UNIVERSITAT DE BARCELONA

Anàlisi de l'atur registrat a Catalunya en els últims anys

ANÀLISIS DE SÈRIES TEMPORALS

LAURA JULIÀ MELIS

07.01.2019

En aquest informe es realitza una anàlisi estocàstica d'una base de dades amb informació relativa a l'atur a Catalunya (en milers de persones aturades, és a dir, sense feina però disponibles i buscant ocupació) amb l'objectiu de fer prediccions de les darreres observacions i comparacions de resultats amb els mètodes deterministes aplicats anteriorment.

Índex

1. Introducció.

En aquest document es recull una anàlisi estocàstica sobre l'atur registrat a Catalunya en els últims anys. Per entendre millor l'objecte d'estudi d'aquest informe, cal entendre bé què s'entén per **atur** i com es mesura. L'**atur** d'un país és el nombre de població activa que no treballa, és a dir, una persona es troba en situació d'atur quan no té un lloc de feina però voldria treballar. Les dades es recullen mirant el nombre d'aturats inscrits a les oficines d'ocupació de Catalunya cada mes.

La informació sobre l'atur registrat a Catalunya (xifres en milers de persones) s'ha obtingut de la pàgina web de l'IDESCAT (https://www.idescat.cat/indicadors/?id=conj&n=10220&col=1) i la font de les dades és el Departament de Treball, Afers Socials i Famílies. Originalment, la base de dades contenia l'atur segregat per sexe i grups d'edat, així com també l'atur total des del gener de l'any 1996 fins a l'agost del 2018. Per al present estudi, només s'ha utilitzat la xifra d'atur dels darrers 6 anys i 8 mesos (01/2012 – 08/2018) i, com que es tracten de dades mensuals, s'ha obtingut finalment una sèrie de 80 registres. A més, s'han dividit dos períodes en les dades: el període mostral conté les dades compreses entre els anys 2012 i 2017, i l'extra mostral, les de l'any 2018.

L'anàlisi en qüestió es farà en diverses passes. En primer lloc, es recuperaran els resultats obtinguts als tests de Daniel i de Kruskal-Wallis per conèixer la tipologia de la sèrie temporal (si té o no tendència i/o estacionalitat). A continuació i depenent dels resultats obtinguts, es solucionaran els problemes adients (si escau) per tal d'obtenir una sèrie temporal estacionaria. Un cop obtingut un procés estocàstic estacionari, es realitzaran la FAS i la FAP per identificar a quin tipus de procés s'assembla més la sèrie en qüestió. Després, es realitzarà l'estimació del model, la seva validació i, finalment, es faran prediccions. Per concloure, es farà una petita discussió de resultats comparant els obtinguts en aquest anàlisi amb els de l'anàlisi determinista fet anteriorment així com una decisió final.

2. Aplicació empírica.

2.1. Tipologia de la sèrie.

Amb el contrast de Daniel es va concloure que la sèrie té tendència (com es pot observar en la representació gràfica adjuntada a continuació) mentre que el contrast de Kruskal-Wallis no va permetre rebutja la hipòtesi nul·la i per tant va concloure que la sèrie no té estacionalitat.

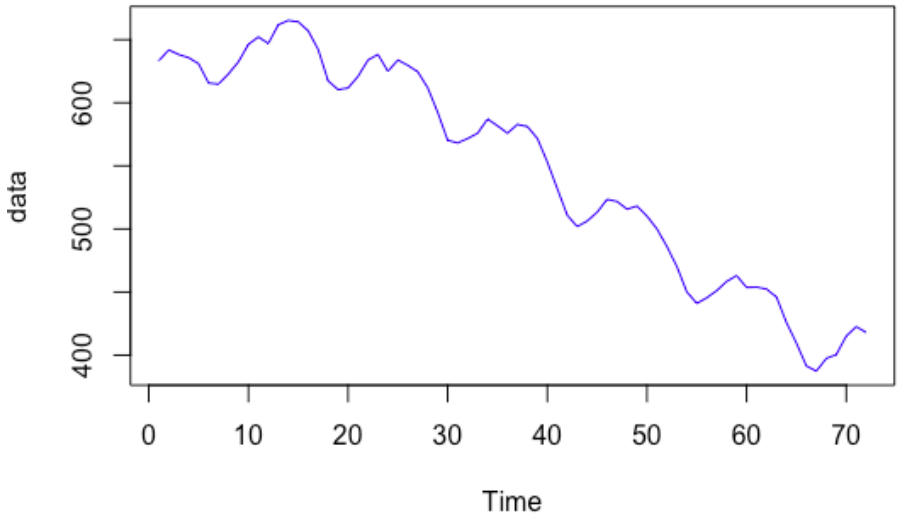

Figura 1: Representació gràfica de la sèrie temporal.

