Jingzhi Zhou
STAT 5550

Analysis of Data “Monthly Beer Production in Australia”

I. Introduction

The data used for analysis are the monthly beer production values in Australia, from January 1956 to August 1995, recorded in megaliters. The production values include ale and stout but do not include beverages with alcohol percentage less than 1.15. The dataset came from kaggle.com. The goal of this analysis is to discover and handle the important features of the data, fit appropriate models to the data and construct optimal forecasts for future data.

In Figure 1, the time series plot of the data suggests an increasing trend from 1956 to 1980 and a slow decreasing trend after 1980. The monthly time series plot reveals an obvious pattern of seasonality. The ACF plot clearly suggests non-stationarity through the slow decay of the sample ACFs.

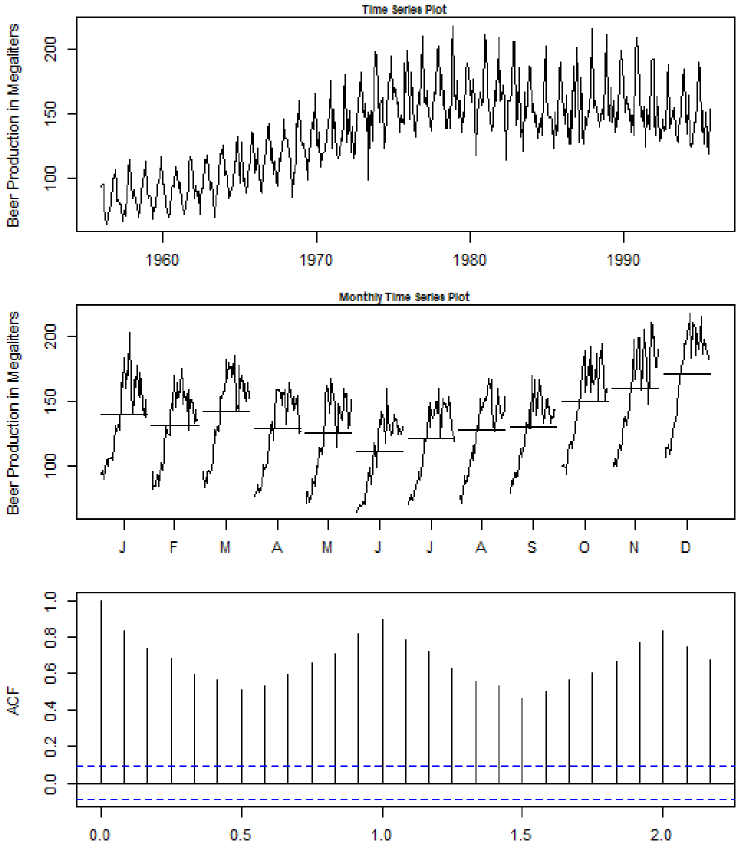

Figure 1 Summary Plot of beer production

In order to eliminate the non-stationary behaviors mentioned above, methods such as differencing and log transformation can be performed. Since the data have a period of twelve, a seasonal differencing with Lag 12 is appropriate to apply. The rapid decay of ACFs suggests that a seasonal differencing works well enough to eliminate the non-stationary features of the data. On the other hand, a first-order differencing on log transformed values of original data has the similar effect on eliminating the trends and the seasonality of the data. The two ACF plots (Figure 1 in appendix) suggests that the time series of the data becomes stationary after using these two methods.

II. Models for Trend and Seasonality

ARMA Modeling

In order to estimate the trend components, two regression models were fitted. The first fitted model is a linear regression model x_t = $ß_0$ + $ß_1$t + w_t, where wt ~ wn (0, $σ^2$), which does not work well for estimating the trend, because the plot of residuals of the data after fitting a linear regression shows a strong evidence of polynomial trend. Therefore, a 3rd-order polynomial regression model x_t = $ß_0$ + $ß_1$t + $ß_2t^2$ + $ß_3t^3$ + w_t, where w_t ~ wn (0, $σ^2$), was then fitted to estimate the trend components. Looking at the summary of model in R, the polynomial regression model has a much better performance in estimating trends. The residual time series plot of detrended data using linear and polynomial regression models separately also suggest this (Figure 2 in appendix).

Log transformation was also applied with linear regression and polynomial regression models. However, transformation does not seem to be better than applying regression techniques only. Thus, log transformation is not considered for detrending.

After deciding to use 3rd-order polynomial model for estimating trend, a model for estimating trend and seasonality together was fitted: x_t = $ß_1$t + $ß_2t^2$ + $ß_3t^3$ + S_t + w_t, where w_t ~ wn (0, $σ^2$) and S_t represents twelve separate seasonal means.

However, the fitted model for estimating trends and seasonality does not appear to fully eliminate all the non-stationarity of the data, which is suggested in the detrended and deseasonalized plot as well as the ACF plot for detrended and deseasonalized data (Figure 3 in appendix). The diagnostic plots for residuals from ARMA (1,1), ARMA (1,2), ARMA (2,2), ARMA (2,1) all suggest lack of fit through the large ACFs and very small p-values for Ljung-Box statistics.

Since a higher order polynomial generates errors and a third order is the highest order to use for estimating trend for this dataset, using an ARMA model for detrended and deseasonalized data is not an appropriate method here. With this in mind, ARIMA method was then used to estimate deseasonalized data.

