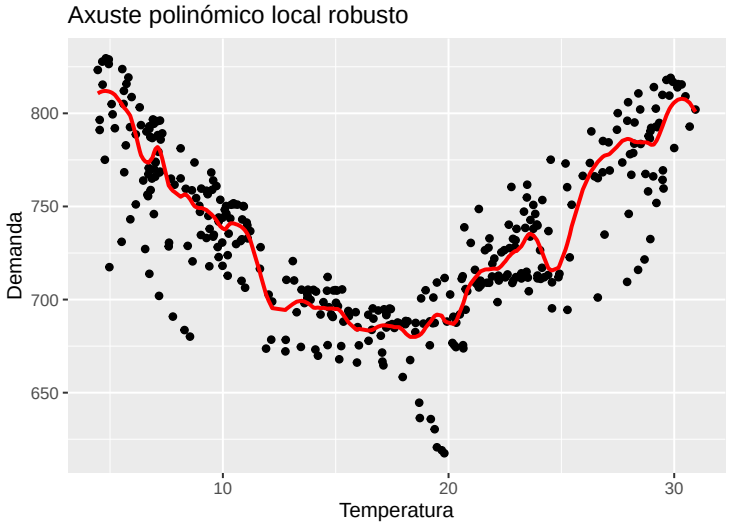


Figura 5.24: Axuste do modelo de regresión polinómico local robusto aos datos de demanda eléctrica e temperatura suavizados.

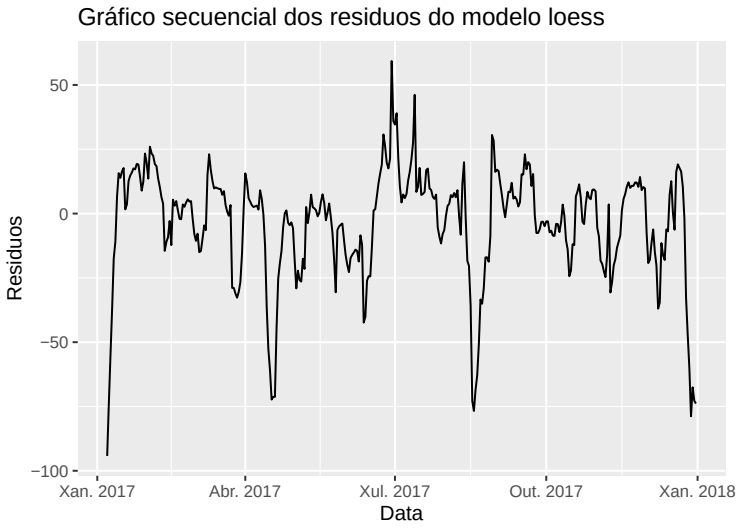


Figura 5.25: Gráfico secuencial dos residuos do modelo *loess*.

Na Figura 5.25 móstrase o gráfico secuencial da serie de residuos do modelo que vimos de axustar. A variabilidade da serie semella constante, non hai patrón repetitivo nin tendencia, polo que é estacionaria. Esta serie correspóndese coa serie de demanda eléctrica limpa do efecto da temperatura.

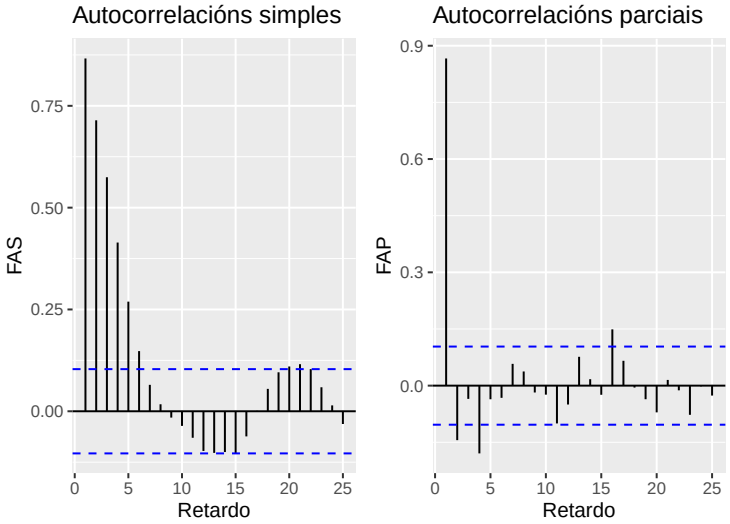


Figura 5.26: Autocorrelacións simples e parciais mostrais dos residuos.

Na Figura 5.26 móstrase a estimación da función autocorrelacións simples e parciais dos residuos do modelo. Se nos fixamos nas autocorrelacións simples mostrais, podemos ver que as primeiras son positivas altas, pero logo van rapidamente a cero e aquelas que saen das bandas teñen moi baixa magnitude. No caso das autocorrelacións parciais mostrais, só tres saen das bandas e teñen baixa magnitude. Así, os residuos presentan estrutura e non son independentes, que era de esperar dado que os datos son medias móbiles, e poden modelarse mediante a metodoloxía ARMA. Polo tanto,

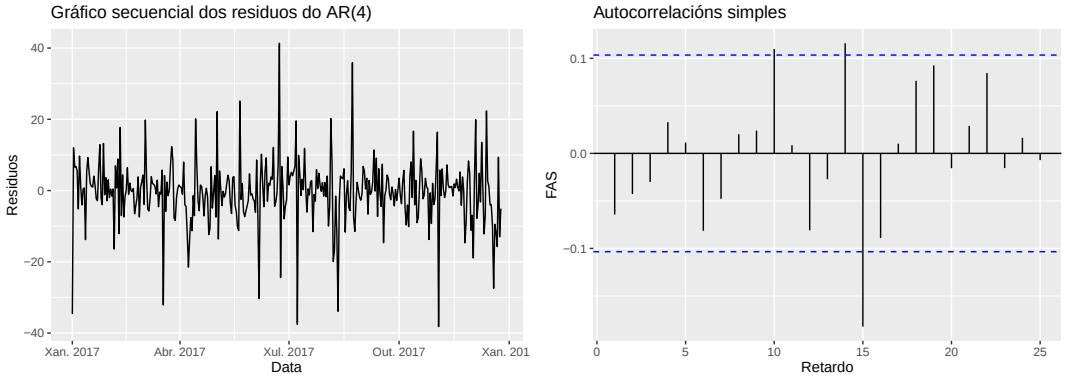
imos tratar de axustar os residuos mediante un modelo ARMA$(p,q)\times(P,Q)$. Á vista dos gráficos de autocorrelacións, podemos pensar nun AR(4), pero é preferible obter un modelo óptimo segundo o criterio AICc. Para isto, aplicamos a función *auto.arima* do paquete *forecast* (Hyndman et al 2020). Esta propón un ARIMA(2,0,3) con media non nula, cuxo AICc é 2590.32. Tamén podemos tratar de axustar un ARIMA(4,0,0) con media nula, dado que o seu AICc é 2591.655 e dista menos de dúas unidades que o AICc do modelo proposto, polo que se pode considerar como un modelo equivalente. Dado que este último é máis sinxelo, pois ten menos parámetros, escollemos este modelo. As estimacións dos coeficientes e as súas desviacións típicas xunto coa desviación típica do ruído branco poden verse na Táboa 5.9. Todos os coeficientes resultaron significativamente distintos de cero. O axuste AR(4) pode escribirse como

$$y_t = 1.14y_{t-1} - 0.25y_{t-2} + 0.23y_{t-3} - 0.24y_{t-4} + a_t.$$

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	a_t
Estimación	1.14	−0.25	0.23	−0.24	
Desviación típica	0.05	0.08	0.08	0.05	8.84

Táboa 5.9: Estimacións dos coeficientes do modelo AR(4) xunto coas súas desviacións típicas e a desviación típica do ruído branco.

Unha vez axustado o modelo AR(4), pasamos a análise dos seus residuos. Recordemos que estes deben ser independentes, ter variabilidade constante e media cero. Tamén sería desexable a normalidade dos mesmos, aínda que non necesario.



(a) Gráfico secuencial. (b) Autocorrelacións simples mostrais.

Figura 5.27: Residuos do modelo *loess*.

Na Figura 5.27 móstranse o gráfico secuencial da serie de residuos (esquerda) e as súas autocorrelacións simples mostrais (dereita). Á vista do gráfico secuencial, non se detecta tendencia nin patrón repetitivo e semella que a súa variabilidade é constante. Se nos fixamos nas autocorrelacións simples, podemos ver que os retardos 10, 14 e 15 saen das bandas, esta última cunha magnitude 0.2. Para

contrastar a independencia dos residuos, aplicamos o test de Ljung-Box, resultando nun p-valor aproximado de 0.08. Se consideramos un nivel de significación α=0.05, non rexeitamos a hipótese nula de independencia.

Unha vez vista a independencia dos residuos, contrastamos se estes teñen media nula mediante o t test. O p-valor asociado ao estatístico t é 0.28, polo que a un nivel $\alpha = 0.05$ non rexeitamos a hipótese nula de media cero. Por último, podemos ver se os residuos son normais a través do test de Shapiro-Wilk. O p-valor resultou ser 8.17×10^{-14}, polo que a un nivel de significación $\alpha = 0.05$, rexeitamos a hipótese nula de normalidade. Dado que os residuos son independentes, teñen varianza constante e media nula, o modelo AR(4) é axeitado.

En definitiva, vimos de axustar o efecto que a temperatura exerce sobre a demanda eléctrica mediante un modelo de regresión polinómico local robusto, cuxos residuos proveñen dun AR(4). Este modelo permítenos obter a demanda de electricidade limpa do efecto da temperatura e, en consecuencia, da compoñente mensual presente nos datos orixinais.

5.2. Análise da demanda eléctrica entre 2015 e 2019

Na Sección 5.1 analizamos a dinámica da demanda de electricidade ao longo dun ano en circunstancias normais. Neste punto, é interesante realizar un estudo da evolución da variable a longo prazo, para tratar de ver se a demanda eléctrica aumentou ou diminuíu durante os anos precedentes ao 2020, no que se produciu un cambio de nivel por mor da pandemia de COVID-19. Para isto, tomamos os datos de demanda eléctrica diaria no territorio nacional dende o ano 2015 ata o 2019. Na Figura 5.28 móstrase o gráfico secuencial desta serie. Nel podemos observar que o nivel da demanda eléctrica durante estes anos non semella aumentar nin diminuír notablemente, senón que oscila en torno a unha horizontal. En xeral, parece que a demanda se mantivo máis ou menos constante.

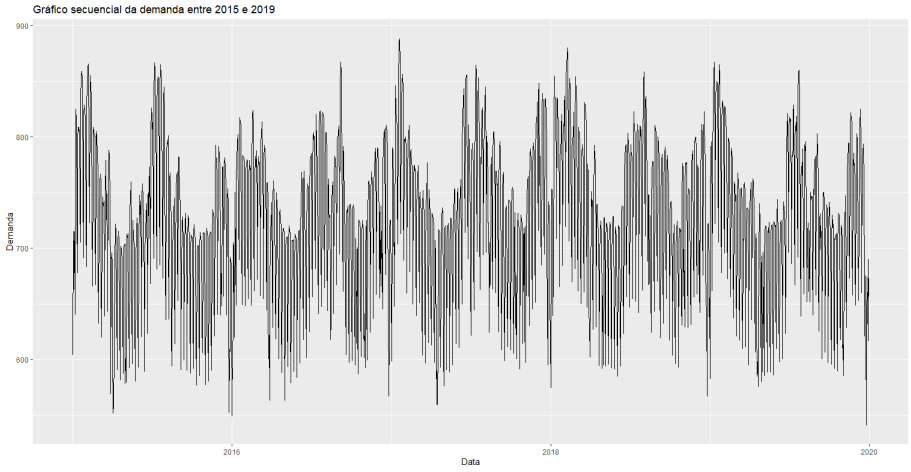
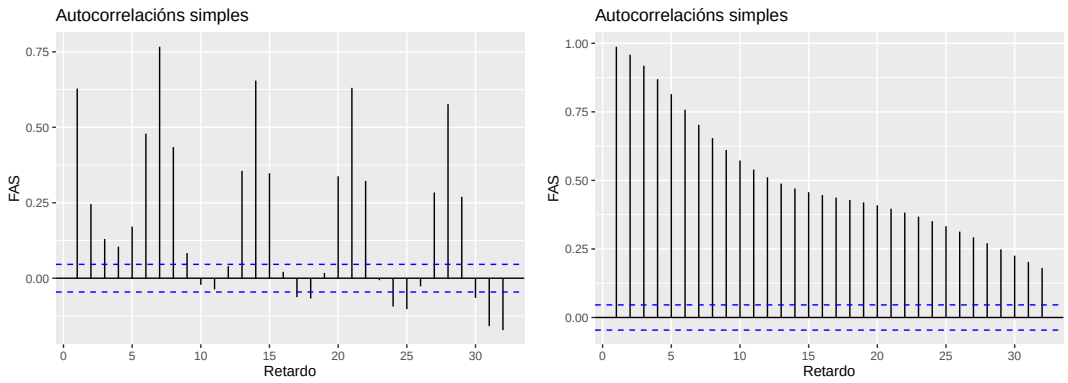


Figura 5.28: Gráfico secuencial da demanda diaria total de electricidade no territorio nacional entre 2015 e 2019.

Doutra banda, podemos notar a presenza de patrón repetitivo semanal (tal e como estudamos na demanda de electricidade ao longo dun ano) e, moi posiblemente, anual e dun efecto mensual. Na Figura 5.29a móstranse as autocorrelacións simples mostrais da serie. Estas presentan correlacións positivas altas nos retardos múltiplos de 7 e, dado que no gráfico secuencial víamos patrón repetitivo semanal, deducimos que a serie presenta compoñente estacional de período $s = 7$. Así, estas capturan



(a) Demanda eléctrica. (b) Demanda eléctrica suavizada.

Figura 5.29: Autocorrelacións simples mostrais.

a compoñente estacional semanal, pero non a de magnitudes superiores.

En conclusión, a serie non é estacionaria, dado que presenta compoñente estacional de período $s = 7$ (patrón repetitivo semanal) e posiblemente anual, que consistiría na repetición do patrón mensual causado polas temperaturas dentro de cada ano. Mentres que a variabilidade da serie semella ser constante e parece que a dita serie non presenta tendencia (o seu nivel non varía ao longo do tempo).

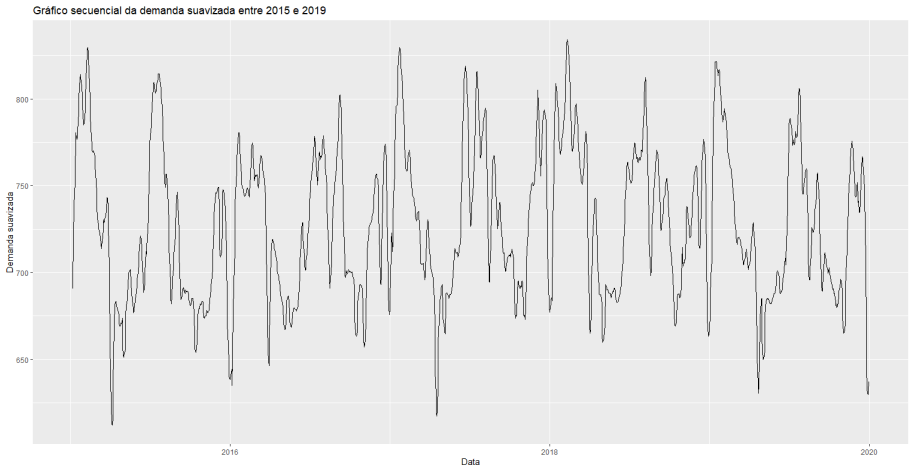


Figura 5.30: Gráfico secuencial da demanda diaria total de electricidade no territorio nacional entre 2015 e 2019 suavizada mediante medias móbiles.

Como xa vimos, a estacionalidade semanal pode aliviarse mediante medias móbiles con xanela $h = 7$. Na Figura 5.30 preséntase o gráfico secuencial da serie de demanda eléctrica suavizada, onde non percibimos o efecto semanal, ao igual que na estimación da función de autocorrelación simples (Figura 5.29b). Este último correlograma indica a presenza de tendencia, que vén dada pola maior demanda nos meses de inverno e verán. Ademais, na Figura 5.31, onde se presenta o diagrama de caixas por meses desta variable, mostra que o patrón mensual ao longo destes anos se corresponde co visto para o 2017.

En conclusión, podemos ver que o nivel da demanda eléctrica diaria ao longo dos anos comprendidos

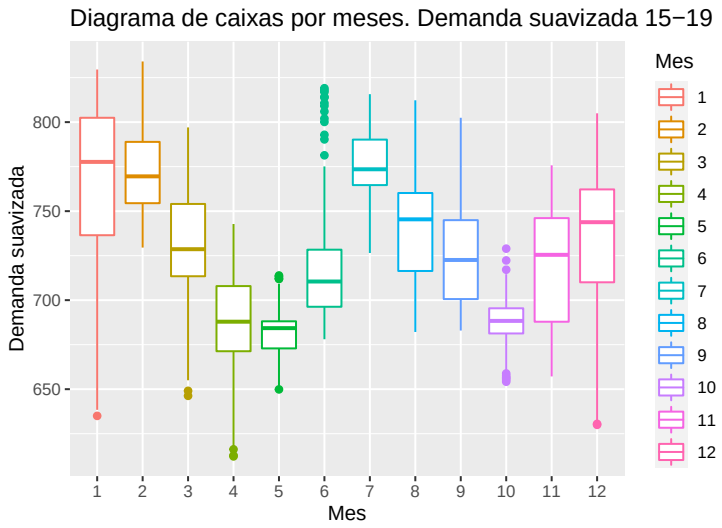


Figura 5.31: Diagrama de caixas por meses da demanda diaria total de electricidade suavizada no territorio nacional entre 2015 e 2019.

entre 2015 e 2019 mantívose sen experimentar cambios bruscos e a tendencia anual é a mesma en todos os casos.

5.3. Análise exploratoria da demanda eléctrica en 2020

O 15 de marzo de 2020 comeza o confinamento en España como medida de contención da COVID-19. En consecuencia, moitas actividades comerciais víronse interrompidas, causando un forte impacto na economía do país. A demanda de electricidade, onde a industria ten un gran peso, tamén sufriu un significativo descenso. Polo tanto, a dinámica da demanda eléctrica ao longo deste ano pode considerarse atípica respecto dos anos predecesores e o seu estudo merece unha sección aparte.

Na Figura 5.32 móstranse os gráficos secuenciais da demanda eléctrica (en XWh) no territorio nacional en 2020 en negro e en 2017 en gris e liña punteada. A parte sombreada correspóndese co confinamento no noso país, dende o 15 de marzo ata o 21 de xuño de 2020. Podemos ver que a serie temporal no 2020 presenta unha posible tendencia e patrón repetitivo semanal, tal e como pasaba nos anteriores anos. Pero a tendencia, neste caso, consta dunha gran baixada a mediados de marzo, coincidindo co inicio do citado confinamento. A mediados de abril, o nivel comeza a medrar levemente e nos meses de verán podemos ver unha maior demanda, debido ás altas temperaturas. Logo, a forma correspóndese aproximadamente coa dos anteriores anos. Así, a gran diferenza respecto destes anos reside no significativo descenso que se produce en marzo de 2020 e que, aínda que despois comeza a subir o nivel, non alcanza o dos anos predecesores. Isto pode verse na gráfica facilmente comparando as dinámicas entre o 2020 e o 2017.

Se nos fixamos na estimación da función de autocorrelacións simples da serie, que se presenta na Figura 5.33, podemos ver que existen repuntes positivos fortes nos múltiplos de 7. Isto, xunto co patrón repetitivo que víamos no gráfico secuencial, é indicativo da presenza de compoñente estacional con período $s = 7$. Polo tanto, concluímos que a serie non é estacionaria. Doutra banda, a variabilidade da serie é constante.

A continuación, na Táboa 5.10 móstranse as medidas de posición da demanda de electricidade neste

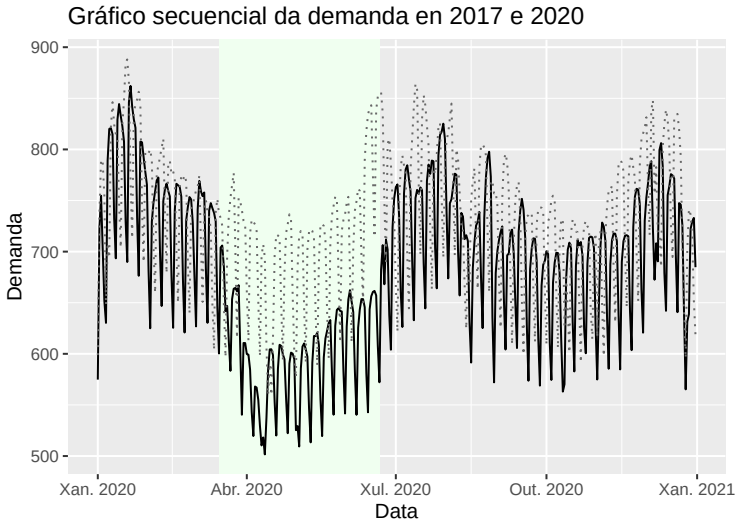


Figura 5.32: Gráficos secuenciais da demanda diaria total de electricidade no territorio nacional no ano 2020 en negro e no ano 2017 en liña punteada. A zona sombreada correspóndese co confinamento no país.

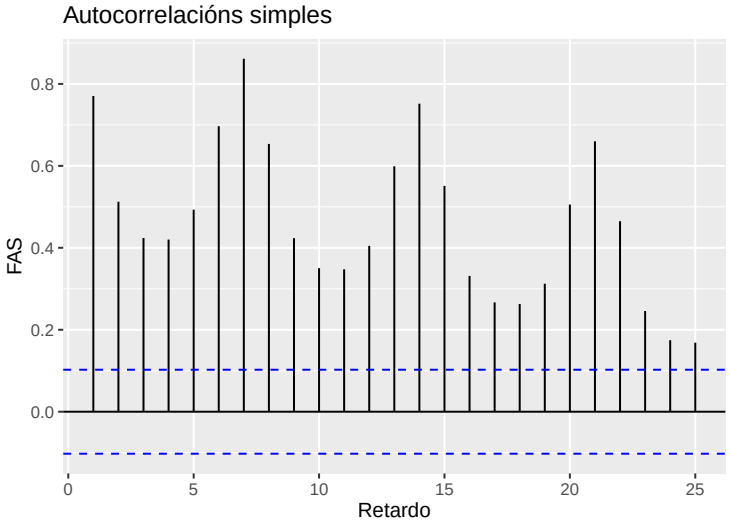


Figura 5.33: Autocorrelacións simples mostrais da serie de demanda eléctrica en 2020.

Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
501.6	625.1	690.5	682.6	743.5	862.1

Táboa 5.10: Medidas de posición da demanda eléctrica en 2020 (en XWh).

ano. Podemos ver que o mínimo é 501.6 XWh e o máximo 862.1 XWh. A mediana é 690.5 XWh e a media é 682.6 XWh. O primeiro cuartil coincide cun valor de demanda de 625.1 XWh e o terceiro con 743.5 XWh. Como era de esperar, todas as medidas son inferiores que as de 2017.

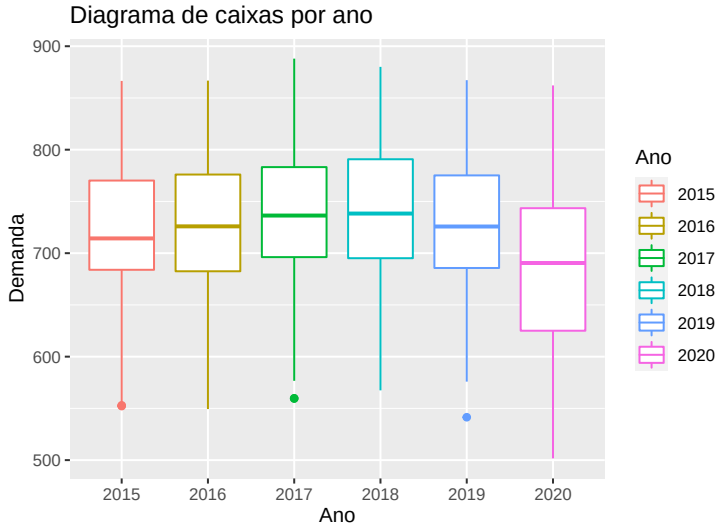


Figura 5.34: Diagrama de caixas da demanda diaria total de electricidade por ano entre 2015 e 2020.

Pode resultar de interese medir canto baixou a demanda eléctrica en 2020 respecto dos anos precedentes. Na Figura 5.34 móstranse os diagramas de caixas por ano da demanda eléctrica entre 2015 e 2020. Podemos notar o gran descenso que se produciu neste último ano por mor da pandemia causada pola COVID-19. A mediana da demanda eléctrica tende a crecer levemente entre 2015 e 2018, de 714 a 738 XWh, respectivamente. No ano 2019 ten un valor de 726 XWh, isto é, dáse unha pequena baixada de nivel e no 2020 volve descender ata 691 XWh. Polo tanto, a mediana da demanda eléctrica baixou en 37 XWh no 2020 respecto da media das medianas da variable entre 2015 e 2019. Ademais, no diagrama de caixas de 2020 existe asimetría negativa (mentres que nos anteriores anos a asimetría é positiva) e non presenta valores atípicos.

Continuando coa análise exploratoria da demanda eléctrica en 2020, mostramos na Figura 5.35, á esquerda, o histograma de frecuencias absolutas e, á dereita, o histograma coa función tipo núcleo da densidade en vermello. Se nos fixamos no histograma, podemos ver que hai un maior volume de datos á esquerda, ao contrario do que pasaba nos anteriores anos. A función tipo núcleo é asimétrica e presenta dúas modas.

Tamén é interesante o estudo desta variable por meses, polo que presentamos o resumo das medidas de posición na Táboa 5.11 e os seus respectivos diagramas de caixas na Figura 5.36. Se nos fixamos na mediana, podemos ver que a demanda diminúe ata maio, onde sube levemente. Nos meses de verán hai unha demanda superior, entre eles sendo maior en xullo. Logo tende a baixar ata novembro. Neste caso, decembro ten un nivel superior a novembro, ao contrario do que pasaba nos anteriores anos. Ademais, podemos ver unha notoria baixada en marzo e abril. Con isto percíbese a tendencia que describíamos ao principio. Nos diagramas de caixas podemos ver que presentan asimetría e xaneiro e xuño amosan observacións atípicas. Tamén podemos ver unha maior dispersión dos datos en marzo,

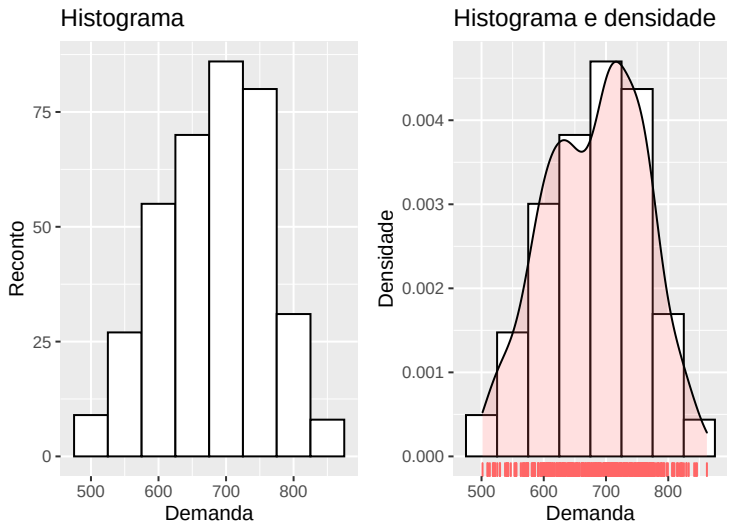


Figura 5.35: Histograma de frecuencias absolutas (esquerda) e histograma e función tipo núcleo da densidade (dereita) da demanda total de electricidade en 2020.