També amb les funcions d'autocorrelació i d'autocorrelació parcial veiem que la sèrie no és estacionaria (sense tendència i estacionalitat) ja que ens veu com la sèrie segueix algun patró (en FAS) i no tendeix ràpidament cap a 0 (en FAP).

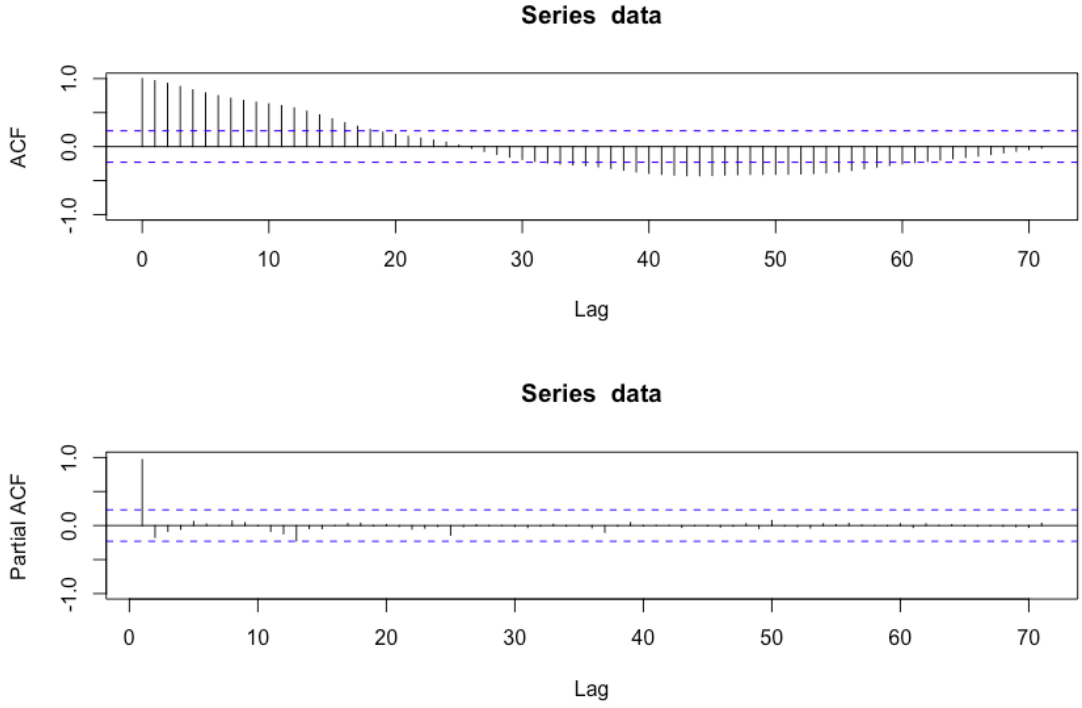

Figura 2: Funció d'autocorrelació (a dalt) i funció d'autocorrelació parcial (a baix).

En conseqüència, la sèrie objecte d'estudi és una sèrie de tipus III (sèrie amb tendència i sense component estacional).

2.2. Anàlisi estocàstica de la sèrie.

2.2.1 Obtenció d'un procés estocàstic estacionari.

Per tal de realitzar una anàlisi estocàstica de les dades sobre l'atur de Catalunya dels darrers anys es considerarà cada un dels valors concrets de la sèrie temporal (la xifra d'atur de cada mes) com una successió de variables aleatòries i.i.d. (independents i idènticament distribuïdes).

Aquesta successió es denomina **procés estocàstic**, però per a l'anàlisi necessitem que aquest procés sigui **estacionari**, és a dir, sense tendència ni estacionalitat i amb variabilitat constant durant tot el procés.

Com que ja s'ha vist que la sèrie en qüestió que es desitja analitzar té tendència, es prendran diferències regulars (ja que no té component estacional) per a eliminar-la.

