UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

ANÀLISI D'INTERVENCIÓ EN SÈRIES TEMPORALS: LA METODOLOGIA DE CHEN I LIU

Autor: Álvaro Martínez López

Director: Dr. Josep Vives
Realitzat a: Departament de Matemàtiques i Estadística

Barcelona, 20 de juny de 2021

Abstract

Explanation of the methodology of C. Chen and L.M. Liu for the detection of outliers in time series and application of intervention analysis. Additionally, it includes a prior description of the G.E.P. Box and G. Jenkins methodology for time series analysis with the definition of the four stages it consists of: identification, estimation, validation and prediction. Finally, two examples of the application of these procedures in real time series carried out with the R program are presented.

Resum

Explicació de la metodologia de C. Chen i L.M. Liu per a la detecció d'outliers en sèries temporals i aplicació de l'anàlisi d'intervenció. Addicionalment, inclou una descripció prèvia de la metodologia de G.E.P. Box i G. Jenkins per a l'anàlisi de sèries temporals amb la definició de les quatre etapes en que consisteix: identificació, estimació, validació i predicció. Per últim, s'exposen dos exemples de l'aplicació d'aquests procediments en sèries temporals reals duts a terme amb el programa R.

2020 Mathematics Subject Classification. 37M10, 62M10, 91B84

Agraïments

Vull agrair a tots els meus companys que m'han acompanyat durant aquest camí de cinc anys. Ha sigut un veritable plaer compartir-los amb ells. Això és per ells.

També a totes les respostes que m'han pogut donar les matemàtiques a totes les preguntes que m'he fet. Aquí n'hi han algunes d'elles.

Índex

1 Introducció

El projecte

Les sèries temporals han estat objecte de numerosos estudis en els darrers anys. El plantejament de tenir certes observacions distribuïdes al llarg del temps ofereix la possibilitat de poder predir situacions específiques en temps futurs. Això fa que aquesta branca de l'estadística prengui una importància vital en temps d'incertesa.

La situació sanitària actual ens brinda una oportunitat inigualable per a treballar amb sèries temporals. La dificultat per saber que pot arribar a passar el dia de demà converteix l'objecte d'aquest treball en una arma de valor incalculable no només per a la comunitat científica, sino per a la ciutadania al complet.

El plantejament del treball serà utilitzar la metodologia de Box-Jenkins per a l'anàlisi de sèries temporals proposada per George E. P. Box i Gwilym Jenkins afegint els coneixements proposats per George E.P. Box i George C. Tiao sobre anàlisi d'intervenció i relacionant-los amb la teoria establerta per Chung Chen i Lon-Mu Liu sobre els outliers a les sèries temporals.

A més, es veurà l'aplicació pràctica de tota aquesta teoria sobre dues sèries temporals corresponents al grau d'ocupació hotelera i l'atur a Catalunya i s'evaluaran els resultats per tal d'estudiar la capacitat real de la metodologia proposada.

Estructura de la Memòria

La memòria seguirà una estructura basada en l'anàlisi de Box-Jenkins. En primer lloc, es definiran alguns conceptes previs per a introduïr al lector en la matèria. A continuació, s'exposarà l'anàlisi clàssica de sèries temporals que servirà com a punt de partida de la metodologia de Box-Jenkins, ja que amb aquest es podran transformar les observacions inicials. Aquestes transformacions es duran a terme per a poder identificar el model com un dels descrits en la posterior secció. Tot seguit, s'explicarà com estimar els paràmetres i seleccionar el model adequat. Les seccions següents correspondran a la validació del model a partir de l'anàlisi dels residus i a la predicció de futures observacions. Per últim, donat que les sèries temporals han patit alguns efectes externs, com per exemple el de la COVID-19, s'explicarà com fer una anàlisi d'intervenció a partir de la detecció d'outliers per a millorar les prediccions. La part final correspondrà a l'aplicació pràctica de la teoria exposada prèviament.

2 Conceptes bàsics

En primer lloc, es portarà a terme la definició d'alguns conceptes per tal de fer una anàlisi posterior de forma clara.

2.1 Processos estocàstics

Definició 2.1. *Un procés estocàstic és una col·lecció de variables aleatòries indexades per un conjunt* $\mathbb{T}$*:*

$$\{X_t, t \in \mathbb{T}\}.$$

En aquest treball, com es fa referència a les sèries temporals, només es centrarà en els processos estocàstics discrets, és a dir on $\mathbb{T} = \mathbb{N}$, i aquest conjunt fa referència al temps.

Pel que fa a la notació, donada una variable aleatòria X, es denotarà per $\mathbb{E}(X)$ l'esperança i $\mathbb{V}(X)$ la variància. Donades dues variables aleatòries X i Y, es referirà a seva la covariància com $\mathbb{C}(X, Y)$.

Definició 2.2. *Un procés estocàstic és de segon ordre quan compleix*

$$\mathbb{E}(X_k^2) < \infty, \quad \forall k \geq 0.$$

Sempre es suposarà que els processos són de segon ordre.

Proposició 2.3. *Un procés estocàstic de segon ordre té una esperança i una variància ben definides i finites.*

Demostració. Per a demostrar que l'esperança està ben definida i és finita s'aplica la desigualtat de Jensen

$$|\mathbb{E}(X_k)| \leq \mathbb{E}|X_k| < \infty.$$

Alternativament, per la variància s'utilitza la desigualtat de Cauchy-Schwarz

$$\mathbb{E}|X_k| \leq (\mathbb{E}(X_k^2))^{1/2} < \infty.$$

Com que $\mathbb{V}(X_k) = \mathbb{E}(X_k^2) - (\mathbb{E}(X_k))^2 < \infty$, queden demostrades ambdues propietats. □

Proposició 2.4. *Per un procés estocàstic de segon ordre, la covàriancia de dues variables aleatòries del procés està ben definida.*

Demostració. S'apliquen les desigualtats de Jensen i de Cauchy-Schwarz

$$\begin{aligned}
|\mathbb{C}(X_j, X_{j+l})| &= |\mathbb{E}((X_j - \mathbb{E}(X_j))(X_{j+l} - \mathbb{E}(X_{j+l})))| \\
&\leq \mathbb{E}(|X_j - \mathbb{E}(X_j)||X_{j+l} - \mathbb{E}(X_{j+l})|) \\
&\leq (\mathbb{E}|X_j - \mathbb{E}(X_j)|^2)^{1/2}(\mathbb{E}|X_{j+l} - \mathbb{E}(X_{j+l})|^2)^{1/2} \\
&= \mathbb{V}(X_j)^{1/2}\mathbb{V}(X_{j+l})^{1/2}.
\end{aligned}$$

□

2.2 Processos estacionaris

Definició 2.5. *Un procés* $\{X_t, t \in \mathbb{Z}\}$ *és estríctament estacionari si per a qualsevols* k_1,...,k_n *i* l*, els vectors*

$$(X_{k_1}, ..., X_{k_n})$$

i

$$(X_{k_1+l}, ..., X_{k_n+l})$$

tenen la mateixa llei.

En particular, per $n = 1$, implica que totes les variables X_k tenen la mateixa llei.

Definició 2.6. *Es defineix la funció d'autocovariància* γ *com*

$$\gamma(l) = \mathbb{C}(X_k, X_{k+l}), \quad \forall k, l \in \mathbb{Z}.$$

La funció d'autocovariància es pot definir en $\mathbb{N}$ ja que $\gamma(-l) = \gamma(l)$. A més, en el cas de $l = 0$

$$\gamma(0) = \mathbb{V}(X_k), \quad \forall k \in \mathbb{Z}.$$

Al valor l se l'anomena *retard*.

Definició 2.7. *Es defineix la funció d'autocorrelació* ρ *com*

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}, \quad \forall k \in \mathbb{Z}.$$

Com $\rho(-l) = \rho(l)$, la funció es pot definir en $\mathbb{N}$.

Definició 2.8. *Un procés és estacionari o estacionari en sentit dèbil quan la seva esperança i variància són constants i la seva funció d'autocovariància depèn només del retard.*

L'estacionarietat estricta implica l'estacionarietat, però l'implicació inversa és generalment falsa. Si el procés és gaussià, això és, que totes les distribucions finites són gaussianes, la estacionarietat en sentit dèbil implica estacionarietat estricta.

A més, per la demostració de la Proposició 2.4 es té

$$|\gamma(l)| \leq \gamma(0), \quad \forall l \geq 0$$

i llavors

$$-1 \leq \rho(l) \leq 1, \quad \forall l \geq 0.$$

La representació gràfica de les autocorrelacions s'anomena *correlograma* i ve donada per

$$\{\rho(l), l \geq 0\}.$$

Aquesta, serà una eina bàsica per a detectar quin és el model que millor s'adapta a les dades.

Per últim, sempre es compleix $\gamma(0) \geq 0$. Si $\gamma(0) = 0$ es té $X_j = \mu, \forall j \geq 1$. Si $\gamma(0) \neq 0$ es té $\rho(0) = 1$.

Es pot concloure que un procés estacionari de segon ordre es pot determinar únicament amb (μ, γ) o (μ, σ, ρ).

Definició 2.9. *Sigui $X = \{X_j, j \in \mathbb{Z}\}$ una sèrie estacionària de segon ordre centrada, es defineix la funció d'autocorrelació parcial com una aplicació*

$$\alpha : \mathbb{N} \to [-1, 1]\,,$$

tal que $\alpha(n) = x_n$ i $\alpha(0) = 1$ i la resta de valors venen donats per la solució del següent sistema

$$R_n \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \rho(1) \\ \vdots \\ \rho(n) \end{bmatrix}$$

on

$$R_n = \begin{bmatrix} \rho(0) & \cdots & \cdots & \rho(n-1) \\ \rho(1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \rho(n-1) & \cdots & \cdots & \rho(0) \end{bmatrix}.$$

A l'observació 6.3, es tornarà a definir el concepte de funció d'autocorrelació d'una forma alternativa amb els nous coneixements adquirits fins aleshores.

2.3 Exemples

Serà necessari introduïr alguns exemples per a la posterior anàlisi de sèries temporals.

- Soroll IID

  Es diu que $\{X_k, k \geq 1\}$ és un soroll i.i.d. si les variables aleatòries són i.i.d. amb mitjana μ i desviació estàndard σ. En aquest cas, $\rho(k) = 1$ si $k = 0$ i $\rho(k) = 0$ si $k > 0$.

- Passeig aleatòri

  Sigui $\{X_k, k \geq 1\}$ un soroll i.i.d.. La sèrie $\{S_k, k \geq 1\}$ amb

  $$S_k = X_1 + ... + X_k, \quad k \geq 1,$$

  s'anomena passeig aleatòri.

- Soroll blanc

  Un procés $\{X_k, k \geq 1\}$ és un soroll blanc si totes les variables tenen esperança μ, variància σ^2 i estan incorrelacionades. Un soroll IID és un cas particular de soroll blanc perquè independència implica incorrelació.

- Soroll blanc gaussià

  Un procés $\{X_k, k \geq 1\}$ és un soroll blanc gaussià si és un soroll blanc i qualsevol vector $(X_{k_1}, ..., X_{k_n})$ segueix una llei normal. En aquest cas, com s'està treballant amb un vector normal, la no correlació implica independència i per tant el procés és també un soroll IID. Per tant un soroll blanc gaussià i un soroll IID gaussià són el mateix objecte. Si les variables aleatòries X_k segueixen una llei normal estàndard, llavors es diu que X és un un soroll blanc gaussià estàndard.

3 Anàlisi clàssica de sèries temporals

El primer pas per analitzar una sèrie temporal és fer la representació gràfica corresponent. A partir d'aquesta es pot treure informació rellevant sobre alguns fenòmens com la tendència, la periodicitat,...

Abans de presentar els models que s'utilitzen per a analitzar sèries temporals, s'exposarà l'anàlisi clàssica o macroscòpic de les mateixes. Aquest s'utilitza previament a definir el model definitiu que s'assignarà a la sèrie.

Es proposa que les sèries estan formades per les següents components:

- Tendència (T): fa referència al comportament o moviment a llarg termini. Pot ser creixent o decreixent.
- Cicles ($C^{(j)}$): són els comportaments recurrents dins la série temporal al llarg d'un període. Aquest període és variable i pot referir-se als anys, als mesos o a les setmanes per exemple. En el cas de ser un cicle anual, assenyalaria que cada mes es produeix un comportament diferent.
- Residus (R): es tracta de la part sobrant un cop s'han eliminat les components que s'han vist prèviament.

Amb aquesta descripció es pot expressar una sèrie de la següent forma:

$$X_i = f(T_i, C_i^{(1)}, ..., C_i^{(k)}, R_i).$$

S'assumeix que els models amb els que es treballaran seran additius, és a dir

$$X_i = T_i + C_i^{(1)} + ... + C_i^{(k)} + R_i.$$

En el cas de que el model que està sent analitzat no segueixi un comportament additiu, aplicant la transformació logarítmica, s'obté un que si ho és. Per tant, aquest aspecte no suposa una gran problemàtica.

La idea general de l'anàlisi és descriure la tendència i els cicles per poder aïllar-los fins que només es distingeixi el component residual. Per a poder finalitzar l'anàlisi, aquest residus han de ser estacionaris, sent més precisos, no han de mostrar cap patró cíclic o estacional ni cap variabilitat creixent ni decreixent.

Existeixen diferents mètodes per aïllar les diferents components, en aquest treball s'exposaran els de diferenciació ja que són els que s'utilitzaran durant l'anàlisi de les sèries. La idea dels mètodes és generar una nova sèrie a partir de l'original, mitjançant la diferenciació, en la que no estigui present la component que fa referència a la tendència o els cicles.

3.1 Tractament de la tendència

En primer lloc, es consideraran els operadors

$$Bx_i = x_{i-1}$$

i

$$\nabla x_i = (Id - B)x_i = x_i - x_{i-1}.$$

Es denotarà la sèrie $\{y_i, i \geq 1\}$ amb $y_i = \nabla x_i$ com la sèrie de les primeres diferències de $\{x_i, i \geq 0\}$ amb la que es pretén eliminar la tendència.

L'operador ∇ es pot aplicar de forma recursiva els cops que sigui necessari per aïllar la tendència obtenint

$$\nabla^j x_i = (Id - B)^j x_i, \quad i \geq 1.$$

Aquest mètode és reversible ja que

$$x_i = x_i - x_{i-1} + x_{i-1} - x_{i-2} + ... + x_1 - x_0 + x_0$$

i per tant

$$x_i = x_0 + \sum_{l=1}^{i} y_l.$$

3.2 Tractament dels cicles

En aquest cas, s'aplica l'operador

$$\nabla_p x_i = (Id - B^p)x_i = x_i - x_{i-p}$$

per obtenir la nova sèrie sense el comportament estacional. De la mateixa forma que en el tractament de la tendència, l'operador es pot aplicar de forma recursiva tants cops com sigui necessàri.

3.3 Transformació logarítmica de les dades

Sovint és necessàri dur a terme una transformació de les dades inicials previament a la seva anàlisi. La transformació més habitual és la logarítmica:

$$y_i = \log x_i, \quad i \geq 1.$$

La transformació es pot aplicar per diferents motius:

1. Transformar els productes dels components de la sèrie en sumes.
2. Transformar les dades positives en dades reals.
3. Neutralitzar les tendències exponencials.
4. Suavitzar la sèrie reduint la variabilitat de les observacions.

4 Tipus de models

Un cop finalitzada l'anàlisi clàssica de la sèrie, el següent pas és ajustar un model a les dades. En aquest capítol es definiràn els models amb els que es treballaran les diferents sèries temporals. Per als models autorregressius i de mitjana mòbil s'enunciaran com a propietats els diferents resultats de l'esperança, la variància, la funció d'autovariància, la funció d'autocorrelació i la funció d'autocorrelació parcial; per a la resta de models s'indicarà posteriorment una forma generalitzada de fer-ho.

4.1 Models autorregressius (AR)

Definició 4.1. *Donat un soroll blanc $Z = \{Z_j, j \in \mathbb{Z}\}$ centrat amb variància $\sigma^2 > 0$ i un nombre natural $p \geq 0$, es defineix un model autorregressiu de dimensió p (AR(p)) com un procés $\{Y_j, j \in \mathbb{Z}\}$ tal que*

$$Y_j = \phi_1 Y_{j-1} + ... + \phi_p Y_{j-p} + Z_j \tag{4.1}$$

on $\phi_1, ..., \phi_p$ són nombres reals.

Observació 4.2. Si es defineix un polinomi

$$\Phi(x) = 1 - \phi_1 x - \phi_2 x^2 - ... - \phi_p x^p$$

i s'utilitza l'operador de retards B definit prèviament, es pot redefinir l'equació del procés com

$$\Phi(B) Y_j = Z_j. \tag{4.2}$$

Definició 4.3. *Un polinomi $\Phi(x)$ és invertible si existeix una sèrie integrable quadràticament $\sum_{i=0}^{\infty} \psi_i x^i$ tal que*

$$\Phi(x) \sum_{i=0}^{\infty} \psi_i x^i = 1. \tag{4.3}$$

Aleshores, si $\Phi(x)$ és invertible es té

$$Y_j = \Phi(B)^{-1} Z_j = \sum_{i=0}^{\infty} \psi_i B^i Z_j = \sum_{i=0}^{\infty} \psi_i Z_{j-i} \tag{4.4}$$

i això serà el que garanteixi la estacionarietat de Y. Llavors s'ha d'estudiar quan un polinomi és invertible.

Teorema 4.4. *Un polinomi $\Phi(z)$ amb $z \in \mathbb{C}$ és invertible si i només si totes les seves arrels estan fora del cercle unitat, és a dir*

$$\{z : \Phi(z) = 0\} \subseteq \{z : |z| > 1\}.$$

Demostració. La demostració es troba a les pàgines 85 i 86 de [1]. □

Propietats 4.5. *Les propietats principals del model són les següents*

1. *Es tracta d'un procés de segon ordre perquè Z ho és.*
2. *És un procés centrat perquè Z és centrat i*

$$\mathbb{E}(Y_j) = \sum_{i=0}^{\infty} \psi_i \mathbb{E}(Z_{j-i}) = 0.$$

3. *Sigui $l \geq 0$, la funció d'autocovariància ve donada per*

$$\begin{aligned}
\gamma(l) = \mathbb{C}(Y_j, Y_{j+l}) &= \mathbb{C}(\sum_{r=0}^{\infty} \psi_r Z_{j-r}, \sum_{s=0}^{\infty} \psi_s Z_{j+l-s}) \\
&= \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \psi_r \psi_s \mathbb{C}(Z_{j-r}, Z_{j+l-s}) \\
&= \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \psi_r \psi_s \sigma^2 \mathbb{1}_{\{j-r=j+l-s\}} \\
&= \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \psi_r \psi_s \sigma^2 \mathbb{1}_{\{s=r+l\}} \\
&= \sum_{r=0}^{\infty} \psi_r \psi_{r+l} \sigma^2.
\end{aligned} \tag{4.5}$$

Per tant, per calcular $\gamma(l)$ es necessita conèixer els coeficients ψ_i. Per a fer-ho es recupera en primer lloc l'equació de la definició d'invertibilitat i es té

$$\begin{aligned}
&(1 - \phi_1 x - \phi_2 x^2 - ... - \phi_p x^p) \sum_{i=0}^{\infty} \psi_i x^i = 1 \\
\iff &\sum_{i=0}^{\infty} \psi_i x^i - \phi_1 \sum_{i=0}^{\infty} \psi_i x^{i+1} - ... - \phi_p \sum_{i=0}^{\infty} \psi_i x^{i+p} = 1 \\
\iff &\sum_{i=0}^{\infty} \psi_i x^i - \phi_1 \sum_{i=1}^{\infty} \psi_{i-1} x^i - ... - \phi_p \sum_{i=p}^{\infty} \psi_{i-p} x^i = 1.
\end{aligned}$$

Llavors s'arriba al següent sistema d'equacions:

$$\begin{aligned}
\psi_0 &= 1 \\
\psi_1 - \psi_0 \phi_1 &= 0 \\
\psi_2 - \psi_1 \phi_1 - \psi_0 \phi_2 &= 0 \\
&... \\
\end{aligned}$$

Per tant, en general

$$\psi_i = \phi_1 \psi_{i-1} + ... + \phi_p \psi_{i-p}, \quad i \geq p.$$

Amb aquest resultats es pot calcular la funció d'autocovariància per a tot l i al no dependre de j es pot concloure que aquest és un procés estacionari. Per tant, com s'ha esmentat prèviament, la condició per a que es tingui un procès $AR(p)$ estacionari és que el polinomi $\Phi(x)$ sigui invertible.

4. *Per a calcular la funció d'autocorrelació es segueix el següent procediment. En primer lloc, es pot representar la funció d'autocovariància com*

$$\begin{aligned}\gamma(l) &= \mathbb{C}(Y_j, \phi_1 Y_{j+l-1} + ... + \phi_p Y_{j+l-p} + Z_{j+l}) \\ &= \phi_1\gamma(l-1) + \phi_2\gamma(l-2) + ... + \phi_p\gamma(l-p), \quad l > 0.\end{aligned}$$

Es divideix per $\gamma(0)$ i s'obtenen les equacions conegudes com Equacions de Yule-Walker

$$\begin{aligned}\rho(l) &= \phi_1\rho(l-1) + ... + \phi_p\rho(l-p), \quad l > 0 \\ \rho(0) &= 1 \\ \rho(-l) &= \rho(l).\end{aligned} \tag{4.6}$$

D'aquest sistema s'obtenen els resultats de les funcions d'autocorrelació.

