TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Models de xarxes neuronals versus models lineals per la predicció de sèries temporals

Autor: Clara Zaldívar Villafranca

Director: Ernest Pons Fanals

Departament: Econometria, Estadística i Economia Aplicada

Convocatòria: Juny 2020

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resum i paraules clau

La predicció de valors futurs és una tècnica que ha causat un interès creixent en molts sectors, per això, s'estan estudiant nous mètodes per tal de poder conèixer les situacions futures de la manera més exacta possible. En aquest treball de fi de grau, es realitzarà una comparativa entre dos mètodes de predicció per dades de sèries temporals, concretament, dades que fan referència al turisme a Espanya des de l'any 2000 fins l'actualitat. Per això s'utilitzaran els models ARIMA i les Xarxes Neuronals. Per fer-ho, s'ha fet una àmplia recerca dels mètodes i un anàlisi de les dades que s'estudien. S'han implementat els mètodes adaptant-los a les dades i tot seguit, s'ha realitzat la comparativa entre els mètodes per veure quin s'ajusta millor a la sèrie i finalment s'ha fet un anàlisi de l'impacte del coronavirus al turisme espanyol.

Les conclusions principals obtingudes mostren que tant els mètodes ARIMA com les xarxes neuronals són vàlids per la predicció de sèries temporals. Per les dades utilitzades els mètodes que millor s'adapten a les dades són els mètodes ARIMA però aquesta conclusió és aplicable únicament per aquestes dades. També s'ha arribat a la conclusió de que cap dels mètodes és capaç de predir un fet extraordinari com ha estat la crisi del coronavirus.

Paraules clau: predicció, sèries temporals, ARIMA, auto.arima, intel·ligència artificial, xarxes neuronals, NNAR, LSTM, turisme, COVID-19.

Abstract and key words

Forecasting future values is a technique that has caused a growing interest in many sectors, so new methods are being studied in order to know future situations as accurately as possible. In this thesis, a comparison will be made between two prediction methods for time series data, specifically, data that refer to tourism in Spain from 2000 to the present. This is why ARIMA models and Neural Networks will be used. To do this, a thorough search of the methods and an analysis of the data under study has been done. The methods have been implemented by adapting them to the data and then the comparison between the methods has been made to see which one best fits the series and finally an analysis of the impact of the coronavirus on the Spanish tourism.

The main conclusions obtained show that both ARIMA methods and neural networks are valid for the prediction of time series. For the data used, the methods that best fit the data are the ARIMA methods, but this conclusion is applicable only to this data. It has also been concluded that none of the methods is capable of predicting an extraordinary event such as the coronavirus crisis.

Key words: prediction, time series, ARIMA, auto.arima, artificial intelligence, neural networks, NNAR, LSTM, tourism, COVID-19.

Agraïments

Abans de començar el treball m'agradaria agrair a l'Ernest Pons, el meu tutor del treball, la seva ajuda, paciència i sobretot la seva implicació. No hauria pogut tenir un tutor millor per un treball tant important com aquest.
També m'agradaria agrair a la meva família per sempre ajudar-me i confiar en mi en tots moments. Per últim, m'agradaria agrair a la meva millor amiga i companya de carrera Cristina per donar-me suport incondicional aquests quatre anys de carrera i els que venen a partir d'ara.

Classificació AMS

37M10 Time series analysis

62M10 Time series, auto-correlation, regression, etc.

62M20 Prediction

62M45 Neural nets and related approaches

92B20 Neural networks, artificial life and related topics

ÍNDEX

ÍNDEX DE TAULES

ÍNDEX DE FIGURES

1. INTRODUCCIÓ

La realització de prediccions per tal de determinar un possible succés futur ha incrementat la seva popularitat al llarg dels anys, ja sigui en l'àmbit econòmic, comercial o mèdic, entre d'altres. Com a conseqüència han anat apareixent nous mètodes i tècniques per tal de realitzar aquesta tasca.

Existeixen molts mètodes per l'anàlisi d'una sèrie temporal, però un dels mètodes més senzills i coneguts és el de *Box i Jenkins* (1970), més conegut com a **Models Autoregressius Integrats de Mitjana Mòbil (ARIMA).** En moltes ocasions s'ha demostrat que aquest mètode univariant és molt eficaç. El problema del mètode ARIMA és que assumeix una estructura lineal de la sèrie temporal (els seus paràmetres són lineals, només poden funcionar si la seva relació és lineal o lineal integrada) i a vegades no es compleix aquesta suposició. Un altre aspecte a remarcar, és que la metodologia ARIMA requereix que la sèrie temporal sigui estacionària. Per vèncer aquestes limitacions existeixen les **Xarxes Neuronals Artificials (ANN)**, aquestes no tenen cap suposició sobre el procés a partir del qual s'ha generat la sèrie temporal i per tant, s'ha demostrat que són útils tant per processos lineals com per no lineals i tant per processos estacionaris com per no estacionaris. Tot i que les Xarxes Neuronals tenen alguns inconvenients, com per exemple que solen ser més costoses computacionalment i poden resultar més difícils d'implementar, són un camp relativament nou i molt extens. En aquest treball s'utilitzaran dos tipus de Xarxes neuronals, les **Xarxes Neuronals Auto-regressives (NNAR)** i les **Long-Short Term Memory (LSTM)**.

Aquest treball té l'objectiu de comparar els models ARIMA, treballats abastament en el Grau, amb els models de Xarxes Neuronals, que en el programa del Grau s'aborden de manera molt bàsica. Un dels principals reptes ha estat aprofundir en el coneixement i implementació d'aquest segon grup de models.

L'exercici comparatiu utilitza sèries de dades de turisme a Espanya entre l'any 2000 fins l'actualitat. A partir del març de 2020 es produeix un fet excepcional que amb tota seguretat modificarà el comportament turístic al país i al món, es tracta de la pandèmia del COVID-19. Considerant aquestes circumstàncies s'ha decidit treballar les dades de forma separada, és a dir, abans i post COVID-19. El fet d'emprar les dades post COVID-19 no estava inclòs en els objectius inicials del treball, sinó que és un treball addicional que pot resultar interessant per futurs treballs encarats a l'avaluació de l'efecte del coronavirus en el sector turístic.

La primera fase del treball inclou el resum dels mètodes a comparar amb especial dedicació a la recerca i estudi detallats dels models de Xarxes Neuronals, models que estan a l'alça des

de fa uns anys i que s'apliquen en molts camps de la ciència i la tecnologia. A més, s'estableixen els criteris de validació que s'utilitzaran per comparar. La segona fase consisteix en un anàlisi exploratori de les dades, la implementació i comparació dels mètodes i finalment un breu anàlisi de l'impacte del COVID-19.

2. METODOLOGIA

A continuació es detalla la teoria de cadascun dels mètodes, separant els ARIMA de les Xarxes Neuronals.

2.1 Metodologia ARIMA

Box i Jenkins (1970)[1] van desenvolupar models que tenen en compte la dependència entre les dades, és a dir, cada observació en un moment donat és modelada en funció dels valors anteriors.

Per treballar amb models ARIMA cal conèixer una sèrie de conceptes bàsics:

- Un ***procés estocàstic*** és un conjunt de variables aleatòries associades a diferents instants del temps. Una sèrie temporal és la realització d'un procés estocàstic.

- El ***soroll blanc*** és una successió de variables aleatòries que es caracteritzen per tenir una esperança constant i igual a zero, igual variància, i a més a més, són independents al llarg de el temps (covariància és zero).

- Un ***camí aleatori*** és un procés estocàstic que es caracteritza perquè la seva primera diferència és un soroll blanc.

- Un ***procés estocàstic*** és ***ergòdic*** si és possible estimar de manera consistent les seves característiques a partir d'una realització seva.

Com s'ha mencionat anteriorment, una de les suposicions principals de la metodologia ARIMA és que les dades han de ser estacionàries, és a dir, la mitjana i la variància han s'han de mantenir constants al llarg del temps i les dades no poden presentar tendència, creixent ni decreixent. En el cas de que alguna d'aquestes suposicions no es compleixi, s'han de transformar les dades, normalment mitjançant diferències o transformació logarítmica.
Per mirar si alguna d'aquestes suposicions no es compleix, cal mirar els gràfics d'autocorrelacions: **Funció d'Autocorrelació Simple (FAS)** i **Funció d'Autocorrelació Parcial (FAP).**

Els models ARIMA es poden separar en tres components **AR** (Autoregressiu), **I** (Integrat) i **MA** (Mitjana mòbil).

[1] George E.P. Box, Gwilym. M. Jenkins (1970). *Time Series Analysis, Forecasting ans Control.* San Francisco: Holden-Day.

El component Autoregressiu, conegut com *AR(p)*, és una combinació lineal de p valors passats de la variable més un terme d'error, que té com a forma general:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

i que mitjançant l'operador retard (B), també es coneix de la forma:

$$\phi(\mathrm{B}) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \qquad \phi(\mathrm{B})\mathrm{y}_t = \delta + \varepsilon_t$$

Un procés autoregressiu és estacionari si les arrels del polinomi B: ($1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$) cauen fora del cercle d'unitat. A més, un procés autoregressiu sempre serà invertible.

El component de Mitjanes Mòbils, conegut com *MA(q),* en que el valor actual es pot predir a partir de la component aleatòria d'aquell moment i en menor mesura, dels impulsos aleatoris anteriors. El model *MA(q)* té la següent forma general:

$$y_t = \delta + \varepsilon_t - \theta_1 y_{t-1} - \theta_2 y_{t-2} - \cdots - \theta_q y_{t-q}$$

Utilitzant l'operador retard (B):

$$\theta(\mathrm{B}) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \qquad y_t = \delta + \theta(\mathrm{B})\varepsilon_t$$

Un procés de mitjanes mòbils, a diferència dels processos autoregressius, sempre és estacionari i es invertible si les arrels del polinomi B: $(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$ cauen fora del cercle d'unitat.

Si es combinen els processos Autoregressius *AR(p)* i els de Mitjanes Mòbils *MA(q)* obtenim un model ***ARMA(p,q)*** que tenen com a forma general:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 y_{t-1} - \theta_2 y_{t-2} - \cdots - \theta_q y_{t-q}$$

Utilitzant l'operador retard (B):

$$\phi(\mathrm{B})\mathrm{y}_t = \delta + \theta(\mathrm{B})\varepsilon_t$$

Un procés *ARMA (p,q)* serà estacionari si la seva component autoregressiva ho és, i serà invertible si la seva component de Mitjanes Mòbils ho és.

Si s'estén un model *ARMA* per permetre que s'incloguin arrels unitàries es fa referència als models *ARIMA* d'ordre *(p,d,q)* on d és el nombre de vegades que s'aplica l'operador diferencia:

$$\phi_p(\mathrm{B})(1 - B)^d \mathrm{y}_t = \mathrm{c} + \theta_q(\mathrm{B})\varepsilon_t$$

On:

$$\phi(\mathrm{B}) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \text{ (Polinomi AR(p))}$$
$$\theta(\mathrm{B}) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \text{(Polinomi (MA)(q))}$$

Els models *ARIMA* són models estocàstics que utilitzen variacions i regressions de dades amb la finalitat de trobar "patrons" per així poder realitzar una predicció futura. Es tracta d'un mètode en que les estimacions futures venen explicades per les dades del passat i no per variables independents.

En els models *ARIMA,* els valors estimats es suposa que són una combinació lineal dels valors passats i dels errors passats.

Sovint, les sèries temporals presenten components estacionals, és a dir, el valor de la sèrie en un moment de temps determinat, vindrà influenciat per aquell període de temps. Per tractar aquestes sèries es poden utilitzar dos opcions:

1. Considerar oscil·lacions estacionals deterministes.
2. Considerar oscil·lacions estacionals estocàstiques.

En aquest estudi es treballarà amb un anàlisi estocàstic, per tant, es consideraran el segon tipus d'oscil·lacions. Per fer-ho, es definirà amb una "s" com el número d'observacions incloses en el cicle estacional complet. Per exemple, si es tractessin amb dades mensuals s=12.

Aquests tipus de processos que contenen un component estacional s'anomenen *SARIMA*
i permeten captar de manera simultània la component "regular" (dependència respecte dels valors adjacents) i la component estacional (dependència respecte de les mateixes estacions d'anys diferents).
Aquest model es denomina com: ***SARIMA(p,d,q)x(P,D,Q)s***. On els primers tres paràmetres fan referència a la part regular i els tres següents a la part estacional. Aquests models tenen la següent forma general:

$$\phi(\mathrm{B})\Phi_s(B)(1-B)^d(1-B^s)^D \mathrm{y}_t = \delta + \theta(\mathrm{B})\Theta_s(\mathrm{B})\varepsilon_t$$

On:

$$\Phi_P B^s = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_2 B^{Ps} \text{ (Polinomi AR(p))}$$
$$\Phi_Q B^s = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_2 B^{Qs} \text{ (Polinomi MA(q))}$$
$$\nabla_S = 1 - B^s \text{(Operador diferència estacional)}$$

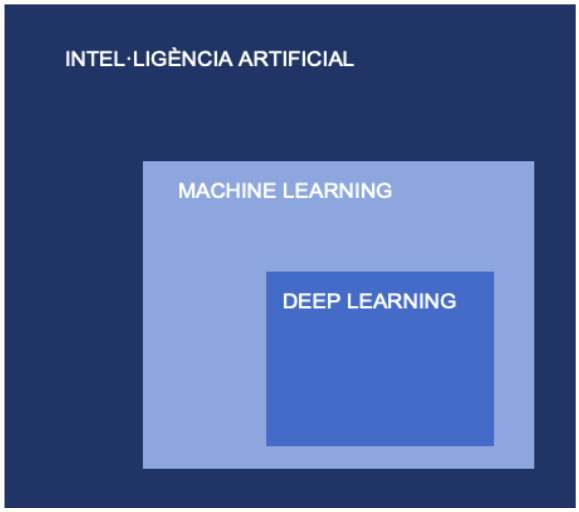
Per la implementació d'aquesta metodologia s'utilitzaran un seguit de passos proposats per *Box i Jenkins* que es detallaran posteriorment. A més a més, també s'utilitzarà dos models diferents per la implementació de la metodologia ARIMA, el model manual i el model automàtic.

Ambdós models s'explicaran durant el cos del treball amb els criteris que s'han aplicat a cadascun.

2.2 Models de Xarxes Neuronals Artificials

Per començar a parlar de Xarxes Neuronals, cal parlar primer d'Intel·ligència Artificial, *Machine Learning i Deep Learning*.

Figura 1 - Introducció a la Intel·ligència Artificial



Font: Elaboració Pròpia

La ***Intel·ligència Artificial*** sorgeix a partir d'alguns treballs publicats a la dècada de 1940 però que no van tenir gran ressò. A partir de 1950 amb l'influent treball d'Alan Turing[2] s'obre una nova disciplina de les ciències de la informació. La intel·ligència artificial estudia les diferents formes en què una màquina interactua amb el món que l'envolta. L'objectiu de la intel·ligència artificial és entrenar a una màquina per tal de resoldre feines que són fàcils pels humans però difícils per les màquines.
A la dècada dels 80 es va començar a treballar sobre aquest camp d'una altra manera. Es va arribar a la conclusió de que en comptes de programar màquines per a que siguin intel·ligents, se'ls donés accés a una gran quantitat de dades i que aquestes trobessin patrons per aprendre per si soles i actuar. Aquest concepte s'anomena ***Machine Learning*** i és una forma analítica de resoldre problemes mitjançant la identificació, classificació o predicció. Per tant, el *Machine Learning* és un subconjunt de la intel·ligència artificial però en aquest cas els algorismes aprenen de les dades que s'han introduït i utilitzen aquest coneixement per treure conclusions sobre dades noves que se li introdueixin.

A partir d'aquesta última tecnologia, l'any 2011 apareix un nou concepte, el ***Deep Learning.*** Aquesta nova tecnologia en termes conceptuals és molt semblant al *Machine Learning* però utilitza altres algorismes. Mentre que el *Machine Learning* utilitza tècniques com la regressió lineal i logística, el *clustering* o els arbres de decisió, el *Deep Learning* utilitza un algorisme que s'encarrega de que la màquina aprengui a classificar la informació de la mateixa manera que el cervell humà, les ***Xarxes Neuronals***.

