TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Mètodes alternatius per la predicció del consum de les llars espanyoles

Autor: Daniel Villalobos Torrejón

Director: Ernest Pons Fanals

Departament: d'Econometria, Estadística i Economia Aplicada

Convocatòria: Gener 2021

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM

En aquest treball s'estudien diferents mètodes de predicció aplicats al consum alimentari. Per a la realització d'aquests, s'utilitzen dades de la població espanyola, a més es mostren els processos necessaris per obtenir les prediccions esmentades. Aquest estudi utilitza algoritmes clàssics, en l'estudi economètric, com els models ARIMA, i altres, com els Support Vector Machines, del camp de la intel·ligència artificial, això ens permetrà comparar l'actuació d'aquests dos mètodes davant una variable econòmica.

Paraules clau: Regressió, Support Vector Machine, ARIMA, Econometria, Predicció.

ABSTRACT

This work studies different predictive methods applied to food consumption. For the realization of these, data of Spanish population are used, in addition the necessary processes are shown to obtain the mentioned predictions. This study uses classical algorithms, in the econometric, such as ARIMA models, and other ones, such as Support Vector Machines, from the Artificial Intelligence field, this will allow us to compare both performances against an economic variable.

Key Words: *Regression, Support Vector Machines, ARIMA, Econometric, Prediction.*

Classificació AMS: 62M20 Prediction

62M10 Time Series, auto-correlation, regression, etc

Abans de començar estrictament amb el treball m'agradaria dedicar unes paraules d'agraïment a aquelles persones que m'han ajudat a la realització d'aquest. En primer lloc, als meus companys per haver-me de suportar durant tots aquests anys. En segon lloc a la meva família per donar-me tot el suport necessari en aquells moments més complicats. I finalment al meu tutor, Ernest, per saber-me guiar en la realització d'aquest treball.

Índex

1. INTRODUCCIÓ

Una de les majors qüestions que la humanitat ha hagut d'afrontar al llarg de la seva historia es i serà la preocupació pel futur. Aquesta preocupació per esdeveniments que ens afectaran en un període més llunyà ha esta sempre present, tot i que ha anat variant amb les èpoques en que ens trobéssim. En la antiguitat, a la prehistòria quan encara no existien civilitzacions, podia ser el que menjaré demà o on dormiré aquesta nit, en canvi en la actualitat pot ser des de quin valor tindrà les accions de la meva empresa a borsa demà, com quan sortirà la vacuna per a certa malaltia, per tant avarca una gran quantitat d'àrees diferents. Es justament per aquest motiu que en els darrers anys, sobretot, han aparegut una sèrie de models matemàtics que a partir de l'anàlisi de dades massives, enteses com a matrius amb una gran quantitat de dades, intenten aproximar-se al màxim a la realitat. Aquests models més moderns estan englobats dintre del que es coneix com a Machine Learning. Però abans del sorgiment d'aquest models, ja hi existien d'altres que feien la mateixa funció com pot ser una regressió lineal simple. Es aquest el motiu pel qual sorgeix aquest treball, es vol comprovar si els models més actuals tenen un major ajust a les dades reals i futures que no pas els models antics, o si pel contrari no es així.

Per triar el temari i fer un estudi que pugui servir en l'actualitat s'ha volgut utilitzar dades, donades pel Ministeri de pesca, agricultura i alimentació del govern espanyol, que recullen els hàbits de consum alimentari de la població espanyola. Ja que es molt probable que donada la situació excepcional que estem vivint a causa del covid-19, les prediccions que es varen realitzar l'any passat cap a aquest no s'hagin apropat a la realitat, i per tant en aquest treball es vol realitzar una predicció amb dades anteriors a l'aparició del coronavirus i intentar que els models que es realitzin s'ajustin a la realitat viscuda, i poder observar possibles canvis en aquest hàbits.

Per aquestes projeccions futures es realitzaran dos models amb les mateixes dades, es farà un exemple amb un model més clàssic com són els models ARIMA (model autoregressiu integrat de mitjanes mòbils), desenvolupats per Box i Jenkins als anys 70 del segle passat, aquest models es caracteritzen per ser un model dinàmic de series temporals. I el segon prototip que s'executarà seran els models de *Support Vector Machines* (SVM, a partir d'aquí), que entren dins del bloc d'algoritmes d'aprenentatge supervisat.

Al tractar-se d'un projecte de final de grau de la doble titulació d'economia i estadística, un dels motius pel qual s'ha triat fer un anàlisi predictiu a partir de models economètrics o models d'aprenentatge supervisat, es que es podria considerar l'econometria com el nexe d'unió entre l'estadística i la economia. El segon motiu pel qual s'ha volgut fer aquest tipus d'anàlisi ha sigut el voler aplicar els coneixements, bastament explicats, sobre els models ARIMA, i la voluntat de voler aprofundir en aquells models els quals no s'han pogut donar d'una manera tant exhaustiva, com son els SVM.

2. MARC TEÒRIC

En aquesta secció es farà una explicació teòrica dels mètodes que utilitzarem per fer l'anàlisi, sobre el qual tracta el treball. En primer lloc elaborarem el model més clàssic, es a dir, el model ARIMA i seguidament es tractaran els models desenvolupats de manera més recent.

2.1 Models ARIMA

Els models ARIMA, o també coneguts com a Box-Jenkins, nom donat pels cognoms dels seus desenvolupadors, de forma molt ràpida i clara són un model estadístic dinàmic que a partir de l'ús de variacions i regressions, en una sèrie temporal, tenen com a finalitat la realització de prediccions cap al futur. Al tractar-se de un model dinàmic implica que les estimacions futures venen explicades per les dades del passat i no pas per variables independents.

Donada aquesta breu introducció continuem per conèixer amb més profunditat aquests models. Per tal de seguir es necessari que s'expliquin uns conceptes bàsics, per tal d'entendre amb claredat el treball, ja que més endavant s'utilitzaran en la realització d'aquest.

La primera idea que cal conèixer es la de ***procés estocàstic***. Es coneix com a procés estocàstic al conjunt de variable aleatòries ordenades segons el subíndex t. En l'anàlisi de sèries temporals la notació es la següent: $\{x_t\}_{t=1}^{T} \equiv x_1, x_2, \dots, x_T$. Per tant una sèrie temporal es la realització d'un procés estocàstic.

El segon concepte important es l'***estacionarietat***. En aquest concepte existeixen dues definicions. La primera es la de ***estacionarietat forta***, que es requereix que la funció de distribució conjunta del procés estocàstic no depengui del temps. De manera algebraica s'expressa de la següent manera: $F(x_1, \dots, x_T) = F(x_{1+h}, \dots, x_{T+h})$. I la segona definició es l'***estacionarietat dèbil,*** que requereix que els moments de primer i segon ordre no depenguin del temps. Les expressions en llenguatge més tècnic són les següents: $E(x_t) = E\,(x_{t+h}) = \mu < \infty; V(x_t) = V(x_{t+h}) = \sigma^2 < \infty; Cov(x_t, x_s) = Cov(x_{t+h}, x_{s+h}) = \gamma_k < \infty, k = |t - s|$. Es important que la sèrie sempre sigui estacionaria per tal de garantir certes condicions dels estimadors, com la consistència, l'eficiència i la no presència de biaix.

Seguidament es parlarà del concepte d'***ergodicitat.*** Un procés estocàstic es ergòdic si es possible estimar de manera consistent les seves característiques a partir d'una realització seva. Per exemple, x_t serà ergòdic per l'esperança si es compleix: $\bar{x} \xrightarrow{p} E(x_t)$. Una relació important entre els dos últims conceptes es, sempre que un procés estocàstics sigui estacionari serà ergòdic, però un procés ergòdic no sempre serà estacionari.

Pels models ARIMA es important que les dades siguin estacionàries, ja que com s'ha mencionat abans per tal de garantir certes condicions dels estimadors, i en cas contrari, que no sigui estacionaria, s'haurà de fer alguna transformació sobre les dades, com per exemple una diferenciació.

Per tal de fer aquesta comprovació es necessari observar els gràfics de la **Funció d'Autocorrelació Simple (FAS)** i la **Funció d'Autocorrelació Parcial (FAP).** La FAS ens ajudarà a observar la dependència dels valors en un determinat període amb els mateixos de k períodes anteriors, es a dir mesurem la tendència d'una sèrie temporal. La forma algebraica d'aquesta funció es: $\rho_k = \frac{\gamma_k}{\sqrt{\gamma_0}\sqrt{\gamma_0}} = \frac{\gamma_k}{\gamma_0}$. La FAP en canvi mesura la "correlació neta", es a dir, la correlació entre dos valors del procés estocàstic una vegada s'ha descomptat la influencia dels membres intermedis. El càlcul de la FAP es realitza de la següent manera: $\tilde{x}_t = \emptyset_{11}\tilde{x}_{t-1} + \varepsilon_t$; $\emptyset_{11} = \frac{d_{x_t}}{d_{x_{t-1}}}$, per al primer component $\emptyset_{11}$, el segon component, $\emptyset_{22}$ es calcula de la següent manera: $\tilde{x}_t = \emptyset_{21}\tilde{x}_{t-1} + \emptyset_{22}\tilde{x}_{t-2} + \varepsilon_t$. Com a dada rellevant, el primer valor de la FAS i de la FAP sempre serà el mateix ja que no existeix cap retard intermedi.

Seguidament desglossarem els models Box-Jenkins en els seus components. Per explicar-los de manera individual, i més endavant agrupar-los tots. El primer component seran els models autoregressius, seguidament s'explicaran els models de mitjanes mòbils i finalment el procés de diferenciació

2.1.1 Models Autoregressius

Aquests models són una representació d'un procés aleatori, en el qual la variable de interès depèn linealment de les observacions passades, representat com $AR(p)$. Formalment es representen de la següent manera:

$$x_t = f(t) + \emptyset_1 x_{t-1} + \cdots + \emptyset_p x_{t-p} + \varepsilon_t; amb\ \varepsilon \sim iid(0, \sigma^2)\ i\ f(t)\begin{cases} 0 \\ \mu \\ \mu + \beta_t \end{cases}$$

Per comprovar que aquest models siguin estacionaris han de complir que la funció determinista no depengui del temps, i que les arrels del polinomi autoregressiu siguin superiors, en valor absolut, a la unitat.

Per la primera condició el valor de la funció determinista ha de ser $f(t) = \begin{cases} 0 \\ \mu \end{cases}$. Per la segona condició mitjançant l'operador retard obtenim el polinomi autoregressiu i comprovem que les arrels siguin superiors a 1. De forma algebraica el polinomi de retards s'especifica així: $\varphi_p(L)x_t = f(t) + \varepsilon_t \rightarrow \varphi_p(L) = (1 - \varphi_1 L - \cdots - \varphi_p L^p)$, i la condició d'estacionarietat és: $\forall |L_i| > 1$. Es poc habitual l'element determinista no depengui del temps, es per això que s'acostuma per observar l'estacionarietat de la sèrie temporal es tingui en compte només els elements estocàstics de la funció, sense comptar els elements deterministes.

Per determinar l'ordre p dels models Autoregressius ens hem de fixar en els primers p coeficients significatius de la FAP.

2.1.2 Models de mitjanes mòbils

Aquests processos es caracteritzen perquè la variable de interès està explicada pel valor actual i per diferents valors d'un terme estocàstic. Això vol dir que el procés es dinàmic a partir del terme de pertorbació, aquests models són reconeguts com $Ma(q)$. De forma matemàtica es defineixen d'aquesta manera:

$$x_t = f(t) + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{(t-q)}$$

Al contrari dels models Autoregressius aquest models son sempre estacionaris, però per contra no sempre son invertibles, només ho són si les arrels del polinomi de retards son superiors a la unitat. Expressat de manera formal: $\frac{1}{\theta_q(L)} = \Psi_\infty(L) x_t = \varepsilon_t;\ \forall |L_i| > 1.$

Per observar el valor q d'una sèrie temporal ens fixarem en el gràfic de la FAS, ja que trobarem q valors significatius.

2.1.3 Integració

Un procés estocàstic diferenciable es aquell procés no estacionari que un cop diferenciat d vegades es transforma en un procés estacionari. Es conegut com $I(d)$, en els models ARIMA, ja que s'anomena com a ordre d'integració. De manera formal s'explica de la següent manera: Tenim el procés estacionari $x_t = x_{t-1} + \varepsilon_t$, i li apliquem $\Delta x_t = x_t - x_{t-1}$, d'aquesta manera el procés seria $x_t = \varepsilon_t$.

Per detectar si una sèrie temporal es pot diferenciar es necessari l'ús de la representació de la FAS, ja que trobarem valors molt propers a la unitat i un decreixement molt lent de la FAS i de la FAP, ja que els valors seran, també, propers a la unitat i significatius.

2.1.4 Model ARIMA

Els últims tres elements explicats conformen els models $ARIMA\ (p, d, q)$, ja que uneixen els models Autoregressius $AR(p)$, els models de mitjanes mòbils $MA(q)$ i l'ordre d'integració $I(q)$. De manera formal es representen:

$$\varphi_p(L) \Delta^d x_t = \theta_q(L) \varepsilon_t$$

On $\varphi_p(L)$, representa el polinomi de retards i θ_q denota el polinomi de mitjanes mòbils, i $\varepsilon_t \sim iid\ (0, \sigma^2)$.

Un model ARIMA serà estacionari sempre i quan el seu component autoregressiu ho sigui també, a més serà invertible en la mesura que la part de mitjanes mòbils també sigui invertible. L'ordre d'integració ens indica quantes vegades cal diferenciar per a que la variable sigui estacionaria.

Moltes variables de caire econòmic moltes vegades es mesuren de manera periòdica, per exemple per trimestres. Aquest fet implica que en algunes d'aquestes variables

aparegui el fenomen de l'estacionalitat. Com la variable amb la que es treballarà es una variable recollida de forma mensual, pot aparèixer aquest fenomen. Arribat a aquest punt es poden donar dos casos i per tant dues línies d'actuació. La primera considerar que el procés estacionals es determinista, i la segona, que el procés estacional es estocàstic. Com el treball tracta sobre un procés estocàstic escollirem la segona línia d'actuació. Per tant a continuació s'explicaran els models *estocàstic* estacionals ARIMA.

2.1.4.1 Model estocàstic estacional ARIMA(p,d,q)(P,D,Q)

Aquests models per tal d'abreviar els anomenarem $SARIMA\ (p,d,q)(P,D,Q)$. La primera part del model denota la part regular del model, mentre que la segona component representa la part estacional del model.

Aquests models es representen de manera formal de la següent manera:

$$\varphi(L)\Phi_\rho(L^s)\Delta_s^D\Delta^d x_t = \theta_q(L)\Theta_Q(L)\varepsilon_t$$

Per a realitzar aquest models Box i Jenkins van establir un seguit de passos a seguir, anomenats, per raons obvies, *metodologia Box i Jenkins.* Consta de cinc fases que explicarem a continuació d'una manera breu i concisa, més endavant a l'hora de realitzar els propis anàlisis es farà més èmfasis:

1) **Diferenciació:** En aquesta fase es pretén garantir l'estacionarietat en variància de la sèrie temporal, en el cas de que no ho sigui es diferenciarà a la part regular del procés.
2) **Identificació**: Aquesta fase es pot dividir en dues més petites
    a. S'analitza l'estacionarietat de la sèrie, en ambdues parts, regular i estacional. Si no ho és s'apliquen d diferències.
    b. Un cop s'ha garantit l'estacionarietat, cal determinar el ordre dels polinomis AR i MA. S'utilitzarà la FAS i la FAP.
3) **Estimació**: Es duu a terme aplicant la màxima versemblança. En cas que no sigui lineal utilitzarem algoritmes matemàtics d'optimització numèrica, per exemple amb mètodes iteratius.
4) **Validació:** S'ha de verificar que es compleixen els supòsits bàsics:
    a. Significació dels paràmetres i bondat de l'ajust
    b. Condicions d'estacionarietat
    c. Terme de pertorbació:
        i. Normalitat
        ii. No Autocorrelació
5) **Predicció**: Es poden realitzar prediccions puntuals i per interval, si aquest es l'objectiu del model.

