Time Series Analysis of Non-Farm Payroll Data - Report

Sam Choi, Eric Xu

4/17/2018

Introduction

The Non Farm Payroll (NFP) Report is a monthly government report tracking the number of payrolls in the U.S., excluding farm workers, private household employees, and non-profit organization employees. It is collected and reported by the the U.S. Bureau of Labor Statistics (BLS), and is an influential economic indicator used to analyze the state of the labor market.

The release of monthly NFP Reports has significant ramifications for economic policy makers and market participants. Thus, it is imperative for both economists and traders to have a general forecast of the next month's NFP Report in order to prepare for the volatility that ensues with the release of the report.

The monthly data is collected by the BLS as a component of the Current Employment Statistics (CES) survey. About a third of all NFP workers are surveyed, and the value reported is the result of additional extrapolation and adjustment.

The objective of this analysis is twofold: To determine the extent to which seasonal and non-seasonal trends can be captured using techniques in univariate time series analysis, and to accurately forecast future NFP Reports.

Data Analysis

Exploratory Data Analysis

We aim to explore PAYNSA, PAYEMS a dataset of non farm payrolls without seasonal adjustments. The data is recorded in thousands of persons, and contains monthly NFP report values from January 1939 to March 2018.

Throughout the 80 years represented in these datasets, various events have occurred that significantly changed the conditions of the economy. For this reason, we will limit our analysis to the years that followed the financial crisis of 2007/2008. By narrowing our scope to 2010-2018, we aim to provide a more telling analysis of the trends associated with the post-recession economic recovery. We first plot the not seasonally adjusted (NSA) data set.

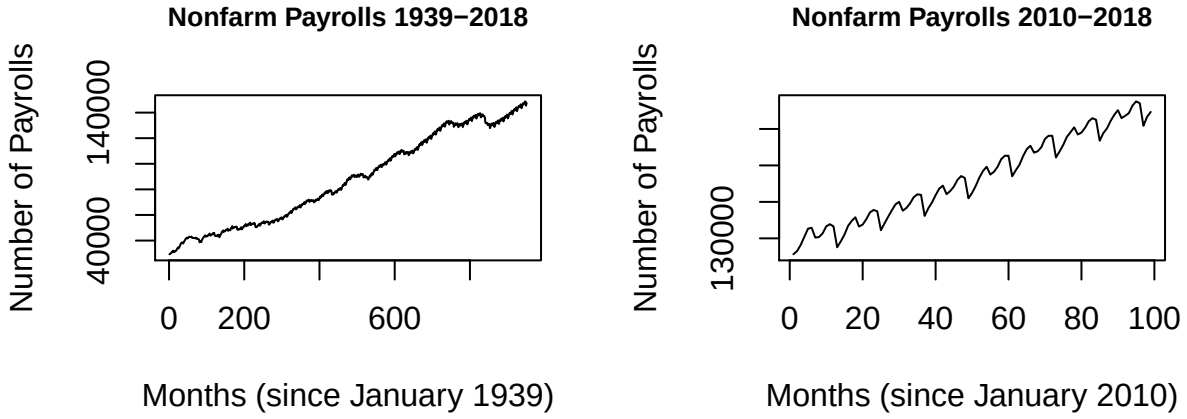

Figure 1. Time series plots of the full NSA data set (left) and the subset used for this analysis (right)

Next we take the first difference and study the distribution of differences using a histogram.

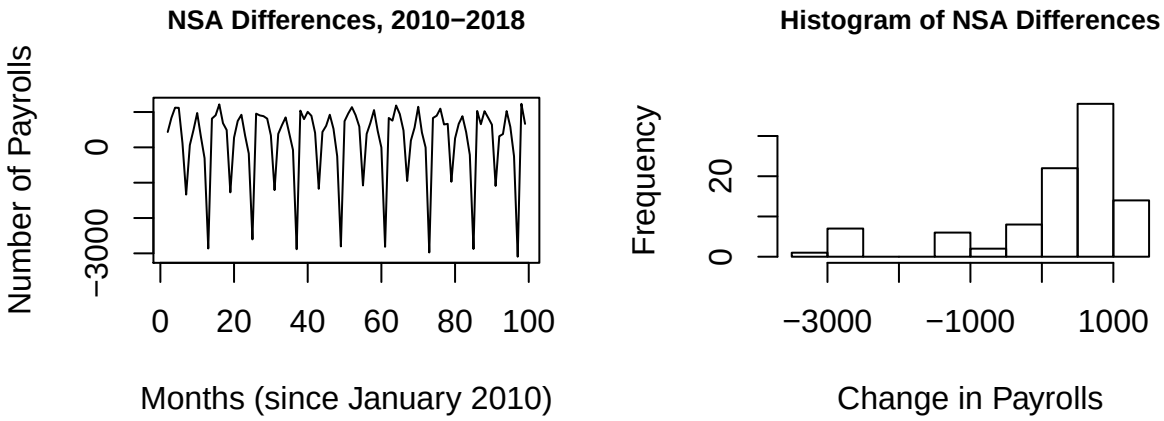

Figure 2. Differenced NSA data and distribution of change in payrolls

Taking the first difference appears to have detrended the original time series, and plotting the histogram indicates that the change in total payrolls may follow a left-skewed distribution, characterized by the long left tail above.