ARIMA Modeling

A seasonal model S_t was used first to estimate the seasonal components by estimating the twelve separate seasonal means. Using the residuals of the seasonal model brings a new data which is the deseasonalized data of original data. From there, ARIMA modeling can be used to analyze the deseasonalized data.

Looking at the sample ACF plots and PACF plots of first order differencing model and second order differencing model (Figure 4 and 5 in appendix), first order and second order differencing have similar effects. The sample PACF values in second order differencing are relatively smaller. But the difference is not so significant to make choice. Further exploration can be done through SARIMA command in R.

The PACF plot clearly suggests an MA(q). It is hard to determine the behavior of model from sample ACF. Thus a few plausible ARIMA models were tested. Table 1 displays summaries of the fitted ARIMA models.

Model	Φ1	Se(Φ1)	Φ2	Se(Φ2)	Θ1	Se(Θ1)	Θ2	Se(Θ2)	AIC
ARIMA(0,1,1)					-0.8694	0.0161			3550.27

ARIMA(0,2,2)					-1.9051	0.0184	0.9383	0.0211	3549.42
ARIMA(1,1,1)	-0.1123	0.0487			-0.8542	0.0189			3547.04
ARIMA(0,1,2)					-1.0232	0.0548	0.1549	0.0520	3543.86
ARIMA(0,2,1)					-1.0000	0.0053			3858.64
ARIMA(1,2,2)	-0.0993	0.0486			-1.8986	0.0180	0.9049	0.0193	3534.68
ARIMA(1,1,2)	0.7639	0.0483			-1.8488	0.0316	0.8954	0.0293	3510.77

Table 1 Fitted ARIMA Models

Commentary:

- ARIMA (0,1,1): The diagnostic plots for residuals suggest lack of fit.
- ARIMA (0,2,2): AIC is a little smaller than that in ARIMA(0,1,1). But the ACF plot are still very large; p-value in Ljung-Box test are nearly zero. It does not worth to increase the order.
- ARIMA (1,1,1): AIC is smaller. But diagnosis is still not as desired.
- ARIMA (0,1,2): The AIC is relatively small, but the difference is not significant enough.
- ARIMA (0,2,1): The AIC value is the largest among all models, suggesting lack of fit. The very small p-values in Ljung-Box Test also indicate the lack of fit.
- ARIMA (1,1,2): AIC is the smallest among all fitted models. Although the ACF plot and Ljung-Box test still suggests lack of fit, it might be the most appropriate ARIMA model.
- ARIMA (1,2,2): The AIC value is relatively small, but is larger than that in ARIMA (1,1,2).
- None of these models have a confidence interval that contains zero.

Therefore, ARIMA(1,1,2) model, that has a first order autoregression, a second order moving average and differences twice, seems to be the most appropriate model for the data. However, because of the large sample ACF value, which suggests non-white noise residuals, and the very small p-values in Ljung-Box test, which suggests failing to reject dependency, the lack of fit of ARIMA model will make the further forecasting not very accurate.

The diagnostic plots for residuals are in Figure 2.

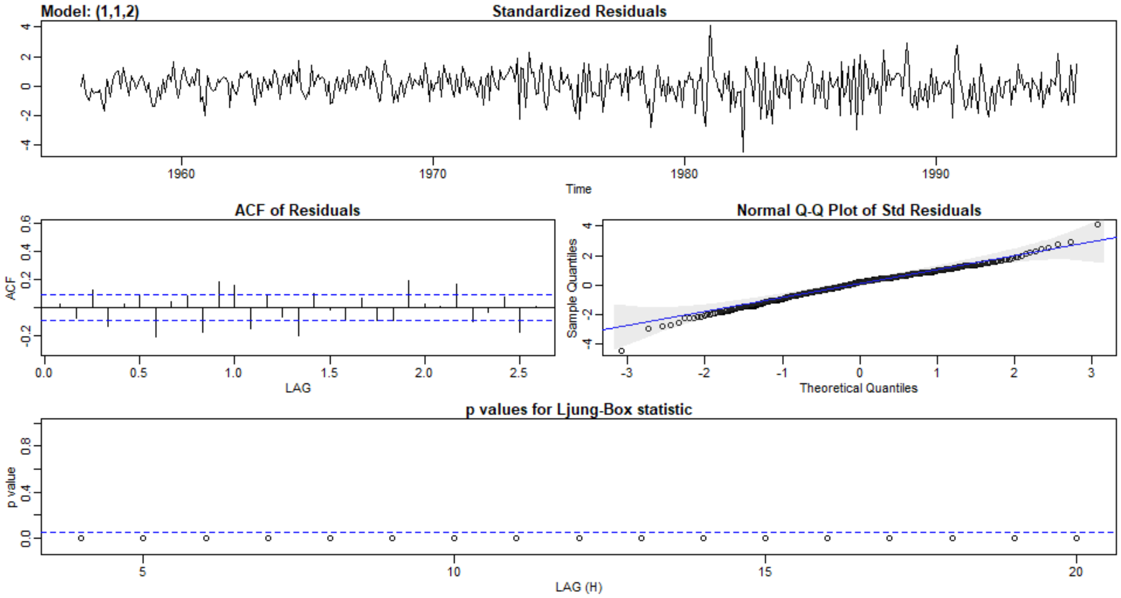

Figure 2 Diagnosis Plot for ARIMA (1,1,2)

The estimated model is:

$$yt = St + xt, \;\; xt = \frac{1 - 1.8488\mathrm{B} + 0.8954B^2}{(1 - 0.7639B)(1 - B)} + \omega t, \;\; \omega t \sim iid\; N(0{,}92.82)$$

$$where\; St\; is\; the\; \text{estimated seasonal effects in the model.}$$

The twelve separate seasonal mean estimates for S_t are in Table 2.

