FORECASTING TIME SERIES

Homework #2

Prof. Lorenzo Pascual Caneiro

Team D: Dennis Pedersen, Eugen Wettstein, Alexandre Bouamama, Milan Medina, Tony Matta, Umut Varol, Vimal Ramakrishnan

Submitted on 11th October 2020

SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

Contents

Table of Figures

1. SUMMARY OF RESULTS

- For this task, Box-Jenkins iterative methodology was used to:
    - Achieve stationarity in the data,
    - Discover a suitable model with white noise residuals,
    - Fit the model and perform forecasting using related datasets, consecutively.
- Time series corresponding to quarterly earnings per share of Coca-Cola Company from Q1-1983 to Q3-2009 observed to be non-stationary in the mean and variance (heteroscedastic).
- Following the Box-Jenkins methodology, log transformation was applied to the original time series to remove heteroscedasticity.
- Suitable models identified for log-transformed data were:

    Assuming Φ_i and θ_j are coefficients of regular AR and MA terms i and j, respectively;
    - **Model 1:** ARIMA (1 , 1 , 5) (2, 1, 0) [4] *with $\theta_i = 0$ for i=1,2,3,4*
    - **Model 2:** ARIMA (4 , 1 , 0) (2, 1, 0) [4] *with $\Phi_i = 0$ for i=3*
    - **Model 3:** ARIMA (8 , 1 , 5) (0, 1, 0) [4] *with $\Phi_i = 0$ for i=1,2,3,4,6,7,8 and $\theta_i = 0$ for i=3*
    - **Model 4:** ARIMA (0 , 1 , 1) (0, 1, 1) [4] *This model was determined by auto.arima function*
- Models' forecasting performance were assessed using MSFE (Mean Squared Forecasting Error) and MAPE (Mean Absolute Percentage Error) criteria based on recursive and rolling out of sample forecasting schemas.
- Model 4 is the best performing forecasting model according to MSFE for horizons from 1 to 4 for both recursive and rolling schemas. According to MAPE, Model 1 outperforms Model 4 for horizons 1 and 2; yet, Model 4 becomes the best model for horizons 3 and 4, again. Considering the problem context; for a 1-year financial forecast (4 quarters), it was decided that Model 4 could be selected as the optimal predictor.
- All models have residuals that are non-normal, hence a GARCH model could be used to predict the future volatility of the Coca Cola Earnings.

		Model 1	**Model 2**	**Model 3**	**Model 4**
		ARIMA (1 , 1 , 5) (2, 1, 0)$_4$ with $\theta_i = 0$ for i=1, 2, 3, 4	ARIMA (4 , 1 , 0) (2, 1, 0)$_4$ with $\Phi_i = 0$ for i=3	ARIMA (8 , 1 , 5) (0, 1, 0)$_4$ *with $\Phi_i = 0$ for i=1,2,3,4,6,7,8 and $\theta_j = 0$ for j=2,3*	ARIMA (0 , 1 , 1) (0, 1, 1)$_4$
Recursive Scheme	**MSFE[1]**	0.00206	0.00225	0.00203	0.00168
	MSFE[2]	0.00391	0.00410	0.00378	0.00311
	MSFE[3]	0.00540	0.00495	0.00451	0.00365
	MSFE[4]	0.00659	0.00516	0.00483	0.00364
	MAPE[1]	4.90549	5.72432	5.53809	5.22456
	MAPE[2]	6.70509	7.12616	7.50393	6.78305
	MAPE[3]	7.65266	7.50182	7.65597	7.20212
	MAPE[4]	8.46339	8.00900	7.26878	6.83613
	MSFE[1]	0.00211	0.00232	0.00201	0.00167

	MSFE[2]	0.00394	0.00413	0.00373	0.00309
	MSFE[3]	0.00546	0.00495	0.00441	0.00362
	MSFE[4]	0.00667	0.00514	0.00467	0.00361
Rolling Scheme	**MAPE[1]**	4.90350	5.77189	5.46916	5.22689
	MAPE[2]	6.68224	7.13422	7.41930	6.79147
	MAPE[3]	7.64676	7.51091	7.60718	7.19900
	MAPE[4]	8.48043	8.00631	7.18177	6.82714

Table 1. Summary Table for Forecasting Performance of the Models

2. METHODOLOGY

In order to find suitable forecasting models for the quarterly earnings per share of the Coca-Cola Company time series, the Box-Jenkins iterative methodology was utilized.

According to this methodology, the first step is to observe and test the stationarity of the time series. In case the time series is not stationary, it must be transformed with the appropriate operation (seasonal and regular differencing, log) before proceeding to the second step of deciding on the best fitting model based on the estimated autocorrelations. In the third step, the parameters of the identified model are estimated with maximum likelihood. In the fourth step, residual diagnostics is performed and checked for white noise. In case the residuals are not white noise, the second iteration over the process starts, until white noise is achieved in the residuals as well. As a final step the point and interval predictions can be calculated and compared among models.

It is worth noting that a 5% level of significance was used for the assessment of all statistical test results (p-value) mentioned. Also, all insignificant coefficients identified in the models were set to zero in order to simplify the models and reduce the computing power and time required. This approach could be discussed since a priori it is not known if some of those variables would be significant with larger sample sizes.

Statistical tests and their respective hypothesis used in the analysis are:

1. **Augmented Dickey-Füller (ADF) Test**
   Augmented Dickey-Füller (ADF) Test was utilized to test for the stationarity in the mean with the following hypothesis:

   H_0: Data is not stationary in the mean, a unit root is present in an autoregressive model

   H_a: Data is stationary in the mean (must be checked for different lags)

2. **Osborn-Chui-Smith-Birchenhall (OSCB) Test**
   Osborn-Chui-Smith-Birchenhall (OSCB) Test was utilized to test for the need of seasonal differencing to achieve stationarity in the mean with the following hypothesis:

   H_0: A seasonal unit root is present in a seasonal autoregressive model

   H_a: No seasonal unit root required

3. **Shapiro-Wilk Test**
   Shapiro-Wilk Test was utilized as a goodness of fit test for normal distribution with the following hypothesis:

   H_0: Data is normally distributed

H_a: Data is not normally distributed

4. **Ljung–Box Test**

   Ljung–Box Test was utilized to check for serial autocorrelations in the data with the following hypothesis:

   H_0: $\rho_1 = \rho_2 = \ldots = \rho_m = 0$ (Data is uncorrelated)

   H_a: At least one is nonzero (Data is correlated)

   where ρ_n is the autocorrelation coefficient, m is the number of lags being tested

5. **Student's t-test**

   Student's t-test was utilized to test for mean being equal to 0 with the following hypothesis:

   H_0: Mean is equal to zero

   H_a: Mean is different than zero

3. DATA ANALYSIS

3.1 Original Data

3.1.1 Visual Assessment

To get an initial overview of the data, the Coca Cola earnings time series, ACF and PACF was plotted (Appendix 1). Appendix 1 gives a strong indication that the Coca Cola Earnings time series is not stationary in the mean (increasing trend) nor in the variance (increase in variance with time). This is also supported by the slow decay to zero in ACF; and ACF and PACF having several lags with significant correlation.