Next we mean-center the differenced NSA time series:

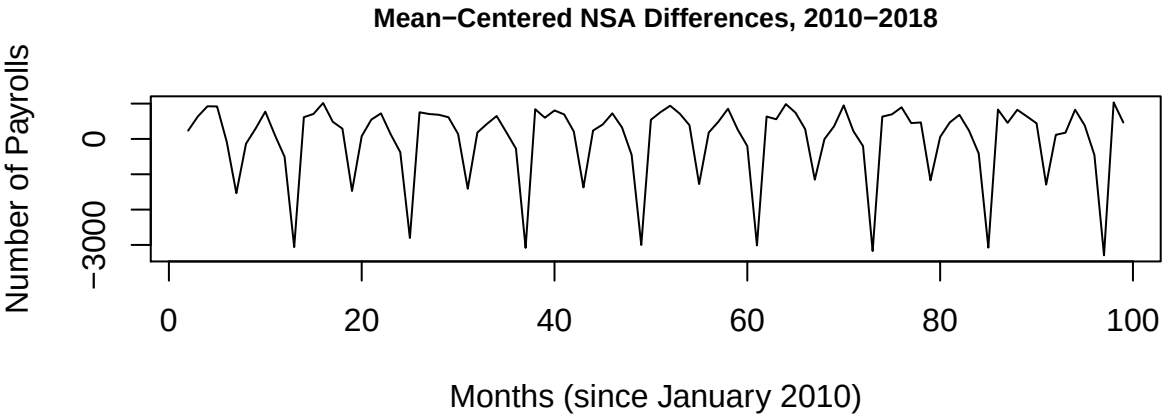

Figure 3. Mean-centered NSA time series

The resulting time series above shows signs of stationarity: The plot appears to have a consistent mean and variance with respect to time.

ARIMA Model Building

We will use the Box-Jenkins method to build an ARIMA model for the non-seasonally adjusted (NSA) NFP data. We begin by transforming the data set to achieve stationarity.

Transforming the dataset to achieve stationarity

We begin by transforming the data set to achieve stationarity. We do so by first using a lagged difference with a period of 12 months, producing a time series that resembles a random walk. We then apply differencing again to achieve stationarity.

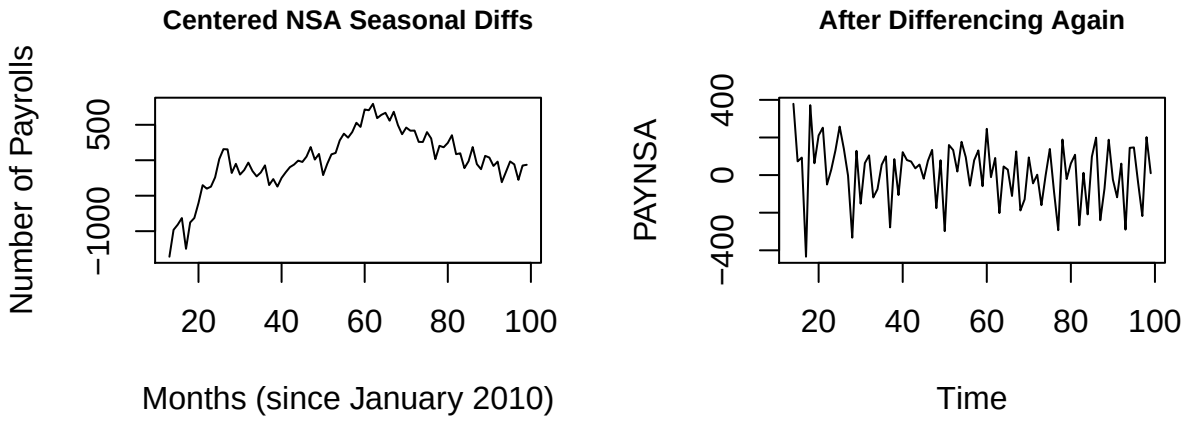

Figure 4. NSA data after lagged difference (left) and after lagged difference plus differencing again (right)

This now resembles white noise with fairly consistent variance, with slightly higher variance in the first 20 months of the data.

Model selection

We now inspect the ACF and PACF plots of the transformed time series.

