Time Series Final Project

Analysis of the monthly users of the Barcelona metro

Quim Bassa, Bernat Chiva & Ferran Ibañez

28/4/2021

Abstract

The aim of this project is to analyze the monthly evolution of the total number of passengers of the Barcelona metro. In particular, we will identify, estimate and validate several models in order to choose the one that suits better to forecast the monthly number of passengers. Specifically, we will se that applying an ARIMA(3,0,0)(0,0,2) to a seasonally differentiated searies we obtain solid and reliable forecasts.

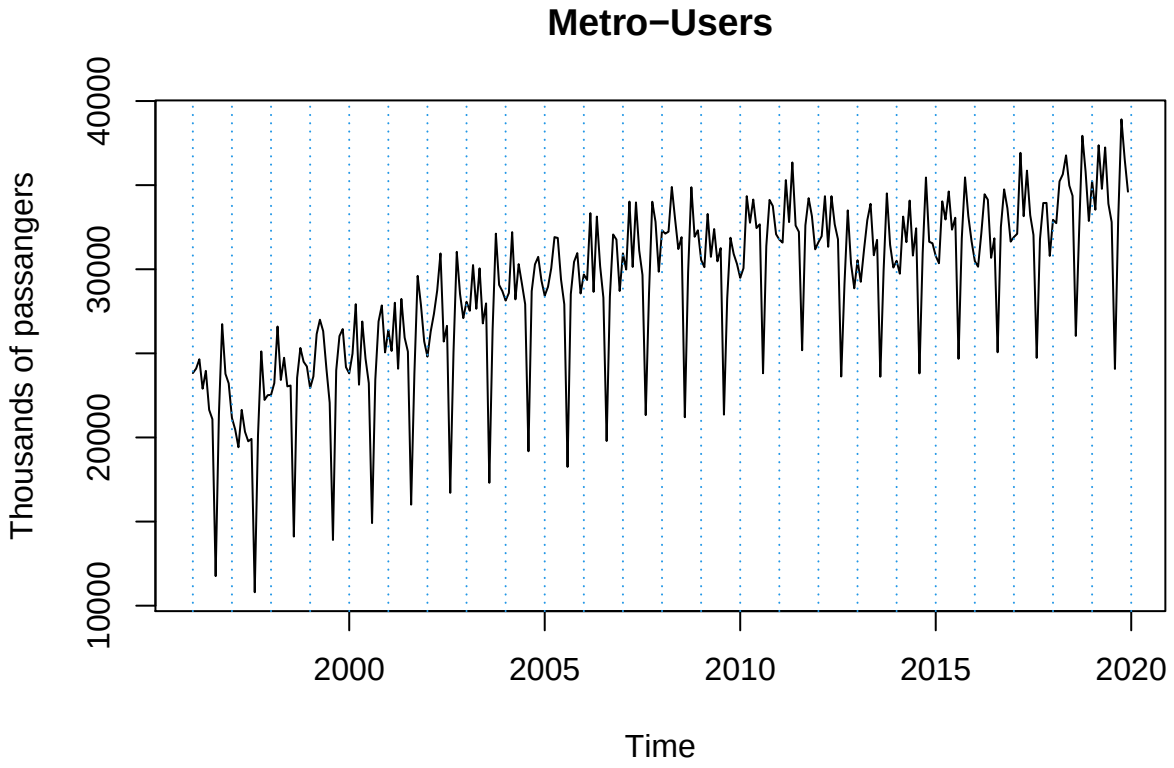
Contents

1 Introduction

This project aims to study the evolution of the number of passengers of Barcelona Metro during the last 20 years. Applying the Box-Jenkins ARIMA methodology to the data given, we expect to reveal the time series analysis and make predictions. First, we will make a quick exploratory data analysis checking the different properties of Time Series. Second, we will propose and fit several models for which we would choose the best one to make predictions and last but not least, we will check the presence of calendar effects and outliers in the series.

2 Dataset and Exploratory Data Analysis

Our study is based in the series of monthly number of passengers of Barcelona Metro during the last 20 years which can be found in the Instituto Nacional de Estadística webpage:
http://www.ine.es/jaxiT3/Tabla.htm?t=20193.



We can see that since the beginning of this series the monthly number of passengers has been increasing since 2020. This may be traduced in no constant mean. Moreover, we can also appreciate a seasonality pattern. There is an oscillation around the trend, the values vary periodically over the months which makes sense according to the reality. We know, for instance, that during summer the number of Metro passengers always drops.

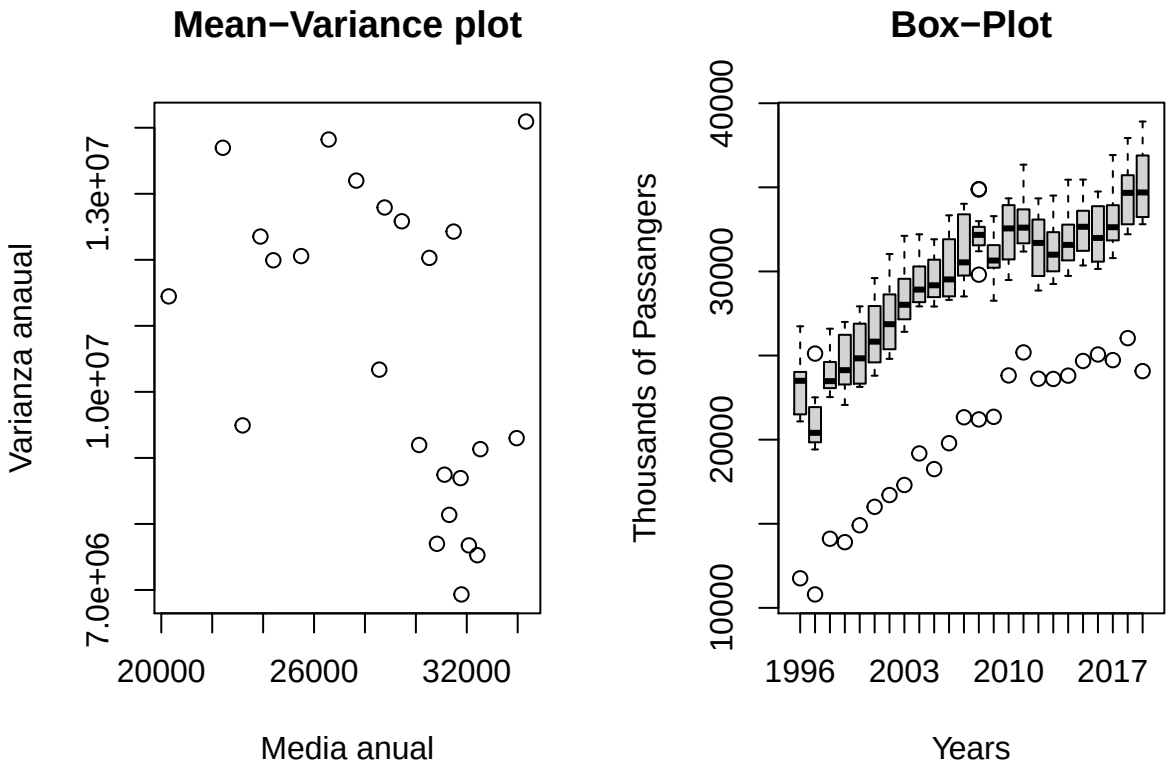
2.1 Stationarity

Below we perform a test to check if the series is stationary and if not we would perform the transformations needed. The series is needed to be stationary so as to obtain consistent parameter

estimates. This is a crucial property that must hold in order to obtain significant conclusion.

2.1.1 Variance Diagnose

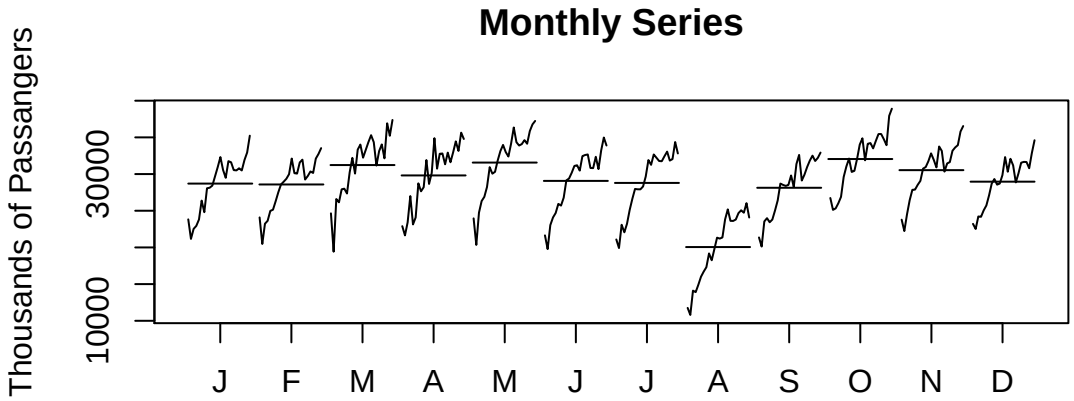
Constant variance is the first property that must hold in order to have stationarity. The verification of this property is based on the Box-plot and the Mean-variance plots.