3.1.2 Formal Tests

Stationarity in the mean

```
> ndiffs(CC_TS, alpha=0.05, test=c("adf"))
[1] 1
```

According to the ADF Test, the required minimum number of regular differences to achieve stationarity in the mean is 1. Thus, in parallel with the visual interpretation; it is concluded that the **data is not stationary in the mean.** Hence, a transformation with d= 1 is needed.

OSCB: Osborn-Chui-Smith-Birchenhall test

```
> nsdiffs(CC_TS,m = 4,test=c("ocsb"))
[1] 1
```

According to the OSCB test, the number of seasonal differences to achieve stationarity is 1. Note that m = 4, because of the nature of the data, quarterly financial data.

3.2 Transformations

3.2.1 Log transformation

In order to achieve stationarity in the mean and the variance in the data, seasonal and regular differencing as well as log transformations are needed. Yet, taking logarithms after differencing causes production of NAs due to negative numbers resulting from differencing. Thus, log transformation was performed in the first step to achieve stationarity in the variance.

As seen in Appendix 2, the time series becomes stationary in the variance but not in the mean. That can be confirmed by running the ADF test on the logged data. This produces the same result as the test performed on the original data, i.e. that a regular differencing of order 1 is needed to achieve stationarity in the mean.

3.2.2 Regular and Seasonal Differencing

In Appendix 3, the time series looks stationary in the mean and in the variance after incorporating the regular differencing (d=1). This is confirmed by the ADF test which outputs the number of differencing needed to achieve stationarity as zero (0). However, as this time series is seasonal by nature (quarterly earnings), this effect can be observed from the significant correlations displayed in ACF and PACF. In parallel, OSCB test outputs the number of seasonal differences needed in lag 4 as 1.

As the seasonal differencing of order 1 (lag-4 differencing) is performed, it can be observed that seasonality effect is eliminated, and the data is stationary in the mean and the variance.

From the initial analysis of the time series it can be concluded that a seasonal model would represent the log-transformed data the best with seasonal term s = 4. As the stationarity was achieved after 1 regular and 1 seasonal differencing; d = 1 and D = 1, it follows that all subsequent models will be defined given the parameters d=1, D=1, s=4. Subsequently, different values for p, q, P, and Q will be tested and evaluated one by one. Appendix 4, which displays the log Coca Cola Earnings after achieving stationarity (d=1, D=1), lays the foundation for choosing the different values (p, q, P, Q).

3.3 Model 1

Looking at the ACF in Appendix 4, lags 1, 4, and 5 are statistically different from zero. Hence, it could be argued that the series can be represented with regular MA terms of order 5 as Q=0 and q=5 since 5 is the last lag significantly different than zero. From the PACF it is observed that lags 1, 4, and 8 are lags with significant coefficients. Thus, it could suggest that seasonal AR terms may exists in the model with order 2: P=2. This leaves p=1 as the final component, and Model 1 can be defined as ARIMA (1, 1, 5)(2, 1, 0)[4].

After the initial fitting of the model, it was observed that p-values for MA1, MA2, MA3, and MA4 indicate insignificance of these features and can be set to zero instead. Thus, Model 1 was then revised to include AR1, MA5, SAR1, and SAR2 (Table 2).

Moreover, the residuals of Model 1 were tested for WN properties. Appendix 5 displays the results of the tests conducted and shows that residuals of this model follow WN. Hence according to the Box-Jenkins methodology, it could be a candidate to perform forecasting for further comparison. The results of the ADF, t-test, Box-test, WN test, and Shapiro test for the residuals are shown in Table 2.

	Ndiffs/p-value	Conclusion
ADF	ndiffs = 0	Residuals are stationary
t-Test	0.8132	Fail to reject H_0, residuals have 0 mean
Ljung-Box	0.3604 (lag 12)	Fail to reject H_0, residuals are uncorrelated
WN	0.1465	Fail to reject H_0, residuals are WN
Shapiro	1.451e-07	Reject H_0, residuals are not normally distributed

Table 2. Model 1 - Residual Diagnostics Tests

3.4 Model 2

Model 2 was defined as ARIMA (4, 1, 0)(2, 1, 0)[4]. The logic for choosing P=2 is the same as for Model 1. When looking at the PACF it was seen that lag 5 was significant so p=5 was the starting point for Model 2, which was changed to p=4 after seeing that AR5 was insignificant. AR3, and SAR1 were removed from the final model as well due to being insignificant. Appendix 6 and Table 4 confirm that the residuals are WN and thus the model can be used for further forecasting.

	Ndiffs/p-value	Conclusion
ADF	ndiffs = 0	Residuals are stationary
t-Test	0.7651	Fail to reject H_0, residuals have 0 mean
Ljung-Box	0.7784 (lag 16)	Fail to reject H_0, residuals are uncorrelated
WN	0.2127	Fail to reject H_0, residuals are WN
Shapiro	4.486e-05	Reject H_0, residuals are not normally distributed

Table 3. Model 2 - Residual Diagnostics Tests

3.5 Model 3

It is also possible to discard the seasonal factors P, and Q. Doing this implies choosing the last lags that are significant in ACF and PACF as regular MA and AR terms, respectively. Again, using Appendix 4's PACF showing that lag 8 as significantly different from zero, we could propose p=8 with P=0. From the ACF having lag 5 still significantly different than zero, we could propose using q=5 and Q=0. Given this logic, Model 3 was defined as ARIMA (8, 1, 5)(0, 1, 0)[4]. Within Model 3 coefficients of AR1, AR2, AR3, AR4, AR6, AR7, AR8, MA2, and MA3 were set to zero as they turned out to be statistically insignificant. Looking at the PACF in Appendix 7; lag 18 could be perceived as significant. However, after performing Ljung-Box test, it was seen that correlations are not significant as shown in Table 5.

	Ndiffs/p-value	Conclusion
ADF	ndiffs = 0	Residuals are stationary
t-Test	0.4504	Fail to reject H_0, residuals have 0 mean
Ljung-Box	0.6691 (lag 18)	Fail to reject H_0, residuals are uncorrelated

WN	0.0965	Fail to reject H_0, residuals are WN
Shapiro	8.469e-06	Reject H_0, residuals are not normally distributed

Table 4. Model 3 - Residual Diagnostics Tests

3.6 Model 4

Model 4 was found by using the auto.arima() function in R to compare the automatic version viz a viz the models found manually. auto.arima() defined Model 4 as ARIMA(0, 1, 1)(0, 1, 1)[4]. Looking at Appendix 8 and Table 6, it was concluded that this model has WN residuals and thus can be used as a potential model for further forecasting performance comparison.

