VNIVERSITAT DE VALÈNCIA

Treball Fi de Grau - Febrer de 2025

Anàlisi predictiva: explorant algorismes i models per a l'estudi de sèries temporals

Autor/a: **Marc González Pastor**

Tutor/a: Ana Navarro Quiles

VNIVERSITAT DE VALÈNCIA

Facultat de Ciències Matemàtiques

Grau en Matemàtiques

Índex

Capítol 1

Introducció

Al llarg de la història, l'ésser humà ha buscat identificar patrons en la natura per entendre i anticipar multitud de fenòmens. Des de l'astronomia babilònica, on s'analitzaven els moviments dels astres per anticipar esdeveniments celestes, fins a les modernes tècniques de modelització de dades, l'ésser humà ha intentat descobrir patrons en la realitat que l'envolta. En aquest sentit, l'anàlisi de dades i la seua modelització matemàtica han estat eines essencials per entendre el món i prendre decisions estratègiques.

En les darreres dècades, l'explosió en la quantitat de dades disponibles i l'avenç tecnològic han transformat completament la manera en què s'aborda l'estudi de sistemes dinàmics. Els problemes que abans requerien mesos de càlcul manual ara es poden resoldre en qüestió de segons gràcies als ordinadors i als algorismes d'aprenentatge automàtic. L'obra *Big Data: A Revolution That Will Transform How We Live, Work, and Think* de Viktor Mayer-Schönberger i Kenneth Cukier [23] posa de manifest com aquesta revolució de les dades està canviant la manera de prendre decisions en àmbits com la salut, la política o l'economia.

No obstant això, l'increment de la disponibilitat de dades també planteja nous reptes. Com es poden utilitzar de manera eficient per fer prediccions precises? Com podem distingir patrons reals de simples coincidències? En aquest context, la modelització matemàtica té un paper primordial, especialment en aquells casos en què les dades evolucionen al llarg del temps. La capacitat de predir la demanda d'un producte, l'evolució de les temperatures o les fluctuacions dels mercats financers depén de la correcta comprensió de com les observacions passades influeixen en les futures.

Aquest és precisament l'objectiu de l'anàlisi de sèries temporals: identificar i modelitzar estructures en dades observades al llarg del temps. Aquest camp ha evolucionat gràcies a models matemàtics com els models autoregressius integrats de mitjana mòbil (ARIMA) i els mètodes de suavitzat exponencial, presentats en obres clàssiques, com [12], i altres més recents com [8].

Darrerament, els progressos en aprenentatge automàtic han revolucionat la manera en què es representen les dades temporals. Models com les xarxes neuronals recurrents (RNN) i les unitats de memòria a llarg termini (LSTM), descrites en [11], han permés millores significatives en aplicacions com el reconeixement de veu, la traducció automàtica i la detecció d'anomalies en sistemes industrials. No obstant això, aquests models solen requerir grans quantitats de dades i una capacitat computacional elevada, la qual cosa els fa menys accessibles en determinats contextos.

Aquest treball se centra en els models clàssics per a l'anàlisi de sèries temporals, amb la missió d'explorar

com es poden aplicar a problemes reals de predicció. A través d'aquesta exploració, es pretén no només comprendre les bases matemàtiques d'aquests models, sinó també reflexionar sobre els seus avantatges, limitacions i possibles extensions.

1.1 Context i objectius de l'anàlisi de sèries temporals

L'anàlisi de sèries temporals és una disciplina estadística i matemàtica que s'ocupa de l'estudi de dades ordenades en el temps. A diferència d'altres formes d'anàlisi de dades, les sèries temporals presenten una estructura en la qual les observacions successives no són independents, sinó que existeixen relacions entre elles que poden revelar patrons amagats. Aquestes dependències fan necessari el desenvolupament d'eines específiques per modelitzar i predir el seu comportament, la qual cosa ha portat al desenvolupament de diversos models i metodologies al llarg de la història.
L'anàlisi de sèries temporals té una aplicabilitat molt àmplia. En economia, per exemple, s'utilitza per predir el creixement del PIB, els tipus d'interés o la inflació. En meteorologia, permet anticipar l'evolució de temperatures i precipitacions. També és essencial en la detecció de fallades en sistemes industrials, la predicció del consum elèctric i el monitoratge de senyals biomèdiques com la freqüència cardíaca o l'activitat neuronal. En cadascun d'aquests casos, la correcta identificació de patrons recurrents pot tindre un impacte significatiu en la presa de decisions.
Els punts centrals de l'anàlisi de sèries temporals es poden classificar en tres grans blocs:

1. **Descripció i exploració de dades.** L'objectiu inicial de qualsevol anàlisi és comprendre les característiques de la sèrie temporal. Això inclou la identificació de components com la tendència (un comportament general de creixement o decreixement), l'estacionalitat (variacions periòdiques que es repeteixen en intervals regulars) i la presència de cicles o patrons a llarg termini. Aquesta fase també requereix l'aplicació d'eines gràfiques i estadístiques per visualitzar la informació de la sèrie i entendre'n la naturalesa.

2. **Modelització i ajustament de models.** Una vegada explorades les dades, el pas següent és la selecció i ajustament d'un model adequat per representar la dinàmica de la sèrie. Els models clàssics, com els processos autoregressius (AR), els de mitjana mòbil (MA) i els seus híbrids ARMA i ARIMA, són àmpliament utilitzats per capturar les estructures subjacents de la sèrie. Aquests models són especialment útils quan la relació entre les observacions successives pot ser descrita per dependències lineals i quan el volum de dades no és excessivament gran. No obstant això, en entorns de dades més complexes, els enfocaments basats en xarxes neuronals han guanyat popularitat per la seua capacitat d'aprendre patrons no lineals i capturar dinàmiques més sofisticades.

3. **Predicció i inferència.** Aquesta capacitat és clau en contextos com la previsió financera, la planificació de la demanda en empreses i la detecció de canvis en sistemes dinàmics. Més enllà de la predicció, l'anàlisi de sèries temporals també permet realitzar inferències sobre la dinàmica del fenomen estudiat, com identificar la influència de factors externs o determinar si una sèrie segueix un patró estacionari o no.

1.2 Estructura del treball

L'organització del present document respon a un enfocament sistemàtic per cobrir els principals aspectes de l'anàlisi de sèries temporals, des de la descripció de les dades fins a la predicció i avaluació dels models. El treball comença al Capítol 2 amb una introducció teòrica al marc conceptual de les sèries temporals, on es presenten les definicions bàsiques i les principals característiques d'aquest tipus de dades. L'ordre en què està estructurada la informació no és casual; s'ha dissenyat per facilitar la comprensió dels conceptes més avançats que es presentaran als capítols posteriors, seguint una construcció lògica i coherent. Seguint aquesta introducció, al Capítol 3 s'exposen alguns dels models clàssics de predicció més utilitzats: els models de suavització exponencial i els models ARIMA, tot i que abans s'estudien els conceptes més rellevants sobre els quals es construeixen aquests models. A continuació, al Capítol 4, es descriuen les tècniques d'avaluació de models i els criteris per seleccionar el model més adequat per a una sèrie temporal donada, garantint així que les prediccions siguen fiables i útils.

Una vegada assentats els fonaments teòrics, la fase següent consisteix en la transició a la pràctica, aprofitant els coneixements adquirits. Al Capítol 5 s'aplicaran els conceptes estudiats a un cas real de predicció del trànsit aeri, materialitzant així els models de suavització exponencial i ARIMA per estimar l'evolució del nombre de passatgers a l'aeroport de València. Donarem una explicació més detallada de les dades a la Secció 5.1.

Per complementar els models clàssics i aprofundir una mica més en la matèria, s'ha decidit incloure un capítol dedicat a models alternatius, el Capítol 6, on s'il·lustren algunes de les tècniques més avançades. El camp de l'anàlisi de dades està en constant evolució, i és fonamental conéixer les últimes tendències i les seues aplicacions per mantindre's al dia en aquest àmbit. Entre aquestes tècniques cal ressaltar l'ús de xarxes neuronals i altres models basats en aprenentatge automàtic, que han demostrat ser molt eficients. A més, es presenta breument el model Prophet, un model desenvolupat per Facebook que ofereix una alternativa interessant als algorismes clàssics.

Finalment, al Capítol 7 es resumeixen les conclusions del treball, destacant els resultats obtinguts i comparant els diferents models utilitzats per acabar amb indicacions sobre els seus casos d'ús i les seues limitacions. També s'identifiquen possibles línies de millora i s'apunten les perspectives de futur per a aquest camp, amb l'objectiu de donar continuïtat a l'estudi i explorar nous enfocaments i tècniques.

Per completar la tasca, s'han inclòs dos apèndixs amb informació addicional. Al primer (Apèndix A), es mostra un extracte del codi utilitzat per dur a terme les anàlisis i les prediccions. El programari s'ha implementat en Python, un dels llenguatges més utilitzats en ciència de dades, i s'ha estructurat de manera clara per facilitar-ne la comprensió. La raó per la qual s'ha optat per Python en comptes de R, un altre llenguatge molt utilitzat en estadística, és la seua flexibilitat i la seua capacitat per integrar diferents llibreries i paquets. Aquest estudi s'ha concebut amb la idea de ser ampliat i reutilitzat en futurs projectes, i Python, per la seua versatilitat, és una elecció idònia per a aquest propòsit. Tot i que alguns programes i funcions puguen semblar complexos, s'han afegit comentaris per explicar el seu funcionament i la seua finalitat. L'estructuració del codi es basa en tres principis: la capacitat d'abstracció, entesa com la possibilitat de processar dades independentment del seu format i procedència (sempre que seguisquen uns estàndards mínims); la modularitat, permetent dividir el programa en fragments més menuts i gestionables, amb la possibilitat d'afegir o eliminar funcions sense afectar la resta del codi; i la reutilització, facilitant la integració de funcions en projectes i programes diferents. Al segon apèndix (Apèndix B), s'hi inclouen les dades amb què s'han fet les prediccions del Capítol 5.

1.3 Eines informàtiques i paquets utilitzats

Durant el desenvolupament d'aquest estudi, s'han utilitzat diverses eines informàtiques per a la manipulació de dades, la implementació de models predictius i la validació dels resultats.
Abans d'entrar en matèria, i com veurem a la Secció 5.2, les dades han sigut sotmeses a un procés de neteja i preparació per minimitzar els possibles errors i inconsistències amb Python, i així evitar que l'extensió del codi siga innecessàriament gran.
No és objecte d'estudi en aquest treball endinsar-se en la manera en què s'han recollit les dades, però sí que és convenient mostrar com s'ha procedit a la seua neteja i quina eina s'ha fet servir. La resposta a aquesta última qüestió ve donada per una ferramenta àmpliament utilitzada en l'àmbit de la ciència de dades, de codi obert i amb una gran comunitat de desenvolupadors: KNIME.
Aquest *software* permet dur a terme tasques de neteja, transformació i anàlisi de dades de manera senzilla i intuïtiva, sense necessitat de saber programar. La raó per la qual s'ha triat KNIME no és una altra que la seua facilitat d'ús i la seua capacitat per a treballar amb grans volums de dades de manera eficient, a més de comptar amb una interfície gràfica amigable, un conjunt d'extensions molt complet i la possibilitat d'observar les transformacions aplicades a les dades en cadascun dels passos del procés, tant de manera gràfica com numèrica.
Com ja s'ha esmentat abans, per a la programació i execució dels models, s'ha fet ús del llenguatge Python. Els avantatges d'haver triat aquest llenguatge enfront de R són la seua versatilitat, l'ampli catàleg de llibreries disponibles i l'escalabilitat, tres premises que són claus per a l'èxit de qualsevol projecte de ciència de dades, sobretot tenint en compte que augmentant el grau d'abstracció d'un treball com aquest es poden abastar més reptes i obtenir resultats més satisfactoris aplicant les tècniques presentades a altres tasques. Els scripts han estat organitzats en diferents mòduls per facilitar la reutilització de codi i la modularitat del projecte.
Pel que fa a les dades, s'han fet servir *datasets* en format CSV, un tipus de fitxer senzill i tabular que permet emmagatzemar informació de manera estructurada i llegir-la fàcilment amb Python o altres ferramentes. Per a l'emmagatzematge dels models entrenats, s'ha utilitzat el format PKL (pickle), que fa possible guardar objectes Python i carregar-los en posteriors execucions, evitant així la necessitat de tornar a entrenar els models cada vegada que es vulguen fer prediccions.

1.3.1 Llibreries de Python

Durant l'anàlisi i la implementació dels models, s'han utilitzat diverses llibreries de Python, entre les quals destaquen:

- `pandas`: Manipulació i anàlisi de dades en format tabular.
- `numpy`: Operacions matemàtiques i vectorials.
- `matplotlib`: Creació de gràfiques.
- `statsmodels`: Funcions i tests estadístics i implementació de Holt-Winters.
- `pmdarima`: Implementació d'ARIMA i AUTO ARIMA.
- `scipy`: Funcions i tests estadístics.
- `prophet`: Implementació del model Prophet.

1.3.2 Estructura del repositori

El codi font i els conjunts de dades utilitzats es poden trobar al repositori de l'autor. El repositori s'ha organitzat de manera modular per facilitar tant la gestió del projecte com la reproducció dels resultats.

- **data/**: Conté els conjunts de dades utilitzats en l'anàlisi i modelització.
- **knime/**: Inclou el flux de treball desenvolupat en KNIME.
- **models/**: Implementació dels models de predicció utilitzats.
- **saved_models/**: Conté els models entrenats en format PKL per a reutilització posterior.
- **tex/**: Arxius LaTeX per a la documentació, organitzats en:
  - **capitols/**: Conté els diferents capítols, en format `.tex`.
  - **altres/**: Inclou taules descriptives i mètriques en format CSV.
  - **imatges/**: Gràfiques de l'anàlisi i les prediccions.
  - **taules/**: Taules en format `.tex` incloses al document principal.
- **utils/**: Mòduls de suport amb funcions per a l'anàlisi (`analysis.py`), preprocessament de dades (`preprocessing.py`), visualització (`visualization.py`) i altres utilitats (`utils.py`).
- **main.py**: Script principal que centralitza l'execució de tot el flux de treball, des de la càrrega de dades fins a la generació de prediccions i mètriques. Aquest script està dissenyat per ser executat directament sense necessitat de modificacions. Les configuracions clau es defineixen en dues estructures:
  - **CONFIG**: Defineix els paràmetres generals del projecte, com la ruta del conjunt de dades, el directori per a les imatges generades, la freqüència temporal de les dades i la columna objectiu per a l'anàlisi.
  - **SECCIONS**: Controla quines parts del procés s'executen i quins resultats es mostren per pantalla, permetent activar o desactivar de manera modular l'anàlisi descriptiva, la visualització de gràfiques, la comprovació de l'estacionarietat i l'execució de models predictius.
- **environment.yml**: Fitxer de configuració de l'entorn Conda amb totes les dependències del projecte. L'ús de Conda garanteix que el treball siga fàcilment reproduïble en qualsevol sistema amb la mateixa configuració d'entorn.
- **README.md**: Descripció general del projecte i instruccions d'ús.

Aquesta estructura permet una gestió clara dels diferents components del projecte, garantint la seua escalabilitat i manteniment. La separació entre configuració i codi dins de `main.py` facilita la personalització dels experiments sense modificar la lògica del programa.

Capítol 2

Fonaments de les sèries temporals

L'anàlisi de dades registrades en diferents moments del temps permet estudiar l'evolució de nombroses variables en àmbits molt diversos. Els mecanismes ací presentats proporcionen un enfocament matemàtic per descriure aquests canvis, identificar regularitats en les observacions i construir models que expliquen el seu comportament i permeten fer prediccions.
Aquest capítol introdueix els conceptes bàsics relacionats amb les sèries temporals i els processos estocàstics, establint així les bases per al seu estudi i aplicació. Es definiran els trets característics d'aquestes sèries, es detallaran les seues propietats i es descriuran les transformacions més habituals per facilitar la identificació de patrons i la construcció de models predictius.
Per a una descripció més detallada, es poden consultar les obres de Peña [27] i Chatfield [8], que desenvolupen aquestes tècniques amb exemples i aplicacions reals.

2.1 Introducció als processos estocàstics

Un procés estocàstic és un conjunt de variables aleatòries indexades en el temps o en algun altre paràmetre, la qual cosa permet modelar tant la incertesa com les dependències entre observacions successives. Aquesta estructura constitueix la base teòrica de molts models pràctics utilitzats en l'anàlisi de sèries temporals.

Definició 2.1.1 (Procés estocàstic)**.** *Un procés estocàstic és un conjunt de variables aleatòries $\{z_t\}$, que poden ser unidimensionals o multidimensionals, indexades per un conjunt T, que pot ser finit o infinit. Si és infinit, T pot ser discret o continu. Quan el conjunt T és un subconjunt dels nombres enters positius, el procés s'anomena discret. En canvi, si T és un subconjunt dels nombres reals, s'anomena continu. Si, a més, T és un subconjunt dels nombres enters positius i el procés és unidimensional, es denomina sèrie temporal.*

Propietats dels processos estocàstics

Les propietats d'un procés estocàstic permeten descriure el seu comportament i quantificar la variabilitat de les seues observacions al llarg del temps o de l'espai; i són la clau per comprendre la dependència entre valors successius i construir models que s'ajusten a la dinàmica de les dades.

- **Mitjana temporal (μ_t).** És una mesura central que descriu el valor esperat de la variable aleatòria en un instant t:

$$\mu_t = \mathbb{E}[z_t],$$

on $\mathbb{E}[\cdot]$ denota l'esperança matemàtica.

- **Variància temporal (σ_t^2).** Indica la dispersió dels valors del procés respecte a la mitjana en un instant t:

$$\sigma_t^2 = V(z_t) = \mathbb{E}[(z_t - \mu_t)^2].$$

- **Autocovariància[1] ($\mathbf{Cov}(z_t, z_{t+h})$).** Mesura la dependència entre els valors d'un procés estocàstic en diferents instants de temps separats per un retard h, t i $t+h$:

$$\mathrm{Cov}(z_t, z_{t+h}) = \mathbb{E}[(z_t - \mu_t)(z_{t+h} - \mu_{t+h})]. \tag{2.1}$$

  Propietats importants:

  - $\mathrm{Cov}(z_t, z_{t+h}) = \mathrm{Cov}(z_{t+h}, z_t)$ (simetria).
  - $\mathrm{Cov}(z_t, z_t) = \mathrm{Var}(z_t)$.

- **Autocorrelació[2] ($\rho(h)$).** Quantifica la força de la dependència lineal normalitzada entre valors del procés separats per un retard h. Es defineix com:

$$\rho(h) = \frac{\mathrm{Cov}(z_t, z_{t+h})}{\sigma_t \sigma_{t+h}}. \tag{2.2}$$

  Els valors de $\rho(h)$ estan compresos entre -1 i 1, on:

  - $\rho(h) = 1$ indica una correlació positiva perfecta.
  - $\rho(h) = -1$ indica una correlació negativa perfecta.
  - $\rho(h) = 0$ indica absència de correlació.

  A més, $\rho(0) = 1$.

2.2 Definició de sèrie temporal

Hem vist que els processos estocàstics són seqüències de variables aleatòries indexades pel temps o un altre paràmetre. Dins d'aquesta categoria, les sèries temporals corresponen a processos en temps discret, en què les observacions es registren a intervals regulars, generalment espaiats de manera uniforme. A continuació, presentem una definició intuïtiva i clara de sèrie temporal.

[1] La diferència entre covariància i autocovariància rau en els termes triats. Mentre que la covariància és més oberta i admet variables aleatòries diferents, l'autocovariància requereix que ambdues variables siguen la mateixa, tot i que en diferents moments temporals.

[2] La diferència entre correlació i autocorrelació és anàloga al cas de la covariància i l'autocovariància.

Definició 2.2.1 (Sèrie temporal, en sentit informal)**.** *Una sèrie temporal és el resultat d'observar els valors d'una variable al llarg del temps en intervals regulars (cada dia, cada setmana, cada mes, cada any, etc.).*

Per formalitzar el concepte, donem una definició més precisa i rigorosa.

Definició 2.2.2 (Sèrie temporal, en sentit formal)**.** *Una sèrie temporal és una successió d'observacions ordenades cronològicament:* $\{z_t\}_{t=1}^{n}$*, on* z_t *és el valor observat de la variable d'interés en l'instant* t *i* n *és el nombre total d'observacions. Aquestes observacions se solen recollir en instants de temps equiespaiats.*

Les sèries temporals són una eina imprescindible per modelitzar l'evolució de diferents fenòmens i extraure informació rellevant sobre el seu comportament i són àmpliament utilitzades en camps com l'economia, l'enginyeria, la salut i la ciència, entre d'altres.
En economia i finances, s'empren per a analitzar mercats, predir tendències i avaluar la volatilitat dels actius financers, com es mostra en diversos estudis [9].
En l'àmbit de la salut pública, ajuden a estudiar l'evolució d'epidèmies i altres processos epidemiològics, tal com es detalla en l'anàlisi espectral aplicada a sèries temporals epidemiològiques [6].
A més, l'anàlisi multivariant permet examinar la interacció entre diferents sèries temporals, una tècnica especialment útil per identificar patrons en dades complexes [2]. No obstant això, els models de diverses variables no seran abordats en aquest treball.
Per a dur a terme l'estudi d'una sèrie temporal en un període concret, sovint es treballa amb finestres temporals.

Definició 2.2.3 (Finestra temporal)**.** *Una finestra temporal és un subconjunt d'una sèrie temporal que conté observacions consecutives dins d'un rang de temps específic. Matemàticament, per a una sèrie temporal* $\{x_t\}_{t=1}^{N}$*, una finestra de longitud* w *es defineix com el conjunt* $\{x_t, x_{t+1}, \ldots, x_{t+w-1}\}$*, on* t *és el primer instant de la finestra.*

L'Exemple 2.2.4 inclou la Figura 2.1, que representa la sèrie temporal del nombre d'hipoteques concedides a l'estat espanyol entre 2003 i 2018, amb observacions mensuals. Les observacions s'han recollit gràcies a l'Institut Nacional d'Estadística (INE), i es poden consultar al seu portal web.

Exemple 2.2.4 (Sèrie temporal de les hipoteques concedides a l'estat espanyol entre 2003 i 2018, per mesos)**.** *Aquesta sèrie temporal mostra l'evolució del nombre d'hipoteques concedides a l'estat espanyol en un període de 15 anys, amb dades mensuals. És un exemple típic de dades temporals, amb variacions periòdiques i tendències a llarg termini.*

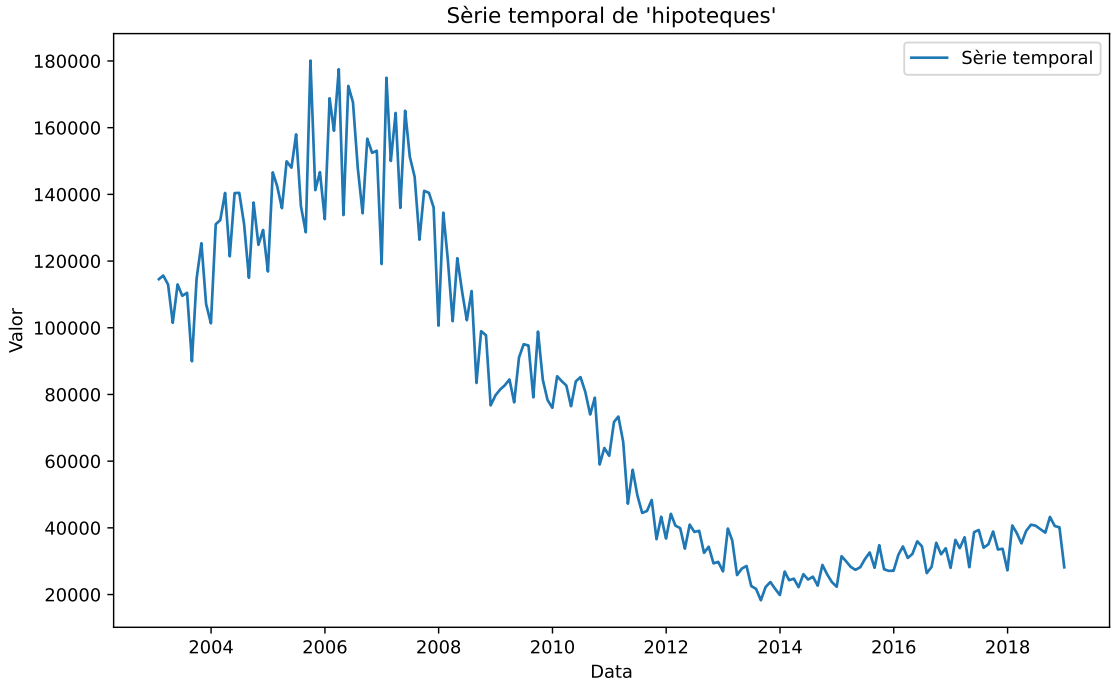

Figura 2.1: Representació de la sèrie temporal de les hipoteques concedides a l'estat espanyol entre 2003 i 2018, per mesos. A l'eix x es mostra l'índex temporal, mentre que a l'eix y es representa el volum de passatgers en cada instant.

Com veiem, les dades mostren una evolució irregular amb fluctuacions periòdiques cada poc temps, i s'hi pot apreciar, a més, una tendència de creixement suau al llarg dels darrers anys, precedida d'un període d'auge i davallada brusca.

2.3 Components de les sèries temporals

Les sèries temporals sovint mostren unes característiques distintives, anomenades components, que són crucials per entendre el seu comportament al llarg del temps. Encara que, de vegades, pot resultar complex distingir i aïllar aquestes components amb exactitud, descriure-les de manera general ajuda a entendre la seua rellevància. Aquesta dificultat deriva de la interacció entre les diferents components —la tendència, l'estacionalitat, la ciclicitat i el soroll blanc—, que poden influir-se mútuament o veure's distorsionades per les particularitats de les dades disponibles.

Tendència

Representa el canvi a llarg termini en el comportament de la sèrie. Consisteix en una evolució contínua en la mitjana, que pot ser ascendent o descendent, sovint associada a factors com el progrés tecnològic, el canvi demogràfic o el creixement econòmic, entre d'altres.

Aquest patró es modelitza habitualment amb funcions lineals, polinòmiques o logarítmiques:

- **Funcions lineals.** Quan la tendència és constant i segueix un patró recte, es pot representar amb una funció lineal:
$$T(t) = a + b \cdot t,$$
on a és l'*intercept* o valor inicial, i b és el pendent, que indica la taxa de canvi. Aquest model és ideal per a tendències simples, com un creixement o decreixement constant al llarg del temps.

- **Funcions polinòmiques.** Quan la tendència no és lineal, es poden utilitzar polinomis de grau superior:
$$T(t) = a_0 + a_1 \cdot t + a_2 \cdot t^2 + \ldots + a_n \cdot t^n.$$
Aquest enfocament és útil quan la tendència mostra acceleració o desacceleració gradual però no exagerada.

- **Funcions logarítmiques.** Les tendències que creixen ràpidament al principi i després es desacceleren (o viceversa) es poden modelitzar amb una funció logarítmica:
$$T(t) = a + b \cdot \ln(t).$$
Aquest tipus de model és freqüent en dades de creixement tecnològic o processos de maduració.

Estacionalitat

Consisteix en moviments repetitius que es produeixen regularment en intervals de temps concrets, com setmanes, mesos o estacions de l'any. Aquesta variació pot atribuir-se a patrons estacionals que influeixen en la sèrie, com ara cicles climàtics o hàbits socials anuals, com les compres de Nadal o les vacances d'estiu.
És important aïllar aquesta component per visualitzar l'evolució a llarg termini, ja que la presència de patrons estacionals pot amagar o distorsionar la tendència o altres característiques rellevants de la sèrie.
L'estacionalitat es pot detectar mitjançant diverses tècniques. Una de les més habituals és la descomposició clàssica, que divideix la sèrie en tendència, estacionalitat i residus, utilitzant mitjanes mòbils o mètodes similars, com veurem al Capítol 3. També es poden utilitzar models SARIMA, que incorporen components estacionals dins del model predictiu, o transformades de Fourier, que analitzen les freqüències per identificar patrons periòdics, tot i que aquesta última tècnica no la tractarem en aquest treball.

Ciclicitat

Els cicles representen oscil·lacions a llarg termini sense una periodicitat fixa, sovint associades a fenòmens econòmics, socials o ambientals que es desenvolupen al llarg de diversos anys. En alguns casos, en la descomposició de sèries temporals es pot introduir una component cíclica per capturar aquests comportaments recurrents amb un període superior a un any. Tanmateix, aquesta component és difícil de distingir de la tendència, ja que la separació entre ambdues no sempre és clara, per la qual cosa no sol incloure's explícitament en l'anàlisi.

A diferència de l'estacionalitat, els cicles no segueixen un patró regular ni es poden predir amb precisió. Per exemple, els cicles econòmics inclouen fases com expansió, auge, contracció i crisi, com la crisi financera del 2008, seguida d'un creixement gradual.
També hi ha cicles mediambientals, com *El Niño*, que afecten el clima global amb intervals variables. La identificació d'aquest tipus de comportaments pot resultar complicada quan els períodes són llargs i les dades disponibles són limitades. S'utilitzen tècniques com l'anàlisi espectral o les transformades d'ones (*wavelet analysis*) per detectar i analitzar aquestes dinàmiques a llarg termini [6].

