UNIVERSITAT DE BARCELONA
Facultat de Matemàtiques i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

SÈRIES TEMPORALS AMB MEMÒRIA LLARGA: MODELS ARFIMA

Autor: Sergio Muñoz Martínez

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

Barcelona, 12 de juny de 2023

Abstract

Time series analysis plays a vital role in understanding various phenomena in a wide variety of fields. Over the years, more complex statistical models have been developed that allow us to analyze time series in a more accurate way. This project studies in depth one of them: the $ARFIMA$ models. These models are very useful for examining long memory time series. The main purpose of this project is to provide a comprehensive analysis of these models, including their theoretical foundations, estimation methods and a practical implementation. In particular, the R programming language is used to evaluate the possibility of modeling the time series that shows the air quality of Barcelona between 2017 and 2019 using an $ARFIMA$ model.

Resum

L'anàlisi de les sèries temporals pren un rol vital a l'hora de comprendre diversos fenomens en disciplines molt variades. Al llarg dels anys s'han anat desenvolupant models estadístics més complexos que ens permeten analitzar d'una manera més acurada les sèries temporals. En aquest treball s'estudia a fons un d'ells: els models $ARFIMA$. Aquests models són de gran utilitat per examinar les sèries temporals que presenten una memòria llarga. L'objectiu d'aquest treball és proporcionar una anàlisi exhaustiva d'aquests models, incloent-ne les bases teòriques matemàtiques, mètodes d'estimació i una implementació pràctica. En concret, es fa ús del llenguatge de programació R per tal d'avaluar la possibilitat de modelitzar la sèrie temporal que ens mostra la qualitat de l'aire de Barcelona entre els anys 2017 i 2019 amb un model $ARFIMA$.

2020 Mathematics Subject Classification. 60G10, 60G22, 62M10

Agraïments

En primer lloc, vull agrair als meus pares tots els sacrificis i esforços que han realitzat per poder donar-me l'oportunitat d'estudiar allò que he volgut. No han sigut uns anys fàcils però, tot i els entrebancs en el camí, sempre han confiat en mi i han estat allà per recolzar-me.

També agraeixo al meu tutor, en Josep Vives, per haver-me orientat en la tria de la temàtica final del treball i pels consells donats en l'elaboració del mateix.

Per últim, vull agrair a tots els meus amics per donar-me suport, celebrant els meus èxits i ajudant-me quan els he necessitat.

Índex

1 Introducció **1**

2 Conceptes previs **3**

2.1 Processos estocàstics i processos estacionaris 3

2.2 Anàlisi clàssica de les sèries temporals 6

2.3 Exemples 8

3 Models lineals de sèries temporals **10**

3.1 Models autorregressius (AR) 11

3.2 Models mitjana mòbil (MA) 12

3.3 Models mixtos (ARMA) 13

3.4 Models integrats (ARIMA) 14

3.5 Models estacionals (SARIMA) 14

4 Models ARFIMA **16**

4.1 Introducció històrica 16

4.2 Memòria llarga 17

4.3 Definició del model 19

4.4 Estacionarietat, causalitat i invertibilitat 19

4.5 Densitat espectral 25

4.6 Funció d'autocovariància 25

4.7 Funció de resposta a l'impuls 26

4.8 Mètodes d'estimació 27

4.8.1 Estimació de l'esperança 28

4.8.2 Estimació de la funció d'autocorrelació 28

4.8.3 Estimació dels paràmetres 28

4.9 Tria del millor model 32

4.10 Predicció 32

4.10.1 Predictors d'un pas endavant 33

4.10.2 Predictors de múltiples passos endavant 34

5 Anàlisi pràctica: qualitat de l'aire de Barcelona **35**

5.1 La sèrie 35

5.2 Les dades 37

5.3 Estimació del model 39

5.4 Validació del model 41

5.5 Altres característiques i propietats del model 43

6 Conclusions **47**

A Codi *R* utilitzat a la secció 5 **49**

1 Introducció

El projecte

En l'àmbit de l'estadística i l'anàlisi de dades, les sèries temporals han estat estudiades des de fa dècades degut a la seva naturalesa dinàmica i la seva aplicabilitat a un nombre de disciplines molt variades com les finances, el medi ambient, l'economia o la biologia. És sobretot a partir de la dècada del 1970 quan apareixen més avenços al respecte, gràcies al desenvolupament dels models $ARIMA$ per part de George E.P. Box i G.M. Jenkins.

Una sèrie temporal és una successió d'observacions mesurades en intervals regulars al llarg del temps i que estan ordenades cronològicament. L'anàlisi de les sèries temporals és fonamental per a poder comprendre i predir el comportament de fenomens relacionats amb camps com els mencionats anteriorment. A mesura que les tècniques i els models estadístics han evolucionat, s'ha tornat crucial desenvolupar plantejaments més sofisticats capaços de capturar la complexitat de les sèries temporals i proporcionar una millor anàlisi i prediccions.

Un d'aquests plantejaments que ha estat desentvolupant-se al llarg de les darreres dècades és el dels models $ARFIMA$. Aquests models ens permeten modelitzar sèries temporals amb memòria llarga, concepte que estudiarem al llarg del treball però que resumidament existeix en aquelles sèries on les observacions passades poden tenir un efecte significatiu en valors futurs.

L'objectiu principal d'aquest projecte és fer un estudi des de la vessant matemàtica dels models $ARFIMA$ i explorar i fer l'anàlisi d'una aplicació a una sèrie temporal real que pugui ser modelitzada per aquest model. Per tal de realitzar aquesta anàlisi final, fer el tractament de les dades i aplicar els mètodes estadístics necessaris, s'utilitzarà el llenguatge de programació R.

Estructura de la Memòria

Aquest treball es pot dividir en dues parts: una part teòrica on s'expliquen incloent formulació matemàtica tots aquells resultats relacionats amb l'àmbit d'estudi i una part pràctica on s'hi apliquen aquests resultats.

Pel que fa a la part teòrica, primerament es defineixen alguns conceptes previs bàsics relacionats principalment amb els processos estocàstics i estacionaris per tal d'introduir al lector en la matèria i tenir una base necessària de cara al gruix del treball. Seguidament, introduirem breument i enunciarem les propietats més importants dels models lineals de sèries temporals més habituals i àmpliament coneguts com els models AR, MA o $ARMA$. La base d'aquests models és necessària per entendre els models $ARFIMA$.

A la secció 4 és on trobem el focus del treball. Es comença fent una petita introducció històrica del desenvolupament dels models $ARFIMA$, així com de les seves aplicacions i els motius de la seva utilitat. Després es troba definit matemàticament el concepte de memòria llarga, propietat principal dels models estudiats. A la resta de la secció es defineix de manera exacta un model $ARFIMA$ i s'estudien a fons moltes de les seves propietats i característiques com la funció d'autocovariància o la densitat espectral. En concret, són de vital importància els teoremes que demostren l'estacionarietat, la causalitat i la invertibilitat del model. Per últim, també es proporcionen tècniques d'estimació, tria del millor model i predicció de futurs valors.

En quant a l'apartat pràctic del treball, s'introdueix la sèrie escollida per a ser ajustada amb un model $ARFIMA$: la qualitat de l'aire a la ciutat de Barcelona entre els anys 2017 i 2019. S'analitzen les dades i es comprova que és una sèrie temporal adient per ser modelitzada amb un model $ARFIMA$. Tot seguit, es segueixen les etapes clàssiques a l'hora de modelitzar una sèrie temporal: l'estimació, l'elecció del millor model i la validació. Per últim, R ens permetrà veure gràficament algunes de les propietats vistes al marc teòric del treball.

Cal mencionar també que aquest treball disposa d'un annex on apareixen totes les llibreries i comandes utilitzades a R per obtenir els resultats i els gràfics de l'anàlisi pràctica.

2 Conceptes previs

Una sèrie temporal és un conjunt d'observacions x_t, on cadascuna d'aquestes observacions està mesurada en un instant de temps determinat t en intervals equiespaiats. Aquests intervals de temps poden ser anys, mesos, dies, hores, etc. En aquesta primera secció definirem alguns conceptes bàsics que seran necessaris per tal d'aprofundir posteriorment. Les referències principals utilitzades són [7], [19], [21] i [22].

2.1 Processos estocàstics i processos estacionaris

Definició 2.1. *Un procés estocàstic és una successió de variables aleatòries que estan indexades per un conjunt $\mathbb{T}$ que indica el temps:*

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

Habitualment els processos estocàstics són a temps discret, amb $\mathbb{T} = \mathbb{N}$ o $\mathbb{T} = \mathbb{Z}$. En cas contrari, si $\mathbb{T} = [0, \infty)$ o $\mathbb{T} = \mathbb{R}$, diem que el procés estocàstic és a temps continu.

En aquest treball només tractarem casos on els processos estocàstics són a temps discret, amb $\mathbb{T} = \mathbb{Z}$. D'aquesta manera i, donada la definició, direm que una sèrie temporal és una realització d'un procés estocàstic.

Per altra banda, donada una variable aleatòria X, denotem la seva esperança per $\mathbb{E}(X)$ i la seva variància per $\mathbb{V}(X)$. Donades dues variables aleatòries X i Y, denotem la seva covariància per $\mathbb{C}(X, Y)$.

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

$$\mathbb{E}(X_t^2) < \infty, \quad \forall t \in \mathbb{Z}.$$

D'ara en endavant suposarem que els processos estocàstics són de segon ordre.

Definició 2.3. *Per a un procés estocàstic de segon ordre podem definir diverses funcions:*

- *La funció mitjana:*
$$\mu_t = \mathbb{E}(X_t), \quad \forall t \in \mathbb{Z}.$$
- *La funció variància:*
$$\sigma_t^2 = \mathbb{V}(X_t), \quad \forall t \in \mathbb{Z}$$
.
- *La funció autocovariància:*
$$\gamma_{t,s} = \mathbb{E}[(X_t - \mu_t)(X_s - \mu_s)] = \mathbb{C}(X_t, X_s), \quad \forall t, s \in \mathbb{Z}.$$
- *La funció autocorrelació:*
$$\rho_{t,s} = \frac{\mathbb{E}[(X_t - \mu_t)(X_s - \mu_s)]}{\sqrt{\sigma_t^2 \sigma_s^2}}, \quad \forall t, s \in \mathbb{Z}.$$

Proposició 2.4. *Un procés estocàstic de segon ordre té una esperança $\mathbb{E}(X_t)$ i una variància $\mathbb{V}(X_t)$ ben definides i finites.*

Demostració. Per a provar que l'esperança està ben definida i que aquesta és finita, utilitzarem la desigualtat de Jensen:

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

Per altra banda, hem d'utilitzar la desigualtat de Cauchy-Schwarz per a provar-ho amb la variància:

$$\mathbb{E}|X_t| \leq \sqrt{\mathbb{E}(X_t^2)} < \infty.$$

D'aquí obtenim que $\mathbb{V}(X_t) = \mathbb{E}(X_t^2) - (\mathbb{E}(X_t))^2 < \infty$. □

Proposició 2.5. *La covariància $\mathbb{C}(X, Y)$ de dues variables aleatòries d'un procés estocàstic de segon ordre està ben definida.*

Demostració. De nou, usarem les desigualtats de Jensen i de Cauchy-Schwarz:

$$\begin{aligned}
|\mathbb{C}(X, Y| &= |\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]| \\
&\leq \mathbb{E}(|X - \mathbb{E}(X)||Y - \mathbb{E}(Y)|) \\
&\leq (\mathbb{E}(|X - \mathbb{E}(X)|^2))^{1/2}(\mathbb{E}(|Y - \mathbb{E}(Y)|^2))^{1/2} \\
&= \mathbb{V}(X)^{1/2}\mathbb{V}(Y)^{1/2}.
\end{aligned}$$

□

Definició 2.6. *L'operador retard B és un operador tal que quan s'aplica a una variable temporal la retarda; és a dir, $BX_t = X_{t-1}$. De manera general, $B^k X_t = X_{t-k}$.*

Definició 2.7. *L'operador diferència regular ∇ és un operador tal que quan s'aplica a una variable temporal la transforma de la següent forma:*

$$\nabla X_t = (1 - B)X_t = X_t - X_{t-1}.$$

De manera general, $\nabla^k X_t = (1 - B)^k X_t$.

Definició 2.8. *L'operador diferència estacional ∇_s és un operador tal que quan s'aplica a una variable temporal la transforma de la següent manera:*

$$\nabla_s X_t = (1 - B^s)X_t = X_t - X_{t-s}.$$

Definició 2.9. *Un procés estocàstic $\{X_t, t \in \mathbb{Z}\}$ és estrictament estacionari si per a qualssevol $k_1, ..., k_n$ i l, els vectors*

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

i els vectors

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

tenen la mateixa llei. És a dir, si realitzem un mateix desplaçament en el temps de totes les variables aleatòries la distribució no varia.

Definició 2.10. *Direm que un procés estocàstic és estacionari o estacionari en sentit feble si compleix les condicions següents:*

1. $\mathbb{E}(X_t) = \mu \in \mathbb{R}, \forall t \in \mathbb{Z}$.

2. $\mathbb{V}(X_t) = \mathbb{E}[(X_t - \mathbb{E}(X_t))^2] = \sigma^2 \in \mathbb{R}, \forall t \in \mathbb{Z}$.

3. $\mathbb{C}(X_t, X_{t+l}) = \mathbb{E}[(X_t - \mathbb{E}(X_t))(X_{t+l} - \mathbb{E}(X_{t+l}))] = \gamma(l), \forall t, l \in \mathbb{Z}$, *on* γ *denota una certa funció definida a* $\mathbb{N}$ *i, degut a la propietat de simetria de la covariància,* $\gamma(-l) = \gamma(l)$.

Definició 2.11. *Definim la funció d'autocovariància* γ *com*

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

Al valor l *l'anomenem retard.*

De la definició de la funció d'autocovariància se'n deriven les següents propietats:

1. $\gamma(0) \geq 0$
2. $\gamma(-l) = \gamma(l), \forall l \in \mathbb{Z}$
3. $|\gamma(l)| \leq \gamma(0), \forall l \in \mathbb{Z}$

Definició 2.12. *Anomenem funció d'autocorrelació* ρ *a la funció:*

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

De la definició de la funció d'autocorrelació se'n deriven les següents propietats:

1. $\rho(0) = 1$, si $\gamma(0) \neq 0$
2. $\rho(-k) = \rho(k), \forall k \in \mathbb{Z}$
3. $|\rho(k)| \leq 1, \forall k \in \mathbb{Z}$

Definició 2.13. *Sigui* $X = \{X_j, j \in \mathbb{Z}\}$ *una sèrie temporal amb autocovariància* γ. *Definim la densitat espectral com*

$$f(\lambda) = \sum_{l=0}^{\infty} \gamma(l) e^{-2\pi i \lambda l},$$

per a $0 < \lambda < \infty$.

Definició 2.14. *Un procés estocàstic és Gaussià si* $\forall t_1, ..., t_n$, *el vector* $(X_{t_1}, ..., X_{t_n})$ *té una llei normal multidimensional.*

A partir de les definicions anteriors, podem fer tot un seguit d'observacions. En primer lloc, és clar que estacionarietat estricta implica estacionarietat. Però, de manera general, l'implicació inversa no és certa.

En canvi, si el procés és Gaussià (totes les distribucions finites són Gaussianes), aleshores estacionarietat sí que implica estacionarietat estricta. Això és degut a que en aquest cas el vector mitjana i la matriu de covariància determinen la llei.

Per altra banda, la representació gràfica de les autocorrelacions ve donada per

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

i s'anomena *correlograma*.

Per últim, veiem que per tal de determinar un procés estacionari de segon ordre és suficient amb tenir la parella (μ, γ) o la tripleta (μ, σ, ρ).

Definició 2.15. *Sigui $X = \{X_j, j \in \mathbb{Z}\}$ una sèrie temporal estacionària de segon ordre centrada. Definim la funció d'autocorrelació parcial com l'aplicació*

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

tal que $\alpha(0) = 1$, $\alpha(1) = \rho(1)$ i per $l \geq 2$:

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

on $\alpha(l)$ es pot interpretar com la correlació entre X_1 i X_{l+1} havent eliminat la influència de les variables $X_2, ..., X_l$.

A partir dels coeficients de la funció d'autocorrelació podem calcular els coeficients de la funció d'autocorrelació parcial. És suficient amb resoldre per a cada n el següent sistema d'equacions i prendre $\alpha(n) = x_n$:

$$\begin{bmatrix} 1 & \rho(1) & \cdots & \rho(n-1) \\ \rho(1) & 1 & \cdots & \rho(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(n-1) & \rho(n-2) & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \\ x_n \end{bmatrix} = \begin{bmatrix} \rho(1) \\ \vdots \\ \\ \rho(n) \end{bmatrix}$$

2.2 Anàlisi clàssica de les sèries temporals

Per tal de poder comprendre millor les sèries temporals, és molt útil fer una representació gràfica d'aquestes. D'aquesta manera, és com clàssicament s'ha arribat a la conclusió que podem separar les sèries en les diferents components que veurem a continuació. Aquesta separació és el pas previ a identificar quin serà el model definitiu que s'assignarà la sèrie, el qual ens servirà per estudiar-la i fer prediccions futures.

Així, segons el punt de vista de l'anàlisi, anterior a la dècada dels 70, les sèries temporals estan formades per tres components: la tendència (T), els cicles ($C^{(j)}$) o components estacionals i els residus o component irregular (R).

La tendència (T) d'una sèrie temporal és la component que representa l'evolució al llarg termini de la sèrie. Aquesta evolució o pauta de comportament estable en llargs períodes de temps pot ser creixent o decreixent.