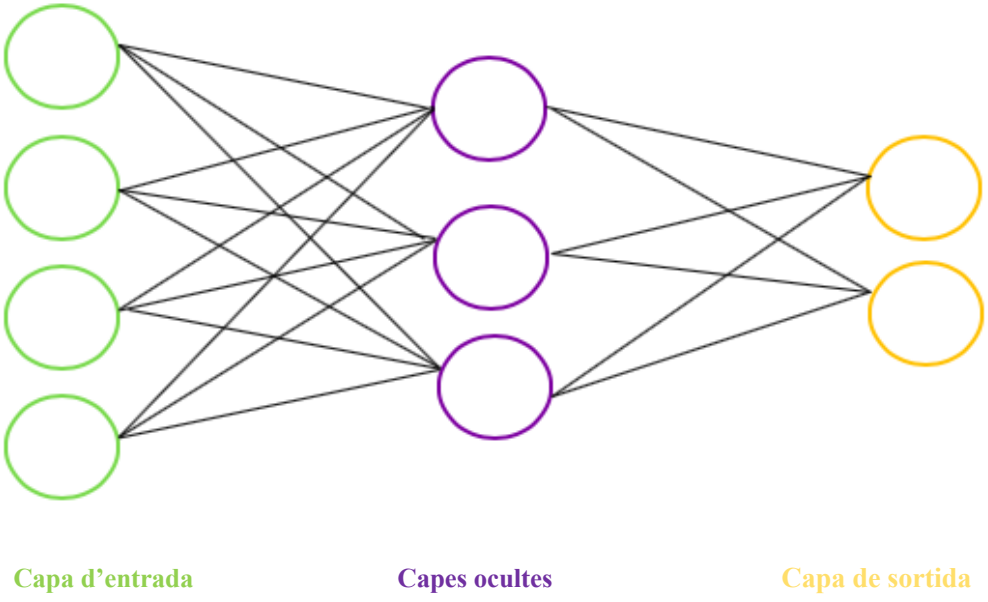
Les ***Xarxes Neuronals*** són models que s'inspiren en el comportament de les neurones i les connexions cerebrals per resoldre problemes de tot tipus. Des de fa uns anys s'ha començat a

[2] A.M, Turing. (Oct 1950). *Computing Machinery and Intelligence.* (Vol. 59). Mind, New Series.

utilitzar de manera massiva (reconeixement de caràcters, imatges, veu, generació de text, prevenció de frau…)

L'estructura bàsica d'una Xarxa Neuronal és la següent:

Figura 2 - Estructura Bàsica d'una Xarxa Neuronal

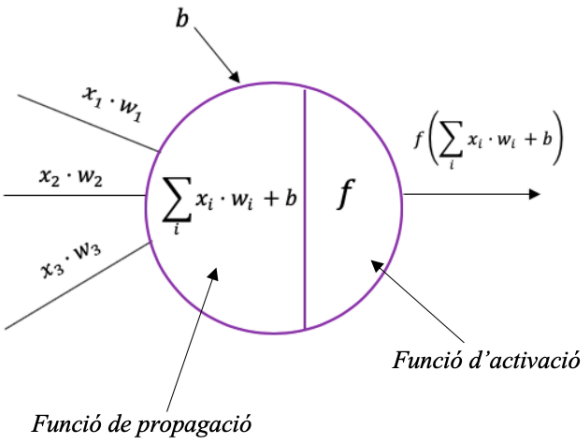


Font: Elaboració Pròpia

La unitat bàsica de processament de la xarxa neuronal és la **neurona**. Aquestes neurones tenen unes connexions d'entrada d'on es rep informació des de l'exterior (**Valors d'entrada o inputs)**, aquesta informació entra per la **capa d'entrada** i pot ser tant binari com continu. Amb aquests valors d'entrada la neurona dins de la **capa o capes ocultes** realitzarà un "càlcul intern" que acabarà traient per la **capa de sortida** un **valor de sortida (*output*).**

Per entendre el funcionament d'una única neurona s'utilitzarà el següent gràfic:

Figura 3 - Funcionament d'una neurona



Font: Elaboració Pròpia

La neurona rep entrades de totes les neurones que estan a la capa anterior i genera una sortida. Cada connexió que rep la neurona està formada per un valor *x* (Valor d'entrada o input) i un valor *w* que fa referència al pes. Cada connexió que arriba a la neurona tindrà associat un pes que servirà per definir amb quina intensitat cada variable d'entrada (*x*) afecta a la neurona. A més a més, existirà un altre valor que fa referència al biaix (*b*). Aquest sumatori s'anomena ***funció de propagació***.

Aquesta funció de propagació passarà per una funció *f* que s'anomena ***funció d'activació*** [3] i aquesta s'encarrega de determinar la sortida que generarà la neurona en funció de la seva entrada. Té com a objectiu acotar els valors de sortida de la neuronal per mantenir-los en certs rangs. Aquesta funció dependrà del tipus de xarxa neuronal amb la què es treballi, existeixen funcions lineals i no lineals. Les funcions d'activació més conegudes són: funció lineal, funció d'activació sigmoide o logística, reLU i tangent hiperbòlica.

Una vegada definida la xarxa, caldrà entrenar-la amb l'objectiu de minimitzar la funció de d'error de les prediccions o més coneguda com **funció de cost**. Aquest procés es du a terme a través de l'algoritme d'optimització del **descens del gradient** [4] que s'encarrega del càlcul dels gradients i juntament amb l'algorisme de propagació enrere o més conegut com ***backpropagation*** [5].

Aquest algorisme de *backpropagation* s'encarrega d'ajustar de manera equitativa els pesos de les connexions de la xarxa. Al calcular les derivades del descens del gradient de la funció de cost respecte a tots els paràmetres, normalment sorgeix un problema ja que hi ha milions de paràmetres en una xarxa neuronal.

L'algorisme *backpropagation* pot ser interpretat com una eina que permet propagar i aproximar les derivades al llarg de la xarxa en base als paràmetres d'aquesta. Aquest algorisme ha permès que es puguin entrenar models molt complexos. Aquest algoritme representa l'error d'una xarxa des de la sortida a l'entrada.

Aquest és el funcionament més bàsic d'una xarxa neuronal, però es tracta d'un camp molt més extens on existeixen molts models i molts més factors a tenir en compte. En aquest estudi es tractaran dos tipus de xarxes neuronals (*feedforward neural networks* i *recurrent neural networks*) que són dos possibles models de xarxes que sembla que s'adapten de manera eficient a les dades de sèries temporals i que es poden ajustar de manera correcta a les dades de l'estudi.

---

[3] Sharma, S. (6 / Set / 2017). *Activation Functions in Neural Networks*. Recollit de Towards Data Science: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

[4] *Gradient Descent*. (s.d). Recollit de Machine Learning Glossary: https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html

[5] Nielsen, M. A. (2015). Neural Networks and Deep Learning Chapter 2 How the Backpropagation Algorith Works. Determination Press.

2.2.1 Xarxes Neuronals Autoregressives (NNAR)

Aquest tipus de Xarxes Neuronals, més conegudes com *Neural Networks Autoregression Model (NNAR)* són un model de xarxes que formen part de les *Feedforward Neural Networks*, aquestes últimes són la classe de xarxes neuronals més simples, en la que la informació només és mou en una direcció: entra pels nodes d'entrada passa pels nodes ocults (si es que hi ha) i surt per els nodes de sortida. És a dir, segueix la estructura més bàsica que s'ha mostrat en la figura anterior ***(Figura 2)***. En aquest model, els valors retardats d'una sèrie temporal s'utilitzen com a entrada del model.

El model NNAR per tant, utilitza també la combinació lineal anomenada funció de propagació definida anteriorment ***(Figura 3)*** i una funció d'activació, que normalment és la sigmoide. Segueix els passos mencionats anteriorment i els pesos de la xarxa s'actualitzen mitjançant l'algorisme de *backpropagation* també mencionat.

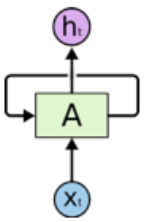
2.2.2 Long-Short Term Memory (LSTM)

Abans de parlar d'aquest model, cal conèixer les **Xarxes Neuronals Recurrents**.

Les Xarxes Neuronals Recurrents són una de les tecnologies més poderoses que hi ha actualment en el món de la intel·ligència artificial. Aquestes xarxes són tan poderoses perquè a diferència d'una xarxa neuronal simple o bàsica, aquestes tenen la capacitat de que la informació no es perdi al llarg del temps. És a dir, tenen memòria.

L'estructura principal d'una Xarxa Neuronal Recurrent és la següent:

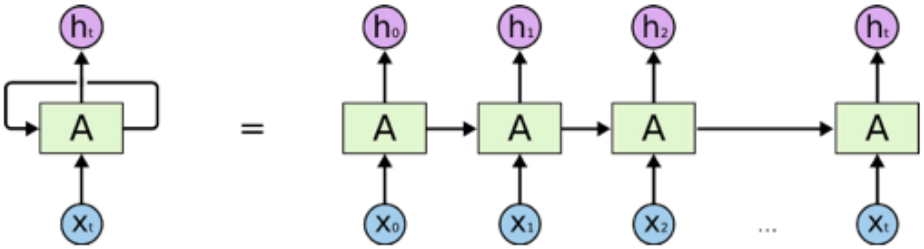
Figura 4 - Estructura d'una Xarxa Neuronal Recurrent



Font: Cristopher Olah, 2015 – Understanding LSTM Networks

On A es una part de la Xarxa Neuronal, x_t és l'entrada i la xarxa produeix un valor h_t i es permet la transferència d'una xarxa a l'altra mitjançant un bucle. Si s'ampliés la figura anterior s'obtindria la següent:

Figura 5 - Estructura d'una Xarxa Neuronal Recurrent Ampliada



Font: Cristopher Olah, 2015 – Understanding LSTM Networks

En aquesta última figura, el que es pot veure és que les neurones s'alimenten a elles mateixes al llarg del temps, i amb això el que s'aconsegueix és, com s'ha mencionat anteriorment, una memòria a curt termini, ja que les neurones poden recordar quin ha sigut el pas anterior. Aquest tipus de xarxes neuronals són molt útils quan es treballa amb seqüències ja que permet a les neurones passar informació a elles mateixes en un futur i analitzar-la.
En aquest estudi, es treballarà amb sèries temporals, i per tant, treballar amb xarxes neuronals recurrents pot ser una bona opció ja que per l'estudi de sèries temporals és important tenir els esdeveniments passats per poder predir esdeveniments futurs.

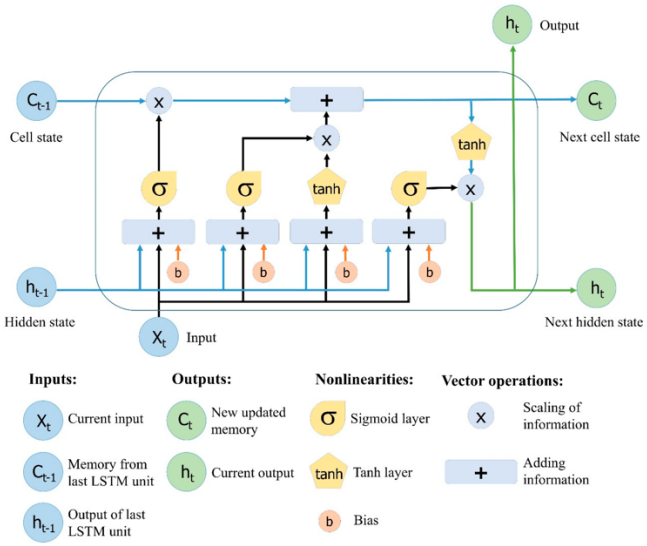
Tot i això s'ha vist que aquestes xarxes tenen ha certs problemes amb la memòria a llarg termini. Els problemes principals d'aquestes xarxes són el *exploding gradient* i *vanishing gradient.*
El problema del *exploding gradient* es dona quan durant l'entrenament de la xarxa es forma un increment excessiu del gradient que acaba provocant un error en el sistema, una solució per aquest problema pot ser marcar un umbral que determini el valor màxim que pot prendre una variable. El *vanishing gradient*, és el contrari, i els components decreixen de manera molt rapida cap a zero, creant una gran dificultat per aprendre correlacions en que la diferència de la informació rellevant i el punt en que es necessiti sigui molt gran, és a dir, creant problemes en quant a memòria a llarg termini.

Per solucionar el problema del *vanishing gradient*, apareixen les ***Long-Short Term Memory*** **(LSTM)**. Aquest tipus de Xarxes funcionen de manera similar a les Xarxes Neuronals Recurrents però permeten una memòria a curt i llarg termini. Es podria dir que les LSTM estan formades per un conjunt de cel·les connectades les unes amb les altres.

A mode resum, les cel·les poden ser vistes com una unitat que pren com a entrada l'estat previ i l'entrada actual. Internament aquestes cel·les decideixen quina informació es manté a la

memòria i quina s'oblida. Després es combina l'estat previ, l'entrada actual i la informació que hi ha a la memòria actual. Així, aquestes xarxes són molt eficients a l'hora de trobar correlacions entre esdeveniments molt llunyans en el temps.

Figura 6 - Estructura d'una Xarxa Long-Short Term Memory



Font: Xuan-Hien Le, Hung Viet Ho, Giha Lee i Sungho Jung, 2019 - Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting

En aquest treball, s'ha decidit no entrar en profunditat dins de l'explicació concreta i detallada. Per més detalls d'aquest mètode, Cristopher Olah va crear l'any 2015 un article molt popular anomenat Understanding LSTM Networks[6].

2.3 Mesures de validació

Una vegada s'hauran implementat els mètodes, caldrà realitzar la comparativa entre ells, per això s'utilitzaran diferents les diferents mesures de validació.

- **Criteri d'informació d'*Akaike* (AIC).**

L'AIC és una mesura de la bondat d'ajust d'un model estadístic. Es pot dir que aquest criteri mesura la relació entre el biaix i la variància del model, o més generalment, mesura l'exactitud i complexitat del model.

6 Olah, C. (2015). *Understanding LSTM Networks*. Recollit de https://colah.github.io/posts/2015-08-Understanding-LSTMs/

$$AIC = 2k - 2\ln(L)$$

On

- k és el número de paràmetres del model
- $\ln(L)$ és el logaritme de la funció de versemblança pel model estadístic

El criteri AIC imposa una penalització per afegir repressores o variables independents. No obstant això, és important tenir en compte que aquest criteri no busca el model correcte, ja que parteix de la premissa que el model veritable pot no estar dins del conjunt de models a avaluar i per tant, el seu objectiu és seleccionar el model que proporcioni les millors prediccions.

Quan es comparen dos models, s'escull aquell que té un valor de l'AIC més petit.

Els següents mètodes de comparació, s'utilitzen una vegada ja s'han realitzat les prediccions. Aquests mètodes s'utilitzen normalment per avaluar la capacitat predictiva d'un model, aquesta, es realitza mitjançant la quantificació dels errors de predicció.

Error = Valor Real - Predicció

- **Error Quadràtic Mig (EQM)**

Aquest estimador, mesura la mitjana de la suma dels errors al quadrat, la diferència entre l'estimador i el que s'estima. l'EQM castiga aquells períodes on la diferència d'errors és més alta que a d'altres períodes.

$$EQM = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}$$

- **Error Absolut Mitjà (EAM)**

Aquest estimador mesura la mitjana de la suma dels errors en valor absolut.

$$EAM = \sum_{i=1}^{n} \frac{|\hat{y}_i - y_i|}{n}$$

El problema d'aquests estimadors és que són adimensionals, és a dir, depenen de les unitats de mesura de la variable que és objecte d'estudi i per tant, per mesurar la capacitat predictiva d'un model, és necessari un estimador que sigui adimensional, aquí és on apareix l'Error Percentual Absolut Mitjà.

- **Error Percentual Absolut Mitjà (EPAM)**

Aquest estimador mesura la mida de l'error absolut en termes percentuals. És la mitjana de l'error absolut o diferència entre la demanda real i la predicció, expressat com un percentatge dels valors reals.

$$EPAM = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| x\ 100$$

Aquest estimador té la següent interpretació:

- **EPAM < 1%:** Molt bona capacitat predictiva
- **1% < EPAM < 3%:** Bona capacitat predictiva
- **3% < EPAM 5%:** Capacitat predictiva regular
- **5%< EPAM:** Capacitat predictiva baixa/molt baixa

- **Root Mean Squared Error (RMSE)**

Aquest estimador es defineix com l'arrel quadrada de l'Error Quadràtic Mitjà:

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

Com l'arrel quadrada d'una variància, RMSE es pot interpretar com la desviació estàndard de la variància inexplicada, i té la propietat útil d'estar en les mateixes unitats que la variable de resposta. Els valors més baixos d'RMSE indiquen un millor ajust. RMSE és una bona mesura de la precisió amb que el model prediu la resposta

2.4 Dades

Per posar a prova l'estudi s'utilitzaran dades sobre els **moviments turístics en frontera**[7] des del gener del 2000 fins l'actualitat.