Aleatorietat o soroll blanc

Aquesta component inclou els elements imprevisibles, habitualment associats a l'atzar. No segueix cap patró determinista o cíclic i pot interferir amb les altres components si no s'analitza adequadament.
Suposarem que el soroll blanc és una seqüència de variables aleatòries gaussianes, independents i idènticament distribuïdes amb mitjana zero i variància constant, és a dir,

$$a_t \sim N(0, \sigma^2).$$

Cadascuna de les variables a_t que s'afigen al procés a cada instant de temps t es coneixen com a innovacions.
Els residus resultants després de modelitzar la tendència, l'estacionalitat i la ciclicitat haurien de ser soroll blanc, complint propietats com la normalitat, la independència, la mitjana zero i la variància constant. Aquesta component ajuda a entendre la part de la variabilitat de la sèrie que no pot ser explicada pels altres components.
A banda d'aquestes quatre components principals, hi ha altres factors que poden influir en el comportament d'una sèrie temporal, tot i que són més específics i més costosos de modelitzar:

- **Efectes de calendari**: variacions relacionades amb factors com el nombre de dies laborables, els festius, o altres peculiaritats del calendari.
- **Efectes de promoció**: alteracions provocades per campanyes publicitàries, ofertes especials o accions de màrqueting.
- **Efectes de canvi de preu**: modificacions associades a pujades o baixades en els preus dels productes o serveis.
- **Efectes de canvi de política**: fluctuacions causades per ajustaments en polítiques econòmiques, socials o legals.

Per il·lustrar la idea de descomposició, l'Exemple 2.3.1 mostra com es pot descompondre una sèrie temporal en tendència, estacionalitat i soroll blanc fent servir la Figura 2.1.

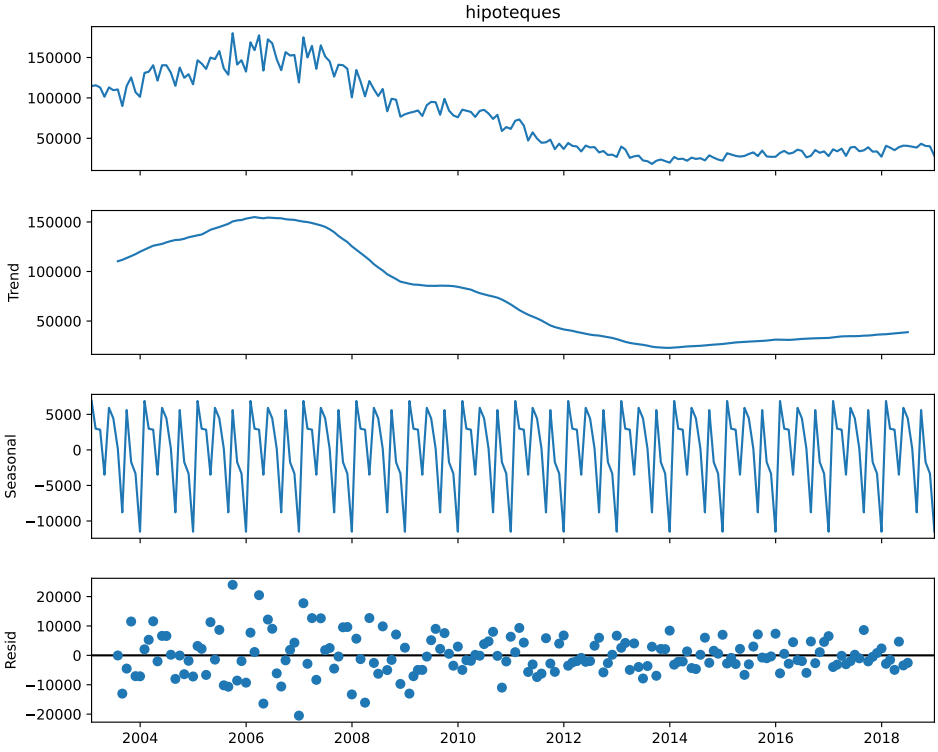

Figura 2.2: Descomposició de la sèrie temporal de les hipoteques concedides a l'estat espanyol entre 2003 i 2018, per mesos. La primera gràfica mostra la sèrie original; la segona, la tendència (*Trend*); la tercera, l'estacionalitat (*Seasonal*); i la quarta, el soroll blanc (*Resid*), que recull les variacions no explicades per les altres components.

Exemple 2.3.1 (Descomposició de la sèrie temporal de les hipoteques concedides a l'estat espanyol entre 2003 i 2018, per mesos.)**.** *La primera de les components mostra un comportament irregular, però sembla estabilitzar-se cap a finals del període. Comença amb una curta pujada lleugera, que ve seguida d'una caiguda abrupta fins a poc més de la meitat de la finestra temporal. Aquest desplomament coincideix amb la crisi econòmica del 2008, que va afectar greument el sector immobiliari, atés que els bancs van reduir dràsticament la concessió de crèdits hipotecaris i en van imposar requisits més estrictes. Per últim, es pot observar una estabilitat relativa fins al final del període, tot i que amb un xicotet signe de recuperació a velocitat moderada. La segona component, l'estacionalitat, permet identificar patrons cíclics que es repeteixen cada poc temps. Els indicis apunten a una estacionalitat trimestral, amb pics màxims a principis i finals d'any, i mínims a l'estiu. Finalment, la tercera component, el soroll blanc, presenta fluctuacions aleatòries de mitjana zero que no segueixen cap patró discernible, tot i que a mesura que avança el temps sembla que la variabilitat disminueix lleugerament.*

2.3.2 Models de sèries temporals

Quan una sèrie temporal es divideix en les seues components principals, es poden utilitzar diversos models matemàtics per explicar com interactuen aquestes parts. Aquests models es defineixen segons la manera en què les components es combinen entre si i es classifiquen en tres categories principals: models additius, models multiplicatius i models mixtos.
Cada tipus de model s'adapta a característiques específiques de les dades, com la magnitud de les fluctuacions o la relació entre les components.

Model additiu

El model additiu és el més senzill i comú per a l'anàlisi de sèries temporals. El valor observat y es pot descompondre en la suma de les seues components: la **tendència** (T), la **component cíclica** (C), la **component estacional** (S) i la **component aleatòria** (I). Aquest model és adequat quan les fluctuacions de la sèrie tenen magnitud constant i les components s'afigen entre elles. L'equació que el defineix és:

$$y = T + C + S + I.$$

Model multiplicatiu

El model multiplicatiu és més complex que el model additiu, però pot ser més adequat per a sèries temporals amb variacions percentuals constants. És especialment apropiat quan les fluctuacions de la sèrie són proporcionals al nivell de la tendència i les components interactuen de manera multiplicativa. En aquest cas, el valor y resulta de la multiplicació de les components, com s'indica a continuació:

$$y = T \times C \times S \times I.$$

Nota: Aquest model es pot expressar com un model additiu mitjançant una transformació logarítmica:

$$y = T \times C \times S \times I \quad \equiv \quad \log(y) = \log(T) + \log(C) + \log(S) + \log(I).$$

Model mixt

El model mixt és una combinació dels dos anteriors, i permet tindre en compte les característiques de cada component de manera més flexible. No hi ha una única manera de definir-lo, ja que pot variar segons les necessitats de l'anàlisi. Dos exemples comuns són:

$$y = (T + C) \times S \times I \quad \text{i}$$

$$y = T + (C \times S) + I.$$

2.3.3 Mesura del pes de la tendència i l'estacionalitat

Per entendre millor el comportament d'una sèrie temporal, pot resultar interessar quantificar la força de la tendència i l'estacionalitat, ja que això facilita la identificació dels factors que afecten la seua evolució.

Aquestes mesures resulten especialment útils quan es treballa amb una col·lecció extensa de sèries temporals i es volen detectar les que presenten una tendència o estacionalitat més marcada. Un estudi més profund d'aquestes mètriques es pot trobar a [31].

Força de la tendència

La força de la tendència es defineix com la proporció de la variació de la sèrie temporal que s'explica per la tendència. Si les dades presenten una tendència forta, les dades ajustades estacionalment haurien de tindre molta més variació que la component residual, de manera que el quocient $V(I)/V(T+C+I)$ hauria de ser relativament menut. Per contra, si la tendència és feble o inexistent, les dues variàncies seran similars, i el quocient serà proper a 1. Així, definim la força de la tendència com:

$$F_T = \max\left(0, 1 - \frac{V(I)}{V(T+I)}\right) \tag{2.3}$$

Això dona una mesura de la força de la tendència entre 0 i 1. Com que la variància del residu de vegades pot ser més gran que la de les dades ajustades estacionalment, establim el valor mínim possible de F_T igual a zero.

Força de l'estacionalitat

La força de l'estacionalitat es defineix de manera similar, però en aquest cas respecte a les dades ajustades per la tendència en lloc de les dades ajustades estacionalment.

$$F_S = \max\left(0, 1 - \frac{V(I)}{V(S+I)}\right) \tag{2.4}$$

Una sèrie amb una força estacional F_S pròxima a 0 presenta poca o cap estacionalitat, mentre que una sèrie amb una estacionalitat forta tindrà F_S pròxima a 1, perquè la variància de la component residual o aleatòria I serà molt més menuda que $V(S+I)$.

2.4 Estacionarietat i transformacions per aconseguir-la

L'estacionarietat és una propietat important en l'anàlisi de sèries temporals, ja que molts models estadístics assumeixen que la sèrie analitzada conserva característiques estables en el temps. Una sèrie és estacionària si les seues propietats estadístiques, com la mitjana i la variància, romanen constants al llarg del temps.
Tanmateix, la majoria de les sèries temporals reals no són estacionàries de manera natural. Això pot ser degut a l'existència de tendències, estacionalitat o altres components. Per aquest motiu, com hem indicat prèviament, és habitual aplicar transformacions com la diferenciació o els ajustos logarítmics per estabilitzar les propietats de la sèrie i aconseguir l'estacionarietat.
En molts casos, és útil definir formalment què significa que una sèrie siga estacionària. En aquest sentit, es distingeixen dues nocions principals: l'estacionarietat en sentit estricte i l'estacionarietat en sentit feble.

Definició 2.4.1 (Procés estacionari (en sentit estricte))**.** *Es diu que un procés estocàstic (en particular, una sèrie temporal) és estacionari en sentit estricte si:*

1. *les distribucions marginals de totes les variables són idèntiques;*
2. *les distribucions finit-dimensionals de qualsevol conjunt de variables depenen només dels retards entre elles.*

L'estacionarietat en sentit estricte és una propietat matemàtica forta que implica que el procés té les mateixes propietats estadístiques al llarg del temps, independentment de l'instant considerat. A la pràctica, però, sovint s'utilitza una definició menys restrictiva coneguda com a estacionarietat en sentit feble o estacionarietat de segon ordre.

Definició 2.4.2 (Procés estacionari (en sentit feble o de segon ordre))**.** *Un procés estocàstic es considera estacionari en sentit feble si:*

1. *L'esperança (o mitjana)* $\mathbb{E}[z_t]$ *és constant i independent del temps.*
2. *La variància* $V(z_t)$ *és finita i constant al llarg del temps.*
3. *La funció d'autocovariància* $Cov(z_t, z_{t+h})$ *depén només del retard* h *i no del temps* t*.*

Com distingir una sèrie estacionària d'una no estacionària

Deduir si una sèrie és estacionària o no és un pas determinant en l'anàlisi de sèries temporals, atés que molts models estadístics assumeixen aquesta propietat. A continuació, es presenten diferents enfocaments per identificar l'estacionarietat:

1. Visualització gràfica. La inspecció visual és una tècnica inicial per detectar patrons estacionaris o no estacionaris, però no sempre és suficient per a confirmar l'estacionarietat amb rigor estadístic. Una sèrie estacionària es caracteritza per fluctuacions al voltant d'una mitjana constant amb una variància estable. En canvi, una sèrie no estacionària pot mostrar:

- Una tendència ascendent o descendent.
- Patrons estacionals recurrents.
- Variacions que augmenten o disminueixen amb el temps.

Aquest tret es pot observar amb més claredat en gràfiques de descomposició, com la que mostra la Figura 2.2.

2. Anàlisi d'autocorrelació. L'autocorrelació mesura la relació entre els valors d'una sèrie en diferents moments del temps:

- En una sèrie estacionària, l'autocorrelació decreix ràpidament amb el retard (*lag*).
- En una sèrie no estacionària, l'autocorrelació pot persistir al llarg de molts retards, indicant dependència a llarg termini.

Això es pot analitzar mitjançant gràfiques d'autocorrelació i d'autocorrelació parcial, que veurem amb detall a la Secció 4.1.1.

3. Proves estadístiques. Tot i que l'anàlisi d'autocorrelació i les gràfiques poden donar indicis sobre l'estacionarietat d'una sèrie, per confirmar-ho amb garanties és recomanable aplicar proves estadístiques. Es poden aplicar proves estadístiques per determinar si una sèrie és estacionària. En aquest treball farem servir el test més utilitzat en estos casos: la prova de Dickey-Fuller augmentada(ADF), que contrasta la hipòtesi nul·la que la sèrie té una arrel unitària (no estacionària) contra l'alternativa que és estacionària. La veurem amb més deteniment a la Secció 4.3.1.

Transformacions per aconseguir estacionarietat

Quan ens trobem davant una sèrie temporal no estacionària, és necessari aplicar transformacions per a convertir-la en estacionària abans de modelitzar-la.
Les transformacions més comunes per aconseguir estacionarietat inclouen:

- **Eliminació de la tendència.** Si la sèrie presenta una tendència clara (ascendent o descendent), es pot eliminar mitjançant dos enfocaments principals, depenent de si la tendència és determinista o estocàstica.

  1. **Substracció d'una funció de tendència:** Es pot ajustar una funció de tendència lineal, polinòmica o logarítmica als valors observats i restar-la de la sèrie original.
  2. **Diferenciació:** Aquest mètode calcula la diferència entre valors consecutius de la sèrie:

     $$\bar{z}_t = z_t - z_{t-1}.$$

     En general, per calcular diferències estacionals de període c, s'utilitza l'operador de diferència ∇_c:

     $$\nabla_c z_t = z_t - z_{t-c}. \tag{2.5}$$

     Aquest procediment genera la sèrie de diferències primeres, que sovint és estacionària. Si no s'aconsegueix estacionarietat amb una sola diferenciació, es pot repetir el procés calculant diferències successives (segones diferències, terceres, etc.) fins a obtenir una sèrie estacionària. Aquesta tècnica és prou habitual en models ARIMA.

- **Desestacionalització.** Quan la sèrie mostra patrons estacionals, aquests es poden eliminar ajustant un model estacional i restant-lo de les dades originals. També es poden utilitzar tècniques de descomposició, com la descomposició clàssica o la de components STL ([7]), que separen la component estacional de la resta de la sèrie. Una vegada eliminada l'estacionalitat, la sèrie residual es pot considerar estacionària.

- **Transformacions de potència.** Si la variància de la sèrie no és constant (heteroscedasticitat), es poden aplicar ajustos que estabilitzen la variància. Els més freqüents són:

  $$\bar{z}_t = \log(z_t), \quad \bar{z}_t = \sqrt{z_t}, \quad \bar{z}_t = \frac{z_t^\lambda - 1}{\lambda} \quad \text{(Transformació Box-Cox)},$$

  on λ és un paràmetre que controla la intensitat de la transformació. Per exemple:

  – $\lambda = 0$: Es converteix en una transformació logarítmica.

- $\lambda = 1$: És equivalent a no aplicar cap transformació.
- Altres valors de λ es poden determinar utilitzant criteris estadístics com la màxima versemblança.

La transformació Box-Cox és especialment útil per tractar amb sèries amb creixements exponencials o variàncies no constants [19].

Capítol 3

Models per a la predicció

Després de revisar les nocions bàsiques, procedim a presentar els models estadístics clàssics més utilitzats en l'anàlisi de sèries temporals. En aquest capítol, es descriuran els models autorregressius (AR), els models de mitjana mòbil (MA) i els models combinats ARMA. Abans, però, es farà una breu introducció a la regressió lineal simple, que servirà com a punt de partida per familiaritzar-se amb conceptes clau d'ajust de models.
Posteriorment, s'afegirà un nivell addicional de complexitat amb els models ARIMA, que seran vinculats amb els processos SARIMA, una extensió més general que inclou la modelització de l'estacionalitat. Finalment, es tractaran els models de suavització exponencial, una alternativa als models ARIMA útil en determinades ocasions, que indicarem més avant, oferint una manera flexible d'ajustar prediccions basades en els valors recents de la sèrie.
El desenvolupament teòric d'aquests models està basat en les referències [27] i [4], encara que també s'han tingut en compte algunes idees de [8] i [33].

3.1 Regressió lineal simple

Abans d'estudiar els models esmentats, és convenient fer un repàs de la regressió lineal simple, una tècnica senzilla que serveix com a base per a la comprensió dels processos més complexos. Aquesta eina estadística permet modelitzar la relació entre una variable dependent y_t i una variable independent x_t mitjançant una funció lineal.
La regressió lineal simple assumeix que la relació entre les dues variables es pot expressar com:

$$y_t = c + bx_t + a_t,$$

on:

- c és l'*intercept* o terme independent, que representa el valor esperat de y_t quan $x_t = 0$.
- b és el pendent o coeficient que determina la magnitud i direcció del canvi en y_t per unitat d'increment en x_t.
- a_t és un terme de soroll blanc.

Els paràmetres c i b es poden estimar mitjançant el mètode dels mínims quadrats, que minimitza la suma dels quadrats dels residus $(y_t - \hat{y}_t)$, assegurant que l'ajust lineal siga òptim en termes d'error quadràtic. Ací, $\hat{y}_t = c + bx_t$ representa el valor estimat de la variable dependent y_t segons el model ajustat, mentre que $y_t - \hat{y}_t$ és el residu que mesura la discrepància entre el valor real y_t i l'estimat $\hat{y}_t$. L'objectiu és optimitzar els paràmetres c i b per reduir aquesta discrepància total al mínim possible. Per il·lustrar aquest concepte, presentem l'Exemple 3.1.1, que mostra com aplicar la regressió lineal simple per a modelitzar la relació entre les hores d'estudi i les notes d'un examen, una situació molt senzilla però útil per a entendre el funcionament d'aquesta tècnica.

Exemple 3.1.1 (Cas senzill de regressió lineal simple)**.** *Suposem que volem analitzar la relació entre les hores d'estudi, x_t, i la nota obtinguda en un examen, y_t. Els resultats es poden modelitzar amb una equació de regressió lineal simple:*

$$y_t = c + bx_t + a_t,$$

on:

- *c és la nota mínima que podria obtindre's sense estudiar.*
- *b és l'augment mitjà de la nota per cada hora d'estudi.*
- *a_t és el soroll blanc, que representa factors aleatoris que poden influir en la nota.*

Hores d'estudi (x_t)	**Nota obtinguda** (y_t)
1	4
2	5
3	6
4	7

Taula 3.1: Exemple de dades per a una regressió lineal simple entre hores d'estudi i nota obtinguda.

Per exemple, utilitzant les dades següents de la Taula 3.1, podem estimar els paràmetres c i b amb el mètode dels mínims quadrats.
Els resultats de l'estimació són:

$$\hat{c} = 3 \quad i \quad \hat{b} = 1.$$

Així, la relació prevista entre hores d'estudi i nota obtinguda seria:

$$\hat{y}_t = 3 + x_t.$$

Per exemple, si s'estudien 6 hores, la nota prevista és:

$$\hat{y}_t = 3 + 6 = 9.$$

NOTA: En aquest exemple no es fa menció explícita al soroll blanc a_t, ja que l'objectiu principal és il·lustrar la relació determinística entre les hores d'estudi x_t i la nota y_t. El soroll blanc es considera com una variabilitat aleatòria que no afecta significativament la interpretació del model ajustat en aquest context simplificat.

Relació amb els processos autorregressius

Si modifiquem l'equació de la regressió lineal simple per permetre que la variable independent x_t depenga dels valors passats de y_t, obtenim un model que s'acosta al concepte d'autorregressió. Això implica canviar x_t per y_{t-1}, transformant l'equació inicial en:

$$y_t = c + \phi y_{t-1} + a_t,$$

on ϕ és el coeficient d'autorregressió que determina la influència del valor passat y_{t-1} en el valor actual y_t.
Aquest model es coneix com a procés autorregressiu de primer ordre AR(1). En el context en què ens trobem, la variable present y_t depén únicament del valor immediatament anterior y_{t-1} i del soroll blanc a_t.
Si es vol tindre en compte més d'una observació passada, només cal afegir més termes a l'equació:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + a_t,$$

obtenint així un procés autorregressiu d'ordre p AR(p), que s'estudiarà a la secció següent. Aquest enfocament és la base de molts models avançats de sèries temporals, com els models ARIMA.

3.2 Models autorregressius

Els processos autorregressius parteixen de la premissa que existeix una relació estrictament lineal entre els valors actuals i els previs, la qual cosa no només simplifica l'anàlisi sinó que també facilita l'aplicació de tècniques estadístiques tradicionals. Aquesta linealitat és un punt de partida crucial per construir un marc predictiu clar i comprensible. Els models autorregressius són especialment útils per a sèries temporals que no són completament aleatòries i que mostren patrons predictibles, com estructures cícliques o fluctuacions estacionàries al voltant d'una mitjana. Es defineixen per un ordre, p, que especifica el nombre de valors passats utilitzats en la predicció. El cas més senzill que estudiarem és el model autorregressiu de primer ordre (AR(1)), que fa servir només una observació passada per a predir el valor actual. Tot seguit, el generalitzarem a ordres superiors (AR(p)).

L'operador de retard

Per simplificar la notació i l'anàlisi dels models autorregressius, s'introdueix l'*operador de retard*. Aquest operador no és més que un operador lineal que, aplicat a una funció temporal, proporciona la mateixa funció retardada un període. De manera general, es defineix així:

$$B[f(t)] \equiv f(t-1).$$

Tot i això, la notació que s'emprarà al llarg d'aquest treball serà la següent:

$$Bz_t = z_{t-1}. \tag{3.1}$$

L'operador de retard satisfà aquestes propietats:

- Si c és una constant, $Bc = c$.
- Si c és una constant i z_t és una sèrie temporal, $Bcz_t = cBz_t = cz_{t-1}$.
- És un operador lineal, és a dir, $B(cz_t + dv_t) = cz_{t-1} + dv_{t-1}$ i $B^k z_t = B \ldots Bz_t = z_{t-k}$.

El procés autorregressiu de primer ordre (AR(1))

Definició 3.2.1 (Procés autorregressiu de primer ordre (AR(1))). *Direm que una sèrie z_t segueix un procés AR(1) si ha sigut generada per:*

$$z_t = c + \phi z_{t-1} + a_t,$$

on c i ϕ són constants a determinar i a_t és un procés de soroll blanc.

NOTA: Perquè el procés siga estacionari, cal que es complisca la condició $-1 < \phi < 1$. Per comprovar-ho, suposem que el procés comença amb $z_0 = h$, sent h un valor fix qualsevol. Els valors posteriors es poden calcular de manera recursiva:

$$\begin{aligned} z_1 &= c + \phi z_0 + a_1 = c + \phi h + a_1, \\ z_2 &= c + \phi z_1 + a_2 = c + \phi(c + \phi h + a_1) + a_2 = c(1+\phi) + \phi^2 h + \phi a_1 + a_2, \\ &\vdots \\ z_t &= c\textstyle\sum_{i=0}^{t-1} \phi^i + \phi^t h + \sum_{i=0}^{t-1} \phi^i a_{t-i}. \end{aligned}$$

Per calcular l'esperança de z_t, cal tindre en compte que l'esperança de les innovacions a_t és zero i que l'esperança d'una constant és la mateixa constant. Així, s'obté:

$$\mathbb{E}[z_t] = c\sum_{i=0}^{t-1} \phi^i + \phi^t h.$$

Es pot observar que l'esperança de z_t depén de t, fet que implica que el procés no és estacionari (en general). Per aconseguir estacionarietat, cal que el primer sumand convergisca a una constant i que el segon siga zero. Això només ocorre quan $-1 < \phi < 1$, ja que la sèrie geomètrica $\sum_{i=0}^{t-1} \phi^i$ convergeix a $\frac{1}{1-\phi}$, fent que el primer sumand tinga com a límit $\frac{c}{1-\phi}$. D'altra banda, el terme $\phi^t h$ tendeix a zero quan $|\phi| < 1$. Per tant, la mitjana del procés és:

$$\mathbb{E}[z_t] = \frac{c}{1-\phi},$$

que és independent de t, complint així una de les condicions necessàries de la Definició 2.4.2. Les altres dues condicions es poden provar de manera similar.
Una manera més compacta d'expressar els processos autorregressius és mitjançant l'operador de retard (Equació 3.1). En el cas d'un procés AR(1), podem escriure:

$$z_t = c + \phi B z_t + a_t.$$

Si definim la sèrie centrada $\tilde{z}_t = z_t - \mu$, sent $\mu = \frac{c}{1-\phi}$ la mitjana del procés, es té que:

$$B\tilde{z}_t = \tilde{z}_{t-1}.$$

Per tant, el procés AR(1) quedaria resumit així:

$$(1 - \phi B)\tilde{z}_t = a_t.$$

Aquesta expressió indica que una sèrie centrada segueix un procés AR(1) si, en aplicar-li l'operador $(1 - \phi B)$, s'obté un procés de soroll blanc. Aquest operador pot ser interpretat com un filtre que elimina la informació autoregressiva, deixant únicament el soroll blanc.

Condició d'estacionarietat

La condició d'estacionarietat del procés AR(1) que acabem de veure es pot analitzar en termes de l'operador de retard. Si considerem l'operador com una equació en B, la condició d'estacionarietat es tradueix en el valor absolut del factor ϕ. Alternativament, podem calcular l'arrel de l'operador igualant a zero:

$$1 - \phi B = 0 \quad \Longrightarrow \quad B = \frac{1}{\phi}.$$

Per tant, perquè la sèrie siga estacionària cal que el valor absolut de l'arrel siga major que 1, és a dir:

$$|B| > 1 \quad \text{o, equivalentment,} \quad |\phi| < 1.$$

El procés autorregressiu general (AR(p))

Definició 3.2.2 (Procés autorregressiu d'ordre p (AR(p)))**.** *Direm que una sèrie temporal z_t estacionària segueix un procés autorregressiu d'ordre p (AR(p)) si:*

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \cdots + \phi_p \tilde{z}_{t-p} + a_t,$$

on $\tilde{z}_t = z_t - \mu$, sent μ la mitjana del procés estacionari z_t i a_t un procés de soroll blanc.

Fent servir la notació d'operadors, l'equació anterior es pot escriure com:

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\tilde{z}_t = a_t.$$

Si es denota $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ al polinomi de grau p en l'operador de retard, el primer terme del qual és 1, es té:

$$\phi_p(B)\tilde{z}_t = a_t,$$

que és l'expressió general d'un procés autorregressiu.

Equació característica

Definició 3.2.3 (Equació característica)**.** *Anomenarem equació característica del procés autorregressiu a l'equació*

$$\phi_p(B) = 0,$$

considerada com a funció de B.

Aquesta equació té p arrels, $G_1^{-1}, \ldots, G_p^{-1}$, en general distintes. Alternativament, es pot escriure:

$$\phi_p(B) = \prod_{i=1}^{p} (1 - G_i B),$$

de manera que els coeficients G_i són els factors de l'equació característica. Es pot provar que el procés és estacionari si $|G_i| < 1 \ \forall i$.

3.3 Models de mitjana mòbil

Tot i que són molt útils per estudiar patrons, els processos AR tenen limitacions a l'hora de representar sèries amb memòria molt curta, on el valor actual de la sèrie depén principalment d'un nombre reduït de valors passats. Aquesta característica dificulta que els processos AR capturen amb precisió les dinàmiques d'aquest tipus de sèries.
Una família de processos que s'adapten millor a aquesta propietat de «memòria molt curta» són els processos de mitjana mòbil (MA, *moving average* en anglés). Aquests processos es basen en un nombre finit, i generalment menut, d'innovacions passades, fent-los més adequats per a situacions en què la influència dels valors anteriors es dissipa ràpidament.

El procés de mitjana mòbil de primer ordre (MA(1))

Definició 3.3.1 (Procés de mitjana mòbil de primer ordre (MA(1)))**.** *Direm que una sèrie temporal z_t estacionària segueix un procés de mitjana mòbil de primer ordre (MA(1)) si:*

$$\tilde{z}_t = a_t - \theta a_{t-1},$$

on $\tilde{z}_t = z_t - \mu$, sent μ la mitjana del procés, θ una constant a determinar i a_t un procés de soroll blanc.

El procés MA(1) es pot escriure de manera més compacta mitjançant la notació d'operadors:

$$\tilde{z}_t = (1 - \theta B)a_t,$$

on B és l'operador de retard definit anteriorment (Equació 3.1). Aquest procés és la suma ponderada de dos termes estacionaris, a_t i $-\theta a_{t-1}$, i, per tant, serà estacionari per a qualsevol valor de θ, a diferència dels processos AR.
Suposarem que $|\theta| < 1$, de manera que la innovació a_t té més pes que a_{t-1} en la generació de z_t. Amb aquesta condició, el procés és *invertible*; és a dir, els valors passats de les innovacions es poden recuperar a partir de la sèrie observada. Aquesta propietat també implica que l'efecte de les innovacions passades decreix exponencialment amb el temps.
Per justificar aquesta propietat, podem desenvolupar iterativament el terme a_{t-1} en la definició del procés, utilitzant la seua relació amb els valors anteriors. Substituint a_{t-1} per la seua expressió en termes de z_{t-1} i a_{t-2}, obtenim:

$$\tilde{z}_t = a_t - \theta(\tilde{z}_{t-1} + \theta a_{t-2}) = a_t - \theta\tilde{z}_{t-1} - \theta^2 a_{t-2}.$$

Repetint aquest procés per a_{t-2} i termes successius, trobem:

$$\tilde{z}_t = a_t - \theta\tilde{z}_{t-1} - \theta^2(\tilde{z}_{t-2} + \theta a_{t-3}) = a_t - \theta\tilde{z}_{t-1} - \theta^2\tilde{z}_{t-2} - \theta^3 a_{t-3}.$$

Continuant aquest procediment, el procés MA(1) es pot expressar com una combinació lineal de les innovacions passades amb pesos que decreixen exponencialment:

$$\tilde{z}_t = a_t - \theta^t a_0 - \sum_{i=1}^{t-1} \theta^i \tilde{z}_{t-i}.$$

Quan $|\theta| < 1$, aquests pesos tendeixen a zero ràpidament, assegurant la invertibilitat. Per contra, si $|\theta| > 1$, l'efecte de les innovacions passades augmentaria exponencialment, cosa que faria el procés no invertible malgrat ser estacionari.
D'ara endavant, suposarem que el procés és invertible. Així, com $|\theta| < 1$, existeix una inversa de l'operador $(1 - \theta B)$, que es pot expressar com:

$$(1 + \theta B + \theta^2 B^2 + \dots)\tilde{z}_t = a_t,$$

que equival a:

$$\tilde{z}_t = -\sum_{i=1}^{\infty} \theta^i \tilde{z}_{t-i} + a_t.$$

El procés de mitjana mòbil general (MA(q))

Definició 3.3.2 (Procés de mitjana mòbil d'ordre q (MA(q)))**.** *Direm que una sèrie temporal z_t estacionària segueix un procés de mitjana mòbil d'ordre q (MA(q)) si:*

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q},$$

on $\tilde{z}_t = z_t - \mu$, sent μ la mitjana del procés, $\theta_1, \dots, \theta_q$ constants a determinar i $a_{t-q}, \dots, a_{t-1}, a_t$ processos de soroll blanc.