5. *Per a la funció d'autocorrelació parcial es té la següent propietat*

$$\alpha(k) = 0, \quad \forall k \geq p+1. \tag{4.7}$$

4.2 Models de mitjana mòbil (MA)

Definició 4.6. *Donat un soroll blanc Z amb variància σ^2 i un nombre natural $q \geq 0$, es defineix un model MA(q) com un procés $\{Y_j, j \in \mathbb{Z}\}$ tal que*

$$Y_j = Z_j + \theta_1 Z_{j-1} + ... + \theta_q Z_{j-q} \tag{4.8}$$

on $\theta_1, ..., \theta_q$ són nombres reals.

Propietats 4.7. *Les propietats principals del model són les següents*

1. *Es tracta d'un procès de segon ordre perquè Z_j ho és.*
2. *És un procés centrat perquè Z és centrat i*

$$\mathbb{E}(Y_j) = \sum_{l=0}^{q} \theta_l \mathbb{E}(Z_{j-l}) = 0. \tag{4.9}$$

3. *Sigui $k \geq 0$, la funció d'autocovariància ve donada per*

$$\begin{aligned}\gamma(k) = \mathbb{C}(Y_j, Y_{j+k}) &= \mathbb{C}(\sum_{r=0}^{q} \theta_r Z_{j-r}, \sum_{s=0}^{q} \theta_s Z_{j+k-s}) \\ &= \sum_{r=0}^{q}\sum_{s=0}^{q} \theta_r\theta_s \mathbb{C}(Z_{j-r}, Z_{j+k-s}) \\ &= \sum_{r=0}^{q}\sum_{s=0}^{q} \theta_r\theta_s\sigma^2 \mathbb{1}_{\{j-r=j+k-s\}} \\ &= \sum_{r=0}^{q}\sum_{s=0}^{q} \theta_r\theta_s\sigma^2 \mathbb{1}_{\{s=r+k\}} \\ &= \sum_{r=0}^{q-k} \theta_r\theta_{r+k}\sigma^2.\end{aligned} \tag{4.10}$$

Per tant és un procés estacionari ja que l'autocovariància entre Y_j i Y_{j+k} no depèn de j. A més quan $k > q$ llavors $\gamma(k) = 0$.

4. *La variància per tant és igual a*

$$\mathbb{V}(Y_j) = \mathbb{C}(Y_j, Y_j) = \sigma^2 \sum_{r=0}^{q} \theta_r^2, \quad \forall j \in \mathbb{Z}. \tag{4.11}$$

5. *La funció d'autocorrelació ve donada per*

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\sum_{r=0}^{q-k} \theta_r \theta_{r+k}}{\sum_{r=0}^{q} \theta_r^2}, \quad 0 \geq k \geq q. \tag{4.12}$$

   Aleshores no és nul·la únicament pels valors $k = 0, 1, ..., q$.

6. *La funció d'autocorrelació parcial dels models MA(q) és decreixent a mesura que augmenta el retard* l.

Observació 4.8. La representació del model a partir de l'operador de retard B és

$$Y_j = Z_j + \theta_1 B Z_j + \theta_2 B^2 Z_j + ... + \theta_q B^q Z_j, \tag{4.13}$$

i sigui $B^0 = Id$ es pot escriure

$$Y_j = \Theta_q(B) Z_j \tag{4.14}$$

on

$$\Theta_q(x) = 1 + \theta_1 x + \theta_2 x^2 + ... + \theta_q x^q \tag{4.15}$$

és un polinomi d'ordre $q \geq 0$.

4.3 Model ARMA

Definició 4.9. *Siguin p i q dos nombres naturals,* $Z \sim WN(0, \sigma^2)$ *i* $\Phi_p(B)$ *i* $\Theta_q(B)$ *dos polinomis invertibles de graus p i q respectivament. Llavors un model ARMA(p,q) és un model que satisfà l'equació*

$$\Phi_p(B) Y_j = \Theta_q(B) Z_j, \quad j \in \mathbb{Z}. \tag{4.16}$$

Observació 4.10. Com Φ_p és un polinomi invertible, el model és causal ja que es pot representar de la següent forma

$$Y_j = \Theta_q(B) \Phi_p(B)^{-1} Z_j. \tag{4.17}$$

Definició 4.11. *Un procés* $\{Y_j, j \in \mathbb{Z}\}$ *és un model ARMA amb mitjana* μ *si*

$$X_j = Y_j - \mu, \quad j \in \mathbb{Z} \tag{4.18}$$

és un model ARMA.

4.4 Model ARIMA

Sigui $\{X_j, j \geq 0\}$ un passeig aleatori definit com

$$X_j = X_{j-1} + Z_j \tag{4.19}$$

on $Z \sim WN(0, \sigma^2)$ i $X_0 = Z_0 = 0$. Aquest no és un procés estacionari perquè

$$\mathbb{V}(X_j) = \mathbb{V}(X_{j-1}) + \sigma^2 = \mathbb{V}(X_{j-2}) + 2\sigma^2 = ... = j\sigma^2. \tag{4.20}$$

Però generant la sèrie amb les primeres diferències s'obté

$$Y_j = X_j - X_{j-1} = Z_j, \quad \forall j \geq 1, \tag{4.21}$$

que en aquest cas si és una sèrie estacionaria.

Aquesta és la idea principal per a entendre els models ARIMA.

Definició 4.12. *Es diu que* $\{X_j, j \geq 0\}$ *és un model ARIMA(p,d,q) amb* $p, d, q \in \mathbb{N}$ *si*

$$Y_j = (Id - B)^d X_j, \quad j \in \mathbb{Z} \tag{4.22}$$

es un model ARMA(p,q).

Per tant, relacionant la definició amb el model presentat a l'equació (4.21), la variable aleatòria $\{Y_j, j \geq 0\}$, es tractaria d'un model ARIMA(0,1,0).

Exemple 4.13. Amb la notació presentada pels models ARMA, un model ARIMA(p,d,q) es podria representar de la següent forma

$$\Phi_p(B)(Id - B)^d X_j = \Theta_q(B) Z_j, \quad j \in \mathbb{Z}. \tag{4.23}$$

Observació 4.14. Lògicament, un model ARIMA(p,0,q) és un model ARMA(p,q) i un passeig aleatòri és un model ARIMA(0,1,0).

4.5 Model SARIMA

Sigui $\{X_j, j \geq 1\}$ una sèrie temporal representada com $X_j = X_{i+(r-1)s}$ amb $i = 1, ..., s$ i $r \geq 1$ on s és el període. Per exemple, en una sèrie de dades mensuals, $s = 12$ i r i i indicarien l'any i el mes de l'observació respectivament.

Un model SARIMA està format per una part regular i una part estacional pura. En aquesta última s'assumeix que només existeix correlació entre les dades del mateix mes. Per tant, en cas de treballar amb sèries mensuals, seria com treballar amb 12 sèries diferents incorrelacionades, una per mes.

Definició 4.15. *La part estacional pura d'un model SARIMA correspon a un model ARMA(a,c), és a dir*

$$X_{i+(r-1)s} - \alpha_1 X_{i-(r-2)s} - ... - \alpha_a X_{i+(r-a-1)s} = Z_{i+(r-1)s} - \beta_1 Z_{i+(r-2)s} - ... - \beta_c Z_{i+(r-c-1)s}$$

amb $\{Z_{i+(r-1)s, r \geq 1, i=1,...,s}\} \sim WN(0, \sigma^2)$*. Aquest model es denota com*

$$ARMA(a, c)_s \times WN(0, \sigma^2).$$

Exemple 4.16. Un model $\text{AR}(1)_{12} \times \text{WN}(0,\sigma^2)$ satisfà

$$X_{i+(r-1)12} - \alpha X_{i+(r-2)12} = Z_{i+(r-1)12}, \quad |\alpha| < 1,$$

o equivalentment,

$$X_j - \alpha X_{j-12} = Z_j, \quad |\alpha| < 1.$$

Observació 4.17. El càlcul d'esperances, variàncies, funcions d'autovariància, autocorrelació i autocorrelació parcial són exactament iguals que els de qualsevol altre model ARMA.

Observació 4.18. El polinomis amb l'operador retard B són de la següent forma

$$A_a(B) = 1 - \varphi_1 B^s - \varphi_2 B^{2\cdot s} - ... - \varphi_a B^{a\cdot s} \tag{4.24}$$

$$C_c(B) = 1 - \vartheta_1 B^s - \vartheta_2 B^{2\cdot s} - ... - \vartheta_c B^{c\cdot s}. \tag{4.25}$$

Definició 4.19. *Utilitzant la notació dels models vista anteriorment, un model*

$$SARIMA(p,d,q)(a,D,c)_s$$

es representa amb la següent equació

$$\Phi_p(B)A_a(B^s)(Id - B)^d(Id - B^s)^D X_j = \mu + \Theta_q(B)C_c(B^s)Z_j, \tag{4.26}$$

per tant, el model és equivalent a un ARIMA(p,d,q) × ARIMA(a,D,c)$_s$ on la primera part faria referència a la part regular, i la segona a la part estacional.

Observació 4.20. Per a que el model sigui causal, és a dir, que es pugui aïllar X_j respecte a a resta d'elements del model, els polinomis $\Phi_p(B)$ i $A_a(B^s)$ han de ser invertibles.

4.6 Selecció del millor model

El criteri de selecció del millor model es basarà en el criteri d'Akaike. La idea principal és escollir el model que tingui un AIC (Crtieri d'Informació d'Akaike) menor. El càlcul d'aquest, mostra una estimació relativa de l'informació perduda amb el model sobre les dades inicials. Per tant, el millor model serà aquell que millor s'ajusti a les dades, penalitzant el sobreajustament dels paràmetres. El càlcul és el següent:

$$AIC = 2k - 2ln\hat{L}, \tag{4.27}$$

on $\hat{L}$ és el màxim valor de la funció de versemblança aplicada a les estimacions de màxima versemblança dels paràmetres del model i k fa referència al nombre de paràmetres d'aquest.

En el cas de les sèries temporals, com es tracten d'una mostra finita d'observacions, es sol utilitzar la següent correcció de l'AIC, la qual denotem com AICc:

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1}, \tag{4.28}$$

on n fa referència al nombre total d'observacions de la mostra.

Aleshores, els càlculs de l'AIC i l'AICc per a un model ARMA(p,q) es farien de la següent forma:

$$AIC = -2ln\hat{L} + 2(p + q + 1),$$
$$AICc = AIC + \frac{2(p + q + 1)^2 + 2(p + q + 1)}{n - p - q - 2}$$

on

$$\hat{L} = L(y_1, ..., y_n, \hat{\mu}, \hat{\sigma}^2, \hat{\phi}_1, ..., \hat{\phi}_q, ..., \hat{\theta}_1, ..., \hat{\theta}_q)$$

i L és la funció de versemblança del model estimat.

4.7 Estimació de paràmetres

A continuació es mostrarà la forma generalitzada d'estimar els diferents paràmetres dels models que s'han vist. Es suposarà que tenim un procès $\{Y_j, j \geq 0\}$ amb la seva corresponent representació causal

$$Y_j = \mu + \sum_{i=0}^{\infty} \psi_i Z_{j-i} \tag{4.29}$$

1. Esperança o mitjana
$$\bar{Y}_n = \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i. \tag{4.30}$$

2. Variància
$$\hat{\sigma}^2 = \hat{\gamma}(0) = \sum_{j=1}^{n} (Y_j - \bar{Y}_n)^2. \tag{4.31}$$

3. Funció d'autocovariància
$$\hat{\gamma}(l) = \sum_{j=1}^{n-l} (Y_j - \bar{Y}_n)(Y_{j+l} - \bar{Y}_n). \tag{4.32}$$

4. Funció d'autocorrelació
$$\hat{\rho}(l) = \frac{\hat{\gamma}(l)}{\hat{\gamma}(0)}. \tag{4.33}$$

Per estimar els paràmetres $\phi_1, ..., \phi_p$ i $\theta_1, ..., \theta_q$ es partirà de la següent representació del procés, suposant que sigui centrat,

$$Z_j = \sum_{i=0}^{\infty} \beta_i Y_{j-i}. \tag{4.34}$$

Llavors per recursivitat es pot definir

$$\begin{aligned} Z_0 &= \beta_0 Y_0 \\ Z_1 &= \beta_1 Y_1 + \beta_0 Y_0 \\ &\vdots \\ Z_n &= \beta_n Y_n + ... + \beta_1 Y_1 + \beta_0 Y_0. \end{aligned} \tag{4.35}$$

S'han de seleccionar els paràmetres adequats que pertanyin a l'interval (-1,1) seguint la recursivitat prèvia tal que minimitzin

$$S(\phi_1, ..., \phi_p, \theta_1, ..., \theta_q) = \sum_{j=1}^{n} Z_j^2. \tag{4.36}$$

En el cas de ser un model SARIMA, es seguiria el mateix procediment tenint en compte que en aquest cas s'afegirien els paràmetres $\varphi_1, ..., \varphi_a$ i $\vartheta_1, ..., \vartheta_c$ a estimar.

5 Tractament dels residus

Modelitzada la sèrie, és necessari validar el model que s'ha pres com a definitiu. Aquesta validacó es centra en verificar algunes condicions sobre els residus del mateix. Llavors, l'anàlisi finalitzarà quan s'hagi comprovat que aquests són un soroll i.i.d. i segueixen una llei determinada.

5.1 Test d'hipòtesi del soroll IID

Teorema 5.1. *Siguin $y_1, y_2, ..., y_n$ observacions d'un soroll i.i.d., sigui*

$$\{\hat{\rho}(j), 1 \leq j \leq h\}$$

el correlograma corresponent fins al retard h. Llavors per una n suficientment gran i una h relativament petita, els valors $\hat{\rho}(j)$ poden ser considerats com una mostra aleatòria de $N(0, \frac{1}{n})$.

Demostració. La demostració es troba a la pàgina 235 de [1]. □

Llavors si la hipòtesi d'i.i.d. és certa, el 95% de les h autocorrelacions observades han d'estar dins l'interval

$$\left[-1.96\frac{1}{\sqrt{n}}, 1.96\frac{1}{\sqrt{n}}\right].$$

Observació 5.2. Si es té una sèrie temporal amb longitud n, és comunament acceptat que els valors de $\hat{\rho}$ fins al retard d'ordre $\frac{n}{3}$ són significatius.

Un mètode més precís basat en la idea prèvia per a saber si una sèrie temporal és un soroll i.i.d. és el test de Ljung-Box(1978). Aquest es basa en l'estadístic

$$Q_{LB} = n(n+2)\sum_{j=1}^{h}\frac{\rho^2(j)}{n-j},$$

on h està prèviament determinada i $Q_{LB}(h) \sim \chi^2_h$ on χ^2_h denota la distribució chi-quadrat amb h graus de llibretat.

Si Q_{LB} és gran, es rebutja la hipòtesi de soroll i.i.d.. Si s'ha obtingut $Q_{LB} = q$,

$$p = P\{Q_{LB} > q\} = P\{\chi^2_h > q\}$$

és el p-valor o risc de rebutjar la hipòtesi quan es certa. Generalment, si $p < 0.05$, es rebutja que la sèrie sigui un soroll i.i.d..

5.2 Test gaussianitat

Un cop se sap que els residus de la sèrie temporal són un soroll i.i.d., es interessant conèixer si són un soroll blanc gaussià. La forma tradicional de fer-ho és amb el test de de Kolmogorov-Smirnov. L'alternativa que s'exposarà i s'utilitzarà per a dur a terme l'anàlisi són el Q - Q-plot i el test de Shapiro-Wilk.

Sigui $Y_1, ..., Y_k$ una mostra d'una llei $N(\mu, \sigma^2)$ i sigui $X_1, ..., X_n$ una mostra d'una llei $N(0, 1)$. Es defineixen $Y_{(1)}, ..., Y_{(k)}$ i $X_{(1)}, ..., X_{(k)}$ com els corresponents estadístics d'ordre i és denoten com $y_{(i)}$ i $x_{(i)}$ les corresponents observacions empíriques.

En primer lloc es té

$$\mathbb{E}(Y_{(i)}) = \mu + \sigma\mathbb{E}(X_{(i)}).$$

Sigui $m_i = \mathbb{E}(X_{(i)})$, es defineix la parella

$$(m_i, y_{(i)}),$$

el gràfic de la qual es coneix com Q - Q-plot i hauria de ser aproximadament lineal. En particular, la correlació entre els dos elements hauria de ser propera a 1. Aquesta és la base del test de Shapiro-Wilk que s'exposarà a continuació.

El valor de m_i s'aproxima per

$$m_i \sim \Phi^{-1}(\frac{i - 0.5}{n}),$$

on Φ és la probabilitat acumulada de la distribució normal estàndard.

A més es té

$$R^2 = \frac{(\sum(Y_{(i)} - \bar{Y}_n)\Phi^{-1}(\frac{i-0.5}{n}))^2}{\sum(Y_{(i)} - \bar{Y}_n)^2 \sum(\Phi^{-1}(\frac{i-0.5}{n}))^2} \in [0, 1]$$

i el p-valor és la probabilitat

$$P\{R^2 < r^2\},$$

on r és la correlació observada. Per tant, si el p-valor és inferior a 0.05 rebutjem la hipòtesi de normalitat.

La distribució de R^2 es coneix com la distribució de Shapiro-Wilk.

6 Prediccions de les sèries temporals

L'últim pas de l'anàlisi seria predir els valors futurs de la sèrie temporal aplicant tota la part prèvia que s'ha dut a terme.

6.1 Predicció òptima i predicció lineal òptima

Es suposa que es tenen n variables $X_1, ..., X_n$. Es pretén fer el càlcul d'una variable aleatòria Y, que en el cas de les sèries temporals correspondria a X_{n+p} amb $p \in \mathbb{N}$. Per a fer-ho es busca una funció

$$f(X_1, ..., X_n),$$

tal que

$$\mathbb{E}\left[(Y - f(X_1, ..., X_n))^2\right]$$

sigui mínim.

Com únicament es disposa de les variables $X_1, ..., X_n$, es definirà la funció anterior com

$$f(X_1, ..., X_n) = \mathbb{E}\left[Y|X_1, ..., X_n\right],$$

aquesta funció s'anomena predictor òptim de Y en termes de $X_1, ..., X_n$. Com el càlcul d'una esperança condicionada pot comportar certes dificultats, es considerarà la funció f com

$$f(X_1, ..., X_n) = b_0 + b_1X_1 + ... + b_nX_n.$$

Per tant, el problema es redueix a trobar els coeficients $\hat{b}_0, \hat{b}_1, ..., \hat{b}_n$ tal que

$$\mathbb{E}\left[(Y - b_0 - b_1X_1 - ... - b_nX_n)^2\right]$$

sigui mínim. Aquesta nova solució s'anomena predictor lineal òptim de Y en termes de $X_1, ..., X_n$ i es representa com

$$P(Y|X_1, ..., X_n) = \hat{b}_0 + \hat{b}_1X_1 + ... + \hat{b}_nX_n.$$

Lògicament, amb aquests conceptes introduïts, els errors de predicció serien

$$\mathbb{E}\left[(Y - \mathbb{E}\left[Y|X_1, ..., X_n\right])^2\right]$$

o

$$\mathbb{E}\left[(Y - P(Y|X_1, ..., X_n))^2\right].$$

Definició 6.1. *En general, l'expressió del predictor lineal òptim d'una variable aleatòria Y en termes d'una altra variable aleatòria X és la següent*

$$P(Y|X) = \mu_Y + \frac{\sigma_{XY}}{\sigma_X^2}(X - \mu_X). \tag{6.1}$$

6.2 Predicció multivariant lineal

Es tenen les mateixes condicions inicials que a la secció prèvia amb les variables aleatòries $X_1, ..., X_n$ i Y. En aquest cas es parteix de la següent igualtat deduïda a partir dels resultats de l'anterior secció

$$b_0 + \mathbb{E}[X_1]\, b_1 + ... + \mathbb{E}[X_n]\, b_n = \mathbb{E}[Y], \tag{6.2}$$

a partir d'aquesta, es dissenya el següent sistema d'equacions

$$\begin{aligned} \mathbb{E}[X_1]\, b_0 + \mathbb{E}\left[X_1^2\right] b_1 + ... + \mathbb{E}[X_1 X_n] &= \mathbb{E}[X_1 Y] \\ &\vdots \\ \mathbb{E}[X_n]\, b_0 + \mathbb{E}[X_n X_1]\, b_1 + ... + \mathbb{E}\left[X_n^2\right] &= \mathbb{E}[X_n Y]. \end{aligned} \tag{6.3}$$

Si es denota $b^t = (b_1, ..., b_n)$, $\mu_x^t = (\mathbb{E}[X_1], ..., \mathbb{E}[X_n])$ i $\mu_Y = \mathbb{E}[Y]$,podem reescriure la igualtat (6.2) com

$$b_0 = \mu_Y - b^t \mu_X$$

i multiplicant per $\mathbb{E}[X_j]$,

$$\mathbb{E}[X_j]\, b_0 = \mathbb{E}[X_j]\, \mu_Y - \mathbb{E}[X_j]\, b^t \mu_X, \quad j = 1, ..., n.$$

Si es substitueix en el sistema (6.3) $\mathbb{E}[X_j]\, b_0$ per l'última igualtat s'obté

$$\mathbb{C}(X_j, X_1) b_1 + ... + \mathbb{C}(X_j, X_n) b_n = \mathbb{C}(X_j, Y), \quad j = 1, ..., n,$$

que es representa amb la següent notació

$$\Sigma b = \mathbb{C}(Y, X), \tag{6.4}$$

on Σ és la matriu de covariàncies del vector X.

