Box-Jenkins methodology on Barcelona's subway passengers

Luis Rojo-González*

June 18, 2020

1 Time series description

Barcelona's subway passengers (thousands of commuters). Monthly data.[1]

Source: Ministerio de Fomento http://www.ine.es/jaxiT3/Tabla.htm?t=20193

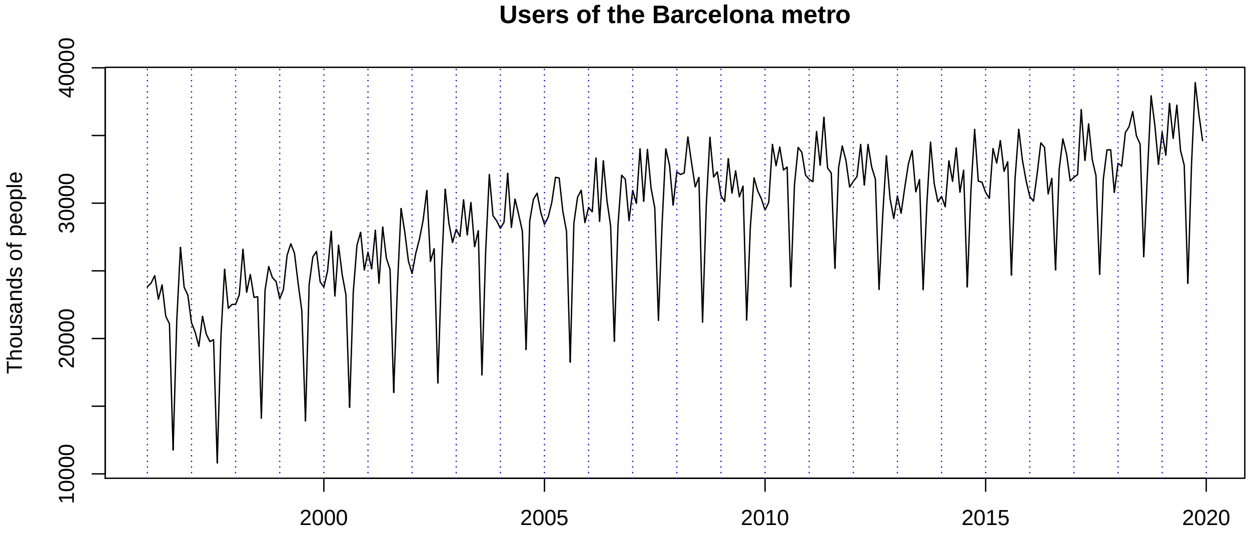


Figure 1: Time serie of Barcelona's subway passengers.

2 Identification

2.1 Stationarity and unit roots analysis

Figure 2 shows there are differences on the variance of the series (see y-axis of scatterplot), so it is conveniently to apply the logarithm to stabilize it.

Applying the logarithm conjoint a seasonal differentiation and up to two regular differentiations to explore the stationarity of the series we get the (Partial)Auto-Correlation Function plots shown in Figure 4, where all of them has short variances, but the first serie (logserie with seasonal differentiation) looks like the best one, but it and the serie with one regular differentiation are so close. Also, we have to note that there are a couple of outliers.

*Universitat Politècnica de Catalunya, luis.rojo.g@usach.cl

[1] You can find the code in https://drive.google.com/drive/folders/1h5FAs--aJiAMflkE0qBRh1gtO32NR4wf?usp=sharing

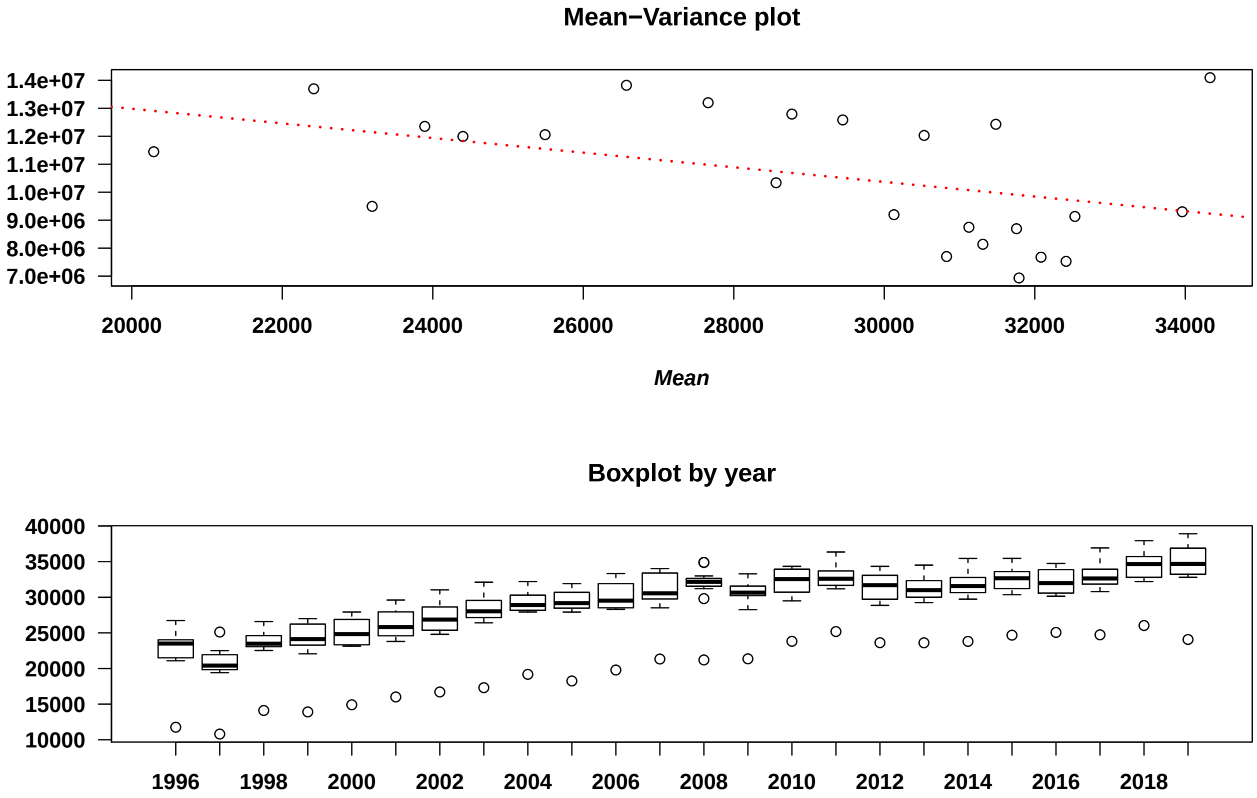


Figure 2: Boxplot and mean-variance plot for anually data.

