Name: **Carlos Felippe Rostand Koetz**
VEE AS: **Time Series – Winter 2014**

Time Series - Student Project: MONTHLY AVERAGE CLAIM COST

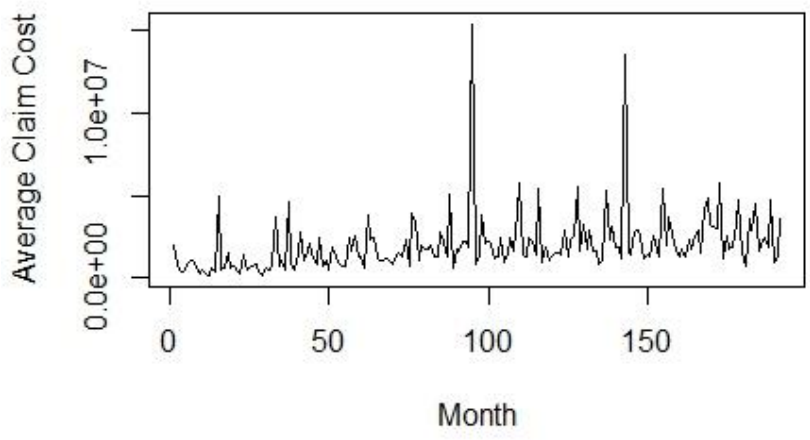
1 INTRODUCTION

The goal of this student project is to model monthly average claim cost (abbreviated as MACC) using the learned techniques of the online Time Series course. At a first glance, this time series seems to have a clear upward long-term pattern, which will be worked with and, by the end, presenting a suitable model for it.

2 DATA

The data series examined is from an interruption insurance business, taken from the following website (http://perso.univ-rennes1.fr/arthur.charpentier/SIN_1985_2000-PE.xls). It contains all claims registered by day of occurrence from Jan 1985 to Dec 2000 (months from 1 to 192). As my interest is to analyse the monthly average claim cost I modeled the data on this sense. Below I present the graph of the monthly average claim cost time series:

Graph 1: Monthly Average Claim Cost Series



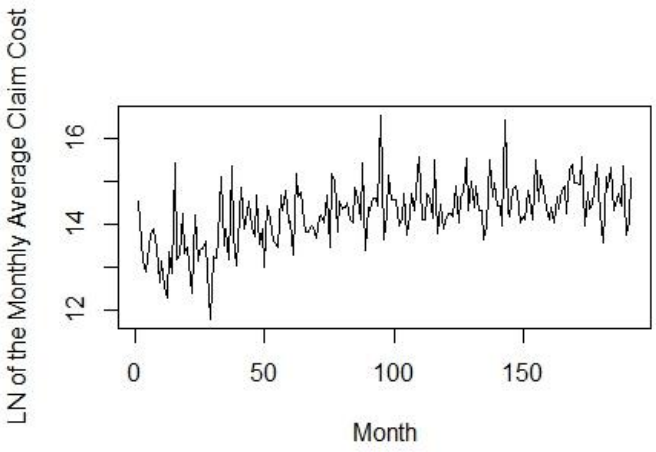
Note that there are upwards spikes at points 95 and 143, which represents extreme events (large losses - catastrophes). Besides those kinds of extreme values, I expect three things from this data:

- As it is the monthly average claim cost, its minimum value is zero and, theoretically, there is no maximum limit;

- As usually claim cost are modeled by positive skewed distributions like Lognormal or Gamma, it is appropriate to work with its natural logarithm instead with its real value; and
- As the series appears to suffer some influence from an inflation factor (that is geometrically applied) the uses of the natural logarithm, as described in the point above, becomes almost mandatory in this case.

After these three considerations above, I present the graph of the natural logarithm of the series as follows:

Graph 2: Natural Logarithm of the Series

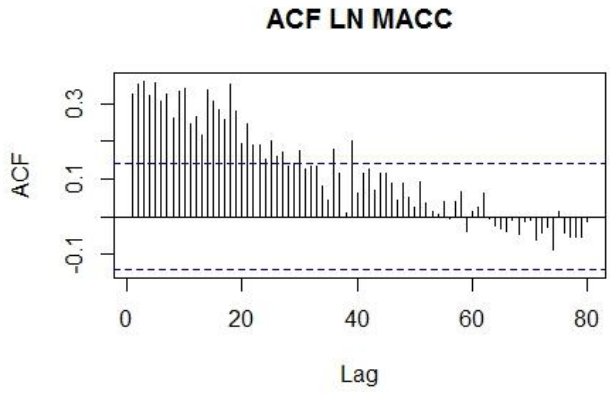


Those spikes, as highlighted before, continues here, but they are smoothed. It can highlight that another spike becomes apparent around month 30, but this one downwards.

