An Analysis of US Employment/Population

Rafael Amaral Porsani

1 Introductory Comments

The last US recession ended in the second-quarter of 2009, and the unemployment rate has since fallen considerably (see Figure 1). Indeed, unemployment is close to the lowest level it's been in the last 40 years, suggesting at first sight the predominance of very healthy labor-market conditions in the country. One could reasonably conjecture, however, that such precipitous drop may potentially be attributed to the simple fact that, more and more, individuals are dropping out of the labor-force. In truth, the unemployment rate - according to the Bureau of Labor Statistics (BLS)[1] - is measured as the estimated number of unemployed individuals (people who are simultaneously jobless, looking for a job, and available for work) divided by the estimated labor force (which is comprised of unemployed plus employed individuals[2]). A decrease in the number of individuals looking for a job leads to a decrease in both the number of unemployed individuals (the numerator of the unemployment rate), and to a similar decrease in the labor force (the denominator of such rate), ultimately producing a decline in the rate itself[3].

Arguably, therefore, a more indicative measure of the overall health of the labor market is the employment-to-population ratio (illustrated in Figure 2). Here, we see that US workers have been, relatively speaking, having a hard time finding jobs - employment is at a much lower level than it was in the late 1990's. The 4% decline seen over the last recession was massive vis-a-vis declines registered in previous recessions, and recovery has been somewhat sluggish - i.e., if compared to the growth in employment seen in the 1980's. Figure 3 highlights in different colors what arguably could be considered different regimes in the employment-to-population series. This figure suggests that such

[1] See https://www.bls.gov/cps/cps_htgm.htm.

[2] In the section entitle "Brief Description of Data", below, we provide a more comprehensive summary of which individuals are considered to be employed by the BLS, and on how employment-related data is collected.

[3] Note that this follows from a basic property of fractions. If a, b and x are non-zero positive integers, with $a < b$, then $\frac{a-x}{b-x} < \frac{a}{b}$.

series grew at a fast pace from about April 1975 to November 2000. In Figure 4, we show linear trends fit to the time periods corresponding to these potential regimes. It should be clear by assessing the slopes of these lines that employment did indeed grow at a stronger pace in the period that goes from 1975 to 2000.

In this study, we model and interpret the time series of employment-to-population from April 1975 to November 2000 (we leave the period that goes from December 2000 to November 2001 as a test set, which we later use to evaluate our forecasts). As we have argued above, this is a period when employment exhibited strong gains. Hopefully, by understanding it better, one could possibly develop recipes/policies for improving the current labor environment in the US. In our view, modeling it is a first step towards gaining such understanding[4].

2 Brief Description of Data

Our data comes from the Federal Reserve Bank of Saint Louis (FRED)[5]. FRED provides, free-of-charge, about 469,000 US and International time series, from numerous sources, in its website. We have actually downloaded our time series from FRED directly into R using the function "getSymbols" from the "quantmod" library. The data we use is monthly, and the observations originally come from the Current Population Survey (CPS) conducted early each month by the BLS. This survey, which measures employment and unemployment in the country, has been conducted in the US every month since 1940 (though the data that is available on FRED starts in 1948). The BLS reports that there are roughly 60,0000 eligible households in the sample for this survey. The sample is selected so as to be representative of the US population. Every month, government employees contact these 60,000 households and ask individuals living in them (aged 16 and above) questions about their labor activities. Individuals are considered employed if they: (i) "did any work for pay or profit during the week when the survey is conducted"; or (ii) have a job but couldn't work because of specific circumstances[6]. Note that part-time workers, according to this definition, are considered to be employed. And also note that

[4]Side note: in Figure 5, we show the estimated total number of employees in the Manufacturing sector. This is included more as a curiosity here. Manufacturing jobs took a hard hit in the last recession, falling by about 30% then. Interestingly, the Trump campaign was apparently successful in identifying this movement - in the sense that they won the electoral vote in a number of states in which the discussion Manufacturing jobs seemed to have been critical.

[5]https://fred.stlouisfed.org/.

[6]In particular, if they have a job but were on vacation; ill; experiencing child care problems; on maternity or paternity leave; prevented from working due to bad weather; involved in a labor dispute; or taking care of personal of family obligation.

if a person has more than one job, he or she will simply show up in the survey as being employed - i.e., s/he won't be "counted twice". The employment-to-population ratio is then defined as the number of employed individuals, as a percentage of the total population (in the surveyed households).

3 Analysis

To model our data, the following steps will be taken: **(i)** we will detrend the data; **(ii)** we'll use spectral analysis to identify cycles/seasonal patterns in the detrended data, and then remove these cycles from it; **(iii)** the residuals (*original data* $-$ *trend* $-$ *seasonal components*) will then be analyzed and modeled as an ARMA process; **(iv)** finally, we'll evaluate how well we have modeled our time series by evaluating the residuals engendered by our approach, and by assessing the quality of our forecasts.

Before implementing these steps, we first assess the autocorrelation (ACF) and partial-autocorrelation functions (PACF) of the original data. On the top portion of Figure 6, we show the ACF and PACF of the original data when all possible lags are used; on the bottom of Figure 6, the ACF and PACF are depicted for a total of 50 lags. We see that the PACF hints towards the existence of cycles in the data; and so does the ACF plot which includes all possible lags. As we'll show below, the cyclical nature of our data will become more apparent once the series has been detrended.

3.1 Detrending the Data

Firstly, we tried fitting a linear regression model to the data. The fitted model takes on the following form: $\hat{Y}_t = -520 + 0.29t$, where $\hat{Y}_t$ denotes the fitted employment-to-population ratio at time t. The $t-$ *statistic* for our slope coefficient is 35.03, indicating that the trend in this period is statistically significant (even at the 1% level of significance). This fitted linear trend suggests that employment-to-population tended to increase by 0.29% per year from April 1975 to November 2000 - a rate which we have argued is higher than those evidenced in neighboring time-periods.

Figure 7 shows the detrended data, when such linear trend is used. Interestingly, this figure suggests that the resulting detrended series is not stationary: the variance on the left-hand site of our plot seems markedly higher than that on the right-hand side. To confirm our visual intuition, we conduct an augmented Dickey-Fuller (ADF) test on the detrended series[7]. The p-value from our test is 0.36 - therefore, we fail to reject the null hypotheses that there exists a unit root in this data. This in turn

[7] The adf.test function in R selected a lag order of 6 for this test.

suggests that the detrended series is indeed non-stationary.

One could attempt to get around this by taking first-differences of the original data; but our main interest here is in the employment-to-population ratio itself, not in first-differences of this quantity. Thus, we will refrain from taking first-differences of our time-series. Instead, we opt to fit a non-linear trend to the data, using a smoothing-spline regression (i.e., a spline regression where the knots are tied to the observations themselves).

We have written a function in R that fits a non-linear trend to our data through smoothing-splines. Note that we could also have utilized the lowess or loess functions in R to fit a non-linear trend here, as well as other methods. One of the advantages of using our own code to fit a non-linear trend is that, by doing so, we know exactly how the trend is being fit - and this allows us to more easily forecast the trend component of our series.

The fitted trend is of the form $f(x) = \alpha_0 + \sum_{j=1}^{p} \alpha_j max(0, x - k_j)$, where x here is equivalent to$t$ (time), and the $k_j's$ are the times associated with the observations in the employment-to-population series[8]. The fitted $\alpha's$ are found by minimizing $\sum_{t=1}^{n} ||y_t - \alpha_0 - \sum_{j=1}^{p} \alpha_j max(0, x_t - k_j)||^2 + \lambda \sum_{j=1}^{p} \alpha_j^2$, with y_t denoting the employment-to-population ratio at time t, x_t is again simply the time t associated with observation y_t; λ is a tuning parameter that penalizes the complexity of our model[9]. In the appendix, we provide a brief derivation of the spline solution.

Figure 8 shows our time series and the non-linear trend. The tuning parameter λ was selected so as to yield a detrended time series that looked stationary, while producing a trend component that was reasonably smooth[10]. The detrended data is shown in Figure 9. We can see that after removing the non-linear trend, we obtain a series that seems stationary. An ADF test was conducted on this time series, producing a very low $p - value$ ($p < 0.01$) [11]- the null hypothesis that there exists a unit root in this series can now be rejected. This is the detrended series that we use in our analysis.

Some comments are in order. Firstly, one could speculate that the overall positive trend seen in this period is associated with an increased integration between the US economy and other economies (globalization), which may have led to increasing exports and the creation of local job openings. It is also possible that an increasing participation of women in the labor-force contributed to the steep increase in employment we saw from 1975 to 2000. Lastly, automation and product innovations may

[8] Note that $p = 308$, the total number of observations.

[9] Also, $n = 308$ and $p = 308$; and k_j denotes, as we mentioned, the time associated with observation j.

[10] We tried different values for λ, ultimately choosing $\lambda = 100$.

[11] Again a lag order of 6 was used in the test.

destroy jobs in some industries, but can also create jobs in others (and note that these can also make companies more efficient, potentially leading to greater shareholder wealth, which can also spur consumption and jobs) - thus, automation by itself (which one could proxy through labor productivity) may also have helped create jobs in this period. We believe these should be subject to scrutiny in further studies.

3.2 Identifying and Removing Cycles (Spectral Analysis)

Figure 10 depicts the ACF and PACF of the detrended data - we can clearly see, from this figure, that the detrended series is influenced by cycles. Figure 11 illustrates the raw periodogram of the detrended data. A large portion of the total variation in the data is explained by the 1.013 frequency (which corresponds to a period of roughly one year). There are also peaks at $\omega = 1.988 \approx 2$ and $\omega = 3$, harmonics of $\omega = 1$. This indicates that the one-year cycle present in the series is not a perfect sinusoid. Also note that there is considerable power at lower frequencies. In particular, we see a small peak at the frequency of 0.118 (which corresponds to a period of 5.33 years, or 64 months). One could associate this peak with long-term business-cycle fluctuations not captured by our non-linear trend.

In Figure 12, we show in detail the yearly seasonal component, which, as we have argued, is responsible for a good portion of the total variation in the detrended series.The graph suggests that there is a strong employment season from March to May, and a weak one from October to January. Generally speaking, April seems to be the best month to find a job, while October is apparently the worst. The yearly seasonal component declines, in general terms, after April, and starts to increase after October. July seems to actually be a stronger month for employment-population than June, August and September. Also note that the annual seasonal component is negative from September to February, staying positive in other months. Summer vacations of hiring managers may be a possible reason for the slowdown in Summer months; a desire by some CFO's to boost calendar-year-end bottom lines[12], coupled with the holiday season, might help to explain the low year-end figures.

We have also computed approximate approximate confidence intervals for the spectral densities associated with the one-year cycle, the cycles corresponding to $\omega = 2$ and $\omega = 3$ (harmonics), and the 5.33 year cycle. The lower values of these intervals are all higher than most of the other periodogram ordinates, indicating that these cycles are all significant.

[12]Note that the calendar year is used as the fiscal year by most publicly-traded US firms.

Figure 13 contains the smoothed periodogram of our detrended series - obtained using a modified Daniell (2,2) kernel. This periodogram is smoother than the raw one, as expected. The peaks occur at exactly the same frequencies we identified previously ($\omega = 0.118, 1.013, 1.988$ and 3). The smoothed periodogram is specially helpful to confirm the low-frequency peak at $\omega = 0.118$ - given in the raw periodogram we saw reasonable power around this frequency.

Moreover, we have utilized an AR spectral estimator to further substantiate our findings. Figure 14 demonstrates that the optimal number of lags to use with such estimator, when using AIC as a criteria, is 39 [13]. Figure 15 shows that this procedure produces peaks at nearly the same frequencies we had identified previously. There are clear peaks at $\omega \approx 1, 2$ and 3, as before. The low frequency peak is now located at $\omega = 0.192$, indicating a cycle with period of roughly 5.2 years - which is very close to the 5.33 years period that was uncovered with the raw and smoothed periodograms.

The ACF of the detrended data, after removal of the one-year cycle, is provided in Figure 16. We can see from this figure that the series still exhibits a cyclical pattern, as expected, when only the one-year cycle is removed from it. Figure 16 also illustrates the ACF of the detrended data after the removal of both the one-year and the 5.33 years (64 months) cycles. Even when influences from these two cycles are removed, we still see that the resulting series possesses a cyclical behavior. Thus, in the analysis that follows, we have opted to remove the influences of all of the four cycles that are evident in our periodograms, and which have found to be significant: the one-year cycle, the 5.33 years one, and the cycles associated with the second and third harmonics of $\omega = 1$ (i.e., $\omega = 2$ and $\omega = 3$; note that removing these helps us to account for the non-sinusoidal behavior of the yearly component).

3.3 Modeling Residuals with ARMA models, Diagnostics and Forecasting

Next, we turn our attention to our residual data (the series obtained after the trend and all seasonal components have been removed). Figure 17 shows the ACF and PACF of the residual data. On the top portion of this figure, we see that the ACF seems to tail off (albeit it alternates between positive and negative territory), while the PACF apparently cuts off after a certain number of lags. Closer inspection (see bottom portion of Figure 17) of the PACF suggests that the estimated partial-autocorrelation function cuts off at about lag 14 or 15. When combined, these findings suggest fitting an AR(14) or AR(15) model to the data. We confirm our visual intuition by assessing what would be

[13] Note that the optimal number of lags, when AIC is used, is not that clear here. There are many lags with AIC values very close to the one obtained with p = 39. We have used other values for p in our analysis, obtaining similar results.

the best AR model, from an AIC standpoint, to fit to the residual data. Figure 18 demonstrates that the lowest AIC (-2.29) is obtained when 14 lags are used, confirming the intuition we developed from evaluating the ACF and PACF of our series.

In Figure 19, model diagnostics are presented for the AR(14) model. For the most part, an AR(14) model does a good job in describing our residual data. An inspection of the normal Q-Q plot and the histogram of residuals produced by this model indicates that model residuals are normally distributed. A Shapiro-Wilk normality test conducted on these yields a p-value of 0.58, further supporting this assertion. The Ljung-Box test, however, suggests there may be serial correlation in the residuals from lags 12 to 16.