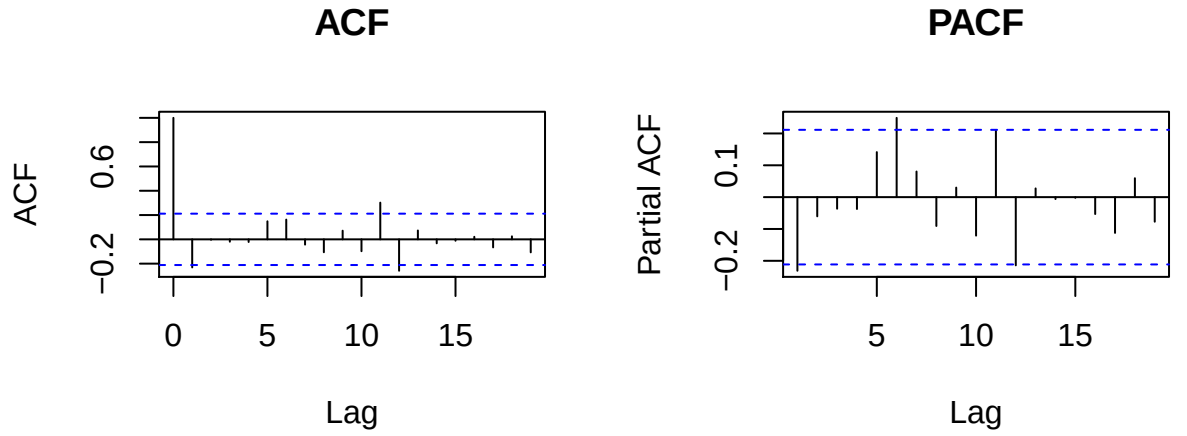

Figure 5. ACF and PACF plots of the transformed time series

Neither the ACF nor PACF appear to be significant at any lag. This further supports the notion that our transformed time series is very close to white noise, or an ARMA(0, 0) model. Therefore, fitting any non-zero order ARMA model and proceeding with parameter estimation will likely produce a model with poor predictive properties including bias. Therefore, instead of fitting an ARMA model we turn to spectral analysis to analyze the seasonal variation of our time series from a frequency perspective.

Spectral Analysis

Analyzing the time series using spectral analysis allows us to identify the key frequencies of variation in the non-seasonally adjusted NFP time series.

We begin by transforming the dataset into a stationary time series. We will compare the estimated spectral densities for two methods of transformation: differencing and curve-fitting. We will refer to the time series transformed by curve-fitting as the detrended time series.

Nonparametric Spectral Estimation

We begin by generating raw periodograms for both the differenced and detrended time series. These can be viewed in Figure A in the Appendix. As expected, both periodograms are bumpy with similar peaks. To reduce the variance of the periodogram and better identify significant peaks, we will apply smoothing to the periodogram using various parameters to find a better estimate of the spectral density. We begin with a daniell kernel with m = 1.

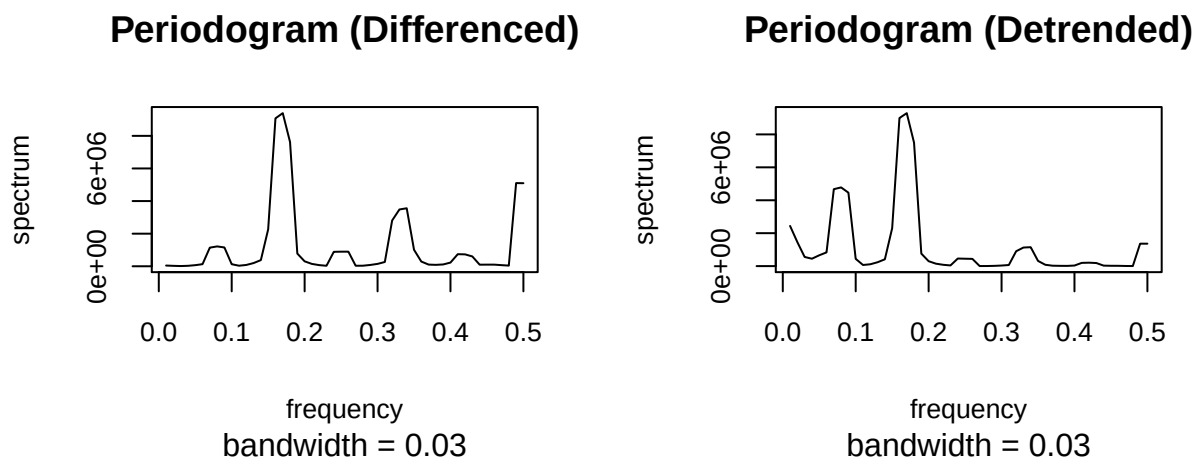

Figure 6. Smoothed periodograms of differenced and detrended time series using a Daniell kernel (m = 1)

This looks quite smooth already. Using m = 2 results in wider peaks shown in Figure B in the Appendix, possibly introducing bias. Therefore, setting m = 1 is an adequate amount of smoothing that allows us to identify significant peaks while minimizing bias.

