The Rain in Davis

Rylan Schaeffer and Aaron Shuler

March 9, 2015

"Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?
Where is the information we have lost in data?
With apologies to T.S. Eliot"

With apologies to Patrice Koehl

Abstract

This project examines daily precipitation in Davis, California, as measured by the experimental farm weather station run by the University of California at Davis. Specifically, this article looks at predicting precipitation by means of regression analysis over large time scales, seasonality, and day-to-day interactions. Various statistical measures are used to ensure significance of the results. This article concludes that, while there appears to be no long term trend in precipitation rates in Davis, seasonality and lag terms can be used to predict future patterns in rainfall.

Introduction

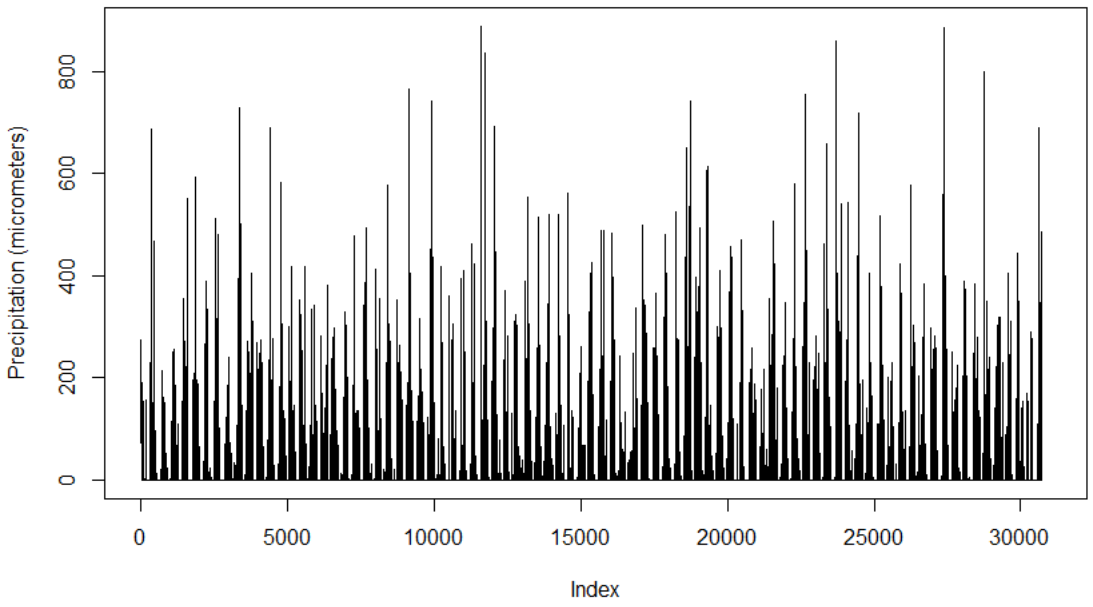
In the 1956 musical, "My Fair Lady," Cockney flower girl Eliza Doolittle struggles to overcome her heavy accent through speech exercises. She fails repeatedly to successfully pronounce the simple phrase, "The rain in Spain stays mainly on the plain," until eventually, at the limits of her endurance, Eliza finally succeeds in mastering the exercise. Just like Eliza, we chose to use rain to demonstrate competency over the subject we had been struggling with: time series analysis.

Our task was to predict daily rainfall (precipitation) in Davis, California. An agricultural research powerhouse, UC Davis performs numerous agriculture experiments for which rainfall plays an critical role. California is currently experiencing its worst drought in over 1200 years (Nuccitelli), a drought so bad that UC Davis researchers (Howitt) estimate has cost California $2.2 billion, 17,000 seasonal jobs and the loss of 5% of the state's irrigated cropland in 2014 alone. UC Davis leads the state and nation in analysis and policy recommendations regarding the drought. UC Davis also created a website (http://drought.ucdavis.edu/) to track all news, research and events related to water shortages. In response, last September, the Davis City Council voted to enact a Stage 3 water shortage emergency (City of Davis), requiring a 30% reduction in water use. The issue of water in our state is critical, and so is predicting when rain will next fall. Rather than predicting statewide rainfall, we limited ourselves to the more manageable task of rainfall in Davis. Predicting rainfall in Davis will give the campus and the town the ability to adjust behavior (i.e. watering crops) accordingly.

