The time series to be analysed in this report is of dry white wine sales in thousands of litres in monthly periods starting from Jan 1980 until Jul 1995. This report is separated into three sections, two sections dedicated to ARIMA modelling of the log sales and reciprocal of sales respectively from Jan 1980 up to December 1954. In both of these sections, the model fitting will be following the Box-Jenkins approach, firstly by identifying the model orders; $(p, d, q) \times (P, D, Q)_{12}$, then estimating the model parameters and finally assessing the fitted model by studying the residuals. The final section will focus on using the fitted model obtained from one of the two transformed datasets to predict the future sales and compare it to the available data of Jan to July 1995.

1 Log Sales Data Analysis

1. The transformed data, `log(x)` is and its autocorrelation function(acf) is plotted in the first row of Figure 1. Looking at (a)(ii), we see that there seems to be an upwards trend in the data, which is differenced once, letting $d = 1$ to obtain the stationary plot below it, (b)(ii). The acf of the differenced series clearly shows that there is a clear seasonal pattern that needs to be differenced to remove it and compare the acf and partial acf (pacf). The order of seasonal deifferencing, D done to obtain plot (c)(ii) is 1. So far, we have that the fitted model is ARIMA$(p, 1, q) \times (P, 1, Q)_{12}$.

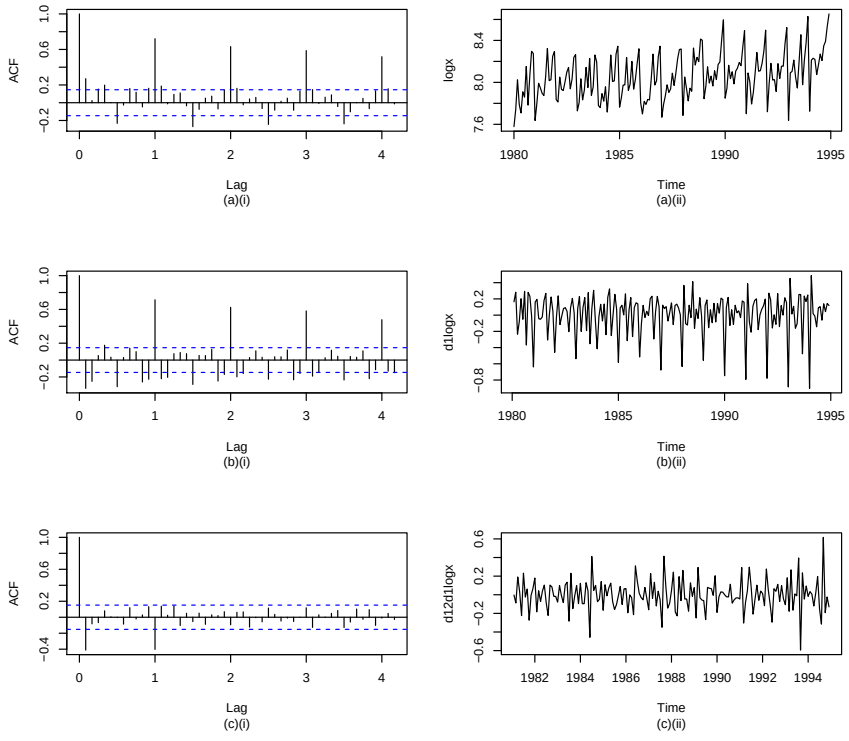

Figure 1: Paired plots of log sales with its acf before and after differencing

2. To finalise the first step of the modelling, we compare the acf and pacf of the differenced data plotted in (c)(ii) of Figure 1 to obtain the values of p, q, P and Q. The plot in Figure 2 shows the acf and pacf next to each other to recognise a decay or cutoff in both plots. There seems to be a decay in the pacf plot, indicating that both $p = P = 0$. The acf plot on the other hand, the cutoff non-seasonally and seasonally happens at lag 1, therefore we have $q = Q = 1$. Altogether, the plots in Figure 1 and 2 implies that the fitted model is

an ARIMA$(0,1,1) \times (0,1,1)_{12}$ which is also known as an airline model.

$$(1-B)(1-B^{12})x_t = (1+\theta_1 B)(1+\Theta_1 B^{12})\epsilon_t$$

where ϵ_t is white noise with mean 0, $B^i x_t = x_{t-i}, i = 1, 2, ...$ and θ_1 and Θ_1 are constant parameters to be estimated.