January	February	March	April	May	June
139.6	130.9	142.3	128.5	125.2	111.8
July	August	September	October	November	December
120.9	127.7	130.5	150.2	159.4	171.4

Table 2 Seasonal Effects

Forecasting:

A forecasting of beer production over 24 months, from September 1995 to August 1997 was performed on the deseasonalized ARIMA model using predict() command. Then the seasonal components were predicted and added back to the predicted nonseasonal factors for the final forecasting data. Forecasts for x_t and y_t are shown in Figure 3. The predicted value for each month with its 95% confidence interval is listed in Table 3. As noted, the forecasts might not be very accurate.

Sep 95	Oct 95	Nov 95	Dec 95	Jan 96	Feb 96	Mar 96	Apr 96	May 96	Jun 96
140.0	160.5	170.4	182.8	151.3	142.9	154.5	140.9	137.6	124.4
[120,159]	[140,180]	[151,190]	[163,202]	[132,170]	[123,162]	[135,174]	[121,160]	[117,157]	[104,144]
Jul 96	Aug 96	Sep 96	Oct 96	Nov 96	Dec 96	Jan 97	Feb 97	Mar 97	Apr 97
133.5	140.4	143.2	163.0	172.2	184.2	152.4	143.7	155.2	141.4
[112,154]	[119,161]	[122,164]	[141,184]	[150,194]	[162,206]	[130,174]	[121,166]	[132,178]	[118,164]
May 97	Jun 97	Jul 97	Aug 97						
138.0	124.6	133.7	140.6						
[114,161]	[101,148]	[110,158]	[116,165]						

Table 3 Predecited value and interval for ARIMA(1,1,2)

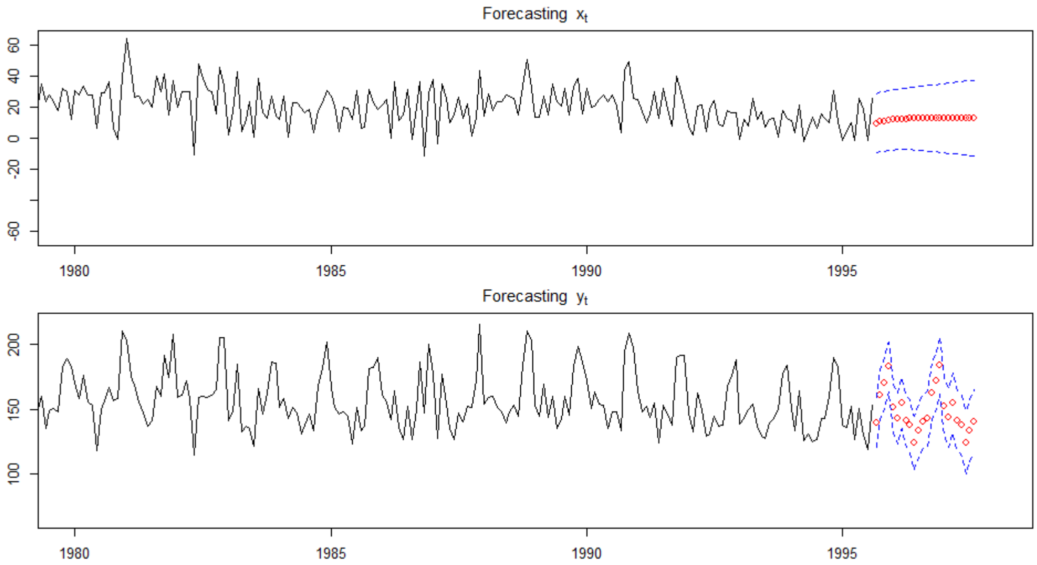

Figure 3 Forecasting Plot for ARIMA (1,1,2)

III. SARIMA Modeling

Unlike using a separate model to estimate seasonality (and trend) before using ARMA or ARIMA modeling, SARIMA model can analysis the data together with its trend and seasonal components. Therefore, a second analysis of the data was performed using SARIMA modeling.

Because of the seasonal elements in the data, a seasonal difference $\nabla_{12} x_t$ was taken. Although the data appears to be stationary, sample ACF values are relatively large at all lags, which means further difference might be needed (Figure 6 in appendix). A nonseasonal difference $\nabla\nabla_{12} x_t$ was applied in addition to the seasonal difference, but it does not bring any improvement. Thus, a second order seasonal difference $\nabla_{12}\nabla_{12} x_t$ was applied to the data (Figure 7 in appendix). Despite it does reduce the sample ACF values at most of the lags, the sample ACF are still considered as large values. Higher order differences were performed, but none of them made much improvement.

Since the sample PACF value of $\nabla_{12}\nabla_{12} x_t$ tails off, an MA_{12} (1) model might be an option for the nonseasonal component. The ACF has a large sample ACF value at the first seasonal lag and relatively small sample ACFs after lag 1. However, the behavior is hard to conclude based on the ACF plot. The sample PACF tails off around lags 12, which suggests an MA (1) model for seasonal components. The ACF and PACF plots are in Appendix.

Several plausible SARIMA models were fitted. The summaries of fitted models are in Table 4.

