Time Series & M-V Analysis STAT8008

Assessment 3

João Viegas

30.04.2018

Student R00157699

Higher Diploma in Data Science and Analytics

Table of Contents

Introduction

In this assignment we conduct a simple experiment on a time series data set. The data set comprises the Consumer Price Index (CPI) monthly values in Ireland, spanning from January 2008 to December 2017.

We will model the the time series, using two different models, Winters Additive and ARIMA(1,0,3)(0,1,1), and we will choose then the one with a superior fit. We will then separate the data in estimation and forecast sets and use the chosen model to forecast the data and match it against the forecast set, assessing its performance. We will then produce a true forecast, covering the following year after the training data set time span, which will be the current year of 2018.

Dataset

This experiment is conducted in a data set about the Consumer Price Index (CPI) monthly values in Ireland, spanning from January 2008 to December 2017, comprising therefore 120 entries.

The dataset summary and descriptive statistics are presented in Table 1 and 2.

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Consumer Price Index	120	100.00%	0	0.00%	120	100.00%

Table 1 - dataset summary

Descriptives				
			Statistic	Std. Error
Consumer Price Index	Mean		146.57	0.25687
	95% Confidence Interval for Mean	Lower Bound	146.0614	
		Upper Bound	147.0786	
	5% Trimmed Mean		146.6954	
	Median		147.7	
	Variance		7.918	
	Std. Deviation		2.81388	
	Minimum		139.5	
	Maximum		151.2	
	Range		11.7	
	Interquartile Range		3.47	
	Skewness		-0.947	0.221
	Kurtosis		-0.139	0.438

Table 2 - dataset descriptive statistics

The figure 1 shows the boxplot for the data set with one outlier.

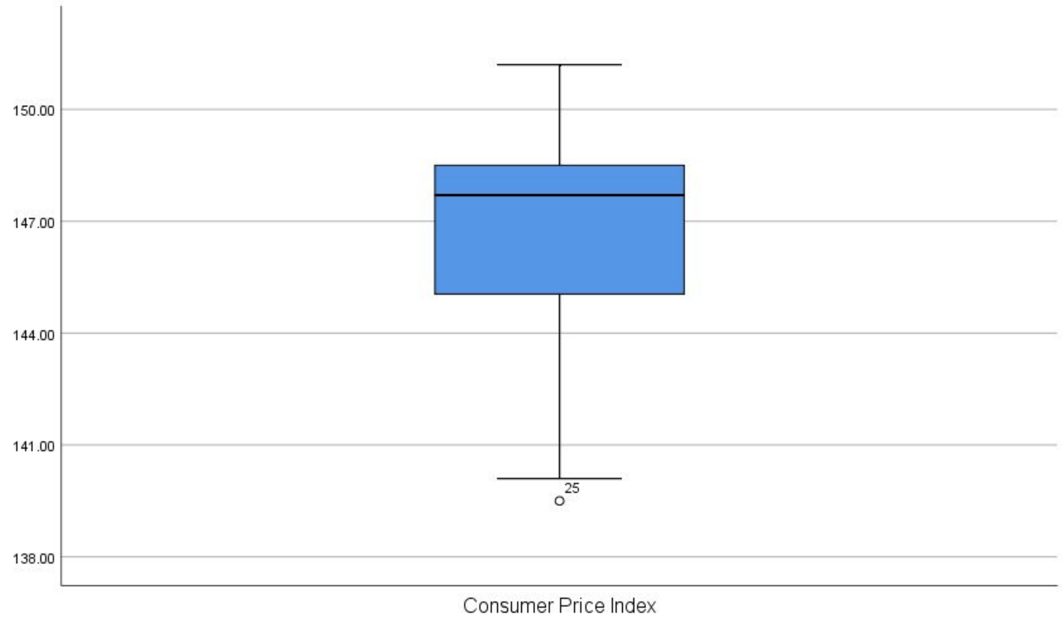

Fig 1 - dataset boxplot

In Fig 2, we can see the plot of the time series dataset, that we'll use in the experiment.

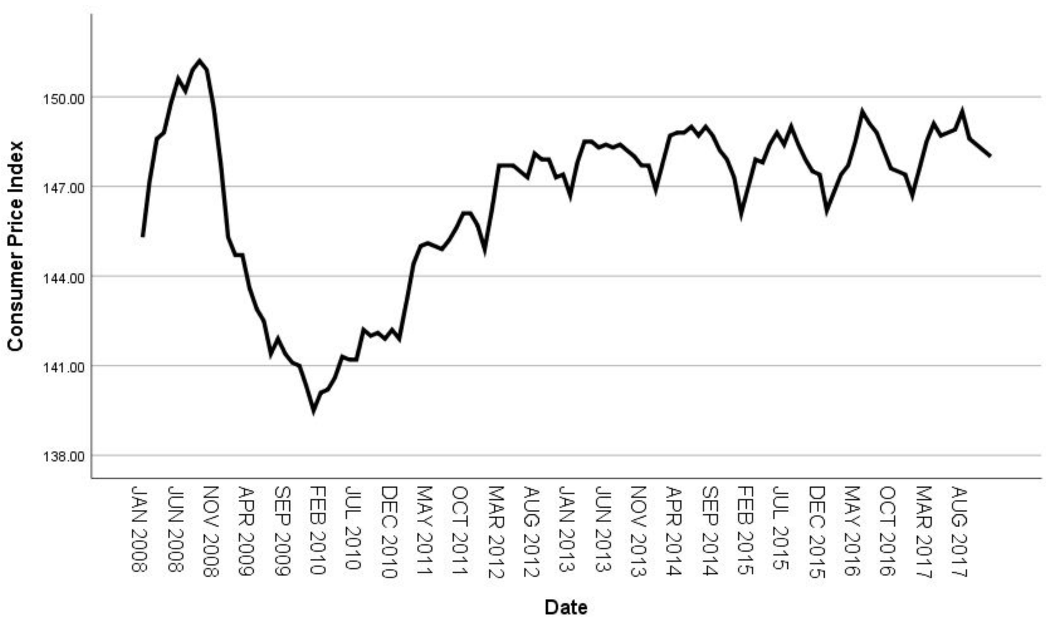

Fig 2 - dataset time series plot

The dataset is somewhat defined by 2 phases, a first one on when there is a strong upward trend and then a disruption taking the shape of a severe dip, a second one where there is a new upward trend and a following stabilization back to a slightly increasing trend with well defined seasonality. The dataset depicts to a substantial extent the runup to the economic crisis of 2008 (strong upward trend), the credit crunch (the dip) the gradual stabilizing of the economic environment (upward trend with also perceptible seasonality) after that and the following limited growth phase (slight upward trend with seasonality).

In fairness, we can detect a constant seasonality in the dataset, although being less perceptible in the graph in the first phases, it is clear, after the dip inflection point, that generally, the values tend to decrease from september until january, and to increase from February to August. It seems to be of constant magnitude, though.

Overall, and taking into account what we know already from the economic cycles, the CPI, is a seasonal series, normally with successive increasing/decreasing trends with an additional cyclical component that is really difficult to model and predict.

In the next sections we will use IBM SPSS Statistics, namely the Time Series Modeler, to fit a model to the series. We will use two models, *Winters Additive* and *Arima(1,0,3)(0,1,1)*, and try to find the best one that we can use for predicting purposes.

Winters Additive model vs Arima(1,0,3)(0,1,1)

To some extent we can say that the data set comprises phases with some linearity in its trend and near constant seasonality magnitude, which might suggest us to start by testing the *Winters Additive* model.

The Arima(1,0,3)(0,1,1) presumes a stationary, autoregressive process of order one and moving average process of order 3, assuming no trend in its non-seasonal component, and in its seasonal component, assuming a trend, a non-stationary process with integration of order one and moving average of order one.

We've then tested and compared these two models using SPSS Time Series Modeler.