As a result, we have tried to include MA terms in our model, fitting different ARMA models to the data. Good results were obtained when an ARMA(14,9) is used (see Figure 20). Residuals from the ARMA(14,9) model look normal[14]; they resemble a white-noise process (see ACF in Figure 20); and Ljung-Box tests provide no evidence of serial correlation in them. Furthermore, the AIC associated with this model was -2.64 - a value that is lower than the one we obtained with the AR(14) model.

Lastly, Figures 21 to 23 show forecasts obtained when an AR(14) model and an ARMA(14,9) are used to model the residual series. Forecasts were computed as the sum of the trend and seasonal components, plus predictions from either the AR(14) or ARMA(14,9) model. For the AR(14) model, we also show one-step-ahead forecasts (Figure 23). Both models seem to perform well in the testing period (December 2000 to November 2001), with most realizations falling withing two-standard errors from our predictions.

4 Concluding Remarks

We speculate that different variables may have contributed to the robust gains in employment evidenced in the period we studied. Firstly, the US economy became more integrated with other economies during this period (globalization). Growing exports may have helped spur job growth over these years. Secondly, it is possible that an increase in the participation of women (and other groups) in the labor force may have also helped boost jobs then. We have also speculated that automation itself (which could be measured through labor productivity, i.e., output per hours worked) may also have contributed to an

[14]The Shapiro-Wilk normality test here yielded a p-value 0f 0.105 - so we fail to reject the null that the data comes from a normal distribution, at a 5% level of significance.

increase in the level of employment-to-population in this period: while automation may destroy jobs, it also has the potential to make companies more efficient, contributing to an increase in shareholder wealth, possibly thus having an impact in household consumption (via a wealth effect - households may "feel richer" and consume more), leading then to more employment. Lastly, this was also a period in which a good number of skilled foreign workers came to the US. These may have helped the economy grow further, imparting their knowledge on local workers, and going on to start-up new companies by themselves. In future studies, it would be interesting to investigate how these forces may have driven employment in this period. This could be done, at first, by for example looking at the cross-spectrum and coherence of employment and other variables, such as exports, labor productivity, and women labor-force participation. Potentially, having an immigration policy that incentivizes skilled workers to join the work force in the US might be a good idea? Maybe creating incentives for women to become entrepreneurs and/or business leaders should be a priority of our government? These questions warrant further scrutiny. Indeed, one of the limitations of our study is that we did not use other time series to explain employment-to-population rates. As a final remark, it is also interesting to see that the detrended series can be largely explained by a one-year seasonal component/cycle. Having more time, we would like to investigate the reasons underscoring the predominance of a weak employment season during the last months of the year (we have speculated as to why this may be the case)[15], and study why hiring in March, April and May is apparently stronger than in other months of the year.

[15]Note that the BLS argues that, in their view, extreme weather conditions do not affect employment by much - though they likely affect the statistic "number of hours worked". In our view, a more likely reason for these low numbers in the last months of the year is financial: CFO's may be looking to improve their end-of-year numbers, thus curtailing hiring in this period.

5 APPENDIX

5.1 Figures

Figure 1: Unemployment Rate in the US - Recessions in Blue. The unemployment rate suggests the economy has almost never been better.

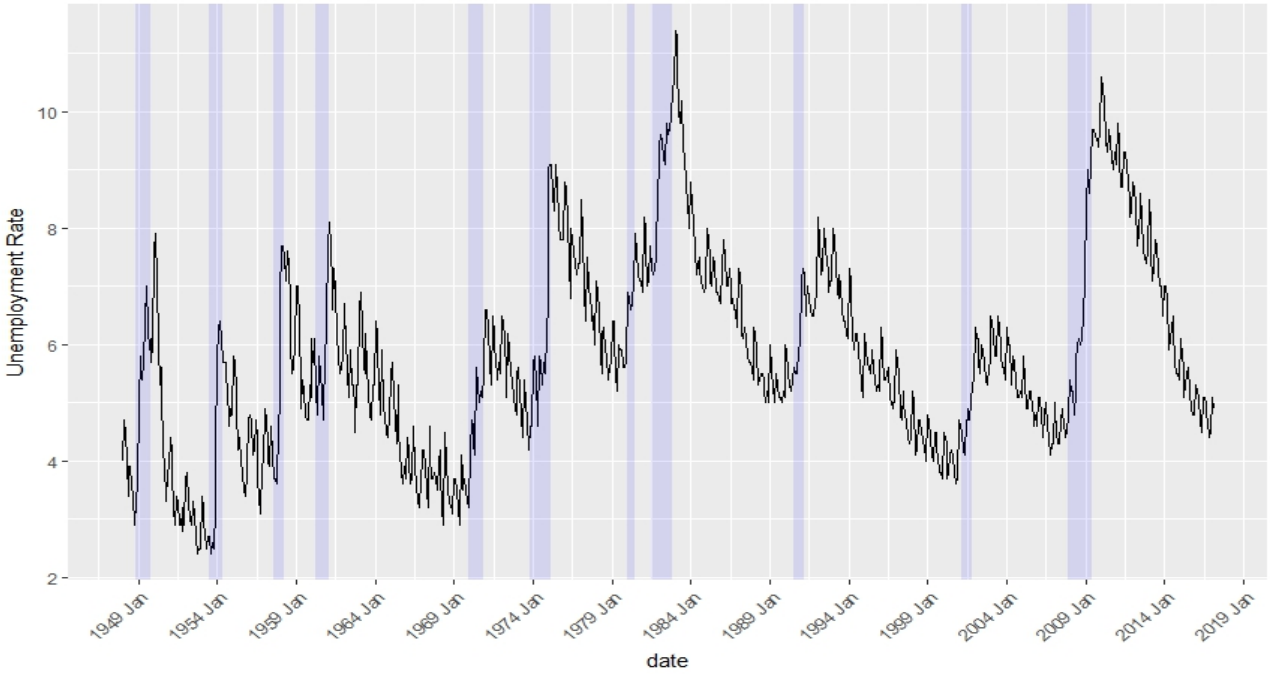


Figure 2: Employment-to-Population Ratio - Recessions in Blue. Employment is at a much lower level than it was in 2001.

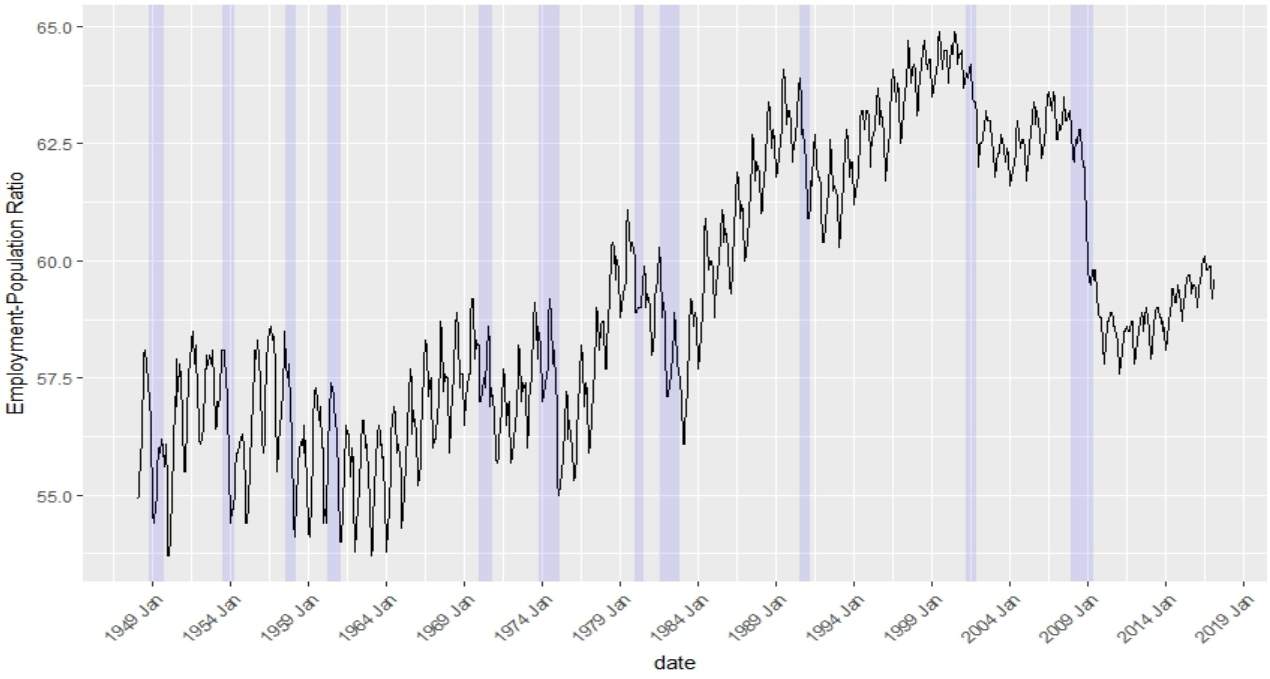


Figure 3: Different Regimes in Employment/Population. We essentially model the data in Green - April 1975 to November 2000 - leaving December 2000 to November 2001 as a test set.

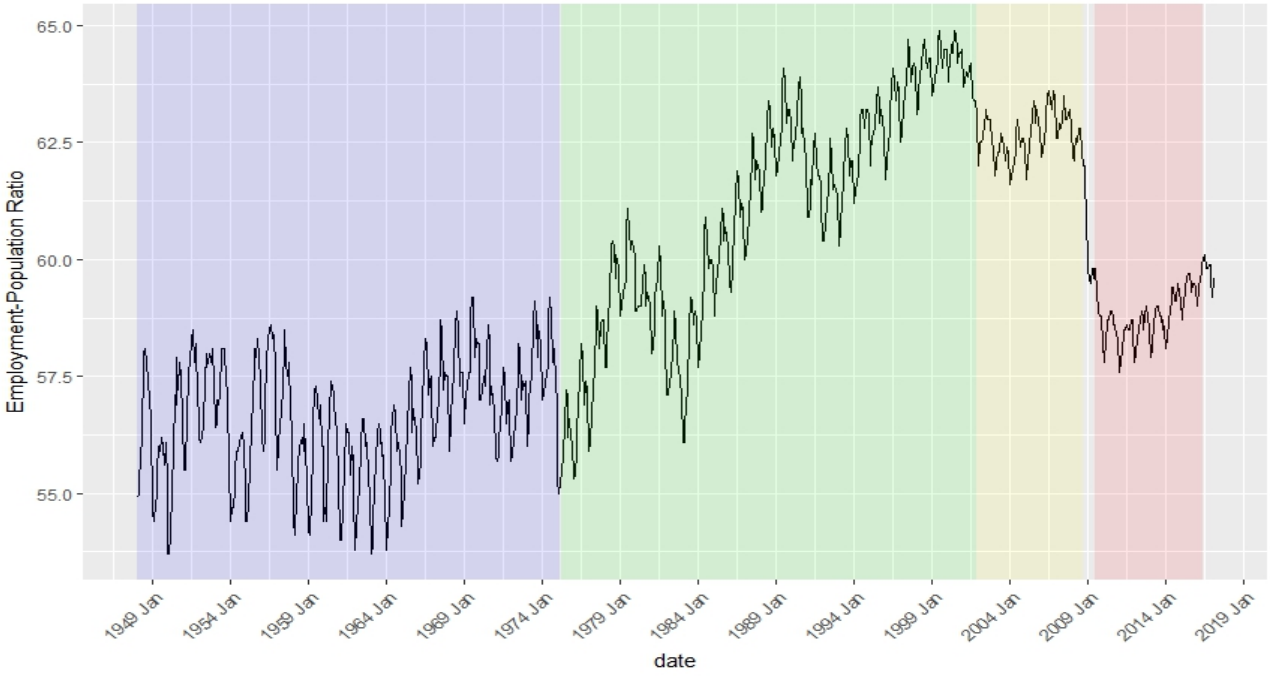


Figure 4: Comparing Trends. There is a steep increase in employment from 1975 to 2000.

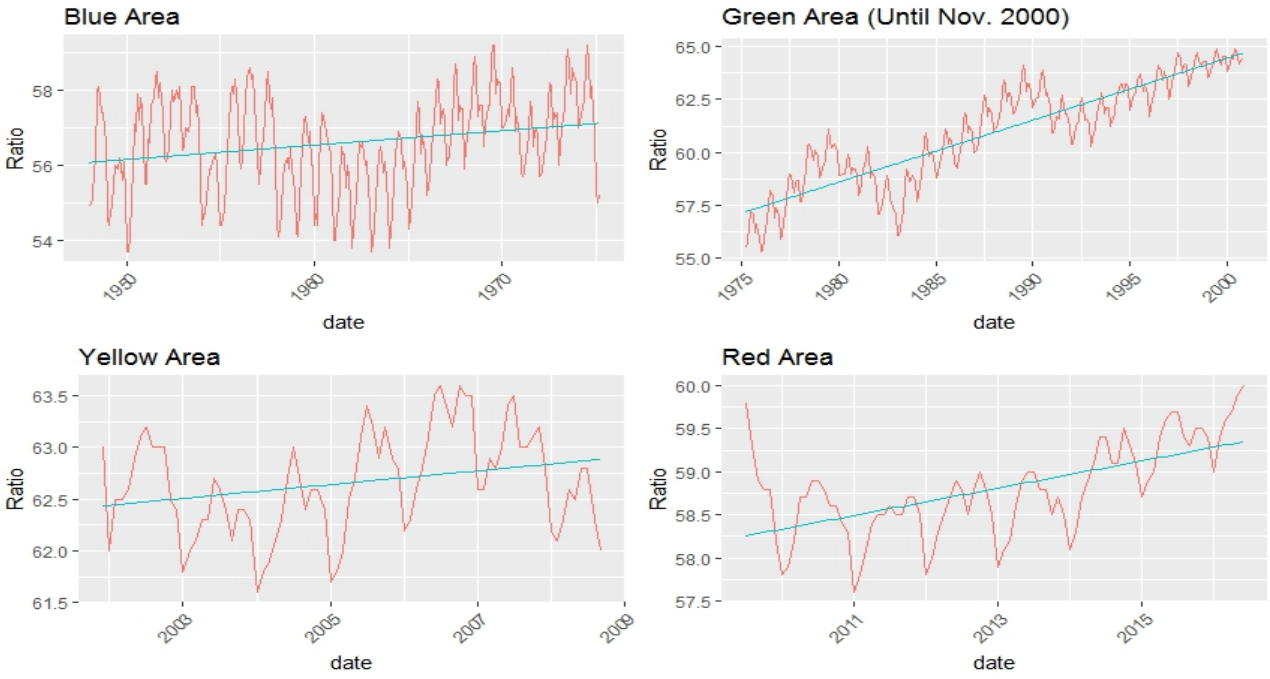


Figure 5: Total Number of Employees in Manufacturing (in thousands)- Recessions in Blue.