To further reduce bias and narrow the peaks, we can apply smoothing with the modified Daniell kernel with m = 1. The periodogram is very smooth, so tapering has little effect. This suggests our data set has very clean and predictable cyclical variation, and we can now find the frequencies of the most significant peaks. We set the taper to 0.1 and overlay the local maxima to show where key frequencies might be.

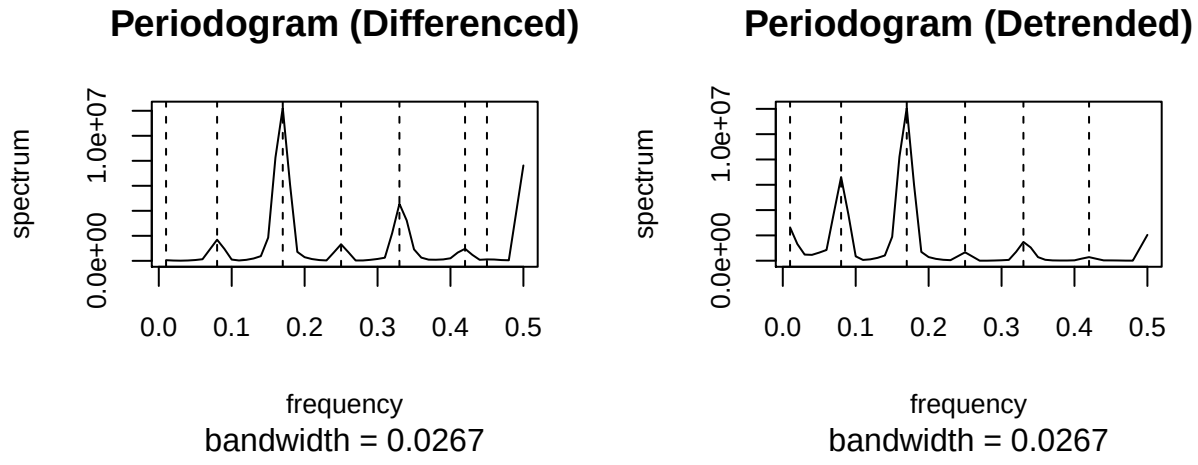

Figure 7. Smoothed and tapered periodograms overlaid with local maxima (dashed lines)

Smoothing using the modified Daniell kernel produces much sharper peaks. There does not appear to be much bumpiness remaining so tapering should not be necessary. To see the smoothed and tapered periodogram, see Figure C in the Appendix.

In the differenced time series, we can identify three major peaks: one at a frequency of about 0.17, corresponding to a semiannual cycle, one at a frequency of about 0.5, corresponding to a two-month cycle, and one at a frequency of about 0.33, corresponding to a quarterly cycle.

The smoothed periodogram of the detrended data has two major peaks: one at a frequency of about 0.08, corresponding to an annual cycle, the largest at a frequency of about 0.17, corresponding to a semiannual cycle, and one at a frequency of 0.08, corresponding to an annual cycle. Two other, less significant peaks appear, one at 0.01 and one at 0.5.

Parametric Spectral Estimation

We can now compare our smoothed and tapered periodograms with the estimated spectral densities generated using parametric spectral estimation.