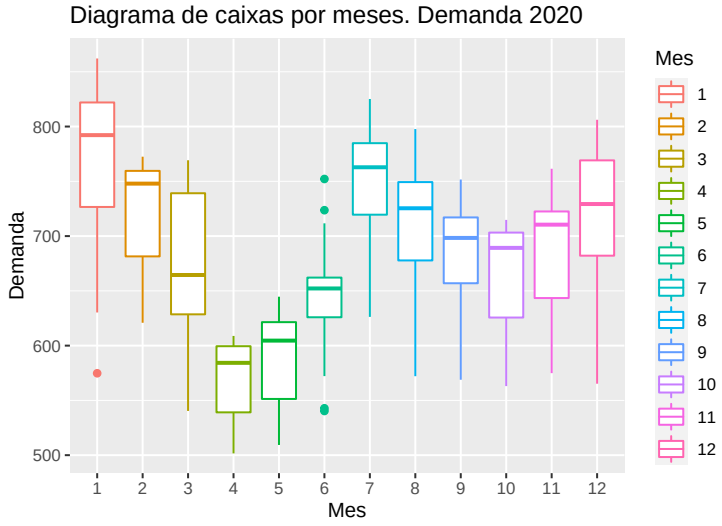


Figura 5.36: Diagrama de caixas por meses da demanda en 2020.

onde dou comezo o confinamento.

Por último, tamén fixemos unha análise da demanda eléctrica segundo os días. Na Táboa 5.12 móstrase o resumo das medidas de posición por días e os correspondentes diagramas de caixas poden verse na Figura 5.37. Se nos fixamos na mediana, podemos ver que entre os días laborais, mércores e xoves teñen lixeiramente unha maior demanda. Ao igual que pasaba nos anteriores anos, a demanda eléctrica nos fins de semana é menor que nos días entre semana. Así o indican, tamén, os diagramas de caixas. Pódese atopar máis información, dende un enfoque descritivo, sobre o cambio no nivel e patrón da demanda eléctrica en España durante o confinamento en Santiago et al (2021).

Mes	Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
Xaneiro	574.8	726.6	792.2	768.6	822	862.1
Febreiro	620.9	681.5	747.9	723.5	759.5	772.5
Marzo	540.4	628.6	664.5	674.9	739.1	769.2
Abril	501.6	539.1	584.3	568.6	599.5	608.9
Maio	509.3	551.3	604.6	590.4	621.4	644.7
Xuño	540.3	625.9	652.2	645	662.1	752.1
Xullo	626.2	719.5	762.8	748.0	784.8	825.2
Agosto	572.1	677.8	725.4	711.6	749.3	797.7
Setembro	568.9	656.9	698.3	683.3	717	751.6
Outubro	563.1	625.6	689.2	666.5	703.2	714.8
Novembro	574.9	643.4	710.4	687.8	722.5	761.5
Decembro	565.2	682.1	729.3	720.5	769.1	806.2

Táboa 5.11: Medidas de posición da demanda eléctrica por meses en 2020 (en XWh).

Día	Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
Luns	537.8	640.2	693.2	691.2	739.9	846.1
Martes	567	670.8	711.2	712.7	755.3	862.1
Mércores	554.9	664.5	718.1	714.8	763.9	841.7
Xoves	529.8	661.2	719.1	714.8	765.8	829.4
Venres	510.4	654.6	712.2	705	759.1	821.1
Sábado	518.2	610.1	638.9	640.5	682	738.1
Domingo	501.6	567.4	602	597.7	634.3	693.4

Táboa 5.12: Medidas de posición da demanda eléctrica por días en 2020 (en XWh).

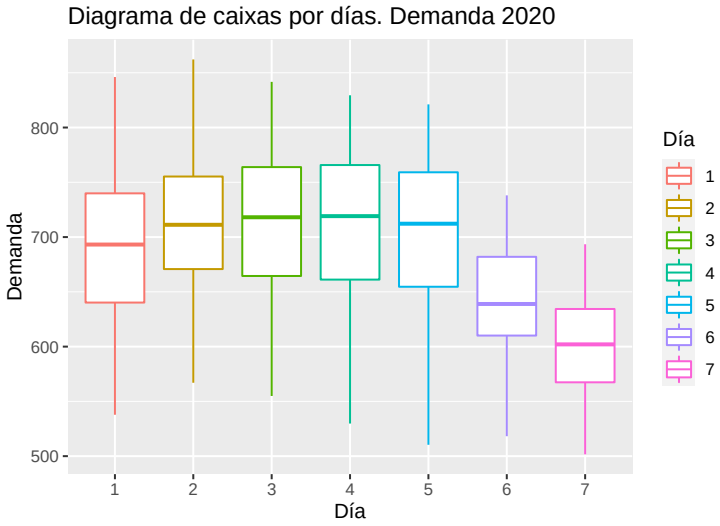


Figura 5.37: Diagrama de caixas por días da demanda eléctrica en 2020.

Día	Luns	Martes	Mércores	Xoves	Venres	Sábado
Martes	0.7130					
Mércores	0.6353	1				
Xoves	0.6353	1	1			
Venres	1	1	1	1		
Sábado	0.0019	2.4×10^{-6}	4.3×10^{-6}	3.8×10^{-6}	3.1×10^{-5}	
Domingo	1.1×10^{-8}	2.1×10^{-11}	3.8×10^{-11}	3.2×10^{-11}	7.6×10^{-10}	0.0019

Táboa 5.13: P-valores asociados ao estatístico de contraste do test de rangos con signo de Wilcoxon.

Sería interesante contrastar o posible efecto semanal. Para esta tarefa precisamos saber se as mostras de demanda por día son independentes ou non. Así, aplicamos o test de independencia τ de Kendall entre pares de días, resultando, a un nivel de significación $\alpha = 0.05$, no rexeitamento en todos os casos da hipótese nula de independencia. En consecuencia, para contrastar a igualdade de medianas da demanda de electricidade por días, aplicamos o test de rangos con signo de Wilcoxon entre pares. Na Táboa 5.13 móstranse os p-valores asociados ao estatístico de contraste. Aos niveis de significación habituais, rexeitamos a hipótese nula tanto en sábados como en domingos, concluíndo na existencia deste efecto.

Vimos de estudar a demanda diaria de electricidade no territorio nacional en 2020 e en anos precedentes. A dinámica atípica da demanda eléctrica neste último ano débese ao considerable descenso a mediados de marzo, como consecuencia do comezo do confinamento en España. O nivel da demanda eléctrica a partir deste mes é inferior aos correspondentes meses nun ano usual. Os niveis de demanda

eléctrica van incrementándose dende mediados de abril, pero aínda non conseguiu alcanzar o nivel dun ano normal. Polo resto, todas as series de demanda sofren o efecto de fin de semana, que pode resolverse aplicando medias móbiles con xanela $h = 7$, tal e como vimos na Sección 5.1.1, e dun efecto mensual provocado aparentemente pola temperatura.

Do estudo da demanda eléctrica no ano 2020 podemos concluír que esta variable capta os cambios bruscos no estado da economía, como, neste caso, o confinamento. Polo tanto, a súa selección para a construción dun primeiro indicador do seguimento da actividade económica semella ser moi axeitada.

Capítulo 6

Limpeza da demanda eléctrica

Lembremos que o noso fin é o desenvolvemento dun índice de alta frecuencia para o seguimento da actividade económica en España e, como punto de partida, tomamos a variable de demanda eléctrica diaria total no territorio nacional para a súa creación. Pero buscamos que a dita variable estea limpa de estacionalidades e de efectos de variables esóxenas, como a temperatura, pois interésanos traballar só coa parte da información da demanda eléctrica relacionada coa actividade económica. No Capítulo 5 estudamos de forma exploratoria a dinámica da demanda de electricidade no país durante 2017, un ano usual no senso de evolución económica, e o 2020, que presenta un comportamento atípico como consecuencia do confinamento debido á pandemia de COVID-19. Vimos que existe un efecto semanal, que pode solucionarse mediante medias móbiles simples con xanela $h = 7$, e un efecto mensual, aparentemente debido á temperatura. Para atallar o problema que supón este último efecto, estudáronse diferentes vías.

Neste capítulo descríbense as principais metodoloxías que consideramos para limpar a demanda eléctrica. Na Sección 6.1 preséntase a corrección de atípicos da demanda de electricidade, dado que é esta variable corrixida a que consideramos para limpar de efectos semanais e mensuais. Logo, na Sección 6.2 presentamos a suavización da demanda eléctrica para eliminar o efecto semanal. O seguinte paso é eliminar o efecto mensual, no que se centra a Sección 6.3, presentando diferentes vías para atallar o problema. Unha vez que corriximos este último efecto, podemos empregar esta demanda eléctrica para a construción dun indicador.

Denominaremos por demanda eléctrica limpa á demanda de electricidade (corrixida de atípicos) unha vez que se lle eliminaron os efectos semanal e mensual. O primeiro pode eliminarse mediante medias móbiles con xanela $h = 7$, tal e como vimos na Subsección 5.1.1, mentres que a corrección do segundo efecto consta dunha maior dificultade, tal e como mostramos na Sección 6.3.

6.1. Substitución de observacións atípicas

No Capítulo 5 fixemos unha análise exploratoria da demanda eléctrica en 2017 e puidemos ver, en particular na Subsección 5.1.2, que existen observacións atípicas motivadas por festividades, que se repiten cada ano. A súa influencia nos resultados fainos pensar na importancia de corrixir as ditas observacións, evitando posibles problemáticas no futuro.

Para a realización desta tarefa botaremos man da función *tsoutliers* do paquete *forecast* (Hyndman et al 2020). Así, partimos dos datos brutos de demanda eléctrica entre 2015 e 2019, cuxo gráfico secuencial pode verse na Figura 6.1 (arriba). Xa estudamos que esta serie presenta estacionalidade, polo que a dita función emprega o método STL para a identificación e estimación de substitutos de datos atípicos. Dado que a súa aplicación á serie completa non capta todos os atípicos que, por exemplo, captaría nun só ano, aplicámola ano a ano e corriximos estas observacións polas propostas

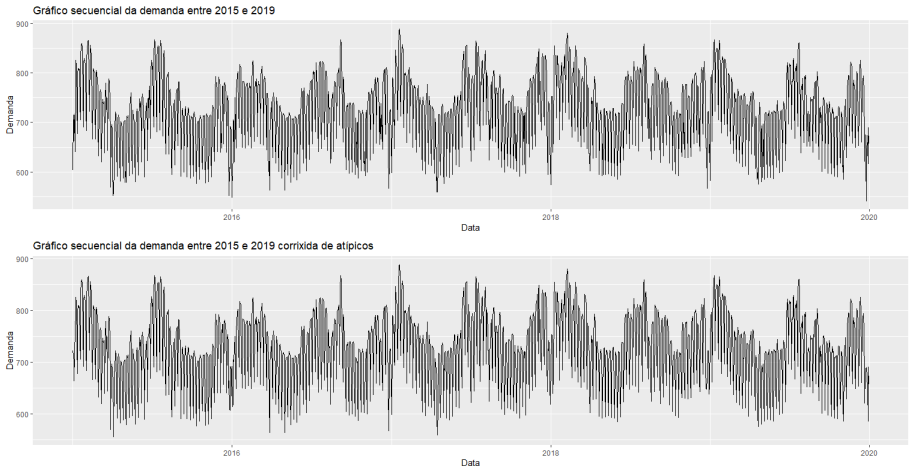


Figura 6.1: Gráfico secuencial da demanda diaria total de electricidade no territorio nacional entre 2015 e 2019 bruta (arriba) e corrixida de observacións atípicas (abaixo).

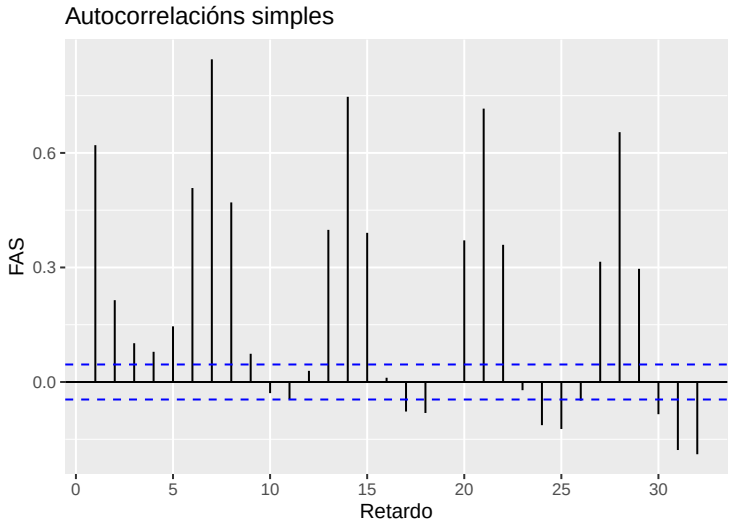


Figura 6.2: Autocorrelacións simples mostrais da demanda eléctrica corrixida de atípicos entre 2015 e 2019.

que ofrece a propia función. Na Figura 6.1 (abaixo) móstrase o gráfico secuencial da serie de demanda eléctrica corrixida de datos atípicos. Se a comparamos coa serie bruta, podemos ver que certas baixadas pronunciadas que se daban, por exemplo, a principios e finais de ano e nas semanas Santas, xa non se atopan na serie corrixida.

Durante este capítulo, a non ser que se indique o contrario, traballaremos coas observacións de demanda eléctrica entre 2015 e 2019 corrixidas de atípicos mediante este método.

6.2. Corrección do efecto semanal

Á vista do gráfico secuencial da demanda eléctrica corrixida de atípicos (Figura 6.1 abaixo), podemos ver que existe un patrón repetitivo. Se analizamos as súas autocorrelacións simples mostrais, que se mostran na Figura 6.2, observamos a existencia de repuntes positivos fortes nos retardos múltiplos de 7. En consecuencia, a serie presenta unha compoñente estacional de período $s = 7$, isto é, existe un efecto semanal, que nos interesa corrixir para obter unha serie máis homoxénea.

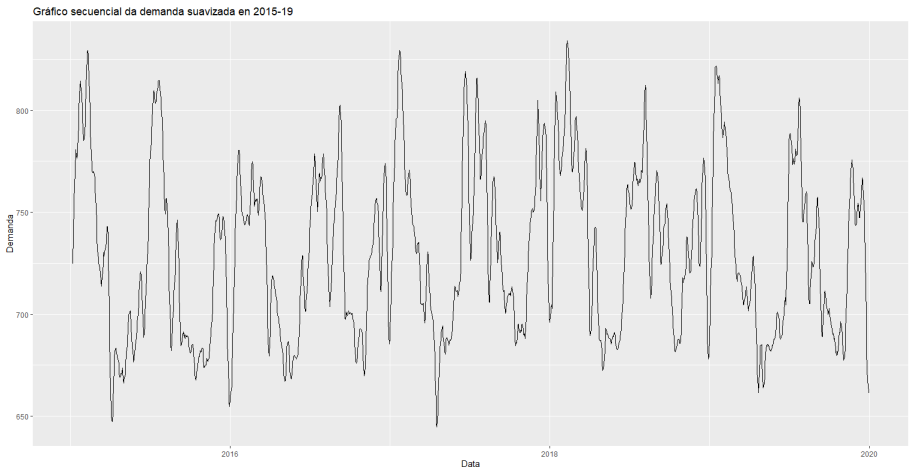


Figura 6.3: Gráfico secuencial da demanda eléctrica (corrixida de atípicos) suavizada entre 2015 e 2019.

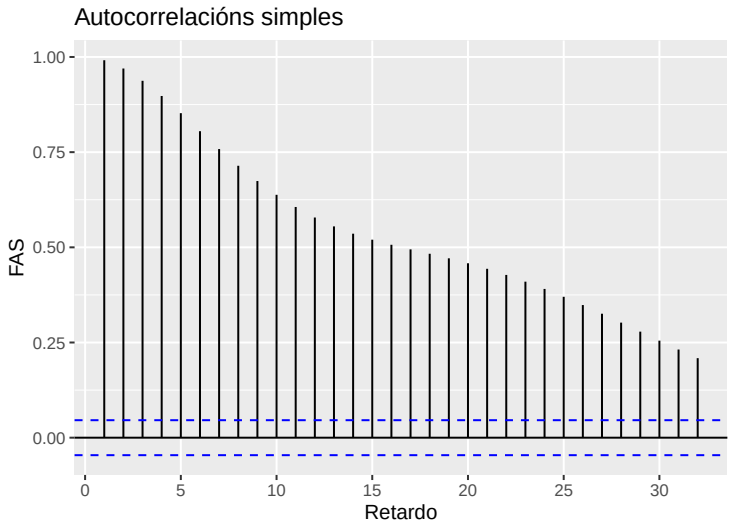


Figura 6.4: Autocorrelacións simples mostrais da demanda eléctrica suavizada entre 2015 e 2019.

Para eliminar a estacionalidade semanal que presenta a demanda eléctrica aplicamos medias móbiles con xanela $h = 7$ e suprimimos os seis primeiros valores por non estar definidos. O resultado pode verse na Figura 6.3. Podemos ver que, efectivamente, non presenta patrón repetitivo semanal. Tamén

o indica a estimación da función de autocorrelacións simples, que se mostra na Figura 6.4, pois non existen repuntes fortes positivos nos múltiplos de 7. Ademais, as autocorrelacións simples mostrais mostran a presenza de tendencia debida ao efecto mensual como causa, en principio, da variación de temperaturas.

En conclusión, mediante a aplicación de medias móbiles con xanela $h = 7$ eliminamos a compoñente estacional semanal que presenta a serie de demanda eléctrica.

6.3. Corrección do efecto mensual

Unha vez que se corrixiu o efecto semanal da demanda eléctrica, podemos ver no seu gráfico secuencial (Figura 6.3) que existe certo efecto mensual, onde segundo en que estación do ano esteamos existe unha maior ou menor demanda. Polo tanto, o que compete neste punto é corrixir o dito efecto, para o que se probaron diferentes vías. Na Subsección 6.3.1 preséntase o primeiro procedemento considerado, onde partimos dun modelo de regresión que explica a relación entre a demanda eléctrica e a temperatura, resultando nun camiño inadecuado para o noso obxectivo. Logo, na Subsección 6.3.2 modelamos a variación da demanda eléctrica e a temperatura co fin de limpar a demanda mediante este modelo, pero o resultado non se correspondía co esperado. Por último, na Subsección 6.3.3 tratamos de corrixir o efecto mensual que presenta a demanda eléctrica mediante a variación porcentual da mediana por meses. Esta metodoloxía proporciona uns resultados convincentes, polo que o seguinte paso, logo da corrección do efecto mensual da demanda eléctrica, é a construción do índice mediante esta variable limpa.

6.3.1. Primeira metodoloxía

Na Subsección 5.1.2 presentamos un modelo de regresión polinómico local robusto para estudar a relación existente entre a demanda eléctrica e a temperatura suavizadas mediante medias móbiles con xanela $h = 7$ (para corrixir o efecto fin de semana no primeiro caso e mitigar as engurras no segundo) no ano 2017. Tomando como referencia este procedemento, imos tratar de obter a demanda eléctrica limpa, isto é, sen estacionalidade semanal nin efecto mensual, considerando este último como causa da variación de temperaturas.

Tomamos os datos de demanda de electricidade diaria total entre 2015 e 2019 (lembremos que excluímos o ano 2020 por ter un comportamento atípico, como vimos na Sección 5.3) corrixida de observacións atípicas, tal e como se explica na Sección 6.1. Logo, suavizamos tanto os datos de demanda eléctrica como de temperatura entre 2015 e 2019 aplicando medias móbiles con xanela $h = 7$, eliminando a compoñente estacional semanal no primeiro caso e aliviando rugosidades no segundo. O seguinte paso é limpar o efecto que a temperatura exerce sobre a demanda. Para isto, axustaremos un modelo de regresión aos datos, considerando como mostra de adestramento as observacións entre 2015 e 2018 e como mostra de test, as observacións no ano 2019. É dicir, axustaremos un modelo de regresión coa mostra de adestramento e avaliaremos o rendemento do mesmo mediante a mostra de test.

Suavizado dos datos

Xa estudamos que a demanda eléctrica presenta compoñente estacional semanal e, ademais, un efecto mensual que cremos que vén dado pola variación da temperatura ao longo do ano. Estes efectos deben corrixirse para obter unha serie homoxénea á hora de analizala. Na Sección 6.2 xa explicamos a corrección do efecto semanal da demanda eléctrica. Centrémonos agora na análise gráfica da serie de temperatura media.

O gráfico secuencial da temperatura media diaria na estación de Madrid Retiro entre 2015 e 2019 móstrase na Figura 6.5. Intúese a presenza de tendencia e así o indica a estimación da función de autocorrelacións simples, que se mostra na Figura 6.6, e un patrón anual, como cabería esperar.

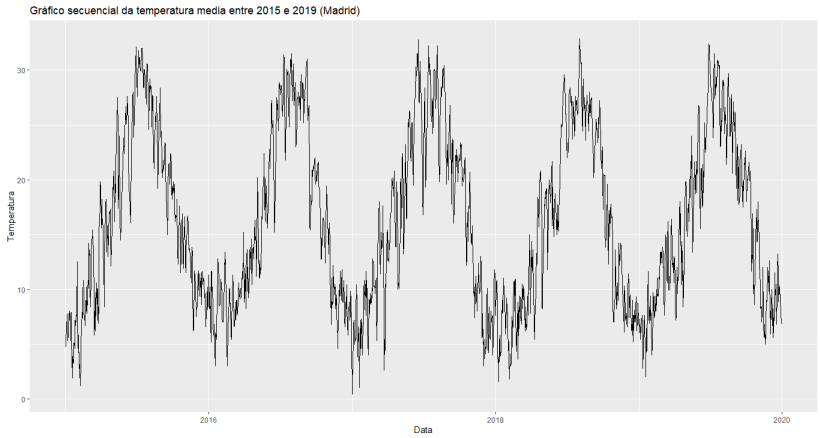


Figura 6.5: Gráfico secuencial da temperatura media diaria na estación Madrid Retiro entre 2015 e 2019.

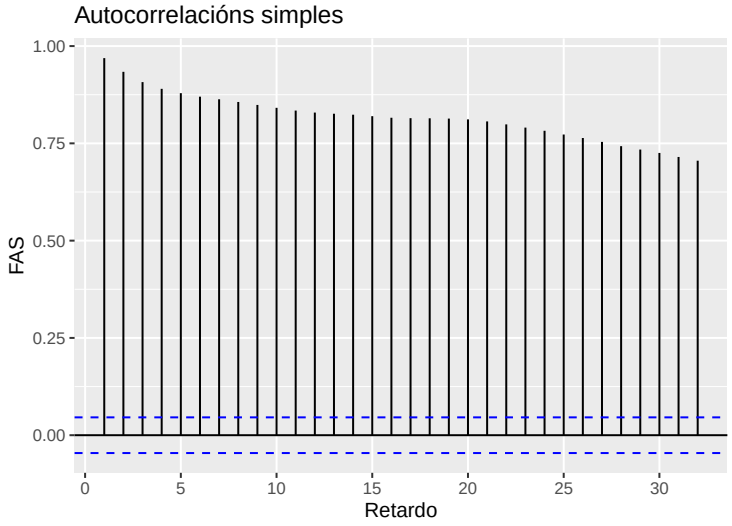


Figura 6.6: Autocorrelacións simples mostrais da serie de temperatura media diaria dende 2015 ata 2019.

Para aliviar as claras rugosidades que presenta a serie, aplicamos medias móbiles con xanela $h = 7$ aos datos de temperatura media diaria entre 2015 e 2019, cuxo gráfico secuencial se mostra na Figura 6.7. Podemos ver que esta presenta tendencia, que tamén nos indican as súas autocorrelacións simples mostrais, que se presentan na Figura 6.8.

Na Figura 6.9 móstrase o diagrama de dispersión dos datos de demanda eléctrica e temperatura media suavizadas mediante medias móbiles con xanela $h = 7$. Existe unha clara relación entre ambas variables, en forma de parábola. O seguinte paso será eliminar o efecto que a temperatura exerce sobre a demanda eléctrica.

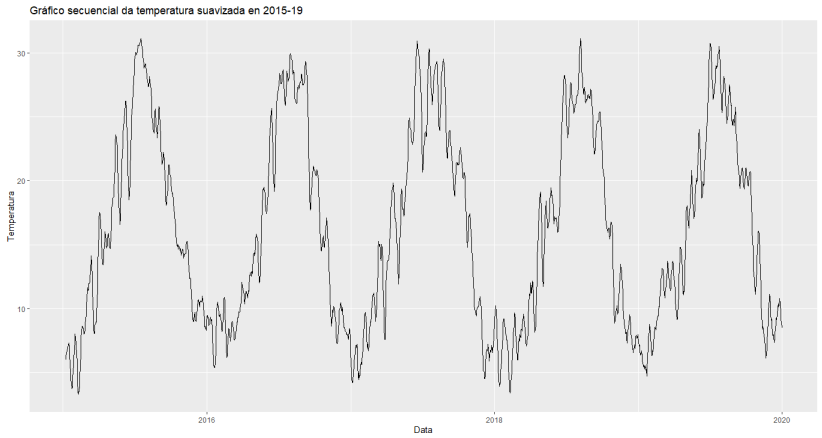


Figura 6.7: Gráfico secuencial da temperatura media diaria suavizada na estación Madrid Retiro entre 2015 e 2019.

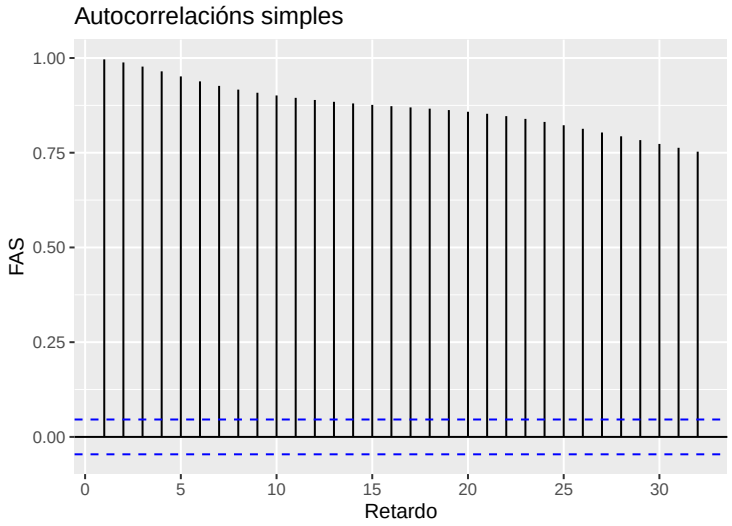


Figura 6.8: Autocorrelacións simples mostrais da temperatura suavizada.

Axuste dun modelo de regresión polinómico local robusto

Unha vez que suavizamos os datos de demanda eléctrica e temperatura, podemos ver no seu diagrama de dispersión (Figura 6.9) que a primeira segue véndose influenciada pola segunda. Polo tanto, trataremos de eliminar o efecto da temperatura sobre a demanda eléctrica mediante o axuste dun modelo de regresión. Para isto, dividimos os datos suavizados en mostras de adestramento e de test. As observacións que se atopen entre 2015 e 2018 constituirán a mostra de adestramento, coa que se axustará o modelo, e as observacións correspondentes a 2019 conformarán a mostra de test, coa que se avaliará o axuste.

Consideramos que a mellor opción para o axuste é tomar un modelo de regresión no ámbito non paramétrico. Aínda que vimos de corrixir as observacións atípicas da demanda eléctrica que capta a función *tsoutliers* da libraría *forecast* (Hyndman et al 2020), no diagrama de dispersión dos datos (Figura 6.9) percibimos certos puntos que se afastan da dinámica esperada. Polo tanto, gustaríanos

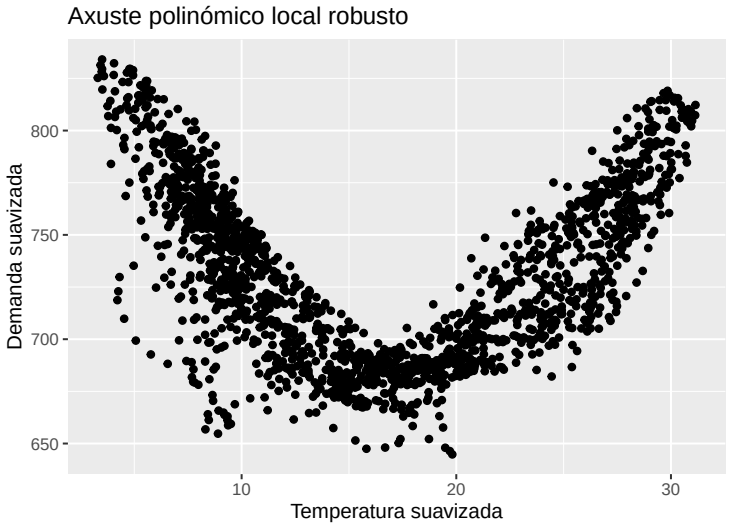


Figura 6.9: Diagrama de dispersión da demanda eléctrica e temperatura suavizadas entre 2015 e 2019.

que o modelo de regresión fose robusto.