[7] INE. (s.d). *Estadística de movimientos turísticos en frontera. Frontur*. Recollit de Instituto Nacional de Estadística: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176996&menu=resultados&secc=1254736195367&idp=1254735576863#!tabs-1254736195367

Aquestes dades les va començar a publicar l'Institut d'Estudis Turístics (*Frontur*)[8] i se'n van fer càrrec fins el mes de setembre de 2015 (inclòs). A partir de llavors l'Institut Nacional d'Estadística (INE)[9] s'encarrega de penjar les dades de manera mensual.

L'Institut Nacional d'Estadística amb la recollida d'aquestes dades té l'objectiu de crear una enquesta contínua amb la finalitat de proporcionar estimacions mensuals i anuals del nombre de visitants no residents a Espanya que arriben al país, així com les característiques dels viatges que realitzen (via d'accés, destí, país de residència, motiu...)[10]

Per recollir aquestes dades es realitzen enquestes de caràcter mensual mitjançant una entrevista directa en els punts de sortida d'Espanya (ports, aeroports, fronteres..).
La unitat de mesura per aquestes dades són els visitants (turistes i excursionistes)

Es defineix ***turista*** com aquella persona que realitza un viatge turístic amb almenys una pernoctació.
Un ***viatge*** és un desplaçament fora del municipi de residència per qualsevol motiu, en el que es pernocta al menys una nit i que finalitza en el període de referència.

Un ***visitant*** es aquella persona que viatja a un destí principal diferent dels seu entorn habitual amb una durada inferior a un any amb qualsevol finalitat principal (oci, negocis o un altre motiu personal) que no sigui empleat d'una entitat resident del país o del lloc visitat. Aquests viatges realitzats pels visitants s'anomenen ***viatges turístics***.

En referència a la població estadística, la població que es objecte d'estudi esta formada per: persones no residents a Espanya que entren o surten del país realitzant o no pernoctació i persones no residents a Espanya que passen pel país en trànsit.

L'àmbit geogràfic s'estén a tot el territori nacional i com s'ha mencionat anteriorment, la mostra es recull en els principals punts d'accés dels viatgers no residents, tant per carretera, aeroports, ports i ferrocarrils.

Per l'estudi en qüestió s'utilitzarà únicament el nombre total de turistes que entren a Espanya. Aquestes dades consten d'una variable (Total de turistes) de 240 observacions (gener 2000 - desembre de 2019). Una vegada s'hagi escollit el millor mètode per la predicció d'aquesta sèrie, es farà ús de les dades de gener a abril de 2020 per calcular l'impacte que ha tingut la crisi del coronavirus en el turisme.

[8]TURESPAÑA. (s.d). *Frontur*. Recollit de http://estadisticas.tourspain.es/es-ES/estadisticas/frontur/Paginas/default.aspx

[9] INE. (s.d). *Instituto Nacional de Estadística*. Recollit de https://www.ine.es

[10] INE. (s.d). *Estadística de movimientos turísticos en frontera. Frontur. Metodologia*. Recollit de INE: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176996&menu=metodologia&idp=1254735576863

2.5 Recursos Informàtics

Per la realització d'aquest estudi s'ha decidit treballar principalment amb l'*Rstudio*, l'eina més coneguda i utilitzada durant tot el Grau d'Estadística. S'ha utilitzat tant per la majoria dels gràfics com per la modelització. Per el mètode *Long-Short Term Memory*, s'ha utilitzat el llenguatge de programació *Python* a través de l'*Rstudio*. Això és possible gràcies al paquet *reticulate* i d'aquesta manera es poden combinar els dos llenguatges. Per implementar LSTM s'han utilitzat dos llibreries de *Deep Learning* anomenades *Keras*[11] i *TensowFlow*[12]. Aquestes llibreries són de codi obert i estan escrites en *Python*. Per últim, s'ha utilitzat l'Excel per realitzar un gràfic dels errors percentuals.

Durant el Grau d'Estadística no s'ha treballat en cap moment amb *Python* ni ninguna de les seves llibreries. Però s'ha decidit que valia la pena dedicar temps a la formació d'aquest llenguatge de programació ja que en el camp de la intel·ligència artificial, *Python* té moltes més eines que l'R i a més la majoria de documentació de la implementació dels diferents models sol estar en aquest llenguatge.

[11] Keras. (s.d). *Keras*. Recollit de https://keras.io

[12] TensorFlow. (s.d). *TensorFlow*. Recollit de https://www.tensorflow.org

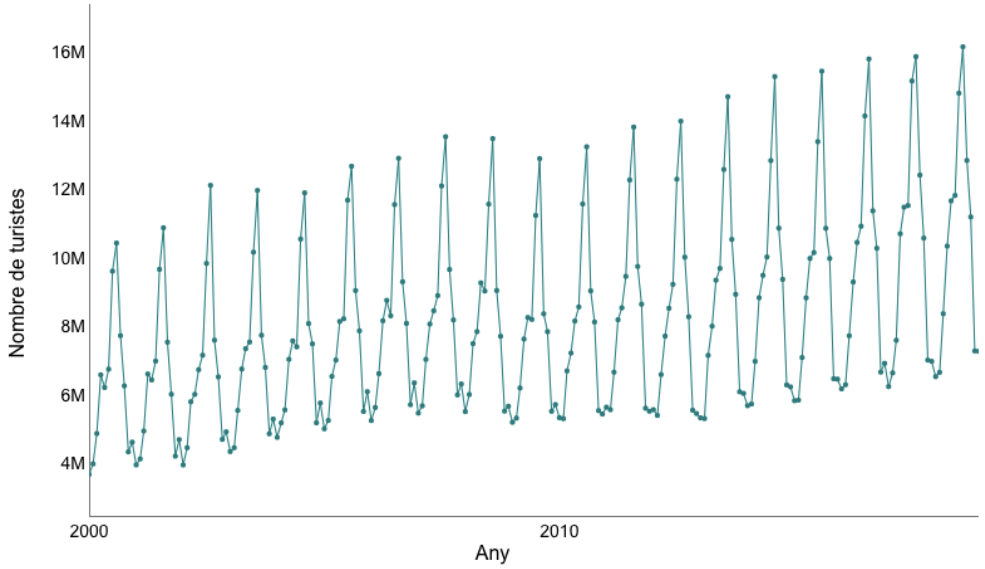
3. COS DEL TREBALL

En aquest apartat es procedirà a posar en pràctica la teoria que s'ha explicat anteriorment i també es faran indicacions dels criteris i paràmetres que s'ha decidit aplicar per cada un dels mètodes.

Com s'ha mencionat anteriorment, s'utilitzaran les dades de gener del 2000 fins desembre de 2019 per realitzar la comparació entre els mètodes. Dins d'aquest interval, s'utilitzarà un 95% (2000-2018) de les dades per entrenar els models i un 5% (2019) per posar a prova aquest model.

3.1 Anàlisi Exploratori de les dades

Figura 7 - Turistes a Espanya (2000-2018)



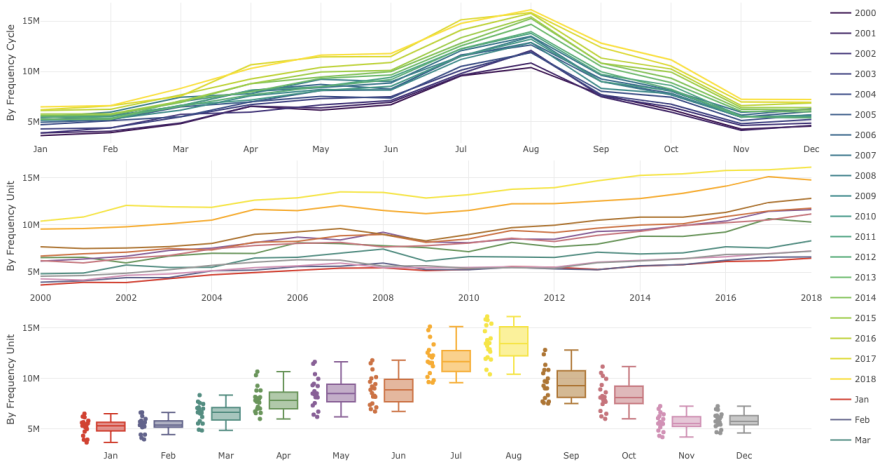
Font: Elaboració Pròpia

S'observa que el comportament de la sèrie és molt similar durant tot el període, és a dir, es registren pujades i baixades similars al llarg dels anys en les diferents estacions de l'any (estacionalitat). Aquest fet té sentit perquè es coneix que hi ha certs mesos de l'any en que el nombre de turistes serà o no major.

A més a més, s'observa que hi ha hagut un creixement del nombre de turistes amb el pas dels anys (tendència). Aquest augment podria ser a causa de les millores en les infraestructures de comunicació i el desenvolupament de les noves tecnologies. Tot i així, dins d'aquest creixement també existeix una certa variació, per exemple, es pot veure que hi ha una baixada del nombre de turistes durant els anys 2008-2012. Si es posa en context es veu que durant

aquells anys es va produir una crisi financera global que va afectar a nivell mundial i per tant, el nombre de turistes es va reduir.

Figura 8 - Estudi del comportament de les dades

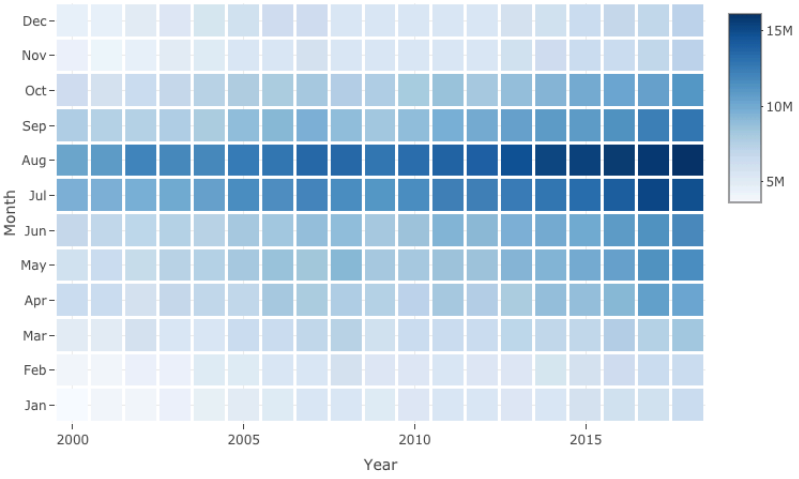


Font: Elaboració Pròpia

D'aquest gràfic es pot destacar que el turisme a Espanya sobretot augmenta a l'estiu, ja que és quan la majoria de treballadors tenen vacances. A més a més, s'observa que a mesura que el bon temps augmenta, el nombre de turistes també augmenten.

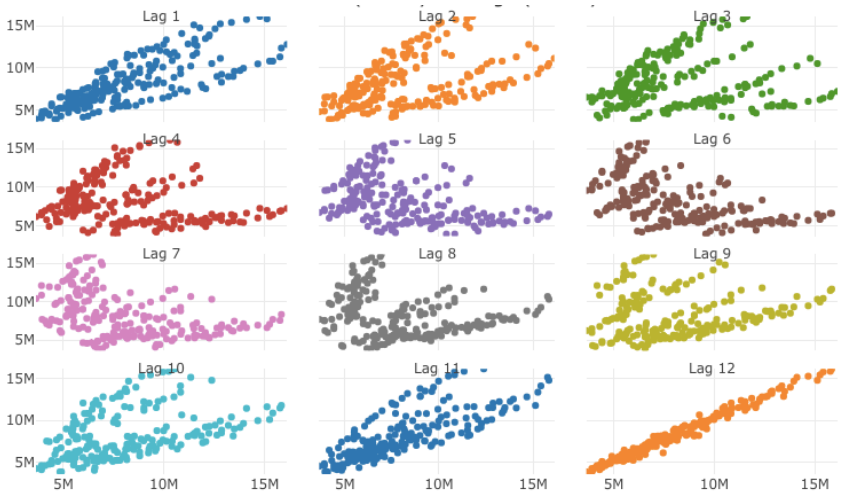
Figura 9 - Mapa de calor

Es reitera el que s'ha observat anteriorment mitjançant aquest *heatmap* (mapa de calor) que indica que als mesos d'estiu és quan el nombre de turistes augmenta. El registre més alt es troba al mes d'agost de 2018 amb 16.121.776 turistes.



Font: Elaboració Pròpia

Figura 10 – Sèrie (Y) versus Retards (X)



Font: Elaboració Pròpia

Aquest gràfic mostra la visualització de la sèrie amb el seus retards i serveix per identificar la correlació entre la sèrie i els retards. S'observa que del primer retard al setè, la relació entre la sèrie i el retard va perdent la seva forma lineal, a partir d'aquest fins l'últim la forma lineal va tornant fins que a l'últim s'hi pot veure dibuixada una línia quasi perfecta.
Amb això el que es vol dir és que s'observa una forta correlació entre el mateix mes al llarg dels anys i també entre els mesos propers.

3.2 ARIMA

Durant l'explicació teòrica d'aquesta metodologia s'ha mencionat que s'utilitzaran dos models diferents per la realització de prediccions:

- Modelització manual:

En aquest cas, com bé diu el nom, s'estudiarà de manera manual la sèrie en qüestió, és a dir, caldrà estudiar la sèrie, mirar si necessita transformacions i decidir els paràmetres que millor s'adapten a les dades per tal de poder fer prediccions.

- Modelització automàtica:

En aquest segon cas, es farà ús d'una funció de l'R que s'encarregarà de fer el mateix procés manual però de manera automàtica. Aquesta funció que prové del paquet *forecast()*[13] de l'R i s'anomena *auto.arima()* [14] és una opció ràpida i en moltes ocasions eficaç per la predicció de sèries temporals.

[13] Hyndman, R. (s.d). *forecast*. Recollit de RDocumentation: https://www.rdocumentation.org/packages/forecast/versions/8.12

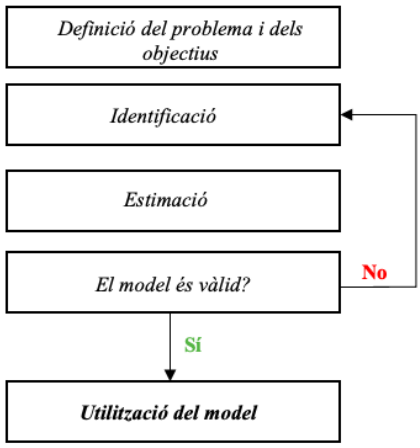
[14] Hyndman, R. (s.d). *auto.arima*. Recollit de RDocumentation: https://www.rdocumentation.org/packages/forecast/versions/8.12/topics/auto.arima

La funció *auto.arima()* s'utilitza de manera que l'usuari únicament s'ha d'encarregar d'aplicar la funció a la sèrie que vol estudiar i aquesta treu una sortida en la que apareix quins són els paràmetres del model que encaixa millor en les dades. Per tal d'escollir el model, la funció escull el model que minimitza l'AIC.

Tot i que aquesta funció sembla que és solució fàcil i ràpida per la predicció de sèries temporals i per tant no seria necessària la modelització manual s'ha de tenir en compte que tal i com diu el nom és automàtica, amb això el que es vol dir és que és una funció que té uns paràmetres establerts i hi ha vegades que pot ser que les dades no expressin clarament el seu comportament i per tant la funció no sigui capaç de triar els paràmetres correctament.

Box i Jenkins proposa una sèrie de passos clau que s'han de seguir per tal de modelar correctament les dades i poder fer una predicció.

Figura 11 - Estructura Box i Jenkins



Font: Elaboració Pròpia

1. Identificació: L'objectiu és identificar un model que de manera versemblant hagi pogut generar els valors observats:

El primer pas per la identificació és crear un gràfic que mostri el comportament de la sèrie, aquest gràfic ja s'ha mostrat anteriorment durant l'anàlisi exploratori de les dades ***(Figura 7)*** i s'ha conclòs que a simple vista s'observa tant una component regular com estacional. Primer de tot, caldrà mirar si hi ha estacionarietat en les dos components, en el cas de que n'hi hagi, caldrà determinar si són necessàries o no transformacions. Per verificar aquesta informació es realitzen els gràfics de la FAS (Funció d'Autocorrelació Simple) i la FAP (Funció d'Autocorrelació Parcial).