Model	$(0,0,1)\times(0,2,1)_{12}$	$(0,0,1)\times(0,1,1)_{12}$	$(0,0,2)\times(1,2,0)_{12}$	$(0,0,2)\times(0,2,1)_{12}$	$(1,1,1)\times(1,1,1)_{12}$
Φ1					-0.1619
se(Φ1)					0.0487
Θ1	-0.1074	0.0251	-0.1024	-0.0778	-0.8865
se(Θ1)	0.0490	0.0513	0.0497	0.0508	0.0185
Θ2			-0.1561	-0.0695	
se(Θ2)			0.0488	0.0500	
S: Φ1			-0.5864		0.1474
S: se(Φ1)			0.0378		0.0555
S: Θ1	-1.0000	-0.3143		-0.9997	-0.8739
S: se(Θ1)	0.0628	0.0504		0.3114	0.0301
AIC	3616.31	3629.02	3785.61	3616.39	3468.29

Table 4 Fitted SARIMA Models

Commentary:

- ARIMA $(0,0,1)\times(0,1,1)_{12}$: A confidence interval that includes 0 indicates the estimated Θ_1 is not significant at the 95% level.
- ARIMA $(0,0,2)\times(0,2,1)_{12}$: Confidence intervals that include 0 indicate the estimated Θ_1 Θ_2 are not significant at the 95% level. A second order nonseasonal MA may not be needed.
- ARIMA $(0,0,2)\times(0,2,1)_{12}$: All estimates are significant. AIC value is reduced compared to the AIC in ARIMA$(0,0,1)\times(0,2,1)_{12.}$ But the sample ACFs are very large.
- Fitting an ARIMA$(0,0,1)\times(0,2,1)_{12}$ model does not generate a good diagnosis plot of residuals, neither does other ARIMA$(p, d, q) \times (P, D, Q)_{12}$ models. The sample ACF values are all large in these SARIMA models; the p-values for Ljung-Box test are nearly zero; AIC values are high (all above 3600). These signs all suggest a lack of fit of these SARIMA models.

Since the ARIMA(0,0,1)×(0,2,1)$_{12}$ model has a better performance in AIC (3616.31) and significance of estimated parameters than other models, it will be used for forecasting. However, due to the lack of fit suggested in the diagnosis plot, the forecast might be accurate.

The ARIMA(0,0,1)×(0,2,1)$_{12}$ model is fitted as:

$$(1-B^{12})^2\ xt = (1-B^{12})(1-0.1074B)\omega t,\ \ \omega t \sim iid\ N(0{,}156.4)$$

Forecasting:

A forecasting of beer production over 24 months, from September 1995 to August 1997 was performed using sarima.for based on model ARIMA(0,0,1)×(0,2,1)$_{12}$. Forecasts are shown in Figure 4. The predicted value for each month, with its 95% confidence interval, is listed in Table 5.

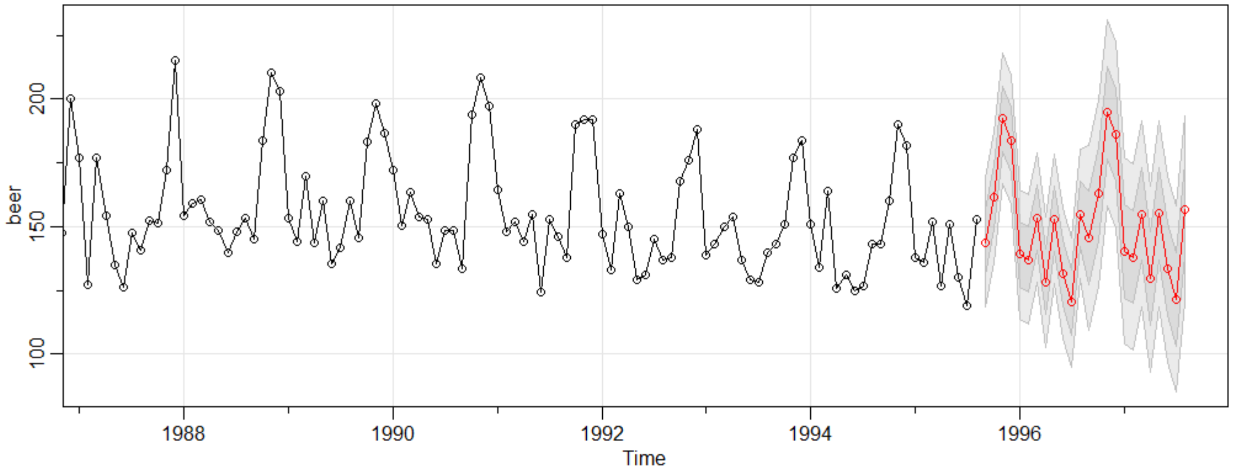

Figure 4 Forecasts on model ARIMA(0,0,1)×(0,2,1)$_{12}$

Sep 95	Oct 95	Nov 95	Dec 95	Jan 96	Feb 96	Mar 96	Apr 96	May 96	Jun 96
143.89	161.56	192.35	183.97	139.15	137.03	153.46	128.28	153.05	131.67
[119,168]	[137,187]	[167,217]	[159,209]	[114,164]	[112,162]	[128,178]	[103,153]	[128,178]	[107,157]
Jul 96	Aug 96	Sep 96	Oct 96	Nov 96	Dec 96	Jan 97	Feb 97	Mar 97	Apr 97
120.25	154.94	145.54	163.13	194.70	185.94	140.30	138.05	154.91	129.56
[95,145]	[130,180]	[110,181]	[127,199]	[159,230]	[150,222]	[105,176]	[102,174]	[119,191]	[94,165]
May 97	Jun 97	Jul 97	Aug 97						
155.11	133.34	121.51	156.88						
[119,191]	[96,169]	[86,157]	[121,193]						

