PETROLEUM CONSUMED BY THE U.S. RESIDENTIAL SECTOR: ESTIMATING AND FORECASTING

Statistics 137

Applied Time Series Analysis

Audrey Chu

999138148

June 2016

Abstract

To estimate the trend and seasonal components, model the stationary subset, and confirm how well the estimated model is able to forecast the data series. With the application of time series modeling methods, an autoregressive moving average (ARMA) model will best fit this real-life dataset of petroleum consumption throughout time.

I. Introduction

This study analyzes the dataset containing the petroleum consumed (in trillion BTU) by the residential sector in the U.S. from January 1984 to December 2015. Petroleum consumption at time t is denoted by Y_t, t = 1, 2, …, n = 384, with the understanding that time 1 corresponds to the month of January 1984 and time 384 corresponds to December 2015. According to the U.S. Energy Information Administration (EIA), petroleum products include transportation fuels, fuel oils for heating and electricity generation, and asphalt and road oil. The current model is

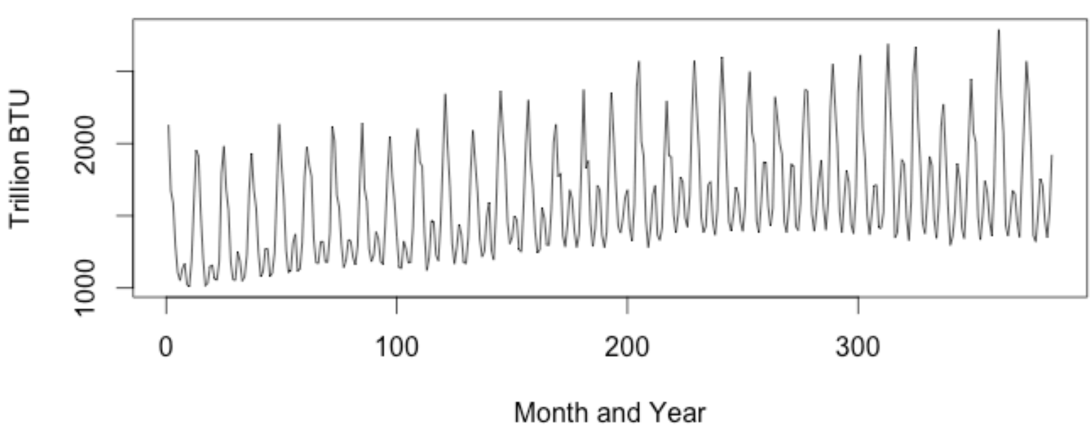
$$Y_t = m_t + s_t + X_t, \; t = 1, \ldots, n = 384 \qquad \text{(a)}$$

where m_t is the trend (smooth part), s_t is the seasonal component, and X_t is the rough part. This report includes justifications of estimations through support of diagnostic plots including autocorrelation (ACF), partial autocorrelation (PACF), and periodograms. The plot of the untransformed petroleum units consumed, shown in Figure 1, has three components of trend, seasonality, and a rough part. Series plot exemplifies non-stationarity through slightly differing variance and mean, and thus will require a Box-Cox or power transformation.

The general trend is upward sloping whereas the seasonality period is roughly between 10 and 15 months. Upward trend is driven by many factors, two of which are the growing number of technology devices as well as the increasing average home size. Trend operates under the assumptions that energy consumption increases as more home entertainment and rechargeable devices are brought into homes. Additionally, the EIA states that the average floor area and home size in the U.S. has also shown an upward trend; specifically, homes built since 1990 are on average 27% larger than homes built in earlier decades (RECS 2009). Since electricity powers heating and cooling equipment and cooking appliances, larger homes suggest a larger burden of energy consumption.

Nature of these seasonal variations is highly correlated with weather due to increase usage of heating equipment for months earlier in the year such as January, explaining high data points. Additionally, cooling units require less energy than heating, explaining low data points (Figure 1.1). Oscillations imply strong presence of seasonality. Autoregressive moving average [ARMA(p, q)] or autoregressive integrated moving average [ARIMA(p, d, q)] may be appropriate here to estimate the rough part of the data.

Figure 1
U.S. Residential Sector Petroleum Consumption



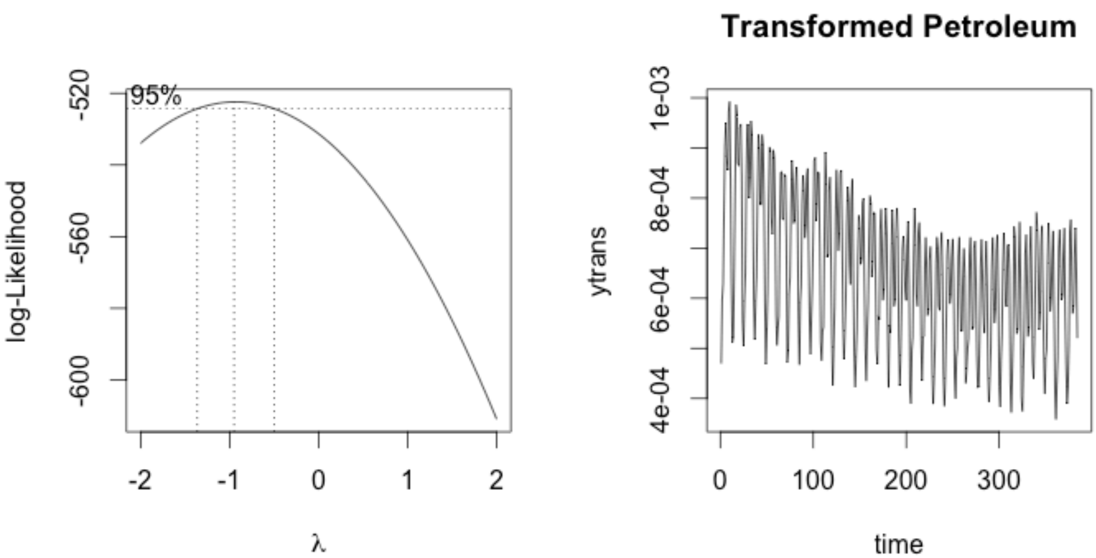
II. METHODOLOGY

Transformation Selection

Power transformations of Y_t^{-1}, $\sqrt{Y_t}$, $1/\sqrt{Y_t}$, and $\ln(Y_t)$ or where λ = -1.0, 0.5, -0.5, and 0, respectively, are provided in graphical plots in Figure 2 and Figure 2.1 of the Appendix. Visually, variance looks equal in either Y_t^{-1} or $\ln(Y_t)$. To further clarify, the Box-Cox transformation plot shows that the optimal value is closest to negative one. Power transformation function gives the optimal value at -0.9299752, supporting the box-cox plot. Although the value is not exactly negative one, setting λ = -1.0 will provide for an optimal transformation considering simplicity.

Additionally, a secondary check of the optimal λ value is obtained through the function 'lamopt' from the 'trndseas' file, and further upholds its proximity to zero. This transformation also has the highest R-squared at 0.1984 and the adjusted R-squared at .1963. Box-Cox transformations measure the maximum likelihood whereas power transformations measure normality. Power transformations are often generalizations of Box-Cox transformations, and can result in different but similar values. In conclusion, the best-transformed model seems to be Y_t^{-1}.

Figure 2.1



Estimation

Analysis will estimate the transformed data, denoted $Ytrans_{t,}$ $[Y_t^{-1}]$, from this point onwards. The trend is estimated using a fitted polynomial of degree 2 with model

$$Ytrans_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t \tag{b}$$

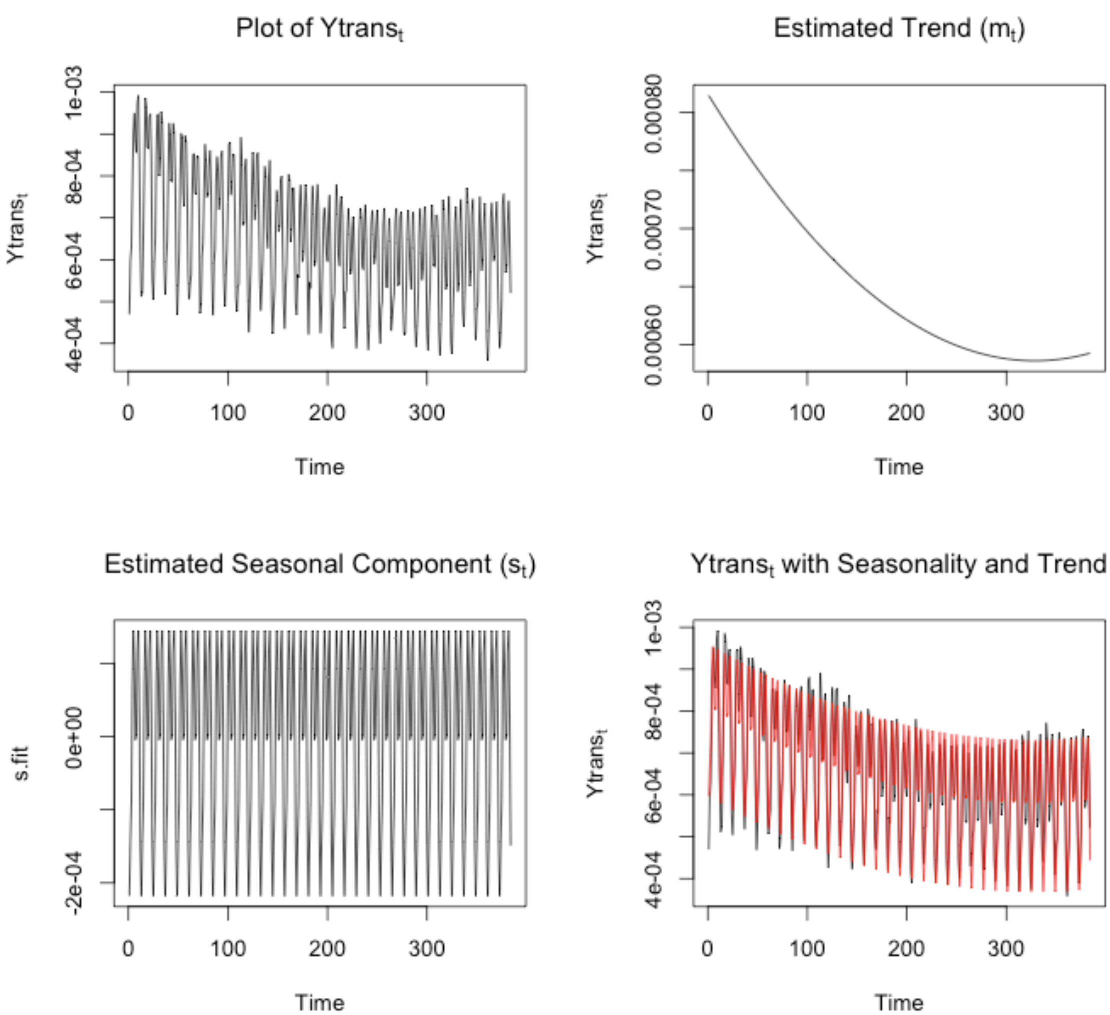
and if $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, are the least square estimators of β_0, β_1, and β_2, then the estimated trend is

$$\hat{m}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2 \tag{c}$$

Degree choice of model is selected through graphical estimations using 'trndseas' function and through a process of increasing stepwise degrees starting from 1. Polynomial with degree 2, otherwise known as the quadratic, and seasonal component of 12 best estimates the upward trend and fluctuations. Seasonal component of period 12 supports earlier predictions of annual variations due to higher energy consumption from heating units during the beginning and ending months and low consumption during the remaining months. Estimation of trend with quadratic model and seasonality with period

12 also shows highest R-squared value at 0.9285714 when compared to models of differing degrees. This model using the transformed data states that the optimal λ is 1, noting that this transformed data is indeed a good adjustment. Figure 3 depicts estimated trend with polynomial degree 2 and estimated seasonality with period length 12. The approximation is represented with a red line and compared with the transformed data.