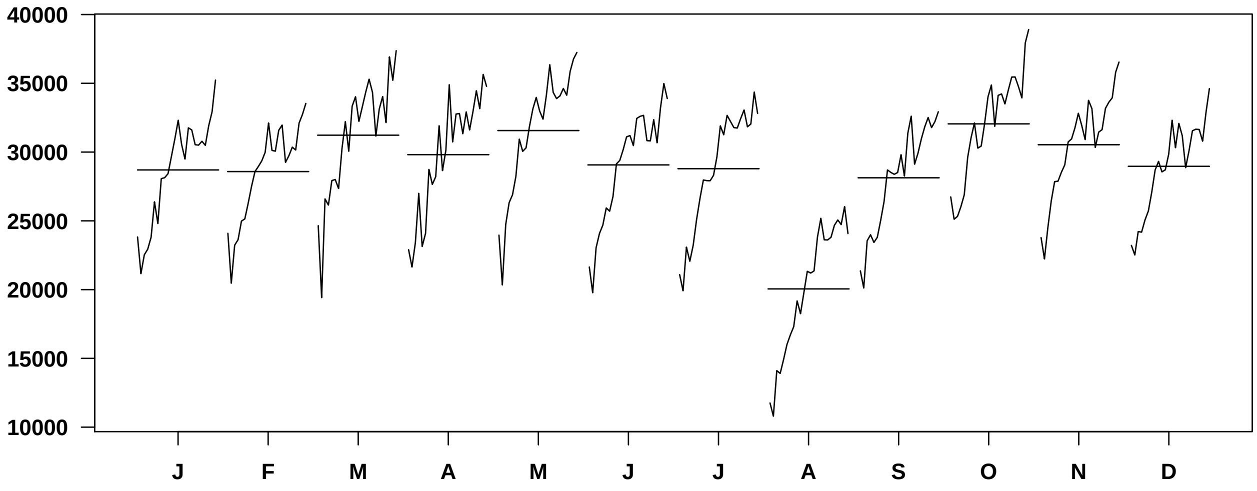
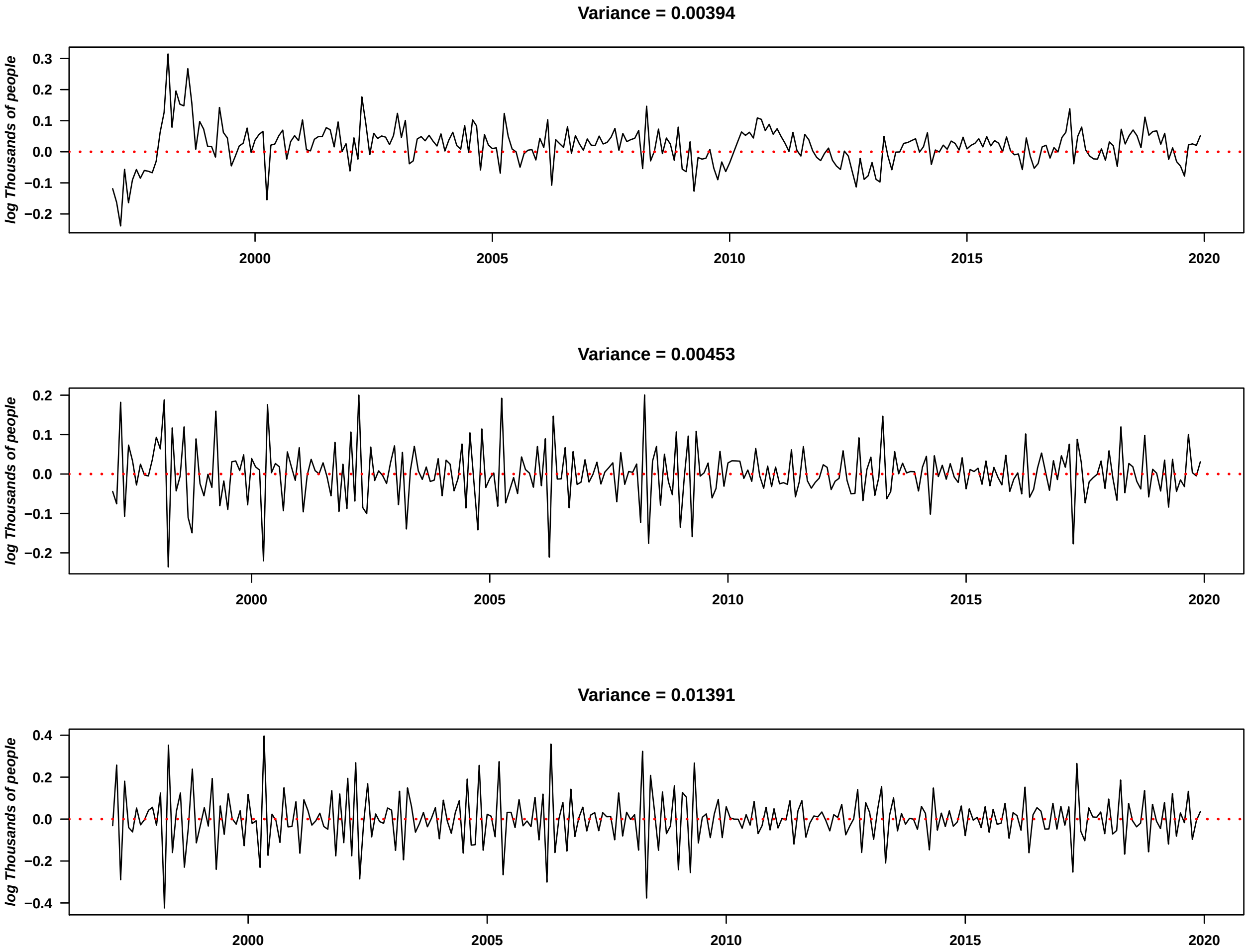


Figure 3: Monthplot.


Variance = 0.00394

Variance = 0.00453

Variance = 0.01391

Figure 4: Differentiated time series plot.

The (Partial)Auto-Correlation function plots for these series seem to indicate the first differentiation (seasonal) would be enough to work with in terms of stationarity such as Figure 5 shows. Nonetheless, this conclusion at first glance might be wrong, so we perform a Ljung-Box test as a more formal tool to see whether the time series has a unit root (null hypothesis) or are stationary (alternative hypothesis).

Obtained results for Ljung-Box test (their p-values) are in Table 1 where we can see all of time series (original and differentiated ones) would be jointly independent, which is a good property towards models chosen[2].

Table 1: Ljung-Box test results for original and differentiated log-time series.

Lag	p.value (d12lnserie)	p.value (d1d12lnserie)	p.value (d1d1d12lnserie)
4	0.0000	0.0000	0.0000
8	0.0000	0.0000	0.0000
12	0.0000	0.0000	0.0000
16	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.0000

Formal tests to prove stationarity are given by Dickey-Fuller (DF), Phillips-Perron (PP) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS); where the presence of unit roots (null hipothesis) against stationarity (alternative hypothesis) are tested for first two tests whereas the last one uses the stationarity as null hypothesis. Table 2 shows p-values for each performed test, where it is clearly to see logged series has an unit root, but from first lag (d12lnserie) we get a stationary time series, thus we can claim that the time series has an unit root and it is stationary at first lag.

Table 2: P-values of Dickey-Fullar, Phillips-Perron and Kwiatkowski-Phillips-Schmidt-Shin tests.

	d12lnserie	d1d12lnserie	d1d1d12lnserie
DickeyFuller	0.0100	0.0100	0.0100
PhillipsPerronShort	0.0100	0.0100	0.0100
PhillipsPerronLong	0.0100	0.0100	0.0100
KPSSShort	0.1000	0.1000	0.1000
KPSSLong	0.1000	0.1000	0.1000

2.2 Descriptive statistics

Some basic descriptive statistics show us the time series we are working with is not normal and has leptokurtic distribution (see Table 3) which is supported by the Jarque-Bera test, also such as Figure 6 shows we can recognize there is an important influent outlier in the time series.

Table 3: Descriptive statistics of the differentiated log-time series.

Min	Q1	Mean	Median	Q3	Sd	Skewnesss	Kurtosis	JarqueBera
-0.24	-0.01	0.02	0.02	0.05	0.06	0.16	6.24	0.00

3 Model for the mean

3.1 Identification

As we saw in Section 2.1, in particular in Figure 2, there are differences on the variance so the logarithm must be applied, then one seasonal difference was applied to find such stationary time series to work with.

[2] Keep in mind that d12ln notation represents differentiation and logarithm transformation.