Table 5 Forecast value and interval Using SARIMA

IV. Model Comparison

Despite both models have obvious sign of lack of fit, the fitted ARIMA model ARIMA(1,1,2), with a seasonal effect model, has a better performance in analyzing this dataset than the fitted SARIMA model ARIMA(0,0,1)×(0,2,1)$_{12}$ does. The fitted ARIMA model has a smaller AIC value of 3510.77 than SARIMA model does, which is 3616.31. Although both AIC values are very big, a difference of 105 is still significant. The residual standard errors in both models provide similar information as AIC values. The estimated σ^2

in ARIMA model is 92.82, while the estimated σ^2 in SARIMA model is 156.4. These evidences all suggest that ARIMA model has a better fit than SARIMA model.

These two models both use differencing and Moving Average to remove the nonstationary elements in the data. The difference is the application of seasonal differencing and seasonal Moving Average in SARIMA model, and an additional autoregressive process of order 1 in ARIMA model.

The forecast data under two different models also have similarity and difference. The forecasting plots of ARIMA and SARIMA models have very similar trend overall. The peaks and troughs appear at the similar time period for these two models. The general trends are similar, and both follow the patterns of past data. However, the predicted values in SARIMA models are more extreme than those in ARIMA models. For example, the predicted production from October 1995 to January 1996 in ARIMA models goes from 160 to 170, and then 182, and finally 151, while the forecasts in SARIMA models goes from 162 to 192 to 184, and then 139. This is more obvious in the prediction intervals of SARIMA model. Although there is no additional data to prove the accuracy of forecasts in these two models, the extremeness of peak and trough data in SARIMA model might be another reason in favoring ARIMA model.

V. Conclusions

In this analysis, three modeling techniques ARMA, ARIMA and SARIMA were used to analyze the beer production in Australia. Because of the failure in estimating the trend components and in removing the nonstationary elements in the data, ARMA modeling is not an appropriate tool to analyze the data. ARIMA and SARIMA modeling can successfully remove the nonstationary elements but fail to generate a model that has a good fit, which is evidenced by the non-white noise residuals in diagnosis plot of both models. ARIMA model, with lower AIC value and smaller residual standard error, shows better performance in analyzing the data. The forecast results might be affected by the lack of fit of models and become not quite accurate. The models used for forecasting were ARMA (1,1,2) and ARIMA$(0,0,1)\times(0,2,1)_{12}$. The forecast results show similar patterns in overall trends and seasonality but differences in the extremeness of peak and trough data. Overall, ARIMA model shows more advantages in analyzing this data set.