Figure 3



Fitted model results in $\hat{\beta}_0 = 8.161081*10^{-04}$, $\hat{\beta}_1 = -5.381070*10^{-04}$, and $\hat{\beta}_2 = 3.145122*10^{-04}$. Note that $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are coefficients of transformed data and will require reverse transformation at the end of the analysis.

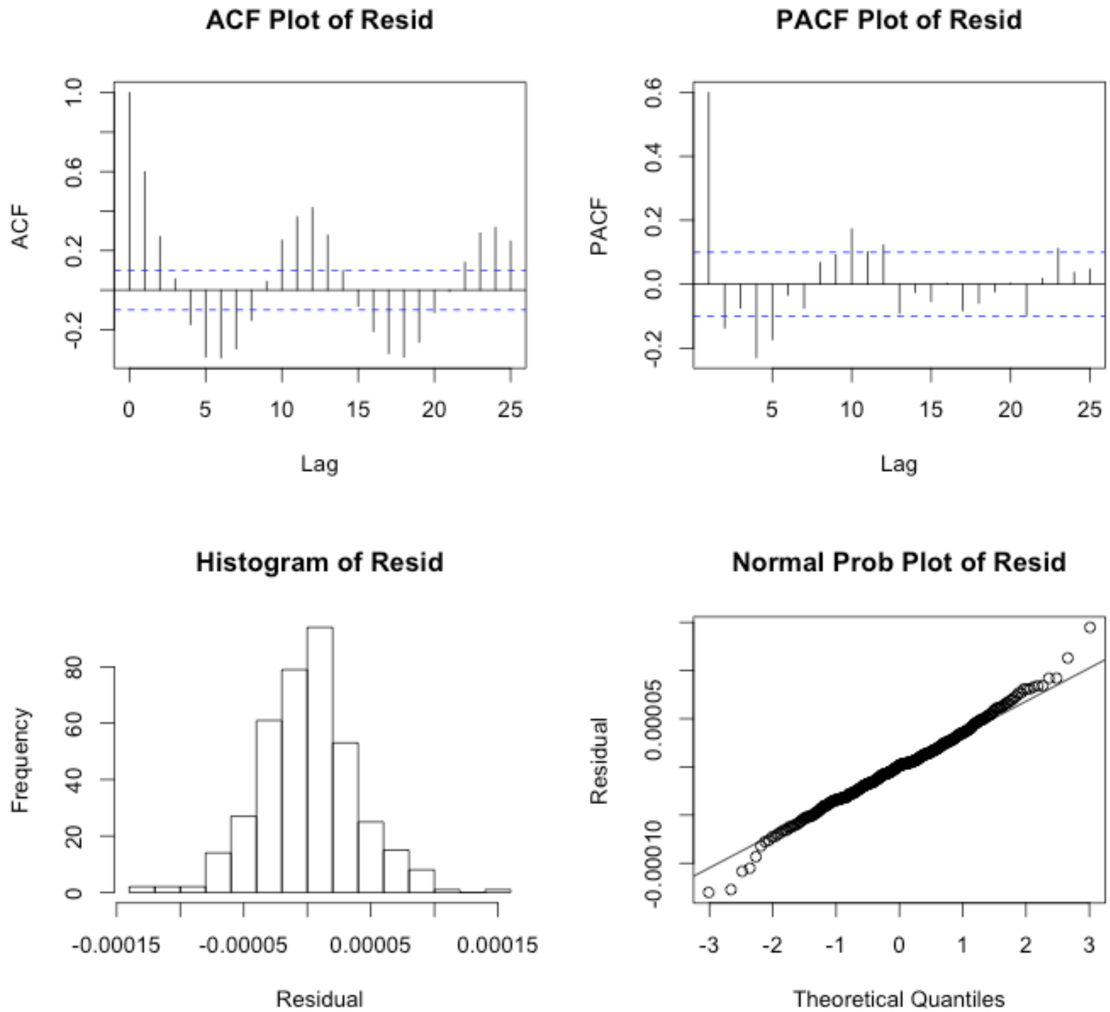
Trend and seasonality estimations allow for modeling of the rough part $\{X_t\}$

$$X_t = Y_t - s_t - m_t, t = 1, \ldots, n = 384 \qquad \text{(d)}$$

The rough part satisfies a weak stationary sequence if it contains three properties: $E[X_t] = \mu$ for all t, $Var(X_t)$ is the same for all t, and $Corr(X_t, X_{t+j}) = \varrho(j)$ is the same for all t and any positive integer *j*. Graphical visuals of transformed data suggest that the rough series is neither independent nor identically distributed (i.i.d.). Box-Ljung test shows a p-value of 2.2e-16, rendering it not independent as well.

Autocorrelation function (ACF) plots of the residuals, $\{X_t\}$, show that the lags oscillate out of the 95% confidence interval where $\alpha = 0.05$ from positive to negative values, starting with a larger deviation. The lags have strong dependence and fluctuate in approximately intervals of lag 5. Partial autocorrelation function (PACF) plots show that the lags also oscillate. It is important t note that the lags of the PACF plot approach the 95% threshold as the lag number increases, with lag 10 begin the last and most obvious lag outside this threshold. Histogram ranges from -0.00015 and 0.00015 and QQ plot show that residuals are relatively normal distributed.

Figure 4



Raw periodogram and smoothed periodogram show a peak roughly at frequency 0.08. The periodogram is a nonparametric estimate of the power of spectral density of a stationary random process and smoothing is achieved through local averaging. Smoothing also eliminates the roughness associated with a raw periodogram. Here, averaging over 7 months is reasonable.

Autoregressive Moving Average [ARMA(p, q)] Modeling

Fitting multiple autoregressive moving average models of order $p = 1, \ldots, 5$ and $q = 4$ or 5, the Akaike's Information Criterion (AICc) is the lowest for ARMA(4, 5) at -6975.393. AICc for ARMA(0, 0) has the largest value at -6727.784 for tested orders, suggesting that $\{X_t\}$ is not i.i.d. In application of ARMA model, it is useful to apply the 'auto.arima()' function since there are many combinations of orders. Function returns the optimal model as ARMA(4, 0, 5) where 0 is the difference in addition to returning a smaller AICc value at -6988.6. Note that ARIMA(p, d, q) has not been tested since estimation is done of data that has already been deseasonalized and detrended. The ARMA(4, 5) model has the form

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \ldots + \phi_4(X_{t-4} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_5\varepsilon_{t-5} \qquad \text{(e)}$$

which can rewritten as

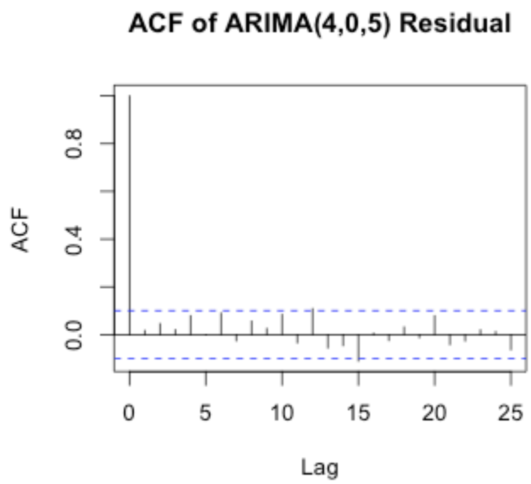
$$Ytrans_t = \beta_0 + \beta_1 X_{t-1} + \ldots + \beta_4 X_{t-4} + \beta_5\varepsilon_{t-1} + \ldots + \beta_9\varepsilon_{t-5} \qquad \text{(f)}$$

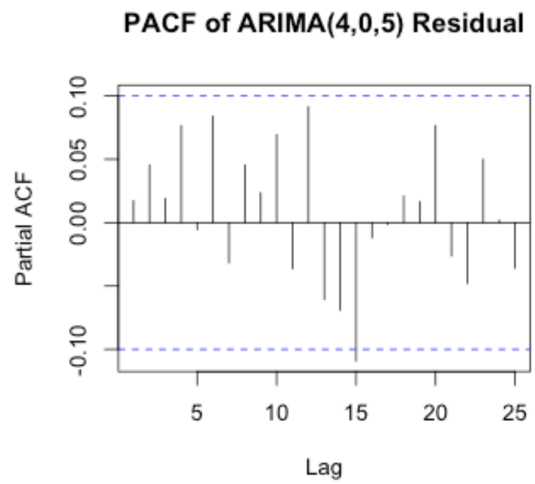
where $\beta_0 = 3.196369e\text{-}07$, $\beta_1 = 1.946475$, $\beta_2 = -2.251592$, $\beta_3 = 1.729474$, $\beta_4 = -.8695894$, $\beta_5 = -1.424588$, $\beta_6 = 1.371674$, $\beta_7 = -7.272223$, $\beta_8 = 4.589010$, and $\beta_9 = 3.808658$

ACF and PACF plots of ARMA(4, 5) model in Figure 5 show that all lags are now insignificant, with the exception of lag 0 for the ACF plot. Lag 0 for ACF plot will always have a correlation of 1 due to its perfect relationship with itself. Histogram of ARMA(3) residuals also shows that residuals are normally distributed in a smaller span. Figure 5.2 in Appendix shows a better representation with a boxplot. Box-Ljung test produces p-value of 0.3322, which fails to reject the null hypothesis $H_0 = \varrho(1) = \ldots = \varrho(j)$