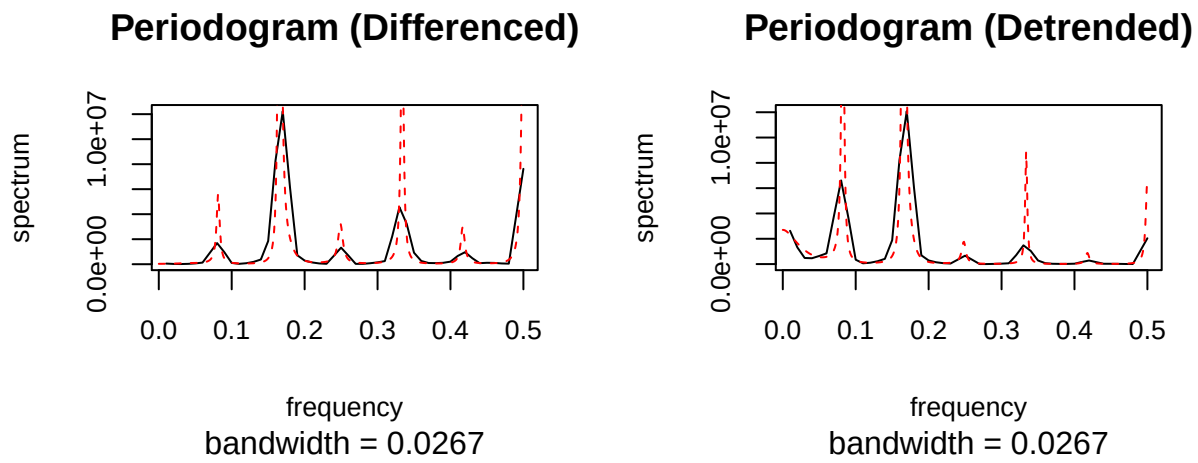

Figure 8. Smoothed and tapered periodograms overlaid with parametric spectral estimator (dashed red)

The parametric spectral estimator has peaks at frequencies 0.08, 0.17, 0.25, 0.33, 0.42, and 0.50. This matches the frequencies of peaks seen in the differenced and detrended time series. However, the size of the peaks at these frequencies differs between the two time series. The peaks have similar magnitude between the parametric and non parametric methods, suggesting that the peaks are indeed signficant.

Forecasting

Using data from January 2010 to March 2017, we aim to forecast nonfarm payrolls for the following 12 months (April 2017 to March 2018). We do this by partitioning our dataset into a training portion that will be used to fit our models, and a testing portion that will be used to measure the accuracy of our forecasts.

ARIMA

We proceed with a non-seasonal ARIMA model, as well as a SARIMA model to fit the general trend of the data, as well as the seasonal trends observed:

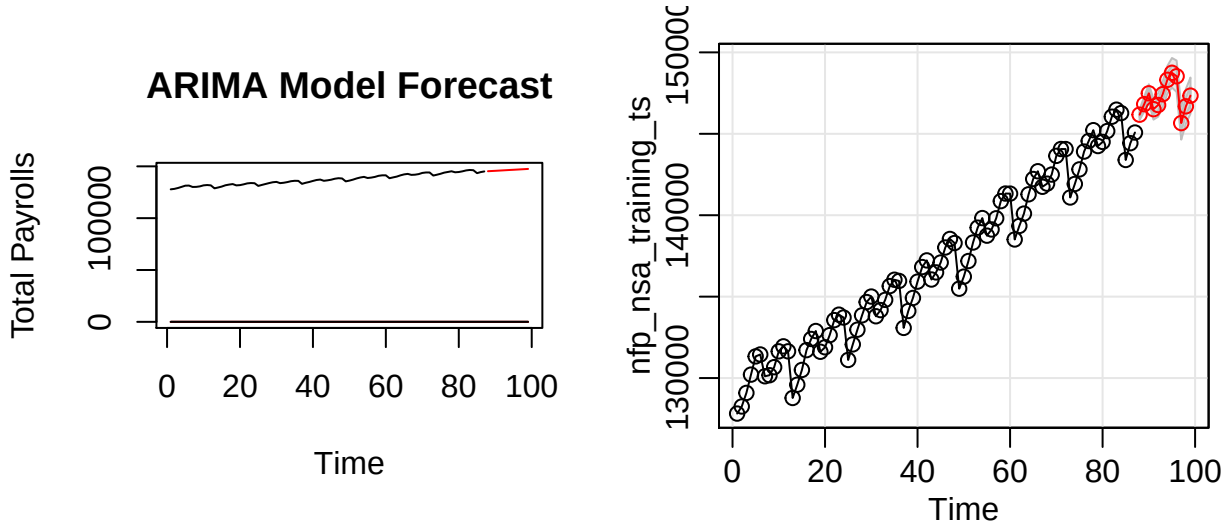

Figure 9. ARIMA model forecast (left) and SARIMA model forecast (right)

We can see that the first graph captures the general direction of increasing payrolls, while the second graph also captures the seasonal behavior of payroll increases and decreases.

We follow up by computing the mean-squared error for both models, and find that the ARIMA forecasts result in a MSE of 1774426, while the SARIMA forecast results in a MSE of 13108. This shows that the SARIMA forecast has not only captured the seasonal trend, but has led to a significantly lower error term (~130 times lower MSE).

Conclusion

Key Takeaways

Our spectral analysis shows that jobs are added and lost according to a highly consistent seasonal pattern. The key frequencies of variation are highly significant in the smoothed periodograms and align with the variation in the time series plot. This is further supported by our process of chasing stationarity, in which a combination of seasonal and regular differencing resulted in a time series resembling white noise. Additionally, our SARIMA model forecast achieves a low MSE in predicting payroll cycles over time, meaning the seasonal trend is consistent.