Optamos por axustar un modelo de regresión polinómico local robusto, cuxa variable resposta é a demanda eléctrica suavizada e a variable explicativa é a temperatura suavizada. A selección da xanela fíxose mediante validación cruzada con criterio de erro a mediana dos erros en valor absoluto, resultando nun valor 0.1. O coeficiente de determinación deste axuste é aproximadamente 0.75, polo que o modelo explica un 75 % da variabilidade da demanda eléctrica. É de esperar que non explique toda a variabilidade, pois a demanda eléctrica tamén dependerá, por exemplo, do desenvolvemento de certas actividades económicas e non só da temperatura. O axuste pode verse na Figura 6.10.

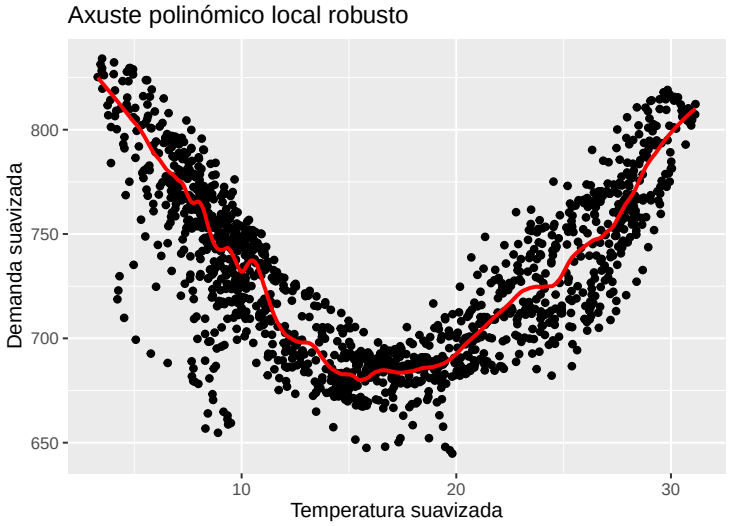


Figura 6.10: Axuste do modelo de regresión polinómico local robusto á mostra de adestramento.

Na Figura 6.11 móstrase o gráfico secuencial dos residuos do axuste, que se corresponderían coa

demanda eléctrica limpa do efecto da temperatura. Podemos ver que a serie de residuos non presenta tendencia nin compoñente estacional e a varianza semella ser constante, polo que a serie é estacionaria. A mesma conclusión obtemos á vista das estimacións das funcións de autocorrelacións simples e parciais, que se mostran na Figura 6.12. Se aplicamos o test de Dickey-Fuller aumentado mediante a función *adf.test* da libraría *tseries* (Trapletti e Hornik 2019), obtemos un p-valor asociado ao estatístico de contraste menor a 0.01, polo que rexeitamos a hipótese de non estacionariedade a un nivel de significación $\alpha = 0.05$. Polo tanto, a serie de residuos é estacionaria, aínda que non atopamos un proceso ARMA que fose un posible xerador válido.

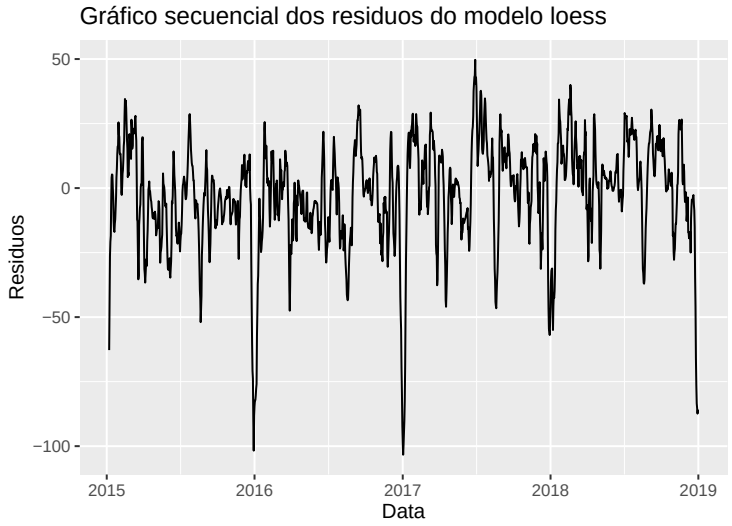


Figura 6.11: Residuos do axuste do modelo de regresión polinómico local robusto.

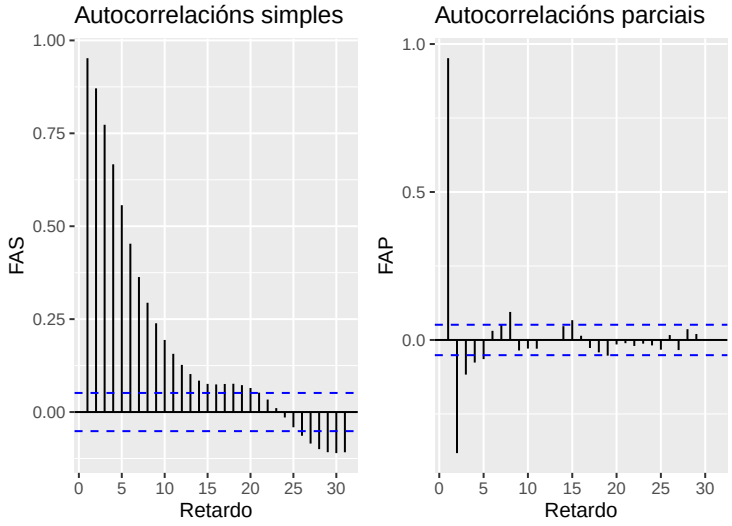


Figura 6.12: Autocorrelacións simples e parciais mostrais dos residuos do axuste.

Na Figura 6.13 móstranse os gráficos secuenciais da demanda eléctrica bruta, suavizada e limpa, de arriba a abaixo respectivamente. Podemos ver, posiblemente mellor comparando a demanda suavizada e limpa, que os residuos do axuste teñen nivel constante, sen presenza do efecto mensual que víamos nas outras dúas series.

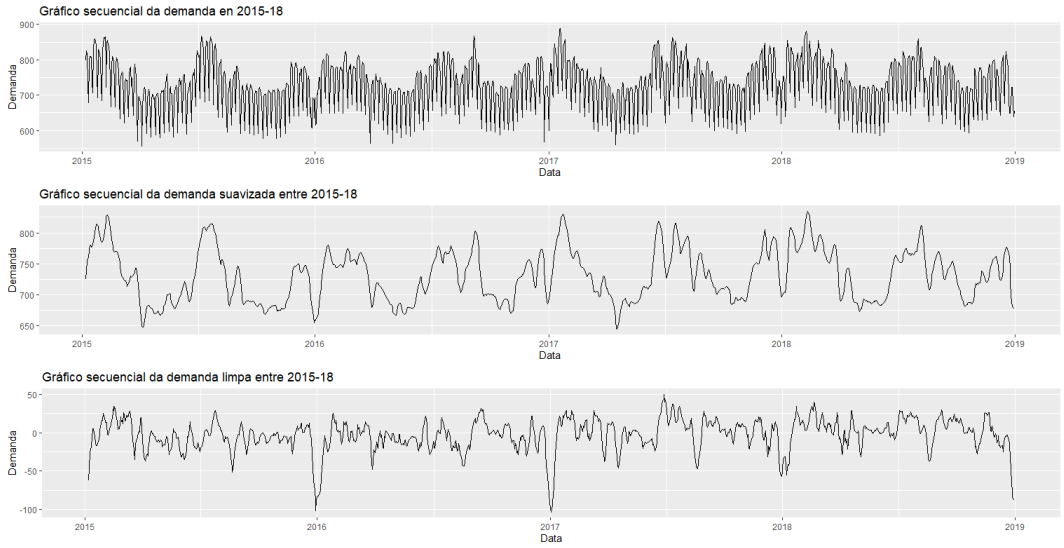


Figura 6.13: Gráficos secuenciais da demanda eléctrica bruta (arriba), suavizada (medio) e limpa do efecto da temperatura (abaixo) entre 2015 e 2018.

Por último, empregamos a mostra de test para avaliar o axuste de regresión polinómico local robusto. Tomamos como medida de precisión das predicións o Pseudo R cadrado, que se define como:

$$\tilde{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y})},$$

onde y_i denota a observación i-ésima da variable Y, $\hat{y}_i$ á predición e $\bar{y}$ á media das observacións. Neste caso, o Pseudo R cadrado ten un valor aproximado de 0.72, indicando unha predición aceptable.

Continuando coa avaliación do modelo, xeramos un gráfico de dispersión de observacións fronte a predicións. No caso de que o axuste fose correcto, estes puntos deberían moverse en torno á recta $y = x$. Na Figura 6.14 móstrase o dito gráfico e podemos ver que os puntos atópanse ao redor da recta, a excepción dunhas poucas observacións.

Na Figura 6.15 móstrase a demanda eléctrica suavizada mediante medias móbiles en 2019 en vermello e a predición do modelo axustado no mesmo ano en azul. Podemos ver que, en xeral, capta axeitadamente a dinámica da serie. Recordemos que o coeficiente de determinación do axuste é 0.75, polo que non podemos esperar que a predición sexa perfecta.

Conclusión

Mediante o procedemento que vimos de describir obtemos unha demanda eléctrica limpa, no senso de que non presenta estacionalidade semanal nin mensual. Unha vez que conseguimos a demanda eléctrica limpa, interésanos partir desta para a construción do índice. Pero neste punto xorde un problema. Os residuos (a demanda eléctrica limpa) móvense en torno a cero, polo que presentan valores

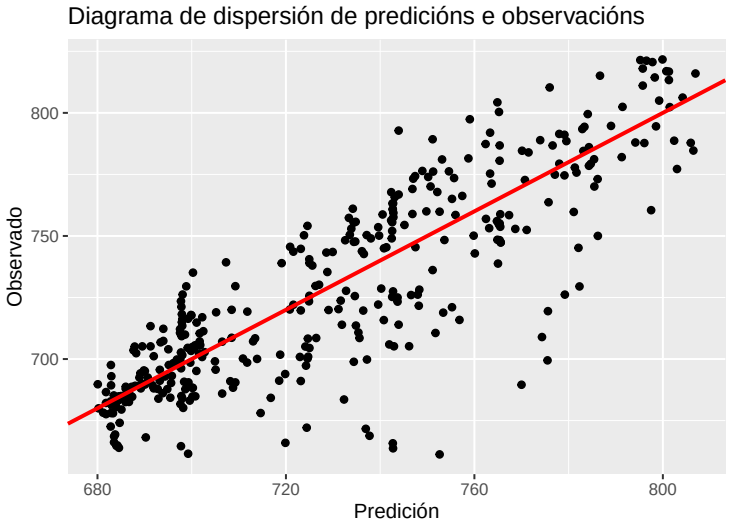


Figura 6.14: Diagrama de dispersión de observacións fronte a predicións.

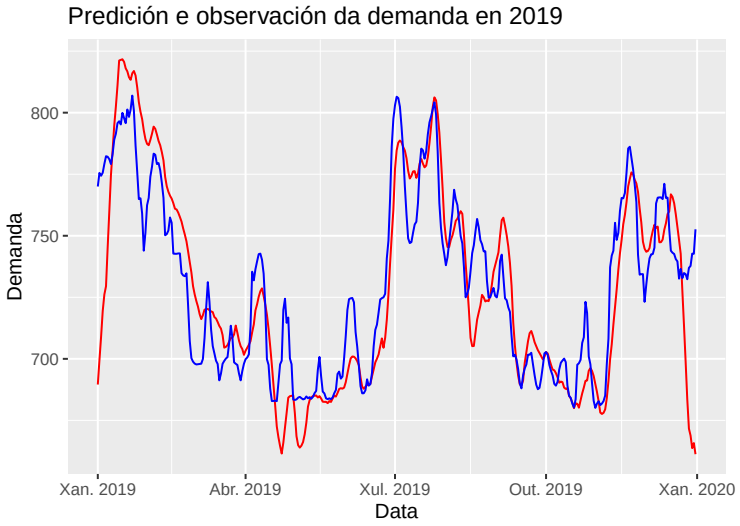


Figura 6.15: Gráfico secuencial da serie de demanda eléctrica suavizada en vermello e da predición do modelo axustado en azul no ano 2019.

negativos. É difícil xustificar que a demanda eléctrica, unha vez eliminado o efecto da temperatura, presente observacións negativas. Ademais, os resultados da construción do índice mediante esta demanda limpa e o cálculo das variacións do mesmo, que poden resultar de interese para medir os cambios ao longo do tempo e para comparar distintos períodos, non teñen rigor debido aos valores negativos. O número índice pode construírse a pesar de considerar unha variable con valores positivos e negativos, pero non pasa o mesmo á hora de calcular as variacións do índice. A serie de variacións debe moverse no intervalo real positivo, pois suponse que se parte de variables positivas. Polo tanto, este camiño queda descartado.

6.3.2. Segunda metodoloxía

O anterior procedemento non resulta válido pola presenza de magnitudes negativas na demanda eléctrica limpa, pois non son fáciles de interpretar. Ademais, esta variable proporciona tamén magnitudes negativas no índice, polo que o cálculo das súas variacións non tería rigor. Polo tanto, buscamos obter a demanda eléctrica limpa cunha escala que teña sentido. Para esta tarefa, enfocamos o problema a través do axuste dun modelo en variacións da demanda eléctrica e da temperatura, para intentar ver que parte da porcentaxe do crecemento da demanda é debido á variación da temperatura.

Dada unha serie de tempo y_t, a diferenza regular $y_t - y_{t-1}$ é un indicador de crecemento absoluto de y_t. Neste caso, consideraremos

$$\log y_t - \log y_{t-1} \approx \frac{y_t - y_{t-1}}{y_{t-1}},$$

que representa a taxa logarítmica de variación dunha variable. É un indicador de crecemento relativo e no caso de que se multiplique por 100, representa a taxa de crecemento porcentual da variable.

Relación entre a variación da demanda eléctrica e da temperatura

Volvemos partir dos datos de demanda eléctrica corrixida de atípicos, tal e como se mostra na Sección 6.1, e suavizados para eliminar o efecto semanal, tal e como se explica na Sección 6.2, entre os anos 2015 e 2019, deixando a un lado o ano 2020 pola súa dinámica atípica. Tamén consideramos a temperatura media entre estes anos e, dado que despois imos aplicar logaritmos e que existen observacións próximas a cero, convertemos estes datos, medidos en graos Celsius, a graos Kelvin[1]. Unha vez feito isto, o seguinte paso é suavizar os datos de temperatura media (en graos Kelvin) mediante medias móbiles con xanela $h = 7$, polas razóns que xa expuxemos anteriormente. A continuación, corresponde aplicar o logaritmo ás dúas variables, demanda e temperatura, suavizadas e, por último, diferenciamos regularmente ambas series de logaritmos. Denominaremos por variación dunha variable á diferenza regular do logaritmo da dita variable. O noso obxectivo reside en modelar a relación, no caso de que exista, das series de variacións de demanda eléctrica e temperatura e, unha vez obtida a variación da demanda limpa do efecto da variación da temperatura, desfaremos os cambios na primeira variable para obter a demanda eléctrica limpa, que esperamos que se mova nunha escala acorde coa da demanda bruta.

Na Figura 6.16 móstranse os gráficos secuenciais das series da diferenza regular do logaritmo da demanda eléctrica (arriba) e temperatura media (abaixo) suavizadas. Podemos ver que no caso da variación da demanda eléctrica non hai tendencia e non parece sufrir un efecto provocado pola temperatura. Ambas series semellan ser estacionarias e se aplicamos o contraste de Dickey-Fuller aumentado, obtemos en ambos casos un p-valor menor que 0.01, polo que rexeitamos a hipótese nula de non estacionariedade.

Vexamos se existe relación entre as variacións de demanda de electricidade e de temperatura media. Para isto, presentamos o diagrama de dispersión correspondente na Figura 6.17. Nel podemos ver que non existe relación entre as ditas variacións. Se nos fixamos na definición de variación que tomamos, diferenciamos regularmente o logaritmo das variables suavizadas mediante medias móbiles con xanela $h = 7$. Polo tanto, a observación i-ésima da demanda eléctrica suavizada correspóndese coa media dos 6 anteriores datos e a propia observación. Nestas circunstancias, se aplicamos unha diferenza regular, a diferenza entre a observación no pasado inmediato e no presente, tense en conta o pasado semanal e non só o destas dúas observacións. En consecuencia, consideraremos diferenzas estacionais de período 7 no lugar das diferenzas regulares.

Así pois, aplicamos unha diferenza estacional de orde 7 ao logaritmo da demanda eléctrica e da temperatura media suavizadas mediante medias móbiles con xanela $h = 7$. Á vista do diagrama de

[1] Un grao Celsius equivale a 274.15 graos Kelvin.

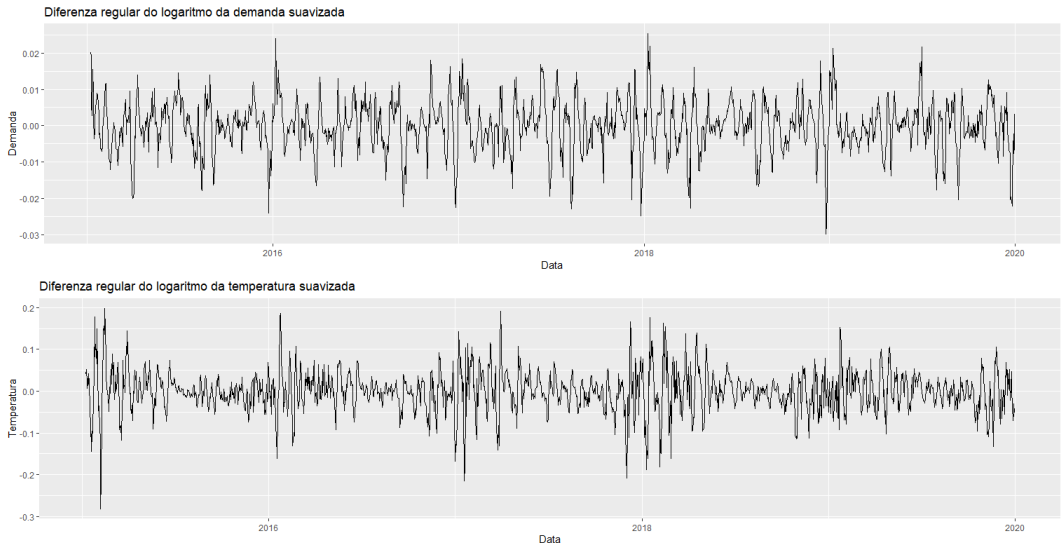


Figura 6.16: Gráficos secuenciais da variación da demanda eléctrica (arriba) e da temperatura media (abaixo).

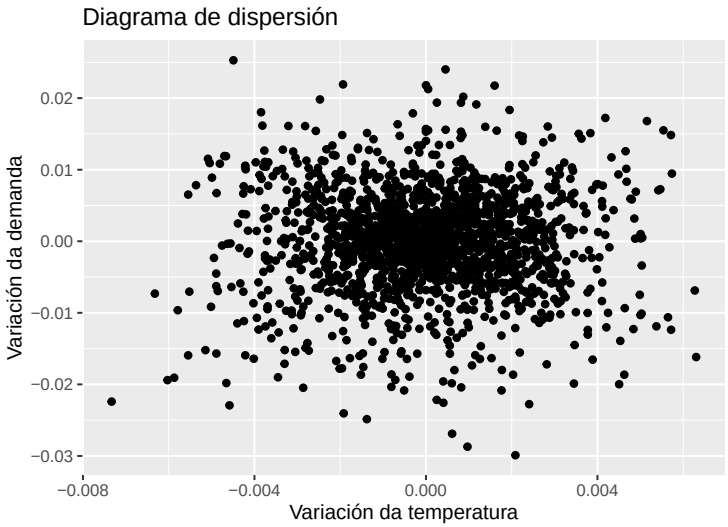


Figura 6.17: Diagrama de dispersión das variacións de demanda eléctrica e temperatura media.

dispersión das variacións (neste caso, denominamos por variación á diferenza estacional de orde 7 do logaritmo da variable suavizada mediante medias móbiles), que se mostra na Figura 6.18, tampouco vemos a existencia de relación.

En realidade, considerar a variación entre días consecutivos ou semanais da temperatura media non ten moito sentido. En cambio, considerar a variación da demanda eléctrica entre días semanais si o tería. Polo tanto, na seguinte subsección mostraremos un estudo da relación entre a diferenza semanal do logaritmo da demanda eléctrica suavizada e a temperatura media bruta en graos Celsius.

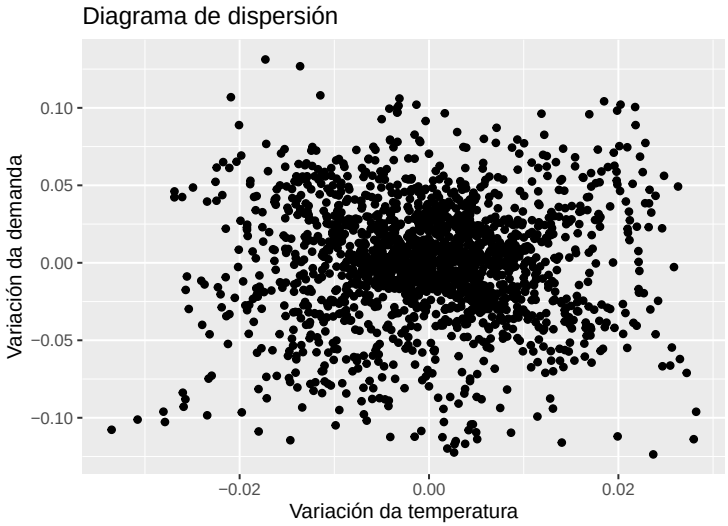


Figura 6.18: Diagrama de dispersión das variacións de demanda eléctrica e temperatura media.

Relación da variación da demanda eléctrica e a temperatura

Centrámonos agora na relación entre a variación da demanda eléctrica e a temperatura media bruta, entendendo por variación a diferenza semanal do logaritmo da variable suavizada por medias móbiles. O gráfico secuencial da variación da demanda eléctrica móstrase na Figura 6.19. Esta serie é estacionaria, polo que non presenta tendencia.

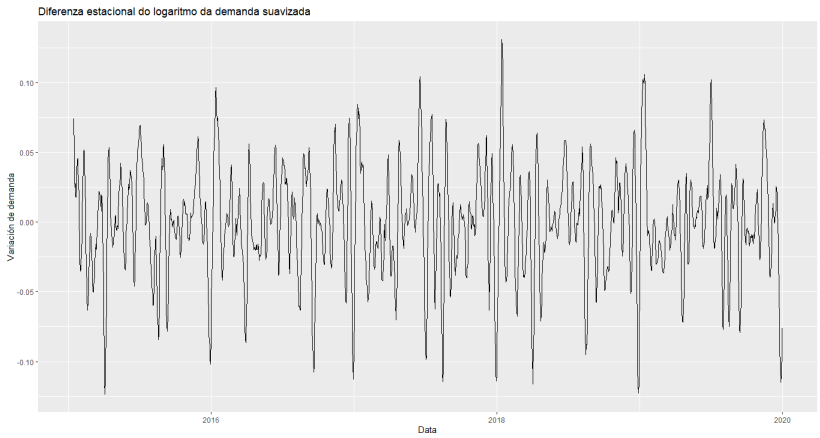
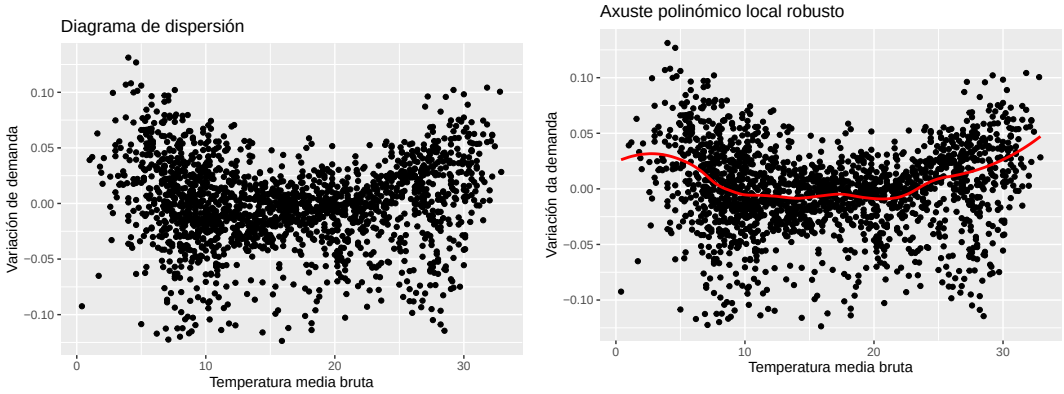


Figura 6.19: Gráfico secuencial da diferenza estacional de orde 7 do logaritmo da demanda eléctrica suavizada.

Na Figura 6.20a móstrase o diagrama de dispersión da variación da demanda eléctrica e a temperatura media bruta. É posible que exista unha débil relación entre ambas variables en forma de bañeira. Podemos tratar de modelala mediante un modelo de regresión polinómico local robusto, cuxa representación pode verse na Figura 6.20b e cuxo coeficiente de determinación é 0.086. O R^2 do axuste é moi baixo, pois podemos ver que no eixo da variación da demanda existe moita variabilidade e só un

0.086 % pode explicarse mediante a temperatura. En xeral, o axuste achégase a describir a tendencia que visualizamos mentalmente á vista do diagrama de dispersión, aínda que ao principio o dito axuste vese afectado polas observacións influíntes que se mostran abaixo á esquerda.



(a) Diagrama de dispersión.

(b) Axuste dun modelo de regresión polinómico local robusto.

Figura 6.20: Variación da demanda eléctrica e temperatura media bruta.

En resumidas contas, dado que a relación entre a variación da demanda eléctrica e a temperatura bruta é pobre, o axuste non presenta boas características. Polo tanto, continuando co propósito de limpar o efecto mensual (considerado froito da variación das temperaturas) da demanda eléctrica, trataremos de medir a relación entre a variación da demanda e a temperatura considerando a diferenza entre unha temperatura de referencia e a temperatura media diaria.

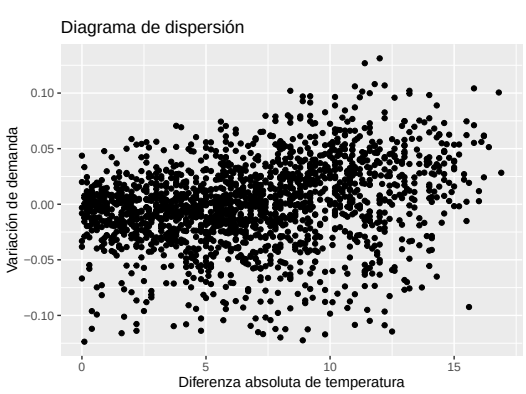
Relación da variación da demanda eléctrica e a diferenza de temperatura

Outra opción é considerar unha temperatura media base e traballar coas diferenzas bruta e absoluta entre a temperatura media diaria e esta referencia, onde no primeiro caso consideraríamos unha corrección diferente segundo o exceso ou falta. A media da temperatura media diaria no país entre 2015 e 2019 é aproximadamente 16.23ºC, polo que podemos considerar como referencia 16ºC.

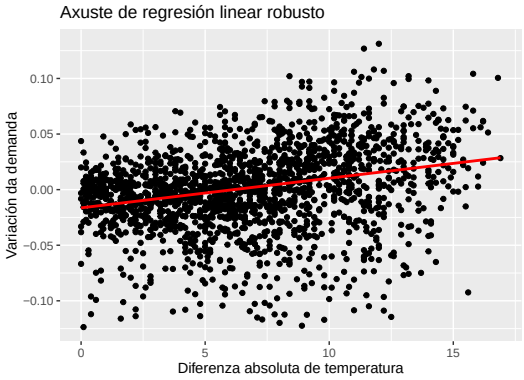
Relación entre a variación da demanda eléctrica e a diferenza absoluta de temperatura. Na Figura 6.21a preséntase o diagrama de dispersión da variación da demanda e a diferenza en valor absoluto entre a temperatura media diaria e a temperatura de referencia. Parece existir certa relación, aínda que pequena. Podemos pensar que a dita relación vén dada por unha recta, polo que axustamos un modelo de regresión linear. Posto que no diagrama de dispersión podemos notar a presenza de certas observacións atípicas, axustamos un modelo de regresión linear robusto mediante a función *rlm* (emprega un M-estimador) da libraría MASS (Venables e Ripley 2002), cuxa representación se mostra na Figura 6.21b e cuxo coeficiente de determinación é 0.066.