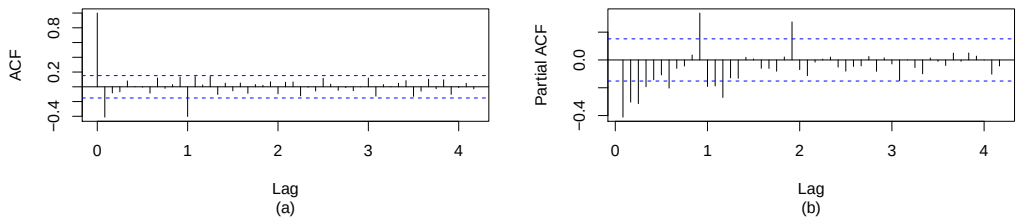

Figure 2: acf plot vs pacf plot of log sales after differencing

3. The parameters can be estimated by the maximum likelihood method using the function `arima` in R as follows, completing the fitted model,

$$(1-B)(1-B^{12})x_t = (1-0.859B)(1-0.684B^{12})\epsilon_t$$

```
> fit = arima(logx, order = c(0,1,1),
    seasonal = list(order = c(0,1,1), period = 12))
> fit$coef
       ma1       sma1
-0.8589073 -0.6838151
```

4. We proceed the modelling by investigating the residuals through diagnostic plots in Figure 3. The main criteria for a reasonable fit that needs to be met is that the residuals must behave like white noise. Based on the plot of the residuals, we notice that there does not seem to be any trend that the series follow, telling us that the mean is constant or independent of time. Another observation that can be raised is the values seem to be symmetrically distributed around a horizontal line $\epsilon = 0$. To check that residuals are uncorrelated the Ljung-Box test for the first k residuals is used,

$$H_0 : r_\epsilon(1) = ... = r_\epsilon(k) = 0 \text{ vs } H_A : r_\epsilon(1) \neq 0 \text{ for some } i = 1, ..., k$$

The test can be evaluated through the plot of the residual acf where if the acf points are within bounds ($\pm 2/\sqrt{n}$, $n = 168$ is the length of time series) dotted in blue, the residual correlation is nonsignificant, which is what we aim for. All the residual correlations are within these bounds, and to further confirm this we can take a look at the plot of the p-values at different lags below it. All the points are over 0.05, showing that there is *not* enough evidence to reject the null hypothesis. This indicates that the residuals are white noise since they are uncorrelated and with mean 0 and fairly identical variance.

The observations made on the variance of the residuals lead us to question its Gaussianity which can be visualised through a Normal Q-Q plot. This is plotted Figure 4, where we see that the data points fairly follow the reference straight line, which is an indication that the residuals do follow a Normal distribution with constant mean and variance. Doing a Shapiro-Wilk test on the null hypothesis that the residuals follow a Normal distribution, the p-value ($0.01742 < 0.05$) obtained opposes the observation at 5% significance level. Therefore, there is enough evidence to reject said null hypothesis.