Fent servir la notació d'operadors, l'equació anterior es pot escriure com:

$$\tilde{z}_t = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q) a_t.$$

Si procedim de manera anàloga al que s'ha fet amb els processos AR i denotem $\theta_p(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$, encara es pot compactar més l'expressió:

$$\tilde{z}_t = \theta_q(B) a_t.$$

És convenient destacar que un procés MA(q) sempre és estacionari, atés que és la suma de processos estacionaris. Direm que el procés és *invertible* si les arrels de l'equació característica ($\theta_q(B) = 0$) són totes de mòdul major que 1. Això garanteix que els coeficients de la representació infinita del procés decreixen exponencialment, evitant que els efectes de les innovacions passades augmenten amb el temps.
Nota: **El procés de mitjana mòbil infinita (MA(∞)).**
Els processos AR i MA esmentats abans són casos particulars d'una representació general de processos estacionaris obtinguda per Herman Wold [25]. L'estadista noruec va demostrar que qualsevol procés estocàstic feblement estacionari, z_t, de mitjana finita, μ, que no continga components deterministes, es pot expressar com una funció lineal de variables aleatòries incorrelades, a_t, com:

$$z_t = \mu + \sum_{i=0}^{\infty} \psi_i a_{t-i} = \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \dots,$$

on $\psi_0 = 1$, $E(z_t) = \mu$, $E(a_t) = 0$, $V(a_t) = \sigma^2$ i $E(a_t a_{t-k}) = 0,\ k > 1$. Suposem que $\sum_{i=1}^{0} \psi_i^2 < \infty$ perquè la sèrie estiga ben definida. Si anomenem $\tilde{z}_t = z_t - \mu$ i utilitzem l'operador de retard, podem escriure:

$$\tilde{z}_t = \psi(B) a_t.$$

3.4 Models combinats ARMA

Després d'haver comprés el funcionament dels models AR i dels models MA, el pas immediatament posterior és combinar-los en un model ARMA. Cal recordar que, tot i que un model AR d'ordre finit es defineix mitjançant uns pocs paràmetres, la seua representació com a procés de mitjana mòbil comporta una sèrie infinita de coeficients que, sota condicions d'estacionarietat, decauen de manera fixa (habitualment exponencial). En contrast, els models MA presenten directament un nombre finit de coeficients no nuls, que poden prendre valors arbitràriament. Els models ARMA integren aquestes dues aproximacions: combinen una component autorregressiva, que confereix un patró de decreixement fix a la representació infinita, amb una component de mitjana mòbil, que aporta una sèrie finita de coeficients determinats per paràmetres lliures. D'aquesta manera, un model ARMA descriu la variable d'interés en funció tant de les observacions passades com dels errors anteriors, sense ser simplement la suma independent d'un procés AR i un procés MA.

El procés ARMA(1,1)

El cas més simple de procés combinat és l'ARMA(1,1), que és la combinació d'un procés AR(1) i un procés MA(1):

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + a_t - \theta_1 a_{t-1}.$$

Fent ús de la notació compacta, l'equació anterior es pot escriure com:

$$(1 - \phi_1 B)\tilde{z}_t = (1 - \theta_1 B)a_t.$$

En aquest cas, cal que $|\phi_1| < 1$ perquè el procés siga estacionari, i que $|\theta_1| < 1$ perquè siga invertible. A més, suposarem que $\phi_1 \neq \theta_1$, atés que si ambdós paràmetres foren idèntics, el producte de $(1-\phi_1 B)^{-1}$ pels dos costats de la igualtat anterior donaria lloc a l'equació $\tilde{z}_t = a_t$, que és un procés de soroll blanc. En la formulació dels models ARMA suposarem sempre que els operadors AR i MA no tenen arrels en comú.

Processos ARMA(p,q)

La rellevància dels processos ARMA rau en el fet que una sèrie temporal estacionària sovint es pot modelitzar de manera més eficient amb un model ARMA que amb un model MA pur o AR individual, ja que requereix un menor nombre de paràmetres.
Generalitzant el model ARMA(1,1), un procés ARMA(p,q) es pot expressar com:

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \cdots + \phi_p \tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q},$$

o, utilitzant l'operador B definit abans:

$$(1 - \phi_1 B - \cdots - \phi_p B^p)\tilde{z}_t = (1 - \theta_1 B - \cdots - \theta_q B^q)a_t.$$

Alternativament, si volem simplificar l'expressió al màxim, es pot fer servir la notació compacta:

$$\phi_p(B)\tilde{z}_t = \theta_q(B)a_t.$$

El procés serà estacionari si les arrels de l'equació característica de l'operador AR, $\phi_p(B) = 0$, són totes de mòdul major que 1, i invertible si les arrels de l'equació característica de l'operador MA, $\theta_q(B) = 0$, compleixen la mateixa condició. És a dir, el procés serà estacionari si ho és el procés AR i invertible si ho és el procés MA.
Suposarem, a més, que els operadors AR i MA no tenen arrels en comú.

3.5 Models ARIMA

Els models vists fins ara es basen en el principi d'estacionarietat, un supòsit fonamental que implica que les característiques estadístiques de la sèrie, com ara la mitjana i la variància, es mantenen constants al llarg del temps, i que la covariància entre valors depén únicament de la distància temporal entre ells, no del moment específic. Tanmateix, moltes sèries temporals del món real no compleixen aquesta condició, ja que solen presentar variacions en la seua estructura, com ara canvis en la variància o la tendència. Quan una sèrie temporal no és estacionària, però esdevé estacionària després d'aplicar-li diferències successives d, es classifica com un procés integrat d'ordre d. Aquest enfocament permet modelar sèries temporals no estacionàries transformant-les en estacionàries i aplicant models com l'ARMA(p, q), que combinen components autoregressius i de mitjana mòbil. La sèrie original, una vegada diferenciada i modelada, s'identifica com un model ARIMA(p, d, q), acrònim d'*Autoregressive Integrated Moving Average*.
El terme *integrat* fa referència precisament a aquest procés de diferenciació i posterior reintegració per reconstruir la sèrie original a partir del model ajustat. Aquest mecanisme és essencial per predir sèries temporals inicialment no estacionàries i ha convertit els models ARIMA en una eina freqüentment utilitzada en molts camps.
En aquest context, p representa el nombre de termes autoregressius, d el grau de diferenciació necessari per a aconseguir l'estacionarietat, i q el nombre de termes de mitjana mòbil. Per exemple, un model ARIMA$(0, d, 0)$ es correspon amb una sèrie temporal que esdevé un soroll blanc després d'haver estat diferenciada d vegades. La seua representació matemàtica és:

$$(1 - B)^d \tilde{z}_t = a_t,$$

on B és l'operador de retard (Equació 3.1) i a_t és un soroll blanc.
El model ARIMA(p, d, q) general es pot expressar com:

$$\phi_p(B)(1 - B)^d \tilde{z}_t = \theta_q(B) a_t,$$

on $\phi_p(B)$ i $\theta_q(B)$ són polinomis que representen les parts AR i MA, respectivament. Si algun dels paràmetres p, d, q és zero, el terme corresponent desapareix de l'equació, simplificant el model.

3.6 Models SARIMA

Moltes sèries temporals presenten patrons estacionals, és a dir, repeticions periòdiques al llarg del temps, com poden ser les variacions anuals en el consum energètic o les vendes que augmenten en determinades èpoques de l'any. La periodicitat d'aquests patrons ve determinada per la longitud del cicle estacional, representada per c, que indica cada quants períodes es repeteix la mateixa estructura.

Quan treballem amb sèries estacionals, abans d'ajustar un model és necessari transformar la sèrie perquè siga estacionària. Una forma habitual d'aconseguir-ho és aplicant diferències estacionals de període c, fent servir l'Equació 2.5.
Si amb aquesta transformació s'aconsegueix eliminar la tendència i l'estacionalitat, podem considerar la sèrie com a estacionària i procedir amb la modelització. En cas contrari, caldria aplicar també diferenciació regular.
Un dels models més utilitzats per a capturar tant la dependència regular com l'estacional és el model ARIMA estacional o SARIMA(p,d,q)(P,D,Q), que combina components autoregressius (AR), de mitjana mòbil (MA) i d'integració (I), incorporant també la dependència estacional. L'expressió general d'aquest model és:

$$\left(1-\sum_{i=1}^{p}\phi_i B^i\right)\left(1-\sum_{j=1}^{P}\Phi_j B^{jc}\right)\nabla^d\nabla_c^D z_t = \left(1-\sum_{k=1}^{q}\theta_k B^k\right)\left(1-\sum_{m=1}^{Q}\Theta_m B^{mc}\right)a_t, \tag{3.2}$$

on B és l'operador retard (Equació 3.1) i a_t representa un terme d'error amb distribució de soroll blanc.

3.7 Suavització exponencial

La suavització exponencial és una tècnica àmpliament utilitzada en l'anàlisi de sèries temporals per a generar prediccions basades en valors històrics. El concepte clau és donar un pes decreixent a les dades a mesura que es tornen més antigues, mantenint una dependència més gran en les observacions recents. Aquesta estratègia resulta especialment útil per tractar amb dades que presenten fluctuacions aleatòries, ja que ajuda a suprimir el soroll i identificar patrons més estables.
La suavització exponencial és una alternativa senzilla i eficient a mètodes més complexos com els ARIMA. Els seus principals avantatges són:

- **Eficiència computacional**: Els càlculs necessaris són senzills i poden ser implementats en temps real, fent-la apta per aplicacions amb dades contínues.
- **Adaptabilitat**: Respon ràpidament a canvis en la dinàmica de les dades, com noves tendències o ruptures en el comportament de la sèrie.
- **Flexibilitat**: Es pot aplicar a una gran varietat de sèries temporals, des de dades amb comportament estàtic fins a aquelles amb tendències o estacionalitats.
- **Simplicitat**: A diferència d'altres models com l'ARIMA, no necessita una gran quantitat de dades històriques.

Tot i els seus avantatges, la suavització exponencial també presenta algunes limitacions que cal considerar:

- **No captura bé les estacionalitats**: Tot i que existeixen extensions com el model de Holt-Winters, la versió bàsica de la suavització exponencial no gestiona de manera efectiva patrons estacionals marcats.

- **Sensibilitat al paràmetre de suavització**: L'elecció del factor d'esmortiment (α) és crítica i pot afectar significativament la qualitat de la predicció. Un valor massa alt pot reaccionar en excés a les variacions recents, mentre que un de massa baix pot perdre informació rellevant.
- **No ofereix informació estructural**: A diferència dels models ARIMA, la suavització exponencial no proporciona una descomposició explícita dels components de la sèrie (tendència, estacionalitat i soroll), limitant la seua capacitat d'interpretació.
- **Prediccions limitades a curt termini**: Com que es basa fortament en les últimes observacions, tendeix a funcionar millor en horitzons de predicció curts i pot no ser fiable per a projeccions a llarg termini.

Model de suavització exponencial simple

La suavització exponencial simple és el model més bàsic dins la família de tècniques de suavització exponencial. Aquest model atorga pesos decreixents de manera exponencial als valors més antics, depenent d'un únic paràmetre, conegut com a factor d'esmortiment (α). Així, les observacions més recents tenen una influència més gran en la predicció, mentre que les més antigues es van atenuant progressivament.
L'equació de recurrència de la suavització exponencial simple és:

$$L_t = \alpha \cdot z_t + (1 - \alpha) \cdot L_{t-1}, \tag{3.3}$$

on:

- L_t és el valor suavitzat en el temps t,
- z_t és el valor observat en el temps t,
- L_{t-1} és el valor suavitzat en el temps $t - 1$,
- α és el factor d'esmortiment, amb $0 \leq \alpha \leq 1$.

Aquest model assumeix que les observacions recents són més rellevants per predir els valors futurs que les més antigues, aplicant un pes exponencial decreixent a mesura que ens allunyem en el temps.

Interpretació del paràmetre α

- Si α és proper a 1, el model respon ràpidament als canvis en les dades.
- Si α és proper a 0, el model genera prediccions més suaus, però pot tardar a adaptar-se als canvis. És a dir, les noves observacions tenen un pes menor en la predicció.

Model de suavització exponencial doble (Holt lineal)

Quan una sèrie temporal presenta una tendència, la suavització exponencial simple no és suficient per a captar la seua dinàmica. Per aquest motiu, es pot utilitzar la suavització exponencial doble, també coneguda com a suavització exponencial de Holt. Aquest mètode amplia la suavització exponencial

simple afegint-hi un component de tendència, que també es va actualitzant de manera exponencial al llarg del temps.
El model es basa en dues components principals:

- L_t, que representa el valor suavitzat de la sèrie en el temps t.
- T_t, que estima la tendència de la sèrie en el temps t.

Les equacions de recurrència són:

$$L_t = \alpha \cdot z_t + (1 - \alpha) \cdot (L_{t-1} + T_{t-1}), \tag{3.4}$$
$$T_t = \beta \cdot (L_t - L_{t-1}) + (1 - \beta) \cdot T_{t-1}, \tag{3.5}$$

on:

- β és el factor de suavització de la tendència, amb $0 \leq \beta \leq 1$.
- La resta de paràmetres són els mateixos que en la suavització exponencial simple.

Interpretació dels paràmetres

- Si β és proper a 1, la tendència estimada pot variar de manera brusca.
- Si β és proper a 0, la tendència estimada és més estable i canvia més lentament.

Model de suavització exponencial triple (Holt-Winters)

Si una sèrie temporal mostra evidències significatives tant de tendència com d'estacionalitat, cal explorar mètodes més avançats que permeten incorporar aquestes components. Per resoldre aquesta limitació, Holt i Winters van desenvolupar el que es coneix com a suavització exponencial triple, una tècnica que combina la suavització del nivell i la tendència amb un component estacional.
Aquest model va ser introduït per Peter Winters el 1960 [32], basant-se en treballs previs de Charles Holt sobre suavització exponencial aplicada a sèries temporals. Winters es va inspirar en mètodes de processament de senyals i en tècniques de filtratge recursiu, adoptant un enfocament on la suavització es repetia tres vegades. Encara que inicialment es va considerar una aproximació més experimental que teòrica, amb el temps es va consolidar com una eina més en la previsió de sèries amb patrons estacionals.
A banda de les components L_t i T_t introduïdes abans, el model assumeix que les dades tenen una estructura cíclica amb període c, S_t, de manera que la component estacional es repeteix cada cert interval de temps. L'estacionalitat pot ser de dos tipus:

- **Additiva**: les fluctuacions estacionals tenen una magnitud constant al llarg del temps.
- **Multiplicativa**: les variacions estacionals són proporcionals al nivell de la sèrie.

Segons aquesta distinció, el model de suavització exponencial triple es pot formular de dues maneres:

- Si l'estacionalitat és **additiva**, les equacions de recurrència són:

$$L_t = \alpha \cdot (z_t - S_{t-c}) + (1 - \alpha) \cdot (L_{t-1} + T_{t-1})), \tag{3.6}$$
$$T_t = \beta \cdot (L_t - L_{t-1}) + (1 - \beta) \cdot T_{t-1}, \tag{3.7}$$
$$S_t = \gamma \cdot (z_t - L_t) + (1 - \gamma) \cdot S_{t-c}. \tag{3.8}$$

- Si l'estacionalitat és de tipus **multiplicatiu**, les equacions són:

$$L_t = \alpha \cdot \left(\frac{z_t}{S_{t-c}}\right) + (1 - \alpha) \cdot (L_{t-1} + T_{t-1}), \tag{3.9}$$
$$T_t = \beta \cdot (L_t - L_{t-1}) + (1 - \beta) \cdot T_{t-1}, \tag{3.10}$$
$$S_t = \gamma \cdot \left(\frac{z_t}{L_t}\right) + (1 - \gamma) \cdot S_{t-c}, \tag{3.11}$$

on:

- γ $(0 \leq \gamma \leq 1)$ és el factor de suavització estacional.
- La resta de paràmetres són els mateixos que en els models anteriors.

Interpretació dels paràmetres

El valor dels paràmetres α, β i γ té un impacte directe en el comportament del model:

- Si γ és proper a 1, els patrons estacionals s'adapten ràpidament als canvis.
- Si γ és proper a 0, es conserva l'estacionalitat detectada en les dades històriques.

Capítol 4

Metodologies d'anàlisi i validació

Després d'haver presentat els principals models per a la predicció de sèries temporals, el pas següent és analitzar les metodologies que permeten aplicar-los de manera efectiva. La selecció i ajust dels models no només depén de la naturalesa de les dades, sinó també del procés seguit per identificar l'estructura predictiva més adequada.
En aquest capítol es descriuen dues metodologies per a la modelització de sèries temporals. En primer lloc, s'aborda la metodologia Box-Jenkins, un enfocament sistemàtic per a la construcció de models ARIMA que inclou les fases d'identificació, estimació i validació del model. Es presenten els criteris per a la selecció dels paràmetres p, d i q, així com les tècniques per avaluar l'ajust del model [5].
A continuació, s'examina la suavització exponencial, una alternativa eficient als models autoregressius en determinats contextos. Com en el cas dels models ARIMA, la qualitat de les prediccions dependrà de l'elecció òptima dels paràmetres de suavització.

4.1 Ajustament de processos ARIMA: Metodologia Box-Jenkins

La metodologia Box-Jenkins, desenvolupada per George Box i Gwilym Jenkins, representa un enfocament sistemàtic per a la modelització i predicció de sèries temporals mitjançant models ARIMA. Aquest mètode és especialment adequat per a sèries temporals amb estructures complexes que no poden ser modelades fàcilment amb tècniques més simples com la regressió o la suavització exponencial. El seu principal avantatge és la capacitat d'adaptació a diferents estructures de dades mitjançant una combinació d'autoregressió, diferenciació i mitjana mòbil, la qual cosa el fa molt potent en aplicacions econòmiques, financeres i industrials [3].
Un dels aspectes clau d'aquesta metodologia és la capacitat d'abordar sèries no estacionàries a través de la diferenciació, la qual cosa permet ampliar els models ARMA cap als ARIMA. No obstant això, els models ARIMA no tenen en compte explícitament la presència de patrons estacionals. Per a aquest tipus de dades, Box i Jenkins van introduir els models SARIMA, que incorporen components estacionals addicionals per captar millor la dinàmica de la sèrie.
El procés Box-Jenkins es compon de tres etapes principals:

1. **Identificació del model**: Determinació dels paràmetres adequats (p, d, q) a partir de l'anàlisi de l'autocorrelació i altres eines estadístiques.

2. **Estimació de paràmetres**: Ajust dels coeficients del model mitjançant tècniques com la màxima versemblança.

3. **Comprovació diagnòstica**: Avaluació dels residus per garantir que el model captura adequadament l'estructura de la sèrie.

Aquest procés és iteratiu i flexible: si un model inicial no proporciona bons resultats, es poden reavaluar els paràmetres i ajustar la identificació o l'estimació fins a trobar una configuració òptima. Aquest enfocament sistemàtic permet construir models robustos capaços de generar prediccions precises.
NOTA: La funció `auto_arima` de la llibreria `pmdarima` en Python facilita l'aplicació d'aquesta metodologia de manera automàtica, seleccionant els valors òptims de p, d i q. També permet ajustar models SARIMA, identificant els paràmetres estacionals P, D i Q més adequats. No obstant això, és recomanable complementar aquesta eina amb una avaluació manual de la sèrie, ja que la selecció òptima de paràmetres pot dependre de criteris específics no sempre captats automàticament.

4.1.1 Identificació del model

El primer pas en la metodologia Box-Jenkins és la identificació del model, que consisteix a determinar els ordres adequats dels components autoregressius (AR), de diferenciació (I) i de mitjana mòbil (MA) per a una sèrie temporal donada. Aquest procés permet establir els valors de p, d i q en un model ARIMA i, en cas de dades amb estacionalitat, també els paràmetres estacionals P, D i Q en un model SARIMA.

Comprovació d'estacionarietat i diferenciació

Abans d'ajustar un model ARIMA, és fonamental verificar si la sèrie és estacionària, és a dir, si les seues propietats estadístiques, com la mitjana i la variància, es mantenen constants al llarg del temps. Si la sèrie no és estacionària, s'aplica la diferenciació fins a aconseguir-ho. L'estacionarietat es pot avaluar de manera visual amb gràfics de la sèrie temporal o mitjançant proves estadístiques com el test de Dickey-Fuller augmentat (ADF), que s'estudiarà amb detall a la Secció 4.3.1.
L'ordre de diferenciació d es determina segons el nombre de diferències requerides per aconseguir estacionarietat. Si la diferenciació s'aplica en excés, es pot introduir soroll addicional en la sèrie, per la qual cosa és important utilitzar tant proves estadístiques com la inspecció visual.

Anàlisi de l'ACF i PACF

Al Capítol 2 es van introduir els conceptes d'autocovariància (Equació 2.1) i d'autocorrelació (Equació 2.2).
A partir d'aquestes definicions, es pot caracteritzar la funció d'autocorrelació simple (ACF), que mesura la correlació entre la sèrie temporal i les seues versions retardades en diferents *lags*. Aquesta funció permet identificar la dependència temporal de les observacions, quantificant la influència que tenen els valors passats en les observacions futures. Si la correlació és positiva, els valors elevats en un instant tendeixen a ser seguits per valors elevats en els instants successius, mentre que, si la correlació és negativa, els valors grans solen ser seguits per valors menuts. La manera més comuna de visualitzar l'ACF és mitjançant un correlograma o gràfica d'autocorrelació (Figura 4.1, esquerra). En el context de la modelització ARIMA, l'ACF resulta especialment útil per determinar l'ordre del component de

mitjana mòbil q. Una caiguda brusca de l'ACF després de pocs retards indica que les dades tenen un ordre autorregressiu finit. A més, si l'ACF mostra un patró sinusoidal o una disminució amortida, això suggereix la presència d'estacionalitat i, per tant, la necessitat de considerar ordres estacionals a més dels ordres no estacionals.
A més de l'ACF, la funció d'autocorrelació parcial (PACF) és una eina clau per identificar la dependència temporal d'una sèrie. A diferència de l'ACF, la PACF mesura la correlació entre dues observacions separades per un determinat retard o *lag*, eliminant l'efecte de les observacions intermèdies. Aquesta funció és útil per determinar l'ordre de la component autoregressiva d'un model ARIMA i es pot visualitzar en un correlograma parcial o gràfica d'autocorrelació parcial (Figura 4.1, dreta).
En el context de la modelització ARIMA, la PACF permet identificar l'ordre del terme autoregressiu p. Si la PACF es redueix dràsticament després d'un determinat retard i els valors posteriors es mantenen dins de l'interval de confiança, això indica la presència d'un model AR d'aquest ordre. Així, el retard on es produeix aquesta caiguda proporciona una estimació del valor de p.

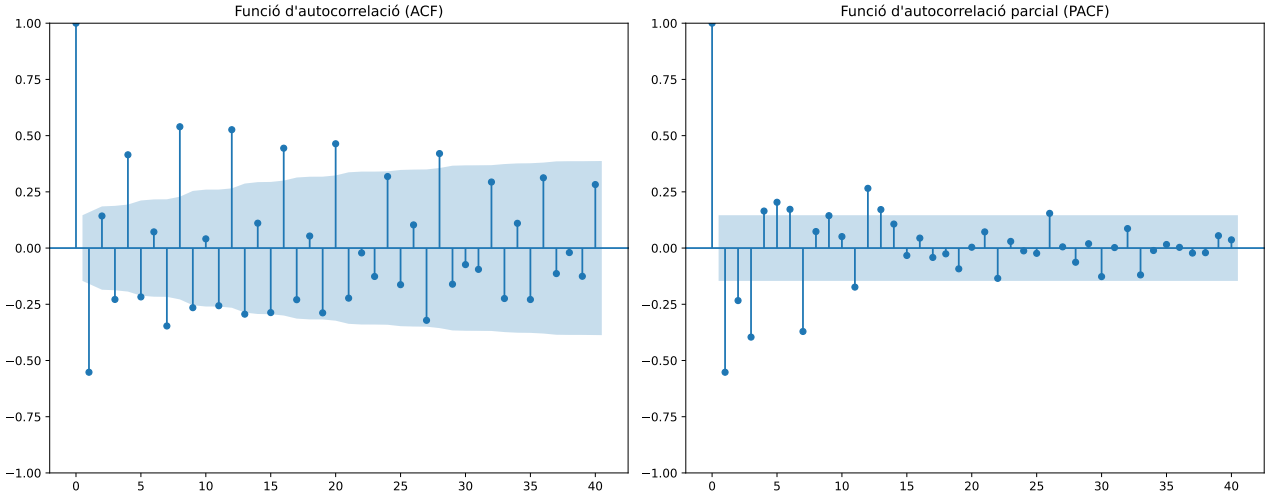

Figura 4.1: ACF i PACF d'una sèrie temporal de les hipoteques concedides a l'estat espanyol entre 2003 i 2018, per mesos.

Aprofitant la gràfica, sembla interessant comentar quines conclusions podem extraure de l'ACF i el PACF. En primer lloc, l'ACF presenta un valor extrem, suggerint un ordre de mitjana mòbil 1. A més, es pot observar un cicle que es repeteix cada quatre passes, un argument més a favor de la presència d'estacionalitat. Per la seua banda, la gràfica del PACF també mostra un valor atípic o extrem, fet que convida a pensar que l'ordre autorregressiu és 1.
Nota: Si l'anàlisi d'estacionarietat indica que cal diferenciar la sèrie, cal fer l'anàlisi posterior fent ús de la sèrie diferenciada, atés que així és com el model ARIMA interpreta la sèrie.
Per determinar els ordres p i q d'un model ARIMA, es poden seguir les regles següents:

- **Identificació de p (ordre autoregressiu, AR):**
  - La PACF presenta valors significatius fins a un retard p i després es tallen bruscament.

  - La ACF decreix gradualment sense un tall brusc.

- **Identificació de q (ordre de mitjana mòbil, MA):**
  - La ACF presenta valors significatius fins a un retard q i després es tallen bruscament.
  - La PACF decreix gradualment sense un tall definit.

- **Identificació de models ARMA(p, q):**
  - Si ni l'ACF ni la PACF presenten un tall definit, però mostren un patró de decreixement lent o oscil·latori, el model pot ser una combinació ARMA(p, q).

En un model AR(p), l'ACF decreix gradualment, mentre que la PACF es trunca en p. En canvi, en un MA(q), la PACF decreix gradualment, mentre que la ACF es trunca en q. En un model ARMA(p, q), ambdues funcions decreixen progressivament sense truncament brusc. La Taula 4.1 pot ajudar a comprendre millor les idees anteriors.

	AR(p)	**MA(q)**	**ARMA(p, q)**
ACF	Decreix gradualment	q coeficients significatius	Decreix gradualment
PACF	p coeficients significatius	Decreix gradualment	Decreix gradualment

Taula 4.1: Patrons de l'ACF i la PACF en processos AR(p), MA(q) i ARMA(p, q).

Tot i que la identificació inicial dels ordres p, d, q es basa en l'anàlisi de l'ACF i la PACF, és recomanable comparar diferents models abans de realitzar l'estimació dels paràmetres. Aquesta selecció es pot fer mitjançant criteris d'informació com el criteri d'informació d'Akaike (AIC) i el criteri d'informació bayesià (BIC), que ajuden a determinar quin model presenta un millor equilibri entre simplicitat i capacitat predictiva. Aquests criteris seran desenvolupats amb més detall a la Secció 4.3.2.
L'objectiu d'aquesta selecció és evitar models sobreajustats o massa complexos. En general, un model amb un AIC/BIC més baix és preferible, però diferències mínimes entre models poden indicar que qualsevol d'ells és una opció viable.

Detecció de l'estacionalitat

Si la sèrie temporal presenta una estructura estacional, cal tindre-la en compte en el model per tal d'evitar que la predicció es veja afectada per patrons estacionals no modelats. L'estacionalitat es manifesta quan es detecten patrons recurrents a intervals fixos, com per exemple un augment de vendes durant determinades èpoques de l'any o variacions diàries en les temperatures.
Hi ha diverses tècniques per identificar la presència d'estacionalitat en una sèrie temporal:

- **Inspecció visual**: L'anàlisi gràfica de la sèrie pot revelar cicles estacionals si es detecten pics recurrents en intervals regulars.

- **Descomposició estacional**: Aquesta tècnica permet desglossar la sèrie en tres components:
  - **Tendència** (moviment a llarg termini).

  - **Estacionalitat** (fluctuacions periòdiques).
  - **Soroll** (variabilitat aleatòria).

  L'ús de mètodes com la descomposició clàssica o STL ([7]) permet aïllar la component estacional per analitzar la seua influència.