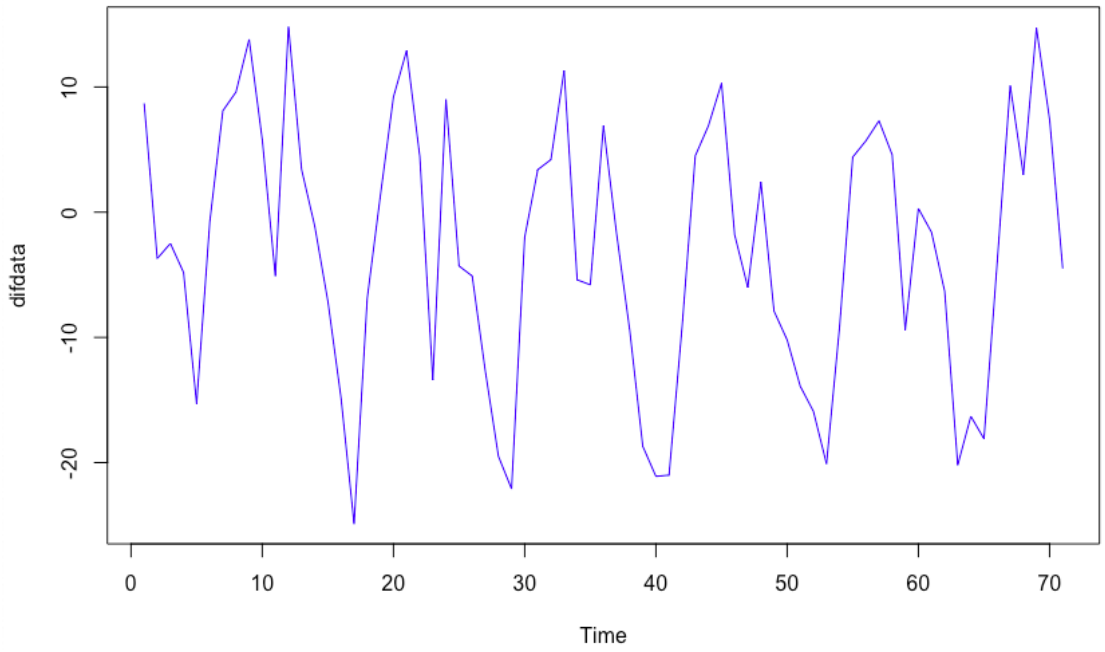

Figura 3: Representació gràfica del procés havent pres diferències.