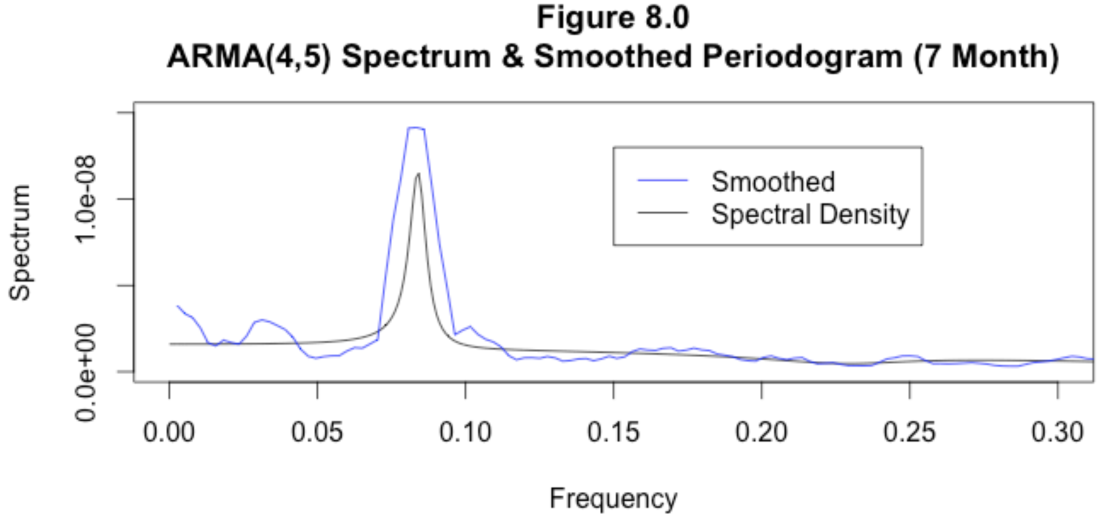
= 0. ARMA(4, 5) residuals consequently acts as white noise and is identically and independently distributed.

Figure 7.0





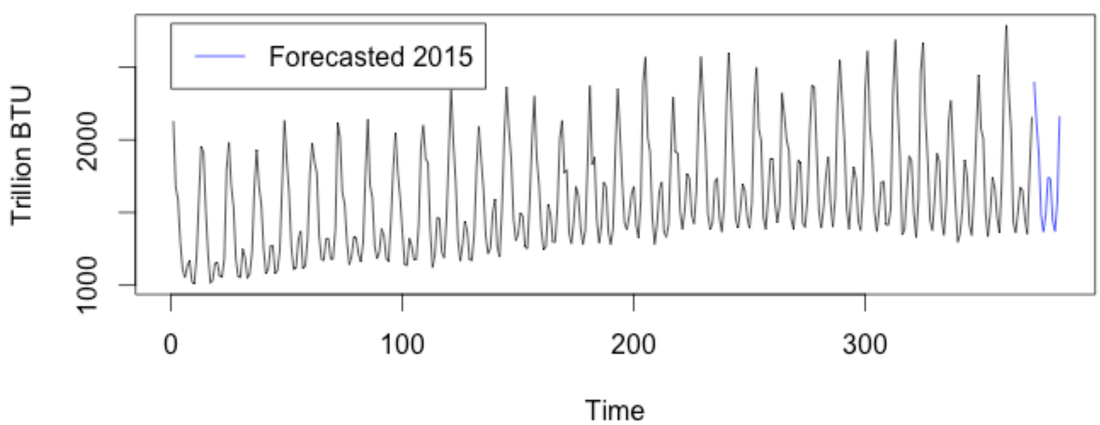
Spectral density of the ARMA(4, 5) is plotted against the smoothed periodogram average over 7 months in Figure 8.0. The estimator with the higher peak is the smoothed periodogram estimate. Points move in similar nature and have similar peaks.



Prediction

Applying values found through trend, seasonality, and rough estimation, a prediction on future values plotted on a graph will show the accuracy of the analysis. Investigating only data from time 1 to time 372, forecasted values of the next year (2015) are plotted with real data points to show the fit. This is obtained after reverse transformation through again taking the transformed data to the power of negative one $[(Y_t^{-1})^{-1}]$ where $Ytrans_t$ contains transformed m_t, s_t, and X_t.

Figure 9.0
Petroleum Prediction from December 2014



III. Conclusion

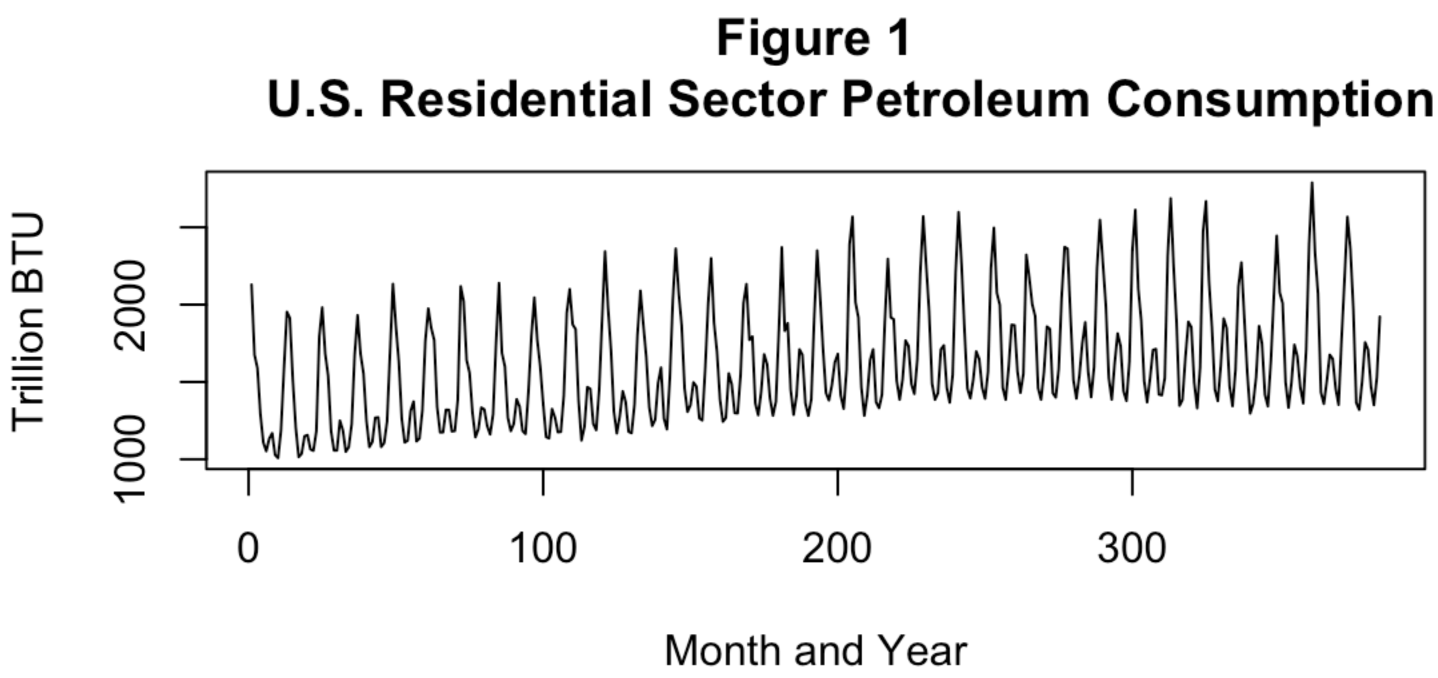
In analyzing petroleum consumption by the U.S. residential sector throughout the duration of January 1984 to December 2015, dataset was grouped into three components: trend (smooth), seasonality, and rough. Since the original dataset showed no sign of stationarity through variable means and differing variance, a power transformation or Box-Cox transformation of $\lambda = -1$ was necessary to produce a model most similar to stationarity and equivariance. Estimation was then produced using the transformed data. Trend was best estimated through a polynomial model of degree 2 and seasonality was best estimated with a period span of 12. This seasonality period is due to higher energy consumption through heating equipment in homes during the months of December and January. The rough part was fit with an autoregressive moving average model of order 4 and order 5, respectively [ARMA (4, 5)]. A series of diagnostic plots and tests showed that the rough fitted through an ARMA process resembles white noise and is identically and independently distributed. It is important to state that this time series analysis relied heavily on visualizations and diagnostics. The calculated predicted values also verify that this model was indeed a precise approximation. Possible improvement of this analysis would be incorporating or implementing an additional check using only deseasonalized data and finding the best ARIMA (p, d, q) model. Ideally, this would produce an equally accurate estimation. Moreover, forecasting the data on future values would provide for an enhanced measure of this analysis since this analysis only provided predicted values based on the whole dataset already containing the predicted time periods.

Appendix: Diagnostic Plots and R Code

Audrey Chu

Original Data

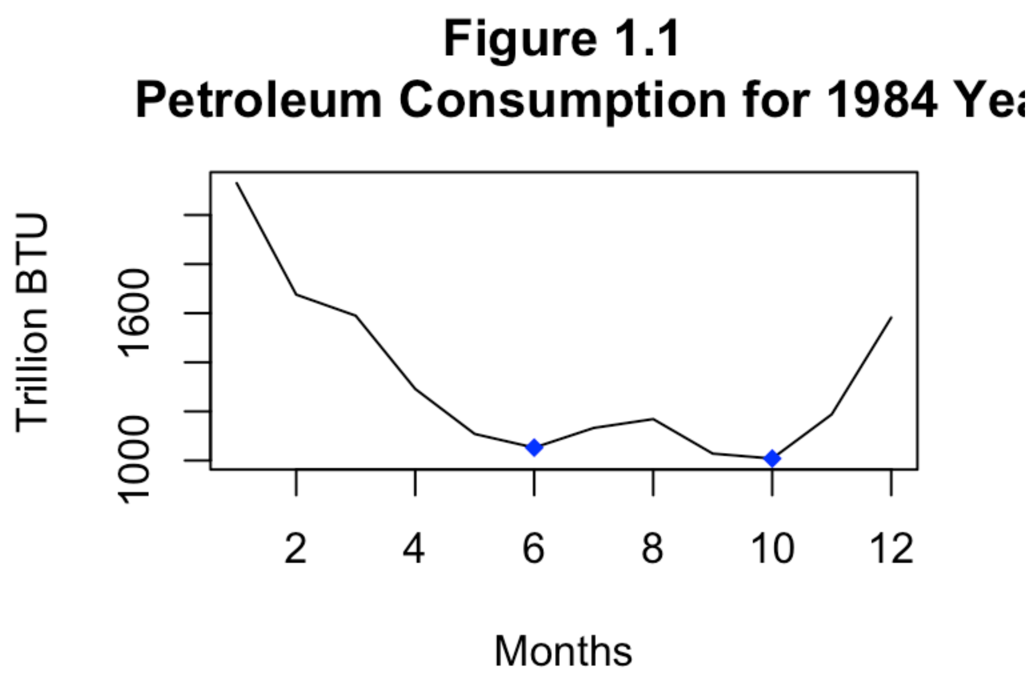
```
time = 1:384
y = read.table("~/Documents/3rd Year/STA137/EnergyConsumption.txt", row.names=NULL)[,
3]
plot(time, y, type='l', main = "Figure 1
     U.S. Residential Sector Petroleum Consumption", ylab="Trillion BTU", xlab="Month
and Year")
```




```
model = lm(y~time)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ time)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -535.3 -278.2 -127.4  241.4  953.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1360.6754    35.4581  38.374  < 2e-16 ***
## time           1.3132     0.1596   8.227 3.05e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 346.7 on 382 degrees of freedom
## Multiple R-squared:  0.1505, Adjusted R-squared:  0.1483
## F-statistic: 67.68 on 1 and 382 DF,  p-value: 3.052e-15
```

```
plot(1:12, y[1:12], type='l', main="Figure 1.1
     Petroleum Consumption for 1984 Year", ylab="Trillion BTU", xlab="Months")
points(6, y[6], col='blue', pch=18)
points(10, y[10], col='blue', pch=18)
```