	Ndiffs/p-value	Conclusion
ADF	ndiffs = 0	Residuals are stationary
t-Test	0.5664	Fail to reject H_0, residuals have 0 mean
Ljung-Box	0.3103/0.1824/0.2788 (lag 5/9/18)	Fail to reject H_0, residuals are uncorrelated
WN	0.3020	Fail to reject H_0, residuals are WN
Shapiro	8.49e-07	Reject H_0, residuals are not normally distributed

Table 5. Model 4 - Residual Diagnostics Tests

4. MODEL COMPARISON AND CONCLUSION

Figure 9 displays the displays the realized values (in black) of the last 24 quarters and the point estimates of 4 different models based on recursive out-of-sample forecasting scheme for horizon 1. Graphs for recursive scheme for horizons 2, 3 and 4 can be found in the Appendix 9-11. Additionally, the plots with a rolling scheme for horizons 1, 2, 3, and 4 can be found in Appendix 12-15, which as it turns out does not change the final conclusion.

Based on the plots obtained, it can be observed that Model 1 and Model 2 are similar, as both deviate after the first peak, as well as for n = 23 (the second to last data point) with a comparable amount for horizon 1. This deviation becomes even higher for horizon 2, 3 and 4, consecutively, for both.

This is also true for Model 3 and Model 4. The difference seems to be that Model 3 and Model 4 deviate slightly less from the original data points, as the direction of the deviations appear to be the same for all models for horizon 1. One can observe clearly that Model 3 and 4 deviates less than Model 1 and 2 as the horizons increase yet, Model 4 yields the best fit among all. Hence, it can be concluded that Model 4 provides more accurate, robust and reliable results for various horizons for the prediction of the past 24 data points of Coca Cola quarterly earnings.

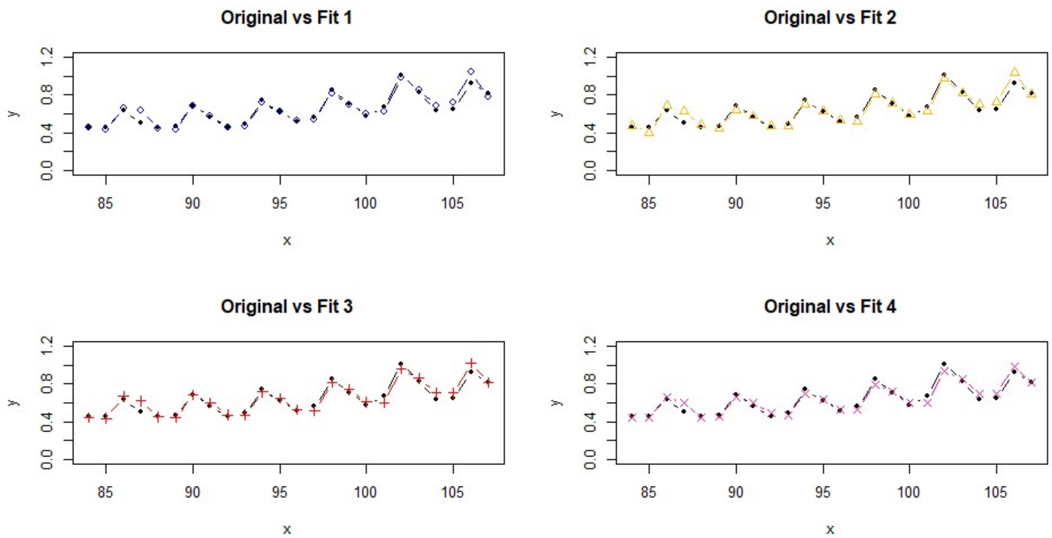

Figure 1. Different Model Predictions vs. Realizations for Last 24 Quarters

Table 6 and 7 display the MSFE and MAPE for the four fitted models with a recursive forecasting scheme. According to the MSFE Model 4 outperforms all other models, even though Model 3 is very similar for all time horizons. Model 2 seems to be the worst performing model in short term (up to 2 quarters ahead), whereas Model 1 performs worse in the longer horizons (3 and 4 quarters ahead). According to the MAPE metric, the results are different where Model 1 outperforms for 1 and 2 quarters ahead (short term), and Model 4 for the three and quarters ahead forecast (long term). Yet, it is worth noting that for both performance criteria, Model 4 provides the best results overall.

MSFE				
Horizon	fit_1	fit_2	fit_3	fit_4
1	0.00206	0.00225	0.00203	0.00168
2	0.00391	0.00410	0.00378	0.00311
3	0.00540	0.00495	0.00451	0.00365
4	0.00659	0.00516	0.00483	0.00364

Table 6. MSFE for Recursive Time Interval

MAPE				
Horizon	fit_1	fit_2	fit_3	fit_4
1	4.90549	5.72432	5.53809	5.22456
2	6.70509	7.12616	7.50393	6.78305
3	7.65266	7.50182	7.65597	7.20212
4	8.46339	8.00900	7.26878	6.83613

Table 7. MAPE for Recursive Time Interval

The same analysis is performed for the rolling scheme as well (Table 8 and 9). It is seen that the overall conclusions are the same even though the actual numbers are slightly different. Hence, it can be concluded that Model 4 is more robust, consistent and produces the best forecasting results. It is thus recommended to make use of Model 4 while forecasting the quarterly earnings of Coca Cola. It should be noted, that all models did not have residuals that were normally distributed and hence it is possible to use a GARCH model in order to predict the future volatility of the Coca Cola earnings. Finally, Appendix 17 shows the point predictions and 95% confidence interval for the next 4 quarters of the Coca Cola earnings. Please keep in mind that the real predictions should be calculated by taking the exp() of the log values obtained.

MSFE				
Horizon	fit_1	fit_2	fit_3	fit_4
1	0.00211	0.00232	0.00201	0.00167
2	0.00394	0.00413	0.00373	0.00309
3	0.00546	0.00495	0.00441	0.00362
4	0.00667	0.00514	0.00467	0.00361

Table 8. MSFE for Rolling Time Interval

MAPE				
Horizon	fit_1	fit_2	fit_3	fit_4
1	4.90350	5.77189	5.46916	5.22689
2	6.68224	7.13422	7.41930	6.79147
3	7.64676	7.51091	7.60718	7.19900
4	8.48043	8.00631	7.18177	6.82714