Data Description

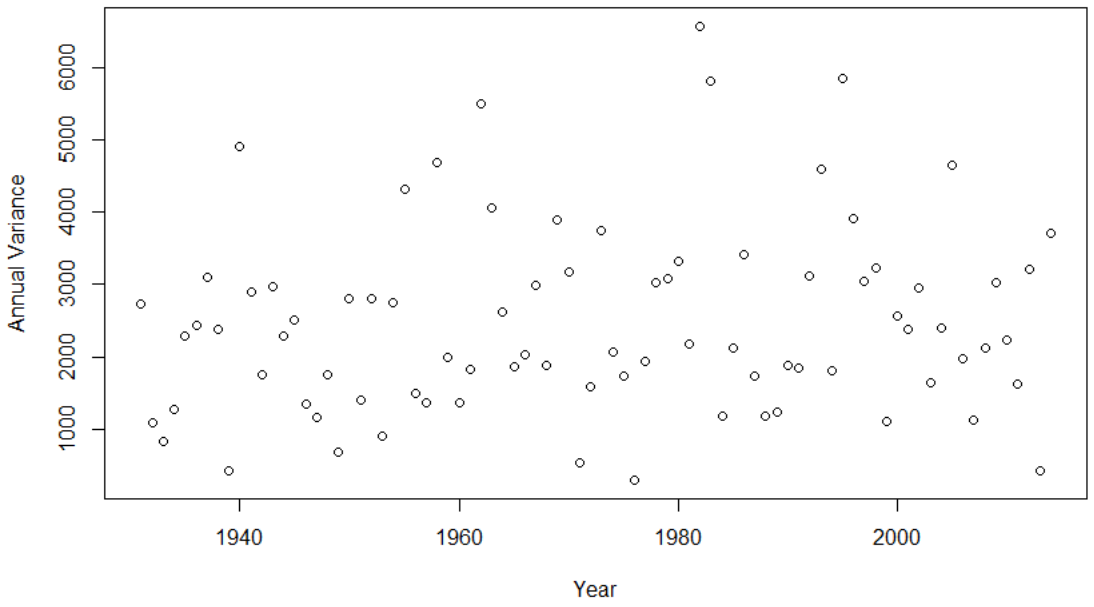
The data consists of daily measurements of precipitation (in micrometers) in Davis. The data comes from the National Oceanic and Atmospheric Administration, which aggregates data from five different sites

in Davis. Since four of these five sites are relatively new, beginning in 2008, 2010, 2010 and 2012, we limited ourselves to data from the oldest site, labeled DAVIS 2 WSW EXPERIMENTAL FARM CA US, which has been measuring daily precipitation since January 1st, 1931. This particular site has 99% coverage from January 1st, 1931 to present, meaning only 1% of daily observations are missing; we identified the missing days and added them to the time series as NA values. With over 30,000 daily observations from over 84 continuous years gathered, this is what our data looked like.