Box-Cox Transformation

Figure 2.0

```
par(mfrow=c(2,2))
plot.ts(y^(-1),ylab=expression(paste("1/",Y[t],")")), main=expression(paste("1/(",Y[t
],")")))
plot.ts(y^.5,ylab=expression(paste("sqrt(",Y[t],")")), main=expression(paste("Plot of
Sqrt(",Y[t],")")))
plot.ts(y^-.5,ylab=expression(paste("1/sqrt(",Y[t],")")), main=expression(paste("Plot
of 1/sqrt(",Y[t],")")))
plot.ts(log(y),ylab=expression(paste("ln(",Y[t],")")), main=expression(paste("Plot of
ln(",Y[t],")")))
```

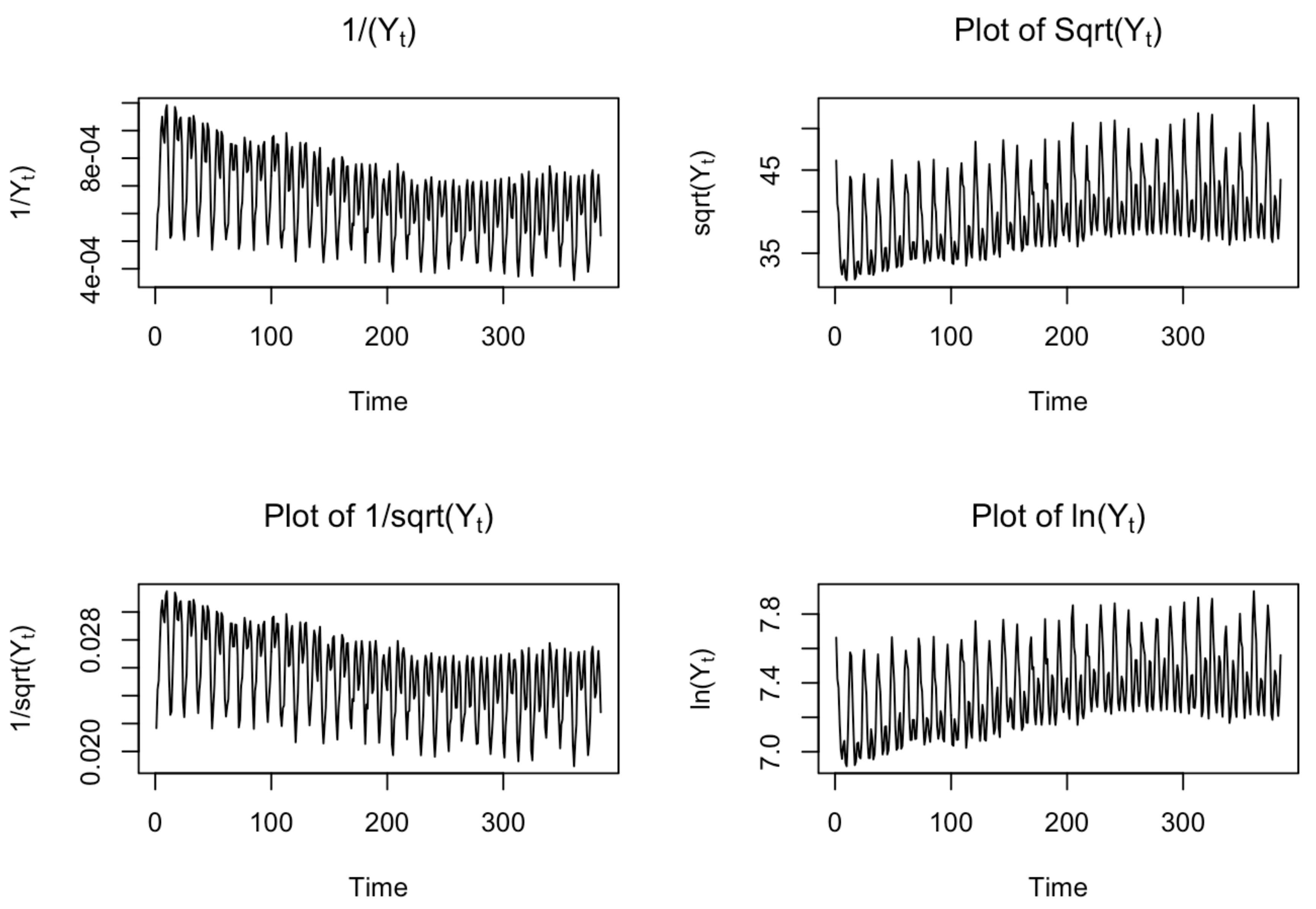
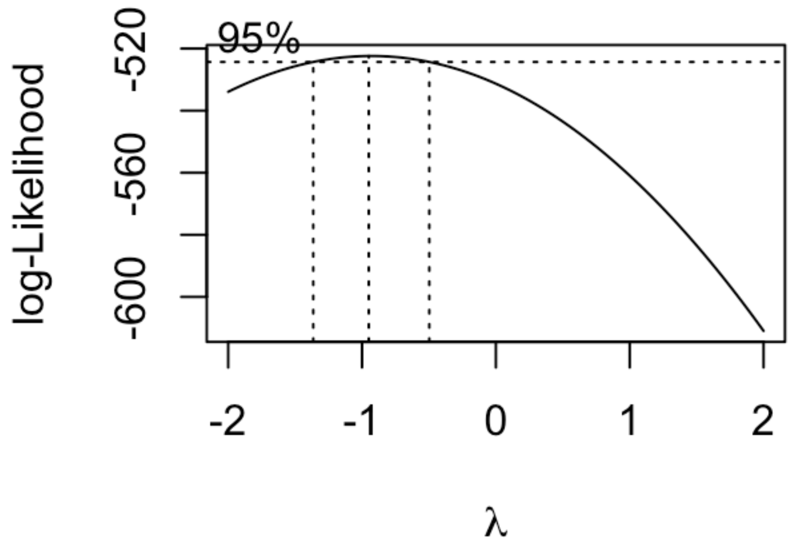


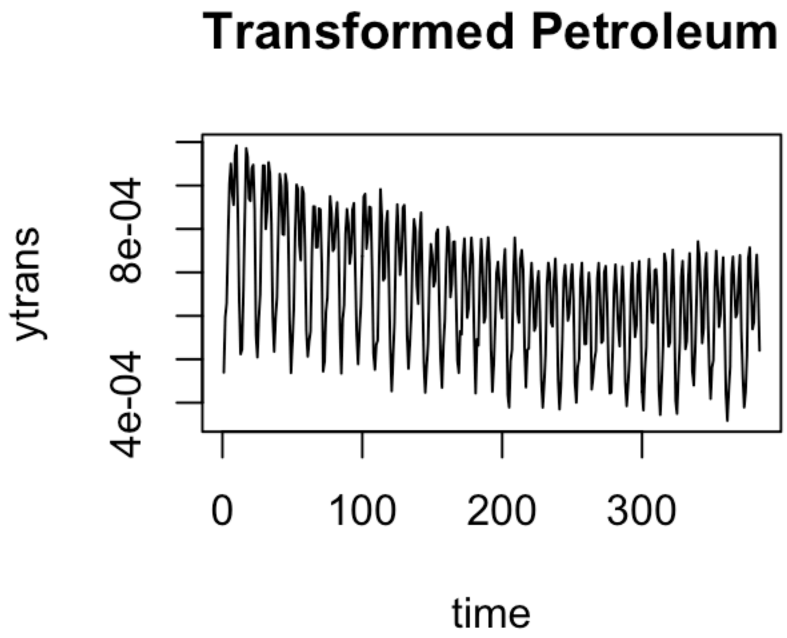
Figure 2.1

```
par(mfrow=c(1,2))
boxcox(model)
powerTransform(model)
```

```
## Estimated transformation parameters
##          Y1
## -0.9299752
```

```
ytrans = y^(-1)
modtrans = lm(ytrans~time)
plot(time, ytrans, type = 'l', main = "Transformed Petroleum")
```