Table 9. MAPE for Rolling Time Interval

Appendix

Appendix 1: Time Series, ACF, PACF Plots for Coca Cola Earnings

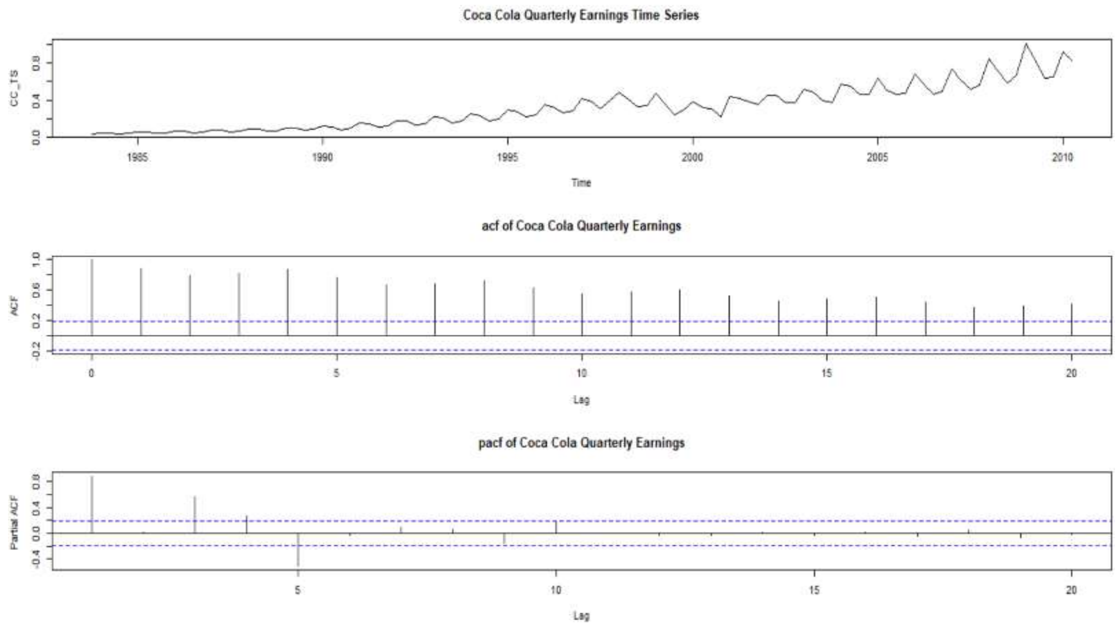

Figure 2. Time Series, ACF and PACF Plots for Coca Cola Earnings

Appendix 2: Time Series, ACF, PACF Plots for Coca Cola Earnings (Log)

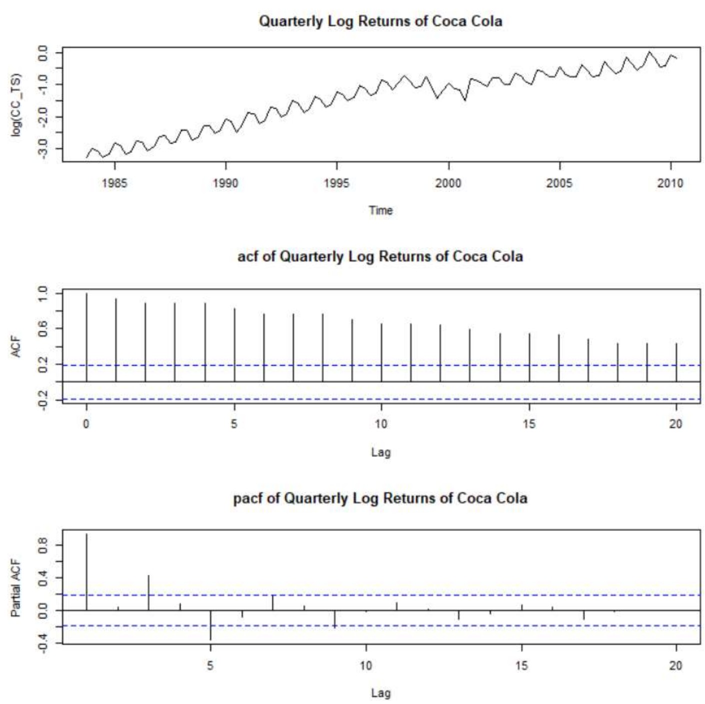

Figure 3. Time Series, ACF and PACF Plots for Coca Cola Earnings (Log)

Appendix 3: Time Series, ACF, PACF Plots for Coca Cola Earnings (Log, d=1)

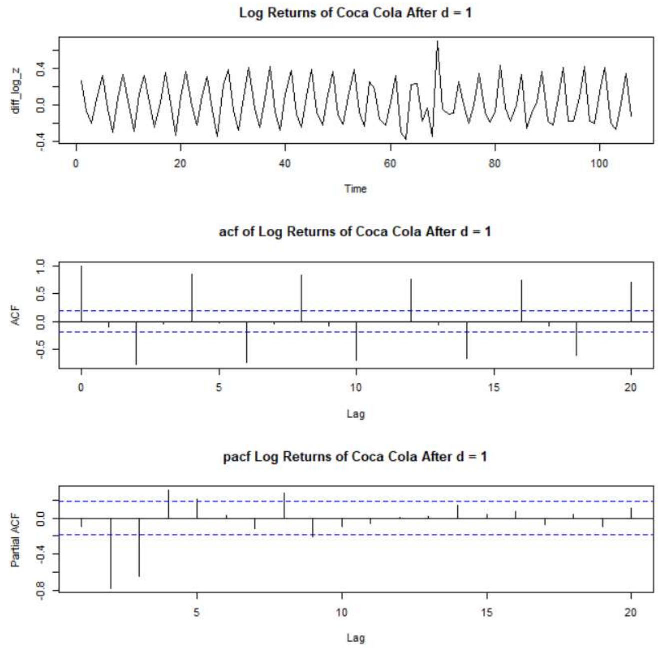

Figure 4. Time Series, ACF and PACF Plots for Coca Cola Earnings (Log, d=1)

Appendix 4: Time Series, ACF, PACF Plots for Coca Cola Earnings (Log, d=1, D=1)

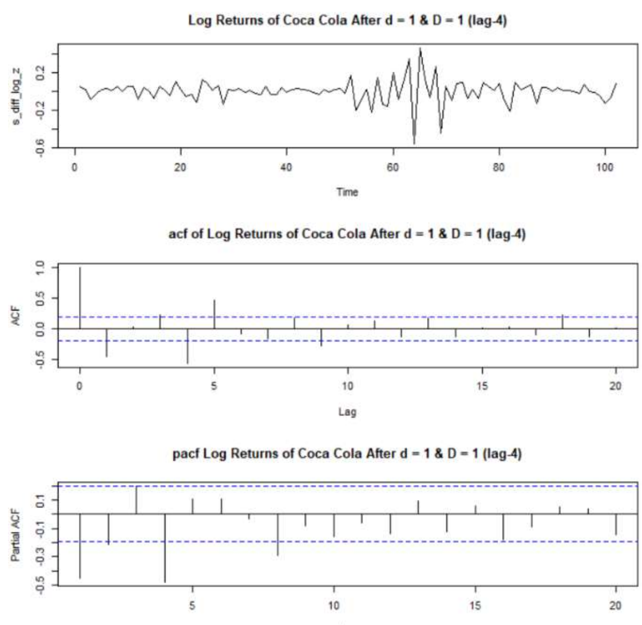

Figure 5. Time Series, ACF and PACF Plots for Coca Cola Earnings (Log, d=1, D=1)