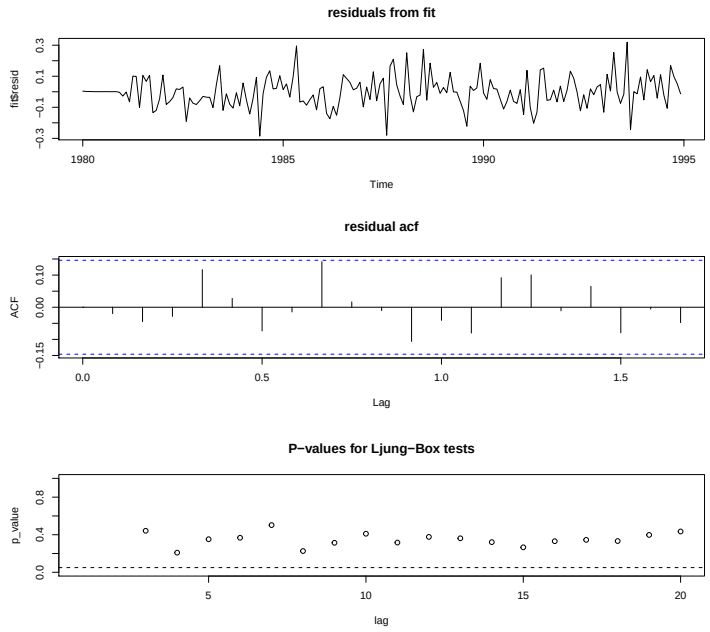

Figure 3: Diagnostic plots of log sales residuals

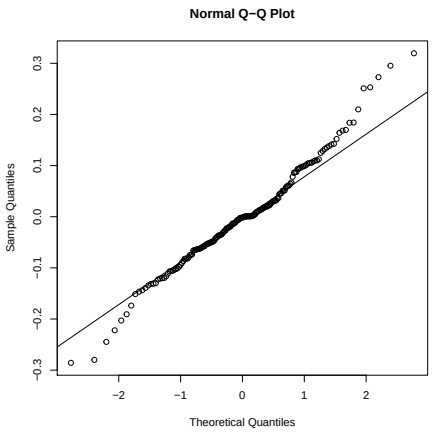

Figure 4: Quantile-quantile plot studying the Normality of residuals

```
> shapiro.test(residfit)

Shapiro-Wilk normality test

data:  residfit
W = 0.98153, p-value = 0.01742
```

2 Reciprocal of Sales Data Analysis

5. The same analysis can be done on a different transformed data, in this case, the reciprocal of sales. With a similar framework as the first section, the final plot (c)(ii) in Figure 5

with the observed downward trend and seasonal pattern removed is done by differencing the data once then further applying a seasonal differencing once as well. From this we get that the $d = D = 1$. The plot of acf and pacf of the data reciprocals in Figure 6 shows a decay in the pacf and cutoffs in the acf. Based on similar observations, the order of the fitted model is similar to the fitted model for log sales, ARIMA$(0, 1, 1) \times (0, 1, 1)_{12}$.

$$(1 - B)(1 - B^{12})x_t = (1 + \theta_1 B)(1 + \Theta_1 B^{12})\epsilon_t$$

where ϵ_t is white noise with mean 0, $B^i x_t = x_{t-i}, i = 1, 2, ...$ and θ_1 and Θ_1 are constant parameters to be estimated.

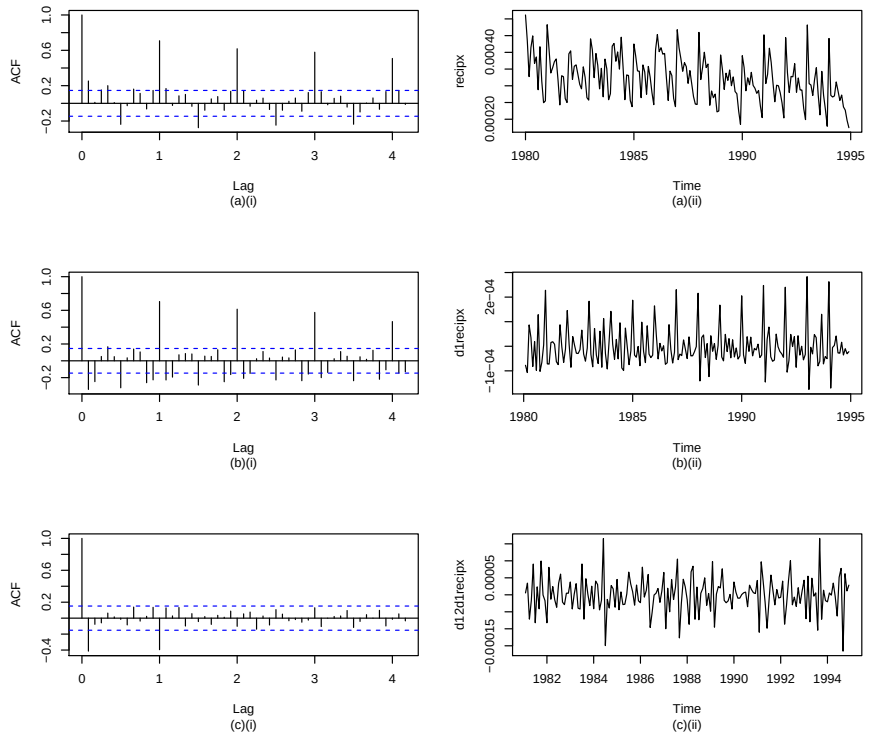

Figure 5: Paired plots of sales reciprocals with its acf before and after differencing

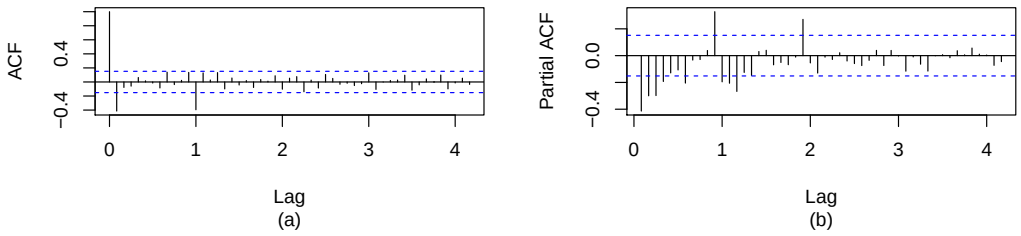

Figure 6: acf plot vs pacf plot of reciprocals of sales after differencing

Similar R codes that was presented for estimating the parameters for log sales can be implemented to the reciprocals.

```
> fitrecip = arima(recipx, order = c(0,1,1),
    seasonal = list(order = c(0,1,1), period = 12))
> fitrecip$coef
       ma1       sma1
-0.8508684 -0.6827482
```

