INDEX:

1. Introduction
2. Descriptive analysis.
    2.1. Description of the series.
    2.2. Decomposition Methods
        2.2.1. Classic Decomposition Methods.
3. Fitting an ARIMA model.
    3.1. Identification.
        3.1.1. Stability in variance.
        3.1.2. Stability in the mean.
        3.1.3. Elimination of the Seasonal Component.
    3.2. Parameter estimation.
    3.3. Model diagnosis.
    3.4. Overfitting.
4. Predictions.
    4.1. ARIMA Model.
5. Conclusion.

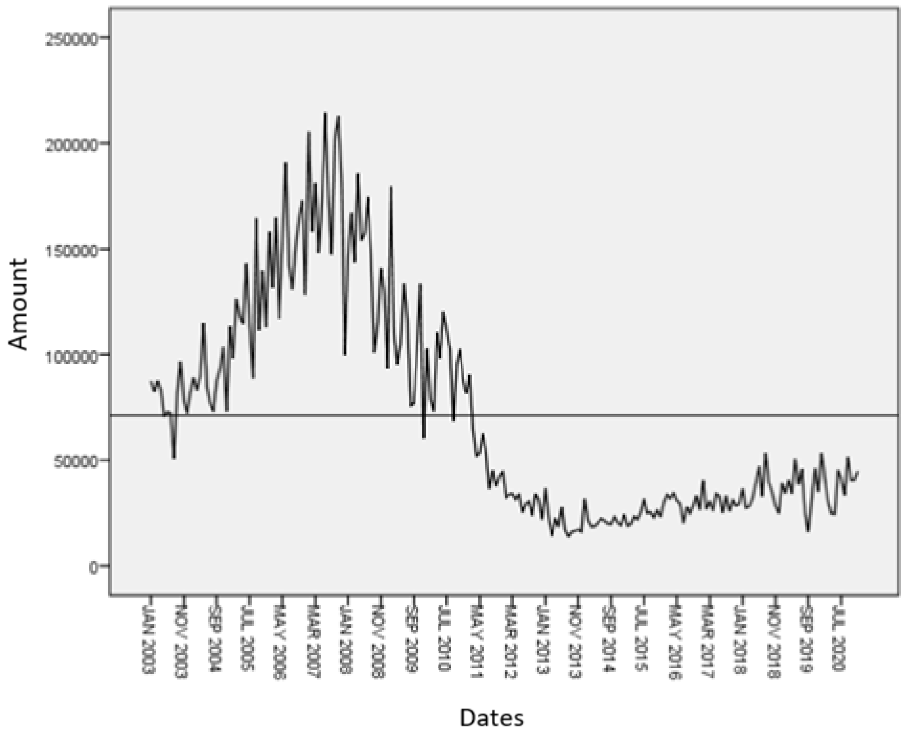
1. Introduction.

In this project we are going to study the serie "Amount of home mortgages in Extremadura since 2003"

In this serie we can observe the amount of home mortgages in thousands of euros. A mortgage consists on a loan made by a bank whose payment is guaranteed by the value of a property. The amount of the loan is the money that they lend you and must be repaid through periodic installments, within a certain period and with the agreed interest rate.

The database from which this series has been taken is the National Institute of Statistics, whose source is the Property Registry.

The behavior of the series, together with its mean is:



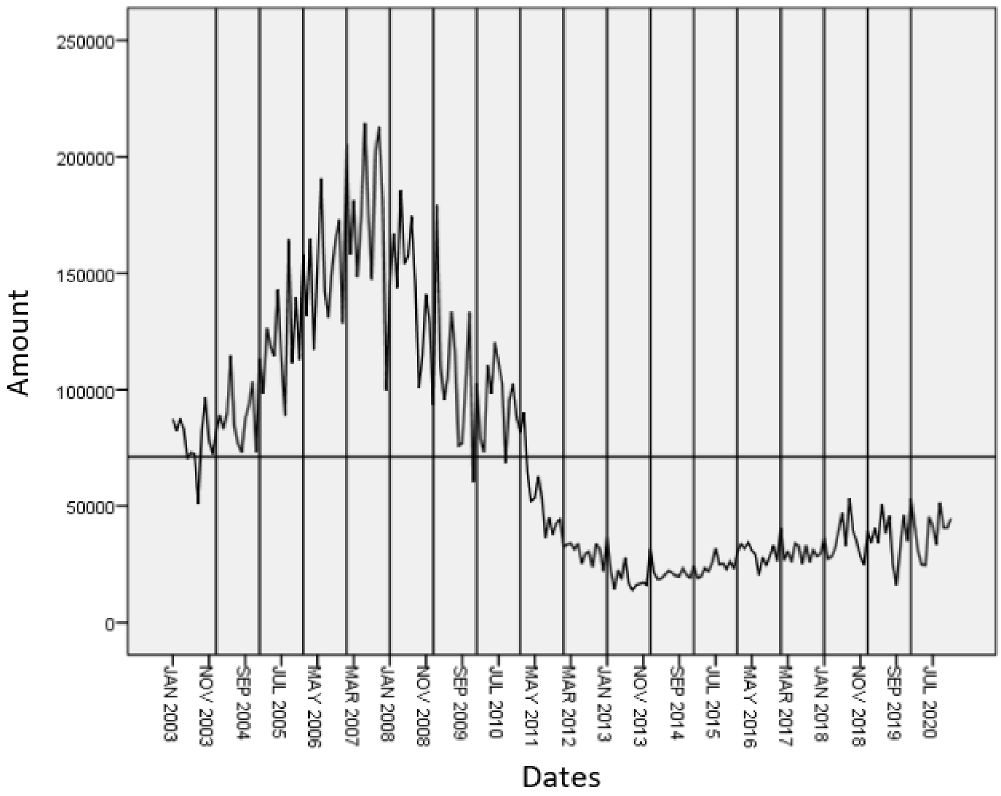
This study is interesting as we can see that in the years before the real estate bubble of 2008 the amount was much higher, mainly due to the mortgage boom, however we can see how from 2008 with the start of the crisis in Spain the amount decreases.

2. Descriptive analysis.

To begin with the descriptive analysis of the serie we need to carry out a descriptive study of it.

2.1 Description of the series.

Lets begin with the representation of the serie.



As we have mentioned before, from 2008 (Horizontal line 5) the amount of mortgages decreases drastically.

In this way we can appreciate three quite differentiated zones:

- A first section (2003-2008) where the series shows an upward trend, as time goes by the amount of mortgages increases, that is, more mortgages are made.
- A second section (2008-2013) where the series shows a tendency to decrease, which is equivalent to less amount, less mortgages.
- A third tranche (2013-2020) where we can see stability, although in recent years it seems that another upward trend in mortgages has begun.

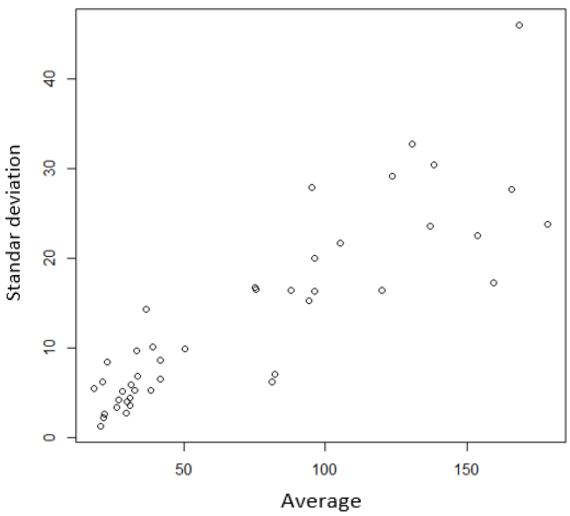
2.2 Decomposition Methods.