É posible que á vista do diagrama de dispersión visualicemos non unha recta, senón máis ben unha curva suave, polo que podemos probar, por exemplo, en axustar un modelo de regresión polinómico local robusto mediante a función *loess* do paquete base de R (R Core Team 2020), cuxa representación se mostra na Figura 6.21c. O coeficiente de determinación deste último axuste é 0.07, polo que este sería preferible ao anterior, ademais de que este é un modelo non paramétrico, polo que non se asumen hipóteses tan restritivas como no primeiro axuste, que poden non cumprirse.

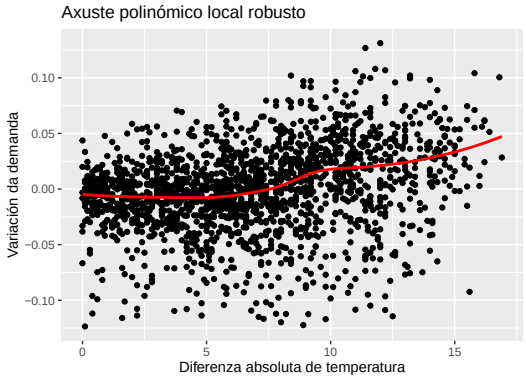
Podemos concluír que a relación entre a variación da demanda eléctrica e a diferenza absoluta de temperatura é moi pobre e así o mostran os coeficientes de determinación dos axustes considerados.



(a) Diagrama de dispersión.



(b) Axuste dun modelo de regresión linear robusto.



(c) Axuste dun modelo de regresión polinómico local robusto.

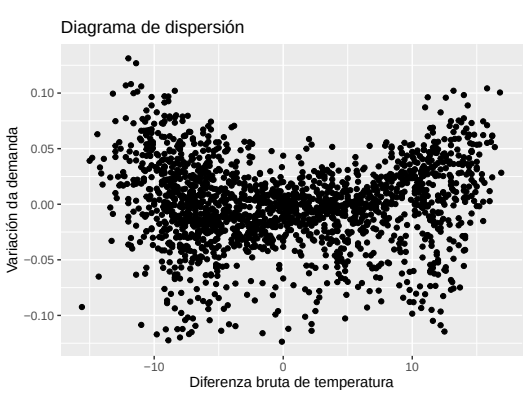
Figura 6.21: Variación da demanda eléctrica e diferenza absoluta de temperatura.

Polo tanto, seguiremos explorando a relación entre a variación da demanda e a temperatura por outros medios, por exemplo, a diferenza bruta de temperatura.

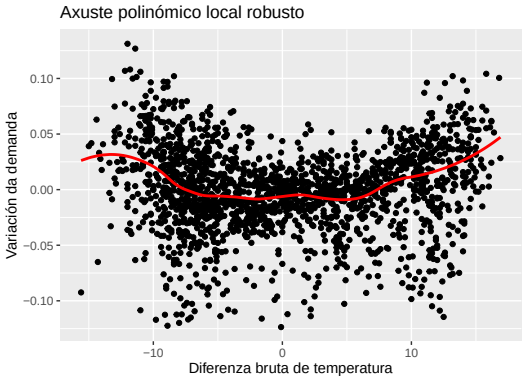
Relación entre a variación da demanda eléctrica e a diferenza bruta de temperatura. Estudemos agora a relación entre a variación da demanda e a diferenza bruta entre a temperatura media diaria e a temperatura de referencia. Na Figura 6.22a móstrase o seu diagrama de dispersión. Vemos que pode existir unha leve relación en forma de bañeira. Para medir a dita relación, axustamos un modelo de regresión polinómico local robusto a estes datos, cuxa representación se mostra na Figura 6.22b, e o seu coeficiente de determinación é 0.086.

Neste caso, podemos estudar por separado aquelas diferenzas positivas e negativas. Así, comezamos analizando a relación entre a variación da demanda e a diferenza bruta positiva (estritamente maior que cero), cuxo diagrama de dispersión se mostra na Figura 6.23a. Podemos notar certa curvatura, polo que axustamos un modelo de regresión polinómico local robusto aos datos, obtendo o axuste da representación da Figura 6.23b e un coeficiente de determinación de aproximadamente 0.085.

Consideramos agora as diferenzas brutas entre a temperatura media e a de referencia negativas ou nulas, obtendo o diagrama de dispersión da Figura 6.24a. De novo, axustamos un modelo de regresión polinómico local robusto, cuxa representación se mostra na Figura 6.24b e cuxo coeficiente

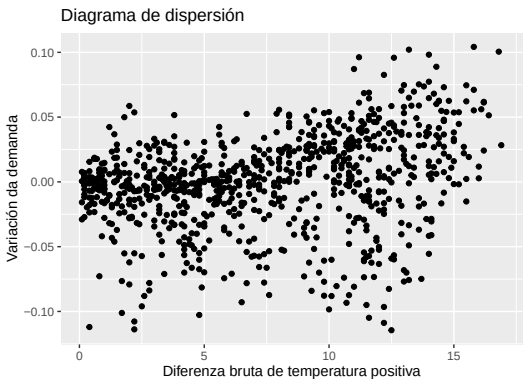


(a) Diagrama de dispersión.

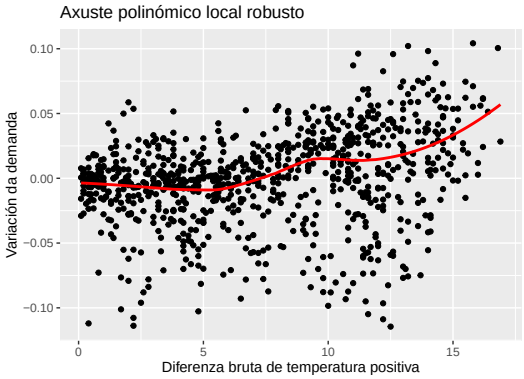


(b) Axuste dun modelo de regresión polinómico local robusto.

Figura 6.22: Variación da demanda eléctrica e diferenza bruta de temperatura.



(a) Diagrama de dispersión.



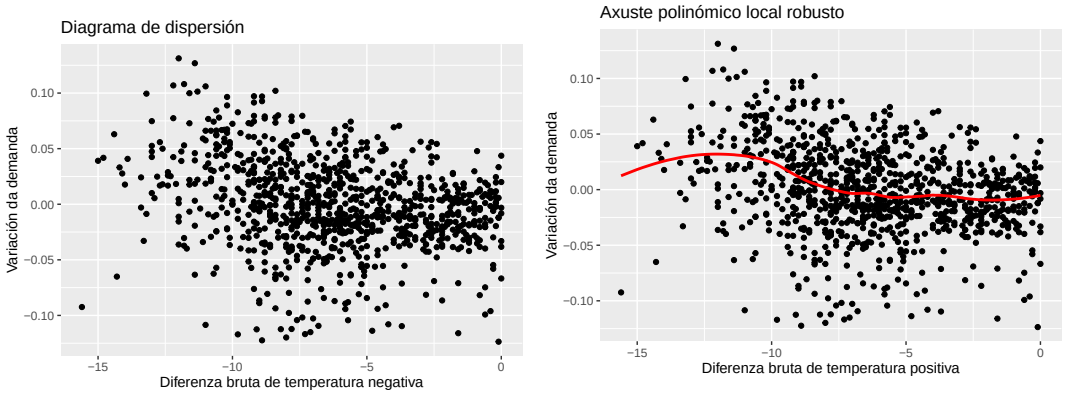
(b) Axuste dun modelo de regresión polinómico local robusto.

Figura 6.23: Variación da demanda eléctrica e diferenza bruta positiva de temperatura.

de determinación é 0.09.

A conclusión nos tres casos é a mesma, a relación existente entre a variación da demanda e a diferenza bruta, positiva e negativa de temperatura é moi débil. En consecuencia, os modelos que axustamos teñen uns coeficientes de determinación case nulos. Polo tanto, consideramos unha última vía para tratar de limpar o efecto da temperatura sobre a demanda eléctrica mediante modelos en variacións. Podemos tratar de modelar as alas da relación entre a variación da demanda eléctrica e a diferenza de temperatura.

Relación entre a variación da demanda eléctrica e a diferenza de temperatura fraccionada. Continuando co estudo da variación da demanda eléctrica e a diferenza bruta de temperatura, cuxa relación semella ter forma de bañeira, tal e como vemos na Figura 6.22a, podemos tratar de modelar só as alas. Así, tomamos un primeiro intervalo onde a diferenza bruta de temperatura é menor ou igual que -5 e o segundo onde a dita diferenza é maior ou igual que 5, que se corresponden coas ditas alas. En ambos casos axustamos un modelo de regresión linear robusto, cuxas representacións se poden ver na Figura 6.25. O coeficiente de determinación do axuste no primeiro intervalo é 0.077 e no segundo é 0.066. Polo tanto, estes axustes tampouco melloran os resultados dos anteriores que vimos de ver,

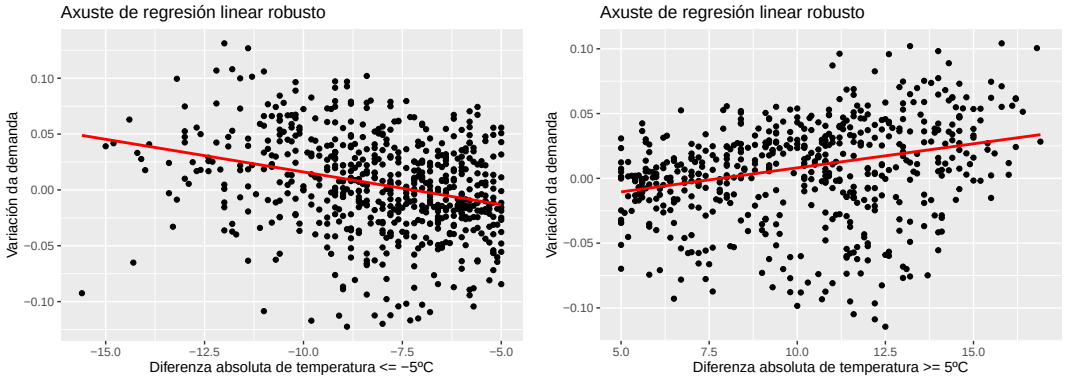


(a) Diagrama de dispersión.

(b) Axuste dun modelo de regresión polinómico local robusto.

Figura 6.24: Variación da demanda eléctrica e diferenza bruta negativa de temperatura.

pois, en realidade, a relación entre as variables é moi leve.



(a) Primeiro intervalo.

(b) Segundo intervalo.

Figura 6.25: Axuste de regresión linear robusto.

Conclusión

Vimos de estudar a relación entre a variación da demanda eléctrica e a temperatura bruta e tamén coa diferenza bruta e absoluta de temperatura. En todos os casos puidemos ver, tanto á vista do diagrama de dispersión dos datos como do coeficiente de determinación dos axustes considerados, que a relación é moi débil. Por exemplo, na Figura 6.22a, onde se presenta o diagrama de dispersión da variación da demanda e a diferenza bruta de temperatura, podemos ver unha relación en forma de bañeira, pero o modelo axustado só explica un 0.086 % da variabilidade da variación da demanda. Certamente, a variación da demanda que estamos considerando é a diferenza estacional de orde 7 do logaritmo da demanda suavizada e corrixida de observacións atípicas. Polo tanto, o efecto da temperatura, que víamos na demanda eléctrica bruta e suavizada, vese atenuado. Ademais, estamos comparando un día co respectivo día da semana anterior, polo que o efecto da temperatura tampouco semella que vaia ser relevante. Así, parece xustificable que a relación sexa tan pobre.

Ben é certo que o noso obxectivo non é explicar ou predicir a variación da demanda de electricidade

segundo a temperatura, senón limpar o seu efecto. A idea consistiría en tomar o modelo con maior coeficiente de determinación dos que vimos de axustar e corrixir a variación da demanda deste pequeno efecto da temperatura e, logo, desfacer os cambios que lle fixemos á demanda suavizada para convertela na variación da demanda (primeiro aplicamos o logaritmo e logo unha diferenza estacional de orde 7), obtendo, así, unha demanda eléctrica suavizada limpa do efecto da temperatura. Pero os resultados deste axuste non entraban dentro do esperado.

En conclusión, ao longo desta sección vimos de analizar a relación, primeiro, entre a variación da demanda eléctrica e da temperatura, cuxo resultado era a inexistencia de relación, e logo entre a variación da demanda e a temperatura e a diferenza bruta e absoluta de temperatura, resultando nunha débil relación como consecuencia da gran variabilidade que presenta a variación da demanda eléctrica e que, ao aplicar o logaritmo e a diferenza estacional de orde 7, o efecto da temperatura sobre a demanda eléctrica suavizada vese moi relaxado. Posto que o coeficiente de determinación dos modelos é moi baixo e os consecuentes resultados non resultaron ser axeitados, parece xustificable que este procedemento quede descartado.

6.3.3. Terceira metodoloxía (proposta final)

Lembremos que o obxectivo do traballo é obter un indicador de alta frecuencia a partir da demanda eléctrica, pero para isto precisamos que esta variable estea limpa dos efectos semanal e mensual. Neste capítulo xa presentamos dúas vías diferentes para a corrección deste último efecto. Na Sección 6.3.1 axustamos un modelo de regresión da demanda de electricidade sobre a temperatura suavizadas, pero rematamos cuns residuos (demanda eléctrica sen o efecto da temperatura) con magnitudes negativas, dificilmente xustificables. Mentres que na Sección 6.3.2 presentamos a modelaxe da relación entre a variación da demanda eléctrica suavizada e a temperatura, pero o resultado non entraba dentro do esperado. Polo tanto, segue aberto o problema da corrección do efecto mensual que sofre a demanda eléctrica. Nesta sección encamiñamos a cuestión mediante a variación porcentual da mediana da demanda de electricidade suavizada mensual respecto da mediana desta variable ao longo dos anos comprendidos entre 2015 e 2019, considerando esta medida, no lugar da media, por ser robusta, pois poderían existir observacións atípicas nos datos que non foron captadas pola función *tsoutliers* (Hyndman et al 2020).

Variación mensual da mediana da demanda eléctrica

Nesta sección presentamos unha alternativa para a corrección mensual da demanda de electricidade por medio de variacións da mediana por meses. Como ata o de agora, partimos dos datos de demanda eléctrica entre 2015 e 2019 corrixidos de observacións atípicas, como se presenta na Sección 6.1, e suavizados para eliminar o efecto semanal, como se mostra na Sección 6.2. Tomamos como referencia a mediana da demanda eléctrica suavizada entre os anos considerados, 726.2 XWh, e calculamos a mediana da demanda suavizada por meses, cuxos valores poden verse na Táboa 6.1.

O seguinte paso é calcular a porcentaxe de variación da mediana mensual en tanto por un respecto da referencia. É dicir, sexa $i \in \{1, \dots, 12\}$ o indicador correspondente ao mes i, entón a variación en tanto por un neste mes vén dada por:

$$\text{Variación no mes } i = \frac{\text{Mediana no mes } i - \text{Referencia}}{\text{Referencia}}.$$

Por último, corriximos a demanda eléctrica suavizada do efecto mensual mediante a seguinte relación:

$$\text{Demanda corrixida no día } j = \text{Demanda suavizada no día } j \cdot (1 - \text{Variación no mes } i),$$

onde o mes i é o correspondente ao día j. Así, se a mediana da demanda eléctrica mensual é superior á referencia (isto é, se a variación é positiva), a demanda resultante baixaría e no caso de que sexa inferior (a variación é negativa), subiría.

Mes	Mediana
Xaneiro	777.8
Febreiro	769.54
Marzo	729.09
Abril	690.65
Maio	684.67
Xuño	710.66
Xullo	773.51
Agosto	745.4
Setembro	722.58
Outubro	688.51
Novembro	725.46
Decembro	747.3

Táboa 6.1: Mediana da demanda eléctrica suavizada por meses (en XWh).

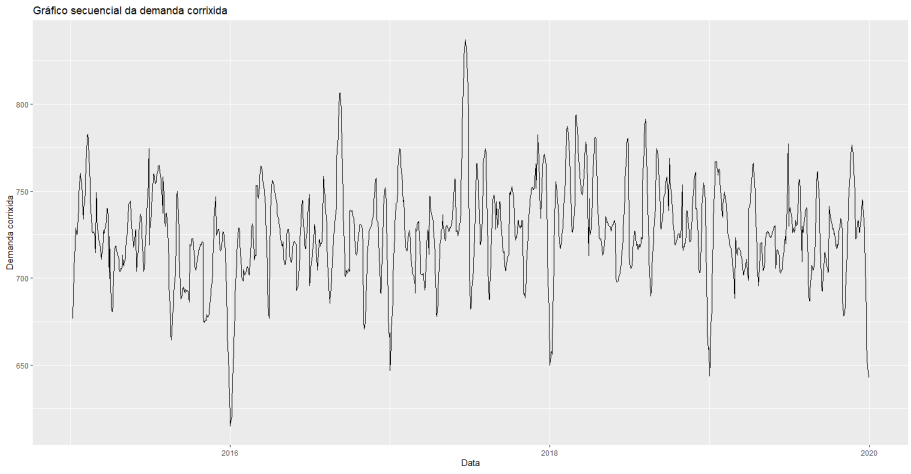


Figura 6.26: Gráfico secuencial da demanda eléctrica suavizada e corrixida do efecto mensual entre 2015 e 2019.

Na Figura 6.26 preséntase o gráfico secuencial da demanda eléctrica suavizada e así corrixida do efecto mensual, cuxo resultado semella ser unha serie estacionaria, sen a tendencia que sufría a demanda

eléctrica suavizada.

Desta forma parece que eliminamos o efecto mensual que producían os cambios de temperatura, pero o que tamén podemos notar é que a finais de cada ano se producen unhas baixas pronunciadas. Para entender este comportamento facemos unha pequena recompilación: partíamos dos datos de demanda de electricidade entre o 2015 e 2019 e mediante a función *tsoutliers* detectamos e substituímos os atípicos polas súas recomendacións. Como exemplo para a análise desta función, tomaremos a serie de demanda eléctrica orixinal en 2017, cuxo gráfico secuencial se mostra na Figura 5.1. Ao final da serie podemos ver que hai unha baixada considerable e o nivel da serie mantense aí, non volve a recuperar o nivel anterior. Así, a función considerada como atípico aquela observación onde o nivel da serie cambia, pero os seguintes datos tómaos como un simple cambio de nivel e non como atípicos. Polo tanto, debido á maior variabilidade que sofre decembro respecto dos restantes meses, parece razoable aplicarlle unha corrección adicional. En consecuencia, partindo das observacións de demanda de electricidade orixinal corrixidas dos atípicos que detecta a función *tsoutliers*, corrixiremos os datos correspondentes á semana de Nadal, do 24 de decembro ata o 1 de xaneiro, mediante a substitución dos mesmos polos valores da demanda na semana anterior, por sinxeleza e para non perder a dinámica semanal. En particular, dado que os días 2 e 3 de xaneiro de 2016 resultan ser fin de semana, o nivel da serie neses instantes é o da semana de Nadal, polo que tamén corrixiremos estes dous datos de forma análoga. Unha vez feita esta corrección, seguiremos o procedemento habitual: aplicación de medias móbiles con xanela $h = 7$, para eliminar o efecto semanal, e a corrección mediante a variación en tanto por un da mediana por meses da variable, para eliminar o efecto mensual, no que está centrada esta sección. Na Figura 6.27 móstranse os gráficos secuenciais da demanda eléctrica así corrixida de atípicos (arriba) e unha vez que se suavizou por medias móbiles (abaixo). É fácil ver que as baixadas a redores de fin de ano non son tan pronunciadas como no anterior caso, tal e como buscábamos. Tamén podemos ver este arranxo no gráfico secuencial da demanda eléctrica suavizada unha vez que se corrixiu o efecto mensual mediante a metodoloxía exposta nesta sección (Figura 6.28). Esta demanda corrixida parece máis acertada que aquela na que non corrixíamos o efecto do Nadal.

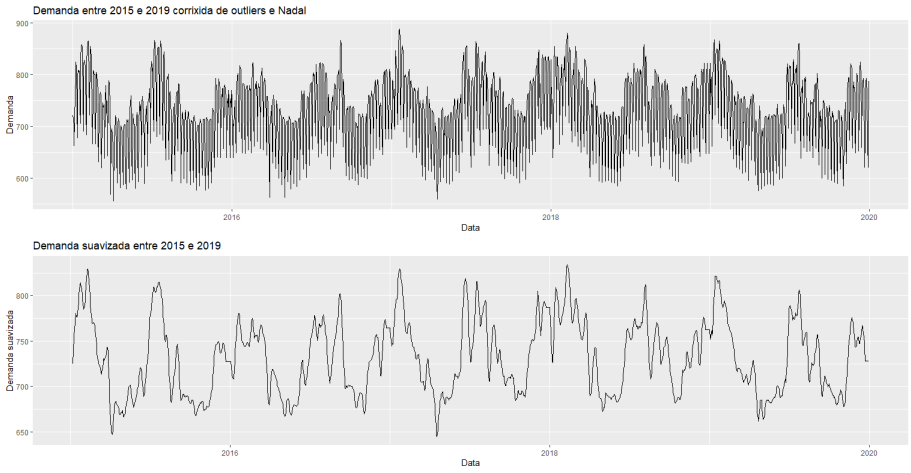


Figura 6.27: Gráfico secuencial da demanda eléctrica corrixida de atípicos e dos dados na semana de Nadal (arriba) e suavizada (abaixo).

Á vista da estimación da función de autocorrelacións simples da serie así corrixida, que se mostra na Figura 6.29, podemos ver que non hai indicios de tendencia nin compoñente estacional. Ademais, a serie é homocedástica, polo que sería estacionaria. Isto tamén pode concluírse ao aplicar o test de

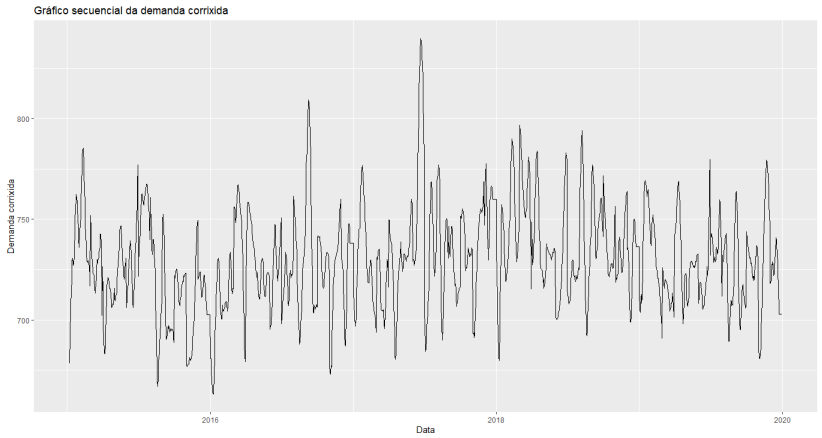


Figura 6.28: Gráfico secuencial da demanda eléctrica corrixida de atípicos (e do Nadal) e dos efectos semanal e mensual.

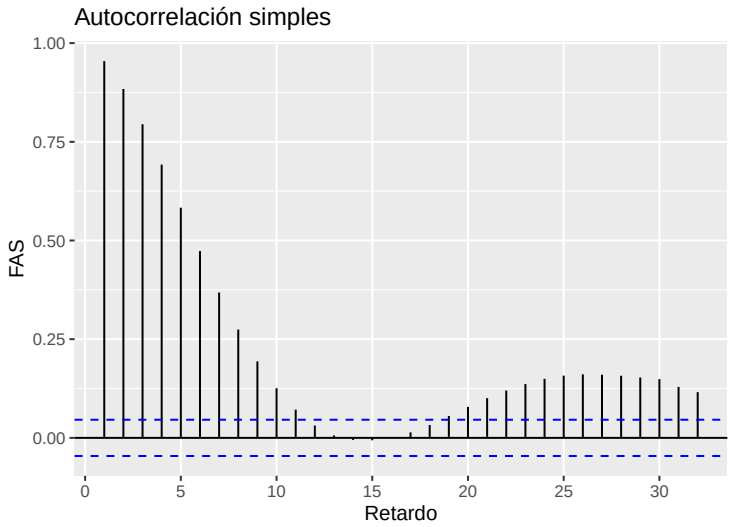


Figura 6.29: Autocorrelacións simples mostrais da demanda eléctrica suavizada e corrixida do efecto mensual entre 2015 e 2019.

Dickey-Fuller aumentado, pois o p-valor asociado resultou ser inferior a 0.01, polo que a un nivel de significación $\alpha = 0.05$, rexeitamos a hipótese nula de non estacionariedade.

A serie de demanda eléctrica suavizada así corrixida non presenta a tendencia da demanda suavizada, cuxo gráfico secuencial se mostra na Figura 6.27 (abaixo), constituída por unha maior demanda en invernos e veráns. Polo tanto, parece que mediante esta corrección eliminamos o efecto mensual da demanda eléctrica suavizada.

Para corroborar este feito, presentamos o diagrama de caixas por meses da demanda suavizada (arriba) e da demanda suavizada e corrixida mediante esta metodoloxía (abaixo) na Figura 6.30. Xa estudamos que a demanda eléctrica suavizada vese afectada pola temperatura, así o reflexa o diagrama de caixas por meses, onde a mediana é superior nos meses de inverno e verán. Doutra banda, o diagrama

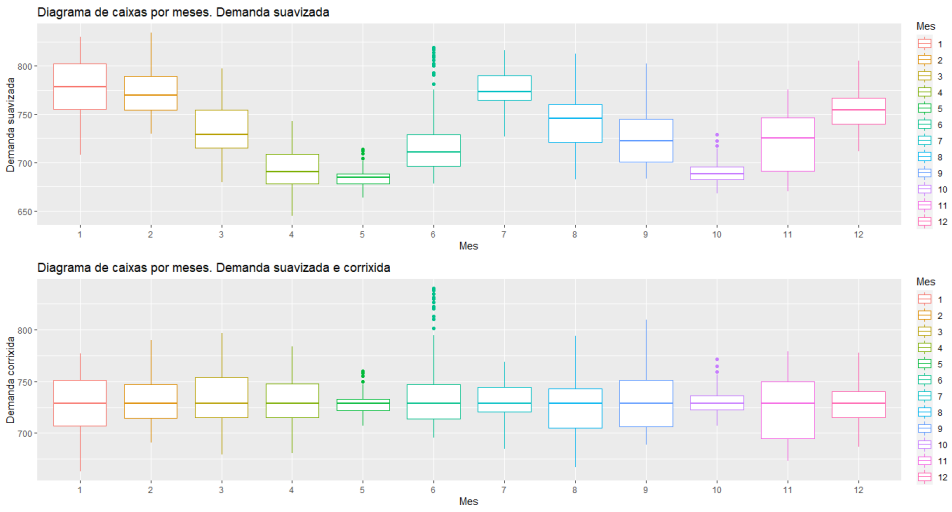


Figura 6.30: Diagrama de caixas por meses da demanda eléctrica suavizada (arriba) e da demanda eléctrica suavizada e corrixida do efecto mensual (abaixo).

de caixas da demanda suavizada e corrixida mostra que a mediana en todos os meses é a mesma, 728.79 XWh (a mediana da demanda eléctrica corrixida de atípicos, incluíndo a semana de Nadal, e suavizada, que empregamos como referencia para a corrección mensual), pois así se fixo a corrección. Á vista do diagrama de caixas e do gráfico secuencial da demanda eléctrica corrixida, podemos concluír que eliminamos o efecto mensual.

Conclusión

Nesta sección consideramos un novo punto de vista para abordar a corrección do efecto mensual, que parece causar a temperatura, da demanda de electricidade e que finalmente foi a solución escollida. Esta corrección fíxose mediante a variación en tanto por un da mediana mensual respecto da referencia, a mediana da demanda eléctrica suavizada entre 2015 e 2019. E vimos, mediante ferramentas gráficas, que esta vía corrixe o efecto mensual e, ademais, a serie resultante atópase na mesma escala que a demanda eléctrica. Polo tanto, xa podemos proceder a construír un índice a partir desta variable.