Figura 12 - Estudi de la FAS i la FAP

Font: Elaboració Pròpia

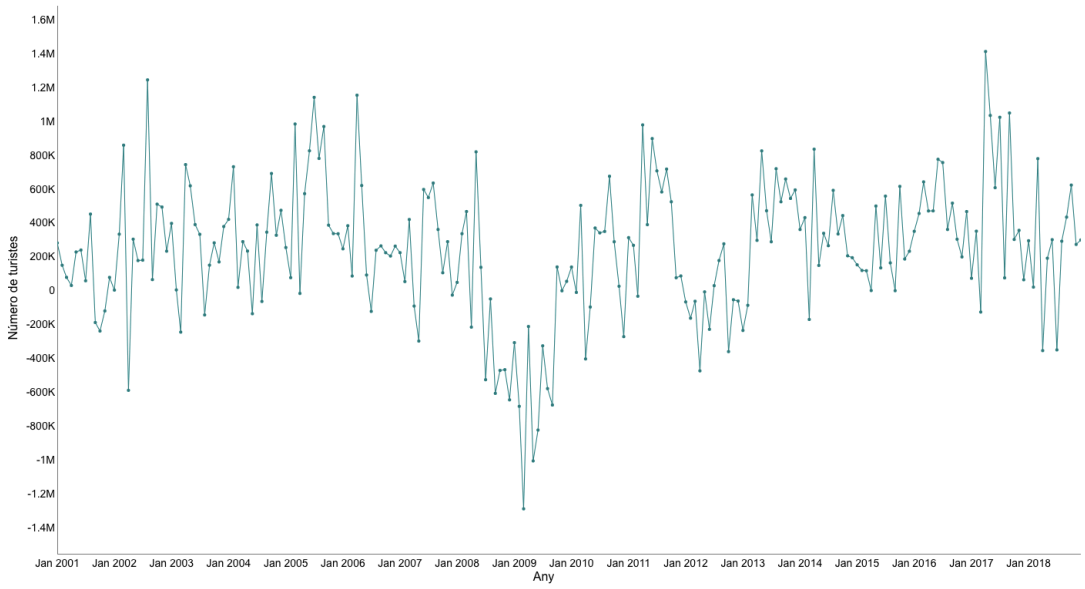
Un procés no estacionari es caracteritzarà per:

- FAS amb coeficients propers a la unitat que decreixen molt lentament.
- FAP amb un coeficient significatiu i proper a la unitat.

Si s'estudia primer la component regular, s'observa que aquestes característiques no es compleixen ja que les correlacions disminueixen força ràpid. En canvi per la component estacional, caldrà mirar les mateixes característiques però observant els coeficients cada 12 mesos, en aquest cas, les correlacions decreixen molt lentament i per tant, s'ha decidit realitzar una diferència estacional.

Es procedeix a aplicar una diferenciació estacional, si una vegada realitzada no s'assegura l'estacionarietat caldrà realitzar una o més d'una diferenciacions:

Figura 13 - Aplicació d'una diferència estacional sobre la sèrie



Font: Elaboració Pròpia

Figura 14 - Estudi de la FAS i la FAP una vegada aplicada la diferència estacional

Font: Elaboració Pròpia

Una vegada aplicada aquesta diferència s'observa en el gràfic de les correlacions que s'ha corregit el problema i que les dades es mantenen mes o menys estables amb una mitjana de zero. Cal destacar que hi ha una forta baixada entre 2008-2010 que com s'ha mencionat anteriorment aquesta baixada és deguda a la crisi global que hi va haver entre els anys 2008-2012.

.2. Estimació:

Es començarà amb el mètode manual per tal d'evitar influenciar-se pel model automàtic, ja que com s'ha mencionat, aquest model automàtic treu un possible model per predir.

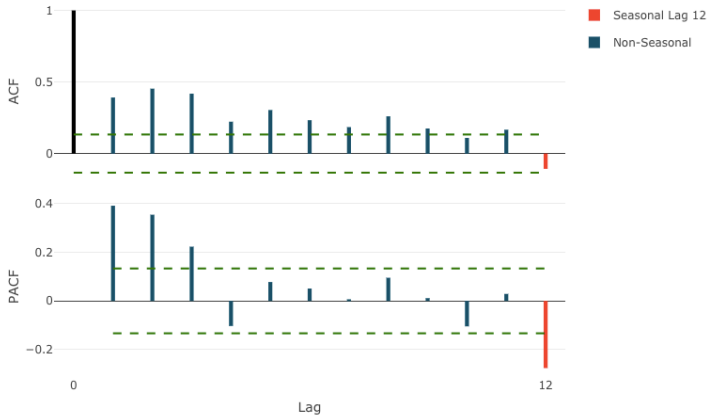
- ***Mètode manual:***

En aquest cas es tractarà amb un model *SARIMA (p,d,q)(P,D,Q)* i per tant s'han de decidir els seus paràmetres, coneixent ja que *d=0 i D=1*.

Com ja s'ha mencionat, els models SARIMA es separen en la part regular i la part estacional.

Per la part estacional es miraran els últims gràfics de la FAS i la FAP ***(Figura 14)*** i per poder determinar el paràmetre de la part regular, el que es pot fer és reduir el número de retards en el gràfic de les autocorrelacions per tal "d'apropar-se" i mirar la dependència respecte als valors adjacents. S'ha utilitzat un retard de 12:

Figura 15 - Estudi de la FAS i la FAP ampliada una vegada aplicada la diferència estacional



Font: Elaboració Pròpia

Una vegada analitzats aquests gràfics es decideix escollir un model ***SARIMA(2,0,2)(1,1,1) [12]*** i s'obtenen els següents resultats:

Taula 1 - Sortida model manual SARIMA (2,0,2)(1,1,1)[12]

```
arima(x = train, order = c(2, 0, 2), seasonal = list(order = c(1, 1, 1)))

Coefficients:
         ar1     ar2      ma1     ma2     sar1     sma1
      0.7418  0.2037  -0.4870  0.0096  -0.0687  -0.2886
s.e.  0.2303  0.2252   0.2266  0.1797   0.1926   0.1912

sigma^2 estimated as 107079375236:  log likelihood = -3050.59,  aic = 6115.18
```

Font: Elaboració Pròpia

- ***Mètode automàtic:***

Per aquest mètode cal utilitzar la funció *auto.arima()*.
S'obté que el millor model és un ***SARIMA(3,0,3)(2,1,0) [12]*** i la següent informació:

Taula 2 - Sortida model automàtic SARIMA (3,0,3)(2,1,0)[12]

```
ARIMA(3,0,3)(2,1,0)[12] with drift

Coefficients:
          ar1     ar2     ar3     ma1     ma2      ma3     sar1     sar2      drift
      -0.2135  0.0528  0.8567  0.5333  0.5206  -0.4386  -0.4643  -0.1582  16962.009
s.e.   0.0672  0.0555  0.0489  0.1117  0.1053   0.1134   0.0825   0.0827   5747.612

sigma^2 estimated as 86160075313:  log likelihood=-2520.66
AIC=5061.32   AICc=5062.62   BIC=5093.25
```

Font: Elaboració Pròpia

En aquest apartat, de moment es mirarà únicament el *Akaike* dels dos models. En aquest cas, sembla que el millor models és el model automàtic ja que es el que té un *Akaike* més petit. Tot i això, encara no es pot afirmar que el model automàtic sigui millor que el model manual, primer de tot cal mirar si aquest és vàlid i també s'utilitzaran altres mesures per la comparació.

3. El model és vàlid?

Una vegada s'han escollit els models, abans de realitzar prediccions, cal mirar si aquests models són vàlids o no.

Primer de tot, es mirarà la significació dels coeficients. En aquest cas, s'ha mirat que el p-valor sigui menor al nivell de significació del 0.05.

Taula 3 - Significació dels coeficients del model manual (esquerra) i model automàtic (dreta)

```
  ar1   ar2   ma1   ma2  sar1  sma1
 TRUE FALSE  TRUE FALSE FALSE FALSE
```

```
  ar1   ar2   ar3   ma1   ma2   ma3  sar1  sar2 drift
 TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

Font: Elaboració Pròpia

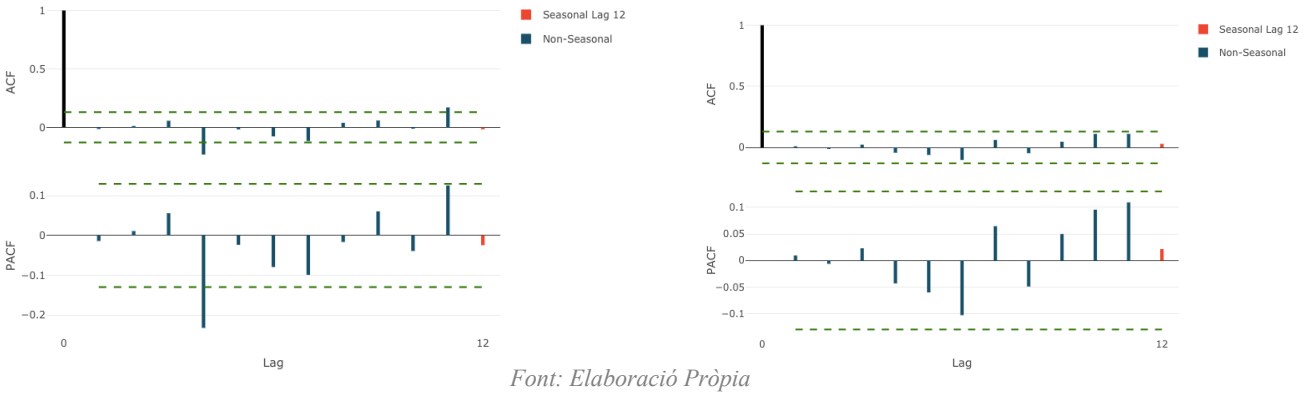
Per que sigui més visual s'han mostrat les dades de tipus booleanes amb vertader (*TRUE*) o fals (*FALSE*). En aquest cas, quan els p-valors de cada un dels coeficients és menor al nivell de significació, és a dir, que els coeficients són significatius, sortirà el valor *TRUE*.

S'observa que no tots els coeficients són significatius. La teoria diu que per que el model sigui vàlid ho haurien de ser, però a la pràctica i a la vida real, que tots els coeficients d'un model

siguin significatius és molt difícil. Tot i que no es compleixi amb aquesta condició, de moment el model es donarà per vàlid ja que en principi no hauria de ser un problema.

A continuació, es realitzaran els gràfics de las FAS i de la FAP pels residus del model per tal de validar-los.

Figura 16 - Estudi de la FAS i la FAP dels residus del model manual (esquerra) i automàtic (dreta)

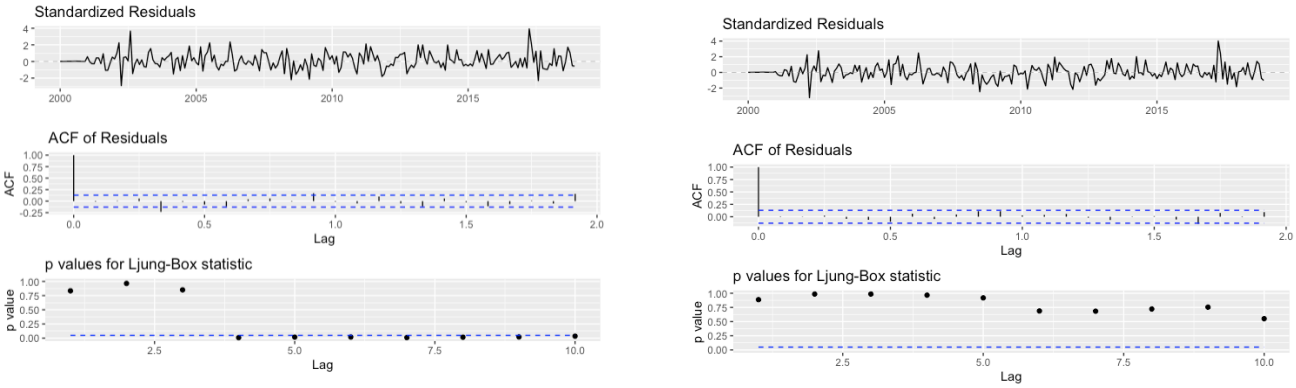


Font: Elaboració Pròpia

S'observa que a partir del quart retard del model manual encara hi queda alguna correlació pendent d'explicar. Una vegada més, cal mencionar que la teoria diu que els residus han d'estar en soroll blanc però a la vida real això es molt difícil, a més la majoria dels residus es troben dins de l'interval i per tant no és un problema greu.

Per mirar si les dades es distribueixen de manera independent, s'utilitzarà l'estadístic *Ljung-Box:*

Figura 17 - Estudi dels residus del model manual (esquerra) i automàtic (dreta)



Font: Elaboració Pròpia

S'observa que sobretot pels residus del model *automàtic*, alguns dels p-valors que corresponen l'estadístic de *Ljnug - Box* són valors que es troben per sobre del 0,05 i això indicaria que no son significatius, per comprovar-ho es realitza un contrast d'hipòtesis amb el test *Ljung-Box*.

$$H_o: Les\ dades\ es\ distribueixen\ de\ manera\ independent$$
$$H_1: Les\ dades\ no\ es\ distribueixen\ de\ manera\ independent$$

Taula 4 – Test de Ljung-Box pel model manual (esquerra) i automàtic (dreta)

```
        Box-Ljung test

data:  model1$residuals
X-squared = 0.044023, df = 1, p-value = 0.8338
```

```
        Box-Ljung test

data:  model2$residuals
X-squared = 0.020187, df = 1, p-value = 0.887
```

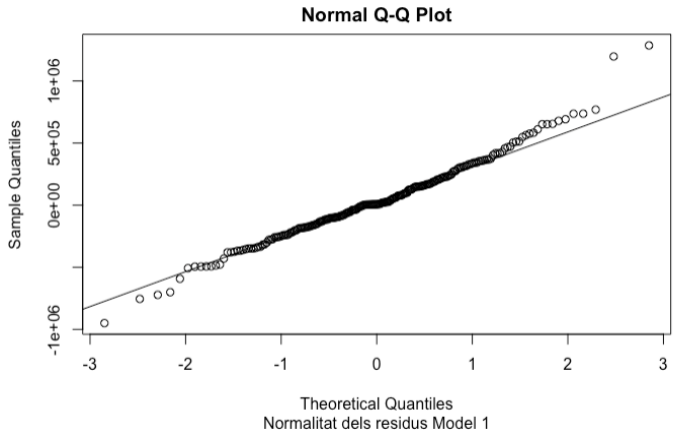
Font: Elaboració Pròpia

En ambdós casos el p-valor és superior al nivell de significació de 0,05 i per tant, no existeixen indicis per rebutjar la hipòtesi nul·la, i per tant, les dades es distribueixen de manera independent.

Per últim, s'estudiarà la normalitat dels residus mitjançant un Q-Q Plot i el test de *Shapiro-Wilk:*

$$H_0: X \sim N\left(\mu, \sigma^2\right) - La\ distribució\ és\ normal$$
$$H_1: X \nsim N\left(\mu, \sigma^2\right) - La\ distribució\ no\ és\ normal$$

Figura 18 - Q-Q Plot model manual



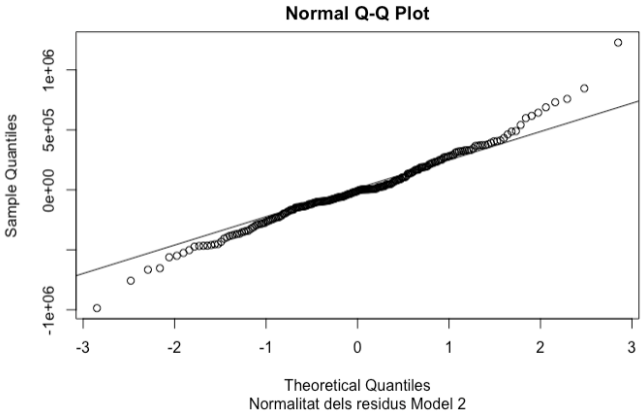
Font: Elaboració Pròpia

Taula 5 - Shapiro-Wilk Test model manual

```
        Shapiro-Wilk normality test

data:  model1$residuals
W = 0.98153, p-value = 0.004486
```

Font: Elaboració Pròpia

Figura 19 - Q-Q Plot model automàtic



Font: Elaboració Pròpia

Taula 6 - Shapiro-Wilk Test model automàtic