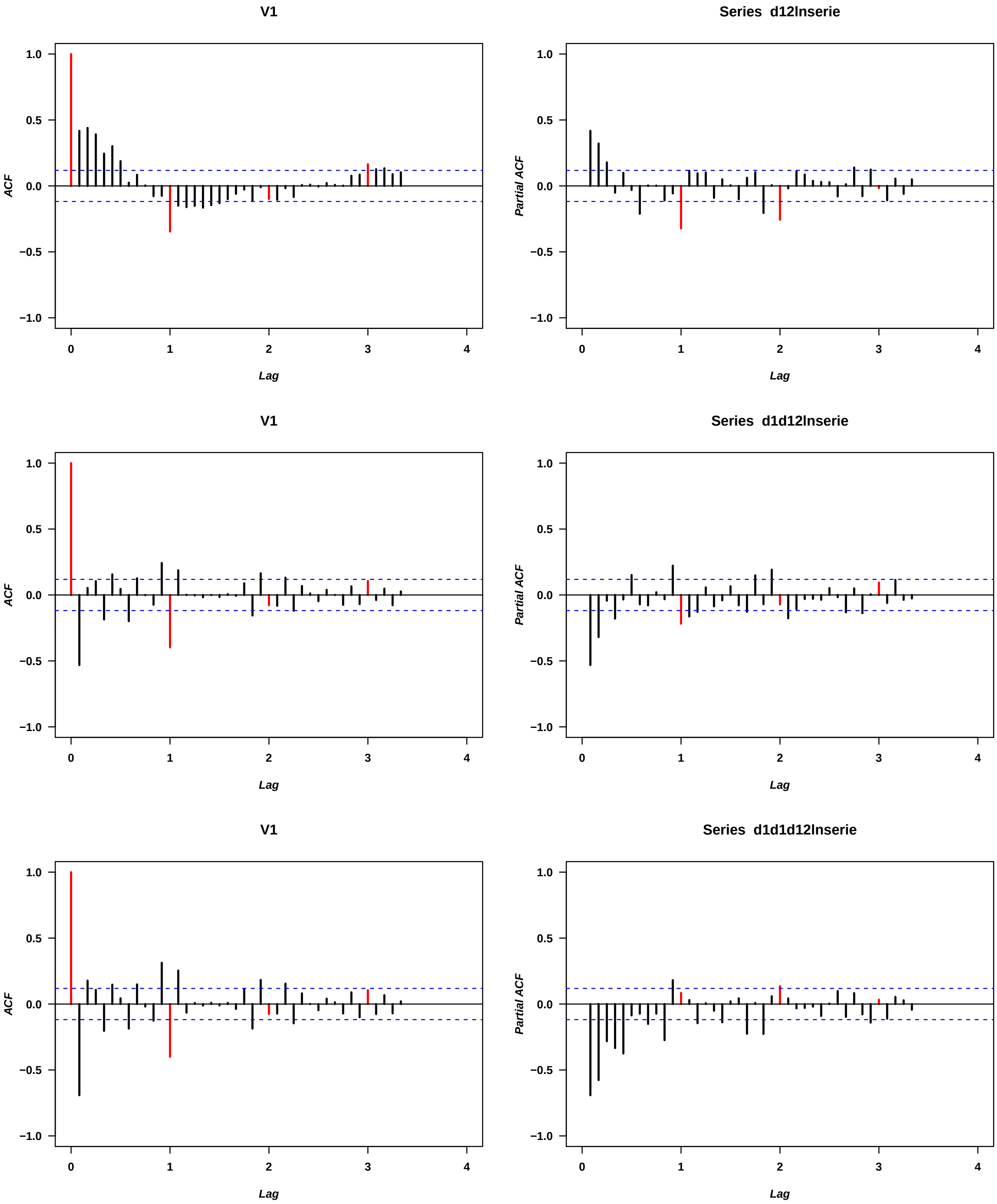


Figure 5: ACF and PACF plots for original time series and differentiated log-time series.

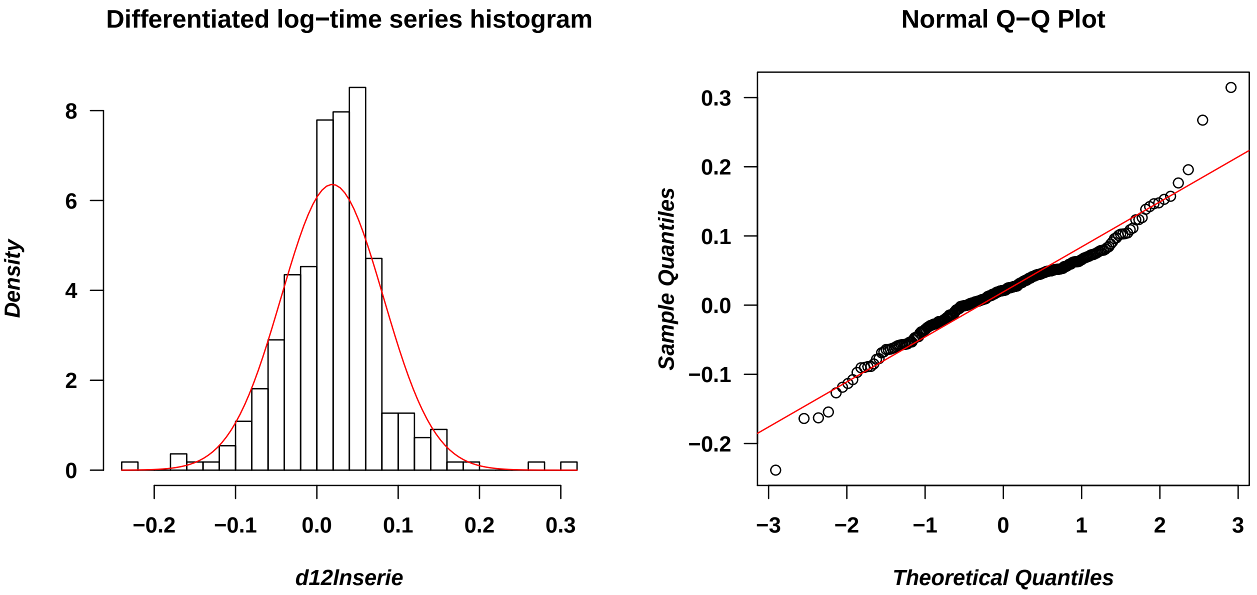


Figure 6: Histogram and QQ-plot for normality assessment.

To help the decision making we take into account the test performed above as well as (Partial)Auto-Correlation Function plots (see Figure 5) which show there would be two *suitable* models: i) SARIMA$(3,0,0)_{(1,1,1)_{12}}$ and ii) SARIMA$(3,0,4)_{(1,1,1)_{12}}$.

3.2 Estimation

The first model to work with is a SARIMA$(3,0,0)_{(1,1,1)_{12}}$, which gives us that the SAR estimation is non-significative so by discarding this we get a SARIMA$(3,0,0)_{(0,1,1)_{12}}$ represented by equation (1). This model looks good in term of the AIC = -3.1722, AICc = -3.1715 and BIC = -3.0966.

$$(1-0.2423_{(0.00)}B-0.2837_{(0.00)}B^2-0.2484_{(0.00)}B^3)(1-B^{12})x_t = 0.0016_{(0.00)} + (1-0.5911_{(0.00)}B^{12})z_t \quad (1)$$

The first model to work with is a SARIMA$(3,0,4)_{(1,1,1)_{12}}$, which gives us some parameters are non-significative so by discarding them on a sequential way we get a SARIMA$(1,0,2)_{(0,1,1)_{12}}$ represented by equation (2). This model looks good in term of the AIC = -3.1620, AICc = -3.1613 and BIC = -3.0864.

$$(1-0.8973_{(0.00)}B)(1-B^{12})x_t = 0.0016_{(0.00)} + (1-0.6619_{(0.00)}B+0.1403_{(0.04)}B^2)(1-0.5945_{(0.00)}B^{12})z_t \quad (2)$$

3.3 Diagnosis

In this section we check the model assumptions taking into account residuals must be gaussian distributed, it means that residuals have zero-mean and constant variance as well as jointly independent white noise considering lags and their π and ψ weights for stationarity (causality) and invertibility, also we check their stability dropping-out some last observations as re-estimating the coefficients.