Appendix

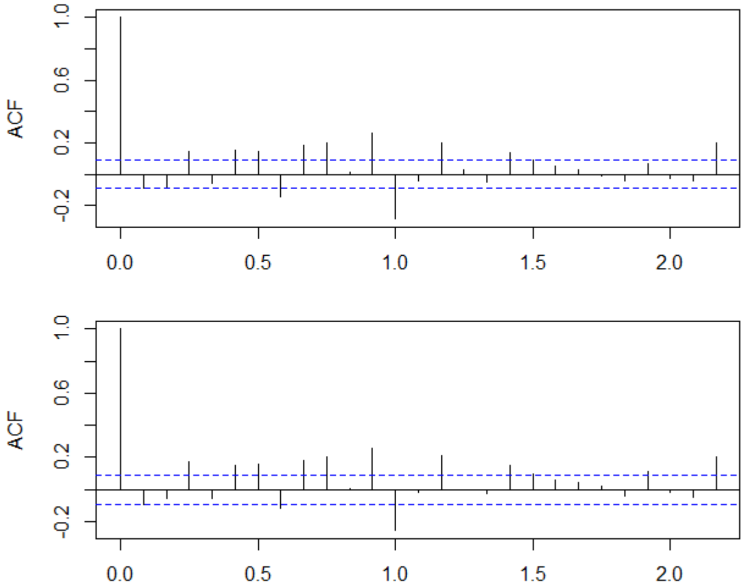

Figure 2 Seasonal Differencing and Log transformed Differencing

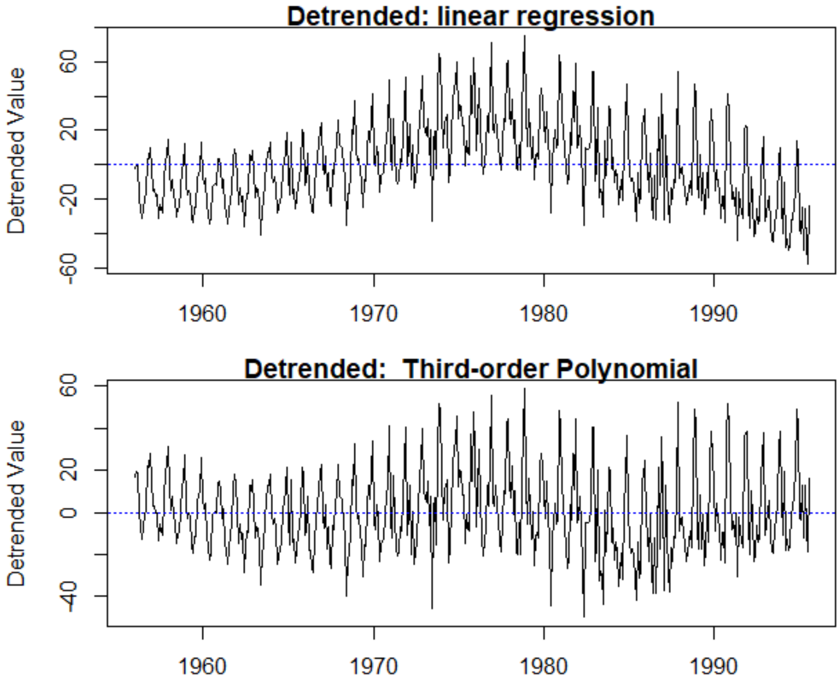

Figure 3 Residual time series plot of detrended data using linear and polynomial regression