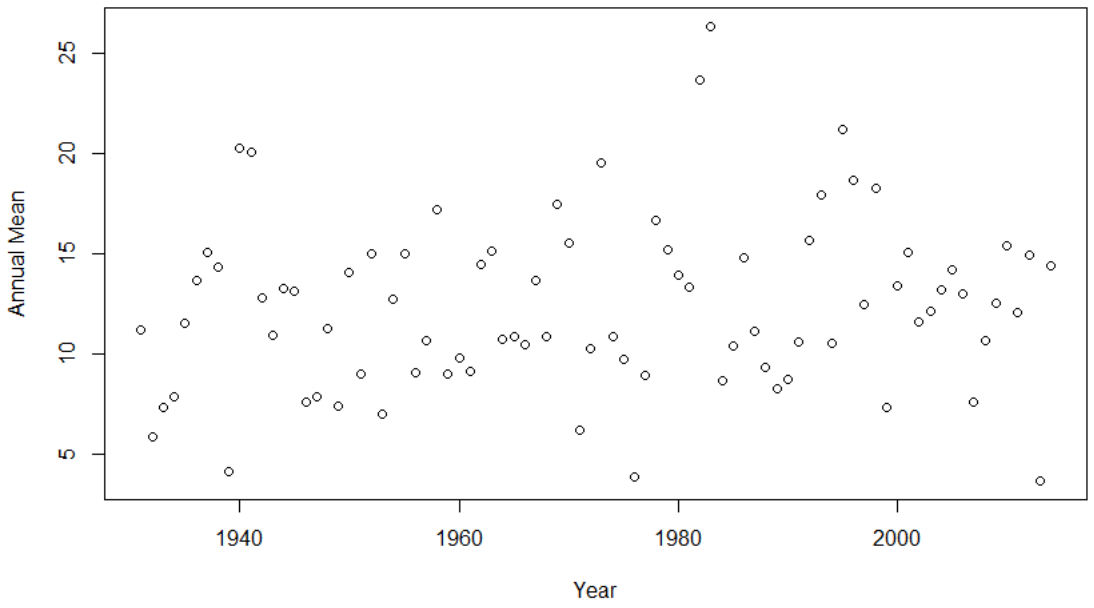


Time-Domain Analysis

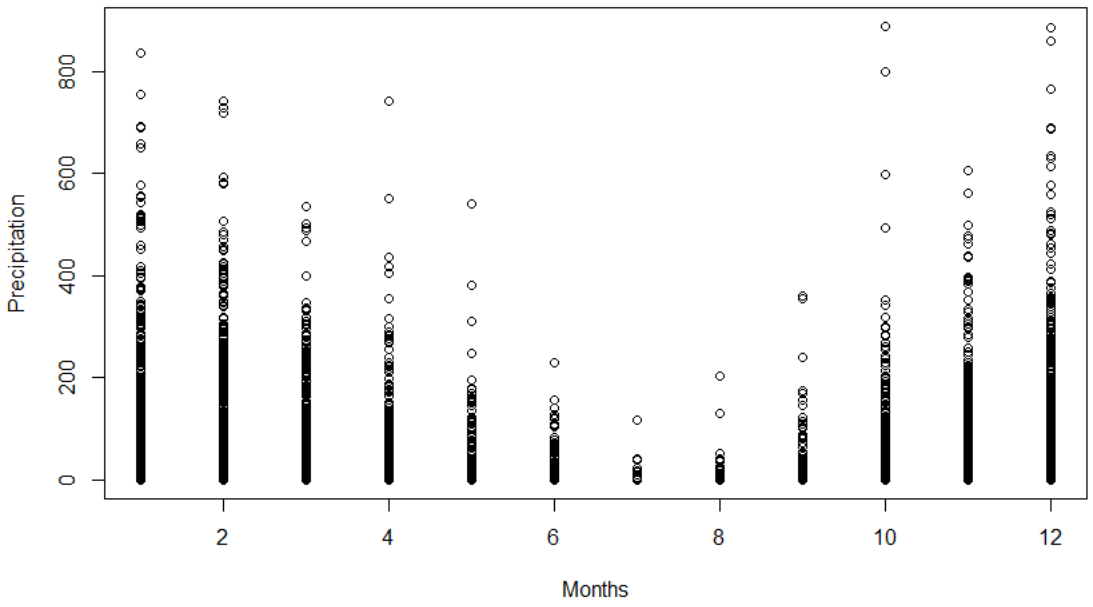
First, we checked whether the variance of the data changes with respect to time. The following is a plot of the variance of daily precipitation in a given year. The line of best fit had a p-value of 0.2415 and $R^2_{adj} = 0.004698$, meaning that the variance is likely not changing over time. This means we do not need to apply a Box-Cox transformation to our data.



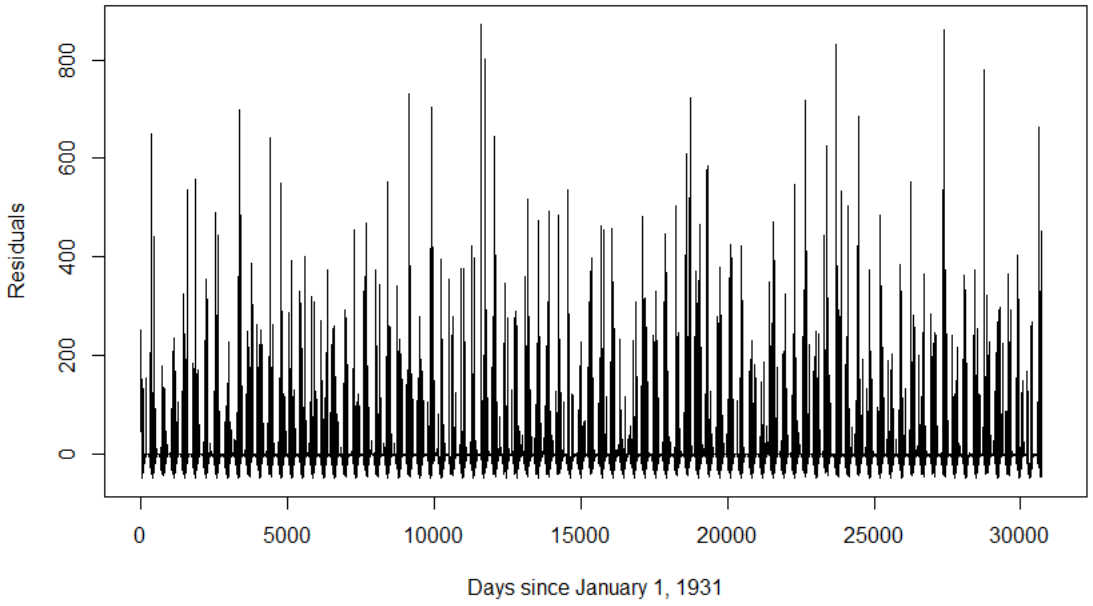
Second, we tested to see whether a mean component (a trend) exists in our data. The following is a plot of the mean value of daily precipitation in a given year. The line of best fit had a slope of 0.02406, a p-value of 0.2132 and $R^2_{adj} = 0.006866$, meaning that the expected value of annual rainfall in Davis is likely not changing over time i.e. the average rainfall per year is approximately constant.



Third, we examined the plot of daily rainfall by month, where month 1 represents January and month 12 represents December, and noticed a clear seasonal component: winter months are more likely to have more precipitation than summer months.