The first model to check is the SARIMA$(3,0,0)_{(0,1,1)_{12}}$ which such as Figure 8 shows looks good in terms of zero-mean and constant variance as well as (Partial)Auto-Correlation Function plots cut-off after

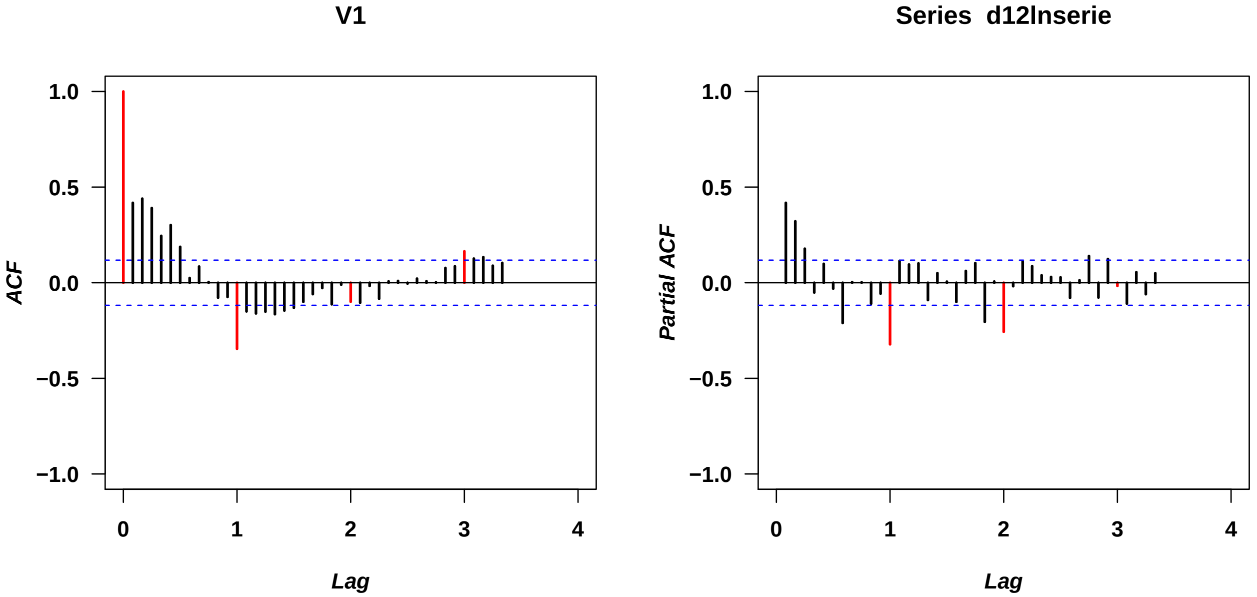


Figure 7: ACF and PACF plots for differentiated log-time series.

first lag and inmediately, respectively. Also, they seem to be normal distributed although some outliers, but are not jointly independent after lag number eight.

To check wheter the model is stationary (causal) and invertible, we rewrite it in their AR(∞) and MA(∞) form, which give us ψ and π-weights. These weights, which are shown in Table 4, are lower than the unit and the root of the model is larger than one, so we can claim there is not invertibility problems and also it is causal.

Modul of AR Characteristic polynomial Roots: 1.133 1.884967 1.884967

Modul of MA Characteristic polynomial Roots: 1.044794 1.044794 1.044794 1.044794 1.044794 1.044794 1.044794 1.044794 1.044794 1.044794 1.044794 1.044794

Table 4: Weight of SARIMA(3, 0, 0)(0, 1, 1)12 as AR and MA infinity.

	1	2	3	4	5	6	7	8	9	10
ψ_j	0.2423	0.3424	0.4001	0.2543	0.2602	0.2346	0.1938	0.1781	0.1564	0.1366
π_j	0.2423	0.2837	0.2484	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000

Finally, the stability assessment is performed using the last 12 observations (to replicate a year). We see that the fitted models, represented in equation (3), has similar estimated coefficients, both significatives, and also similar AIC (-3.1722 and -3.1558), AICc (-3.1715 and -3.1550) and BIC (-3.0966 and -3.0778) values, thereby we can say this model is stable.

$$(1-0.2312_{(0.00)}B-0.2881_{(0.00)}B^2-0.2628_{(0.00)}B^3)(1-B^{12})x_t=0.0016_{(0.00)}+(1-0.5912_{(0.00)}B^{12})z_t \quad (3)$$

The second model to check is the SARIMA$(1,0,2)_{(0,1,1)_{12}}$ which such as Figure 9 shows looks good in terms of zero-mean and constant variance as well as (Partial)Auto-Correlation Function plots cut-off after first lag and inmediately, respectively. Also, they seem to be normal distributed although some outliers, but are not jointly independent after lag number six.

To check wheter the model is stationary (causal) and invertible, we rewrite it in their AR(∞) and MA(∞) form, which give us ψ and π-weights. These weights, which are shown in Table 5, are lower than the unit and

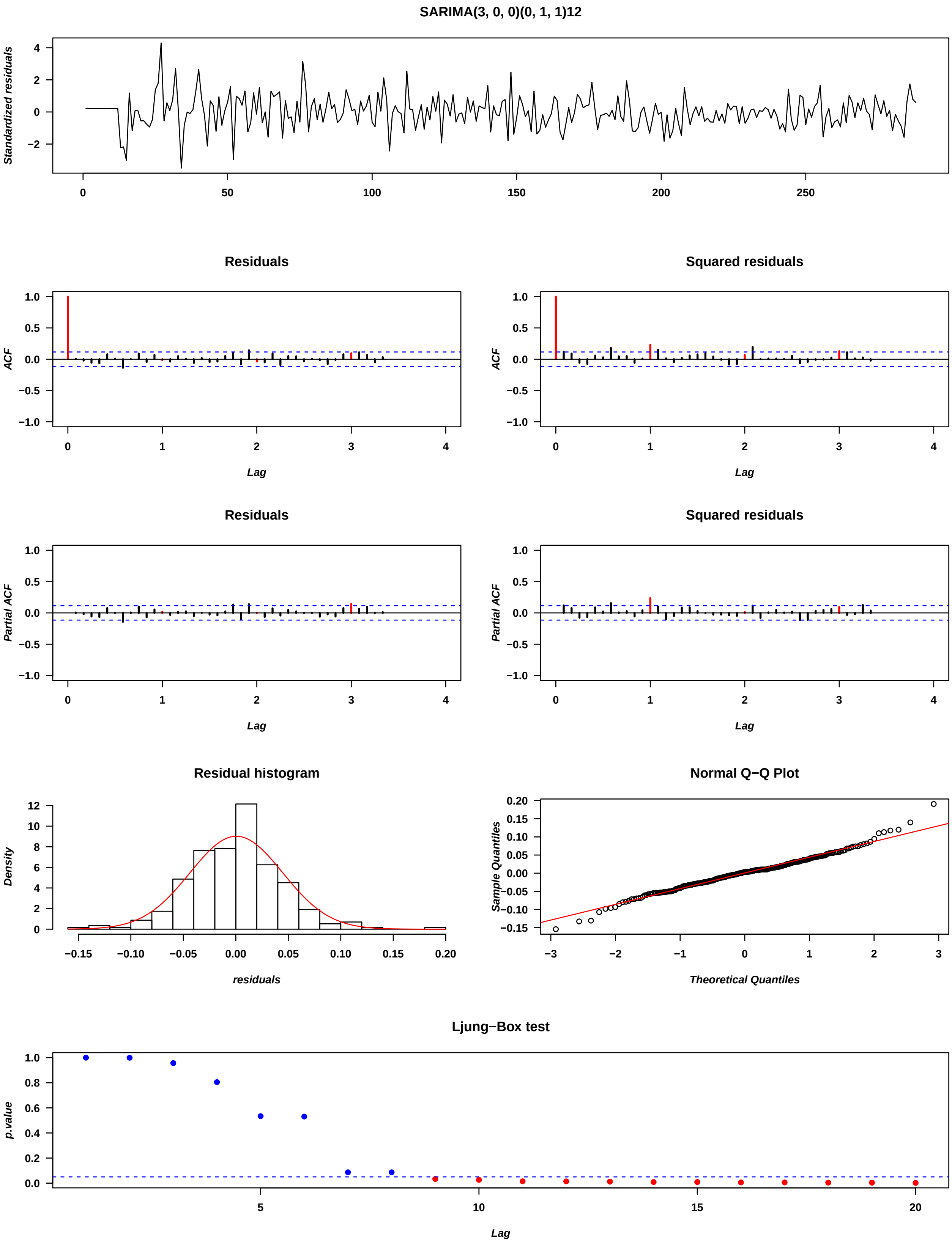


Figure 8: SARIMA(3, 0, 0)(0, 1, 1)12 residuals’ analysis.