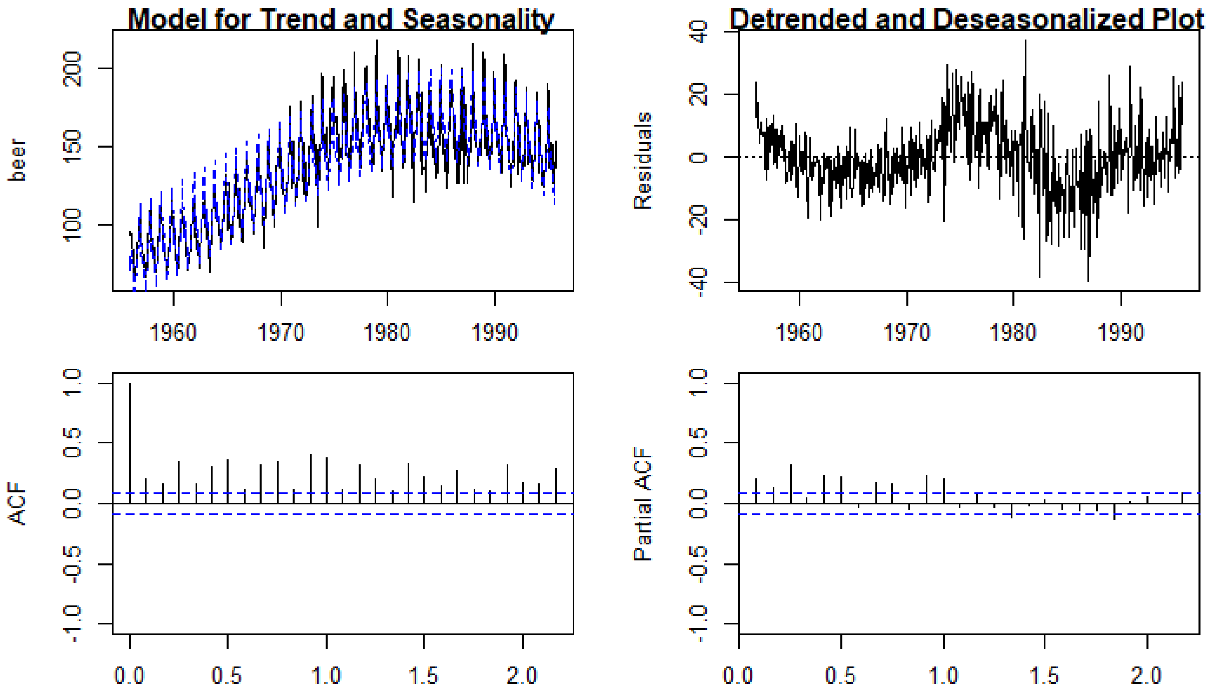

Figure 4 Analysis plots for detrended and deseasonalized data

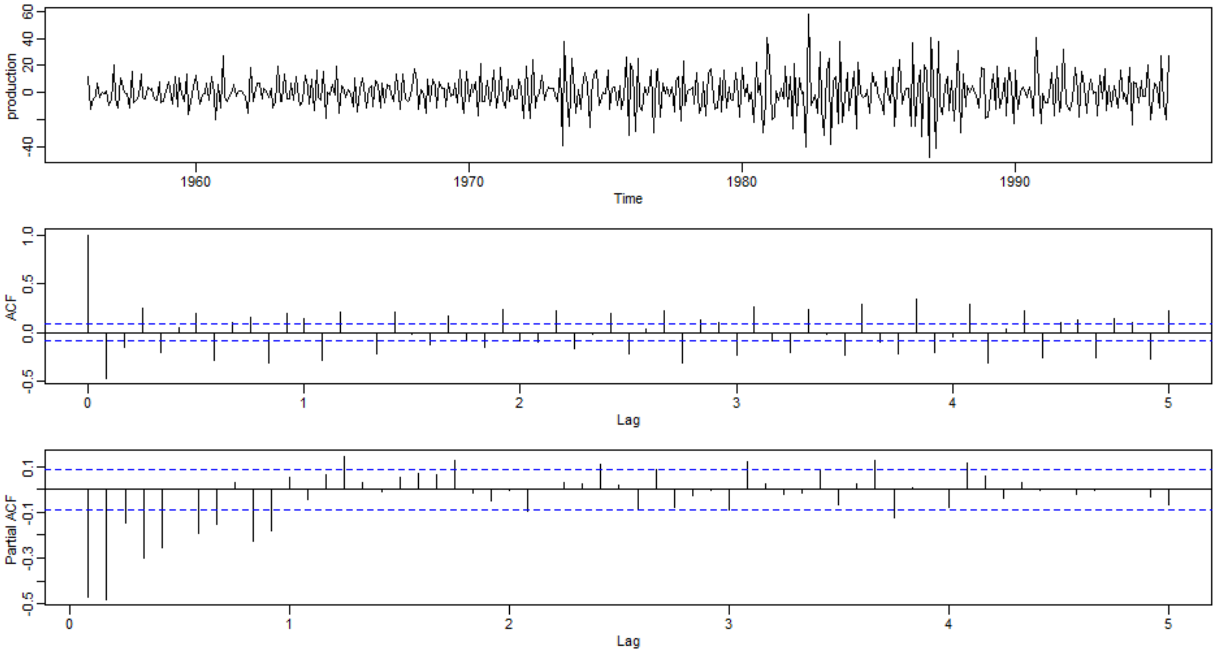

Figure 5 ACF and PACF for differencing order 1

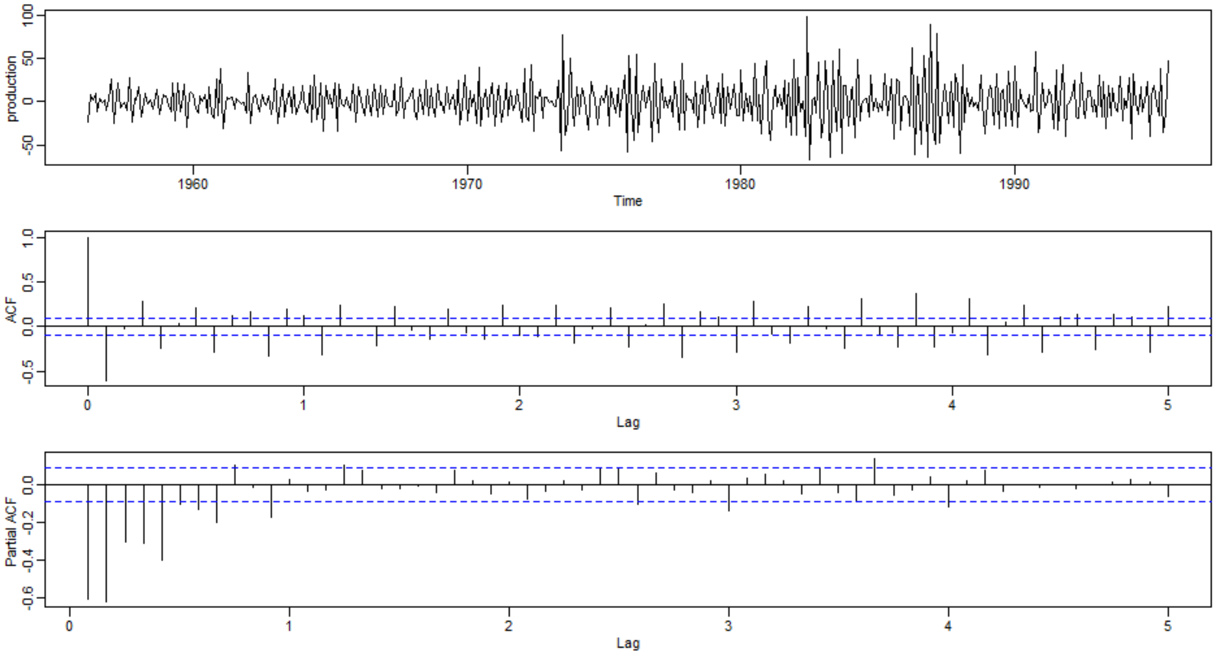

Figure 6 ACF and PACF of differencing order 2