3 MODEL SPECIFICATION

The first step of model specification is to look at the ACF graph of the natural logarithm of the series, as follows:

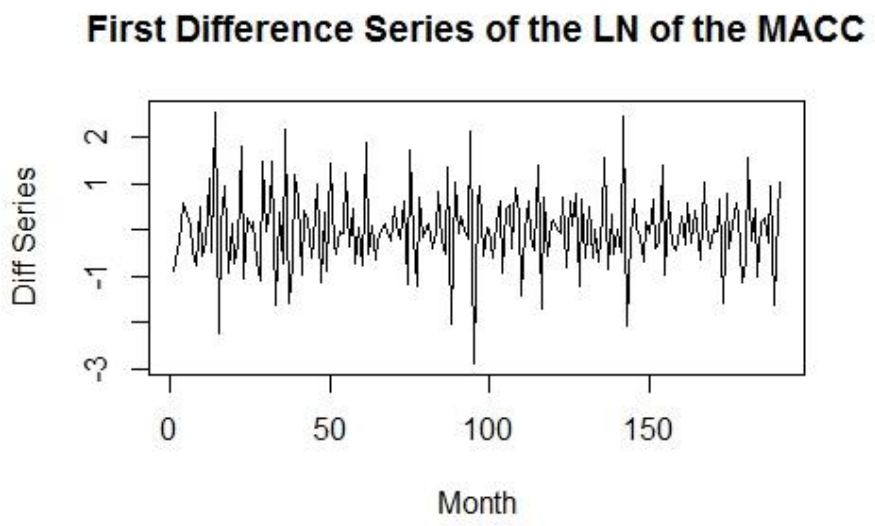
Graph 3: ACF of the LN MACC



It is usual to present few legs (around 20), but here it was increased the number presented in the chart to see clearly its behavior, so it was used leg equals 80. Analysing the ACF of

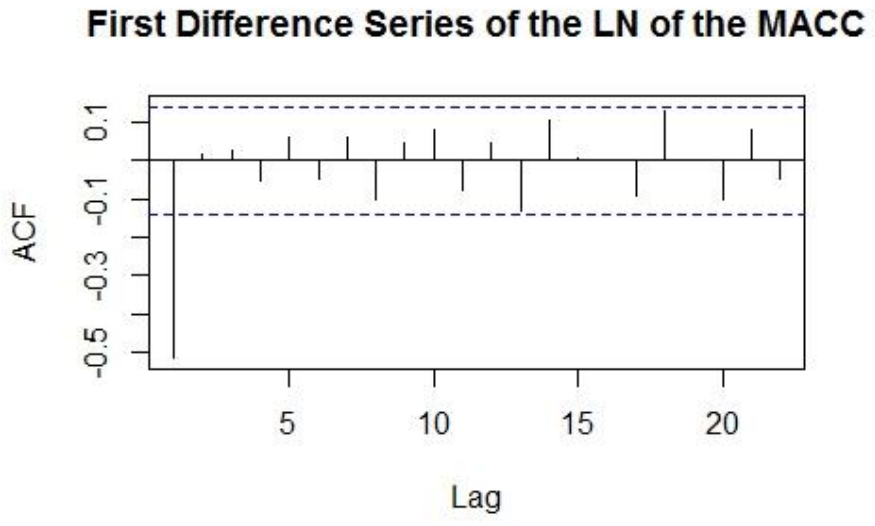
the series, it can be inferred that seasonality is not present and it is possible to see an upward general long-term trend in the series, which can be removed by taking its first difference and letting it stationary. The following chart is the first difference of the LN of the MACC:

Graph 4: First Difference of the LN of the MACC



The series now appears to oscillate around zero, it seem to have constant variance, and without any apparent pattern. Therefore, the natural logarithm and the first difference is enough to becomes the series stationary. Next, it is shown the ACF and PACF to determine the type of model to be used to model de series.