We are going to study more precisely the components of variation of the series, for this we will use the decomposition methods.

We know that in our series the seasonal period is 1 year, let's see if an additive model or a multiplicative model is more appropriate, for them we look at the variation of the seasonal fluctuations of the series with respect to its level.

In this case we can see, thanks to the graph of the series, that as the level of the series changes, the seasonal fluctuations change, so we will use a multiplicative model.

To make sure of this we can see the average-scatter plot of it.



As we can see the average dispersion graph for a complete cycle, the first values are closer, while the last observations are more dispersed, which prompts us to use a multiplicative model and affirm that the series is not stable in variance (We will see later).

However, to be completely sure, we are going to make use of a hypothesis test, so that: H0: =0 and H1: ≠0. If this contrast is significant, we will reject H0 and therefore we will need to use a multiplicative model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-11.697  -2.070  -1.055   2.437  15.367

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.48956    1.37894   0.355    0.724
dv$media     0.17872    0.01587  11.259 4.04e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indeed, the result of the test is significant, so we reject the null hypothesis, and therefore we will need to use a multiplicative model. Furthermore, the series is not stable in variance.
Let $X_t = f(T_t, S_t, I_t)$ where Xt is the observation of the series, f an arbitrary function and T_t, S_t, I_t the respective trend-cycle, seasonal and irregular components of the series. In our case, since it is a multiplicative model, we have that $X_t = T_t * S_t * I_t$.

2.2.1 Classic Decomposition Methods.

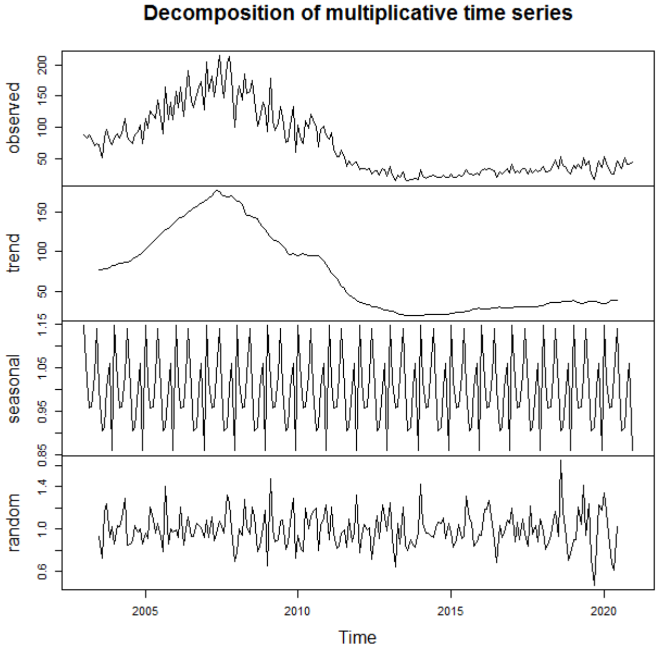
Let's look at the seasonal indices of our serie:

Seasonal factors	
Serie: Amount	
Period	Seasonal factor (%)
1	114,4
2	100,4
3	96,9
4	96,3
5	103,1
6	115,0
7	100,7
8	88,3
9	92,1
10	100,1
11	106,1
12	86,5

As we can see, the months with the least seasonal index are December and August, that is, these months are the ones in which the lowest amounts are registered for home mortgages in Extremadura.

On the other hand, the months with the highest seasonal index are January and June, that is, they are the months in which the highest mortgage amounts are recorded.

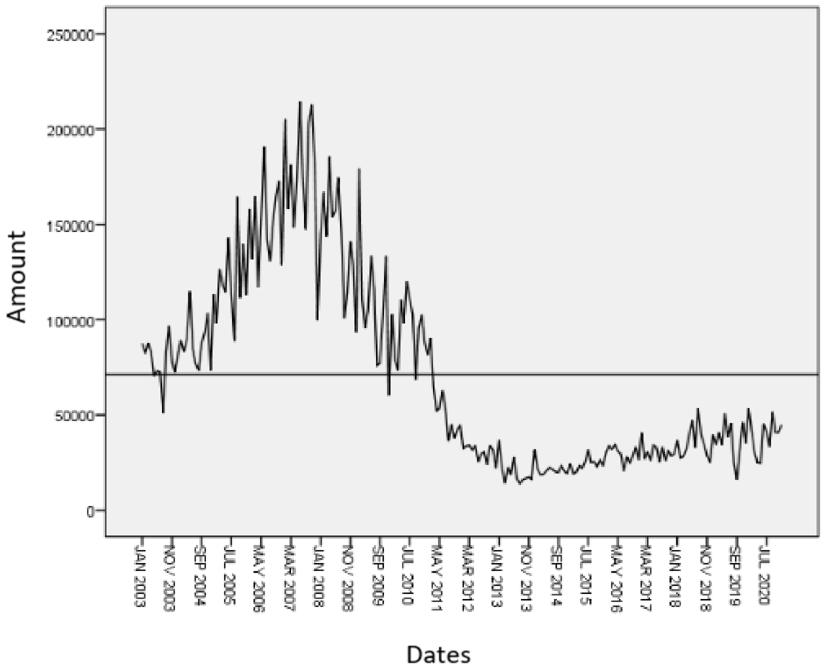
We proceed to jointly represent the variation components of the series together with it.



3. Fitting an ARIMA model.

3.1 Identification.

To carry out the identification, we first proceed to represent the series again.

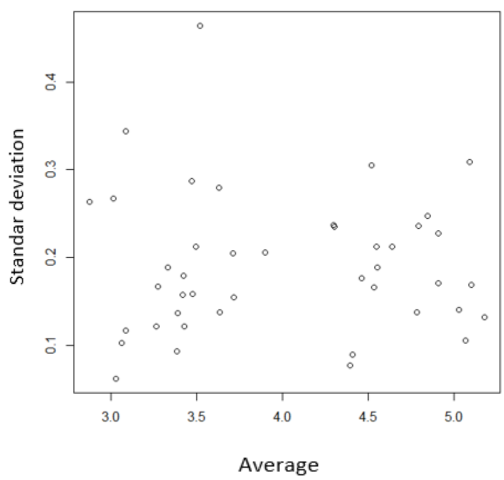


Let us see if the series is stationary. For this we need it to be stable in mean, stable in variance and not present a seasonal component.

3.1.1 Stability in variance.

Let's start with the stability in variance, for this we would make use of the mean-dispersion graph and the hypothesis test of the previous section, if we see that in the graph the points are not homogeneously distributed throughout (as it happens), or that the result of the hypothesis test is significant our series will not be stable in variance and we will need a logarithmic transformation.

Indeed, as we have already seen, the result of the test is significant and we will need to use the logarithmic transformation to have a stable series in variance.Let us see that indeed the serie $x'_t = log x_t$ is stable in variance.