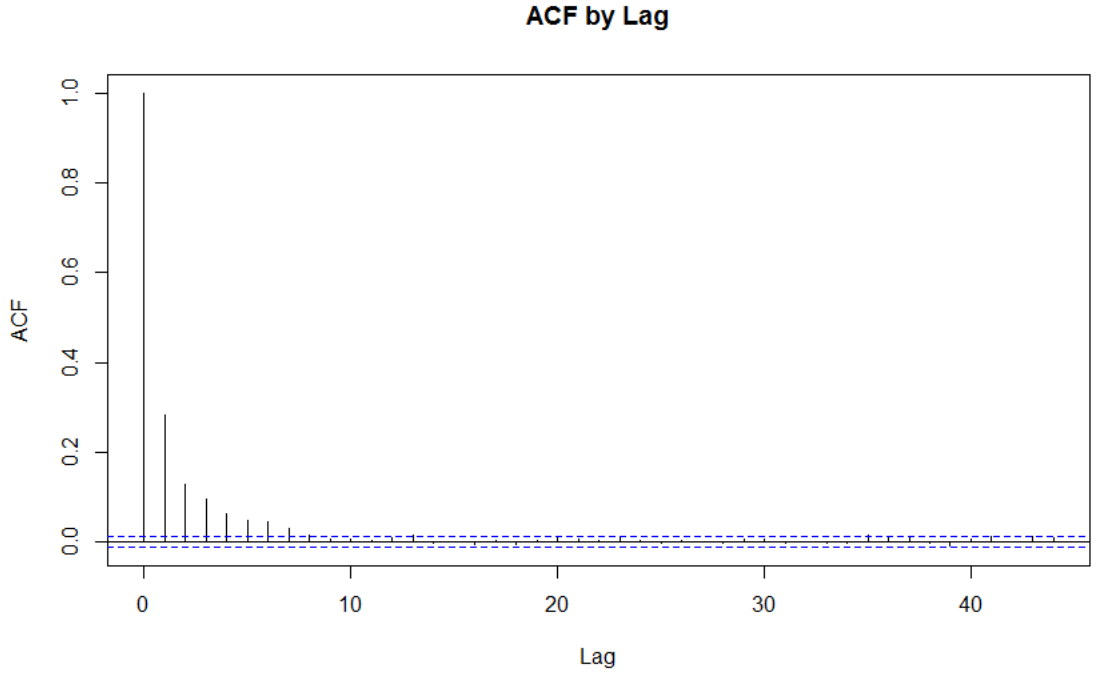


Hence, the trend and seasonal components need to be estimated and removed. To do so, we used the small trend estimation method by formulating the time series as a two dimensional matrix of 85 years (rows) by 365 days (columns). We then subtracted the row mean and column mean from each observation. The following is a plot of the detrended, deseasonalized data (the residuals).

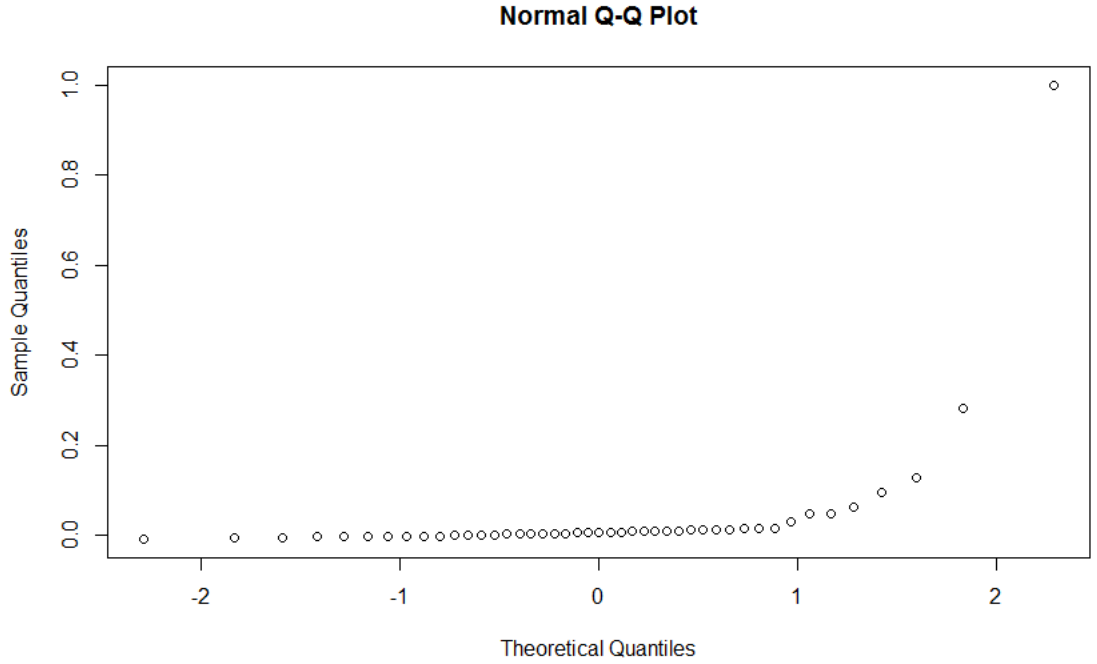


At this point, we needed to determine whether the residuals are independent and identically distributed (i.i.d.) random variables because if they are, then no dependence exists and nothing more can be said regarding this data. If the residuals are i.i.d., then the sample Autocorrelation Function (ACF) should be approximately normal distributed with zero mean and variance $1/n$, where n is the sample size (Theorem