Definició 6.2. *La fórmula general del predictor lineal òptim de Y en termes de $X_1, ..., X_n$ és la següent*

$$P(Y|X_1, ..., X_n) = (\mu_Y - \hat{b}^t \mu_X) + \hat{b}^t X = \mu_Y + \hat{b}^t (X - \mu_X) \tag{6.5}$$

i a partir de la igualtat (6.4) es té

$$\hat{b} = \Sigma^{-1} \mathbb{C}(Y, X).$$

Llavors, l'error de predicció resultant seria el següent:

$$\begin{aligned} \mathbb{E}\left[(Y - P(Y|X_1, ..., X_n))^2\right] &= \mathbb{E}\left[(Y - \mu_Y - \hat{b}^t (X - \mu_X))^2\right] \\ &= \mathbb{V}(Y) + \mathbb{E}\left[(\hat{b}^t (X - \mu_X))^2\right] - 2\mathbb{E}\left[(Y - \mu_Y)\hat{b}^t (X - \mu_X)\right] \\ &= \mathbb{V}(Y) + \hat{b}^t \Sigma \hat{b} - 2\mathbb{C}(Y, X)^t \hat{b} \\ &= \sigma_Y^2 - \mathbb{C}(Y, X)^t \hat{b}. \end{aligned}$$

6.3 Predicció d'una sèrie temporal

Sigui $X = \{X_j, j \in \mathbb{Z}\}$ una sèrie temporal de segon ordre estacionària amb mitjana μ i funció d'autocovariància γ. Considerem les variables aleatòries $X_1, ..., X_n$ representades com $X = X_1, ..., X_n$. Aplicant els conceptes explicats prèviament, es pretén calcular

$$P(X_{n+1}|X_1, ..., X_n) = \mu + \hat{b}^t(X - \mu) \tag{6.6}$$

amb

$$\hat{b} = \Sigma^{-1}\mathbb{C}(X_{n+1}, X). \tag{6.7}$$

Es tenen les igualtats

$$\mathbb{C}(X_{n+1}, X)^t = (\gamma(n), \gamma(n-1), ..., \gamma(1))$$

i

$$\Sigma = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(n-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \cdots & \cdots & \gamma(0) \end{bmatrix}.$$

Ara, dividint (6.7) entre $\gamma(0)$ s'obté

$$R\hat{b} = \begin{bmatrix} \rho(n) \\ \vdots \\ \rho(1) \end{bmatrix}, \tag{6.8}$$

tenint en compte que

$$R = \begin{bmatrix} \rho(0) & \cdots & \cdots & \rho(n-1) \\ \rho(1) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \rho(n-1) & \cdots & \cdots & \rho(0) \end{bmatrix}.$$

Finalment s'ha obtingut

$$\hat{b} = R^{-1} \begin{bmatrix} \rho(n) \\ \vdots \\ \rho(1) \end{bmatrix},$$

que aplicat a la igualtat (6.6) s'arriba a la conclusió que

$$P(X_{n+1}|X_1, ..., X_n) = \mu + (\rho(n), ..., \rho(1))R^{-1} \begin{bmatrix} X_1 - \mu \\ \vdots \\ X_n - \mu \end{bmatrix}.$$

Observació 6.3. Amb els nous coneixements d'aquesta secció es pot definir la funció d'autocorrelació parcial com una funció tal que

$$\alpha(0) = 1$$
$$\alpha(1) = \rho(1)$$
$$\alpha(l) = \rho(X_1 - P(X_1|X_2, ..., X_l), X_{l+1} - P(X_{l+1}|X_2, ..., X_l)),$$

on donades dues variables aleatòries X i Y

$$\rho(X, Y) = \frac{\mathbb{C}(X, Y)}{[\mathbb{V}(X)\mathbb{V}(Y)]^{1/2}}.$$

Per tant, $\alpha(l)$ fa referència la correlació entre X_1 i X_{l+1} després de descomptar la influència de les variables $X_2, ..., X_l$.

A més, recuperant la definició 2.9 de la mateixa i la igualtat 6.8, podem concloure que

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \hat{b}_n \\ \vdots \\ \hat{b}_1 \end{bmatrix}$$

i s'obté que $\alpha(n) = x_n = \hat{b}_1$.

7 Anàlisi d'intervenció

En alguns casos, després d'aplicar l'anàlisi de Box-Jenkins que s'ha explicat en les seccions anteriors, les prediccions no s'ajusten a la realitat per motius externs a la sèrie temporal. Sovint, es produeixen fets a l'entorn que afecten directament a les observacions futures de la sèrie i provoquen un impacte que canvia la morfologia del procés. Exemples vàlids serien l'impacte de la crisi financera del 2008 o, un més actual, el de la COVID-19. En aquest capítol s'exposarà com tractar aquests fets amb la tècnica ideada per Box i Tiao coneguda com anàlisi d'intervenció. Aquesta buscarà quan es van produïr els primers efectes de l'impacte i com i durant quant temps han afectat els mateixos.

Suposem que es té un model SARIMA $X = \{X_j, j \in \mathbb{Z}\}$ sobre el que es pretén aplicar l'anàlisi. S'afegiran una o diverses sèries deterministes $M = \{M_j, j \in \mathbb{Z}\}$ anomenades intervencions per a generar un nou model amb la següent representació

$$Y_j = X_j + M_j, j \in \mathbb{Z} \tag{7.1}$$

Ara, es procedirà definir els diferents models d'intervenció M que existeixen.

7.1 Models d'intervenció

Existeixen dues variables principals a partir de les quals es generen els diferents models d'intervenció.

Definició 7.1. *La variable esglaó produeix un efecte permanent a partir del temps t_0. Es representa de la següent forma:*

$$S_t^{(t_0)} = \begin{cases} 0, & t < t_0 \\ 1, & t \geq t_0 \end{cases} \tag{7.2}$$

Definició 7.2. *La variable impuls produeix un efecte únic en el temps t_0. Es representa de la següent forma:*

$$P_t^{(t_0)} = \begin{cases} 0, & t \neq t_0 \\ 1, & t = t_0 \end{cases} \tag{7.3}$$

Proposició 7.3. *Les variables esglaó i impuls estan relacionades mitjançant la següent expressió*

$$P_t^{(t_0)} = (1 - B)S_t^{(t_0)}, \tag{7.4}$$

on B és l'operador de retards $BS_t^{(t_0)} = S_{t-1}^{(t_0)}$.

Demostració. S'evalua la igualtat per els casos $t < t_0$, $t = t_0$ i $t > t_0$.

- $t < t_0$
$$P_t^{(t_0)} = S_t^{(t_0)} - S_{t-1}^{(t_0)} = 0 - 0 = 0$$
- $t = t_0$
$$P_t^{(t_0)} = S_t^{(t_0)} - S_{t-1}^{(t_0)} = 1 - 0 = 1$$
- $t > t_0$
$$P_t^{(t_0)} = S_t^{(t_0)} - S_{t-1}^{(t_0)} = 1 - 1 = 0$$

□

Amb cada variable es poden donar diferents models d'intervenció. Per descriure'ls s'agruparan en dos grups principals:

1. El model més simple és aquell que l'impacte és fixe i d'una magnitud específica. Es representa per a les diferents variables com

$$M_t = \omega B^b S_t^{(t_0)} \tag{7.5}$$

o

$$M_t = \omega B^b P_t^{(t_0)} \tag{7.6}$$

on $\omega \in \mathbb{R}$.

2. El model més complex és aquell que l'impacte té un començament gradual. Es representa per a les diferents variables com

$$M_t = \frac{\omega B^b}{1 - \delta B} S_t^{(t_0)} \tag{7.7}$$

o

$$M_t = \frac{\omega B^b}{1 - \delta B} P_t^{(t_0)} \tag{7.8}$$

on $0 \leq \delta \leq 1$. En el cas de $\delta = 0$, els resultats (7.7) i (7.8) es simplificarien respectivament als resultats (7.5) i (7.6). D'altra banda, si $\delta = 1$, el model (7.8) seria equivalent al (7.5) per el resultat obtingut de la proposició 7.3.

Exemple 7.4. Models amb la variable esglaó.

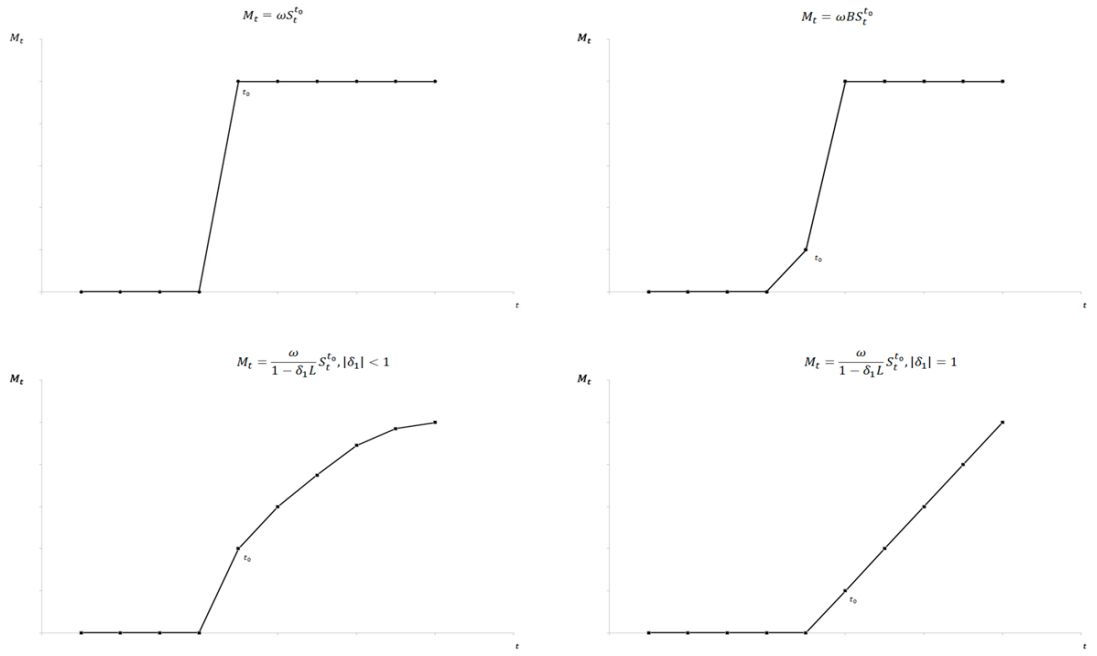

Figura 1: Diferents models amb variable esglaó

Exemple 7.5. Models amb variable impuls.

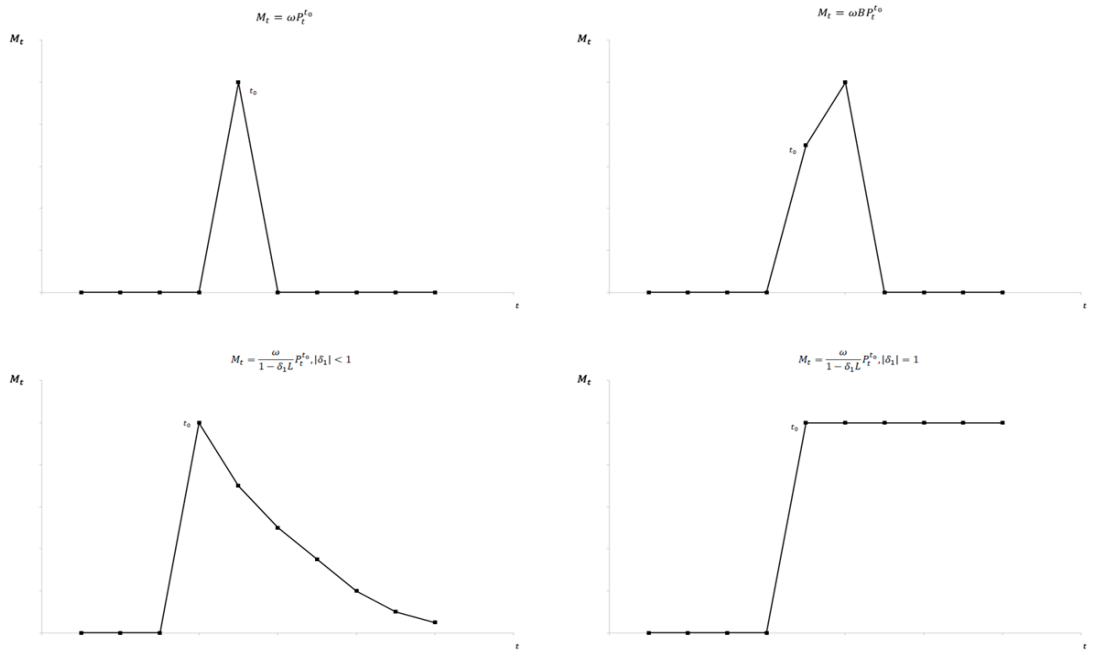

Figura 2: Diferents models amb variable impuls

Observació 7.6. En alguns casos, es pot tindre un procés amb diversos impactes de diferents models. Gràcies al resultat de la proposició 7.3, es poden representar de diferents maneres. A continuació un exemple

$$M_t = \left(\frac{\omega_1 B}{1-\delta B} + \frac{\omega_2 B}{1-B}\right) P_t^{(t_0)} = \left(\frac{\omega_1 B}{1-\delta B} P_t^{(t_0)} + \omega_2 B S_t^{(t_0)}\right).$$

Per tant, com a conclusió, sigui $I_t^{t_0}$ una variable impuls o esglaó, tots els models proposats en aquesta secció es poden generalitzar amb la següent expressió:

$$M_t = \nu(B) I_t^{t_0} \tag{7.9}$$

on

$$\nu(B) = \frac{\omega(B) B^b}{\delta(B)} \tag{7.10}$$

amb

$$\omega(B) = \omega_0 - \omega_1 B - \omega_2 B^2 - ... - \omega_s B^s \tag{7.11}$$

$$\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - ... - \delta_r B^r \tag{7.12}$$

7.2 Anàlisi d'outliers

Per a saber quan aplicar un model d'intervenció, es buscaran els outliers o les observacions atípiques de la sèrie que marcaran l'inici de l'impacte. En aquest apartat s'explicaran els diferents tipus d'outliers que existeixen i els efectes que produeixen sobre el procés.

Suposarem que es té un procés $\{X_t, t \in \mathbb{Z}\}$ ARMA(p,q) representat amb la següent equació

$$X_t = \frac{\Theta(B)}{\Phi(B)} Z_t$$

on

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$$
$$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q$$

amb $Z_t \sim WN(0, \sigma^2)$. Aquest serà el procés lliure d'outliers sobre el que s'afegiran els mateixos per a generar un nou proces $\{Y_t, t \in \mathbb{Z}\}$. Tot i així, tots el resultats que es veuran es poden extendre fins a un model SARIMA(p,d,q)(P,D,Q)$_s$.

A continuació s'exposaran els diferents tipus d'outliers que es consideraran:

1. Outlier additiu (AO): un outlier additiu es un succés que afecta a la sèrie en un únic instant ($t = t_0$). L'impacte és equivalent al de l'efecte més simple de la variable impuls i s'expressa com
$$Y_t = X_t + \omega I_t^{(t_0)} \tag{7.13}$$

2. Outlier innovacional (IO): aquest outlier correspon a un succès que es propaga conforme el model del procés afectant a totes les observacions posteriors a la seva aparició. Aquesta es la seva equació
$$Y_t = X_t + \frac{\Theta(B)}{\Phi(B)} \omega I_t^{(t_0)} \tag{7.14}$$

3. Canvi de nivell (LS): es tracta d'un succès que té un impacte sobtat i es manté al llarg del temps. La seva representació és la següent
$$Y_t = X_t + \frac{1}{1-B} \omega I_t^{(t_0)} \tag{7.15}$$

4. Canvi temporal (TC): els successos produïts per aquest tipus d'outliers generen un efecte inicial que decau exponencialment fins no tindre gairebé rellevància sobre la sèrie. Es descriuen com
$$Y_t = X_t + \frac{1}{1-\delta B} \omega I_t^{(t_0)}, \ per \ 0 < \delta < 1 \tag{7.16}$$

Observació 7.7. Si es relacionen els outliers amb els models d'intervenció que s'han vist anteriorment, partint de l'equació (7.9) es tindria:

$$\nu(B) = \begin{cases} 1, \ per\ un\ AO \\ \frac{\Theta(B)}{\Phi(B)}, \ per\ un\ IO \\ \frac{1}{1-B}, \ per\ un\ LS \\ \frac{1}{1-\delta B}, \ per\ un\ TC \end{cases} \tag{7.17}$$

Es important recalcar algunes característiques dels efectes dels outliers que s'han proposat:

1. Els efectes dels outliers són independents de l'estructura del procés sobre el que fan efecte, excepte en el cas dels IO.
2. Els outliers AO i LS són casos límit dels TC. El primer correspon al cas $\delta = 0$ i el segon a $\delta = 1$.
3. L'outlier AO produeix un efecte immediat i únic sobre la sèrie en el període t_0 de magnitud ω.
4. En el cas dels outliers TC, s'efectua un impacte inicial de magnitud ω en el període t_0 i aquest decau gradualment amb un factor d'amortiment δ. Generalment s'utilitza el resultat $\delta = 0.7$.
5. Si es té un outlier de tipus LS, s'introdueix un canvi sobtat de magnitud ω que es manté al llarg del recorregut de la sèrie.
6. Els outliers IO produeixen un efecte de magnitud $\omega\psi_j$, on ω fa referència a la magnitud de l'efecte inicial i ψ_j correspon al coeficient j-èssim del polinomi $\Psi(B)$. $\Psi(B)$ compleix la igualtat $\Psi(B) = \frac{\Theta(B)}{\Phi(B)}$.

Els outliers també generen un efecte sobre els residus de la sèrie. La seva estimació és la següent:

$$\hat{Z}_t = \Pi(B)Y_t \tag{7.18}$$

on

$$\Pi(B) = \frac{\Phi(B)}{\Theta(B)} = 1 - \pi_1 B - \pi_2 B^2 - ... - \pi_g B^g \tag{7.19}$$

Desenvolupant la igualtat (7.18) s'arriba al següent resultat:

$$\hat{Z}_t = \Pi(B)X_t + \Pi(B)M_t = Z_t + \omega\nu(B)\Pi(B)I_t^{(t_0)} = Z_t + \omega o_{it} \tag{7.20}$$

i segons els tipus d'outliers que es presentin existeixen el següents residus

$$\begin{aligned} AO: \ & \hat{Z}_t = Z_t + \omega\Pi(B)I_t^{(t_0)} \\ IO: \ & \hat{Z}_t = Z_t + \Pi(B)\Psi(B)I_t^{(t_0)} = Z_t + I_t^{(t_0)} \\ LS: \ & \hat{Z}_t = Z_t + \frac{\omega}{1-B}\Pi(B)I_t^{(t_0)} \\ TC: \ & \hat{Z}_t = Z_t + \frac{\omega}{1-\delta B}\Pi(B)I_t^{(t_0)} \end{aligned} \tag{7.21}$$

Per últim, es pot representar el resultat de o_{it} de la igualtat (7.20) de la següent forma

$$o_{it} = \begin{cases} 0,\ t < t_0,\ \forall j \\ 1,\ t = t_0,\ \forall j \\ -\pi_j,\ t = t_0 + j(j = 1, 2, ..., T - t_0),\ i = 1(AO) \\ 0,\ t = t_0 + j(j = 1, 2, ..., T - t_0),\ i = 2(IO) \\ 1 - \sum_{h=1}^{j} \pi_h,\ t = t_0 + j(j = 1, 2, ..., T - t_0),\ i = 3(LS) \\ \delta^j - \sum_{h=1}^{j-1} \delta^{j-h}\pi_h - \pi_j,\ t = t_0 + j(j = 1, 2, ..., T - t_0),\ i = 4(TC) \end{cases} \tag{7.22}$$

i pel mètode dels mínims quadrats es pot estimar l'efecte d'un outlier en $t = t_0$ com

$$\begin{aligned} \hat{\omega}_{AO}(t_0) &= \frac{\sum_{t=t_0}^{n} \hat{Z}_t o_{1t}}{\sum_{t=t_0}^{n} o_{1t}^2} \\ \hat{\omega}_{IO}(t_0) &= \hat{Z}_t \\ \hat{\omega}_{TC}(t_0) &= \frac{\sum_{t=t_0}^{n} \hat{Z}_t o_{3t}}{\sum_{t=t_0}^{n} o_{3t}^2} \\ \hat{\omega}_{LS}(t_0) &= \frac{\sum_{t=t_0}^{n} \hat{Z}_t o_{4t}}{\sum_{t=t_0}^{n} o_{4t}^2} \end{aligned} \tag{7.23}$$

cal esmentar que en el cas de $t_0 = n$, $\hat{\omega}_{AO}(n) = \hat{\omega}_{IO}(n) = \hat{\omega}_{TC}(n) = \hat{\omega}_{LS}(n) = \hat{Z}_t$.

7.3 Detecció i tractament dels outliers

Per a trobar els outliers i tractar-los de la manera adequada, s'aplicarà la metodologia més moderna dissenyada per Chen i Liu al 1990.