2.2 Support Vector Machine

Seguidament es realitzarà l'explicació teòrica de la segona part del treball, els ***Support Vector Machine*** o en català Maquines de Vector Suport, des de aquest punt es parlarà

d'aquesta metodologia com SVM. Aquesta metodologia va ser desenvolupada per Vladimir Vapnik i el seu equip al 1996 i, pertany al grup d'algoritmes d'aprenentatge supervisat, això vol dir que a partir d'unes dades d'entrenament es dedueix una funció, i estan molt relacionats amb problemes de classificació i regressió.

Vapnik a l'hora de realitzar el seu algoritme es va basar en un algoritme de classificació anterior desenvolupat per Rosenblatt, el perceptró, aquest es el model matemàtic més bàsic d'una neurona. L'estructura d'una neurona es la següent té multitud d'entrades i un únic canal de sortida. Aleshores el perceptró utilitza una matriu per representar les xarxes neuronals i un discriminador, un vector binari, que traça l'entrada cap a un vector de sortida a partir de la matriu. D'aquesta manera el valor del vector de sortida, que serà o -1 o 1, es pot classificar com una classe o una altra. A partir d'aquí va definir el següent problema d'optimització:

$$Min\ D(\beta, \beta_0) = -\sum y_i(x_i'\beta + \beta_o)$$

Aquest problema serveix per definir l'hiperplà òptim de separació entre les dues classes minimitzant la distància dels punts mal classificats sobre aquest hiperplà. Però aquest algoritme té una limitació molt important, només convergirà l'algoritme si les classes son separables, si pel contrari no son separables no convergirà.

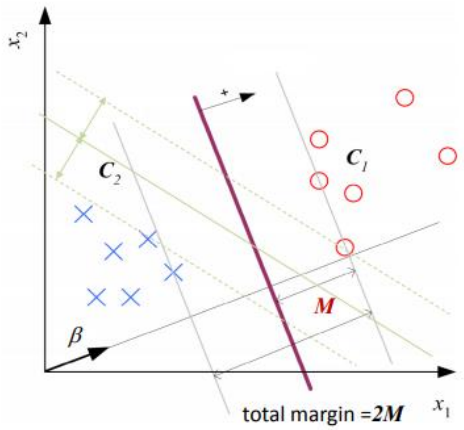
Aleshores Vapnik va buscar la solució a aquesta limitació i va busca l'hiperplà tal que la distància fos màxima als punts més propers de cada classe. De forma algebraica s'expressa així:

$$\max_{\beta} M$$

$$s.a\ \frac{1}{||\beta||} y_i(x_i'\beta + \beta_0) \geq M\ \forall i$$

Vapnik va definir la distancia dels punts més propers al hiperplà com a marge, per tant a la funció que s'ha expressat més amunt es pot comprovar com tots els punts estaran a una distància mínima del marge, aquest hiperplà es el que minimitzarà l'error generalització. De forma gràfica l'idea es la següent:

Figura 1 Marge SVM



Definint $\left|\left|\beta\right|\right| = \frac{1}{M}$, es pot formular de nou el problema de la següent manera:

$$\min_{\beta} \frac{1}{2}\left|\left|\beta\right|\right|^2$$
$$s.a\ y_i(x_i'\beta + \beta_0) \geq 1$$

Utilitzant el Lagrangià sobre aquest problema i donant-se les condicions de Karush-Karun-Tucker obtenim el següent resultat:

$$\min_{\beta,\beta_0} L = \frac{1}{2\beta'\beta} - \sum^{n} \lambda_i(y_i(x_i'\beta + \beta_0) - 1)\ \ \lambda_i \geq 0\ \forall i \rightarrow$$
$$\rightarrow \lambda_i(y_i(x_i'\beta + \beta_0) - 1) = 0\ \forall\ i$$

Aleshores podem obtenir els punts que pertanyen al marge:

$$\begin{cases} \lambda_i > 0 \rightarrow x_i, & pertany\ al\ marge. \\ \lambda_i = 0 \rightarrow x_i, & no\ pertany\ al\ marge \end{cases}$$

Per definir la funció de discriminació dels SVM, Vapnik, va utilitzar les derivades del Lagrangià sobre els paràmetres β.

$$\begin{cases} \frac{dL}{d\beta} \rightarrow \beta = \sum^{n} \lambda_i y_i x_i \\ \frac{dL}{d\beta_0} \rightarrow 0 = \sum^{n} \lambda_i y_i \\ \forall x\ \in R^p \end{cases}$$

La funció de discriminació es formula de la següent manera:

$$S_{SVM}(x) = sign(\hat{\beta}x + \hat{\beta}_0) = sign\,(\sum_{i \in SV} \lambda_i y_i x_i' x + \hat{\beta}_0)$$

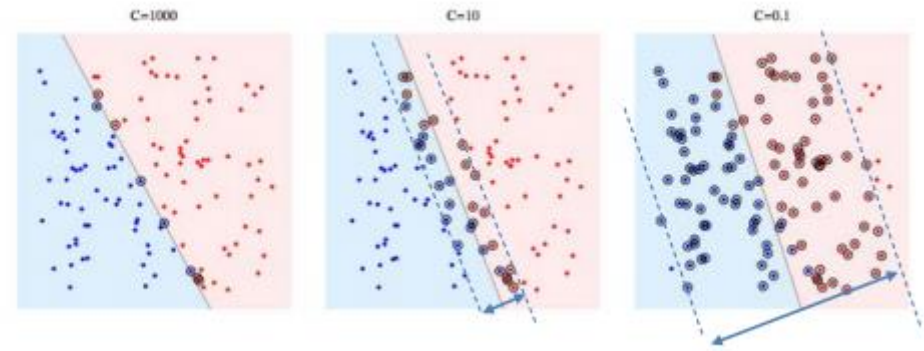
D'aquesta funció obtindrem el resultat de la classificació dels diferents punts, si son superiors a zero pertanyen a una classe, mentre que si en son superiors en pertanyen a l'altre. D'aquesta manera es defineixen els vectors de suports que son, aquells punts que estan en el hiperplà òptim de separació. Aquets punts són els més difícils de de classificar, però per una altra banda, també són aquells que aporten més informació.

El principal problema d'aquest algoritme es que el marge "dur" es sensible a outliers, això vol dir que no es un algoritme robust. A més, amb aquest algoritme si les classes es solapen el problema en qüestió no es separable, per tant, Vapnik i el seu equip, van decidir definir el marge "suau". Per aquests punts mal classificats es calcula la distància des de on està localitzat fins a on hauria d'estar localitzat, es a dir el marge, aquest càlcul es defineix amb la variable, ξ. I al problema de de minimització inicial se li afegeix un paràmetre de penalització per aquells valors mal classificats, C. Aleshores el problema queda definit de la següent manera:

$$\min \frac{1}{2}||\beta||^2 + C\sum^n \xi_i$$
$$s.a\ y_{i(x_i'\beta+\beta_0)} \geq 1 - \xi$$

El rol que executa la variable C es la de definir la tolerància respecte els punts mal col·locats, si li donem un valor petits existiran més valors incorrectes, mentre que si tenim una tolerància molt baixa, i per tant li donem un valor molt elevat estaríem en el cas del marge dur. Aquest problema es veu representat en la següent imatge:

Figura 2 Rol del paràmetre C



Un cop tenim definit aquest problema s'ha de definir la funció discriminant del marge "suau". Com s'ha fet anteriorment s'utilitzarà el lagrangià per poder definir la solució i poder obtenir els vectors suport i els punts que pertanyen al marge.

$$\min_{\beta.\beta_0} L = \frac{1}{2}\beta'\beta + C\sum_{}^{n} \xi_i - \sum_{}^{n} \lambda_i(y_i\,(x_i'\beta + \beta_0) - (1 - \xi_i)) - \sum_{}^{n} \mu_i\xi_i$$

Aleshores els punts de suport es defineixen com:

$$\begin{cases} \lambda_i = 0 \rightarrow y_i(x_i'\beta + \beta_0) > 1 \rightarrow no\ pertany\ al\ marge \\ 0 < \lambda_i < C \rightarrow y_i(x_i'\beta + \beta_0) = 1 \rightarrow\ pertany\ al\ marge\ i\ es\ SV \\ \lambda_i = C \rightarrow y_i(x_i'\beta + \beta_0) < 1 \rightarrow no\ pertany\ al\ marge\ i\ es\ SV \end{cases}$$

S'ha d'aclarir que aquells punts que $\lambda_i \neq 0$, només n'hi ha uns pocs que siguis punts de suport i es dona quan $y_i(x_i'\beta + \beta_0) = 1 - \xi_i$, per tant els únics que es romandran al marge son aquells que $\xi_i = 0$, mentre que si $\xi_i > 0$ el punt estarà mal col·locat.

La solució a la funció discriminant serà:

$$\begin{cases} \dfrac{dL}{d\beta} \rightarrow \beta - \displaystyle\sum^{n} \lambda_i y_i x_i = 0 \\ \dfrac{dL}{d\beta_o} \rightarrow \displaystyle\sum^{n} \lambda_i y_i = 0 \\ \dfrac{dL}{d\xi_i} \rightarrow C - \lambda_i - \mu_i = 0 \end{cases}$$

La funció de discriminació es formula de la següent manera:

$$S_{SVM}(x) = sign(\hat{\beta}x + \hat{\beta}_0) = sign(\sum^{n} \lambda_i y_i x_i' x + \hat{\beta}_0)$$

Com amb la funció discriminant anterior, la del marge dur, els valors que siguin superiors a zero seran classificats com una classe, mentre que els que el seu valor sigui inferior a zero seran classificats a l'altre.

2.2.2 Support Vector Regression

Per el nostre treball no es necessari classificar cap tipus de punts, ja que no es tracta de cap problema de classificació, sinó que es tracta d'un problema de regressió, i per tant hem de definir la metodologia apropiada per aquest tipus de qüestió. Però els SVM ens permeten definir un problema d'optimització per aquest tipus de exercici, els Support Vector Regression, SVR a partir d'aquest punt. Però l'idea d'aquest s'entendrà millor un cop s'han explicats els SVM.

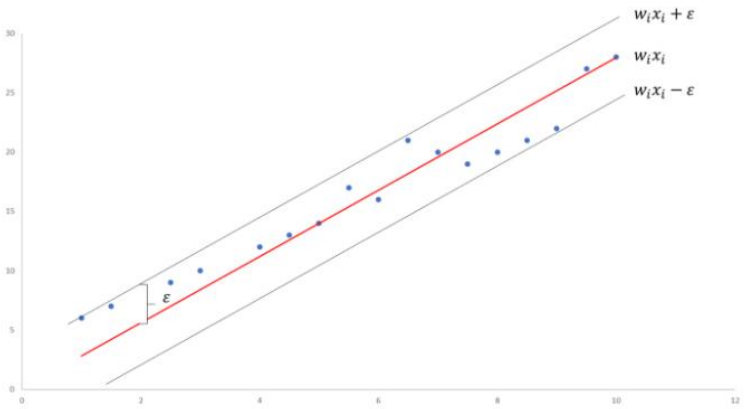
Amb la mateixa idea que els problemes de classificació, en els problemes de regressió ens permeten definir quina quantitat d'error es acceptable per modelar les nostres dades i trobar la línia de regressió apropiada, o en dimensions més elevades l'hiperplà.

Aquest models a diferencia dels mínims quadrats ordinaris, utilitzats en les regressions lineals simples, els models de SVR, el que busquen es minimitzar els coeficients i no pas els errors. L'error que estem disposats a assumir es gestiona en les constriccions del model, es pot modificar el terme d'error per obtenir la precisió desitjada. El model es representa de la següent manera:

$$\begin{cases} \min \frac{1}{2} ||w||^2 \\ s.a\ |y_i - w_i x_i| \leq \varepsilon \end{cases}$$

De forma gràfica es veuria així:

Figura 3 Marge SVR



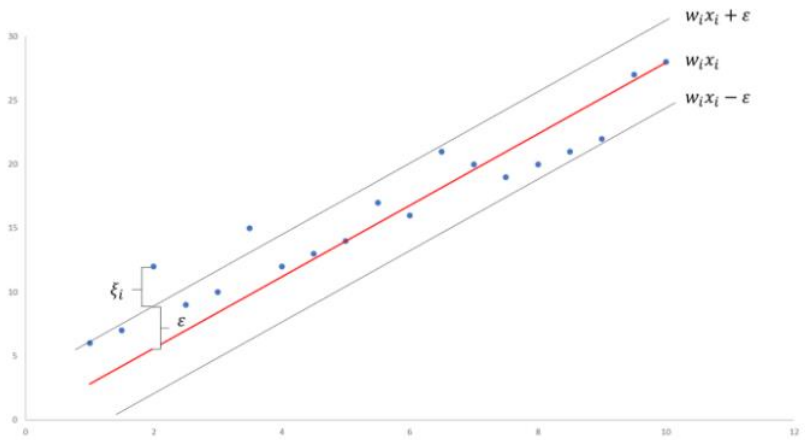
D'aquesta manera obtindrem una línia de regressió i uns marges en els que es veuran acotats els punts de la nostra base de dades, però que passa amb aquells punts que els errors son superiors a ε? Per aquests punts es pot definir una altre paràmetre que reculli aquells punts que es veuen a fora del marge, aquest paràmetre serà ξ. Estaríem definint el marge suau dels problemes de classificació.

Aleshores per afegir aquests valors a la funció objectiu li hem de donar una penalització, i serà, com amb els SVM, C, que definirà la tolerància d'assumir aquests errors de punts que es troben a fora del marge. El model es veurà modificat de la següent manera:

$$\begin{cases} \min \frac{1}{2}||w||^2 + C\sum^{n}|\xi_i| \\ s.a\ |y_i - w_i x_i| \leq \varepsilon + |\xi_i| \end{cases}$$

Per a que es vegi de manera més clara, es mostra de manera gràfica.

Figura 4 Rol paràmetre C



2.3 Criteris de Validació

Un cop haguem realitzats els models caldrà comprovar que aquests s'ajusten d'una manera correcta i per tant realitzen una bona predicció de les dades. A continuació s'explicaran els diferents criteris que s'utilitzen per validar aquest tipus de model.

Seguidament explicarem una sèrie de mètodes que s'utilitzen per avaluar la capacitat predictiva del model. Tots els mètodes tenen com a característica principal que mesuren l'error que ha realitzat el model a l'hora de predir els valors reals. Per tant, per utilitzar aquests criteris hem d'haver realitzar una predicció.

El primer es **l'error quadràtic mig (EQM)**: Aquest estimador mesura la mitjana dels errors realitzats en la predicció al quadrat. Per tant penalitza aquells períodes on els errors són molt elevats en comparació a altres. De manera algebraica la seva expressió es:

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

El segon criteri, bastant semblant al primer, és **l'error absolut mitjà (EAM)**. A diferència del primer aquest criteri fa la mitjana de la suma de la resta dels errors en valor absolut. De forma matemàtica s'expressa així:

$$EAM = \frac{1}{n}\sum^{n} |\widehat{y_i} - y_i|$$

Seguidament explicarem **l'arrel quadrada de l'error quadràtic mitjà (RMSE)**, les sigles venen donades pel seu nom anglès *root mean squared error*. Es pot interpretar com la desviació estàndard de la variància no explicada. El càlcul es el següent:

$$RMSE = \sqrt{\frac{1}{n}\sum^{n} (\widehat{y_i} - y_i)^2}$$

Com més proper sigui aquest valor a 0 millor es l'ajust del model. Però a la practica no s'arriba gairebé mai, per tant ens quedarem amb els valor més petits pòsibles.