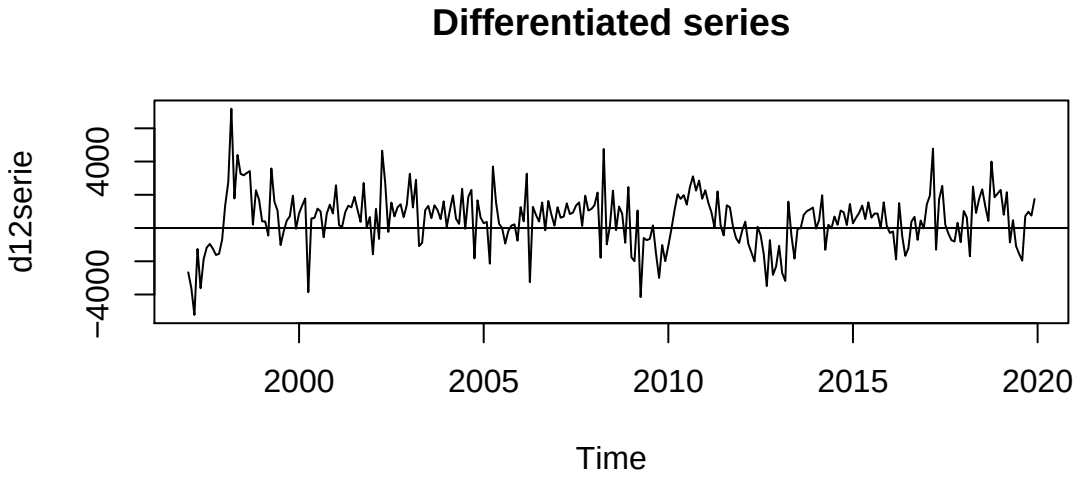


In our case, the plots reveal that there is constant variance. We do not see an increase of variance for high values of the mean, and the length of the boxes remain more or less constant through the different levels.

2.1.2 Seasonality

The verification of seasonality is based in the monthplot.





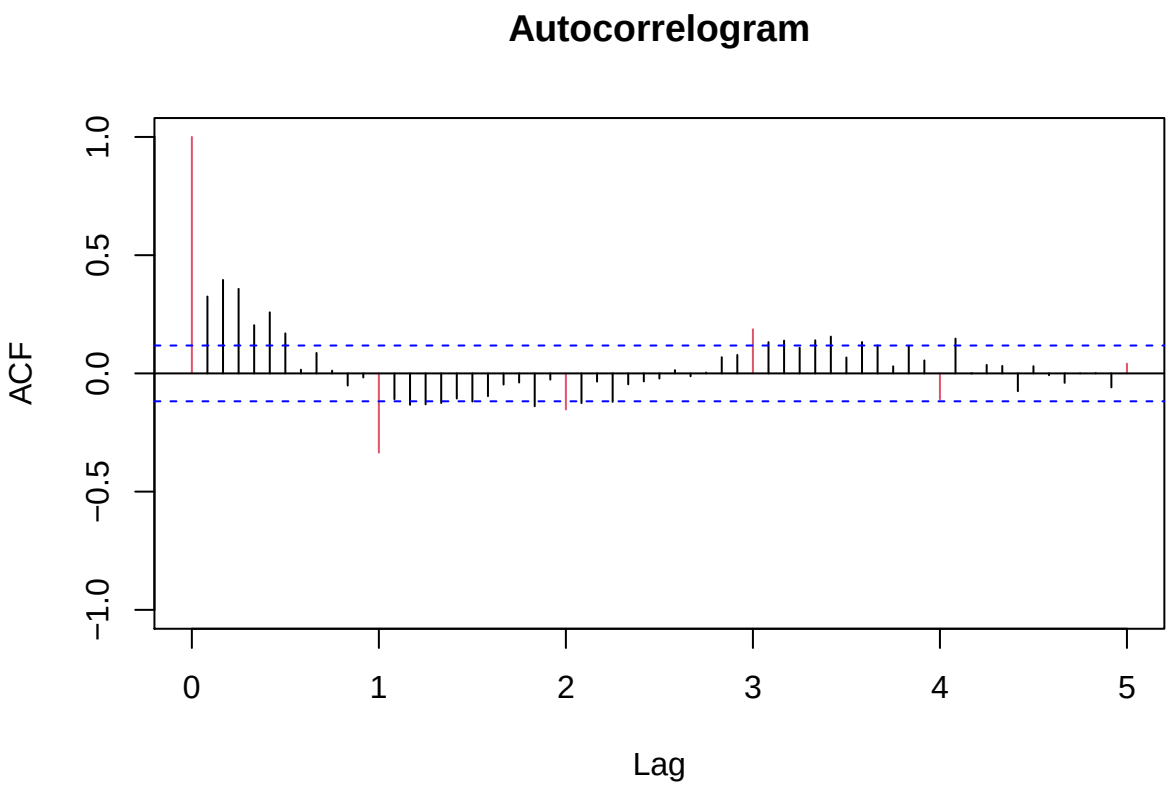
In the beggining of this section we mentioned that there was a clear seasonality patter present in the series. Now, the Mont-Plots reveals that every year during August the number of passangers drops considerably. Therefore, it would be interesting to take into account a seasonality of order 12.

2.1.3 Constant Mean

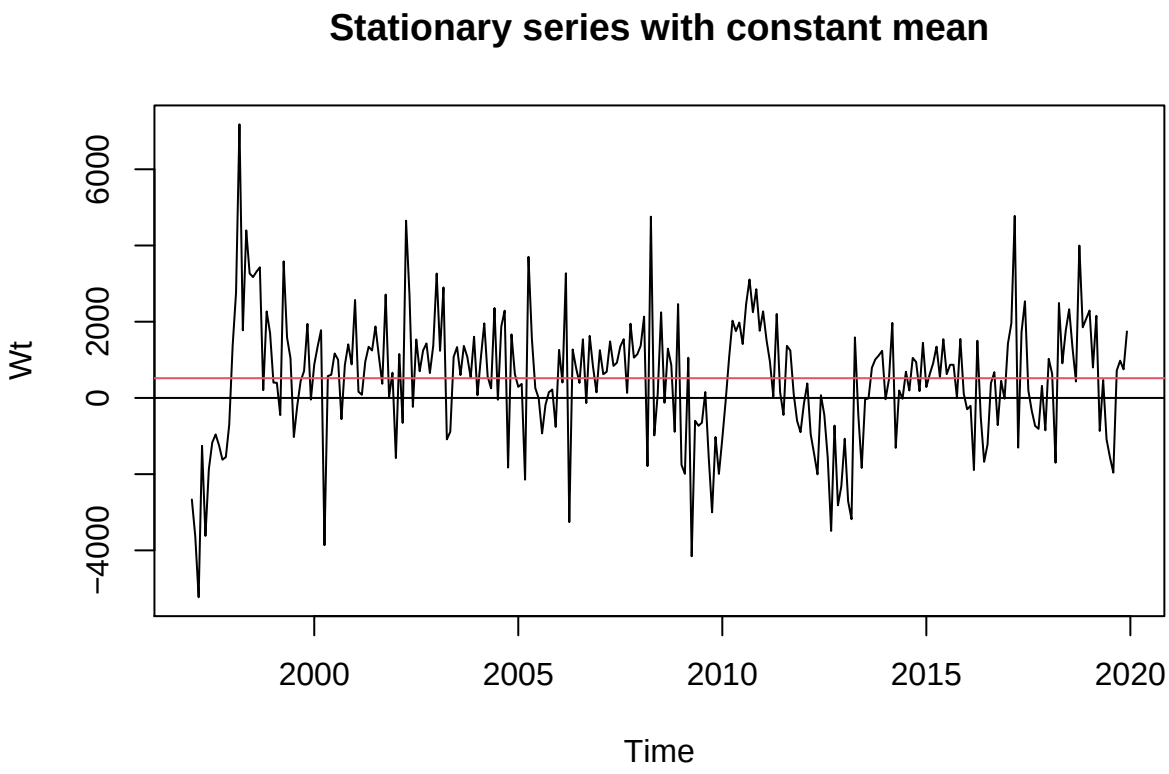
Now we will check whether the mean of the series is constant. Since it is not straight forward to deduce if the mean is constant or not from the plot of the seasonal differentiated series, we will validate it using the ACF.