Appendix 5: Model 1 – ACF and PACF for the Residuals

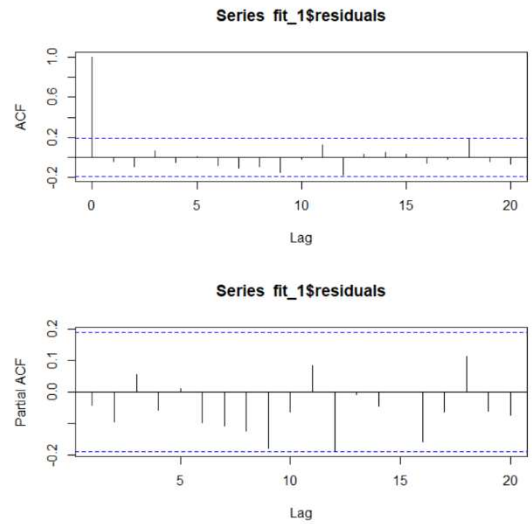

Figure 6. Model 1 - ACF and PACF for the Residuals

Appendix 6: Model 2 – ACF and PACF for the Residuals

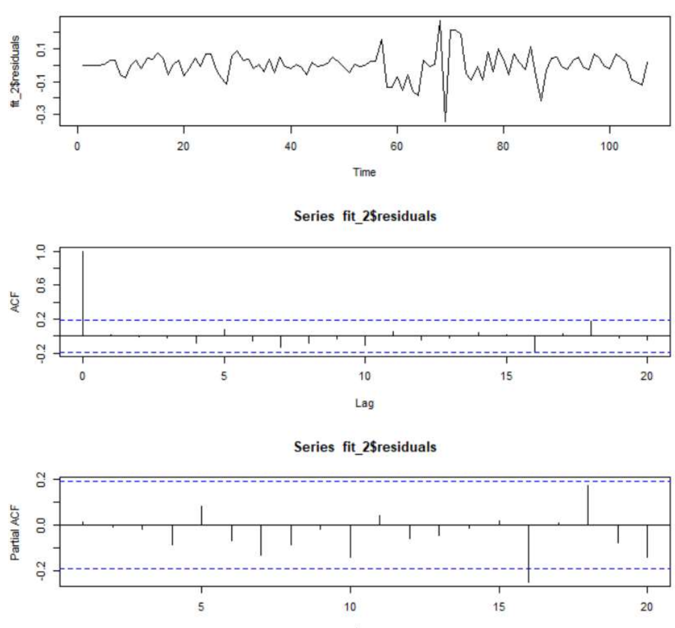

Figure 7. Model 2 - ACF and PACF for the Residuals

Appendix 7: Model 3 – ACF and PACF for the Residuals

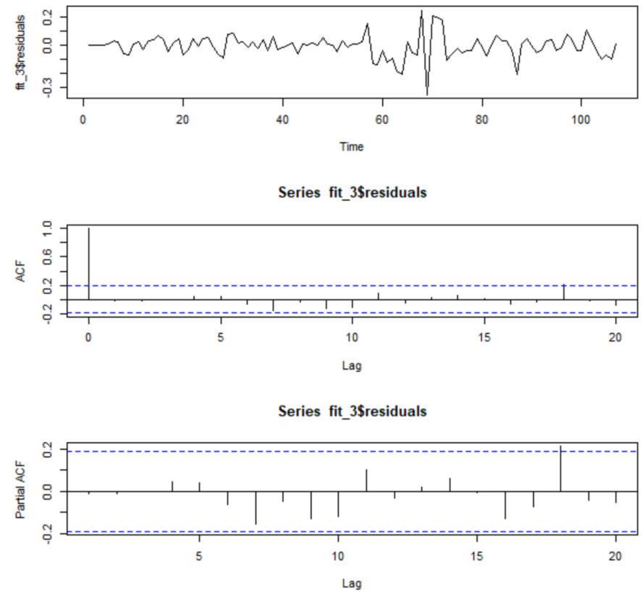

Figure 8. Model 3 - ACF and PACF for the Residuals

Appendix 8: Model 4 -ACF and PACF for the Residuals

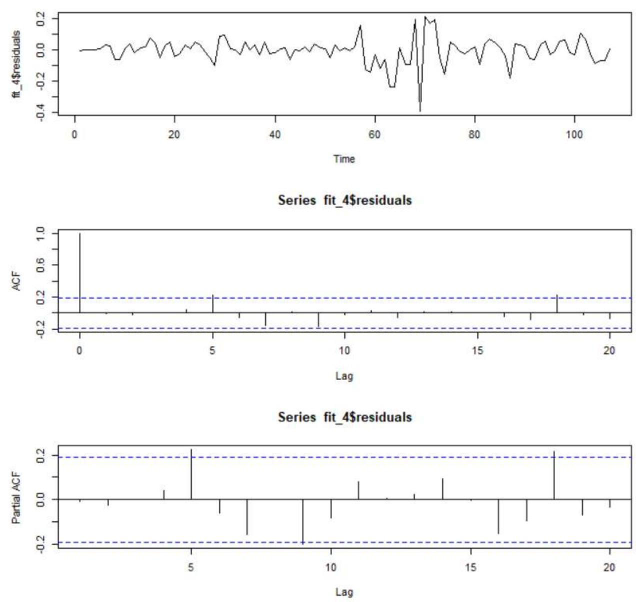