Aquest tres criteris que hem explicat són dos mesures de precisió dependents d'escala, això vol dir que depenen de les unitats amb que calculem la variable objectiu, per tant es necessari establir un criteri que no es vegi influenciat per les magnituds que estan mesurades les dades. En aquets context trobem dos criteris més, un que utilitza els errors de predicció, com els que hem explicat fins ara, i un altre que no els utilitza.

Primer explicarem l'**error percentual absolut mitjà (EPAM).** D'una manera breu, aquest criteri mesura la dimensió del error en percentatge. Es calcula de la següent forma:

$$EPAM = \frac{1}{n}\sum^{n} \left|\frac{\widehat{y_i} - y_i}{y_i}\right| * 100$$

Aquest valor té la següent interpretació:

$$\begin{cases} EPAM < 1\%: Molt\ bona\ predicció \\ EPAM \in (1\%, 3\%]: Capacitat\ predictiva\ bona \\ EPAM \in (3\%, 5\%]: Capacitat\ predictiva\ mediocre \\ EPAM > 5\%: Capacitat\ predictiva\ pobre \end{cases}$$

3. Anàlisi Previ

En el següent apartat es farà un breu anàlisi descriptiu per tal de conèixer la base de dades que s'utilitzarà en aquesta tasca. També es definiran uns conceptes abans de seguir amb l'estudi per tal d'evitar futures confusions.

3.1 Dades

Aquest treball es realitzarà fent ús d'una combinació de bases de dades, molt completes, que realitza el ministeri d'agricultura pesca i alimentació del govern espanyol[1], sobre el consum d'aliments de les llars espanyoles. Aquestes bases de dades es realitzen des de l'any 1999, però s'ha decidit utilitzar només les dades recollides a partir de l'any 2014, ja que des de llavors fins l'actualitat s'ha utilitzat el mateix cens i s'han recollit les mateixes variables. Per a la recollida de dades s'utilitza una mostra de 12500 llars que registren les compres diàriament amb un lector òptic, utilitzen més de 2000 punts de sondeig per recollir les mostres. L'univers sobre el qual es realitza l'estudi són les llars del territori peninsular, i tant les illes balears com les illes canàries, s'exclouen les ciutats autònomes de Ceuta i Melilla.

Als informes es defineix el concepte de **llar** com aquella persona o conjunt de persones que ocupen en comú un habitatge familiar o part d'aquesta i consumeixen aliments i altres bens sota un mateix pressupost.

Per realitzar aquest estudi s'utilitzaran les variables de **consum per càpita**, en kilograms o litres(Kg/L) , i la **despesa per càpita**, mesurada en euros (€), i a **preus corrents**, com a variables resposta. Utilitzarem les variables totals per realitzar aquest treball.

En l'àmbit territorial, no es tindrà en compte la segmentació geogràfica per comunitats autònomes que es realitza el ministeri, i es farà un estudi sobre el total nacional.

Abans de començar amb l'anàlisi de les variables escollides, cal fer una definició d'aquestes, ja que poden semblar conceptes similars però no es així.

El primer concepte que definirem es **consum per càpita,** aquesta idea fa referencia a la relació entre el consum total d'una població i la seva població. Aleshores cal definir que es el consum, traduint de la *Real Academia Española* es defineix com acció i efecte de consumir i consumir fa referència a l'acció d'utilitzar o gastar un producte o servei per satisfer necessitat.

El segon concepte despesa per càpita fa referencia a la relació entre la despesa total d'una població, respecte aquesta població. Fent el mateix procés que el consum per càpita, arribem a la definició de gastar que diu de forma traduïda però literal, utilitzar els diners en quelcom.

[1] Ministerio de Agricultura, pesca y alimentación. *Series anuales del consumo de hogares*. https://www.mapa.gob.es/es/alimentacion/temas/consumo-tendencias/panel-de-consumo-alimentario/series-anuales/default.aspx

Per tant tot i que en certs camps, com l'elèctric, siguin concepte sinònims, em pogut comprovar que son diferents. Quan parlem, en aquest treball de consum parlarem de magnituds físiques com els kilograms o els litres, mentre que si parlem de despesa parlarem en termes monetaris. Tot i que ambdues variables estan molt relacionades entre si.

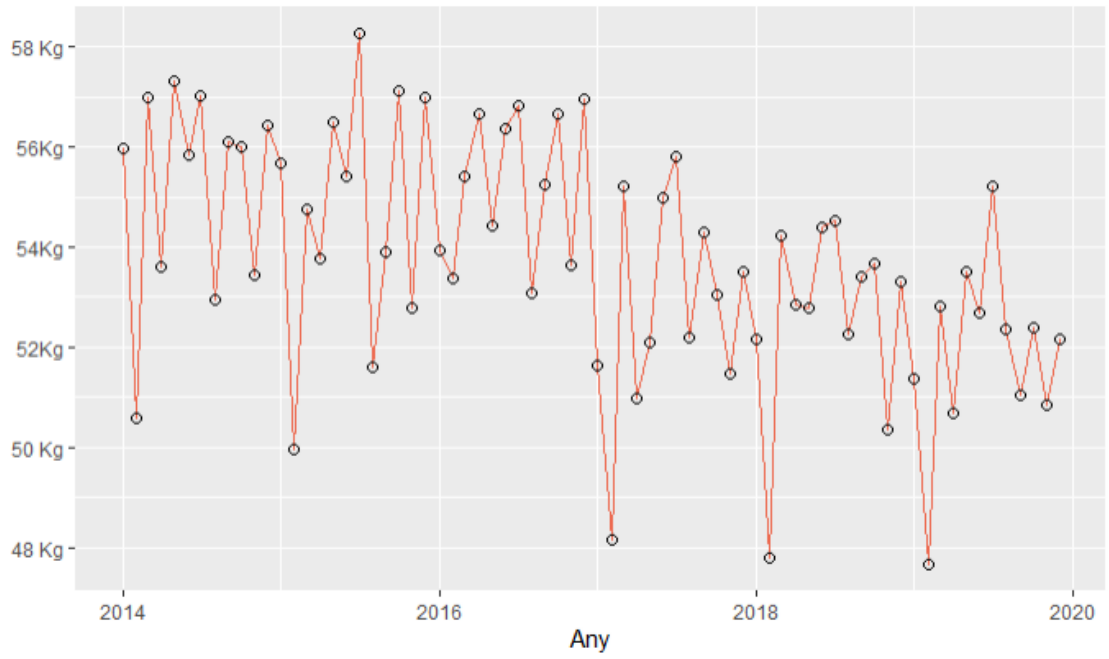
3.2 Anàlisi univariant

Com s'ha mencionat amb anterioritat, a continuació es realitzarà un breu anàlisi descriptiu de les dues variables d'estudi del treball per tal de conèixer possibles valors anòmals i observar el comportament de les dades.

Aquesta exploració de les dades es realitzarà en dos apartats diferenciats al tractar-se de variables diferents entre elles.

3.2.1 Anàlisi del consum per càpita

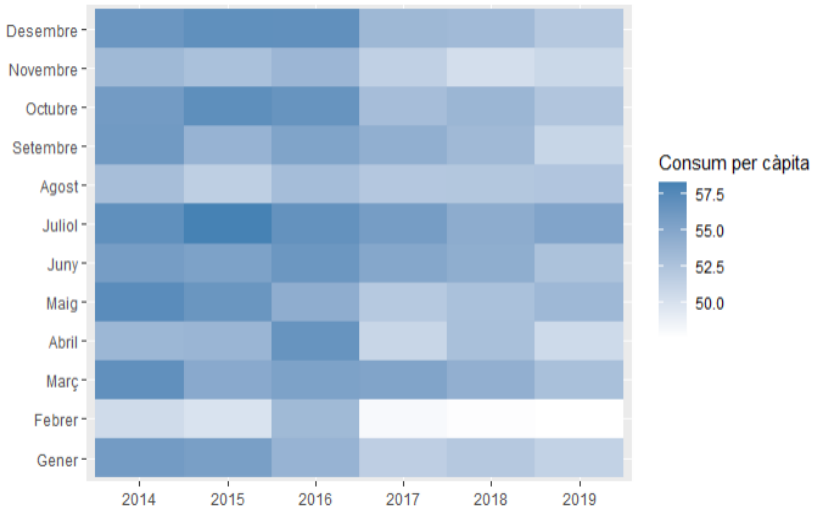
Figura 5 Consum per càpita alimentari d'Espanya (2014-2019)



Com es pot observar al gràfic, el consum per càpita de les llars espanyoles ha anat disminuint amb els anys, es pot comprovar com a l'any 2015, concretament al juliol, obtenim el valor més elevat a la sèrie, i pel contrari el valor mínim el trobem al 2019, específicament al febrer. Aquest comportament decreixent pot donar-se per dos motius, el primer una augment poblacional major que la quantitat d'aliments consumits, o bé simplement hi ha hagut un decrement en la quantitat de menjar que la població espanyola consumeix.

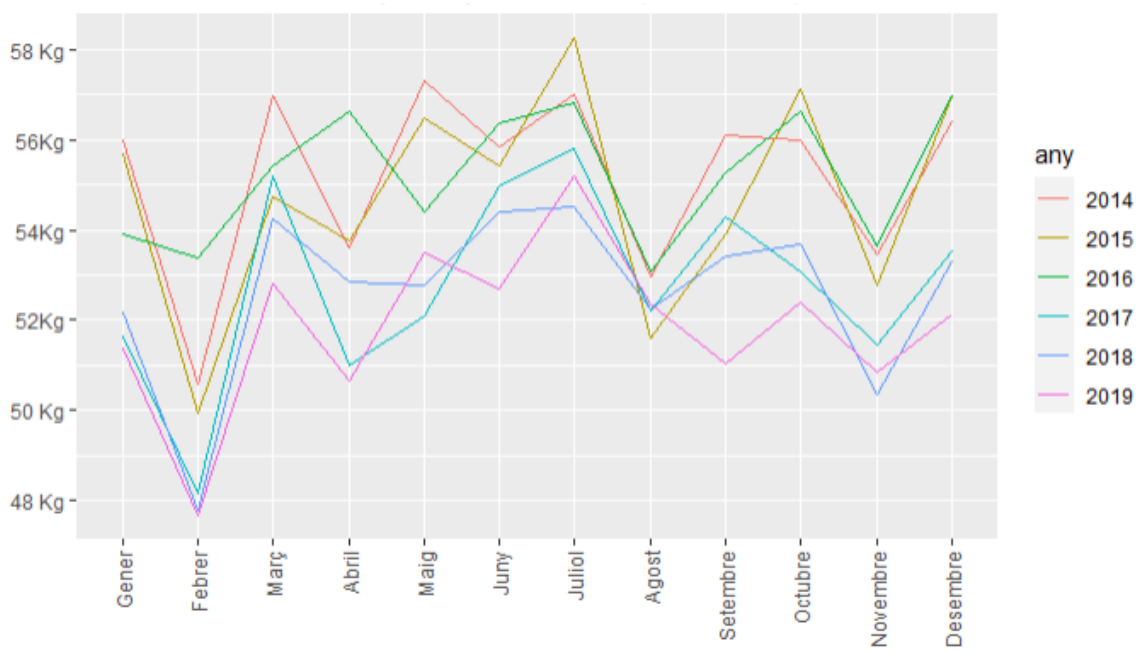
També es pot observar un comportament periòdic molt clarament. Si ens hi fixem bé, al mes de febrer ens hi trobem una baixada bastant accentuada del consum per càpita. Aquest comportament es veu explicat al mapa de calor que es mostra a continuació. On els mesos pintats de forma més clara es el mes de febrer.

Figura 6 Heatmap del Consum alimentari per càpita



Aquest comportament s'observa amb més claredat al gràfic que es mostra a continuació. On s'observa com de forma general, el més de febrer hi ha una baixada molt pronunciada en tots els anys, a excepció del 2016. Degut a que les festivitats de Nadal es troben als mesos de desembre i gener, el mes de febrer pot veure's afectat i que es consumeixi menys.

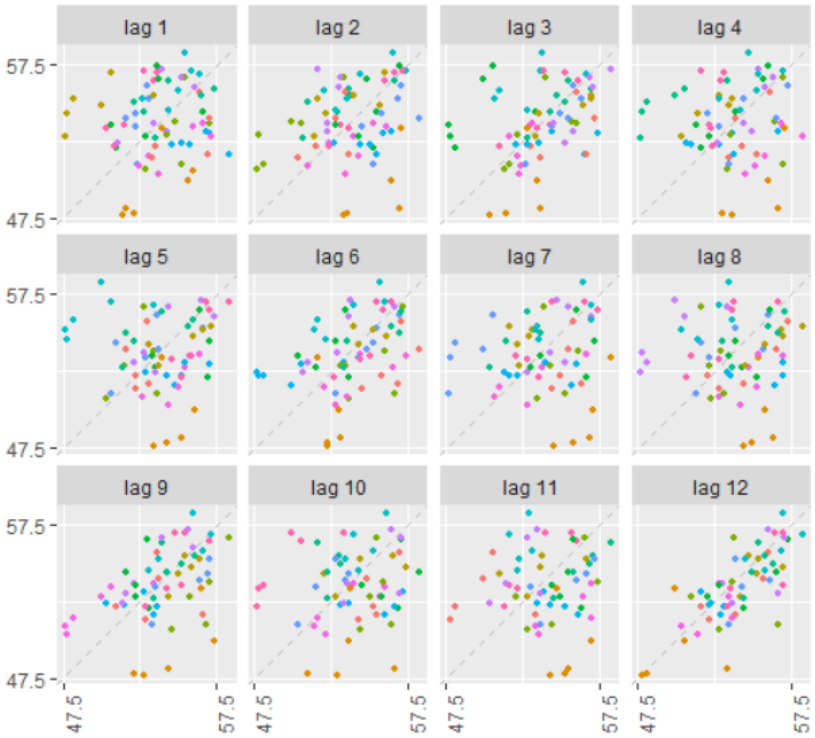
Figura 7 Consum alimentari per càpita menusal, desagregat anualment



Aquest gràfic ens mostra la correlació de la sèrie temporal amb els seus retards, s'utilitza per identificar la correlació entre aquestes. Podem observar com des de el primer retard fins l'onzè no s'observa cap tipus de relació, si més no, son núvols de punts sense cap tipus de forma, però en el dotzè retard veiem com si existeix una relació lineal. Això

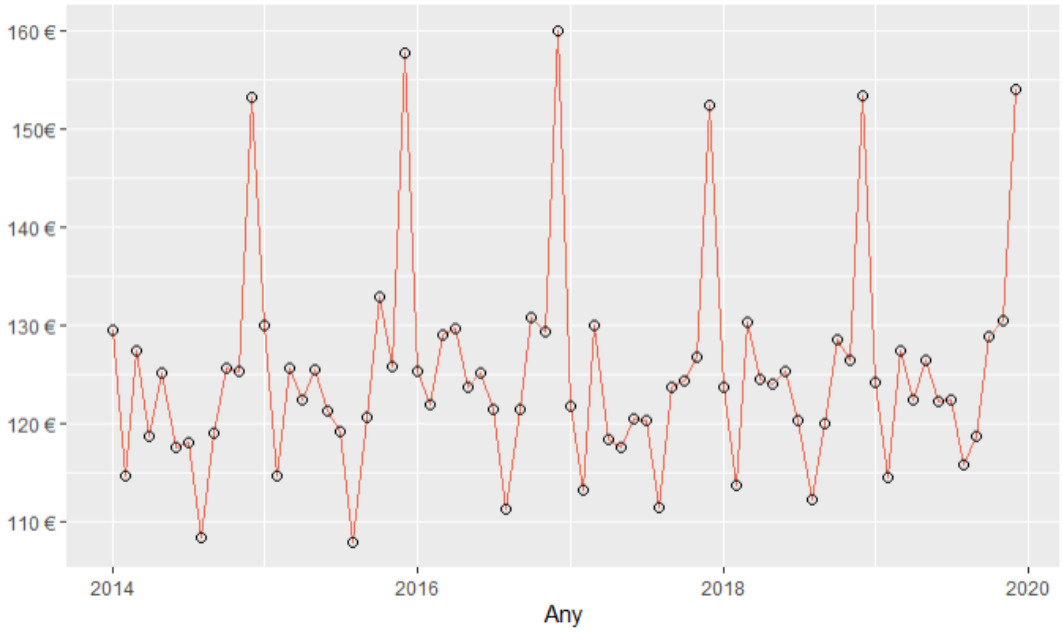
indica que existeix una forta relació entre el mateix mes al llarg dels anys i per tant tenen un comportament similar durant aquests.