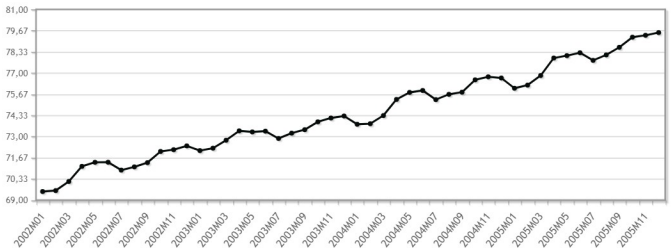

Figura 1: Evolució mensual de l'Índex de Preus de Consum (IPC) a Espanya entre el 2002 i el 2005. Font: Instituto Nacional de Estadística.

En la sèrie anterior, per exemple, veiem una clara tendència a l'alça de l'IPC a Espanya entre els anys 2002 i 2005.

Els cicles ($C^{(j)}$) recolleixen els moviments oscil·latoris per sobre i per sota la tendència i que es mantenen en un període de temps determinat. La durada del cicle es mesura des

d'un pic fins el següent o des d'una vall fins la següent. Aquesta durada no té perquè mantenir-se constant. Aquests cicles o estacionalitat poden venir donats per diferents raons en funció de la naturalesa de la sèrie, però els motius més comuns són de tipus físic-natural (temps meteorològic, cicles biològics, etc) o de tipus institucional (vacances, festes, etc).

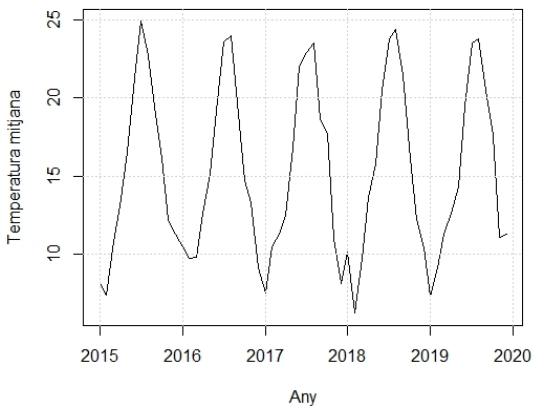

Figura 2: Mitjana de les temperatures mínimes mensuals (en graus centígrads) de la comarca del Barcelonès entre els anys 2015 i 2019. Font: Institut d'Estadística de Catalunya.

Les sèries temporals referents a les temperatures mitjanes mensuals són un clar exemple on podem apreciar la presència de la component estacional. En aquest cas, la sèrie pren uns valors més alts als mesos d'estiu (juliol i agost) i uns valors més baixos als mesos d'hivern (entre desembre i febrer).

Per últim, la component irregular R es refereix a les variacions esporàdiques de la sèrie que no estan presents ni en la tendència ni en els cicles, tenint per tant un caràcter residual. Es podria descompondre aquesta component en dues parts: una d'aleatòria que recull petits efectes accidentals o no explicats i una d'erràtica, que és conseqüència d'efectes no previsibles però que poden ser explicats a posteriori com per exemple catàstrofes naturals o una pandèmia.

Un cop presentades aquestes components, podem representar una sèrie temporal mitjançant l'esquema següent:

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

En funció de la interacció entre aquestes components, podem trobar-nos amb un esquema additiu, un esquema multiplicatiu o un de mixte. De manera general, al llarg del treball utilitzarem un esquema de tipus additiu:

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

És important recordar que un model no additiu pot ser transformat en un que sí ho sigui aplicant una transformació logarítmica a les dades. També cal tenir en compte

que en una sèrie temporal no hi tenen perquè estar presents totes les components. Per exemple, en la sèrie de la Figura 2 no hi apreciem una tendència creixent o decreixent.

L'objectiu d'aquesta anàlisi clàssica de les sèries temporals és tenir una bona descripció de la tendència i els cicles per tal de poder aïllar-los fins que tan sols es pugui distingir la component residual. Per fer una anàlisi correcta, aquests residus han de ser estacionaris, sense cap patró cíclic o estacional ni cap variabilitat creixent ni decreixent.

Hi ha diferents mètodes que ens permeten aïllar aquestes components, els quals no entrarem a explicar en detall, però els mencionarem. En primer lloc, existeixen tres maneres d'identificar la tendència d'una sèrie. Una és ajustar un cert tipus de funció a les dades, com per exemple un polinomi. Un altre mètode és filtrar la sèrie per tal d'obtenir una funció més regular que pugui ser interpretada com una tendència. I una última manera és utilitzar l'operador retard i l'operador diferència regular definits a la pàgina 4 per tal d'eliminar la tendència de la sèrie.

Per altra banda, si volem eliminar la component estacional podem usar de nou tres mètodes: la identificació directa dels cicles, fer un filtratge per tal d'obtenir una sèrie sense estacionalitat (que llavors restem a l'original), i per últim també podem utilitzar l'operador diferència estacional definit a la pàgina 4.

Per acabar, sovint és necessari aplicar una transformació a les dades originals abans de la seva anàlisi. La transformació més emprada és la logarítmica:

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

Aquesta tranformació pot ser necessària per diferents motius: convertir els productes de les components de la sèrie en sumes, transformar les dades positives en dades reals, neutralitzar les tendències exponencials o suavitzar reduint la variabilitat de les observacions.

2.3 Exemples

Seguidament, mostrarem alguns casos de tipus sèries temporals que són habituals en l'estudi i anàlisi d'aquestes.

1. Soroll IID.
   Aquest és potser el model més simple de sèrie temporal ja que no tenim ni tendència ni component estacional. Diem, doncs, que $\{X_k, k \geq 1\}$ és un soroll IID si les variables aleatòries són independents i idènticament distribuides (IID) amb mitjana μ i desviació estàndard σ. En aquest cas, $\rho(k) = 1$ si $k = 0$ i $\rho(k) = 0$ si $k > 0$. En aquest model les sèries temporals són purament aleatòries.

2. Passeig aleatori.
   Sigui $\{X_k, k \geq 1\}$ un soroll IID. Aleshores la sèrie $\{S_k, k \geq 1\}$ on:

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

   s'anomena passeig aleatori.
   Podem veure que aquestes sèries no són estacionàries en sentit feble. Si escrivim

   $$S_k = S_{k-1} + X_k.$$

Aleshores,

$$\mathbb{E}[S_k] = \mathbb{E}[S_{k-1}] + \mathbb{E}[X_k]$$

i per tant la funció mitjana és constant només si $\mathbb{E}[X_k] = 0$. En tal cas i, veient que X_k és independent de S_{k-1}, tenim

$$\mathbb{V}[S_n] = \mathbb{V}[S_{n-1}] + \mathbb{V}[X_n].$$

Llavors:

$$\mathbb{V}[S_n] = \mathbb{V}[S_{n-1}] \iff \mathbb{V}[X_n] = 0.$$

Això passa si $X_n \equiv 0$. Ara, si $\mathbb{E}[X_n] = \mu$ i $\mathbb{V}[X_n] = \sigma^2$, tenim $\mathbb{E}[S_n] = n\mu$ i $\mathbb{V}[S_n] = n\sigma^2$. Per altra banda, si $n < m = n + p$,

$$\begin{aligned}
\mathbb{C}(S_n, S_m) &= \mathbb{C}(S_n, S_{n+p}) \\
&= \mathbb{C}(S_n, S_n) + \mathbb{C}(S_n, X_{n+1} + ... + X_{n+p}) \\
&= V[S_n] \\
&= n\sigma^2.
\end{aligned}$$

Aleshores, com que $\mathbb{C}(S_n, S_{n+p})$ depèn de n, la sèrie no és estacionària.

3. Soroll blanc.
   Sigui $\{X_k, k \geq 1\}$ una sèrie de variables aleatòries no correlacionades, cadascuna d'elles amb esperança μ i variància σ^2. Clarament X_k és estacionària i s'anomena de manera comú com a soroll blanc (*White Noise* en anglès). Se sol indicar amb la notació

   $$\{X_t\} \sim WN(\mu, \sigma^2).$$

   Un soroll IID és un cas particular d'un soroll blanc.

4. Soroll blanc Gaussià.
   Diem que un procés $\{X_k, k \geq 1\}$ és un soroll blanc Gaussià si és un soroll blanc i qualsevol vector $(X_{k_1}, ..., X_{k_n})$ té llei normal. Aquest procés també és un soroll IID, perquè, sota Gaussianitat, no correlació implica independència. Per tant, com hem vist, un soroll blanc Gaussià és un cas particular d'un soroll IID, i aquest és un cas particular d'un soroll blanc.

3 Models lineals de sèries temporals

A partir de la dècada dels 70, el focus d'estudi de les sèries temporals va passar a ser el de trobar models que, ajustant-se a les dades, aconseguissin descriure i predir futures observacions fent les correccions estacionals o de tendència adients. En aquesta secció, doncs, veurem les definicions i propietats més importants d'aquests models més habituals i que ens ajudaran a entendre millor els models que ens ocupen en el treball: els models ARFIMA. Els conceptes que s'explicaran s'han extret de [6], [10], [19], [22]. En aquesta bibliografia també s'hi explica com seleccionar el millor model o com fer una estimació dels paràmetres. Ara bé, com que en el nostre treball ens centrarem en els models ARFIMA, deixarem aquesta explicació per a més endavant.

Abans d'entrar en detall amb els models en qüestió, donarem un seguit de definicions i teoremes que ens seran de gran utilitat.

Definició 3.1. *Diem que un procés estocàstic és lineal quan el podem escriure com el resultat d'una transformació lineal d'un procés soroll blanc; és a dir,*

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

on $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, $\psi_0 = 1$ *i* $\sum_{j=0}^{\infty} \psi_j^2 < \infty$.

Teorema 3.2. ***(Teorema de representació de Wold)****. Qualsevol procés estocàstic estacionari en sentit feble es pot representar com la suma de dos processos:*

$$y_t = z_t + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

on z_t *és un procés purament determinista,* $\{\psi_j, j \geq 0\}$ *compleixen les condicions de la Definició* 3.1 *i* ε_t *és un procés soroll blanc amb*

$$\mathbb{E}[\varepsilon_t] = 0,$$

$$\mathbb{V}[\varepsilon_t] = \sigma_\varepsilon^2,$$

$$\mathbb{C}(\varepsilon_t, \varepsilon_{t-k}) = 0, \quad \forall k > 0.$$

Demostració. La demostració es pot trobar a les pàgines 187, 188 i 189 de [6]. □

Definició 3.3. *Direm que un procés estocàstic* $\{y_t, t \in \mathbb{Z}\}$ *és causal si i només si existeixen* $\{\psi_j, j \geq 0\}$ *que compleixen:*

1. $\sum_{j=0}^{\infty} |\psi_j| < \infty$
2. $y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$

És a dir, que podem escriure $y_t = \psi(B)\varepsilon_t$.

Definició 3.4. *Direm que un procés estocàstic* $\{y_t, t \in \mathbb{Z}\}$ *és invertible si i només si existeixen* $\{\pi_j, j \geq 0\}$ *que compleixen:*

1. $\sum_{j=0}^{\infty} |\pi_j| < \infty$

2. $\varepsilon_t = \sum_{j=0}^{\infty} \pi_j y_{t-j}$

És a dir, que podem escriure $\varepsilon_t = \pi(B) y_t$.

3.1 Models autorregressius (AR)

Definició 3.5. *Sigui* $Z = \{Z_j, j \in \mathbb{Z}\}$ *un soroll blanc centrat amb variància* $\sigma^2 > 0$ *i donat un nombre natural* $p \geq 0$*, definim un model autorregressiu d'ordre* p *com un procés estacionari* $\{Y_j, j \in \mathbb{Z}\}$ *que satisfà l'equació:*

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

on $\phi_1, ..., \phi_p$ *són nombres reals. En aquest cas direm que* $Y_j \sim AR(p)$.

Podem redefenir l'equació anterior d'una altra manera. Si definim el polinomi

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

i usem l'operador retard B, aleshores obtenim:

$$\Phi(B) Y_j = Z_j, \quad \forall j \in \mathbb{Z}.$$

Definició 3.6. *Diem que un polinomi* $\Phi(x)$ *és invertible si existeix una sèrie* $\sum_{i=0}^{\infty} \psi_i x^i$ *tal que:*

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

on $\sum_{i=0}^{\infty} \psi_i^2 < \infty$.

En tal cas, veiem que si el polinomi $\Phi(x)$ és invertible, aleshores tenim que:

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

i aquest fet ens garanteix el caràcter causal i estacionari de Y. El següent teorema ens mostrarà quan un polinomi és invertible.

Teorema 3.7. *Un polinomi* $\Phi_p(z)$*, amb* $z \in \mathbb{C}$*, és invertible si i només si totes les seves arrels es troben fora del cercle unitat; és a dir, si i només si*

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

Demostració. La demostració es pot trobar a les pàgines 86 i 87 de [6]. □

Observacions 3.8. Seguidament, farem un seguit d'observacions en relació a les funcions d'autocovariància, d'autocorrelació i d'autocorrelació parcial.

1. Com que el model $AR(p)$ és causal, tenim immediatament que és centrat i, si $l \geq 0$, la seva funció d'autocovariància és:

$$\gamma(l) = \mathbb{C}(Y_j, Y_{j+l}) = \sigma^2 \sum_{r=0}^{\infty} \psi_r \psi_{r+l}.$$

   Veiem, doncs, que per tal de calcular $\gamma(l)$ necessitem conèixer els coeficients ψ_i. Si recuperem l'equació de la definició d'invertibilitat, acabem obtenint que:

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

   D'aquesta manera podem concloure que es tracta d'un procés estacionari, ja que no depén de j, i que si tenim la condició de que el polinomi $\Phi(x)$ és invertible aleshores $AR(p)$ és efectivament estacionari.

2. Si volem determinar la funció d'autocorrelació, notem que, per $l > 0$,

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

   i per tant,

$$\gamma(l) = \phi_1 \gamma(l-1) + ... + \phi_p \gamma(l-p).$$

   Si dividim per $\gamma(0)$ (per trobar la funció d'autocorrelació), obtenim:

$$\rho(l) = \phi_1 \rho(l-1) + ... + \phi_p \rho(l-p),$$

$$\rho(0) = 1,$$

$$\rho(-l) = \rho(l).$$

   Aquest sistema rep el nom d'*equacions de Yule-Walker* i la seva solució és la funció d'autocorrelació del model $AR(p)$.

3. Pel que fa a la funció d'autocorrelació parcial, tenim que

$$\alpha(k) = 0, \qquad \forall k \geq p+1.$$

3.2 Models mitjana mòbil (MA)

Definició 3.9. *Sigui $Z = \{Z_j, j \in \mathbb{Z}\}$ un soroll blanc centrat amb variància $\sigma^2 > 0$ i donat un nombre natural $q \geq 0$, definim un model mitjana mòbil d'ordre q com un procés estacionari $\{Y_j, j \in \mathbb{Z}\}$ que satisfà l'equació:*

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

on $\theta_1, ..., \theta_q$ són nombres reals. En aquest cas direm que $Y_j \sim MA(q)$.

Observacions 3.10.

1. Si $k \geq 0$, calculem la funció d'autocovariància:

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

   Com que l'autocovariància entre Y_j i Y_{j+k} no depèn de j, tenim que és un procés estacionari.

2. La funció d'autocorrelació és:

$$\rho(k) = \frac{\sum_{r=0}^{q-k} \theta_r\theta_{r+k}}{\sum_{r=0}^{q} \theta_r^2}, \qquad 0 \leq k \leq q.$$

3. La funció d'autocorrelació parcial és decreixent a mesura que augmenta el retard l.

4. Pel que fa a la representació del model en termes de l'operador retard B, tenim

$$Y_j = Z_j + \theta_1 B Z_j + \theta_2 B^2 Z_j + ... + \theta_q B^q Z_j$$

   i si $B^0 = Id$, aleshores escrivim

$$Y_j = \Theta_q(B) Z_j,$$

   on

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

   és un polinomi d'ordre $q \geq 0$.

3.3 Models mixtos (ARMA)

Definició 3.11. *Siguin p i q dos nombres naturals fixats, $Z \sim WN(0, \sigma^2)$ i $\Phi_p(\cdot)$, $\Theta_q(\cdot)$ dos polinomis invertibles de grau p i q, respectivament, sense cap arrel en comú. Un model $ARMA(p, q)$ és un model que satisfà l'equació:*

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

Observacions 3.12.

1. Com que $\Phi_p(\cdot)$ és invertible, podem escriure

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

   i això prova que els models $ARMA$ són lineals i causals.

2. Els processos $ARMA$ tenen una funció d'autocorrelació com la de la seva part AR i una funció d'autocorrelació parcial com la de la seva part MA. Aquestes funcions decreixen exponencialment en valor absolut cap a zero.

3.4 Models integrats (ARIMA)

Sigui $\{Y_j, j \geq 0\}$ un passeig aleatori que el definim com

$$Y_j = Y_{j-1} + Z_j,$$

on $Z \sim WN(0, \sigma^2)$ i $Y_0 = Z_0 = 0$. Aquest procés no és estacionari perquè la variància

$$\mathbb{V}(Y_j) = \mathbb{V}(Y_{j-1}) + \sigma^2$$

i per tant,

$$\mathbb{V}(Y_j) = j\sigma^2.$$

Ara, si calculem les primeres diferències obtenim la sèrie

$$X_j := Y_j - Y_{j-1} = Z_j, \quad \forall j \geq 1,$$

que sí és estacionaria. Així doncs, hem vist com podem convertir una sèrie no estacionària en una d'estacionària. Aquesta és la base de com generalitzar models $ARMA$ en models $ARIMA$. Una altra manera de veure-ho és que extenem el model $ARMA$ per a que permeti incloure arrels unitàries.

Definició 3.13. *Siguin $p, d, q \in \mathbb{N}$. Aleshores $\{X_j, j \geq 0\}$ és un model $ARIMA(p, d, q)$ si*

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

és un model $ARMA(p, q)$.