Model Parameters

a) Winters Additive model

Model			**Estimate**	**SE**	**t**	**Sig.**
Consumer Price Index-Model_1	No Transformation	Alpha (Level)	1	0.091	10.95	0
		Gamma (Trend)	0.102	0.029	3.525	0.001
		Delta (Season)	0.999	9361.466	0	1

Table 3 - Winters Additive model parameters

In the Winters Additive model, the Delta parameter, season related, is nonsignificant.

b) ARIMA(1,0,3)(0,1,1)

ARIMA (1,0,3)(0,1,1) Model Parameters				Estimate	SE	t	Sig.
Consumer Price Index	No Transformation	AR	Lag 1	0.964	0.037	25.869	0
		MA	Lag 1	-0.591	0.093	-6.322	0
			Lag 2	-0.248	0.112	-2.216	0.029
			Lag 3	-0.272	0.103	-2.635	0.01
		Seasonal Difference		1			
		MA, Seasonal	Lag 1	0.89	0.224	3.973	0

Table 4 - ARIMA (1,0,3)(0,1,1) model parameters

In the Arima Model all the parameters, non-seasonal and seasonal are significant.

Observed vs Fit Values

a) Winters Additive model

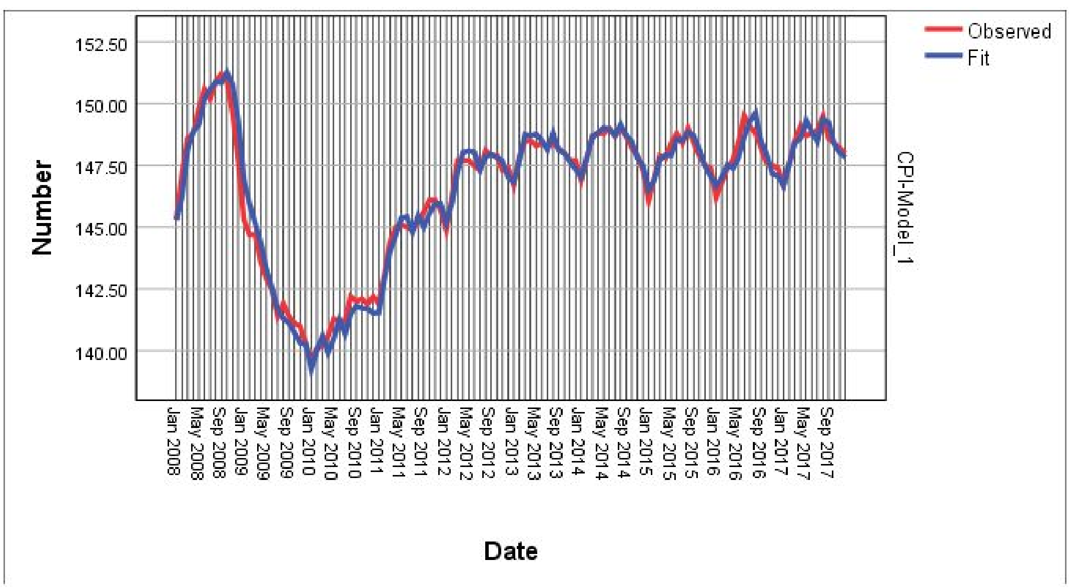

Fig 3 - Winters Additive model: Observed vs Fit data

Looking into the sequence chart showing the series and the fit values, we can say that, in general, the fitted data follows the original data very well in the *Winters Additive* model.

b) Arima(1,0,3)(0,1,1)

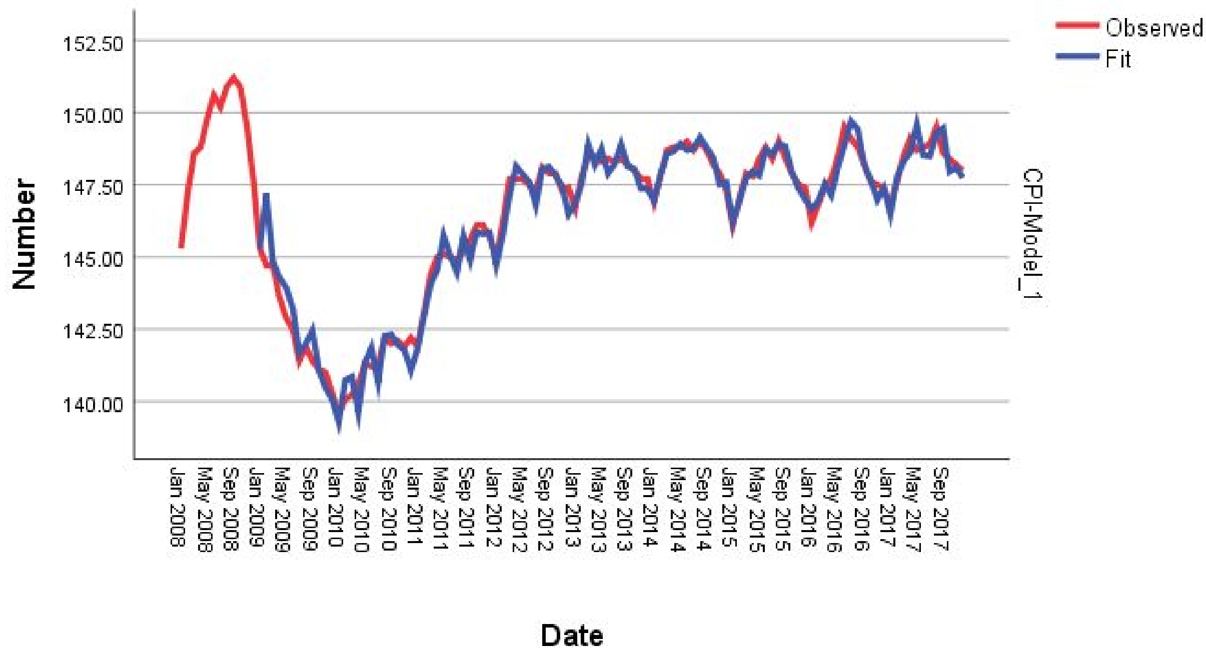

Fig 4 - ARIMA(1,0,3)(0,1,1) model: Observed vs Fit data

In the ARIMA model, the fit values also follow the observed ones quite closely.

Residuals

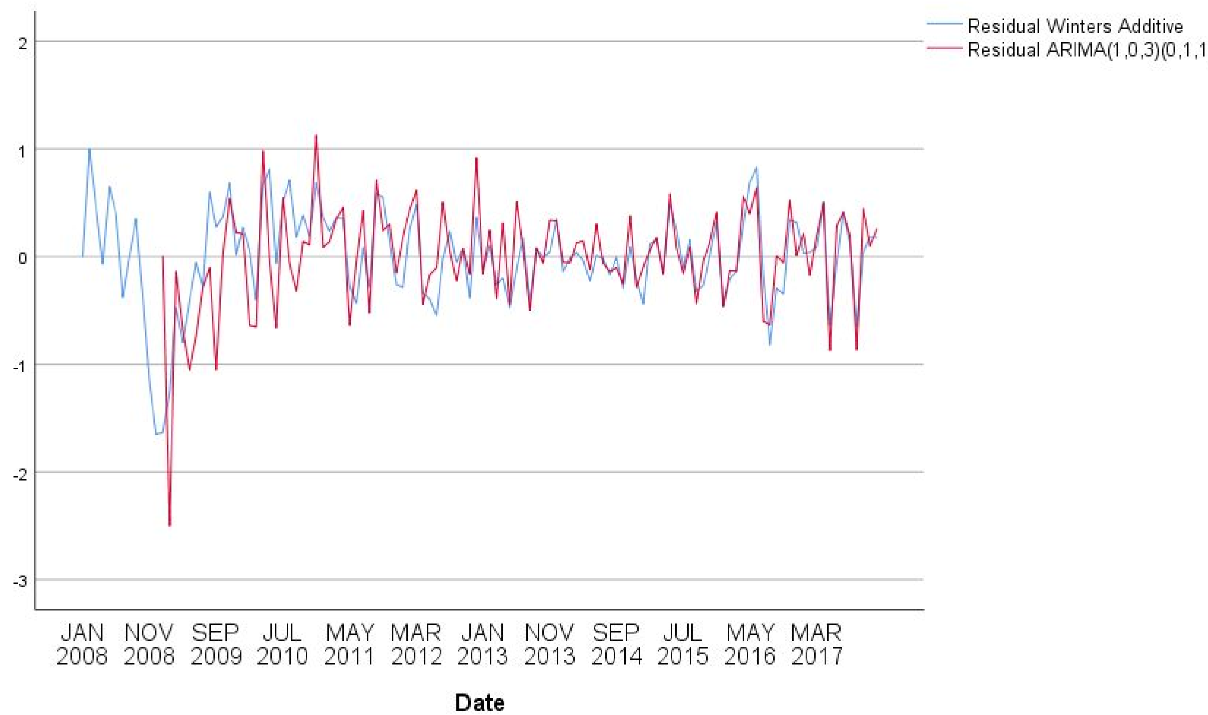

Fig 5 - Residuals: Winters Additive vs ARIMA(1,0,3)(0,1,1)