Es parteix del model general dels outliers deduït a partir dels coneixements de l'anterior apartat. Es tracta del següent:

$$Y_t = X_t + M_t = X_t + \sum_{j=1}^{k} \omega_j \nu_j(B) I_{t,j}^{(t_0)} \tag{7.24}$$

i la corresponent expressió dels residus

$$\hat{Z}_t = \Pi(B) Y_t = Z_t + \sum_{j=1}^{k} \omega_j \Pi(B) \nu_j(B) I_{t,j}^{(t_0)} \tag{7.25}$$

ja que $Z_t = \Pi(B) X_t = \frac{\Phi(B)}{\Theta(B)} X_t$.

Prèviament a comentar el desenvolupament del mètode, cal examinar el valor màxim dels estadístics estandaritzats dels efectes dels outliers que a continuació s'exposen

$$\begin{aligned}
\hat{\tau}_{AO}(t_0) &= \left(\frac{\hat{\omega}_{AO}(t_0)}{\hat{\sigma}_u} \right) \left(\sum_{t=t_0}^{n} o_{1t}^2 \right)^{1/2} \\
\hat{\tau}_{IO}(t_0) &= \frac{\hat{\omega}_{IO}(t_0)}{\hat{\sigma}_u} \\
\hat{\tau}_{TC}(t_0) &= \left(\frac{\hat{\omega}_{TC}(t_0)}{\hat{\sigma}_u} \right) \left(\sum_{t=t_0}^{n} o_{3t}^2 \right)^{1/2} \\
\hat{\tau}_{LS}(t_0) &= \left(\frac{\hat{\omega}_{LS}(t_0)}{\hat{\sigma}_u} \right) \left(\sum_{t=t_0}^{n} o_{4t}^2 \right)^{1/2}
\end{aligned} \tag{7.26}$$

on $\sigma_u^2 = \frac{1}{T} \sum_{t=1}^{T} (\hat{Z}_t - \bar{Z})$ on $\bar{Z}$ fa referència a la mitjana dels $\hat{Z}_t$.

Es defineix una constant positiva $C \in [3, 4.5]$ que determinarà la presencia o no d'outliers, generalment s'escull $C = 3$. A partir d'aquí el procediment segueix tres etapes que es descriuran a continuació:

- Etapa 1: Estimació de paràmetres i detecció dels outliers

    1. S'estimen els paràmetres del model de la sèrie i els residus tal com s'ha vist en els capítols previs.
    2. Per $t = 1, ..., n$ es calculen $\hat{\tau}_{AO}(t)$, $\hat{\tau}_{IO}(t)$, $\hat{\tau}_{TC}(t)$ i $\hat{\tau}_{LS}(t)$ utilitzant els residus obtinguts en el punt (1). Es defineix $\eta_t = max\{|\hat{\tau}_{AO}(t)|, |\hat{\tau}_{IO}(t)|, |\hat{\tau}_{TC}(t)|, |\hat{\tau}_{LS}(t)|\}$. Si $\eta_t > C$, llavors es té un outlier del tipus corresponent.
    3. Si s'ha trobar algun outlier al pas previ, s'elimina l'efecte del mateix sobre el les observacions i els residus del model com s'ha vist als apartats previs i es repeteix el pas (2) per a revisar si existeixen més outliers. En cas contrari, es passa al pas (4).
    4. Si no s'ha trobat cap outlier, la sèrie està lliure d'aquests i no cal fer un anàlisi d'intervenció. Si s'han trobat outliers, es retorna al pas (1) i es tornen a estimar els paràmetres. Si quan es tornen a fer els passos (2) i (3) no es troben outliers ja es pot passar a la segona etapa del mètode.

- Etapa 2: Estimació dels efectes dels outliers i els paràmetres del model

    1. Es suposa que s'han trobat m moments $t_1, t_2, ..., t_m$ identificats com possibles outliers. S'estimen els efectes dels outliers ω_j amb els resultats de (7.23) i les estimacions dels residus i les observacions del model de la primera etapa.
    2. Ara es calculen els $\hat{\tau}_j = \frac{\hat{\omega}_j}{std(\hat{\omega}_j)}$. Si $min_j|\hat{\tau}_j| = \hat{\tau}_\nu \leq C$, es deixa de tenir en compte l'outlier del temps t_ν i es torna a fer el pas (1) d'aquesta segona etapa. En un altre cas es procedeix al següent pas.
    3. Amb els últims resultats de ω_j es descompten els efectes dels outliers a la sèrie.
    4. Per últim s'estimen els paràmetres del model tal com s'ha vist als capítols previs aplicant tots els resultats obtinguts en aquesta etapa.

- Etapa 3: Detecció d'outliers basant-se amb les últimes estimacions dels paràmetres

    1. Amb les últimes estimacions del model de la segona etapa es calculen els residus del model.
    2. Amb els residus calculats en última instància es repeteixen les etapes 1 i 2 tenint en compte que les estimacions dels paràmetres de la primera etapa són les fixades en el punt (4) de la segona etapa i que es poden ometre els punts (3) i (4) de la segona etapa. Les estimacions del punt (1) de la segona etapa son els efectes definitius dels outliers.

8 Anàlisi pràctic de sèries temporals

8.1 Introducció

L'anàlisi que es durà a terme per a les sèries temporals que es mostraran a continuació seguirà la metodologia de Box-Jenkins i l'anàlisi d'intervenció a partir dels outliers proposat per Chen-Liu. L'eina que es farà servir serà l'Rstudio i els scripts utilitzats es mostraran al final de la memòria a més de trobar-se en la carpeta `scripts` entregada conjuntament. Totes les funcions i els paquets esmentats es troben explicats amb detall a la documentació d'R adjunta a la bibliografia.

L'organització correspondrà als punts següents:

1. Identificació: s'estudiarà el comportament de la sèrie temporal i s'avaluarà quines seran les transformacions vistes a l'anàlisi clàssica que es duran a terme per tal que el model sigui estacionari. Un cop fetes les transformacions pertinents, es buscarà quin és el model més òptim per a les dades de tots els que s'han vist al llarg de la memòria. Per a evaluar de quin es tracta, s'utilitzarà com a punt de partida la funció `auto.arima` del paquet `forecast` i s'aniran fent diferents proves fins arribar a uns resultats correctes.

2. Estimació: a partir del model seleccionat es faran les estimacions de tots els paràmetres. Per a fer-ho es farà ús de la funció `arima` del paquet `stats`.

3. Validació: es validarà el model mitjançant el tractament dels residus vist amb anterioritat. Les funcions utilitzades seran `Box.test`, `shapiro.test` i `ks.test` del paquet `stats`.

4. Predicció: es farà una predicció de les observacions futures de la sèrie temporal. Per a fer-la es s'utilitzarà la funció `predict` del paquet `stats`.

5. Anàlisi d'intervenció: si es creu necessàri, es buscaran els outliers del procés i es farà el tractament adequat per a tornar a modelar la sèrie amb els nous paràmetres i tornar a fer les prediccions dels futurs valors. Es necessitarà l'aplicació d'algunes funcions del paquet `tsoutliers`.

8.2 Sèrie temporal d'establiments hotelers a Catalunya

La primera sèrie temporal correspondrà al grau d'ocupació mensual dels establiments hotelers en milers de persones a Catalunya. Les observacions daten des del Gener de 1999 fins al Gener de 2021 i són les següents.

Anys	Gener	Febrer	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre
2021	161.3	-	-	-	-	-	-	-	-	-	-	-
2020	1017.7	1101.8	438.7	0	9.6	143.8	742	1058.6	607.2	314.3	143.2	219.2
2019	959.8	1097	1362.3	1786.9	1997.4	2300.3	2436.5	2586.1	2064.5	1824.6	1204.4	1132.6
2018	927	1021	1386.7	1647.4	1956.3	2168.2	2374	2467.5	2092.5	1790	1131.4	1083.4
2017	900.5	1023.7	1318.8	1835.4	1910.6	2181.4	2471	2434.9	2063.4	1661.4	993.6	977.1
2016	829.6	986.3	1330.5	1483	1869.1	2030.2	2403	2438.9	1968.1	1731	1013.5	1011.4
2015	747.4	868.8	1106.3	1486.7	1812.2	1856.6	2126.6	2365.4	1794.4	1627	961.2	933.8
2014	682.9	806.2	1063.8	1406.4	1646.8	1830.5	2026.6	2264	1744.2	1483.1	893.3	925.7
2013	637.2	757.7	1139.1	1236	1603.8	1750.5	1991.1	2187	1688.8	1440.2	915.4	860.9
2012	703.2	846.7	1085.3	1455.1	1549.8	1716.8	2048.1	2096.9	1727.9	1366.5	867.2	773.4
2011	676.4	844.6	1057.6	1442.8	1536.3	1819.7	2108.2	2170.4	1713.9	1468.7	865.5	802.2
2010	664.9	811.7	1011.1	1348.6	1586.7	1619.4	1975.4	2054.8	1546.6	1450.3	879.8	830.5
2009	601.7	710.1	823.7	1221	1432.4	1468.1	1690.3	1858.1	1445.3	1246.1	783.9	755.8
2008	660.8	797.6	1081.2	1122.8	1463.4	1548.8	1757.4	1847.5	1445.8	1230	767.4	689.1
2007	675.3	741.5	973.2	1287.6	1394.4	1586.4	1721.8	1839.3	1493.3	1263.8	875.4	718.5
2006	626.4	735.7	887.6	1296.6	1410.3	1514.6	1797	1845.6	1524.5	1233.7	789.7	659.7
2005	540.5	612.1	783.6	930.8	1233.3	1288.8	1488.6	1578.1	1281.9	1096.6	682.6	635.9
2004	483.8	590	772.3	998.6	1152.7	1208.8	1359	1465.6	1184.4	1035.9	664.7	622.1
2003	450.5	530	702.1	894.7	1056.4	1227.8	1282.9	1390.8	1095.4	952.1	592	597.1
2002	452.1	540.7	732.7	862	1066.9	1098.3	1289.5	1387.8	1103.9	942.8	588.4	544.9
2001	462.5	541	639.6	902.9	976.5	1111.9	1261.8	1308.1	1138.6	849.1	530	496.8
2000	452.2	500.2	673.9	916.7	971.6	1140.2	1236.7	1294.8	1099.6	861.6	544.1	517.3
1999	403.5	507.2	584.7	826.6	1017.1	1041	1216.8	1306.7	1070.1	838.4	503.3	474.4

Taula 1: Establiments hotelers. Nombre de viatjers (en milers)

8.2.1 Identificació

La representació gràfica de la sèrie és la següent:

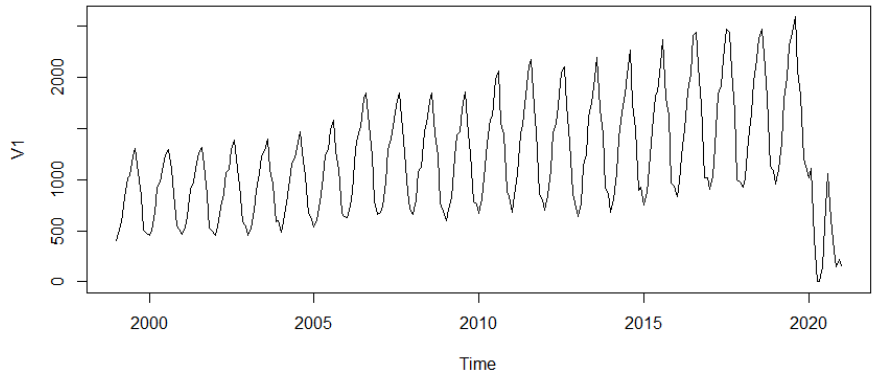

Figura 3: Gràfic de la sèrie temporal corresponent al grau d'ocupació hotelera

En primer lloc, s'aprecia certa variabilitat de les dades. A mesura que passa el temps, degut a l'augment del pic màxim anual de les observacions, aquesta creix. Per a suavitzar

aquest efecte, es calcularan els logaritmes de totes les observacions de la sèrie i es passarà a treballar amb aquests nous resultats. Cal assenyalar que per a fer possible aquest càlcul, la dada corresponent a l'Abril de 2020 s'ha substitït per 1 quan prèviament era 0. Les observacions obtingudes són les següents:

Anys	Gener	Febrer	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre
2021	5.083266	-	-	-	-	-	-	-	-	-	-	-
2020	6.925300	7.006423	6.060057	0	2.261763	4.968423	6.609349	6.964703	6.408858	5.750348	4.964242	5.389985
2019	6.866725	7.000334	7.216930	7.488238	7.599602	7.740795	7.798318	7.857906	7.632643	7.509116	7.093737	7.032271
2018	6.831954	6.928538	7.234682	7.406954	7.578810	7.681653	7.772332	7.810961	7.646115	7.489971	7.031211	6.987860
2017	6.802950	6.931179	7.184478	7.515018	7.555173	7.687722	7.812378	7.797661	7.632110	7.415416	6.901335	6.884589
2016	6.720944	6.893961	7.193310	7.301822	7.533212	7.615890	7.784473	7.799302	7.584824	7.456455	6.921165	6.919091
2015	6.616601	6.767113	7.008776	7.304314	7.502297	7.526502	7.662280	7.768702	7.492426	7.394493	6.868183	6.839262
2014	6.526348	6.692332	6.969603	7.248789	7.406589	7.512344	7.614115	7.724888	7.464051	7.301890	6.794922	6.830550
2013	6.457084	6.630288	7.037994	7.119636	7.380131	7.467657	7.596443	7.690286	7.431773	7.272537	6.819361	6.757978
2012	6.555641	6.741346	6.989612	7.282830	7.345881	7.448217	7.624668	7.648215	7.454662	7.220008	6.765270	6.650796
2011	6.516785	6.738863	6.963757	7.274341	7.337132	7.506427	7.653590	7.682667	7.446527	7.292133	6.763307	6.687358
2010	6.499637	6.699131	6.918794	7.206822	7.369412	7.389811	7.588526	7.627934	7.343814	7.279526	6.779695	6.722028
2009	6.399759	6.565406	6.713806	7.107425	7.267107	7.291724	7.432661	7.527310	7.276072	7.127774	6.664281	6.627777
2008	6.493451	6.681607	6.985827	7.023581	7.288518	7.345236	7.471591	7.521589	7.276418	7.114769	6.643008	6.535386
2007	6.515157	6.608675	6.880590	7.160535	7.240219	7.369223	7.451126	7.517140	7.308744	7.141878	6.774681	6.577166
2006	6.439989	6.600822	6.788521	7.167501	7.251558	7.322907	7.493874	7.520560	7.329422	7.117773	6.671653	6.491785
2005	6.292495	6.416896	6.663899	6.836044	7.117449	7.161467	7.305591	7.363977	7.156099	6.999970	6.525909	6.455041
2004	6.181672	6.380123	6.649373	6.906354	7.049862	7.097383	7.214504	7.290020	7.076992	6.943026	6.499336	6.433101
2003	6.110358	6.272877	6.554076	6.796488	6.962622	7.112979	7.156878	7.237634	6.998875	6.858670	6.383507	6.392085
2002	6.113903	6.292865	6.596736	6.759255	6.972513	7.001519	7.162010	7.235475	7.006605	6.848854	6.377407	6.300602
2001	6.136647	6.293419	6.460843	6.805612	6.883975	7.013826	7.140295	7.176331	7.037555	6.744177	6.272877	6.208188
2000	6.114125	6.215008	6.513082	6.820780	6.878944	7.038959	7.120202	7.166112	7.002702	6.758791	6.299133	6.248623
1999	6.000176	6.228905	6.371099	6.717321	6.924711	6.947937	7.103980	7.175260	6.975507	6.731495	6.221186	6.162051

Taula 2: Establiments hotelers. Logaritme del nombre de viatjers (en milers)

Amb la corresponent representació gràfica:

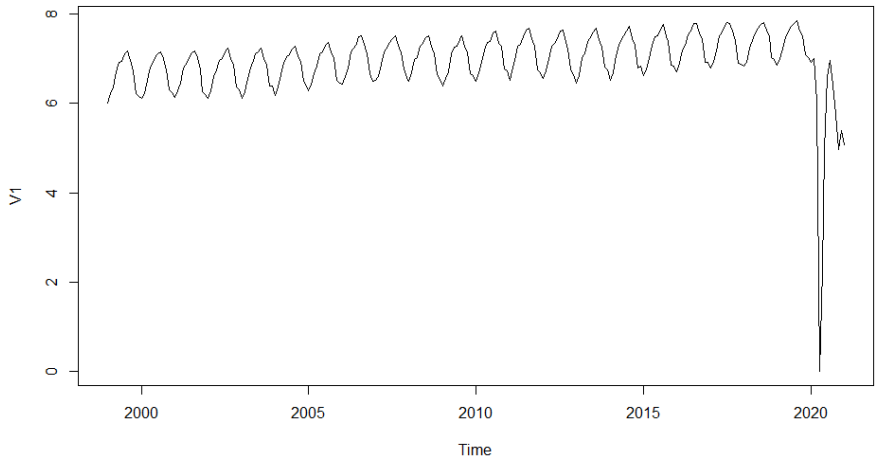

Figura 4: Gràfic de la sèrie temporal corresponent al grau d'ocupació hotelera aplicant logaritmes

S'ha conseguit reduïr la variabilitat de les dades al llarg del temps, però el descens del 2020 produït per l'efecte de la COVID-19, s'ha pronunciat encara més.

D'altra banda, pot apreciar clarament un comportament estacional, en els mesos corresponents a estiu les observacions són significativament superiors a la resta de mesos. També s'aprecia una tendència creixent que s'atura amb l'arribada de la COVID-19, la qual provoca un descens fulgurant del turisme a Catalunya. Per tant, a primera vista sembla que la sèrie no serà estacionària ni en mitjana ni en variància. Per a ser més precisos en aquest aspecte, es mostren a continuació els correlogrames de la funció d'autocorrelació simple (ACF) i la funció d'autocorrelació parcial (PACF).

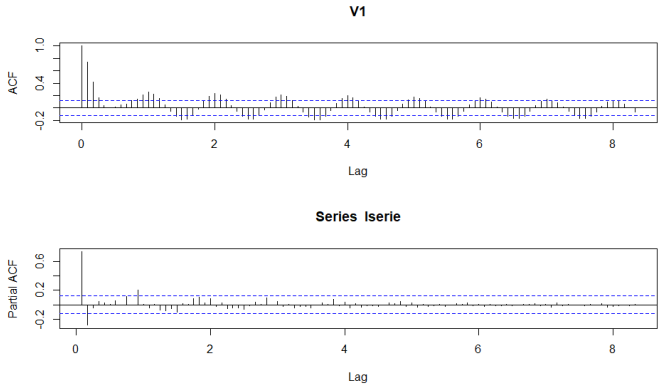

Figura 5: ACF i PACF de la sèrie corresponent al grau d'ocupació hotelera després d'aplicar logaritmes

Com era d'esperar, la sèrie no es estacionària ja que els valors de l'ACF no descenen exponencialment cap a 0.

Llavors, seguint les instruccions explicades a l'anàlisi clàssica de sèries temporals, s'aplicaran diferències regulars i estacionals amb l'objectiu d'obtindre unes obervacions que en aquest cas, si que siguin estacionàries. Un cop s'ha fet aquest procediment s'obté la següent representació de les observacions.

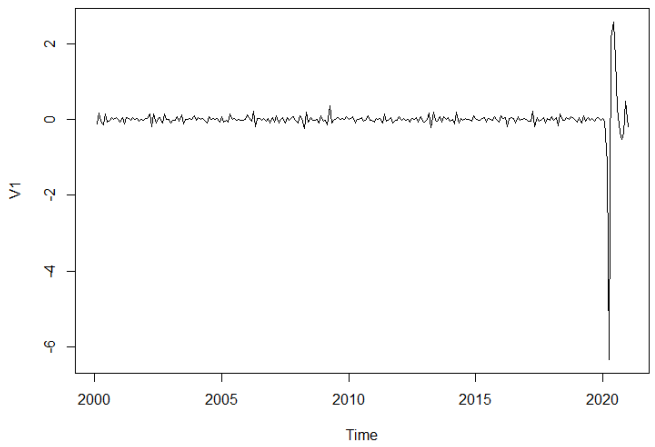

Figura 6: Gràfic de la sèrie corresponent al grau d'ocupació hotelera amb logaritmes i diferències regulars i estacionals

Ara, les dades oscil·len al voltant del 0 amb una variància molt petita i es pot veure un augment evident d'aquesta al 2020. Tot i que s'hagi obtingut possiblement una sèrie estacionària, aquest fet podria ser problemàtic. L'anàlisi d'intervenció podria ser una possible solució.

Per a corroborar l'estacionarietat de la sèrie diferenciada es mostraran de nou l'ACF i la PACF.