Capítulo 7

Construción do índice

Co obxectivo de construír un índice de alta frecuencia que nos permita rastrexar a actividade económica do país, sen ter que agardar, por exemplo, tres meses para coñecer os datos do PIB, consideramos a demanda eléctrica diaria no territorio nacional. Esta variable conta con efectos semanal e mensual, onde o primeiro se pode atallar mediante a aplicación de medias móbiles con xanela $h = 7$, tal e como se mostra na Sección 6.2. Na Subsección 6.3.3 presentamos unha vía para corrixir o efecto mensual da demanda eléctrica suavizada mediante a variación porcentual da mediana mensual da variable respecto da mediana da dita variable entre 2015 e 2019, resultando nun camiño axeitado para o caso que nos compete. Estas correccións fixémolas deixando de lado o ano 2020, por ser este un ano atípico, pero resulta de gran interese consideralo para a análise do comportamento do índice, que construiremos ao longo deste capítulo.

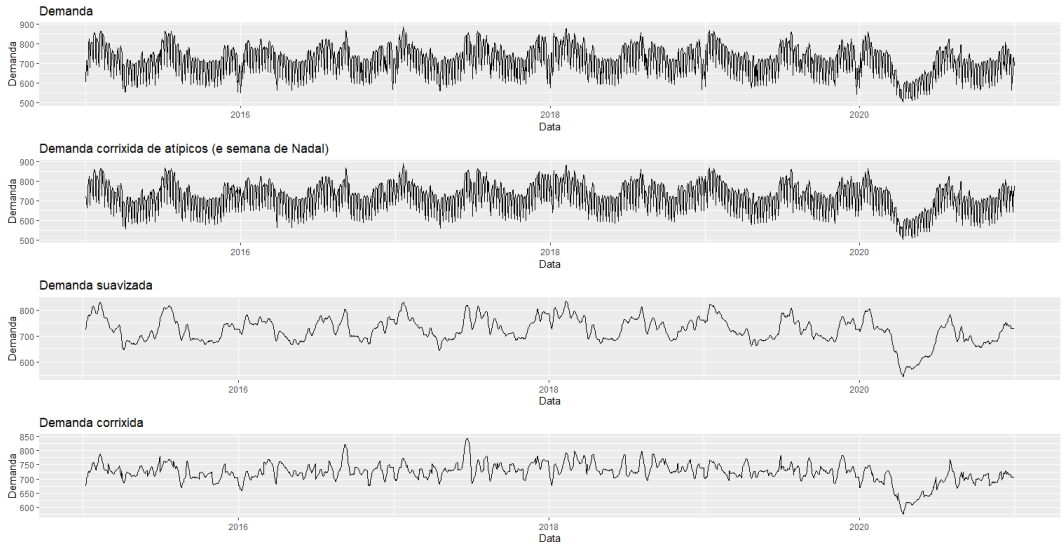


Figura 7.1: Gráfico secuencial da demanda eléctrica bruta entre 2015 e 2020, corrixida de atípicos e da semana de Nadal, suavizada mediante medias móbiles e corrixida do efecto mensual (de arriba a abaixo).

Polo tanto, partimos da demanda eléctrica bruta entre 2015 e 2020, corriximos os atípicos mediante

a función *tsoutliers*[1] e as semanas de Nadal. Logo, eliminamos o efecto semanal mediante medias móbiles con xanela $h = 7$. Por último, para limpar o efecto mensual, corriximos a demanda eléctrica suavizada mediante a variación en tanto por un da mediana mensual desta variable respecto da mediana da mesma entre 2015 e 2020[2]. Na Figura 7.1 móstrase este proceso. Podemos ver que, finalmente, logo de facer as ditas correccións, obtemos unha demanda eléctrica limpa, sen efectos semanais nin mensuais. Partindo desta variable, construiremos un índice para analizar a dinámica da economía no país. Neste capítulo presentamos un indicador de frecuencias diaria (Sección 7.1), semanal (Sección 7.2) e, por último, mensual (Sección 7.3). Tamén compararemos a súa dinámica coa dalgúns agregados macroeconómicos e na Sección 7.4 discutimos as fortalezas do noso indicador.

Para a construción dos índices empregamos a media, que é a referencia tradicional. Tamén os calculamos en base á mediana, pero os cambios non difiren moito, polo que presentamos só os resultados dos indicadores feitos considerando a medida usual, a media. Ademais, tomamos como período base o 2015 dado que é o ano de referencia da gran maioría das variables macroeconómicas.

7.1. Indicador diario

Comezamos presentando o índice en base 100 de frecuencia diaria, tomando como referencia a media da demanda de electricidade limpa en 2015. Sexa i o indicador do día, este índice vén dado por:

$$\text{Índice no día } i = 100 \cdot \frac{\text{Consumo eléctrico no día } i}{\text{Referencia}}.$$

Na Figura 7.2 preséntase o índice diario así calculado. Podemos ver que captura a gran baixada que se produciu a mediados de marzo de 2020, como resultado do inicio do confinamento en España, e a súa lenta recuperación, sen acadar aínda o nivel de demanda eléctrica nos anos anteriores.

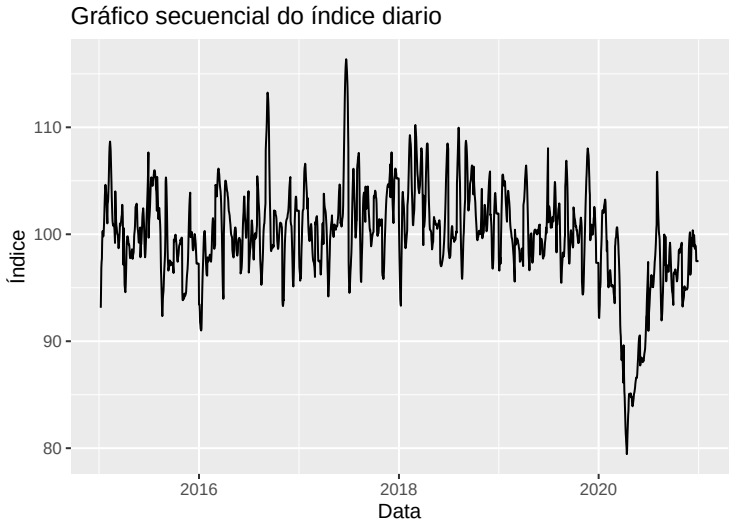


Figura 7.2: Gráfico secuencial do índice diario en base 100 baseado na demanda eléctrica limpa.

Neste punto, resulta de gran interese comparar o comportamento do índice diario que vimos de calcular con algún indicador macroeconómico tradicional. Por exemplo, podemos considerar o IPI e o PIB.

[1] A función *tsoutliers* non considera como atípicos aqueles debidos ao confinamento.

[2] Dado que consideramos a mediana, non se producen diferenzas significativas na corrección mensual considerando esta medida entre 2015 e 2019 ou ata o 2020.

7.1.1. Índice diario e o IPI

Primeiro, comparamos o noso indicador diario co Índice de Produción Industrial (IPI), que presentamos no Capítulo 4. Na Figura 7.3 preséntase o gráfico secuencial do IPI corrixido de calendario e estacionalidade, entre 2015 e 2020, con base 2015. A súa dinámica consiste nun leve crecemento dende 2015 ata o 2018, mantense constante ata o 2020, onde ao redor de marzo pega un considerable descenso, para logo tratar de recuperar o anterior nivel.

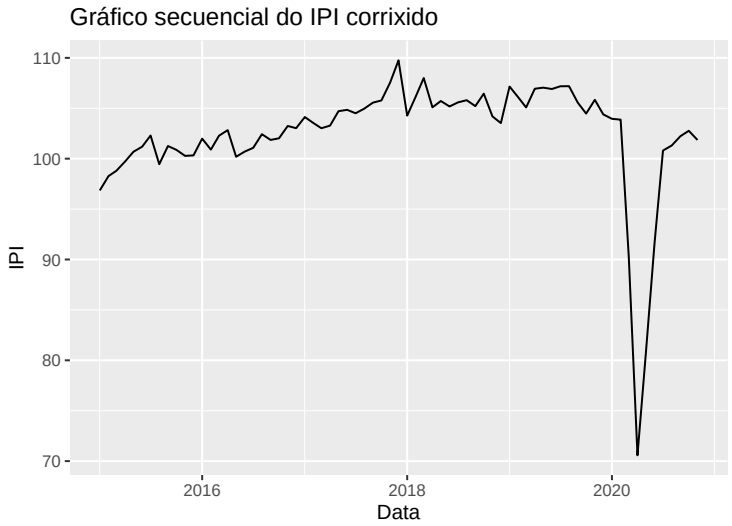


Figura 7.3: Gráfico secuencial do Índice de Produción Industrial corrixido de calendario e estacionalidades con base 2015 entre 2015 e 2020.

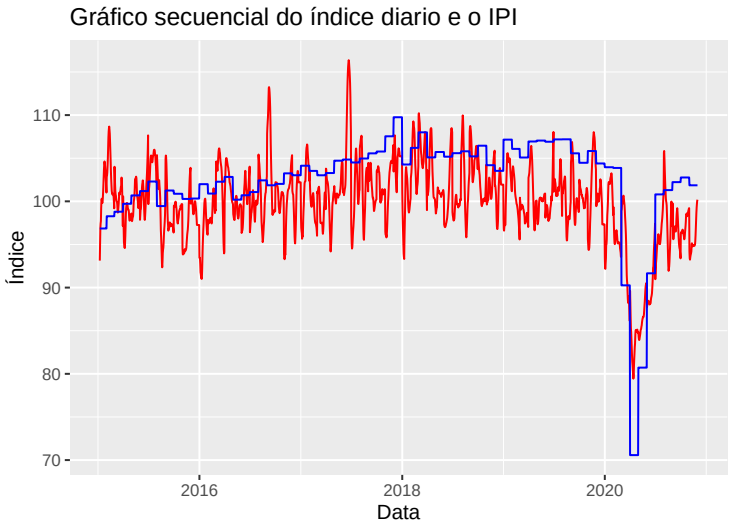


Figura 7.4: Gráficos secuenciais do índice diario (en vermello) e do IPI corrixido (en azul).

Na Figura 7.4 móstranse os gráficos secuenciais do índice diario en vermello e do IPI corrixido de calendario e estacionalidade en azul. Podemos ver que existe certa relación entre ambos índices. En especial, semella que o noso índice diario capta considerablemente ben a dinámica do ano 2020. Debemos notar que o IPI baséase na actividade industrial, mentres que o noso indicador construíuse mediante a demanda eléctrica total, que conta coa industria, o sector servizos e os fogares, polo tanto non é esperable que a relación sexa perfecta. Ademais, volvendo ao 2020, podemos ver que o descenso no IPI é maior que no noso indicador, dado que no confinamento paráronse todas as actividades non esenciais, incluíndo o sector industrial, e a demanda eléctrica referente aos fogares aumentou, pois tivemos que permanecer na casa. Para unha mellor comparación, xa que o índice diario presenta moita variabilidade, máis da que nos gustaría, imos suavizar o dito índice.

7.1.2. Suavizado do indicador

Para a suavización do índice diario aplicamos diferentes métodos, como medias móbiles, o núcleo de Nadaraya-Watson, *LOWESS*, *splines* de suavización, entre outros, pero o que mellor recollía o comportamento do indicador foi o axuste linear local considerando a xanela dada polo método *plug-in* (Ruppert et al 1995), cuxo gráfico secuencial se mostra na Figura 7.5.

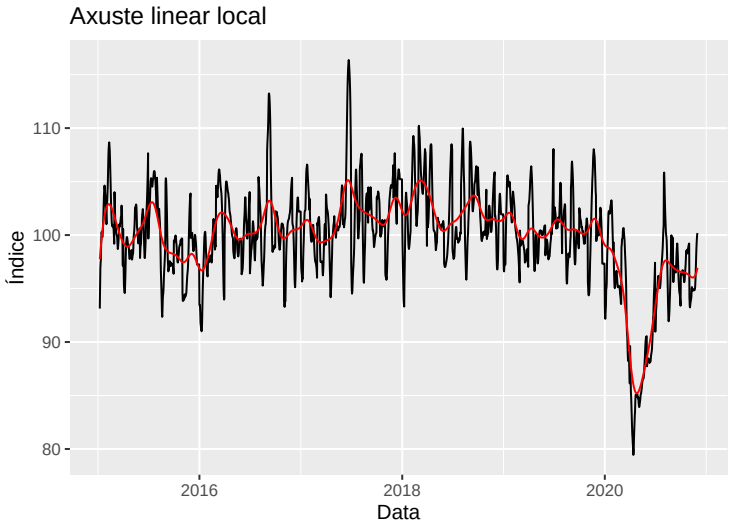


Figura 7.5: Gráficos secuenciais do índice diario bruto (en negro) e suavizado (en vermello).

Considerando o índice diario suavizado, estudamos agora o seu comportamento respecto do IPI corrixido. Na Figura 7.6 móstrase o gráfico secuencial destes dous indicadores. O noso indicador non se axusta tan ben ao IPI ata o 2020, onde recolle a caída provocada polas medidas gobernamentais para reducir os contaxios e a súa lenta recuperación. Debemos ter en conta que o noso índice baséase nunha soa variable, polo que non podemos esperar que a súa dinámica sexa exactamente a mesma que a do IPI, un agregado macroeconómico. É posible que existan outras variables que expliquen mellor estes anos, pero o noso é suficiente para adiantar certas situacións tan abruptas como a pandemia de COVID-19. Polo tanto, a contribución principal do noso indicador é que permite detectar grandes cambios nas actividades económicas día a día, mentres que o IPI tardaría un mes, e os correspondentes días de atraso na súa publicación, para mostrar este suceso.

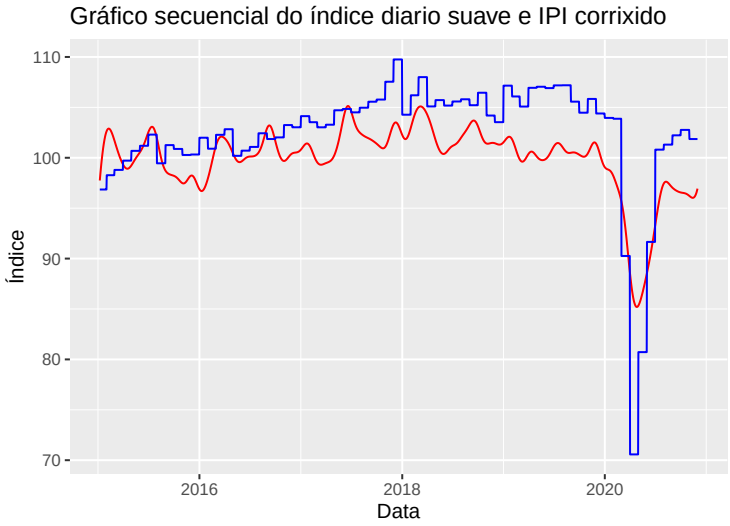
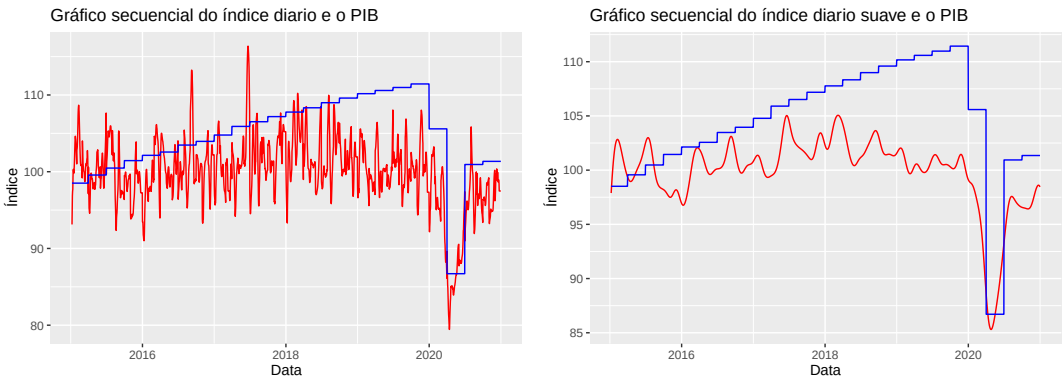


Figura 7.6: Gráficos secuenciais do índice diario suavizado (en vermello) e do IPI corrixido (en azul).

7.1.3. Índice diario e o PIB

Tamén é interesante comparar o índice diario que vimos de construír co indicador máis empregado para o seguimento do estado da economía no país, o Produto Interior Bruto (PIB), que se presenta no Capítulo 4.



(a) Índice diario e PIB.

(b) Índice diario suave e PIB.

Figura 7.7: Gráficos secuenciais do índice diario (en vermello) e do PIB (en azul).

Nas Figuras 7.7a e 7.7b móstranse os gráficos secuencias do PIB, en azul, e do indicador diario bruto e suavizado, en vermello, de esquerda a dereita. Parello ao que víamos co IPI, o noso índice non segue a mesma dinámica de crecemento nos primeiros anos, pero unha vez que chega o 2020, podemos ver que captura o descenso que se produciu en marzo e como, pouco a pouco, comeza a subir. Polo tanto, este índice permítenos adiantar esta grande baixada nun corto período de tempo, sen agardar pola publicación do PIB, que tardaría os 3 meses correspondentes á súa frecuencia e case outros dous para a súa divulgación. Ademais, neste caso, cando o PIB anuncia a caída causada polo confinamento, o noso índice xa nos anticipa a recuperación das actividades económicas.

7.2. Indicador semanal

Nesta sección presentamos o índice en base 100 de frecuencia semanal, tomando como referencia a media dos datos semanais, calculados como a media dos sete días de cada semana, da demanda de electricidade limpa do ano 2015. Isto é, se denotamos por i ao indicador semanal, o índice vén dado por:

$$\text{Índice na semana } i = 100 \cdot \frac{\text{Media na semana } i}{\text{Referencia}}.$$

Na Figura 7.8 móstrase o gráfico secuencial do índice semanal que vimos de calcular, no que destacamos a caída da demanda eléctrica a partir de mediados de marzo de 2020.

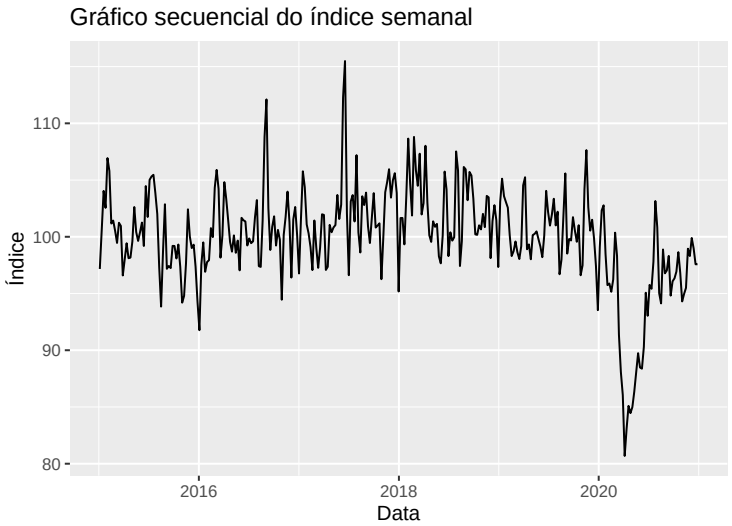


Figura 7.8: Gráfico secuencial do índice semanal en base 100 baseado na demanda eléctrica limpa.

7.3. Indicador mensual

Por último, imos construír un índice de frecuencia mensual en base 100, tomando como referencia a media das medias da demanda de electricidade limpa mensual no ano 2015. Se denotamos por i ao indicador dos meses, o índice vén dado por:

$$\text{Índice no mes } i = 100 \cdot \frac{\text{Media no mes } i}{\text{Referencia}}.$$

Na Figura 7.9 móstrase o gráfico secuencial do índice mensual. Ao igual que os indicadores de maior frecuencia, captura a caída da demanda eléctrica a redores de marzo de 2020 e a súa posterior e lenta recuperación.

A continuación, imos analizar o comportamento deste índice mensual e os agregados económicos tradicionais, o IPI e o PIB, e o indicador IRE.

7.3.1. Índice mensual e o IPI

Unha vez construído o índice mensual, interésanos comparar a súa dinámica coa do Índice de Produción Industrial (IPI). Para isto, presentamos na Figura 7.10 os gráficos secuenciais do noso

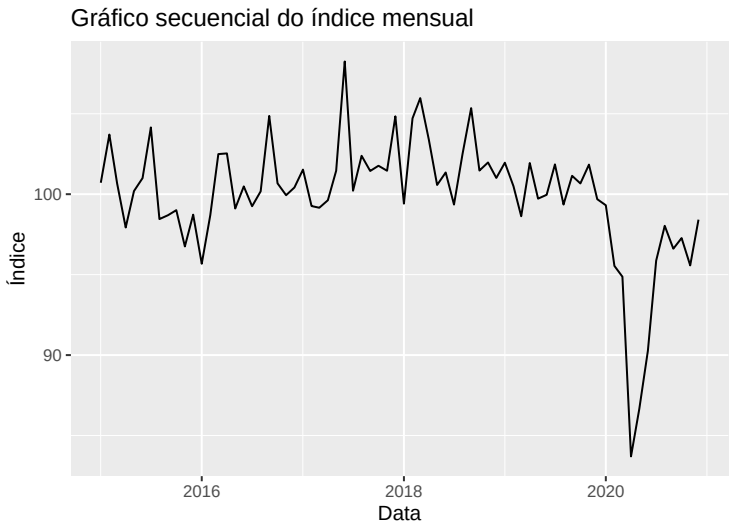


Figura 7.9: Gráfico secuencial do índice mensual en base 100 baseado na demanda eléctrica limpa.

indicador mensual, en vermello, e do IPI corrixido de estacionalidades e calendario, en azul. Parece que ambos índices gardan certa relación, especialmente na caída de 2020. Dende logo, non podemos esperar que o noso índice mensual, baseado só na demanda de electricidade, poida seguir á perfección a tendencia do IPI, pois existe incerteza e, ademais, os índices macroeconómicos vense influenciados por máis dunha variable. Amais, o noso indicador ten en conta a demanda eléctrica dos sectores industrial e servizos e dos fogares, mentres que o IPI só conta coa produción industrial. Coa creación do noso índice buscamos unha primeira aproximación sinxela ao comportamento da actividade económica no país e semella que, ante notables cambios, o noso indicador é de grande utilidade.

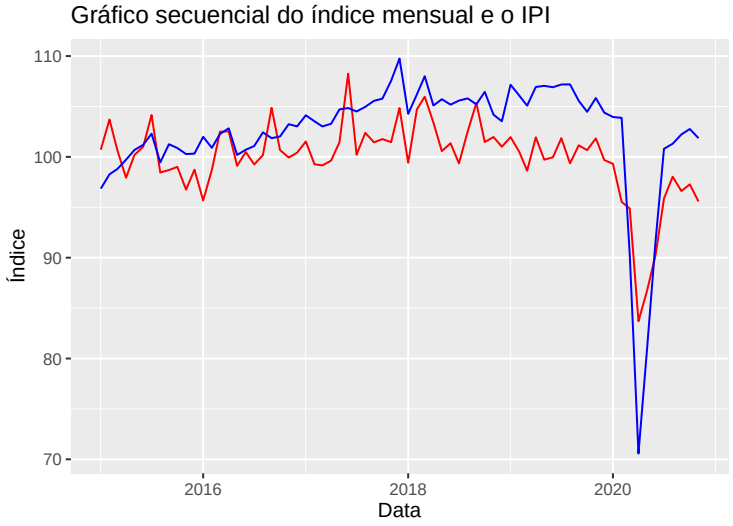


Figura 7.10: Gráficos secuenciais do índice mensual (en vermello) e do IPI corrixido (en azul).

Medidas de correlación

Na Figura 7.10 albiscamos certa relación entre o índice mensual que vimos de calcular e o IPI corrixido. Nesta subsección adicarémonos ao estudo desta relación. Por exemplo, podemos comezar debuxando o diagrama de dispersión dos datos, cuxo resultado se mostra na Figura 7.11. Nel podemos ver que non nos equivocamos ao vaticinar a existencia de correlación.

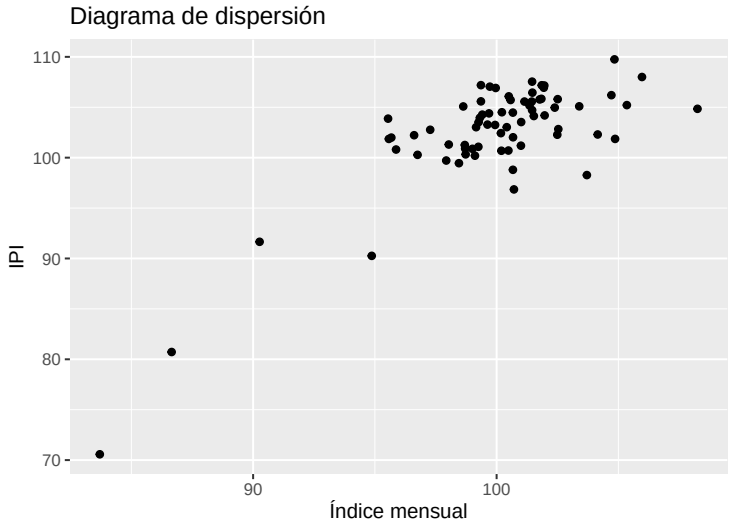


Figura 7.11: Diagrama de dispersión do índice mensual e o IPI corrixido.

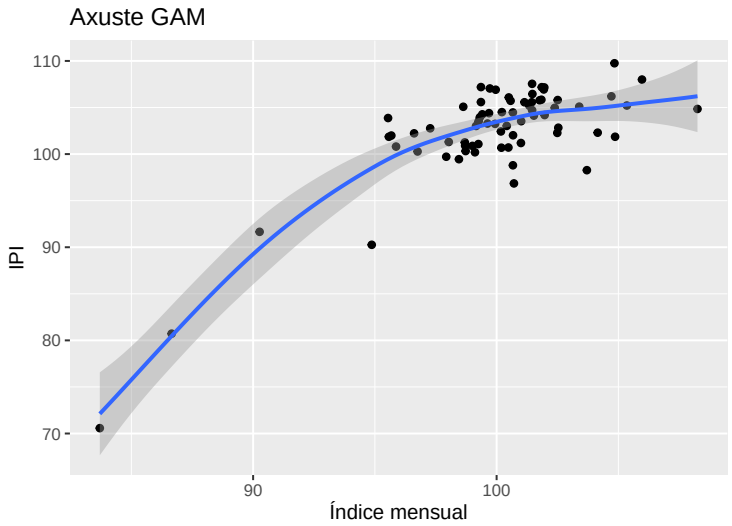


Figura 7.12: Axuste do modelo aditivo ao IPI corrixido e ao índice mensual.

A correlación mostral de Pearson entre ambos indicadores é 0.79, indicando unha relación positiva linear bastante forte entre ambos. Ademais, podemos aplicar algún contraste de correlación non pa-

ramétrico, como os tests de Spearman e Kendall, no que en ambos casos se rexeita a hipótese nula de independencia.

Tendo en conta os resultados precedentes, aparte de ser de interese, podería ser tamén factible estimar indicadores como o IPI (de frecuencia mensual) a partir do valor do indicador proposto (de frecuencia mensual). Para esta tarefa, consideramos o axuste de modelos GAM, que permiten a estimación de efectos suaves das variables explicativas. Deste xeito poderemos tamén caracterizar o efecto do noso indicador sobre o IPI, identificando, por exemplo, se é linear, sigmoide ou mesmo de tipo asintótico, entre outros tipos. Especificamente, podemos axustar un modelo aditivo xeneralizado, onde o IPI corrixido xoga o papel da variable resposta e o índice mensual o da variable explicativa. Consideramos distintas bases e números de elementos da mesma para a suavización da curva da variable independente, resultando en que o mellor axuste, en base ao criterio do coeficiente de determinación axustado, vén dado mediante a elección das denominadas bases *gaussian process* e tomando 20 elementos para esta. O axuste fíxose mediante a función *gam* do paquete *mgcv* (Wood 2011). Na Figura 7.12 móstrase a saída gráfica do axuste resultante, cuxo coeficiente de determinación axustado é 0.85 e explica un 88.3 % da variabilidade dos datos.