1.2.1). Looking at the ACF plot, $\hat{\rho}$ is centered at mean 0.0413. This value is relatively close to zero, so we consider the variance of $\hat{\rho}$ as well. $\text{Var}(\hat{\rho}) = 0.0236$, which is over 700 times larger than the approximate expected variance of 1/30000. This means that the sample ACF's variance is too large for the ACF to be distributed approximately normal.

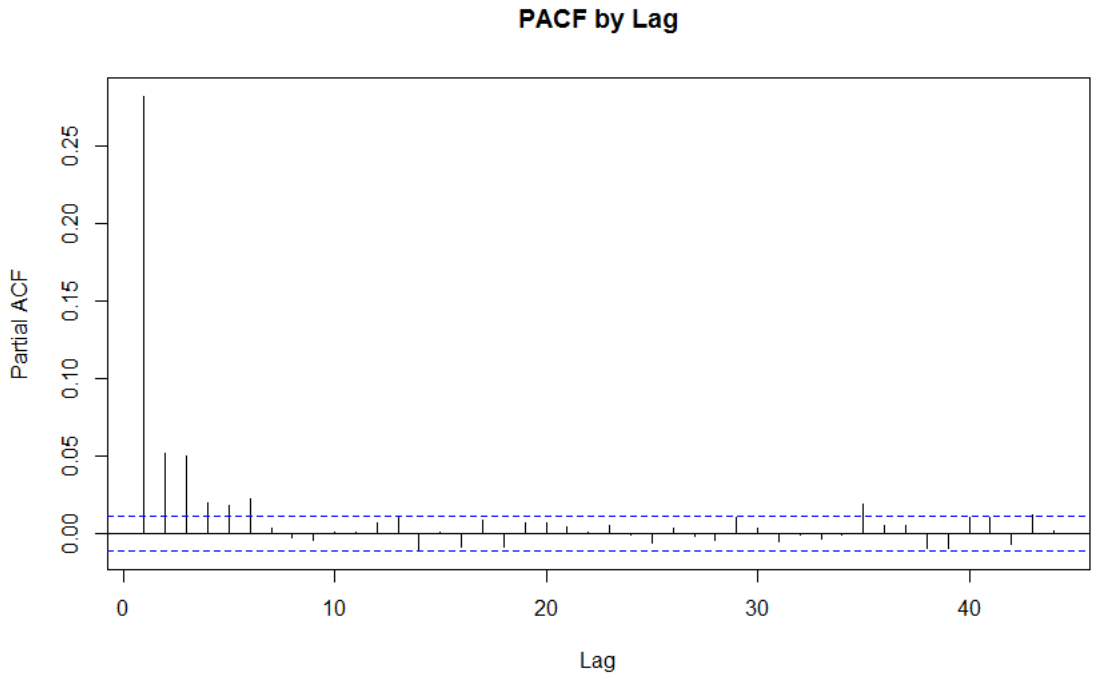


Using a QQ plot, we can see that the sample ACF values $\hat{\rho}$ do not follow a normal distribution. This means that the residuals are not i.i.d. and we may proceed with the analysis.



We next plotted the Partial Autocorrelation Function (PACF) to determine which time series model best fits the data we observe. Judging by the ACF and PACF plots, it is clear that an AR model would be

appropriate.



We also calculated the Akaike Information Criterion corrected (AIC_c) values for all autoregressive-moving-average processes (ARMA(p,q)), $0 \leq p \leq 7$, $0 \leq q \leq 7$. This would allow us to compare which of the sixty four ARMA processes has the lowest AIC_c value, and hence, is the best process to make predictions with. The resulting values are as follows:

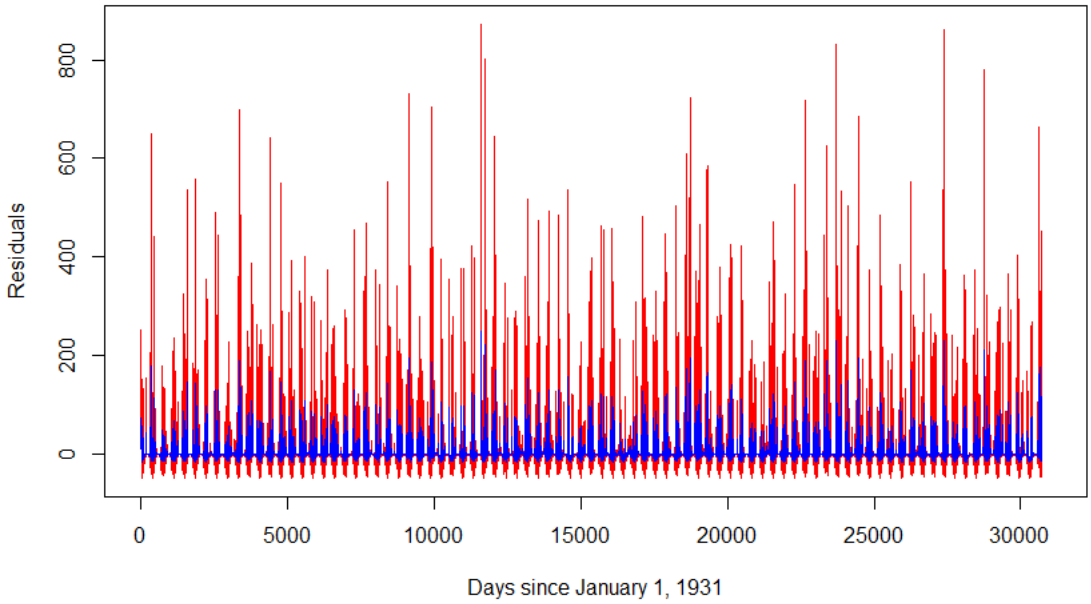
p\q	0	1	2	3	4	5	6	7
0	323979.7	321772.4	321513.8	321367.5	321322.4	321305.3	321271.9	321258.8
1	321438.5	321317.6	321253.4	321255.4	321256.0	321258.0	321253.9	321258.8
2	321358.3	321258.3	321255.3	321257.1	321259.2	321259.6	321257.1	321257.8
3	321283.1	321256.1	321257.1	321259.1	321259.7	321259.9	321259.8	321260.0
4	321273.4	321278.0	321259.5	321258.7	321260.3	321261.2	321261.0	321261.8
5	321264.9	321256.6	321260.7	321260.4	321263.3	321261.0	321262.9	321263.8
6	321251.9	321254.9	321257.1	321259.2	321260.3	321262.5	321263.6	321265.3
7	321253.5	321255.4	321257.4	321259.3	321261.3	321263.4	321265.1	321267.3

Time-Domain Discussion

The ACF plot appears to decay exponentially, suggesting an autoregressive (AR) process. The PACF plot agrees, cutting off somewhere between 3 and 6 lags, suggesting something between an AR(3) and AR(6) process. We were not sure whether the PACF values at lags 4, 5, and 6 are statistically significant due to how close they are to the boundary. To make a decision, we looked at the AIC_c values. The smallest AIC_c value of 321251.9 at p=6 and q=0 suggests an AR(6) process. The AIC_c value at p=3 and q=0 was 321283, which is slightly larger than the AIC_c value at p=6 and q=0, clarifying that the optimal model is the AR(6) process.

Using the calculated coefficients for ϕ_1 through ϕ_6, our AR(6) equation has the following form: $X_t = 0.2605X_{t-1} + 0.0340X_{t-2} + 0.0409X_{t-3} + 0.0116X_{t-4} + 0.0105X_{t-5} + 0.0192X_{t-6} + z_t$. This equation means that, excluding the mean and seasonal component, the amount of rainfall on a particular day depends on the rainfall from the previous six days. In general, the dependency of X_t on X_{t-h}, where t is the time index and h is the lag such that $1 \leq h \leq 6$, decreases as h moves from 1 to 6; in layman's terms, this means that the amount of rain yesterday provides more information regarding the amount of rain today than the amount