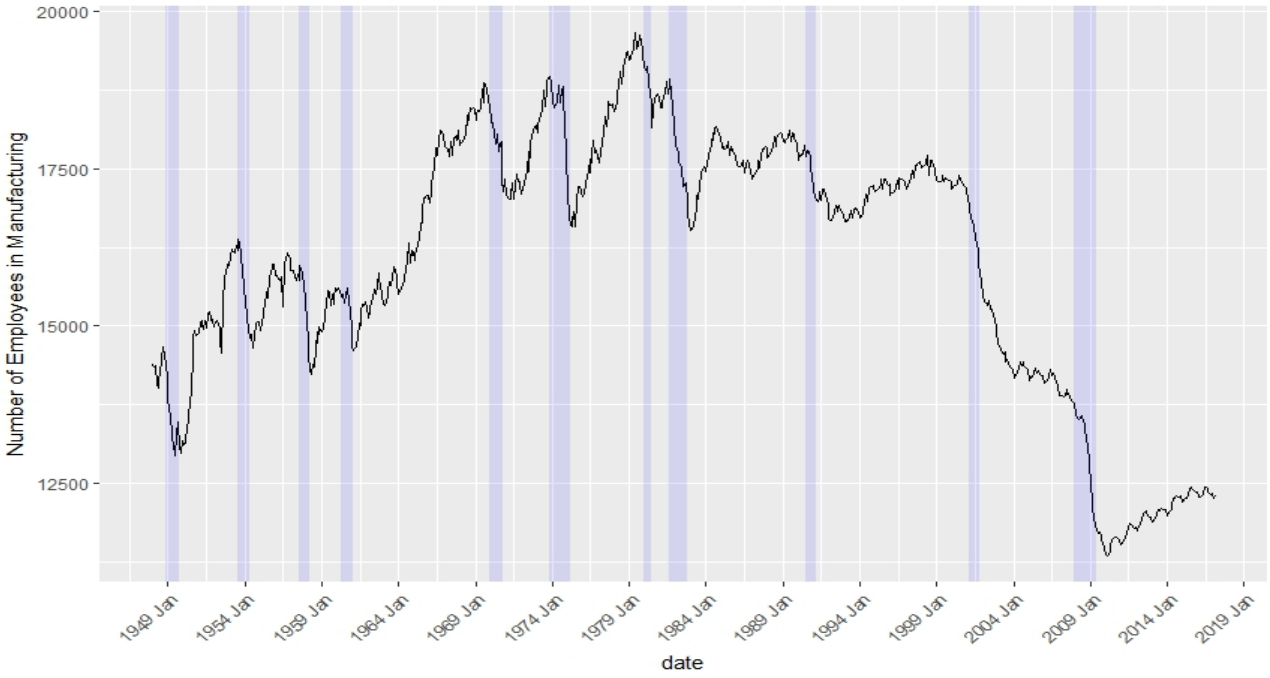


Figure 6: ACF/PACF - Original Data.

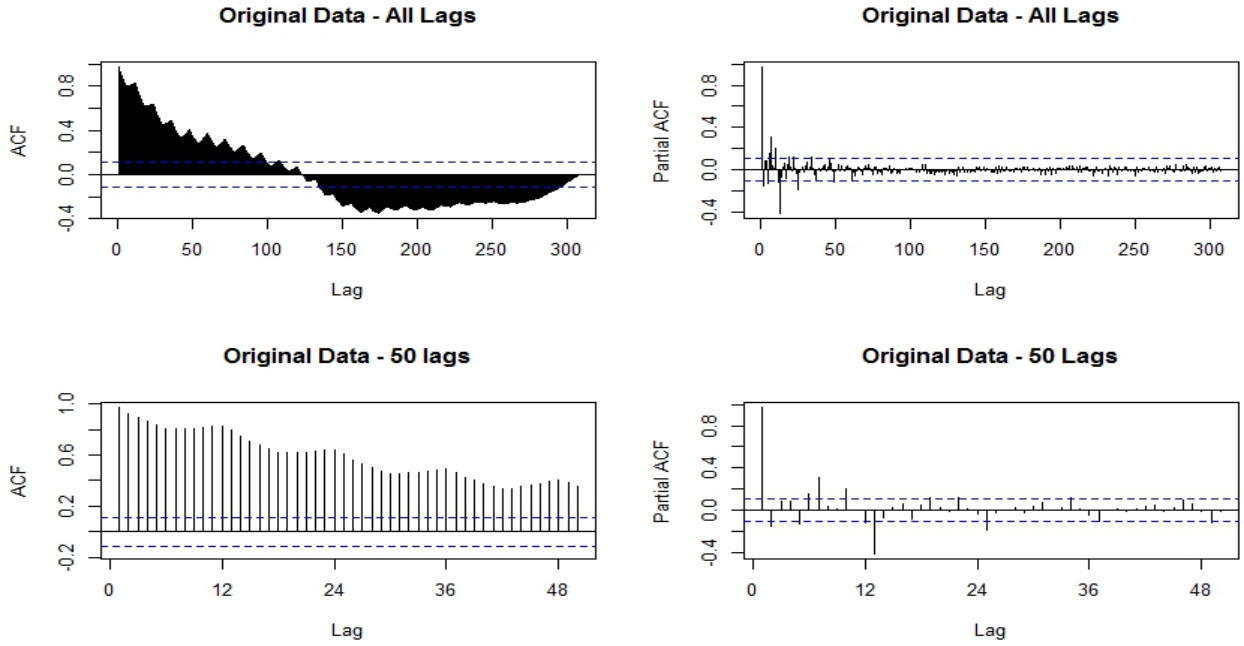


Figure 7: Detrended Data - Linear Trend. This figure suggests that the detrended data (when a linear trend is used) is non-stationary - its variance is not constant. Trend: $Y = -520 + 0.29t$; Augmented Dickey-Fuller Test - p-value = 0.36.

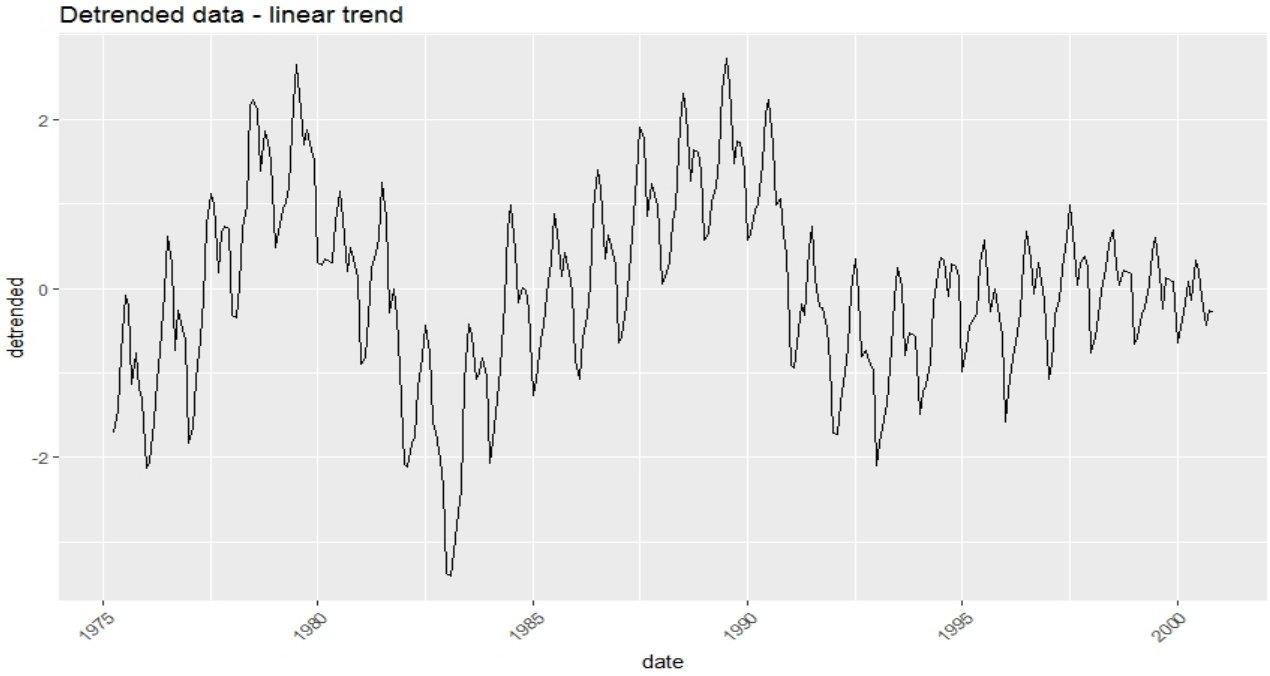


Figure 8: Data + Non-Linear Trend. We estimate a non-linear trend using smoothing splines: $f(x) = \alpha_0 + \sum_{j=1}^{p} \alpha_j max(0, x - k_j)$. The solution is found by minimizing: $\sum_{i=1}^{n} ||y_i - \alpha_0 - \sum_{j=1}^{p} \alpha_j max(0, x_i - k_j)||^2 + \lambda \sum_{j=1}^{p} \alpha_j^2$. Observations themselves are the $k_j's$.

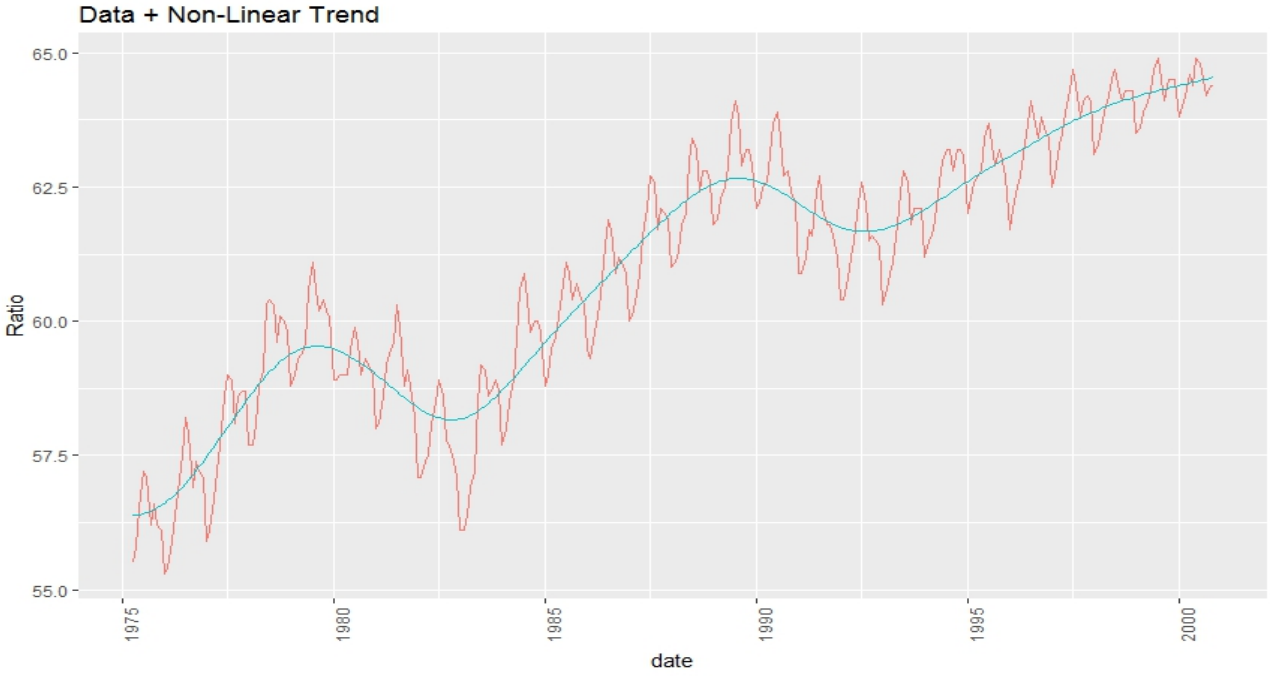


Figure 9: Detrended Data - Non-Linear Trend. The data now looks stationary. Also: Augmented Dickey-Fuller Test - p-value < 0.01.