- **Anàlisi de l'ACF**: L'estacionalitat es pot detectar mitjançant l'ACF, identificant pics significatius en retards que corresponen a la periodicitat del fenomen estacional. Per exemple, si una sèrie mensual mostra correlacions fortes en retards de 12 mesos, això suggereix una component estacional anual.

- **Tests estadístics**: Es poden utilitzar proves específiques per confirmar la presència d'estacionalitat. Un dels més comuns és el test de Kruskal-Wallis ([18]), que s'utilitza per a comparar més de dos grups independents i determinar si provenen de la mateixa distribució. Sota la hipòtesi nul·la, s'assumeix que totes les mostres procedeixen d'una mateixa distribució. Aquest test és especialment útil en el context de sèries temporals per detectar diferències estacionals entre diferents períodes de temps. Més formalment, el test de Kruskal-Wallis contrasta les hipòtesis següents:

  - H_0: Les medianes de totes les mostres són iguals, és a dir, no hi ha diferències significatives entre els grups.
  - H_1: Almenys una de les medianes és diferent, indicant que almenys un grup té una distribució diferent.

Quan es detecta estacionalitat, cal ajustar el model adequadament. En aquests casos, s'utilitza un model SARIMA, que incorpora components estacionals addicionals per capturar la dinàmica recurrent de la sèrie. Un model SARIMA es defineix pels paràmetres estacionals (P, D, Q, s), on:

- P: Nombre de termes autoregressius estacionals.
- D: Nombre de diferenciacions estacionals necessàries.
- Q: Nombre de termes de mitjana mòbil estacionals.
- s: Període de la component estacional (per exemple, 12 per a dades mensuals, 4 per a dades trimestrals, etc.).

Aquestes tècniques permeten una selecció inicial dels paràmetres estacionals, però és important destacar que el procés Box-Jenkins és iteratiu.

4.1.2 Estimació de paràmetres

Una vegada identificats els ordres p, d, q del model ARIMA, el següent pas en la metodologia Box-Jenkins és l'estimació dels paràmetres del model. Aquest procés consisteix a determinar els coeficients autoregressius (ϕ) i de mitjana mòbil (θ), així com qualsevol altre paràmetre necessari per ajustar el model a les dades observades.

L'estimació dels paràmetres d'un model ARIMA es realitza habitualment mitjançant el mètode de màxima versemblança (MLE), que troba els valors dels coeficients que maximitzen la probabilitat d'observar les dades donades. Aquest procés s'executa mitjançant optimització numèrica, la qual cosa permet ajustar els coeficients de manera eficient per obtenir el millor ajust del model.

Ajust i validació de l'estimació

Després de l'estimació dels paràmetres, cal validar els resultats per a assegurar-se que el model ajustat és fiable. Aquesta validació inclou:

- **Verificació de la significació dels coeficients**: Es comprova si els coeficients estimats són estadísticament significatius mitjançant els seus p-valors. Coeficients amb valors p alts poden indicar que no tenen una influència rellevant en el model.
- **Comparació de prediccions amb les dades reals**: S'avalua si el model és capaç de fer prediccions precises comparant els valors previstos amb els observats mitjançant mètriques d'error com RMSE o MAPE.

Automatització de l'estimació

L'estimació de paràmetres en models ARIMA pot ser automatitzada mitjançant eines computacionals. En aquest treball s'ha utilitzat `auto_arima()` de `pmdarima`, que selecciona automàticament els millors valors per a p, d, q i P, D, Q basant-se en criteris d'informació i optimització numèrica.
A més, altres biblioteques com `statsmodels` permeten ajustar models ARIMA mitjançant la funció `fit()`, que requereix definir prèviament els paràmetres del model. Tot i que `auto_arima()` és més convenient per a una selecció automàtica, l'ús de `fit()` pot ser útil quan es vol tindre un major control sobre l'estimació.

4.1.3 Comprovació diagnòstica

La fase de comprovació diagnòstica és essencial en la metodologia Box-Jenkins, ja que permet avaluar si el model ajustat és adequat. Aquesta validació es realitza principalment mitjançant l'anàlisi dels residus, que són les diferències entre els valors observats i els valors predits pel model. L'objectiu és assegurar-se que els residus es comporten com a soroll blanc, és a dir, que no presenten estructura ni correlació.

Anàlisi dels residus

Els residus han de complir els criteris següents per considerar que el model ARIMA està ben ajustat:

- **Distribució normal**: Han de seguir una distribució gaussiana amb mitjana zero.
- **Variància constant**: La dispersió s'ha de mantindre estable al llarg del temps.
- **Absència d'autocorrelació**: Els errors no han de mostrar dependència temporal.

Per verificar si compleixen les pautes anteriors, es poden utilitzar diverses gràfiques:

- **Histogrames i gràfics de densitat**: Permeten comprovar visualment si els residus segueixen una distribució normal.
- **Gràfics Q-Q**: Comparen la distribució empírica dels residus amb una distribució normal teòrica.
- **Gràfics ACF i PACF dels residus**: Ajuden a detectar si els errors encara contenen estructura temporal.

Autocorrelació i test de Ljung-Box

Un dels requisits indispensables de la validació és assegurar-se que els errors del model no presenten autocorrelació. Això es pot avaluar mitjançant:

- **Gràfics ACF i PACF dels residus**: Si els errors mostren valors significatius en diversos retards, significa que encara contenen informació que el model no ha capturat adequadament.
- **Test de Ljung-Box**: Aquest test comprova si els errors estan correlacionats en diferents *lags*. La hipòtesi nul·la del test estableix que els residus són soroll blanc, és a dir, no presenten dependència temporal. Si el p-valor del test és alt, no es pot rebutjar la hipòtesi nul·la, fet que indica que els residus són acceptablement aleatoris. Si el test de Ljung-Box indica la presència d'autocorrelació en els errors, caldrà revisar la selecció del model i considerar ajustos en els paràmetres p, d, q. Es pot consultar més informació a [20].

Sobreajustament i refinament del model

Un model ARIMA pot ser massa complex, capturant soroll aleatori de les dades d'entrenament i mostrant un rendiment deficient en dades fora de mostra. Aquest fenomen és conegut com a sobreajustament. Alguns indicadors de sobreajustament són:

- L'error en entrenament és molt baix, però l'error en validació és alt.
- Els coeficients estimats són molt grans o propers a 1, indicant una possible inestabilitat.
- L'AIC o BIC del model és molt baix, però les prediccions no són robustes.

Per evitar el sobreajustament, es poden seguir algunes estratègies:

- **Reduir la complexitat del model**: Disminuir p i q per evitar que el model siga massa flexible.
- **Ajustar millor la diferenciació**: Si la sèrie encara presenta tendència o estacionalitat, augmentar d o D.
- **Comparar diferents models**: Utilitzar criteris com l'AIC o el BIC per triar el model amb millor equilibri entre simplicitat i capacitat predictiva.

En cas de detectar problemes en la validació, es pot repetir el procés iteratiu Box-Jenkins, revisant la selecció inicial dels paràmetres i realitzant ajustos per millorar la qualitat del model.

4.2 Suavització exponencial

La suavització exponencial és una família de tècniques de modelització de sèries temporals basada en la combinació ponderada de les observacions passades, en què els valors més recents tenen una influència major en la predicció. Aquest mètode és especialment útil per a sèries temporals en què la influència dels valors històrics disminueix amb el temps, i ofereix una alternativa més flexible i adaptativa que les mitjanes mòbils tradicionals.
A diferència de la metodologia Box-Jenkins, que es basa en models autoregressius i de mitjana mòbil, la suavització exponencial es fonamenta en la idea que les prediccions han de ser una combinació progressiva de les observacions anteriors, amb un pes decreixent exponencialment. Aquest enfocament permet adaptar-se de manera eficient a les variacions estructurals de la sèrie sense necessitat de diferenciació explícita.

4.2.1 Model de suavització exponencial simple

El model de suavització exponencial simple és una tècnica de predicció utilitzada per modelar sèries temporals sense tendència ni estacionalitat. Aquest mètode assigna un pes decreixent a les observacions passades, fent que el valor més recent tinga una influència major en la predicció. A diferència d'altres enfocaments basats en mitjanes mòbils, la suavització exponencial permet una adaptació més flexible als canvis en la sèrie, preservant una estructura matemàtica senzilla i computacionalment eficient.
Per aplicar el model, cal establir un valor inicial L_0, que es pot determinar de diverses maneres:

- Fixant $L_0 = z_0$, assignant-li el primer valor de la sèrie.
- Utilitzant la mitjana dels primers N valors: $L_0 = \frac{1}{N}\sum_{i=1}^{N} z_i$.

A banda d'aquest valor inicial, el model de suavització exponencial simple requereix fixar el factor d'esmortiment $0 < \alpha < 1$, que determina la quantitat de pes que s'assigna a les observacions anteriors. Tradicionalment, el valor d'α es triava de manera que es minimitzaren els errors de predicció d'un pas:

$$\min_{\alpha} \frac{1}{n}\sum_{t=1}^{n}(z_t - \hat{z}_t)^2.$$

Si bé aquesta estratègia permet ajustar el model als valors observats, el valor de la condició inicial L_0 també pot influir en els resultats finals, especialment en sèries curtes. Per tant, una aproximació alternativa consisteix a estimar simultàniament L_0 i α:

$$\min_{L_0,\alpha} \frac{1}{n}\sum_{t=1}^{n}(z_t - \hat{z}_t)^2.$$

El model assumeix que la sèrie segueix l'estructura següent:

$$z_t = L_{t-1} + a_t,$$

on L_{t-1} representa el nivell estimat de la sèrie fins al moment $t-1$ i a_t és un terme de soroll blanc de mitjana 0 i variància σ^2, seguint una distribució normal. Atés que el model no incorpora una

tendència explícita, el nivell en l'instant t s'actualitza mitjançant una combinació ponderada de l'última observació i el nivell anterior, com mostra l'Equació 3.3.
Aquest esquema de funcionament fa que el model de suavització exponencial simple actualitze el nivell de manera successiva a mesura que es registren noves observacions. Una vegada s'han estimat L_0 i α, les prediccions per a passos futurs es mantenen constants en l'últim nivell estimat:

$$\hat{z}_{n+k} = L_n, \quad k = 1, 2, \ldots, h,$$

4.2.2 Model de suavització exponencial doble (Holt lineal)

El model de Holt és una extensió del model de suavització exponencial simple que permet capturar la tendència en la sèrie temporal. A més d'estimar el nivell, incorpora un component de tendència que s'actualitza dinàmicament a mesura que es registren noves observacions.
Com en el model simple, el nivell inicial L_0 s'estableix segons els criteris prèviament descrits. A més, és necessari definir la tendència inicial T_0, que es pot estimar de diferents maneres:

- Mitjançant la diferència mitjana entre els primers valors: $T_0 = \frac{z_c - z_0}{c}$.
- Ajustant una regressió lineal als primers c valors i utilitzant la pendent estimada.

Una vegada definides les condicions inicials, els paràmetres de suavització es determinen minimitzant els errors de predicció d'un pas:

$$\min_{\alpha,\beta,L_0,T_0} \frac{1}{n} \sum_{t=1}^{n} (z_t - \hat{z}_t)^2.$$

Aquest enfocament és equivalent a suposar que la sèrie segueix el model generatiu següent:

$$z_t = L_{t-1} + T_{t-1} + a_t,$$

on L_{t-1} i T_{t-1} representen, respectivament, el nivell i la tendència estimats en l'instant $t-1$, i a_t és un terme de soroll blanc amb mitjana 0 i variància σ^2. A diferència del model simple, on només es recalibra el nivell, ací també s'actualitza la tendència, com es mostra en les equacions d'actualització (Equació 3.4 i Equació 3.5).
Una vegada estimats els paràmetres del model (α, β) i les condicions inicials (L_0, T_0), la projecció per a futurs instants temporals incorpora la tendència i es calcula com:

$$\hat{z}_{n+k} = L_n + k \cdot T_n, \quad k = 1, 2, \ldots, h,$$

4.2.3 Model de suavització exponencial triple (Holt-Winters)

El model de Holt-Winters incorpora una component estacional, permetent modelitzar tant la tendència com la variació periòdica de la sèrie temporal. A més de definir els valors inicials de nivell i tendència, aquest model requereix establir els components estacionals inicials $S_{-c}, \ldots, S_0$, que es poden determinar de diverses maneres:

- Assignant els primers c valors com a factors estacionals normalitzats:

$$S_t = z_t - L_0 \quad \text{(cas additiu)}, \quad S_t = \frac{z_t}{L_0} \quad \text{(cas multiplicatiu)}.$$

- Estimant la mitjana de cada posició estacional en els primers cicles de la sèrie.

Una vegada establides les condicions inicials, els paràmetres de suavització es determinen minimitzant els errors de predicció d'un pas:

$$\min_{\alpha,\beta,\gamma,L_0,T_0,S_{-c},\ldots,S_0} \frac{1}{n}\sum_{t=1}^{n}(z_t-\hat{z}_t)^2.$$

Aquest enfocament és equivalent a suposar que la sèrie segueix el model generatiu següent:

$$z_t = L_{t-1} + T_{t-1} + S_{t-c} + a_t,$$

on L_{t-1}, T_{t-1} i S_{t-c} representen, respectivament, el nivell, la tendència i l'efecte estacional estimats en l'instant $t-1$, i a_t és un terme de soroll blanc amb mitjana 0 i variància σ^2. Segons les equacions d'actualització descrites prèviament, el nivell (Equació 3.6), la tendència (Equació 3.7) i l'efecte estacional (Equació 3.8) són ajustats en cada instant t emprant la informació proporcionada per z_t.
Si l'efecte estacional és multiplicatiu, el model generatiu es modifica així:

$$z_t = (L_{t-1} + T_{t-1}) \cdot S_{t-c} + a_t.$$

En aquest cas, el nivell (Equació 3.9), la tendència (Equació 3.10) i l'efecte estacional multiplicatiu (Equació 3.11) s'actualitzen de manera diferent, fent que l'efecte estacional siga més pronunciat en valors elevats de la sèrie.
Una vegada estimats els paràmetres del model (α,β,γ) i les condicions inicials, les prediccions per a instants futurs es calculen com:

$$\hat{z}_{n+k} = \begin{cases} L_n + k\cdot T_n + S_{n+k-c}, & \text{(cas additiu)} \\ (L_n + k\cdot T_n)\cdot S_{n+k-c}, & \text{(cas multiplicatiu)} \end{cases} \tag{4.1}$$

per a $k = 1, 2, \ldots, h$, on h representa l'horitzó de predicció, que sol coincidir amb la longitud del cicle estacional c.

4.3 Eines i criteris de selecció

En aquesta secció, es tractarà d'analitzar la qualitat dels diferents models estudiats als Capítol 3 i Capítol 4. En primer lloc, es presentaran algunes eines que seran d'ajut per a polir els models i fer-los més precisos. L'objectiu d'aquesta fase serà seleccionar individualment els models i sotmetre'ls a una sèrie de proves per a determinar quin d'ells és el més adequat per a la predicció de la sèrie temporal. D'altra banda, quan s'haja aconseguit una col·lecció de processos ben perfilats, el pas següent serà comparar-los per a determinar quin d'ells és el més eficaç. Tot i que existeixen moltes maneres d'enfrontar els models per prendre una decisió, en aquest capítol només se'n presentaran algunes de les més comunes. La tria d'un mètode o altre dependrà de les necessitats de l'usuari i de la naturalesa de les dades.

4.3.1 Detecció de l'estacionarietat

Un dels dubtes més freqüents en l'anàlisi de sèries temporals està relacionat amb una de les seues propietats més importants: l'estacionarietat. Com ja hem vist a la Definició 2.4.2, una sèrie temporal es considera estacionària si la seua mitjana i variància es mantenen constants en el temps i l'autocovariància no depén d'aquest, i la manera més habitual d'aconseguir-ho és mitjançant la diferenciació. No obstant això, com podem saber quan és necessari diferenciar una sèrie temporal per garantir l'estacionarietat? Per a determinar-ho, es poden utilitzar diversos tests estadístics. Entre els més comuns es troben el test de Dickey–Fuller augmentat (ADF), que comprova la presència d'una arrel unitària. A més, existeixen altres proves, com el test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS), que planteja la hipòtesi inversa, assumint que la sèrie és estacionària i provant si hi ha evidència en contra; i el test de Phillips-Perron (PP), que és una variant del test ADF amb correccions per heteroscedasticitat i autocorrelació en els errors.

Test ADF

En aquest treball, s'ha optat per utilitzar el test ADF per a comprovar l'estacionarietat de les sèries temporals, atesa la seua simplicitat i eficàcia en la majoria dels casos, fet que el converteix en una eina àmpliament emprada en aquest tipus d'anàlisi. Aquesta prova té com a hipòtesi nul·la que la sèrie no és estacionària, és a dir, que presenta una arrel unitària. Per contra, la hipòtesi alternativa afirma que la sèrie és estacionària i, per tant, no requereix més diferenciació.

$$\begin{cases} H_0 : \text{La sèrie temporal no és estacionària; és a dir, cal aplicar diferenciació.} \\ H_1 : \text{La sèrie temporal és estacionària.} \end{cases}$$

Així, si el p-valor obtingut en aquest test supera el nivell crític de significació (0.05), es dedueix que la sèrie no és estacionària i, en conseqüència, necessita ser diferenciada. Si no és el cas, es rebutja la hipòtesi nul·la i no es requereix diferenciar la sèrie.
L'estadístic ADF, utilitzat en el test, és un nombre negatiu. Com més menut és, més forta és la rebutja de la hipòtesi que hi ha una arrel unitària en algun nivell de confiança.

4.3.2 Comparació entre models

Per a avaluar la precisió i l'eficàcia dels models predictius, és necessari utilitzar mètriques d'error que quantifiquen la desviació entre les prediccions i els valors reals. Aquesta comparació permet identificar quin model s'ajusta millor a les dades segons els criteris establerts en l'anàlisi.
Entre les mètriques més utilitzades es troben l'error percentual absolut mitjà (MAPE) i les seues variants ponderades (WAPE i WMAPE), que mesuren l'error relatiu en termes percentuals. També es considera l'error quadràtic mitjà (RMSE), que dona més pes als errors grans, així com criteris d'informació com l'AIC i el BIC, que són útils per a la selecció de models.
A continuació, s'analitzen aquestes mètriques en detall, destacant els seus avantatges, limitacions i aplicacions en l'àmbit de les sèries temporals.

MAPE, WAPE i WMAPE

L'error percentual absolut mitjà (MAPE, per les seues sigles en anglés) mesura el percentatge d'error de les prediccions respecte als valors reals. Com que calcula l'error mitjà al llarg del temps (o una

altra variable), no fa distinció entre els instants; és a dir, no assigna cap preferència a certs moments. Es calcula com:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{z_t - \hat{z}_t}{z_t}\right| \times 100,$$

Una mesura similar és l'error percentual absolut pesat (WAPE, per les seues sigles en anglés), que és una versió ponderada del MAPE. Quan els valors són baixos o molt diferents, el WAPE pot oferir una millor representació de l'error global. Es calcula com:

$$\text{WAPE} = \frac{\sum_{t=1}^{n}|z_t - \hat{z}_t|}{\sum_{t=1}^{n}|z_t|} \times 100,$$

Tant el MAPE com el WAPE tracten tots els elements de la sèrie de manera uniforme, sense considerar diferències de prioritat entre moments. Per a solucionar aquest inconvenient, es pot utilitzar el MAPE ponderat (WMAPE, per les seues sigles en anglés), que assigna un pes w_t[1] a cada observació en funció de la seua importància relativa. Es calcula com:

$$\text{WMAPE} = \frac{\sum_{t=1}^{n} w_t\,|z_t - \hat{z}_t|}{\sum_{t=1}^{n} w_t\,|z_t|} \times 100,$$

Per a il·lustrar les diferències entre les mètriques MAPE, WAPE i WMAPE, a l'Exemple 4.3.3 es mostren els valors de cada mètrica aplicats a un conjunt de dades simulat. Això permet observar com cadascuna pondera els errors i com els resultats poden variar segons la metodologia utilitzada.

Exemple 4.3.3 (Comparació entre MAPE, WAPE i WMAPE)**.** *Per a comprendre millor aquestes diferències, considerem un conjunt de dades de vendes simulades i les prediccions en tres dies diferents, respectivament. La Taula 4.2 mostra els valors obtinguts per a cadascuna d'elles.*

	Venda	**Predicció**	**MAPE**	**WAPE**	**WMAPE**
Dilluns	20	25	25%		0.7
Dimarts	24	24	0%		0.2
Dimecres	4	3	25%		0.1
Total	48	52	16.67%	12.5%	25%

Taula 4.2: Comparació entre les mètriques MAPE, WAPE i WMAPE en funció de les vendes i les prediccions diàries.

És important assenyalar que la columna MAPE representa el percentatge d'error absolut entre la venda real i la predicció, calculat individualment per a cada dia. D'altra banda, la columna WMAPE mostra els pesos atorgats a cada dia en el càlcul d'aquesta mètrica, mentre que el WAPE es calcula a nivell agregat, motiu pel qual només apareix en la fila total.
A partir dels resultats de la Taula 4.2, es poden extraure diverses conclusions sobre el comportament de les mètriques analitzades.

[1] El pes assignat és arbitrari i depén únicament de la concepció que es tinga de les dades.

- *El MAPE reflecteix la diferència percentual absoluta entre la venda real i la predicció per a cada dia. Tant el dilluns com el dimecres mostren un error del 25%, mentre que el dimarts, on la predicció coincideix exactament amb la venda real, l'error és del 0%. No obstant això, com que el MAPE no pondera les observacions segons el seu pes relatiu, pot donar una percepció esbiaixada de l'error global.*

- *El WAPE, que sí que té en compte el volum total de vendes, proporciona una mesura més representativa de l'error agregat. En aquest cas, el WAPE total és del 12.5%, un valor inferior al MAPE agregat (16.67%), la qual cosa indica que els errors en dies amb menys vendes tenen un impacte menor en el resultat final.*

- *El WMAPE, que assigna pesos específics a cada dia, dona més importància als dies amb major volum de vendes. En aquest exemple, el WMAPE és del 25%, cosa que suggereix que els errors en els dies amb més vendes han tingut un pes rellevant en el resultat final.*

Aquest enfocament és especialment útil en situacions en què no tots els dies o productes tenen la mateixa importància i es vol ajustar la mètrica a aquesta realitat.

RMSE

L'error quadràtic mitjà (RMSE, per les seues sigles en anglés) és una mètrica àmpliament utilitzada per avaluar el rendiment de models predictius, especialment en l'anàlisi de sèries temporals. Representa la desviació mitjana de les prediccions respecte als valors observats, donant un pes més gran als errors elevats. Es calcula com:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i - \hat{z}_i)^2},$$

on:

- z_i són els valors observats,
- $\hat{z}_i$ són els valors predits pel model i
- N és el nombre total d'observacions.

Un RMSE més baix indica un millor ajust del model als valors observats. No obstant això, com que depén de la magnitud de les dades, és important interpretar-lo en el context específic de cada conjunt de dades. A més, penalitza els errors grans de manera més severa que altres mètriques, fet que pot reduir la seua robustesa en presència de valors extrems o *outliers*.
En l'anàlisi de sèries temporals, el RMSE és útil per comparar diferents models predictius. Per exemple, un model ARIMA pot obtindre un RMSE inferior a un SARIMA si aquest últim no ha ajustat correctament els components estacionals.
El RMSE eleva al quadrat les diferències entre els valors observats i predits abans de calcular la mitjana, cosa que amplifica els errors grans. Aquesta característica el fa especialment adequat per a models on els errors segueixen una distribució normal o gaussiana. En aquests casos, minimitzar el RMSE equival a maximitzar la versemblança del model.

Comparat amb el MAPE, el RMSE té l'avantatge de mantindre les mateixes unitats que la variable d'interés, la qual cosa facilita la interpretació de l'error en termes absoluts. Tanmateix, a diferència del MAPE, el RMSE no expressa l'error en termes percentuals i, per tant, pot ser menys intuïtiu quan es treballa amb dades de diferents escales. Mentre que el MAPE pondera els errors en funció dels valors observats i proporciona una mesura relativa, el RMSE amplifica els errors més grans, fent-lo més sensible a fluctuacions extremes.

Criteris d'informació

Després d'haver identificat l'ordre del model ARIMA, és a dir, els valors de p, d i q, cal estimar-ne els paràmetres. Quan el paquet `auto_arima` obté un model, utilitza l'estimació per màxima versemblança (MLE). Aquesta tècnica busca els valors dels paràmetres que maximitzen la probabilitat d'obtindre les dades observades.
En el cas dels models ARIMA, el MLE és equivalent a les estimacions per mínims quadrats, ja que els paràmetres es determinen minimitzant la suma dels quadrats dels residus:

$$\min \sum_{t=1}^{T} \varepsilon_t^2.$$

Per a determinar l'ordre òptim, es poden utilitzar diversos criteris d'informació. Un dels més emprats és el criteri d'informació d'Akaike (AIC), que també es fa servir en regressió lineal per a seleccionar predictors:

$$\text{AIC} = -2\log(L) + 2(p + q + k + 1),$$

on L és la versemblança de les dades, i el terme k pren el valor 1 si $c \neq 0$ i 0 en cas contrari. L'últim terme entre parèntesis representa el nombre de paràmetres del model, incloent-hi σ^2, la variància dels residus.
Quan la mostra és reduïda, es recomana utilitzar l'AIC corregit (AICc), que introdueix un terme addicional per evitar el biaix en mostres xicotetes:

$$\text{AICc} = \text{AIC} + \frac{2(p + q + k + 1)(p + q + k + 2)}{T - (p + q + k + 1) - 1}.$$

Un altre criteri d'informació rellevant és el criteri bayesià d'informació (BIC), que es defineix com:

$$\text{BIC} = \text{AIC} + [\log(T) - 2](p + q + k + 1).$$

Els models més adequats són aquells que minimitzen l'AIC, l'AICc o el BIC, tot i que l'AIC és el més utilitzat en la pràctica. No obstant això, cal destacar que aquests criteris no són adequats per a seleccionar el grau de diferenciació d, ja que aquest pas transforma les dades i altera la forma en què es calcula la versemblança. Per aquesta raó, és recomanable determinar d prèviament mitjançant anàlisis d'estacionarietat i, posteriorment, seleccionar els paràmetres p i q mitjançant l'AIC.

Capítol 5

Aplicació pràctica

El món de les matemàtiques, i en particular el de l'estadística, no tindria sentit sense la seua aplicació pràctica en la resolució de problemes reals. En aquest capítol, es presenta un cas d'ús de l'anàlisi de sèries temporals fent servir un conjunt de dades reals. No obstant això, atés que estem parlant la realitat, cal tindre en compte que l'obtenció de les dades és també un pas important a l'hora de fer un estudi estadístic. No només es tracta d'un procés més o menys tècnic de recollida, sinó que prèviament cal plantejar-se un seguit de qüestions que permeten raonar sobre els objectius i la finalitat de la feina. A més, aquesta primera tasca d'aconseguir la informació que considerem útil per a l'estudi ve seguida d'un conjunt de processos de neteja i estandardització que cal complir amb el màxim rigor possible per garantir uns resultats adequats i coherents. Per tancar el capítol, es mostren els resultats obtinguts amb els models de predicció vistos als capítols anteriors, així com la seua interpretació i la comparació entre ells.

5.1 Descripció de les dades reals

Per a la transformació del contingut teòric previ en resultats tangibles s'han utilitzat les dades del trànsit comercial nacional de l'aeroport de València entre els anys 2000 i 2019, agrupat per mesos. Aquestes dades s'han obtingut d'Aeroports Espanyols i Navegació Aèria (AENA), l'organisme encarregat de la gestió dels aeroports de l'estat espanyol. Es pot consultar la informació actual al web. La raó d'haver tallat les dades en aquest moment és perquè, a partir de l'any 2020, la pandèmia de la COVID-19 va alterar significativament els patrons de trànsit aeri, cosa que podria distorsionar els resultats de l'anàlisi. Amb això, es disposa d'un total de 240 observacions, suficients per a dur a terme un estudi estadístic detallat. Per facilitar les tasques d'anàlisi, s'ha dividit el conjunt de dades en dues parts: una d'entrenament, que comprén les dades fins a desembre de 2018, i una altra de prova, que inclou les dades mensuals de l'any 2019. Tots aquests valors estan recollits en un fitxer de tipus CSV (`passengers.csv`), que es pot llegir fàcilment amb eines com R o Python. Es poden consultar al repositori de l'autor, i també se'n pot veure un extracte a la Taula B.1.

5.2 Processos de neteja i preparació de dades

Abans de procedir a l'anàlisi de les dades, és necessari dur a terme uns passos previs d'adequació de les dades al context en què es vol treballar. La manera més senzilla d'estructurar les dades per facilitar la feina és pensar en el final. *Quina seria la millor manera de treballar amb aquests resultats?* La resposta a aquesta pregunta ajuda a dissenyar una taula simple però ben estructurada i ordenada, que permeta visualitzar de manera clara les més que possibles errades en les dades. Així, es poden identificar i corregir els valors nuls, duplicats o erronis, fent servir les estratègies més adequades per a cada cas.
En situacions com aquesta, el més usual és treballar amb les dades recollides en un fitxer CSV, tot i que en entorns de producció es podria optar per una base de dades relacional o inclús una no relacional, segons les necessitats del projecte. No obstant això, per al present, un fitxer CSV és suficient per a dur a terme les tasques necessàries.
Com ja s'ha comentat a la Secció 1.3, la feina prèvia a la cerca de models és la neteja i la preparació de les dades, que ha sigut possible gràcies al *software* KNIME.
Tot seguit, amb el fitxer CSV processat i netejat, es pot procedir a l'anàlisi de les dades i a l'aplicació dels models de predicció corresponents per mitjà de Python i les llibreries creades per a aquest propòsit.