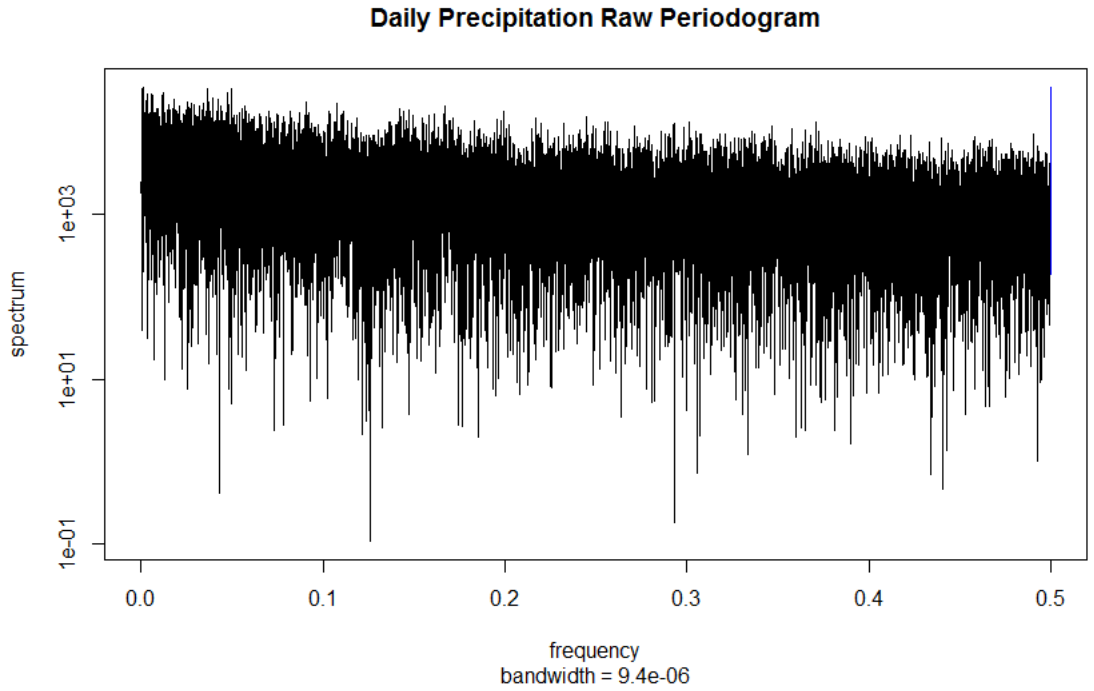
of rain from four, five or six days ago. This also means that the amount of rainfall seven or more days prior adds no new information to predicting the amount of rainfall on a given day beyond what information is provided by the six prior days. Plotted in blue against the red detrended, deseasonalized observations, this is what the model looks like.



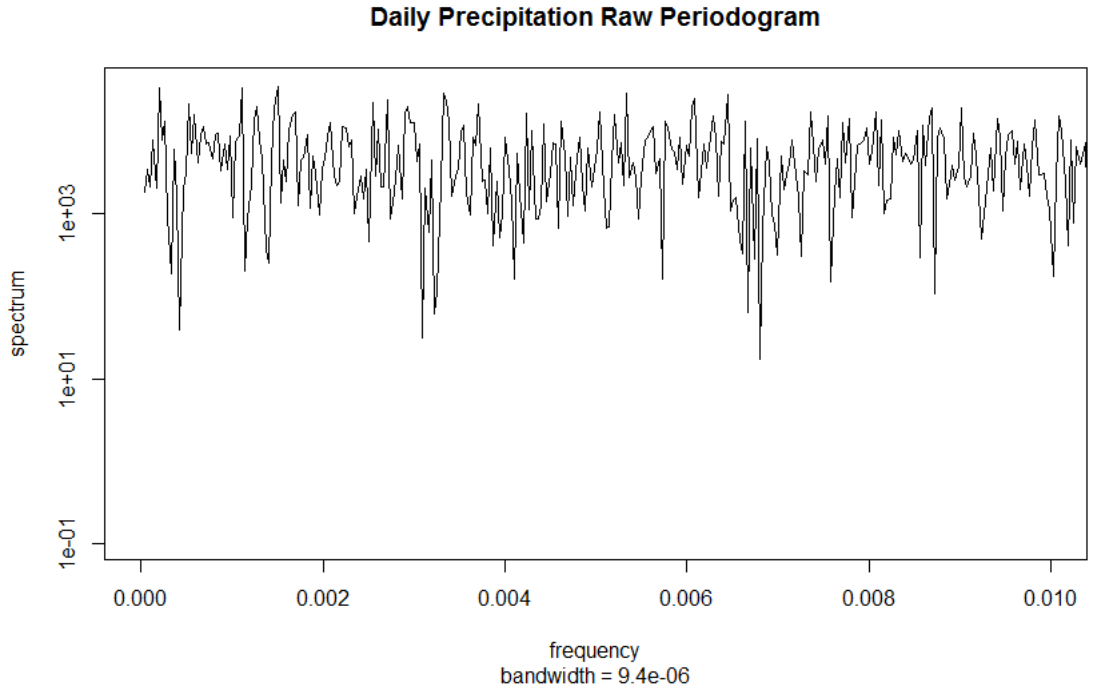
Interestingly, there are two small exception: ϕ_3 and ϕ_6. $\phi_3 = 0.0409$ is greater than $\phi_2 = 0.0340$, meaning that the amount of rain three days prior provides more information towards predicting the amount of rain today than does the amount of rain two days prior. Similarly, $\phi_6 = 0.0192$ is larger than $\phi_4 = 0.0116$ and $\phi_5 = 0.0105$, meaning that the amount of rain six days prior provides more information regarding the amount of rain today than the amount of rain from four days prior and from five days prior. As the quote at the beginning of our paper hints at, we can make no statement on why this is the case.