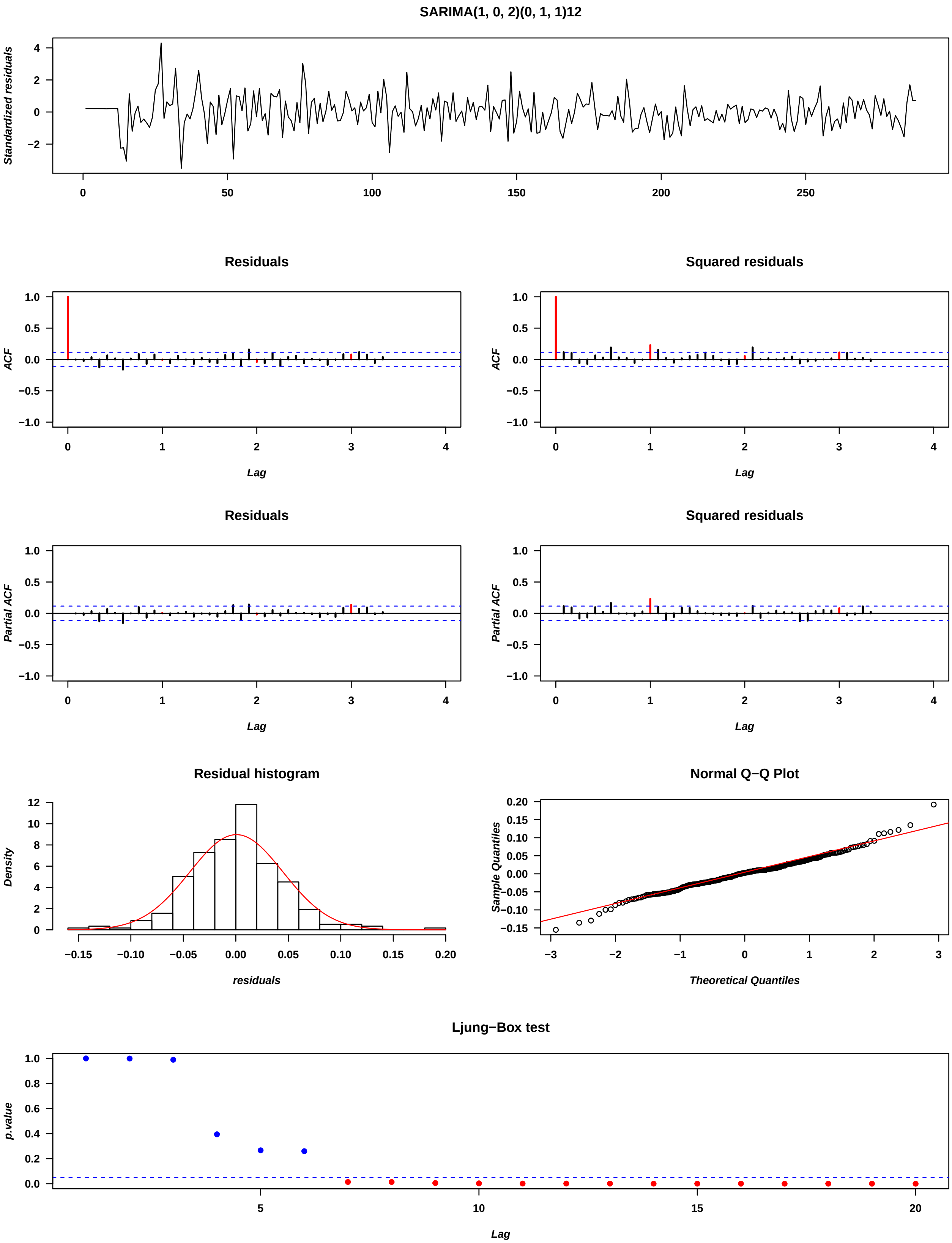


Figure 9: SARIMA(1, 0, 2)(0, 1, 1)12 residuals' analysis.

the root of the model is larger than one, so we can claim there is not invertibility problems and also it is causal.

Modul of AR Characteristic polynomial Roots: 1.114403

Modul of MA Characteristic polynomial Roots: 1.044296 1.044296 1.044296 1.044296 1.044296 1.044296 1.044296 1.044296 1.044296 1.044296 1.044296 1.044296 2.669968 2.669968

Table 5: Weight of SARIMA(1, 0, 2)(0, 1, 1)12 as AR and MA infinity.

	1	2	3	4	5	6	7	8	9	10
ψ_j	0.2355	0.3516	0.3155	0.2831	0.2540	0.2280	0.2046	0.1836	0.1647	0.1478
π_j	0.2355	0.2961	0.1630	0.0663	0.0210	0.0046	0.0001	-0.0006	-0.0004	-0.0002

Finally, the stability assessment is performed using the last 12 observations (to replicate a year). We see that the fitted models, represented in equation (4), has similar estimated coefficients, both significatives, and also similar AIC (-3.1620 and -3.1455), AICc (-3.1613 and -3.1447) and BIC (-3.0864 and -3.0675) values, thereby we can say this model is stable.

$$(1-0.9054_{(0.00)}B)(1-B^{12})x_t = 0.0016_{(0.00)} + (1-0.6846_{(0.00)}B+0.1479_{(0.03)}B^2)(1-0.5941_{(0.00)}B^{12})z_t \quad (4)$$

We have seen both models have good properties and behaviours, nevertheless their residuals are not jointly independent further than eight and six lags, respectively. Also, we know that the AIC for $\text{SARIMA}(3,0,0)_{(0,1,1)_{12}}$ and $\text{SARIMA}(1,0,2)_{(0,1,1)_{12}}$ are -3.1722 and -3.1620, respectively, which would enable us to choose in favor of $\text{SARIMA}(3,0,0)_{(0,1,1)_{12}}$ model; so, both models has good properties to work with and any of them could be a good choise, but the criterias support a bit more in favor of first fitted model.

3.4 Forecasting

As we know the most suitable model among fitted ones corresponds to $\text{SARIMA}(3,0,0)_{(0,1,1)_{12}}$. Thus, if we consider this model to get forecasting during the next period, that is to say for the next 12 months we have that curve such as Figure 10 shows including with ± 2 prediction error bounds, and Table 6 shows these values.

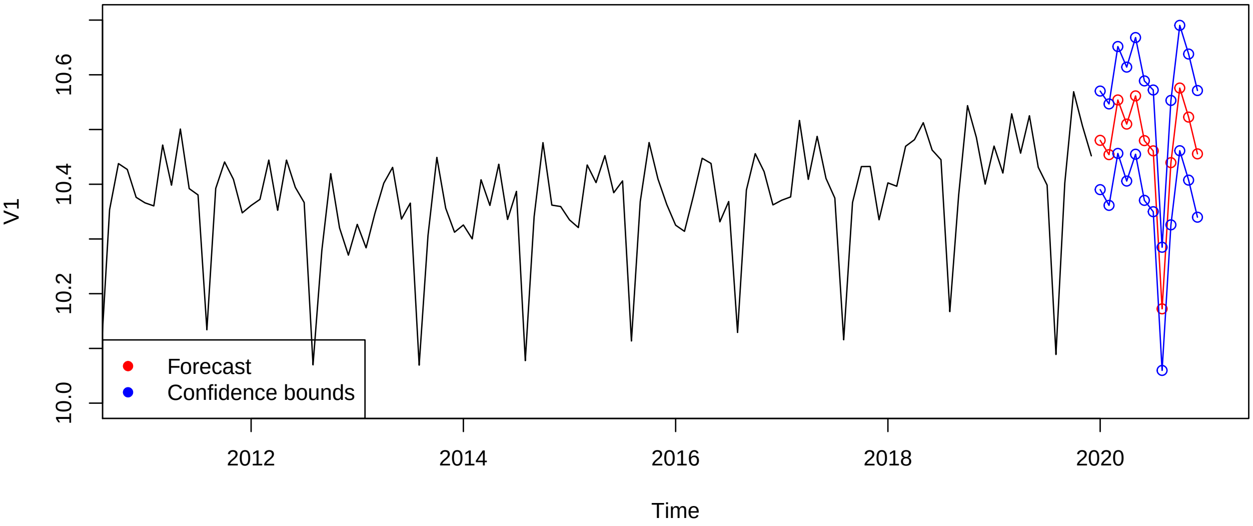


Figure 10: Forecasting for the next 12 months.