5.3 Resultats

Com ja hem comentat al llarg dels capítols anteriors, aquest treball s'ha centrat en l'estudi d'una sèrie temporal de dades reals, amb l'objectiu de predir el comportament futur d'aquesta sèrie. Seguint la mateixa línia de treball, i just després d'haver completat els processos de neteja i preparació de les dades, el pas següent és posar en pràctica els models de predicció vistos, sense oblidar la importància d'una exploració prèvia que permeta entendre millor la sèrie i les seues característiques.

5.3.1 Anàlisi exploratòria de les dades

Atés que el concepte de sèrie temporal ha sigut el primer introduït en l'estudi, resulta coherent il·lustrar-lo amb una representació gràfica.
L'evolució del trànsit comercial nacional a l'aeroport de València entre els anys 2000 i 2019, representada a la Figura 5.1, mostra una dinàmica variable amb diferents fases de creixement i decreixement al llarg del temps.
Inicialment, entre el 2000 i el 2004, el volum de passatgers es manté relativament estable sense una tendència clara. A partir del 2004 fins al 2008, s'observa un període de creixement significatiu, possiblement influït per un augment de l'activitat econòmica i del turisme. No obstant això, aquesta expansió es veu interrompuda per una fase de decreixement entre el 2008 i el 2014, la qual podria estar relacionada amb la crisi econòmica global iniciada el 2008 i les seues repercussions en el transport aeri.
A partir del 2014, la tendència torna a revertir-se, registrant un creixement sostingut fins al final del període analitzat el 2019, possiblement per una recuperació econòmica i un augment de la connectivitat de l'aeroport.
A més d'aquestes variacions a llarg termini, la sèrie temporal presenta un patró estacional clar, amb fluctuacions recurrents que podrien correspondre a períodes de major activitat turística, com els mesos d'estiu, i mínims en mesos de menor demanda.

Aquest comportament suggereix que qualsevol modelització haurà de tindre en compte tant els canvis de tendència com l'estacionalitat per tal de realitzar prediccions precises. Més avant, a la Subsecció 5.3.2, es durà a terme una anàlisi més detallada de les components de la sèrie per caracteritzar millor aquestes variacions i entendre la seua influència en les prediccions futures.

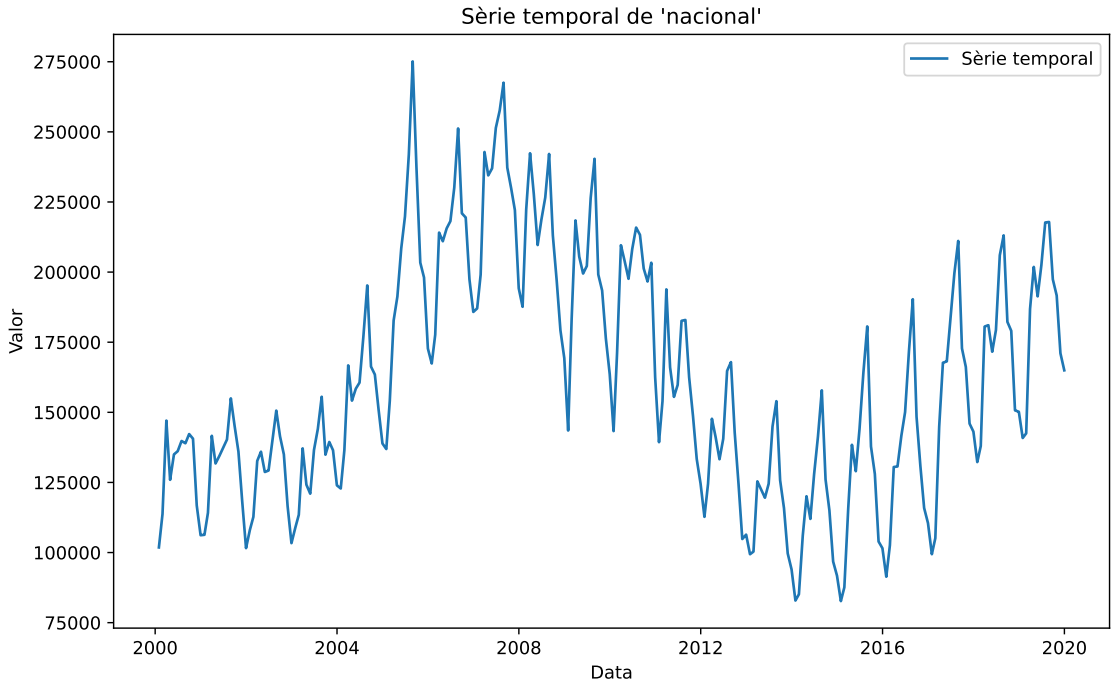

Figura 5.1: Representació de la sèrie temporal del trànsit comercial nacional a l'aeroport de València entre els anys 2000 i 2019. A l'eix x es mostra l'índex temporal, mentre que a l'eix y es representa el volum de passatgers en cada instant.

Aprofitant la introducció a les dades, és pertinent incloure una taula descriptiva amb els principals estadístics, com la mitjana, la desviació estàndard, els valors mínim i màxim, així com els percentils 25, 50 i 75.
Aquesta informació es pot trobar a la Taula 5.1, que proporciona una visió general del comportament del trànsit comercial nacional durant el període analitzat. S'observa que el volum mitjà de passatgers mensuals és de 161.008, amb una desviació estàndard de 42.259, indicant una variabilitat considerable al llarg del temps. Els valors extrems, amb un mínim de 82.678 i un màxim de 275.123 passatgers, reflecteixen possibles fluctuacions estacionals o tendències de creixement a l'aeroport. A més, els percentils mostren una distribució asimètrica, amb un 50% de les observacions per sota de 152.455 passatgers i un 75% amb menys de 195.592, suggerint la presència de pics de trànsit en determinats períodes.

	count	mean	std	min	25%	50%	75%	max
nacional	240	161008.05	42258.76	82678	130186	152455	195592	275123

Taula 5.1: Estadístiques descriptives del trànsit comercial nacional a l'aeroport de València (2000-2019).

A més, com que les dades són mensuals, pot resultar interessant analitzar la distribució del trànsit comercial nacional per mesos de l'any. La Taula 5.2 mostra la descriptiva dels passatgers mensuals, permetent identificar patrons estacionals.

Mes	count	mean	std	min	25%	50%	75%	max
Gener	20	124726.95	31195.81	82678	101218.25	117775	141452.25	187615
Febrer	20	136883.05	38204.13	85125	110812.5	130486.5	158327.25	222332
Març	20	168337.55	41686.42	106051	136033	157225	197780	242832
Abril	20	165769.75	38307.77	120078	131484.75	160094	202360.25	234473
Maig	20	163365.65	38055.09	111982	132190.75	156958	198066.5	236999
Juny	20	171553.7	38762.26	124535	137131.5	160188	203935.25	251415
Juliol	20	186619.6	39054.50	139697	144744.25	179861.5	219874.25	257596
Agost	20	198039.55	41563.38	138962	157275.5	192789.5	223498.25	275123
Setembre	20	171763.65	36882.34	125784	142050	164389.5	199678.75	237589
Octubre	20	162690.15	36016.53	115020	133932.25	156085.5	194229.5	229908
Novembre	20	146893.7	38923.82	96792	116391.75	141226.5	177020	222117
Desembre	20	135453.3	33046.63	91746	105462.5	131705.5	163950.75	194235

Taula 5.2: Estadístiques descriptives del trànsit comercial nacional a l'aeroport de València (2000-2019) desglossades per mesos.

En general, s'observa que els mesos d'estiu registren els valors més alts de trànsit de passatgers. Agost és el mes amb la mitjana més elevada (198.039 passatgers), seguit de juliol i juny. Això indica un pic estacional que probablement es correspon amb l'augment de la demanda de vols durant les vacances estivals. De fet, el valor màxim de tota la sèrie es registra a l'agost (275.123 passatgers).
D'altra banda, els mesos amb menor volum de passatgers són gener i febrer, amb mitjanes de 124.726 i 136.883, respectivament. Aquest fet suggereix que l'hivern és la temporada de menor activitat, possiblement a causa de la reducció del turisme i dels desplaçaments per oci en comparació amb els mesos d'estiu.
Les diferències entre mesos també es reflecteixen en la desviació estàndard, que és més alta als mesos d'agost i març - aquest últim, marcat per la festivitat de Falles-, la qual cosa indica una major variabilitat interanual en aquests períodes. En canvi, gener i febrer presenten desviacions més moderades, la qual cosa podria apuntar a una estabilitat relativa en el trànsit aeri durant aquests mesos.
Aquestes observacions reforcen la necessitat d'incloure la component estacional en l'anàlisi de la sèrie, fet que es desenvoluparà amb més detall al Subsecció 5.3.2.

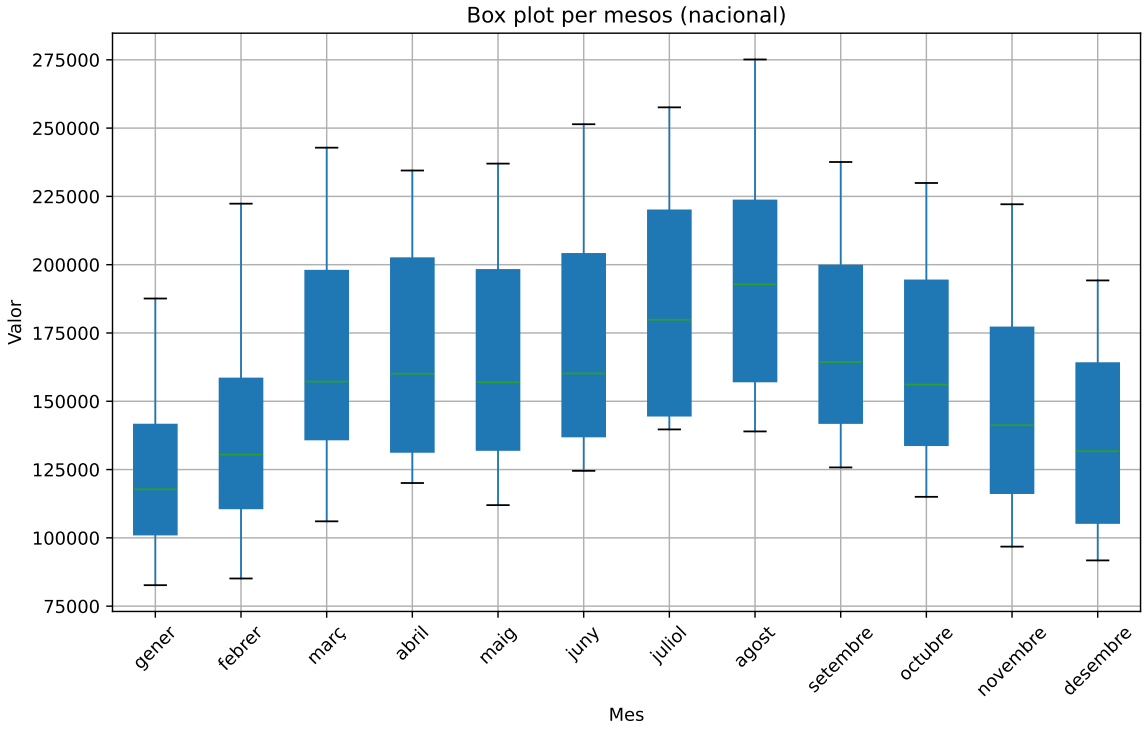

Figura 5.2: Diagrama de caixes del trànsit comercial nacional mensual a l'aeroport de València entre els anys 2000 i 2019, mostrant la distribució dels passatgers per mes.

La Figura 5.2 presenta el diagrama de caixes del trànsit comercial nacional mensual, permetent visualitzar la distribució i la variabilitat dels passatgers per cada mes de l'any.
S'observa clarament la presència d'una estacionalitat marcada, amb els mesos d'estiu (juliol i agost) registrant les medianes més altes i una dispersió considerable, fet que confirma el pic de trànsit en aquests mesos. Agost destaca especialment com el mes amb el màxim registre de passatgers i una elevada variabilitat interanual.
En canvi, els mesos d'hivern, especialment gener i febrer, presenten les medianes més baixes i un rang interquartílic més reduït, indicant una menor activitat en aquest període. A més, es pot notar que aquests mesos tenen menys variabilitat respecte als d'estiu, la qual cosa suggereix una demanda més estable en temporada baixa.
Setembre i octubre mostren una transició entre l'alt trànsit de l'estiu i el descens cap als mesos d'hivern, mentre que març i abril també presenten valors elevats, possiblement a causa de períodes vacacionals com la Setmana Santa.
Aquest patró reforça la importància d'incloure una component estacional en qualsevol model predictiu, ja que les variacions estacionals tenen un impacte significatiu en el comportament del trànsit comercial nacional.

5.3.2 Descomposició de la sèrie temporal

Amb les dades ja preparades i una visió més clara de la sèrie, procedim a descompondre-la atenent el que s'ha vist a la Secció 2.3.
En aquest estudi, suposarem que la sèrie temporal segueix un model additiu, on la tendència, l'estacionalitat i el component residual es combinen de manera lineal. Aquesta decisió es basa en l'observació que l'amplitud de la variabilitat estacional és relativament constant al llarg del temps, cosa que suggereix que no depén del nivell de la sèrie. En cas contrari, un model multiplicatiu podria ser més adequat, ja que assumiria que les fluctuacions són proporcionals al valor de la sèrie. Si la variabilitat canviara de manera no uniforme, un model mixt també podria ser considerat.
La Figura 5.3 mostra aquesta descomposició, permetent visualitzar com cada component contribueix a la sèrie global. Aquesta separació resulta essencial per entendre millor l'estructura de les dades i facilitar una modelització més precisa.

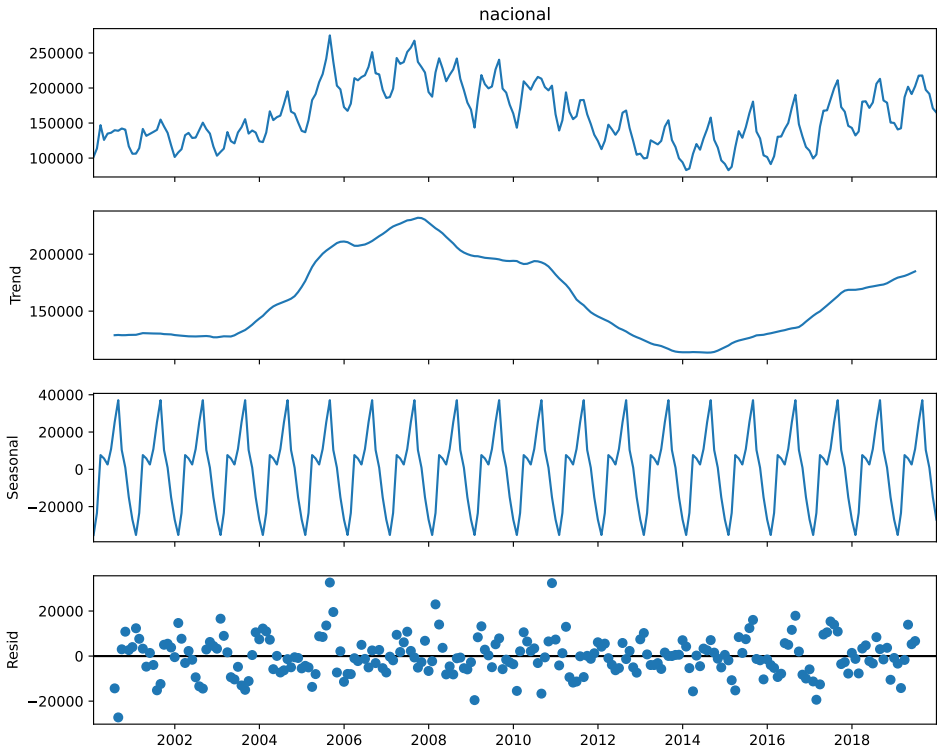

Figura 5.3: Descomposició de la sèrie temporal del trànsit comercial nacional mensual a l'aeroport de València entre els anys 2000 i 2019. La primera gràfica mostra la sèrie original; la segona, la tendència (*Trend*); la tercera, l'estacionalitat (*Seasonal*); i la quarta, el soroll blanc (*Resid*), que recull les variacions no explicades per les altres components.

Tendència

La tendència de la sèrie temporal no presenta un comportament clarament definit, ja que experimenta diverses fases diferenciades al llarg del temps.
Inicialment, es manté relativament constant, però poc després comença un període de creixement sostingut, possiblement a causa d'un efecte de crida o *boom* en el trànsit aeri nacional. No obstant això, aquest augment es veu interromput per un descens pronunciat al voltant del 2008, coincidint amb la crisi econòmica global, que va afectar significativament el sector del transport aeri.
Després d'aquest període de contracció, la tendència mostra signes de recuperació progressiva en els últims anys de la sèrie, indicant una possible estabilització o retorn als nivells anteriors. Malgrat això, el comportament general de la tendència no és prou estable per a ser modelitzat fàcilment mitjançant tècniques tradicionals com les vistes a la Secció 2.3.
Si s'ampliara la finestra temporal uns anys més, podria observar-se un patró més clar, possiblement ajustable a una funció polinòmica o un model més elaborat que capture millor els canvis estructurals en la sèrie.

Estacionalitat

L'estacionalitat de la sèrie temporal mostra un patró cíclic clar que es repeteix de manera regular. Per identificar el període d'aquesta estacionalitat, resulta convenient utilitzar un diagrama de caixes com el de la Figura 5.2, que confirma la variabilitat mensual del trànsit aeri nacional.
Com era d'esperar, els mesos amb més volum de passatgers corresponen a l'època estival, destacant especialment juliol i agost, quan es registra el màxim trànsit de l'any. Per contra, els mesos de desembre, gener i febrer presenten una reducció significativa en el nombre de desplaçaments, tot i que al desembre la caiguda no és tan acusada a causa de les festivitats nadalenques.
Aquest efecte vacacional és una particularitat del trànsit aeri que no sempre queda ben capturat pels models clàssics, atés que les fluctuacions poden variar lleugerament d'un any a un altre en funció de les dates específiques de les festivitats i períodes vacacionals. Per abordar aquesta qüestió, s'han desenvolupat metodologies específiques, com el model Prophet, que permeten una gestió més flexible d'aquest tipus d'estacionalitat. A la Secció 6.1, es farà una introducció a aquest model i el seu enfocament en la predicció de sèries temporals amb estacionalitat complexa.

Soroll blanc

Per últim, el soroll blanc es pot observar en les fluctuacions aleatòries que no segueixen cap patró discernible. Aquesta component recull la variabilitat de la sèrie que no ha pogut ser explicada per la tendència ni per l'estacionalitat, representant així els elements imprevisibles del sistema.
En aquest treball, suposarem que el soroll blanc és una seqüència de variables aleatòries normals, independents i idènticament distribuïdes amb mitjana zero i variància constant, és a dir,

$$a_t \sim N(0, \sigma^2).$$

L'anàlisi del soroll blanc és una etapa important per avaluar la qualitat de la descomposició de la sèrie. Els residus resultants després de modelitzar la tendència i l'estacionalitat haurien de complir les propietats esmentades a la Secció 4.1.3: normalitat, mitjana zero, variància constant i manca d'autocorrelació. Si es detectara autocorrelació en aquests residus, significaria que encara hi ha

estructura en les dades que el model no ha capturat adequadament, fet que indicaria la necessitat d'ajustar-lo o incorporar components addicionals.
Els tests estadístics realitzats mostren que els residus no s'ajusten perfectament a una distribució normal. En concret, el test de Jarque-Bera ha retornat un p-valor de 0.0001 amb els següents valors:

- Asimetria (*skewness*): 0.3699
- Excés de curtosi: 4.2259

Aquests valors indiquen una lleugera asimetria cap a la dreta i una major presència de valors extrems respecte a una distribució normal. Per la seua banda, el test de Shapiro-Wilk ha donat un p-valor de 0.0176, també suggerint una desviació respecte a la normalitat.
A la Figura 5.4, es pot observar l'histograma dels residus juntament amb una distribució normal ajustada. S'aprecia que els residus segueixen una distribució en forma de campana, tot i que les cues són més llargues del que es podria esperar en una distribució normal perfecta, fet que concorda amb la curtosi elevada detectada en el test de Jarque-Bera.

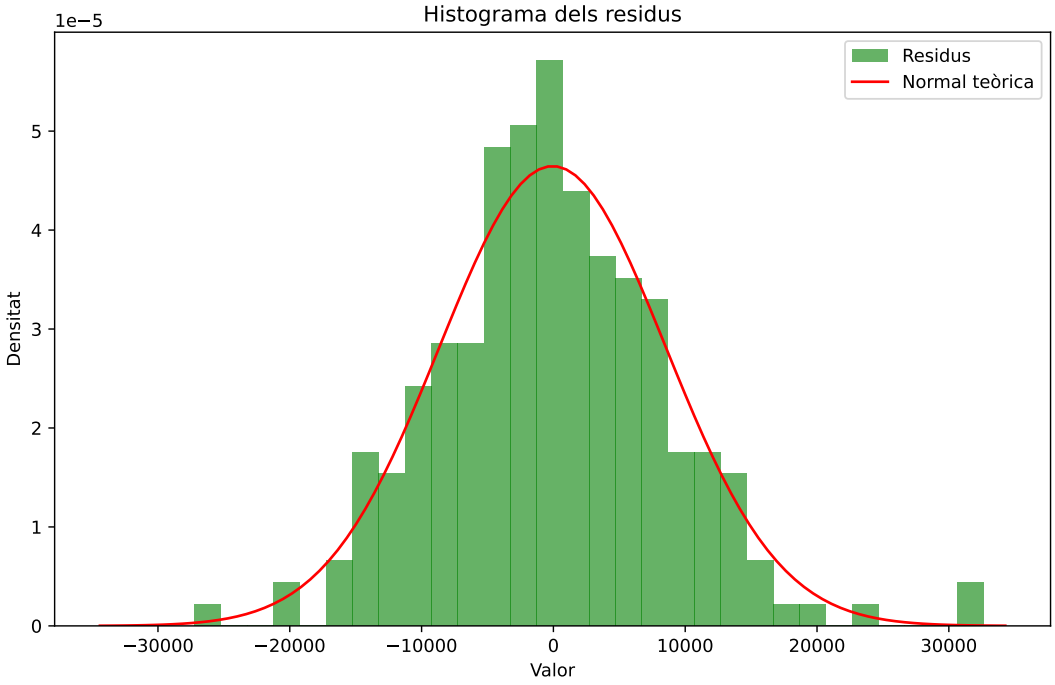

Figura 5.4: Histograma dels residus de la sèrie temporal del trànsit comercial nacional mensual a l'aeroport de València (2000-2019) amb ajust a una distribució normal.

D'altra banda, la Figura 5.5 mostra la gràfica Q-Q, que permet analitzar visualment la normalitat dels residus. Els punts segueixen aproximadament la línia central, encara que hi ha desviacions en les cues, la qual cosa confirma l'excés de curtosi observat en els tests estadístics.

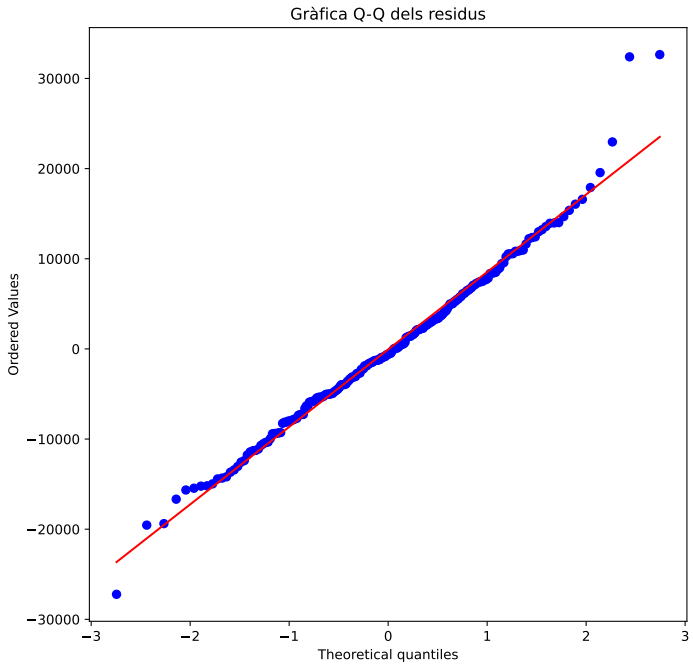

Figura 5.5: Gràfica Q-Q dels residus de la sèrie temporal del trànsit comercial nacional mensual a l'aeroport de València (2000-2019).

Malgrat aquestes desviacions, assumirem que els residus són aproximadament normals, ja que no s'ha detectat cap estructura evident que requerisca una modelització addicional. No obstant això, en futurs estudis podria ser interessant analitzar si aquestes desviacions tenen alguna causa específica, com la influència de factors exògens o la necessitat d'incloure una altra component en el model.

Mesura del pes de la tendència i l'estacionalitat

Per a quantificar la influència de la tendència i l'estacionalitat en la sèrie temporal analitzada, s'han calculat les corresponents mesures de força. Aquests valors permeten determinar en quin grau aquestes components expliquen la variabilitat de la sèrie.
Per quantificar aquests indicadors, s'han fet servir l'Equació 2.3 i l'Equació 2.4, respectivament. Els resultats obtinguts són els següents:

- **Força de la tendència**: 0.95.
- **Força de l'estacionalitat**: 0.85.

La tendència és especialment dominant, amb un pes del 95%, fet que suggereix que el trànsit nacional a l'aeroport de València ha seguit una evolució sostinguda al llarg del temps. Així mateix, la força de

l'estacionalitat, amb un valor del 85%, reflecteix la presència d'un patró periòdic clar, amb fluctuacions recurrents atribuïbles a la demanda estacional del transport aeri.
Aquestes mesures són especialment útils per comprendre millor la naturalesa de la sèrie i permetran, en anàlisis posteriors, relacionar-les amb el rendiment dels diferents models de predicció que s'aplicaran. En particular, es calcularà el coeficient de determinació R^2 per a cada model, la qual cosa permetrà avaluar fins a quin punt la tendència i l'estacionalitat contribueixen a la capacitat predictiva dels models.

Detecció de l'estacionarietat i transformacions per aconseguir-la

Per a determinar si la sèrie temporal és estacionària, s'han aplicat proves estadístiques que analitzen la presència d'arrels unitàries. En concret, s'han utilitzat el test de Dickey-Fuller augmentat (ADF) per a la diferenciació regular, vist a la Secció 4.3.1; i el test de Kruskal-Wallis per a la diferenciació estacional, introduït a la Secció 4.1.1.
Inicialment, amb les dades originals ($d = 0$), s'ha obtingut un p-valor de 0.6097, fet que indica que la sèrie no és estacionària i, per tant, requereix diferenciació. Després d'aplicar una diferenciació primera ($d = 1$), el p-valor ha descendit fins a 0.01, la qual cosa permet rebutjar la hipòtesi nul·la i concloure que la sèrie diferenciada és estacionària. D'altra banda, el test de Kruskal-Wallis ha retornat un p-valor de $5.312 * 10^{-8}$ amb $D = 0$, la qual cosa ens porta a rebutjar H_0. ixò indica que hi ha diferències significatives entre els grups comparats. Per contra, amb $D = 1$ s'ha obtingut un p-valor de 0.999, la qual cosa significa que no hi ha evidència per rebutjar H_0. Això suggereix que els grups tenen distribucions similars i que no hi ha diferències significatives.
En resum, els valors òptims de diferenciació són:

- Diferenciació regular: $d = 1$.
- Diferenciació estacional: $D = 1$.

Els resultats indiquen que la sèrie requereix una diferenciació d'ordre 1 per eliminar la tendència i aconseguir estacionarietat, a més d'una diferenciació estacional d'ordre 1 per eliminar l'estacionalitat. Per a complementar aquesta anàlisi, s'han representat les funcions d'autocorrelació (ACF) i autocorrelació parcial (PACF) de la sèrie original. La Figura 5.6 mostra com la funció d'autocorrelació decreix de manera lenta en la sèrie original, un patró típic d'una sèrie no estacionària. Per contra, després d'aplicar la diferenciació, les correlacions disminueixen ràpidament, suggerint que la sèrie diferenciada presenta propietats més pròximes a l'estacionarietat.

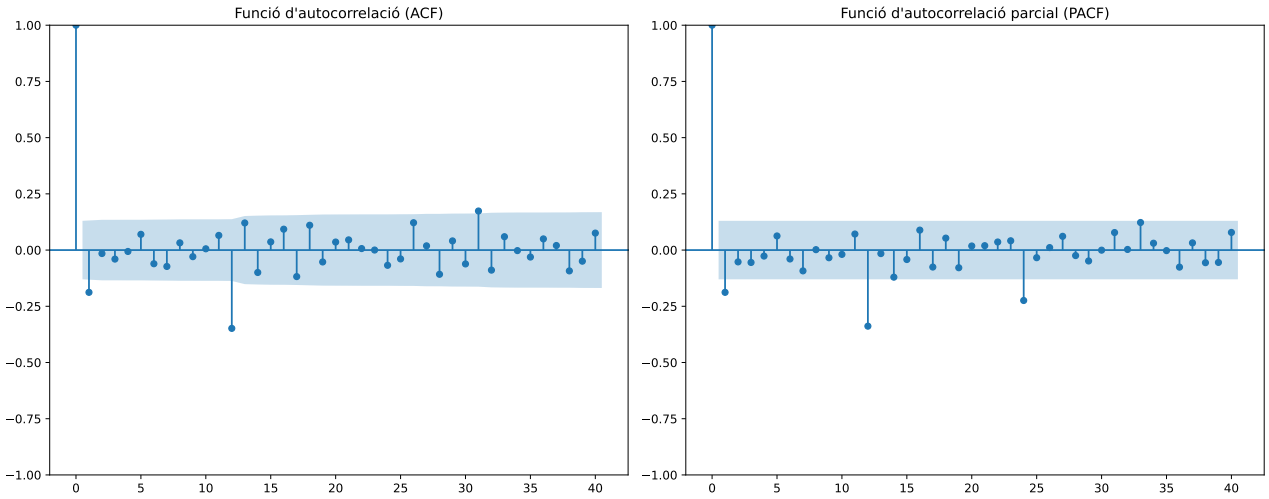

Figura 5.6: Funcions d'autocorrelació i d'autocorrelació parcial de la sèrie temporal del trànsit comercial nacional mensual a l'aeroport de València (2000-2019).