The residuals plot is not clear on what model is actually the best fit, both of them though are less accurate in the first phase of the data, which was already expected, and tend to improve before experiencing a slight increase in the last phase of the time series range.

Fit and Measure statistics

a) Winters Additive model

Number of Predictors		0
Model Fit statistics	Stationary R-squared	0.659
	R-squared	0.974
	RMSE	0.454
	MAPE	0.227
	MAE	0.332
	MaxAPE	1.124
	MaxAE	1.649
	Normalized BIC	-1.461
Ljung-Box Q(18)	Statistics	65.146
	DF	15
	Sig.	0
Number of Outliers		0

Table 5 - Winters Additive model: Model Statistics

The models fit statistics tell us that *Stationary R-Squared*, a measure that is preferable to the *R-squared* when there is a trend or a seasonal pattern, which is clearly our case, is positive and shows us that the model explains a good part of the overall variation of the time series.

The *Mean Absolute Error (MAE)*, with reference to our CPI values, is 0.332, which is not that substantial.

The *Mean Absolute Percentage Error (MAPE)* is 0.227, which is again quite low.

The *Maximum Absolute Error (MaxAE)* is the largest forecast error, and in this case is 1.649, which again if we take into account the values range [139.5, 151.2], it is not a substantial error, which is underlined by the *Maximum Absolute Percentage Error (MaxAPE)* which is 1.124.

The Root Mean Squared Error (RMSE) can be thought as the standard deviation of the error terms and we got 0.454, which is indicative of a good fit. The Ljung-Box statistic is not significant.

b) Arima(1,0,3)(0,1,1)

Number of Predictors		0	Winters Additive
Model Fit statistics	Stationary R-squared	0.975	0.659
	R-squared	0.969	0.974
	RMSE	0.492	0.454
	MAPE	0.237	0.227
	MAE	0.345	0.332
	MaxAPE	1.728	1.124
	MaxAE	2.5	1.649
	Normalized BIC	-1.202	-1.461
Ljung-Box Q(18)	Statistics	12.342	65.146
	DF	13	15
	Sig.	0.5	0
Number of Outliers		0	0

Table 6 - ARIMA(1,0,3)(0,1,1) model: Model Statistics

Both *Stationary R-Squared*, and *R-squared* are positive and explain in a substantial measure the variation in our series with values of 0.975 and 0.969 respectively, thus having better values than the Winters Additive model.

The *Mean Absolute Error (MAE)*, the *Mean Absolute Percentage Error (MAPE)* and the *Root Mean Squared Error (RMSE)*, are all slightly above the values of the Winters Additive model.

The *Mean Absolute Percentage Error (MAPE)*, and the *Maximum Absolute Error (MaxAE)* are a bit above the Winters Additive model, but not to a critical extent whereas the *Normalized BIC* suggests the Winters Additive to be the best fit, as its value is the lowest.

The *Mean Absolute Percentage Error (MAPE)* is 0.227, which is again quite low.

The *Maximum Absolute Error (MaxAE)* is the largest forecast error, and in this case is 1.649, which again if we take into account the values range [139.5, 151.2], it is not a substantial error, which is underlined by the *Maximum Absolute Percentage Error (MaxAPE)* which is 1.124.

The Root Mean Squared Error (RMSE) can be thought as the standard deviation of the error terms and we got 0.454, which is indicative of a good fit. The Ljung-Box statistic is though, on the limit of significance, but the autocorrelation plots, in the next section, somehow clarify the autocorrelation hypothesis.

Autocorrelation and Partial Autocorrelation of residuals

a) Winters Additive model

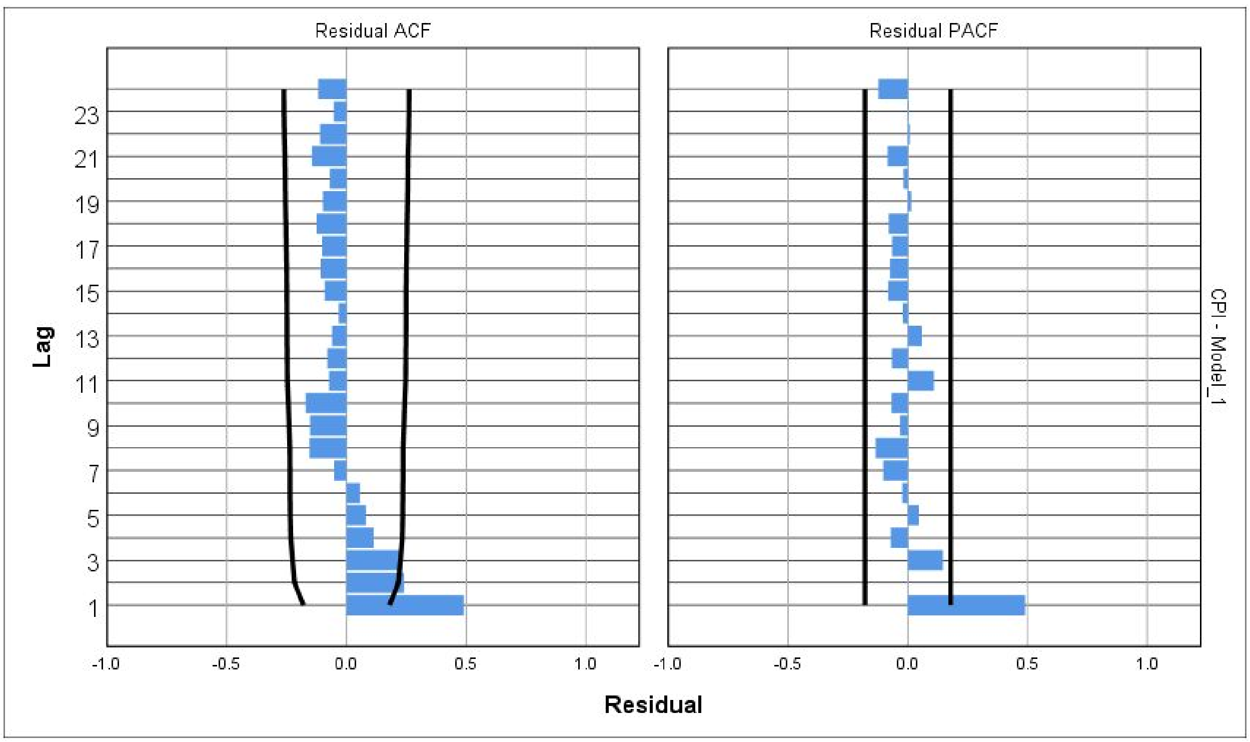

Fig 6 - Winters Additive: Residuals Autocorrelation and Partial Autocorrelation plots

In the 24 lags presented, the Autocorrelation function graph (ACF) shows significant autocorrelation in the first 3 lags, of those only the first is confirmed by the partial autocorrelation function (PACF), basically telling us that there is probably significantly auto correlated errors. So the Winters Additive model is probably failing to factor in some component seasonality, but then again this might be effect of the initial phase in such a short time series.