Variance	serie	d12serie	d1d12serie
Value	2.4231993×10^7	2.7102416×10^6	3.6286313×10^6

Table 1: Series variance

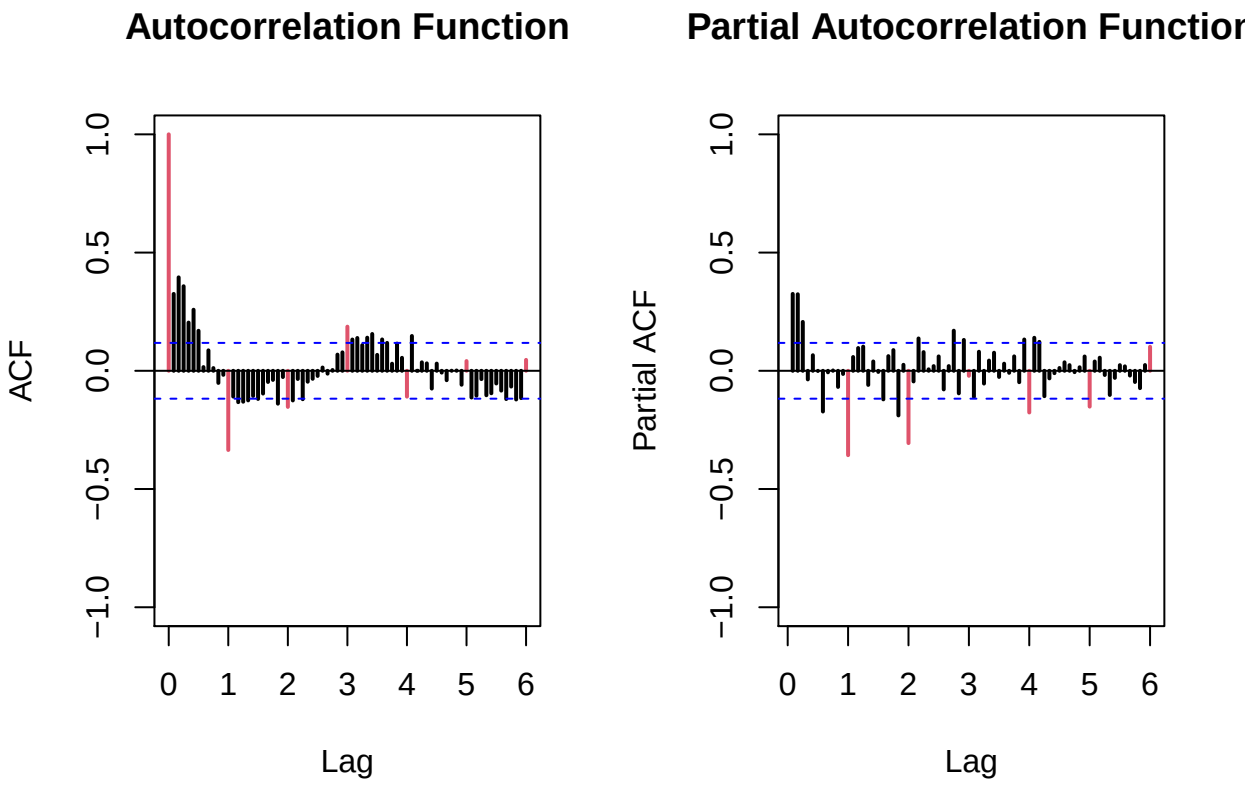


The ACF suggests that the series is already stationary since the ACF decays fast towards zero. Therefore the dependence structure will only depend on the lags and not on the time origin. Moreover, if we take the variance of the different series we can see that the series is already stationary with the seasonal differentiated series since an extra regular difference makes the variance increase. Then can conclude then, that the seasonal differentiated series of the number of Metro passengers is stationary with constant mean equal to 517.86.



2.2 Model Identification

In this section we will analyze the ACF and PACF of the stationary series to identify several plausible models.



Observing the plots we can identify at least two model for each component:

For the regular part we can assume AR(3) or and ARMA(1,1).

- AR(3): The ACF shows a exponential decreasing pattern, and the PACF has no significant lag after p=3. Following the parsimony principle the lag 6 has been not considered as true significant.
- ARMA(1,1): We consider a exponential decreasing pattern for both, ACF and PACF.

For the seasonal part we can assume MA(1) or ARMA(1,1).

- MA(1): The PACF shows a exponential decreasing pattern, and the ACF has no significant lag after q=1.
- AR(2): The ACF shows a sinusoidal decreasing pattern, and the PACF has no significant lag after p=2.

3 Model estimation

In the previous section we have identified different models for the regular and seasonal part. From all, we are going to select only four combinations to perform the estimation. Those models selected are the following:

- ARIMA$(3,0,0)(0,0,1)_{12}$: $X_t(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)=Z_t(1+\Theta_1 B^{12})$
- ARIMA$(3,0,0)(2,0,0)_{12}$: $X_t(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(1-\Phi_1 B^{12}-\Phi_2 B^{24})=Z_t$
- ARIMA$(1,0,1)(1,0,1)_{12}$: $X_t(1-\phi_1 B)(1-\Phi_1 B^{12})=Z_t(1+\theta_1 B)(1+\Theta_1 B^{12})$
- ARIMA$(3,0,0)(0,0,2)_{12}$: $X_t(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)=Z_t(1+\Theta_1 B^{12}+\Theta_2 B^{24})$

Table 2 displays the parameter estimates. We can observe that the intercept is significant for all the models. Moreover, almost all the coefficients are significant at a 95% confidence. Only the seasonal AR1 parameter for the ARIMA$(1,0,1)(1,0,1)_{12}$ and the seasonal MA2 parameter for the ARIMA$(3,0,0)(0,0,2)_{12}$, are non-significant. In terms of AIC, the ARIMA$(3,0,0)(0,0,1)_{12}$ and the ARIMA$(3,0,0)(0,0,2)_{12}$ are the ones with the lowest value.

Table 2: Results

	ARIMA(3,0,0)(0,0,1)			
	d12serie			
	ARIMA(3,0,0)(0,0,1)	ARIMA(3,0,0)(2,0,0)	ARIMA(1,0,1)(1,0,1)	ARIMA(3,0,0)(0,0,2)
	(1)	(2)	(3)	(4)
ar1	0.230***	0.141**	0.948***	0.234***
	(0.060)	(0.058)	(0.026)	(0.060)
ar2	0.330***	0.297***		0.330***
	(0.057)	(0.056)		(0.057)
ar3	0.250***	0.292***		0.240***
	(0.059)	(0.059)		(0.059)
ma1			−0.652***	
			(0.058)	
sma1	−0.794***		−0.850***	−0.723***
	(0.052)		(0.053)	(0.071)
sar1		−0.570***	0.090	
		(0.059)	(0.080)	
sar2		−0.387***		
		(0.060)		
sma2				−0.103
				(0.075)
intercept	520.937***	534.509***	510.084***	513.526***
	(88.883)	(138.724)	(96.122)	(79.091)
Observations	276	276	276	276
Log Likelihood	-2,345.401	-2,352.702	-2,347.907	-2,344.431
σ^2	1,346,408.000	1,449,496.000	1,365,784.000	1,332,533.000
Akaike Inf. Crit.	4,702.801	4,719.403	4,707.814	4,702.862