S'observa com la tendència decreixent que tenia la sèrie ha desaparegut i, a més, sembla que la variabilitat és constant. Es comprovarà amb el test de Dickey-Fuller si el procés és estacionari.

```
        Augmented Dickey-Fuller Test

data:  difdata
Dickey-Fuller = -4.6869, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Figura 4: Sortida d'R del test de Dickey-Fuller.

Com que el p-valor és inferior a 0.05, es diu que no hi ha evidències suficients per rebutjar la hipòtesi alternativa de estacionarietat.

2.2.2. Identificació del model.

Un cop obtingut el procés estocàstic estacionari es procedeix a seleccionar el model més adequat.

A partir de les funcions d'autocorrelació i d'autocorrelació parcial es veu que hi ha diverses línies significants que ens permeten concloure que els residus no són aleatoris. Sembla que un model bo per aquest procés seria un procés mixt (una combinació de procés autoregressiu, AR, i de mitjana mòbil, MA) ja que en les dues funcions hi ha línies fora de la zona de no significació.

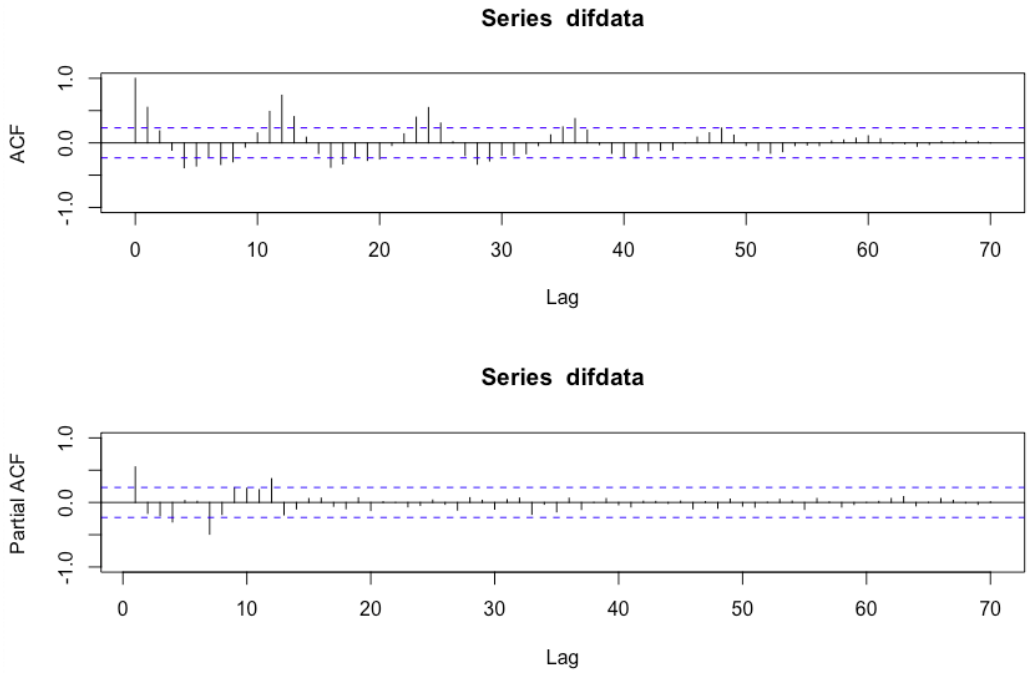

Figura 5: Funció d'autocorrelació i funció d'autocorrelació parcial del procés estocàstic estacionari.

La funció `auto.arima` del paquet **forecast** de R ens serà útil per identificar el millor model ARIMA, el que s'ajusti millor a les nostres dades. El criteri de selecció que es seguirà serà l'AIC (Criteri d'informació d'Akaike) i per tant, ens triarà el model amb el valor d'AIC més petit. La sortida obtinguda ha sigut la següent:

```
Series: difdata
ARIMA(2,0,2) with non-zero mean

Coefficients:
         ar1      ar2      ma1     ma2     mean
      0.8742  -0.7313  -0.4542  0.8654  -3.0503
s.e.  0.1172   0.1027   0.0817  0.2104   1.5068

sigma^2 estimated as 64.26:  log likelihood=-247.05
AIC=506.1   AICc=507.42   BIC=519.68
```

Figura 6: Sortida d'R de la funció auto.arima.

Els resultats indiquen que el model més adequat seria un ARIMA(2,0,2), amb un AIC = 507,42. Aquest és un model que té sentit ja que si mirem les gràfiques anteriors es pot veure una funció d'autocorrelació amb un patró semblant al d'una AR(2) i la gràfica de la funció d'autocorrelació parcial que tendeix a 0 amb valors positius i negatius semblant-se més o menys a una MA(2).

2.2.3. Estimació i validació del model.

El model s'ha estimat fent servir la funció `arima`:

```
Call:
arima(x = difdata, order = c(2, 0, 2))

Coefficients:
         ar1      ar2      ma1     ma2  intercept
      0.8742  -0.7313  -0.4542  0.8654    -3.0503
s.e.  0.1172   0.1027   0.0817  0.2104     1.5068

sigma^2 estimated as 59.73:  log likelihood = -247.05,  aic = 504.1
```

Figura 7: Sortida d'R de la funció arima.

A continuació s'ha dut a terme la validació del model, necessària per confirmar que el model ajustat és correcte i que ens permetrà realitzar bones prediccions més endavant. Per fer-ho, s'analitzaran els residus del model i la significació dels coeficients.

i. Anàlisi dels residus.

En primer lloc es mirarà que els anàlisis segueixin una distribució normal a partir del test de normalitat de Shapiro-Wilk. Com es pot observar en el quadre adjuntat sota aquestes línies, s'ha obtingut un p valor molt superior a 0.05 i per tant, no es pot rebutjar la hipòtesi nul·la de normalitat dels residus.

```
        Shapiro-Wilk normality test

data:  model$residuals
W = 0.98606, p-value = 0.6208
```

Figura 8: Sortida d'R del test de normalitat de Shapiro-Wilk.

Després, ha sigut necessari confirmar la independència dels residus amb el test de Box-Pierce, en el qual també s'ha obtingut un p valor superior a 0.05.