Diem que el procés Y és el procés X al que se li ha aplicat l'operador diferència regular d vegades. Per tant, seguint la notació vista amb els models $ARMA$, un model $ARIMA(p, d, q)$ es pot representar com:

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

En el cas que un procés $ARIMA$ tingui tendència, tenim

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

3.5 Models estacionals (SARIMA)

Com hem vist anteriorment, en moltes ocasions ens podem trobar amb sèries temporals que presenten un clar comportament estacional. Sigui $\{X_j, j \geq 0\}$ una sèrie temporal. Podem escriure-la de la forma $X_j = X_{i+(r-1)s}$, on $i = 1, ..., s$ indica el mes d'observació i $r \geq 1$ indica l'any si, per exemple, tenim una sèrie amb periodicitat mensual on tindrem $s = 12$.

Els models $SARIMA$ estan dividits en dues parts: una de regular i una d'estacional. Aquesta última part suposa que tan sols hi ha correlació entre dades del mateix mes, mentre que les dades de diferents mesos estan incorrelacionades. Això significa que tenim (en el cas de períodes anuals) 12 sèries temporals, una per mes, incorrelacionades entre elles.

Definició 3.14. *La part estacional d'un model $SARIMA$ és un model $ARMA(a,c)$ de la següent manera:*

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

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

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

Observació 3.15. Es poden calcular les funcions d'autovariància, autocorrelació i autocorrelació parcial de la mateixa manera que en els models ARMA.

Definició 3.16. *Siguin els polinomis amb l'operador retard B de la següent forma:*

$$A_a(B) = 1 - \varphi_1 B^s - \varphi_2 B^{2s} - ... - \varphi_a B^{as}$$

$$C_c(B) = 1 - \vartheta_1 B^s - \vartheta_2 B^{2s} - ... - \vartheta_c B^{cs}$$

aleshores un model

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

es pot representar amb l'equació

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

i per tant el model és equivalent a un $ARIMA(p,d,q) \times ARIMA(a,D,c)_s$. La primera part es refereix a la part regular del model i la segona, a l'estacional.

4 Models ARFIMA

Un cop definits els conceptes previs necessaris i presentats els diferents tipus de models bàsics, arribem al nucli del treball on estudiarem els tipus de models més adients per tal de treballar amb sèries temporals que presenten una memòria llarga (concepte que definirem en les pàgines següents). Aquests models són els anomenats $ARFIMA$ o $FARIMA$, de l'anglès *Autoregressive fractionally integrated moving average*. Els podem trobar representats amb les dues nomenclatures, però per facilitar la notació i evitar confusions d'ara en endavant els anomenarem sempre $ARFIMA$.

Primerament farem una petita introducció històrica, seguida de la definició de conceptes bàsics referents a les sèries temporals amb memòria llarga. Després, introduirem els models $ARFIMA$, explicant-ne totes les seves característiques i propietats. I finalitzarem presentant diversos mètodes d'estimació dels paràmetres, de predicció i de selecció del millor model.

Els conceptes que s'explicaran s'han extret en major part de [17]. També s'han utilitzat [6], [8], [21] i [22].

4.1 Introducció històrica

Tot i que la major part de l'estudi sobre el tema ha tingut lloc en les darreres dècades, podem trobar evidència empírica sobre dades amb memòria llarga a la dècada dels 50, concretament en el treball de Hurst [15] en el camp de l'hidrologia. Anys d'estudi més tard, Beran [3] assenyala que s'han trobat evidències de memòria llarga en sèries temporals relacionades amb una gran varietat d'àmbits com les finances, l'astronomia, la química o la meteorologia.

Pérez i Ruiz [18] indiquen que la propietat de memòria llarga acostuma a estar relacionada amb la persistència que mostren les autocorrelacions mostrals d'algunes sèries temporals estacionàries. Aquestes autocorrelacions decreixen en un ritme molt lent fins finalment convergir a zero. Si recordem, aquest comportament, d'una banda no és compatible amb els models autorregressius AR i els models $ARMA$ ja que les autocorrelacions decreixen exponencialment. D'altra banda, tampoc és compatible amb l'extrem grau de persistència dels models $ARIMA$.

Granger i Joyeux [13] adverteixen dels perills d'aplicar diferències en una sèrie temporal que aparentment no és estacionària per tal d'obtenir estacionarietat. Aquesta pràctica pot tenir conseqüències negatives en la correcta modelització de la sèrie, ja que a aquesta sèrie a la qual se li han aplicat diferències se li elimina la component de baixes freqüències, que és de gran importància per a les prediccions a llarg termini. Els autors assenyalen que per a modelitzar aquest tipus de sèries, la diferenciació entera és excessiva (sobrediferenciació), però tampoc és idoni no diferenciar (subdiferenciació).

Davant d'aquesta problemàtica, Granger i Joyeux [13] i Hosking [14] proposen una mena de processos intermitjos en els quals l'ordre d'integració és fraccionari. En aquests tipus de models, la memòria de la sèrie es troba entremig de la memòria curta que presenten els models $ARMA$ i la memòria persistent dels models $ARIMA$. El que provoca això és que les "innovacions" d'observacions que presentin aquestes sèries tinguin efectes transitoris que perdurin durant molt de temps, però que finalment acabin desapareixent. Si ho comparem, de nou, amb els models estudiats anteriorment, aquest comportament és diferent al que presenten. Per un cantó, en els models $ARMA$ l'efecte de la "innovació"

desapareix ràpidament de forma exponencial. Per l'altre, degut a la seva naturalesa, en els models $ARIMA$ les "innovacions" tenen efectes permanents.

Així es com neixen els models autorregressius i de mitjana mòbil fraccionalment integrats, que segueixen la notació $ARFIMA(p, d, q)$, on d és un nombre real i per tant pot ser fraccionari. Aquest paràmetre d descriu les propietats dinàmiques al llarg termini, mentre que l'estructura de dependència al curt termini és explicada pels paràmetres de la component $ARMA(p, q)$ del model.

4.2 Memòria llarga

Tal i com hem vist al Teorema 3.2. (Teorema de representació de Wold), qualsevol procés estocàstic estacionari en sentit feble es pot escriure com

$$X_j = \sum_{i=0}^{\infty} \psi_i Z_{j-i} + f(t),$$

on $f(t)$ és un procés purament determinista i Z és un soroll blanc centrat amb variància σ^2. A partir d'ara, suposarem que el nostre procés X és lineal, que recordem que significa que és una sèrie temporal estacionària centrada purament no determinista que es pot representar

$$X_j = \sum_{i=0}^{\infty} \psi_i Z_{j-i},$$

amb $\sum_{i=0}^{\infty} \psi_i^2 < \infty$.

La funció d'autocorrelació d'aquests processos és

$$\rho(l) = \frac{\sum_{r=0}^{\infty} \psi_r \psi_{r+l}}{\sum_{r=0}^{\infty} \psi_r^2}.$$

D'aquí podem veure immediatament que $\rho(l)$ convergeix a 0 quan el retard l tendeix a ∞.

Podem fitar

$$\sum_{r=0}^{\infty} |\psi_r||\psi_{r+l}| \leq (\sum_{r=0}^{\infty} |\psi_r|^2)^{1/2} (\sum_{r=0}^{\infty} |\psi_{r+l}|^2)^{1/2}.$$

El primer terme del producte de la dreta és una constant finita i el segon terme el podem reescriure com

$$(\sum_{j=l}^{\infty} |\psi_j|^2)^{1/2},$$

que convergeix a 0 quan $l \to \infty$ perquè $\sum_{r=0}^{\infty} |\psi_r|^2$ és finita.

Definició 4.1. *Sigui $\rho(l)$ la funció d'autocorrelació amb retard l d'un procés estacionari $\{Y_j, j \in \mathbb{Z}\}$ que satisfà*

$$\sum_{l=0}^{\infty} |\rho(l)| < \infty.$$

Diem que en tal cas el procés estacionari té memòria curta. En cas contrari, si

$$\sum_{l=0}^{\infty} |\rho(l)| = \infty,$$

diem que té memòria llarga.

Observació 4.2. Altres autors també defineixen la memòria llarga utilitzant la funció d'autocovariància:

Definició 4.3. *Sigui $\gamma(l) = \mathbb{C}(Y_j, Y_{j+l})$ la funció d'autocovariància amb retard l d'un procés estacionari $\{Y_j, j \in \mathbb{Z}\}$. El procés té memòria llarga si satisfà*

$$\sum_{l=0}^{\infty} |\gamma(l)| = \infty.$$

Si ens remuntem als processos $ARMA$, podem veure que aquests tenen memòria curta. Com hem comentat anteriorment, la seva funció d'autocorrelació decreix exponencialment. De fet, es pot provar que

$$|\rho(l)| \leq \frac{c}{|r|^l},$$

on $c \geq 0$ i r és l'arrel més petita del polinomi AR i que, per estar fora del cercle unitat, $|r| > 1$.

Ara bé, no tots els processos estacionaris tenen memòria curta. A la primera observació de les Observacions 3.8. vam indicar que

$$\gamma(l) = \mathbb{C}(Y_j, Y_{j+l}) = \sigma^2 \sum_{r=0}^{\infty} \psi_r \psi_{r+l}.$$

Aleshores, per models lineals causals de segon ordre tenim

$$\begin{aligned}
\sum_{l=0}^{\infty} |\gamma(l)| &\leq \sigma^2 \sum_{l=0}^{\infty} \sum_{r=0}^{\infty} |\psi_r||\psi_{l+r}| \\
&= \sigma^2 \sum_{r=0}^{\infty} \sum_{s=r}^{\infty} |\psi_r||\psi_s| \\
&\leq \sigma^2 \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} |\psi_r||\psi_s| \\
&= \sigma^2 (\sum_{r=0}^{\infty} |\psi_r|)^2.
\end{aligned}$$

Per tant, $\sum_{r=0}^{\infty} |\psi_r| < \infty$ és una condició suficient per garantir memòria curta. Tot i així,

$$\sum_{k=0}^{\infty} |\psi_k|^2 < \infty \tag{4.1}$$

és una condició més feble que

$$\sum_{k=0}^{\infty} |\psi_k| < \infty. \tag{4.2}$$

Per tant, efectivament podem tenir processos estacionaris amb memòria llarga. Existeixen models lineals causals estacionaris de segon ordre tals que $\sum_{l=0}^{\infty} |\gamma(l)|$ i $\sum_{r=0}^{\infty} |\psi_r|$ són infinites.

4.3 Definició del model

Aquests models són els models $ARFIMA$, que definirem a continuació. Com hem dit, són models lineals causals estacionaris de segon ordre i, per tant, satisfan (4.1). Però, al tenir memòria llarga, no satisfan (4.2).

Definició 4.4. *Un procés $\{X_t, t \in \mathbb{Z}\}$ és un model $ARFIMA(p,d,q)$ amb $-1 < d < \frac{1}{2}$ si i només si $(Id - B)^d X$ és un procés $ARMA(p,q)$. És a dir, X és un procés estacionari que satisfà*

$$\Phi_p(B)(Id - B)^d X = \Theta_q(B) Z, \tag{4.3}$$

on $\Phi_p(B) = 1 + \phi_1 B + ... + \phi_p B^p$ i $\Theta_q(B) = 1 + \theta_1 B + ... + \theta_q B^q$ són els operadors autorregressius i mitjana mòbil, respectivament, i no tenen arrels en comú.
Notem que podem reescriure (4.3) com

$$\Phi_p(B) X = \Theta_q(B)(Id - B)^{-d} Z \tag{4.4}$$

on $(Id - B)^{-d}$ és un operador de diferències fraccionàries definit pel desenvolupament binomial

$$(Id - B)^{-d} = \sum_{i=0}^{\infty} \psi_i B^i = \psi(B),$$

on

$$\psi_i = \frac{\Gamma(i+d)}{\Gamma(i+1)\Gamma(d)} \tag{4.5}$$

per $d < \frac{1}{2}$, $d \neq 0, -1, -2, ...$ i Z és un soroll blanc centrat amb variància σ^2.
Per tant, podem interpretar un model $ARFIMA$ com un model $ARMA(p,q)$ amb un soroll fraccionari donat per $(Id - B)^{-d} Z$.

Observació 4.5. Recordem que la funció Gamma es defineix, si $x > 0$,

$$\Gamma(x) := \int_0^{\infty} t^{x-1} e^{-t}\, dt.$$

D'una manera més general, si definim

$$\psi(z) = (1 - z)^{-d},$$

aleshores aquesta funció és analítica al disc unitat obert $\{z : |z| < 1\}$ i analítica al disc unitat tancat $\{z : |z| \leq 1\}$ per d negatives. Sota aquestes circumstàncies podem escriure $\psi(\cdot)$ com el desenvolupament de Taylor

$$\psi(z) = \sum_{i=0}^{\infty} \psi_i z^i.$$

4.4 Estacionarietat, causalitat i invertibilitat

A continuació veurem una sèrie de teoremes que provaran l'existència d'una solució estacionària dels processos $ARFIMA$ definits a l'equació (4.4). En concret, veurem la unicitat, la causalitat i la invertibilitat. Començarem, però, amb un seguit de definicions i teoremes previs necessaris per provar els teoremes finals.

Definició 4.6. *Sigui $\{T_k, k \geq 1\}$ una sèrie de variables aleatòries de quadrat integrable. Diem que $\{T_k, k \geq 1\}$ convergeix en L^2 (o en mitjana quadràtica) a una variable aleatòria quadrat integrable T si i només si*

$$\mathbb{E}[(T - T_k)^2] \xrightarrow{n\to\infty} 0.$$

Definició 4.7. *Diem que una funció mesurable $\ell : (0, +\infty) \to (0, +\infty)$ és una funció de variació lenta si, $\forall a > 0$,*

$$\lim_{x\to\infty} \frac{\ell(ax)}{\ell(x)} = 1.$$

Exemples 4.8. Uns exemples de funcions de variació lenta poden ser $\ell(x) = log(x)$ o $\ell(x) = b$, on b és una constant positiva.

Teorema 4.9. *Sigui $\{X_t, t \in \mathbb{Z}\}$ un procés estacionari amb esperança 0 i sigui $\{\varphi_j, j \in \mathbb{Z}\}$ una successió absolutament sumable. Aleshores,*

1. *El procés $\{Y_t, t \in \mathbb{Z}\}$ definit per*

$$Y_t = \sum_{j=0}^{\infty} \varphi_j X_{t-j} = \varphi(B) X_t$$

   és estacionari amb $\mathbb{E}(Y_t) = 0$.

2. *Suposem que el procés estacionari $\{X_t\}$ està donat per l'equació*

$$X_t = \sum_{j=0}^{\infty} \eta_j Z_{t-j} = \eta(B) Z_t.$$

   Aleshores, podem escriure

$$Y_t = \varphi(B)\eta(B)Z_t = \eta(B)\varphi(B)Z_t = \psi(B)Z_t,$$

   on aquestes equacions convergeixen en mitjana quadràtica i $\psi(B) = \eta(B)\varphi(B)$.

Lema 4.10. *Sigui $\{Y_t, t \in \mathbb{Z}\}$ un procés estacionari amb densitat espectral f i desenvolupament mitjana mòbil convergent en L^2*

$$Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} = \psi(B) Z_t.$$

Definim l'operador $\pi(B) = \frac{1}{\psi(B)}$. Si la densitat espectral f satisfà les següents condicions:

1. *$f(\lambda) \sim |\lambda|^{-2d}\ell(1/|\lambda|)$ quan $\lambda \to 0$, on ℓ és una funció de variació lenta i $d \in (-1, \frac{1}{2})$.*
2. *f està fitada a $[\varepsilon, \pi], \forall \varepsilon > 0$.*
3. *f^{-1} és localment integrable a $(0, \pi]$.*

Aleshores el desenvolupament

$$Z_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j} = \pi(B) Y_t$$

convergeix en L^2.

Ara sí, podem veure els teoremes que ens provaran existència, unicitat, causalitat i invertibilitat d'una solució estacionària dels processos $ARFIMA$. També veurem quines són les funcions d'autocovariància, d'autocorrelació i d'autocorrelació parcial.