Polo tanto, existe unha relación considerablemente alta entre o noso índice mensual e o IPI corrixido, sendo esta non linear, de tipo asintótico.

7.3.2. Índice mensual e o PIB

Nesta subsección imos comparar o índice mensual co PIB. Na Figura 7.13 móstrase o gráfico secuencial de ambas series, onde volvemos ver que a dinámica concorda ao longo do 2020, facendo dos resultados do noso indicador moi prometedores. Aínda tendo unha frecuencia mensual, este índice permítenos predicir o comportamento da economía aproximadamente catro meses antes que o PIB, polo que a súa labor é de notable importancia.

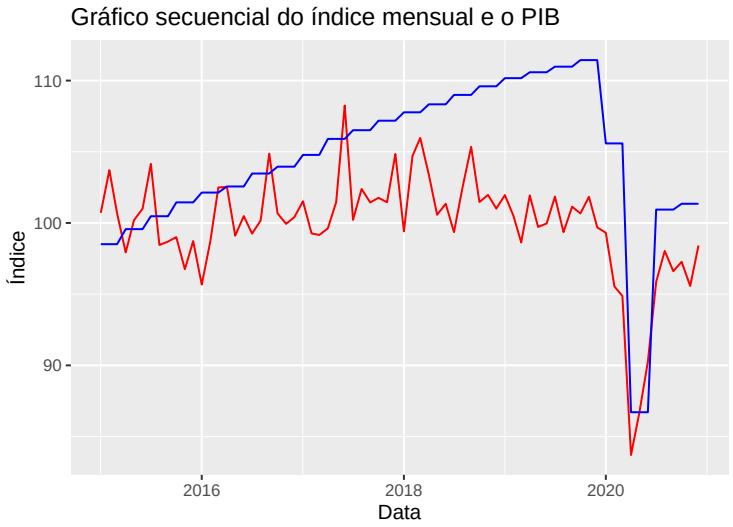


Figura 7.13: Gráficos secuencias do índice mensual (en vermello) e do PIB (en azul).

No lugar de facer esta comparación mes a trimestre, podemos pensar en mensualizar o PIB, pois de trimestralizar o noso índice mensual perderíamos información. Ademais, tamén podemos pensar en comparar non só o nivel da serie, senón os seus cambios ao longo do tempo, para o que consideraremos as variacións interanuais. Dada unha serie temporal $y_1, \dots, y_T$ que ten frecuencia f, calcúlase a variación

interanual da observación i-ésima como:

$$\frac{y_{i+f}}{y_i} - 1.$$

Estas variacións fan posible a cuantificación do crecemento (ou decrecemento) que sufriu a serie respecto a un período de referencia. No caso de multiplicarse por cen, representaría a variación interanual en tanto por cen.

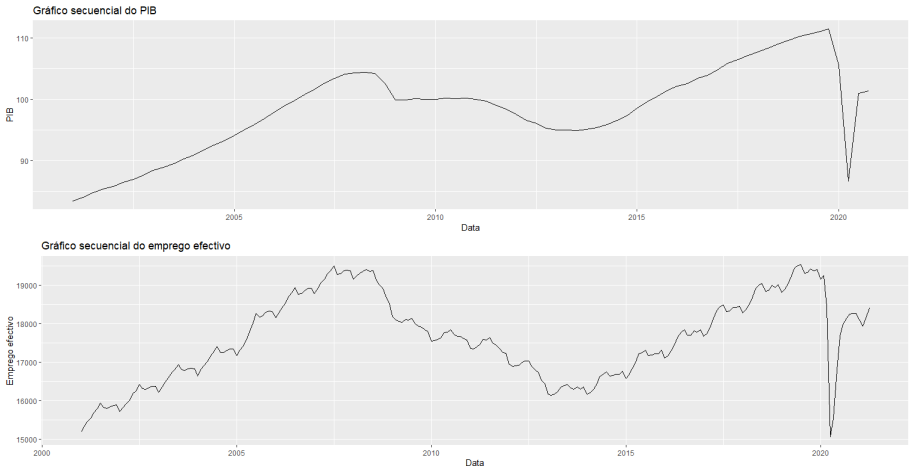


Figura 7.14: Gráficos secuenciais do PIB (arriba), de frecuencia trimestral, e do emprego efectivo (abaixo), de frecuencia mensual, dende 2001 ata 2021.

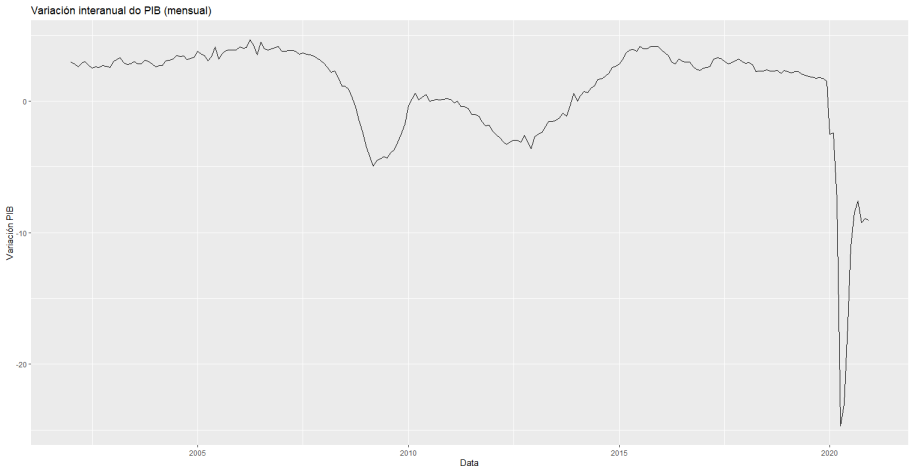


Figura 7.15: Gráfico secuencial da variación interanual do PIB mensualizada entre 2002 e 2021.

A función *td* do paquete *tempdisagg* (Sax e Steiner 2013) permite desagregar ou interpolar unha serie de baixa frecuencia a unha de maior frecuencia, onde a media desta última é consistente coa primeira. A función permite realizar esta tarefa con ou sen variables explicativas e ten implementados distintos procedementos dos que, no noso caso, empregaremos o método uniforme. O emprego efectivo, definido como o número de afiliados menos aqueles que están en ERTE (Expediente de Regulación Temporal

de Emprego), parece ter relación co Produto Interior Bruto, tal e como pode verse mediante os seus gráficos secuenciais, que se mostran na Figura 7.14. Podemos ver que a tendencia de ambas series é moi semellante. Polo tanto, usamos a dita variable para desagregar o PIB, ambas en variacións interanuais. Para unha descrición máis detallada tanto do proceso de desagregación dunha serie temporal como do paquete *tempdisagg*, pode consultarse Sax e Steiner (2013).

Así pois, desagregamos a variación interanual do PIB por medio da variación interanual do emprego efectivo, cuxo resultado se mostra a Figura 7.15.

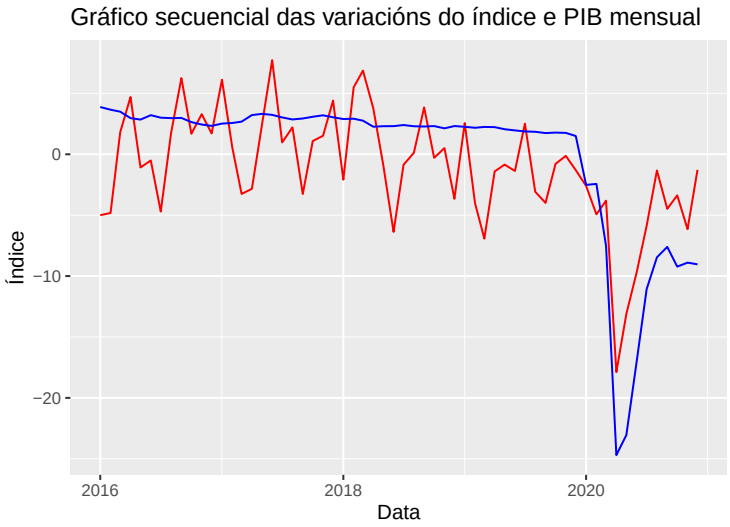


Figura 7.16: Gráficos secuenciais da variación interanual do índice mensual (en vermello) e do PIB mensualizada (en azul).

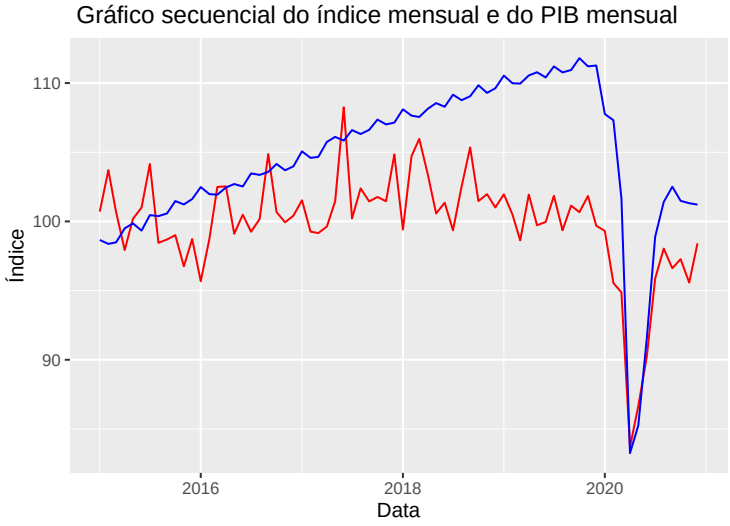


Figura 7.17: Gráficos secuenciais do índice mensual (en vermello) e do PIB mensualizado (en azul).

O seguinte paso é calcular as variacións interanuais do noso índice mensual e pintalas xunto coas do PIB. Na Figura 7.16 móstrase o gráfico destas variacións entre 2016 e 2020. Dado que os cambios no PIB dun trimestre a outro case non varían, a excepción dos que se producen en situacións tan abruptas como a pandemia provocada pola COVID-19, e que o noso índice ten bastante ruído, estas variacións están en diferentes escalas. Polo tanto, debuxaremos mellor o indicador mensual e o PIB mensual, que deberían estar na mesma escala. Con este obxectivo, desfacemos as variacións interanuais mensuais do PIB que vimos de calcular tomando como puntos iniciais os valores do PIB mensuais ao longo de 2001, que se desagregaron mediante a función *td*, sen variables explicativas e co método uniforme. Na Figura 7.17 móstranse os gráficos secuencias do indicador mensual e o do PIB así mensualizado. De novo, destacamos a dinámica de 2020, onde claramente o noso índice capta moi ben o descenso e a seguida recuperación. Tamén podemos ver que o noso indicador mostra un restablecemento máis rápido que o PIB. Isto pode deberse a que a demanda eléctrica se recobra logo do confinamento, mentres que o PIB aínda queda lastrado, por exemplo, polo turismo.

7.3.3. Índice mensual e o IRE

Por último, imos comparar as variacións interanuais do índice mensual e do Índice da Rede Eléctrica (IRE), que definimos ao longo do Capítulo 4. Neste caso, empregaremos o IRE corrixido de efectos de calendario e da evolución das temperaturas, que xa vimos ao longo deste traballo que ten un efecto sobre a demanda de electricidade.

Parece interesante comparar o crecemento que experimentaron ambos indicadores, o noso índice mensual e o IRE, ao longo do tempo. Así, empregamos as variacións interanuais, que xa definimos na Subsección 7.3.2. Na Figura 7.18 móstranse os gráficos secuenciais de ambas variacións (dende 2016 ata o 2020). En xeral, podemos ver que a evolución é moi semellante ao longo dos anos, con especial énfase no descenso que se produciu en 2020.

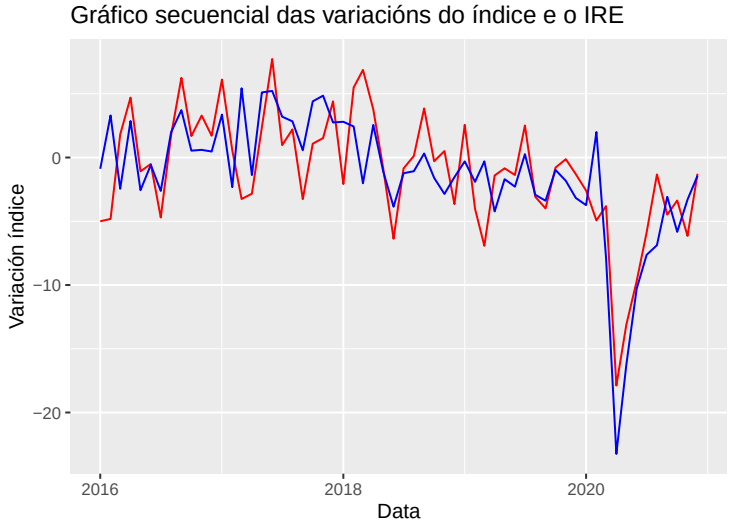
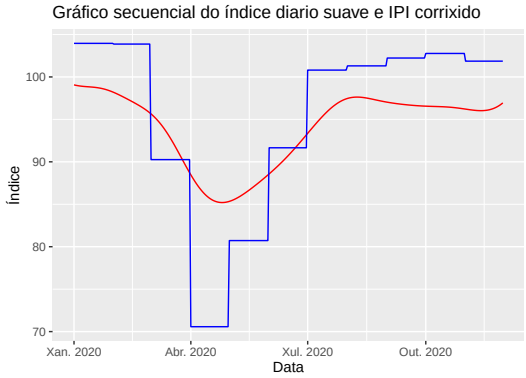


Figura 7.18: Gráficos secuenciais das variacións interanuais do índice mensual (en vermello) e do IRE (en azul).

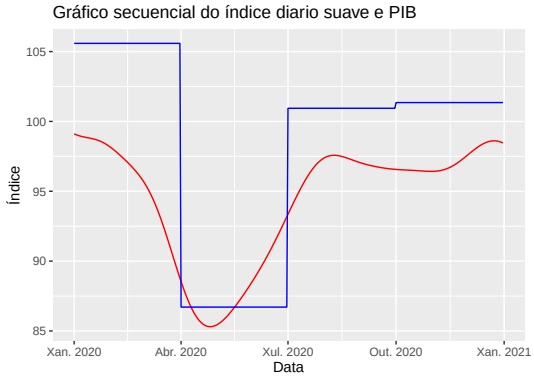
O coeficiente de correlación de Pearson entre as variacións destes dous indicadores é 0.75, indicando unha relación linear positiva considerablemente forte. Polo tanto, podemos concluír, tanto á vista do gráfico secuencial como do coeficiente de correlación mostral, que existe unha marcada relación entre o indicador que propoñemos e o IRE.

7.4. Capacidade de anticipación

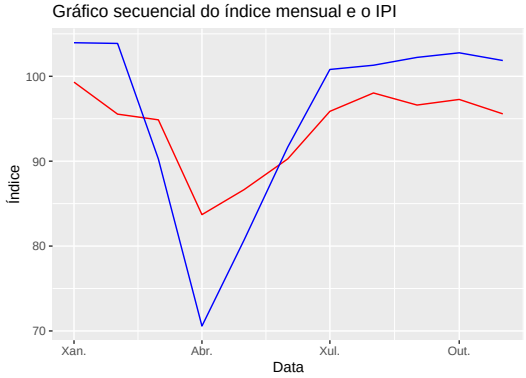
Ao longo deste capítulo presentamos un indicador de frecuencias diaria, semanal e mensual baseados na demanda de electricidade limpa e comparamos a dinámica do primeiro e do último respecto do IPI corrixido e do PIB, e en todos os casos víamos unha dinámica semellante: entre os anos 2015 e



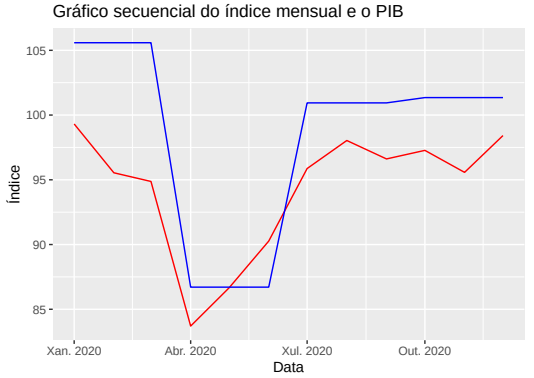
(a) Índice diario suave e IPI corrixido.



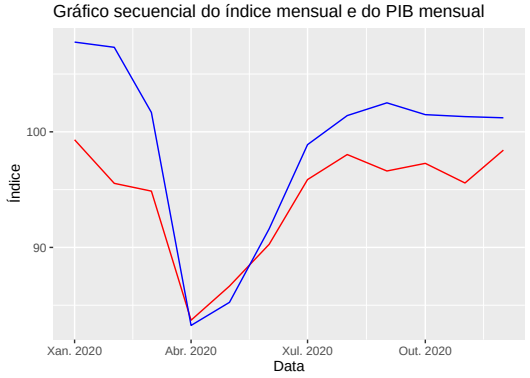
(b) Índice diario suave e PIB.



(c) Índice mensual e IPI corrixido.



(d) Índice mensual e PIB.



(e) Índice mensual e PIB mensualizado.

Figura 7.19: Gráficos secuenciais do noso índice en vermello e dos indicadores tradicionais en azul ao longo de 2020.

2019 o noso indicador parece non axustarse adecuadamente. É posible que existan outras variables que

permitan explicar mellor a evolución da economía ao longo desa etapa. Pero o indicador construído mediante a demanda eléctrica recolle aceptablemente ben o efecto que a pandemia da COVID-19 causou sobre a economía no 2020. Na Figura 7.19 móstranse os diferentes gráficos secuencias nos que comparabamos os índices diario e mensual co IPI e co PIB centrados no 2020. En todos eles, especialmente nos que se comparan o noso índice e o PIB, podemos ver que, efectivamente, o indicador proposto captura perfectamente o descenso da actividade económica como consecuencia do confinamento e a súa continua recuperación.

Ademais, debemos salientar que os indicadores tradicionais considerados, o IPI e o PIB, teñen unha frecuencia mensual e trimestral, respectivamente, e aínda tardan un período considerable de tempo en publicarse os resultados. Polo tanto, os indicadores usuais non nos indicarían que se produciu este grande impacto a partires de marzo de 2020 ata ben pasado uns meses despois. Mentres que os indicadores que vimos de construír permiten adiantar este suceso de xeito case inmediato.

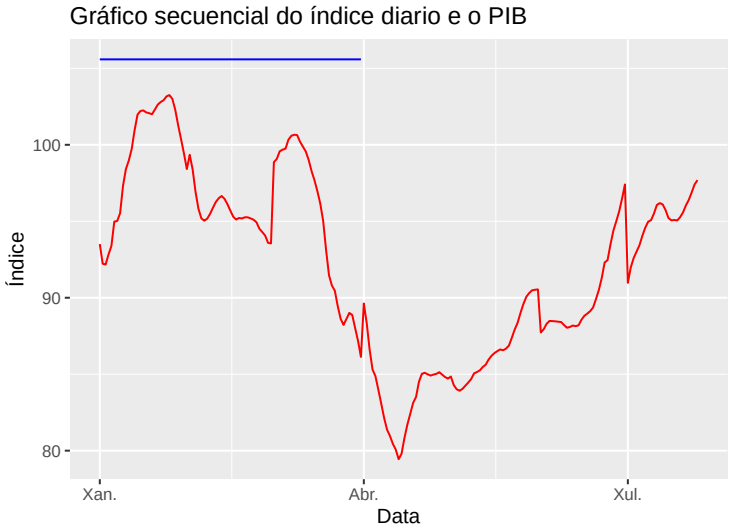


Figura 7.20: Gráficos secuenciais do indicador diario (vermello) e do PIB (azul) entre o comezo do 2020 ata o 25 de xullo dese ano, onde só dispoñemos do dato do primeiro trimestre do PIB.

Por exemplo, consideremos o rango de datas entre o comezo de 2020 ata o 25 de xullo, que abrangue o confinamento. A este último día, contamos co noso indicador diario ata o día anterior, o 24 de xullo, mentres que só dispoñemos do dato do primeiro trimestre do PIB. Podemos ver esta comparación na Figura 7.20, onde en vermello se presenta o indicador diario e en azul o PIB. Polo tanto, a 25 de xuño o noso índice xa nos indica o gran descenso que se produciu na actividade económica como consecuencia do confinamento e incluso o inicio dunha leve recuperación, mentres que se só dispoñemos da serie trimestral do PIB a ese día aínda non coñecemos o gran impacto que causou a pandemia. E aínda cando coñecemos o segundo dato do PIB, correspondente ao segundo trimestre, que se publica o 28 de xullo e se mostra na Figura 7.21, no que se presenta o brusco cambio no estado da economía, o noso indicador xa dispón de 25 observacións máis, correspondentes a xullo, e sinala unha recuperación, que non se verá no PIB ata catro meses aproximadamente despois (os tres meses correspondentes á súa frecuencia e o mes que tarda en publicarse).

Polo tanto, o noso indicador proporciona información sobre grandes impactos na economía en tempo real, esquivando a espera doutros indicadores tradicionais de frecuencias máis baixas, como o PIB. Amais, como puidemos ver, o noso índice tamén nos ofrece unha medida da baixada do PIB antes de que se publique o dato correspondente ao trimestre.

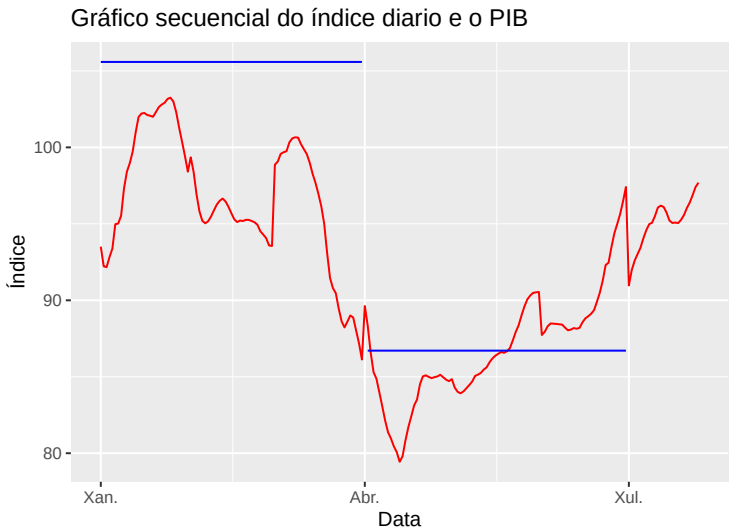


Figura 7.21: Gráficos secuenciais do indicador diario (vermello) e do PIB (azul) entre o comezo do 2020 ata o 25 de xullo dese ano, engadindo o dato do segundo trimestre do PIB.

7.5. Conclusión

Durante este capítulo presentamos o desenvolvemento dun indicador de frecuencias diaria, semanal e mensual baseado na demanda eléctrica limpa de efectos semanais e mensuais, tal e como se describiu no Capítulo 6. Comparamos a súa dinámica coa de agregados macroeconómicos máis tradicionais, como o IPI e o PIB, resultando nun comportamento non tan semellante entre os anos 2015 e 2019, pero unha vez que chega o 2020 e co comezo do confinamento, o noso indicador recolle os efectos deste impacto de forma moi aceptable. Tamén comparamos as variacións interanuais do indicador mensual e o IRE e puidemos ver que a súa dinámica é moi semellante dende o 2016 ata o 2020, especialmente neste último.

Por último, poñemos en valor o indicador que vimos de construír, pois aporta información instantánea de cambios tan bruscos, como a pandemia de COVID-19, no estado económico sen ter que esperar polas observacións doutros indicadores de frecuencias máis baixas, como o PIB.

En conclusión, calculamos un indicador a partir da demanda de electricidade limpa, que permite anticipar en tempo real grandes impactos nas actividades económicas, aínda que non é tan fino ante cambios pequenos. Así, no caso de monitorizarse, cumpriría a importante función dun detector de anomalías, coa vantaxe engadida de poder anticipar esta alarma como mínimo un mes antes que calquera dos indicadores tradicionais (PIB, IPI). Polo tanto, a súa contribución é dobre, por unha banda proporciona un medio para a detección de anomalías na actividade económica e, por outra, anticipa o aviso con respecto a outros indicadores, podendo servir de alternativa para a detección de anomalías temperás.

Capítulo 8

Aplicación en Shiny

Como se fixo mención ao longo do traballo, a parte práctica desenvolveuse mediante o software estatístico libre R (R Core Team 2020). Así, construímos un indicador de frecuencias mensual e inferiores mediante a demanda eléctrica limpa, empregando a linguaxe de programación do dito software. Agora ben, non todos os especialistas na área económica de ABANCA coñecen este soporte lóxico. Polo tanto, resulta de grande utilidade a creación dunha aplicación na que todo usuario poida acceder aos resultados, sen necesidade de indagar en R. Ademais, dado que os datos empregados para a construción do indicador son diarios, é importante ter automatizada a descarga dos mesmos, sen necesidade de acudir todos os días á páxina web da Rede Eléctrica de España.

Esta aplicación realizouse mediante a libraría *Shiny* (Chang et al 2020), cuxa estrutura básica se describe na Sección 8.1. Na Sección 8.2 descríbese o proceso de automatización da descarga dos datos de demanda eléctrica diaria total no territorio nacional. Por último, o funcionamento e a aparencia da aplicación creada móstrase na Sección 8.3.

8.1. Estrutura básica dunha aplicación en Shiny

A libraría *Shiny* permite crear aplicacións interactivas empregando código R e poden expoñerse abertamente nun web. Nesta sección presentamos brevemente a estrutura de construción básica dunha aplicación mediante esta libraría.

As aplicacións *Shiny*, como vimos de mencionar, son interactivas, por exemplo, o usuario pode escoller os valores de entrada (*inputs*) nun rango dado, polo que os ditos valores poden cambiar ao longo do tempo e, en consecuencia, os valores de saída (*outputs*) deben ser actualizados inmediatamente, reflectindo os cambios feitos. Así, *Shiny* emprega unha base de programación reactiva para automatizar a actualización das saídas cando as entradas se modifican.

Estas aplicacións constan de dúas compoñentes fundamentais, que interactúan entre elas:

- Interface gráfica (IG). Nesta parte defínese o aspecto e o deseño da aplicación. Nela poden actualizarse os valores de entrada e móstranse visualmente os valores de saída.
- Servidor (código). Nesta parte defínese o traballo para que a aplicación funcione. Constitúese por código R no que as variables poden cambiar ao gusto do usuario (valores de entrada) mediante a interface gráfica. Dado que poden ser modificadas, estes elementos son reactivos. Aqueles valores que se queren mostrar ao usuario coñécense como valores de saída e deben ser enviados á interface gráfica. Así, o servidor é a parte na que se reciben os valores de entrada e que se transforman mediante cálculos en R para acadar os valores de saída.

En resumidas contas, a interface gráfica le os valores de entrada que introduce o usuario. Logo, envíanse ao servidor, onde se realizan as modificacións necesarias para obter obxectos reactivos. Aqueles valores que interesa ensinar ao usuario (valores de saída) envíanse á interface gráfica con este

obxectivo. Cada vez que o usuario cambia un valor de entrada, os valores de saída, que dependen deste, actualízanse.

A aplicación constitúese dunha carpeta que contén o *script* de código R, *app.R*, composto pola interface gráfica e o servidor. Nesta carpeta tamén poden gardarse recursos adicionais que resultan necesarios para o funcionamento da aplicación. O ficheiro *app.R* tería a seguinte estrutura base:

```
# Cargamos a libraría Shiny
library(shiny)

# Interface gráfica
ui <- fluidPage( )

# Servidor
server <- function(input, output) { }

# Execución da aplicación
shinyApp(ui = ui, server = server)
```

Tamén se podería construír a aplicación separando ambas compoñentes, interface gráfica e servidor, en dous arquivos diferentes.

Unha vez visto isto, centrámonos no esquema do funcionamento da aplicación. Comezamos coa lectura de valores de entrada na interface gráfica. Todas as funcións adicadas á definición dos valores de entrada constan de dous argumentos principais. O primeiro, `inputID`, é o nome que identifica ao valor de entrada e pódese ler no servidor mediante `input$nome`, onde `nome` é o identificador do valor de entrada. O segundo argumento é o que se coñece como `label` e ten como fin crear unha etiqueta lexible para o usuario. Os restantes argumentos son específicos para o control. Por exemplo, pode fixarse un valor por defecto para o valor de entrada.