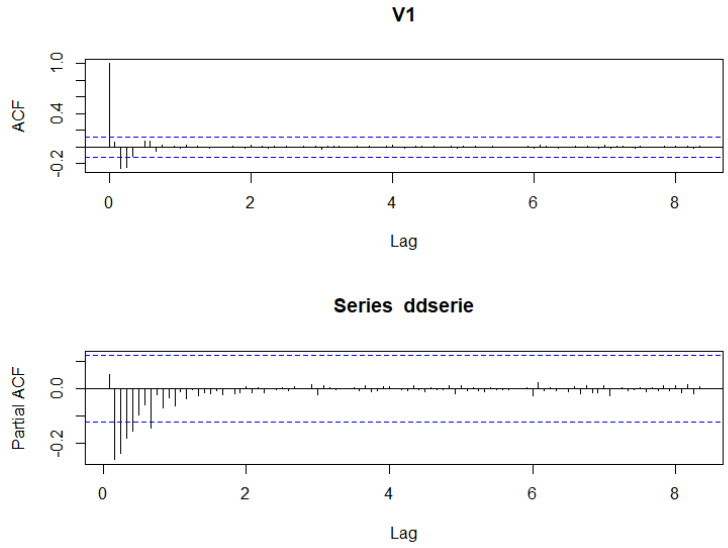

Figura 7: ACF i PACF de la sèrie corresponent al grau d'ocupació hotelera diferenciada

En efecte, sembla una sèrie estacionària ja que els retards de l'ACF descenen exponencialment cap a 0. Per a seleccionar el model més òptim, s'utilitzarà com s'ha comentat prèviament, la funció `auto.arima` amb el paràmetre `ic=‘‘aicc’’` per a indicar que el criteri de selecció sigui el de minimitzar l'AICc, com a punt de partida. Finalment, després de fer algunes proves amb diferents models, s'ha conclòs que el que millor s'ajusta a les dades és un $ARIMA(2,1,1)(0,1,1)_{12}$, amb un AICc de valor 296.42.

8.2.2 Estimació

Per estimar els paràmetres del model s'utilitza la comanda `arima` amb els paràmetres `order=c(2,1,1), seasonal=list(order=c(0,1,1),period=12)`, per indicar el tipus de procès al que pertany la sèrie, i `method=‘‘CSS-ML’’`, que indica que es faci una primera estimació minimitzant la suma dels residus al quadrat, com s'ha comentat en la secció 4.7, i a partir d'aquesta torni a fer una estimació per tal de maximitzar la funció de log-versemblança $ln\hat{L}$. Els resultats finals obtinguts dels procediments són els següents:

$$(1-\phi_1 B-\phi_2 B^2)(1-B)(1-B^{12})X_j = Z_j(1+\theta_1 B)(1+\vartheta_1 B^{12})$$

amb $\phi_1 = 0.7557$, $\phi_2 = -0.3265$, $\theta_1 = -0.84605$ i $\vartheta_1 = -0.2977$.

8.2.3 Validació

Sabent que la sèrie temporal al ser un model SARIMA amb els paràmetres definits a l'anterior secció, és estacionària i invertible, s'estudiarà la validació dels residus amb els tests de Ljung-Box i Shapiro-Wilk.

Els residus obtinguts a partir del model definit són els següents:

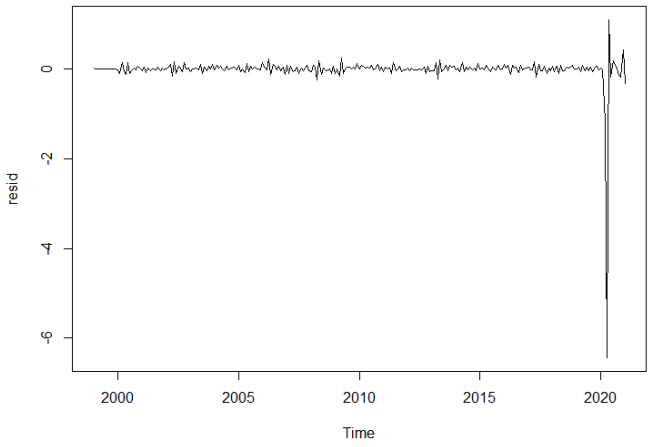

Figura 8: Residus del model definit per a la sèrie corresponent al grau d'ocupació hotelera

Amb els corresponents correlogrames de les ACF i PACF:

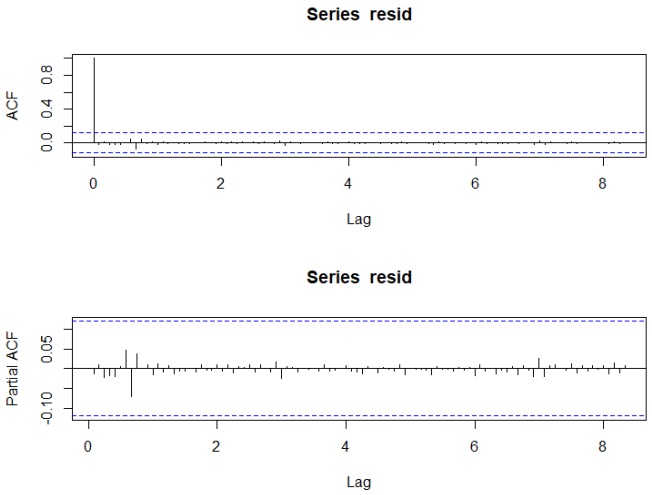

Figura 9: ACF i PACF dels residus del model definit per a la sèrie corresponent al grau d'ocupació hotelera

En primer lloc, mirant els correlogrames, sembla que els residus es tracten d'un soroll blanc ja que no s'aprecia cap pic significatiu al llarg dels retards. No obstant, el gràfic previ mostra amb claredat que tot i que la majoria de les dades oscil·len al voltant del 0 de forma homogènia, les observacions correponents al 2020 difereixen significativament de la resta.

Fent els tests especificats amb anterioritat, s'obté un p-valor de 0.8195 per al Test de Ljung-Box i un molt proper al 0 pel de Shapiro-Wilk. Per tant, tot i ser un soroll IID, aquests residus no són gaire bons perquè el fet d'augmentar la variància d'una forma tant descarada en les observacions que corresponen al 2020, fa que no se'ls pugui assignar una distribució adequada.

Així que, com s'ha vist al llarg de l'anàlisi, l'impacte de la COVID-19 està molt present en aquesta sèrie i l'anàlisi d'intervenció sembla més que necessàri. Tot i això, es duran a terme les prediccions amb aquest model per a ser comparades posteriorment amb les obtingudes després d'aplicar l'anàlisi d'intervenció.

8.2.4 Predicció

Es fa l'estimació de la resta de l'any 2021 a partir de la comanda `predict` sobre el model que s'ha dissenyat en l'anterior secció i sobtenen les següents prediccions:

	Febrer	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre
2020	5.1487545	4.598639	0.4598252	2.1023891	4.0417909	5.2093178	5.4693239	5.0153306	4.5127336	3.8323374	4.1158484

Taula 3: Prediccions de la sèrie corresponent al grau d'ocupació hotelera

Amb les corresponents representacions gràfiques amb els intervals de confiança del 95%:

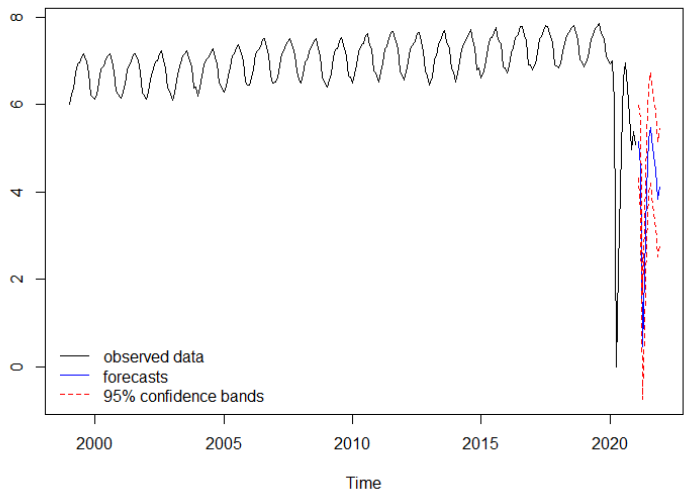

Figura 10: Gràfic de les prediccions de la sèrie corresponent al grau d'ocupació hotelera

Tot i que, com s'ha comentat a la secció prèvia, el model no s'ajusta del tot bé degut a l'impacte de la COVID-19, les prediccions obtingudes suggereixen que es tornarà a produïr un descens a l'Abril i al Maig de les reserves, millorant lleugerament als mesos corresponents a l'estiu. De totes formes, els resultats seguirien sent pitjors que els de l'any passat, que ja van ser dolents.

Cal dir, que al no aplicar l'anàlisi d'intervenció quan és evident la seva necessitat, el model està prenent els resultats de l'any 2020 com una evolució natural de la sèrie i per això les prediccions del 2021 tenen aquestes similituds amb les observacions de l'any anterior. Ara, es procedirà amb l'aplicació de l'anàlisi d'intervenció i es veurà com canvien els paràmetres del model i en conseqüència, les prediccions del 2021.

8.2.5 Anàlisi d'intervenció

El primer pas per a dur a terme l'anàlisi d'intervenció és detectar on es troben els outliers i quina és la seva afectació sobre la sèrie. Per a fer-ho s'utilitzarà la comanda `tso`, que fa automàticament tot el procés de recerca i estimació dels outliers proposats al capítol 7 de la memòria. Aquesta, utilitza per a fer la primera etapa del procediment la funció `locate.outliers.oloop`, amb la qual a partir de les observacions de la sèrie i el model definit prèviament, fa una primera estimació dels outliers. Per a la segona etapa, a l'hora de descartar els outliers, la funció fa ús d'una altra anomenada `discard.outliers`. S'introdueix el paràmetre `method = ''en-masse''` per a que compari els outliers amb el valor de C i s'eliminin d'aquesta forma els que no tinguin un efecte significatiu. El valor de C per als procediments és $3+0.0025*(n-50)$, sigui n el nombre total d'observacions de la sèrie. Per últim, es torna a fer l'estimació dels paràmetres i es repeteix el procès, com s'ha de procedir en l'etapa 3. Per a dur a terme l'estimació dels paràmetres correctament, s'afegeix el paràmetre `tsmethod = ''arima''` i s'indica dins el paràmtre `args.tsmethod` el tipus de model ARIMA amb el que es treballa. D'aquesta forma, torna a estimar els paràmetres utilitzant la funció `arima` feta servir amb anterioritat i a conseqüència d'això, s'obté un mateix tipus de model ARIMA que el que s'ha definit prèviament. Per a que la funció detecti els tipus d'outliers que s'han definit en seccions prèvies, es fa ús del paràmetre `types = c(''TC'',''AO'',''LS'',''IO'')`.

En el cas de la sèrie temporal amb la que s'està treballant, s'ha indicat que es tracta d'un model $ARIMA(2,1,1)(0,1,1)_{12}$, com s'ha definit en seccions anteriors, i l'efecte total dels outliers ha estat el següent:

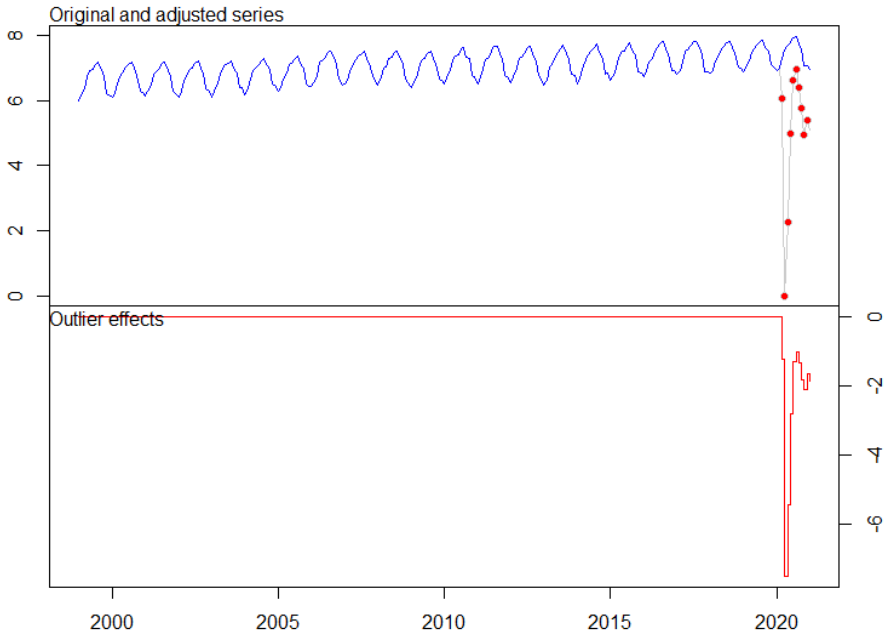

Figura 11: Efecte conjunt dels outliers sobre la sèrie dels establiments hotelers

Es pot veure com fins a l'arribada de la COVID-19 no s'ha trobat cap outlier, però en arribar, es produeix un outlier en cada observació. L'impacte conjunt d'aquests afecta de forma negativa tenint un pic clar al més d'Abril quan hi havia confinament a Catalnuya. D'altra banda, aquest impacte negatiu es suavitza durant els mesos d'estiu quan es van reduïr les restriccions i es van permetre desplaçaments per tota la península. A més, com es pot veure a la part superior del plot, el model suggeria uns resultats lleugerament millors als dels anys anteriors, establint un màxim absolut al mes d'Agost, una situació lògica amb la tendència creixent de les observacions. Ara, es definiran un a un els 10 outliers trobats des del mes de Març de 2020 fins al Desembre de 2021.

- Outlier 1: Outlier additiu situat a l'observació 255 que correspon al mes de Març de 2020. Suposa el primer impacte de la COVID-19 i només afecta en aquesta observació. Té un coeficient de magnitud -1.2374, pel que redueix el valor de la sèrie en aquest mes.
- Outlier 2: Outlier innovacional situat a l'observació 256 que correspon al mes d'Abril de 2020. Té un coeficient molt elevat de valor -7.5423, que produeix un primer impacte de magnitud -7.5423 al mes d'Abril de 2020 i en les següents observacions s'estabilitza al voltant del -2.4.
- Outlier 3: Canvi temporal situat a l'observació 257 que correspon al mes de Maig de 2020. Aquest provoca un primer impacte de valor -3.4606 que, sumat a l'outlier 2, tenen un efecte conjunt de -5.443 per al mes de Maig de 2020. A mesura que passa el temps, l'efecte de l'outlier decreix.

- Outlier 4: Canvi de nivell situat a l'observació 258 que correspon al mes de Juny de 2020. El seu coeficient és de 2.5783 i provoca un impacte d'aquesta magnitud sobre totes les observacions següents. Per tant, contrarresta d'alguna forma els outliers d'efecte negatiu previs per a suavitzar l'efecte del virus. Tot i així, al mes de Juny, l'impacte global dels outliers és de -2.8103, que suposa un descens dels efectes externs cap a la sèrie, comparat amb els mesos anteriors, i en conseqüència uns valors més grans d'aquesta.

- Outlier 5: Outlier innovacional situat a l'observació 259 que correspon al mes de Juliol de 2020. Té un coeficient de 0.2657 pel que suposa una petita millora en les observacions a partir d'aquest mes. La seva afectació és similar a la del outlier 2, té un primer impacte més elevat, de magnitud 0.2657, i els corresponents als mesos següents s'estabilitzen al voltant d'un valor, en aquest cas de 0.8. La suma de l'efecte d'aquest i la resta d'outliers previs es tradueix en un efecte total de -1.3064, lleugerament superior al del mes passat.

- Outlier 6: Canvi temporal situat a l'observació 260 que correspon al mes d'Agost de 2020. El seu coeficient de 0.0739 provoca un impacte molt lleuger de forma positiva que va decreixent en quant a intensitat. Gràcies a aquesta petita millora, el mes d'Agost representa el mes amb uns millors resultats pel que fa a l'ocupació hotelera a Catalunya al 2020 ja que la magnitud de l'impacte del virus és del -1.012432.

- Outlier 7: Outlier innovacional situat a l'observació 261 que correspon al mes de Setembre de 2020. Té un coeficient de -0.7647 que provoca un decreixement de nou en la sèrie. Aquest, igual que els outliers 2 i 5 té un primer impacte més gran de valor -0.7647 i per a les següents observacions s'estabilitza al voltant del -0.25. Aquest outlier, juntament amb els anteriors, provoca un efecte negatiu de -1.3465 sobre la sèrie per al mes de Setembre.

- Outlier 8: Outlier additiu situat a l'observació 262 que correspon al mes d'Octubre de 2020. Provoca un únic impacte negatiu de -1.2940 que agreuja l'impacte dels outliers sobre la sèrie en aquesta observació. Aquest mes, els outliers provoquen un efecte total de -1.8398.

- Outlier 9: Canvi de nivell situat a l'observació 263 que correspon al mes de Novembre de 2020. Aquest, provoca un impacte de -1.6490 que es manté al llarg de la sèrie i fa que l'impacte conjunt segueixi agreujant-se i els valors de la sèrie segueixin decreixent. Tenint en compte això i l'efecte de la resta dels outliers, al mes de Novembre l'impacte dels outliers és de -2.1195.

- Outlier 10: Outlier additiu situat a l'observació 264 que correspon al mes de Desembre de 2020. El seu coeficient de 0.2878 provoca un únic impacte d'aquesta magnitud en aquesta observació que redueix en valor absolut l'impacte conjunt dels outliers. Tenint això en compte, al mes de Desembre es té un impacte total dels outliers de -1.665404.

Cal afegir que al mes de Gener de 2021, es mantenen els efectes de tots els outliers excepte dels de tipus additiu i es té una afectació conjunta de -1.879230, lleugerament superior en valor absolut a la del mes anterior.

Un cop definits els outliers i els seus efectes, el model resultant és el descrit a continuació:

$$Z_t = \Pi(B)X_t + \Pi(B)\sum_{j=1}^{10}\omega_j\nu_j(B)I_{t,j}^{t_0}$$

on

$$\Pi(B) = \frac{(1-B)(1-B^{12})(1-\phi_1 B-\phi_2 B^2)}{(1+\theta_1 B)(1+\vartheta_1 B^{12})} = \frac{(1-B)(1-B^{12})(1+0.0949B-0.0654B^2)}{(1-0.6661B)(1-0.9997B^{12})}$$

$$\nu_1(B) = \nu_8(B) = \nu_{10}(B) = 1$$
$$\nu(B)_2 = \nu_5(B) = \nu_7(B) = \Pi(B)$$
$$\nu_4(B) = \nu_9(B) = \frac{1}{1-B}$$
$$\nu_3(B) = \nu_6(B) = \frac{1}{1-0.7B}$$

amb $t_0 \in \{255, 256, 257, 258, 259, 260, 261, 262, 263, 264\}$ i els valors dels coeficients ω_j corresponen als definits previament durant l'explicació dels outliers.

Per a validar el model, s'analitzaran els residus, els quals es poden representar gràficament de la següent forma:

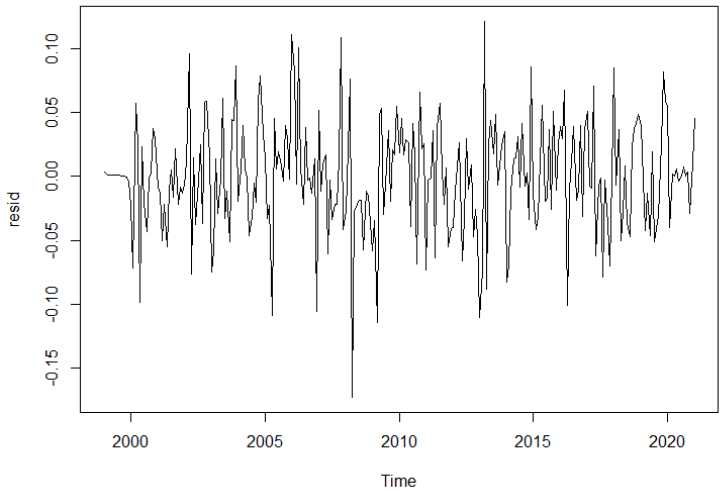

Figura 12: Residus del model de la sèrie corresponent al grau d'ocupació hotelera després d'aplicar l'anàlisi d'intervenció

Es evident que s'ha fet desaparèixer l'efecte de la COVID-19 i ara si que sembla que es podrà ajustar una distribució a diferència del cas exposat amb anterioritat.

Aplicant els tests de Ljung-Box i Shapiro-Wilk s'han obtingut uns p-valors de 0.9591 i 0.07393 respectivament que indiquen que en aquest cas, els residus es tracten d'un soroll IID gaussià.