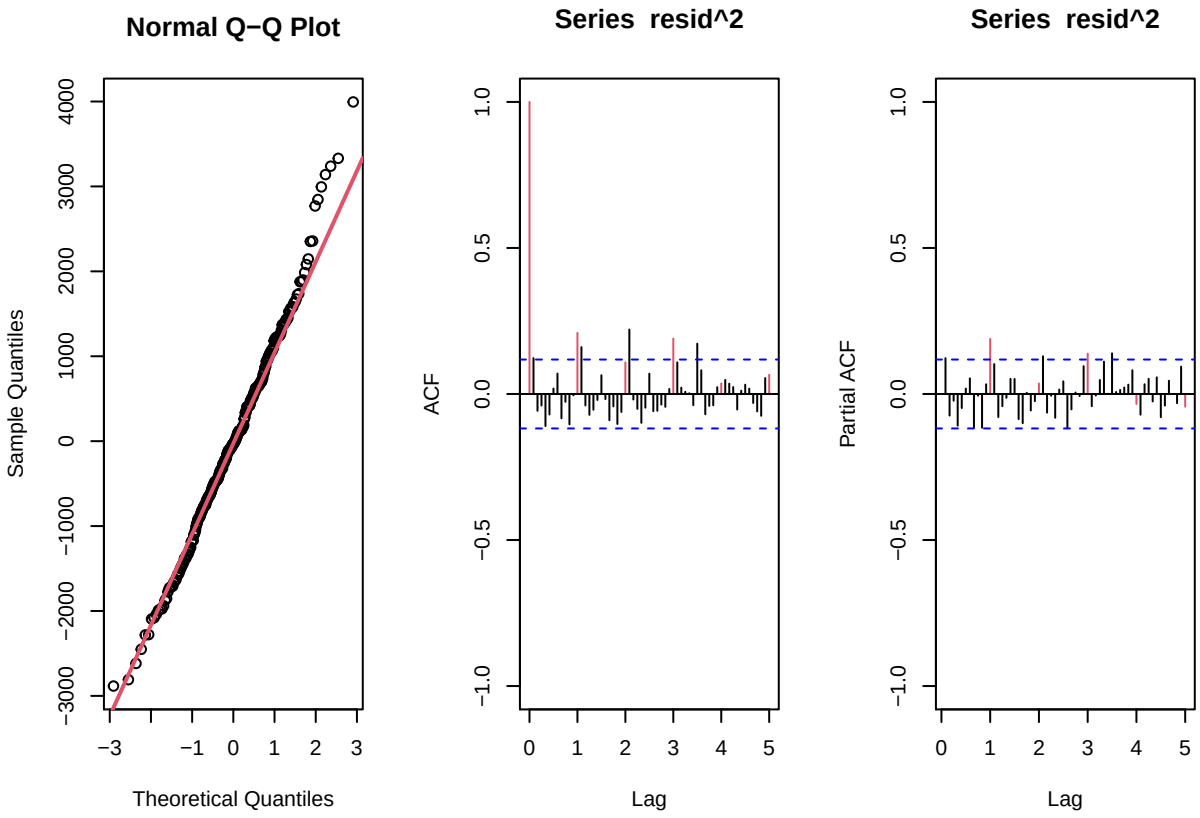
Note: *p<0.1; **p<0.05; ***p<0.01

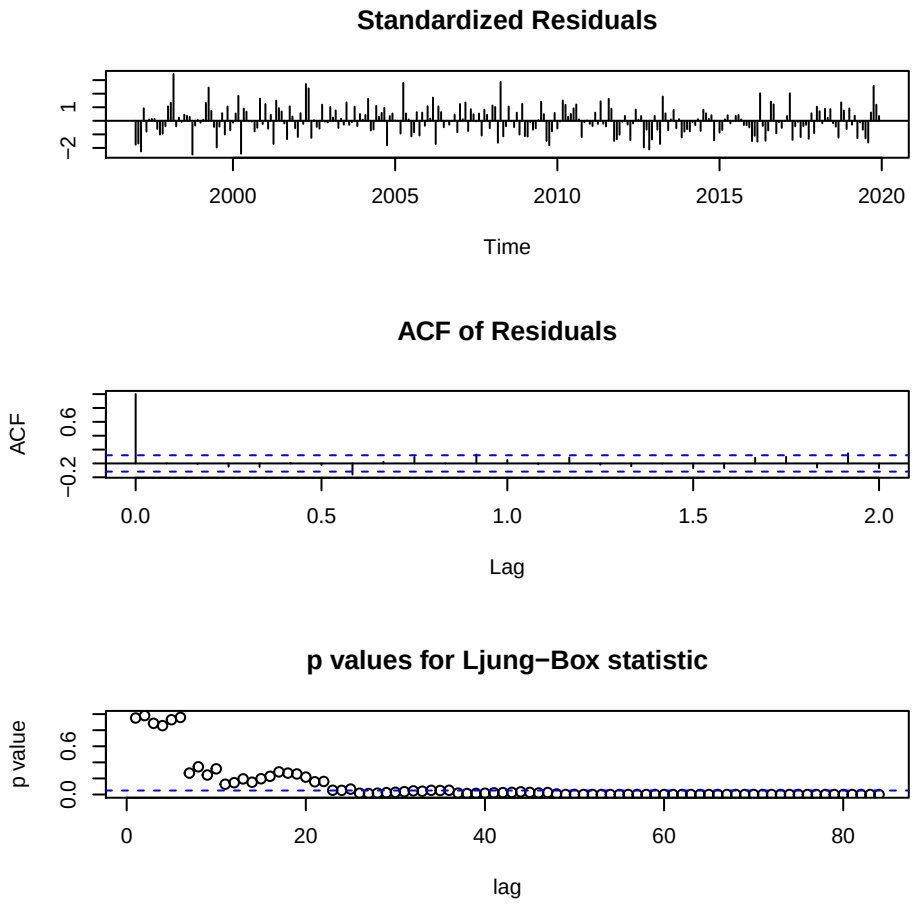
4 Model Validation

In this section we are going to provide a residual analysis, check if all the assumptions hold, evaluate the model capability for prediction and select the best model to forecast. The evaluation will be done and presented for each model.

4.1 ARIMA(3,0,0)(0,0,1)

4.1.1 Residual Analysis





- Normality of the residuals: Using a QQplot and the histogram of the residuals we should be able to see that the errors are normally distributed around 0. This is the case of Model 1, both the QQplot and the histogram show no deviation from a theoretical normal distribution, we only see some outliers in the right side of the distribution.
- Homocedasticity of the residuals: For validation of the model the errors should have a constant variance. This can be checked by plotting the residuals in a timeline, both "Residuals" and "Square Root of Absolute Residuals" show a linear trend with no big oscillations.
- Independence of the residuals: The third assumption we have to verify is that the errors are independent from each other, one of the ways to detect it is using the Ljung-Box statistic, that computes a hypothesis test for each error. Plotting the p-values of the statistic we should see that these value are over the significance line of 0.05, this is not entirely true in Model 1, where we see that from lag 23, independence is not followed. One possible solution would be to re-identify or add another parameter to the model.

4.1.2 Causality and Invertibility

If we take a look to the invertibility and causality of the model, we need to compute the module of all roots which should be greater than 1:

- Modul of AR Characteristic polynomial Roots: 1.107838 1.900311 1.900311. So the model is invertible.

- Modul of MA Characteristic polynomial Roots: 1.019415 1.019415 1.019415 1.019415 1.019415 1.019415 1.019415 1.019415 1.019415 1.019415 1.019415 1.019415 1.019415 . So the model is causal.

4.1.3 Capability of prediction

A model will be suitable to make forecast if it satisfies the condition of stability. To evaluate stability, we are going to perform to model estimations, one with the full sample and the other one without the last 12 observations. Then, we will compare wheter the results are similar en sign, magnitud and significance and if so, it would mean that the model is stable and suitable for making predictions.

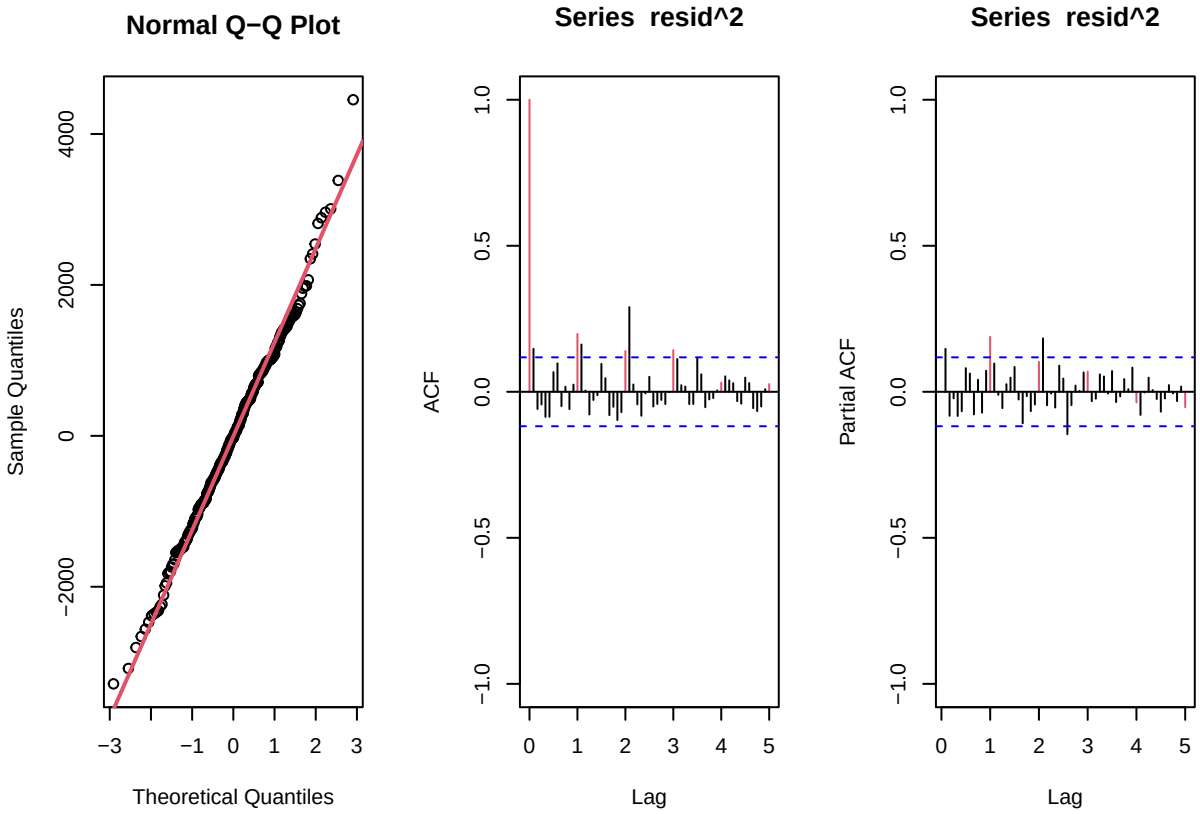
Table 3: ARIMA(3,0,0)(0,0,1)