5.3.3 Models de predicció

Amb la sèrie temporal descomposta i les dades preparades, es pot procedir a l'aplicació dels models de predicció vistos en els capítols anteriors. Seguint els principis de modularitat i escalabilitat del codi, s'han implementat les funcions necessàries per a l'ajust i la predicció de cada model en un mòdul específic, que es pot importar fàcilment en qualsevol entorn de treball. Els mòduls implementats, juntament amb les funcions corresponents, es poden trobar al directori `models/` del repositori de l'autor. Cadascun d'aquests mòduls conté les funcions necessàries per a l'ajust i la predicció dels models corresponents, així com els paràmetres òptims obtinguts en l'etapa de selecció de models. Els models que s'han implementat són el Holt-Winters i l'ARIMA, tot i que aquest últim s'ha ajustat de dues maneres diferents: mitjançant la selecció manual dels paràmetres (Secció 5.3.3) i amb l'ajuda de la funció `auto_arima` (Secció 5.3.3).

ARIMA

La funció `ajustar_arima` (Secció A.1) té com a objectiu trobar el millor model ARIMA per a una sèrie temporal donada mitjançant una cerca exhaustiva sobre un conjunt de possibles paràmetres. Aquesta cerca es basa en la minimització de l'AIC, que permet seleccionar el model òptim en termes de complexitat i ajust a les dades.
El procés d'ajustament es duu a terme seguint els passos següents:

1. **Generació de combinacions de paràmetres:** Es defineixen intervals per als paràmetres del model ARIMA (p, d, q) i per als components estacionals (P, D, Q), incloent-hi la periodicitat estacional m. A partir d'aquests intervals, es generen totes les combinacions possibles de paràmetres.

2. **Estimació del model:** Per cada combinació (p, d, q) i (P, D, Q), es construeix un model ARIMA i s'ajusta a les dades d'entrenament.

3. **Càlcul de l'AIC:** Després d'ajustar el model, es calcula el valor de l'AIC. Aquest valor és un indicador de la qualitat del model, penalitzant aquells que tenen més paràmetres per evitar sobreajustament.

4. **Selecció del millor model:** Es comparen els valors d'AIC de tots els models ajustats, i es guarda el model amb el menor valor d'AIC com el millor model.

Durant l'execució, la funció mostra en pantalla els resultats de cada model ajustat, indicant els paràmetres utilitzats, el valor de l'AIC obtingut i el temps emprat en l'estimació. Finalment, es retorna el model òptim seleccionat, mostrat a l'Figura 5.7.
Aquest procés permet una selecció sistemàtica i objectiva del model ARIMA més adequat per a la sèrie temporal, assegurant que s'obté un bon equilibri entre la simplicitat i la capacitat predictiva.

Resultats d'ARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	228
Model:	SARIMAX(0, 2, 1)x(1, 1, 1, 12)	Log Likelihood	-2266.153
Date:	Mon, 10 Feb 2025	AIC	4542.306
Time:	11:55:44	BIC	4559.136
Sample:	01-31-2000	HQIC	4549.106
	- 12-31-2018		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	5.7994	4.777	1.214	0.225	-3.563	15.162
ma.L1	-0.9969	0.078	-12.716	0.000	-1.151	-0.843
ar.S.L12	0.2678	0.120	2.223	0.026	0.032	0.504
ma.S.L12	-0.6934	0.102	-6.794	0.000	-0.893	-0.493
sigma2	9.126e+07	1.53e-07	5.95e+14	0.000	9.13e+07	9.13e+07

Ljung-Box (L1) (Q):	3.82	Jarque-Bera (JB):	1.72
Prob(Q):	0.05	Prob(JB):	0.42
Heteroskedasticity (H):	0.59	Skew:	0.01
Prob(H) (two-sided):	0.03	Kurtosis:	3.44

Figura 5.7: Resum estadístic del model ARIMA ajustat a la sèrie temporal del trànsit comercial nacional mensual a l'aeroport de València (2000-2019).

El model seleccionat incorpora una diferenciació d'ordre 2 en la component no estacional, fet que suggereix que la sèrie necessitava ser diferenciada dues vegades per aconseguir estacionarietat. A més,

presenta components estacionals amb un terme autoregressiu i un terme de mitjana mòbil estacional, la qual cosa indica la presència d'una estructura repetitiva al llarg del temps.
L'anàlisi dels coeficients mostra que la majoria són significatius, ja que els seus p-valors són inferiors a 0.05, fet que confirma la seua rellevància en la modelització de la sèrie temporal. No obstant això, el p-valor associat a l'*intercept* és de 0.225, fet que suggereix que aquest paràmetre podria no ser estrictament necessari des d'un punt de vista estadístic. Tot i això, darrere d'aquesta aparent incongruència hi ha la sospita que puga minimitzar una mica els criteris d'informació, indicant un millor compromís entre simplicitat i capacitat predictiva del model.
Amb això, si recordem l'equació general dels models SARIMA (Equació 3.2), el model òptim obtingut es pot expressar com:

$$\left(1-\Phi_1 B^{12}\right)\left(\nabla^2\nabla_{12}^1 z_t - I\right) = \left(1-\theta_1 B^1\right)\left(1-\Theta_1 B^{12}\right) a_t$$
$$\left(1-0.2678 B^{12}\right)\left(\nabla^2\nabla_{12}^1 z_t - 5.7994\right) = \left(1-(-0.9969) B^1\right)\left(1-(-0.6934) B^{12}\right) a_t$$

Pel que fa a l'avaluació del model, el test de Ljung-Box proporciona un valor de Q de 3.82 amb una probabilitat associada de 0.05. Això permet deduir que, encara que els residus es comporten pràcticament com soroll blanc, hi ha una certa evidència que podria quedar alguna estructura no modelada. Aquest fet no implica necessàriament que el model siga inadequat, però suggereix que es podrien explorar lleugeres variacions en la seua especificació per avaluar possibles millores.
Una vegada vistes les característiques del model ARIMA, procedirem a analitzar-ne el rendiment en la sèrie temporal d'estudi. La Figura 5.8 mostra la predicció completa generada pel model, diferenciant clarament les dades d'entrenament, el conjunt de test i la predicció.

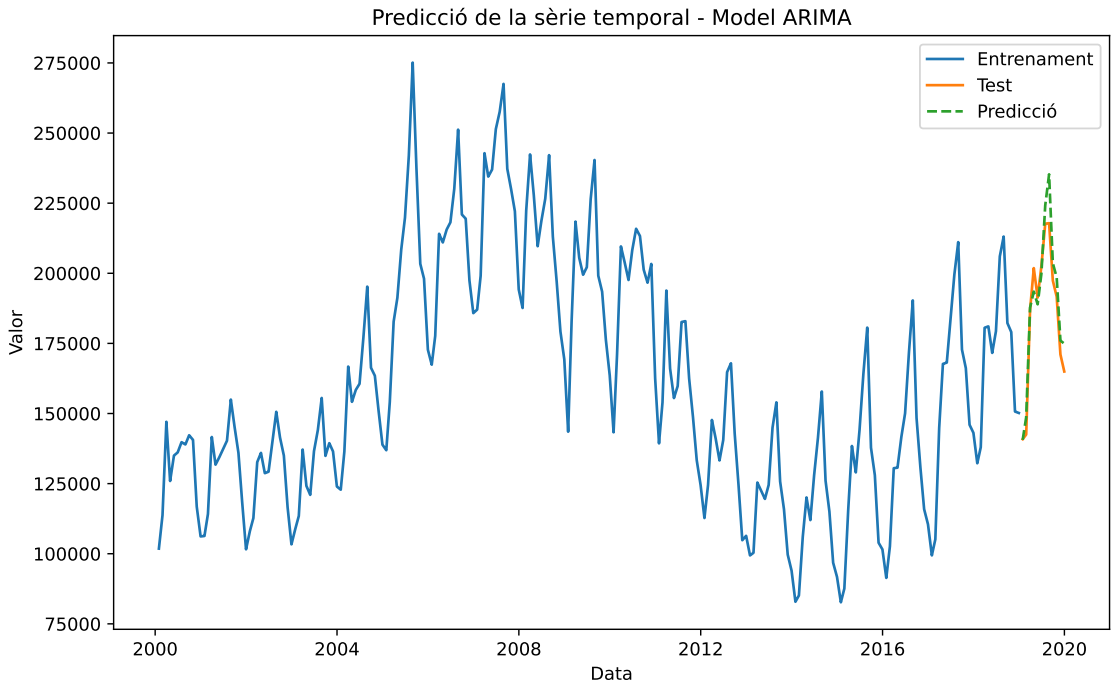

Figura 5.8: Predicció del trànsit comercial nacional mensual a l'aeroport de València (2000-2019) amb el model ARIMA.

Per quantificar l'exactitud del model, s'ha elaborat una taula comparativa que recull els valors reals, les prediccions d'ARIMA, la diferència absoluta i l'error percentual. Aquesta informació permet identificar patrons en les desviacions del model i determinar en quines situacions funciona millor o presenta majors discrepàncies.

L'anàlisi de la Taula 5.3 revela que l'error percentual varia entre valors propers al 0.16% i desviacions més significatives, com el cas d'agost de 2019, on l'error arriba al -8.00%. Aquesta discrepància pot indicar que ARIMA no ha capturat completament un efecte particular d'aquell període, com un canvi atípic en la tendència o una influència no modelitzada, com esdeveniments no recurrents o efectes econòmics inesperats. Per contra, en mesos com març i juny, l'error és inferior al 1.5%, fet que suggereix que el model aconsegueix ajustar-se bé a la dinàmica estacional de la sèrie.

Data	Valor real	Predicció	Diferència	Error (%)
2019-01	140839	140609	230	0.16
2019-02	142500	148681	-6181	-4.34
2019-03	186876	187740	-864	-0.46
2019-04	201851	193478	8373	4.15
2019-05	191335	188929	2406	1.26
2019-06	202493	199819	2674	1.32
2019-07	217684	224217	-6533	-3.00
2019-08	217863	235286	-17423	-8.00
2019-09	197404	203605	-6201	-3.14
2019-10	191649	198495	-6846	-3.57
2019-11	171059	175991	-4932	-2.88
2019-12	164973	174912	-9939	-6.02

Taula 5.3: Comparació entre els valors reals i les prediccions del model ARIMA per al trànsit comercial nacional a l'aeroport de València (2019).

La Figura 5.9 il·lustra la comparació entre els valors reals i les prediccions, facilitant la identificació de desviacions significatives.

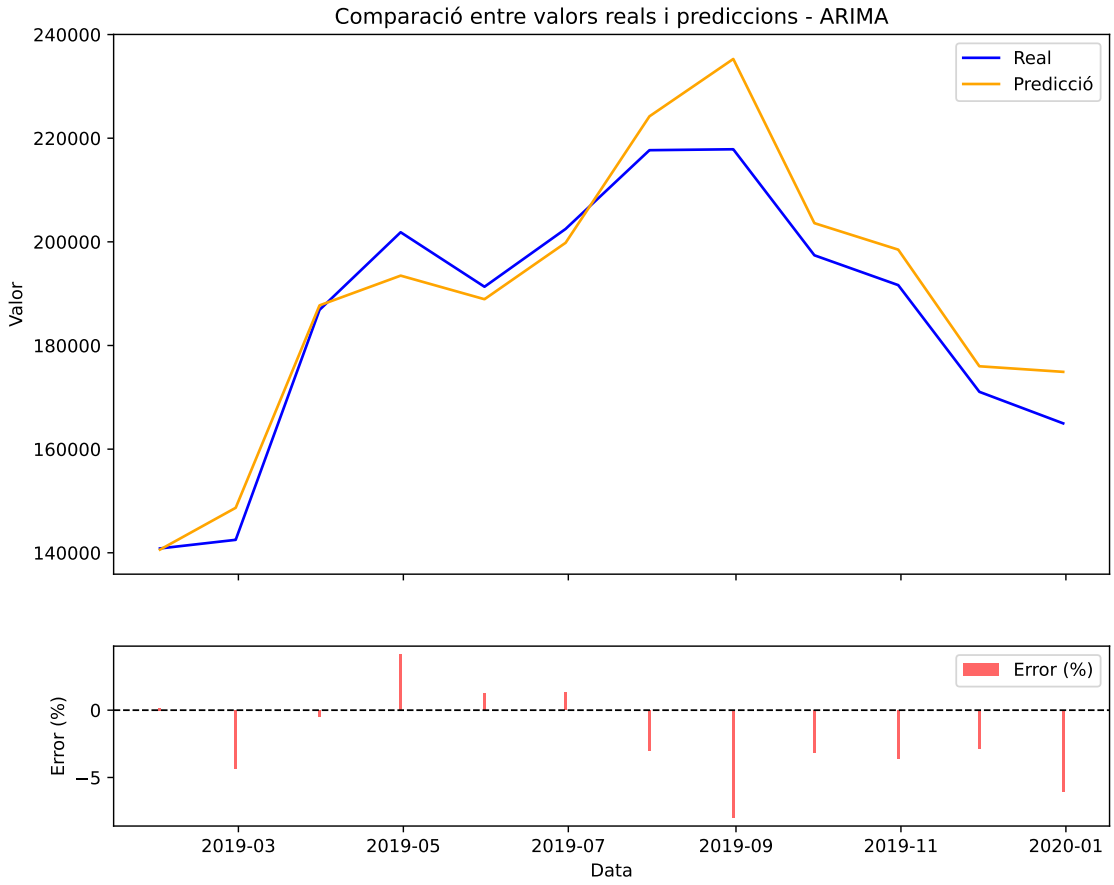

Figura 5.9: Comparació entre els valors reals i les prediccions del model ARIMA per al trànsit comercial nacional mensual a l'aeroport de València (2000-2019).

En general, el model ARIMA ha aconseguit capturar les tendències generals i l'estacionalitat de la sèrie, però en determinats punts presenta desviacions més marcades. Aquestes diferències podrien ser analitzades en comparació amb altres models per determinar quin enfocament proporciona millors resultats en termes d'error global i estabilitat predictiva.

AUTO_ARIMA

El model ARIMA també es pot ajustar automàticament mitjançant la funció `auto_arima` de la llibreria `pmdarima`, que optimitza la selecció dels paràmetres basant-se en el criteri d'informació d'Akaike (AIC). Aquest enfocament evita la necessitat d'explorar manualment combinacions de paràmetres, fent que el procés siga més eficient i adaptatiu a la naturalesa de la sèrie temporal.
La funció `ajustar_auto_arima`, definida al mòdul `auto_arima` inclòs a la Secció A.2, implementa aquest

procés explorant diferents valors per als paràmetres d'integració (d, D) i permetent una optimització automàtica dels valors (p, q) i (P, Q).
El procés d'ajustament es realitza seguint els passos següents:

1. **Selecció dels valors de diferenciació:** Es prova cada combinació possible per als paràmetres d i D fins a un valor màxim especificat.

2. **Ajust dels models:** Per cada combinació de diferenciació, es genera un model ARIMA amb diferents valors de (p, q) i (P, Q) dins dels límits definits.

3. **Càlcul de l'AIC:** Cada model ajustat s'avalua en funció del seu criteri d'informació d'Akaike (AIC).

4. **Selecció del millor model:** Es compara l'AIC de cada model i es guarda el que minimitza aquest valor.

A diferència de la selecció manual d'ARIMA (Secció 5.3.3), aquest mètode optimitza la cerca dels millors valors dels paràmetres estacionals i no estacionals sense necessitat d'intervenció manual. A més, el procés incorpora tècniques de validació creuada interna per garantir que els paràmetres seleccionats siguen estables i robustos.
Aquest enfocament és especialment útil per a sèries temporals on la diferenciació òptima no és evident o quan es vol una exploració més extensa de les possibles configuracions. En la secció següent, s'analitzen els resultats obtinguts amb aquest mètode.
A la Figura 5.10 es pot revisar el resum estadístic del model seleccionat per la funció `auto_arima`.

Resultats d'AUTO ARIMA

SARIMAX Results

Dep. Variable:	y	No. Observations:	228
Model:	SARIMAX(1, 1, 0)x(1, 2, [1], 12)	Log Likelihood	-2206.959
Date:	Mon, 10 Feb 2025	AIC	4421.919
Time:	14:42:59	BIC	4435.171
Sample:	01-31-2000	HQIC	4427.280
	- 12-31-2018		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.1245	0.147	-0.845	0.398	-0.413	0.164
ar.S.L12	-0.2835	0.160	-1.770	0.077	-0.597	0.030
ma.S.L12	-0.8584	0.198	-4.335	0.000	-1.247	-0.470
sigma2	2.875e+08	3e-10	9.58e+17	0.000	2.88e+08	2.88e+08

Ljung-Box (L1) (Q):	0.12	Jarque-Bera (JB):	1.55
Prob(Q):	0.72	Prob(JB):	0.46
Heteroskedasticity (H):	0.50	Skew:	-0.02
Prob(H) (two-sided):	0.01	Kurtosis:	3.43

Figura 5.10: Resum estadístic del model AUTO ARIMA ajustat a la sèrie temporal del trànsit comercial nacional mensual a l'aeroport de València (2000-2019).

El model seleccionat per `auto_arima` és un SARIMA(1,1,0)(1,2,1)[12] amb un valor d'AIC de 4421.919, fet que indica una millor adequació a les dades en comparació amb el model SARIMA seleccionat manualment (Secció 5.3.3), que tenia un AIC de 4542.306. Aquesta diferència en l'AIC suggereix que el model determinat automàticament ofereix una millor capacitat predictiva mantenint una complexitat més ajustada.
L'estructura del model mostra que la diferenciació aplicada és d'ordre 1 en la component no estacional i d'ordre 2 en la component estacional. Això indica que, si bé la sèrie necessitava una única diferenciació per aconseguir estacionarietat en l'àmbit general, la component estacional requeria una correcció més forta per eliminar patrons anuals més pronunciats. A més, el model inclou un únic terme autoregressiu (AR(1)), amb un coeficient de -0.1245 i un p-valor de 0.398, fet que suggereix un impacte reduït d'aquesta component en la predicció. Pel que fa als termes estacionals, es compta amb un terme autoregressiu (AR(12)) amb un coeficient de -0.2835 i un p-valor de 0.077, la qual cosa indica que podria tindre certa rellevància malgrat no ser significatiu al 5%. En canvi, el terme de mitjana mòbil estacional (MA(12)) presenta un coeficient de -0.8584 i un p-valor pràcticament nul, confirmant així la seua importància en la modelització de la sèrie.

L'equació del model es pot expressar com:

$$\left(1-\phi_1 B^{12}\right)\left(1-\Phi_1 B^{12}\right)\left(\nabla^1\nabla_{12}^2 z_t\right) = \left(1-\Theta_1 B^{12}\right) a_t$$
$$\left(1-(-0.1245) B^{12}\right)\left(1-(-0.2835) B^{12}\right)\left(\nabla^1\nabla_{12}^2 z_t\right) = \left(1-(-0.8584) B^{12}\right) a_t$$

L'anàlisi dels residus proporciona indicadors que suggereixen una bona qualitat del model. La prova de Ljung-Box dona un valor $Q = 0.12$ amb una probabilitat associada de $p = 0.72$, indicant que no hi ha evidència d'autocorrelació significativa en els residus. Això suggereix que el model ha capturat bé l'estructura de la sèrie i que els errors es comporten de manera aleatòria.
Amb el model seleccionat per `auto_arima`, s'ha generat la predicció sobre la sèrie temporal, diferenciant clarament les dades d'entrenament, el conjunt de test i la predicció final. La Figura 5.11 mostra aquesta predicció.

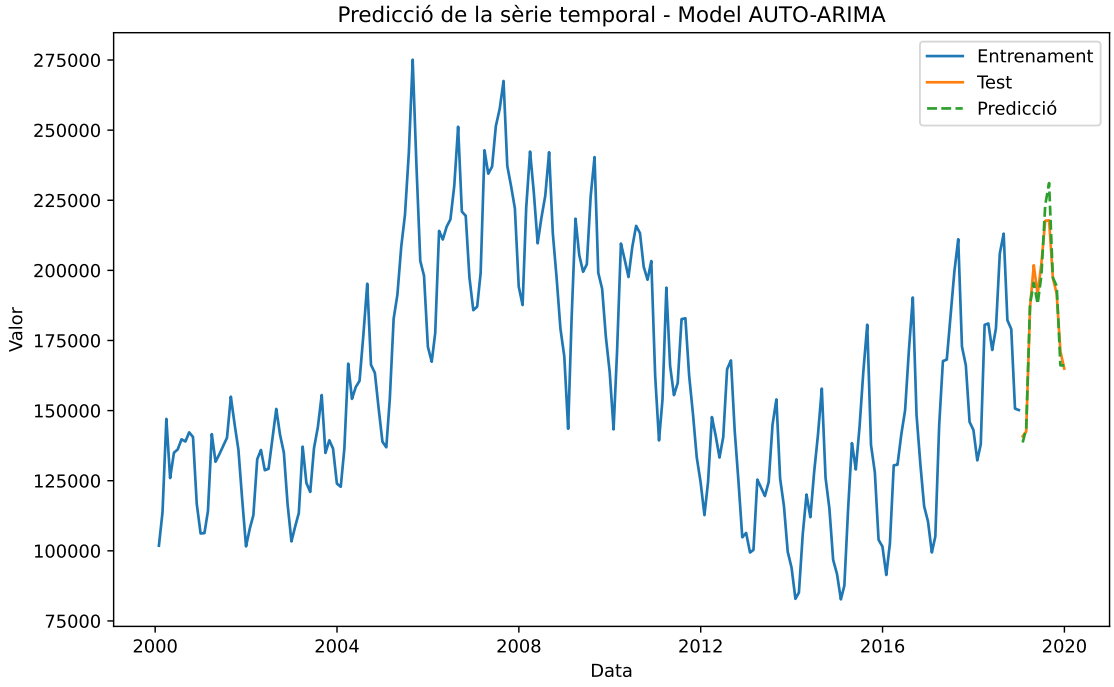

Figura 5.11: Predicció del trànsit comercial nacional mensual a l'aeroport de València (2000-2019) amb el model AUTO_ARIMA.

Per avaluar la precisió del model, la Taula 5.4 recull una comparació detallada entre els valors reals i les prediccions del model, juntament amb els errors percentuals.

Data	Valor real	Predicció	Diferència	Error (%)
2019-01	140839	138629	2210	1.57
2019-02	142500	143584	-1084	-0.76
2019-03	186876	188148	-1272	-0.68
2019-04	201851	195596	6255	3.10
2019-05	191335	188035	3300	1.72
2019-06	202493	198156	4337	2.14
2019-07	217684	223312	-5628	-2.59
2019-08	217863	231132	-13269	-6.09
2019-09	197404	197461	-57	-0.03
2019-10	191649	194378	-2729	-1.42
2019-11	171059	166125	4934	2.88
2019-12	164973	166114	-1141	-0.69

Taula 5.4: Comparació entre els valors reals i les prediccions del model AUTO_ARIMA per al trànsit comercial nacional a l'aeroport de València (2019).

L'anàlisi de la taula mostra que l'error percentual varia entre un mínim de -0.03% i un màxim de -6.09%, amb valors generals més baixos que els observats en el model ARIMA seleccionat manualment. En mesos com març i desembre, el model ha realitzat prediccions molt ajustades, amb errors inferiors a l'1%. En canvi, en mesos com agost de 2019, la desviació arriba al -6.09%, indicant que el model no ha capturat perfectament les variacions estacionals d'aquest període.
El rendiment global d'`auto_arima` suggereix que aquest enfocament automàtic pot generar millors prediccions que la selecció manual, ja que el seu AIC és menor i la distribució d'errors sembla més estable.
Finalment, la Figura 5.12 visualitza la comparació entre els valors reals i les prediccions del model, facilitant la detecció de desviacions significatives.

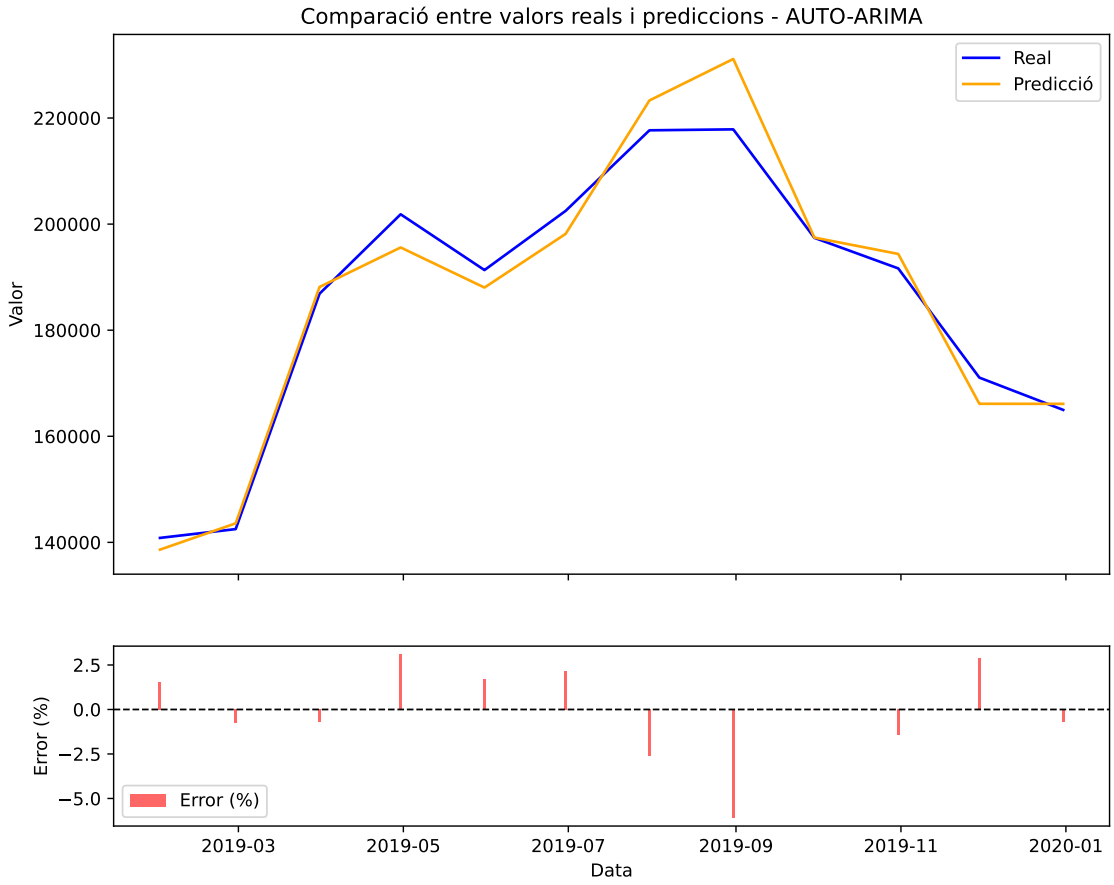

Figura 5.12: Comparació entre els valors reals i les prediccions del model AUTO_ARIMA per al trànsit comercial nacional mensual a l'aeroport de València (2000-2019).

En conjunt, el model seleccionat per `auto_arima` ha millorat l'ajust respecte a l'ARIMA manual, aconseguint un compromís més equilibrat entre simplicitat i capacitat predictiva. Tot i que en alguns mesos concrets presenta desviacions més altes, globalment ofereix una millor estabilitat en les prediccions, la qual cosa el converteix en una alternativa robusta i eficient per a la modelització de la sèrie temporal.

Suavització exponencial

A diferència dels models autoregressius com ARIMA, el mètode de Holt-Winters es basa en la suavització exponencial per modelitzar la tendència i l'estacionalitat de la sèrie temporal. Aquest enfocament és especialment útil quan la sèrie presenta patrons estacionals clars i una tendència que pot ser capturada mitjançant components additius o multiplicatius.

La funció `ajustar_holt_winters`, definida a la funció `ajustar_holt_winters` del paquet `holt_winters`, present a la Secció A.3, permet ajustar aquest model considerant una tendència additiva i una component estacional configurable. Aquesta implementació utilitza el mètode `ExponentialSmoothing` de la llibreria `statsmodels`, amb l'objectiu d'optimitzar els paràmetres de la suavització de manera automàtica.
L'ajustament del model segueix els passos següents:

1. Es defineix si la component estacional serà *additiva* o *multiplicativa*, segons la naturalesa de la sèrie.

2. Es fixa una tendència additiva per capturar els canvis de nivell a llarg termini.

3. S'incorpora un component estacional amb una periodicitat especificada per l'usuari.

4. El model es calcula mitjançant optimització interna, ajustant automàticament els paràmetres de suavització òptims.

Aquest enfocament permet capturar tant la tendència com l'estacionalitat de la sèrie sense necessitat de diferenciació explícita, com en el cas d'ARIMA. A més, el model de Holt-Winters és particularment eficient per a sèries amb fluctuacions estacionals regulars, sent una alternativa més interpretable i flexible en escenaris on la tendència i l'estacionalitat són prominents.
A la Figura 5.13 es mostra el resum estadístic del model ajustat amb Holt-Winters.