```
        Box-Pierce test

data:  model$residuals
X-squared = 0.87544, df = 1, p-value = 0.3495
```

Figura 9: Sortida d'R del test de independència de Box-Pierce.

Es confirmen aquets resultats (independència i aleatorietat dels residus) amb les funcions d'autocorrelació i autocorrelació parcial dels residus:

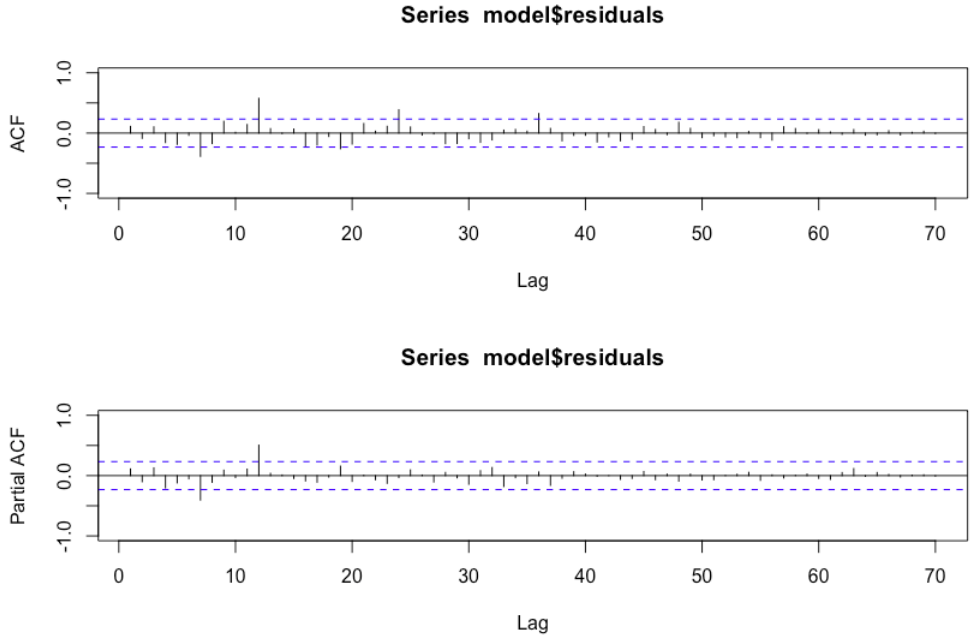

Figura 10: Funció d'autocorrelació i funció d'autocorrelació parcial dels residus del model.

ii. Significació dels coeficients.

Tant el terme constant com els altres 4 coeficients del model tenen un p valor inferior a 0.05 i per tant, es pot concloure que són tots significatius.

ar1	ar2	ma1	ma2	intercept
4.461960e-14	5.373829e-13	1.365477e-08	1.958256e-05	2.146568e-02

Figura 11: Sortida d'R del test de Dickey-Fuller.

2.2.4. Prediccions.

En aquest apartat es treballarà principalment amb el període extramostral (les dades de l'atur dels 8 primers mesos de l'any 2018).

En primer lloc s'ha realitzat un plot amb les diferències regulars de les dades originals (en blau) i els residus del model ajustat (en vermell).

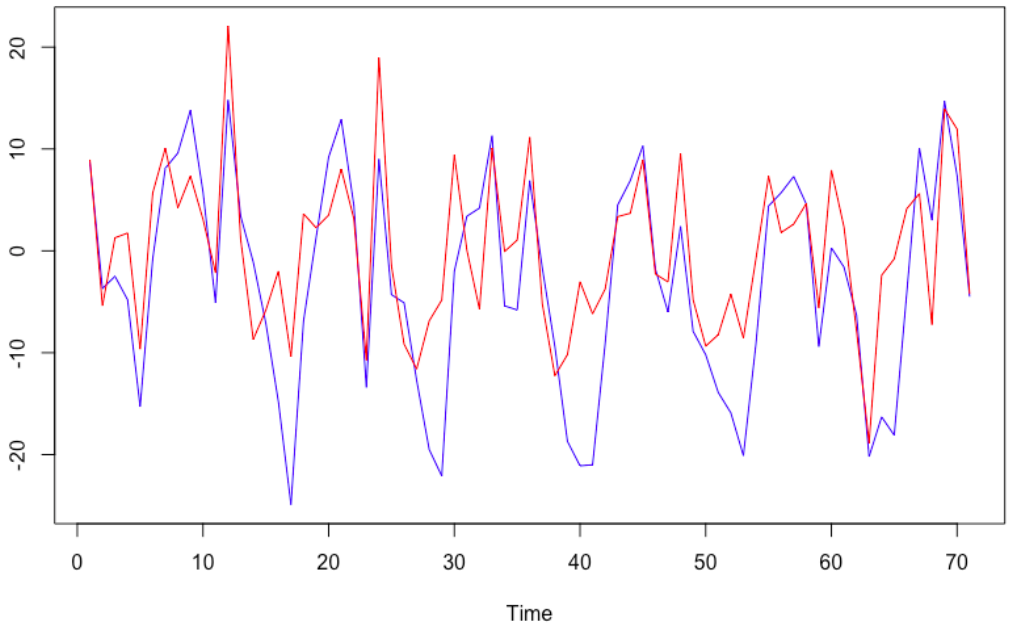

Figura 12: Representació gràfica del procés havent pres diferències enfront els residus.

A continuació es mostra un taula amb les dades extramostrals observades, les prediccions i l'error causat. La fila 72 es correspon amb el mes de gener del 2018 i així successivament fins a la 79, que fa referència al registre d'agost de 2018.

	DadesObservades	Prediccions	error
72	422.9	417.6978	5.202190
73	418.2	425.6193	-7.419315
74	411.5	423.4127	-11.912666
75	398.9	416.6826	-17.782616
76	385.6	402.2329	-16.632946
77	370.2	387.3380	-17.138012
78	369.1	371.8965	-2.796454
79	380.7	371.9265	8.773490

Figura 13: Taula de dades observades enfront les prediccions amb l'error comès.