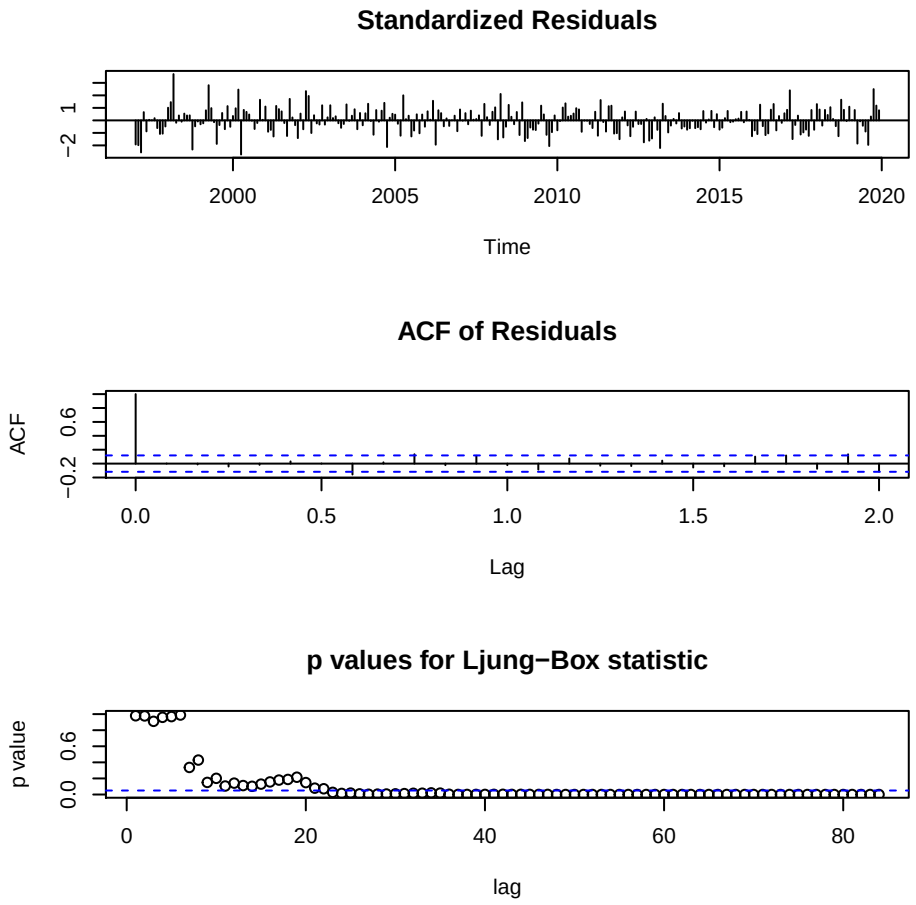
	Full Sample	Shrinked Sample
	(1)	(2)
ar1	0.230^{***}	0.204^{***}
	(0.060)	(0.061)
ar2	0.330^{***}	0.335^{***}
	(0.057)	(0.057)
ar3	0.250^{***}	0.278^{***}
	(0.059)	(0.060)
sma1	-0.794^{***}	-0.774^{***}
	(0.052)	(0.047)
intercept	520.937^{***}	515.485^{***}
	(88.883)	(99.985)
Observations	276	264
Log Likelihood	−2,345.401	−2,241.496
σ^2	1,346,408.000	1,328,907.000
Akaike Inf. Crit.	4,702.801	4,494.993
Note:		*p<0.1; **p<0.05; ***p<0.01

Clearly in this case the stability is fullfiled, this means that the correlation structure has not changed in the last year, and that the use of the complete series for making predictions is reliable.

4.2 ARIMA(3,0,0)(2,0,0)

4.2.1 Residual Analysis





- Normality of the residuals: Using a QQplot and the histogram of the residuals we should be able to see that the errors are normally distributed around 0. This is the case of Model 1, both the QQplot and the histogram show no deviation from a theoretical normal distribution, we only see some outliers in the right side of the distribution.
- Homocedasticity of the residuals: For validation of the model the errors should have a constant variance. This can be checked by plotting the residuals in a timeline, both "Residuals" and "Square Root of Absolute Residuals" show a linear trend with no big oscillations.
- Independence of the residuals: The third assumption we have to verify is that the errors are independent from each other, one of the ways to detect it is using the Ljung-Box statistic, that computes a hypothesis test for each error. Plotting the p-values of the statistic we should see that these value are over the significance line of 0.05, this is not entirely true in Model 2 again, the lags from lag 23 onwards are dependent.

4.2.2 Causality and Invertibility

- Modul of AR Characteristic polynomial Roots: 1.0404 1.150185 1.0404 1.724141 1.724141 .

They are all greater than 1, so Model 2 is invertible and causal.

4.2.3 Capability of prediction

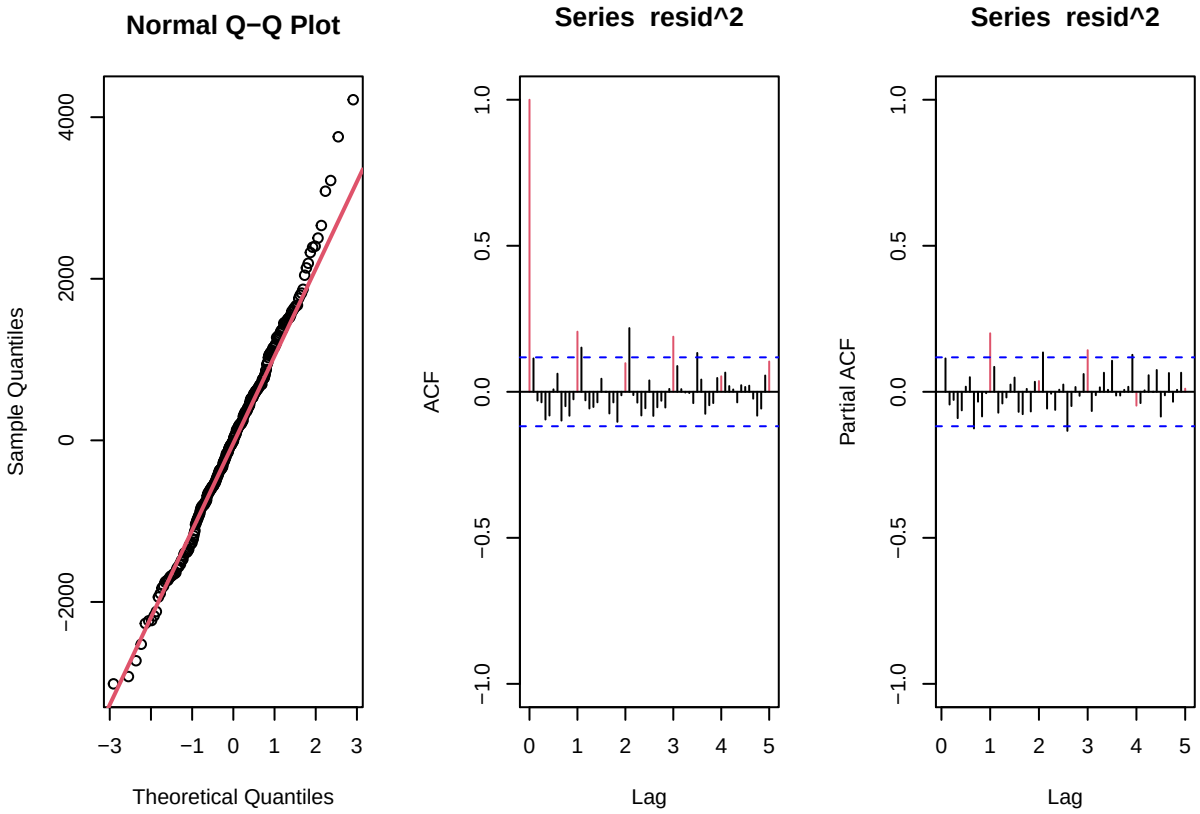
Table 4: ARIMA(3,0,0)(2,0,0)