Figura 8 Sèrie (y) vs Retards (x)



3.2.2 Anàlisi de la despesa per càpita

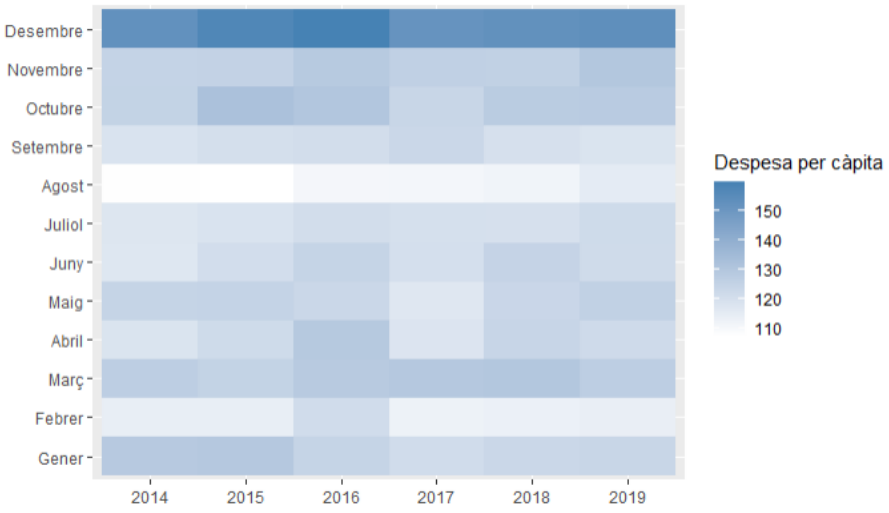
Figura 9 Despesa per càpita en aliments a Espanya (2014-2019)



Es pot observar com el gràfic de la despesa no presenta cap tendència global, ni decreixent ni creixent, el que si podem observar es com els mesos de desembre de cada

any tenen un valor molt superior a tots els demes. Això pot ser degut a que la majoria de festivitats es concentren en aquests mesos. A més, podem comprovar com els mesos d'agost de forma general tenen una despesa menor que tots els altres.

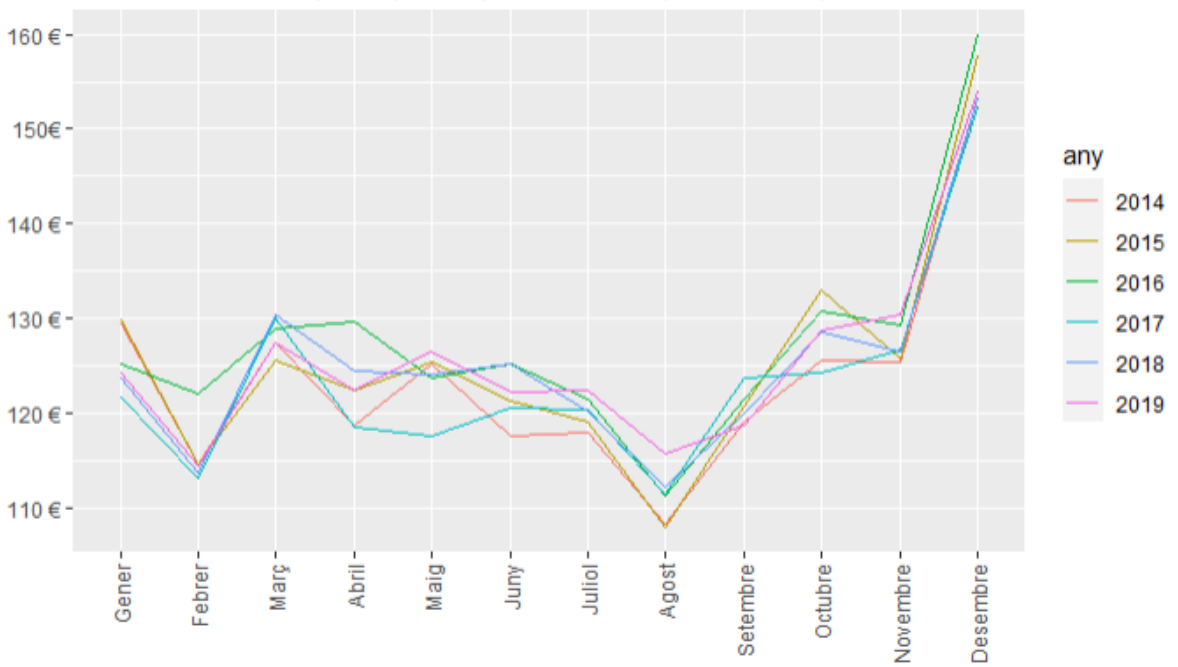
Figura 10 Heatmap Despesa per càpita alimentaria



El comportament de la variable queda recollit al *heatmap* que es mostra a continuació. Es un signe d'estacionalitat de la despesa per càpita. Podem observar un color molt més fosc al mes de desembre, i per contra un color gairebé blanc als mesos d'agost.

També observem aquest comportament de froma generalitzada al gràfic que mostrem a continuació. On veiem com la despesa es molt superior l'últim mes de l'any, a més podem comporovar com, a excepció de l'any 2016, el comportament de la variable es molt semblant any rere any .

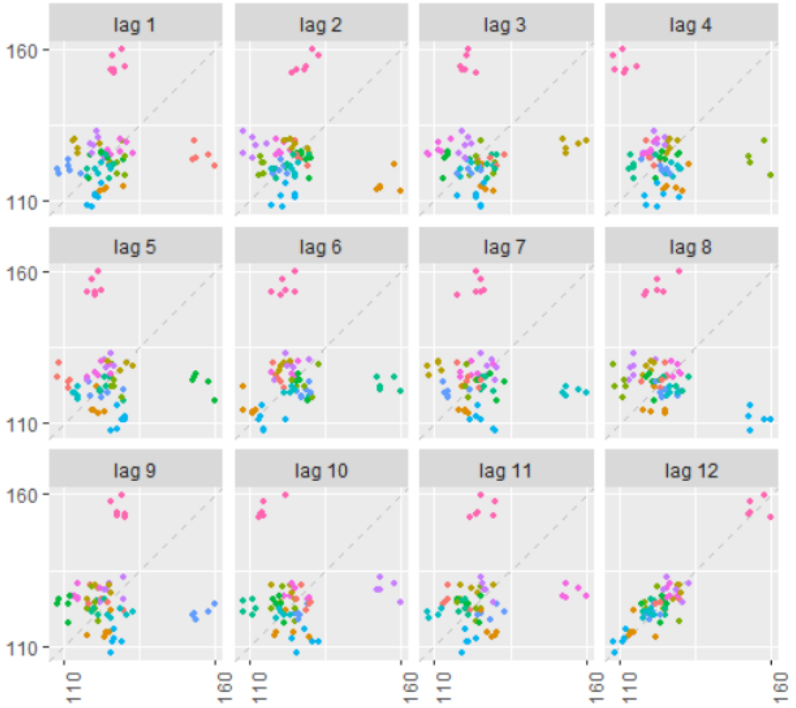
Figura 11 Despesa per càpita en aliments a Espanya, desagregat anualment



Com s'ha dit amb anterioritat aquest gràfic ens mostra la correlació de la sèrie temporal amb els seus retards. Es pot observar un comportament molt similar a la variable del consum per càpita, ja que s'observa una relació lineal al dotzè retard que no s'observa

en cap altre mes. Això ens indica una forta correlació forta entre els mateixos mesos de l'any

Figura 12 Sèrie (y) vs Retards (x



4. Anàlisi ARIMA

Seguidament aplicarem la metodologia ARIMA, explicada amb anterioritat al *Marc Teòric* d'aquest treball. Farem ús de les cinc fases establertes per Box i Jenkins que es coneix com metodologia Box i Jenkins.

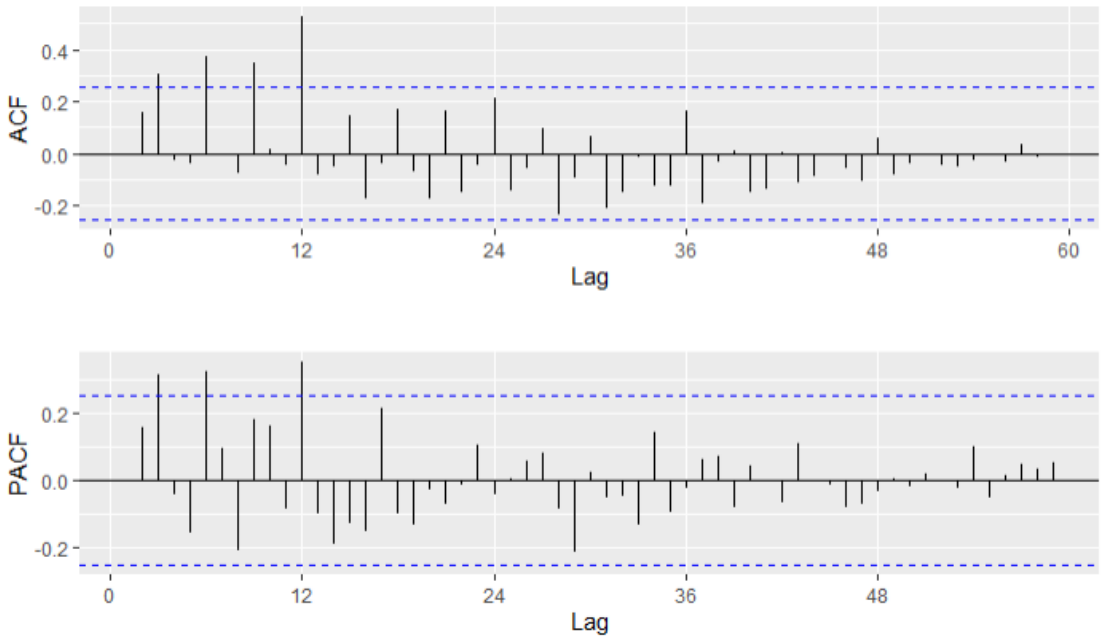
Com en aquest treball s'utilitzen dues variables respostes es dividirà en dos subapartats i es tractaran de manera diferenciada. Iniciarem aquest anàlisi amb el consum per càpita. En tots dos casos utilitzarem les dades dels anys 2014 fins l'any 2018, amb l'objectiu de predir l'any 2019 i observar els errors realitzats pel model predit. Les primeres dades, les compreses entre el 2014 i 2018 seran les dades de test, mentre que les dades de l'any 2019 seran de validació.

Per realitzar aquest anàlisi utilitzarem el software estadístic *R*, i de la llibreria *forecast* que s'utilitza per l'anàlisi de series temporals.

4.1 Anàlisi ARIMA: Consum per càpita

El primer pas per tal de realitzar aquest anàlisi es la de identificar el model. Per això es necessari mostrar el comportament de les dades, mostrat amb antelació a l'anàlisi univariant de les dades, (*Figura 1).* Com s'ha pogut observar en l'anàlisi univariant s'observa un comportament regular i estacional. Per poder contrastar aquesta informació cal realitzar una representació gràfica de la ***FAS*** i la ***FAP*.**

Figura 13 Correlograma del Consum per càpita alimentari d'Espanya.



Es pot observar com la component regular de la sèrie temporal té les característiques pròpies de un procés estacionari, ja que les correlacions disminueixen ràpidament cap a zero. Aquest comportament també s'observa a la part estacional. No cal aplicar cap diferenciació a cap de les dues components del procés.

Per donar validesa a que la sèrie es estacionaria es realitzarà un test augmentat de Dickey-Fuller, per tal de donar més robustesa a aquests resultats. Fent ús de la funció *adf.test.* Aquest test té les següents hipòtesis :

$$H_0: La\ sèrie\ no\ es\ estacionaria. \equiv Té\ arrel\ unitaria$$
$$H_1: La\ sèrie\ es\ estacionaria \equiv No\ té\ arrel\ unitaria$$

Figura 14 Test de Dickey Fuller augmentat

```
        Augmented Dickey-Fuller Test

data:  cons
Dickey-Fuller = -3.9911, Lag order = 3, p-value = 0.01609
alternative hypothesis: stationary
```

Com es pot observar en els resultats obtinguts el p-valor es inferior al 5% de significació per tant rebutgem la hipòtesis nul·la, per això tenim indicis per pensar que la sèrie es estacionaria.

Seguidament es realitzarà la primera estimació del model, com ja es va explicar en el *Marc teòric*, el procés amb el que estem treballant sembla ser que es estacional, per tant es tracta d'un model $SARIMA(p,d,q)(P,D,Q)$, i haurem de decidir quin paràmetres té.

El primer paràmetre que li podem assignar es l'odre d'integració, tant per la part estacional, com per la part regular, ja que anteriorment hem vist que no cal diferenciar, per tant els paràmetres d i D, els hi assignem un valor de 0.

Per als demés paràmetres s'ha decidit, degut al comportament de la *FAS* i la *FAP*, que serà un model. $SARIMA(1,0,1)(1,0,1)_{12}$

Els resultats obtinguts amb la implementació d'aquest model son els següents.

Figura 15 Sortida model SARIMA (1,0,1)(1,0,1)[12]

```
Call:
arima(x = cons, order = c(1, 0, 1), seasonal = list(order = c(1, 0, 1)))

Coefficients:
         ar1      ma1    sar1     sma1  intercept
      0.9542  -0.7813  0.9531  -0.5877    53.9230
s.e.  0.0500   0.1062  0.0547   0.2214     2.5691

sigma^2 estimated as 2.045:  log likelihood = -114.27,  aic = 238.54
```

Per poder validar el model, abans de realitzar les prediccions d'aquest es necessari validar el model. Per tal de fer aquesta comprovació, es necessari, en primer lloc mirar la significació dels coeficients.