Un cop s'ha determinat la distribució que segeuixen els residus del model, es poden fer les prediccions tenint en compte l'efecte futur dels outliers. Per a fer-ho se li afegeix a la comanda `predict` el paràmetre `newxreg` amb els valors corresponents als efectes futurs dels outliers. En aquest cas, s'han obtingut les següents prediccions:

	Febrer	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre
2021	5.270468	5.570998	5.849728	6.025098	6.125699	6.263745	6.327089	6.114055	5.9486	5.475646	5.416053

Taula 4: Prediccions de la sèrie corresponent al grau d'ocupació hotelera després d'aplicar l'anàlisi d'intervenció

Amb la corresponent representació gràfica conjuntament amb l'interval de confiança del 95%

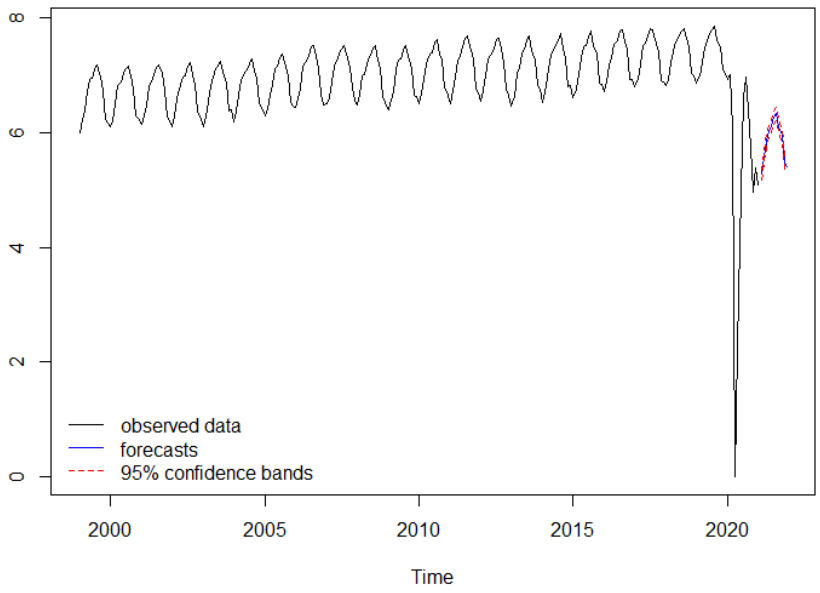

Figura 13: Gràfic de les prediccions de la sèrie corresponent al grau d'ocupació hotelera després d'aplicar l'anàlisi d'intervenció

A la vista està que els resultats després de fer l'anàlisi d'intervenció són més optimistes. Això es degut a que han modelat l'efecte de la COVID-19 com un efecte amb un pic màxim al mes de Març, que redueix la seva intensitat a mesura que passen els mesos. Per entendre això es mostrarà a continuació l'efecte dels outliers en els mesos de les observacions predites.

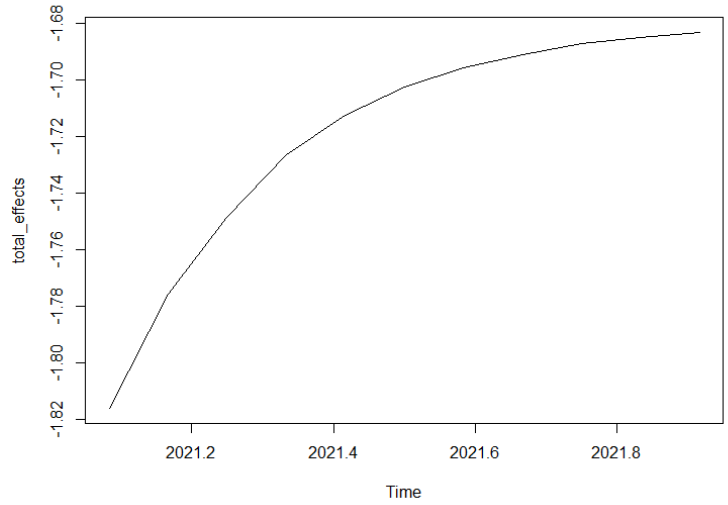

Figura 14: Efecte dels outliers en les prediccions de la sèrie corresponent al grau d'ocupació hotelera després d'aplicar l'anàlisi d'intervenció

Com es pot apreciar a la gràfica, a mesura que passa el temps, els efectes dels outliers són menors en valor absolut, demostrant el fenòmen comentat amb anterioritat.

Per a tindre una visió més global amb la que evaluar els resultats de les prediccions del model, es desfaran els logarítmes de les prediccions i es veuran els valors reals de les observacions predites. Els resultats obtinguts han estat:

	Febrer	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre
2021	194.5069	262.6961	347.1401	413.682	457.4644	525.182	559.5253	452.1685	383.2164	238.8048	224.9894

Taula 5: Prediccions definitives de la sèrie corresponent al grau d'ocupació hotelera després de desfer els logaritmes

La representació gràfica d'aquestes prediccions és la següent:

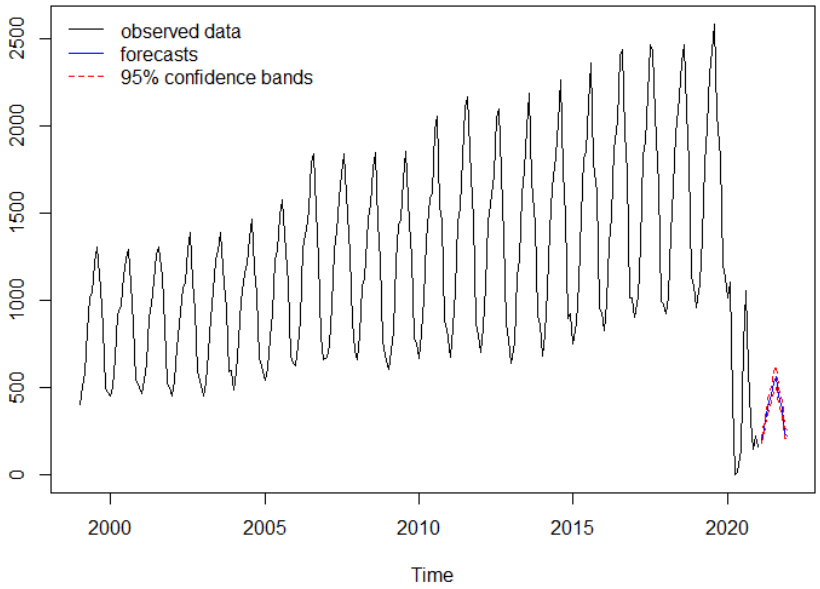

Figura 15: Gràfic de les prediccions definitives de la sèrie corresponent al grau d'ocupació hotelera després de desfer els logaritmes

En primer lloc, s'evidencia que l'efecte de la COVID-19 ha reduït de forma clara els valors de les observacions predites. El màxim absolut del 2021 es trobaria al mes d'Agost amb una aproximació de 560 milers de persones reservant un hotel a Catalunya. Aquests resultats suposarien l'estiu amb menys ocupació hotelera registrats al país. També cal remarcar que els resultats obtinguts segueixen una distribució més semblant a la d'anys previs al 2020.

8.3 Sèrie temporal de l'atur

La segona sèrie que s'analitzarà serà la corresponent a les dades de persones a l'atur a Catalunya desde Gener de 1996 a Febrer de 2021. En la següent taula es mostren totes les observacions:

Anys	Gener	Febrer	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre
2021	508081	512290	-	-	-	-	-	-	-	-	-	-
2020	393682	395214	417047	467810	483149	485019	469349	480642	478201	484559	484748	497611
2019	398376	396642	395740	381598	371091	357272	358830	371418	372623	387267	390182	388124
2018	422866	418181	411461	398946	385568	370192	369124	380718	380344	391197	394405	392907
2017	453923	452342	446017	425751	409490	391388	387313	397385	400373	415071	422462	418018
2016	518080	510237	499991	486123	470205	450060	441016	445440	451081	458406	462979	453645
2015	582769	581124	571655	552974	531899	510947	501785	506306	513187	523528	521660	515668
2014	633871	629586	624467	611822	592304	570214	568231	571616	575812	587133	581652	575948
2013	661817	665176	664050	656995	642166	617288	610429	611658	620911	633832	638344	624872
2012	633210	641948	638247	635721	630932	615576	614792	622882	632457	646306	652091	646956
2011	589623	602611	611269	601541	595342	576394	570869	584648	600930	615558	615669	614244
2010	583883	597287	604038	593656	574736	554261	539191	555894	557268	566496	564541	562673
2009	455757	479487	498352	505262	505041	488247	495911	519129	531352	543603	555405	561761
2008	282897	290912	291640	299387	305247	311422	321964	342082	354215	377999	402836	423232
2007	261348	257583	254720	248659	244813	245490	255575	267473	258234	258012	259170	265789
2006	269904	273062	270573	259989	253686	250757	257969	264325	255144	256166	257771	260749
2005	282286	280486	275029	263200	252704	247969	259417	269054	266392	262588	263825	262605
2004	283377	278239	269601	263186	256636	257228	264085	273921	274027	273715	276589	274294
2003	271002	268461	266284	257374	253137	252482	256266	261991	258845	267590	274844	279586
2002	250880	253059	248054	246022	244082	240531	249127	254959	251027	257350	265571	268924
2001	222028	219873	217049	209559	210021	208518	217032	225567	222175	231487	239077	242891
2000	228594	226146	220003	210156	203626	202026	209921	210399	209577	212723	217027	215449
1999	256447	252321	245287	238668	228196	222604	221074	221769	222063	225662	230235	225145
1998	306334	300372	293345	279001	268936	261679	257285	257390	262612	263413	264162	255965
1997	352521	350838	341483	332396	323362	314639	305100	302898	310224	313283	315240	307443
1996	385124	382067	377616	365648	355684	347288	341048	337871	346363	351641	355321	350990

Taula 6: Persones en l'atur desde 1996

Degut a la l'extensió del treball, l'anàlisi amb la metodologia de Box-Jenkins d'aquesta segona sèrie estarà explicada als annexes. Aquesta seguirà els mateixos apartats que l'exposada prèviament en la sèrie corresponent al grau d'ocupació hotelera. Per tant, seguint l'esquelet principal del projecte, es procedirà a mostrar l'anàlisi d'intervenció sobre la el procés.

8.3.1 Anàlisi d'intervenció

Per dur a terme l'anàlisi d'intervenció, s'han empleat els mateixos procediments que amb l'exemple previ. Seguint-los, s'ha fet una recerca dels outliers amb la que s'han trobat únicament dos. La seva afectació global sobre la sèrie és la següent:

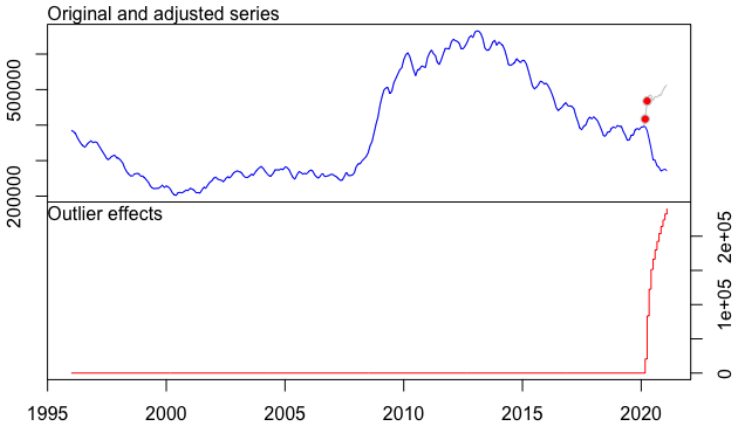

Figura 16: Afectació i representació dels outliers sobre la sèrie corresponent a l'atur

Seguint el model ARIMA definit amb anterioritat, era esperable un decreixement en el nombre d'aturats per al 2020. Ha passat just el contrari i per això s'han trobat dos outliers als mesos de Març i Abril de 2020. Resulta evident que el motiu d'aquests ha estat l'impacte de la COVID-19.

Es procedirà a fer una explicació més específica dels dos outliers exposats:

- Outlier 1: outlier additiu situat a l'observació 291 que fa referència al mes de Març de 2021. Té un únic efecte de valor 20674, sobre la sèrie en aquesta observació. La corresponent representació seria la següent:

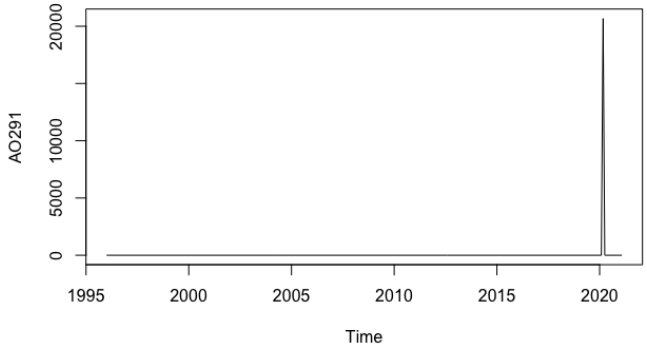

Figura 17: Representació de l'efecte del primer outlier de la sèrie corresponent a l'atur

- Outlier 2: outlier innovacional situat a l'observació 292 que fa refèrencia al mes d'Abril de 2021. Té un primer impacte de magnitud 83443 que va augmentant a mesura que passa el temps. Per tant, això ha provocat i seguirà provocant un augment en el valor en les observacions. La representació gràfica del seu efecte és la següent:

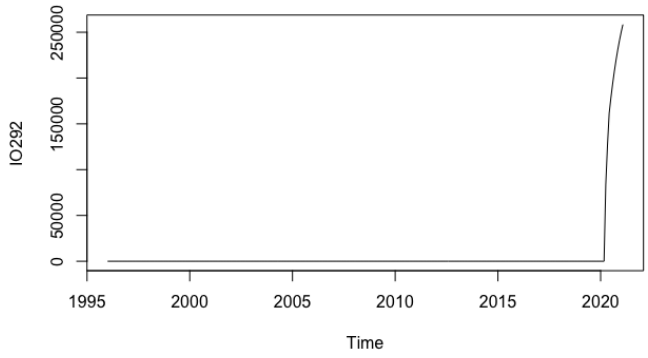

Figura 18: Representació de l'efecte del segon outlier de la sèrie corresponent a l'atur

Un cop exposats els outliers, el model resultant, que és com s'ha vist en la secció d'identificació un $ARIMA(1,1,3)(0,1,2)$, es representa de la següent manera:

$$Z_t = \Pi(B)X_t + \Pi(B)\sum_{j=1}^{2}\omega_j\nu_j(B)I_{t,j}^{t_0}$$

on

$$\begin{aligned}\Pi(B) &= \frac{(1-B)(1-B^{12})(1-\phi_1 B)}{(1+\theta_1 B+\theta_2 B^2+\theta_3 B^3)(1+\vartheta_1 B^{12}+\vartheta_1 B^{24})} \\ &= \frac{(1-B)(1-B^{12})(1-0.9050B)}{(1-0.3910B-0.0425B^2-0.1820B^3)(1-0.3796B^{12}-0.1248B^{24})}\end{aligned}$$

$$\begin{aligned}\nu_1 &= 1 \\ \nu_2 &= \Pi(B) \\ \omega_1 &= 20674 \\ \omega_2 &= 83443 \\ t_0 &\in 291, 292\end{aligned}$$

Abans de fer les prediccion, serà necessari validar el model. Per fer-ho, s'analitzaran els residus com s'ha fet amb anterioritat amb la resta de models. Primerament és mostrarà el gràfic corresponent als mateixos:

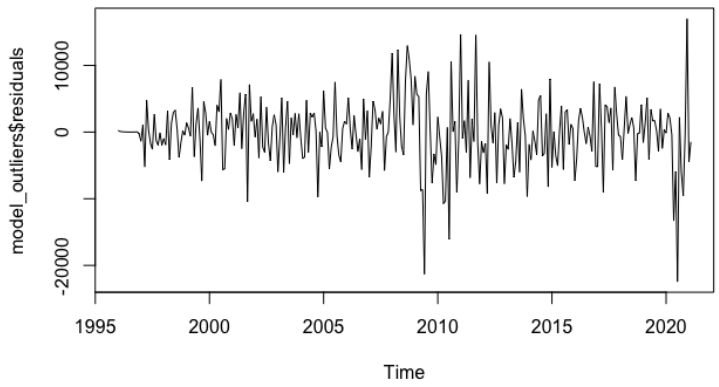

Figura 19: Gràfic dels residus de la sèrie corresponent a l'atur després d'aplicar l'anàlisi d'intervenció

Es pot apreciar com s'ha reduït la variabilitat en comparació amb els residus de la sèrie previs a l'aplicació de l'anàlisi d'intervenció. L'efecte de la COVID-19, tot i seguir present, no afecta de forma tant clara als residus. Aquest fet, farà possible trobar una distribució que s'ajusti adequadament a aquestes dades.

En quant a l'apartat dels tests, el de Box-Ljung ofereix un p-valor de 0.7914 i el de Kolmogorov-Smirnov un altre de 0.11. Per tant, podem concloure que els residus es tracten d'un soroll blanc gaussià i el model és vàlid.

Per últim, després d'haver dut a terme tots aquests procediments, serà interessant calcular una predicció de les observacions futures i comparar-les amb el model prèviament definit sense l'aplicació de l'anàlisi d'intervenció. Els resultats obtinguts han estat els següents:

	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre
2021	531130	572623	581865	583131	575885	589624	591509	602355	605715	614071

Taula 7: Prediccions de la sèrie corresponent a l'atur després de fer l'anàlisi d'intervenció

Amb la corresponent representació gràfica amb els intervals de confiança del 95%

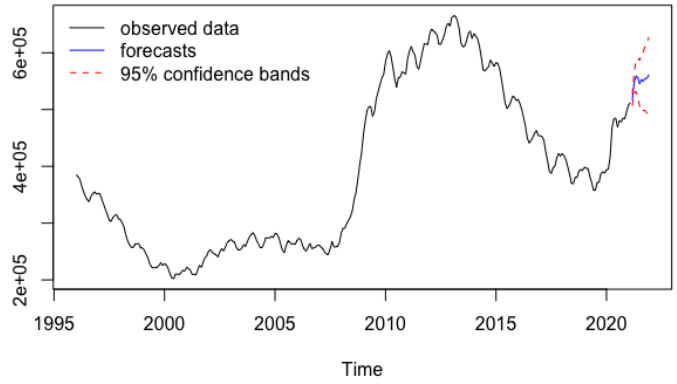

Figura 20: Gràfic de les prediccions de la sèrie corresponent a l'atur després d'aplicar l'anàlisi d'intervenció

Com s'ha comentat durant l'anàlisi dels outliers, l'efecte del segon outlier provoca un creixement en les observacions predites. Per això s'han obtingut resultats més alts que en les prediccions calculades amb anterioritat. Tot i això, no s'han produït alteracions tan grans com amb la sèrie corresponent a l'ocupació hotelera. De fet, els resultats de les dues prediccions d'aquesta sèrie segueixen la mateixa distribució com es pot veure a continuació:

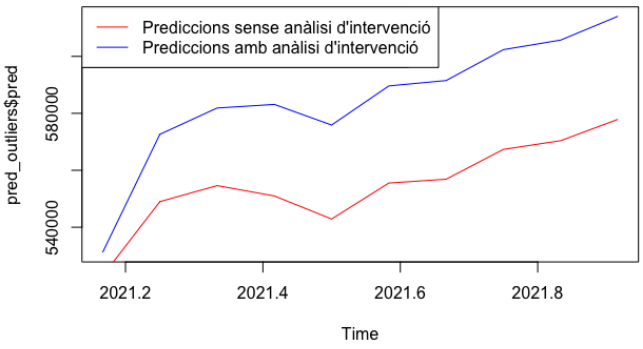

Figura 21: Gràfic de les prediccions de la sèrie corresponent a l'atur abans i després d'aplicar l'anàlisi d'intervenció

9 Conclusions

En primer lloc, en una situació com l'actual, on el COVID-19 ha afectat en gairebé la majoria de sectors tant econòmics com socials, resulta indispensable tindre una tècnica complementària a la metodologia habitual de Box-Jenkins per a poder analitzar de manera diferent les sèries temporals on qualsevol efecte extern a elles estigui present. La forma més adequada per a fer-ho ha estat la metodologia de Chen i Liu per a la recerca i tractament dels outliers.

Per això, era molt important portar a la pràctica tota la teoria proposada i poder apreciar l'abast real de les tècniques sobre sèries temporals amb dades reals. La selecció d'aquestes dues sèries temporals no ha estat aleatòria. En ambdues, tot i tindre outliers provocats per l'efecte de la COVID-19, l'impacte d'aquests és molt diferent. Mentre que per al grau d'ocupació hotelera, els outliers provoquen una reducció global en el valor de les observacions i una petita recuperació, per a l'atur generen únicament un augment en el valor de les dades.

Retornant a la qüestió de la necessitat de l'aplicació d'aquesta metodologia, aquest fet es pot apreciar en la secció de validació dels dos models. Obtindre uns residus amb valors extranys tan allunyats de la resta d'observacions provoca que no es pugui ajustar una distribució a les dades. I això suposa que el model, tot i que els seus residus siguin un soroll IID, no sigui del tot vàlid com s'ha explicat en la secció corresponent al tractament dels residus.

Un altre aspecte molt rellevant és la definició del model, ja que la recerca dels outliers és farà sobre el tipus de model ARIMA proposat. Per tant, canviarà el valor dels paràmetres pero mai la quantitat. Centrant-nos en les sèries analitzades, respecte a la que fa referència al grau d'ocupació hotelera, es pot veure que, com l'efecte dels outliers provoca alteracions tan grans sobre la sèrie, el valor dels coeficients ARIMA canvia de forma dràstica. D'altra banda, en l'anàlisi corresponent a la sèrie de l'atur, en el valor dels coeficients ARIMA s'han produït únicment petits canvis. Lògicament, quant més exagerats siguin els efectes dels outliers, més difícil serà ajustar correctament el model. Per això és més senzill obtindre resultats satisfactoris amb una sèrie com la segona.

Per últim, és important parlar de les prediccions. Aquestes, com s'ha vist, estan subjectes als outliers que s'hagin trobat i el model que s'hagi determinat com el que millor s'ajusta a les dades. En sèries com la corresponent a l'atur, on el valor dels coeficients del model ARIMA no ha canviat excessivament, l'efecte dels outliers és el que determina principalment el canvi en el valor de les noves observacions predites. Alternativament, en sèries com la corresponent al grau d'ocupació hotelera, on el valor dels paràmetres ARIMA ha canviat tant, les modificacions en les observacions resulten una conseqüència tant per l'efecte dels outliers com per els nous paràmetres calculats. Per això, per a series com aquesta última, és primordial haver fet una identificació correcta del model, ja que d'una altra forma, aquestes alteracions tan nombroses a la sèrie podrien provocar obtindre un efecte dels outliers semblant, però unes prediccions de futures observacions significativament diferents.