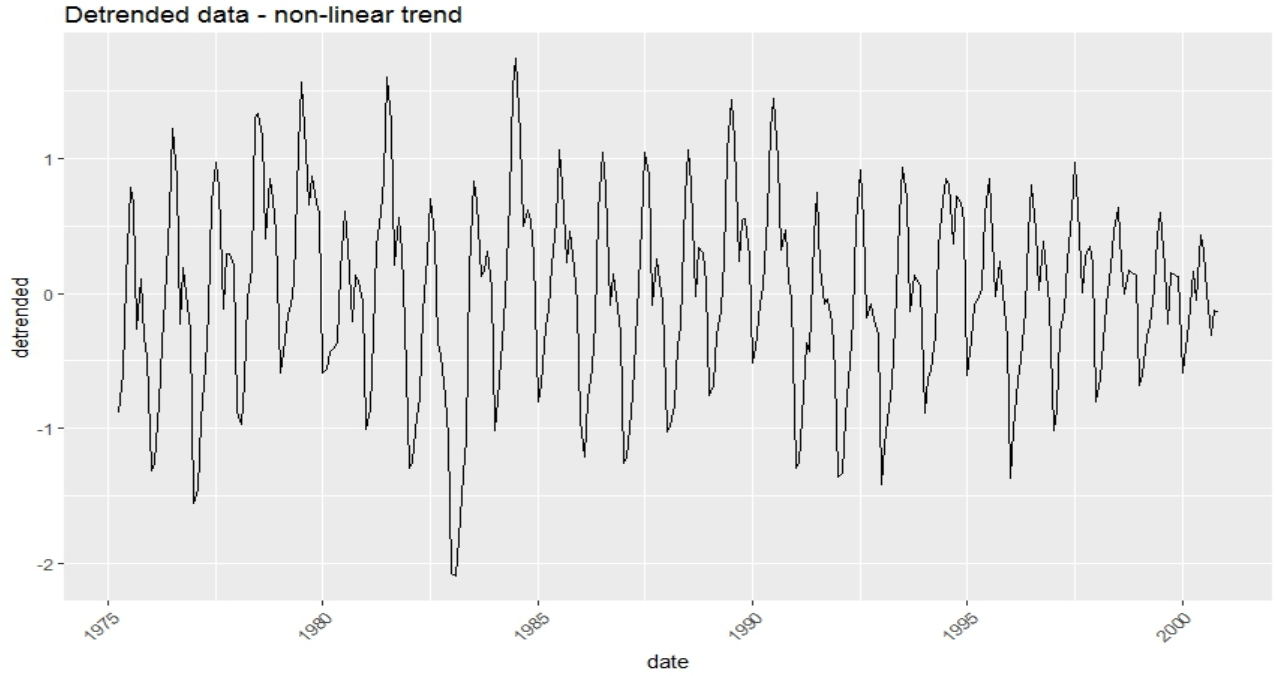


Figure 10: ACF/PACF - Detrended Data.

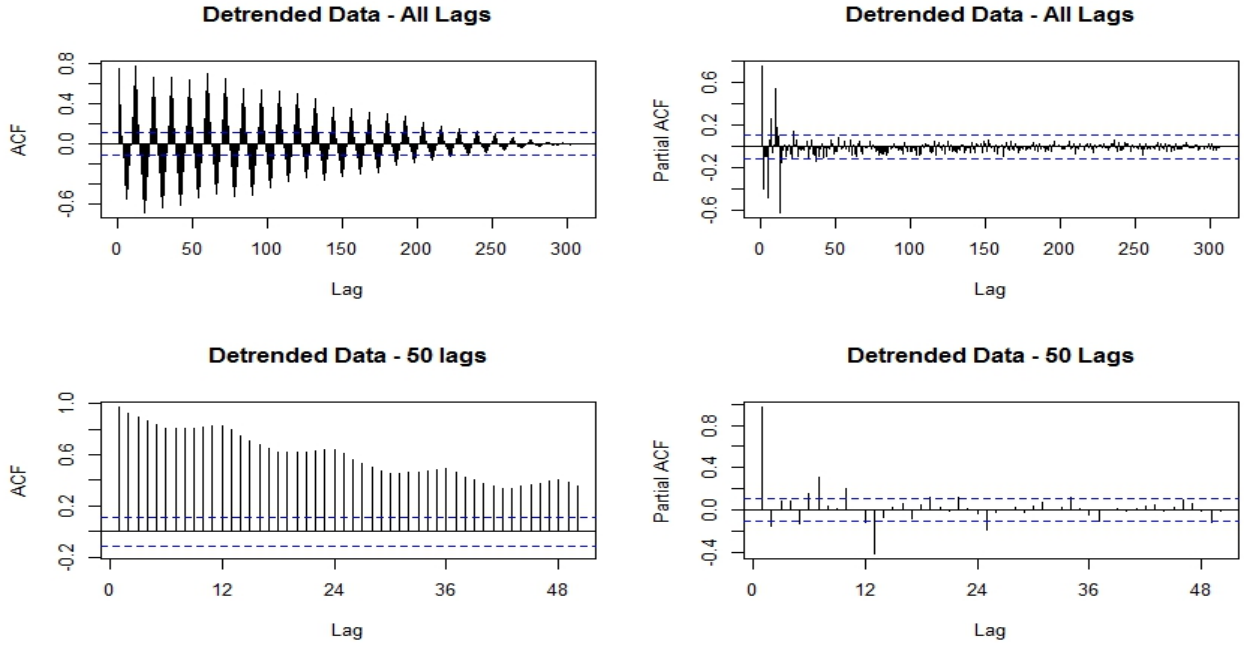


Figure 11: Raw Periodogram. We see a one Year Cycle + 5.3 Year Cycle, and Harmonics of $\omega = 1$. Red dotted line indicates lower bound of the confidence interval for the spectrum corresponding to $\omega = 1$. Peaks are located at $\omega = 0.188, 1.013, 1.988, 3$. These correspond to the following periods (in years), respectively: $5.33, 0.99, 0.50, 0.33$.

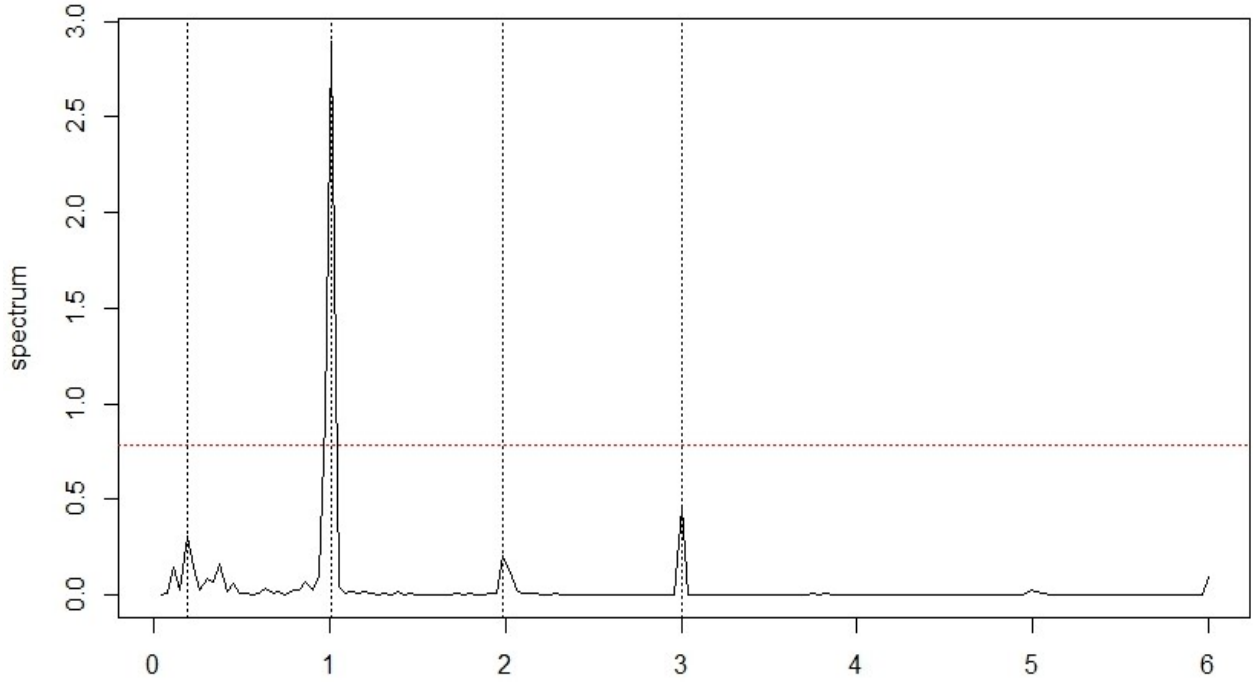


Figure 12: Yearly Seasonal Component.

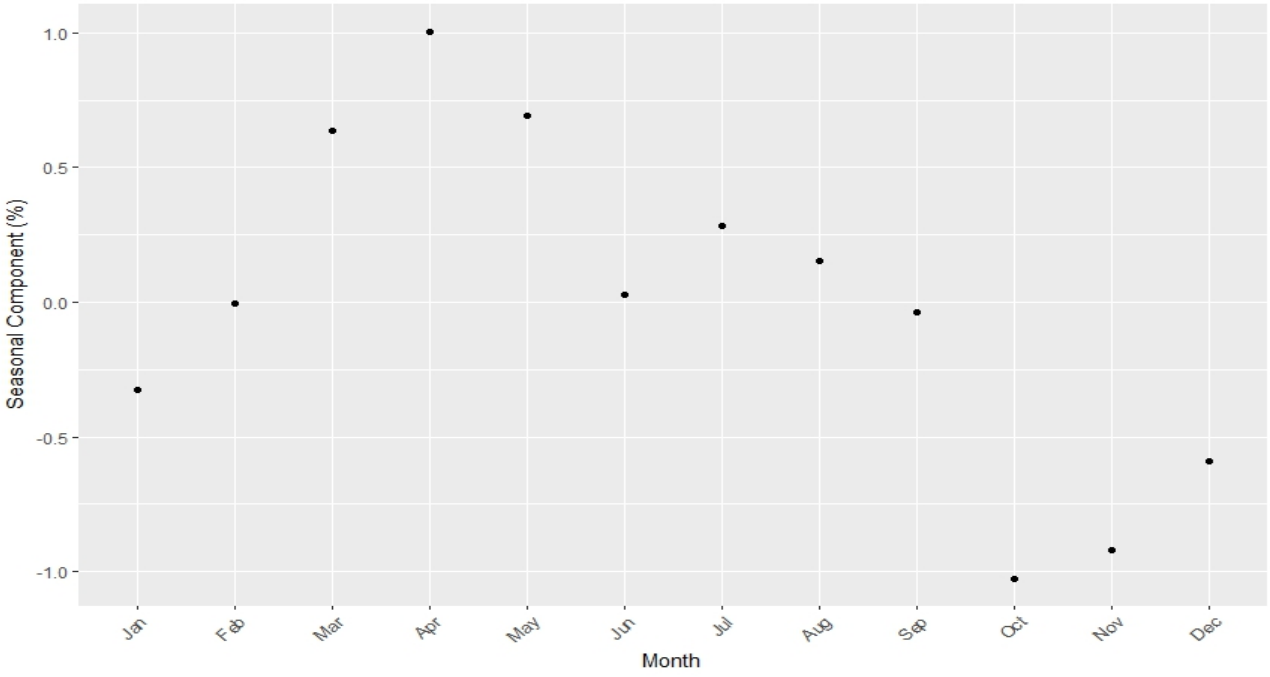


Figure 13: Smoothed Periodogram - Modified Daniell(2,2). Low frequency peak at $\omega = 0.188$ - same as in the raw periodogram.

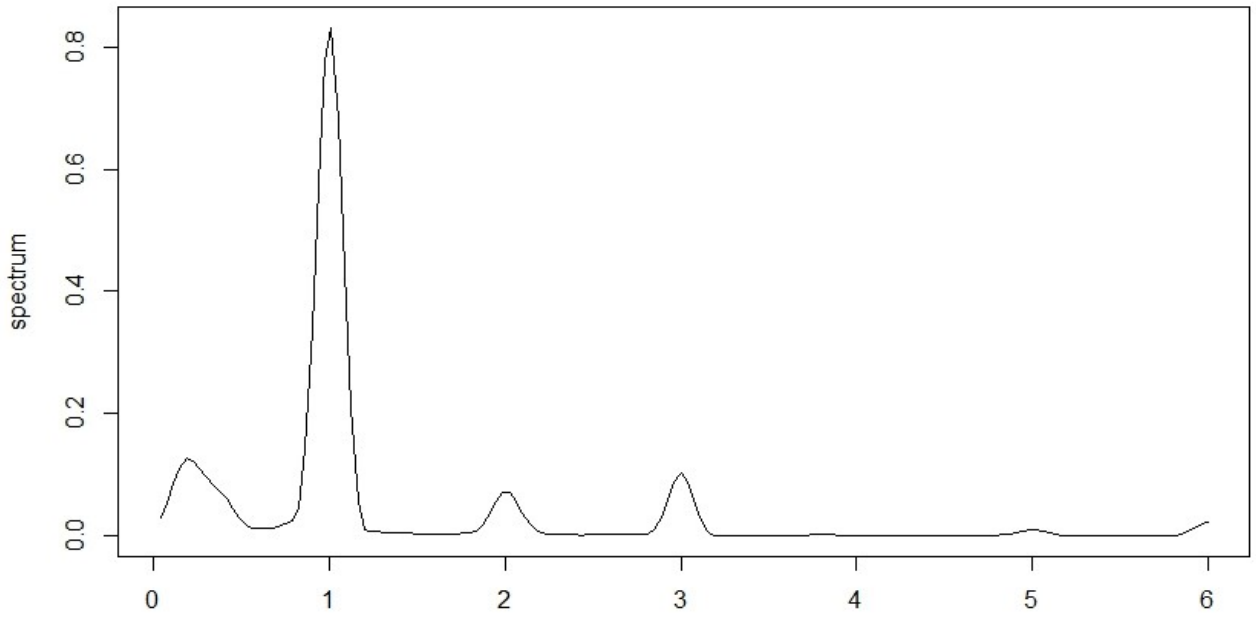


Figure 14: Autoregressive Spectral Estimator (1). Lowest AIC: p = 39.