Teorema 4.11. *Sigui Z un soroll blanc centrat amb variància σ^2 i sigui el procés $ARFIMA$ definit a* (4.4)*. Si $-1 < d < \frac{1}{2}$, si els polinomis $\Phi_p(\cdot)$ i $\Theta_q(\cdot)$ no tenen arrels en comú i si les arrels de $\Phi_p(\cdot)$ es troben fora del cercle unitat tancat $\{z : |z| \leq 1\}$, aleshores existeix una única solució estacionària de* (4.4) *donada per*

$$X_j = \sum_{i=0}^{\infty} \psi_i Z_{j-i}, \tag{4.6}$$

on

$$\psi(z) = (1-z)^{-d} \frac{\Theta_q(z)}{\Phi_p(z)}.$$

Demostració. En primer lloc, considerem el procés $Y_j = \sum_{i=0}^{\infty} \eta_i Z_{j-i}$, on els coeficients η_i són de la forma

$$\eta_i = \frac{\Gamma(i+d)}{\Gamma(i+1)\Gamma(d)} = \prod_{k=1}^{i} \frac{k-1-d}{k}.$$

Aquests coeficients compleixen

$$\sum_{i=0}^{\infty} \eta_i^2 < \infty$$

i per tant el polinomi trigonomètric $\sum_{i=0}^{n} \eta_i e^{\mathbf{i}\lambda i}$ ($\mathbf{i}$ de imaginari) convergeix a $\eta(e^{\mathbf{i}\lambda}) = (1-e^{\mathbf{i}\lambda})^{-d}$. En conseqüència, $\{Y_j\}$ és un procés estacionari ben definit com a (4.6). Com que $\Phi_p(z)$ no té arrels a $|z| = 1$, existeix $\delta > 1$ tal que podem escriure la sèrie de Laurent convergent

$$\varphi(z) = \sum_{i=0}^{\infty} \varphi_i z^i = \frac{\Theta_q(z)}{\Phi_p(z)},$$

per a $\delta^{-1} < |z| < \delta$. Per tant, pel Teorema 4.9.1,

$$X_j = \varphi(B) Y_j = \sum_{i=0}^{\infty} \varphi_i Y_{j-i}$$

és un procés estacionari. Donat que $\{\varphi_i\}$ és absolutament sumable i que $\{Y_j\}$ és estacionari, pel Teorema 4.9.2 podem escriure $\{X_j\}$ com

$$X_j = \psi(B) Y_j = \sum_{i=0}^{\infty} \psi_i Z_{j-i}, \tag{4.7}$$

on $\psi(B) = \varphi(B)\eta(B)$. Si multipliquem (4.7) per $\Phi_p(B)$ i apliquem el Teorema 4.9, obtenim

$$\Phi_p(B) = X_j = \Phi_p(B)\varphi(B)\eta(B) Z_j = \Theta_q(B)\eta(B) Z_j.$$

Això mostra que $\{X_j\}$ és un procés estacionari que satisfà (4.4). Ara provarem la unicitat. Sigui $\{W_j\}$ un procés estacionari que satisfà (4.4); és a dir,

$$\Phi_p(B) W_j = \Theta_q(B)\eta(B) Z_j. \tag{4.8}$$

Com que $\Phi_p(z) \neq 0$ per a $|z| = 1$, podem escriure la sèrie de Laurent absolutament convergent

$$\frac{1}{\Phi_p(z)} = \sum_{i=0}^{\infty} h_i z^i = h(z),$$

on $\delta^{-1} < |z| < \delta$ per a $\delta > 1$. Per tant, multiplicant (4.8) per $h(B)$ i aplicant el Teorema 4.9, obtenim

$$W_j = h(B)\Theta_q(B)\eta(B)Z_j = \varphi(B)\eta(B)Z_j = \psi(B)Z_j,$$

que coincideix amb $\{X_j\}$ en mitjana quadràtica. □

Un cop demostrat el teorema, podem notar que les funcions d'autocovariància, d'autocorrelació i d'autocorrelació parcial són

$$\gamma(0) = \frac{\Gamma(1-2d)\sigma^2}{\Gamma^2(1-d)},$$

$$\rho(l) = \frac{\Gamma(1-d)\Gamma(i+d)}{\Gamma(1+i-d)\Gamma(d)} = \prod_{k=1}^{l} \frac{k-1+d}{k-d}, \quad l \geq 1$$

i

$$\alpha(l) = \frac{d}{l-d}, \quad l \geq 1.$$

D'acord amb la fórmula de Stirling, per x prou gran, tenim

$$\Gamma(x) \sim \sqrt{2\pi} e^{1-x}(x-1)^{x-\frac{1}{2}}.$$

Usant aquesta fórmula podem determinar de manera precisa la velocitat de convergència dels coeficients de la sèrie causal i les autocorrelacions. Si j és prou gran, els coeficients es comporten com

$$\psi_j \sim \frac{1}{j^{1-d}\Gamma(d)}.$$

En efecte,

$$\begin{aligned}
\frac{\Gamma(i+d)}{\Gamma(i+1)\Gamma(d)} &= \frac{e^{1-d}(i+d-1)^{i+d-\frac{1}{2}}}{\Gamma(d) i^{i+\frac{1}{2}}} \\
&= \frac{1}{\Gamma(d)} e^{1-d}(1+\frac{d-1}{i})^i \frac{(i+d-1)^{d-\frac{1}{2}}}{i^{\frac{1}{2}}} \\
&= \frac{i^{d-1}}{\Gamma(d)} e^{1-d}(1+\frac{d-1}{i})^i \frac{(i+d-1)^{d-\frac{1}{2}}}{i^{d-\frac{1}{2}}} \\
&= \frac{i^{d-1}}{\Gamma(d)} e^{1-d}(1+\frac{d-1}{i})^i (1+\frac{d-1}{i})^{d-\frac{1}{2}} \\
&\sim \frac{i^{d-1}}{\Gamma(d)}.
\end{aligned}$$

Fent estimacions similars obtenim

$$\rho(j) \sim \frac{\Gamma(1-d)}{\Gamma(d)} \frac{1}{j^{1-2d}}.$$

De fet, de

$$\rho(j) = \frac{\Gamma(1-d)}{\Gamma(d)} \frac{\Gamma(j+d)}{\Gamma(1+j-d)}$$

només hem de comprovar

$$\frac{\Gamma(j+d)}{\Gamma(1+j-d)} \sim \frac{1}{j^{1-2d}}.$$

Fem-ho:

$$\begin{aligned}
\frac{\Gamma(j+d)}{\Gamma(1+j-d)} &= \frac{e^{1-2d}(j+d-1)^{j+d-\frac{1}{2}}}{(j-d)^{j-d+\frac{1}{2}}} \\
&= \frac{e^{1-2d}(j-d+2d-1)^{j-d+\frac{1}{2}+2d-1}}{(j-d)^{j-d+\frac{1}{2}}} \\
&= \frac{(j-d+2d-1)^{j-d+\frac{1}{2}} e^{1-2d}(j+d-1)^{2d-1}}{(j-d)^{j-d+\frac{1}{2}}} \\
&= (1+\frac{2d-1}{j-d})^{j-d} e^{1-2d}(j+d-1)^{2d-1}(1+\frac{2d-1}{j-d})^{\frac{1}{2}} \\
&= (j+d-1)^{2d-1}(1+\frac{2d-1}{j-d})^{\frac{1}{2}} \\
&= \frac{(j+d-1)^{2d-\frac{1}{2}}}{(j-d)^{\frac{1}{2}}} \\
&= \frac{(j-d+2d-1)^{2d-\frac{1}{2}}}{(j-d)^{2d-\frac{1}{2}}(j-d)^{1-2d}} \\
&= \frac{1}{(j-d)^{1-2d}}(1+\frac{2d-1}{j-d})^{2d-\frac{1}{2}} \\
&= \frac{1}{j^{1-2d}}(1-\frac{d}{j})^{2d-1}(1+\frac{2d-1}{j-d})^{2d-\frac{1}{2}}.
\end{aligned}$$

I com que $(1-\frac{d}{j})^{2d-1}$ i $(1+\frac{2d-1}{j-d})^{2d-\frac{1}{2}}$ convergeixen a 1, tenim el resultat.

Observacions 4.12.

1. La sèrie de ψ_j és divergent en valor absolut per qualsevol $d \in [0,1)$. Si $d < 0$ la sèrie és convergent i el model té memòria curta.

2. Pel que fa al coeficient $|\rho(l)|$, veiem que per a $d \in [0, \frac{1}{2})$ la successió $\rho(l)$ convergeix a 0 en valor absolut, però $1-2d \in (0,1]$ i per tant

$$\sum_{l=0}^{\infty} |\rho(l)| = \infty.$$

3. Quan d és propera a $\frac{1}{2}$ l'efecte de la memòria llarga es fa més gran. Pel contrari, si $d < 0$ el model té memòria curta. També podem observar que si $d \geq \frac{1}{2}$ el model és no estacionari (veure [20]).

Teorema 4.13. *Sigui el procés $ARFIMA$ definit a (4.4). Suposem que els polinomis $\Phi_p(\cdot)$ i $\Theta_q(\cdot)$ no tenen arrels en comú i que $d \in (-1, \frac{1}{2})$. Aleshores,*

(a) Si les arrels de $\Phi_p(\cdot)$ es troben fora del cercle unitat tancat $\{z : |z| \leq 1\}$, aleshores la solució $\{X_j\}$ de (4.6) és causal.

(b) Si les arrels de $\Theta_q(\cdot)$ es troben fora del cercle unitat tancat $\{z : |z| \leq 1\}$, aleshores la solució $\{X_j\}$ de (4.6) és invertible.

Demostració.

(a) Suposem que $\Phi_p(z)$ no té arrels al disc unitat tancat $\{z : |z| \leq 1\}$. En tal cas, existeix $\varepsilon > 0$ tal que

$$\frac{1}{\Phi_p(z)} = \sum_{j=0}^{\infty} g_j z^j = g(z),$$

per a $|z| < 1+\varepsilon$. Aleshores, per qualsevol $\alpha \in (0, \varepsilon)$, la successió $g_j(1+\alpha)^j \xrightarrow{j\to\infty} 0$. Per tant, existeix una constant $K > 0$ tal que

$$|g_j| \leq K(1+\alpha)^{-j}, \quad j \geq 0.$$

Així, la successió $\{g_j\}$ és absolutament sumable. Ara, utilitzant el Teorema 4.9.1, tenim

$$Y_t = \alpha(B)X_t,$$

on $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ és un procés soroll fraccionari i on ψ_j són els coeficients donats a (4.5). Per altra banda, $\alpha(B) = g(B)\Theta_q(B) = \sum_{j=0}^{\infty} \alpha_j B^j$. Com que $\{\alpha_j : j \geq 0\}$ és absolutament sumable i X_t és estacionari, pel teorema 4.9 Y_t ve donat per

$$Y_t = \psi(B)Z_t,$$

on $\psi(B) = \alpha(B)\eta(B) = \sum_{j=0}^{\infty} \psi_j B^j$ és un procés causal ben definit i convergeix en L^2.

(b) Suposem que $\Theta_q(z)$ no té arrels al disc unitat tancat $\{z : |z| \leq 1\}$ i sigui $\omega_t = \varphi(B)Y_t$. Pel teorema 4.9.1, ω_t és un procés estacionari ben definit i amb densitat espectral

$$f_\omega(\lambda) = \frac{\sigma^2}{2\pi}|1 - e^{i\lambda}|^{-2d}|\theta(e^{i\lambda})|^2.$$

Veiem que f_ω compleix les condicions del Lema 4.10. Aleshores el desenvolupament

$$Z_t = \sum_{j=0}^{\infty} \tilde{\pi}\omega_{t-j} = \tilde{\pi}(B)\omega_t$$

convergeix en L^2, on $\tilde{\pi}(B) = \frac{1}{\eta(B)\theta(B)}$. Per últim, pel Teorema 4.9.2, obtenim la sèrie $L^2 - convergent$

$$Z_t = \sum_{j=0}^{\infty} \pi_j Y_{t-j} = \pi(B)Y_t,$$

on $\pi(B) = \tilde{\pi}(B)\varphi(B)$.

□

Observació 4.14. En tots els teoremes hem indicat que $d \in (-1, \frac{1}{2})$. De manera habitual s'acostumava a indicar aquest rang en $d \in (-\frac{1}{2}, \frac{1}{2})$. Per exemple, a [6], [14] utilitzen $d \in (-\frac{1}{2}, \frac{1}{2})$. L'extensió a -1 és deguda a [4].

4.5 Densitat espectral

Sota les hipòtesis del Teorema 4.13, la densitat espectral de

$$\Phi_p(B)X = \Theta_q(B)(Id - B)^{-d}Z$$

es pot escriure com

$$\begin{aligned} f(\lambda) &= \frac{\sigma^2}{2\pi}|1 - e^{-i\lambda}|^{-2d}\frac{|\Theta_q(e^{-i\lambda})|^2}{|\Phi_p(e^{-i\lambda})|^2} \\ &= \frac{\sigma^2}{2\pi}[2sin(\frac{\lambda}{2})]^{-2d}\frac{|\Theta_q(e^{-i\lambda})|^2}{|\Phi_p(e^{-i\lambda})|^2}. \end{aligned} \tag{4.9}$$

I aleshores

$$f(\lambda) \sim \frac{\sigma^2}{2\pi}\frac{|\Theta_q(1)|^2}{|\Phi_p(1)|^2}|\lambda|^{-2d},$$

quan $|\lambda| \to 0$, ja que $sin(\lambda) \sim \lambda$.

4.6 Funció d'autocovariància

Seguidament, es mostrarà un mètode per a calcular de manera exacta la funció d'autocovariància d'un procés $ARFIMA(p, d, q)$ utilitzant l'anomenat algoritme de Sowell. Considerant les parametritzacions de la funció d'autocovariància que es deriva d'escriure la densitat espectral (4.9) en termes dels paràmetres del model definit, la funció d'autocovariància d'un procés $ARFIMA(p, d, q)$ ve donada per

$$\gamma(h) = \frac{1}{2\pi}\int_0^{2\pi} f(\lambda)e^{-i\lambda h}\, d\lambda. \tag{4.10}$$

En particular, tal i com hem vist anteriorment, les funcions d'autocovariància i autocorrelació d'un procés $ARFIMA(0, d, 0)$ són, respectivament

$$\gamma(h) = \sigma^2\frac{\Gamma(1-2d)}{\Gamma(1-d)\Gamma(d)}\frac{\Gamma(h+d)}{\Gamma(1+h-d)}$$

i

$$\rho(h) = \frac{\Gamma(1-d)}{\Gamma(d)}\frac{\Gamma(h+d)}{\Gamma(1+h-d)}.$$

Aleshores, el polinomi $\Phi_p(B)$ definit a (4.3) el podem escriure com

$$\Phi_p(B) = \prod_{i=1}^{p}(1 - \rho(i)B).$$

Sota la hipòtesi que totes les arrels de $\Phi_p(B)$ tenen multiplicitat 1, podem deduir de (4.10) que

$$\gamma(h) = \sigma^2\sum_{i=-q}^{q}\sum_{j=1}^{p}\psi(i)\xi_j C(d, p+i-h, \rho(j)),$$

amb

$$\xi_j = \left[\rho(j)\prod_{i=1}^{p}(1 - \rho(i)\rho(j))\prod_{k\neq j}(\rho(j) - \rho(k))\right]^{-1},$$

$$C(d,h,\rho) = \frac{\gamma(h)}{\sigma^2}[\rho^{2p}\beta(h) + \beta(-h) - 1],$$
$$\beta(h) = F(d+h, 1, 1-d+h, \rho),$$
$$F(a,b,c,x) = 1 + \frac{a \cdot b}{c \cdot 1}x + \frac{a \cdot (a+1) \cdot b \cdot (b+1)}{c \cdot (c+1) \cdot 1 \cdot 2}x^2 + \cdots,$$

on aquesta funció $F(a,b,c,x)$ és la funció hipergeomètrica Gaussiana. El terme $\psi(i)$ és

$$\psi(i) = \sum_{k=max(0,i)}^{min(q,q+i)} \theta_k \theta_{k-i}.$$

En absència de paràmetres AR, la fórmula per a la funció d'autocovariància es redueix a

$$\gamma(h) = \sigma^2 \sum_{i=-q}^{q} \psi(i) \frac{\Gamma(1-2d)\Gamma(h+d-i)}{\Gamma(1-d)\Gamma(d)\Gamma(1+i-d-h)}.$$

Per altra banda, podem descriure el comportament asimptòtic de la funció d'autocovariància com

$$\gamma(h) \sim c_\gamma |h|^{2d-1}, \tag{4.11}$$

on

$$c_\gamma = \frac{\sigma^2}{\pi} \frac{|\Theta_q(1)|^2}{|\Phi_p(1)|^2} \Gamma(1-2d) sin(\pi d),$$

per a $|h|$ gran.

Ara, sigui $\{y_1, y_2, ..., y_n\}$ una mostra del procés definit a (4.4) i sigui $\bar{y}$ la mitjana mostral. La variància exacta de $\bar{y}$ ve donada per l'expressió

$$\mathbb{V}(\bar{y}) = \frac{1}{n}\left[2\sum_{j=1}^{n-1}\left(1 - \frac{j}{n}\right)\gamma(j) + \gamma(0)\right].$$

Per a n gran, per (4.11), tenim que de manera asimptòtica la fórmula de la variància és

$$\mathbb{V}(\bar{y}) \sim \frac{c_\gamma}{d(2d+1)} n^{2d-1}.$$

4.7 Funció de resposta a l'impuls

La funció de resposta a l'impuls (IRF) és l'eina més utilitzada per avaluar l'efecte de les pertorbacions en les sèries temporals. A continuació mostrarem una de les diverses aproximacions que es poden fer per a trobar la IRF d'un procés $ARFIMA(p,d,q)$.

Suposem que les arrels dels polinomis $\Phi_p(B)$ i $\Theta_q(B)$ es troben fora del cercle unitat tancat i que $d \in (-1, \frac{1}{2})$. Aleshores el procés és estacionari (Teorema 4.11), causal i invertible (Teorema 4.13). En aquest cas, tenim l'aproximació asimptòtica dels coeficients MA

$$\psi_j \sim \left(\frac{1 + \sum_{i=1}^{q} \theta_i}{1 - \sum_{i=1}^{p} \phi_i}\right) \frac{j^{d-1}}{\Gamma(d)},$$

quan $j \to \infty$.

Com a cas particular, tenim que per a un procés $ARFIMA(p,d,q)$ els coeficients MA són

$$\psi_j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)}.$$

Ara, usant que

$$(Id-B)^{-d} = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)} B^j = \sum_{j=0}^{\infty} \eta_j B^j$$

i la representació de Wold

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

tenim que el procés definit a (4.4) es pot escriure com

$$(Id-B)^{-d} X_t = \sum_{j=0}^{\infty} \eta_j X_{t-j} = \sum_{j=0}^{\infty} R_j Z_{t-j},$$

on R_j és l'anomenada IRF i ve donada per

$$R_j = \sum_{i=0}^{j} \psi_i \eta_{j-i}. \tag{4.12}$$

Ara, usant que

$$\eta_j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)},$$

els termes η_j es poden expressar recursivament com $\eta_j = \eta_{j-1}(1+\frac{d-1}{j})$, per a $j \geq 1$ i $\eta_0 = 1$. Utilitzant que per a j grans, l'expressió asimptòtica és

$$\eta_j \sim \frac{j^{d-1}}{\Gamma(d)},$$

i suposant que $\sum_{j=0}^{\infty} \psi_j < \infty$, tenim la representació asimptòtica següent:

$$R_j \sim \frac{j^{d-1}}{\Gamma(d)} \sum_{i=0}^{\infty} \psi_i, \tag{4.13}$$

quan $j \to \infty$ i $\frac{\psi_j}{j^{d-1}} \to 0$.