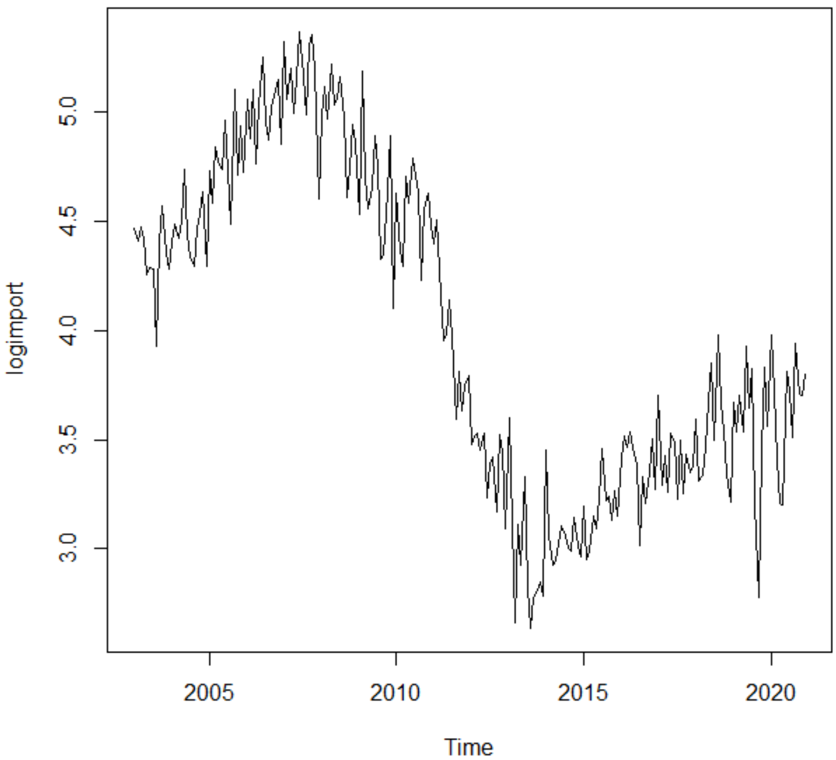
As we can see now, it does seem that the data is homogeneously distributed throughout the graph. If we perform the hypothesis test, we get:

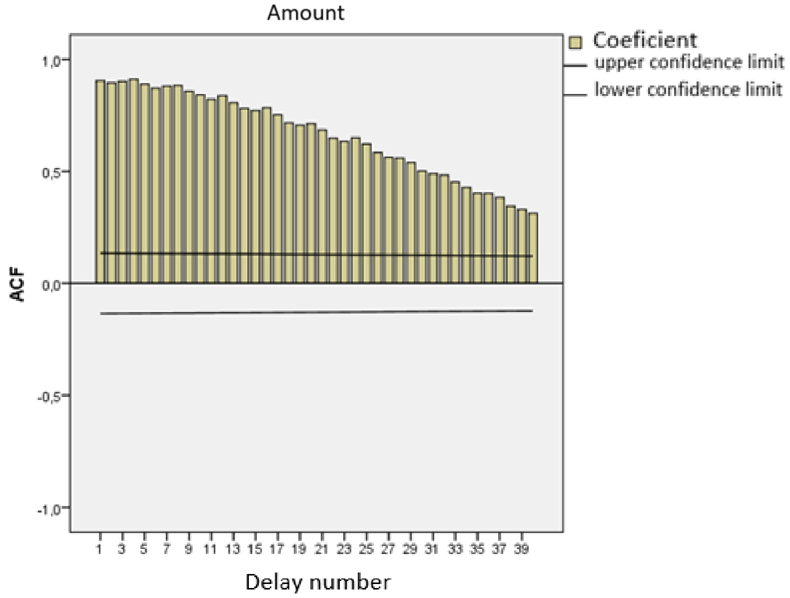
```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.12926 -0.05381 -0.01415  0.04513  0.27244 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.1945623  0.0690047   2.820  0.00737 **
dv$media    -0.0009539  0.0169816  -0.056  0.95548   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

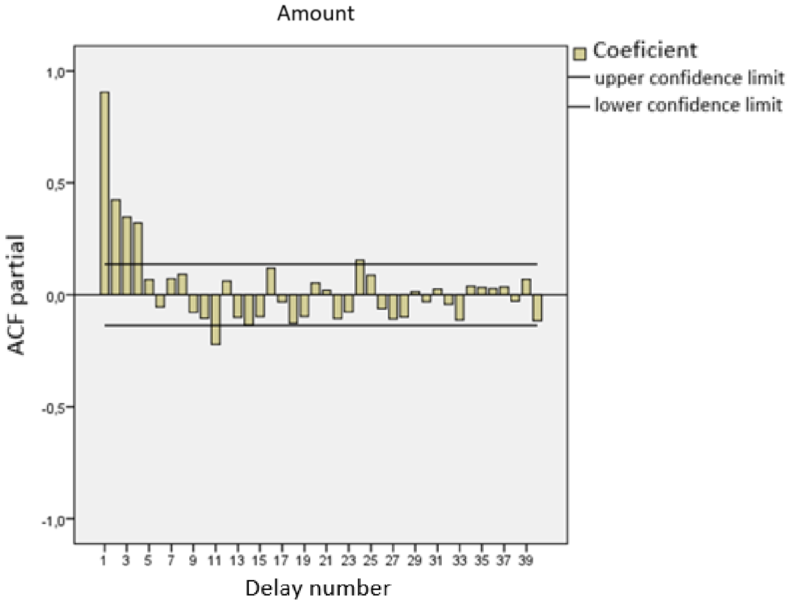
Indeed, the result of the test is not significant, so we do not have enough evidence to reject the null hypothesis and the series $x'_t = log x_t$ is stable in variance.

3.1.2 Stability in the mean.

Now let's see if our series $x'_t = log x_t$ is stable in mean, for this we represent the graph of our series together with its functions fas and fap, so that if we observe in the graph of the fas a slow decrease to zero and the graph of the fap that the first delay is very close to one, our series will not be stable in mean and we will need to make use of an ordinary differentiation.

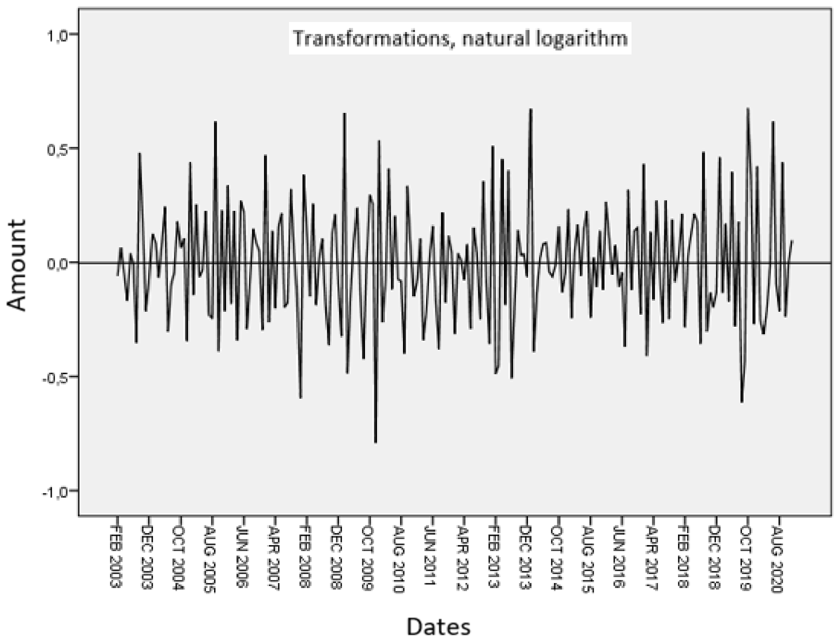


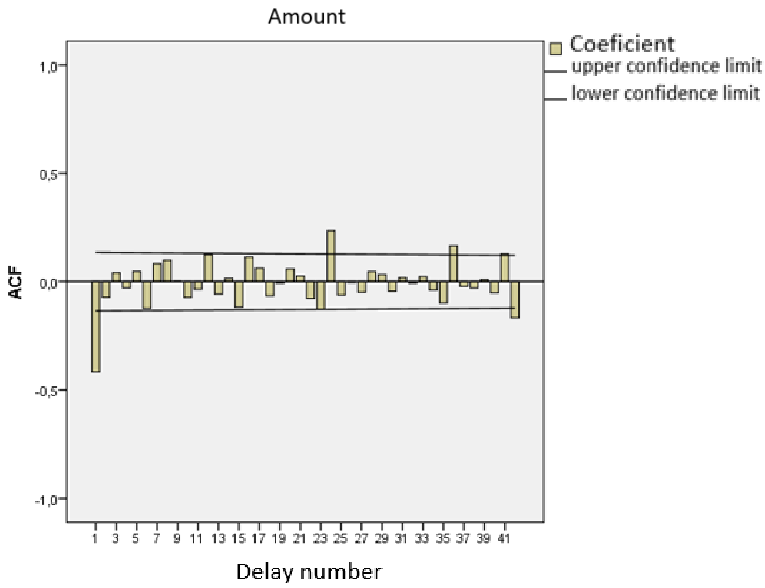


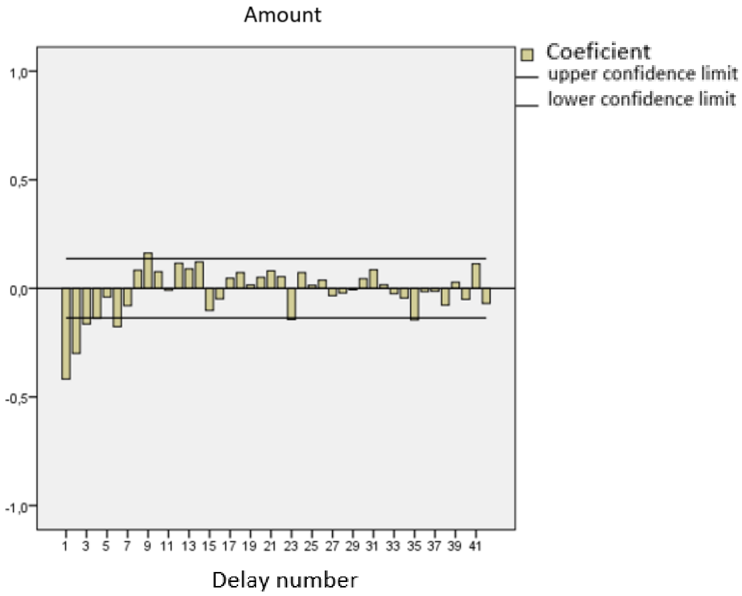


Indeed, we see that the aforementioned occurs, then the series $x_t^{'} = log x_t$ is not stable in mean and we need to make use of an ordinary differentiation so that we obtain a transformed series $x_t^{'} = (1 - B) log x_t$.

We are going to represent the series $x_t^{'} = (1 - B) log x_{t'}$, along with its fas and fap to observe that we indeed have a stable series in mean.


Transformations, natural logarithm
Amount
Dates
1,0
0,5
0,0
-0,5
-1,0
FEB 2003
DEC 2003
OCT 2004
AUG 2005
JUN 2006
APR 2007
FEB 2008
DEC 2008
OCT 2009
AUG 2010
JUN 2011
APR 2012
FEB 2013
DEC 2013
OCT 2014
AUG 2015
JUN 2016
APR 2017
FEB 2018
DEC 2018
OCT 2019
AUG 2020

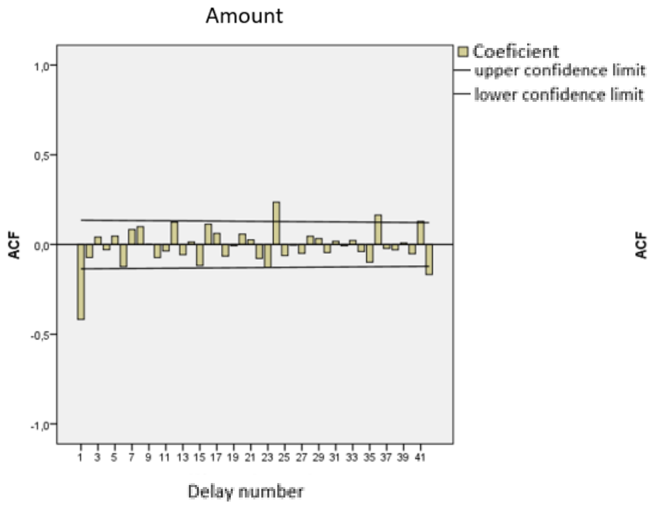

Amount
Coeficient
upper confidence limit
lower confidence limit
ACF
Delay number
1,0
0,5
0,0
-0,5
-1,0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41

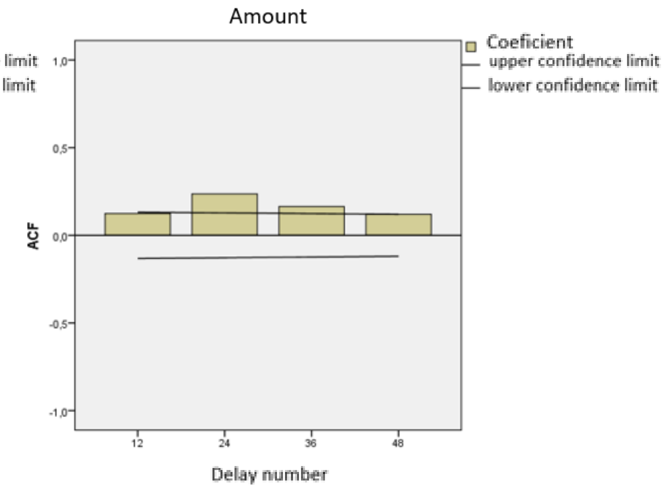


As we can see, we have a series that does not present any type of trend and that, in turn, is stable in variance.

3.1.3 Elimination of the Seasonal Component.

We now represent the fas in the multiple lags of the seasonality of the series without trend and stable in variance $x'_t = (1 - B)logx_t$.

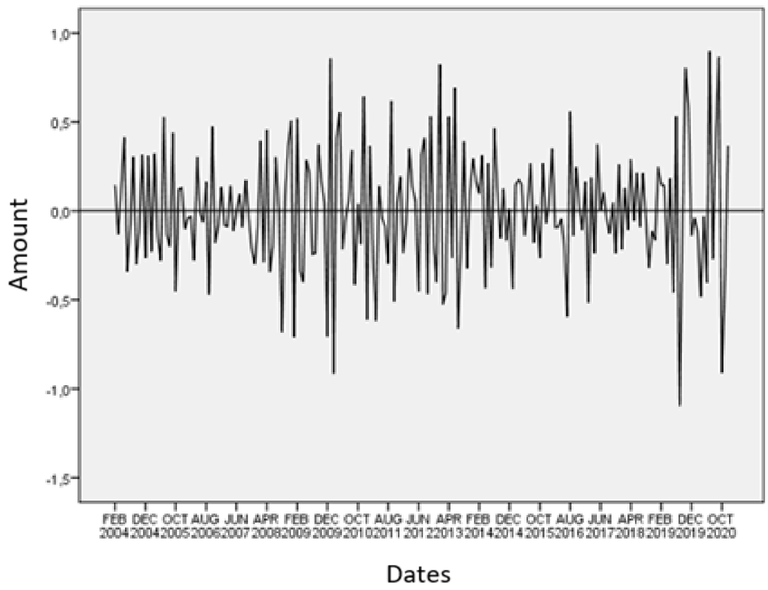




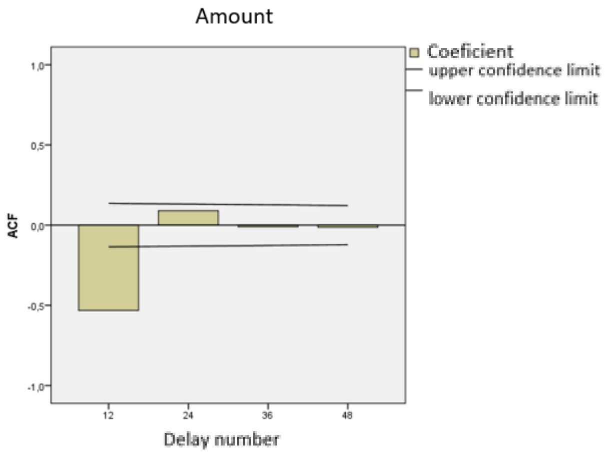
If we look at the phase of the series $x'_t = (1 - B)logx_t$ we can see that the multiple lags of the seasonality stand out from the rest and slowly decrease to zero, so we will need a cycle differentiation to finally obtain a stationary series.

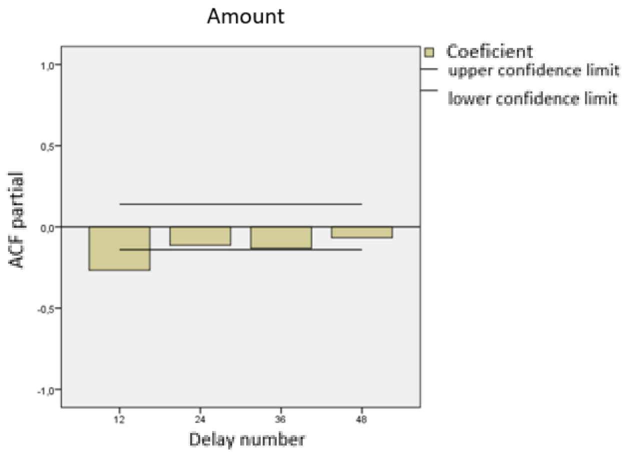
Let us see that the transformed series $x'_t = (1 - B)(1 - B^{12})logx_t$ does not have a seasonal component.

If we represent it we have:



Transformations: natural logarithm, difference, seasonal difference, period





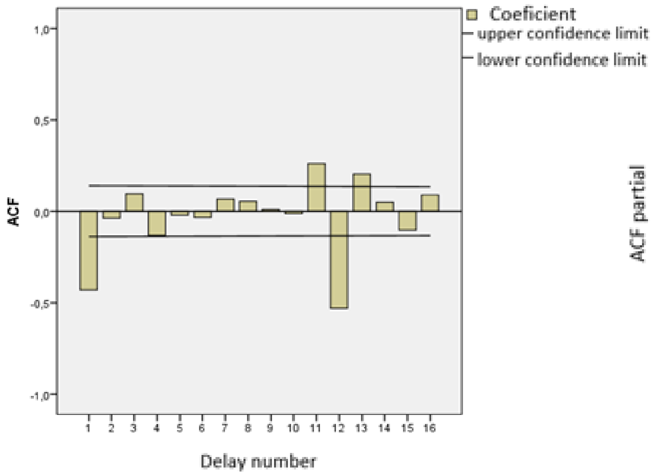
Indeed, the transformed series $x'_t = (1 - B)(1 - B^{12})logx_t$ is already stationary.

3.2. Parameter estimation.

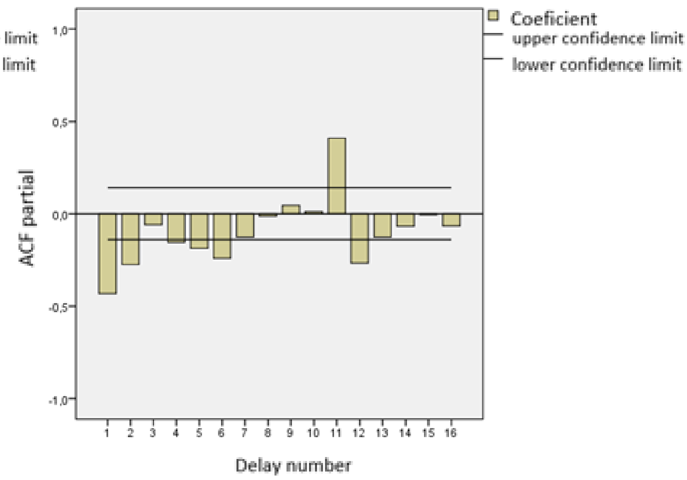
Once we have a seasonal series, we are going to proceed to identify the orders of the autoregressive and moving average parts of both the seasonal and ordinary parts.