Os elementos reactivos, aqueles que se constrúen no servidor mediante código R e que empregan información dos valores de entrada, que non se queren mostrar na interface gráfica defínense mediante a función `reactive` e para chamar ao dito obxecto emprégase o seu nome seguido de `()`. Mentres que se queremos que estes obxectos se mostren na interface gráfica (valores de saída), debemos definilos mediante `output$nome`, onde `nome` é o nome do valor de saída, e coa función `render*`. O asterisco `*` en `render*` fai referencia ao tipo de valor de saída, por exemplo, `renderPlot` definiría un gráfico reactivo. Para mostrar os valores de saída na interface gráfica deben enviarse a esta compoñente. Con este obxectivo existen múltiples funcións segundo o tipo do valor de saída, por exemplo, `plotOutput` cando é un gráfico.

Exemplo

A continuación, mostramos un pequeno exemplo para ilustrar o funcionamento base dunha aplicación *Shiny*. Partimos da serie de tempo `AirPassengers`, que contén o número mensual de pasaxeiros en avión en miles entre os anos 1949 e 1960. O obxectivo é o deseño dunha aplicación na que o usuario especifique un ano de inicio e outro de fin entre 1949 e 1960 e que devolva o gráfico secuencial da serie comprendida entre os anos requiridos.

```
# Cargamos a libraría Shiny
library(shiny)

# Cargamos os datos
data <- AirPassengers

# Interface gráfica
ui <- fluidPage(
```

```
# Lectura dos valores de entrada
sliderInput(inputId = "rango",label = "Escolla as datas",
value =c(1949,1960), min = 1949, max = 1960),
# Mostramos o valor de saída, neste caso un gráfico
plotOutput("plot_rango")

)

# Servidor
server <- function(input, output) {
# Elemento reactivo, recorte da serie segundo os anos (valores de entrada)
serie_recortada<-reactive({
window(data,start=c(input$rango[1],1),end=c(input$rango[2],12))
})
# Valor de saída, neste caso o gráfico da serie recortada
output$plot_rango<-renderPlot({
autoplot( serie_recortada() )
})

}

# Executamos a aplicación
shinyApp(ui = ui, server = server)
```

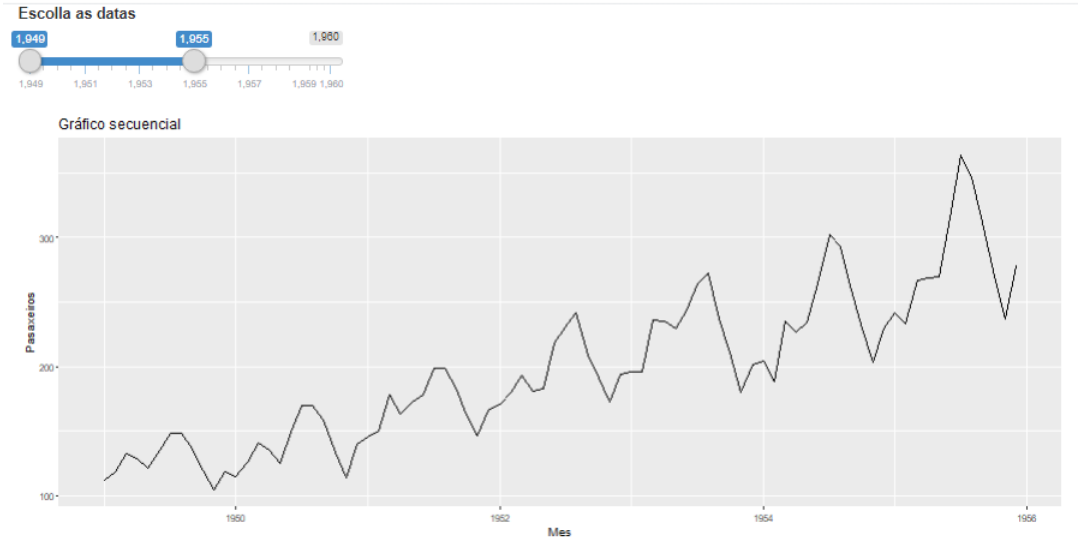


Figura 8.1: Aplicación do exemplo.

Así, comezamos por ler os valores de entrada. Para que o usuario poida seleccionar os anos de inicio e fin emprégase un selector escorregado `sliderInput`. Especifícase o nome deste valor de entrada no primeiro argumento, `rango`, e a etiqueta que se mostra na interface gráfica sobre o selector, `Escolla as datas`, no segundo. No terceiro defínense os valores por defecto do ano inicial, `1949`, e

final, `1960`, e fixamos estes anos como o mínimo e o máximo, respectivamente. Para chamar a este valor de entrada, constituído por dous valores (os anos de inicio e fin), no servidor emprégase o comando `input$rango`. En particular, se queremos chamar á primeira compoñente do valor de entrada, empregaremos `input$rango[1]` e analogamente para a segunda, `input$rango[2]`.

No servidor construímos un elemento reactivo, pois cambia segundo o valor do valor de entrada, que recorta a serie temporal de `AirPassengers` segundo os anos escollidos na interface gráfica, `serie_recortada`. Por último, xeramos o valor de saída que devolve o gráfico secuencial da `serie_recortada`, que chamamos por `serie_recortada()`, mediante a función `renderPlot` e nomeámolo por `plot_rango`.

Por último, para visualizar o gráfico secuencial da serie recortada na interface da aplicación, envíamos o valor de saída a esta compoñente. Como o valor de saída se construíu coa función `renderPlot`, para mostralo na interface gráfica empregaremos a función `plotOutput`, especificando o nome do valor de saída entre comiñas, `plot_rango`.

Na Figura 8.1 móstrase a aplicación resultante deste exemplo. Vemos que mediante un selector escorregado podemos escoller os anos de inicio e fin entre 1949 e 1960. Neste caso, está seleccionado 1949 e 1955, polo que o gráfico secuencial se mostra para este período de tempo.

No caso de que se busque información máis detallada e que trate en profundidade a construción de aplicacións en *Shiny* con múltiples exemplos, pode consultarse Wichkham (2021).

8.2. Automatización da descarga de datos

A demanda eléctrica coa que estivemos traballando ao longo deste proxecto ten unha frecuencia diaria. Para que non teñamos que acceder todos os días á páxina web da Rede Eléctrica de España para actualizar os datos, imos automatizar a súa descarga.

A Rede Eléctrica de España (REE) dispón dunha API[1], REData API, que permite acceder aos datos da entidade, posibilitando o seu emprego para intereses propios. Na propia páxina da REE, `https://www.ree.es/es/apidatos`, preséntanse as instrucións do uso da súa API e tamén se mostran algúns exemplos.

As solicitudes da súa API teñen a seguinte estrutura:

`GET /{lang}/datos/{category}/{widget}?[query]`,

onde `lang` define a lingua da resposta da consulta, `category` é a categoría xeral, `widget` é o *widget* particular e `query` é o conxunto de parámetros empregados para filtrar os datos que especificamos. Os diferentes valores que aceptan estes argumentos poden verse no web oficial da REE. No noso caso, especificamos que a lingua fose o inglés, `en`, a categoría `demanda` e o *widget* `evolucion`. O seguinte paso consiste en especificar os valores dos parámetros da consulta. Posto que queremos tomar a demanda eléctrica diaria ao longo dun ano, tomamos como `start_date` o primeiro día do dito ano e como `end_date` o último día no caso de que o ano concluíra. En caso contrario, tomaremos como data fin aquela anterior ao día presente, `Sys.Date()-1`. Dado que estamos interesados nos datos de frecuencia diaria, definimos `time_trunc=day`. Estas datas deben darse en formato ISO 8601, isto é `YYYY-MM-DDTHH:MM`. O seguinte argumento, `geo_trunc`, define o ámbito xeográfico dos datos solicitados e só admite o valor `electric_system`. Os dous últimos argumentos, `geo_limit` e `geo_ids`, precisan o sistema eléctrico da consulta e o seu ID, respectivamente. Por exemplo, se nos interesa a Península especificaríamos `geo_limit=peninsular` e `geo_ids=8741`. As respostas ás consultas atópanse en formato JSON.

No seguinte código exemplifícase a consulta das observacións diarias da demanda eléctrica na Península nun ano rematado.

```
# Datos Península
```

[1] API é o acrónimo de *Application Programming Interface* e está composto por un conxunto de funcións que permite que as aplicacións accedan a datos e interactúen con elementos externos.

```
# Concatenamos a consulta segundo o ano que especifiquemos
url<-paste("https://apidatos.ree.es/en/datos/demanda/evolucion?start_date=",
ano,
"-01-01T00:00&end_date=",
ano,
"-12-31T23:59&time_trunc=day&geo_trunc=electric_system&geo_limit=peninsular&geo_ids=8741",
sep="")
# Chamamos aos datos en formato json e convertímolos en obxecto R
dat<-jsonlite::fromJSON(url)
```

No noso caso, descargamos os datos diarios na Península, Illas Canarias, Illas Baleares, Ceuta e Melilla, cuxa suma resulta na demanda eléctrica total no país, ao longo dos anos 2015 a 2020 e gardamos os ficheiros nun arquivo *RData*. As observacións de 2021 ata o día precedente ao actual descárganse automaticamente unha vez que executemos a aplicación que construímos.

8.3. Aplicación

Nas Figuras 8.2 e 8.3 móstranse capturas de dúas das xanelas da aplicación que fixemos. Guiándonos por estas imaxes, describiremos o interior de cada unha das xanelas, especialmente da que ten por nome `Indicador baseado na demanda eléctrica`, pois céntrase no indicador que desenvolvemos ao longo deste traballo.

Se pinchamos sobre `Indicador baseado na demanda eléctrica` atopámonos con diferentes xanelas. Na `Introdución` relátase a construción do dito indicador e o que podemos atopar nesta aplicación: os gráficos secuenciais da demanda eléctrica bruta, sen observacións atípicas e sen efectos semanais nin mensuais (`Demanda eléctrica`), os gráficos secuencias do indicador de frecuencias diaria, semanal e mensual (`Indicador`) e a súa comparación con algúns agregados macroeconómicos, como o PIB, IPI e o IRE (`Indicador diario vs agregados macroeconómicos` e `Indicador mensual vs agregados macroeconómicos`).

Na xanela `Lectura de datos` o usuario debe indicar as datas nas que está interesado mediante un selector despregable. A mínima data é o inicio de 2015 e a máxima o día anterior ao actual. Segundo as datas especificadas, móstranse os gráficos secuencias citados, onde o usuario pode interactuar con elas. Por exemplo, facendo zoom nun rango de datas en particular ou empregando o cursor para coñecer o valor da serie no punto onde se colocou. Estes gráficos fixéronse mediante a función *dygraph* da libraría *dygraphs* (Vanderkam et al 2018), onde, cando se quere debuxar series de tempo, o argumento `data` que se lle pasa debe estar en formato *xts* (Ryan e Ulrich 2018). Esta función permite xogar con varias opcións que fan do gráfico interactivo. Por exemplo, a función *dyRangeSelector* engade un selector de rango baixo o gráfico, que permite facer zoom no rango especificado polo usuario. E a función *dyHighlight* permite personalizar como é o aspecto do resaltado da serie temporal cando situamos o cursor enriba dela.

Esta aplicación permite, polo tanto, ver os gráficos secuencias da demanda eléctrica e do indicador construído mediante ela. Pero é interesante poder dispoñer das observacións do indicador nun arquivo Excel, o formato que empregan os usuarios en ABANCA, para rexistrar os resultados e empregalos para diferentes fins, como análises económicas, presentacións en comités, etc. Así, na xanela `Descarga de datos` permítese a descarga dos datos correspondentes aos indicadores segundo a súa frecuencia nun ficheiro Excel.

Doutra banda, a xanela `Outros indicadores adiantados` está pensada para engadir diferentes indicadores baseados en variables de alta frecuencia, como a mobilidade. Na seguinte xanela, `Variables macroeconómicas`, inclúense diferentes variables coas que traballan os usuarios de ABANCA, como o Produto Interior Bruto e o Índice de Produción Industrial. Por último, na xanela `Memoria TFM`

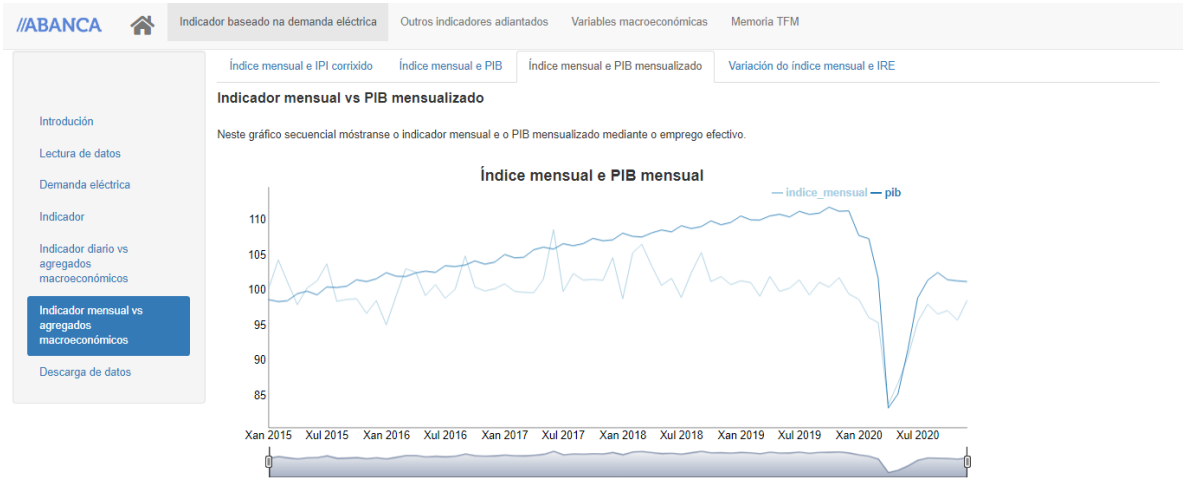
(a) Xanela `Portada`.

(b) Xanela `Demanda eléctrica`.

Figura 8.2: Captura dalgunhas xanelas da aplicación.

incluímos a documentación deste traballo, pois pode resultar de utilidade como complemento da aplicación.

(a) Xanela `Indicador`.



(b) Xanela `Indicador mensual vs agregados macroeconómicos`.

Figura 8.3: Captura dalgunhas xanelas da aplicación.

Capítulo 9

Conclusións e liñas futuras de investigación

A inesperada pandemia provocada pola COVID-19 puxo de manifesto a necesidade de dispoñer de indicadores de alta frecuencia para o seguimento do estado económico do país, pois os indicadores máis usuais, como o Produto Interior Bruto (PIB), acostuman ter unha frecuencia trimestral e, a maiores, existen atrasos na súa publicación. Así, este traballo céntrase no comezo do desenvolvemento dun indicador de frecuencia inferior á do PIB, que permitiría adiantar a dinámica do estado económico nun corto período de tempo. Para a súa construción consideramos como primeira variable a demanda de electricidade diaria total no país.

Comezamos describindo anualmente o comportamento da demanda eléctrica ao longo do Capítulo 5. Primeiro, analizamos o 2017 como mostra dun ano normal en termos de evolución económica. Continuamos cun estudo a longo prazo desta variable, tomando como período temporal os anos comprendidos entre 2015 e 2019, e puidemos ver que o patrón que estudamos no ano 2017 se mantén. Pero esta dinámica xa non é a mesma ao longo do 2020, senón que hai un descenso significativo provocado polo confinamento. Disto dedúcese que o comportamento da demanda eléctrica reflexa o estado das actividades económicas fronte a cambios bruscos. Tamén vimos que a demanda eléctrica presenta atípicos nas festividades nacionais, compoñente estacional semanal e un efecto mensual. Mediante a aplicación de medias móbiles con xanela $h = 7$ eliminamos a estacionalidade semanal (Sección 6.2), mentres que a resolución do efecto mensual precisa dun maior estudo.

O seguinte paso, como adiantamos, é eliminar o efecto mensual que permanece na variable suavizada por medias móbiles. Para esta tarefa probáronse diferentes metodoloxías, expostas na Sección 6.3. Finalmente, chegamos á conclusión de que a mellor opción, dentro do noso interese, para a eliminación do efecto mensual é corrixir a demanda eléctrica suavizada mediante a variación en tanto por un da mediana mensual desta variable respecto da mediana da dita variable ao longo dos anos considerados.

Unha vez que dispoñemos da demanda eléctrica limpa, tanto do efecto semanal como mensual, construímos un índice de diferentes frecuencias. Tamén comparamos a súa dinámica co Índice de Produción Industrial (IPI) e co Produto Interior Bruto (PIB). O comportamento do noso índice nos anos anteriores ao 2020 non é moi semellante ao destes agregados, pero debemos ter en conta que o noso índice foi construído en base a unha soa variable, mentres que os indicadores macroeconómicos teñen en conta múltiples factores. Agora ben, en canto ao 2020, o noso indicador presenta un comportamento moi similar ao dos índices tradicionais. Polo tanto, acadamos un indicador de alta frecuencia que permite identificar grandes cambios na evolución da economía, como a pandemia de COVID-19, nun curto período de tempo, mentres que, por exemplo, o PIB sinalaríao meses despois. De feito, o noso indicador diario xa anuncia a leve recuperación da economía a mediados de abril de 2020 cando saen os resultados do PIB, notificando o gran descenso froito do confinamento. En consecuencia, cabe salientar a potencial utilidade do indicador que vimos de construír.

Tamén comparamos a dinámica das variacións interanuais do indicador mensual e o Índice Rede Eléctrica (IRE), resultando nun comportamento moi semellante, pois ambos se basean na mesma variable.

Ademais, desenvolvemos unha aplicación en *Shiny* (Chang et al 2020), onde cada vez que se executa a aplicación se descargan automaticamente os datos de demanda eléctrica diaria total en España e se permite visualizar os gráficos secuencias desta variable e do indicador que construímos. Tamén se posibilita a baixada dos datos do indicador nun arquivo Excel.

Para abordar o problema proposto por ABANCA utilizamos unha ampla variedade de técnicas estatísticas, dende a análise exploratoria ao axuste de modelos non paramétricos, pasando polo uso de técnicas computacionais para elaborar cadros de mando, e destacando especialmente a aplicación de ferramentas para o estudo e modelización de series temporais. Polo tanto, no presente traballo aplicamos unha boa parte de todos os conceptos adquiridos no Máster en Técnicas Estatísticas para cumprir cos obxectivos marcados por ABANCA, no marco da resolución de problemas no ámbito empresarial.

O índice que presentamos neste traballo só emprega unha variable, a demanda eléctrica, pero o problema do que xorde este proxecto enmárcase no mundo multivariante. Polo tanto, o seguinte paso reside en engadir máis variables de alta frecuencia, como a mobilidade, os pagos con tarxetas e as tendencias extraídas de Google, entre outras variables, para construír un indicador ou indicadores que inclúan unha información máis completa da actividade económica e que anticipen os seus cambios, sexan estes grandes ou pequenos.

Apéndice A

Análise da demanda eléctrica

Neste capítulo preséntanse os resultados da análise da demanda de electricidade nos anos 2018 e 2019, complementando o visto no Capítulo 5. Este estudo fíxose coa intención de comprobar se o comportamento da demanda eléctrica é a mesma ao longo dun ano, considerando que este estea baixo condicións normais en termos de evolución económica.

A.1. Análise exploratoria da demanda eléctrica en 2018

Nesta sección mostramos os resultados da análise exploratoria da demanda diaria de electricidade (en XWh) no territorio nacional en 2018. Na Figura A.1 móstrase o gráfico secuencial da serie de

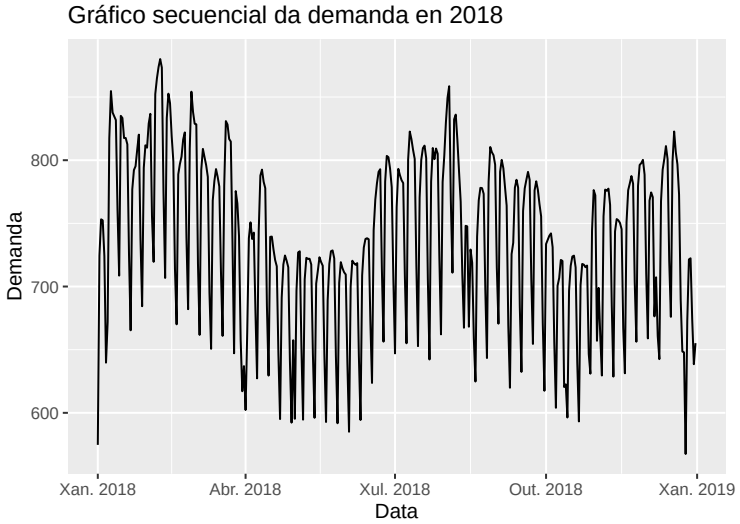


Figura A.1: Gráfico secuencial da demanda diaria total de electricidade no territorio nacional no ano 2018.

demanda eléctrica no territorio nacional no ano 2018. Podemos ver que a dinámica da serie é moi semellante á da demanda eléctrica en 2017 (Figura 5.1), cunha posible presenza de tendencia e patrón repetitivo semanal. A serie ten forma, que consiste nunha subida ata mediados de febreiro (baixa máis tarde que no 2017), para que logo comece a baixar ata mediados de abril. Mantense constante durante

unhas semanas e comezar a subir de novo en xuño. A serie continúa subindo ata redores de setembro, onde hai unha baixada. O nivel volve subir en novembro e pega un descenso a finais de decembro, posiblemente motivado polos festivos de Nadal. Na Figura A.2 móstranse as autocorrelacións simples mostrais desta serie. Neste percibimos unha compoñente estacional de período $s = 7$, dado que no gráfico secuencial víamos patrón repetitivo e as correlacións mostrais nos múltiplos de 7 son positivas altas e tardan en reducir a súa magnitude. Este comportamento tamén o víamos na serie do ano anterior. Así, a serie non é estacionaria como consecuencia desta estacionalidade semanal, mentres que a varianza da serie parece ser constante.

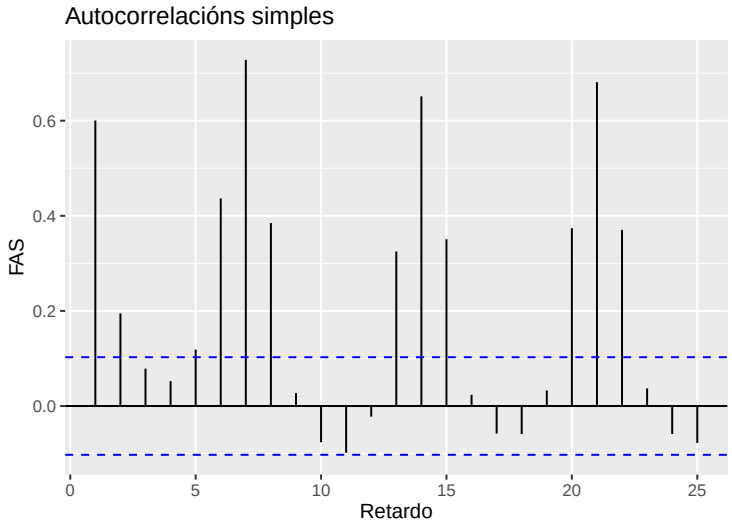


Figura A.2: Autocorrelacións simples mostrais da serie de demanda eléctrica en 2018.

Na Táboa A.1 móstranse as medidas de posición da demanda de electricidade en 2018. Podemos ver que o mínimo é 567.4 XWh (superior ao do ano pasado) e o máximo 880.1 XWh (inferior ao do ano anterior). A mediana é 738.3 XWh (superior á do ano pasado) e a media 736.7 XWh (tamén superior). O primeiro cuartil coincide cun valor de demanda de 695.1 XWh (levemente inferior) e o terceiro cuartil con 790.8 XWh (superior). En xeral, podemos dicir que a demanda de electricidade total neste ano é levemente superior respecto do anterior.

Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
567.4	695.1	738.3	736.7	790.8	880.1

Táboa A.1: Medidas de posición da demanda eléctrica en 2018. (en XWh).

Na Figura A.3 móstrase o diagrama de caixas da demanda eléctrica neste ano. Podemos ver unha lixeira asimetría á dereita e non hai atípicos. Doutra banda, na Figura A.4 mostramos, á esquerda, o histograma de frecuencias absolutas e, á dereita, o histograma xunto coa función de densidade en vermello. Tanto no histograma como na densidade podemos ver asimetría. Esta última presenta tres modas.

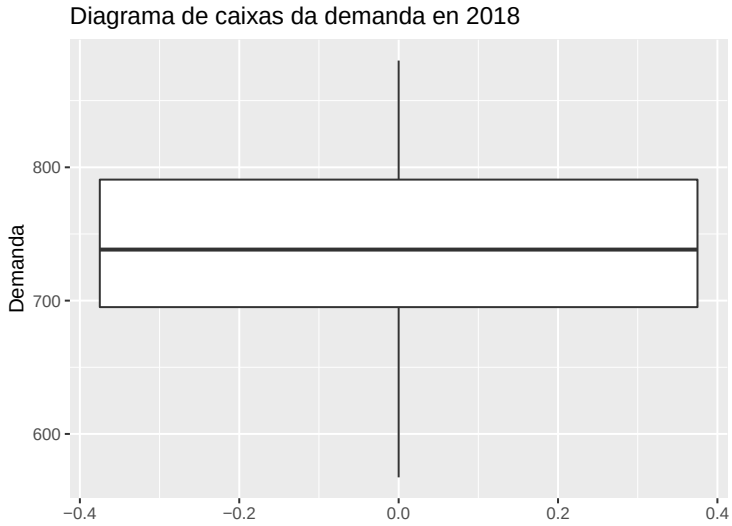


Figura A.3: Diagrama de caixas da demanda total de electricidade en 2018.

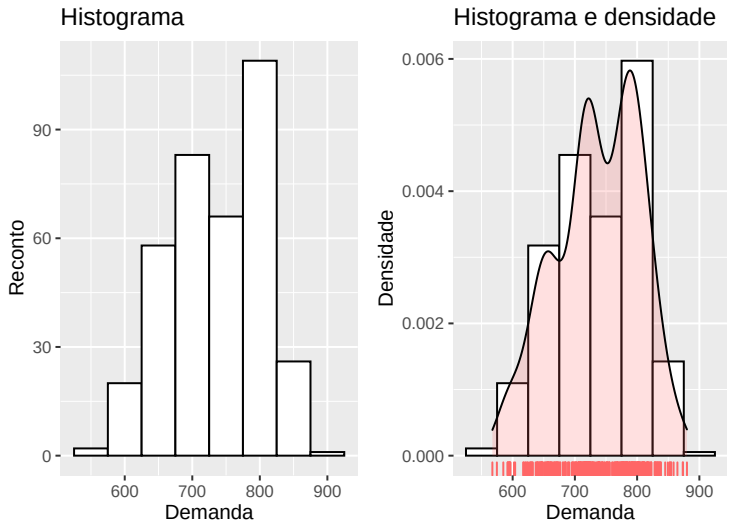


Figura A.4: Histograma de frecuencias absolutas (esquerda) e histograma e función tipo núcleo da densidade (dereita) da demanda total de electricidade en 2018.

Dado que no gráfico secuencial da serie se intuía unha tendencia dada por un efecto mensual, imos estudar a demanda eléctrica por meses. O diagrama de caixas por meses desta variable móstrase na Figura A.5. Podemos ver que todos os meses son asimétricos e que, ademais, xaneiro e xuño presentan valores atípicos. Amais, podemos notar que existe unha maior demanda nos meses de inverno e de verán, sinalando este efecto mensual provocado polas temperaturas.

Na Táboa A.2 mostramos as medidas de posición da variable por meses. Se nos fixamos na mediana, podemos ver que hai unha subida en febreiro, para logo ir baixando ata maio. En xuño hai unha subida

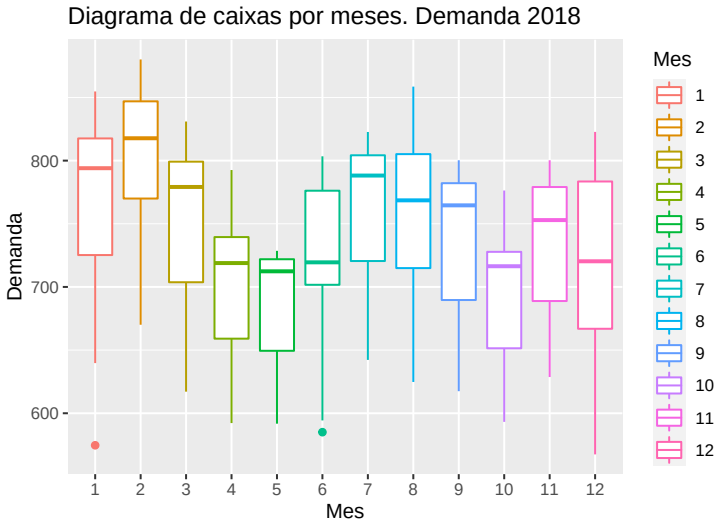


Figura A.5: Diagrama de caixas por meses da demanda eléctrica en 2018.