Altogether,

$$(1-B)(1-B^{12})x_t = (1-0.851B)(1-0.683B^{12})\epsilon_t$$

is the model suggested to fit the time series. It is worth noting that the estimated parameters are very close to when the data was initially log transformed, with less than 0.01 difference for both parameters. This suggests that the fitted model is close to the true model of the time series.

To assess the model, diagnostic plots of the residuals are explored. The first plot in Figure 7 tells us a little bit about the mean and variance that the residuals follow. We see very clearly that the residuals seem to be symmetrically distributed along $\epsilon = 0$ which suggests that that is the value of the mean. To test that they are uncorrelated, the acf and p-value plot is used. In the acf plot, at lag 0.667 (the 8th month), the value exceeds the boundary, implying that at that lag the residual correlation is significant. However, the p-value plot shows no significant evidence to reject the null hypothesis that the residuals are uncorrelated at the 5% level. It is also useful to check the Gaussianity of the residuals through a Normal Q-Q plot. The data points in Figure 8 form a fairly straight line along the reference line, telling us that there is evidence that the residuals do follow a Normal distribution. This is further confirmed by the Shapiro-Wilk test, where the p-value ($0.08617 > 0.05$) *is* significant at 5% level.

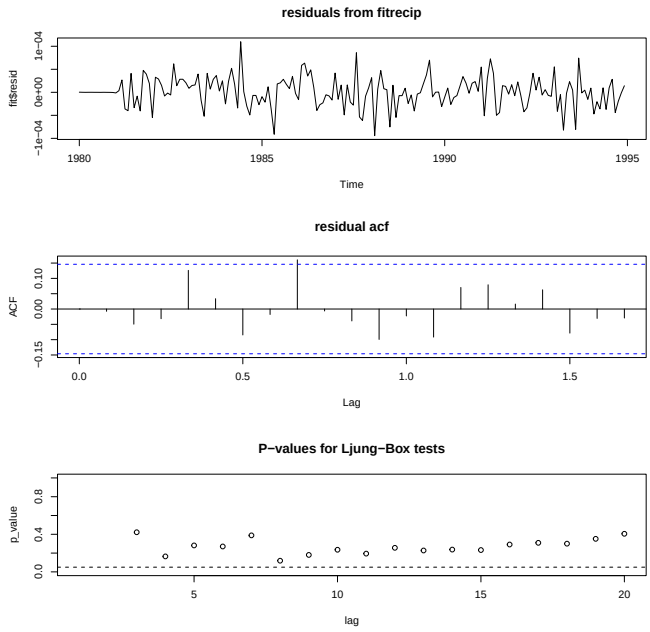

Figure 7: Diagnostic plots of sales reciprocals residuals

```
> shapiro.test(residfitrecip)
```

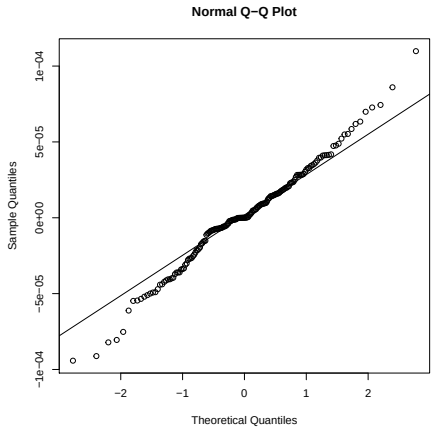

Figure 8: Quantile-quantile plot studying the Normality of residuals

```
Shapiro-Wilk normality test

data:  residfitrecip
W = 0.98667, p-value = 0.08617
```

3 Prediction for period Jan 1995 to July 1995

6. In this final and short section of this report, we use the fitted model to predict "future" data for Jan to July 1995, the final 7 values available in the dry white wine dataset. We apply the `predict()` function in R to the fitted model of the reciprocal transform of the data. This is because its fitted model has the higher log-likelihood value of 1475.236 compared to the fit of the log sales with only 135.9041.

```
> predrecip = predict(fitrecip, n.ahead = 7)
> actual = c(2367, 3819, 4067, 4022, 3937, 4365, 4290)
> cbind(1/predrecip$pred, 1/(predrecip$pred-1.96*predrecip$se),
    1/(predrecip$pred+1.96*predrecip$se), actual, 1/predrecip$pred - actual)
```

The final code returns the predicted values, its corresponding upper and lower bound of the 95% prediction interval using the standard error obtained from `predict()`, as well as the true values available in the dataset alongside its difference with the predicted values which are all summarised in Table 1.