Let's start with the identification of the orders of the ordinary part, for this we represent the fas and the fap of the seasonal series $x'_t = (1 - B^{12})(1 - B)logx_t$.

Amount



Amount

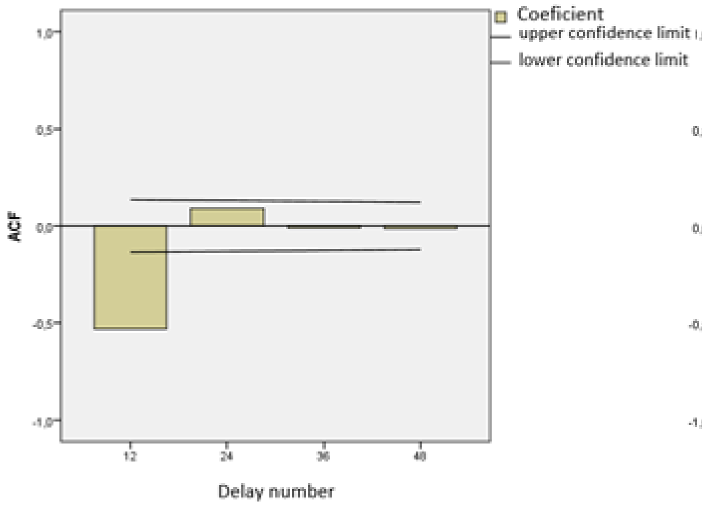


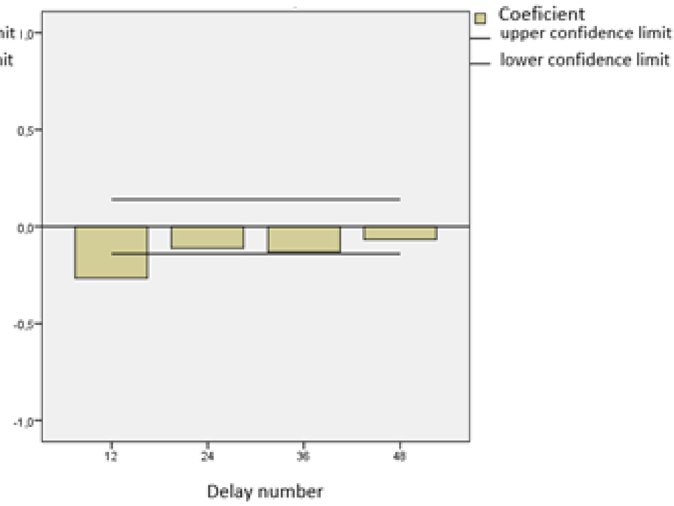
Amount

Amount

We can see that it can be an MA(1) since after the first delay the fas becomes zero and the fap decreases sinusoidally to zero or it can also be an AR(2) since after the first two delays the fap is goes to zero and also fas decreases to zero.

Let us now see the orders of the seasonal part, for this we look at the fap and the fa of the delays that are multiples of the seasonality and we observe that:





As we can see, it can be a MA(1) because the first delay of the fas is clearly significant or an AR(1) because the first delay of the fap is significantly different from zero.
Then we have 4 possible models

1. **ARIMA(0,1,1) x ARIMA(0,1,1)$_{12}$**
2. **ARIMA(2,1,0) x ARIMA(0,1,1)$_{12}$**
3. **ARIMA(0,1,1) x ARIMA(1,1,0)$_{12}$**
4. **ARIMA(2,1,0) x ARIMA(1,1,0)$_{12}$**

We now proceed to the estimation of the parameters.

For this, we use SPSS, and we will model the time series with each of the previous models, including the search for outliers, since it can cause the parameter estimates to change.

In order not to lengthen the process too much, I will only include the estimate of the constant in the case in which it is significant.

Carry out the process explained with the first model, for the others I will only put the estimates of the parameters.

Model Fit

Fit Statistic	Mean	SE	Minimum	M
R squared stationary	,622	.	,622	,622
R squared	,878	.	,878	,878
RMSE	18665,218	.	18665,218	18665,218
MAPE	17,840	.	17,840	17,840
MaxAPE	94,840	.	94,840	94,840
MAE	11982,524	.	11982,524	11982,524
MaxAE	95642,109	.	95642,109	95642,109
Standardized BIC	19,747	.	19,747	19,747

As we can see, its normalized BIC is 19,747 unless the BIC is more interesting and the model will be the one that best estimates the parameters.
It will serve as a decision tool.

Model statistics

Model	Number of predictors	Model fit statistics	Ljung-Box Q(18)			Number of atypical values
		R squared stationary	Statistics	DF	Sig.	
Amount-Model 1	0	,622	20,835	16	,185	1

Ljung-Box test not significant

ARIMA model parameters								
					Estimation	SE	t	Sig.
Amount -Model _1	Importe	Natural logarithm	Difference		1			
			MA	Delay 1	,603	,057	10,603	,000
			Seasonal difference		1			
			Seasonal MA	Delay 1	,879	,067	13,057	,000

All parameters are significantly different from zero, so the model is valid

Atypical values						
			Estimation	SE	t	Sig.
Amount-Model_1	Sep 2019	Aditive	-,769	,187	-4,122	,000

ARIMA(2,1,0) x ARIMA(0,1,1)$_{12}$

Model fit				
Fit Statistic	Mean	SE	Minimum	Maximum
R squared stationary	,613	.	,613	,613
R squared	,877	.	,877	,877
RMSE	18817,046	.	18817,046	18817,046
MAPE	17,997	.	17,997	17,997
MaxAPE	95,614	.	95,614	95,614
MAE	12120,550	.	12120,550	12120,550
MaxAE	96422,650	.	96422,650	96422,650
Normalized BIC	19,790	.	19,790	19,790

Model statistics						
Model	Number of Predictors	Model Fit Statistics	Ljung-Box Q(18)			Number of atypical values
		R squared stationary	Statistics	DF	Sig.	
Amount-Model_1	0	,613	22,726	15	,090	1

ARIMA model parameters								