Mes	Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
Xaneiro	574.6	725.2	794	768.5	817.6	854.7
Febreiro	670.1	770	817.7	801.7	846.9	880.1
Marzo	617.1	703.7	779.1	751.7	799.2	830.9
Abril	592.2	659	718.9	702.9	739.5	792.6
Maio	591.7	649.4	712.4	687.4	721.9	728.6
Xuño	585	701.7	719.5	720.4	776.2	803.4
Xullo	642.2	720.5	788.2	763.3	804.2	822.7
Agosto	624.8	714.9	768.6	758.4	805.1	858.6
Setembro	617.5	689.6	764.6	737.4	782.1	800.3
Outubro	593.2	651.4	716.4	696.1	727.8	776.3
Novembro	628.7	688.8	752.9	735.1	779.1	800.3
Decembro	567.4	666.9	720.3	721.7	783.5	822.7

Táboa A.2: Medidas de posición da demanda eléctrica por meses en 2018 (en XWh).

livián e podemos notar que a demanda en xullo é superior á dos restantes meses de verán. A variable continúa baixando ata novembro, onde sube, para en decembro volver baixar. Podemos ver que esta dinámica se corresponde coa forma que describíamos no gráfico secuencial (Figura A.1). Ademais, esta correspóndese esencialmente coa do anterior ano. Todo apunta, polo tanto, á existencia dun efecto mensual, que cremos que vén dado pola variación da temperatura nas diferentes estacións do ano.

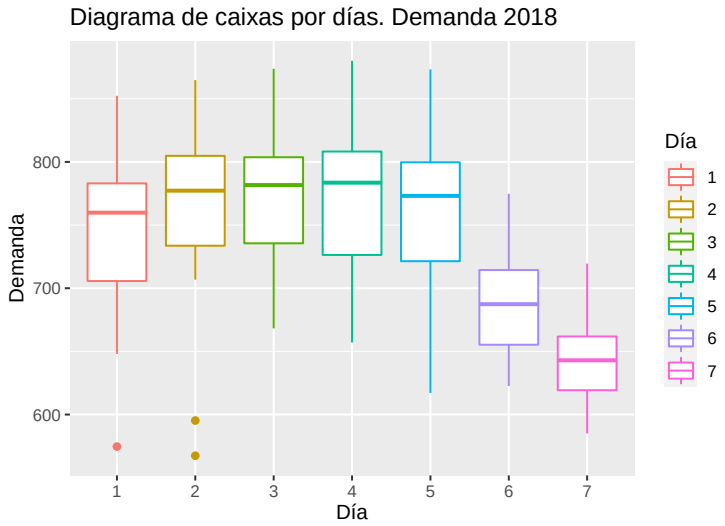


Figura A.6: Diagrama de caixas por días da demanda eléctrica en 2018.

Día	Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
Luns	574.6	705.7	759.8	749.6	783	852.4
Martes	567.4	733.7	777.2	770.2	804.8	864.7
Mércores	668.2	735.6	781.7	773.8	803.7	873.8
Xoves	657	726.3	783.6	769.6	808.2	880.1
Venres	617.1	721.4	773.1	762.9	799.7	873.3
Sábado	622.6	655.3	687.4	689.2	714.4	774.8
Domingo	585	619.3	642.9	641.3	661.8	719.5

Táboa A.3: Medidas de posición da demanda eléctrica por días en 2018 (en XWh).

Tamén fixemos un estudo da demanda por días para estudar o patrón repetitivo semanal que presenta a serie. Na Figura A.6 preséntase o diagrama de caixas por días. Podemos notar que os

Día	Luns	Martes	Mércores	Xoves	Venres	Sábado
Martes	0.23					
Mércores	0.16	1				
Xoves	0.38	1	1			
Venres	1	1	1	1		
Sábado	1.9×10^{-7}	2.1×10^{-11}	1.6×10^{-12}	1.6×10^{-11}	3×10^{-10}	
Domingo	9.8×10^{-14}	9.6×10^{-15}	$< 2\times10^{-16}$	3.4×10^{-16}	5.9×10^{-15}	5.9×10^{-7}

Táboa A.4: P-valores asociados ao estatístico de contraste do test de rangos con signo.

sábados e domingos teñen unha menor demanda, sendo menor neste último caso. No fin de semana tamén hai unha menor dispersión dos datos. Os luns e venres semellan ter unha menor demanda entre os días laborais. Por último, podemos ver que en todos os casos hai asimetría negativa e os luns e martes presentan observacións atípicas, posiblemente como resultado de días festivos.

Na Táboa A.3 móstranse as medidas de posición da demanda por días. Fixándonos na mediana, vemos que a demanda é maior, entre os días laborais, os mércores e xoves, seguidos de preto polos martes (semellante ao que pasaba no ano 2017: a demanda nos días laborais é superior á dos fins de semana e entre eles a demanda nos martes, mércores e xoves é maior). Logo, a demanda decrece considerablemente no sábado e séguео facendo no domingo. Así, podemos pensar que existe un efecto de fin de semana. Sería interesante contrastalo, pero para isto precisamos saber se as mostras de demanda por días son independentes ou dependentes. Para isto aplicamos o test de independencia τ de Kendall por pares e, a un nivel $\alpha = 0.05$, rexeitase a hipótese nula de independencia en todos os casos. Dado que as mostras son dependentes, aplicamos o test de rangos con signo de Wilcoxon entres pares para contrastar a igualdade de medianas. Os p-valores asociados ao estatístico de contraste móstranse na Táboa A.4. A súa vista, rexeitamos a hipótese nula a un nivel $\alpha = 0.05$ nos sábados e domingos. Polo tanto, existe un efecto semanal.

A.2. Análise exploratoria da demanda eléctrica en 2019

Nesta sección presentamos os resultados da análise exploratoria da demanda diaria de electricidade (en XWh) no territorio nacional en 2019. Na Figura A.7 móstrase o gráfico secuencial da serie. Parece que a serie presenta tendencia e patrón repetitivo, ao igual que nos anteriores anos. A forma da serie consiste nunha subida ata principios do segundo mes Febreiro (comeza a baixar aproximadamente como en 2017), para que logo baixe ata finais de maio. Continúa cunha subida leve ata xullo e a demanda volve baixar ao redor de setembro. A serie continúa baixando ata novembro, punto no que comezará a subir. Ao final hai unha baixada pronunciada. Podemos ver que a dinámica concorda coa dos anos anteriores. Ademais, a varianza da serie é constante.

Se nos fixamos nas autocorrelacións simples mostrais da serie, que se mostran na Figura A.8, podemos ver que hai repuntes fortes positivos nos múltiplos de 7 tardando en baixar. En consecuencia disto e do patrón repetitivo que vimos, a serie presenta compoñente estacional con período $s = 7$. Polo tanto, a serie non é estacionaria.

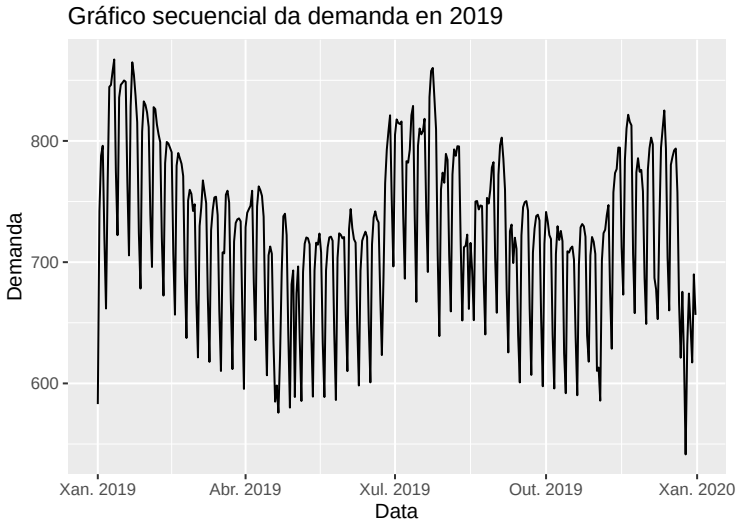


Figura A.7: Gráfico secuencial da demanda diaria total de electricidade no territorio nacional no ano 2019.

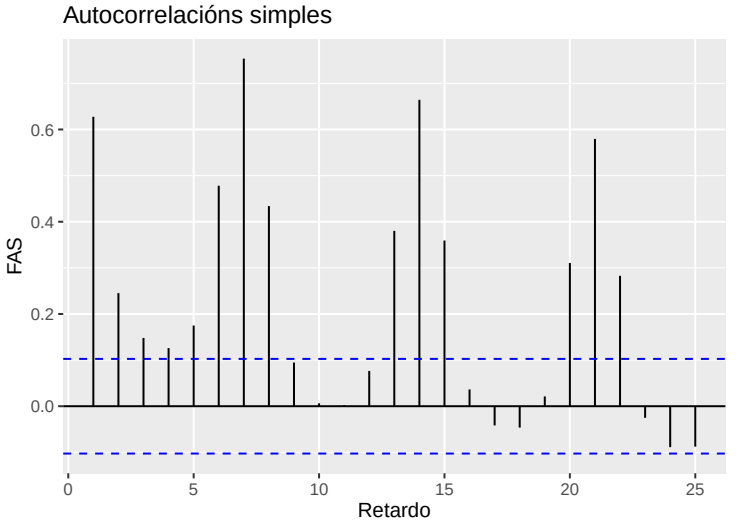


Figura A.8: Autocorrelacións simples mostrais da serie de demanda eléctrica en 2019.

O resumo das medidas de posición da variable móstrase na Táboa A.5. Podemos ver que o mínimo é 541.5 XWh (inferior ao dos anteriores anos) e o máximo 867.2 XWh (tamén inferior ao dos anteriores anos). A mediana é 725.8 XWh (inferior) e a media 725.1 XWh (inferior). O primeiro cuartil correspóndese cunha demanda de 685.7 XWh (inferior) e o terceiro con 775.2 XWh (inferior). Así, a demanda total neste ano foi lixeiramente inferior que a dos anteriores anos. No diagrama de caixas, que se mostra na Figura A.9, podemos ver unha leve asimetría á dereita e presenta un atípico.

Na Figura A.10 móstranse, á esquerda, o histograma de frecuencias absolutas e, á dereita, o histo-

Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
541.5	685.7	725.8	725.1	775.2	867.2

Táboa A.5: Medidas de posición da demanda eléctrica en 2019 (en XWh).

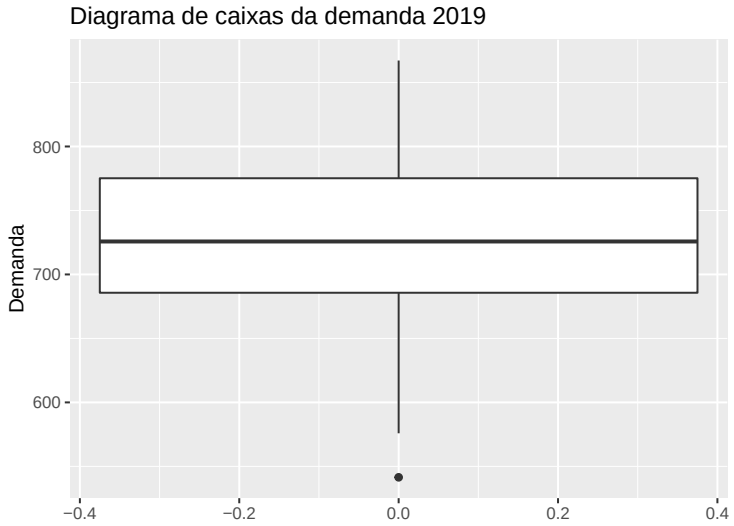


Figura A.9: Diagrama de caixas da demanda total de electricidade en 2019.

grama xunto coa función tipo núcleo da densidade. Tanto o histograma como a función de densidade presentan asimetría. Esta última só presenta unha moda.

A continuación, móstranse os diagramas de caixas por meses na Figura A.11 e as súas respectivas medidas de posición na Táboa A.6. Se nos fixamos na mediana, podemos ver que a demanda tende a baixar ata maio, onde comeza a subir. A demanda entre os meses de verán é maior en xullo. Esta continúa baixando ata novembro, onde se da unha subida. En decembro a demanda diminúe. Esta dinámica correspóndese coa dos anteriores anos, parece depender do mes e segundo a temperatura que se dea no mesmo. As diferenzas entre as medianas tamén se poden ver no diagrama de caixas. Nel destaca a dispersión no mes de decembro respecto dos anteriores meses. Este sofre asimetría positiva mentres que os outros meses presentan asimetría negativa. Por último, podemos ver que xaneiro, febreiro e xullo presentan valores atípicos. Percibimos, así, a existencia dun efecto mensual provocado, aparentemente, pola temperatura.

Tamén fixemos unha análise da demanda de electricidade por días. Comezamos introducindo os diagramas de caixas por días na Figura A.12. Podemos ver que en todos os casos hai asimetría positiva e martes, mércores e venres presentan atípicos (posiblemente por mor de festivos). Tamén podemos ver que hai unha maior demanda nos días laborais que nos fins de semana. Isto tamén pode verse na mediana por días, cuxos valores se presentan na Táboa A.7 xunto con outras medidas de posición. Podemos ver que entre os días laborais, a demanda é maior os mércores e xoves. A demanda decrece no fin de semana, sendo o sábado o que presenta unha maior demanda. Así, semella que existe un

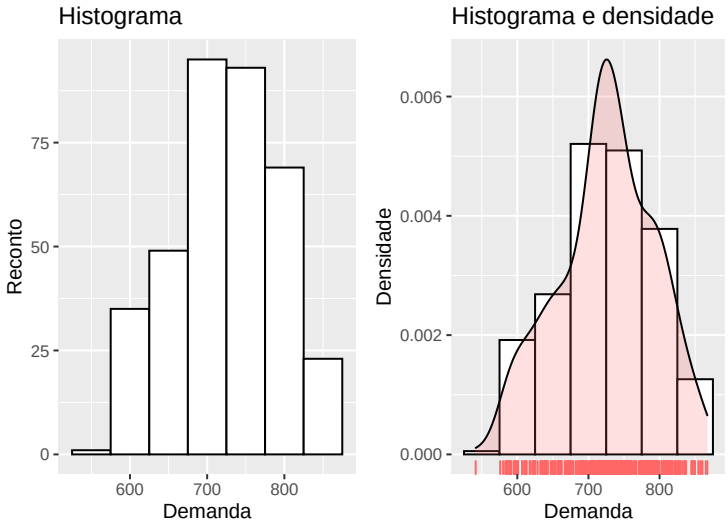


Figura A.10: Histograma de frecuencias absolutas (esquerda) e histograma e función tipo núcleo da densidade (dereita) da demanda total de electricidade en 2019.

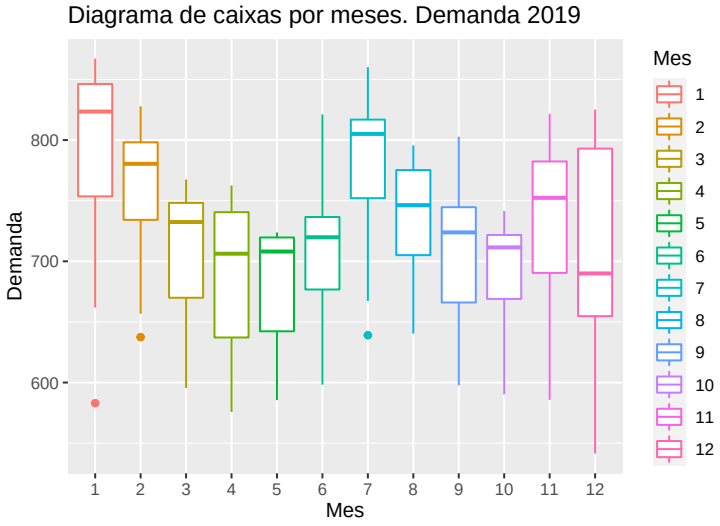


Figura A.11: Diagrama de caixas por meses da demanda eléctrica en 2019.

efecto de fin de semana, como nos anteriores anos.

Sería interesante contrastar o efecto semanal. Primeiro, necesitamos saber se as mostras de demanda por días son ou non independentes. Para isto, aplicamos o test de independencia τ de Kendall entre pares de días, resultando no rexeitamento en todos os casos da hipótese nula de independencia a un nivel $\alpha = 0.05$. Dado que as mostras son dependentes, tomamos o contraste de rangos con signo de Wilcoxon entre pares para ver se as medianas por días son iguais. Os p-valores asociados ao estatístico de contraste móstranse na Táboa A.8. En base a isto, rexeitamos a hipótese nula de igualdade de

Mes	Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
Xaneiro	582.9	753.6	823.5	792.5	846.2	867.2
Febreiro	637.5	734.2	780.4	760	798.1	827.9
Marzo	595.5	669.9	732.4	707.6	748.2	767.5
Abril	575.9	637.2	706.3	689.7	740.5	762.5
Maio	585.5	642.3	708.1	681.8	719.7	723.8
Xuño	598.3	676.7	719.8	708.6	736.5	821.2
Xullo	639.1	752.1	805.1	781.4	816.9	860.2
Agosto	640.4	705.1	746.3	732.6	775.3	795.6
Setembro	597.6	666	723.8	709.9	744.6	802.7
Outubro	590.2	669	711.5	692.1	721.7	741.4
Novembro	585.7	690.4	752.3	733.3	782.3	821.6
Decembro	541.5	654.8	690	713.3	793	825.1

Táboa A.6: Medidas de posición da demanda eléctrica por meses en 2019 (en XWh).

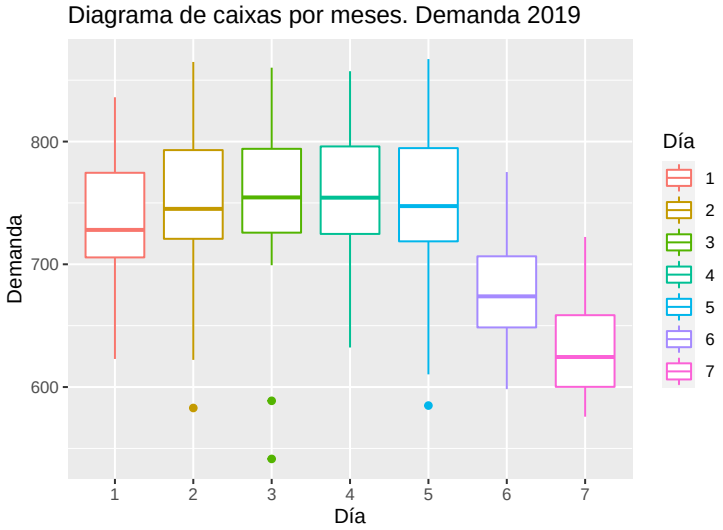


Figura A.12: Diagrama de caixas por días da demanda eléctrica en 2019.

Día	Mínimo	Primeiro cuartil	Mediana	Media	Terceiro cuartil	Máximo
Luns	622.9	705.6	728	739.4	774.6	836.1
Martes	582.9	720.8	745.2	753.4	793	864.9
Mércores	541.5	725.8	754.5	757.4	794.1	860.2
Xoves	632.2	724.8	754.3	760	796	857.3
Venres	584.9	718.7	747.4	752.1	794.7	867.2
Sábado	598.3	648.6	673.9	680.5	706.5	775.2
Domingo	575.9	600.2	624.5	632.4	658.6	722.3

Táboa A.7: Medidas de posición da demanda eléctrica por días en 2019 (en XWh).

Día	Luns	Martes	Mércores	Xoves	Venres	Sábado
Martes	0.52					
Mércores	0.12	1				
Xoves	0.12	1	1			
Venres	0.58	1	1	1		
Sábado	7.6×10^{-8}	1.4×10^{-9}	1.4×10^{-10}	1.9×10^{-10}	2.2×10^{-9}	
Domingo	3.4×10^{-15}	9.1×10^{-15}	1.1×10^{-14}	2×10^{-15}	1.2×10^{-14}	9.8×10^{-7}

Táboa A.8: P-valores asociados ao estatístico de contraste dos rangos con signo.

medianas nos sábados e domingos. Así, podemos concluír que existe un efecto de fin de semana.

A.3. Conclusión

Xa analizamos a demanda diaria de electricidade no territorio nacional nos anos 2017, 2018 e 2019 por separado e chegamos a resultados moi semellantes. Tal e como podemos ver na Figura A.13, a súa dinámica consiste nunha maior demanda durante os meses de verán e inverno, posiblemente motivada polo uso de aire acondicionado e calefacción, respectivamente. Así, parece existir un efecto mensual causado pola temperatura. Ademais, a serie de demanda eléctrica presenta un efecto semanal, que puidemos contrastar nos tres escenarios mediante o test de rangos con signo de Wilcoxon.

Respecto da media da demanda eléctrica nestes tres anos, vimos que o 2018 presenta un maior

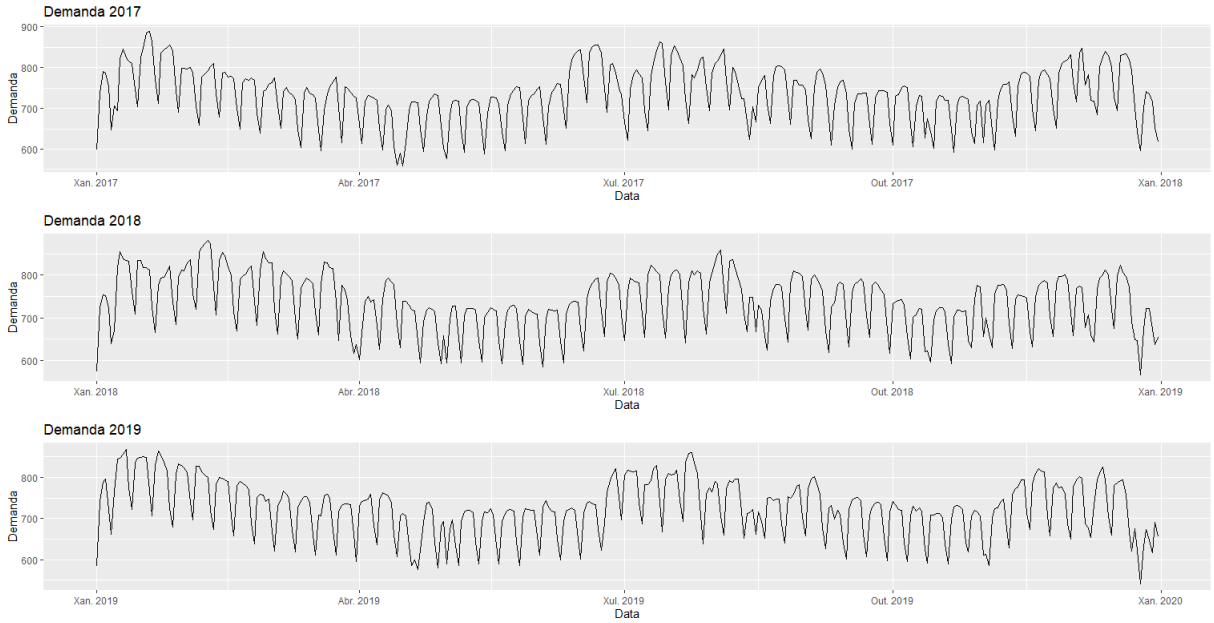


Figura A.13: Gráficos secuenciais da demanda diaria total de electricidade no territorio nacional nos anos 2017, 2018 e 2019 (de arriba a abaixo).

nivel, seguido do 2017, mentres que a demanda de electricidade no 2019 viuse reducida.

Bibliografía

[1] http://www.aemet.es/es.

[2] Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control. Holden-Day, San Francisco.

[3] Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J (2020) shiny: Web Application Framework for R. R package version 1.4.0.2. URL: https://CRAN.R-project.org/package=shiny.

[4] Chislett W (2021) Challenges and opportunities for Spain in times of COVID-19. Working Paper, Real Instituto Elcano, Madrid.

[5] Cleveland RB, Cleveland WS, McRae JE, Terpenning I (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess. Journal of Official Statistics, Vol. 6, No. 1: 3-73.

[6] Cleveland WS (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association, Vol. 74, No. 368: 829-836.

[7] Cryer JD, Chan K (2010) Time Series Analysis. With Applications in R. Springer, Nova York.

[8] Dorin F, Perrotti D, Goldszier P (2020) Index numbers and their relationship with the economy. ECLAC Methodologies, Santiago.

[9] Fernández de Castro BM, Fernández Sotelo MA (2004) Técnicas de Investigación Social. Relaciones Laborales. Unidixital, Santiago de Compostela.

[10] de la Fuente A (2020) The economic consequences of Covid in Spain and how to deal with them. Emerald Publishing Limited, Fundación de Estudios de Economía Aplicada, Madrid.

[11] Härdle W, Müller M, Sperlich S, Werwatz A (2004) Nonparametric and Semiparametric Models. Springer, Berlín.

[12] Hollander M, Wolfe AD, Chicken E (2015) Nonparametric Statistical Methods. Wiley, Novo Jersey.

[13] Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2020) forecast: Forecasting functions for time series and linear models. R package version 8.12, URL: http://pkg.robjhyndman.com/forecast.

[14] https://www.ine.es/.

[15] Jarque CM, Bera AK (1987) A Test for Normality of Observations and Regression Residuals. International Statistical Review, Vol. 55, No. 2: 163-172.

[16] Johnson RA, Wichern DW (2007) Applied Multivariate Statistical Analysis. Pearson Education, Nova Jersey.

[17] Lourenço N, Rua A (2020) The DEI: tracking economic activity daily during the lockdown. Working Paper, Banco de Portugal, Lisboa.

[18] Peña D (2005) Análisis de series temporales. Alianza Editorial, Madrid.

[19] Peña D, Romo J (1997) Introducción a la Estadística para las Ciencias Sociales. McGraw-Hill, Madrid.

[20] Pizarro M (2020) climaemet (R Climate AEMET Tools). URL: https://CRAN.R-project.org/package=climaemet.

[21] R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[22] https://www.ree.es/es.

[23] Ruppert D, Sheather SJ, Wand MP (1995) An effective bandwidth selector for local least squares regression. Journal of the American Statistical Association, No. 90: 1257-1270.

[24] Ryan JA, Ulrich JM (2020) xts: eXtensible Time Series. R package version 0.12.1. URL: https://CRAN.R-project.org/package=xts.

[25] Santiago I, Moreno-Munoz A, Quintero-Jiménez P, Garcia-Torres F, Gonzalez-Redondo MJ (2021) Electricity demand during pandemic times: The case of the COVID-19 in Spain. Energy Policy, Vol. 148.

[26] Sax C, Steiner P (2013) Temporal Disaggregation of Time Series. The R Journal, Vol. 5, No. 2: 80-87. URL: https://doi.org/10.32614/RJ-2013-028.

[27] Shapiro SS, Wilk MB (1965) An Analysis of Variance Test for Normality (Complete Samples). Biometrika, Vol. 52, No. 3/4: 591-611.

[28] Trapletti A, Hornik K (2019) tseries: Time Series Analysis and Computational Finance. R package version 0.10-47. URL: https://CRAN.R-project.org/package=tseries.

[29] Vanderkam D, Allaire JJ, Owen J, Gromer D, Thieurmel B (2018) dygraphs: Interface to 'Dygraphs' Interactive Time Series Charting Library. R package version 1.1.1.6. URL: https://CRAN.R-project.org/package=dygraphs.

[30] Venables WN, Ripley BD (2002) Modern Applied Statistics with S. Fourth Edition. Springer, Nova York.

[31] Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, Nova York.

[32] Wickham H (2021) Mastering Shiny. Build Interactive Apps, Reports and Dashboards Powered by R. O'Reilly Media, USA.

[33] Wickham H, François R, Henry L, Müller K (2021) dplyr: A Grammar of Data Manipulation. R package version 1.0.4. URL: https://CRAN.R-project.org/package=dplyr.

[34] Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society, Vol. 73: 3-36.