Through the EDA process, we have seen that job growth from 2010 to 2018 has been very closely linear. This trend aligns with GDP and population growth rate, which have stayed fairly flat over the same time period.

Interesting Findings

Cycles of variation every quarter, six months, and each year align with the seasons. This means that the season is a key predictor of payroll performance, providing support for the idea that many financial markets are closely tied to the time of year.

Differencing and detrending the NSA data using curve-fitting resulted in different key frequencies being highlighted as most significant by the spectral estimation process. Looking at both allowed and comparing with parametric spectral estimation allowed us to determine the three most significant frequencies of variation.

Future Analysis

Our analysis could be extended by widening our scope of analysis to include more data while adjusting for outliers such as those caused by recessions. This would allow us to analyze data in periods where growth was not so closely linear.

Multivariate time series analysis comparing payroll data with factors such as GDP growth, population growth, education levels, etc. could help reveal dependencies between job creation over longer time scales and economic and demographic attributes, helping determine what policy decisions improve employment rates.

Appendix

Figures

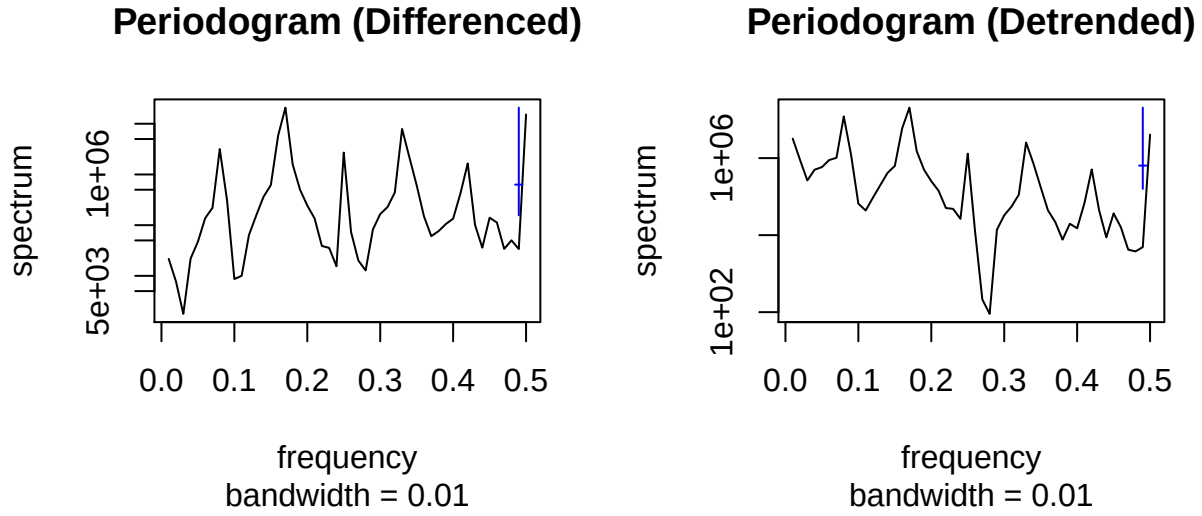

Figure A. Raw periodograms of differenced (left) and detrended (right) NSA data

The differenced time series produces a raw periodogram with a much lower baseline in the 0.0 to 0.1 frequency range and the detrended time series produces a raw periodogram with a lower baseline in the 0.26 to 0.28 frequency range. This means in the differenced time series, the peak at a frequency of about 0.08 is more significant, and in the detrended time series, the peaks at 0.25 and 0.33 are more significant.