En el següent gràfic es pot veure representat l'atur en milers de persones a Catalunya entre els mesos de gener de 2012 a agost de 2018 (línia blava) així com també les prediccions realitzades (en vermell). S'observa com les prediccions segueixen la mateixa forma que les dades observades encara que en tots els mesos s'ha sobreestimat una mica la xifra d'aturats.

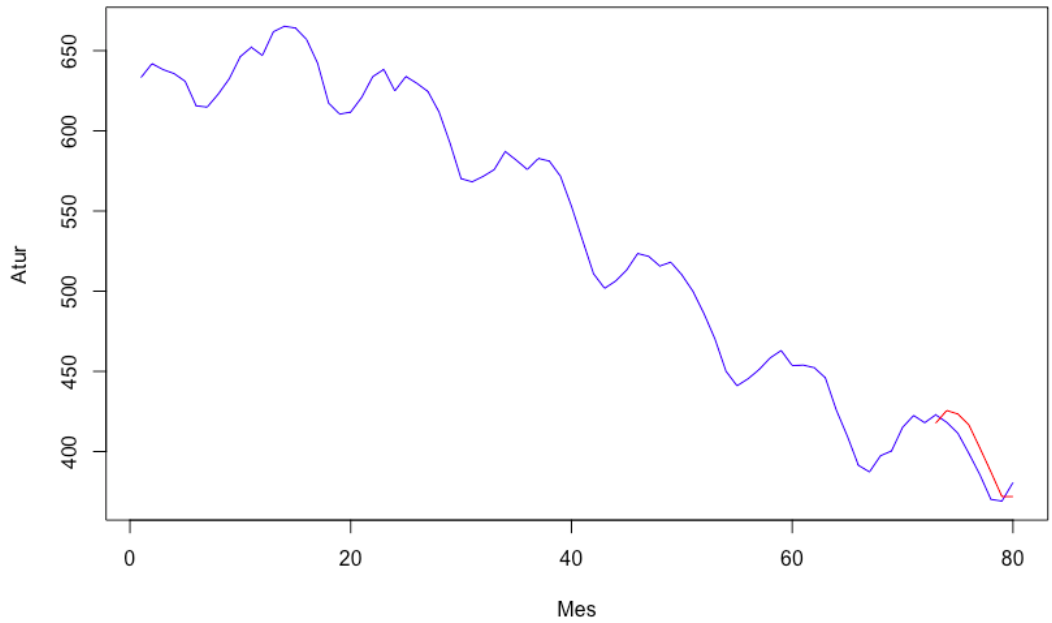

Figura 14: Representació gràfica de l'atur juntament amb les prediccions de 2018..

Finalment, s'ha avaluat la capacitat predictiva del model mesurant l'error quadràtic mitjà, l'error absolut mità i l'error percentual absolut mità. Els resultats són:

EAM	EQM	EPAM
10.95721	149,4253	2,795%

3. Comparativa i conclusions.

Per tal de comparar els resultats del present estudi amb l'estudi determinista realitzant anteriorment, es recuperaran els resultats obtinguts. Cal recordar que el mètode elegit com el millor fou el de la tendència lineal.

MÈTODE	EAM	EQM	EPAM
Ingenu	24,638	943,526	6,49%
DMM	15,842	359,302	3,90%
AEH	14,319	293,344	3,54%
Tendència lineal	11,081	155,026	2,79%
Model ARIMA(2,0,2)	10.95721	149,4253	2,795%

Amb aquesta taula resum de resultats es pot veure com els millors mètodes han estat el de la tendència lineal (mètode determinista) i el model ARIMA (2,0,2) obtingut realitzant l'anàlisi estocàstica. Ambdós tenen un EPAM molt similar i inferior al 3%. Per tant, la capacitat predictiva és força bona i al fer prediccions amb qualsevol dels dos mètodes, aquestes seran correctes. A més, encara que els EQM i EAM dels dos mètodes són també similars, el model ARIMA aconsegueix uns errors una mica inferiors.

Amb tot, l'anàlisi estocàstica que ens a portat a ajustar un model ARIMA(2,0,2) ens ha ofert les estimacions menys dolentes d'entre tots els mètodes realitzats.

4. Annex.