4.8 Mètodes d'estimació

A continuació es mostrarà com estimar, en primer lloc l'esperança i la funció d'autocorrelació ρ. Seguidament, s'exposaran una sèrie de mètodes que ens permetran estimar els paràmetres dels models amb memòria llarga. Per a fer aquestes estimacions, a diferència de la secció anterior, prendrem $d \in (-\frac{1}{2}, \frac{1}{2})$.

4.8.1 Estimació de l'esperança

Sigui $\{X_t, t \in \mathbb{Z}\}$ un procés $ARFIMA(p,d,q)$ causal i invertible definit (com a (4.3)) per:

$$\Phi_p(B)(Id - B)^d(X_t - \mu) = \Theta_q(B)Z_t, \tag{4.14}$$

on $Z_t \sim WN(0, \sigma^2)$ i $d \in (-\frac{1}{2}, \frac{1}{2})$.
Un estimador natural de l'esperança $\mathbb{E}(X_t) = \mu$ és la mitjana mostral

$$\overline{X}_n = \frac{X_1 + ... + X_n}{n}.$$

Com que la funció d'autocorrelació ρ de $\{X_t\}$ compleix que $\rho(h) \to 0$ quan $h \to \infty$, aleshores

$$\mathbb{E}(\overline{X}_n - \mu)^2 \to 0 \;\; si \;\; n \to \infty,$$

i

$$n\mathbb{E}(\overline{X}_n - \mu)^2 \to \begin{cases} 0 \; si \; -\frac{1}{2} < d < 0, \\ \infty \; si \; 0 < d < \frac{1}{2} \end{cases}$$

L'estimador mitjana mostral, però, no és asimptòticament normal perquè la suma dels valors absoluts de les autocorrelacions no és convergent.

4.8.2 Estimació de la funció d'autocorrelació

La funció d'autocorrelació ρ s'acostuma a estimar a través de la funció d'autocorrelació mostral $\hat{\rho}$. Si $-\frac{1}{2} < d < 0$, $\{X_t\}$ té la representació mitjana mòbil

$$X_t = \mu + \sum_{j=0}^{\infty} \psi_j Z_{t-j},$$

on $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Si $0 < d < \frac{1}{2}$ la situació és més complicada ja que el comportament asimptòtic de l'estimador no està clar.

4.8.3 Estimació dels paràmetres

Hi ha una gran varietat de mètodes que ens permeten estimar els paràmetres d, ϕ i θ dels models $ARFIMA$. A continuació presentarem alguns dels més utilitzats: la màxima versemblança, les aproximacions autorregressives, l'estimador de Whittle i un mètode de regressió.

(A) Màxima versemblança. La versemblança Gaussiana de $X = (X_1, ..., X_n)$ del procés definit a (4.14) amb $\mu = 0$ la podem expressar de la següent manera:

$$\mathcal{L}(\beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}}(r_0 \cdots r_{n-1})^{-\frac{1}{2}} exp\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(X_j - \hat{X}_j)^2/r_{j-1}\},$$

on $\beta = (d, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q)$, $\hat{X}_j, j = 1, ..., n$ són els predictors i $r_{j-1} = \sigma^{-2}\mathbb{E}(X_j - \hat{X}_j)^2, j = 1, ..., n$. Els estimadors màxim-versemblants $\hat{\beta}$ i $\hat{\sigma}^2$ es poden trobar maximitzant $L(\beta, \sigma^2)$ respecte β i σ^2. Acabem obtenint que

$$\hat{\sigma}^2 = \frac{S(\hat{\beta})}{n}$$

on

$$S(\hat{\beta}) = \sum_{j=1}^{n} (X_j - \hat{X}_j)^2 / r_{j-1}$$

i $\hat{\beta}$ és el valor de β que minimitza

$$\ell(\beta) = log\left(\frac{S(\beta)}{n}\right) + \frac{\sum_{j=1}^{n} log(r_{j-1})}{n}.$$

(B) Aproximacions autorregressives. Donada la complexitat computacional de calcular exactament els estimadors màxim-versemblants, molts autors han considerat fer ús d'aproximacions autorregressives per tal d'accelerar el procés de càlcul d'estimacions dels paràmetres.

Sigui $\{Y_t, t \in \mathbb{Z}\}$ un procés amb memòria llarga definit pel desenvolupament autorregressiu

$$Y_t = Z_t + \pi_1(\theta)Y_{t-1} + \pi_2(\theta)Y_{t-2} + ...,$$

on $\pi_j(\theta)$ són els coeficients de $\phi(B)\theta^{-1}(B)(Id - B)^d$. Com que a la pràctica només tenim disponible un nombre finit d'observacions $\{Y_1, ..., Y_n\}$, definirem el següent model truncat

$$Y_t = \tilde{Z}_t + \pi_1(\theta)Y_{t-1} + \pi_2(\theta)Y_{t-2} + ... + \pi_m(\theta)Y_{t-m},$$

on $m < t \leq n$. Aleshores l'estimador màxim-versemblant aproximat $\hat{\theta}_n$ l'obtenim minimitzant la funció

$$\mathcal{L}_1(\theta) = \sum_{t=m+1}^{n} [Y_t - \pi_1(\theta)Y_{t-1} - \pi_2(\theta)Y_{t-2} - ... - \pi_m(\theta)Y_{t-m}]^2.$$

Hi ha diversos mètodes proposats relacionats amb les aproximacions autorregressives que permeten obtenir unes millors estimacions, com el mètode Haslet-Raftery o el de Beran. Aquests i altres poden ser consultats a [17].

(C) L'estimador de Whittle. Hem de començar per definir què és un *periodograma*.

Definició 4.15. *La densitat espectral d'un procés $\{Y_t, t \in \mathbb{Z}\}$ és estimada pel periodograma*

$$I(\lambda) = \frac{1}{2\pi n}\left|\sum_{j=1}^{n} Y_j e^{i\lambda j}\right|^2.$$

Un cop definit, podem explicar aquesta metodologia per obtenir estimadors màxim-versemblants aproximats, que està basada en el càlcul del periodograma mitjançant la transformada ràpida de Fourier (FFT) i l'ús de l'anomenada aproximació de Whittle de la funció log-versemblant Gaussiana. Com que el càlcul de la FFT té una complexitat numèrica d'ordre $\mathcal{O}[nlog_2(n)]$, aquest enfocament produeix algoritmes molt ràpids per calcular estimacions de paràmetres. Suposarem que el vector mostral $Y = (y_1, ..., y_n)$ té distribució normal amb esperança 0 i variància Γ_θ. Aleshores, la funció log-versemblant dividida per la mida de la mostra és

$$\mathcal{L}(\theta) = -\frac{1}{2n} log\, det\Gamma_\theta - \frac{1}{2n} y^\intercal \Gamma_\theta^{-1} y. \tag{4.15}$$

La matriu de variàncies i covariàncies Γ_θ s'expressa en termes de la densitat espectral del procés $f_\theta(\cdot)$ de la següent manera:

$$(\Gamma_\theta)_{ij} = \gamma_\theta(i-j),$$

on

$$\gamma_\theta(k) = \int_{-\pi}^{\pi} f_\theta(\lambda) exp(i\lambda k)\, d\lambda.$$

Per tal d'aplicar el mètode de Whittle, s'han de fer dues aproximacions. Donat que

$$\frac{1}{n} log\, det\Gamma_\theta \to \frac{1}{2\pi}\int_{-\pi}^{\pi} log[2\pi f_\theta(\lambda)]\, d\lambda,$$

quan $n \to \infty$, el primer terme de (4.15) s'aproxima per

$$\frac{1}{2n} log\, det\Gamma_\theta \approx \frac{1}{4\pi}\int_{-\pi}^{\pi} log[2\pi f_\theta(\lambda)]\, d\lambda.$$

Per altra banda, el segon terme de (4.15) s'aproxima per

$$\begin{aligned}
\frac{1}{2n} y^\intercal \Gamma_\theta^{-1} y &\approx \sum_{l=1}^{n}\sum_{j=1}^{n} y_l \{\frac{1}{8\pi^2 n}\int_{-\pi}^{\pi} f_\theta^{-1}(\lambda) exp[i\lambda(l-j)]\, d\lambda\} y_j \\
&= \frac{1}{8\pi^2 n}\int_{-\pi}^{\pi} f_\theta^{-1}(\lambda) \sum_{l=1}^{n}\sum_{j=1}^{n} y_l y_j exp[i\lambda(l-j)]\, d\lambda \\
&= \frac{1}{4\pi}\int_{-\pi}^{\pi} \frac{I(\lambda)}{f_\theta(\lambda)}\, d\lambda,
\end{aligned}$$

on

$$I(\lambda) = \frac{1}{2\pi n}\left|\sum_{j=1}^{n} y_j e^{i\lambda j}\right|^2$$

és el periodograma de la sèrie $\{Y_t\}$.

Per tant, la funció log-versemblança és aproximada per

$$\mathcal{L}_2(\theta) = -\frac{1}{4\pi}\left[\int_{-\pi}^{\pi} log f_\theta(\lambda)\, d\lambda + \int_{-\pi}^{\pi} \frac{I(\lambda)}{f_\theta(\lambda)}\, d\lambda\right].$$

Si substituim les integrals per sumes de Riemann per facilitar la computació, obtenim l'expressió alternativa

$$\mathcal{L}_3(\theta) = -\frac{1}{2n}\left[\sum_{j=1}^{n} log f_\theta(\lambda_j) + \sum_{j=1}^{n} \frac{I(\lambda_j)}{f_\theta(\lambda_j)}\right].$$

(D) Un mètode de regressió. Podem escriure la densitat espectral d'un procés estacionari com

$$f(\lambda) = f_0(\lambda)\left[2sin\left(\frac{\lambda}{2}\right)\right]^{-2d}. \tag{4.16}$$

Geweke i Porter-Hudak [12] proposen el següent mètode de regressió per a estimar paràmetres.

Prenent logaritmes als dos costats de l'equació (4.16) i avaluant la densitat pectral en les freqüències de Fourier $\lambda_j = 2\pi j/n$, tenim

$$logf(\lambda_j) = logf_0(0) - dlog\left[2sin\left(\frac{\lambda_j}{2}\right)\right]^2 + log\left[\frac{f_0(\lambda_j)}{f_0(0)}\right]. \tag{4.17}$$

Per altra banda, el logaritme del periodograma $I(\lambda_j)$ és

$$logI(\lambda_j) = log\left[\frac{I(\lambda_j)}{f(\lambda_j)}\right] + logf(\lambda_j). \tag{4.18}$$

Si combinem (4.17) i (4.18) obtenim

$$logI(\lambda_j) = logf_0(0) - dlog\left[2sin\left(\frac{\lambda_j}{2}\right)\right]^2 + log\left\{\frac{I(\lambda_j)[2sin(\lambda/2)]^{2d}}{f_0(0)}\right\}.$$

Si definim $y_j = logI(\lambda_j), \alpha = logf_0(0), \beta = -d, x_j = log\left[2sin\left(\frac{\lambda_j}{2}\right)\right]^2$ i

$$\varepsilon_j = log\left\{\frac{I(\lambda_j)[2sin(\lambda/2)]^{2d}}{f_0(0)}\right\},$$

obtenim la següent equació de regressió:

$$y_j = \alpha + \beta x_j + \varepsilon_j.$$

L'estimació per mínims quadrats del paràmetre d ve donada per

$$\hat{d}_m = -\frac{\sum_{j=1}^m (x_j - \overline{x})(y_j - \overline{y})}{\sum_{j=1}^m (x_j - \overline{x})^2},$$

on $m \ll n$, $\overline{x} = \frac{\sum_{j=1}^m x_j}{m}$ i $\overline{y} = \frac{\sum_{j=1}^m y_j}{m}$.

Per acabar aquesta secció, mostrarem un algoritme que ens permetrà estimar el paràmetre d usant el mètode de regressió explicat anteriorment. Sigui X_t el procés definit a (4.3). Aleshores recordem que $U_t = (Id - B)^d X_t$ és un procés $ARMA(p,q)$ i $Y_t = \frac{\Phi_p(B)}{\Theta_q(B)} X_t$ és un procés $ARFIMA(0,d,0)$. L'algoritme és el següent:

1. Estimem d del model $ARIMA(p,d,q)$ i denotem l'estimació per $\hat{d}$.
2. Calculem $\hat{U}_t = (Id - B)^{\hat{d}} X_t$.
3. Usem el procediment Box-Jenkins (veure [5]) o el criteri AIC (veure[2]) per tal d'identificar i estimar els paràmetres ϕ i θ del procés $ARMA(p,q)$, $\phi(B)\hat{U}_t = \theta(B)Z_t$.
4. Calculem $\hat{Y}_t = \frac{\hat{\Phi}_p(B)}{\hat{\Theta}_q(B)} X_t$.
5. Estimem d del model $ARFIMA(0,d,0)$, $(Id - B)^{\hat{d}}\hat{Y}_t = Z_t$. El valor de $\hat{d}$ obtingut en aquest pas és ara la nova estimació de d.
6. Repetim els passos del 2 al 5, fins que les estimacions dels paràmetres d, ϕ i θ convergeixin.

4.9 Tria del millor model

Un cop estimats els paràmetres, ens podem trobar amb que tenim un conjunt de possibles models candidats que es poden ajustar a les nostres dades. En aquesta secció veurem diferents criteris que ens permetran seleccionar el millor de tots aquests models. El funcionament serà el mateix per a cada criteri d'informació: el model que presenti un valor més baix serà el model que s'adequarà millor a les nostres dades. En cas que els criteris d'informació de diversos models siguin similars, es triarà el que hagi d'estimar un menor nombre de paràmetres. Dit això, considerarem els valors i criteris següents:

(A) Log-versemblança: podem mesurar la bondat d'ajustament d'un model mitjançant el valor logarítmic de versemblança d'un model regressiu. Aquesta log-versemblança l'hem vista a la secció anterior (estimació de paràmetres). En aquest cas, un valor més gran implica un millor ajustament del model. Aquest valor es troba entre menys infinit i més infinit.

(B) Criteri d'informació d'Akaike (AIC): és un estimador de la predicció de l'error i, en conseqüència, de la qualitat relativa del model. Quan un model estadístic s'utilitza per a representar el procés que ha generat les dades, aquesta representació gairebé mai serà exacta, per tant hi haura informació que es perdrà usant el model en qüestió. El que fa l'AIC és estimar la quantitat relativa d'informació perduda per un model determinat. Com menys informació perdi el model, millor serà la qualitat d'aquest. Per tal d'estimar aquest valor, l'AIC recompensa la bondat de l'ajustament i a més també inclou una penalització, que és una funció creixent dels paràmetres estimats. Aquesta penalització té en compte el risc del sobreajustament; és a dir, augmentar voluntàriament el nombre de paràmetres lliures en el model per tal de millorar la bondat de l'ajustament.

Definim, doncs, el valor AIC com:

$$AIC = 2k - 2log(\hat{L}),$$

on k és el nombre de paràmetres estimats del model i $\hat{L}$ és el màxim valor de la funció versemblança pel model estimat.

(C) Criteri d'informació de Bayes (BIC): també conegut com a criteri d'informació de Schwarz (SIC), recompensa la bondat de l'ajustament i fa que hi hagi una penalització per anar augmentant de manera excessiva el nombre de paràmetres. Les principals diferències amb l'AIC són que en el BIC la penalització és major i que aquest criteri d'informació no és assimptòticament eficient.

Definim el valor BIC com la funció:

$$BIC = klog(n) - 2log(\hat{L}),$$

on n és el nombre d'observacions, k és el nombre de paràmetres estimats del model i $\hat{L}$ és el màxim valor de la funció versemblança.

4.10 Predicció

Per últim, proposarem un parell de procediments que ens serviran per predir els valors futurs d'una sèrie temporal: els predictors d'un pas endavant i els predictors de múltiples passos endavant.

4.10.1 Predictors d'un pas endavant

Sigui $\{y_t\}$ un procés lineal regular invertible amb representació de Wold

$$y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

i amb l'expansió $AR(\infty)$

$$y_t - \sum_{j=1}^{\infty} \pi_j y_{t-j} = \varepsilon_t,$$

on $\mathbb{V}(\varepsilon_t) = \sigma^2$.

Ara, sigui $\{y_1, ..., y_t\}$ un conjunt finit d'observacions. Denotem $\mathcal{P}_t$ com l'espai de Hilbert generat per les observacions $\{y_1, ..., y_t\}$. Aleshores, el millor predictor lineal d'una observació y_{t+1} basat en un passat finit ve donat per l'expressió

$$\tilde{y}_{t+1} = \mathbb{E}[y_{t+1}|\mathcal{P}_t] = \phi_{t1} y_t + ... + \phi_{tt} y_1,$$

on $\phi_t = (\phi_{1t}, ..., \phi_{tt})^\intercal$ és l'única solució de l'equació lineal

$$\Gamma_t \phi_t = \gamma_t,$$

amb $\Gamma_t = [\gamma(i-j)]_{i,j=1,...,t}$ i $\gamma_t = [\gamma(1), ..., \gamma(t)]^\intercal$. El càlcul d'aquests coeficients es pot dur a terme mitjançant diferents algoritmes, com el de Durbin-Levinson (veure pàgines 66 i 67 de [17]). A més, la variància de l'error de predicció del predictor $\tilde{y}_{t+1}$,

$$\nu_t = \mathbb{E}[y_{t+1} - \tilde{y}_{t+1}]^2,$$

es calcula utilitzant les equacions recursives

$$\nu_t = \nu_{t-1}(1 - \phi_{tt}^2),$$

per a $t \geq 1$. Els coeficients ϕ_{tt} són els coeficients d'autocorrelació parcial i $v_0 = \gamma(0)$. Per tant, podem escriure v_t com

$$v_t = \gamma(0) \prod_{j=1}^{t} (1 - \phi_{jj}^2).$$

Per altra banda, és difícil trobar expressions explícites per a les autocorrelacions parcials de models amb memòria llarga. Tot i així, aquestes autocorrelacions parcials tendeixen a zero en els processos $ARFIMA(p, d, q)$, tal i com mostra [16].