Trend and Seasonal Estimation

Figure 3.0

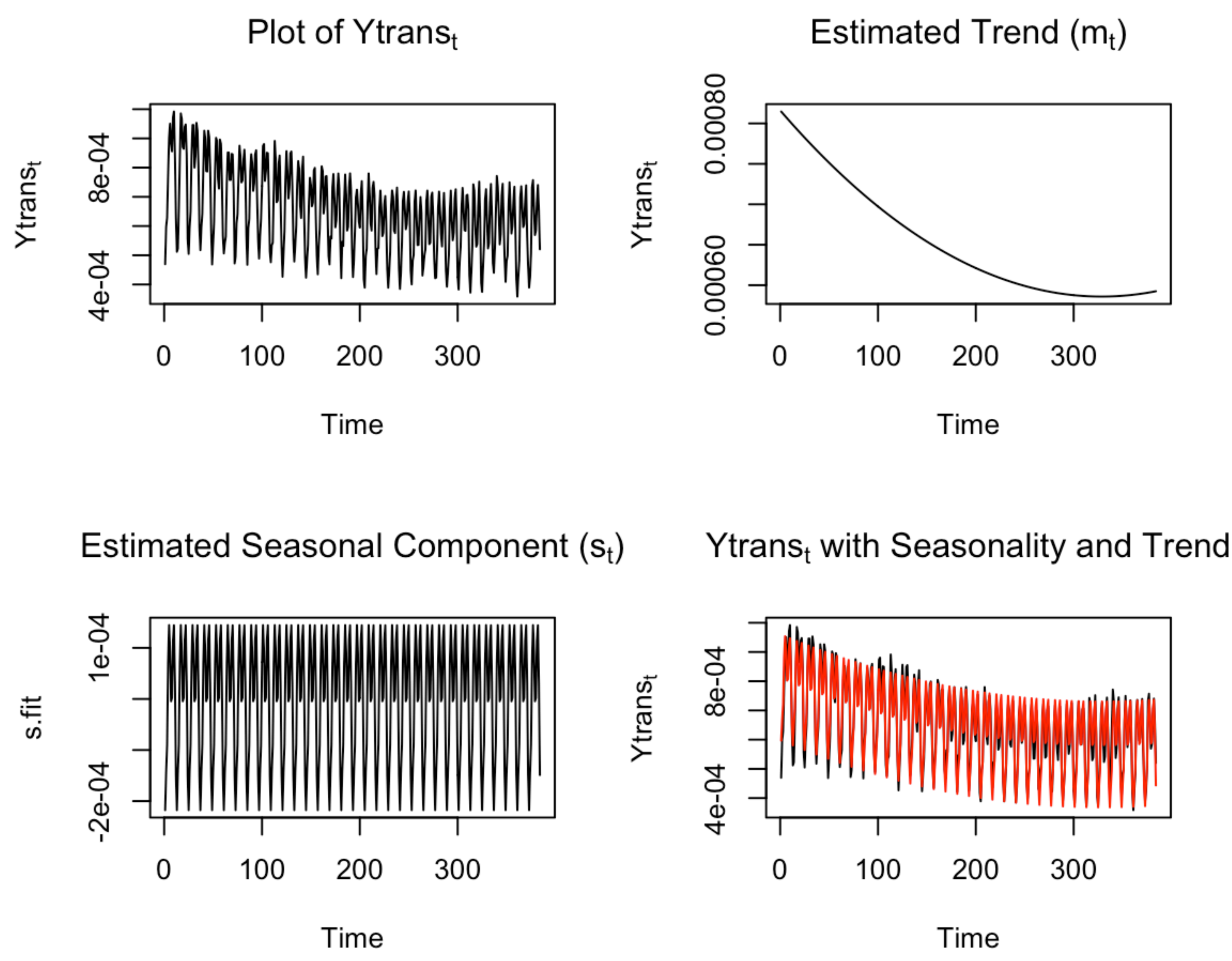
```
mod = trndseas(ytrans,degtrnd=2, seas=12)
mod$rsq
```

```
## [1] 0.9285714
```

```
mod$lamopt
```

```
## [1] 1
```

```
m.fit = mod$trend
s.fit = rep(mod$season,length.out=384)
par(mfrow=c(2,2))
plot.ts(ytrans ,ylab=expression(paste("",Ytrans[t],"")), main=expression(paste("Plot
of ",Ytrans[t],"")))
plot.ts(m.fit, ylab=expression(paste("",Ytrans[t],"")), main=expression(paste("Estima
ted Trend (",m[t],")")))
plot.ts(s.fit,main=expression(paste("Estimated Seasonal Component (",s[t],")")))
plot.ts(ytrans,ylab=expression(paste("",Ytrans[t],"")), main=expression(paste("",Ytra
ns[t]," with Seasonality and Trend")))
points(mod$fit,type='l',col='red')
```

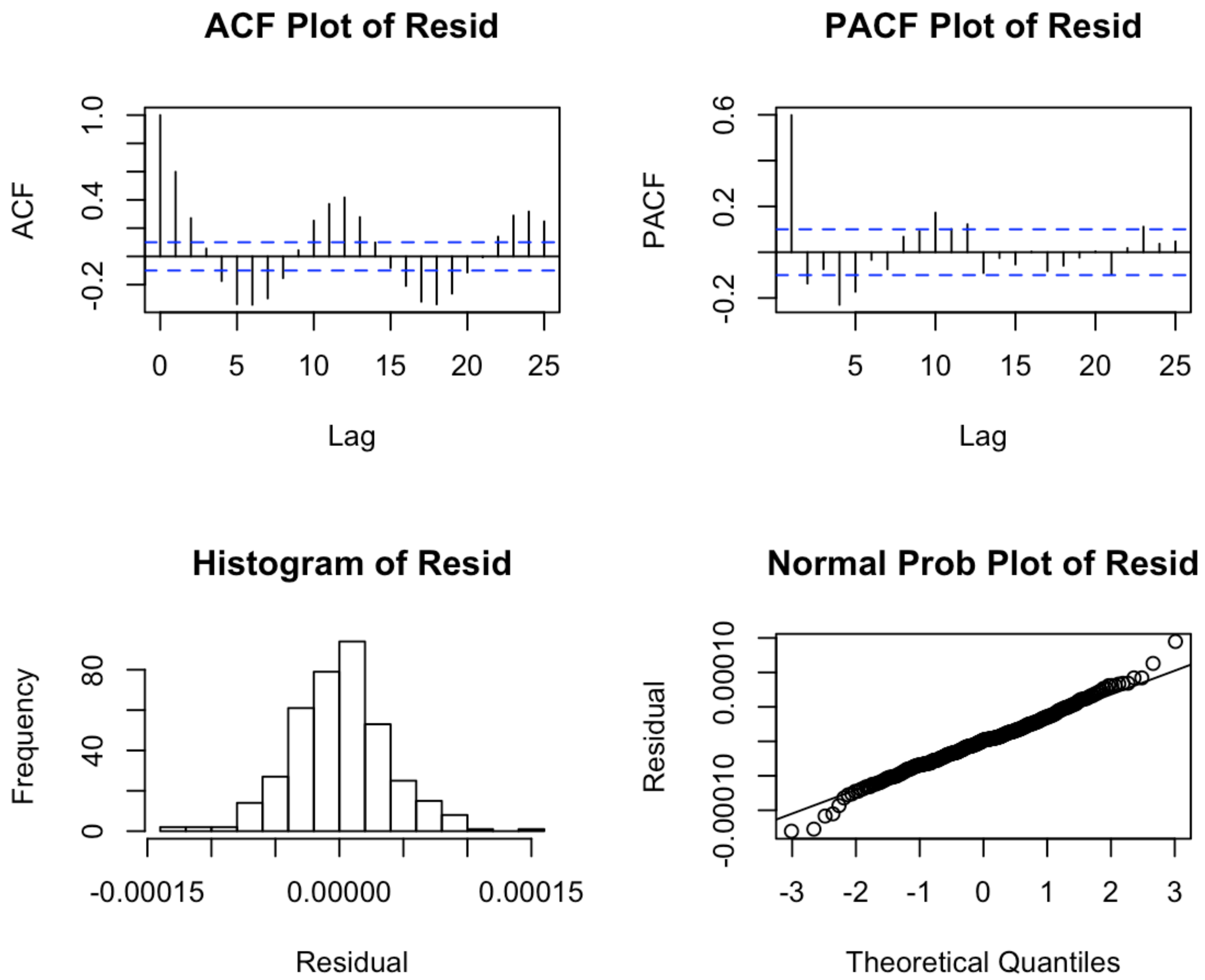



```
mod$coef
```

```
##                [,1]
##        8.161081e-04
## 1     -5.381070e-04
## 2      3.145122e-04
## x21   -2.177859e-04
## x22   -1.440324e-04
## x23   -8.967068e-05
## x24    7.204337e-05
## x25    1.446329e-04
## x26    9.314625e-05
## x27   -4.488970e-06
## x28    2.180358e-06
## x29    1.147087e-04
## x210   1.442814e-04
## x211   3.402836e-05
```

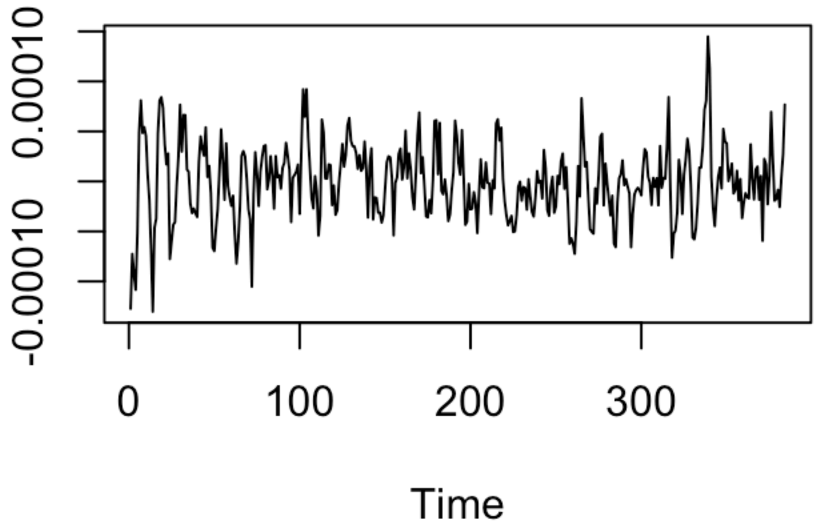
Rough Estimation

```
x = ytrans - m.fit - s.fit
par(mfrow=c(2,2))
acf(x, main= "ACF Plot of Resid", ylab = "ACF", xlab ="Lag")
pacf(x, main= "PACF Plot of Resid", ylab = "PACF", xlab = "Lag")
hist(x, main= "Histogram of Resid", ylab = "Frequency", xlab = "Residual")
qqnorm(x, main= "Normal Prob Plot of Resid",ylab="Residual"); qqline(x)
```




```
par(mfrow=c(1,1))
plot.ts(x, main = "Figure 4.1
        Plot of Residuals", ylab = "")
```

Figure 4.1
Plot of Residuals



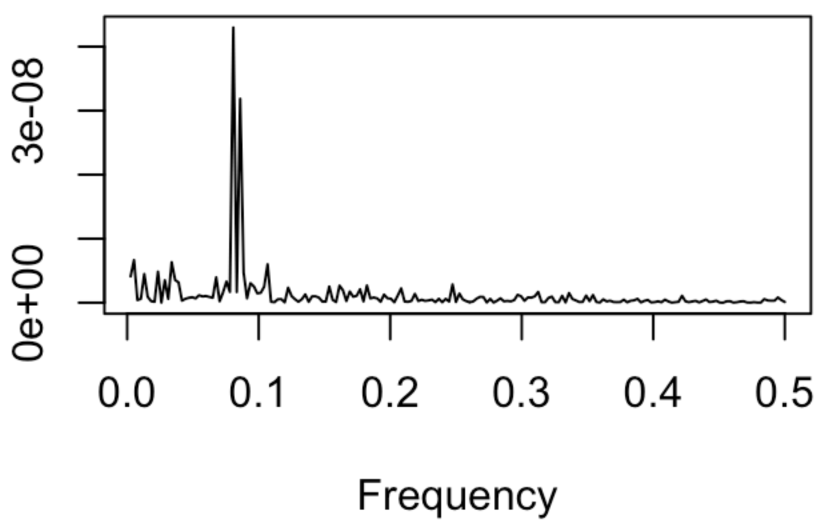

```
Box.test(x,lag=10,'Ljung-Box')
```

```
##
##  Box-Ljung test
##
## data:  x
## X-squared = 341.056, df = 10, p-value < 2.2e-16
```