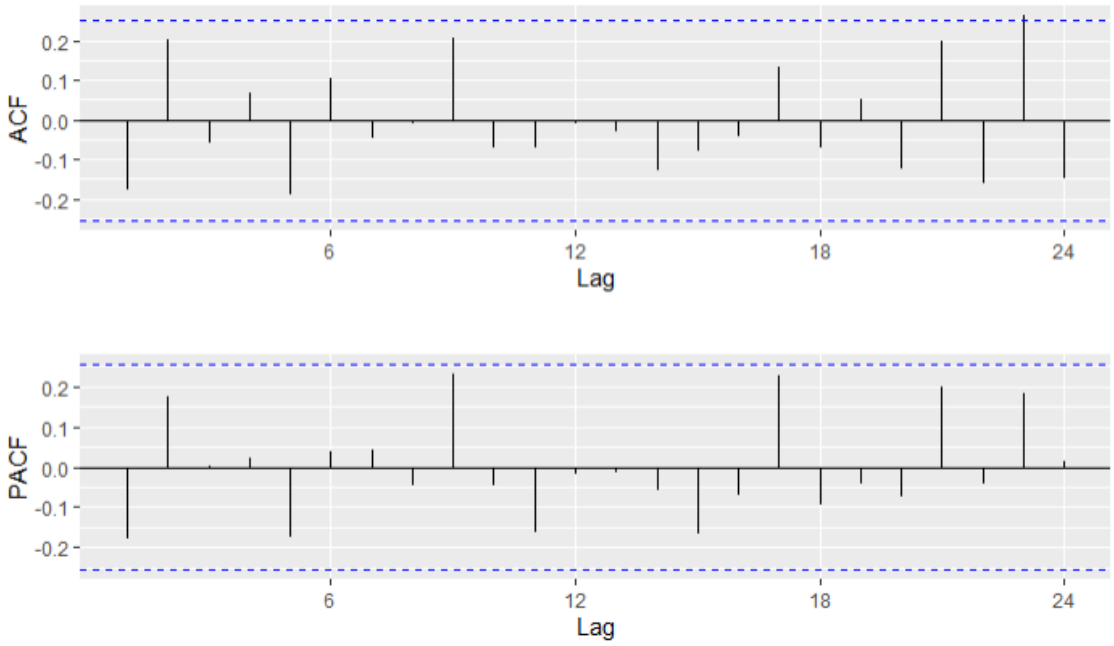
Taula 1 Test de Coeficients model SARIMA(1,0,1)(1,0,1)[12]

```
z test of coefficients:

           Estimate Std. Error z value  Pr(>|z|)
ar1        0.954197   0.050006 19.0815 < 2.2e-16 ***
ma1       -0.781304   0.106233 -7.3547 1.914e-13 ***
sar1       0.953119   0.054746 17.4098 < 2.2e-16 ***
sma1      -0.587710   0.221444 -2.6540  0.007955 **
intercept 53.923042   2.569089 20.9892 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Com podem observar tots els coeficients són significatius per un nivell del 5%, aquest es un cas bastant estrany a la pràctica ja que no acostumem a ser tots els coeficients significatius. Com es obvi el model es vàlid al tenir tots els coeficients significatius.

Podem observa amb els correlogrames dels residus que presenten un comportament de soroll blanc, per tant no hi ha cap correlació entre els residus. Fet que valida encara més el model.

Figura 16 Correlograma dels residus, model SARIMA(1,0,1)(1,0,1)[12]

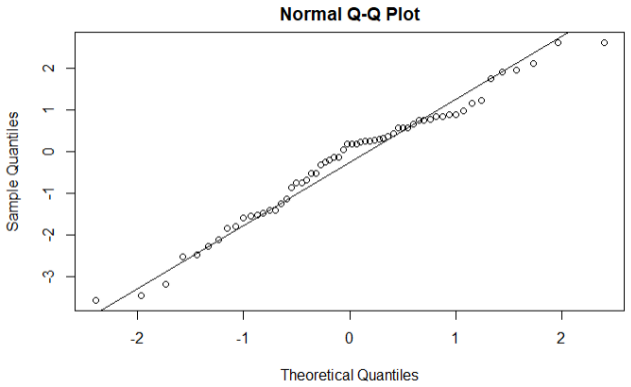


Finalment cal estudiar que els residus es distribueixin de forma normal. Per obtenir aquesta informació cal representar un Q-Q plot i realitzar el test de Shapiro-Wilk. La hipòtesi del test de normalitat que utilitzarem es la següent

$$H_0: X \sim N(\mu, \sigma^2) \rightarrow La\ distribució\ es\ normal$$
$$H_1: X \nsim N(\mu, \sigma^2) \rightarrow La\ distribució\ no\ es\ normal$$

Es pot observar com en el gràfic els residus es distribueixen de manera normal, la distribució es simètrica, tot i que cap al centre s'allunya de la línia recta. Però tot i no seguir la línia recta el test de Shapiro-Wilk, ens mostra com no rebutgem la hipòtesi nul·la i per tant podem concloure que els residus es distribueixen de forma normal.

Figura 17 Estudi de la normalitat dels residus del model




```
        Shapiro-Wilk normality test

data:  arima1$residuals
W = 0.9738, p-value = 0.223
```

En definitiva el model sembla vàlid en tots els aspectes. A continuació es farà la predicció de l'any 2019, per poder observar si les prediccions del model s'ajusten correctament o si pel contrari el model tot i ser vàlid proporciona unes males prediccions.

Els resultats de les prediccions realitzades son els següents:

Taula 2 Valors predits vs reals

PREDICCIÓ	VALOR REAL	ERROR
5,179083	5,135814	0,043269
4,855074	4,766034	0,08904
5,365328	5,282891	0,082437
5,212235	5,066566	0,145669
5,257076	5,350009	-0,092933
5,392657	5,270051	0,122606
5,459245	5,519906	-0,060661
5,166132	5,237182	-0,07105
5,329849	5,102492	0,227357
5,36241	5,240342	0,122068

Es pot observar com les prediccions en general semblen ser molt bones i els errors son bastant petits.

Per comprovar la magnitud d'aquests errors es farà a partir dels criteris que es van explicar al *Marc Teòric*.

Taula 3 Criteris de Validació ARIMA

CRITERI	VALOR
Error Quadràtic Mig	1,355778
Error Absolut Mitjà	1,037225
Arrel quadrada del error quadràtic mitjà	1,164379
Error percentual mitjà	2,005198 %

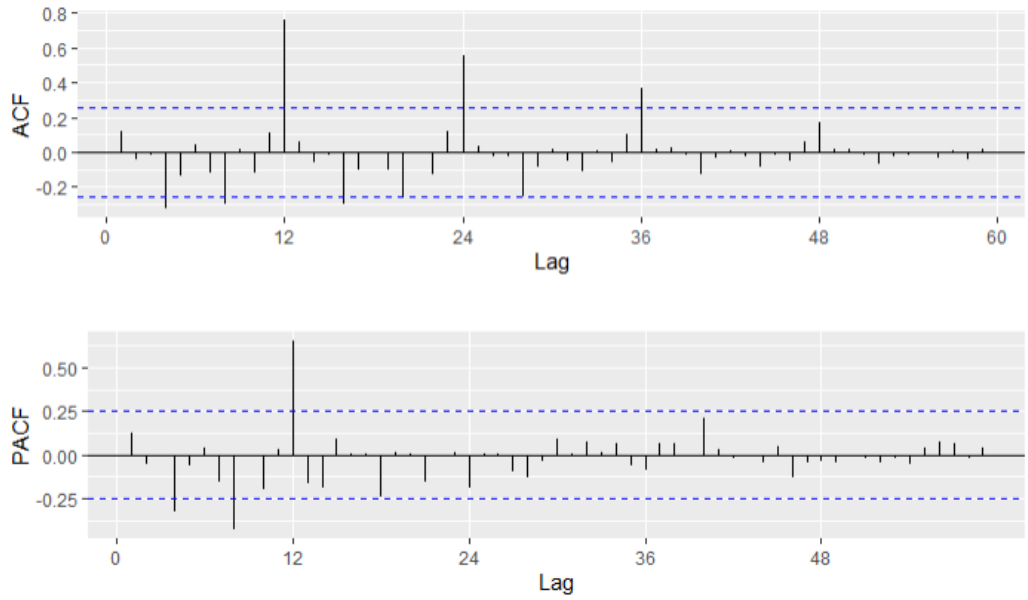
En general podem observar com les prediccions son bones amb errors molt petits i propers a 0, per tant podem dir que el model realitza una bona predicció.

4.2 Anàlisi ARIMA: Despesa per càpita

Per l'anàlisi de la despesa utilitzarem les mateixes eines que s'han utilitzat anteriorment amb el consum.

Primerament cal identificar el model del qual disposem. Podem observar un comportament estacional, com ja s'havia constatat en l'anàlisi univariant. També es pot comprovar com amb els correlogrames ens presenten aquest comportament estacional.

Figura 18 Correlogrames Despesa per càpita



Cal també comprovar que aquesta sèrie sigui estacionaria, pel comportament del correlogrma tenim indicis per pensar que efectivament es estacionaria, ja que els coeficients disminueixen ràpidament i es torben per sota del llindar de significació. Ja

sigui a la part estacional com a la regular. Però es realitzarà un test augmentat de Dickey-Fulley per tal de donar més solidesa.

Figura 19 Test de Dickey-Fuller augmentat de la Despesa per càpita

```
        Augmented Dickey-Fuller Test

data:  desp
Dickey-Fuller = -4.9183, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary
```

Com s'observa en el test el p-valor es inferior al nivell de significació del 5%, per tant rebutgem la hipòtesi nul·la, es per aquest motiu que hi ha indicis per pensar que la sèrie es estacionaria.

A continuació es farà l'ajust del model $SARIMA(p,d,q)(P,D,Q)$. Pel que fa l'ordre d'integració es pot observar com es zero, ja que la sèrie es estacionària i per tant no cal diferenciar per tal de fer la sèrie estacionaria. Observant el comportament del correlograma podem decidir que els paràmetres del model seran: $SARIMA(1,0,1)(1,0,1)_{12}$.

Els resultats obtinguts d'aquest model son els següents:

Figura 20 Sortida model SARIMA (1,0,1)(1,0,1)[12]

```
Call:
arima(x = desp, order = c(1, 0, 1), seasonal = list(order = c(1, 0, 1)))

Coefficients:
         ar1      ma1    sar1     sma1  intercept
      0.7126  -0.5691  0.9942  -0.6561   124.8262
s.e.  0.2624   0.3020  0.0101   0.2699     4.5975

sigma^2 estimated as 10.9:  log likelihood = -174.27,  aic = 358.54
```

Cal, però validar el model, per això es necessari veure quins coeficients del model estimat son significatius.

Taula 4 Test de significació dels coeficients del model SARIMA(1,0,1)(1,0,1)[12]

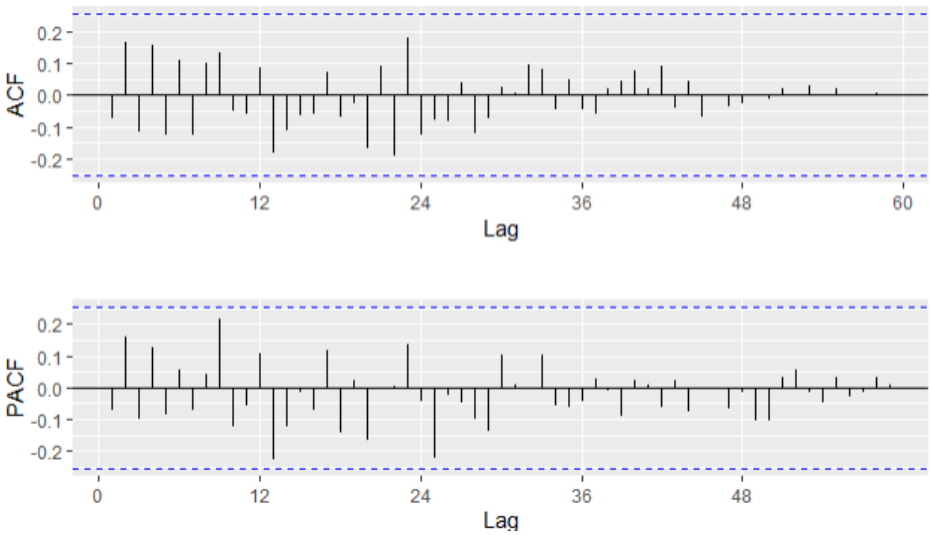
```
z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)
ar1          0.712559   0.262399  2.7156  0.006617 **
ma1         -0.569069   0.301961 -1.8846  0.059487 .
sar1         0.994249   0.010092 98.5227 < 2.2e-16 ***
sma1        -0.656099   0.269851 -2.4313  0.015043 *
intercept 124.826230   4.597517 27.1508 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podem comprovar com tots els coeficients son significatius per un nivell de significació del 5%, a excepció del coeficient de mitjana mòbils de la part regular, que no es

significatiu a un nivell del 5% per poquet. Tot i que els coeficients del model no siguin en la seva totalitat significatius, de moment podem donar el model per vàlid.

Seguirem amb la validació, per poder validar el model cal que els residus siguin estacionaris i no mostrin correlació. Per observar aquest comportament es farà ús dels correlogrames d'aquests.

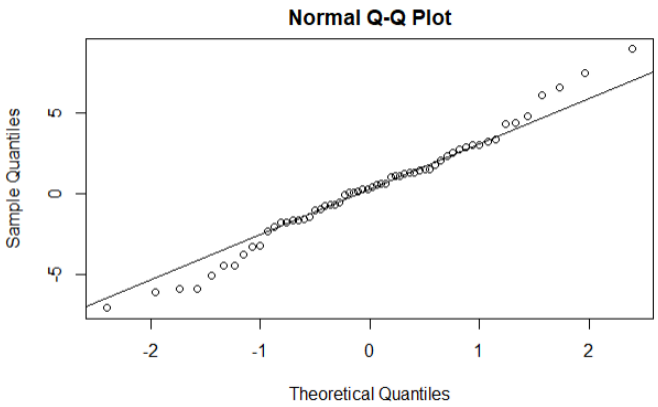
Figura 21 Correlograma dels residus



Com s'observa als correlogrames no hi ha cap de significatiu, per tant es tracta d'un procés de soroll blanc. Això es un bon indici per tal de poder acabar de validar el model.

A continuació es realitzarà el test de normalitat dels residus utilitzant el Q-Q plot i el test de Shapiro-Wilk. Com es pot observar en els gràfics i el test que s'adjunten a continuació els residus presenten un comportament normal, al no rebutjar el test de Shapiro-Wilk i presentar un comportament simètric en el Q-Q plot.

Figura 22 Estudi de la normalitat dels residus




```
        Shapiro-Wilk normality test

data:  arima2$residuals
W = 0.98477, p-value = 0.658
```

Per tant un cop hem validat el model queda realitzar la predicció de l'any 2019, per observar com s'ajusta el model. Els resultats obtinguts es mostren a continuació:

Taula 5 Valors predits vs reals

PREDICCIÓ	VALOR REAL	ERROR
124,6261	124,1854	0,4407
115,1867	114,4591	0,7276
128,9003	127,3227	1,5776
122,9442	122,4124	0,5318
122,6627	126,3848	-3,7221
122,7434	122,1981	0,5453
120,1155	122,3382	-2,2227
111,1469	115,7633	-4,6164
121,1563	118,7122	2,4441
128,0257	128,7444	-0,7187

Per comprovar la magnitud d'aquests errors es farà a partir dels criteris que es van explicar al *Marc Teòric*.

Taula 6 Criteris de Validació ARIMA

CRITERI	VALOR
Error Quadràtic Mig	5,309812
Error Absolut Mitjà	1,777477
Arrel quadrada del error quadràtic mitjà	2,304303
Error percentual mitjà	1,446909 %

De forma general la predicció que realitza el model estimat es bona, amb valors propers a zero. Tot i que l'error quadràtic mig es superior a cinc, es deu a que està afectat per les magnituds de la variable resposta.

5. Anàlisi SVR

Seguidament en aquesta part del treball es realitzarà l'anàlisi del consum i de la despesa per càpita amb l'algoritme d'aprenentatge supervisat dels SVM. Per fer aquesta regressió es farà ús del software estadístic *Rstudio,* aprofitant que existeixen llibreries ja implementades per tal de fer aquest tipus d'anàlisi.