```
library(MASS)
library(forecast)
library(TSA)
library(tseries)

## IMPORTACIÓ DE LA BASE DE DADES.
dades<- ts(read.table("dades.txt"))
data<-dades[-c(73:80)] #Període mostral
dataExtra<-dades[c(73:80)] #Període extra-mostral

## 2.1. TIPOLOGIA DE LA SÈRIE.
plot.ts(data,col=c(4)) # Sèrie amb tendència i sense estacionalitat.

par(mfrow=c(2,1))
plot(acf(data,lag=100, plot=F), ylim=c(-1,1))
pacf(data,ylim=c(-1,1),lag=100)

## 2.2.1 OBTENCIÓ D'UN PROCÉS ESTOCÀSTIC ESTACIONARI.
## Té tendència. Agafem diferències (regulars) per a solucionar-ho.
difdata<-diff(data)
par(mfrow=c(1,1))
plot.ts(difdata,col=4) # Hem eliminat tendència

# S'ha solucionat el problema d'estacionarietat?
adf.test(difdata) # Test procés estacionari: sí és estacionari.

## 2.2.2 IDENTIFICACIÓ DEL MODEL.
par(mfrow=c(2,1))
plot(acf(difdata,lag=100, plot=F), ylim=c(-1,1))
pacf(difdata,ylim=c(-1,1),lag=100)

auto.arima(difdata) # millor model ARIMA

## 2.2.3. ESTIMACIÓ I VALIDACIÓ DEL MODEL.
model<-arima(difdata,order = c(2,0,2))
model # Estimació del model

shapiro.test(model$residuals) # normalitat residus
Box.test(model$residuals) # independencia residus
par(mfrow=c(2,1)) # Anàlisi dels residus.
plot(acf(model$residuals,lag=100), ylim=c(-1,1))
pacf(model$residuals,ylim=c(-1,1),lag=100)

pnorm(c(abs(model$coef)/sqrt(diag(model$var.coef))), mean=0, sd=1,
lower.tail=FALSE) # Significació dels coeficients

## 2.2.4. PREDICCIONS.
ts.plot(difdata,model$residuals,col=c(4,2))
prediccions<-tail(dades,12)-(predict(model,n.ahead=8)$pred)

plot(dades,type="l",xlab = "Mes",ylab ="Atur", col=4)
lines(prediccions,type="l",x=c(73:80),col="red")

# Capacitat predictiva
e <- dataExtra-prediccions
EQM  <- sum(e^2)/8 # Error quadràtic mitjà
EAM  <- sum(abs(e))/8  # Error absolut mitjà
EPAM <- (sum(abs(e)/dataExtra)/8)*100
cbind(DadesObservades= dataExtra, Prediccions=prediccions, error= e)
```