Teorema 4.16. *Sigui $\{y_t\}$ un procés $ARFIMA(p, d, q)$ amb $0 < d < \frac{1}{2}$. Aleshores, els coeficients d'autocorrelació parcial ϕ_{tt} satisfan*

$$|\phi_{tt}| \sim \frac{d}{t}, \tag{4.19}$$

quan $t \to \infty$.

Per últim, introduirem el predictor aproximat. L'algoritme de Durbin-Levinson per a calcular els coeficients ϕ_{tj} és d'ordre $\mathcal{O}(n^2)$. Per tant, per mostres amb moltes observacions ens convindria un algoritme més ràpid per obtenir prediccions d'observacions finites. Una manera de fer-ho és aproximant els coeficients de regressió ϕ_{tj} per π_j, basant-nos en el lema següent:

Lema 4.17. *Si $\{y_t\}$ és un procés estacionari, aleshores $\phi_{tj} \xrightarrow{t\to\infty} \pi_j$.*

Amb aquesta aproximació, podem definir el predictor de mostra simple com

$$\check{y}_{t+1} = \sum_{j=1}^{t} \pi_j y_{t+1-j},$$

on la variància de l'error de predicció és

$$\mathbb{V}(y_{t+1} - \check{y}_{t+1}) = \sigma^2 + r_t,$$

on $r_t = \mathbb{V}\left(\sum_{j=t+1}^{\infty} \pi_j y_{t+1-j}\right)$.

Veiem que la variància de l'error de predicció d'aquest predictor aproximat és més gran que la variància de l'error de predicció de $\tilde{y}_{t+1}$. Tot i així, a mesura que t creix, aquests dos predictors es fan similars.

4.10.2 Predictors de múltiples passos endavant

Sigui h el nombre de passos endavant. Com abans, sigui $\{y_1, ..., y_t\}$ un conjunt finit d'observacions i denotem $\mathcal{P}_t$ com l'espai de Hilbert generat per les observacions $\{y_1, ..., y_t\}$. El millor predictor lineal de y_{t+h} basat en un passat finit $\mathcal{P}_t$ és

$$\tilde{y}_t(h) = \phi_{t1}(h)y_t + ... + \phi_{tt}(h)y_1,$$

on $\phi_t(h) = [\phi_{t1}(h), ..., \phi_{tt}(h)]^\intercal$ satisfà

$$\Gamma_t \phi_t(h) = \gamma_t(h),$$

amb $\gamma_t(h) = [\gamma(h), ..., \gamma(t+h-1)]^\intercal$.

De manera anàloga a la predicció d'un pas endavant, podem utilitzar les aproximacions de mostra finita de h passos endavant, que venen donades per l'expressió

$$\check{y}_t(h) = \sum_{j=0}^{t} \pi_j(h) y_{t-j},$$

que té variància de l'error de predicció

$$\mathbb{V}(y_{t+h} - \check{y}_t(h)) = \sigma^2 + r_t(h),$$

on $r_t(h) = \mathbb{V}\left(\sum_{j=t+1}^{\infty} \pi_j(h) y_{t-j}\right)$.

5 Anàlisi pràctica: qualitat de l'aire de Barcelona

Un cop presentats i analitzats tots aquells aspectes més importants relacionats amb les sèries temporals amb memòria llarga i, en particular, dels models $ARFIMA$, en aquesta darrera secció durem a terme l'anàlisi d'una sèrie temporal real. En aquesta anàlisi veurem si la sèrie escollida presenta un comportament de memòria llarga i n'ajustarem el model $ARFIMA$ més adient. Seguidament, procedirem a validar el model i veurem i calcularem la major part dels conceptes explicats a la secció anterior.

Per tal de poder realitzar l'anàlisi, utilitzarem el llenguatge de programació amb aplicacions de caire estadístic R (en concret farem servir l'entorn de desenvolupament integrat *RStudio*). Totes les llibreries i comandes utilitzades es poden trobar a l'annex del treball. Cal destacar que la llibreria principal utilitzada que ens permetrà ajustar el model i calcular i dibuixar gràficament molts dels aspectes explicats és la llibreria *afmtools*. Aquesta llibreria fou desenvolupada per Contreras-Reyes, Goerg i Palma i les seves funcions i explicacions es troben a [9].

5.1 La sèrie

En primer lloc, presentarem la sèrie escollida i n'explicarem alguns detalls per a que, abans de fer-ne una anàlisi matemàtica, coneguem amb detall de què estem parlant i què és el que estem analitzant. La major part de la informació presentada ha estat extreta de [1] i [11].

La sèrie temporal que estudiarem es pot titular com la qualitat de l'aire de Barcelona entre els anys 2017 i 2019. A què ens referim amb això? Una intensa activitat industrial, un elevat volum de tràfic a la ciutat de Barcelona i a la seva àrea metropolitana i un clima mediterrani on de mitjana plou poc al llarg de l'any són factors que afavoreixen la concentració d'elements contaminants a l'aire que respirem, fet que afecta perjudicialment la nostra salut. Per tal de mesurar aquests contaminants, Catalunya disposa de l'índex català de qualitat de l'aire (ICQA), el qual es determina a partir de la quantitat a l'aire de vuit contaminants clau i descriu la situació de la qualitat de l'aire a cada estació de control en un moment determinat.

Els llindars dels contaminants per al càlcul de l'ICQA són les concentracions en $\mu g/m^3$ (excepte pel CO, pel qual són en mg/m^3). Per a obtenir cada valor, s'utilitza com a base temporal de càlcul la mitjana horària per a tots els contaminants, a excepció del $PM10$ i el $PM2,5$, on per a cada hora es mostra el valor de la mitjana en les 24 hores anteriors.

Per a que ens fem una idea dels llindars dels valors d'alguns dels contaminants i què significa cada interval de valors, mostrem la següent taula resum.

Contaminants (base temporal de càlcul)		Nivell de ICQA (concentració en µg/m³, excepte pel CO que són en mg/m³)					
		Bona	Raonablement bona	Regular	Desfavorable	Molt desfavorable	Extremadament desfavorable
Diòxid de nitrogen (NO_2)	1-hora	0-40	41-90	91-120	121-230	231-340	>340
Partícules en suspensió PM10	24-hores	0-20	21-40	41-50	51-100	101-150	>150
Partícules en suspensió PM2,5	24-hores	0-10	11-20	21-25	26-50	51-75	>75
Ozó troposfèric (O_3)	8-horari	0-50	51-100	101-130	131-240	241-380	>380
Diòxid de sofre (SO_2)	1-hora	0-100	101-200	201-350	351-500	501-750	>750
Monòxid de carboni (CO)	1-hora	0-2	3-5	6-10	11-20	21-50	>50

Figura 3: Intervals dels possibles valors del nivell de ICQA de diferents contaminants. Font: Web de l'apartat de medi ambient i sostenibilitat del Gencat, veure [11].

La interpretació de cada nivell de l'ICQA és la següent:

- Bona: les concentracions de partícules perjudicials per a la salut són molt baixes i no hi hauria d'haver riscos.
- Raonablement bona: només les persones extremadament sensibles es podrien veure afectades i haurien de limitar els seus esforços a l'aire lliure.
- Regular: les persones amb malalties respiratòries o grups concrets de població sensible podrien presentar símptomes i dificultats respiratòries lleus.
- Desfavorable: la qualitat de l'aire és dolenta i els grups sensibles i persones amb malalties respiratòries com l'asma es veuran afectades. Han d'evitar fer esforços a l'aire lliure.
- Molt desfavorable: la qualitat de l'aire és molt dolenta. No han de realitzar activitats a l'aire lliure cap persona sensible o amb problemes respiratoris. La resta de la població ha de reduir les activitats a l'aire lliure.
- Extremadament desfavorable: condició d'emergència. Tota la població es pot veure afectada greument i ha d'evitar qualsevol activitat a l'aire lliure.

Els dos elements contaminants que es troben de manera més abundant a Barcelona són el diòxid de nitrogen (NO_2) i les partícules en suspensió de menys de 10 micres $PM10$. Tot i que no són visibles als nostres ulls, aquests contaminants són perjudicials per a la salut de tothom i especialment per a grups vulnerables com les persones amb malalties cardiovasculars i respiratòries, els infants o la gent gran.

Com que tan sols analitzarem una sèrie temporal, el contaminant escollit ha estat el NO_2. El diòxid de nitrogen és un gas irritant que ataca als pulmons i que es produeix per la combustió de vehicles de motor, el transport marítim o les calefaccions. Aproximadament un 60% del total prové del trànsit.

Feta aquesta introducció explicatòria, procedirem a comentar com hem obtingut les dades de la sèrie temporal i quines observacions hem triat exactament. Les dades es poden trobar a [11]. Dins d'aquest portal podem filtrar totes les dades disponibles pel nom de l'estació, el contaminant i les dates d'inici i fi. Per a la nostra sèrie temporal, hem triat l'estació de Barcelona - Eixample (Barcelona té 11 estacions de control i n'havíem de triar una). El contaminant escollit ha estat el NO_2, tal i com hem comentat, i la mostra agafada comença l'1 de gener del 2017 i finalitza el 31 de desembre del 2019. Per altra banda, com veiem al web, disposem d'un valor per a cada hora del dia. El que hem fet és triar el valor corresponent a les 12h del migdia per a cada dia. Així, la nostra sèrie temporal té freqüència diària i té un total de 1095 observacions.

Com a últim fet a destacar, hem observat que els valors de l'ICQA per al NO_2 i les $PM10$ estan correlacionats; és a dir, quan els valors del nivell de NO_2 a l'aire són alts, els de les $PM10$ també ho són, i viceversa. Per tant, hauríem obtingut una sèrie temporal similar havent triat com a contaminant les $PM10$.

5.2 Les dades

Ara que coneixem bé la sèrie amb la qual treballarem, començarem amb l'anàlisi d'aquesta. Primerament, farem una petita descriptiva de les dades de les quals disposem. La següent taula mostra alguna de la informació més rellevant de la sèrie.

Mínim	Màxim	Mitjana	Mediana	1r Quartil	3r Quartil
12.00	187.00	57.37	53.00	38.00	72.00

Taula 1: Principal informació referent a les dades de la sèrie temporal.

Com podem veure, la mitjana i la mediana tenen uns valors de 57.37 i 53, respectivament. Aquests valors es troben a l'interval d'un nivell de l'ICQA raonablement bo. Per tant, podem afirmar que de mitjana la qualitat de l'aire de Barcelona entre els anys 2017 i 2019 pel que fa a la concentració de NO_2 és prou bona. Per altra banda, podem veure quantes observacions es troben en cada un dels intervals definits a la Figura 3.

Nivell de ICQA	Número d'observacions	Percentatge del total
Bona (0-40)	318	29.04%
Raonablement bona (41-90)	662	60.46%
Regular (91-120)	91	8.31%
Desfavorable (121-230)	24	2.19%
Molt desfavorable (231-340)	0	0%
Extremadament desfavorable (340)	0	0%

Taula 2: Distribució de les observacions segons el nivell de ICQA.

Es pot observar que la major part de les observacions es troben en els dos primers intervals (un 89.5% del total entre tots dos), mentre que els nivells de ICQA regulars i desfavorables ocupen un 10.5% del total. No hi ha cap observació que es trobi en els nivells de ICQA més preocupants. Per tant, podem confirmar que la qualitat de l'aire pel que fa a la concentració de NO_2 és en general raonablement bona en els anys triats i assoleix nivells desfavorables en comptades ocasions. Tot i així aquestes observacions que

tenen pics tan alts en relació a la mitjana de la sèrie podrien tenir efectes importants a l'hora de modelitzar-la. Com hem vist al llarg del treball, en els models $ARMA$ els efectes d'aquestes "innovacions" desapareixen ràpidament i en els models $ARIMA$ tenen efectes permanents. És per això que per a aquest tipus de sèrie pot ser interessant plantejar un model $ARFIMA$ que la descrigui.

Abans d'estimar un model per a la nostra sèrie, mostrarem un parell de figures més que ens ajudaran a entendre-la millor i a concloure que efectivament el model que millor s'hi pot adaptar és un model $ARFIMA$. En primer lloc, a la Figura 4 podem veure la representació gràfica de la sèrie. Hem considerat aplicar una transformació logarítmica a la sèrie, ja que la variància no és del tot constant. Però no s'aprecia ni estacionalitat ni una tendència, i havent fet una prova amb la sèrie transformada logarítmicament hem obtingut resultats molt similars. Per tant seguirem l'anàlisi amb la sèrie original.

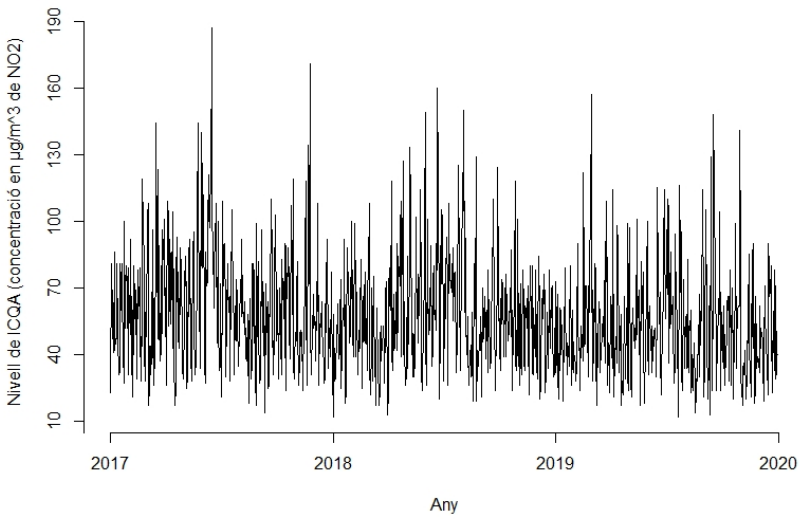

Figura 4: Representació gràfica del nivell de ICQA de NO_2 a Barcelona (Eixample) entre l'1 de gener de 2017 i el 31 de desembre de 2019.

Per altra banda, a la Figura 5 podem trobar un correlograma de la funció d'autocorrelació (ACF). Analitzant el correlograma és on clarament ens adonem que la sèrie presenta la propietat estudiada de memòria llarga, ja que les ACF decauen de manera lenta i exponencial fins a convergir a zero. Cal destacar que al correlograma presentat hi ha un retard de $l = 1000$, fet pel qual veiem que la sèrie té una memòria molt llarga i acabem de confirmar que un model $ARFIMA$ serà adient per a modelitzar-la.

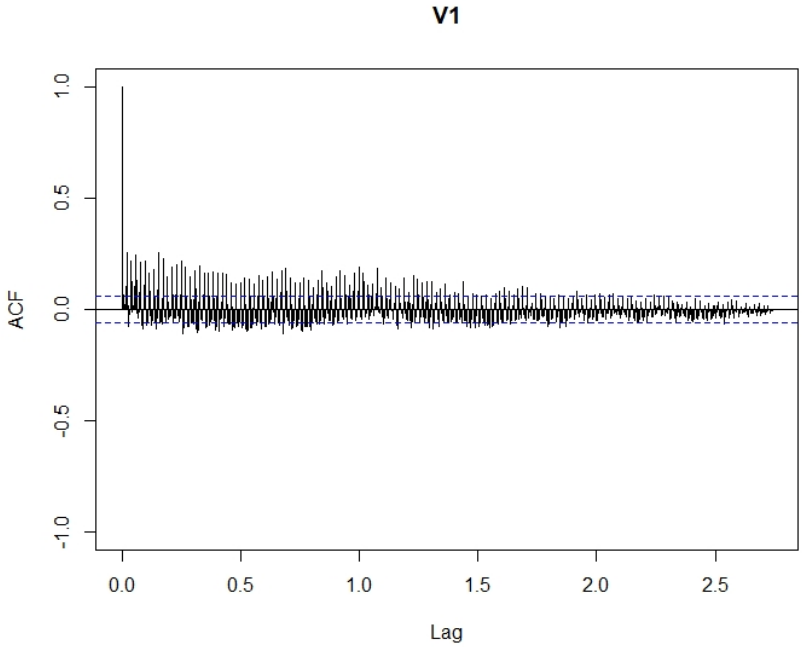

Figura 5: ACF de la sèrie temporal.

5.3 Estimació del model

El següent pas en la nostra anàlisi és estimar el model $ARFIMA(p, d, q)$ que millor s'adapti a la sèrie; és a dir, hem de trobar els valors de p, d i q que facin que el model s'ajusti millor a les dades. Per a fer-ho, utilitzarem la funció `arfima.whittle` de la llibreria *afmtools*, la qual estima els paràmetres del model aplicant el mètode de Whittle explicat a les pàgines 29 i 30.

Com a exemple, la següent taula mostra l'estimació dels paràmetres per a diferents models $ARFIMA$ ajustats a la nostra sèrie.

Model	Paràmetre	Estimació
$ARFIMA(0, d, 0)$	d	0.2047
$ARFIMA(1, d, 0)$	d	0.1153
	ϕ	0.1495
$ARFIMA(0, d, 1)$	d	01116
	θ	0.1699
$ARFIMA(1, d, 1)$	d	0.1330
	ϕ	-0.2543
	θ	0.3982

Taula 3: Paràmetres estimats per a diferents models ARFIMA.

Seguidament, hem de decidir amb quin model ens quedem per tal de seguir l'anàlisi amb el model escollit. Usant els estimadors de Whittle, aplicarem el criteri d'informació d'Akaike per a determinar el millor model. Com més baix sigui el valor de l'AIC, millor serà el model estimat. Hem estimat tots els models $ARFIMA$ possibles on $p, q \in \{0, 1, 2\}$.