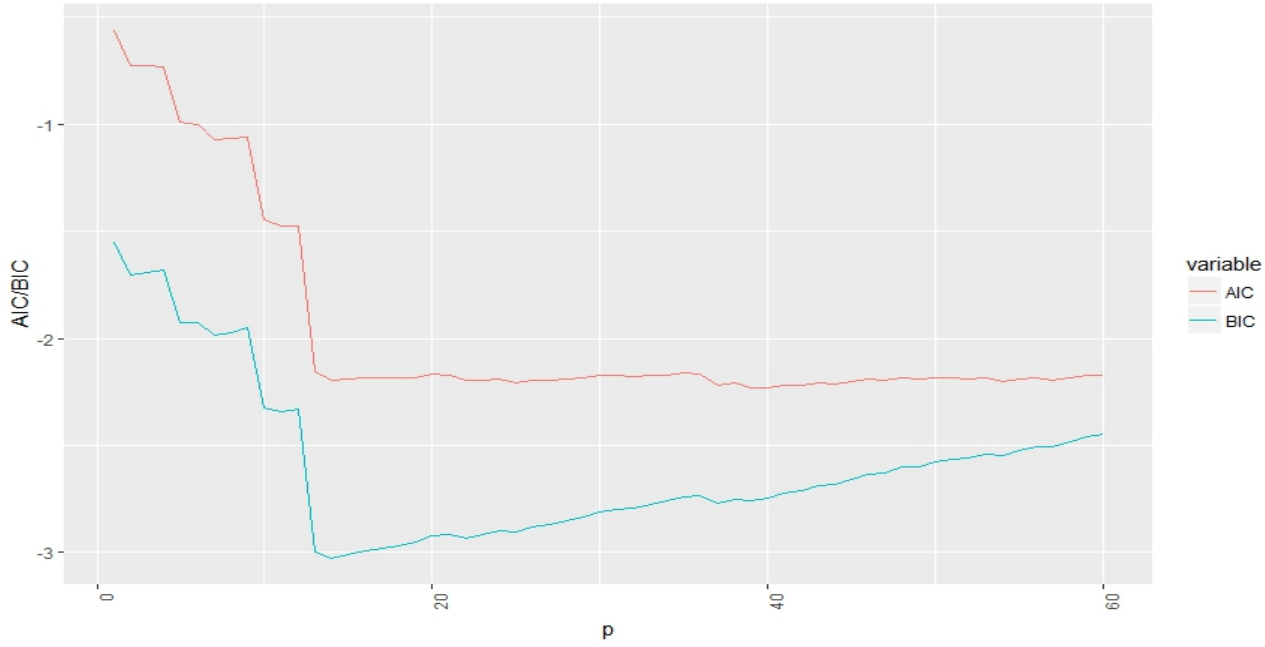


Figure 15: Autoregressive Spectral Estimator (2). Low frequency peak at $\omega = 0.192$; Corresponds to period of 5.2 years.

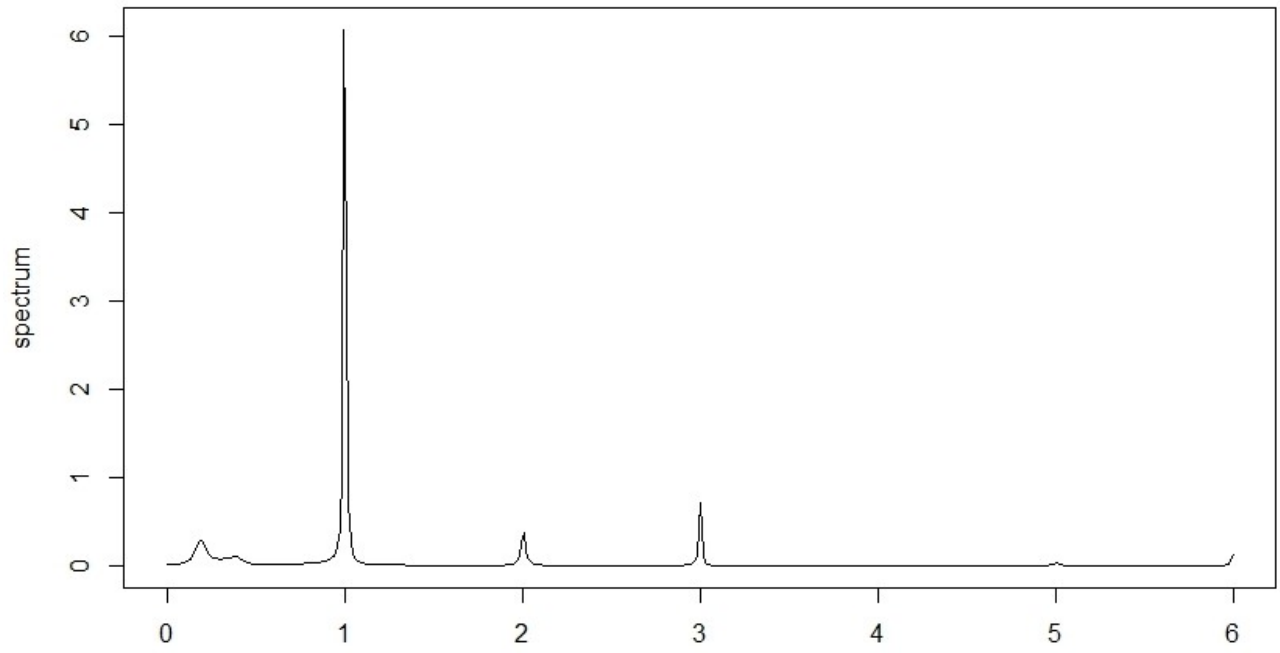


Figure 16: ACF and PACF of Detrended Data, After Removal of Select Seasonal Components. ACF still indicates there are seasonal patters left in the data, if we don't remove all cycles from it.

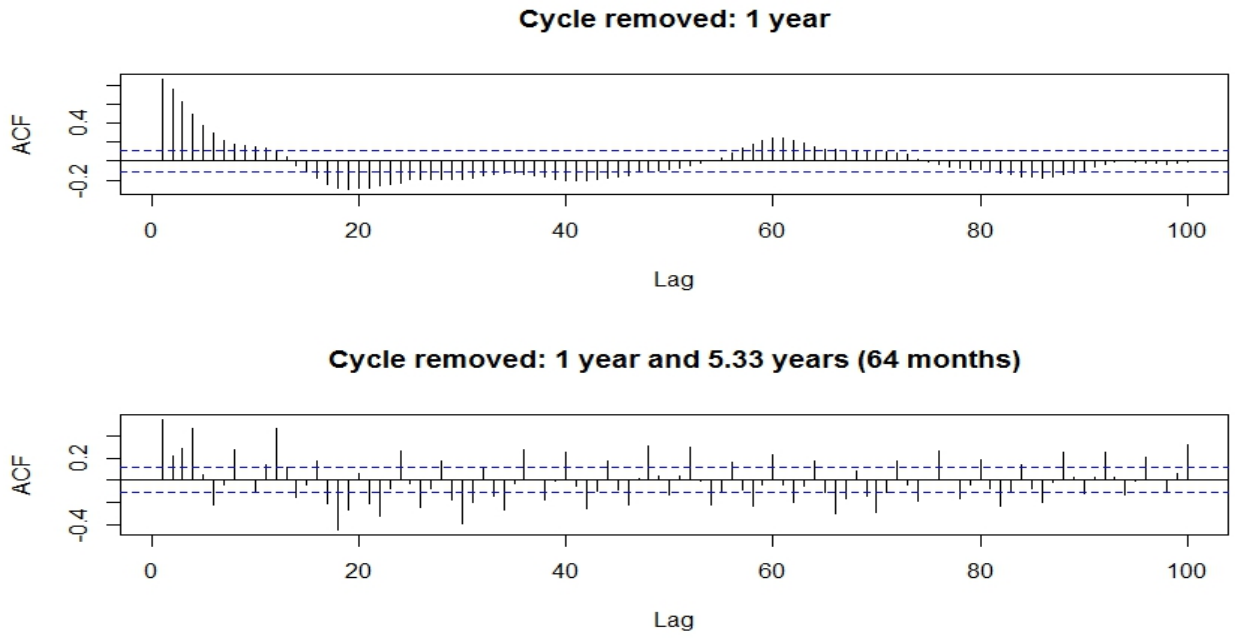


Figure 17: ACF and PACF of Residual Data (Trend and All Seasonal Components Have been Removed). ACF seems to tail off and PACF seems to cut off at lag 14 or 15, suggesting fitting either an AR(14) or AR(15) model to the residual data.

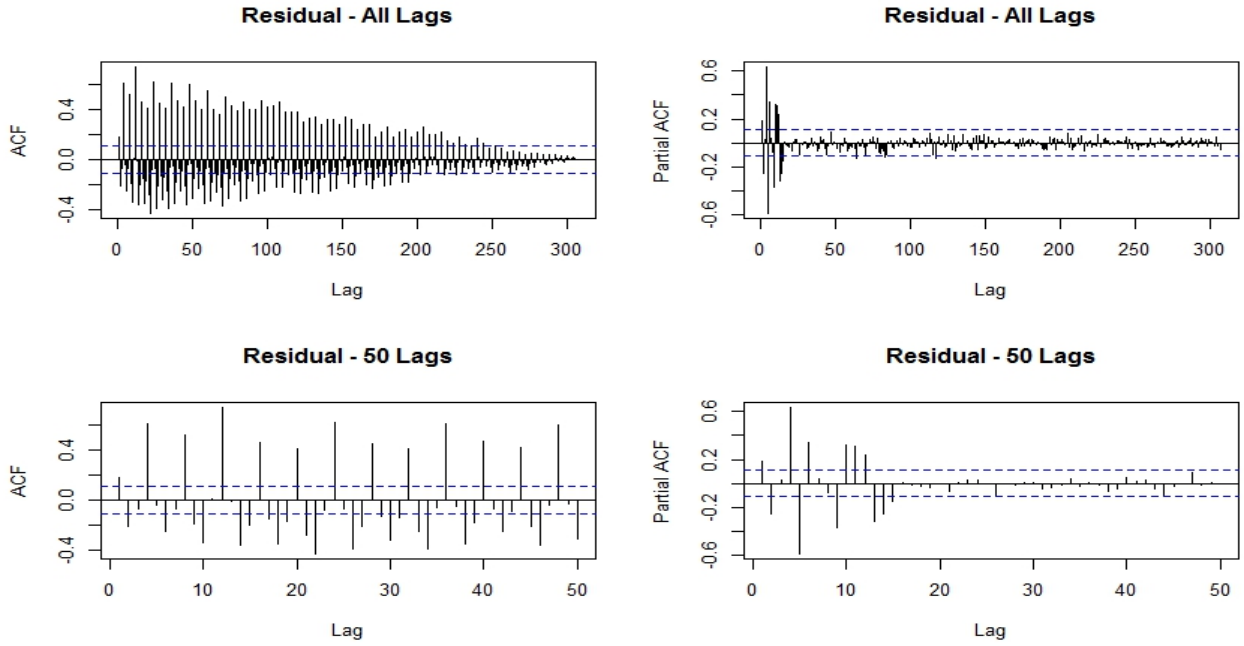


Figure 18: AIC and BIC of possible AR models. Minimum AIC at p = 14 (AIC = -2.29).

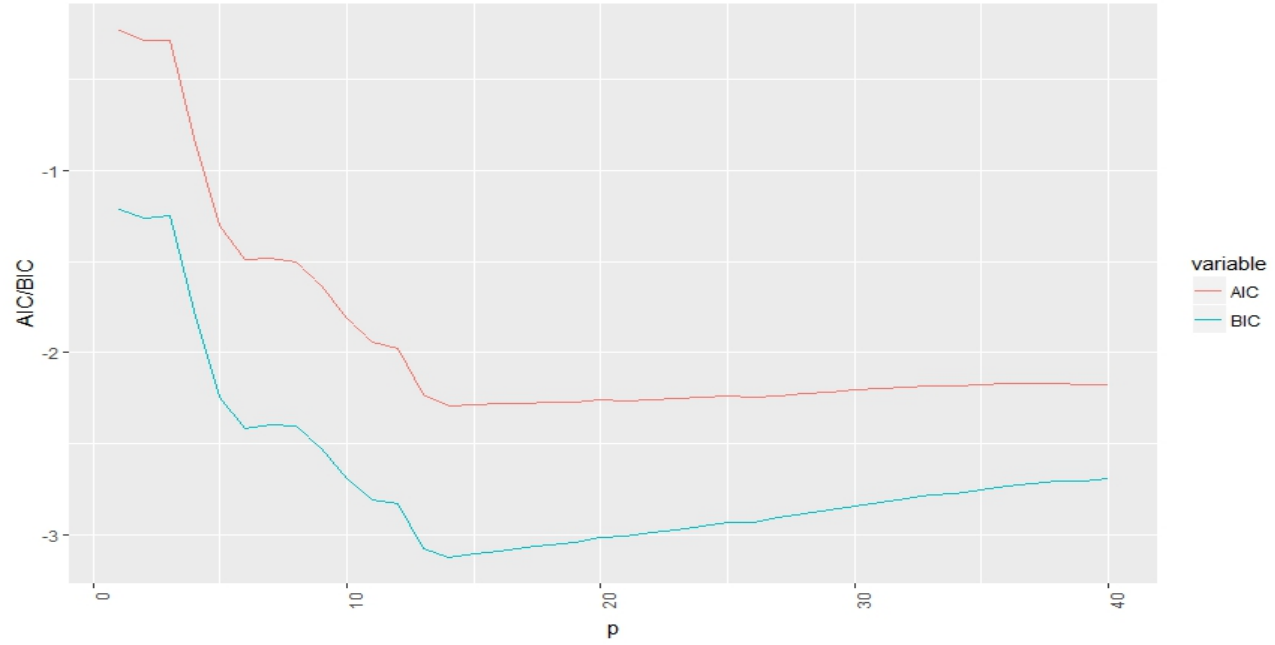


Figure 19: Diagnostics - AR(14) Model. Looks like we are mostly good, except for a bit of serial correlation in the residuals.