```
        Shapiro-Wilk normality test

data:  model2$residuals
W = 0.98099, p-value = 0.003688
```

Font: Elaboració Pròpia

Observant els gràfics *Q-Q Plot* ***(Figura 18 i 19)*** d'ambdós models sembla que els residus estan distribuïts normalment però si que és veritat que s'observa que la distribució es simètrica però més apuntada del que hauria de ser, és a dir les cues no segueixen del tot la línia recta que haurien de seguir. Segurament, és per això que no s'accepta la hipòtesi de normalitat en el test de *Shapiro* en cap dels dos models, ja que el p-valor de l'estadístic es inferior al nivell de significació.

Com sembla sobretot un problema de les cues, i segons la teoria és la situació que menys preocupa per la no normalitat i afecta relativament poc a les distribucions de l'estadístic de contrast, i per tant, no és un problema greu, es decideix que el model continuarà sent vàlid.

Finalment, es decideix donar ambdós models com a vàlids ja que no hi ha cap indici excessivament rellevant que ens faci rebutjar la validació del model. Si que és veritat que hi ha algunes condicions que no es compleixen segons la teoria i després si realment són importants es veurà reflexat a les prediccions, però com ja s'ha mencionat, són problemes molt típics a la vida real i hi ha condicions que són molt difícils de complir.

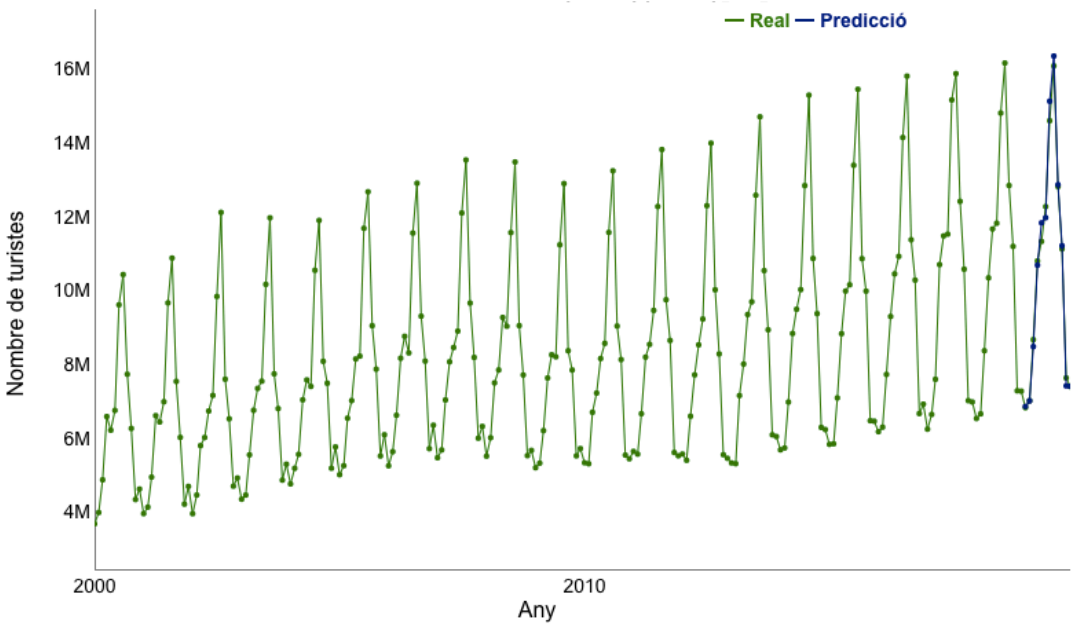
A més a més, el fet de que aquestes suposicions no es compleixin del tot també són un bon incentiu per provar altres mètodes que no necessiten cap tipus de suposició, com bé són les Xarxes Neuronals.

4. Prediccions i errors

Una vegada s'ha validat el model, és l'hora de posar-lo a prova i estudiar la seva capacitat predictiva. Com s'ha mencionat anteriorment, es realitzarà les prediccions de tot l'any 2019.

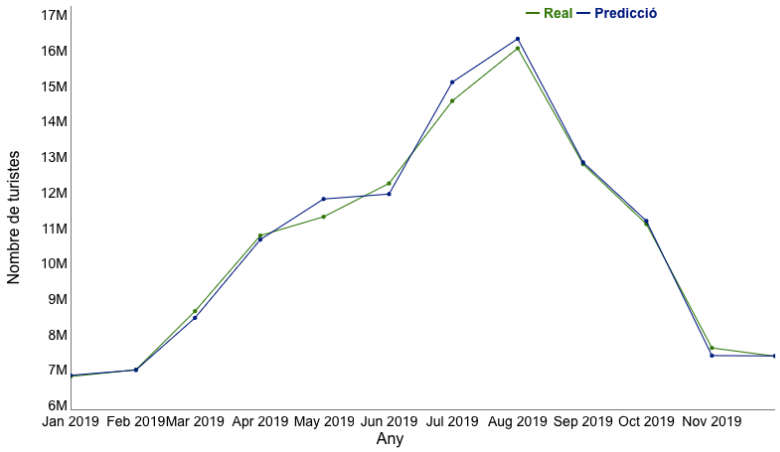
- Mètode manual:

Figura 20 – Prediccions model manual SARIMA(2,0,2)(1,1,1)[12]



Font: Elaboració Pròpia

Figura 21 - Prediccions any 2019 model manual SARIMA(2,0,2)(1,1,1)[12]

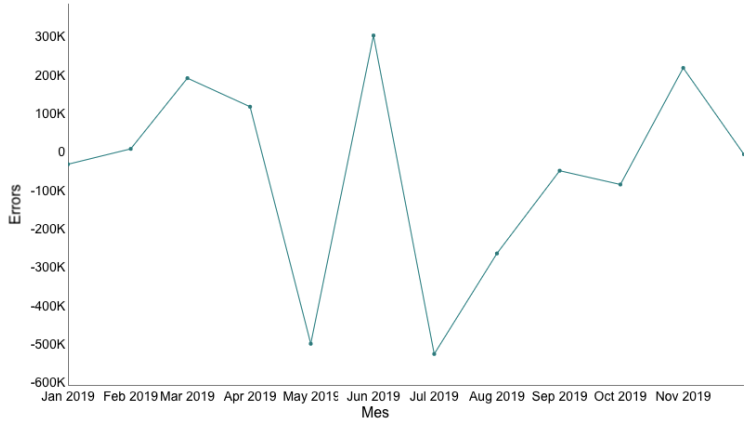


Font: Elaboració Pròpia

Observant el gràfic amb totes les dades ***(Figura 20),*** sembla que les prediccions segueixen aquesta tendència creixent amb el pas dels anys que es mencionava a l'anàlisi exploratori, a més a més, també s'observa que les prediccions segueixen una estacionalitat similar a la resta de les dades, és a dir, depenent del mes les prediccions son majors o menors. Si s'observa en concret l'any 2019 ***(Figura 21),*** sembla que les prediccions s'ajusten força bé als valors reals. De totes maneres, és important no fixar-se únicament en el que es veu a simple vista i analitzar les prediccions de manera numèrica, amb els errors i les mesures de validació.

A continuació, s'observaran els errors d'aquestes prediccions de manera gràfica. Cal recordar que els errors mostren la diferència entre els valors reals i els valors que s'han predit amb el model seleccionat:

Figura 22 - Errors de predicció any 2019 model manual SARIMA(2,0,2)(1,1,1)[12]



Font: Elaboració Pròpia

Per aquest model s'observa que els errors estan en un rang d'entre 300.000 i -600.000 turistes. Aquest rang no es pot quantificar com a bo o dolent sense fixar un problema concret.

El mètode gràfic no es suficient per decidir si el model té un bon ajust o no, caldrà analitzar els següents criteris de validació:

Taula 7 - Criteris de validació model manual SARIMA(2,0,2)(1,1,1)[12]

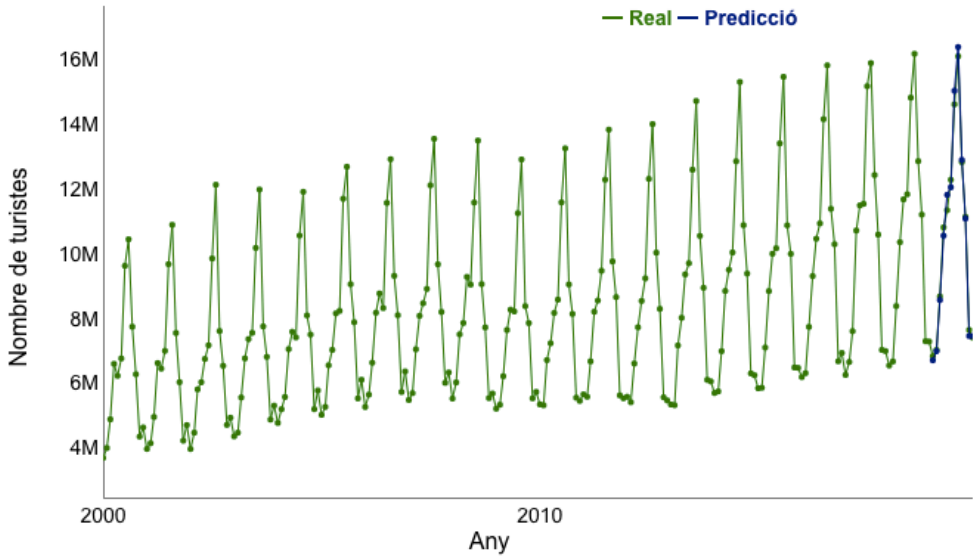
	AIC	***EQM***	***EAM***	***EPAM***	***RMSE***
ARIMA MANUAL	6.115,18	6663223827	192097,5	1,68%	258.132,2

Font: Elaboració Pròpia

Per estudiar individualment el model, s'estudiarà la capacitat predictiva mitjançant el EPAM o MAPE ja que és l'únic estimador adimensional i per tant es pot comparar de manera individual. Com s'ha mencionat a la teoria quan es parlava dels criteris de validació, quan 1% < EPAM < 3% com passa en aquest cas amb un EPAM del 1,68%, es pot considerar que el model té una bona capacitat predictiva.

- ***Mètode automàtic***

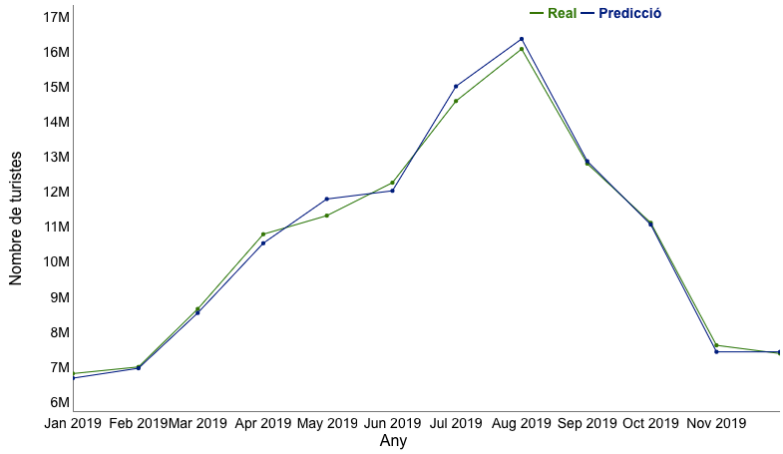
Figura 23 - Prediccions model automàtic SARIMA(3,0,3)(2,1,0)[12]



Font: Elaboració Pròpia

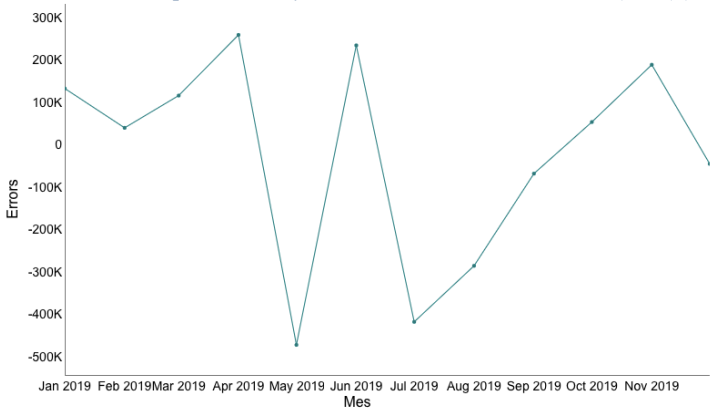
Figura 24 - Prediccions any 2019 model automàtic SARIMA(3,0,3)(2,1,0)[12]

A simple vista observant tant el gràfic amb totes les dades com el gràfic que es centra en l'any 2019 les prediccions semblem força similars al que s'obté amb el mètode manual i per tant l'anàlisi és molt similar. Sembla que les prediccions s'ajusten prou bé als valors reals ja que a primera vista no hi ha cap valor que s'escapi de manera estrepitosa del nombre de turistes reals.



Font: Elaboració Pròpia

Figura 25 - Errors de predicció any 2019 model automàtic SARIMA(3,0,3)(2,1,0)[12]



Font: Elaboració Pròpia

Per aquest model s'observa que els errors estan en un rang d'entre uns 300.000 i -500.000 turistes.

Per aquest model, obtenim els següents criteris de validació:

Taula 8 - Criteris de validació model automàtic SARIMA(3,0,3)(2,1,0)[12]

	AIC	***EQM***	***EAM***	***EPAM***	***RMSE***
AUTO ARIMA	6.090,14	5675696927	192526,2	1.76%	238237,2

Font: Elaboració Pròpia

En aquest cas s'obté un EPAM amb un valor de 1.76% i ja que es troba dins de l'interval entre l'1% i el 3%, es pot dir que el model té una bona capacitat predictiva.

Durant la validació d'aquests models s'ha vist que hi havia condicions que no es complien de manera teòrica però s'ha decidit donar per vàlid el model de totes maneres ja que normalment a la pràctica és molt difícil que es compleixin totes les condicions. Una vegada realitzades les prediccions s'observa que es tracta d'un bon ajust dels dos models, amb això el que es vol remarcar es que s'ha vist que tot i que no es compleixin totes les condicions de manera estricta es pot obtenir un bon ajust del model.

3.3 Xarxes Neuronals Autoregressives

Aquest mètode s'implementa a l'R mitjançant la funció *nnetar()*[15] que ve donada pel paquet *forecast()*.

Aquesta funció retorna un model del tipus $NNAR(p,P,k)_m$ on p indica els retards que s'utilitzen per la predicció, P indica el component estacional, k indica els nodes ocults i m indica la periodicitat (mensual, trimestral..).

Un model $NNAR(p,P,k)_m$ té $(y_{t-1}, y_{t-2}, \dots, y_{t-p}, y_{t-m}, y_{t-2m}, y_{t-Pm})$ entrades i k neurones a la capa oculta.

La funció *nnetar()* si no se li especifiquen els paràmetres s'encarrega automàticament. En aquest estudi, s'utilitzarà la funció *nnetar()* de manera automàtica i es deixarà que l'R s'encarregui de fixar els paràmetres.
A més a més aquesta funció té un paràmetre *lambda* que s'encarrega de la transformació de Box-Cox. Aquesta transformació s'encarrega de transformar variables dependents no normals en una forma normal.

[15] Hyndman, R. (s.d). *nnetar*. Recollit de RDocumentation: https://www.rdocumentation.org/packages/forecast/versions/8.12/topics/nnetar

Cox va desenvolupar un procediment per identificar un exponent adequat (lambda) per transformar les dades per millorar la seva normalitat. El valor lambda indica la potència a la qual s'han de recollir totes les dades. Per això, la transformació de potència Box-Cox busca diversos valors de lambda mitjançant diversos mètodes fins que es trobi el millor valor.
En aquest estudi s'utilitzarà paràmetre com lambda="auto". Això voldrà dir que es selecciona el millor paràmetre de lambda mitjançant una funció de l'R anomenada *BoxCox.lambda()*[16] que la seva funció principal és la selecció automàtica del paràmetre de transformació de Box Cox. Aquesta última funció té per defecte un mètode que s'anomena Guerrero (1993)[17] en el que lambda minimitza el coeficient de variació per a subsèries de dades.

Una vegada s'implementa el model, obtenim la següent sortida:

Taula 9 - Sortida model NNAR(5,1,4)[12]

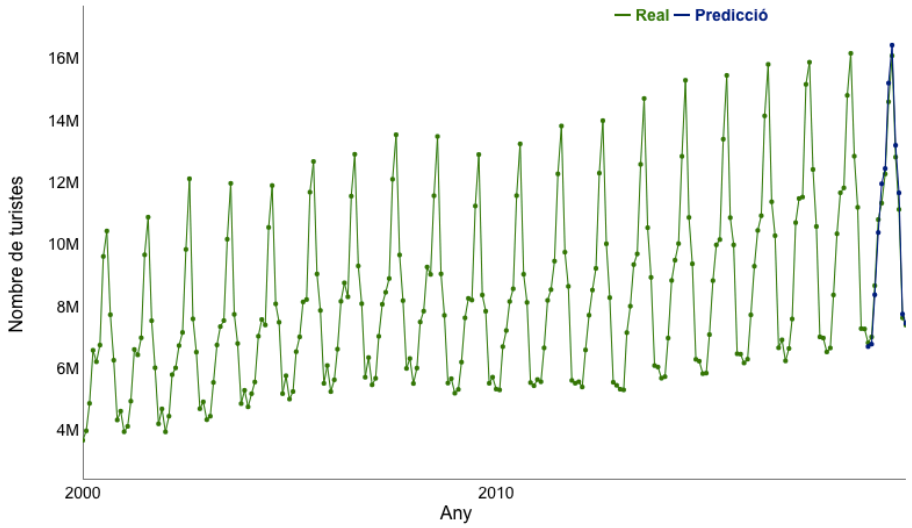
```
Model:  NNAR(5,1,4)[12]
Call:   nnetar(y = train, lambda = "auto")