Amb tot això, podem dir que la metodologia de Chen i Liu, és una eina molt útil que en casos com els de les sèries analitzades, millora clarament l'ajustament del model. Tot i així, per a sèries com la que correspon al grau d'ocupació hotelera, on es produeixen grans modificacions, resulta molt difícil fer unes bones prediccions, però si que l'efecte dels outliers es pot descriure amb claredat. I la clau és aquesta, l'afectació dels outliers es

dibuixa amb certa efectivitat però si aquesta és massa gran, un petit error en l'identificació del model pot suposar grans errors en les prediccions.

Abans d'acabar, també resulta rellevant explicar que les prediccions obtingudes segur que no seran iguals a les que es donaran a la realitat. Però això està lluny del que es busca. Les prediccions han de servir com a estudi per entendre com pot evolucionar una sèrie, mai han de fer-se servir com a eina per estimar amb exactitud unes dades. D'aquesta forma, els resultats obtinguts podrien haver estat un suport per a organismes oficials que han de pendre decisions en quant aspectes tant importants com l'economia o el turisme.

10 Annexes

10.1 Anàlisi de Box-Jenkins de la sèrie corresponent a l'atur

10.1.1 Identificació

En primer lloc, es mostrarà la representació gràfica de les observacions presentades

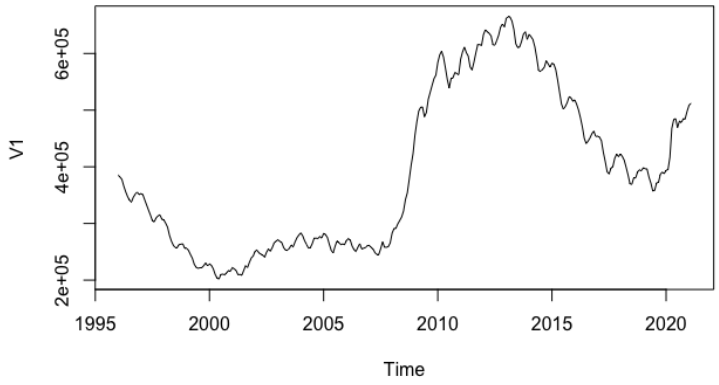

Figura 22: Gràfic amb les observacions de la sèrie corresponent a l'atur

amb els corresponents correlogrames de l'ACF i la PACF

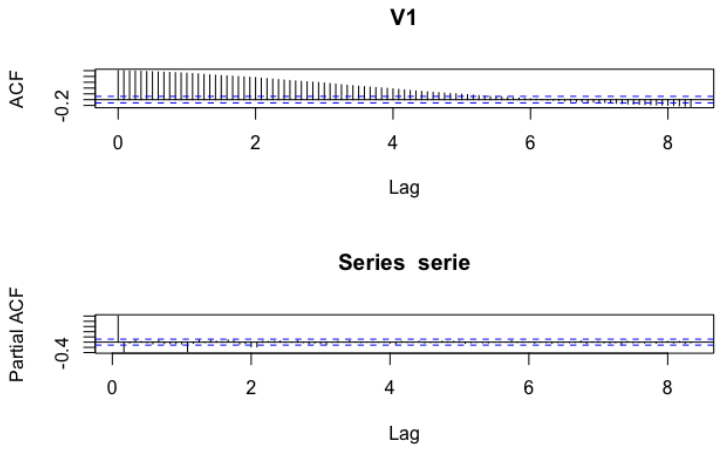

Figura 23: ACF i PACF de la sèrie corresponent a l'atur

En primer lloc, s'observa una estacionalitat clara de les dades, ja que als mesos corresponents a l'estiu, aquestes es redueixen significativament. D'altra banda, tot i que es produeixen creixements i decreixements de les observacions al llarg del seu recorregut, s'aprecia una tendència creixent si evaluem la sèrie de forma global de principi a fi.

Amb aquesta informació i els correlogrames exposat, sembla evident que la sèrie no és estacionària. S'aplicaran diferències regulars i estacionals per a resoldre aquest aspecte.

Un cop diferenciada la sèrie, s'obté la següent representació gràfica:

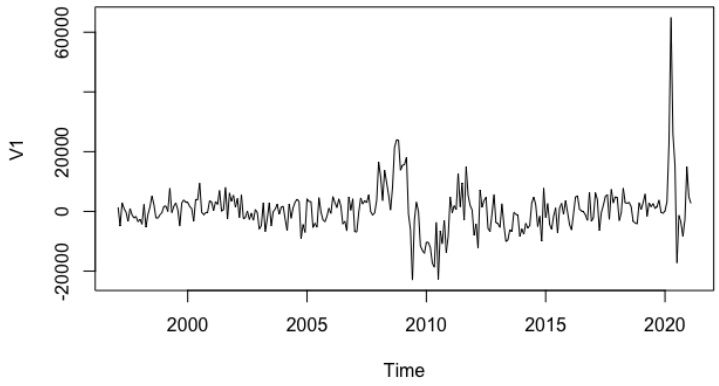

Figura 24: Representació gràfica de la sèrie corresponent a l'atur diferenciada

amb els corresponents correlogrames de l'ACF i el PACF

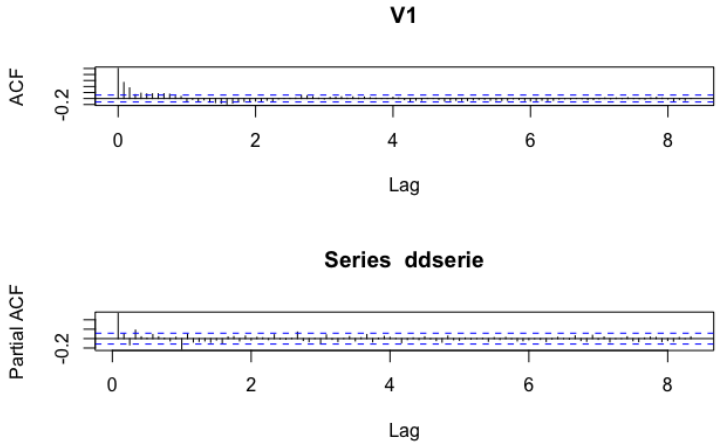

Figura 25: ACF i PACF de la sèrie corresponent a l'atur diferenciada

En primer lloc, la nova sèrie, tot i tindre uns resultats al voltant del 0, pateix un canvi de variància entre el 2008-2010 i té un pic destacadíssim a l'inici de 2020. Pel resultat dels correlogrames sembla estacionària, però aquestes variacions comentades podrien suposar un problema. S'estudiaran amb més detall aquests aspectes en la part corresponent a la validació del model.

En quant al model seleccionat, després d'evaluar l'ajustament d'alguns amb diferents paràmetres dels models ARIMA, s'ha conclòs que el que s'ajusta millor a les dades és un $ARIMA(1,1,3)(0,1,2)_{12}$ amb un AICc de valor 5866.2.

10.1.2 Estimació

El model estimat, com s'ha comentat prèviament, es tracta d'un $ARIMA(1,1,3)(0,1,2)_{12}$ amb els següents paràmetres:

$$(1-\phi_1 B)(1-B)(1-B^{12})X_t = Z_t(1+\theta_1 B+\theta_2 B^2+\theta_3 B^3)(1+\vartheta_1 B^{12}+\vartheta_2 B^{24})$$

on $\phi_1 = 0.9346$, $\theta_1 = -0.3389$, $\theta_2 = -0.1019$, $\theta_3 = -0.3165$, $\vartheta_1 = -0.4140$ i $\vartheta_3 = -0.1246$.

10.1.3 Validació

Per a validar el model, s'analitzaran els residus. La respresentació d'aquests és la següent:

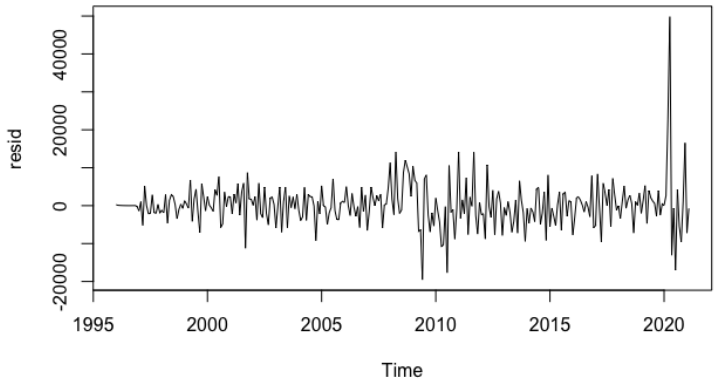

Figura 26: Gràfic corresponent als residus del model escollit per a la sèrie de l'atur

De la mateixa forma que amb la sèrie del grau d'ocupació hotelera, existeixen un conjunt de valors extranys a l'inici del 2020, que tot indica que poden ser una conseqüència de l'impacte de la COVID-19, que impedeixen ajustar una distribució a als residus. Tot i obtindre un p-valor pel Test de Shapiro-Wilk de 0.7162 indicant que es tracten d'un soroll IID, els tests de normalitat no es verifiquen i no s'ajusta cap distribució a les dades. Una possible solució al problema seria l'anàlisi d'intervenció.

10.1.4 Prediccions

Tot i que no s'hagi trobat una distribució que s'ajusti correctament als residus, a continuació és mostrarà la predicció de les observacions corresponents a la resta del 2021. Aquestes, seran interessants per a ser comparades posteriorment amb les obtingudes després d'aplicar l'anàlisi d'intervenció.

	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre
2021	522910	548986	554609	550988	542883	555522	556868	567336	570387	577880

Taula 8: Prediccions del nombre d'aturats dels mesos de Març fins Desembre de 2021

Que tenen la següent representació gràfica amb els intervals de confiancça del 95%

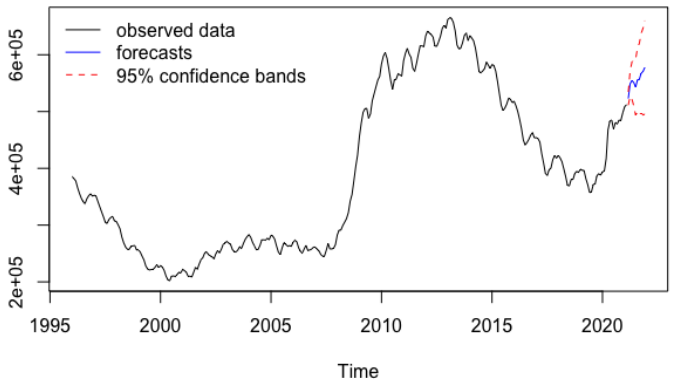

Figura 27: Gràfic corresponent les prediccions del model escollit per a la sèrie de l'atur

10.2 Script de la sèrie corresponent al grau d'ocupació hotelera

```
# Llibreries ------------------------------------------------------------

library(timeDate)
library(timeSeries)
library(fBasics)
library(aTSA)
library(forecast)
library(tseries)
library(ggplot2)
library(lmtest)
library(fUnitRoots)
library(FitARMA)
library(strucchange)
library(reshape)
library(Rmisc)
library(tsoutliers)
library(fitdistrplus)
library(car)

# Lectura de dades ------------------------------------------------------

setwd("/Users/alvaro.marlo/Documents/TFG")
par(mfrow=c(1,1))
serie <- ts(read.table("establiments_hotelers_ts.txt"),
            start = 1999, frequency = 12)
n <- length(serie)

# Anàlisi inicial -------------------------------------------------------

# Gràfic sèrie:
plot.ts(serie)

# Anàlisi gràfic:
plot(decompose(serie))

# Tendència:
plot(aggregate(serie))

# Estacionalitat:
boxplot(serie~cycle(serie))

# ACF i PACF:
par(mfrow=c(2,1))
acf(serie, lag= 100)
pacf(serie, lag=100)

# Recomanacions de diferenciacions:
```

```
ndiffs(serie)
nsdiffs(serie)

# Sèrie no estacionària
# En primer lloc cal aplicar logarítmes ja que la variància augmenta
# al llarg del temps
# Clara estacionalitat i tend?ncia positiva
# S'aplicaran diferencies regulars i estacionals per a tindre
# una sèrie estacionària


# Logaritmes ------------------------------------------------------------
lserie <- log(serie)
lserie

# Gràfics
par(mfrow=c(1,1))
plot(lserie)
plot(decompose(lserie))

# Tendència:
plot(aggregate(lserie))

# Estacionalitat:
boxplot(serie~cycle(lserie))

# ACF i PACF
par(mfrow=c(2,1))
acf(lserie, lag=100)
pacf(lserie, lag=100)

# Recomanacions de diferenciacions:
ndiffs(lserie)
nsdiffs(lserie)

# No és estacionària
# Segueix havent tendència positiva i estacionalitat
# S'aplicaran per tant diferències regulars i estacionals

# Diferències regulars --------------------------------------------------
dserie <- diff(lserie)

# Gràfics
par(mfrow=c(1,1))
plot(dserie)
plot(decompose(dserie))

# ACF i PACF
par(mfrow=c(2,1))
```

```
acf(dserie, lag=100)
pacf(dserie, lag=100)

# S'ha eliminat la tendència
par(mfrow=c(1,1))
plot(aggregate(dserie))

# La sèrie segueix sense ser del tot estacion?ria
# Però ja no tenim tendència

# Diferències estacionals ------------------------------------------------
ddserie <- diff(dserie, lag = 12)

# Gràfics
par(mfrow=c(1,1))
plot(ddserie)
plot(decompose(ddserie))

# ACF i PACF
par(mfrow=c(2,1))
acf(ddserie, lag=100)
pacf(ddserie, lag =100)

# S'ha eliminat la estacionalitat
par(mfrow=c(1,1))
boxplot(ddserie~cycle(ddserie))

# Eliminada tendència i estacionalitat
# Sembla estacionària per el ACF i el PACF

# Selecció model ---------------------------------------------------------

# S'utilitza la funció auto.arima per estimar quin podria ser el millor model
model <- auto.arima(lserie, ic = "aicc", trace = TRUE,
                    d = 1, D = 1, method = "CSS-ML")
model

# Després de provar amb altres models es decideix quedar-se
# l'ARIMA(2,1,1)(0,1,1)[12] ja que és el que oferia un AICc menor
# després d'aplicar l'anàlisi d'intervenció

# L'estimació dels paràmetres és la següent
# AR1 = 0.7557
# AR2 = -0.3265
# MA1 = -0.8460
# SMA1 = -0.2977
model <- arima(lserie,
               order = c(2,1,1),
               seasonal = list(order = c(0,1,1), period =12),
```

```
               method = "CSS-ML")
model

# Residus ----------------------------------------------------------------
resid <- model$residuals

# Residus tipificats
resid_tip <- scale(resid, center = T, scale = T)

# Gràfics
par(mfrow=c(1,1))
plot(resid)
tsdiag(model)

# ACF i PACF
par(mfrow=c(2,1))
acf(resid, lag = 100)
pacf(resid,lag = 100)

# Soroll IID: SI
Box.test(resid, type="Ljung-Box")

# Normalitat: NO
qqnorm(resid)
qqline(resid)
shapiro.test(resid)

# Histograma
par(mfrow=c(1,1))
hist(resid_tip, prob = TRUE, breaks = 50)
x <- seq(min(resid_tip), max(resid_tip), length = n)
f <- dnorm(x, mean = mean(resid_tip), sd = sd(resid_tip))
lines(x, f, col = "red", lwd = 2)

# Test Kolmogorov Smirnov amb t-Student (No s'ajusta)
pvalues = vector()
for (i in (1:30)) {
  test <- ks.test(resid_tip, "pt", df = i)
  pvalues <- append(pvalues, test$p.value)
}
pvalues
barplot(pvalues)

# Prediccions ------------------------------------------------------------

pred <- predict(model, n.ahead =11)

# Valors predits
p <- pred$pred
```

```
# Intervals 95%
p_upper <- pred$pred + 1.96 * pred$se
p_lower <- pred$pred - 1.96 * pred$se

# Gràfic
par(mfrow=c(1,1))
plot(cbind(lserie, p_lower), plot.type = "single", ylab = "", type = "n")
lines(lserie)
lines(p, type = "l", col = "blue")
lines(p_upper, type = "l", col = "red", lty = 2)
lines(p_lower, type = "l", col = "red", lty = 2)
legend("bottomleft", legend = c("observed data",
                                "forecasts", "95% confidence bands"),
       lty = c(1,1,2,2), col = c("black", "blue", "red", "red"), bty = "n")

# S'ha vist que en les últimes observacions hi han potencials outliers així que
# es farà un anàlisi d'intervenció

# Recerca outliers ----------------------------------------------------

# Detecció outliers
outliers <- tso(lserie, types = c("TC", "AO", "LS", "IO"), tsmethod = "arima",
                discard.method = "en-masse",
                args.tsmethod = list(order = c(2, 1, 1),
                                     seasonal = list(order = c(0, 1, 1),
                                                     period = 12)))
outliers
plot(outliers) # representació gràfica
outliers$outliers # outliers trobats
outliers$fit # model

# Vector amb els tipus d'outliers
outliers_type <- outliers$outliers$type
# Outlier 1 -> Outlier additiu (AO)
# Outlier 2 -> Outlier innovacional (IO)
# Outlier 3 -> Canvi temporal (TC)
# Outlier 4 -> Canvi de nivell (LS)
# Outlier 5 -> Outlier innovacional (IO)
# Outlier 6 -> Canvi temporal (TC)
# Outlier 7 -> Outlier innovacional (IO)
# Outlier 8 -> Outlier additiu (AO)
# Outlier 9 -> Canvi de nivell (LS)
# Outlier 10 -> Outlier additiu (AO)


# Vector amb els index a la sèrie temporal dels outliers
outliers_ind <- outliers$outliers$ind
# Outlier 1 -> Observació 255
```

```
# Outlier 2 -> Observació 256
# Outlier 3 -> Observació 257
# Outlier 4 -> Observació 258
# Outlier 5 -> Observació 260
# Outlier 6 -> Observació 261
# Outlier 7 -> Observació 262
# Outlier 8 -> Observació 263
# Outlier 9 -> Observació 264
# Outlier 10 -> Observació 265

# Vector amb el temps dels outliers
outliers_time <- outliers$outliers$time
# Outlier 1 -> Març del 2020
# Outlier 2 -> Abril del 2020
# Outlier 3 -> Maig del 2020
# Outlier 4 -> Juny del 2020
# Outlier 5 -> Juliol del 2020
# Outlier 6 -> Agost del 2020
# Outlier 7 -> Setembre del 2020
# Outlier 8 -> Octubre del 2020
# Outlier 9 -> Novembre del 2020
# Outlier 10 -> Desembre del 2020


# Anàlisi outliers ----------------------------------------------------

# OUTLIER 1
outlier_1 <- outliers(outliers_type[1], outliers_ind[1])
o1 <- outliers.effects(outlier_1, n)

# Coeficient de l'outlier
coefhat_o1 <- outliers$outliers$coefhat[1]

# Efecte sobre la sèrie temporal de l'outlier
o1_effect <- coefhat_o1*o1

# Convertim l'efecte en una sèrie temporal
o1_effect_ts <- ts(o1_effect, frequency = frequency(serie),
                   start = start(serie))

# Plot de l'efecte
par(mfrow=c(1,1))
plot(o1_effect_ts)

# Outlier additiu amb coeficient -1.2374 que afecta al Març de 2020

# OUTLIER 2
outlier_2 <- outliers(outliers_type[2], outliers_ind[2])
o2 <- outliers.effects(outlier_2, n, pars = coefs2poly(outliers$fit))
```

```
# Coeficient de l'outlier
coefhat_o2 <- outliers$outliers$coefhat[2]

# Efecte sobre la sèrie temporal de l'outlier
o2_effect <- coefhat_o2*o2

# Convertim l'efecte en una sèrie temporal
o2_effect_ts <- ts(o2_effect, frequency = frequency(serie),
                   start = start(serie))

# Plot de l'efecte
plot(o2_effect_ts)

# Outlier innovacional de coeficient -7.5423 que comença el seu efecte
# a l'Abril del 2020

# OUTLIER 3

outlier_3 <- outliers(outliers_type[3], outliers_ind[3])
o3 <- outliers.effects(outlier_3, n)

# Coeficient de l'outlier
coefhat_o3 <- outliers$outliers$coefhat[3]

# Efecte sobre la sèrie temporal de l'outlier
o3_effect <- coefhat_o3*o3

# Convertim l'efecte en una sèrie temporal
o3_effect_ts <- ts(o3_effect, frequency = frequency(serie),
                   start = start(serie))

# Plot de l'efecte
plot(o3_effect_ts)

# Outlier de canvi temporal de coeficient -3.4606 que comença el seu efecte
# al Maig del 2020

# OUTLIER 4

outlier_4 <- outliers(outliers_type[4], outliers_ind[4])
o4 <- outliers.effects(outlier_4, n)

# Coeficient de l'outlier
coefhat_o4 <- outliers$outliers$coefhat[4]

# Efecte sobre la sèrie temporal de l'outlier
o4_effect <- coefhat_o4*o4
```

```
# Convertim l'efecte en una sèrie temporal
o4_effect_ts <- ts(o4_effect, frequency = frequency(serie),
                   start = start(serie))

# Plot de l'efecte
plot(o4_effect_ts)

# Outlier de canvi de nivell de coeficient 2.5783 que comença el seu efecte
# al Juny de 2020

# OUTLIER 5

outlier_5 <- outliers(outliers_type[5], outliers_ind[5])
o5 <- outliers.effects(outlier_5, n, pars = coefs2poly(outliers$fit))

# Coeficient de l'outlier
coefhat_o5 <- outliers$outliers$coefhat[5]

# Efecte sobre la sèrie temporal de l'outlier
o5_effect <- coefhat_o5*o5

# Convertim l'efecte en una sèrie temporal
o5_effect_ts <- ts(o5_effect, frequency = frequency(serie),
                   start = start(serie))

# Plot de l'efecte
plot(o5_effect_ts)

# Outlier innovacional de coeficient 0.2657 que comença el seu efecte
# al Juliol del 2020.

# OUTLIER 6

outlier_6 <- outliers(outliers_type[6], outliers_ind[6])
o6 <- outliers.effects(outlier_6, n)

# Coeficient de l'outlier
coefhat_o6 <- outliers$outliers$coefhat[6]

# Efecte sobre la sèrie temporal de l'outlier
o6_effect <- coefhat_o6*o6

# Convertim l'efecte en una sèrie temporal
o6_effect_ts <- ts(o6_effect, frequency = frequency(serie),
                   start = start(serie))

# Plot de l'efecte
plot(o6_effect_ts)
```

```
# Outlier de canvi temporal de coeficient 0.0739 que comença el seu efecte
# a l'Agost del 2020.

# OUTLIER 7

outlier_7 <- outliers(outliers_type[7], outliers_ind[7])
o7 <- outliers.effects(outlier_7, n, pars = coefs2poly(outliers$fit))

# Coeficient de l'outlier
coefhat_o7 <- outliers$outliers$coefhat[7]

# Efecte sobre la sèrie temporal de l'outlier
o7_effect <- coefhat_o7*o7

# Convertim l'efecte en una sèrie temporal
o7_effect_ts <- ts(o7_effect, frequency = frequency(serie),
                   start = start(serie))

# Plot de l'efecte
plot(o7_effect_ts)

# Outlier innovacional de coeficient -0.7647 que comença el seu efecte
# al Setembre del 2020.

# OUTLIER 8

outlier_8 <- outliers(outliers_type[8], outliers_ind[8])
o8 <- outliers.effects(outlier_8, n)

# Coeficient de l'outlier
coefhat_o8 <- outliers$outliers$coefhat[8]

# Efecte sobre la sèrie temporal de l'outlier
o8_effect <- coefhat_o8*o8

# Convertim l'efecte en una sèrie temporal
o8_effect_ts <- ts(o8_effect, frequency = frequency(serie),
                   start = start(serie))

# Plot de l'efecte
plot(o8_effect_ts)

# Outlier additiu de coeficient -1.2940 que afecta al Octubre del 2020.

# OUTLIER 9

outlier_9 <- outliers(outliers_type[9], outliers_ind[9])
o9 <- outliers.effects(outlier_9, n)
```

```
# Coeficient de l'outlier
coefhat_o9 <- outliers$outliers$coefhat[9]

# Efecte sobre la sèrie temporal de l'outlier
o9_effect <- coefhat_o9*o9

# Convertim l'efecte en una sèrie temporal
o9_effect_ts <- ts(o9_effect, frequency = frequency(serie),
                   start = start(serie))

# Plot de l'efecte
plot(o9_effect_ts)

# Outlier canvi de nivell de coeficient -1.649 que comença el seu efecte
# al Novembre del 2020.

# OUTLIER 10

outlier_10 <- outliers(outliers_type[10], outliers_ind[10])
o10 <- outliers.effects(outlier_10, n)

# Coeficient de l'outlier
coefhat_o10 <- outliers$outliers$coefhat[10]

# Efecte sobre la sèrie temporal de l'outlier
o10_effect <- coefhat_o10*o10

# Convertim l'efecte en una sèrie temporal
o10_effect_ts <- ts(o10_effect, frequency = frequency(serie),
                    start = start(serie))

# Plot de l'efecte
plot(o10_effect_ts)

# Outlier additiu de coeficient 0.2878 que afecta al Desembre del 2020.

# Disseny nou model ----------------------------------------------------

# S'ha estimat que és un model SARIMA(2,1,1)(0,1,1)
# AR1 = -0.0949
# AR2 = 0.0654
# MA1 = -0.6661
# SMA1 = -0.9997
model_outliers <- outliers$fit
model_outliers

# Comprovació residus --------------------------------------------------

resid <- model_outliers$residuals
```

```
# Residus tipificats
resid_tip <- scale(resid, center = T, scale = T)

# Gràfics
par(mfrow=c(1,1))
plot(resid)
tsdiag(model)

# ACF i PACF
par(mfrow=c(2,1))
acf(resid, lag = 100)
pacf(resid,lag = 100)

# Soroll IID: S?
Box.test(resid, type="Ljung-Box")

# Normalitat: NO
qqnorm(resid)
qqline(resid)
shapiro.test(resid_tip)
ks.test(resid_tip, "pnorm")

# Predicció ------------------------------------------------------------

# Efecte dels outliers en les prediccions (11 futures observacions)

# Outlier 1:
forecast_o1 <- outliers.effects(outlier_1,n+11)[266:276]
forecast_o1

# Outlier 2:
forecast_o2 <- outliers.effects(outlier_2,n+11,
                                pars = coefs2poly(model_outliers))[266:276]
forecast_o2

# Outlier 3:
forecast_o3 <- outliers.effects(outlier_3,n+11)[266:276]
forecast_o3

# Outlier 4:
forecast_o4 <- outliers.effects(outlier_4,n+11)[266:276]
forecast_o4

# Outlier 5:
forecast_o5 <- outliers.effects(outlier_5,n+11,
                                pars = coefs2poly(model_outliers))[266:276]
forecast_o5
```

```
# Outlier 6:
forecast_o6 <- outliers.effects(outlier_6,n+11)[266:276]
forecast_o6

# Outlier 7:
forecast_o7 <- outliers.effects(outlier_7,n+11,
                                pars = coefs2poly(model_outliers))[266:276]
forecast_o7

# Outlier 8:
forecast_o8 <- outliers.effects(outlier_8,n+11)[266:276]
forecast_o8

# Outlier 9:
forecast_o9 <- outliers.effects(outlier_9,n+11)[266:276]
forecast_o9

# Outlier 10:
forecast_o10 <- outliers.effects(outlier_10,n+11)[266:276]
forecast_o10

# Efecte futur dels outliers
regressors <- cbind(forecast_o1, forecast_o2, forecast_o3, forecast_o4,
                    forecast_o5, forecast_o6, forecast_o7, forecast_o8,
                    forecast_o9, forecast_o10)

# Utilitzem predict
pred <- predict(outliers$fit, n.ahead = 11, newxreg = regressors)

# Valors predits
p <- pred$pred

# Intervals 95%
p_upper <- pred$pred + 1.96 * pred$se
p_lower <- pred$pred - 1.96 * pred$se

# Gràfic
par(mfrow=c(1,1))
plot(cbind(lserie, p_lower), plot.type = "single", ylab = "", type = "n")
lines(lserie)
lines(p, type = "l", col = "blue")
lines(p_upper, type = "l", col = "red", lty = 2)
lines(p_lower, type = "l", col = "red", lty = 2)
legend("bottomleft", legend = c("observed data",
                                "forecasts", "95% confidence bands"),
       lty = c(1,1,2,2), col = c("black", "blue", "red", "red"), bty = "n")

# Prediccions finals (exp)
p <- exp(p)
```

```
p_upper <- exp(p_upper)
p_lower <- exp(p_lower)

# Gràfic definitiu
par(mfrow=c(1,1))
plot(cbind(serie, p_lower), plot.type = "single", ylab = "", type = "n")
lines(serie)
lines(p, type = "l", col = "blue")
lines(p_upper, type = "l", col = "red", lty = 2)
lines(p_lower, type = "l", col = "red", lty = 2)
legend("topleft", legend = c("observed data",
                             "forecasts", "95% confidence bands"),
       lty = c(1,1,2,2), col = c("black", "blue", "red", "red"), bty = "n")

# Ampliació ------------------------------------------------------------
forecast_effects_o1 <- coefhat_o1 * forecast_o1
forecast_effects_o2 <- coefhat_o2 * forecast_o2
forecast_effects_o3 <- coefhat_o3 * forecast_o3
forecast_effects_o4 <- coefhat_o4 * forecast_o4
forecast_effects_o5 <- coefhat_o5 * forecast_o5
forecast_effects_o6 <- coefhat_o6 * forecast_o6
forecast_effects_o7 <- coefhat_o7 * forecast_o7
forecast_effects_o8 <- coefhat_o8 * forecast_o8
forecast_effects_o9 <- coefhat_o9 * forecast_o9
forecast_effects_o10 <- coefhat_o10 * forecast_o10

forecast_effects <- cbind(forecast_effects_o1,
                          forecast_effects_o2,
                          forecast_effects_o3,
                          forecast_effects_o4,
                          forecast_effects_o5,
                          forecast_effects_o6,
                          forecast_effects_o7,
                          forecast_effects_o8,
                          forecast_effects_o9,
                          forecast_effects_o10)


forecast_effects
total_effects <- rowSums(forecast_effects)
total_effects <- ts(total_effects, frequency = 12, start = c(2021,2))
plot(total_effects)
```