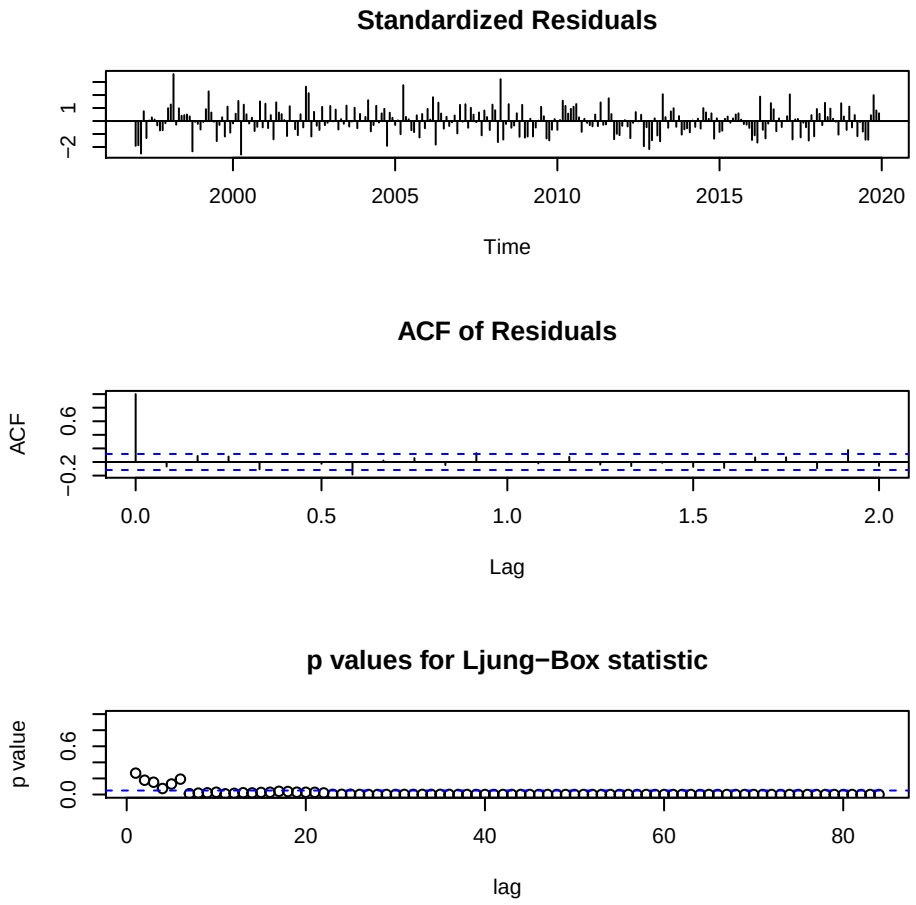
	Full Sample	Shrinked Sample
	(1)	(2)
ar1	0.141**	0.122**
	(0.058)	(0.058)
ar2	0.297***	0.302***
	(0.056)	(0.056)
ar3	0.292***	0.325***
	(0.059)	(0.060)
sar1	−0.570***	−0.595***
	(0.059)	(0.059)
sar2	−0.387***	−0.432***
	(0.060)	(0.061)
intercept	534.509***	518.985***
	(138.724)	(144.749)
Observations	276	264
Log Likelihood	−2,352.702	−2,246.854
σ^2	1,449,496.000	1,403,273.000
Akaike Inf. Crit.	4,719.403	4,507.708
Note:		*p<0.1; **p<0.05; ***p<0.01

Clearly in this case the stability is fullfiled, this means that the correlation structure has not changed in the last year, and that the use of the complete series for making predictions is reliable.

4.3 ARIMA(1,0,1)(1,0,1)

4.3.1 Residual Analysis





- Normality of the residuals: Using a QQplot and the histogram of the residuals we should be able to see that the errors are normally distributed around 0. This is the case of Model 1, both the QQplot and the histogram show no deviation from a theoretical normal distribution, we only see some outliers in the right side of the distribution.
- Homocedasticity of the residuals: For validation of the model the errors should have a constant variance. This can be checked by plotting the residuals in a timeline, both "Residuals" and "Square Root of Absolute Residuals" show a linear trend with no big oscillations.
- Independence of the residuals: The third assumption we have to verify is that the errors are independent from each other, one of the ways to detect it is using the Ljung-Box statistic, that computes a hypothesis test for each error. Plotting the p-values of the statistic we should see that these value are over the significance line of 0.05, this is not true in Model 3 and we can conclude that the residuals are not independent.

4.3.2 Causality and Invertibility

4.3.3 Capability of prediction

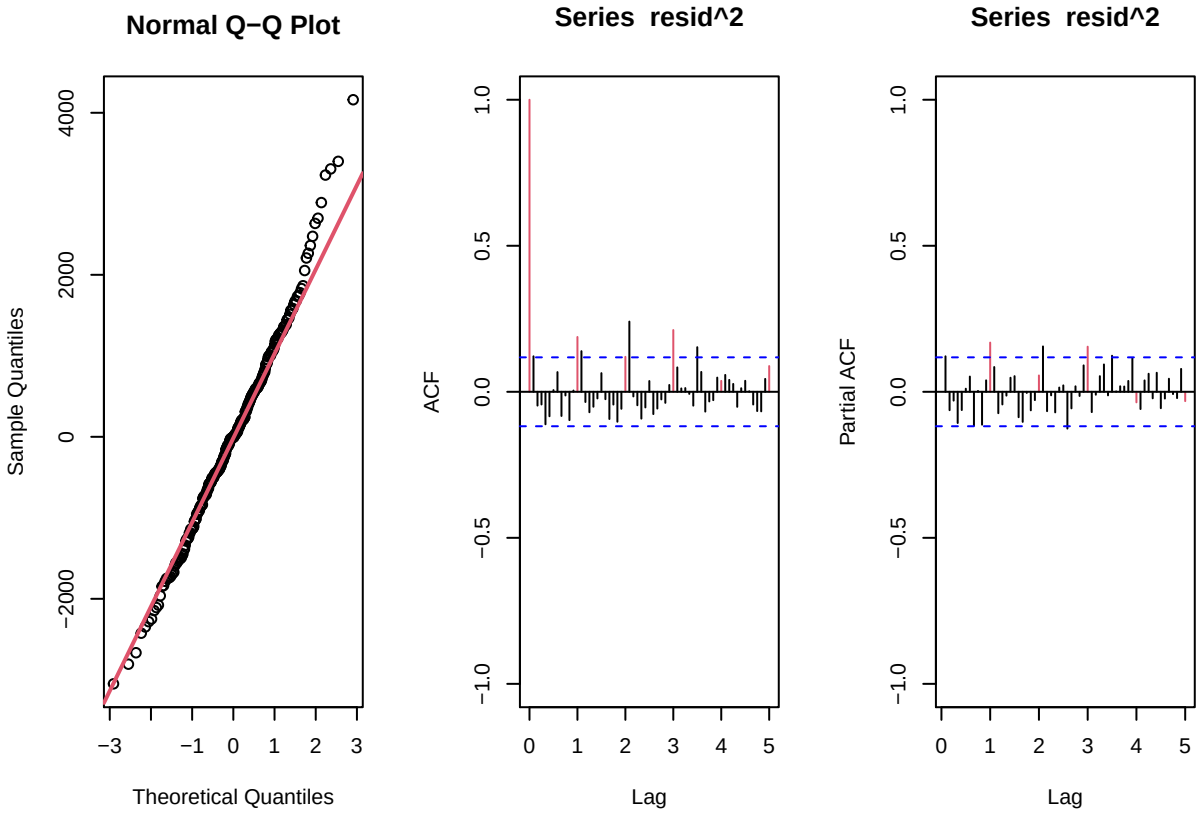
Clearly in this case the stability is fullfiled, this means that the correlation structure has not changed in the last year, and that the use of the complete series for making predictions is reliable.