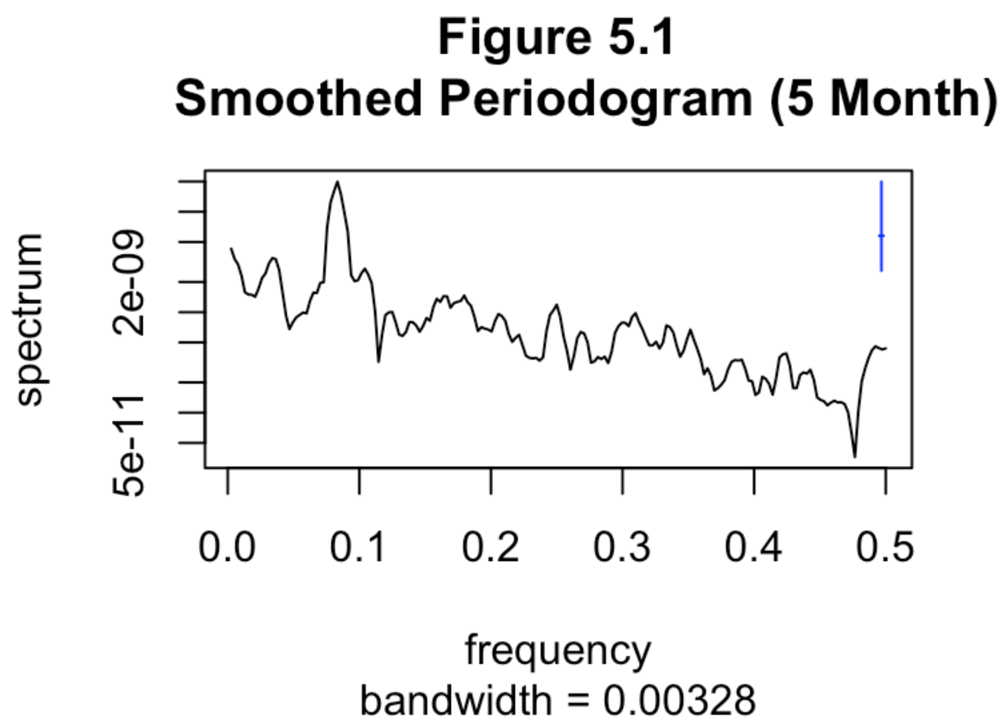
Box-Ljung test shows a very low p-value, which can be intepreted as significant. This means that the rough series is not independent.

```
xpgrm = spec.pgram(x,log='no',plot=F)
plot(xpgrm$freq,xpgrm$spec,type='l',xlab='Frequency',ylab='', main = 'Figure 5.0
    Raw Periodogram')
```

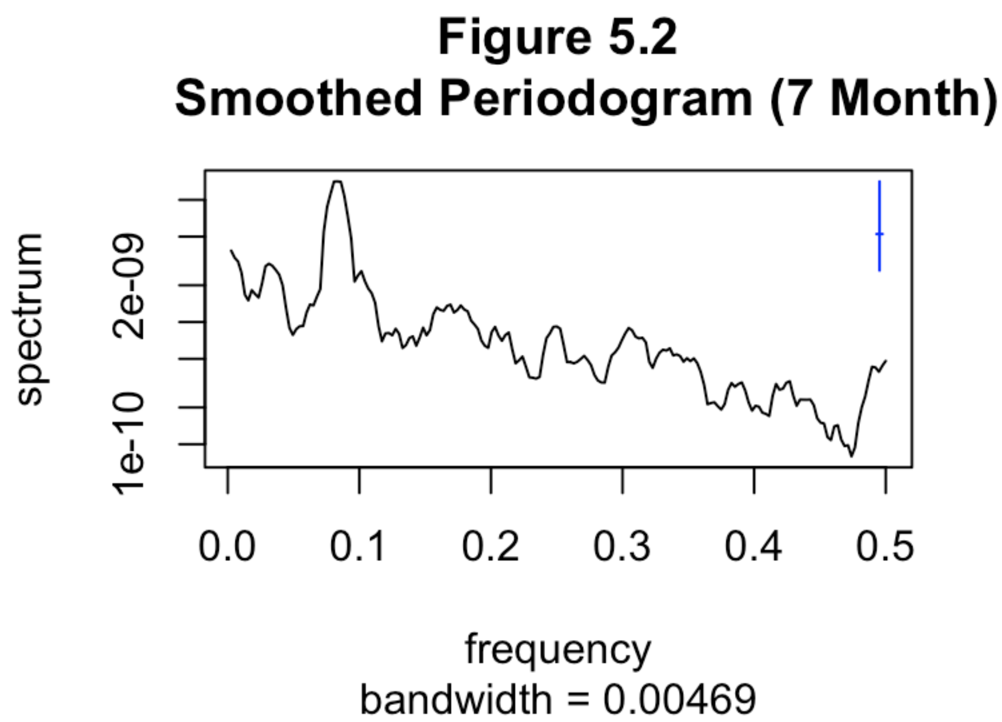
Figure 5.0
Raw Periodogram



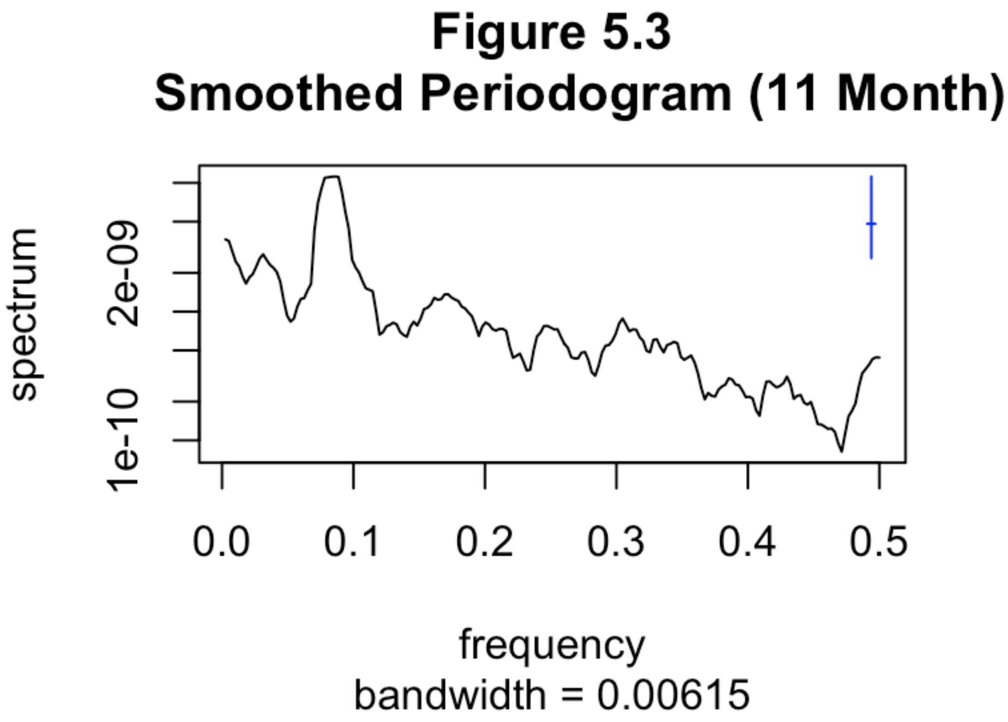
```
xpgrm5 = spec.pgram(x,spans=5, main = "Figure 5.1
    Smoothed Periodogram (5 Month)")
```



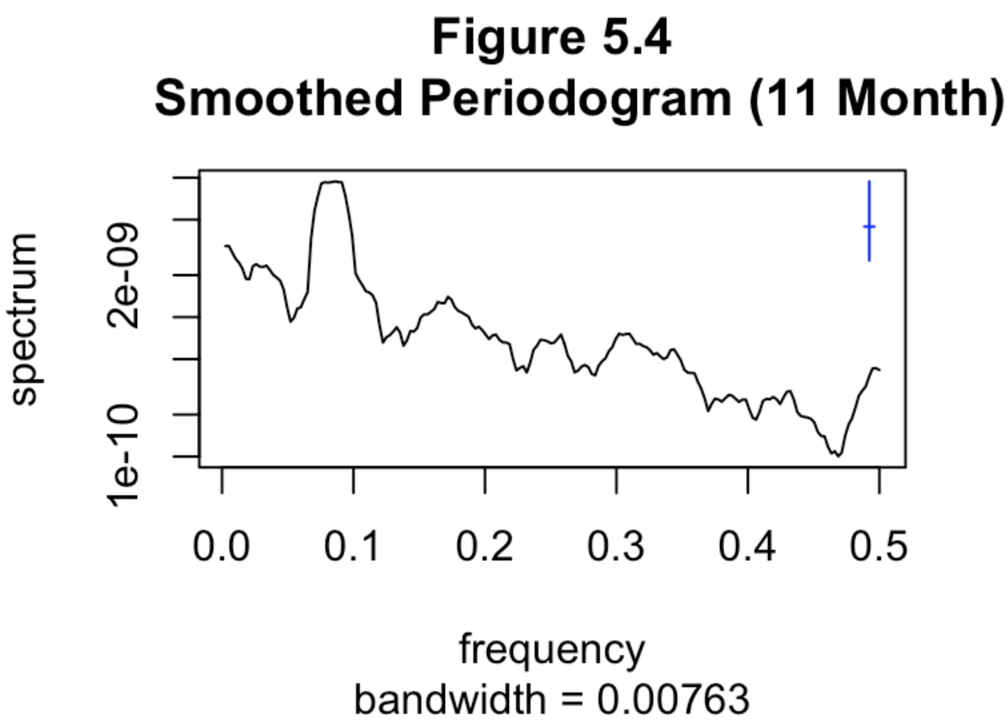

```
xpgrm7 = spec.pgram(x,spans=7, main="Figure 5.2
    Smoothed Periodogram (7 Month)")
```




```
xpgrm9 = spec.pgram(x,spans=9, main="Figure 5.3
    Smoothed Periodogram (11 Month)")
```




```
xpgrm11 = spec.pgram(x,spans=11, main="Figure 5.4
    Smoothed Periodogram (11 Month)")
```