Figure 9. Model 4 - ACF and PACF for the Residuals

Appendix 9: Recursive Schema (horizon= 2) plots

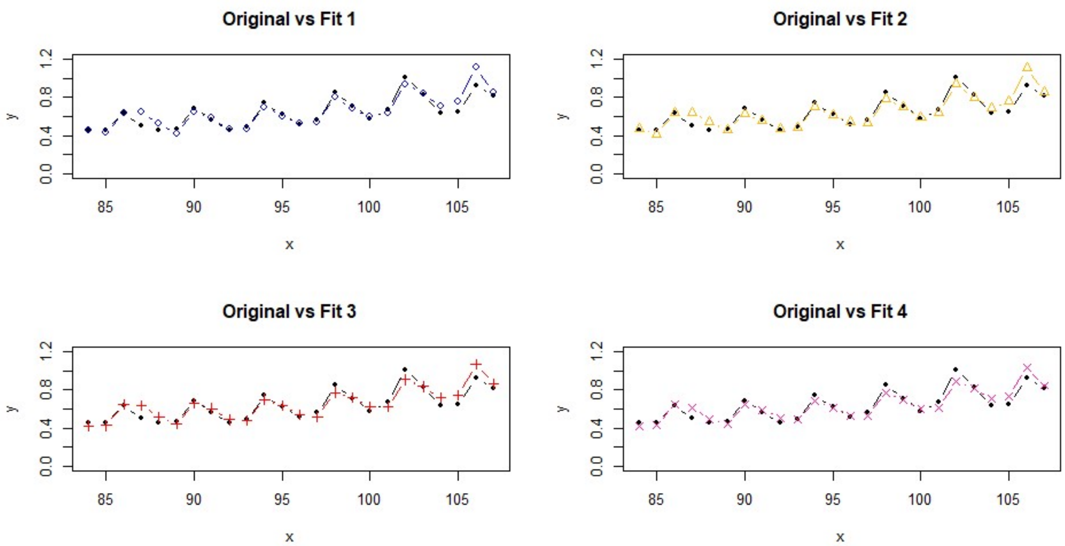

Figure 10. Recursive Schema (horizon=2) plots

Appendix 10: Recursive Schema (horizon=3) plots

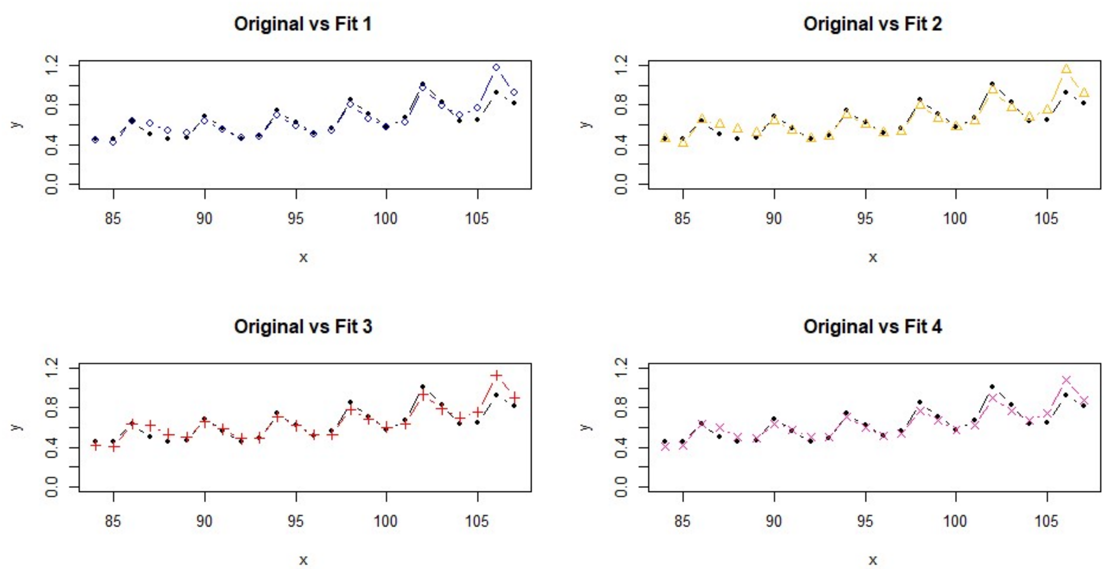

Figure 11. Recursive Schema (horizon=3) plots

Appendix 11: Recursive Schema (horizon=4) plots

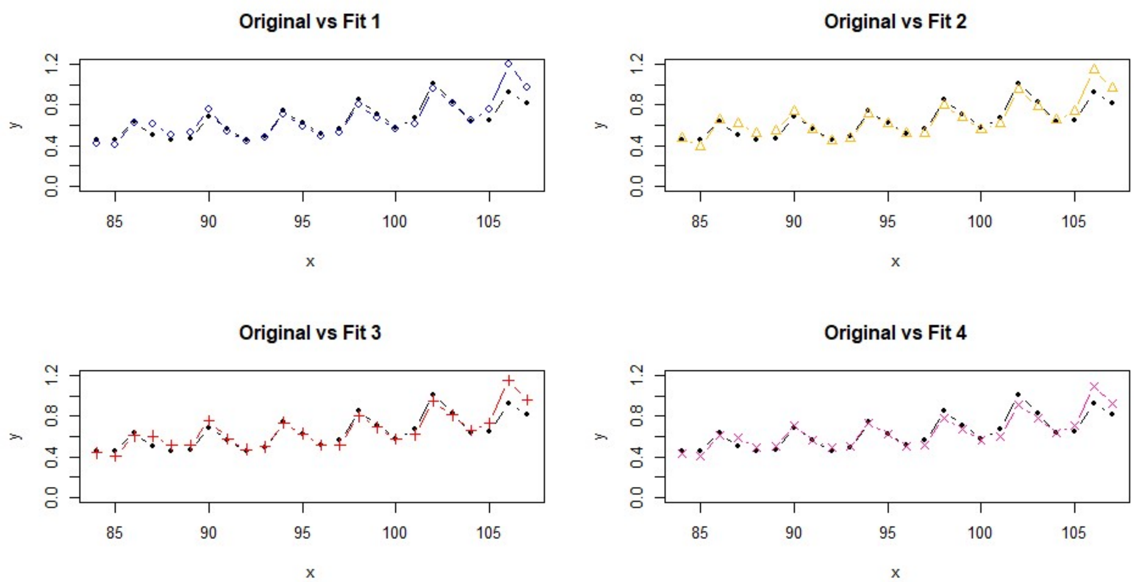

Figure 12. Recursive Schema (horizon=4) plots

Appendix 12: Rolling Schema (horizon=1) plots