b) Arima(1,0,3)(0,1,1)

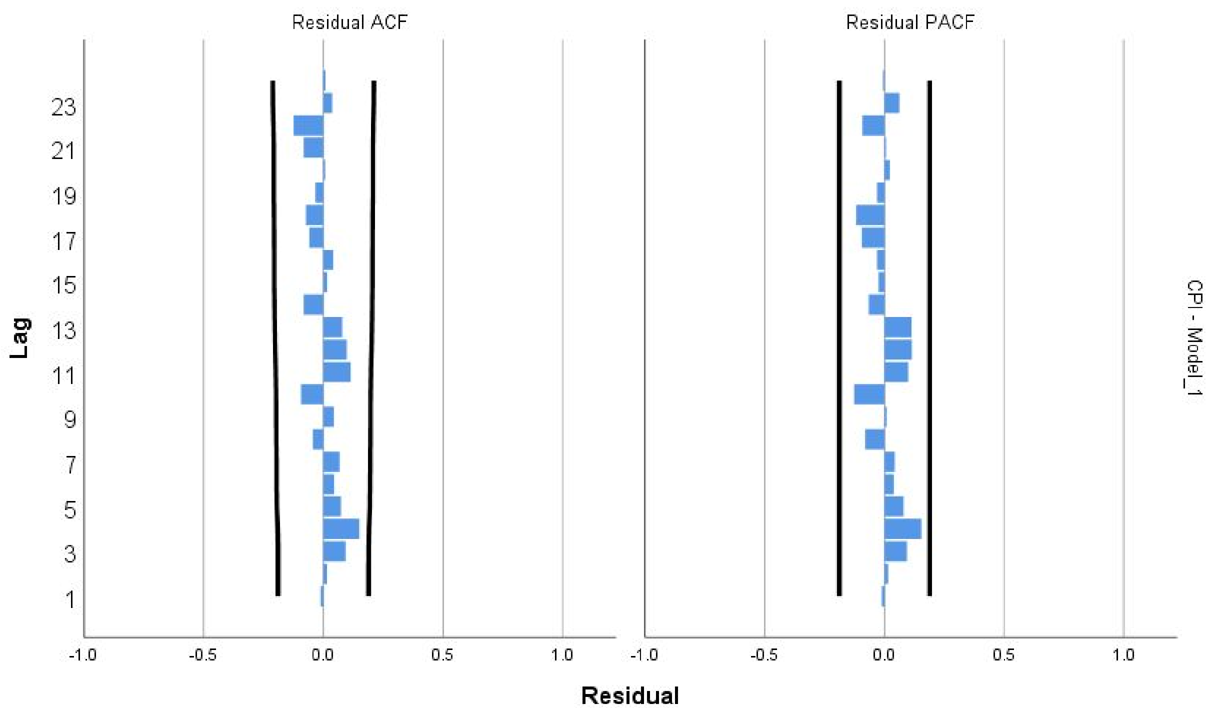

Fig 7 - ARIMA(1,0,3)(0,1,1): Residuals Autocorrelation and Partial Autocorrelation plots

In the ARIMA case there is no sight of significant autocorrelation between periodically lagged values, so the ARIMA model has passed this test.

Normality of Residuals

a) Winters Additive model

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual Winters Additive	0.07	120	.200*	0.951	120	0

[a] Lilliefors Significance Correction

Table 7 - Winters Additive: Residuals normality tests

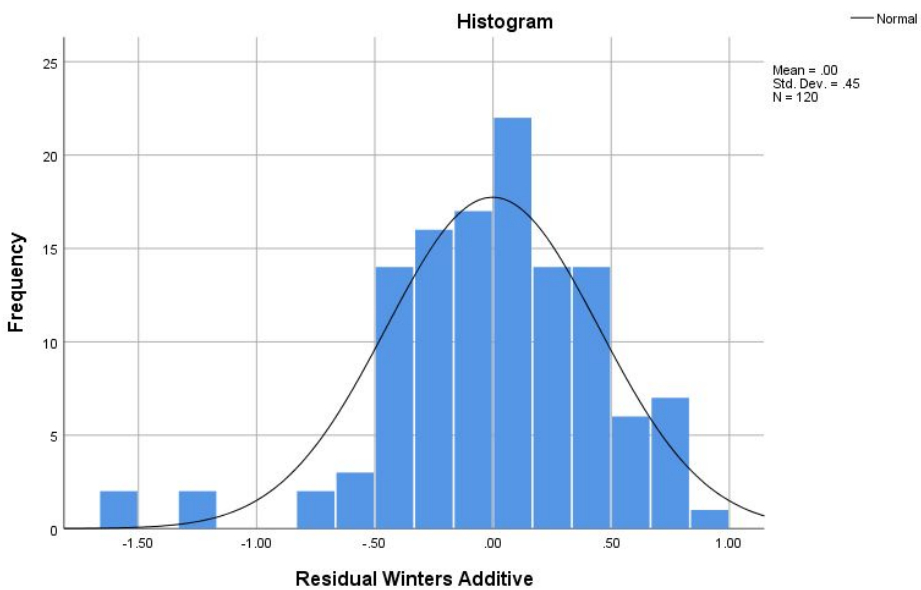

Fig 8 - Winters Additive: Residuals histogram

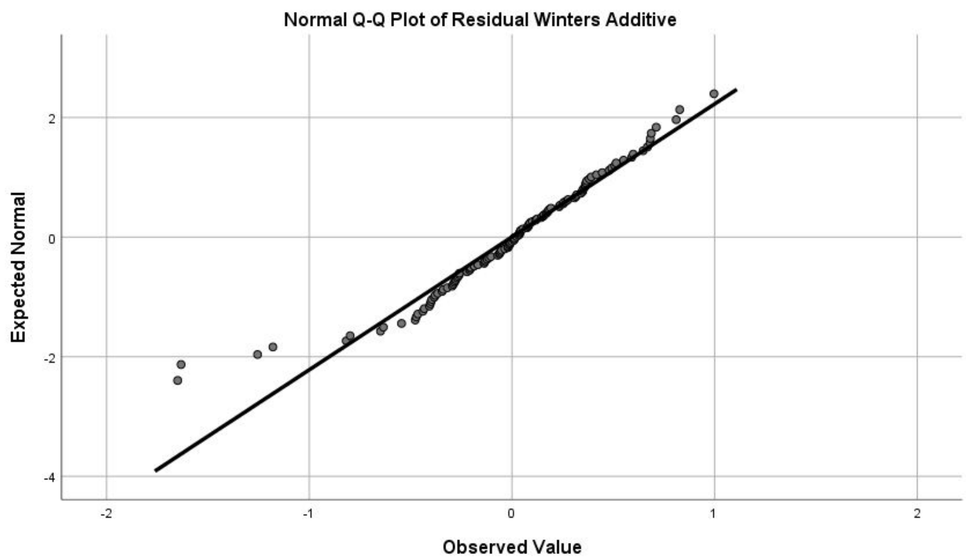

Fig 9 - Winters Additive: Residuals Normal Q-Q plot

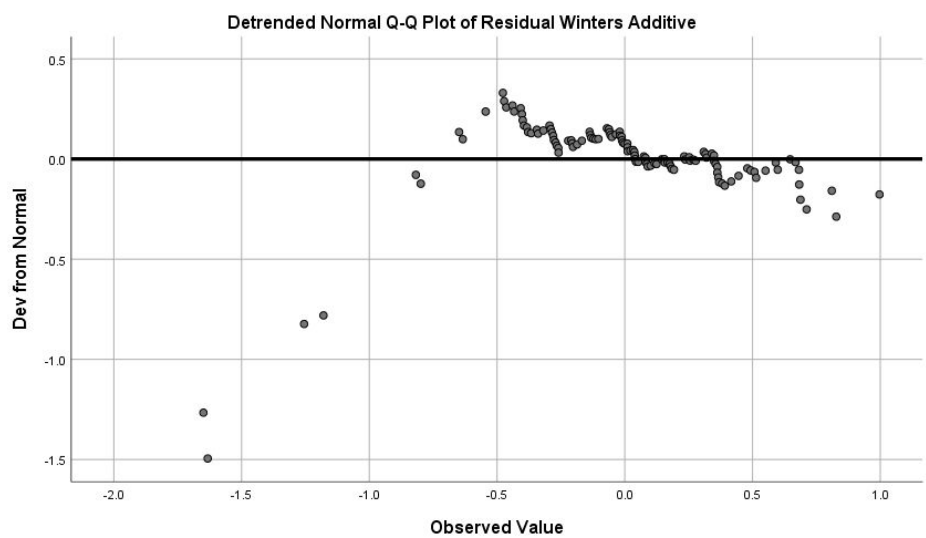

Fig 10 - Winters Additive: Residuals Detrended Normal Q-Q plot

The Kolmogorov test suggests we should accept the null hypothesis of normality, but the Shapiro-Wilk clearly indicates that we should reject that null hypothesis, and therefore the normality of the residuals.