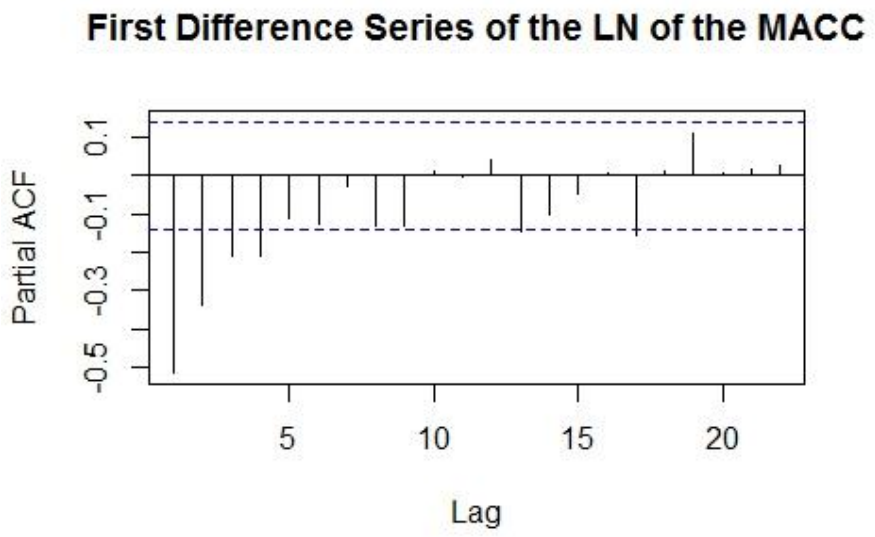
Graph 5: ACF of the First Difference of the LN of the MACC



The sample ACF suggests the existence of a MA(1) component because the autocorrelation cuts out after leg-1 and the remaining autocorrelations becomes close to zero as legs

increases; the alternating pattern between positive and negative values suggests a parameter θ close to 1.

Graph 6: PACF of the First Difference of the LN of the MACC



The PACF suggests that the series can be a MA (1). To confirm this model, it is presented the EACF table of this series, which can conclusively assist in the choice of the model:

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o  o  o  o
1 x x o o o o o o o o  o  o  o
2 x x o o o o o o o o  o  o  o
3 x x x o o o o o o o  o  o  o
4 x x x x o o o o o o  o  o  o
5 x x x x o o o o o o  o  o  o
6 x x o o o o x o o o  o  o  o
7 x x x x x o x x o o  o  o  o
```

The EACF table confirms that the first difference of the natural logarithm of the MACC is a MA (1). Therefore, the model is an ARIMA (0,1,1) (or an IMA(1,1)) of the natural logarithm of the MACC. The main equation is:

$$\nabla[lnY_t] = e_t - \theta e_{t-1}$$

4 MODEL FITTING

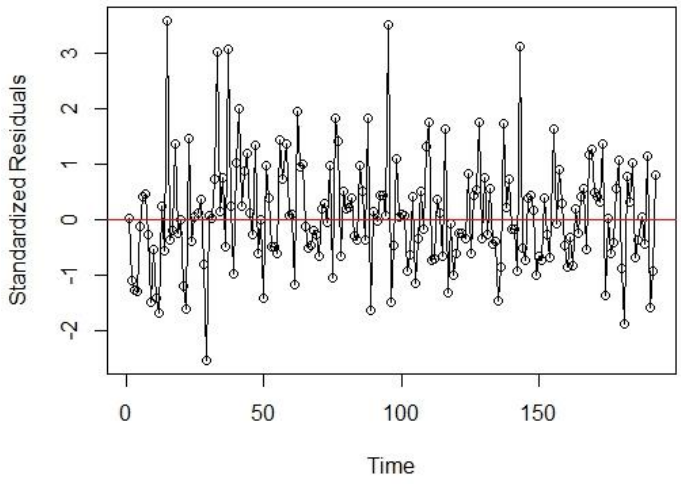
The model is fitted using R, which provides the parameter estimation by the method of maximum likelihood, as follows:

```
Call:
arima(x = db$log_ac, order = c(0, 1, 1))
Coefficients:
          ma1
      -0.9118
s.e.   0.0249
sigma^2 estimated as 0.3833:  log likelihood = -180.32,  aic = 362.64
```

5 MODEL DIAGNOSTICS

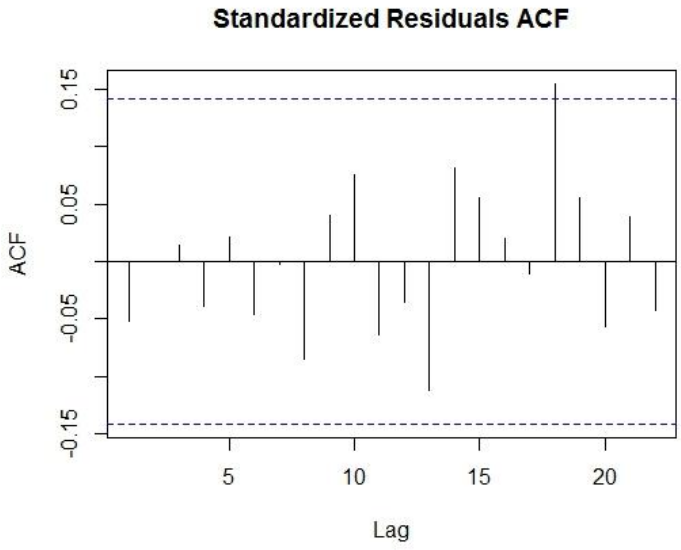
To verify the overall quality of the present model, is necessary to analyse its residuals. It is started by plotting the residuals:

Graph 7: Standardized Residuals



Even though, it looks to have few outliers (around 6 discrepant points among 191, that is, around 3%). The plot does not indicate any irregularities in the model. The next step in the model diagnostics is to look into the ACF of residuals:

Graph 8: Standardized Residuals ACF

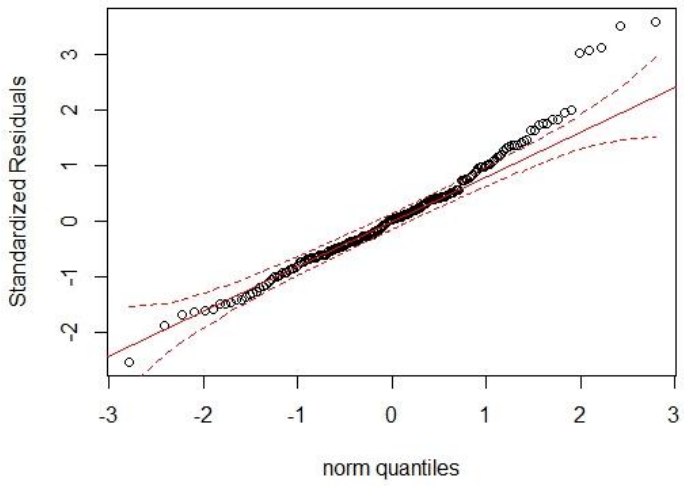


The sample ACF of the residuals shows that there is no significant correlation within the residuals, which confirms the selected model as adequate to the data set. To give a numerical value to this analysis it is proceed with the Ljung-Box test in R:

```
Box-Ljung test
data:  rstandard(m1.macc)
X-squared = 10.1042, df = 15, p-value = 0.8131
```

This leads to a p-value of 0.8131. So, there is no evidence to reject the null hypothesis that the error terms are uncorrelated. The last point to be examined is the normality of the residuals. It is started by plotting the Q-Q Plot with its interval of confidence:

Graph 9: Q-Q Plots of the Standardized Residuals



It is possible to see that the residuals present a heavy upper tail, probably because the extreme values in the data, and because of that, the normality of residuals might be breached. The Shapiro-Wilk test of normality was ran:

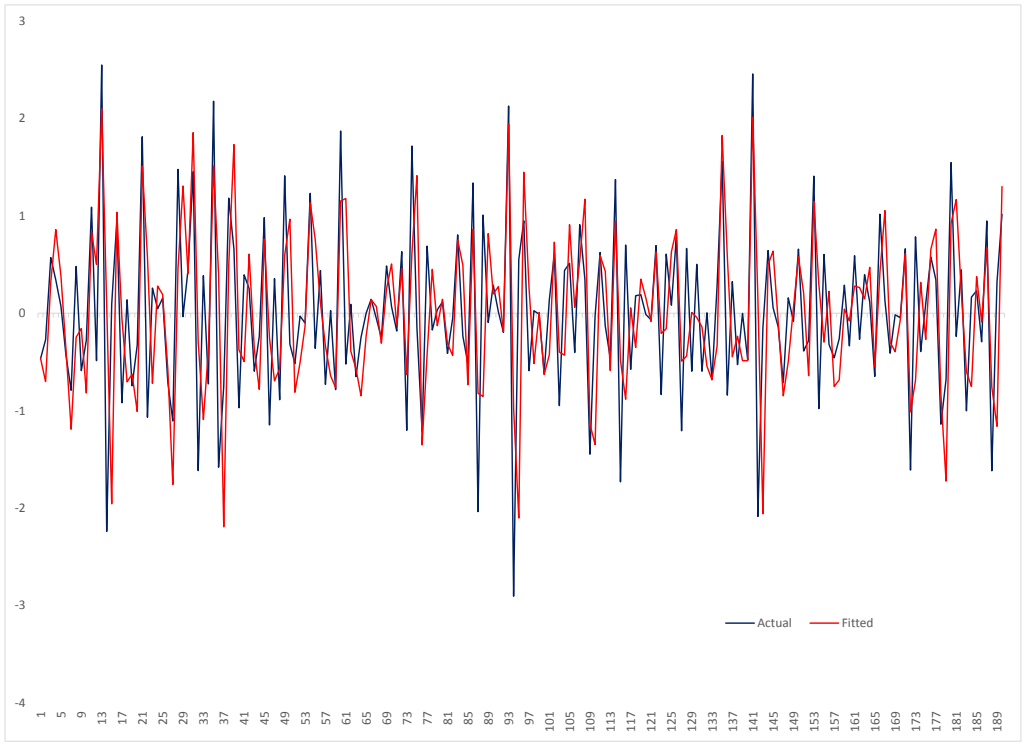
```
Shapiro-Wilk normality test
data:  rstandard(m1.macc)
W = 0.9658, p-value = 0.0001246
```

Therefore, based on the result of the test, that has the null hypothesis that the residuals are normally distributed, the normality is rejected at 5% of significance, and there is no evidence that the residuals are normally distributed.

6 ACTUAL VS FITTED

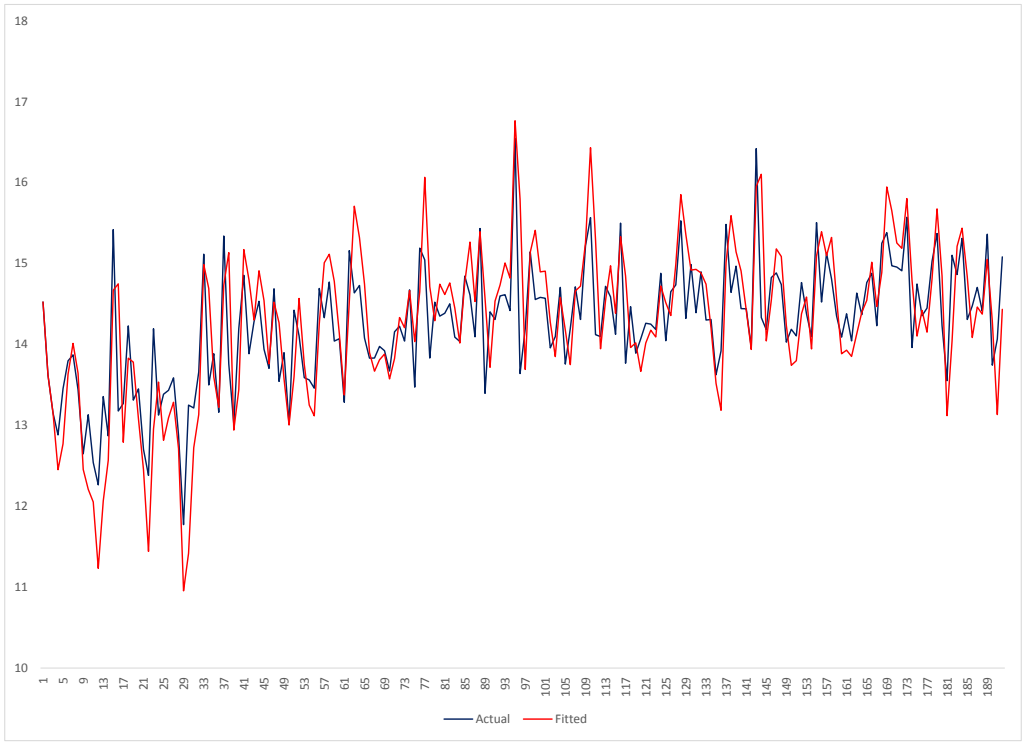
Notwithstanding, the residuals were rejected for normality, but I keep assuming that the model constructed so far works. So, will presented below the three stages of the fitted model. The first graph represents the MA (1) model over the difference of the natural logarithm of the time series:

Graph 10: Actual vs Fitted series of the Difference of the Log



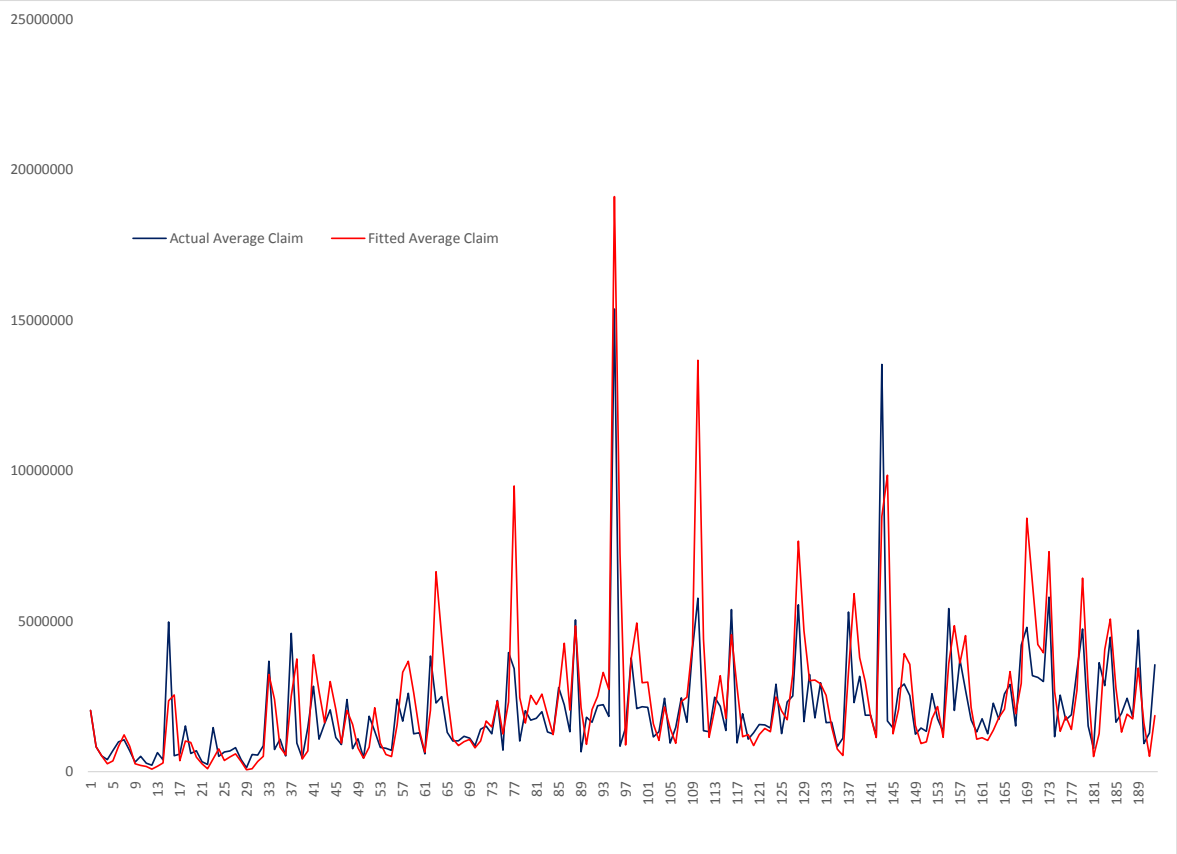
It is possible to see that the estimated MA (1) model captures the dependence structure of the difference of the natural logarithm of the MACC time series quite well. The second graph illustrates the comparison between the actual and the fitted time series of the natural logarithm (after integrate the differenced series), which sustain a good fitting:

Graph 11: Actual vs Fitted time series of the Natural Logarithm



Finally, the third illustration represents the original MACC time series against its fitted values using all the techniques described on this student project (this series is taken by applying the e number on the LN series shown in the previous graph):

Graph 12: MACC actual vs fitted



Although the fitting is not perfect, the result is satisfactory. I call the attention to the suspicious points (outliers) in the actual data values, that in the real life represents the extreme values. Those points makes the residuals of the transformed series not be normally distributed, but the overall model captures the essence of the data set.

7 CONCLUSION

The transformation on the original series was needed, because its nature that works with money and presents exponential growth. The model diagnostics of the transformed series indicates that the choice of an ARIMA (0,1,1) well represents the data, while remains relatively simple, and thus uses the principle of parsimony. As properly presented, the residuals of the transformed data are not normal, that can be understood by the kind of data that is worked here. Those high peaks can be interpreted as extreme values or large claims (usually covered by a reinsurance contract in real life), and without a depth analysis of it (that the material of the online course does not cover), it is need to keep them as they are presented in the data set. By the end, the model proposed seems to capture the essence of the data and fit it quite well, even with the extreme values inside it.