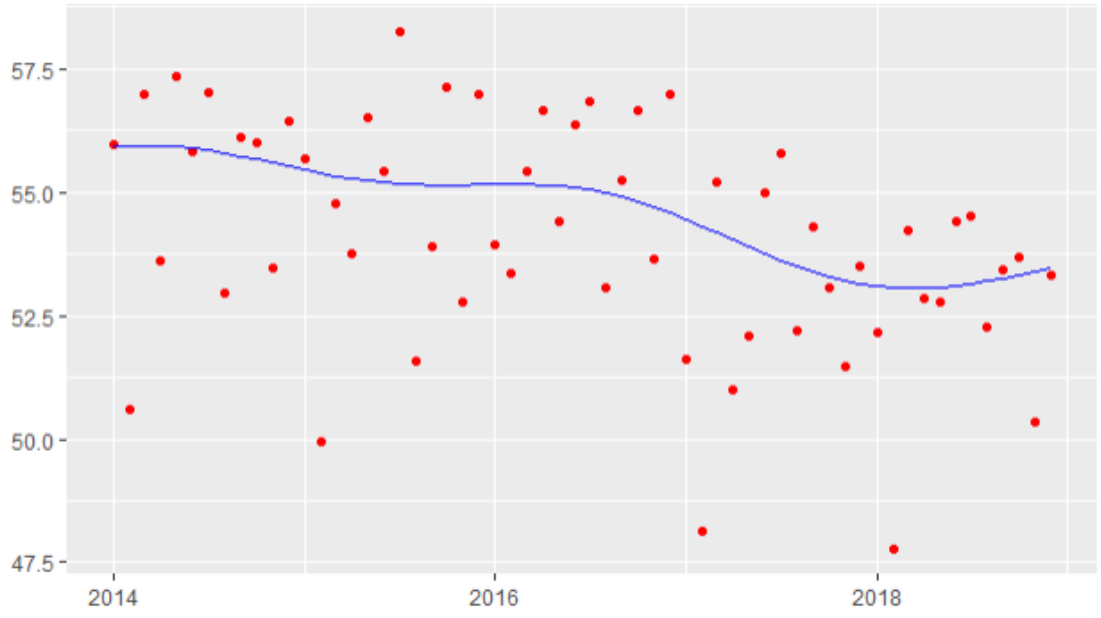
Com s'ha fet amb anterioritat es dividirà en dues parts, primer es farà l'anàlisi del consum per càpita i més endavant es durà a terme la regressió de la despesa. A més, s'utilitzarà el mateix criteri que s'ha seguit en els models ARIMA, i es realitzarà un model amb les mateixes dades d'entrenament, per més endavant validar els resultats.

5.1 Anàlisi SVR: Consum per càpita

Per començar aquesta el primer pas es definir el model que seguirem i per tal de comparar els resultats amb els models ARIMA es realitzarà un primera predicció fent una regressió del consum per càpita respecte el temps, no utilitzarem cap vector de β, de moemnt, tal i com es va fer en primera estància amb els models de Box-Jenkins.

El gràfic que es mostra a continuació representa en forma de punts el valor real del consum per càpita respecte el temps, i la línia mostra la regressió realitzada per el model.

Figura 23 Model SVR del Consum per càpita

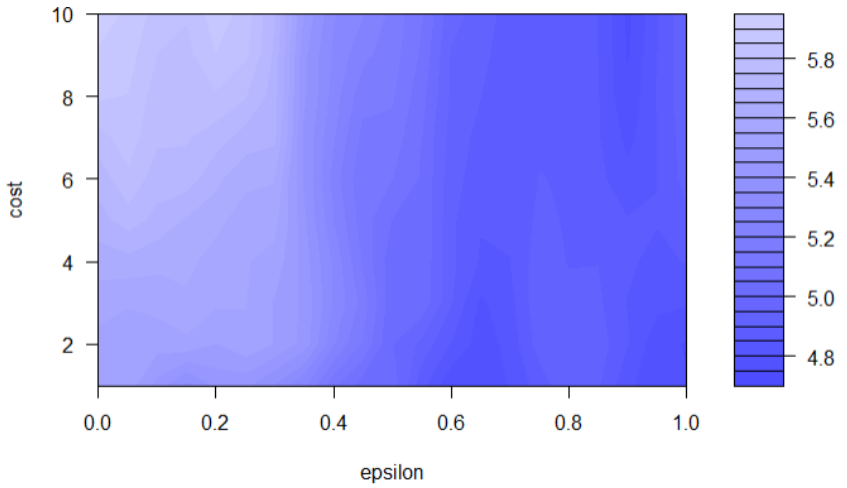


Es pot observar en primer lloc que la regressió no es bona, tot i que la tendència de les dades si aconsegueix calcular-la, es veu clarament com la estacionalitat de les dades no la detecta.

Això ens pot portar a una primera conclusió i es que els SVR no son tant potents per detectar el comportament estacional d'una variable econòmica com ho son els models ARIMA.

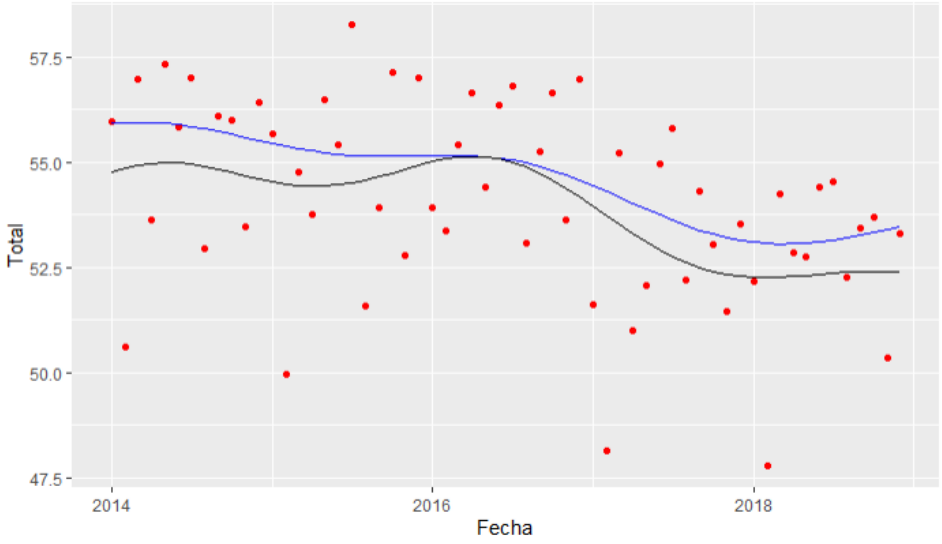
Però abans de treure conclusions precipitades modificarem els paràmetres del nostre model per tal de millorar l'ajust que aquest ens dona. Per realitzar aquesta tasca existeix la funció *tune* ja implementada a l'Rstudio. A continuació adjuntarem un mapa de calor que ens mostrarà les zones on millor prediu les dades el nostre model.

Figura 24 Millor model SVR



Aquest gràfic s'ha d'interpretar de la següent manera, la zona amb un color més fosc es allà on el model te un error de MSE més petit, i com podem observar aquest valor es troba amb un ε al voltant de 0,9 i un cost de aproximadament 10, el podem observar a la part més alta del gràfic. De manera que realitzarem la representació gràfica d'aquest model per observar si hem millorat els resultats del model anterior.

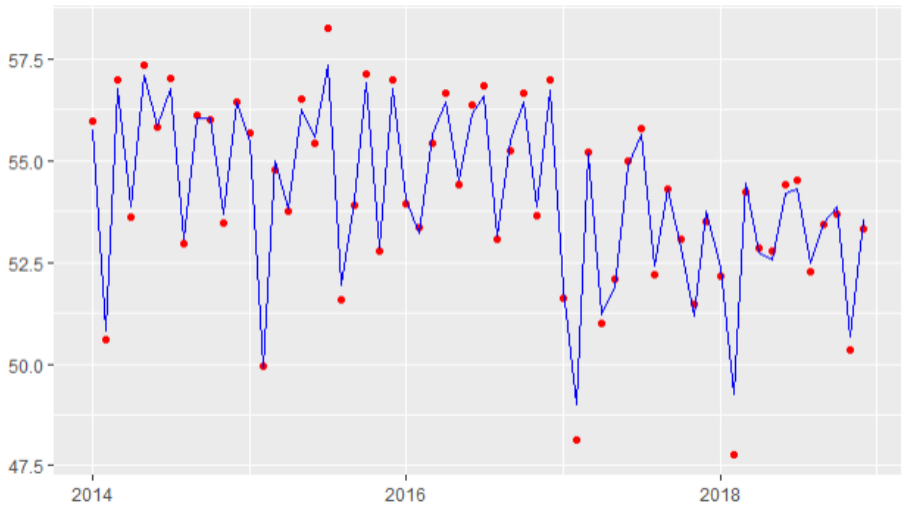
Figura 25 Millor model vs Primer model del Consum per càpita (SVR)



Com es pot observar el nou model millora l'ajust, ja que si detecta la estacionalitat de les dades però no ajusta bé els valors de les dades, esta molt allunyat de ser un bon ajust, es podria dir que li falta informació. Es per això que utilitzarem un vector de β per tal d'entrenar el model i poder afitar més la tendència del nostre model. Amb aquesta idea a continuació es tornarà a calcular un nou model, però, en aquest, utilitzarem un

seguit de variables que estan recollides en la mateixa base de dades, i que bàsicament es una desagregació del consum per càpita, per exemple el consum total per càpita d'aigua. Utilitzarem un conjunt de 23 variables que representen de forma global un 95% del consum per càpita total.

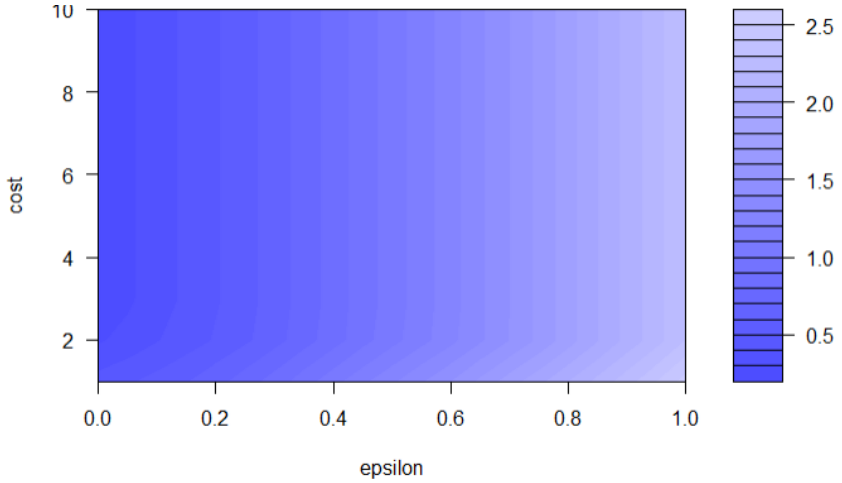
Figura 26 Nou model SVR, amb més información, del Consum per càpita



Com es pot observar en el nou model fent ús del vector de β, la regressió millora moltíssim i el model si es capaç de traçar una línia de tendència més acurada y també detecta els pics d'estacionalitat de les dades.

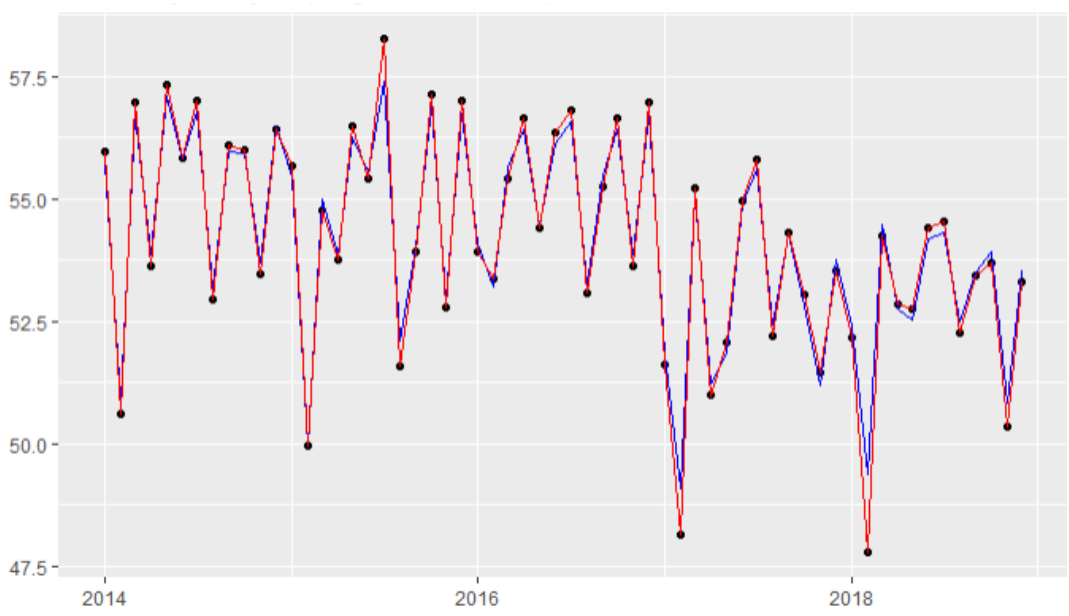
Ara bé encara es podria millorar aquest model, com s'ha fet amb anterioritat, en el primer model realitzat, ja que el que acabem de mostrar s'ha calculat amb els valors de ε i de cost predeterminades a la funció de l'*RStudio*. Es per això que tornarem a calcular el ε i el cost de la funció òptims per aquesta nova regressió. Utilitzarem el mateix mètode i es farà una representació en un mapa de calor on observarem quina es la zona que minimitza el MSE.

Figura 27 Millor model SVR



Com es pot observar la zona que minimitza l'error es troba en un valor de ε al voltant de 0 i en un rang de cost acotat entre el 2 i el 6. Per a casos com aquest en que no es pot distingir a primera vista i com a mètode de comprovació es pot demanar a la funció que ens retorni el millor model. En aquest cas es troba exactament en els valors de ε i de cost de 0 i 4, respectivament, es a dir estem fent coincidir el marge estricte amb el marge suau, al imposar un ε de 0. A continuació mostrem els resultats d'aquest nou model.

Figura 28 Millor nou model vs Nou model



Es pot observar una petita millora respecte el model anterior, sobretot respecte els punts mínims i màxims de les dades, com poden ser els mesos de febrer, ja que abans el model es quedava una mica curt. Aquesta millora es pot comprovar en la taula següent on es mostren els resultats obtinguts.

Taula 7 Criteris de Validació SVR

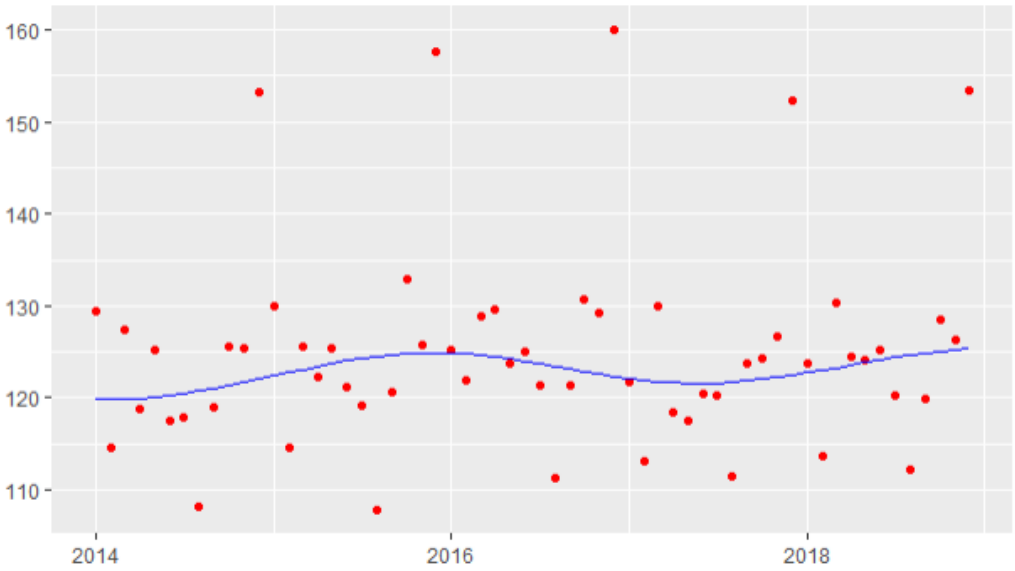
CRITERI	VALOR
Error Quadràtic Mig	0.2834
Error Absolut Mitjà	0.4425
Arrel quadrada del error quadràtic mitjà	0.5324
Error percentual mitjà	0.8656%

Com es pot observar els valors dels errors son molt petits, el model pràcticament prediu els valors exactes de la nostra mostra.