The histogram resembles a bell-shaped curve and shows some skew to the left, that can also be spotted in the Normal Q-Q plot, but the normality is not that evident in the Detrended Normal Q-Q plot.

b) Arima(1,0,3)(0,1,1)

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual ARIMA(1,0,3)(0,1,1)	0.13	108	0	0.921	108	0

[a] Lilliefors Significance Correction

Table 8 - ARIMA(1,0,3)(0,1,1): Residuals normality tests

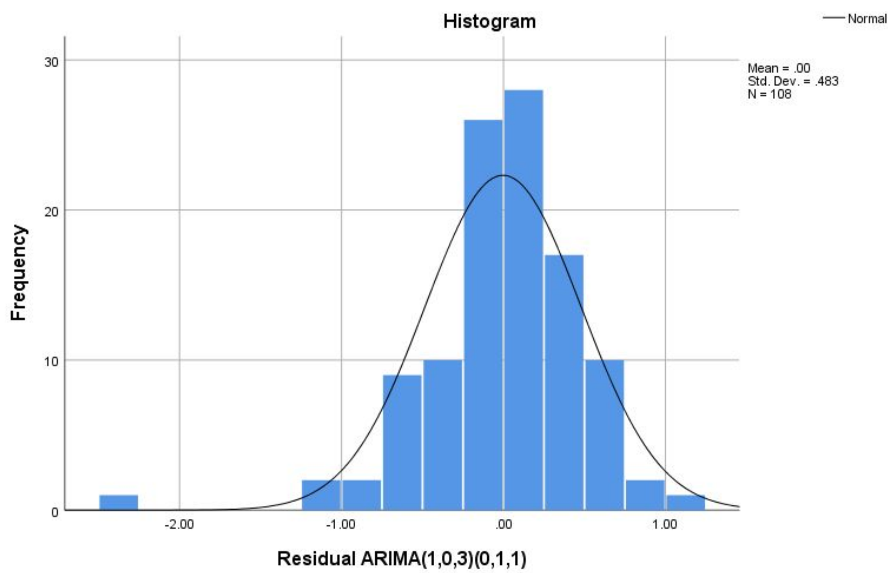

Fig 11 - ARIMA(1,0,3)(0,1,1): Residuals histogram

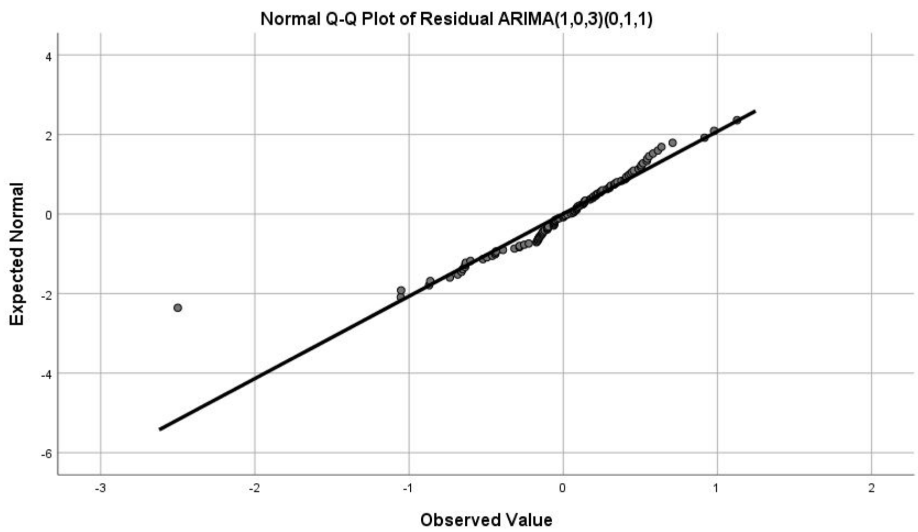

Fig 12 - ARIMA(1,0,3)(0,1,1): Residuals Normal Q-Q plot

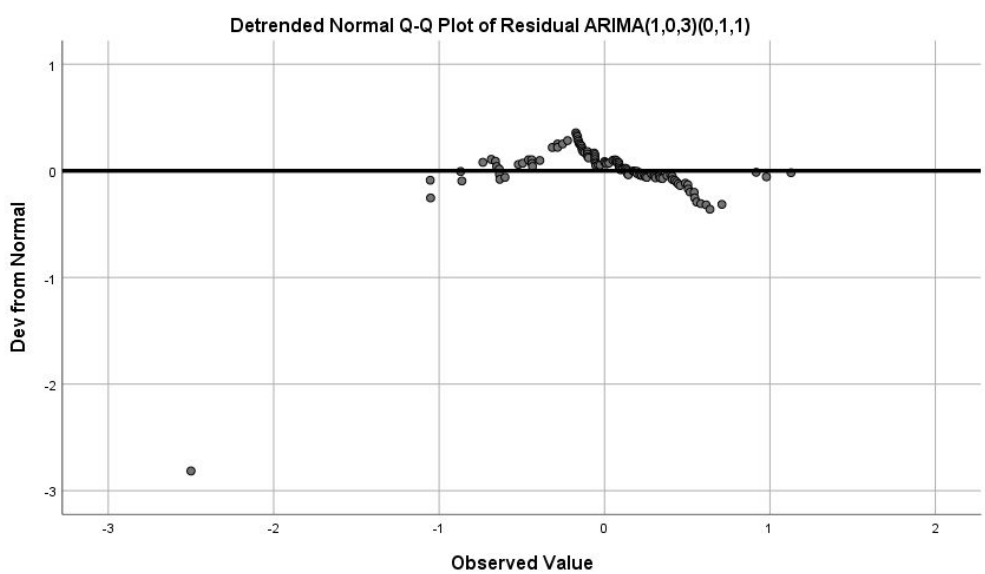

Fig 13 - ARIMA(1,0,3)(0,1,1): Residuals Detrended Normal Q-Q plot

Both normality tests indicate that we should reject the null hypothesis and therefore don’t accept the normality assumption of the residuals. While the histogram and the Normal Q-Q plot might suggest normality, with an outlier on the left, the Detrended Normal Q-Q plot does not confirm that suggestion. Overall we will assume the residuals as having a normal distribution, mostly due to the histograms and the Normal Q-Q plots, taking into account that the number of cases is only 120.

Model choice

The models analysis revealed that they are not that much different, in term of errors produced. Both of the models struggled with the first phase of the time-series, the steep climb of the CPI on the run-up to the crisis and the following dip are not that well modeled by both approaches and generate the most deviations in their residuals, even if the analysis couldn't find significant outliers. Also as the time series ages, both the models show a slight increase in their residuals. So the plot of *observed vs fit* values (Fig 5) shows similar accuracy on both cases. Also, the residuals normality analysis is not that conclusive, still allowing us the assumption of normality on both cases.