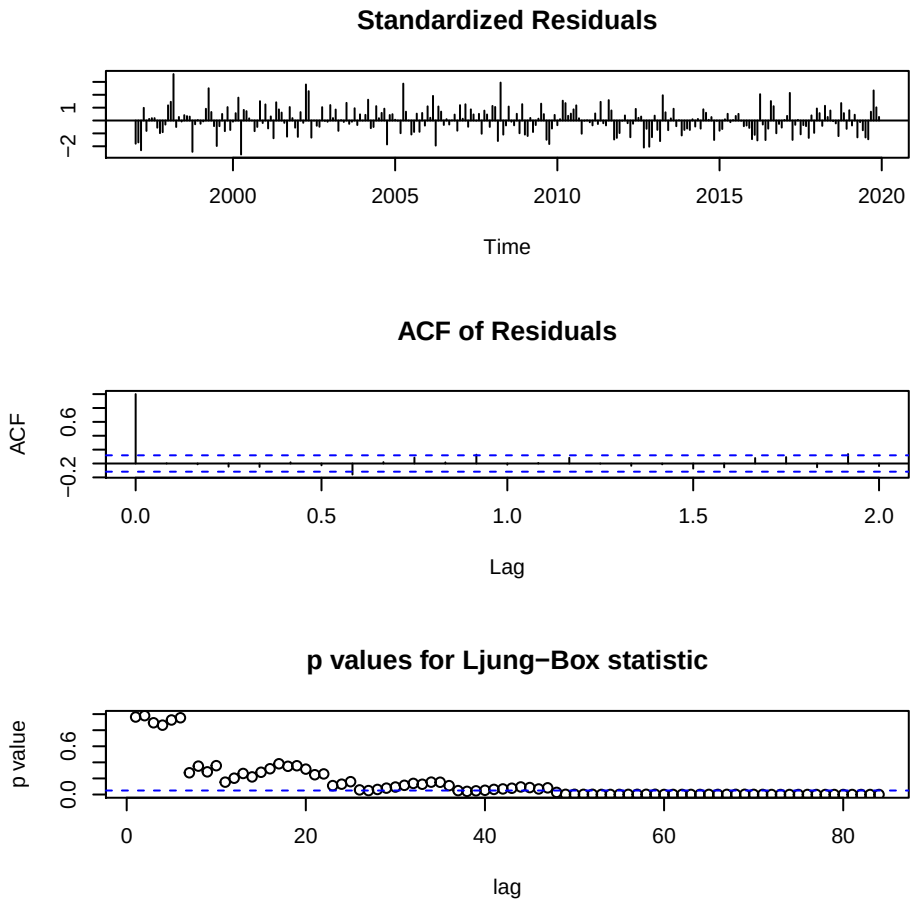
Table 5: ARIMA(1,0,1)(1,0,1)

	Full Sample	Shrinked Sample
	(1)	(2)
ar1	0.948^{***}	0.953^{***}
	(0.026)	(0.024)
ma1	-0.652^{***}	-0.668^{***}
	(0.058)	(0.055)
sar1	0.090	0.052
	(0.080)	(0.079)
sma1	-0.850^{***}	-0.815^{***}
	(0.053)	(0.050)
intercept	510.084^{***}	511.194^{***}
	(96.122)	(114.974)
Observations	276	264
Log Likelihood	−2,347.907	−2,245.322
σ^2	1,365,784.000	1,365,738.000
Akaike Inf. Crit.	4,707.814	4,502.645
Note:		*p<0.1; **p<0.05; ***p<0.01

4.4 ARIMA(3,0,0)(0,0,2)

4.4.1 Residual Analysis





- Normality of the residuals: Using a QQplot and the histogram of the residuals we should be able to see that the errors are normally distributed around 0. This is the case of Model 1, both the QQplot and the histogram show no deviation from a theoretical normal distribution, we only see some outliers in the right side of the distribution.
- Homocedasticity of the residuals: For validation of the model the errors should have a constant variance. This can be checked by plotting the residuals in a timeline, both "Residuals" and "Square Root of Absolute Residuals" show a linear trend with no big oscillations.
- Independence of the residuals: The third assumption we have to verify is that the errors are independent from each other, one of the ways to detect it is using the Ljung-Box statistic, that computes a hypothesis test for each error. Plotting the p-values of the statistic we should see that these value are over the significance line of 0.05, this is true for Model 4, as we do not see dependence until lag 43.

4.4.2 Causality and Invertibility

If we take a look to the invertibility and causality of the model, we need to compute the module of all roots, and should be all greater than 1:

- Modul of AR Characteristic polynomial Roots: 1.113293 1.934532 1.934532

- Modul of MA Characteristic polynomial Roots: 1.014151 1.014151 1.014151 1.014151 1.014151 1.014151 1.014151 1.014151 1.014151 1.191581 1.191581 1.014151 1.014151 1.014151 1.014151 1.191581 1.191581 1.191581 1.191581 1.191581 1.191581 1.191581 1.191581 1.191581 1.191581

Since all roots are greater than 1, we can conclude that the model is invertible and causal. Thus, will be able to perform forecasting.

4.4.3 Capability of prediction

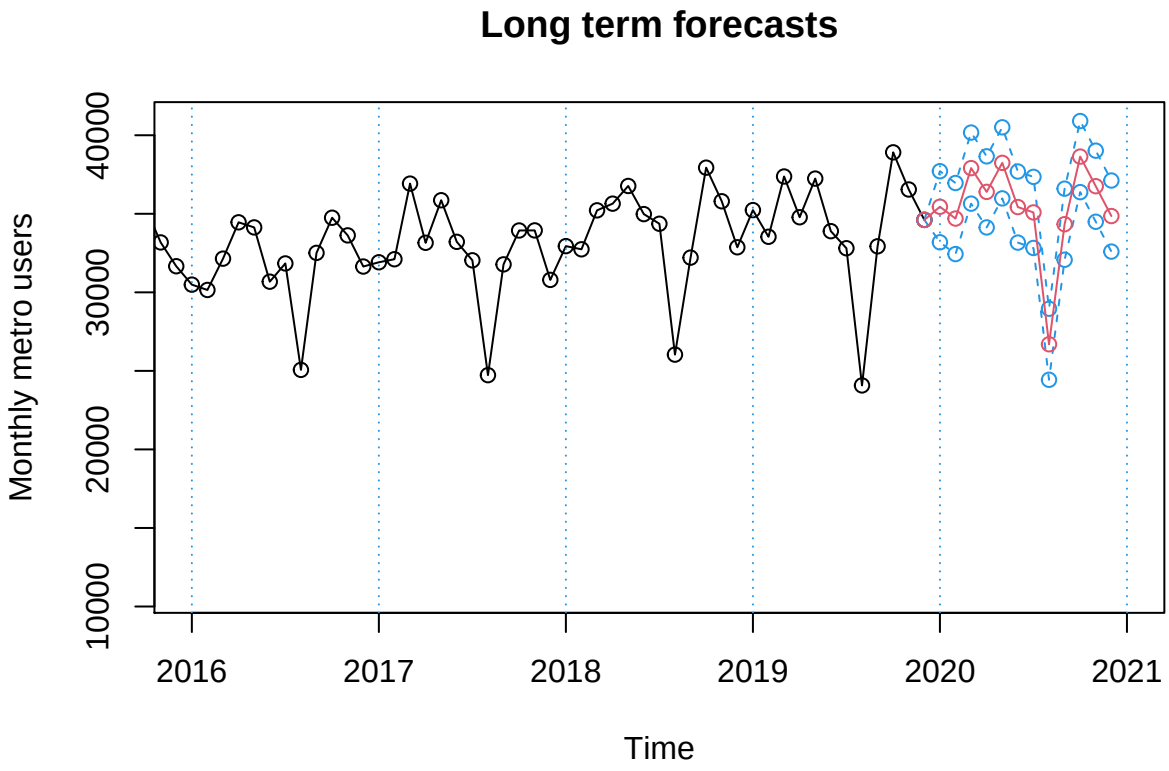
Table 6: ARIMA(3,0,0)(0,0,2)

	Full Sample	Shrinked Sample
	(1)	(2)
ar1	0.948^{***}	0.953^{***}
	(0.026)	(0.024)
ma1	-0.652^{***}	-0.668^{***}
	(0.058)	(0.055)
sar1	0.090	0.052
	(0.080)	(0.079)
sma1	-0.850^{***}	-0.815^{***}
	(0.053)	(0.050)
intercept	510.084^{***}	511.194^{***}
	(96.122)	(114.974)
Observations	276	264
Log Likelihood	−2,347.907	−2,245.322
σ^2	1,365,784.000	1,365,738.000
Akaike Inf. Crit.	4,707.814	4,502.645
Note:	*p<0.1; **p<0.05; ***p<0.01	

The stability property is clearly fulfilled implying that the correlation structure has not changes in the last year and the use of the stationary differentiated series is reliable for making predictions.

5 Forecasting

In this section we are going to provide a plot with the long term forecast for the next twelve months with the corresponding confidence bands. The model we selected to perform the forecast is the $\text{ARIMA}(3,0,0)(0,0,2)_{12}$



6 Outlier Treatment and Calendar Effects

In this section we are going to analyze whether the Calendar Effects are significant and ajust the possible utiliers that may appear.