5.2 Anàlisi SVR: Despesa per càpita

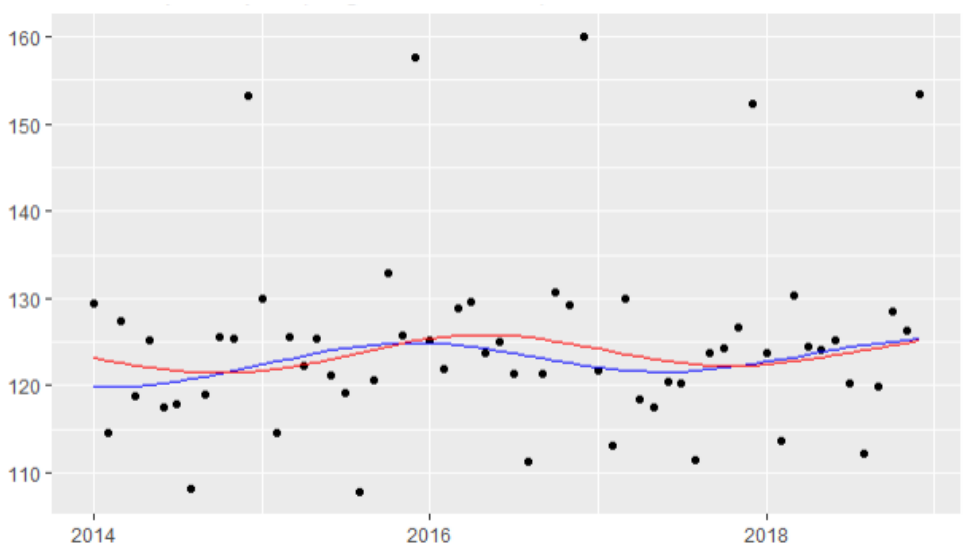
Seguidament s'analitzarà el comportament dels models de SVM amb la variable de la despesa per càpita, per a fer-ho es realitzarà la mateixa estratègia que amb el consum i per tant es començarà fent un una regressió simple de la despesa per càpita total respecte el temps, i a partir dels resultats obtinguts es marcarà una estratègia o una altra. Tot i que tractant-se de una variable estacional com el consum podem esperar un resultat similar.

Figura 29 Primer model SVR, Despesa per càpita



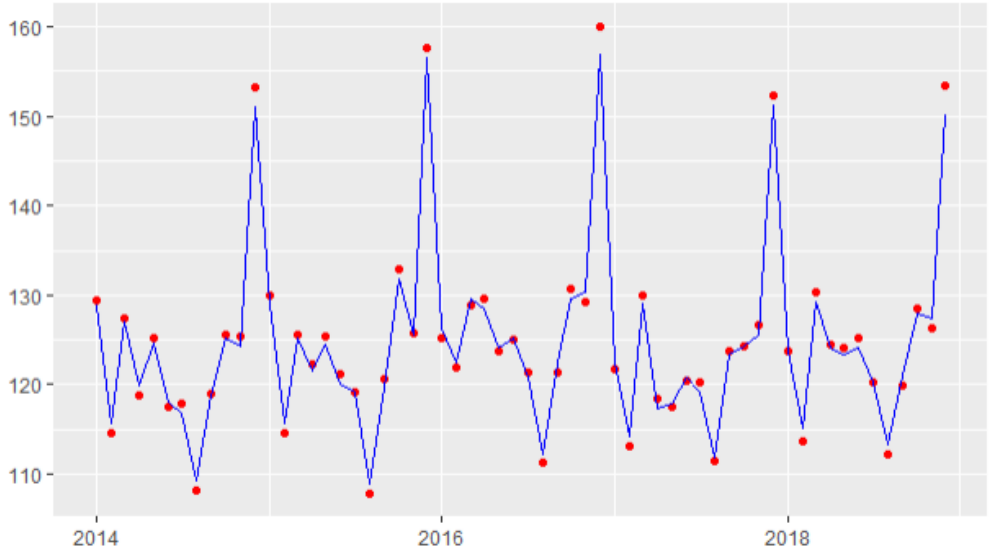
Com es pot observar en el gràfic el model no detecta l'estacionalitat de la variable, no obstant ja esperàvem aquest resultat. Seguint amb la mateixa estratègia calcularem els millors paràmetres del model i el tornarem a representar, no es mostrarà el procés ja que d'aquesta manera es faria massa repetitiu. El valor d'ε es de 0.4 i el cost de 1,

Figura 30 Millor model vs Primer moder SVR, Despesa per càpita



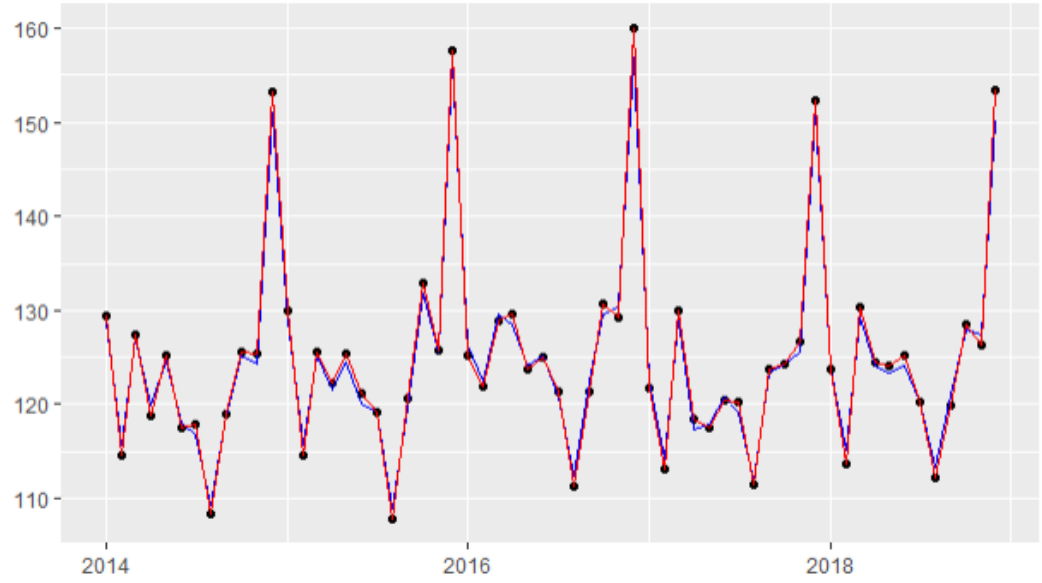
D'aquesta manera el model segueix sense poder predir l'estacionalitat i els "outliers" que es donen en els mesos de desembre no els detecta i els deixa fora del marge. Per tant no es capaç de fer una bona regressió només amb el temps, per això afegirem, com s'ha fet anteriorment, un seguit de variables que representin un 95% de la despesa total, d'aquesta manera esperem que el model si sigui capaç de detectar els valors de desembre i faci una bona regressió de les dades.

Figura 31 Nou primer model SVR.



Com es pot comprovar al gràfic la millora es considerable y la regressió augmenta en qualitat respecte la primera realitzada per aquesta variable. Ara bé si disminuïm l'error que estem disposats a assumir com hem fet amb el consum i busquem els millors paràmetres podem millorar encara més la predicció realitzada. Segons la funció *tune* els millors paràmetres són un ε de 0 i un cost de 3 i el resultat es encara millor.

Figura 32 Nou primer model vs Nou millor model SVR



Com s'observa els resultats son molt bons, amb els nous paràmetres i això es reflexa en els resultats obtinguts segons els criteris de validació que s'han utilitzat en aquest treball. Es pot observar com les prediccions son molt bones i per tant tenim valors molt bons.

Taula 8 Criteris de Validació SVR

CRITERI	VALOR
Error Quadràtic Mig	3.2483
Error Absolut Mitjà	1.4229
Arrel quadrada del error quadràtic mitjà	1.8023
Error percentual mitjà	1.1801%

6. Predicció de l'impacte del coronavirus

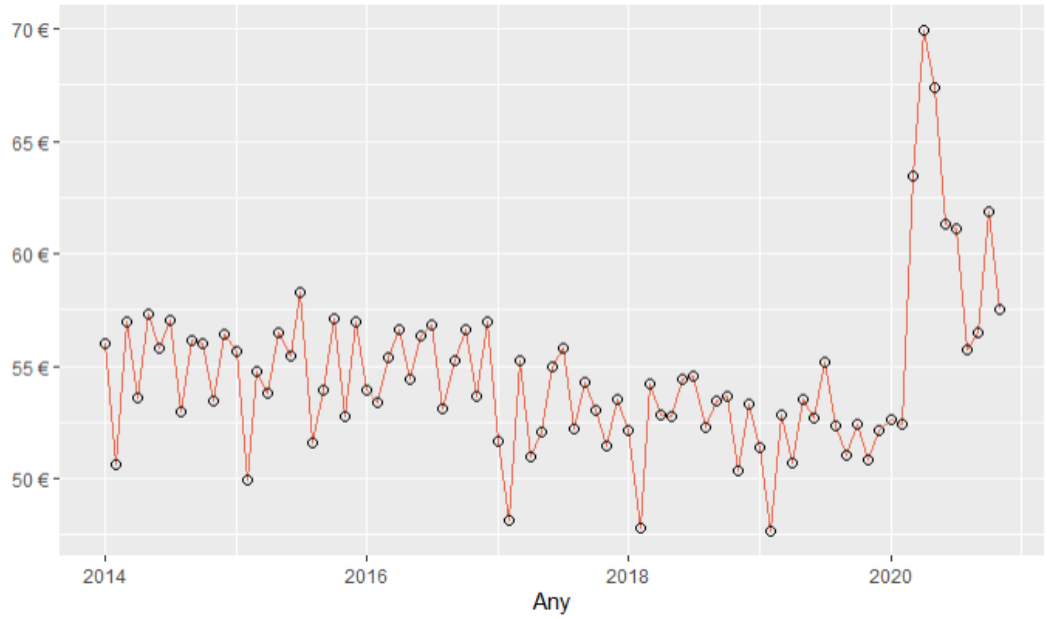
En aquest apartat, a partir dels millors models realitzats, per a cadascuna de les dos variables estudiades al llarg d'aquest treball, el consum i la despesa per càpita, es realitzarà la predicció de l'any 2020, i analitzarem la capacitat predictiva dels models construïts davant d'un shock inesperat com ha sigut la pandèmia provocada per la covid-19.

Per començar amb aquesta avaluació dels models iniciarem amb el consum per càpita per tal de seguir la mateixa tònica de l'assaig. La manera en que es procedirà en aquest punt del treball serà el següent, primer es farà un anàlisi amb els models ARIMA, es continuarà amb el model dels SVR i finalment es realitzarà una comparativa dels dos resultats obtinguts per tal de decidir quin dels dos models ha predit millor aquest xoc totalment inesperat.

6.1 Predicció Coronavirus: Consum per càpita

Abans de començar, però, cal representar les dades de l'any 2020, ja que fins aquest moment no s'han pogut observar els canvis en el consum respecte anys anteriors.

Figura 33 Consum per càpita alimentari d'Espanya (2014-2020)



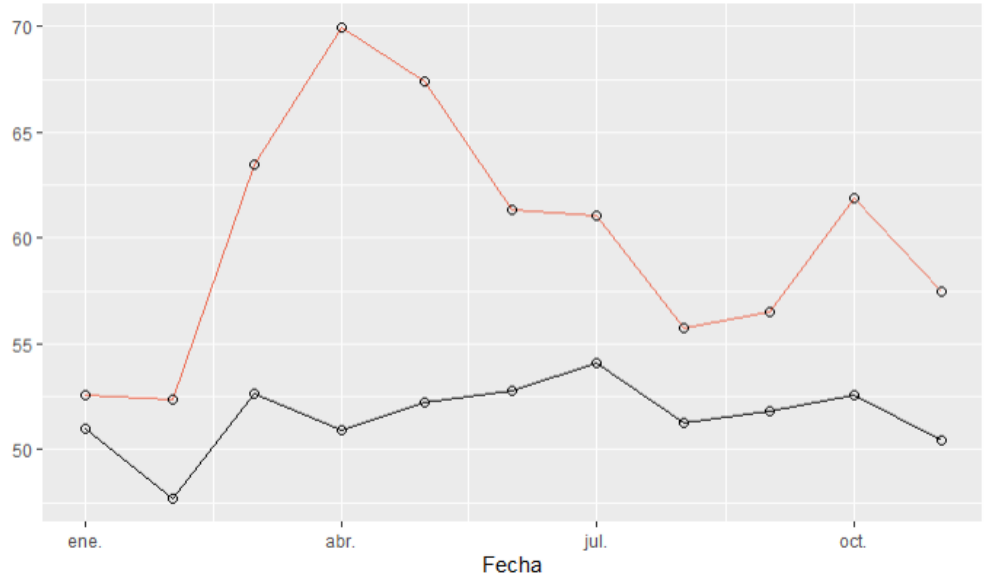
Com es pot observar l'any 2020 s'ha vist molt afectat degut a la irrupció del coronavirus, on el mes de març es veu un augment molt elevat del consum per càpita, i encara major al mes d'abril, possiblement degut al confinament i per la por de que aquest dures més temps del que estava estipulat.

6.1.1 Predicció Coronavirus: Consum per càpita ARIMA

Seguint amb l'anàlisi començarem amb el model ARIMA per tal d'intentar predir aquest xoc. Com a mode de recordatori el model que es va escollir en primer lloc va ser un $SARIMA(1,0,1)(1,0,1)_{12}$.

Implementarem aquest model amb les dades des de l'any 2014 fins l'any 2019 i seguidament es realitzarà la predicció de l'any 2020.