Resultat de Holt Winters

```
================================================================================
Dep. Variable:                 nacional   No. Observations:                  228
Model:             ExponentialSmoothing   SSE                 20604438178.643
Optimized:                         True   AIC                         4208.829
Trend:                         Additive   BIC                         4263.699
Seasonal:                      Additive   AICC                        4212.102
Seasonal Periods:                    12   Date:                Tue, 18 Feb 2025
Box-Cox:                          False   Time:                        20:37:21
Box-Cox Coeff.:                    None
================================================================================
                          coeff                 code              optimized
--------------------------------------------------------------------------------
smoothing_level        0.6060714                alpha                   True
smoothing_trend           0.0001                 beta                   True
smoothing_seasonal     0.3063889                gamma                   True
initial_level          1.2833e+05                  l.0                   True
initial_trend           200.34343                  b.0                   True
initial_seasons.0      -20865.774                  s.0                   True
initial_seasons.1      -14173.556                  s.1                   True
initial_seasons.2       10350.840                  s.2                   True
initial_seasons.3       1830.6632                  s.3                   True
initial_seasons.4       364.36111                  s.4                   True
initial_seasons.5       4954.2986                  s.5                   True
initial_seasons.6       11658.038                  s.6                   True
initial_seasons.7       20261.361                  s.7                   True
initial_seasons.8       10742.851                  s.8                   True
initial_seasons.9       7040.3819                  s.9                   True
initial_seasons.10     -9180.4618                 s.10                   True
initial_seasons.11     -22983.003                 s.11                   True
--------------------------------------------------------------------------------
```

Figura 5.13: Resum estadístic del model Holt-Winters ajustat a la sèrie temporal del trànsit comercial nacional mensual a l'aeroport de València (2000-2019).

El model ajustat segueix un esquema Holt-Winters additiu amb una periodicitat estacional de 12 mesos. El criteri d'informació d'Akaike (AIC) obtingut és de 4210.812, inferior al dels models ARIMA i AUTO ARIMA, cosa que indica un millor ajust a les dades.
Dels coeficients estimats, destaca un valor de suavització del nivell ($\alpha = 0.606$), la qual cosa implica que el model respon ràpidament als canvis en el nivell de la sèrie. En canvi, la suavització de la tendència ($\beta = 0.0001$) és pràcticament nul·la, indicant que el model considera que la tendència ha de ser quasi constant. La suavització de l'estacionalitat ($\gamma = 0.306$) té un valor moderat, reflectint un ajust equilibrat als patrons estacionals.
La resta de coeficients inicials (L_0, T_0 i $S_0, \ldots, S_{11}$) defineixen els valors inicials de les components de nivell, tendència i estacionalitat, respectivament. Amb això, es pot construir de manera recursiva la sèrie temporal predita a partir de les dades d'entrenament, fent ús de l'Equació 4.1. En aquest cas, i segons el que s'ha especificat a l'hora de definir el model, s'ha optat per una tendència additiva i una

estacionalitat additiva amb una periodicitat de 12 mesos, de manera que caldrà triar la primera de les equacions anteriors.
És important assenyalar que la fase d'entrenament ha generat un avís de convergència, indicant que l'optimització no ha convergit completament. Això pot suggerir que la selecció inicial de paràmetres o la naturalesa de les dades ha dificultat trobar una solució òptima, fet que podria resoldre's ajustant manualment alguns hiperparàmetres.

Anàlisi del model Holt-Winters

Després de l'ajustament del model Holt-Winters, s'ha generat la predicció de la sèrie temporal, diferenciant clarament les dades d'entrenament, el conjunt de test i la predicció. La Figura 5.14 mostra aquesta predicció.

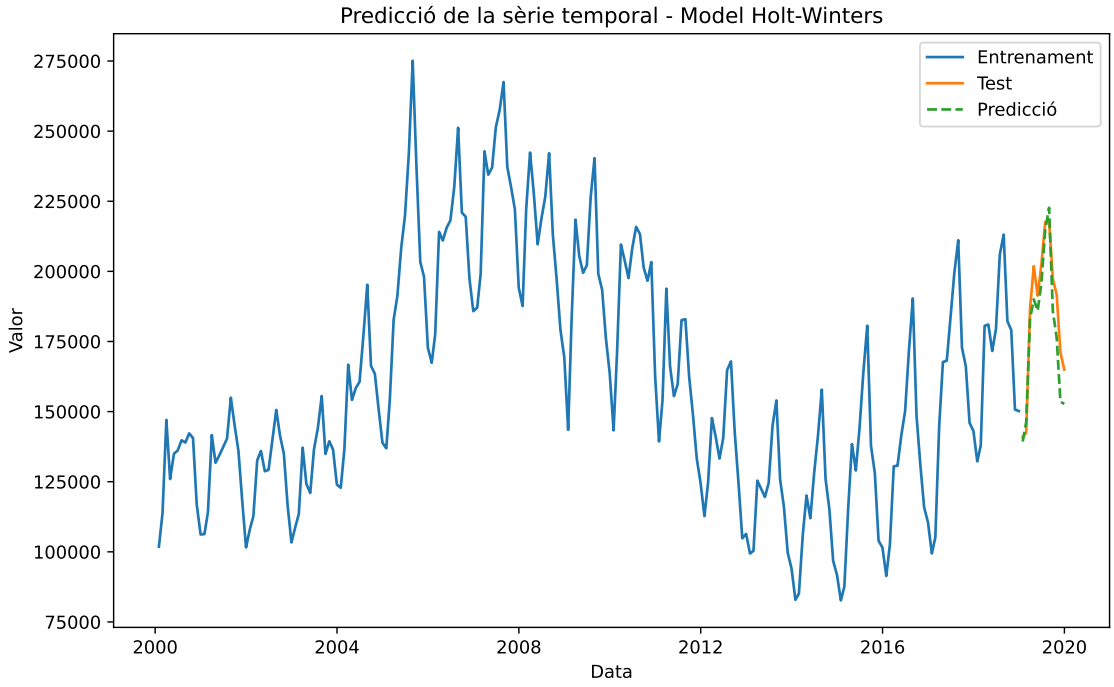

Figura 5.14: Predicció del trànsit comercial nacional mensual a l'aeroport de València (2000-2019) amb el model Holt-Winters.

Per avaluar la precisió del model, la Taula 5.5 recull una comparació detallada entre els valors reals i les prediccions del model, juntament amb els errors percentuals.

Data	Valor real	Predicció	Diferència	Error (%)
2019-01	140839	139235	1604	1.14
2019-02	142500	145958	-3458	-2.43
2019-03	186876	182640	4236	2.27
2019-04	201851	189425	12426	6.16
2019-05	191335	185170	6165	3.22
2019-06	202493	194755	7738	3.82
2019-07	217684	213698	3986	1.83
2019-08	217863	221419	-3556	-1.63
2019-09	197404	184456	12948	6.56
2019-10	191649	174626	17023	8.88
2019-11	171059	151701	19358	11.32
2019-12	164973	150674	14299	8.67

Taula 5.5: Comparació entre els valors reals i les prediccions del model Holt-Winters per al trànsit comercial nacional a l'aeroport de València (2019).

A l'anàlisi dels resultats s'observa que, en general, el model Holt-Winters captura l'estacionalitat de la sèrie, però presenta alguns desajustos en mesos concrets. Els errors percentuals es mantenen en valors moderats per a la major part del període de test, amb desviacions properes al 1%-3% en mesos com gener, febrer o juliol. No obstant això, es detecten errors significativament més alts en mesos com octubre i novembre, on la desviació supera el 10%. Això pot indicar que el model no ha capturat completament certes variacions estacionals o tendències puntuals.
El rendiment global del model suggereix que Holt-Winters és una bona elecció per a sèries amb patrons estacionals clars, però pot ser més sensible a canvis sobtats en la tendència. La Figura 5.15 il·lustra les desviacions entre els valors reals i les prediccions, facilitant la detecció de punts crítics amb majors errors.

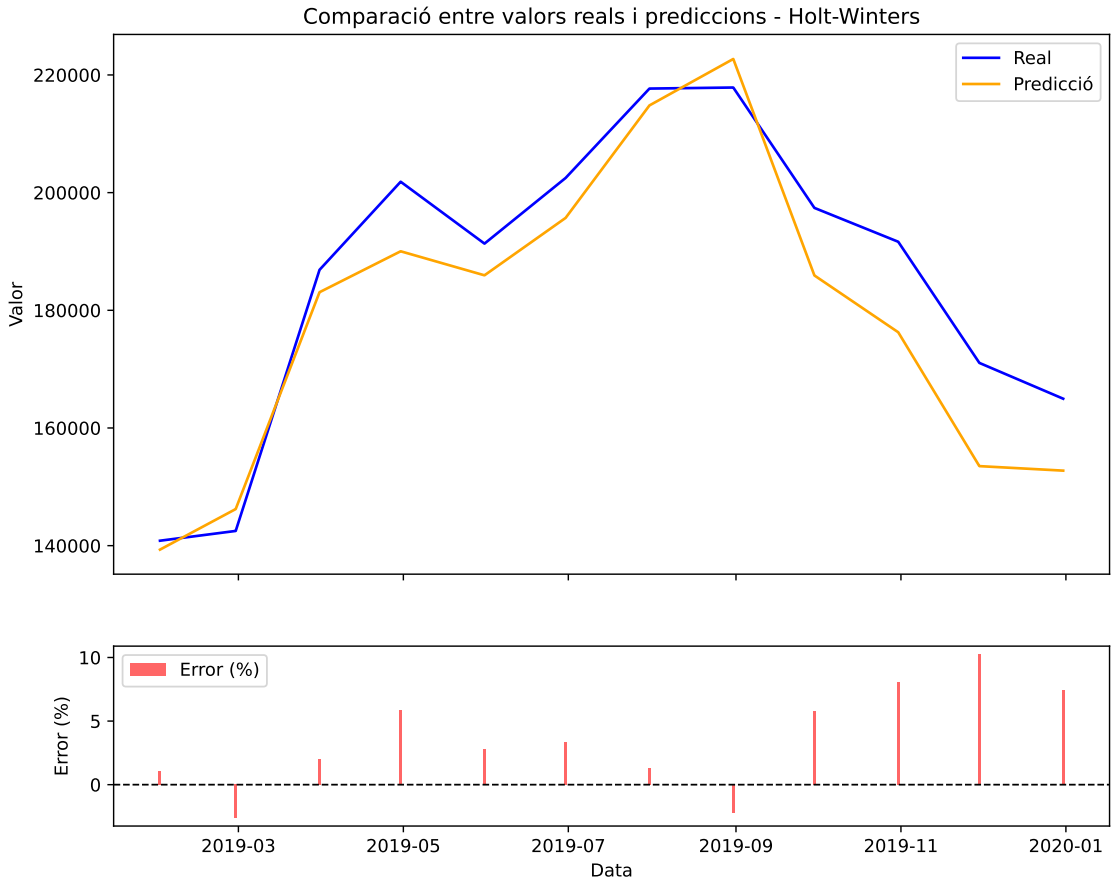

Figura 5.15: Comparació entre els valors reals i les prediccions del model Holt-Winters per al trànsit comercial nacional mensual a l'aeroport de València (2000-2019).

En conjunt, el model Holt-Winters ofereix una aproximació flexible per a la modelització de sèries estacionals, amb un rendiment raonable en la major part del període de test. No obstant això, la presència de desviacions més altes en determinats mesos suggereix que podria beneficiar-se d'un ajustament més detallat dels seus paràmetres o de tècniques complementàries per millorar la precisió en punts crítics.

5.4 Validació i interpretació dels resultats

A partir dels resultats anteriors, es poden extraure diverses conclusions rellevants sobre la seua precisió i adequació. La Taula 5.6 presenta un resum de les mètriques d'avaluació vistes a la Subsecció 4.3.2, comparant l'error quadràtic mitjà (RMSE) i l'error percentual absolut mitjà (MAPE) per a cadascun

dels models considerats.

	ARIMA	**AUTO_ARIMA**	**Holt-Winters**
RMSE	7509.50	5142.39	10606.32
MAPE	3.19	1.97	4.83

Taula 5.6: Avaluació de l'error dels models ARIMA, AUTO_ARIMA i Holt-Winters en la predicció del trànsit comercial nacional a l'aeroport de València (2019) mitjançant RMSE i MAPE.

L'enfocament basat en AUTO_ARIMA ha mostrat els millors resultats, amb un RMSE de 5142.39 i un MAPE de 1.97%. Aquesta millora respecte a la selecció manual de paràmetres suggereix que la cerca automàtica optimitza el compromís entre ajust i generalització, evitant tant la sobreparametrització com la infrarepresentació de patrons en les dades.
El model ARIMA seleccionat manualment ha obtingut un RMSE de 7509.50 i un MAPE de 3.19%. Malgrat oferir una estructura interpretable, els seus resultats són inferiors als d'AUTO_ARIMA, fet que indica que l'optimització manual dels paràmetres pot no ser tan efectiva en conjunts de dades amb estacionalitat i tendències complexes.
Pel que fa al model Holt-Winters, tot i ser útil en escenaris amb estacionalitat marcada, ha sigut el menys precís amb un RMSE de 10606.32 i un MAPE de 4.83%. L'error més elevat en comparació amb els altres models suggereix que la suavització exponencial no ha sigut capaç de capturar completament la dinàmica de la sèrie, especialment en punts de canvi de tendència o fluctuacions no sistemàtiques.

5.4.1 Perspectives i models alternatius

Els resultats obtinguts confirmen que, tot i la seua solidesa, els models estadístics tradicionals poden presentar dificultats a l'hora de capturar patrons temporals més complexos. La selecció automàtica d'ARIMA ha demostrat ser una eina eficient per millorar la precisió de les prediccions, però els models basats en supòsits lineals no sempre són suficients per a reflectir tota la dinàmica subjacent de les dades.
Aquesta situació planteja la necessitat d'explorar models alternatius que permeten una major flexibilitat en la modelització i una millor adaptació als canvis en la sèrie temporal. En determinats contextos, enfocaments més avançats poden oferir millores significatives, ja siga mitjançant tècniques capaces de capturar no-linealitats, models híbrids o sistemes automatitzats d'optimització de paràmetres. Alguns d'aquests models s'exploraran al Capítol 6, on s'analitzarà el seu potencial i les seues aplicacions en la predicció de sèries temporals.

Capítol 6

Models alternatius

L'anàlisi de sèries temporals ha estat tradicionalment dominada per models estadístics com ARIMA i Holt-Winters, que han demostrat ser eines robustes per a la predicció en nombrosos contextos. No obstant això, parteixen de certs supòsits sobre l'estructura de les dades, com l'estacionarietat o la linealitat en les relacions temporals, fet que pot restringir-ne l'ús en situacions més complexes. Els avenços en intel·ligència artificial i l'increment de la capacitat computacional han donat pas a nous enfocaments que busquen superar aquestes limitacions mitjançant metodologies més flexibles i adaptatives.

Aquest capítol es dedica a explorar alguns d'aquests models alternatius que han guanyat popularitat en els darrers anys. L'objectiu no és substituir els models tradicionals, sinó entendre en quins contextos poden resultar més eficients o aportar un valor afegit a les tècniques ja establertes. A més, sotmetrem el primer model avaluat, Prophet, a una anàlisi comparativa amb els models clàssics per a determinar-ne el rendiment en la sèrie temporal d'estudi. Aquest model està dissenyat per gestionar sèries temporals amb tendències canviants i estacionalitats marcades. Destaca per la seua flexibilitat, facilitat d'ús i capacitat per tractar dades amb absències o *outliers*, fet que el converteix en una eina atractiva per a analistes i científics de dades.

També introduirem alguns models basats en xarxes neuronals, centrant-nos en les xarxes neuronals recurrents (RNN) i les Long Short-Term Memory (LSTM), que permeten capturar dependències temporals complexes. Tot i el seu potencial, requereixen grans volums de dades d'entrenament i tenen un cost computacional elevat, factors a tindre en compte en la seua aplicació.

Finalment, inclourem una breu explicació sobre l'ús d'`AutoTS`, una llibreria que automatitza la selecció i comparació de models de predicció, facilitant la cerca del model òptim sense intervenció manual intensiva. Permet avaluar tant models estadístics com tècniques d'aprenentatge automàtic, ajustant els hiperparàmetres per maximitzar la precisió de les prediccions.

L'ús d'aquests models comporta tant beneficis com reptes, que cal considerar abans d'optar per una metodologia específica. D'una banda, la seua capacitat per modelitzar relacions no lineals permet captar dinàmiques més complexes dins de les dades, la qual cosa pot ser útil en sèries temporals amb canvis sobtats o estructures difícils de modelitzar amb un enfocament purament estadístic. A més, eines com AutoTS redueixen la càrrega computacional associada a la selecció de models, facilitant el procés per a usuaris sense experiència prèvia en la sintonització de models predictius.

Tanmateix, també presenten inconvenients que cal tindre en compte. La majoria requereixen grans

volums de dades per a entrenar-se adequadament, especialment en el cas de les xarxes neuronals, que necessiten seqüències llargues per captar patrons de llarg abast. Això pot ser un problema en escenaris amb dades escasses o sèries temporals curtes. A més, la seua interpretabilitat és menor en comparació amb models com SARIMA o Holt-Winters, on els paràmetres tenen un significat més intuïtiu i clar. En aplicacions crítiques, aquesta manca d'interpretabilitat pot ser un obstacle per a la seua adopció generalitzada.
Pel que fa a les línies d'investigació actuals en aquest camp, una de les àrees més actives se centra en la millora de l'eficiència i la capacitat d'aprenentatge de les xarxes neuronals aplicades a sèries temporals. Es treballa en l'optimització d'arquitectures i algoritmes per reduir el cost computacional i millorar la capacitat d'extracció de patrons. A més, es desenvolupen models híbrids que combinen les propietats dels models tradicionals amb els avantatges de les tècniques basades en aprenentatge profund, amb l'objectiu de trobar un equilibri entre interpretabilitat i capacitat predictiva. Paral·lelament, hi ha una creixent atenció en el desenvolupament d'eines d'AutoML (*Automated Machine Learning*) per a la selecció òptima de models i hiperparàmetres sense necessitat d'una intervenció manual constant.
Aquest capítol oferirà una visió general d'aquests models, posant èmfasi en les seues aplicacions pràctiques i en la comparació amb els enfocaments tradicionals.

6.1 Prophet

Prophet és un model de regressió no lineal desenvolupat per Meta [30], dissenyat inicialment per a la predicció de dades diàries amb estacionalitat setmanal i anual, així com per capturar l'efecte dels dies festius. A diferència dels models tradicionals com SARIMA o Holt-Winters, que assumeixen una estacionalitat fixa i no consideren explícitament l'impacte de dies festius o altres esdeveniments irregulars, Prophet ofereix una solució més flexible que permet incorporar aquests efectes de manera natural dins del model.
Una de les grans limitacions dels models estadístics clàssics és la dificultat per gestionar patrons estacionals que no es repeteixen amb una freqüència constant o que es veuen afectats per factors externs, com ara canvis en el calendari comercial o variacions en dies laborables i festius. Prophet resol aquest problema introduint variables específiques per als dies festius, que permeten ajustar la predicció en funció d'aquests esdeveniments, un aspecte fonamental en moltes aplicacions empresarials on la demanda de productes o serveis es veu afectada per festivitats locals o nacionals.
El model de Prophet es pot expressar mitjançant la següent equació:

$$y_t = g(t) + s(t) + h(t) + \varepsilon_t,$$

on:

- $g(t)$ representa la tendència a llarg termini, modelada com una funció a trams lineals (*piecewise-linear*) o, opcionalment, mitjançant una funció logística per imposar un límit superior a la tendència.
- $s(t)$ captura els patrons estacionals, que s'expressen en termes de sèries de Fourier. Per defecte, el model utilitza components d'ordre 10 per a l'estacionalitat anual i d'ordre 3 per a l'estacionalitat setmanal.

- $h(t)$ incorpora els efectes de dies festius mitjançant variables indicadores (*dummy variables*), fet que el diferencia clarament dels models convencionals. Aquesta capacitat permet que Prophet ajuste la predicció en funció de dies no laborables o festivitats específiques, adaptant-se millor a contextos comercials o socials on aquestes dates tenen un impacte significatiu.

- ε_t és un terme d'error aleatori de soroll blanc.

A més d'aquest tractament explícit dels dies festius, Prophet destaca per la seua capacitat per detectar automàticament punts de canvi (*changepoints*) en la tendència sense necessitat d'especificar-los manualment, un aspecte especialment útil en sèries temporals on la dinàmica subjacent pot experimentar alteracions sobtades. Així mateix, el model es basa en un enfocament bayesià, que facilita una estimació flexible i automatitza la selecció de diversos paràmetres, inclosos els *changepoints* i altres propietats del model.
Aquest enfocament el converteix en una eina especialment adequada per a entorns on la predicció ha de considerar factors externs variables, com ara el comerç minorista, el transport o la demanda energètica, on els dies festius i altres esdeveniments poden tindre un impacte considerable en les dades històriques i futures.

6.1.1 Aplicació pràctica

Una vegada presentades les característiques del model Prophet, procedirem a analitzar-ne el rendiment en la sèrie temporal d'estudi. La Figura 6.1 mostra la predicció completa generada pel model, diferenciant clarament les dades d'entrenament, el conjunt de test i la predicció.

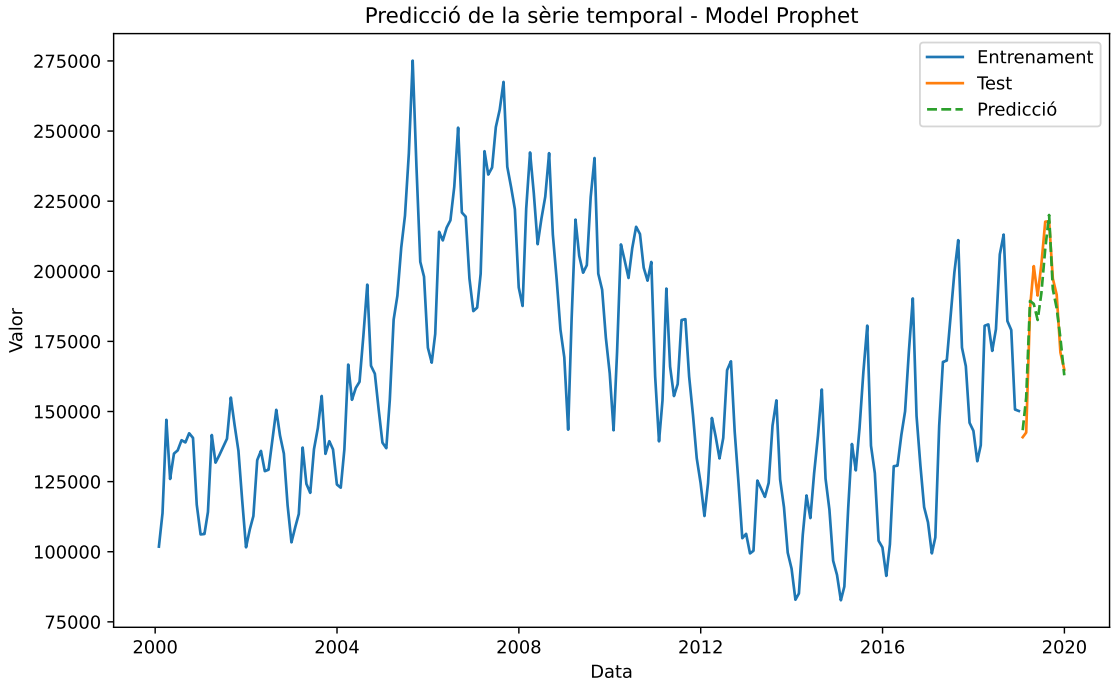

Figura 6.1: Predicció del trànsit comercial nacional mensual a l'aeroport de València (2000-2019) amb el model Prophet.

Comparació entre valors reals i prediccions

Per quantificar l'exactitud del model, s'ha elaborat una taula comparativa que recull els valors reals, les prediccions de Prophet, la diferència absoluta i l'error percentual. Aquesta informació permet identificar patrons en les desviacions del model i determinar en quines situacions funciona millor o presenta majors discrepàncies.

L'anàlisi de la Taula 6.1 revela que l'error percentual varia entre valors propers al 1% i desviacions més significatives, com el cas de febrer de 2019, on l'error arriba al -8.72%. Aquesta discrepància pot indicar que Prophet no ha capturat completament un efecte particular d'aquell període, com un canvi atípic en la tendència o una influència no modelitzada, com esdeveniments no recurrents o efectes econòmics inesperats. Per contra, en mesos com març i agost, l'error és inferior al 2%, fet que suggereix que el model aconsegueix ajustar-se bé a la dinàmica estacional de la sèrie.

És important destacar que Prophet gestiona bé les tendències a llarg termini, però en algunes ocasions pot presentar desviacions puntuals en mesos específics. Aquest aspecte es pot atribuir a la forma en què el model estima els canvis de tendència o a la influència de components estacionals no periòdics. En qualsevol cas, aquestes diferències haurien de ser contrastades amb altres models per avaluar quin enfocament proporciona millors resultats en termes d'error global i estabilitat predictiva.

Data	Valor real	Predicció	Diferència	Error (%)
2019-01	140839	143336	-2497	-1.77
2019-02	142500	154931	-12431	-8.72
2019-03	186876	189434	-2558	-1.37
2019-04	201851	188363	13488	6.68
2019-05	191335	182621	8714	4.55
2019-06	202493	192212	10281	5.08
2019-07	217684	208006	9678	4.45
2019-08	217863	220085	-2222	-1.02
2019-09	197404	193633	3771	1.91
2019-10	191649	186803	4846	2.53
2019-11	171059	176122	-5063	-2.96
2019-12	164973	162911	2062	1.25

Taula 6.1: Comparació entre els valors reals i les prediccions del model Prophet per al trànsit comercial nacional a l'aeroport de València (2019).

La Figura 6.2 il·lustra la comparació entre els valors reals i les prediccions, facilitant la identificació de desviacions significatives.

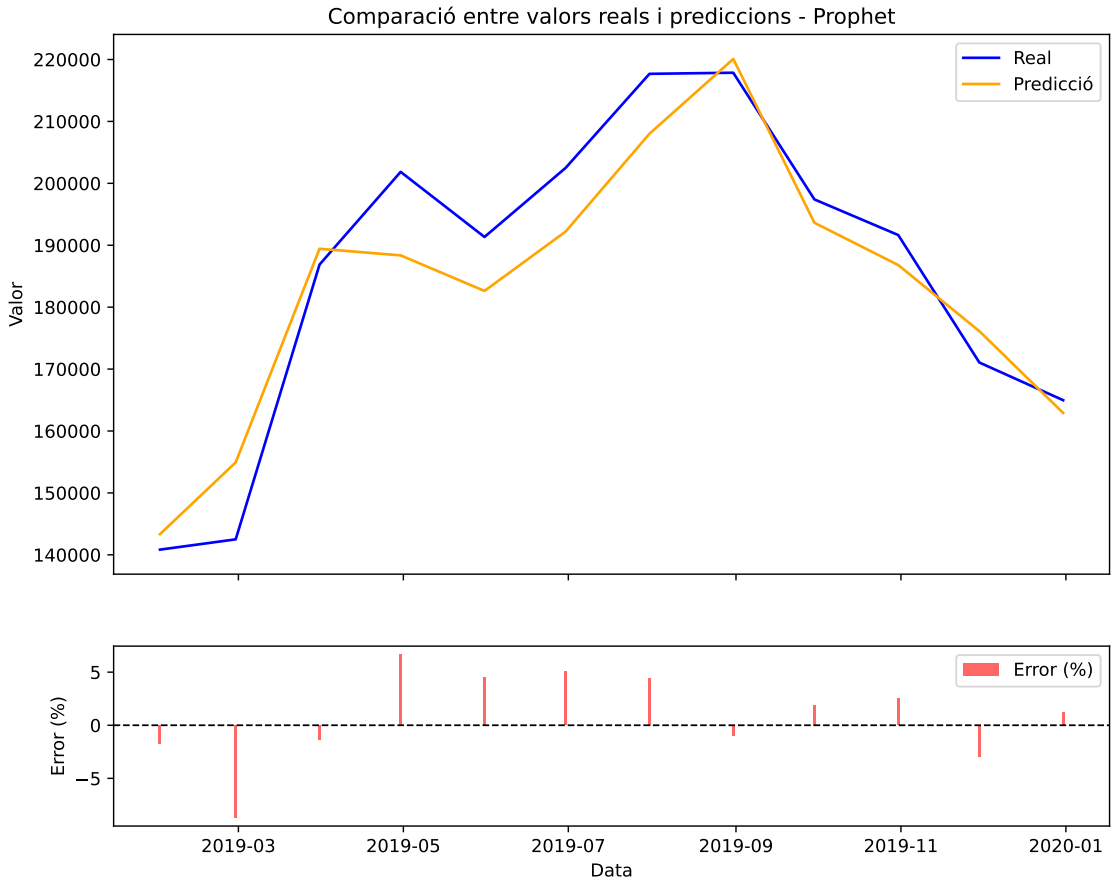

Figura 6.2: Comparació entre els valors reals i les prediccions del model Prophet per al trànsit comercial nacional mensual a l'aeroport de València (2000-2019).

Rendiment del model

Per quantificar l'ajust del model, s'han calculat diverses mètriques d'error, incloent-hi l'error quadràtic mitjà (RMSE) i l'error absolut mitjà percentual (MAPE). Aquestes mètriques permeten comparar Prophet amb altres models utilitzats en aquest estudi.

Mètrica	ARIMA	AUTO_ARIMA	Holt-Winters	Prophet
RMSE	7509.50	5142.39	10606.32	7621.57
MAPE (%)	3.19	1.97	4.83	3.52

Taula 6.2: Extensió de la Taula 5.6 amb la incorporació del model Prophet en la comparació de mètriques d'error (RMSE i MAPE) per al trànsit comercial nacional a l'aeroport de València (2019).

Com es pot observar, Prophet millora els resultats obtinguts amb Holt-Winters, però no aconsegueix superar AUTO ARIMA. Això suggereix que, mentre Prophet és una eina robusta en la modelització de tendències canviants i l'impacte de dies festius, en sèries altament estacionàries pot ser més eficient un enfocament purament estadístic.