					Estimation	SE	t	Sig.
Amount-Model_1	Amount	Natural Logarithm	AR	Delay 1	-,593	,070	-8,469	,000
				Delay 2	-,232	,070	-3,323	,001
			Difference		1			
			Seasonal Difference		1			
			Stationary MA	Delay 1	,857	,063	13,504	,000

Atypical values						
			Estimación	SE	t	Sig.
Amount-Model_1	Sep 2019	Aditiv e	-,824	,185	-4,450	,000

ARIMA(0,1,1) x ARIMA(1,1,0)$_{12}$

Ajuste del modelo

Fit Statistic	Mean	SE	Minimum	Maximum
R squared stationary	,556	.	,556	,556
R squared	,869	.	,869	,869
RMSE	19393,969	.	19393,969	19393,969
MAPE	19,367	.	19,367	19,367
MaxAPE	121,997	.	121,997	121,997
MAE	12671,788	.	12671,788	12671,788
MaxAE	80523,448	.	80523,448	80523,448
Normalized BIC	19,824	.	19,824	19,824

Estadísticos del modelo						
Model	Number of Predictors	Model Fit Statistics	Ljung-Box Q(18)			Number of atypical values
		R squared stationary	Statistics	DF	Sig.	
Amount-Model_1	0	,556	25,825	16	,057	1

ARIMA model parameters								
					Estimation	SE	t	Sig.
Amount-Model_1	Amount	Natural Logarithm	Difference		1			
			MA	Delay 1	,654	,055	11,922	,000
			Seasonal AR	Delay 1	-,542	,063	-8,624	,000
			Seasonal Difference		1			

Atypical values						
			Estimation	SE	t	Sig.
Amount-Model_1	Sep 2019	Aditive	-,846	,194	-4,368	,000

ARIMA(2,1,0) x ARIMA(1,1,0)$_{12}$

Model fit				
Fit Statistic	Mean	SE	Minimum	Maximum
R squared stationary	,578	.	,578	,578
R squared	,871	.	,871	,871
RMSE	19422,381	.	19422,381	19422,381
MAPE	18,746	.	18,746	18,746
MaxAPE	87,113	.	87,113	87,113
MAE	12430,303	.	12430,303	12430,303
MaxAE	70344,620	.	70344,620	70344,620
Normalized BIC	19,932	.	19,932	19,932

Model statistics						
Model	Number of Predictors	Model Fit Statistics	Ljung-Box Q(18)			Number of atypical values
		R squared stationary	Statistics	DF	Sig.	
Amount-Model_1	0	,578	21,706	15	,116	2

ARIMA model parameters								
					Estimation	SE	t	Sig.
Amount-Model_1	Amount	Natural Logarithm	AR	Delay 1	-,685	,067	-10,285	,000
				Delay 2	-,385	,067	-5,732	,000
			Difference		1			
			Seasonal AR	Delay 1	-,557	,064	-8,679	,000
			Seasonal Differencee		1			

Atypical values							
				Estimation	SE	t	Sig.
Amount-Model_1	Nov 2018	Transitory	Magnitude	-,613	,162	-3,795	,000
			Decrease factor	,698	,177	3,947	,000
	Ago 2019	Transitory	Magnitude	-,907	,176	-5,161	,000
			Decrease factor	,656	,135	4,866	,000

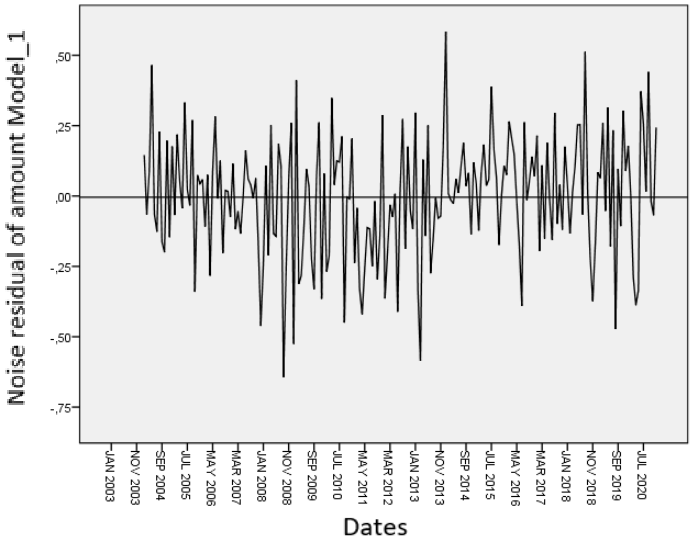
We are going to summarize the models by means of their normalized BIC, and we will carry out the diagnosis for the model with the lowest BIC, since it will be the preferable model.

Model	normalized BIC
ARIMA(0,1,1) x ARIMA(0,1,1)$_{12}$	19,747
ARIMA(2,1,0) x ARIMA(0,1,1)$_{12}$	19,790
ARIMA(0,1,1) x ARIMA(1,1,0)$_{12}$	19,824
ARIMA(2,1,0) x ARIMA(1,1,0)$_{12}$	19,932

As we can see, the model with the lowest normalized BIC is **ARIMA(0,1,1)xARIMA(0,1,1)$_{12}$**, so we are going to carry out the diagnosis for said model.

3.3 Model diagnosis.

Let's start with the Diagnosis, for this let's see the residuals of the model, these have to be stable in variance, uncorrelated, of null mean and that they follow a normality.
Let us see that they are stable in variance.



We can see that the residuals are stable in mean and variance, since the data fluctuate around zero and we can also adjust bands above and below so that the residuals do not leave them.