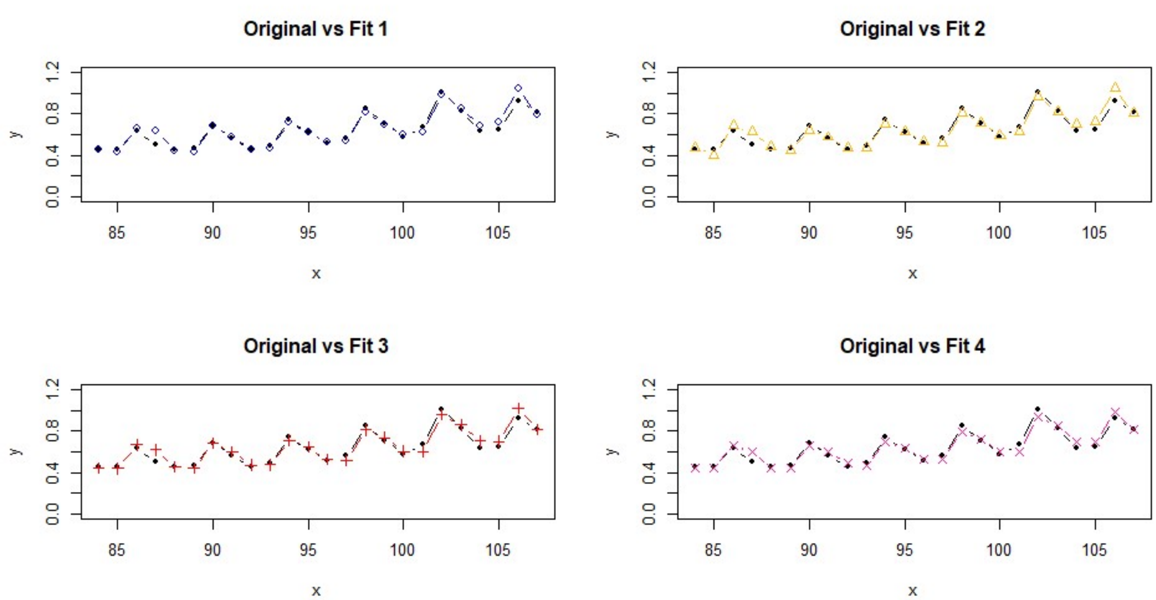

Figure 13. Rolling Schema (horizon=1) plots

Appendix 13: Rolling Schema (horizon=2) plots

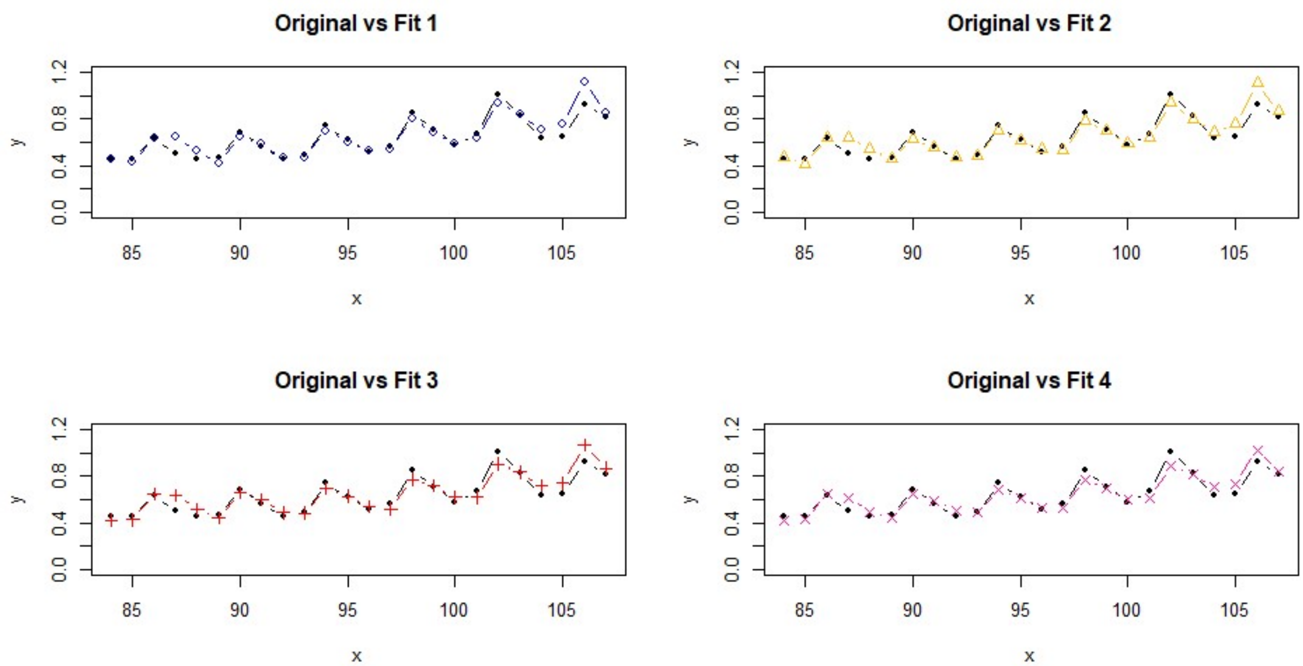

Figure 14. Rolling Schema (horizon=2) plot

Appendix 14: Rolling Schema (horizon=3) plots

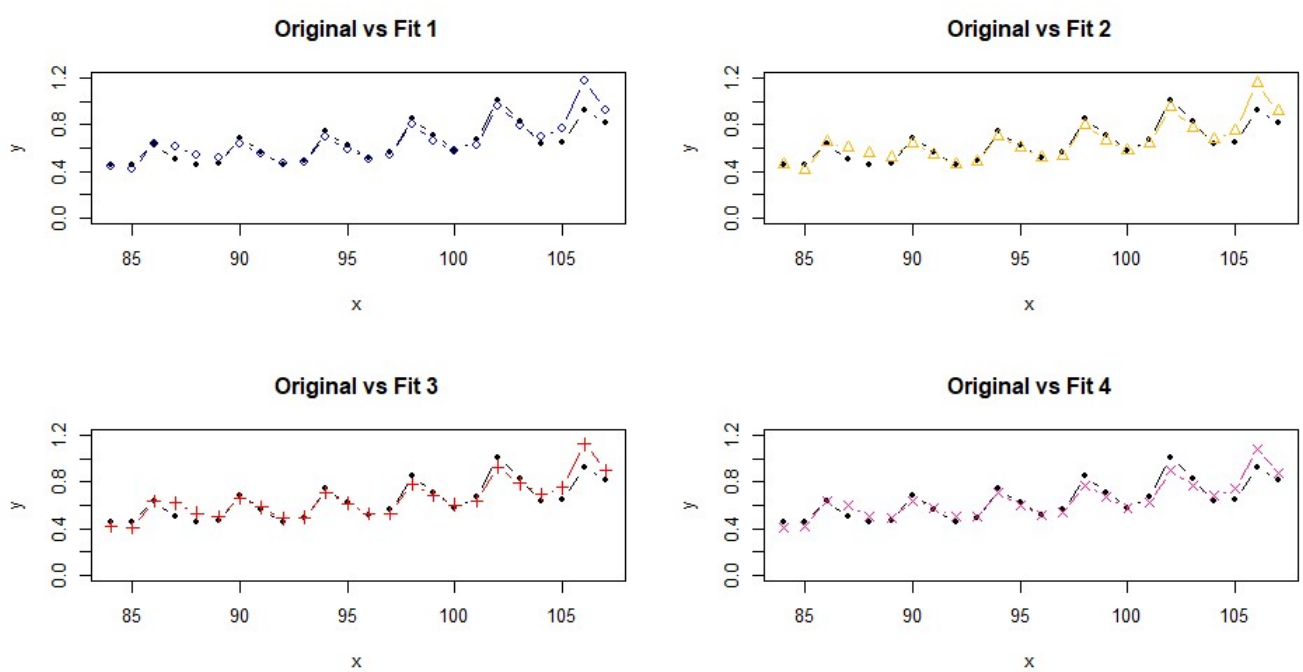

Figure 15. Rolling Schema (horizon=3) plots

Appendix 15: Rolling Schema (horizon=4) plots

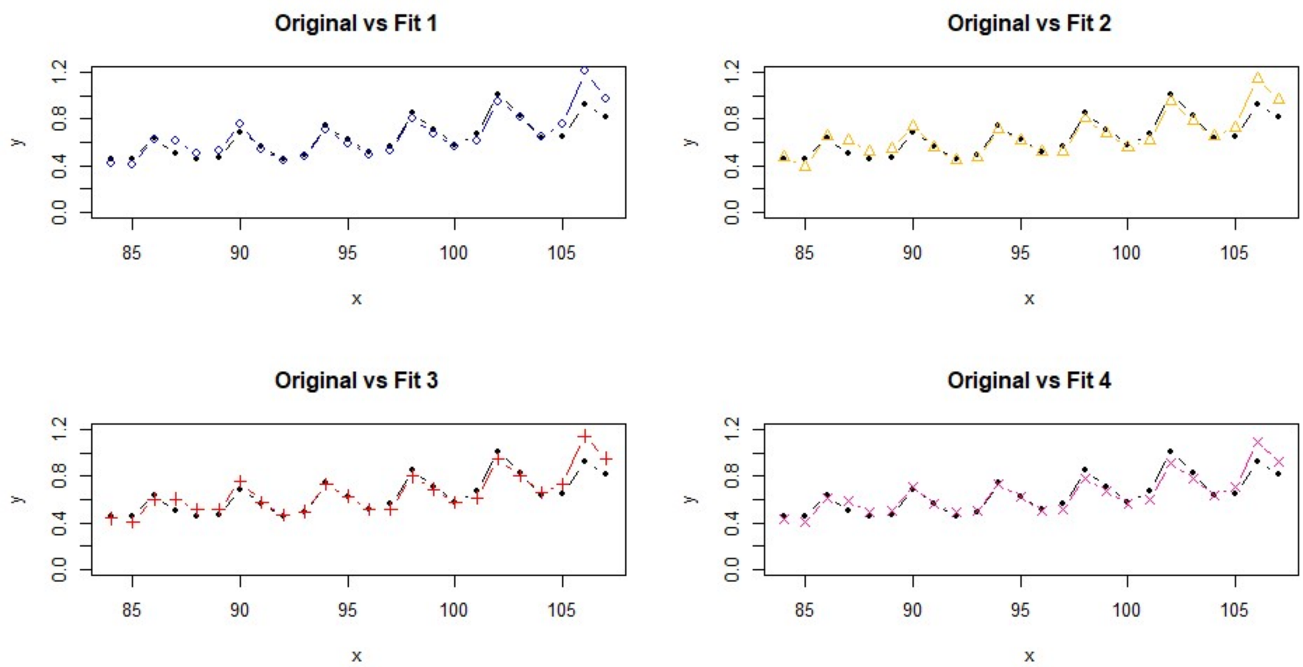

Figure 16. Rolling Schema (horizon=4) plots

Appendix 16: Decomposition of Additive Logged Coca Cola Quarterly Earnings