Average of 20 networks, each of which is
a 6-4-1 network with 33 weights
options were - linear output units

sigma^2 estimated as 0.09261
```

Font: Elaboració Pròpia

S'obté un model amb 5 retards, component estacional i 4 nodes ocults. El valor de la transformació de Box-Cox (lambda) per aquestes dades és de 0.12458.

Si s'ajusta el model mitjançant la funció *forecast()* s'obtenen els següents resultats:

Figura 26 – Prediccions model NNAR(5,1,4)[12]



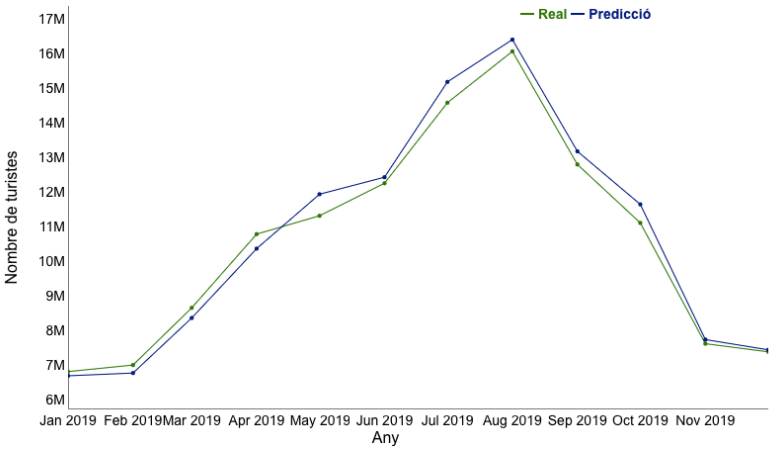
Font: Elaboració Pròpia

[16] Hyndman, R. (s.d). *BoxCox.lambda*. Recollit de RDocumentation: https://www.rdocumentation.org/packages/forecast/versions/8.12/topics/BoxCox.lambda

[17] Hyndman, R. (s.d). *Guerrero's method for Box Cox lambda selection*. Recollit de feasts: https://feasts.tidyverts.org/reference/guerrero.html

Figura 27– Prediccions any 2019 model NNAR(5,1,4)[12]

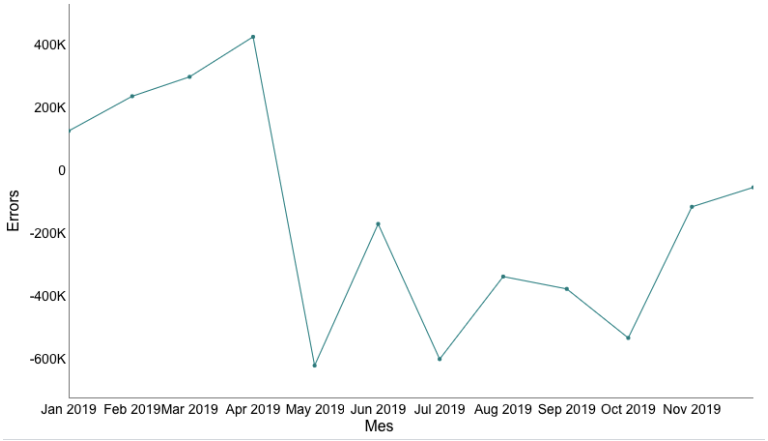
A simple vista, sembla que el model NNAR ajusta de manera correcta les dades ja que no s'hi veu cap diferència extremadament significativa entre els valors reals i els valors predits. Tot i així, si s'observa el gràfic únicament de l'any 2019 sembla que prediccions sobreestimen durant quasi tot el període ja que van lleugerament per sobre dels valors reals.



Font: Elaboració Pròpia

Figura 28 - Errors de predicció any 2019 model NNAR(5,1,4)[12]

Com s'ha vist, s'observa que les prediccions són generalment majors al valor real de turistes que venen a Espanya durant tot l'any, excepte el mes d'abril que resulta que el nombre de turistes és major a l'esperat. En aquest cas els errors de predicció es mouen entre els 400.000 i -600.000 turistes.



Font: Elaboració Pròpia

Per aquest model, obtenim els següents criteris de validació:

Taula 10 - Criteris de validació model NNAR(5,1,4)[12]

	AIC	***EQM***	***EAM***	***EPAM***	***RMSE***
NNAR		139884081245	325025.1	2,98%	374.010.8

Font: Elaboració Pròpia

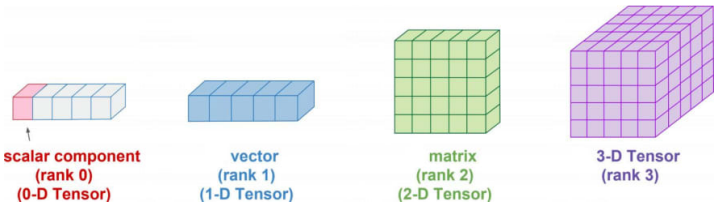
Per aquest model el valor de l'EPAM és d'un 2,98% per tant el model té una bona capacitat predictiva tot i que s'apropa molt al 3%.

3.4 Long-Short Term Memory

Aquest mètode, com s'ha mencionat a l'apartat de recursos informàtics, s'ha implementat amb el llenguatge de programació *Python* i amb les llibreries *Keras* i *TensorFlow*.

En *Machine Learning,* tots els mètodes utilitzen **tensors** com la seva estructura bàsica. Un **tensor**[18] és un objecte matemàtic que guarda valors numèrics de diferents dimensions:

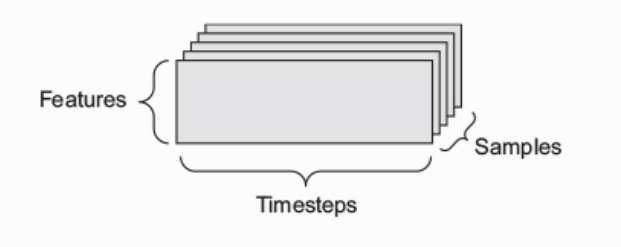
Figura 29 - Dimensions d'un tensor



Font: Juan Carlos, 2019 - https://dev.to/juancarlospaco/tensors-for-busy-people-

A la figura s'observa que els tensors principals són els objectes matemàtics que es coneixen (escalar, vector, matriu, cub) i que es poden veure amb l'ull humà. Depenent de les dades amb les que es vulgui treballar, s'utilitzaran un tipus de tensors o uns altres.

Figura 30 - Tensor 3 Dimensions (3D)



Font: François Chollet i Joseph J. Allaire – Deep Learning with R

Les sèries temporals univariants utilitzen tensors de 3 dimensions (*3D tensors*) perquè una de les dimensions (*Timesteps*) correspondrà al les propietats de cada pas del temps i després un tensor de 2 dimensions que el formen les mostres (*Samples*) i les característiques (*Features*)

Per tant, les sèries temporals tenen un tensor de 3 dimensions formats per:

- *Samples*: correspon a les observacions (files) de les dades.

[18] Kan, C. N. (11 / Des / 2018). *Quick ML Concepts: Tensors*. Recollit de Medium, Towards Data Science: https://towardsdatascience.com/quick-ml-concepts-tensors-eb1330d7760f

- *Timesteps*: són les observacions passades que s'utilitzen per predir valors futurs, són equivalents a les variables de retard.
- *Features*: Les columnes de les dades.

$$Imput\_shape = [samples, timesteps, features]$$

El primer que s'ha de fer és preparar les dades. Per això es començarà escalant la sèrie, per fer-ho s'utilitzarà la funció *MinMaxScaler*[19] del paquet *scikt-learn*. Aquesta funció realitza el següent càlcul per cada una de les observacions de la sèrie:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Les LSTM són força sensibles a l'escala de les dades i per això es molt recomanable reescalar les dades en un rang de 0 a 1. Una vegada s'han introduït les dades al model es realitza la inversa d'aquesta escala per tal de tornar a tenir l'escala original de la sèrie.

1. Definir el model

El següent pas a realitzar, serà definir el tensor. Com s'ha mencionat, per dades de sèries temporals s'utilitzarà un tensor de tres dimensions i en aquest cas s'utilitzarà un tensor [204, 12 , 1] on el primer número són el nombre d'observacions de la mostra amb la que s'entrenarà el model, s'utilitzaran 12 observacions de retard (*timesteps*) i una única *feature* (ja que tenim una sèrie temporal univariant).

$$Imput\_shape = [204, 12, 1]$$

En aquest mètode, a diferència dels altres utilitzats, s'ha decidit utilitzar un 90% de les dades per entrenament, això deixa 24 observacions per la prova (test), s'ha decidit fer així perquè després de diferents proves s'ha vist que el nombre d'observacions de retard que funcionava millor pel model era 12. Al tenir aquest nombre de *timesteps*, el model entén que s'han utilitzar dotze observacions anteriors per tal de predir un valor futur, i per obtenir un any de prediccions caldrà utilitzar un 90% d'entrenament. Per tant, el nombre d'observacions d'entrenament inicial serà de 216 i el de prova de 24, però una vegada aplicat el *timesteps*=12 el nombre d'observacions serà 204 per l'entrenament i una sortida de 12 prediccions.

Les Xarxes neuronals es defineixen a *Keras* com una seqüència de capes, és a dir, es van definint capes una rere l'altra. Per ajuntar aquestes capes, primerament s'afegeix una funció anomenada *Sequential*() i a partir d'aquí es poden anar afegint capes al model.

[19] learn, s. (s.d). *MinMaxScaler*. Recollit de scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

Per la sèrie temporal de l'estudi s'ha decidit utilitzar dos capes, la primera una capa recurrent anomenada LSTM a la que s'han afegit el nombre de neurones a la capa oculta, juntament amb les mides del tensor i després una capa de sortida amb una neurona anomenada *Dense*().

Taula 11 - Sortida model LSTM

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 10)                480
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 11
=================================================================
Total params: 491
Trainable params: 491
Non-trainable params: 0
_________________________________________________________________
```

Font: Elaboració Pròpia

2. Compilar el model

Una vegada definida la Xarxa Neuronal, caldrà compilar-la. Al compilar el model el que s'està fent es transformar aquestes capes simples que s'han afegit al model en transformacions de matrius molt més avançades. Per la compilació s'han de definir principalment dos paràmetres[20]:

- ***L'algorisme d'optimització*** que s'ha explicat a la teoria: els més coneguts són *Adam, Stochastic Gradient Descent i RMSprop*. En aquest treball s'ha utilitzat *Adam* [21].
- ***La funció de pèrdua*** que en aquest cas s'utilitzarà l'error absolut mitjà (*mae*).

3. Ajustar el model

Per últim, caldrà ajustar el model, això voldrà dir adaptar els pesos a les dades d'entrenament. La xarxa s'entrena mitjançant l'algorisme de *Backpropagation* i s'optimitza mitjançant l'algorisme d'optimització i la funció de pèrdua.

L'algorisme de *Backpropagation* requereix que la xarxa estigui entrenada durant un nombre específic d'èpoques o cicles o més conegudes com ***epochs***. Els *epochs* són el nombre de vegades que el conjunt sencer de dades s'exposa a la xarxa. Una xarxa LSTM pot estar entrenada per desenes, centerars o inclus milers d'*epochs*. Cada *epoch* es pot dividir en grups de lots o més coneguts com *batch*. Un *batch* és el nombre de dades que té cada iteració d'un cicle o *epoch*.

[20] Calvo, D. (10 / Des / 2018). *Función de coste – Redes neuronales*. Recollit de Diego Calvo: https://www.diegocalvo.es/funcion-de-coste-redes-neuronales/

[21] Brownlee, J. (3 / Jul / 2017). *Gentle Introduction to the Adam Optimization Algorithm for Deep Learning*. Recollit de Machine Learning Mistery: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

Un conjunt de dades d'entrenament es pot dividir en un o més *batch*:

Batch size = 1: Els pesos s'actualitzen després de cada mostra i el procediment s'anomena **Stochastic Gradient Descent**.

Batch size = 32: Els pesos s'actualitzen després d'un nombre especificat de mostres i el procediment s'anomena descens de **Mini-Batch Gradient Descent**. Els valors comuns són 32, 64 i 128, adaptats a l'eficiència i a la velocitat actualitzades del model.

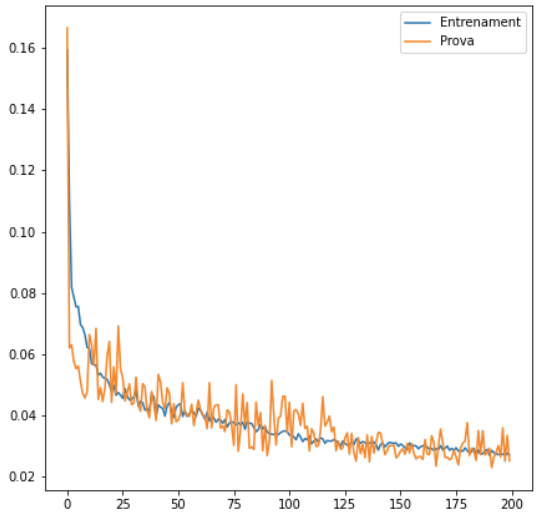
Batch size = n: on n és el nombre de mostres del conjunt de dades d'entrenament. Els pesos s'actualitzen al final de cada època i el procediment s'anomena **Batch Gradient Descent**.

Per aquest treball s'utilitzilitzarà el **Stochastic Gradient Descent,** és a dir, un *batch Size* =1 i un número d'*epochs* = 200.

L'elecció òptima d'aquests paràmetres és molt difícil i no existeix una manera d'escollir-los, en aquest cas s'han decidit escollir aquests paràmetres després de realitzar proves amb diferents valors i veient com es comportava la sèrie depenent dels valors dels paràmetres.

4. Evaluació i predicció del model

Figura 31 - Funcions de pèrdua

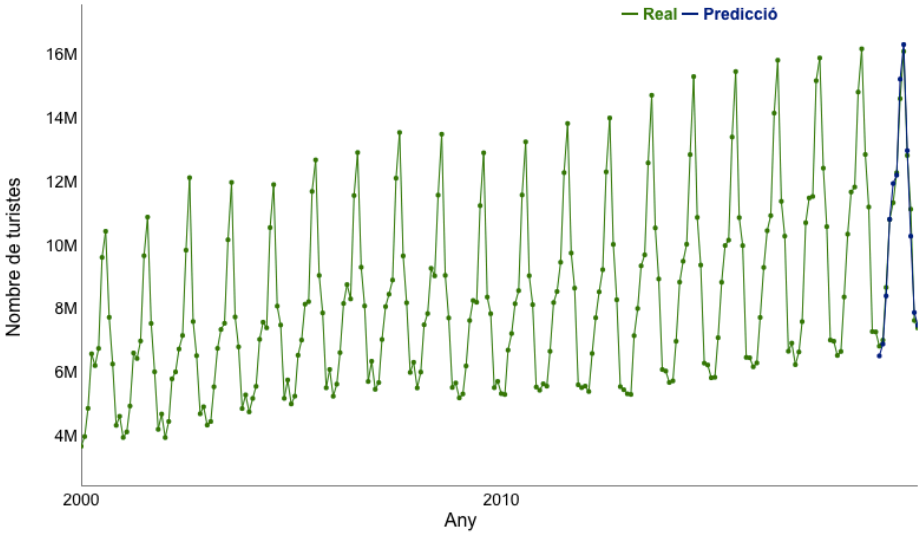


Font: Elaboració Pròpia

Una vegada ajustat el model, a la figura s'observen les dues funcions de pèrdua, una per l'entrenament i l'altra per la prova, cal recordar que s'han calculat amb l'error absolut mitjà. S'observa a més que en cap moment el error del test comença a crèixer, això es un bon indici per veure que no existeix un sobreajustament (*overfitting*) del model. Un problema de sobreajustament sorgeix del sobreentrenament de l'algorisme d'aprenentatge, és a dir, el model s'ajustarà molt bé a les dades conegudes però tindrà problemes per predir a partir de dades que el model no ha vist mai.