Com veiem a la Taula 4, el model amb un menor AIC és el model simple $ARFIMA(0, d, 0)$, que és el que hauríem d'escollir. Però per tal que la posterior anàlisi sigui més interessant, hem decidit triar el model $ARFIMA(1, d, 1)$, on $d = 0.1330$, per a que tinguem una part autorregressiva i una part mitjana mòbil. Aquest model presenta el quart AIC més baix. També hem calculat les log-versemblances dels models mitjançant la funció `arfima.whittle.loglik` i el model escollit presenta una de les més altes (17.84). Recordem que en el cas de la log-versemblança, un major valor d'aquesta implica un millor ajustament del model.

p	q	AIC	$\hat{d}$	p-valor
0	0	-33.68	0.2047	0
0	1	-31.68	0.1116	0.0246
0	2	-29.68	0.1716	0.0005
1	0	-31.68	0.1153	0.0201
1	**1**	**-29.68**	**0.1330**	**0.0073**
1	2	-27.68	0.3993	0
2	0	-29.68	0.1649	0.0009
2	1	-27.68	0.3685	0
2	2	-25.68	0.1985	0

Taula 4: AIC de diferents models i estimació del paràmetre d amb el seu p-valor.

Un cop triat el model, podem obtenir un petit resum d'aquest amb la funció `summary`. Hi trobem l'estimació dels paràmetres d, ϕ i θ amb els seus respectius errors estàndard, el valor de la prova t de Student i el p-valor. En tots els paràmetres obtenim un valor de la t superior a 2 en valor absolut (valor llindar que hem de superar) i amb un p-valor inferior a 0.05. En aquests casos podem rebutjar la hipòtesi nul·la i concloure que tots els paràmetres són estadísticament diferents de 0. Per últim, el valor *sd.innov* és la ràtio entre l'espectre teòric i el periodograma.

```
$call
arfima.whittle(series = serie, nar = 1, nma = 1, fixed = NA)

$coefmat
          Estimate  Std. Error   t value     Pr(>|t|)
d        0.1330399  0.04962895  2.680692 7.347009e-03
phi 1   -0.2543382  0.03230849 -7.872179 3.552714e-15
theta 1  0.3982253  0.03591352 11.088452 0.000000e+00

$sd.innov
[1] 25.98771
```

Figura 6: Resum del model $ARFIMA(1, 0.1330, 1)$.

Amb aquests paràmetres estimats, si ens remuntem a la definició d'un model $ARFIMA$ (Definició 4.4) i si definim la nostra sèrie com un procés $\{X_t\}$, podem escriure el nostre model a través de l'equació

$$(1 - 0.2543B)(1 - B)^{0.1330}X = (1 + 0.3982B)Z.$$

5.4 Validació del model

Ara que ja tenim el model estimat passarem a l'etapa de validació. En primer lloc comprovarem que es compleixen les condicions d'estacionarietat, causalitat i invertibilitat aplicant el Teorema 4.11 i el Teorema 4.13. Per altra banda, analitzarem també el comportament dels residus del model.

Una manera de veure l'estacionarietat del procés és fixar-se en el correlograma de l'ACF de la sèrie de la Figura 5. Com hem comentat anteriorment, els valors de l'ACF decreixen exponencialment fins a 0, fet que ens indica estacionarietat. Ara bé, podem utilitzar una prova més robusta per detectar tant l'estacionarietat com la causalitat i la invertibilitat.

Com demostràvem als teoremes 4.11 i 4.13, donat el nostre procés $ARFIMA$, si $-1 < d < \frac{1}{2}$ i els polinomis $\Phi_p(\cdot)$ i $\Theta_q(\cdot)$ no tenen arrels en comú, aleshores:

(a) Si les arrels de $\Phi_p(\cdot)$ es troben fora del cercle unitat tancat $\{z : |z| \leq 1\}$, llavors existeix una única solució estacionària de (4.4), i aquesta solució és causal.

(b) Si les arrels de $\Theta_q(\cdot)$ es troben fora del cercle unitat tancat $\{z : |z| \leq 1\}$, llavors la solució de (4.4) és invertible.

La primera condició és fàcil de comprovar. La nostre d estimada és $d = 0.1330 \in (-1, \frac{1}{2})$. Per a comprovar que els dos polinomis no tenen arrels en comú i que es troben fora del cercle unitat tancat $\{z : |z| \leq 1\}$, usarem la funció `check.parameters.arfima` de la llibreria *afmtools*. Aquesta funció està incorporada a la comanda `plot` junt amb altres gràfics i el resultat que obtenim es pot trobar al marge superior esquerre de la Figura 7. Com podem veure, les arrels dels polinomis AR i MA no coincideixen i es troben fora del cercle unitat (a la imatge se'ns mostra la inversa de les arrels). Per tant es compleixen les condicions d'estacionarietat, causalitat i invertibilitat.

Com a observació, també podem destacar una manera més senzilla de veure que les arrels de $\Phi_p(\cdot)$ i $\Theta_q(\cdot)$ es troben fora del cercle unitat tancat i que, per tant, la solució és causal i invertible. Com que en el nostre model triat $p = q = 1$, és suficient comprovar que els valors obtinguts de ϕ i θ són menors que 1 en valor absolut. En el nostre cas, efectivament $|\phi| = 0.2543 < 1$ i $|\theta| = 0.3982 < 1$.

Pel que fa a la resta de figures, al marge inferior esquerre hi trobem l'espectre teòric (línea vermella) i l'espectre empíric (els punts). A la columna de la dreta veiem els correlogrames de les autocorrelacions de la sèrie i dels residus. Al no ser un nombre de retards molt elevat, s'hi aprecia l'efecte de la memòria llarga de la sèrie, ja que hi podem veure la persistència de les autocorrelacions.

Per acabar d'analitzar els residus farem ús de la funció `tsdiag`, la qual ens dibuixa els residus estandarditzats, l'ACF dels residus (de nou) i els p-valors de l'estadístic de Ljung-Box en funció del retard. En aquesta darrera interpretació hi pararem atenció. El test de Ljung-Box es basa en l'estadístic

$$Q = n(n+2)\sum_{k=1}^{h} \frac{\hat{\rho}^2(k)}{n-k},$$

on n és la mida de la mostra, $\hat{\rho}(k)$ és l'autocorrelació de la mostra amb retard k i h és el nombre de retards que s'estan provant. Q segueix una distribució khi quadrat amb h graus de llibertat. Aleshores podem definir el test de Ljung-Box de la següent manera:

H_0: Les dades es distribueixen de forma independent.

H_a: Les dades no es distribueixen de forma independent.

Aplicant la prova de Ljung-Box als residus del nostre model, obtenim un valor de l'estadístic $Q = 4.4106$ amb un p-valor de $0.9269 > 0.05$. Per tant, no podem rebutjar la hipòtesi nul·la i podem concloure que les dades es distribueixen de manera independent. Això es pot veure al tercer gràfic de la Figura 8.

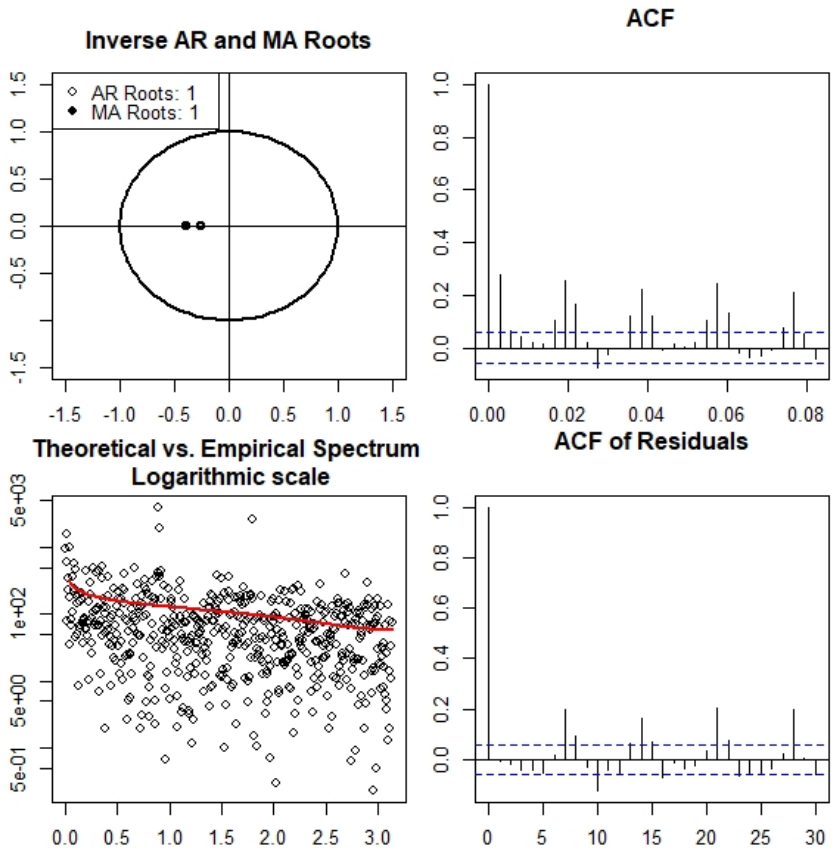

Figura 7: Representacions gràfiques obtingudes amb la comanda `plot`.

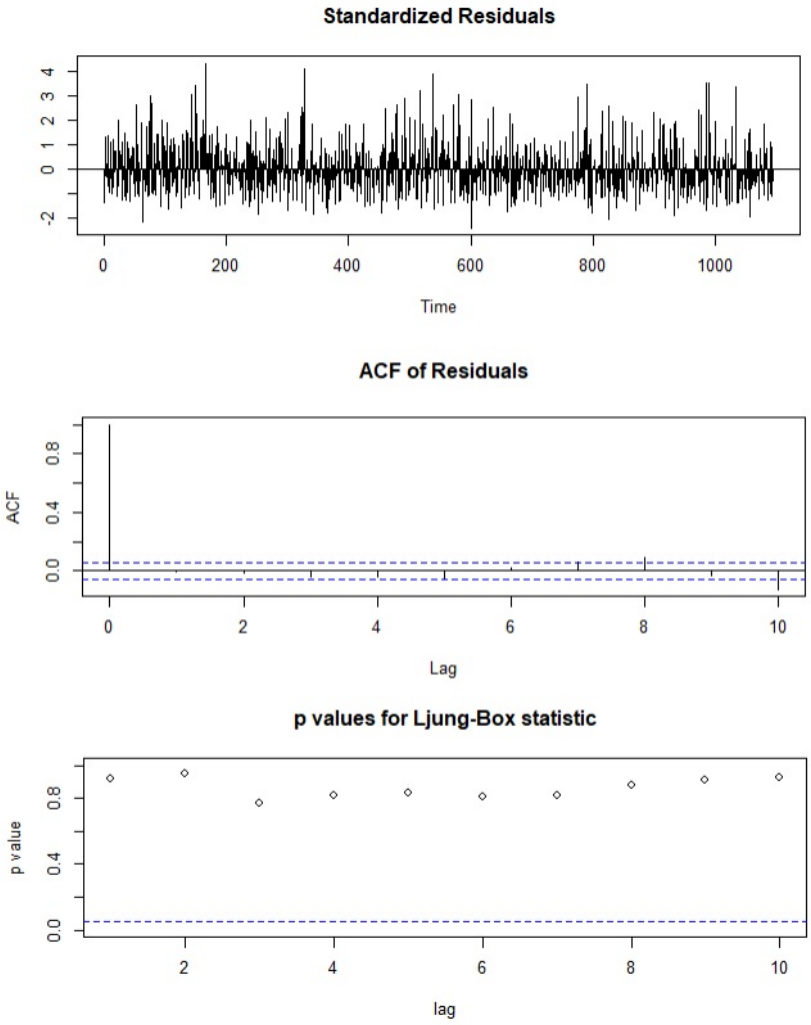

Figura 8: Gràfiques de l'anàlisi de residus del model utilitzant la comanda `tsdiag`.

5.5 Altres característiques i propietats del model

Per acabar amb la nostra anàlisi, en aquest últim apartat veurem i representarem gràficament molts dels aspectes definits al llarg del treball aplicats al model $ARFIMA(1, 0.1330, 1)$ de la nostra sèrie temporal.

Primerament, utilitzem la funció `ir.arfima` per tal de calcular la funció R_j de l'IRF (funció de resposta a l'impuls, veure pàgines 26 i 27). Aquesta funció es pot veure representada a la Figura 9. Podem observar com decreix exponencialment de manera ràpida, concretament amb una taxa de variació de j^{d-1} perquè aquestes funcions tenen el comportament dels η_j. Aquest comportament és comú en els models $ARFIMA$.

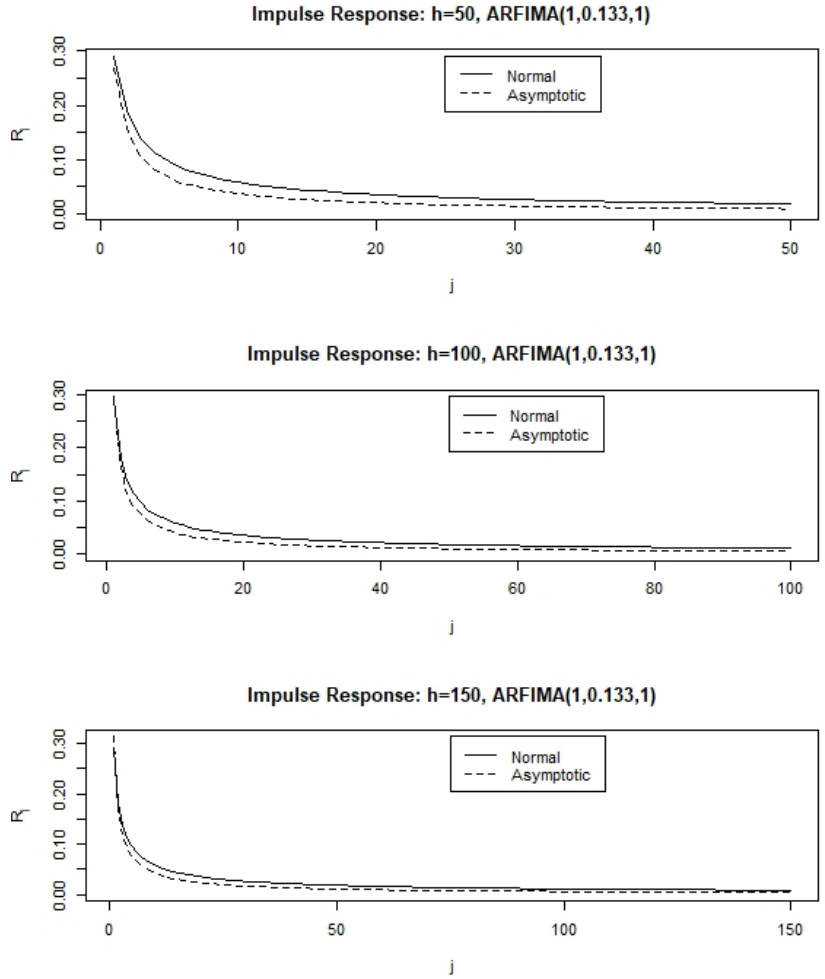

Figura 9: Gràfiques de les funcions $log_{10}(R_j)$ amb retards $h = 50, 100$ i 150.

Hi podem veure representades, doncs, 3 corbes de R_j associades amb el nostre model en funció del retard h. S'ha utilitzat el mètode asimptòtic de la fórmula (4.13), la línea de punts a les gràfiques, i el mètode de la fórmula (4.12), la línea contínua. Podem veure com per valors grans de $j \approx 50$ tots dos mètodes convergeixen. A més, l'aproximació asimptòtica tendeix a ser més similar al mètode *Normal* a mesura que el retard h és més gran.

En segon lloc, podem veure la funció de la densitat espectral del nostre procés $ARFIMA$ a la Figura 10. Si $d < 0$, la densitat espectral s'anul·la a l'origen i hi predominen les freqüències altes. En aquests casos diem que el procés $ARFIMA$ presenta una dependència negativa o antipersistència. Però per valors de $d \in (0, \frac{1}{2})$, com és el nostre cas, la densitat espectral és una funció decreixent no fitada a l'origen i hi predominen les freqüències baixes. Això mostra la relació directa que hi ha entre la persistència de les autocorrelacions i la dinàmica de l'espectre en freqüències baixes.

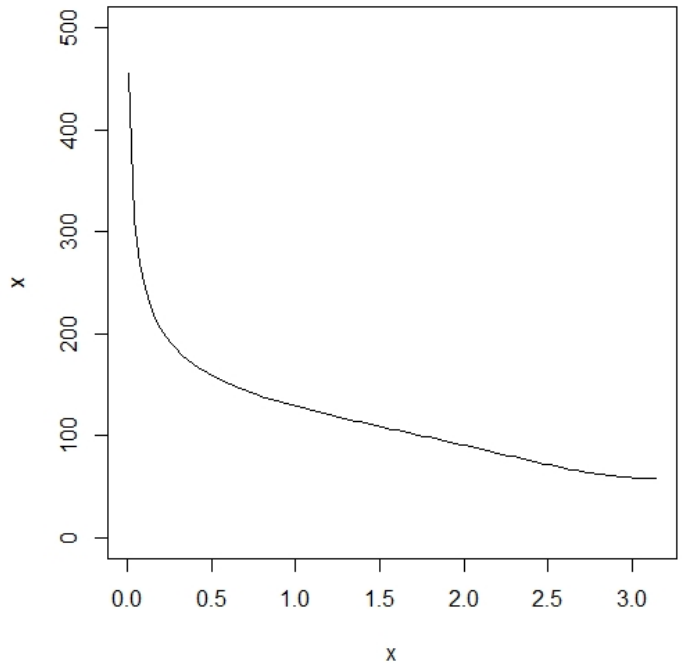

Figura 10: Funció densitat espectral del nostre model.