Table 6: Forecasting for next 12 months.

	1	2	3	4	5	6	7	8	9	10	11	12
$\log(x_t)$	10.48	10.45	10.55	10.51	10.56	10.48	10.46	10.17	10.44	10.58	10.52	10.46
$+2se$	10.57	10.55	10.65	10.61	10.67	10.59	10.57	10.28	10.55	10.69	10.64	10.57
$-2se$	10.39	10.36	10.46	10.41	10.45	10.37	10.35	10.06	10.33	10.46	10.41	10.34

4 Outlier Treatment

Given the prior knowledge on this time series, it is natural to think on calendar effects are presents, e.g. weekends and non-labour days. On the other hand, given the above results on residual where some outliers are shown in residual qqplot (see Figure 8) ouliers treatment must be carried out.

4.1 Calendar effects

By estimating a model using both trading days and eastern days we get that these coefficients are significative almost at any confidence level; thus, each linearized time series is such as Figure 11 shows, where we can clearly see the time series does not change at all despite of we applied these kind of outliers treatment. Even so, it is a starting point toward the automatic detection of outliers as we shows in the next Section.

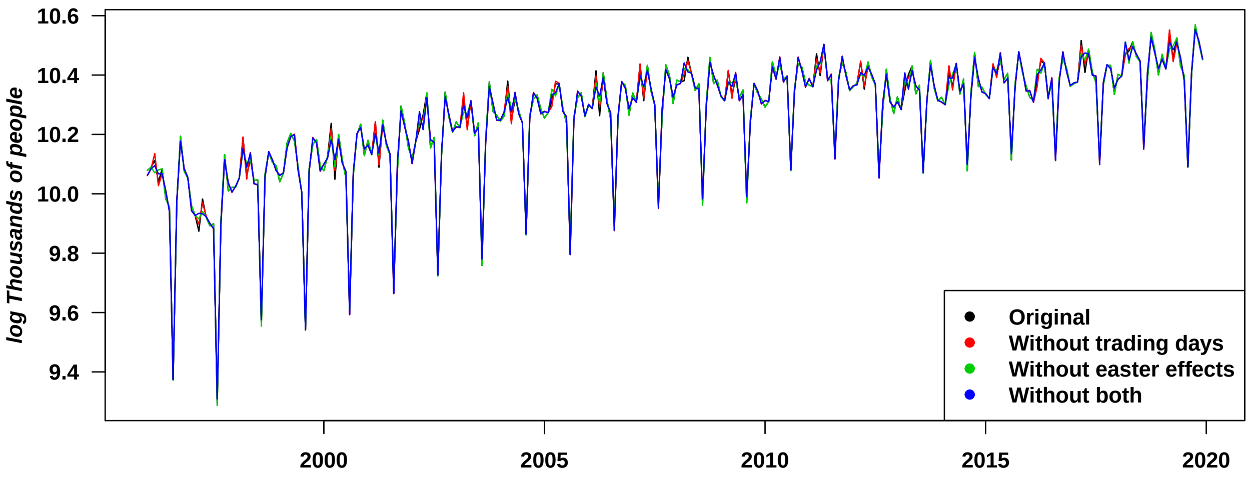


Figure 11: Linearized time series.

4.2 Automatic detection of outliers

Once we have obtained a first approach desired linearized time series it is turn to detect other possible outliers we should take into account. On this way, we consider three kind of outliers: i) Level Shift, ii) Additive Outlier and iii) Transitory Change. If we do so considering a criteria equal to 2.5 we get the linearized time series and the residual by considering it as Figure 12 shows, also Table 7 shows the identified outliers and the kind as they are considered.

Now, as we have a linearized time series a model identification must be doing again. Such as we saw above, the first seasonal differentiation was enough to obtain a stationary time series, so we expect that the same occur here. Thus, such as Figure 13 shows the identified model is the same as before, that is to say it is a SARIMA$(3,0,0)_{(0,1,1)_{12}}$, where its mathematical expression is as equation (5) shows. This model looks good in term of the AIC = -4.4874 and the AICc = -4.4866 and BIC = -4.4117.

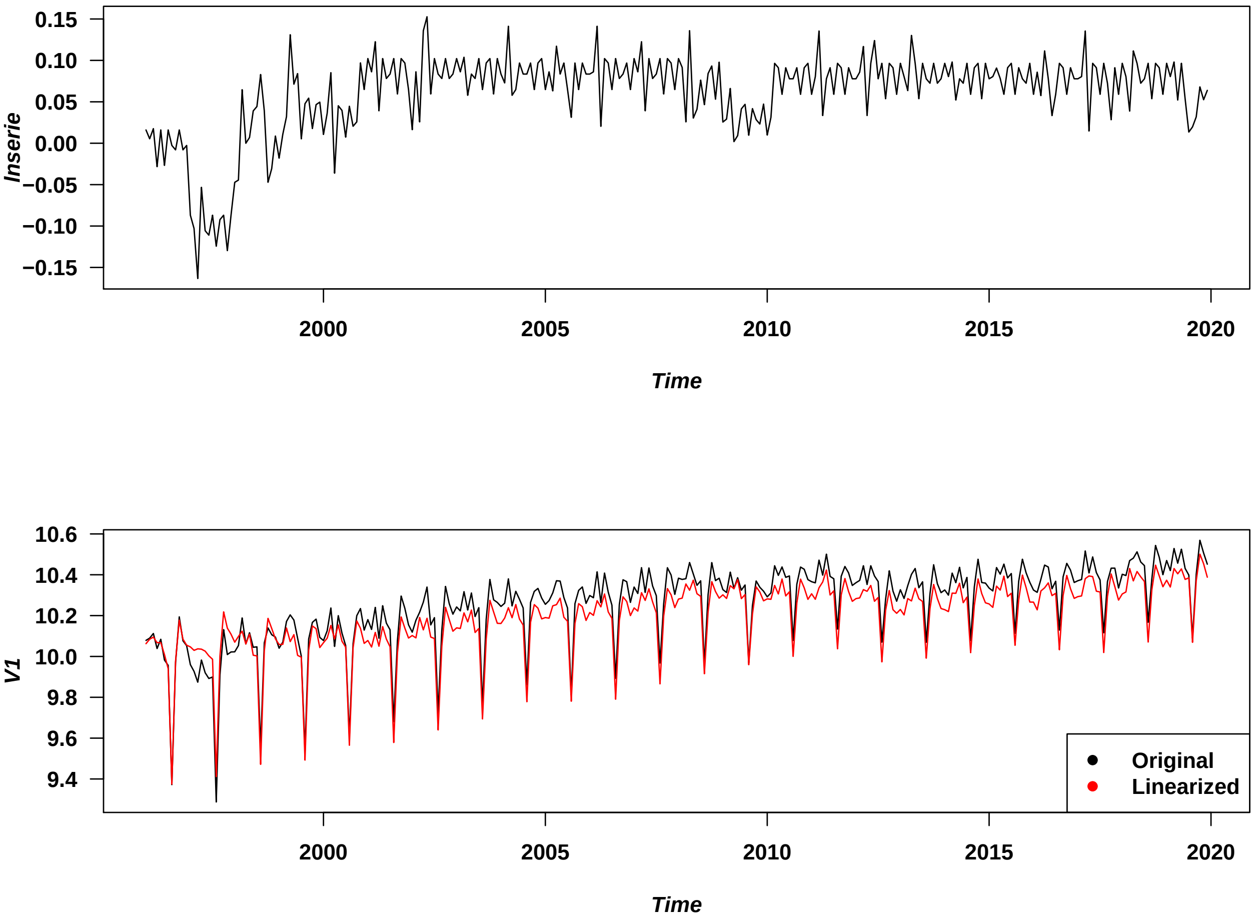


Figure 12: Linearized time series adding outlier detection.