6.1.2 Conclusió

Els resultats mostren que Prophet ofereix un bon ajust a la sèrie temporal, amb errors comparables als models tradicionals. La seua principal fortalesa rau en la flexibilitat per modelitzar patrons estacionals canviants i integrar l'efecte de dies festius, aspecte que models com ARIMA o Holt-Winters no contemplen de manera explícita.
Tot i això, la comparació de mètriques indica que en sèries on la component estacional és més estable, models com AUTO ARIMA poden resultar més precisos en termes d'error absolut. Així, Prophet destaca com una eina potent en escenaris amb canvis en la tendència o patrons estacionals no trivials, però pot no ser l'opció òptima en contextos amb una estacionalitat ben definida i previsible.

6.2 Models basats en xarxes neuronals

L'ús de xarxes neuronals artificials (*Artificial Neural Network*, ANN) en la predicció de sèries temporals ha experimentat un creixement notable en els darrers anys. A diferència dels models estadístics tradicionals com ARIMA o Holt-Winters, les xarxes neuronals ofereixen una aproximació flexible i no lineal, capaç de captar patrons complexos i relacions ocultes en les dades. Aquesta característica les fa especialment útils en contextos en què les sèries temporals presenten comportaments no estacionaris o estan influïdes per múltiples factors externs [29].

6.2.1 Xarxes neuronals recurrents (RNN)

Les xarxes neuronals recurrents (RNN) són una de les arquitectures més populars per al tractament de dades seqüencials, com les sèries temporals. A diferència de les xarxes neuronals convencionals, que assumeixen que cada entrada és independent, les RNNs incorporen connexions recurrents que els permeten mantindre una memòria de les dades prèvies. Això facilita la captura de patrons temporals i millora la capacitat predictiva en seqüències de dades llargues.
Tanmateix, les RNN tradicionals pateixen el problema de la desaparició del gradient, que dificulta l'aprenentatge de dependències a llarg termini. Per resoldre aquesta limitació, s'han desenvolupat variants més avançades, com les xarxes de memòria a llarg termini (*Long Short-Term Memory*, LSTM) i les unitats recurrents amb porta (*Gated Recurrent Unit*, GRU).

6.2.2 LSTM i GRU: millores en l'aprenentatge seqüencial

Les xarxes LSTM van ser introduïdes per abordar el problema de la desaparició del gradient. Aquestes xarxes inclouen tres tipus de portes:

- **Porta d'oblit**: Decideix quina informació de l'estat anterior s'ha d'eliminar.
- **Porta d'entrada**: Determina quina nova informació s'ha d'afegir a l'estat intern.
- **Porta de sortida**: Filtra la informació que es transferirà com a entrada per a la següent iteració.

Aquest mecanisme permet que les LSTM gestionen dependències de llarg termini de manera més eficient que les RNN convencionals, fent-les especialment útils per a la predicció de sèries temporals amb patrons complexos i dades volàtils.
D'altra banda, les xarxes GRU són una simplificació de les LSTM, però mantenen una capacitat similar per capturar relacions temporals. A diferència de les LSTM, les GRU utilitzen només dues portes (*reset* i *update*), la qual cosa redueix la complexitat computacional i accelera el procés d'entrenament sense perdre massa precisió.
Segons diversos estudis, les GRU sovint superen les LSTM en termes d'eficiència computacional, especialment quan es treballa amb grans volums de dades o en contextos en què l'entrenament ha de ser ràpid i eficient [29].

6.2.3 Consideracions i aplicacions

Les xarxes neuronals recurrents i les seues variants han estat àmpliament utilitzades en la predicció de sèries temporals financeres, meteorològiques i de demanda energètica, entre altres camps. No obstant això, aquests models tenen alguns desafiaments importants:

- **Necessitat d'un gran volum de dades**: L'efectivitat de les xarxes neuronals depèn en gran part de la disponibilitat de dades d'entrenament suficients i representatives.
- **Cost computacional elevat**: Els models profunds requereixen recursos computacionals considerables, especialment quan es treballa amb xarxes complexes com les LSTM.
- **Dificultat d'interpretació**: A diferència dels models estadístics tradicionals, les xarxes neuronals funcionen com a caixes negres, la qual cosa pot dificultar la interpretació dels resultats i la justificació de les prediccions.

Malgrat aquestes limitacions, els models basats en xarxes neuronals són una eina potent en la predicció de sèries temporals i continuen sent un camp d'investigació actiu, amb noves tècniques que busquen millorar-ne l'eficiència i la interpretabilitat.

6.3 AutoTS: Automatització en la predicció de sèries temporals

L'ús de l'AutoML en la predicció de sèries temporals ha guanyat popularitat gràcies a la seua capacitat per automatitzar la selecció del model més adequat per a cada conjunt de dades. En aquest context, `AutoTS` és una llibreria de Python dissenyada per generar prediccions precises de manera eficient i

escalable. Aquesta eina ha demostrat la seua efectivitat en competicions com el concurs M6 del 2023, on va aconseguir el millor rendiment en la predicció de mercats financers [22].
A diferència d'altres enfocaments, AutoTS no es limita a un únic tipus de model, sinó que proporciona un entorn unificat per a l'avaluació de diferents estratègies predictives, incloent-hi:

- **Models estadístics**: ARIMA, Holt-Winters, VAR, DynamicFactor, entre altres.
- **Regressió**: RollingRegression, WindowRegression, UnivariateRegression, etc.
- **Aprenentatge profund**: GluonTS, NeuralProphet, PytorchForecasting.
- ***Ensembling***: combinació de models amb tècniques de mosaïc i ensembles horitzontals.

6.3.1 Característiques principals

Un dels avantatges d'AutoTS és la seua capacitat per gestionar múltiples sèries temporals (*multivariate forecasting*) i generar prediccions probabilístiques amb intervals de confiança. La llibreria s'integra directament amb `pandas`, permetent treballar amb *dataframes* sense conversió de formats.
A més, inclou més de 30 transformacions específiques per a sèries temporals, aplicables seguint l'estil `sklearn` mitjançant `fit`, `transform` i `inverse_transform`. Aquest conjunt d'eines millora la qualitat predictiva en diferents tipus de dades.
AutoTS també incorpora un sistema d'AutoML basat en algorismes genètics, capaç de trobar automàticament la millor combinació de models, preprocessaments i estratègies d'*ensembling*, reduint la necessitat d'experimentació manual.

6.3.2 Selecció de models i *ensembling*

L'usuari pot definir un subconjunt específic de models a provar, o bé seleccionar configuracions predefinides per optimitzar el procés:

- `probabilistic`: models que ofereixen intervals de confiança en les prediccions.
- `multivariate`: enfocament per a múltiples sèries temporals simultànies.
- `fast` o `superfast`: models optimitzats per a velocitat.
- `all`: utilitza tota la gamma de models disponibles.

Un dels punts forts d'AutoTS és la seua arquitectura d'*ensembling*, que assigna el model més adequat a cada sèrie temporal mitjançant estratègies horitzontals i de mosaïc. Això permet obtenir prediccions més precises i millorar l'escalabilitat en conjunts de dades de gran volum.
AutoTS ofereix una solució avançada per a la predicció automatitzada de sèries temporals, combinant models estadístics, regressió i aprenentatge automàtic en un marc flexible i optimitzat. La seua capacitat per treballar amb sèries multivariants, fer prediccions probabilístiques i automatitzar la selecció de models la converteix en una eina valuosa per a investigadors i professionals que necessiten prediccions precises amb una mínima intervenció manual.
Tanmateix, la qualitat de les prediccions depén de la informació disponible i de l'adequada configuració dels paràmetres d'entrenament, fet que subratlla la importància d'una anàlisi acurada de les dades d'entrada.

Capítol 7

Conclusions

Des de l'inici, la finalitat de l'estudi ha estat analitzar i avaluar diferents mètodes de predicció aplicats a una sèrie temporal real. A partir d'una revisió detallada dels conceptes matemàtics teòrics i de la implementació de models clàssics, com ARIMA i Holt-Winters, així com alternatives més modernes com Prophet, s'ha tractat d'identificar quins enfocaments són més efectius per capturar les tendències i l'estacionalitat del fenomen estudiat.
Aquest capítol presenta, en primer lloc, un resum dels resultats obtinguts, on es comparen els diferents models utilitzats a partir de les seues mètriques de predicció i capacitat d'ajust a les dades històriques. Posteriorment, es discuteixen les limitacions trobades en aquests models tradicionals i es proposen possibles línies de millora que podrien abordar aquestes limitacions en futurs estudis.

7.1 Resum dels resultats obtinguts

Els resultats d'aquest estudi han permés analitzar les fortaleses i debilitats dels diferents mètodes aplicats a l'anàlisi de sèries temporals. Per avaluar-ne el rendiment, s'han utilitzat diverses tècniques de validació, comparant la capacitat predictiva de cadascun mitjançant mesures d'error habituals com el RMSE i el MAPE.
Dels mètodes implementats, `AUTO_ARIMA` ha destacat per oferir un bon equilibri entre simplicitat i precisió. La seua optimització automàtica facilita l'ajust sense necessitat d'una selecció manual detallada dels paràmetres, fet que redueix la complexitat del procés i minimitza els errors en moltes situacions. Gràcies a aquesta facilitat d'ús, ha estat l'alternativa més eficient en termes de precisió, superant els resultats obtinguts amb l'ajust manual d'ARIMA, Holt-Winters i Prophet.
Pel que fa a ARIMA ajustat manualment, ha proporcionat una bona aproximació, però ha requerit una selecció acurada dels paràmetres per aconseguir resultats satisfactoris. Tot i que aquest enfocament pot ser útil quan es busca un control més detallat de l'estructura del model, el procés d'optimització pot resultar costós en termes de temps. En aquest cas, `AUTO_ARIMA` ha ofert una alternativa més eficient, evitant la necessitat d'exploracions paramètriques exhaustives.
D'entre els mètodes estudiats, Holt-Winters ha mostrat una menor precisió. Malgrat ser adequat per a dades amb patrons estacionals ben definits, pot no ser la millor opció quan la tendència és canviant o presenta fluctuacions més complexes. No obstant això, la seua simplicitat i interpretació el converteixen en una alternativa viable en escenaris on es prioritza la rapidesa de càlcul per damunt de la precisió.

Finalment, Prophet ha demostrat ser una opció atractiva gràcies a la seua capacitat d'incorporar factors exògens com dies festius i punts de canvi en la tendència. Aquest enfocament pot ser especialment útil en contextos en què les interrupcions o variacions externes tenen un paper rellevant en l'evolució de la sèrie. Tot i ser un model relativament nou, els seus resultats han estat prometedors, fet que justifica explorar més a fons el seu potencial mitjançant tècniques d'optimització addicionals o la incorporació de més dades.
En conjunt, els resultats suggereixen que l'elecció del mètode depén de la naturalesa de la sèrie temporal i dels objectius específics de l'estudi. Si es busca una solució generalista i eficaç sense ajust manual, `AUTO_ARIMA` és una opció recomanable. En canvi, Prophet pot ser més adequat en contextos amb canvis sobtats o influències externes significatives. Quan es treballa amb dades fortament estacionals, Holt-Winters pot resultar útil, mentre que l'ajust manual d'ARIMA és recomanable per a situacions en què es disposa de coneixement expert i es vol una personalització més precisa.
No existeix una solució universal, per la qual cosa és recomanable provar diverses opcions i comparar-ne els resultats abans de prendre una decisió final. A més de les mètriques de precisió, factors com la disponibilitat de recursos computacionals, la facilitat d'implementació i la interpretació dels resultats també poden ser determinants en l'elecció més adequada per a cada situació.

7.2 Propostes de millora i perspectives futures

Els resultats obtinguts en aquest estudi han demostrat l'eficàcia dels models estadístics tradicionals en la predicció de sèries temporals. No obstant això, el camp de l'anàlisi predictiva evoluciona constantment i ofereix noves oportunitats per millorar la precisió, l'eficiència computacional i la capacitat d'adaptació dels models. Amb aquesta perspectiva, és interessant explorar diversos camins que podrien portar la predicció de sèries temporals a un nivell superior.
Un primer pas cap a una millor modelització és la incorporació de models multivariants que permeten tenir en compte múltiples fonts d'informació. En situacions en què una sèrie temporal està influenciada per factors externs, com poden ser variables macroeconòmiques, canvis en la regulació o esdeveniments puntuals, els models tradicionals univariants poden resultar insuficients. Alternatives com els models autoregressius vectorials (*Vector Autoregressive Models*, VAR) o extensions com SARIMAX permeten integrar aquestes influències externes, millorant la capacitat predictiva.
D'altra banda, el desenvolupament de tècniques més avançades en aprenentatge automàtic ha obert la porta a nous enfocaments que combinen la solidesa dels models estadístics amb la flexibilitat dels algorismes d'aprenentatge automàtic. Algorismes com *Gradient Boosting* (XGBoost, LightGBM) han demostrat ser especialment útils en contextos en què les relacions no són purament lineals, mentre que les xarxes neuronals recurrents (LSTM, GRU) permeten capturar patrons seqüencials de manera més eficaç que els models clàssics.
L'optimització computacional també és un aspecte clau per garantir que aquests models siguen escalables i eficients en entorns amb grans volums de dades. Implementacions en paral·lel, ús de GPU i adaptació a arquitectures de computació distribuïda poden reduir dràsticament el temps d'entrenament i inferència, permetent l'ús d'aquests models en aplicacions crítiques on la rapidesa de resposta és essencial.
Una altra línia de millora important és la capacitat de realitzar prediccions en temps real. En sectors com la gestió de la demanda energètica, la logística o el comerç electrònic, la capacitat de fer prediccions contínues basades en dades que s'actualitzen en temps real pot aportar un avantatge competitiu

significatiu. Tecnologies com Apache Kafka o TensorFlow Serving poden ser integrades per garantir que les prediccions es generen i s'adapten a mesura que arriben noves dades.
Finalment, perquè aquests models siguen aplicables en entorns de producció, és fonamental integrar-los amb sistemes de bases de dades especialitzats en sèries temporals, com TimescaleDB o InfluxDB. Aquesta integració permet una gestió eficient de les dades històriques i facilita la implementació de sistemes de monitoratge i reentrenament automàtic, assegurant que els models es mantenen actualitzats sense necessitat d'intervenció manual constant.
En resum, el futur de la predicció de sèries temporals passa per la combinació de models més avançats, l'optimització computacional i la integració en sistemes en temps real. Aquestes millores no sols augmentarien la precisió de les prediccions, sinó que també garantirien que els models siguen àgils, adaptatius i escalables per a entorns cada vegada més exigents.

Apèndix A

Codi

A.1 ARIMA

```
from pmdarima.arima import ARIMA
import itertools
import time

def ajustar_arima(train, p_range=(0,2), d_range=(0,2), q_range=(0,2), P_range=(0,1), D_range=(0,1),
↪  Q_range=(0,1), m=12):
    """
    Ajusta un model ARIMA provant diferents valors dels paràmetres i seleccionant el millor segons
    ↪  AIC.

    Arguments:
    - train: Sèrie temporal d'entrenament.
    - p_range, d_range, q_range: Rangs per als paràmetres ARIMA.
    - P_range, D_range, Q_range: Rangs per als paràmetres estacionals SARIMA.
    - m: Periodicitat estacional.

    Retorna:
    - El millor model ARIMA segons AIC.
    """

    millor_model = None
    millor_aic = float("inf")
    millor_ordre = None
    millor_ordre_estacional = None

    print("Realitzant cerca pas a pas per minimitzar l'AIC")

    combinacions_parametres = list(itertools.product(range(p_range[0], p_range[1] + 1),
```

```
                                                range(d_range[0], d_range[1] + 1),
                                                range(q_range[0], q_range[1] + 1)))
    combinacions_estacionals = list(itertools.product(range(P_range[0], P_range[1] + 1),
                                                      range(D_range[0], D_range[1] + 1),
                                                      range(Q_range[0], Q_range[1] + 1)))

    for ordre in combinacions_parametres:
        for ordre_estacional in combinacions_estacionals:
            try:
                inici_temps = time.time()
                model = ARIMA(
                    order=ordre,
                    seasonal_order=ordre_estacional + (m,),
                    suppress_warnings=True
                ).fit(train)
                temps_execucio = time.time() - inici_temps

                aic = model.aic()
                ordre_str = f"ARIMA{ordre}{ordre_estacional + (m,)}"
                temps_str = f"Temps={temps_execucio:.2f} segons"
                print(f" {ordre_str:<35}: AIC={aic:.3f}, {temps_str}")

                if aic < millor_aic:
                    millor_aic = aic
                    millor_model = model
                    millor_ordre = ordre
                    millor_ordre_estacional = ordre_estacional

            except Exception as e:
                print(f"Error amb ARIMA{ordre}{ordre_estacional + (m,)}: {e}")
                continue

    print(f"\nMillor model seleccionat: ARIMA{millor_ordre} Seasonal{millor_ordre_estacional + (m,)}
    ↪  | AIC={millor_aic:.3f}")

    return millor_model
```

A.2 AUTO_ARIMA

```
from pmdarima import auto_arima
import time

def ajustar_auto_arima(train, m=1):
    """
    Ajusta un model ARIMA automàtic provant tots els valors de d i D fins als màxims definits.

    Arguments:
    - train: Sèrie temporal d'entrenament.
    - m: Període d'estacionalitat.

    Retorna:
    - El millor model ARIMA segons AIC.
    """

    max_d = 2
    max_D = 2

    millor_model = None
    millor_aic = float("inf")
    millor_ordre = None
    millor_ordre_estacional = None

    print("Iniciant la cerca del millor model ARIMA...")

    for d in range(0, max_d + 1):
        for D in range(0, max_D + 1):
            print(f"Provant model amb d={d}, D={D}...")
            try:
                inici_temps = time.time()
                model = auto_arima(
                    train,
                    start_p=0, start_q=0,
                    max_p=2, max_q=2,
                    d=d, start_P=0, D=D, start_Q=0,
                    max_P=1, max_Q=1,
                    m=m, seasonal=True,
                    error_action='warn',
                    trace=True,
                    suppress_warnings=True,
                    stepwise=False,
                    random_state=20,
```

```
                n_fits=50
            )
            temps_execucio = time.time() - inici_temps

            aic = model.aic()
            ordre = model.order
            ordre_estacional = model.seasonal_order

            print(f"El model amb d={d}, D={D} té un AIC de {aic:.3f} | Temps={temps_execucio:.2f}
            ↪  segons")

            if aic < millor_aic:
                millor_aic = aic
                millor_model = model
                millor_ordre = ordre
                millor_ordre_estacional = ordre_estacional

        except Exception as e:
            print(f"S'ha produït un error amb d={d}, D={D}: {e}")
            continue

print(f"Model òptim seleccionat: ARIMA{millor_ordre}{millor_ordre_estacional}[{m}] |
↪  AIC={millor_aic:.3f}")

return millor_model
```

A.3 Holt_Winters

```python
from statsmodels.tsa.holtwinters import ExponentialSmoothing

def ajustar_holt_winters(train, seasonal="add", seasonal_periods=1, trend="add", damped_trend=False):
    """
    Ajusta un model Holt-Winters (Suavització Exponencial).

    Arguments:
    - train: Sèrie temporal d'entrenament.
    - seasonal: Tipus d'estacionalitat ("add" o "mul").
    - seasonal_periods: Període d'estacionalitat.
    - trend: Tipus de tendència ("add", "mul" o None).
    - damped_trend: Indica si la tendència ha d'estar esmorteïda.

    Retorna:
    - Model Holt-Winters ajustat o None si hi ha un error.
    """
    try:
        model = ExponentialSmoothing(
            train,
            seasonal=seasonal,
            seasonal_periods=seasonal_periods,
            trend=trend,
            damped_trend=damped_trend
        ).fit(optimized=True, use_boxcox=None, remove_bias=True)

        print(
            f"Model Holt-Winters ajustat correctament amb estacionalitat {seasonal}, període
            ↪ {seasonal_periods}, tendència {trend} i damped_trend {damped_trend}.")
        return model

    except Exception as e:
        print(f"S'ha produït un error en ajustar Holt-Winters: {e}")
        return None
```

A.4 Prophet

```python
from prophet import Prophet
import pandas as pd

def ajustar_prophet(train, m=1, d=0):
    """
    Ajusta un model Prophet amb estacionalitat segons el valor de m i diferenciació si cal.

    Arguments:
    - train: Sèrie temporal d'entrenament.
    - m: Període d'estacionalitat.
    - d: Nombre de diferenciacions aplicades abans de l'entrenament.

    Retorna:
    - Model Prophet ajustat.
    """
    df_train = train.reset_index().rename(columns={train.index.name: "ds", train.columns[0]: "y"})

    model = Prophet(yearly_seasonality=True, changepoint_prior_scale=0.05)

    if m == 7:
        model.add_seasonality(name="daily", period=7, fourier_order=5)
    elif m == 12:
        model.add_seasonality(name="monthly", period=12, fourier_order=15)
    elif m == 52:
        model.add_seasonality(name="weekly", period=52, fourier_order=20)

    model.fit(df_train)

    return model

def predir_prophet(model, periods, freq="M", train=None, d=0):
    """
    Genera prediccions amb un model Prophet i assegura que es corresponen amb el test.
    Si s'ha aplicat diferenciació (d > 0), es reintegra la predicció a l'escala original.

    Arguments:
    - model: Model Prophet ajustat.
    - periods: Nombre de períodes a predir.
    - freq: Freqüència de la predicció (per defecte "M" per mensual).
    - train: Conjunt d'entrenament original (necessari per a reintegració si d > 0).
    - d: Nombre de diferenciacions aplicades.
```

```
    Retorna:
    - Sèrie de prediccions amb l'índex corregit.
    """
    future = model.make_future_dataframe(periods=periods, freq=freq)

    forecast = model.predict(future)
    forecast["ds"] = pd.to_datetime(forecast["ds"])

    prediccions = forecast.set_index("ds")["yhat"].iloc[-periods:]
    prediccions.index.name = "data"
    prediccions.index = prediccions.index + pd.offsets.MonthEnd(0)

    if d > 0 and train is not None:
        ultim_valor_train = train.iloc[-1, 0]
        prediccions = prediccions.cumsum() + ultim_valor_train

    return prediccions.rename("Predicció")
```

Apèndix B

Dades

B.1 Conjunt de dades *passengers.csv*

Any	Mes	Vols nacionals
2000	gener	101824
2000	febrer	113734
2000	març	147049
2000	abril	125925
2000	maig	134994
2000	juny	136136
2000	juliol	139786
2000	agost	138962
2000	setembre	142252
2000	octubre	140590
2000	novembre	116807
2000	desembre	106169
...		
2019	gener	140839
2019	febrer	142500
2019	març	186876
2019	abril	201851
2019	maig	191335
2019	juny	202493
2019	juliol	217684
2019	agost	217863
2019	setembre	197404
2019	octubre	191649
2019	novembre	171059
2019	desembre	164973

Taula B.1: Registre mensual del trànsit comercial nacional a l'aeroport de València (2000-2019).

B.2 Conjunt de dades *hipoteques.csv*

Any	Mes	Hipoteques
2003	gener	114571
2003	febrer	115671
2003	març	112951
2003	abril	101474
2003	maig	113010
2003	juny	109578
2003	juliol	110490
2003	agost	89933
2003	setembre	114686
2003	octubre	125353
2003	novembre	107123
2003	desembre	101325

...

Any	Mes	Hipoteques
2018	gener	40717
2018	febrer	38529
2018	març	35331
2018	abril	39060
2018	maig	40950
2018	juny	40644
2018	juliol	39571
2018	agost	38545
2018	setembre	43286
2018	octubre	40537
2018	novembre	40170
2018	desembre	28168

Taula B.2: Registre mensual de les hipoteques concedides a l'estat espanyol (2003 i 2018).

Índex de figures

Índex de taules

Bibliografia

[1] J Scott Armstrong i Fred Collopy. «Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons». A: *International Journal of Forecasting* 8 (1992), pàg. 69 - 80.

[2] Juan Bógalo, Pilar Poncela i Eva Senra. «Understanding fluctuations through Multivariate Circulant Singular Spectrum Analysis». A: *Expert Systems With Applications* 251 (2024), pàg. 123827. DOI: 10.1016/j.eswa.2024.123827. URL: https://doi.org/10.1016/j.eswa.2024.123827.

[3] George E.P. Box, Gwilym M. Jenkins i Gregory C. Reinsel. *Time series analysis: Forecasting and control: Fourth edition*. 2013. DOI: 10.1002/9781118619193.

[4] Peter J. Brockwell i Richard A. Davis. *Introduction to Time Series and Forecasting*. 3rd. Springer International Publishing, 2016. ISBN: 978-3-319-29852-8. DOI: 10.1007/978-3-319-29854-2.

[5] Jason Brownlee. *Introduction to Time Series Forecasting with Python*. v1.9. Machine Learning Mastery, 2020.

[6] Bernard Cazelles et al. «Time-dependent spectral analysis of epidemiological time-series with wavelets». A: (). DOI: 10.1098/rsif.2007.0212.

[7] Robert B Cleveland et al. «STL: A Seasonal-Trend Decomposition Procedure Based on Loess (with Discussion)». A: *Journal of Official Statistics* 6 (1990).

[8] Chris Chatfield. *The analysis of time series: An introduction, sixth edition*. 2016.

[9] Cathy W S Chen. «A review of threshold time series models in finance». A: *Statistics and Its Interface* 4 (2011), pàg. 167 - 181.

[10] David A. Dickey i Wayne A. Fuller. «Distribution of the Estimators for Autoregressive Time Series With a Unit Root». A: *Journal of the American Statistical Association* 74 (366 juny de 1979), pàg. 427. ISSN: 01621459. DOI: 10.2307/2286348.

[11] Ian Goodfellow, Yoshua Bengio i Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[12] James D Hamilton. *Time Series Analysis*. 1994. ISBN: 0691042896.

[13] Timothy O Hodson. «Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not». A: *Geosci. Model Dev* 15 (2022), pàg. 5481 - 5487. DOI: 10.5194/gmd-15-5481-2022. URL: https://doi.org/10.5194/gmd-15-5481-2022.

[14] Charles C Holt. «Forecasting seasonals and trends by exponentially weighted moving averages». A: (). DOI: 10.1016/j.ijforecast.2003.09.015. URL: www.elsevier.com/locate/ijforecast.

[15] Rob J Hyndman i Yeasmin Khandakar. *Journal of Statistical Software Automatic Time Series Forecasting: The forecast Package for R*. Inf. tèc. 2008. URL: http://www.jstatsoft.org/.

[16] Rob J. Hyndman i George Athanasopoulos. «Forecasting: principles and practice, 3rd edition». A: *International Journal of Forecasting* 22 (4 2021). ISSN: 01692070.

[17] Sungil Kim i Heeyoung Kim. «A new metric of absolute percentage error for intermittent demand forecasts». A: *International Journal of Forecasting* 32 (2016), pàg. 669 - 679. DOI: 10.1016/j.ijforecast.2015.12.003. URL: http://creativecommons.org/licenses/by/4.0/.

[18] William H. Kruskal i W. Allen Wallis. «Use of Ranks in One-Criterion Variance Analysis». A: *Journal of the American Statistical Association* 47 (260 1952). ISSN: 1537274X. DOI: 10.1080/01621459.1952.10483441.

[19] Michael Kutner et al. *Applied Statistical Linear Models*. 2005.

[20] G. M. Ljung i G. E.P. Box. «On a measure of lack of fit in time series models». A: *Biometrika* 65 (2 1978). ISSN: 00063444. DOI: 10.1093/biomet/65.2.297.

[21] J Humberto Lopez. «The power of the ADF test». A: *Economics Letters* 57 (1997), pàg. 5 - 10.

[22] Spyros Makridakis et al. «The M6 forecasting competition: Bridging the gap between forecasting and investment decisions». A: *International Journal of Forecasting* (2024). DOI: 10.1016/j.ijforecast.2024.11.002. URL: http://creativecommons.org/licenses/by/4.0/.

[23] V Mayer-Schönberger i K Cukier. *Big Data: A Revolution that We Transform How We Live, and Think*. 2013.

[24] Andrew V. Metcalfe i Paul S.P. Cowpertwait. *Introductory Time Series with R*. Springer New York, 2009. DOI: 10.1007/978-0-387-88698-5.

[25] J. N. i Herman Wold. «A Study in Analysis of Stationary Time Series.» A: *Journal of the Royal Statistical Society* 102 (2 1939). ISSN: 09528385. DOI: 10.2307/2980009.

[26] Klaus Neusser. *Springer Texts in Business and Economics*. Inf. tèc. URL: http://www.springer.com/series/10099.

[27] Daniel Peña. *Análisis de series temporales*. 2a ed. Alianza, 2010. ISBN: 9788420669458.

[28] Joaquin Amat Rodrigo i Javier Escobar Ortiz. *skforecast*. Febr. de 2024. DOI: 10.5281/zenodo.8382788. URL: https://skforecast.org/.

[29] Kady Sako, Berthine Nyunga Mpinda i Paulo Canas Rodrigues. «Neural Networks for Financial Time Series Forecasting». A: (2022). DOI: 10.3390/e24050657. URL: https://doi.org/10.3390/e24050657.

[30] Sean J. Taylor i Benjamin Letham. «Forecasting at Scale». A: *American Statistician* 72 (1 gen. de 2018), pàg. 37 - 45. ISSN: 15372731. DOI: 10.1080/00031305.2017.1380080.

[31] Xiaozhe Wang i Rob Hyndman. «Characteristic-Based Clustering for Time Series Data». A: *Data Mining and Knowledge Discovery* 13 (2006), pàg. 335 - 364. DOI: 10.1007/s10618-005-0039-x.

[32] Peter R. Winters. ≪Forecasting Sales by Exponentially Weighted Moving Averages≫. A: *Management Science* 6 (3 1960). ISSN: 0025-1909. DOI: 10.1287/mnsc.6.3.324.

[33] Alain Zemkoho. ≪A Basic Time Series Forecasting Course with Python≫. A: *Operations Research Forum* 4 (1 des. de 2022), pàg. 2. ISSN: 2662-2556. DOI: 10.1007/s43069-022-00179-z.