6.1 Calendar Effects

We try to detect and analyze the outliers for predicted events such as the Easter holidays or the proportion of trading days.

Due to we are dealing with monthly data we want to verify if either the variation of the number of labor days in a month or the calendar movement of the Easter week, affects the series.

In this case due to the intercept is significant, we execute the calendar effect treatment on the original series, but when we consider the auxiliary varibles in order to get the linearised time series, this variables must be in the same difference applied on original serie.

Besides, we also have considered an intervention analysis. We want to study the effect of the creation of integrated tickets in 2001. The integrated tickets are also known as T-10, T-Mes, or T-Jove, are a set of special tickets created by the government to promote the use of public transport. They allow the user to combine multiple modes of transport in one journey and use the same ticket multiple times.

We also consider the effect of this policy in the series. It makes sense to wonder if these tickets produced an actual increase in the number of passengers.

Looking the result displayed in table 7, we can assume that the best model is the one with the calendar effects but without the Intervention Analysis because it has the lowest AIC with all the parameters significant. Besides, we can conclude that the creation of the integrated tickets didn't have any significant effect on the number of passengers.

6.2 Outlier Detection

One the series has taken into account the calendar effects, it's time to check the presence of outliers. Table shows that only 2 outliers were found in the series: An additive outlier, whihc only affects on one period, and a Transitory Change, which affects on one period and its effect decreases in the next periods. knowing that, we are prepared to linearize our series taking into account the effects of the Calendar Effects and those two outliers.

6.3 Forecasting

Before carrying on with the forecasting, we need to check again the stability of the updated model.

Table 9 shows that the model is still stable because the estimates between the Full sample and shrinked sample are very similar so we can proceed to do the forecasts.

Table 7: Test of Calendar Effects

	No CE	Trading Days	Easter Effect	Easter-Trading Days	Easter, IA
	(1)	(2)	(3)	(4)	(5)
ar1	0.234***	0.373***	0.403***	0.585***	0.582***
	(0.060)	(0.062)	(0.060)	(0.060)	(0.060)
ar2	0.330***	0.340***	0.246***	0.198***	0.191***
	(0.057)	(0.061)	(0.063)	(0.070)	(0.070)
ar3	0.240***	0.111*	0.172***	0.056	0.043
	(0.059)	(0.060)	(0.060)	(0.061)	(0.062)
sma1	−0.723***	−0.829***	−0.494***	−0.551***	−0.555***
	(0.071)	(0.069)	(0.063)	(0.066)	(0.066)
sma2	−0.103	0.066	−0.323***	−0.201***	−0.198***
	(0.075)	(0.075)	(0.063)	(0.067)	(0.067)
intercept	513.526***	529.064***	501.423***	512.850***	780.432***
	(79.091)	(92.947)	(79.966)	(89.359)	(255.466)
d12wTradDays		159.301***		150.126***	150.069***
		(16.545)		(13.477)	(13.392)
d12wEast			−2,421.482***	−2,303.662***	−2,303.159***
			(190.048)	(153.324)	(153.589)
Itickets					−319.422
					(290.254)
Observations	276	276	276	276	276
Log Likelihood	−2,344.431	−2,311.125	−2,291.309	−2,246.891	−2,246.371
σ^2	1,332,533.000	1,052,649.000	909,872.500	666,380.700	663,724.300
Akaike Inf. Crit.	4,702.862	4,638.251	4,598.619	4,511.783	4,512.742

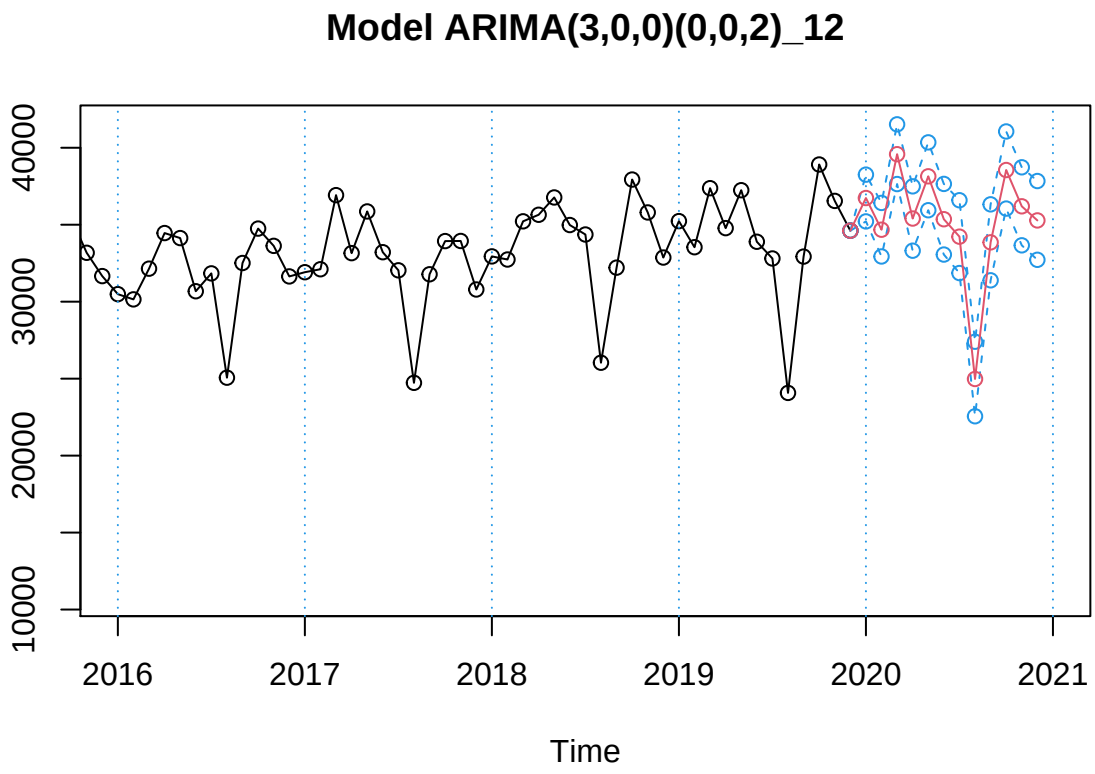
Note: *p<0.1; **p<0.05; ***p<0.01

	Obs	Type_detected	W_coeff	ABS_L_Ratio	Date	Perc.Obs
2	152	AO	1664.357	3.515739	Ago 2008	1664.357
1	271	TC	-2679.472	3.451213	Jul 2018	-2679.472

Table 8: Outliers Detected

Table 9: Stability Test

	Full Sample	Shrinked Sample
	(1)	(2)
ar1	0.558***	0.558***
	(0.060)	(0.060)
ar2	0.253***	0.253***
	(0.069)	(0.069)
ar3	0.056	0.056
	(0.062)	(0.062)
sma1	−0.611***	−0.611***
	(0.067)	(0.067)
sma2	−0.200***	−0.200***
	(0.067)	(0.067)
intercept	515.886***	515.886***
	(84.999)	(84.999)
d12wTradDays	153.184***	153.184***
	(13.101)	(13.101)
d12wEast	−2,302.526***	−2,302.526***
	(153.137)	(153.137)
Observations	276	276
Log Likelihood	−2,233.918	−2,233.918
σ^2	600,531.900	600,531.900
Akaike Inf. Crit.	4,485.835	4,485.835
Note:		*p<0.1; **p<0.05; ***p<0.01



In the plot above the red line indicated the estimated, forecasted, number of metro passengers for the next months, and in blue is the 95% confidence band. We can observe that the model still keeps the upward trend that exists from previous years and the seasinality with the drop in the month of August. Overall the forecasts seems to adjust really accurate with thin confidence bands, indicating good performance by the model.

7 Conclusions

We have been able throughout all the study to analyse the series of metro passangers, a dataset with more than 20 years of information and we have been able to do forecasting with it. It is paramount at this stage of the study to point out the complexity of real-life studies of this caliber, the amount of information that it is needed and the complexity of the models increases exponentially. However, as a first practical analysis with time series data we got to the results we set to accomplish in the beginning.

From a first descriptive analytics and with the use of informal tools, such as graphs and formal tools such as hypothesis testing we choose the best SARIMA model possible for the metro data. After careful analysis of the stationarity and making sure we would be able to make predictions, we have proceed with outlier and calendar effects treatment to better adjust the model. Once done, we have been able to proceed with the desired forecast for the next year.

All this process has been followed with strict focus on scientific and statistic results and we are proud of the outcome. The results can be used for planning purposes for the Public Sector when dealing with the metro schedule for the year 2020. Furthermore, the model could be reuses, after some regular checking, for the years to come and can be perfectioned with even more data. We are sure that the pandemic has truncated this model, making all of the forecast not usable in practice, although the model is still useful.