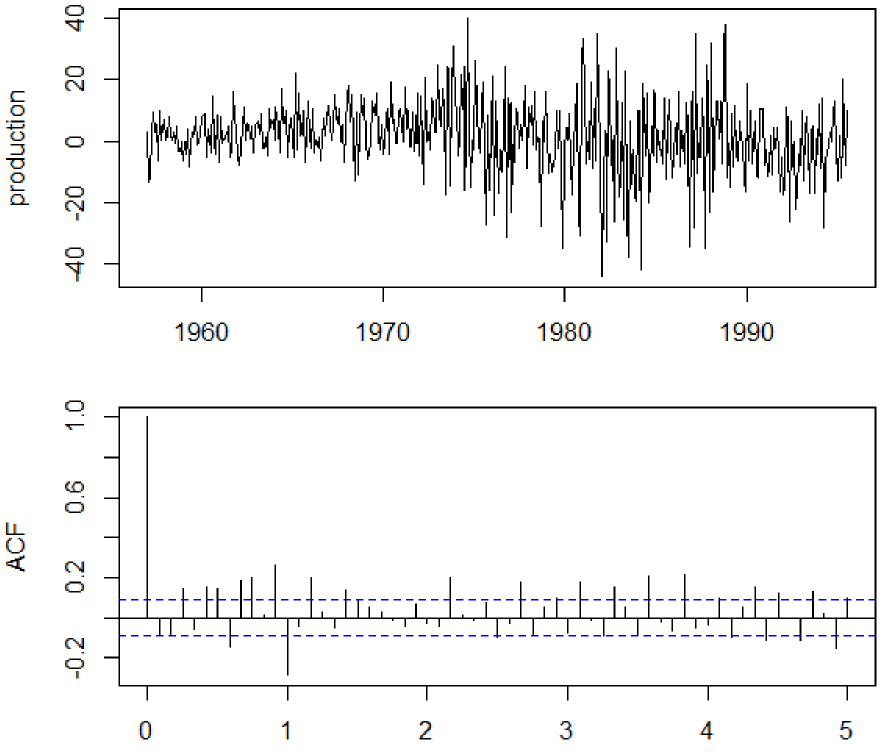

Figure 6 Seasonal Difference and ACF plot

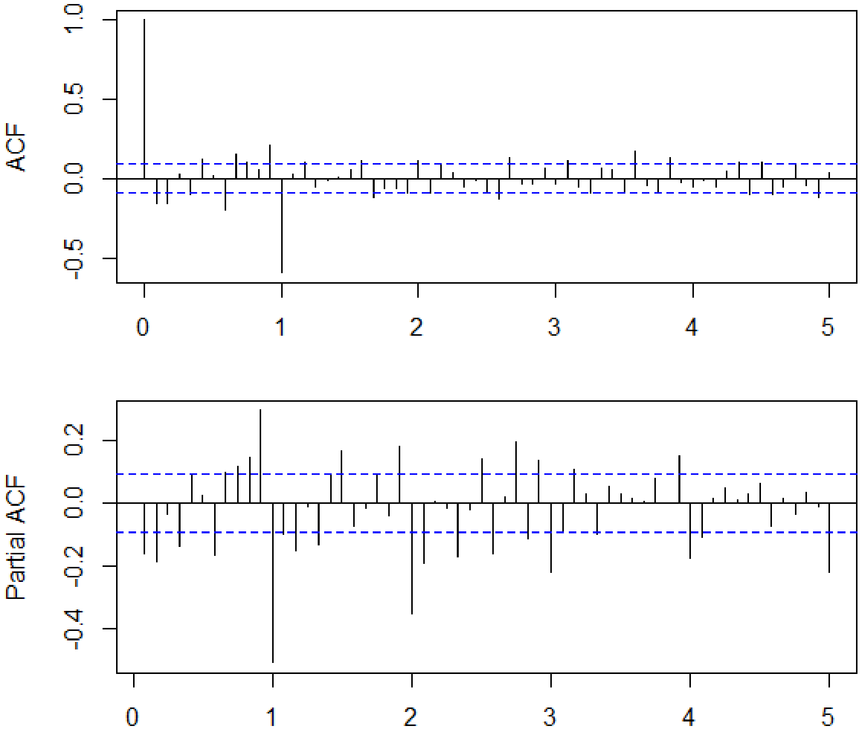

Figure 7 Second Order Seasonal Difference

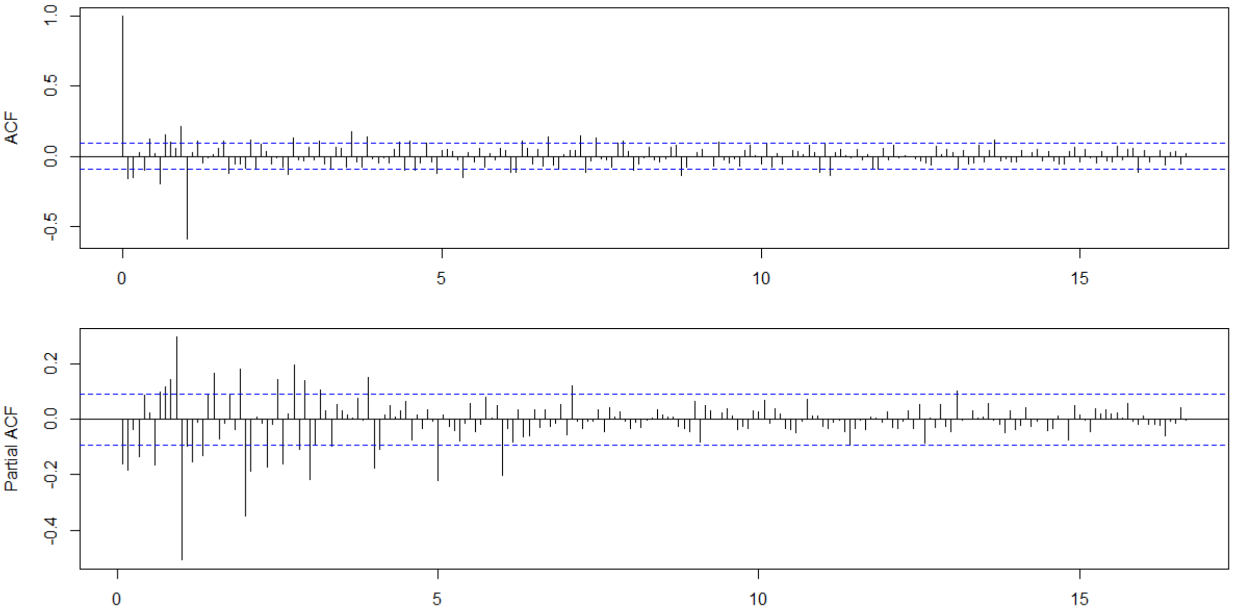

Figure 8 Difference order 2 ACF &PACF