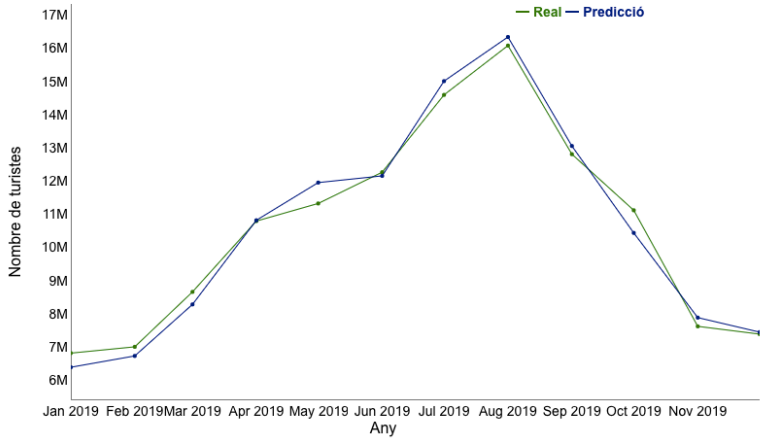
A continuació es mostren les els gràfics de les prediccions:

Figura 32 - Prediccions model LSTM



Font: Elaboració Pròpia

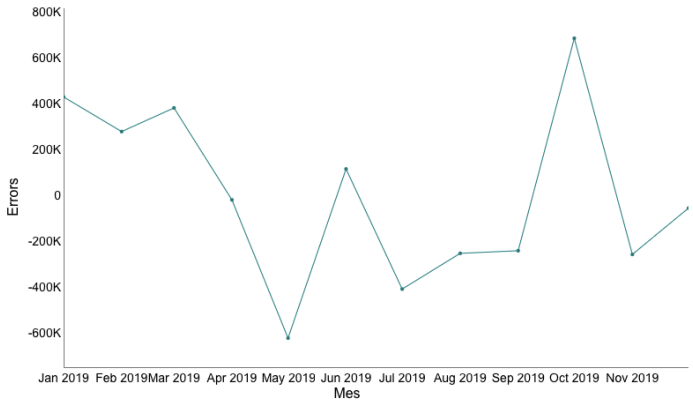
Figura 33 - Prediccions any 2019 model LSTM



Font: Elaboració Pròpia

A simple vista, les prediccions semblen que s'ajusten de manera correcta a les dades, segueixen amb la tendència creixent d'un augment del nombre de turistes i a més a més també mostren aquest caràcter estacional en el que depenent del mes o estació de l'any el nombre de turistes augmenta o disminueix.

Figura 34 - Errors de predició model LSTM



Font: Elaboració Pròpia

Per aquest model, obtenim els següents criteris de validació:

Taula 12 - Criteris de validació model LSTM

	AIC	*EQM*	*EAM*	*EPAM*	*RMSE*
LSTM		148344947766	295274	2,85%	385155,7

Font: Elaboració Pròpia

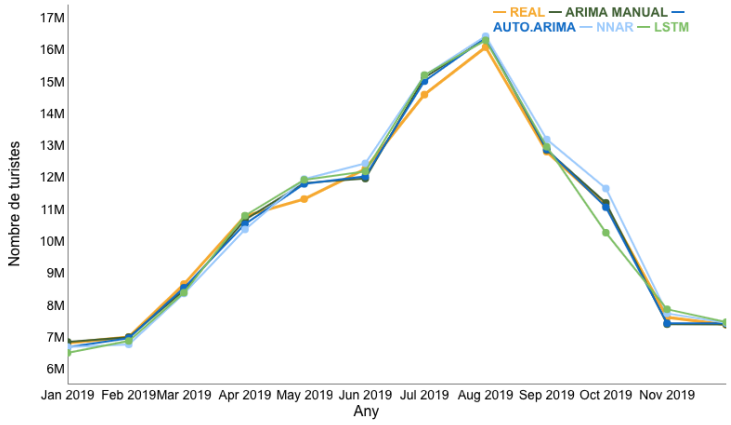
Per aquest model el valor de l'EPAM és d'un 2,85% per tant el model té una bona capacitat predictiva.

3.5 Comparació

Una vegada s'han avaluat tots els models, el següent pas és realitzaran les comparacions principals entre ells per tal de veure quin model s'ajusta millor a les dades. Per comparar-los s'utilitzaran els criteris de validació que s'han mencionat anteriorment.

Primer de tot, es mostrarà un gràfic amb totes les prediccions i el valor real:

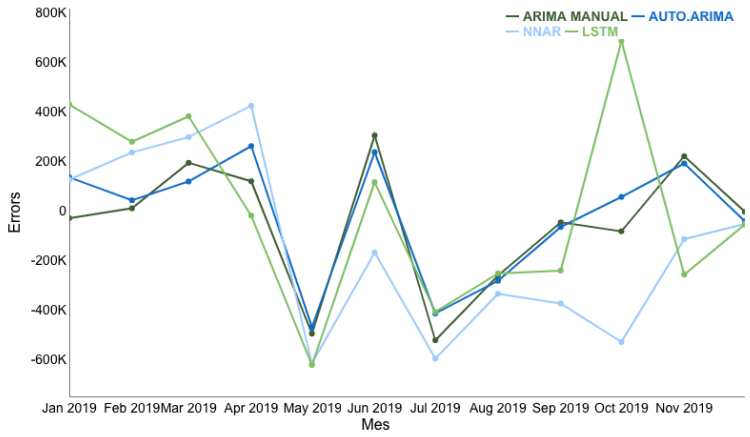
Figura 35 - Prediccions amb tots els models



Font: Elaboració Pròpia

Com ja s'ha vist durant l'estudi individual dels mètodes, s'observa que totes les prediccions han aconseguit un ajust força acceptable, en cap dels mètodes es destaca ninguna predicció fora de l'usual ni cap anomalia. El mateix passa si s'observen els errors de predicció de tots els mètodes:

Figura 36 - Errors de predicció de tots els models



Font: Elaboració Pròpia

La taula següent, mostra els criteris de validació utilitzats per cada un dels mètodes per tal de poder fer una comparació entre ells:

Taula 13 - Criteris de comparació per tots els models

	AIC	***EQM***	***EAM***	***EPAM***	***RMSE***
ARIMA MANUAL	6.115,18	66632238273	192097,5	1,68%	258132,2
AUTO ARIMA	6.090,14	56756969274	192526,2	1,76%	238237,2
NNAR		13988408124	325025,1	2,98%	374010,8
LSTM		148344947766	295274	2,85%	385155,7

Font: Elaboració Pròpia

Per la comparació de tots els mètodes sobretot caldrà observar els valors de l'Error Percentual Mitjà (*EPAM*) i de l'arrel de l'Error Quadràtic Mig (*RMSE*).

Sembla ser que qualsevol dels mètodes ha resultat ser vàlid per la predicció de turistes a Espanya. Encara que òbviament, són mètodes diferents i per tant s'obtindran resultats diferents, sobretot si es comparen els models *ARIMA* amb Xarxes Neuronals.
De fet, els models *ARIMA* es van dissenyar específicament per treballar amb sèries temporals, mentre que les Xarxes Neuronals Recurrents estan dissenyades per treballar amb dades seqüencials. Amb això es vol recalcar de nou que els dos són vàlids per la predicció, però sembla ser que diferents situacions i dades requeriran diferents aproximacions. És a dir, depenent de la índole de les dades (estacionalitat, tendència, mida de les dades..) es requeriran unes aproximacions o unes altres.

En aquest problema, s'ha vist que els models *ARIMA* tenen uns errors de predicció ***(Taula 13)*** més baixos que els models de Xarxes Neuronals. Això pot ser degut a que les Xarxes Neuronals requereixen una quantitat de dades d'entrenament molt més elevada i quan aquestes dades mostren patrons complexos al llarg dels anys. A més, sembla ser que els models *ARIMA* treballen millor que les Xarxes amb dades que mostren comportaments estacionals, tendència, correlacions...com amb les dades que s'està treballant.

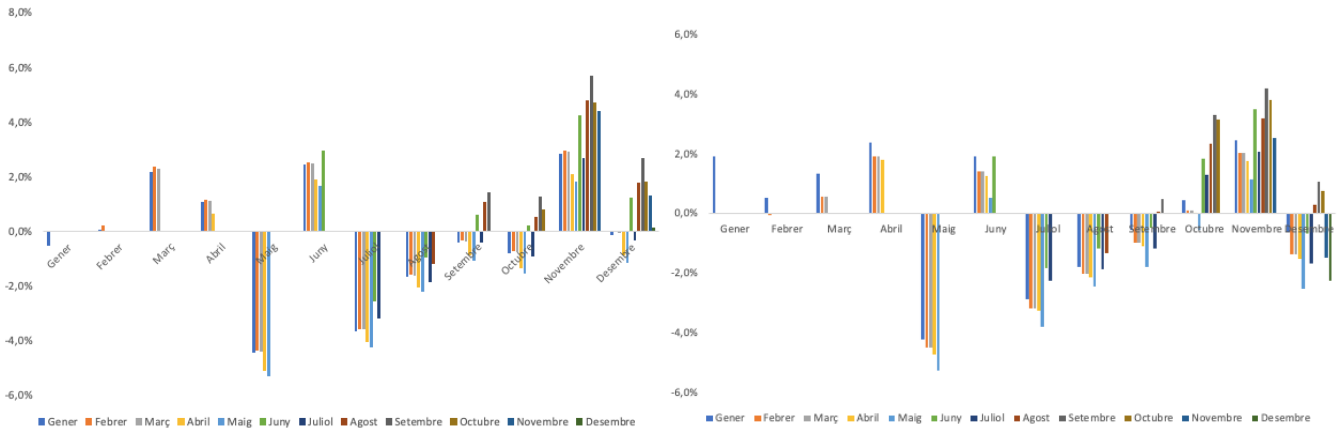
Per tant, per aquest treball s'escolliran els models *ARIMA* per sobre de les Xarxes Neuronals però és molt important recalcar que aquesta idea no es pot generalitzar.

Ara, caldrà decidir quin dels dos models de la metodologia ARIMA és el millor. Per això, s'utilitzarà els resultats analítics ***(Taula 13)***, en els que s'observa que l'EPAM del model manual és més baix que el model automàtic però el RMSE del model automàtic és més baix i a més el AIC també és més baix, per tant, ambdós models són vàlids però en aquest cas s'escollirà el model automàtic.

També mitjançant els dos models d'ARIMA, s'han realitzat prediccions de 2019 afegint a les dades d'entrenament cada vegada un mes més, amb la finalitat de mirar si els models tenen la capacitat d'adaptar-se a les dades a mesura que reben noves observacions, és a dir, nova informació.

Per això s'han obtingut dos taules equivalents a les prediccions i als errors per els dos models d'ARIMA (que es troben a l'annex) i els següents gràfics realitzats amb l'Excel que mostren els errors en termes percentuals:

Figura 37 - Errors percentuals model manual (esquerra) i automàtic (dreta)



Font: Elaboració Pròpia

En general, no sembla que ningun dels dos models s'adapti millor a les dades a mesura que passa el temps, per tant, es seguirà amb l'elecció del model automàtic.
En el cas de que es vegués que un dels dos models s'ajustés millor es plantejaria juntament amb els resultats analítics l'elecció d'aquell model. Ja que, que un model sigui capaç d'ajustar-

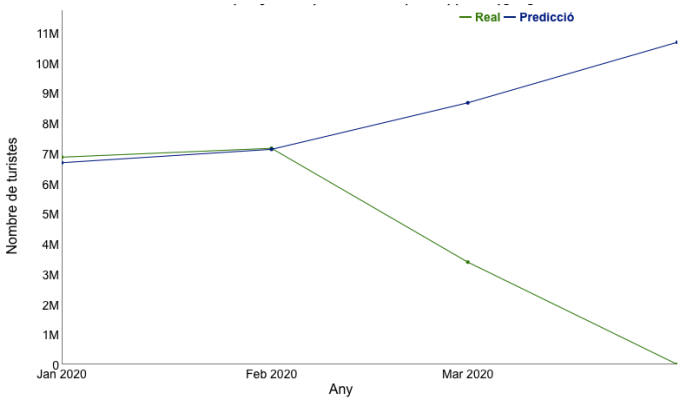
se millor a les dades amb més informació voldrà dir que aquell model tindrà una millor capacitat d'aprendre per ell mateix a mesura que li entra nova informació. Un fet que si que es important remarcar i que es podria estudiar, és que s'observa que hi ha mesos de l'any en que independentment del model de predicció que s'utilitzi, els errors són més grans que altres mesos, per exemple, el mes de maig o el mes de novembre, les prediccions estan menys ajustades que altres mesos, aquest és un fet curiós que es podria estudiar en un futur.

3.6 Impacte COVID-19 al turisme a Espanya

En aquest apartat es realitzarà un estudi de l'impacte de la crisi del coronavirus al turisme a Espanya. L'objectiu d'aquest apartat es estudiar com aquesta pandèmia ha afectat a un sector tant important a Espanya com és el turisme. Actualment, el turisme és el sector que mes riquesa aporta a l'economia espanyola amb un total de 176 milions d'euros anuals que representen un 14% del PIB i a més a més aporta 2,8 milions de llocs de treball[22].

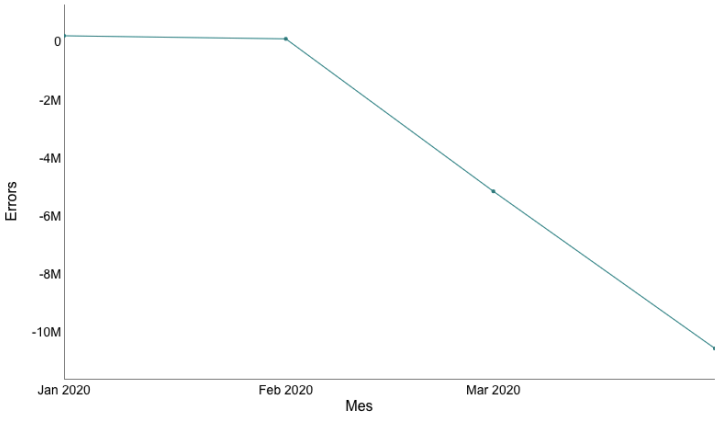
Per fer-ho es realitzaran prediccions de gener a abril de 2020 amb el mètode escollit durant la comparació de mètodes de l'apartat anterior, és a dir, amb el model ARIMA automàtic. Una vegada realitzades aquestes prediccions el que es vol estudiar són els errors, és a dir, es vol estudiar la diferència entre els turistes que han arribat a Espanya en comparació amb els que s'esperava que arribessin.
A la figura i taula següent s'observarà les prediccions realitzades durant els quatre primers mesos de 2020, el seu valor real i els errors de predicció:

Figura 39 - Prediccions del nombre de turistes (gener 2020-abril 2020)



Font: Elaboració Pròpia

Figura 38 - Errors de predicció turistes (gener 2020-abril 2020)



Font: Elaboració Pròpia

[22] Canalis, X. (30 / Ago / 2019). El turismo es el sector que más riqueza aporta a la economía española. *Hosteltur*.

Taula 14 - Valor real / Predicció / Error del nombre de turistes

	REAL	***PREDICCIÓ***	***ERROR***
Gener 2020	6.872.255	6.688.238	184.017,26
Febrer 2020	7.166.067	7.131.598	34.469,24
Març 2020	3.391.448	8.674.906	-5.283.457,96
Abril 2020	0	10.681.771	-10.681.770,90

Font: Elaboració Pròpia

Observant tant els gràfics com la taula ja s'observa una caiguda estrepitosa del nombre de turistes que arriben a Espanya durant aquests mesos. Aquesta mala predicció sembla lògica ja que és impossible o pràcticament impossible predir un esdeveniment similar com aquest. Sembla que els dos primers mesos el nombre de turistes es va mantenir estable però és al mes de gener quan hi ha ja una caiguda de 5 milions de turistes. Aquesta caiguda comença al mes de març coincidint amb l'arribada massiva del coronavirus a Espanya i l'estat d'alarma[23] decretat el 14 de març de 2020[24] pel govern Espanyol. A partir d'aquest moment va quedar prohibida l'entrada o sortida de ciutadans excepte casos excepcionals. Per tant, s'atribueixen aquests 3.391.448 turistes a la primer quinzena del març de 2020. Com s'observa durant el mes d'abril el nombre de turistes és 0 quan s'esperaven 10.681.771 coincidint amb Setmana Santa.

Amb les dades que es tenen fins ara no es poden obtenir unes grans conclusions, però si que es pot veure que el turisme ha estat un dels sectors més afectats per aquesta crisi que això comporta unes repercussions molt grans pel conjunt de l'economia espanyola ja que com s'ha mencionat el turisme és el sector que més riquesa aporta a l'economia espanyola.