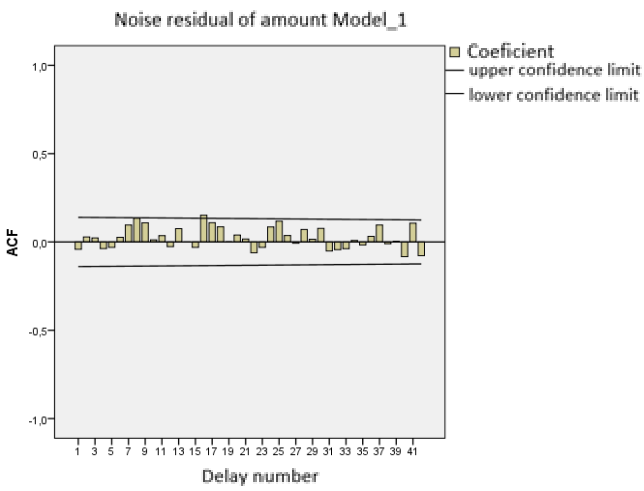
Autocorrelations

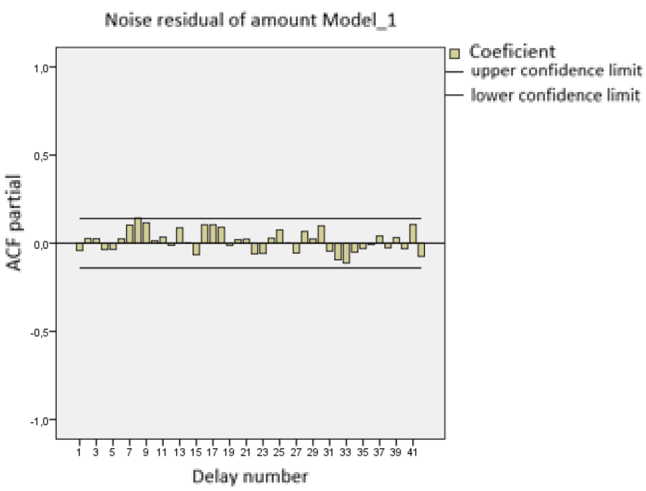
Series: Noise Residue Amount-Model_1

Delay	Autocorrelation	Standar error[a]	Box-Ljung statistic		
			Value	gl	Sig.[b]
1	-,042	,070	,360	1	,548
2	,028	,069	,527	2	,768
3	,023	,069	,636	3	,888
4	-,037	,069	,925	4	,921
5	-,031	,069	1,122	5	,952
6	,026	,069	1,266	6	,974
7	,096	,069	3,227	7	,863
8	,134	,068	7,059	8	,530
9	,108	,068	9,557	9	,388
10	,012	,068	9,586	10	,478
11	,036	,068	9,865	11	,543
12	-,026	,068	10,015	12	,615
13	,075	,068	11,259	13	,589
14	-,001	,067	11,260	14	,666
15	-,030	,067	11,462	15	,719
16	,151	,067	16,538	16	,416
17	,109	,067	19,175	17	,319
18	,086	,067	20,835	18	,288

19	,000	,066	20,835	19	,346
20	,039	,066	21,178	20	,387
21	,016	,066	21,239	21	,444
22	-,062	,066	22,111	22	,453
23	-,031	,066	22,330	23	,500
24	,085	,066	24,021	24	,460
25	,118	,065	27,250	25	,343
26	,036	,065	27,559	26	,381
27	-,007	,065	27,571	27	,433
28	,071	,065	28,761	28	,425
29	,015	,065	28,817	29	,475
30	,077	,064	30,226	30	,454
31	-,051	,064	30,849	31	,474
32	-,043	,064	31,304	32	,502
33	-,038	,064	31,658	33	,534
34	,008	,064	31,675	34	,582
35	-,017	,064	31,747	35	,626
36	,032	,063	31,998	36	,659
37	,096	,063	34,294	37	,597
38	-,010	,063	34,319	38	,640
39	,004	,063	34,324	39	,683
40	-,083	,063	36,093	40	,647
41	,106	,062	38,993	41	,560
42	-,078	,062	40,548	42	,535
a. The assumed underlying process is independence (white noise).					
b. It is based on the asymptotic chi-square approximation.					

We can see that all the p-values are greater than 0.1, so the residuals are uncorrelated. Indeed, if we look at their fas and fap, we can see that there are no significant appearances that they are different from zero.





Let's see if it has zero mean and normality.

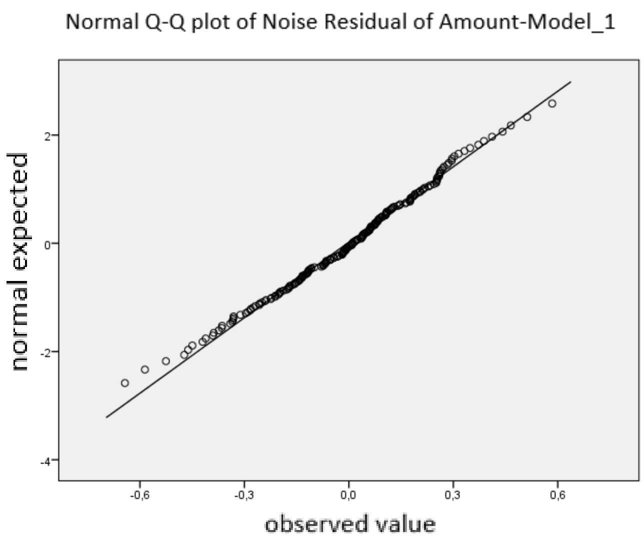
Descriptives

			Statistical	Standard error
Noise residue from Amount-Model_1	Mean		,00	,015
	95% confidence interval for the mean	Inferior limit	-,03	
		Superior limit	,03	
	Media cropped at 5%		,00	
	Median		,01	
	Variance		,046	
	Standard deviation		,215	
	Minimum		-1	
	Maximum		1	
	Range		1	
	Interquartile range		0	
	Asymmetry		-,226	,171
	Curtosis		,092	,340

Normality tests

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistical	gl	Sig.	Statistical	gl	Sig.
Noise residue from Amount-Model_1	,058	203	,093	,994	203	,641

Indeed, we can admit that the residuals have a null mean, since 0 is included in the 95% confidence interval for the mean and since the result of the normality tests is not significant (taking a p-value less than 0.05 as significant) we also admit the normality of the residuals.



3.4. Overfitting.

Now let's see if the models:

- **ARIMA(1,1,1) x ARIMA(0,1,1)$_{12}$**
- **ARIMA(0,1,2) x ARIMA(0,1,1)$_{12}$**
- **ARIMA(0,1,1) x ARIMA(1,1,1)$_{12}$**
- **ARIMA(0,1,1) x ARIMA(0,1,2)$_{12}$**

They are viable and if they are, let's see if it is better than the one we already have (**ARIMA(0,1,1)xARIMA(0,1,1)$_{12}$**).

Let's start with the **ARIMA(1,1,1) x ARIMA(0,1,1)$_{12}$**.

ARIMA model parameters								
					Estimation	SE	t	Sig.
Amount-Model_1	Amount	Natural logarithm	AR	Delay 1	-,078	,118	-,664	,507
			Difference		1			
			MA	Delay 1	,558	,097	5,730	,000