The Winters Additive model, having a non-significant parameter, the season related Delta, and some debatable autocorrelation of residuals in the first lag, which might be the reason for the Ljung-Box being on the verge of significance, still shows the better, lower, error outcomes, RMSE, MAPE, MAE, MaxAPE and MaxAE, but the differences are tiny, not substantial, while also a slightly better Normalized BIC.

The Arima(1,0,3)(0,1,1) having slightly higher errors and BIC, all its parameters are significant, and nevertheless has no sign of residuals autocorrelation, and last but not the least shows a substantial difference in its R-square stationary, which is 0.975 compared with 0.659 of the Winters Additive model, showing that it can explain better the variation of the time series.

For these reasons we choose the Arima(1,0,3)(0,1,1) model:

				Estimate	SE	t	Sig.
Consumer Price Index	No Transformation	AR	Lag 1	0.964	0.037	25.869	0
		MA	Lag 1	-0.591	0.093	-6.322	0
			Lag 2	-0.248	0.112	-2.216	0.029
			Lag 3	-0.272	0.103	-2.635	0.01
		Seasonal Difference		1			
		MA, Seasonal	Lag 1	0.89	0.224	3.973	0

Table 9 - chosen model: ARIMA(1,0,3)(0,1,1) parameters, all significant

Estimation and Forecasting on splitted data

In this section we split the sample data in order to do estimation of model parameters in the first part, and forecast training on the second part, enabling us to verify the forecasting performance against observed data. The estimation dataset was defined to include the date range Jan 2008 - Dec 2016, and the training dataset was defined with the date range Jan 2017 - Dec 2017.

We then computed the model ARIMA(1,0,3)(0,1,1), and we got the following parameters:

				Estimate	SE	t	Sig.
Consumer Price Index	No Transformation	AR	Lag 1	0.964	0.043	22.158	0
		MA	Lag 1	-0.679	0.096	-7.087	0
			Lag 2	-0.338	0.119	-2.846	0.005
			Lag 3	-0.338	0.106	-3.197	0.002
		Seasonal Difference		1			
		MA, Seasonal	Lag 1	0.948	0.612	1.548	0.125

Table 10 - Forecast ARIMA(1,0,3)(0,1,1) model parameters with separate estimation data

We must note the fact that the moving average (MA) parameter, of the seasonal component, is not significant, thus not explaining significantly data variation.

The following plot shows that the forecast data, follows closely the observed data.

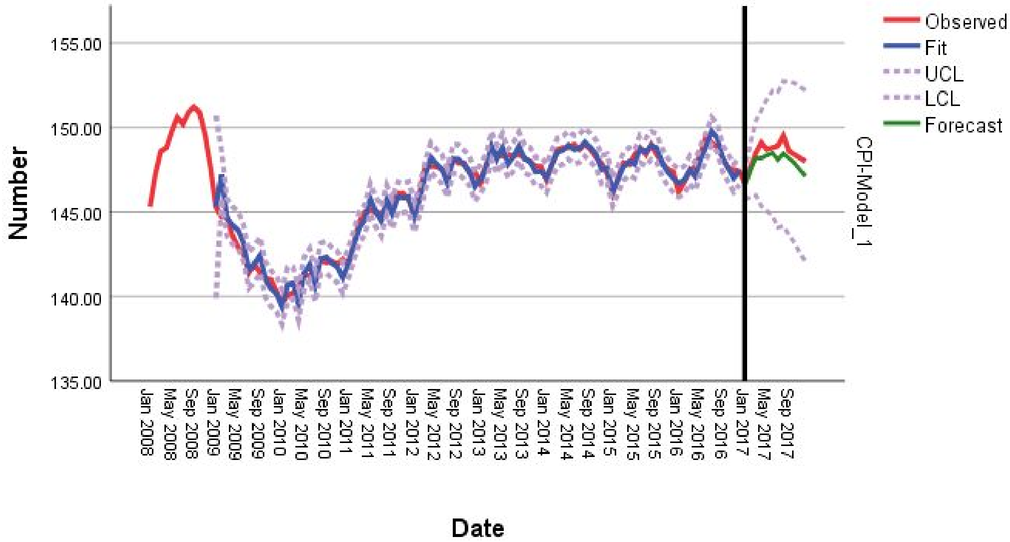

Fig 14 - Forecast ARIMA(1,0,3)(0,1,1) model Observed data against fit and Forecast data

The fit measures are presented in the following Table 11.

Number of Predictors		0	Selection Phase
Model Fit statistics	Stationary R-squared	0.977	0.975
	R-squared	0.97	0.969
	RMSE	0.498	0.492
	MAPE	0.232	0.237
	MAE	0.338	0.345
	MaxAPE	1.728	1.728
	MaxAE	2.5	2.5
	Normalized BIC	-1.156	-1.202
Ljung-Box Q(18)	Statistics	14.144	12.342
	DF	13	13
	Sig.	0.364	0.5
Number of Outliers		0	0

Table 11 - Forecast ARIMA(1,0,3)(0,1,1) model Fit measures with separate estimation data

The fit measures are generally slightly better, when compared to the ARIMA model computed in the previous selection phase. The Ljung-Box statistic is in this case clearly not significant, indicating no autocorrelation, and the Normalized BIC is slightly higher.

The analysis of the residual autocorrelations and partial autocorrelations does not show any significant autocorrelation, in any lag.

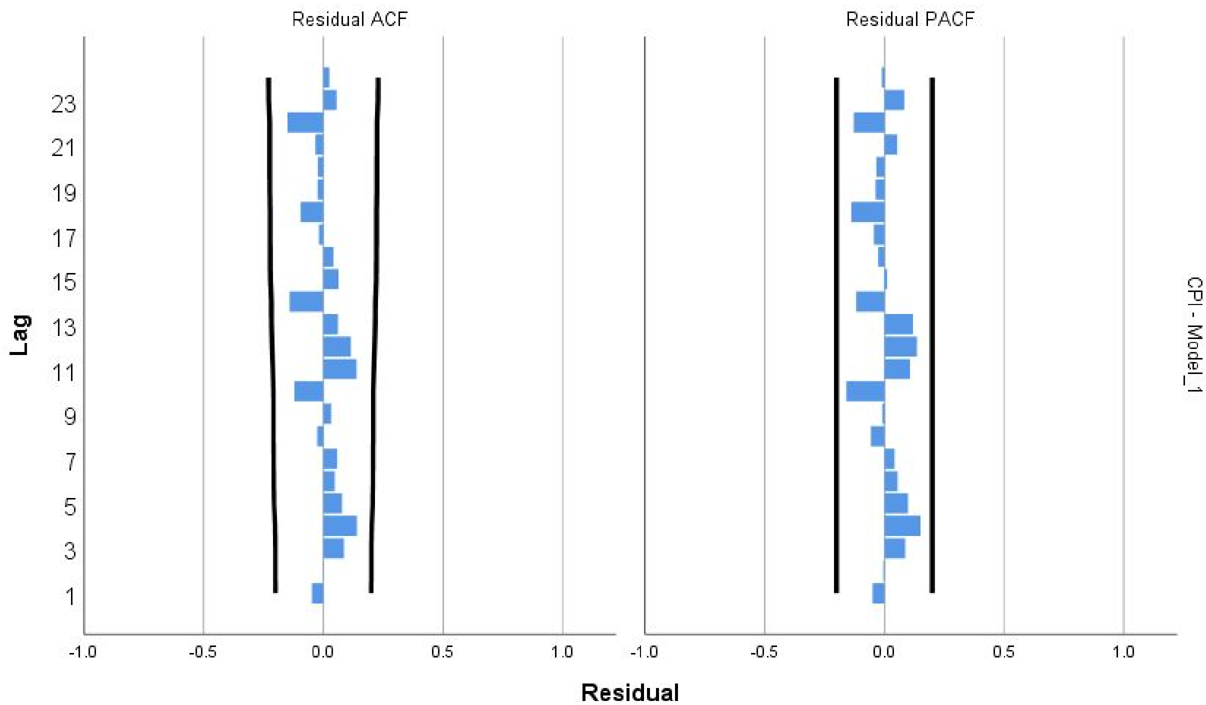

Fig 15 - Forecast ARIMA(1,0,3)(0,1,1) model residual autocorrelations and partial autocorrelations

Next we've studied the assumption of normality on the residuals.