Es considera que podria ser molt interessant l'estudi a partir d'ara ja que a partir del 21 de juny es decretarà el final de l'estat d'alarma. Com a conseqüència els ciutadans espanyols podran viatjar a una altra comunitat autònoma així com a altres països que formin part de l'espai *Shengen*, amb l'excepció de Portugal, que es podrà a partir de l'1 de juliol. A més a més, també es permetrà l'entrada de turistes estrangers a Espanya. Per últim, el govern espanyol ha destinat 4.262 milions d'euros a reactivar el turisme[25].
Per tant, a partir d'aquest 21 de juny pot ser molt interessant estudiar com evoluciona la recuperació d'un sector tant important com el turisme i veure no solament Espanya sinó tot el món en general com es recupera d'aquesta crisi mundial.

[23] Alarma, E. d. (s.d). Recollit de Wikipedia: https://es.wikipedia.org/wiki/Estado_de_alarma

[24] La Moncloa. (2020). *El Gobierno decreta el estado de alarma para hacer frente a la expansión de coronavirus COVID-19.*

[25] BLÁZQUEZ, P. (18 / Jun / 2020). El Gobierno destina 4.262 millones de euros a relanzar el turismo. *La Vanguardia*.

4. CONCLUSIONS

Finalment i tenint en compte els resultats obtinguts, en aquest apartat es recullen les conclusions i es proposen línies d'interès per continuar treballant en el mateix sentit.

En referència a la recerca i aprofundiment de l'aplicabilitat dels mètodes que utilitzen les sèries temporals ARIMA i xarxes neuronals:

- Després de la consulta de múltiples fonts bibliogràfiques s'ha constatat que el mètode tradicionalment emprat pel tractament de dades de sèries temporals és el mètode ARIMA. Pels mètodes ARIMA s'han utilitzat dos models: el model manual, en els que s'han realitzat transformacions i s'han escollit els paràmetres i el model automàtic, que s'ha implementat amb la funció *auto.arima.*

- Existeixen altres mètodes que també es poden utilitzar per aquest tipus de sèries entre els quals destaquen les xarxes neuronals que s'estan utilitzant actualment en estudis de predicció de diversos àmbits i disciplines. Aquestes xarxes són complexes d'entendre i aplicar, justament per la seva complexitat, s'ha demostrat que tenen un gran potencial i una gran capacitat d'autoaprenentatge. Per aquest estudi s'han utilitzat dos tipus de xarxes neuronals: les xarxes neuronals *feedforward* i les xarxes neuronals recurrents (LSTM).

Respecte a la comparació dels mètodes:

- Tant els resultats obtinguts amb els mètodes ARIMA com les Xarxes Neuronals s'ajusten de manera correcta a les dades.

- S'ha comprovat que els mètodes ARIMA ofereixen prediccions mes ajustades a la realitat que els mètodes de xarxes neuronals. Aquesta conclusió no es pot generalitzar per totes les sèries de dades, el parer general és que les xarxes neuronals requereixen una quantitat de dades d'entrenament molt més elevada i poden funcionar millor quan aquestes dades mostren patrons complexos al llarg dels anys.

- Dins dels mètodes ARIMA, mitjançant criteris de validació s'ha escollit el model automàtic com el model que ajusta millor. A l'introduir noves dades en aquests models per veure com s'adapten a mesura que passa el temps, s'ha constatat que cap dels dos destaca per adaptar-se de manera significativa a les dades, però s'ha vist que hi ha diversos mesos de l'any en que l'error és més gran que en altres de manera sistemàtica.

Pel que fa al tractament d'un esdeveniment extraordinari en el interval temporal de la sèrie:

- S'ha pogut constatar que la sèrie de dades ha seguit fins l'actualitat un patró en el que el nombre de turistes era creixent al llarg dels anys i amb un component estacional. Aquestes dades abasten el període de gener de l'any 2000 fins desembre de l'any 2019.

- La pandèmia del COVID-19 ha provocat un canvi substancial en les dades de turisme a Espanya i al món. Aquest succés i qualsevol altre esdeveniment extraordinari és impossible de predir per qualsevol mètode de predicció.

- A causa de l'efecte de la pandèmia sobre el comportament turístic s'ha considerat necessari tractar de forma independent les dades pre i post pandèmia. Les dades que fan referència a la pandèmia van de gener fins a abril de l'any 2020.

Per últim, en referència als reptes identificats per continuar investigant:

- En el cas de disposar d'un termini més ampli i més recursos per finalitzar aquest treball seria interessant continuar fent recerca i implementant les xarxes neuronals. L'objectiu seria aconseguir unes prediccions mes ajustades a les dades, millorant els models de xarxes neuronals.

- Una nova línia de treball pot ser esbrinar les causes de les diferències dels errors entre mesos, a l'hora d'estudiar com s'adapta el model a les dades a mesura que s'afegeix informació.

- El fet de disposar només de quatre mesos de dades fa que no hagi estat possible poder fer un estudi extens de l'impacte del coronavirus. En aquest, es podria avaluar com es recupera el turisme espanyol, a més, es podria reajustar el model per preparar-lo per possibles situacions similars.

5. BIBLIOGRAFIA

A.M, Turing. (Oct 1950). *Computing Machinery and Intelligence.* (Vol. 59). Mind, New Series.

Alarma, E. d. (s.d). Recollit de Wikipedia: https://es.wikipedia.org/wiki/Estado_de_alarma

BLÁZQUEZ, P. (18 / Jun / 2020). El Gobierno destina 4.262 millones de euros a relanzar el turismo. *La Vanguardia*.

Brownlee, J. (3 / Jul / 2017). *Gentle Introduction to the Adam Optimization Algorithm for Deep Learning*. Recollit de Machine Learning Mistery: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

Brownlee, J. (s.d). *Long Short Term Memory with Python*. Recollit de https://vel.life/阅《long-short-term-memory-networks-with-python》/long-short-term-memory-networks-with-python.pdf: Machine Learning Mastery

Calvo, D. (10 / Des / 2018). *Función de coste – Redes neuronales*. Recollit de Diego Calvo: https://www.diegocalvo.es/funcion-de-coste-redes-neuronales/

Canalis, X. (30 / Ago / 2019). El turismo es el sector que más riqueza aporta a la economía española. *Hosteltur*.

George E.P. Box, Gwilym. M. Jenkins (1970). *Time Series Analysis, Forecasting ans Control.* San Francisco: Holden-Day.

Gradient Descent. (s.d). Recollit de Machine Learning Glossary: https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html

Hyndman, R. (s.d). *auto.arima*. Recollit de RDocumentation: https://www.rdocumentation.org/packages/forecast/versions/8.12/topics/auto.arima

Hyndman, R. (s.d). *BoxCox.lambda*. Recollit de RDocumentation: https://www.rdocumentation.org/packages/forecast/versions/8.12/topics/BoxCox.lambda

Hyndman, R. (s.d). *forecast*. Recollit de RDocumentation: https://www.rdocumentation.org/packages/forecast/versions/8.12

Hyndman, R. (s.d). *Guerrero's method for Box Cox lambda selection*. Recollit de feasts: https://feasts.tidyverts.org/reference/guerrero.html

Hyndman, R. (s.d). *nnetar*. Recollit de RDocumentation: https://www.rdocumentation.org/packages/forecast/versions/8.12/topics/nnetar

INE. (s.d). *Estadística de movimientos turísticos en frontera. Frontur*. Recollit de Instituto Nacional de Estadística: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176996&menu=resultados&secc=1254736195367&idp=1254735576863#!tabs-1254736195367

INE. (s.d). *Estadística de movimientos turísticos en frontera. Frontur. Metodologia*. Recollit de INE: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176996&menu=metodologia&idp=1254735576863

INE. (s.d). *Instituto Nacional de Estadística*. Recollit de https://www.ine.es

Kan, C. N. (11 / Des / 2018). *Quick ML Concepts: Tensors*. Recollit de Medium, Towards Data Science: https://towardsdatascience.com/quick-ml-concepts-tensors-eb1330d7760f

Keras. (s.d). *Keras*. Recollit de https://keras.io

La Moncloa. (2020). *El Gobierno decreta el estado de alarma para hacer frente a la expansión de coronavirus COVID-19.*

learn, s. (s.d). *MinMaxScaler*. Recollit de scikit learn: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
Nielsen, M. A. (2015). Neural Networks and Deep Learning Chapter 2 How the Backpropagation Algorith Works. Determination Press.
Olah, C. (2015). *Understanding LSTM Networks*. Recollit de https://colah.github.io/posts/2015-08-Understanding-LSTMs/
Recessió global 2008-2012. (s.d). Recollit de Wikipedia: https://ca.wikipedia.org/wiki/Recessió_global_2008-2012
Sharma, S. (6 / Set / 2017). *Activation Functions in Neural Networks*. Recollit de Towards Data Science: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6
TensorFlow. (s.d). *TensorFlow*. Recollit de https://www.tensorflow.org
TURESPAÑA. (s.d). *Frontur*. Recollit de http://estadisticas.tourspain.es/es-ES/estadisticas/frontur/Paginas/default.aspx

6. ANNEX

6.1 Codi

En el següent enllaç, es troba tot el codi utilitzat durant l'estudi tant amb *R* com amb *Python*. També estan a disposició les dades utilitzades:

https://github.com/czaldivar18/Treball-de-Fi-de-Grau---Codi

6.2 Taules

En aquest apartat s'afegiran alguns resultats que no s'han vist durant el cos del treball, com són les taules amb els resultats obtinguts de prediccions de cada un dels mètodes i també les taules de les prediccions i errors calculats mes a mes.

- Mètode manual:

	Predicció	*Valor Real*	*Error*
Gener	6817370	6782568	-34.802
Febrer	6967088	6972244	5.156
Març	8437781	8626912	189.131
Abril	10644387	10759139	114.752
Maig	11791347	11289498	-501.849
Juny	11930554	12230598	300.044
Juliol	15087373	14558858	-528.515
Agost	16310903	16043994	-266.909
Setembre	12825012	12773287	-51.725
Octubre	11171722	11084163	-87.559
Novembre	7373481	7589213	215.732
Desembre	7364014	7355018	-8.996

- Mètode automàtic:

	Predicció	Valor Real	Error
Gener	6652010	6782568	130558
Febrer	6934583	6972244	37661
Març	8513071	8626912	113841
Abril	10502127	10759139	257012
Maig	11764181	11289498	-474683
Juny	11998272	12230598	232326
Juliol	14979171	14558858	-420313
Agost	16332056	16043994	-288062
Setembre	12843709	12773287	-70422
Octubre	11032803	11084163	51360
Novembre	7402643	7589213	186570
Desembre	7402522	7355018	-47504

- Xarxes Neuronals Autoregressives (NNAR):

	Predicció	Valor Real	Error
Gener	6661358	6782568	121210
Febrer	6740754	6972244	231490
Març	8333858	8626912	293054
Abril	10339033	10759139	420106
Maig	11912901	11289498	-623403
Juny	12404450	12230598	-173852
Juliol	15161913	14558858	-603055
Agost	16384967	16043994	-340973
Setembre	13153414	12773287	-380127
Octubre	11619938	11084163	-535775
Novembre	7708695	7589213	-119482
Desembre	7412790	7355018	-57772

- Long-Short Term Memory

	Predicció	Valor Real	Error
Gener	6782568	6357048	-425520
Febrer	6972244	6697915	-274329
Març	8626912	8249266	-377646
Abril	10759139	10783452	24313
Maig	11289498	11918265	628767
Juny	12230598	12119783	-110815
Juliol	14558858	14973279	414421
Agost	16043994	16302192	258198
Setembre	12773287	13019992	246705
Octubre	11084163	10401952	-682211
Novembre	7589213	7852130	262917
Desembre	7355018	7415357	60339

Taules de prediccions mes a mes:

- Mètode manual

	Gener	Febrer	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre	REAL
Gener	6817370												6782568
Febrer	6967088	6957881											6972244
Març	8437781	8423393	8427256										8626912
Abril	10644387	10632084	10638196	10687823									10759139
Maig	11791347	11779244	11784589	11865005	11885628								11289498
Juny	11930554	11918873	11924180	11995692	12024942	11867946							12230598
Juliol	15087373	15076218	15081461	15149092	15177200	14931668	15023057						14558858
Agost	16310903	16299852	16304923	16372877	16396407	16196223	16339553	16231902					16043994
Setembre	12825012	12814751	12819490	12885454	12911058	12696737	12826194	12634685	12590456				12773287
Octubre	11171722	11161644	11166282	11230845	11253395	11058701	11184111	11023826	10944424	10992449			11084163
Novembre	7373481	7363703	7368303	7429428	7451028	7267957	7385930	7225566	7156752	7230864	7253473		7589213
Desembre	7364014	7354570	7359043	7418070	7438568	7264546	7377619	7224859	7157173	7220747	7258011	7343052	7355018

- Errors Percentuals mètode manual

	Gener	Febrer	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre
Gener	-0,5%											
Febrer	0,1%	0,2%										
Març	2,2%	2,4%	2,3%									
Abril	1,1%	1,2%	1,1%	0,7%								
Maig	-4,4%	-4,3%	-4,4%	-5,1%	-5,3%							
Juny	2,5%	2,5%	2,5%	1,9%	1,7%	3,0%						
Juliol	-3,6%	-3,6%	-3,6%	-4,1%	-4,2%	-2,6%	-3,2%					
Agost	-1,7%	-1,6%	-1,6%	-2,0%	-2,2%	-0,9%	-1,8%	-1,2%				
Setembre	-0,4%	-0,3%	-0,4%	-0,9%	-1,1%	0,6%	-0,4%	1,1%	1,4%			
Octubre	-0,8%	-0,7%	-0,7%	-1,3%	-1,5%	0,2%	-0,9%	0,5%	1,3%	0,8%		
Novembre	2,8%	3,0%	2,9%	2,1%	1,8%	4,2%	2,7%	4,8%	5,7%	4,7%	4,4%	
Desembre	-0,1%	0,0%	-0,1%	-0,9%	-1,1%	1,2%	-0,3%	1,8%	2,7%	1,8%	1,3%	0,2%

- Mètode automàtic:

	Gener	Febrer	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre	REAL
Gener	6652010												6782568
Febrer	6934583	6972548											6972244
Març	8513071	8576677	8576739										8626912
Abril	10502127	10551441	10551262	10563900									10759139
Maig	11764181	11797512	11797177	11821511	11882794								11289498
Juny	11998272	12056247	12056319	12075139	12166776	11998353							12230598
Juliol	14979171	15022916	15022682	15034810	15113703	14824387	14887780						14558858
Agost	16332056	16365949	16365732	16389142	16436732	16231885	16345926	16254593					16043994
Setembre	12843709	12896398	12896354	12912166	13002149	12833187	12920619	12766360	12708479				12773287
Octubre	11032803	11071928	11071476	11083872	11144467	10880691	10940745	10823121	10716710	10734829			11084163
Novembre	7402643	7434008	7433818	7455464	7503135	7324443	7430921	7347717	7270018	7300517	7395405		7589213
Desembre	7402522	7454106	7454139	7467949	7541705	7403557	7478974	7333114	7276787	7299926	7464684	7519698	7355018

- Errors Percentuals mètode automàtic:

	Gener	Febrer	Març	Abril	Maig	Juny	Juliol	Agost	Setembre	Octubre	Novembre	Desembre
Gener	1,9%											
Febrer	0,5%	0,0%										
Març	1,3%	0,6%	0,6%									
Abril	2,4%	1,9%	1,9%	1,8%								
Maig	-4,2%	-4,5%	-4,5%	-4,7%	-5,3%							
Juny	1,9%	1,4%	1,4%	1,3%	0,5%	1,9%						
Juliol	-2,9%	-3,2%	-3,2%	-3,3%	-3,8%	-1,8%	-2,3%					
Agost	-1,8%	-2,0%	-2,0%	-2,2%	-2,4%	-1,2%	-1,9%	-1,3%				
Setembre	-0,6%	-1,0%	-1,0%	-1,1%	-1,8%	-0,5%	-1,2%	0,1%	0,5%			
Octubre	0,5%	0,1%	0,1%	0,0%	-0,5%	1,8%	1,3%	2,4%	3,3%	3,2%		
Novembre	2,5%	2,0%	2,0%	1,8%	1,1%	3,5%	2,1%	3,2%	4,2%	3,8%	2,6%	
Desembre	-0,6%	-1,3%	-1,3%	-1,5%	-2,5%	-0,7%	-1,7%	0,3%	1,1%	0,7%	-1,5%	-2,2%