També podem dibuixar el periodograma (veure Figura 11) tal i com està definit a la Definició 4.15, mitjançant la funció `per.arfima`. Per altra banda, podem calcular exactament la funció d'autocovariància del nostre model $ARFIMA(1, 0.1330, 1)$ utilitzant la funció `rho.sowell`, que es basa en l'aplicació de l'algoritme de Sowell explicat a la secció 4.6 del treball. A la Figura 12 hi trobem comparats el correlograma de l'ACF usant Sowell (esquerra) i el correlograma de l'ACF de la nostra sèrie (dreta). Podem apreciar, tot i no ser exactament iguals, la similitud entre l'ACF del model i l'ACF de la sèrie.

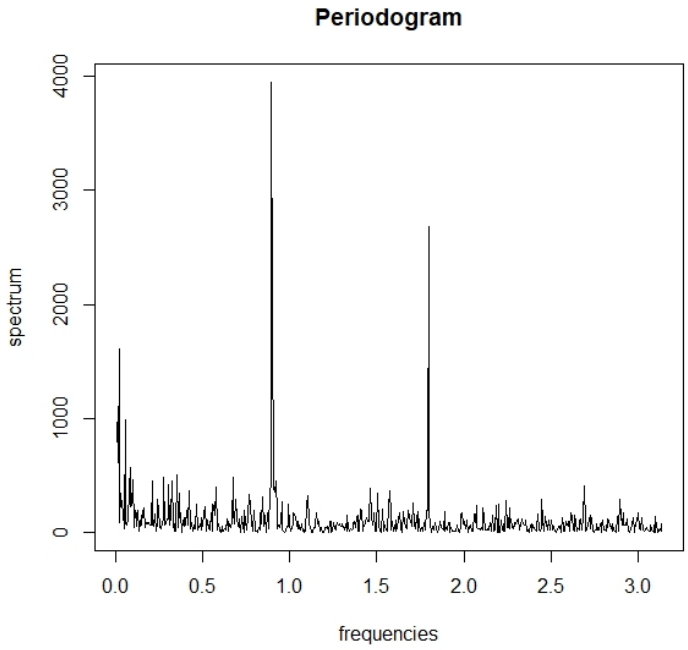

Figura 11: Periodograma de la sèrie.

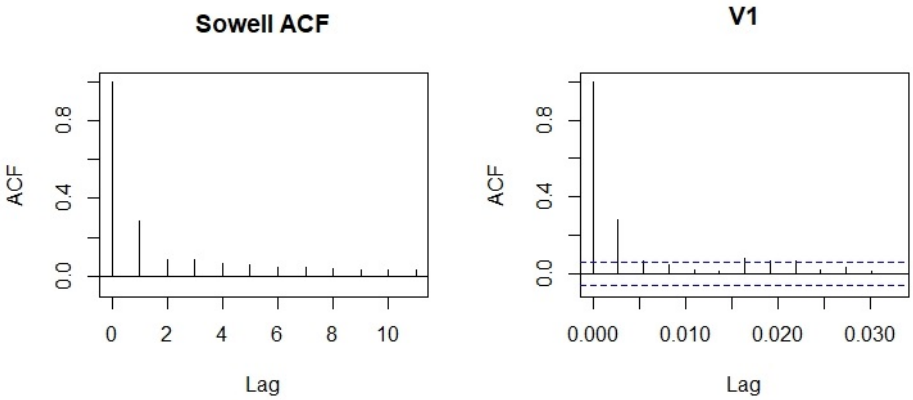

Figura 12: ACF del model usant l'algoritme de Sowell (esquerra) i ACF de la sèrie amb nombre de retard $h = 12$.

Per últim, podem calcular la variància de la mitjana mostral (veure també la secció 4.6), de manera exacta i de manera asimptòtica, fent servir la funció `smv.afm`. Els valors obtinguts, respectivament, han estat de 0.2384 i 1.3192, valors relativament similars.

6 Conclusions

Amb el treball finalitzat, aquesta secció final té com a objectiu recollir totes aquelles conclusions que se n'extreuen del mateix, tant a nivell acadèmic com a nivell personal, i veure si s'han assolit els objectius marcats a l'inici.

En primer lloc, pel que fa a la part més teòrica del treball considero que ha estat un èxit, ja que s'han desenvolupat tots els objectius que es van marcar en el plantejament del projecte. Hi trobem un preàmbul que recull tots aquells conceptes bàsics que són de vital importància per poder seguir desenvolupant la teoria. Seguidament, hem vist la història i la necessitat del naixement dels models $ARFIMA$ i com aquests models prenen cada cop més importància en l'anàlisi de les sèries temporals. Això és degut a que proporcionen una alternativa més flexible a l'hora d'investigar una sèrie temporal que la que donen els mètodes tradicionals com els models $ARMA$ o $ARIMA$. En concret, hem vist com els models $ARFIMA$ s'adapten millor als efectes de les "innovacions" de les observacions en una sèrie temporal i fan que aquestes tinguin efectes transitoris que perduren durant un cert temps, però que acaben desapareixent. A nivell conceptual, això és la propietat coneguda com a memòria llarga, propietat que hem pogut definir de manera precisa i matemàtica.

Per altra banda, hem pogut enunciar i desenvolupar gran part de la teoria matemàtica que hi ha darrere dels models $ARFIMA$. Hem demostrat que, si es compleixen unes certes condicions inicials, els nostres models presenten les propietats d'estacionarietat, causalitat i invertibilitat. A més, hem mostrat un algoritme que permet calcular la funció d'autocovariància de manera exacta (l'algoritme de Sowell) i hem especificat quines són les equacions de la funció de densitat espectral i la funció de resposta a l'impuls d'un procés $ARFIMA$. Per acabar amb la part teòrica, s'han exposat diversos mètodes per estimar els paràmetres dels models, per seleccionar-ne el millor i per a fer futures prediccions. Amb els mètodes d'estimació de paràmetres és on hem pogut enllaçar la part teòrica amb la part pràctica del treball, ja que un dels mètodes (l'estimador de Whittle) és el que hem fet servir per a ajustar un model $ARFIMA$ a la nostra sèrie temporal triada.

Gràcies a la part pràctica del treball hem pogut veure l'aplicabilitat dels models $ARFIMA$ a un cas d'una sèrie temporal real: la qualitat de l'aire de la ciutat de Barcelona entre el 2017 i el 2019. S'ha vist que aquesta és una sèrie adient per ser modelitzada amb un model $ARFIMA$, ja que presenta uns pics elevats de tant en tant i aquests han de ser tractats amb cura per a que els seus efectes ni desapareguin ràpidament ni tinguin efectes permanents. S'ha ajustat, doncs, un model $ARFIMA(1, 0.1330, 1)$ a la sèrie i s'han aplicat la majoria d'aspectes estudiats al llarg del treball mitjançant R.

Degut al límit d'extensió del treball, hi ha alguns aspectes que no han aparegut que serien interessants d'estudiar en una ampliació futura del tema. Per la vessant teòrica del projecte, seria d'interès conèixer altres mètodes d'estimació dels paràmetres i de predicció més desenvolupats. A més, també existeixen altres models derivats dels models $ARFIMA$, com els models $ARFIMA$ estacionals ($SARFIMA$) o els $ARFIMA - GARCH$. Es podrien buscar les diferències d'aquests models amb l'estudiat, buscar-ne la seva aplicabilitat i trobar alguna sèrie temporal que s'hi adapti millor. En quant a la part pràctica del treball, el punt que no s'ha pogut realitzar és el de la predicció de futurs valors de la sèrie. Malauradament, la llibreria que més s'ha utilitzat (*afmtools*) i amb la qual s'estimava i es creava el model no disposa d'una eina per a predir valors futurs com a tal, sinó que tan sols té un test que compara la capacitat predictiva entre dos models $ARFIMA$. Com que

a la nostra anàlisi ens hem centrat amb un sol model, no s'ha considerat necessari utilitzar aquest test. De cara a futures ampliacions, s'intentaria buscar una manera d'obtenir futures prediccions utilitzant un model $ARFIMA$ amb R, o es triarien múltiples models per tal d'avaluar la capacitat predictiva.

Per últim, a nivell personal l'elaboració d'aquest treball m'ha aportat molts aspectes positius. Com a faceta més important, he pogut expandir el meu coneixement sobre les sèries temporals i la seva modelització. L'estudi de les sèries temporals ha estat de les matèries que més m'han interessat al llarg del grau i poder fer un treball que m'ha permès extendre el que havia après ha sigut molt gratificant. Per altra banda, he hagut d'aplicar alguns dels conceptes apresos al llarg del grau i he millorat la meva capacitat de redactar de manera formal i matemàtica, utilitzant les tècniques de demostració vistes aquests anys. A més, també he pogut fer ús del llenguatge R, l'ús del qual està molt estès en el camp de l'estadística, he après a cercar bibliografia de caire científic i a buscar bases de dades adients al tema en qüestió.

A Codi *R* utilitzat a la secció 5

```
##############################################################
#####                                                    #####
#####     CAS PRACTIC: QUALITAT DE L'AIRE DE BARCELONA    #####
#####                                                    #####
##############################################################

## Carreguem llibreries i serie

library(fracdiff)
library(longmemo)
library(hypergeo)
library(sandwich)
library(polynom)
library(afmtools)

serie <- ts(read.table("no2.txt"),start=2017,frequency=365)

## Coneixer la base de dades

View(serie) # Veure la base de dades
dim(serie) # Dimensio
summary(serie) # descriptiva univariant

# veiem quantes observacions hi ha a cada interval de l'icqa

icqa_bona <- sum(serie >=0 & serie <= 40)
icqa_bona
icqa_raonablbona <- sum(serie >=41 & serie <= 90)
icqa_raonablbona
icqa_regular <- sum(serie >=91 & serie <= 120)
icqa_regular
icqa_desfavorable <- sum(serie >=121 & serie <= 230)
icqa_desfavorable
icqa_moltdesfavorable <- sum(serie >=231 & serie <= 340)
icqa_moltdesfavorable
icqa_extrdesfavorable <- sum(serie >=341)
icqa_extrdesfavorable

# mirem si hi ha missings

apply(apply(serie,2,is.na),2,sum)

barplot(table(serie),main = "Barplot del nivell ICQA de NO2")

# histograma

hist(serie)
```

```
# grafics

table(serie)
pie(table(serie), col = c("turquoise", "white"),
main="Pie del nivell ICQA de NO2")

# grafica de la serie

plot.ts(serie,xlab="Any",ylab="Nivell de ICQA (concentracio en
microg/m^3 de NO2)",
col=1,axes = FALSE)
axis(side = 2, at = c(0,10, 40, 70, 100, 130, 160, 190),
labels = c(0,10, 40, 70, 100, 130, 160, 190))
axis(side = 1, at = seq(2017, 2020, by = 1),
labels = seq(2017, 2020, by = 1))

# autocorrelacio (acf) i autocorrelacio parcial (pacf)

acf(serie,ylim=c(-1,1),lag=50) #lag 50
pacf(serie,ylim=c(-1,1),lag=50) #lag 50
acf(serie,ylim=c(-1,1),lag=150) #lag 150
pacf(serie,ylim=c(-1,1),lag=150) #lag 150
acf(serie,ylim=c(-1,1),lag=500) #lag 500
pacf(serie,ylim=c(-1,1),lag=500) #lag 500
acf(serie,ylim=c(-1,1),lag=1000) #lag 1000
pacf(serie,ylim=c(-1,1),lag=1000) #lag 1000

### veiem com l'acf decreix exponencialment fins a convergir a 0

## Estimacio model ARFIMA amb Whittle

# estimacio parametres de models amb p,q = {0,1}

mod = arfima.whittle(serie, nar = 1, nma = 1, fixed = NA)
# anar modificant
summary(mod)

# provem tots els models amb p,q = {0,1,2} i calculem AIC
i log-versemblanca

i = 1 # modificar 0,1,2
j = 1 # modificar 0,1,2
mod = arfima.whittle(serie, nar = i, nma = j)
theta = c(mod$d, mod$ar, mod$ma)
mll = arfima.whittle.loglik(theta, series = serie, nar = i, nma = j)
AIC = - 2 * ( mll$L - (1+i+j) )
coef = summary.arfima(mod)$coefmat[1, 4]
res = c(i, j, AIC, mod$d, coef)
```

```
res

# el model amb menor AIC es el p=q=0

# pero triarem el p=q=1 per fer analisi amb model amb part AR i MA

# Calculem la log-versemblanca

mod = arfima.whittle(serie, nar = 1, nma = 1, fixed = NA)
summary(mod)
mod$ma
param=c(0.1330399, -0.2543382, 0.3982253)
loglik=arfima.whittle.loglik(theta=param, serie, nar=1, nma=1)
loglik$L
loglik$sigma

# d = 0.1330399, phi = -0.2543382, theta = 0.3982253

### Validacio del model

plot(mod)
res=residuals(mod)
acf(res)
Box.test(res, 10, type="Ljung")
tsdiag(mod)

par(mfrow=c(1,1))
check.parameters.arfima(d=mod$d,ar=mod$ar,ma=mod$ma,plot=TRUE)

### Altres aspectes vistos durant el treball

# Funcio de resposta a l'impuls

ir.arfima(50, mod, plot.it = TRUE)
ir.arfima(100, mod, plot.it = TRUE)
ir.arfima(150, mod, plot.it = TRUE)

# densitat espectral

par(mfrow=c(1,1))
x=spectrum.arfima(d=0.1330399, ar=-0.2543382,
ma=0.3982253, sd.innov = 25.98771)
plot(x, xlim=c(0.01, pi), ylim=c(0,500))

# Periodograma

per=per.arfima(serie)
plot(per$freq, per$spec, type="l",
xlab="frequencies",ylab="spectrum", main="Periodogram")
```

```
cbind(per$spec,per$freq)

# Calcul de l'autocovariancia exacta amb l'algoritme de Sowell

par(mfrow=c(1,2))
rho.sowell(mod, 12, mod$ar, mod$ma, mod$d , mod$sd.innov)
acf(serie,lag=12)

# Calcul de la variancia de la mitjana mostral

smv.afm(object=mod,comp=TRUE)
smv.afm(object=mod,comp=FALSE)
```

Referències

[1] Ajuntament de Barcelona, *Qualitat de l'aire*. [Online]. Available: https://ajuntament.barcelona.cat/qualitataire/ca/.

[2] Akaike, H. (1973). *Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models*. Biometrika, Vol. 60, No. 2, pp. 255-265.

[3] Beran, J. (1992). *Statistical Methods for Data with Long-Range Dependence*. Statistical Science, 7, pp. 404-416.

[4] Bondon, P.; Palma, W. (2007). *A Class of Antipersistent Processes*. Journal of time series analysis, Vol. 28, No. 2, pp. 261-273.

[5] Box, George E.P.; Jenkins, Gwylim M.; Reinsel Gregory C. (2008). *Time Series Analysis: Forecasting and Control*, 4a ed. Wiley.

[6] Brockwell, P.J.; Davis, R.A. (1991). *Time Series: Theory and Methods*, 2a ed. Springer.

[7] Brockwell, P.J.; Davis, R.A. (1996). *Introduction to Time Series and Forecasting*, 2a ed. Springer.

[8] Castaño, E. (2016). *Identificación de modelos ARFIMA*. Revista de la Facultad de Ciencias Universidad Nacional de Colombia, Vol. 5, No. 1, pp. 12-37.

[9] Contreras-Reyes, J.E.; Goerg, G.M.; Palma, W. (2013). *Estimation, Diagnostic and Forecasting Functions for ARFIMA models. Package 'aftmtools'*. [Online]. Available: https://www.researchgate.net/publication/300842241_Estimation_Diagnostic_and_Forecasting_Functions_for_ARFIMA_models.

[10] Cowpertwait, P.S.P; Metcalfe, A.V. (2009). *Introductory Time Series with R*, 1a ed. Springer.

[11] Gencat: Medi Ambient i Sostenibilitat, *Descàrrega de dades automàtiques*. [Online]. Available: https://mediambient.gencat.cat/ca/05_ambits_dactuacio/atmosfera/qualitat_de_laire/vols-saber-que-respires/descarrega-de-dades/descarrega-dades-automatiques/.

[12] Geweke, J.; Porter-Hudak, S. (1983). *The Estimation and Application of Long Memory Time Series Models*. Journal of Time Series Analysis, Vol. 4, No. 4, pp. 221-238.

[13] Granger, C.W.J.; R. Joyeux. (1980). *An introduction to long-memory time series models and fractional differencing*. Journal of Time Series Analysis, Vol. 1, No. 1, pp. 15-29.

[14] Hosking, J. R. M. (1981): *Fractional differencing*. Biometrika, Vol. 68, No. 1, pp. 165-176.

[15] Hurst, H.E. (1951). *Methods for storage capacity of reservoirs*. Trans. Am. Soc. Civil Engineers, 116, pp. 770-799.

[16] Inoue, A. (2002). *Asymptotic behavior for partial autocorrelation functions of fractional ARIMA processes*. The annals of Applied Probability, Vol. 12, No. 4, pp. 1471-1491.

[17] Palma, W. (2007). *Long-memory time series: Theory and Methods.* Wiley.

[18] Pérez Espartero, A.; Ruiz Ortega, E. (2002). *Modelos de memoria larga para series económicas y financieras.* Investigaciones económicas, Vol. 26, No.3, pp. 395-445.

[19] Pons i Fanals, E.: Apunts de l'assignatura Anàlisi de sèries temporals, Grau d'Estadística, Universitat de Barcelona.

[20] Velasco, C. (1999). *Non-stationarity log-periodogram regression.* Journal of Econometrics, Vol. 91, pp. 325-371.

[21] Shumway, Robert H.; Stoffer, David S. (2017). *Time Series and Its Applications with R examples.* Springer.

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