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual Forecast ARIMA(1,0,3)(0,1,1)	0.12	96	0.002	0.913	96	0

[a] Lilliefors Significance Correction

Table 12 - Forecast ARIMA(1,0,3)(0,1,1) model residuals normality tests

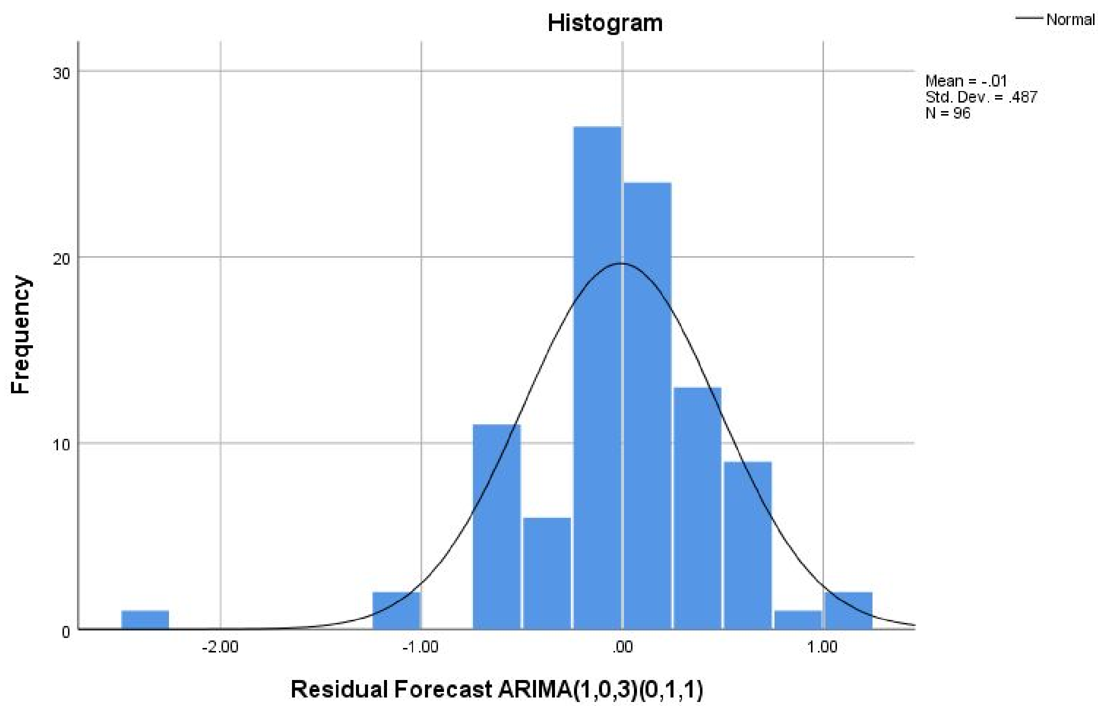

Fig 16 - Forecast ARIMA(1,0,3)(0,1,1) model residuals histogram

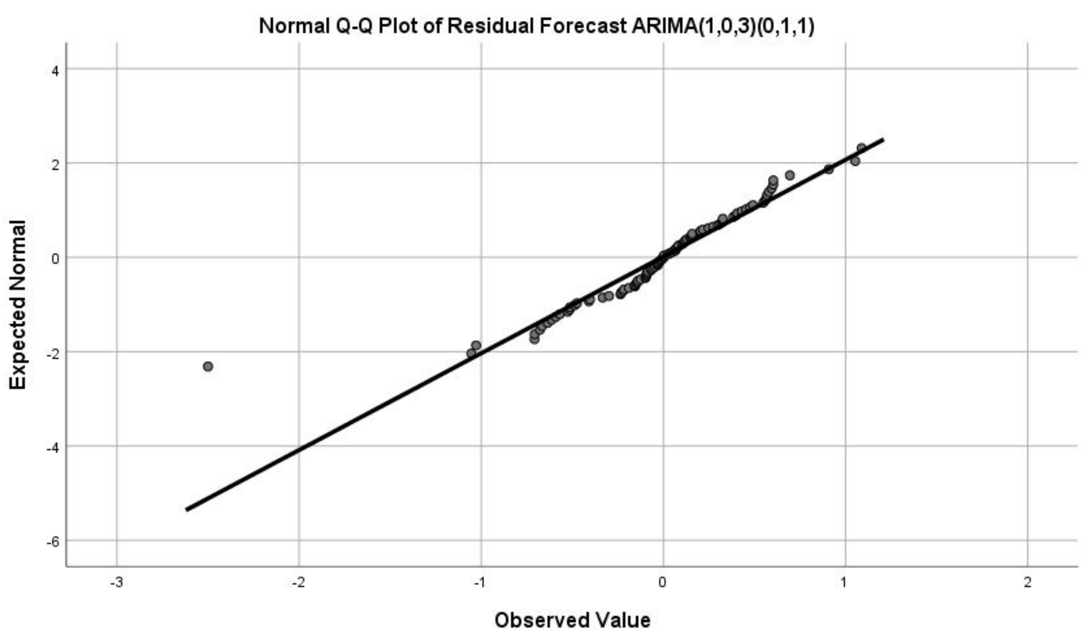

Fig 17 - Forecast ARIMA(1,0,3)(0,1,1) model residuals Normal Q-Q plot

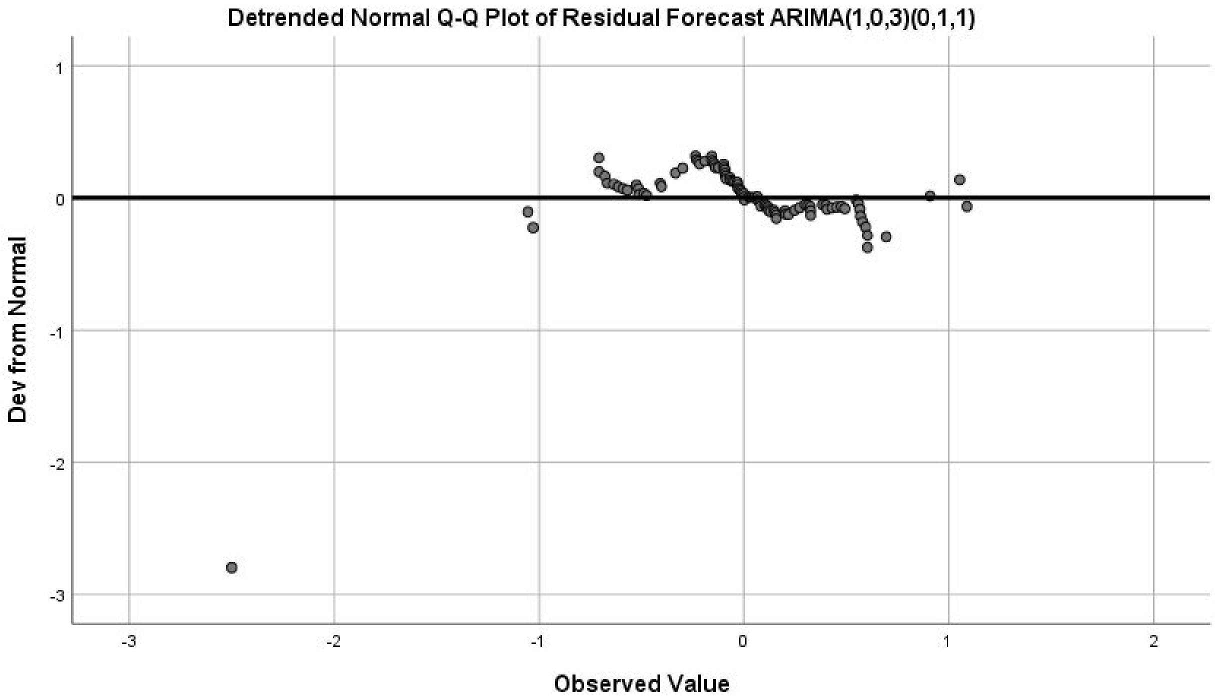

Fig 18 - Forecast ARIMA(1,0,3)(0,1,1) model residuals Detrended Normal Q-Q plot

Both normality tests statistics are non-significant, so we should reject the null hypothesis of normality based on it, however, from the histogram and Normal Q-Q plot, we will assume the residuals to be reasonably normally distributed.