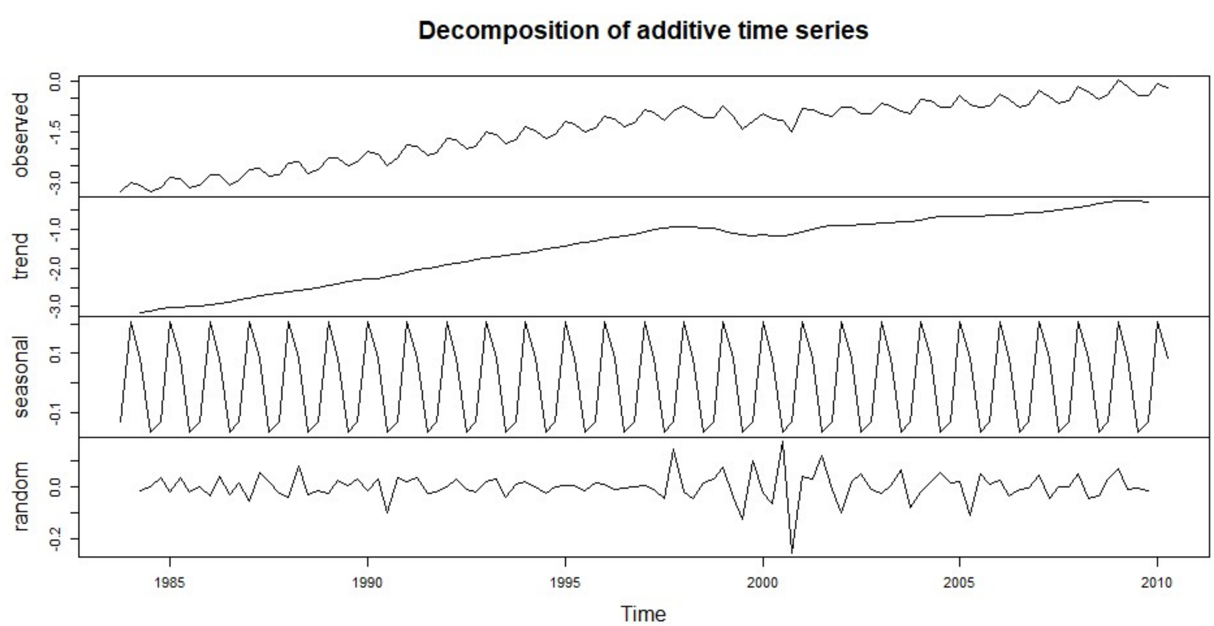

Figure 17. Decomposition of Additive Logged Coca Cola Quarterly Earnings

Appendix 17: Predictions for the next 4 Quarters for the Logged Coca Cola Earnings (Model 4)

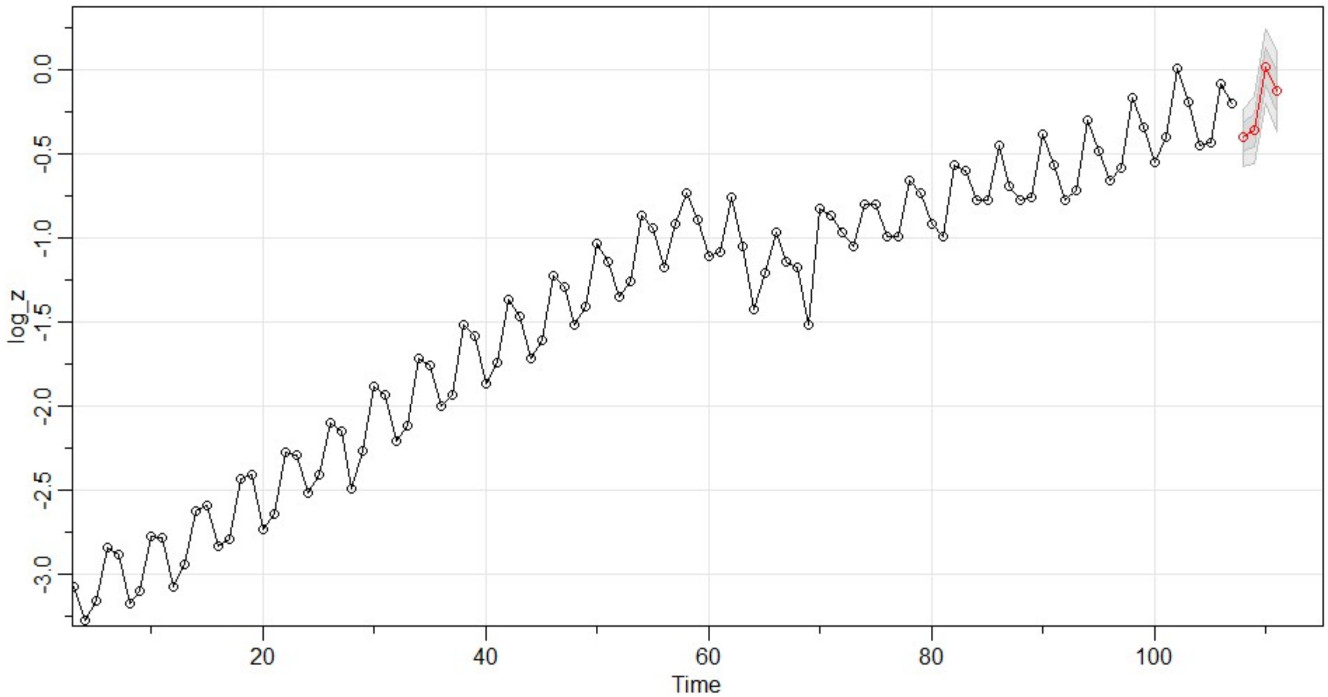

Figure 18. Prediction for the next 4 Quarters for the Logged Coca Cola Earnings (Model 4)

NB: Note that these are not the real predictions. Real predictions should be calculated by taking the exp() of these values obtained. Was put just to show the logged predictions.