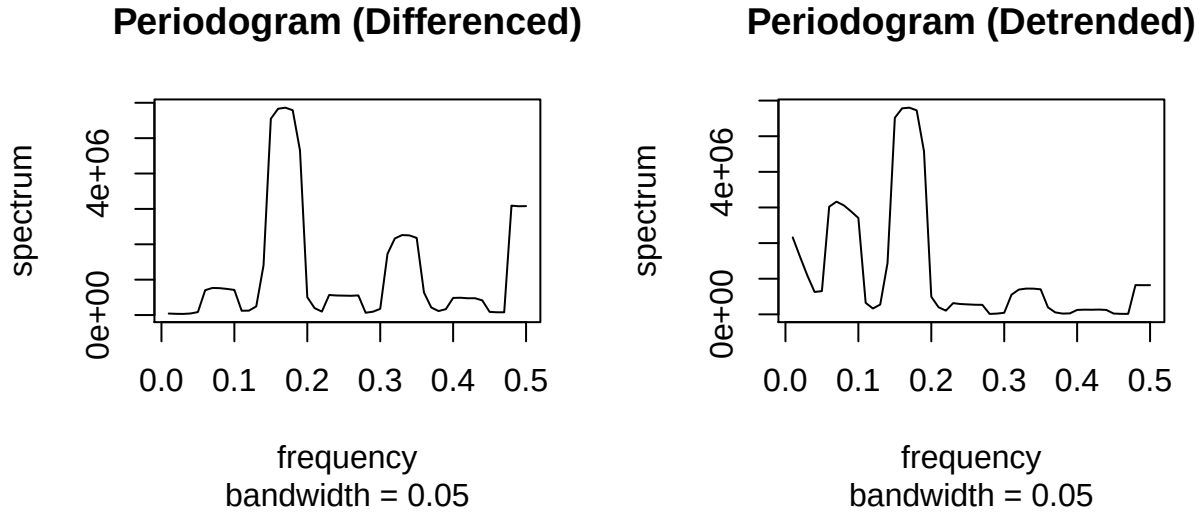

Figure B. Smoothed periodograms of differenced and detrended time series using a Daniell kernel (m = 2)

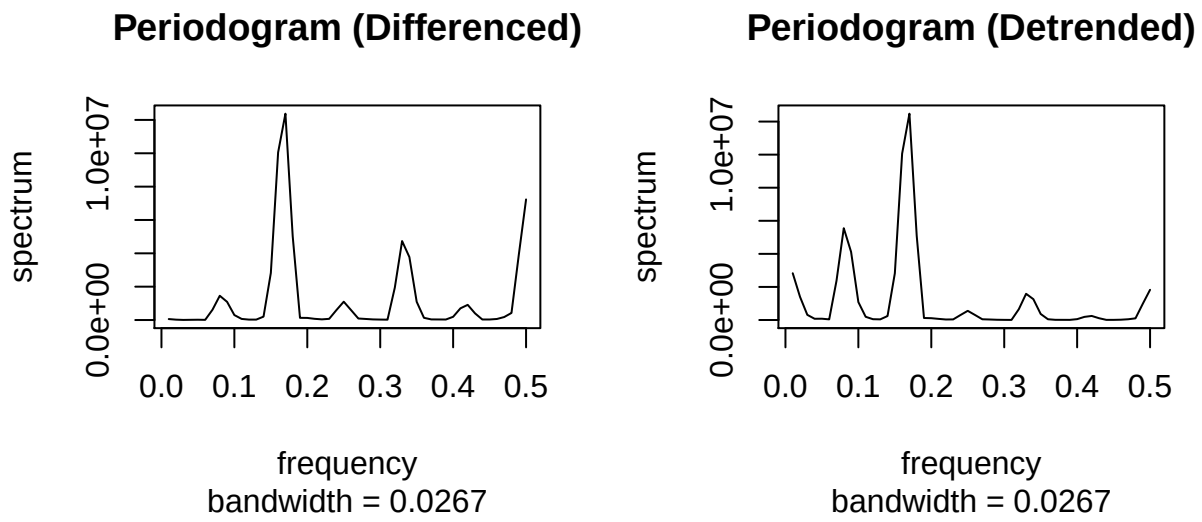

Figure C. Smoothed and tapered periodograms of differenced and detrended time series using a modified Daniell kernel (m = 1)

Code

EDA

```
PAYNSA <- read.csv(file = "PAYNSA.csv", header = TRUE, sep = ",")
nfp_nsa_ts <- ts(PAYNSA[2])
plot.ts(nfp_nsa_ts, main = "Total Nonfarm Payrolls (Not Seasonally Adjusted)",
        xlab = "Months (since January 1939)", ylab = "Number of Payrolls")

nfp_nsa_ts_2010_2018 <- ts(PAYNSA[853:951, ][2])
plot.ts(nfp_nsa_ts_2010_2018, main = "Total Nonfarm Payrolls (NSA), 2010-2018",
        xlab = "Months (since January 2010)", ylab = "Number of Payrolls")

NSA_diff <- diff(nfp_nsa_ts_2010_2018, lag = 1, differences = 1)
plot.ts(NSA_diff, main = "NSA Differences, 2010-2018", xlab = "Months (since January 2010)",
        ylab = "Number of Payrolls")

hist(as.numeric(unlist(NSA_diff)), breaks=15,  main = "Histogram of NSA Payroll changes",
     xlab = "Number of Payrolls")

NSA_mean <- mean(NSA_diff, na.rm = TRUE)
NSA_mean
centered_NSA_diff <- NSA_diff - SA_mean
plot.ts(centered_NSA_diff, main = "Centered NSA Diffs, 2010-2018",
        xlab = "Months (since January 2010)", ylab = "Number of Payrolls")
```