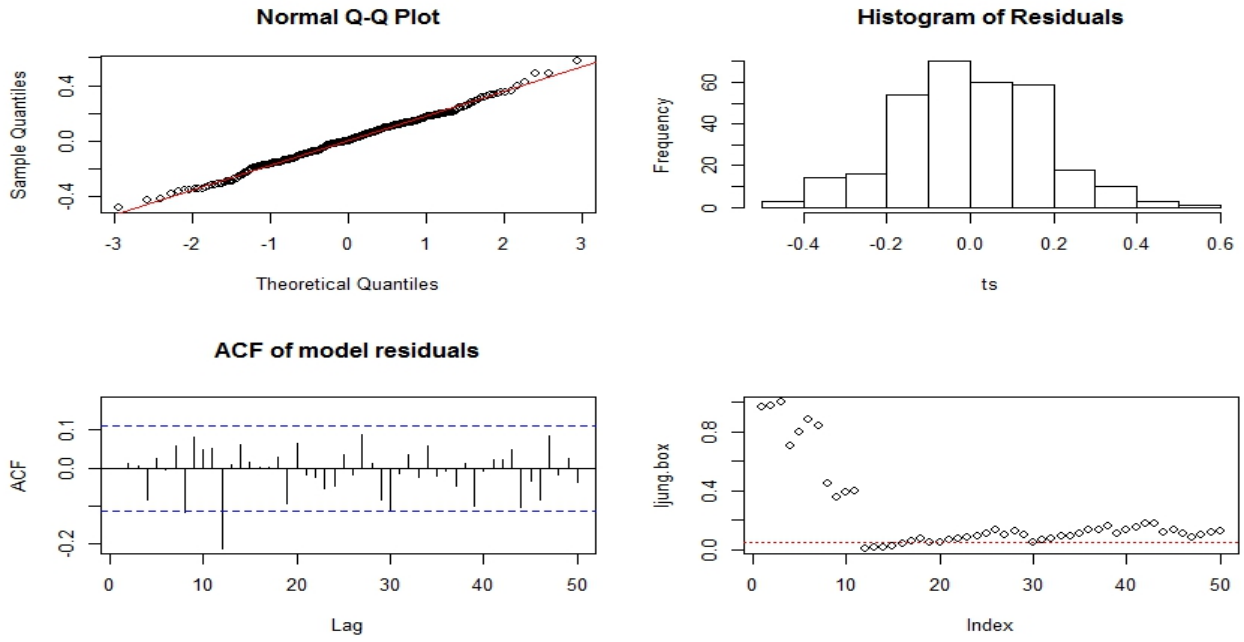


Figure 20: Diagnostics - ARMA(14,9) Model. Residuals resemble normally-distributed white noise. AIC = -2.64.

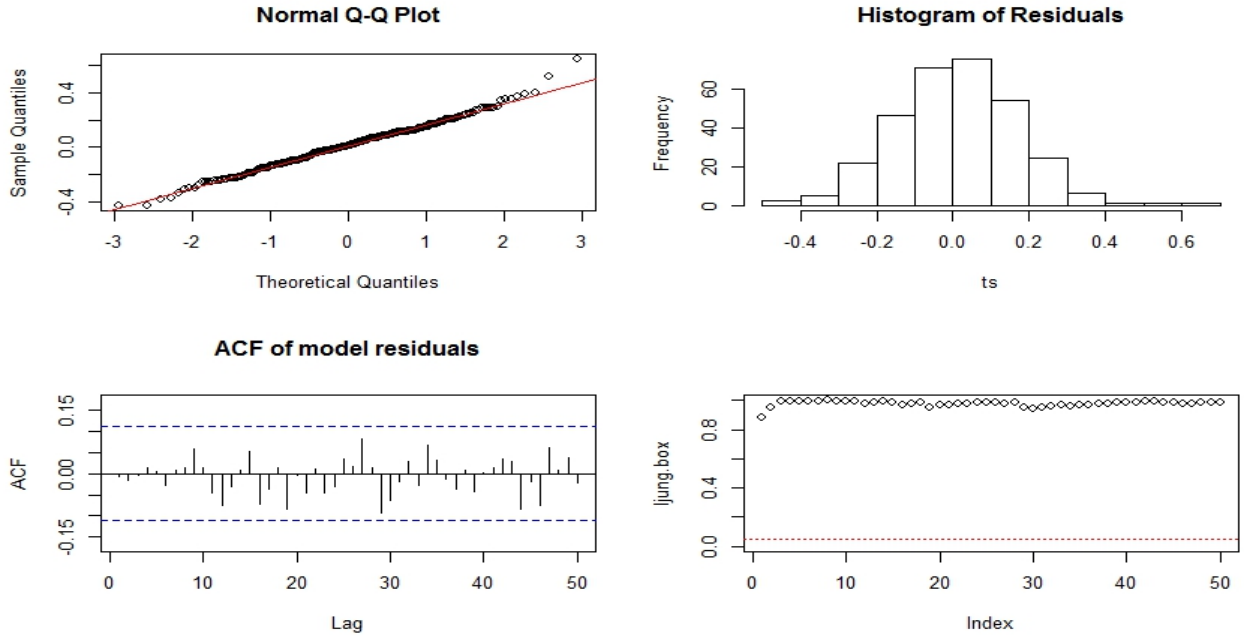


Figure 21: Forecasts using AR(14):red line indicates forecasts; blue lines indicate 2 standard-error bounds.

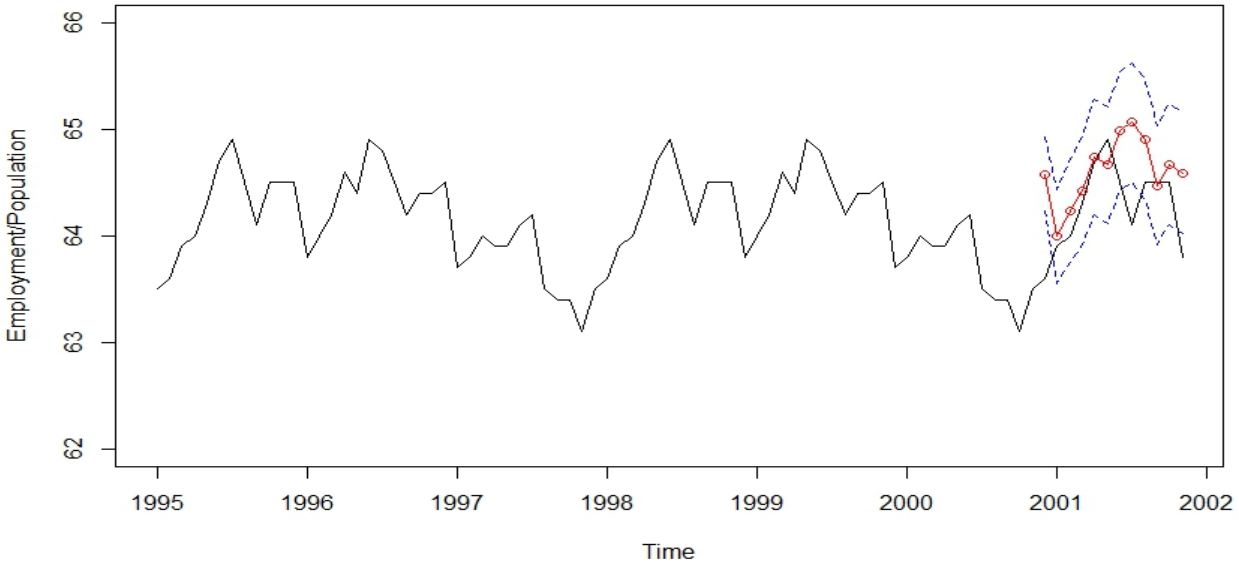


Figure 22: Forecasts using ARMA(14,9):red line indicates forecasts; blue lines indicate 2 standard-error bounds.

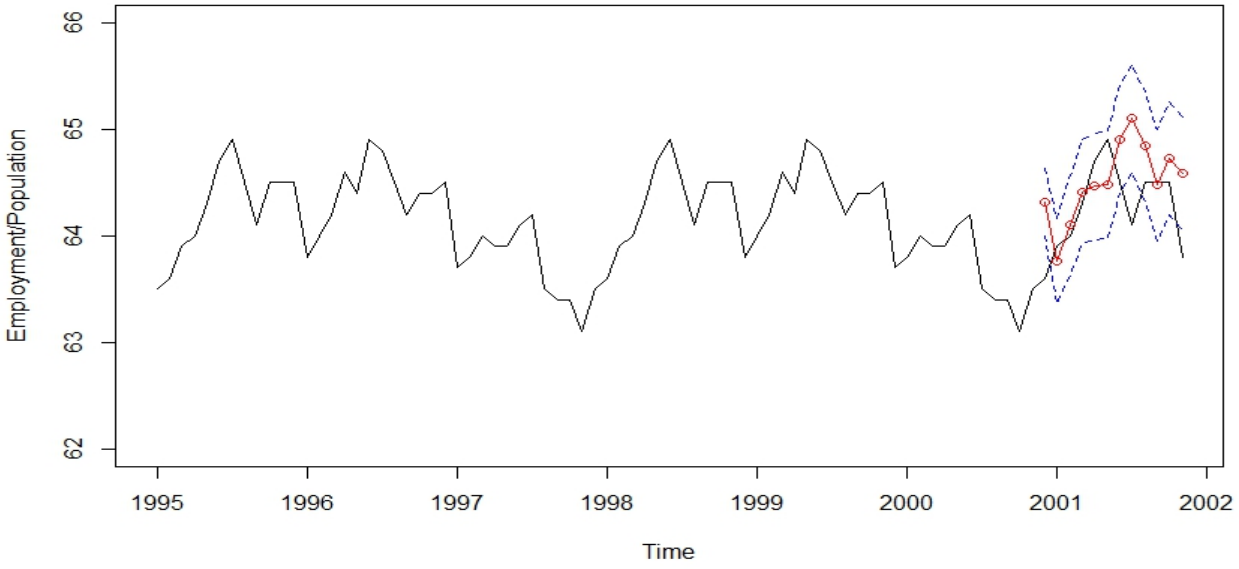
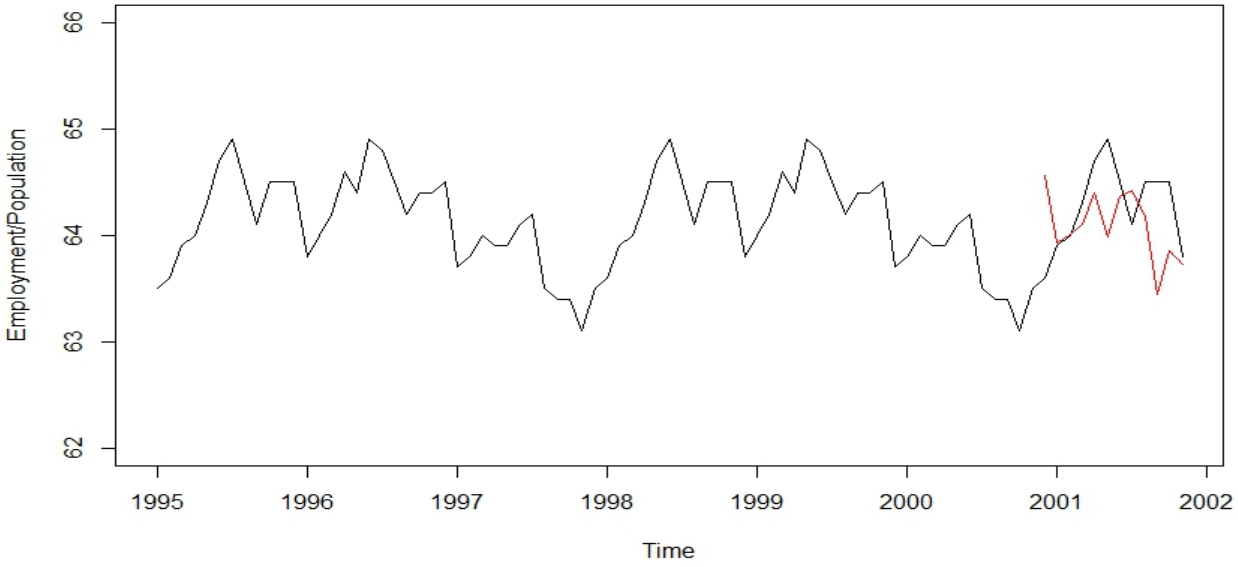


Figure 23: One-Step-Ahead Forecasts using AR(14):red line indicates forecasts.



5.2 Math Appendix - Derivation of Solution to Penalized Spline Regression

Suppose we have training examples of the type (y_i, x_i), $i = 1, ..., n$. Let x be one-dimensional (for example, x can be represent $t = time$). Also assume we have knots k_j, $j = 1, ..., p$. Suppose we try to fit a linear spline of the form $f(x) = \alpha_0 + \sum_{j=1}^{p} \alpha_j max(0, x - k_j)$ to the data, by minimizing $\sum_{i=1}^{n} ||y_i - \alpha_0 - \sum_{j=1}^{p} \alpha_j max(0, x_i - k_j)||^2 + \lambda \sum_{j=1}^{p} \alpha_j^2$.

We look for:

$$\hat{\alpha} = argmin_{\alpha_0, \{\alpha_j\}_{j=1...p}} \sum_{i=1}^{n} ||y_i - \alpha_0 - \sum_{j=1}^{p} \alpha_j max(0, x_i - k_j)||^2 + \lambda \sum_{j=1}^{p} \alpha_j^2$$

Let $\hat{\alpha}$ now denote the $(p+1) \times 1$ vector containing our parameter estimates.