Preliminary ARMA(p, q) or ARIMA(p, d, q) Model

```
fitARMA0 = arima(x, order=c(0,0,0))
fitARMA14 = arima(x, order=c(1,0,4))
fitARMA24 = arima(x, order=c(2,0,4))
fitARMA34 = arima(x, order=c(3,0,4))
fitARMA44 = arima(x, order=c(4,0,4))
fitARMA54 = arima(x, order=c(5,0,4))
fitARMA45 = arima(x, order=c(4,0,5))
aicc(fitARMA0)
```

```
## [1] -6727.784
```

```
aicc(fitARMA14)
```

```
## [1] -6914.465
```

```
aicc(fitARMA24)
```

```
## [1] -6916.829
```

```
aicc(fitARMA34)
```

```
## [1] -6975.005
```

```
aicc(fitARMA44)
```

```
## [1] -6972.908
```

```
aicc(fitARMA54)
```

```
## [1] -6975.393
```

```
aicc(fitARMA45)
```

```
## [1] -6967.331
```

```
auto = auto.arima(x, max.p = 8, max.q =8, max.d = 2); auto
```

```
## Warning in auto.arima(x, max.p = 8, max.q = 8, max.d = 2): Unable to fit
## final model using maximum likelihood. AIC value approximated
```

```
## Series: x
## ARIMA(4,0,5) with zero mean
##
## Coefficients:
##          ar1      ar2     ar3      ar4      ma1     ma2      ma3     ma4
##       1.8990  -1.9961  1.3589  -0.6564  -1.3891  1.1768  -0.5111  0.0139
## s.e.  0.1223   0.2237  0.1910   0.0862   0.1211  0.1852   0.1442  0.0831
##          ma5
##       0.3097
## s.e.  0.0579
##
## sigma^2 estimated as 6.914e-10:  log likelihood=3504.84
## AIC=-6989.19   AICc=-6988.6   BIC=-6949.69
```

```
fitARIMA405 = arima(x, order=c(4, 0, 5))
aicc(fitARIMA405)
```

```
## [1] -6967.331
```

```
par(mfrow=c(1,1))
res = fitARIMA405$res
ts.plot(res, main = "Figure 6.0
  Plot of Residual of ARIMA(4,0,5)")
h=12
n = 372
fcast = predict(fitARIMA405,n.ahead=h)
fc = fcast$pred
upper = fc+qnorm(0.975)*fcast$se
lower = fc-qnorm(0.975)*fcast$se
polygon(x=c(n+1:h,n+h:1),y=c(upper,rev(lower)),col='lightblue',border=NA)
lines(x=n+(1:h),y=fc,col='blue')
```

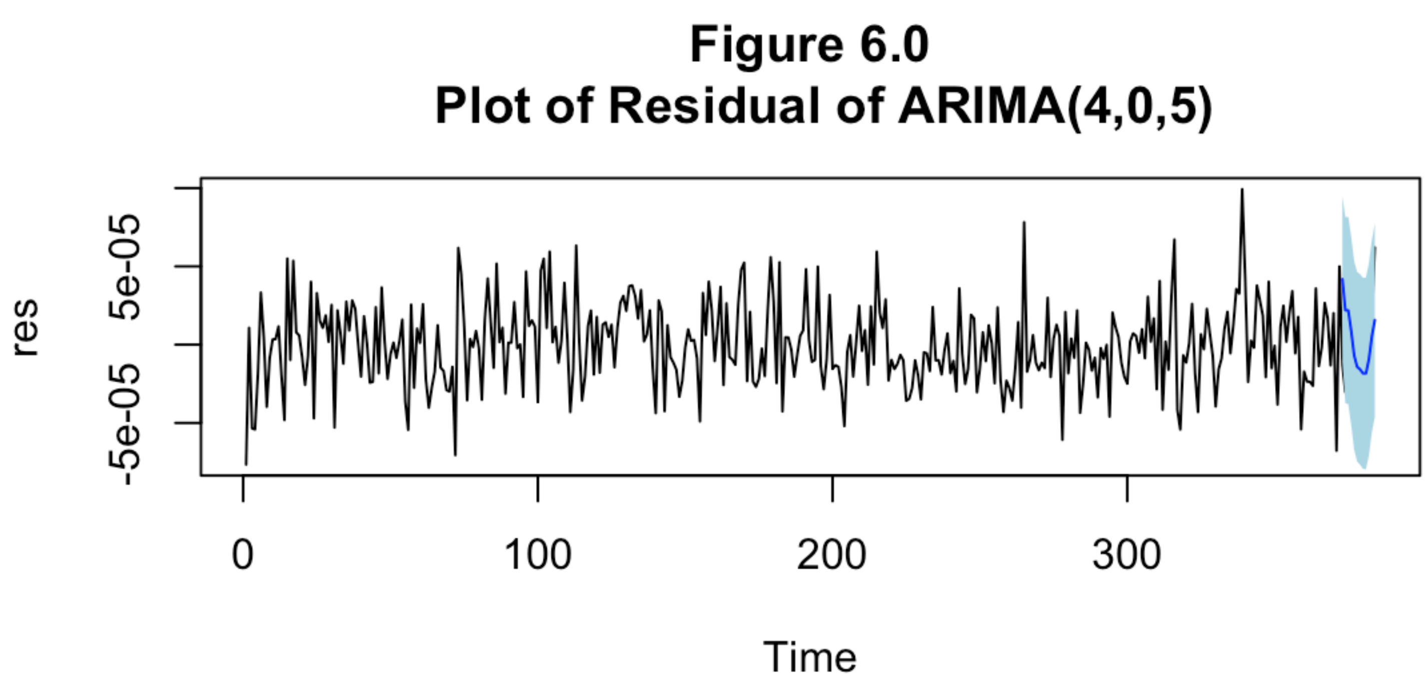


Figure 7.0

```
par(mfrow=c(1,2))
acf(res, main = "ACF of ARIMA(4,0,5) Residual")
pacf(res, main = "PACF of ARIMA(4,0,5) Residual")
```

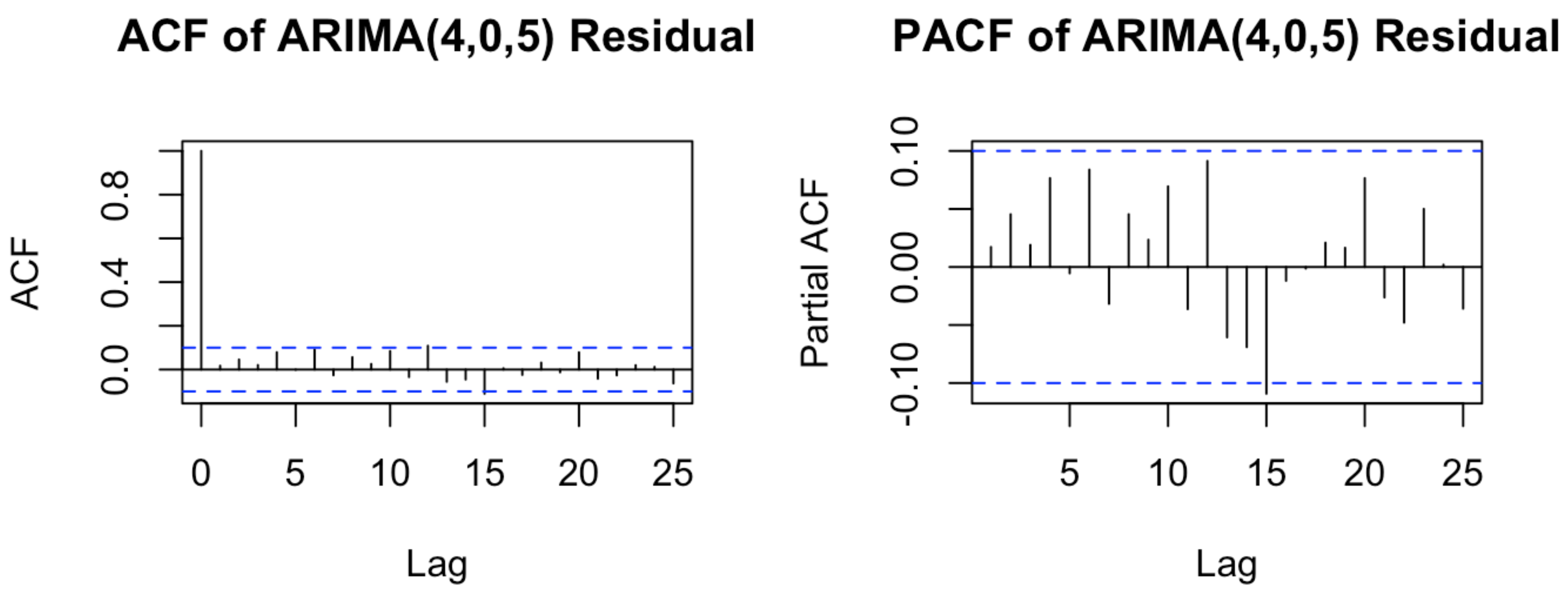
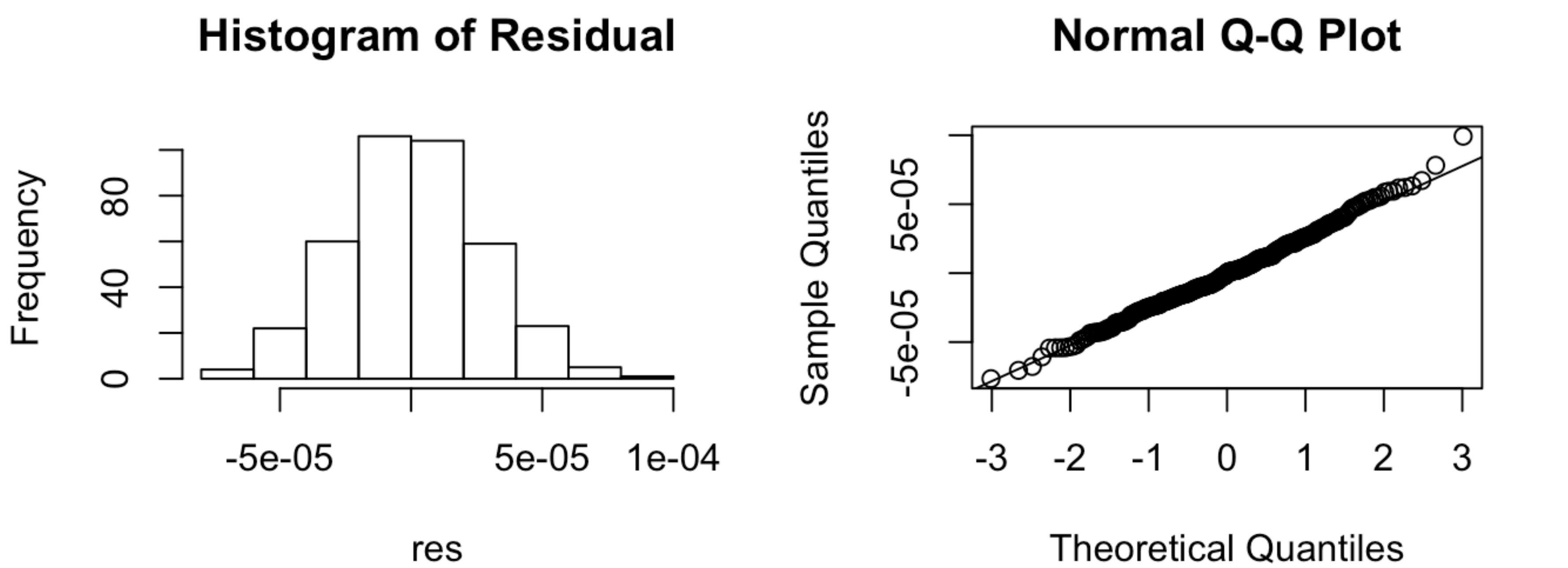
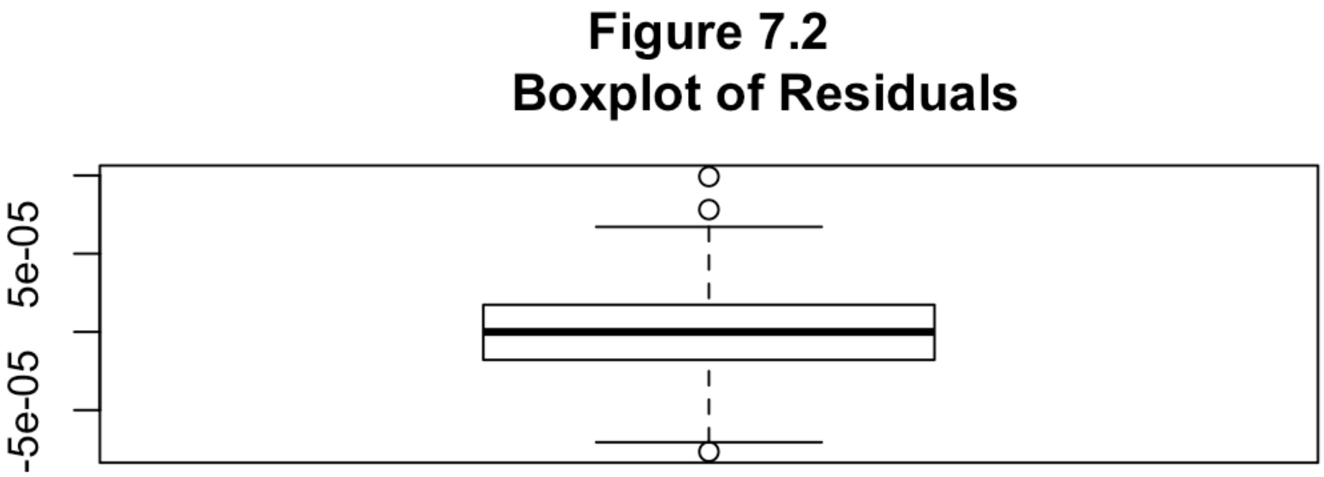