ARIMA Modeling

```
PAYNSA <- read.csv(file = "PAYNSA.csv", header = TRUE, sep = ",")
nfp_nsa_ts_2010_2018 <- ts(PAYNSA[853:951, ][2])
plot.ts(nfp_nsa_ts_2010_2018, main = "Non Seasonally Adjusted NFP Data")

NSA_sdiff <- diff(nfp_nsa_ts, lag = 12, differences = 1)
NSA_smean <- mean(NSA_sdiff)
centered_NSA_sdiff <- NSA_sdiff - NSA_smean
plot.ts(centered_NSA_sdiff, main = "Centered NSA Seasonal DIffs, 2010-2018",
        xlab = "Months (since January 2010)", ylab = "Number of Payrolls")
```

```r
NSA_sdiff2 <- diff(NSA_sdiff, lag = 1, differences = 1)
plot.ts(NSA_sdiff2, main = "NSA Seasonal Diffs Differenced Again")

acf(NSA_sdiff2, main = "ACF Plot of NSA Seasonal Diffs Differenced Again")
pacf(NSA_sdiff2, main = "PACF Plot of NSA Seasonal Diffs Differenced Again")
```

Spectral Analysis

```r
mvspec(differenced_NSA, detrend = FALSE)
mvspec(nfp_nsa_ts_2010_2018)

pgram_diffed <- mvspec(differenced_NSA, kernel("modified.daniell", 1), log = "no")
pgram_detrend <- mvspec(nfp_nsa_ts_2010_2018, kernel("modified.daniell", 1), log = "no")

mvspec(differenced_NSA, kernel("daniell", 1), log = "no")
mvspec(nfp_nsa_ts_2010_2018, kernel("daniell", 1), log = "no")

mvspec(differenced_NSA, kernel("daniell", 2), log = "no")
mvspec(nfp_nsa_ts_2010_2018, kernel("daniell", 2), log = "no")

mvspec(differenced_NSA, kernel("modified.daniell", 1), log = "no")
mvspec(nfp_nsa_ts_2010_2018, kernel("modified.daniell", 1), log = "no")

mvspec(nfp_nsa_ts_2010_2018, kernel("modified.daniell", 1), log = "no", taper = 0.2)
mvspec(differenced_NSA, kernel("modified.daniell", 1), log = "no", taper = 0.2)

key_freq_ind <- c(1, which(diff(sign(diff(pgram_detrend$spec)))==-2) + 1)
key_freq <- pgram_detrend$freq[key_freq_ind]
abline(v=key_freq, lty=2)

key_freq_ind <- c(1, which(diff(sign(diff(pgram_diffed$spec)))==-2) + 1)
key_freq <- pgram_diffed$freq[key_freq_ind]
abline(v=key_freq, lty=2)

pgram_ar <- spec.ar(differenced_NSA, plot=F)
lines(pgram_ar$freq, pgram_ar$spec, lty=2, col="red")

mvspec(differenced_NSA, kernel("modified.daniell", 1), log = "no",
       main = "Periodogram (Differenced)")
pgram_ar <- spec.ar(differenced_NSA, plot=F)
lines(pgram_ar$freq, pgram_ar$spec, lty=2, col="red")

t <- 1:length(nfp_nsa_ts_2010_2018)
fit <- lm(nfp_nsa_ts_2010_2018 ~ t)
detrended_NSA <- fit$residuals

mvspec(nfp_nsa_ts_2010_2018, kernel("modified.daniell", 1), log = "no",
       main = "Periodogram (Detrended)")
pgram_ar <- spec.ar(detrended_NSA, plot=F)
lines(pgram_ar$freq, pgram_ar$spec, lty=2, col="red")
```

Forecasting

```r
PAYNSA <- read.csv(file = "PAYNSA.csv", header = TRUE, sep = ",")
nfp_nsa_ts <- ts(PAYNSA[2])
nfp_nsa_training_ts <- ts(PAYNSA[853:939,][2])
```

```r
nfp_nsa_testing_ts <- ts(PAYNSA[940:951,][2])

nobs <- length(nfp_nsa_training_ts)
fit <- arima(nfp_nsa_training_ts, order=c(0, 1, 0), xreg=1:nobs)
n_pred <- length(nfp_nsa_testing_ts)
forecast <- predict(fit, n_predictions, newxreg=(nobs+1):(nobs+n_pred))

mse <- function(ts1, ts2) {
  sum_squares = 0
  for (i in 1:length(ts1)) {
    diff <- ts1[i] - ts2[i]
    sum_squares = sum_squares + diff^2
  }
  sum_squares/length(ts1)
}

par(mfrow=c(2,1))
ts.plot(nfp_nsa_training_ts, forecast$pred)
sfit <- sarima.for(nfp_nsa_training_ts, n_pred, p=0, d=1, q=0, P=0, D=1, Q=0, S=12)

mse(forecast$pred, nfp_nsa_testing_ts)
mse(sfit$pred, nfp_nsa_testing_ts)
```