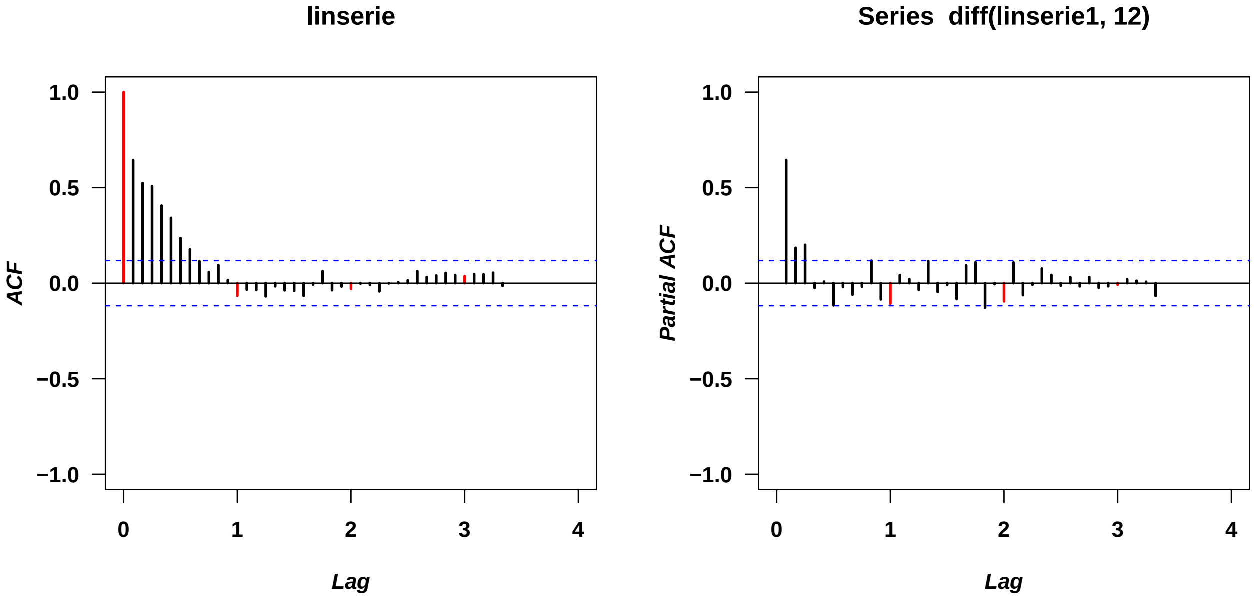


Figure 13: ACF and PACF plots for linearized log-time series.

Table 7: Detected ouliers. Level Shift (LS), Additive Outlier (AO) and Transitory Change (TC).

Obs	type_detected	W_coeff	ABS_L_Ratio
13	LS	-0.10	4.13
25	LS	0.06	2.80
27	LS	0.07	3.27
32	AO	0.08	3.62
34	TC	-0.07	3.22
40	TC	0.10	4.00
43	AO	-0.05	2.81
59	LS	0.06	2.81
73	AO	-0.09	4.00
77	AO	0.05	2.81
116	AO	-0.07	3.45
149	TC	-0.05	2.70
157	LS	-0.05	2.81
171	LS	0.05	2.62
198	AO	0.05	2.83
207	AO	0.04	2.58
246	AO	-0.06	3.00
262	AO	-0.05	2.81
283	TC	-0.08	3.15

$$(1-0.4749_{(0.00)}B-0.1162_{(0.08)}B^2-0.2036_{(0.00)}B^3)(1-B^{12})x_t=0.0014_{(0.00)}+(1-0.1790_{(0.01)}B^{12})z_t \quad (5)$$

4.3 Diagnosis

On the other hand, Figure 14 shows the model validation stage where a clearly improvement is obtained related to qqplot and Ljung-Box test, where in this last test we get a jointly independent residuals until 15 lag against the eight lags obtained before. Also this model is both causal and invertible due to ψ and π weights shown in Table 8 are lower than 1.

Modul of AR Characteristic polynomial Roots: 1.143877 2.072093 2.072093

Modul of MA Characteristic polynomial Roots: 1.15416 1.15416 1.15416 1.15416 1.15416 1.15416 1.15416 1.15416 1.15416 1.15416 1.15416 1.15416

Table 8: Weight of SARIMA(3, 0, 0)(0, 1, 1)12 as AR and MA infinity.

	1	2	3	4	5	6	7	8	9	10
ψ_j	0.4749	0.3417	0.4211	0.3364	0.2782	0.2570	0.2228	0.1923	0.1695	0.1482
π_j	0.4749	0.1162	0.2036	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000	-0.0000

Finally, the stability assessment is performed using the last 12 observations (to replicate a year). We see that the fitted models, represented in equation (6), has similar estimated coefficients, both significatives, and also similar AIC (-4.4874 and -4.4752), AICc (-4.4866 and -4.4744) and BIC (-4.4117 and -4.3972) values, thereby we can say this model is stable.

$$(1-0.4657_{(0.00)}B-0.1156_{(0.08)}B^2-0.2164_{(0.00)}B^3)(1-B^{12})x_t=0.0013_{(0.01)}+(1-0.1593_{(0.01)}B^{12})z_t \quad (6)$$