Figure 7.1

```
par(mfrow=c(1,2))
hist(res, main = "Histogram of Residual")
qqnorm(res); qqline(res)
```




```
par(mfrow=c(1,1))
boxplot(res, main = "Figure 7.2
        Boxplot of Residuals")
```



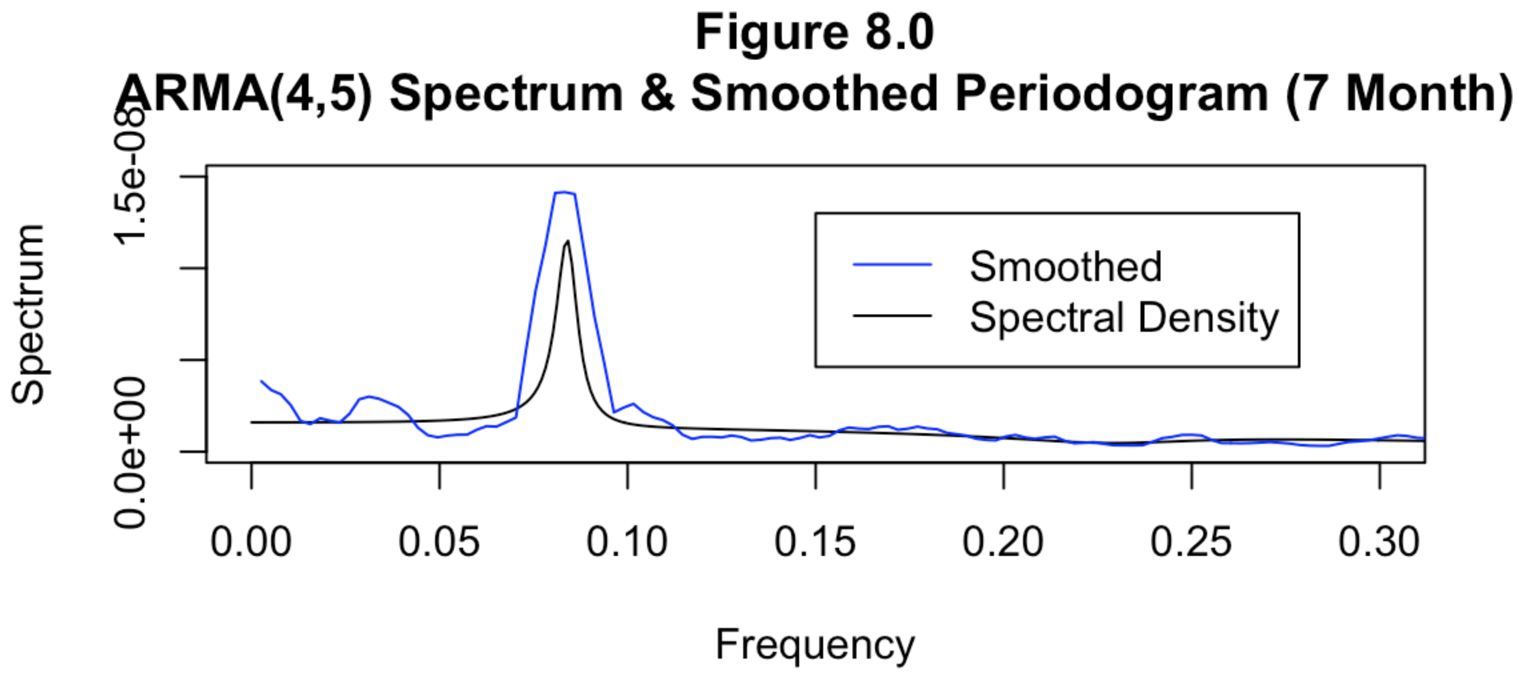

```
Box.test(res,lag=10,'Ljung-Box')
```

```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 11.3325, df = 10, p-value = 0.3322
```

Spectral Density

```
coef.ar = auto$coef[1:4]
coef.ma = auto$coef[5:9]
sigma2 = auto$sigma2
```

```
mod1spec = arma.spec(ar=coef.ar, ma=coef.ma, var.noise=sigma2, log='no', main="Figure 8.0
ARMA(4,5) Spectrum & Smoothed Periodogram (7 Month)", xlim=c(0.0,0.3), ylab="Spectrum
", xlab="Frequency", ylim=c(0,15.0e-09))
points(xpgrm7$freq, xpgrm7$spec, type='l', col='blue')
legend(0.15, 13.0e-09, c("Smoothed", "Spectral Density"), lty=c(1,1), col=c('blue', '
black'))
```



Prediction

```
# Prediction of time 372 to 384
y372 = y[1:372]
n = 372
h=12
# Forecast the trend
deg = 2
coef = mod$coef[1:(deg+1)]
time = (n+(1:h))/n; time
```

```
##  [1] 1.002688 1.005376 1.008065 1.010753 1.013441 1.016129 1.018817
##  [8] 1.021505 1.024194 1.026882 1.029570 1.032258
```

```
predmat = matrix(rep(time,deg)^rep(1:deg,each=h),nrow=h,byrow=F)
predmat = cbind(rep(1,h),predmat); predmat
```

```
##       [,1]     [,2]     [,3]
##  [1,]    1 1.002688 1.005384
##  [2,]    1 1.005376 1.010782
##  [3,]    1 1.008065 1.016194
##  [4,]    1 1.010753 1.021621
##  [5,]    1 1.013441 1.027062
##  [6,]    1 1.016129 1.032518
##  [7,]    1 1.018817 1.037988
##  [8,]    1 1.021505 1.043473
##  [9,]    1 1.024194 1.048972
## [10,]    1 1.026882 1.054486
## [11,]    1 1.029570 1.060014
## [12,]    1 1.032258 1.065557
```

```r
m.fc = predmat %*% coef
# Forecast the seasonality
s.fc = rep(mod$season,length.out=n+h)
s.fc = s.fc[-(1:n)]
# Forecast the rough part
fcast = predict(fitARIMA405,n.ahead=h)
x.fc = fcast$pred
# Combine forecasts
y.fc = ( m.fc + s.fc + x.fc)^(-1)
y.fc
```

```
## Time Series:
## Start = 385
## End = 396
## Frequency = 1
##           [,1]
##  [1,] 2399.258
##  [2,] 2123.120
##  [3,] 1903.211
##  [4,] 1483.402
##  [5,] 1367.182
##  [6,] 1485.714
##  [7,] 1742.031
##  [8,] 1728.739
##  [9,] 1446.591
## [10,] 1369.519
## [11,] 1575.130
## [12,] 2162.323
```

```r
par(mfrow=c(1,1))
oldy=y[1:372]
plot.ts(oldy,xlim=c(0,n+h), main = "Figure 9.0
Petroleum Prediction from December 2014", ylab= "Trillion BTU", pch=1)
points(x=n+1:h, y=y.fc, col='blue',type='l',pch=8)
legend(0, 2800, c("Forecasted 2015"), lty=c(1,1), col=c('blue'))
```

Figure 9.0
Petroleum Prediction from December 2014