Figura 34 Consum per càpita alimentari 2020 (Predit ARIMA vs Real)



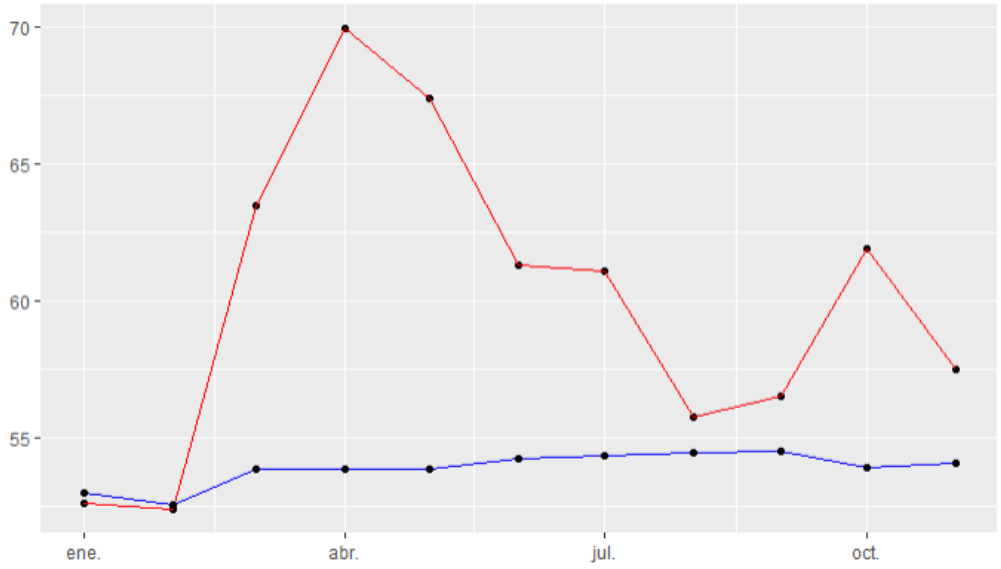
Com es pot comprovar al gràfic la predicció no es gens bona, ja que el model no es capaç de predir aquest gran xoc produït per la pandèmia, tot i que de cara a la segona part de l'any si que es capaç de reproduir la mateixa tendència que les dades originals.

6.1.2 Predicció Coronavirus: Consum per càpita SVR

Seguidament farem el mateix procediment que hem realitzat amb els models ARIMA, però amb els SVR, entrenarem el model amb dades compreses des de 2014 a 2019 i a continuació predirem les dades de l'any 2020. Els paràmetres que s'han escollit són els mateixos que s'han utilitzat en el millor model creat amb anterioritat, ε de 0 i cost de 4.

A continuació es mostren els resultats:

Figura 35 Consum per càpita alimenatri 2020 (Predit SVR vs Real)



Com podem observar la predicció que realitza el nostre model no es bona, el model no te la capacitat de reaccionar al impacte produït pel virus i realitza una predicció poc acurada als resultats observats.

6.1.3 Predicció Coronavirus: Consum per càpita, Resultats.

En aquest apartat es recolliran de forma numèrica els resultats obtinguts de cada model en una taula per tal d'observar quin dels models s'ajusta millor, tot i que ja hem observat que no es un bon ajust.

Taula 9 Criteris de Validació ARIMA vs SVR

CRITERI	ARIMA	SVR
Error Quadràtic Mig	85.78	59.39
Error Absolut Mitjà	7.68	5.72
Arrel quadrada del error quadràtic mitjà	9.26	7.71
Error percentual mitjà	12.29	8.90

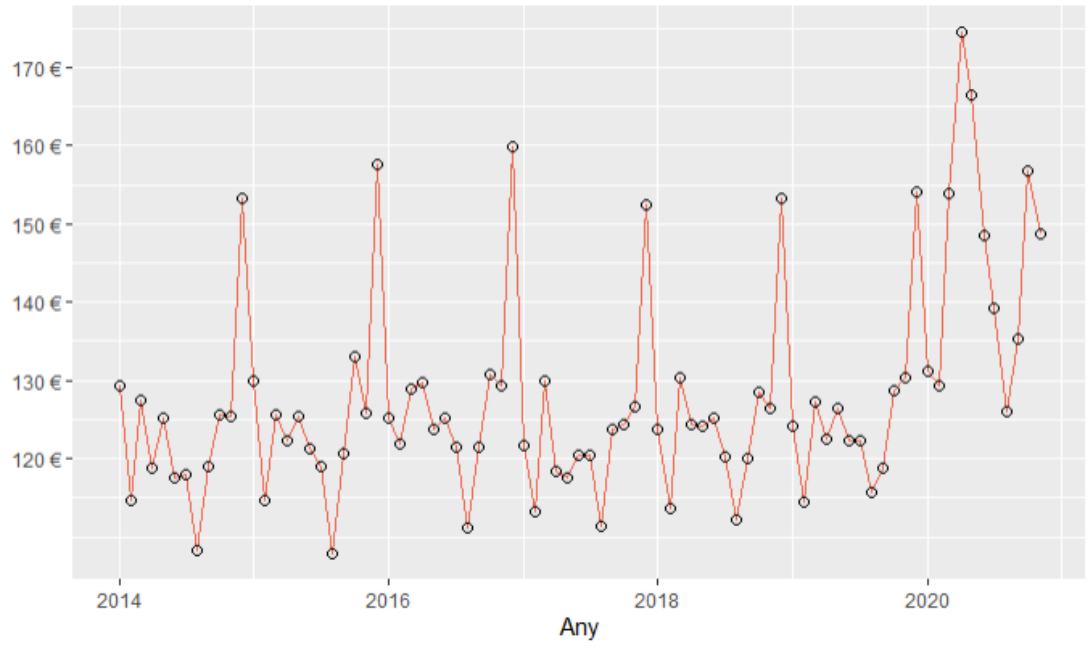
Com es pot comprovar a la taula els resultats no son acurats, com si ho eren anteriorment amb la predicció de l'any 2019, a cap criteri ens trobem en un llindar acceptable d'error, però això ens ho podíem imaginar observant el gràfic.

Seguidament es realitzarà el mateix procés, però en aquest cas amb la despesa en aliments per càpita per observar si el comportament d'ambdós models es diferent al observat en el consum.

6.2 Predicció Coronavirus: Despesa per càpita

Com hem fet amb anterioritat primer representarem les dades de l'any 2020, per poder observar en primera instància com es distribueixen els valors d'aquest any i sí segueixen l'estacionalitat que ha caracteritzat aquest conjunt de dades.

Figura 36 Despesa per càpita alimentaria d'Espanya (2014-2020)



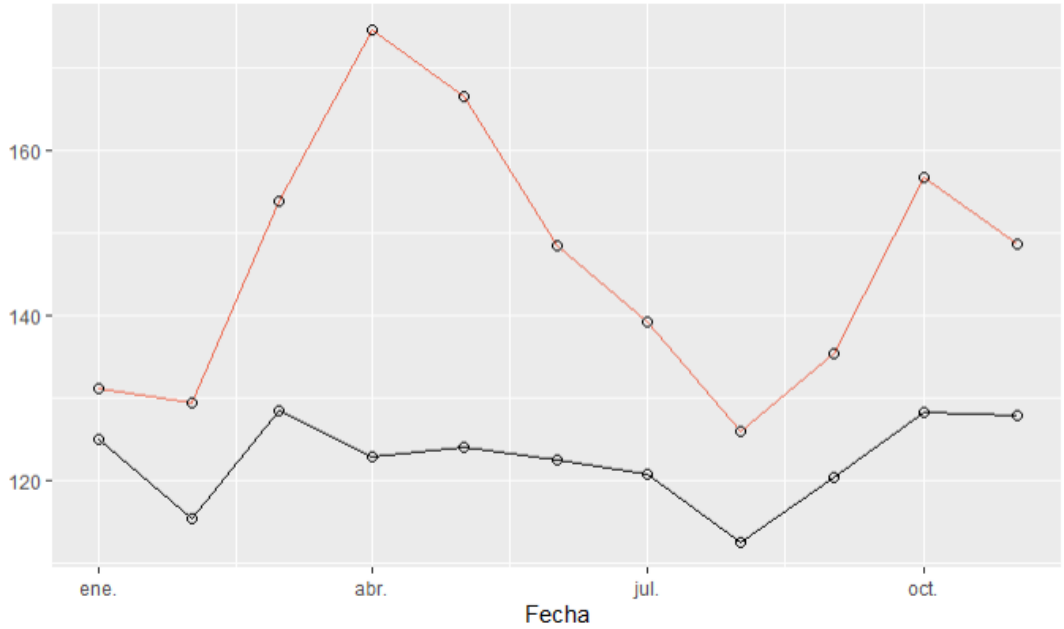
A primera vista veiem com tota l'estacionalitat s'ha perdut, ja que l'impacte produït pel coronavirus va obligar a canviar els patrons de consum de la població espanyola, aquest fet ja l'havíem observat amb anterioritat amb el consum per càpita.

6.2.1 Predicció Coronavirus: Despesa per càpita ARIMA

Iniciarem l'anàlisi amb els models ARIMA, seguint la mateixa metodologia que amb s'ha utilitzat amb el consum per càpita, i farem ús del model construït al apartat 5, aquest model es tracta d'un $SARIMA(1,0,1)(1,0,1)_{12}$.

S'implementarà aquest model amb les dades fins l'any 2019 i es farà la predicció sobre l'any 2020, i seguidament es mostraran els resultats.

Figura 37 Despesa per càpita 2020 (Predit ARIMA vs Real)

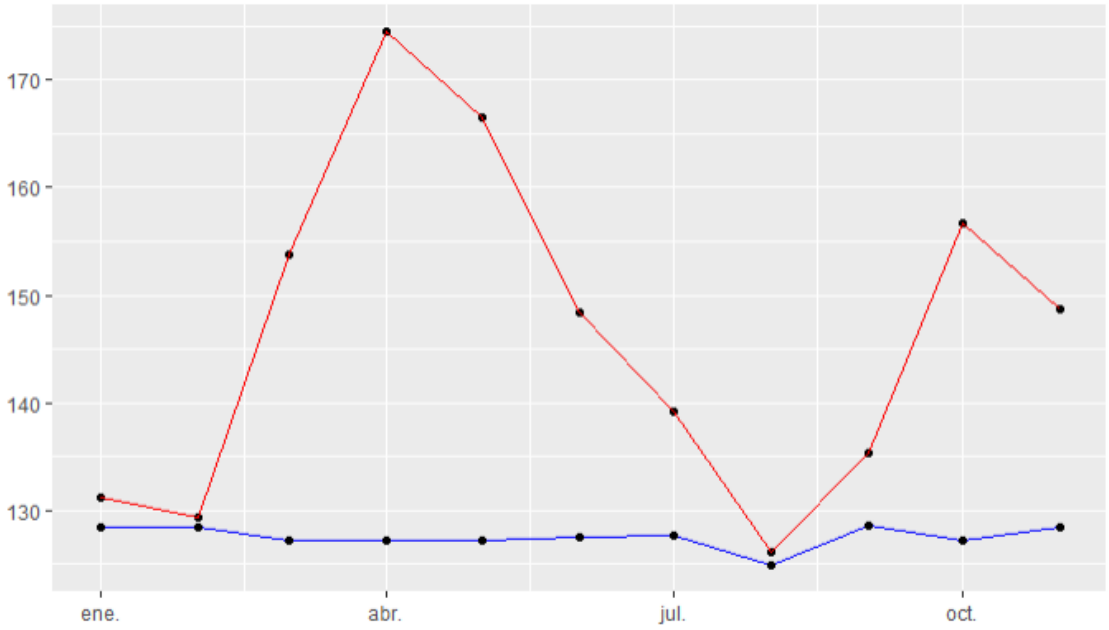


Veiem com passava amb el consum per càpita aquest model no es capaç de predir, tampoc, la despesa per càpita. Tot i que la tendència de les dades si que es mes o menys capaç de predir-la.

6.2.2 Predicció Coronavirus: Despesa per càpita SVR

Seguint amb l'anàlisi realitzat ens tocarà implementar el model realitzat amb anterioritat a la despesa per càpita de l'any 2020. Es seguirà la mateixa manera de fer que amb el consum per càpita, on s'entrenarà el model fins l'inici de l'any 2020 amb les dades de 2014 fins 2019 i seguidament és realitzarà la predicció convenient. Els resultats es mostren a continuació al gràfic:

Figura 38 Despesa per càpita 2020 (Predit SVR vs Real)



Com era d'esperar seguint la tendència dels models ARIMA i del model de SVR amb el consum ens podíem esperar un model poc precís, i així ha sigut.

6.2.3 Predicció Coronavirus: Despesa per càpita, Resultats.

Com s'ha realitzat amb el consum és recolliran en una taula els resultats de les prediccions per tal de veure'ls d'una manera més objectiva.

Taula 10 Criteris de Validació ARIMA vs SVR

CRITERI	ARIMA	SVR
Error Quadràtic Mig	665.10	534.68
Error Absolut Mitjà	21.76	17.28
Arrel quadrada del error quadràtic mitjà	25.79	23.12
Error percentual mitjà	14.28	10.98

Com ja podíem inferir els resultats no son gaire bons i cap dels dos models es un bon predictor de l'any 2020 per la despesa per càpita en aliments.

7. CONCLUSIONS

Per últim, un cop ja s'han obtingut tots els resultats necessaris per aquest anàlisis, es farà un recull de conclusions d'aquests i un seguit de idees que han sorgit mentre es realitzava aquest treball.

L'objectiu principal d'aquest treball era realitzar una comparativa entre dos models estadístics per la predicció de variables econòmiques, en el cas que ens interessa, el consum i la despesa alimentària per càpita de la població espanyola, i poder observar com es comportaven aquest models. Els models escollits van ser els models ARIMA, una metodologia tradicional molt utilitzada, i que encara ara es segueixen utilitzant i els *Support Vector Machine,* que es van desenvolupar amb anterioritat i es una metodologia més moderna amb menys recorregut en aquest àmbit. A més, en conseqüència de la situació tan poc usual viscuda arreu del món, degut a l'aparició del coronavirus, s'ha volgut estudiar com aquestes dues metodologies es comportaven respecte el xoc que ha provocat el virus.

En primer lloc, es va voler posar a prova els dos models amb dades recollides entre els anys 2014 i 2019, per tal d'analitzar com els models es comportaven amb dades més estàndards. D'aquí es poden treure diferents conclusions. La primera es que tots dos models serveixen per predir els resultats de ambdues variables, però en el cas dels SVM és necessària l'obtenció de més informació, a més informació faig referencia a que per a fer una predicció bona amb els SVM es va haver d'utilitzar un seguit de variables per tal d'entrenar el model i obtenir uns bon resultats, en canvi, en el cas del mètode de Box-Jenkins només va ser necessari per fer una bona predicció els valors totals i les dates de les dades. Per tant, totes dos metodologies són completament vàlides per la funció que s'està testejant en aquest treball, ja que les diferencies entre models son molt petites, però si es cert que per els SVM és necessita més informació, per tant es podrien utilitzar els dos mètodes en funció de les dades que es disposin.

En segon lloc, cap dels dos models utilitzats ha servit per realitzar una predicció davant d'un xoc tan gran, com el provocat pel coronavirus. No son capaços d'adaptar-se i anticipar l'impacte que aquesta situació, totalment extraordinària, ha causat. S'ha pogut comprovar com els valors dels criteris de validació utilitzats són molt elevats i en cap cas s'apropen a una predicció acceptable de la realitat observada.

En tercer i últim lloc, un cop vist el potencial, i tenint les dades necessàries a l'abast, dels dos models queda pendent cap al futur l'estudi de les diferents comunitats per poder establir també un component geogràfic a l'estudi, a més del temporal. També podria ser interessant realitzar un estudi de l'estructura alimentaria de la població espanyola, entesa com la distribució dels diferents aliments que es consumeixen, i inferir cap al futur com aquesta estructura pot variar respecte a un punt concret del temps.

8. BIBLIOGRAFIA

- SURIÑACH, J. [et al.]. *Análisis económico regional : nociones básicas de la teoría de la cointegración*. Barcelona : Fundació Bosch i Gimpera : Antoni Bosch, 1995

- NOVALES, A. *Econometría.* 2a ed. Madrid : McGraw-Hill, 1993

-Gunn, Steve R. "Support vector machines for classification and regression." *ISIS technical report* 14.1 (1998): 5-16.

ANNEX

El codi d'R i bases de dades utilitzats per la realització d'aquest treball, de forma integra, es troben de forma publica en el següent enllaç: https://github.com/danivillalobostorrejon/TFG. D'aquesta manera es poden reproduir els resultats obtinguts al llarg d'aquets treball.