Notice that we can write this problem as:

$$\hat{\alpha} = argmin_{\alpha} ||Y - Z\alpha||^2 + \lambda \alpha^T D \alpha,$$

Where D is a $(p+1) \times (p+1)$ matrix, with 1 in all its diagonal elements aside from the first one, which is 0, and with remaining elements equal to 0. In other words, $D_{rc} = 1$ if both $r = c$ and $r, c \neq 1$; and 0 otherwise (for r,c$\in\{1...p+1\}$):

$$D = \begin{bmatrix} 0 & 0 \\ 0 & I_p \end{bmatrix}$$

And Z is a $n \times (p+1)$ matrix, whose entries are:

$$Z = \begin{bmatrix} 1 & max(0, x_1 - k_1) & max(0, x_1 - k_2) & ... & max(0, x_1 - k_j) \\ 1 & max(0, x_2 - k_1) & max(0, x_2 - k_2) & ... & max(0, x_2 - k_j) \\ ... & ... & ... & ... & ... \\ 1 & max(0, x_{n-1} - k_1) & max(0, x_{n-1} - k_2) & ... & max(0, x_{n-1} - k_j) \\ 1 & max(0, x_n - k_1) & max(0, x_n - k_2) & ... & max(0, x_n - k_j) \end{bmatrix}$$

Taking the derivative with respect to α and setting it to zero, we get the FOC:

$$-2Z^T(Y - Z\alpha) + 2\lambda D\alpha = 0$$

$$-Z^TY + Z^TZ\alpha + \lambda D\alpha = 0$$

$$(Z^TZ + \lambda D)\alpha = Z^TY$$

Which implies:

$$\hat{\alpha} = (Z^TZ + \lambda D)^{-1}Z^TY$$

5.3 Appendix: R Code

```
1
2
3  library(ggplot2) # nice-looking plots
4  library(quantmod) # reads in data from Fred
5  library(scales) # customize x axis in ggplots
6  library(reshape2) # plot multiple ts with ggplot
7  library(gridExtra) # plot multiple graphs with ggplot (like par(mfrow) command)
8  library(tseries)
9  library(forecast) # Acf plots that ommit lag 0
10 library(data.table)
11
12
13 employment = getSymbols('LNU02300000',src='FRED', auto.assign=F)
14 employment.df = data.frame(date=time(employment), coredata(employment) )
15 employment.df[,1] = as.Date(employment.df[,1])
16 colnames(employment.df)[2] = "Employment"
17
18 unemployment = getSymbols('UNRATENSA',src='FRED', auto.assign=F)
19 unemployment.df = data.frame(date=time(unemployment), coredata(unemployment) )
20 colnames(unemployment.df)[2] = "Unemployment"
21
22 manufacturing = getSymbols('CEU3000000001',src='FRED', auto.assign=F)
23 manufacturing.df = data.frame(date=time(manufacturing), coredata(manufacturing) )
24 colnames(manufacturing.df)[2] = "Manufacturing"
25 manufacturing.df = subset(manufacturing.df, date >= "1948-01-01" )
26
27
28 # reading-in recession dates - We will just copy-paste the dates
29 # here from Fred's website, for simplicity pourposes
30 recessions.df = read.table(textConnection(
31 "Peak, Trough
32 1857-06-01, 1858-12-01
33 1860-10-01, 1861-06-01
34 1865-04-01, 1867-12-01
35 1869-06-01, 1870-12-01
36 1873-10-01, 1879-03-01
37 1882-03-01, 1885-05-01
38 1887-03-01, 1888-04-01
39 1890-07-01, 1891-05-01
```

```
1893-01-01, 1894-06-01
1895-12-01, 1897-06-01
1899-06-01, 1900-12-01
1902-09-01, 1904-08-01
1907-05-01, 1908-06-01
1910-01-01, 1912-01-01
1913-01-01, 1914-12-01
1918-08-01, 1919-03-01
1920-01-01, 1921-07-01
1923-05-01, 1924-07-01
1926-10-01, 1927-11-01
1929-08-01, 1933-03-01
1937-05-01, 1938-06-01
1945-02-01, 1945-10-01
1948-11-01, 1949-10-01
1953-07-01, 1954-05-01
1957-08-01, 1958-04-01
1960-04-01, 1961-02-01
1969-12-01, 1970-11-01
1973-11-01, 1975-03-01
1980-01-01, 1980-07-01
1981-07-01, 1982-11-01
1990-07-01, 1991-03-01
2001-03-01, 2001-11-01
2007-12-01, 2009-06-01"), sep=',',
colClasses=c('Date', 'Date'), header=TRUE)

recessions.trim = subset(recessions.df, Peak >= "1948-01-01" )



# Creating initial plots
# Employment-Population
emp = ggplot(employment.df) + geom_line(aes(x=date, y=Employment))  +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
geom_rect(data=recessions.trim, aes(xmin=Peak, xmax=Trough, ymin=-Inf, ymax=+Inf),
fill='blue', alpha=0.1)  +ylab("Employment-Population Ratio")

emp +scale_x_date(date_breaks = '5 years', date_labels  = "%Y %b")

# Unemployment
unemp = ggplot(unemployment.df) + geom_line(aes(x=date, y=Unemployment)) +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
geom_rect(data=recessions.trim, aes(xmin=Peak, xmax=Trough, ymin=-Inf, ymax=+Inf),
fill='blue', alpha=0.1) +ylab("Unemployment Rate")

unemp + scale_x_date(date_breaks = '5 years', date_labels  = "%Y %b")

# Manufacturing
manuf = ggplot(manufacturing.df) + geom_line(aes(x=date, y=Manufacturing)) +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
geom_rect(data=recessions.trim, aes(xmin=Peak, xmax=Trough, ymin=-Inf, ymax=+Inf),
fill='blue', alpha=0.1) +ylab("Number of Employees in Manufacturing")

manuf +scale_x_date(date_breaks = '5 years', date_labels  = "%Y %b")



# Dividing-up data into regimes
```

```
window1 = data.frame(Beggining = as.Date("1948-01-01"), End = as.Date("1975-03-01"))
window2 = data.frame(Beggining = as.Date("1975-04-01"), End = as.Date("2001-11-01"))
window3 = data.frame(Beggining = as.Date("2001-12-01"), End = as.Date("2008-09-01"))
window4 = data.frame(Beggining = as.Date("2009-7-01"), End = as.Date("2016-06-01"))


emp_piecewise = ggplot(employment.df) + geom_line(aes(x=date, y=Employment))  +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
geom_rect(data=window1, aes(xmin=Beggining, xmax=End, ymin=-Inf, ymax=+Inf),
fill='blue', alpha=0.1)  +
geom_rect(data=window2, aes(xmin=Beggining, xmax=End, ymin=-Inf, ymax=+Inf),
fill='green', alpha=0.1)  +
geom_rect(data=window3, aes(xmin=Beggining, xmax=End, ymin=-Inf, ymax=+Inf),
fill='yellow', alpha=0.1)  +
geom_rect(data=window4, aes(xmin=Beggining, xmax=End, ymin=-Inf, ymax=+Inf),
fill='red', alpha=0.1)  +
ylab("Employment-Population Ratio")

emp_piecewise +scale_x_date(date_breaks = '5 years', date_labels  = "%Y %b")


# fitting trend lines
window1.index = employment.df$date >= window1$Beggining &
employment.df$date <= window1$End

window2.index = employment.df$date >= window2$Beggining &
employment.df$date <= window2$End

window3.index = employment.df$date >= window3$Beggining &
employment.df$date <= window3$End

window4.index = employment.df$date >= window4$Beggining &
employment.df$date <= window4$End


emp_ts1 = ts(employment.df[window1.index,2], start = c(1948,1),end = c(1975,3),frequency = 12)

emp_ts2 = ts(employment.df[window2.index,2], start = c(1975,4),end = c(2000,11),frequency = 12)

emp_ts3 = ts(employment.df[window3.index,2], start = c(2001,12),end = c(2008,09),frequency = 12)

emp_ts4 = ts(employment.df[window4.index,2], start = c(2009,7),end = c(2016,6),frequency = 12)


summary(fit1 <- lm(emp_ts1~time(emp_ts1))) #
summary(fit2 <- lm(emp_ts2~time(emp_ts2))) #
summary(fit3 <- lm(emp_ts3~time(emp_ts3))) #
summary(fit4 <- lm(emp_ts4~time(emp_ts4))) #


# Plotting linear trends
data.plot <- data.frame(date = time(emp_ts1), Employment_Population = as.numeric(emp_ts1),Trend = fit1$fitted.values)

data_long <- melt(data.plot, id="date")
plot1 = ggplot(data=data_long,
aes(x=date, y=value, colour=variable)) +
theme(legend.position="none",axis.text.x = element_text(angle = 45, hjust = 1)) +
geom_line() +
```

```
labs(title="Blue Area") +ylab("Ratio")

data.plot <- data.frame(date = time(emp_ts2), Employment_Population = as.numeric(emp_ts2),Trend = fit2$fitted.values)

data_long <- melt(data.plot, id="date")
plot2 = ggplot(data=data_long,
aes(x=date, y=value, colour=variable)) +
theme(legend.position="none",axis.text.x = element_text(angle = 45, hjust = 1))+
geom_line() +
labs(title="Green Area (Until Nov. 2000)") +ylab("Ratio")

data.plot <- data.frame(date = time(emp_ts3), Employment_Population = as.numeric(emp_ts3),Trend = fit3$fitted.values)

data_long <- melt(data.plot, id="date")
plot3 = ggplot(data=data_long,
aes(x=date, y=value, colour=variable)) +
theme(legend.position="none",axis.text.x = element_text(angle = 45, hjust = 1))+
geom_line() +
labs(title="Yellow Area")+ylab("Ratio")

data.plot <- data.frame(date = time(emp_ts4), Employment_Population = as.numeric(emp_ts4),Trend = fit4$fitted.values)

data_long <- melt(data.plot, id="date")
plot4 = ggplot(data=data_long,
aes(x=date, y=value, colour=variable)) +
theme(legend.position="none",axis.text.x = element_text(angle = 45, hjust = 1))+
geom_line() +
labs(title="Red Area")+ylab("Ratio")



grid.arrange(plot1, plot2,plot3,plot4,nrow = 2, ncol=2)


# From now on we focus on the 4-1975 to 11-2000 period
# redefining our variables
window = data.frame(Beggining = as.Date("1975-04-01"), End = as.Date("2000-11-01"))
emp_ts = emp_ts2

# ACF and PACF
par(mfrow = c(2,2))

# Maximum number of lags
Acf(emp_ts,length(emp_ts),main = "Original Data - All Lags")
Pacf(emp_ts,length(emp_ts),main = "Original Data - All Lags")

# ACF and PACF
# Maximum number of lags
Acf(emp_ts,50,main = "Original Data - 50 lags")
Pacf(emp_ts,50,main = "Original Data - 50 Lags")

## The detrended data
detrended = resid(fit2)

# Plotting detrended data
data.plot <- data.frame(date = time(emp_ts), detrended = detrended)
ggplot(data.plot) + geom_line(aes(x=date, y=detrended))  +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
labs(title="Detrended data - linear trend")
```

```
214
215
216 adf.test(detrended)
217
218 # using smoothing splines to extract trend
219 # we will use a function that we have created
220
221 # Creating the knots, note that we will use the xi's
222 # themselves as the knots
223 x = as.numeric(time(emp_ts))
224 knots = x
225 n = length(x)
226 Y = as.numeric(emp_ts)
227
228 filloutZ = function(x,knots){
229
230 p = length(knots)
231
232 Z = matrix(nrow = n,ncol = p+1)
233 ZMax = matrix(nrow = n,ncol = p)
234
235 for (i in 1:n){
236
237 for (j in 1:p){
238
239 ZMax[i,j] = max(0,x[i]-knots[j])
240 }
241
242 }
243
244 Z = as.matrix(cbind(1,ZMax))
245
246 return(Z)
247 }
248
249
250 # Next, we declare a function that solves for alpha.hat
251 solveforalpha.hat = function(Z,Y,lambda){
252
253 ncolZ = dim(Z)[2]
254
255 D = diag(ncolZ)
256 D[1,1] = 0
257
258 return(solve(t(Z)%*%Z + lambda*D,t(Z)%*%Y))
259
260 }
261
262
263 # Let's fit the model for a fixed level of lambda
264 # We'll manually tweak lambda, until we get a
265 # trend that is smooth enough, and a detrended
266 # series that looks stationary
267 lambda = 100
268
269 Z = filloutZ(x,knots)
270 alpha.hat = solveforalpha.hat(Z,Y,lambda)
271 Y.hat = Z%*%alpha.hat
```

```r

detrended = as.numeric(emp_ts) - as.numeric(Y.hat)

# Plotting non-linear trend
data.plot <- data.frame(date = x, Employment_Population = Y,Trend = Y.hat)

data_long <- melt(data.plot, id="date")
ggplot(data=data_long,
aes(x=date, y=value, colour=variable)) +
theme(legend.position="none",axis.text.x = element_text(angle = 90, hjust = 1)) +
geom_line() +
labs(title="Data + Non-Linear Trend") +ylab("Ratio")


# Plotting detrended data
data.plot <- data.frame(date = x, detrended = detrended)
ggplot(data.plot) + geom_line(aes(x=date, y=detrended))  +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
labs(title="Detrended data - non-linear trend")


#adf test
adf.test(detrended)

detrended = ts(detrended,start = c(1975,4),end = c(2000,11),frequency = 12)

par(mfrow = c(2,2))
# ACF and PACF
# Maximum number of lags
Acf(detrended,length(detrended),main = "Detrended Data - All Lags")
Pacf(detrended,length(detrended),main = "Detrended Data - All Lags")

# ACF and PACF
# Maximum number of lags
Acf(emp_ts,50,main = "Detrended Data - 50 lags")
Pacf(emp_ts,50,main = "Detrended Data - 50 Lags")


# Creating Periodograms
par(mfrow = c(1,1))

# RAW ##############################################################################
detrended.per = spec.pgram(detrended, taper=0, log="no",detrend = FALSE,main = "")

frequencies = detrended.per$freq
spectrum = detrended.per$spec
rank(-spectrum)

# The frequencies corresponding to the
# largest spectral densities values

f = NULL

f[1] = frequencies[rank(-spectrum) == 1]
f[2] = frequencies[rank(-spectrum) == 2]
f[3] = frequencies[rank(-spectrum) == 3]
f[4] = frequencies[rank(-spectrum) == 4]
f[5] = frequencies[rank(-spectrum) == 5]
```

```
f[6] = frequencies[rank(-spectrum) == 6]

# these correspond to the following periods:
1/f

# conf intervals - returned value:
U = qchisq(.025,2) # 0.05063
L = qchisq(.975,2) # 7.37775
l1 = 2*detrended.per$spec[frequencies == f[1]]/L
u1 = 2*detrended.per$spec[frequencies == f[1]]/U
l2 = 2*detrended.per$spec[frequencies == f[3]]/L
u2 = 2*detrended.per$spec[frequencies == f[3]]/U
l3 = 2*detrended.per$spec[frequencies == f[4]]/L
u4 = 2*detrended.per$spec[frequencies == f[4]]/U
l4 = 2*detrended.per$spec[frequencies == f[5]]/L
u4 = 2*detrended.per$spec[frequencies == f[5]]/U


abline(v=f[1], lty="dotted")
abline(v=f[3], lty="dotted")
abline(v=f[4], lty="dotted")
abline(v=f[5], lty="dotted")

abline(h = l1,lty="dotted", col = 'red')
#abline(h = l2,lty="dotted", col = 'red')
#abline(h = l3,lty="dotted", col = 'red')
#abline(h = l4,lty="dotted", col = 'red')

# The spectrums of relevance:
spectrum_rel = NULL
spectrum_rel[1] = detrended.per$spec[frequencies == f[1]]
spectrum_rel[2] = detrended.per$spec[frequencies == f[3]]
spectrum_rel[3] = detrended.per$spec[frequencies == f[4]]
spectrum_rel[4] = detrended.per$spec[frequencies == f[5]]

freq = c(f[1],f[3],f[4],f[5])
period = 1/freq
summary = as.matrix(cbind(freq,period))


# Smoothed Periodogram#############################################################

k = kernel("modified.daniell", c(2,2))

smooth = spec.pgram(detrended, k, taper=0, log="no", main = "Smoothed Periodogram - Modified Daniell(2,2)")
frequencies = smooth$freq
spectrum = smooth$spec
rank(-spectrum)

# The frequencies corresponding to the
# largest spectral densities values

f = NULL

f[1] = frequencies[rank(-spectrum) == 1]
f[2] = frequencies[rank(-spectrum) == 2]
f[3] = frequencies[rank(-spectrum) == 3]
f[4] = frequencies[rank(-spectrum) == 4]
```

```
f[5] = frequencies[rank(-spectrum) == 5]
f[6] = frequencies[rank(-spectrum) == 6]
f[7] = frequencies[rank(-spectrum) == 7]
f[8] = frequencies[rank(-spectrum) == 8]
f[9] = frequencies[rank(-spectrum) == 9]
f[10] = frequencies[rank(-spectrum) == 10]
f[11] = frequencies[rank(-spectrum) == 11]
f[12] = frequencies[rank(-spectrum) == 12]

# these correspond to the following periods:
1/f

#Autoregressive Spectral Estimator####################################

####AIC BIC#########################################################

###ACF PACF#########################################################################
par(mfrow = c(2,2))

AIC = rep(0, 60) -> BIC
for (k in 1:60){
fit = ar(detrended, order=k, aic=FALSE)
sigma2 = var(fit$resid, na.rm=TRUE)
BIC[k] = log(sigma2) + (k*log(n)/n)
AIC[k] = log(sigma2) + ((n+2*k)/n) }
IC = cbind(AIC,BIC)

# Plotting AIC/BIC
data.plot <- data.frame(p = seq(1,60,1), AIC = AIC,BIC = BIC)

data_long <- melt(data.plot, id="p")
ggplot(data=data_long,
aes(x=p, y=value, colour=variable)) +
theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
geom_line() +
labs(title="") +ylab("AIC/BIC")

lag = which.min(AIC)

par(mfrow=c(1,1))

sp.ar = spec.ar(detrended, log="no",order=lag, main = "Autoregressive Spectral Estimator")

frequencies = sp.ar$freq
spectrum = sp.ar$spec
rank(-spectrum)

# The frequencies corresponding to the
# largest spectral densities

f = NULL

f[1] = frequencies[rank(-spectrum) == 1]
f[2] = frequencies[rank(-spectrum) == 2]
f[3] = frequencies[rank(-spectrum) == 3]
f[4] = frequencies[rank(-spectrum) == 4]
f[5] = frequencies[rank(-spectrum) == 5]
f[6] = frequencies[rank(-spectrum) == 6]
```

```
f[7] = frequencies[rank(-spectrum) == 7]
f[8] = frequencies[rank(-spectrum) == 8]
f[9] = frequencies[rank(-spectrum) == 9]

# these correspond to the following periods:
1/f

#Removing Seasonal Components
data = data.table(detrended = as.numeric(detrended),index1 = seq(1,12,1),index2 = seq(1,4,1),index3 = seq(1,64,1),index4 = seq
      (1,6,1))
seasonality1 = data[,mean(detrended),by = index1]
seasonality2 = data[,mean(detrended),by = index2]
seasonality3 = data[,mean(detrended),by = index3]
seasonality4 = data[,mean(detrended),by = index4]



data_merged = merge(data,seasonality1, all.x = TRUE,by = 'index1', sort = 'FALSE')
data_merged = merge(data_merged,seasonality2, all.x = TRUE,by = 'index2', sort = 'FALSE')
data_merged = merge(data_merged,seasonality3, all.x = TRUE,by = 'index3', sort = 'FALSE')
data_merged = merge(data_merged,seasonality4, all.x = TRUE,by = 'index4', sort = 'FALSE')

data_merged = as.data.frame(data_merged)
residual1 = data_merged$detrended - data_merged[,c(6)] # remove 1 year
residual2 = data_merged$detrended - rowSums(data_merged[,c(6,8)]) # remove 1 year and 64 months
residual3 = data_merged$detrended - rowSums(data_merged[,c(6,7,9)]) # remove 1 year, 4 months, 6 months

par(mfrow = c(2,1))
Acf(residual1,100, main = 'Cycle removed: 1 year')
Acf(residual2,100,main = 'Cycle removed: 1 year and 5.33 years (64 months)')

residual = data_merged$detrended - rowSums(data_merged[,c(6,7,8,9)])

adf.test(residual)

# plotting one-year cycle
data.plot <- data.frame(date = c("Jan","Feb","Mar","Apr","May","Jun",
"Jul","Aug","Sep","Oct","Nov","Dec"), value = seasonality1$V1)
ggplot(data.plot) + geom_point(aes(x=factor(date, levels=unique(date)),y=value))  +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
ylab("Seasonal Component (%)") +xlab("Month")



#ACF - PACF of residuals
par(mfrow=c(2,2))
acf.data = Acf(residual,length(residual),main = "Residual - All Lags")
pacf.data = Pacf(residual,length(residual),main = "Residual - All Lags")
acf.data = Acf(residual,50, main = "Residual - 50 Lags")
pacf.data = Pacf(residual,50,main = "Residual - 50 Lags")

# PACF seems to cut off after 14, 15 lags,
# ACF tails off
# suggests AR(14) or AR(14)


###ACF PACF#####################################################################
#figuring out best AIC for AR model

AIC = rep(0, 40) -> BIC
```

```
for (k in 1:40){

fit = ar(residual, order=k, aic=FALSE)
sigma2 = var(fit$resid, na.rm=TRUE)
BIC[k] = log(sigma2) + (k*log(n)/n)
AIC[k] = log(sigma2) + ((n+2*k)/n) }


IC = cbind(AIC,BIC)

lag = which.min(AIC)

# Plotting AIC/BIC
data.plot <- data.frame(p = seq(1,40,1), AIC = AIC,BIC = BIC)

data_long <- melt(data.plot, id="p")
ggplot(data=data_long,
aes(x=p, y=value, colour=variable)) +
theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
geom_line() +
labs(title="") +ylab("AIC/BIC")


#Fitting AR model
model_ar = arima(residual, order = c(14, 0, 0),include.mean = FALSE,method="ML")

ts = model_ar$residuals

par(mfrow = c(2,2))
qqnorm(ts)
qqline(ts, col=2)
hist(ts, br=12, main = 'Histogram of Residuals')
shapiro.test(ts)
Acf(ts,lag = 50, main = 'ACF of model residuals')

ljung.box = NULL
for (i in 1:50){
ljung.box[i] =  Box.test(ts,lag = i, type = c("Ljung-Box"))$p.value
}

# we have serial correlation from lags 12 to 16
plot(ljung.box)
abline(h = 0.05,lty="dotted", col = 'red')

# fitting an arma model to data
k = 9
model_arima = arima(residual, order = c(14, 0, k),include.mean = FALSE,method = 'ML',optim.control = list(maxit = 1000))
sigma2 = var(model_arima$residuals, na.rm=TRUE)
AIC = log(sigma2) + ((n+2*k)/n)

ts = model_arima$residuals


par(mfrow = c(2,2))
qqnorm(ts)
qqline(ts, col=2)
hist(ts, br=12, main = 'Histogram of Residuals')
shapiro.test(ts)
```

```
Acf(ts,lag = 50, main = 'ACF of model residuals')

ljung.box = NULL
for (i in 1:50){
ljung.box[i] =  Box.test(ts,lag = i, type = c("Ljung-Box"))$p.value
}

plot(ljung.box,ylim = c(0,1))
abline(h = 0.05,lty="dotted", col = 'red')


# forecasting employment#############################################################
# AR(14) forecasts
regr = ar.ols(residual, order=14, demean=FALSE, intercept=FALSE)
fore = predict(regr, n.ahead=12)

# Seaonality forecast

s4 = c(3,4,5,6,1,2,3,4,5,6,1,2)
s3 = seq(53,64)
s2 = rep(c(1,2,3,4),3)
s1 = c(9,10,11,12,seq(1,8,1))

data = data.table(index1 = s1,index2 = s2,index3 = s3,index4 = s4)

data_merged = merge(data,seasonality1, all.x = TRUE,by = 'index1', sort = 'FALSE')
data_merged = merge(data_merged,seasonality2, all.x = TRUE,by = 'index2', sort = 'FALSE')
data_merged = merge(data_merged,seasonality3, all.x = TRUE,by = 'index3', sort = 'FALSE')
data_merged = merge(data_merged,seasonality4, all.x = TRUE,by = 'index4', sort = 'FALSE')

data_merged = as.data.frame(data_merged)
seasonal_forecast = rowSums(data_merged[,5:8])

# Trend forecast

filloutZ_test = function(x,knots){

p = length(knots)
n = length(x)
Z = matrix(nrow = n,ncol = p+1)
ZMax = matrix(nrow = n,ncol = p)

for (i in 1:n){

for (j in 1:p){

ZMax[i,j] = max(0,x[i]-knots[j])
}

}

Z = as.matrix(cbind(1,ZMax))

return(Z)
}
```

```r

time_test = x[length(x)]+seq(1,12,1)/12
Z_test = filloutZ_test(time_test,knots)
Y.hat = Z_test%*%alpha.hat

# Adding up all components
forecast = as.numeric(Y.hat) + as.numeric(seasonal_forecast)+as.numeric(fore$pred)
fore$pred = ts(forecast, start = c(2000,12), end = c(2001,11), frequency = 12)
fore$se = ts(as.numeric(fore$se), start = c(2000,12), end = c(2001,11), frequency = 12)



window.plot = data.frame(Beggining = as.Date("1999-01-01"), End = as.Date("2001-11-01"))
window.index = employment.df$date >= window.plot$Beggining &
employment.df$date <= window.plot$End
emp.plot = ts(employment.df[window.index,2], start = c(1995,1),end = c(2001,11),frequency = 12)

par(mfrow = c(1,1))
dataplot = window(emp.plot, c(1995,1), c(2001, 11))
ts.plot(dataplot, fore$pred, col=1:2,
ylab="Employment/Population",ylim = c(62,66))
lines(fore$pred, type="p", col=2)
lines(fore$pred+2*fore$se, lty="dashed", col=4)
lines(fore$pred-2*fore$se, lty="dashed", col=4)



##forecasts using ARIMA(14,9)#################################################################

fore = predict(model_arima,n.ahead = 12)

# Adding up all components
forecast = as.numeric(Y.hat) + as.numeric(seasonal_forecast)+as.numeric(fore$pred)
fore$pred = ts(forecast, start = c(2000,12), end = c(2001,11), frequency = 12)
fore$se = ts(as.numeric(fore$se), start = c(2000,12), end = c(2001,11), frequency = 12)



window.plot = data.frame(Beggining = as.Date("1999-01-01"), End = as.Date("2001-11-01"))
window.index = employment.df$date >= window.plot$Beggining &
employment.df$date <= window.plot$End
emp.plot = ts(employment.df[window.index,2], start = c(1995,1),end = c(2001,11),frequency = 12)

par(mfrow = c(1,1))
dataplot = window(emp.plot, c(1995,1), c(2001, 11))
ts.plot(dataplot, fore$pred, col=1:2,
ylab="Employment/Population",ylim = c(62,66))
lines(fore$pred, type="p", col=2)
lines(fore$pred+2*fore$se, lty="dashed", col=4)
lines(fore$pred-2*fore$se, lty="dashed", col=4)




## One-step ahead forecast using an AR(14) model#############################

# Test Data#################################################################
trend.season.forecast = as.numeric(Y.hat) + as.numeric(seasonal_forecast)

window.plot = data.frame(Beggining = as.Date("2000-12-01"), End = as.Date("2001-11-01"))
window.index = employment.df$date >= window.plot$Beggining &
employment.df$date <= window.plot$End
```

```
677
678 test = ts(employment.df[window.index,2]-trend.season.forecast,
679 start = c(2000,12),end = c(2001,11),frequency = 12)
680
681 # One-step-forecasts
682 fit <- Arima(residual,order = c(14,0,0),include.mean = FALSE)
683 fit_test <- Arima(c(residual,test), model=fit)
684 onestep <- fitted(fit_test)[309:320] + trend.season.forecast
685 onestep = ts(onestep, start = c(2000,12), end = c(2001,11), frequency = 12)
686
687 # plot
688 ts.plot(dataplot, onestep, col=1:2,
689 ylab="Employment/Population",ylim = c(62,66))
```