We then plotted the residuals in a sequence chart.

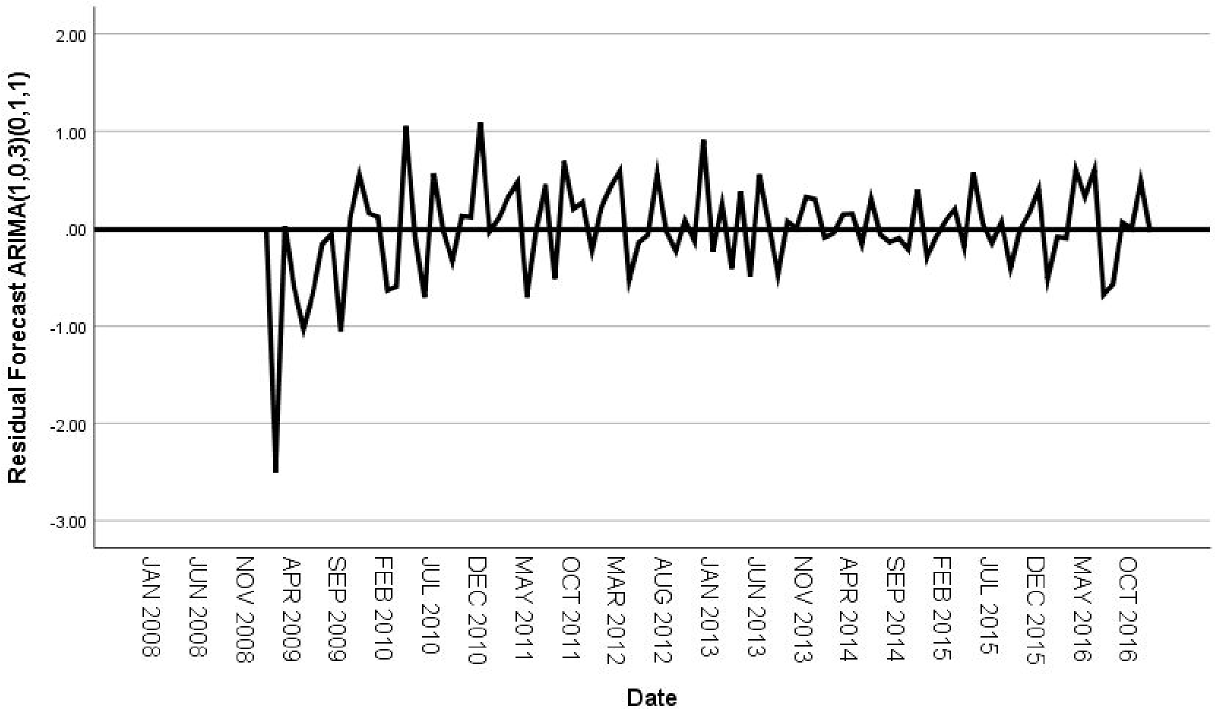

Fig 19 - Forecast ARIMA(1,0,3)(0,1,1) model residuals sequence chart

Despite the lack of randomness in the initial stages of the time series, that we've already seen before, related to the runup and dip of the crisis, after that, the residuals look reasonably random, and we can assume that the model has captured all the substantial components of trend and seasonality present in the time series, so we can in some extent, take the forecast as valid.

True Forecast

In this section we will produce a forecast from the moment our data sample finishes into the future, we will predict the monthly values for the whole year of 2018.

For this we will compute the model using all the available observed data.

Unsurprisingly, the model parameters are the exact same as the ones in the selection phase, as the sample data is the same.

				Estimate	SE	t	Sig.
Consumer Price Index	No Transformation	AR	Lag 1	0.964	0.037	25.869	0
		MA	Lag 1	-0.591	0.093	-6.322	0
			Lag 2	-0.248	0.112	-2.216	0.029
			Lag 3	-0.272	0.103	-2.635	0.01
		Seasonal Difference		1			
		MA, Seasonal	Lag 1	0.89	0.224	3.973	0

Table 13 - ARIMA(1,0,3)(0,1,1) model parameters

The model fit measures, autocorrelation and residuals analysis are therefore the same as presented before in the selection phase, and so all the considerations and remarks made there regarding the model, do apply here.

Next we plot the observed, fit and forecast data.

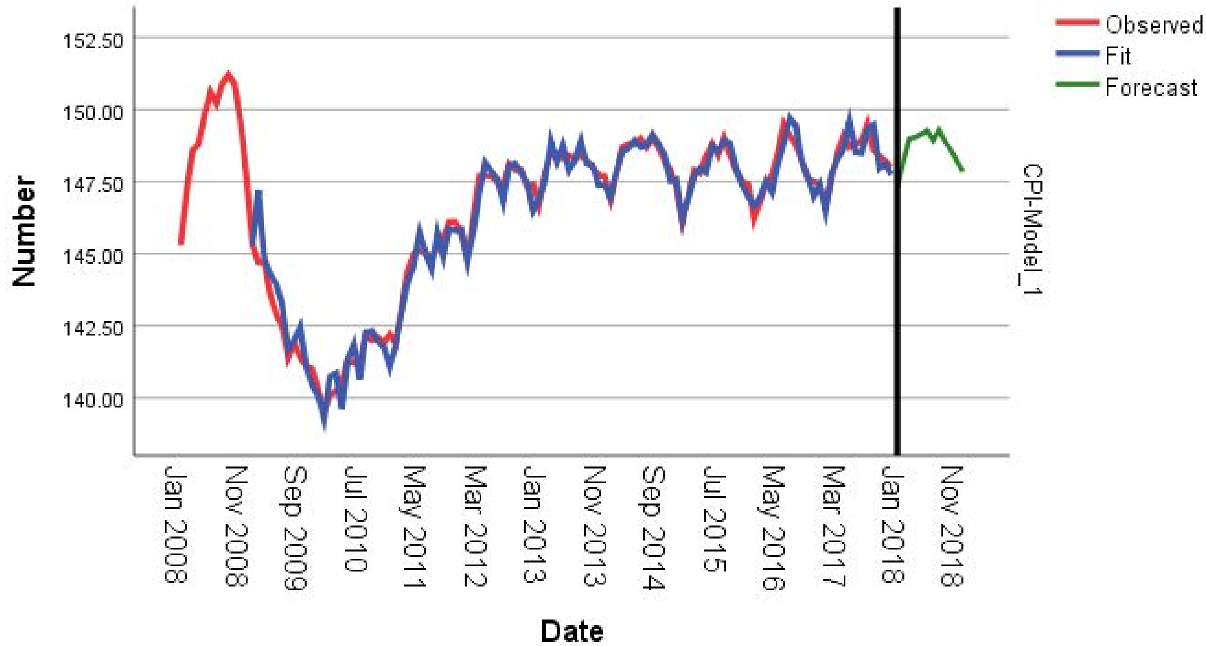

Fig 20 - ARIMA(1,0,3)(0,1,1) model: Observed vs Fit vs Forecast data

The forecast data seems to replicate correctly the seasonality and slight upbeat trend seen in the latest stages of the observed data.

Based on the back test we've made, we can be reasonably confident that the model is able to adapt to inflections of the real data, provided we don't predict too much into the future and by constantly incorporating the observed data and related residuals into the model.