Period	Predicted sales	95% Prediction Interval	Actual Sales	Difference
Jan 1995	2439.504	(2918.455, 2095.594)	2367	72.50397
Feb 1995	3606.755	(4779.172, 2896.252)	3819	-212.24526
Mar 1995	3863.350	(5260.652, 3052.550)	4067	-203.65003
Apr 1995	3637.318	(4867.423, 2903.531)	4022	-384.68205
May 1995	3458.191	(4566.885, 2782.652)	3937	-478.80886
Jun 1995	3412.958	(4502.743, 2747.893)	4365	-952.04245
Jul 1995	4226.683	(6061.641, 3244.516)	4290	-63.31663

Table 1: Predicted and true sales reciprocals by period and its corresponding 95% prediction interval

Overall, the values of the predicted sales are very close to the true sales with the largest difference is 952.04 thousand litres for June 1995. This is evident in the fact that all the actual sales fall within the 95% prediction interval. It seems that the fitted model *is* suitable for predicting future values.

All in all, it is safe to say that the ARIMA model suggested in the second section of this discussion is a good fit for the dry white wine time series by looking at the distribution of its residuals of the initial reciprocal transform of the data as well as its ability to produce prediction intervals that include all of the true values of the sales.

Appendix

```
source("tsdiags.R")
x = scan("drywhitewine.txt")
x = ts(x, start = c(1980, 1), end = c(1994, 12), frequency = 12)

logx = log(x)
par(mfrow = c(3,2))
acf(logx, lag = 50, sub = "(a)(i)", main = "")
plot(logx, sub = "(a)(ii)")
d1logx = diff(logx)
acf(d1logx, lag = 50, sub = "(b)(i)", main = "")
plot(d1logx, sub = "(b)(ii)")
d12d1logx = diff(d1logx, lag=12)
acf(d12d1logx, lag = 50, sub = "(c)(i)", main = "")
plot(d12d1logx, sub = "(c)(ii)")

#2
par(mfrow = c(1,2))
acf(d12d1logx, lag = 50, main = "", sub = "(a)")
pacf(d12d1logx, lag = 50, main = "", sub = "(b)")

#3
fit = arima(logx, order = c(0,1,1), seasonal = list(order = c(0,1,1),
    period = 12))
fit$coef

#4
par(mfrow = c(1,1))
tsdiags(fit)

par(mfrow = c(1,1))
residfit = resid(fit)
qqnorm(residfit);qqline(residfit)
shapiro.test(residfit)

#5
recipx = 1/x
par(mfrow = c(3,2))
acf(recipx, lag = 50, sub = "(a)(i)", main = "")
plot(recipx, sub = "(a)(ii)")
d1recipx = diff(recipx)
acf(d1recipx, lag = 50, sub = "(b)(i)", main = "")
plot(d1recipx, sub = "(b)(ii)")
d12d1recipx = diff(d1recipx, lag=12)
```

```
acf(d12d1recipx, lag = 50, sub = "(c)(i)", main = "")
plot(d12d1recipx, sub = "(c)(ii)")

par(mfrow = c(1,2))
acf(d12d1recipx, lag = 50, sub = "(a)", main = "")
pacf(d12d1recipx, lag = 50, sub = "(b)", main = "")

fitrecip = arima(recipx, order = c(0,1,1), seasonal = list(order = c(0,1,1),
    period = 12))
fitrecip$coef

tsdiags(fitrecip)

par(mfrow = c(1,1))
residfitrecip = resid(fitrecip)
qqnorm(residfitrecip); qqline(residfitrecip)
shapiro.test(residfitrecip)

#6
fit$loglik
fitrecip$loglik
predrecip = predict(fitrecip, n.ahead = 7)
actual = c(2367, 3819, 4067, 4022, 3937, 4365, 4290)
cbind(1/predrecip$pred, 1/(predrecip$pred-1.96*predrecip$se),
    1/(predrecip$pred+1.96*predrecip$se), actual, 1/predrecip$pred - actual)
```