10.3 Script de la sèrie corresponent a l'atur

```
# Llibreries ------------------------------------------------------------

library(timeDate)
library(timeSeries)
library(fBasics)
library(aTSA)
library(forecast)
library(tseries)
library(ggplot2)
library(lmtest)
library(fUnitRoots)
library(FitARMA)
library(strucchange)
library(reshape)
library(Rmisc)
library(tsoutliers)
library(fitdistrplus)
library(actuar)
library(car)

# Lectura de dades ------------------------------------------------------

setwd("/Users/alvaro.marlo/Documents/TFG")
par(mfrow=c(1,1))
serie <- ts(read.table("atur_ts.txt"), start = 1996, frequency = 12)
n <- length(serie)

# Anàlisi inicial -------------------------------------------------------

plot.ts(serie)

# Anàlisi gràfic
plot(decompose(serie))

# Pendent
plot(aggregate(serie))

# Estacionalitat
boxplot(serie~cycle(serie))

# ACF i PACF
par(mfrow=c(2,1))
acf(serie, lag= 100)
pacf(serie, lag=100)

# Recomanacions
ndiffs(serie)
```

```
nsdiffs(serie)

# Sèrie no estacionària
# Clara estacionalitat i pendent positiu
# S'aplicaran diferencies regulars i estacionals per
# a tindre una sèrie estacionària

# Diferències regulars ------------------------------------------------
dserie <- diff(serie)

# Gràfics
par(mfrow=c(1,1))
plot(dserie)
plot(decompose(dserie))

# ACF i PACF
par(mfrow=c(2,1))
acf(dserie, lag=100)
pacf(dserie, lag=100)

# S'ha eliminat la tendència
par(mfrow=c(1,1))
plot(aggregate(dserie))

# La sèrie segueix sense ser estacionària
# Però ja no tenim tendència
# Presència d'un pic cap al 2008

# Diferències estacionals ---------------------------------------------

# Fem diferències estacionals per extreure el cicle
ddserie <- diff(dserie, lag = 12)
par(mfrow=c(1,1))
plot(ddserie)

# ACF i PACF
par(mfrow=c(2,1))
acf(ddserie, lag=100)
pacf(ddserie, lag =100)

# Eliminació estacionalitat
par(mfrow=c(1,1))
boxplot(ddserie~cycle(ddserie))

# Recomanacions
ndiffs(ddserie)
nsdiffs(ddserie)
auto.arima(ddserie, ic = 'aicc')
```

```
# Eliminada tendència i estacionalitat
# Sembla estacionària per el ACF i el PACF

# Presència d'outliers i pics cap al 2009 i 2020
par(mfrow=(c(1,1)))
plot(ddserie)
plot(aggregate(ddserie))
boxplot(ddserie~cycle(ddserie))

# Selecció model ----------------------------------------------------------

# S'utilitza la funció auto.arima per estimar quin podria ser el millor model
model <- auto.arima(serie, ic = "aicc", trace = TRUE,
                    d = 1, D = 1, method = "CSS-ML")
model

# SARIMA(1,1,3)(0,1,2)[12] amb par?metres
# AR1 = 0.9346
# MA1 = -0.3389
# MA2 = -0.1019
# MA3 = -0.3165
# SMA1 = -0.4140
# SMA2 = -0.1246
model <- Arima(serie,
               order = c(1,1,3),
               seasonal = list(order = c(0,1,2), period =12),
               method = "CSS-ML")

model

# Residus -----------------------------------------------------------------
resid <- model$residuals

# Residus tipificats
resid_tip <- scale(resid, center = T, scale = T)

# Gràfics
par(mfrow=c(1,1))
plot(resid)
tsdiag(model)

# ACF i PACF
par(mfrow=c(2,1))
acf(resid, lag = 100)
pacf(resid,lag = 100)

# Soroll IID: SI
Box.test(resid, type="Ljung-Box")
```

```
# Normalitat: NO
qqnorm(resid)
qqline(resid)
shapiro.test(resid)

# Histograma
par(mfrow=c(1,1))
hist(resid_tip, prob = TRUE, breaks = 50)
x <- seq(min(resid_tip), max(resid_tip), length = n)
f <- dnorm(x, mean = mean(resid_tip), sd = sd(resid_tip))
lines(x, f, col = "red", lwd = 2)

# Test Kolmogorov Smirnov amb t-Student (No s'ajusta)
pvalues = vector()
for (i in (1:30)) {
  test <- ks.test(resid_tip, "pt", df = i)
  pvalues <- append(pvalues, test$p.value)
}
pvalues
barplot(pvalues)

# Prediccions ------------------------------------------------------------

pred <- predict(model, n.ahead = 10)
pred

# Valors predits
p <- pred$pred

# Intervals 95%
p_upper <- pred$pred + 1.96 * pred$se
p_lower <- pred$pred - 1.96 * pred$se

# Gràfic
par(mfrow=c(1,1))
plot(cbind(serie, p_lower), plot.type = "single", ylab = "", type = "n")
lines(serie)
lines(p, type = "l", col = "blue")
lines(p_upper, type = "l", col = "red", lty = 2)
lines(p_lower, type = "l", col = "red", lty = 2)
legend("topleft", legend = c("observed data",
                             "forecasts", "95% confidence bands"),
       lty = c(1,1,2,2),col = c("black", "blue", "red", "red"), bty = "n")

# S'ha vist que hi han potencials outliers cap al 2008 i a l'inici del 2020

# Recerca outliers -------------------------------------------------------

# Busquem els outliers de la sèrie
```

```
outliers <- tso(serie, types = c("TC", "AO", "LS", "IO"), tsmethod = "arima",
                discard.method = "en-masse",
                args.tsmethod = list(order = c(1,1,3),
                                     seasonal = list(order = c(0,1,2),
                                                     period =12)))
model
outliers # representació gràfica
plot(outliers) # outliers trobats
outliers$fit # model

# Vector amb els tipus d'outliers
outliers_type <- outliers$outliers$type
# Outlier 1 -> Canvi temporal (TC)
# Outlier 2 -> Outlier innovacional (IO)

# Vector amb els index a la sèrie temporal dels outliers
outliers_ind <- outliers$outliers$ind
# Outlier 1 -> Observació 162
# Outlier 2 -> Observació 292

# Vector amb el temps dels outliers
outliers_time <- outliers$outliers$time
# Outlier 1 -> Juny del 2009
# Outlier 2 -> Abril del 2020

# Anàlisi outliers ------------------------------------------------------

# OUTLIER 1

outlier_o1 <- outliers(outliers_type[1], outliers_ind[1])
o1 <- outliers.effects(outlier_o1, n)

# Coeficient de l'outlier
coefhat_o1 <- outliers$outliers$coefhat[1]

# Efecte sobre la s?rie temporal de l'outlier
o1_effect <- coefhat_o1*o1

# Convertim l'efecte en una s?rie temporal
o1_effect_ts <- ts(o1_effect, frequency = frequency(serie),
                   start = start(serie))

# Plot de l'efecte
par(mfrow=c(1,1))
plot(o1_effect_ts)

# Outlier additiu amb coeficient 20674 que comença el seu efecte
#al Març de 2020.
```

```
# OUTLIER 2

outlier_o2 <- outliers(outliers_type[2], outliers_ind[2])
o2 <- outliers.effects(outlier_o2, n, pars = coefs2poly(outliers$fit))

# Coeficient de l'outlier
coefhat_o2 <- outliers$outliers$coefhat[2]

# Efecte sobre la s?rie temporal de l'outlier
o2_effect <- coefhat_o2*o2

# Convertim l'efecte en una sèrie temporal
o2_effect_ts <- ts(o2_effect, frequency = frequency(serie),
                   start = start(serie))

# Plot de l'efecte
plot(o2_effect_ts)

# Outlier innovacional de coeficient 83443 que comença el seu efecte
# a l'Abril del 2020

# Disseny nou model ----------------------------------------------------

# S'ha estimat que és un model SARIMA(1,1,3)(0,1,2)
outliers

# Creació model

model_outliers <- outliers$fit

# Comprovació residus --------------------------------------------------

resid <- model_outliers$residuals

# Residus tipificats
resid_tip <- scale(resid, center = T, scale = T)

# Gràfics
par(mfrow=c(1,1))
plot(resid)
tsdiag(model)

# ACF i PACF
par(mfrow=c(2,1))
acf(resid, lag = 100)
pacf(resid,lag = 100)

# Soroll IID: S?
Box.test(resid, type="Ljung-Box")
```

```
# Normalitat: NO
qqnorm(resid)
qqline(resid)
shapiro.test(resid_tip)
ks.test(resid_tip, "pnorm")

# Predicció ------------------------------------------------------------
# Predicció TC
forecast_o1 <- outliers.effects(outlier_o1,n+12)[303:312]

# Predicció IO
forecast_o2 <- outliers.effects(outlier_o2,n+12,
                                pars = coefs2poly(model_outliers))[303:312]

regressors = cbind(forecast_o1, forecast_o2)
regressors

pred_outliers <- predict(model_outliers, n.ahead = 10, newxreg = regressors)
pred

# Valors predits
p <- pred_outliers$pred

# Intervals 95%
p_upper <- pred_outliers$pred + 1.96 * pred_outliers$se
p_lower <- pred_outliers$pred - 1.96 * pred_outliers$se

# Gràfic
par(mfrow=c(1,1))
plot(cbind(serie, p_upper), plot.type = "single", ylab = "", type = "n")
lines(serie)
lines(p, type = "l", col = "blue")
lines(p_upper, type = "l", col = "red", lty = 2)
lines(p_lower, type = "l", col = "red", lty = 2)
legend("topleft", legend = c("observed data",
                             "forecasts", "95% confidence bands"),
       lty = c(1,1,2,2), col = c("black", "blue", "red", "red"), bty = "n")
pred

# Comparació prediccions -----------------------------------------------

par(mfrow=c(1,1))
plot(pred_outliers$pred, col = "blue")
lines(pred$pred, col = "red")
legend("topleft", legend = c("Prediccions sense anàlisi d'intervenció",
                             "Prediccions amb anàlisi d'intervenció"),
       lty=c(1,1), col = c("red","blue"))
```

Referències

[1] Brockwell, P.J.; Davis, R.A.: *Time Series: Theory and Methods*, 2a ed., Nova York, Springer, 1991, ISBN 978-1-4419-0320-4.

[2] Cowpertwait, P.S.P; Metcalfe, A.V.: *Introductory Time Series with R*, 1a ed., Nova York, Springer, 2009, ISBN 978-0-387-88698-5.

[3] Box, G.E.P.; Tiao G.C.: Intervention Analysis with Applications to Economic and Environmental Problems, *Journal of the American Statistical Association*, vol. 70, no. 349,Maig de 1975, pp. 70-79, `www.jstor.org/stable/2285379`.

[4] Chen, C; Liu, L.M.: Joint Estimation of Model Parameters and Outlier Effects in Time Series, *Journal of the American Statistical Association*, vol. 88, no. 421, Maig de 1993, pp. 284-297, `www.jstor.org/stable/2290724`.

[5] López de Lacalle, J.: *tsoutliers: Detection of Outliers in Time Series*, 2019, `cran.r-project.org/web/packages/tsoutliers/tsoutliers.pdf`.

[6] Hyndman, R; Athanasopoulos, G; Bergmeir, C; Caceres, G; Chhay L; O'Hara-Wild, M; Petropoulos, F; Razbash, S; Wang, E; Yasmeen, F: *forecast: Forecasting functions for time series and linear models*, 2021, `cran.r-project.org/web/packages/forecast/forecast.pdf`.

[7] Jaimes Berrios, A.A.; Quintanilla Aparicio, M.I.: *Análisis de series temporales con outliers e intervenciones y sus aplicaciones(TFG)*, Universidad de El Salvador, Facultad de Ciencias Naturales y Matemática, Març de 2008, `ri.ues.edu.sv/id/eprint/12816/1/19200760.pdf`.

[8] Vives Santa Eulàlia, J.: *Apunts de l'assignatura Anàlisi de Sèries Temporals*, Màster de Fonaments de la Ciència de Dades, Universitat de Barcelona.