			Seasonal Difference		1			
			MA, seasonal	Delay 1	,873	,066	13,197	,000

As we can see the AR1 component is not significantly different from zero, so the model is not valid.

Now let's see the model **ARIMA(0,1,2) x ARIMA(0,1,1)$_{12}$**

ARIMA model parameters								
					Estimation	SE	t	Sig.
Amount-Model_1	Amount	Natural logarithm	Difference		1			
			MA	Delay 1	,634	,072	8,810	,000
				Delay 2	-,044	,072	-,617	,538
			Seasonal difference		1			
			MA, seasonal	Delay 1	,873	,066	13,172	,000

The MA2 component is not significantly different from zero, so the model is invalid.

Now let's see the model **ARIMA(0,1,1) x ARIMA(1,1,1)$_{12}$**

ARIMA model parameters								
					Estimation	SE	t	Sig.
Amount-Model_1	Amount	Natural logarithm	Difference		1			
			MA	Delay 1	,602	,058	10,370	,000
			AR, seasonal	Delay 1	-,019	,093	-,206	,837
			Seasonal difference		1			
			MA, seasonal	Delay 1	,866	,078	11,154	,000

The SAR1 component is not significant, so the model is invalid.

Finally let's see the model **ARIMA(0,1,1) x ARIMA(0,1,2)$_{12}$**.

ARIMA model parameters								
					Estimation	SE	t	Sig.
Amount-Model _1	Amo unt	Natural logarithm	Difference		1			
			MA	Delay 1	,602	,05 8	10,382	,00 0
			Seasonal difference		1			
			MA, seasonal	Delay 1	,883	,08 2	10,816	,00 0
				Delay 2	-,012	,08 1	-,143	,88 7

The SMA2 component is not significantly different from zero, so the model is invalid.

Therefore, the best model is **ARIMA(0,1,1) x ARIMA(0,1,1)$_{12}$** and it is with him that we will make the predictions.

We then recall the model coefficients and the outlier:

ARIMA(0,1,1) x ARIMA(0,1,1)$_{12}$					
Estimation			SE	t	Sig.
MA	Delay 1	,603	,057	10,603	,000
MA, seasonal	Delay 1	,879	,067	13,057	,000

Atypical values						
			Estimation	SE	t	Sig.
Amount-Model_1	Sep 2019	Aditiv o	-,769	,187	-4,122	,000

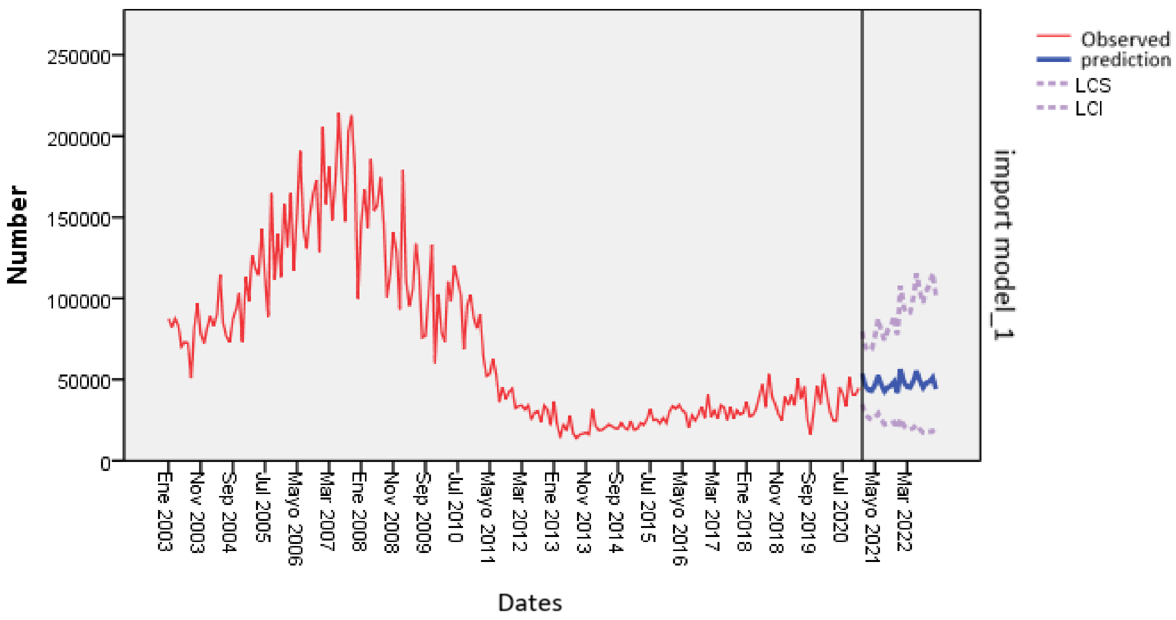
And the model equation is:

$$\log log\left(X_t\right) =- 0,769I_t^{81} + \frac{(1-0,603B)(1-0,879B^{12})}{(1-B)(1-B^{12})}Z_t$$

4. Predictions.

4.2. ARIMA Model .

Now let's see the predictions of the ARIMA(0,1,1) x ARIMA(0,1,1)12 model:



Date	Prediction
JAN 2021	53834
FEB 2021	46119
MAR 2021	43493
APR 2021	43020
MAY 2021	46713
JUN 2021	52474
JUL 2021	46749
AUG 2021	42749
SEP 2021	45439
OCT 2021	46206
NOV 2021	48676
DEC 2021	41894
JAN 2022	56370
FEB 2022	48335
MAR 2022	45623
APR 2022	45167
MAY 2022	49088
JUN 2022	55191
JUL 2022	49214
AUG 2022	45043
SEP 2022	47920
OCT 2022	48773
NOV 2022	51426
DEC 2022	44301

5. Conclusion.

In summary, we have analyzed a time series with the learning obtained in class and we have been able to comment on predictions about it.
Also as a curiosity you can see perfectly the effect that the crisis produced in Spain.
In the first place we have studied the components of our series, then we have carried out a study to see if it is stationary and the necessary transformaions to obtain a stationary series from our original series.
With the series already stationary, we have fitted a model of the ARIMA family to the data, adding the study of outliers to it.
Finally, after carrying out the ARIMA method, we have seen the predictions and we can see that in the coming years we will experience a slight increase in the amount of mortgages in Extremadura.