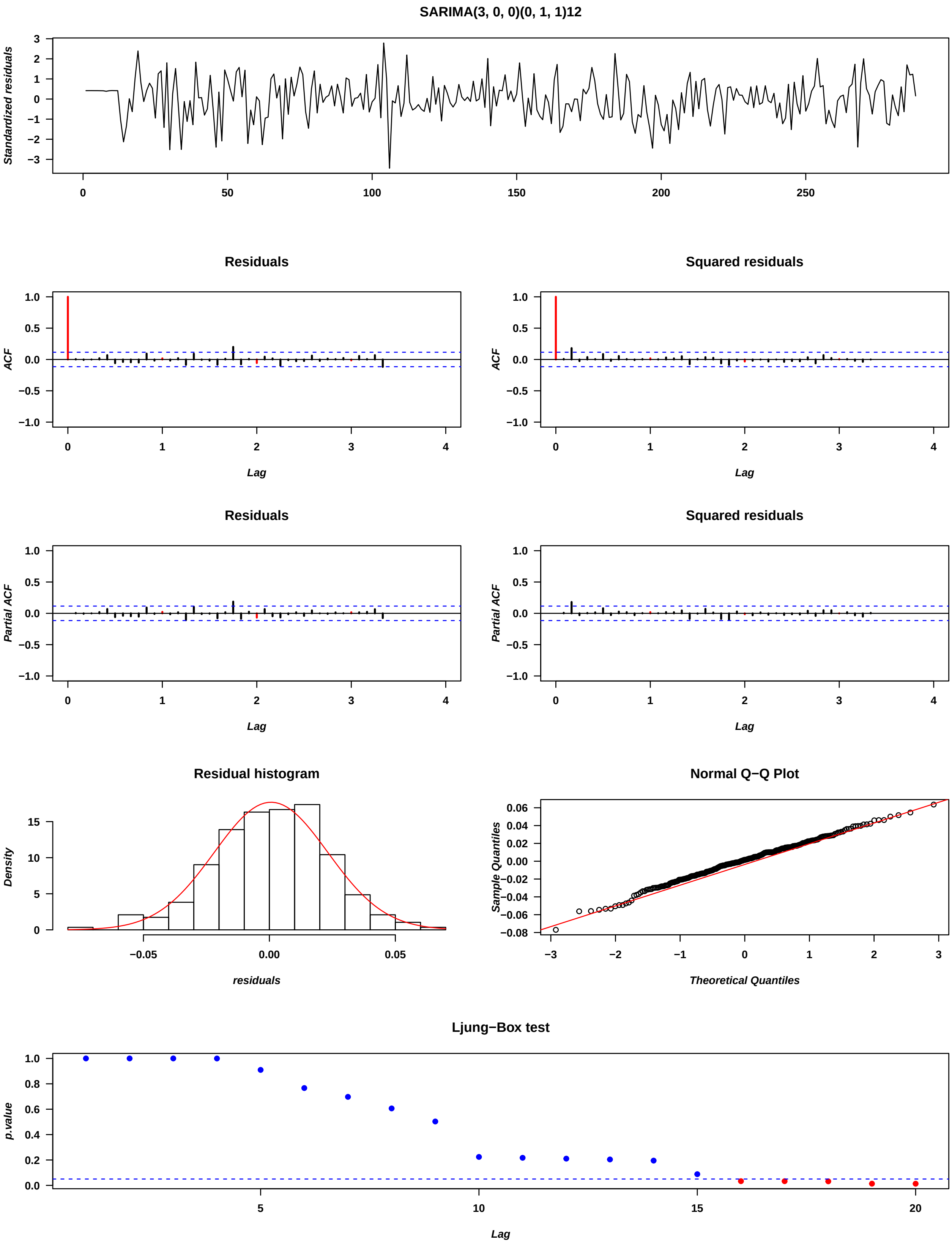


Figure 14: SARIMA(3, 0, 0)(0, 1, 1)12 residuals' analysis.

4.4 Forecasting

Following with the proposed methodology, but now considering the linearized time series we get the forecasting for the next 12 months such as Figure 15 shows conjoint the obtained confidence intervals where the values are shown in Table 9.

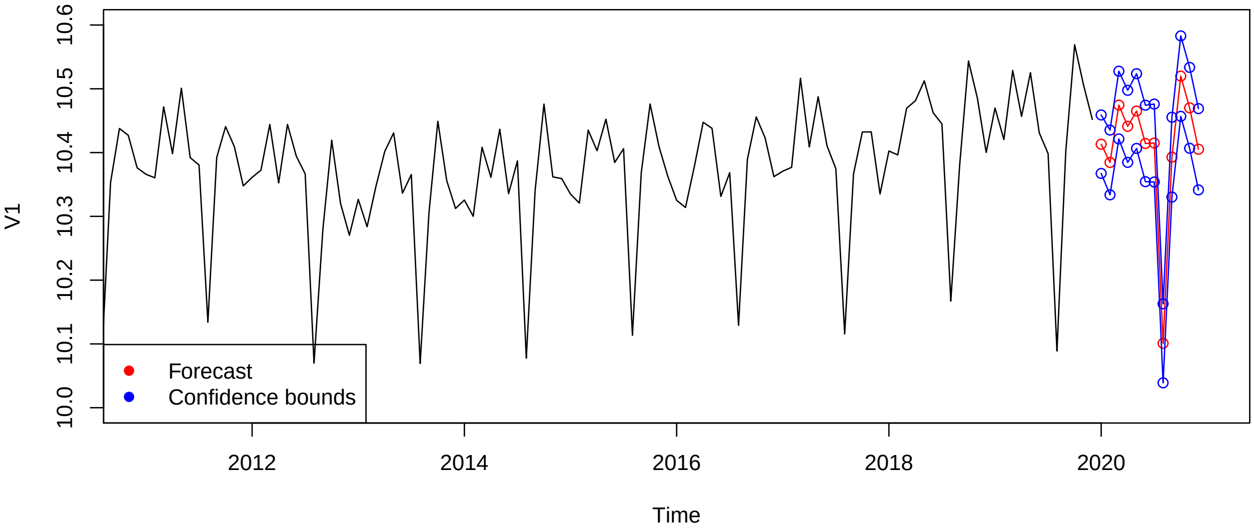


Figure 15: Forecasting for the next 12 months using the linearized time series.

Table 9: Forecasting for the next 12 months using the linearized time series.

	1	2	3	4	5	6	7	8	9	10	11	12
$\log(x_t)$	10.41	10.38	10.47	10.44	10.47	10.41	10.41	10.10	10.39	10.52	10.47	10.41
$+2se$	10.46	10.44	10.53	10.50	10.52	10.47	10.48	10.16	10.46	10.58	10.53	10.47
$-2se$	10.37	10.33	10.42	10.38	10.41	10.35	10.35	10.04	10.33	10.46	10.41	10.34

4.5 Model selection

Finally, we already have two possible forecast for next 12 months; so we have to compare these two models. Both forecast are shown in Figure 16 where it is clearly to see that the model which works with linearized time series gives us lower values against the non-linearized one.

We can consider those performance indicators such as Table 10 shows which are obtained when we not consider the last 12 observation to check the stability of the models. Thus, we have criterion on which model selection would be based on. Therefore, according to these criterias, that model estimated on linearized time series is only worst in terms of the number of estimated parameters, nonetheless the parsimony criterias such as AIC and BIC are better in both cases; so we can claim that the second model which includes the outliers treatment is the best option among the estimated models.

Table 10: Performance indicators for estimated models.

	par	Sigma2	AIC	BIC	RMSPE	MAPE	meanLength
Non-linearized	5	0.0021	-3.1455	-3.0675	0.0040	0.0032	0.2152
Linearized	24	0.0005	-4.4752	-4.3972	0.0025	0.0023	0.1166

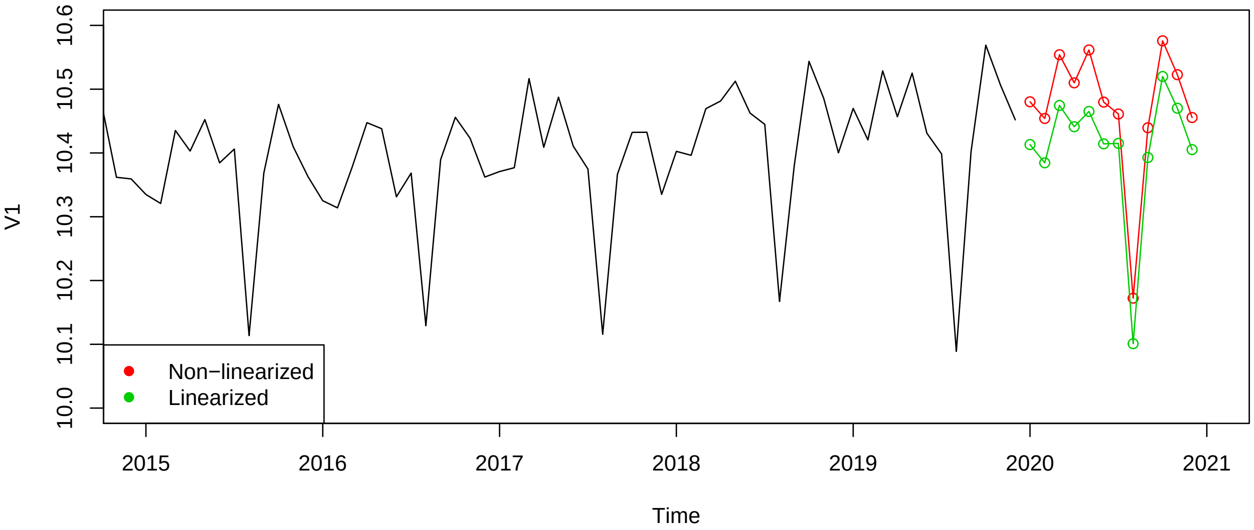


Figure 16: Comparison on forecasting for the next 12 months using the non-linearized and linearized time series.