Frequency-Domain Analysis

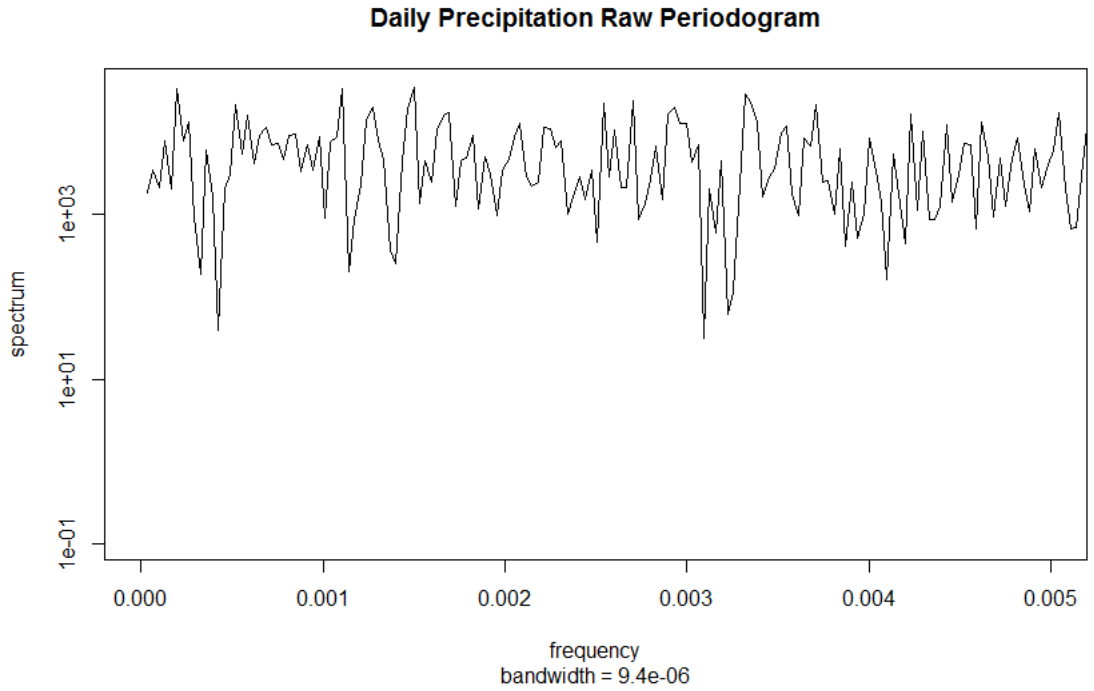
We decided to attempt analyzing our data set from a frequency-domain perspective in addition to a time-domain perspective. The first spectrum we plotted was meaningless, as there were too many peaks.



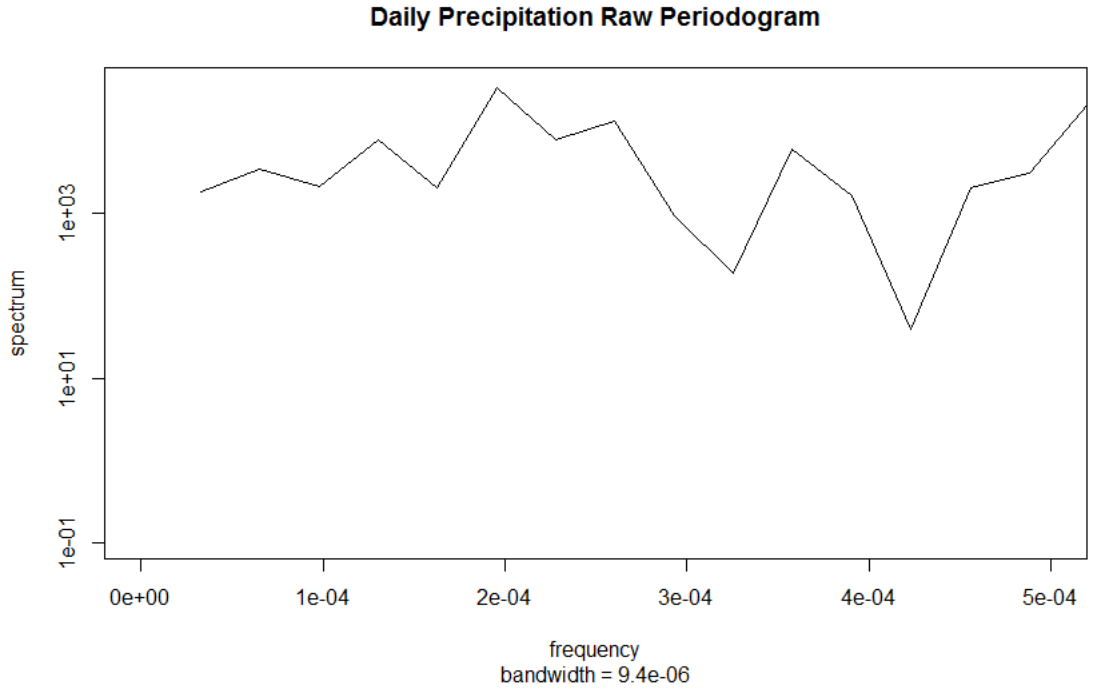
The reason why there are so many peaks is because the frequencies displayed range from $1/n$ to $1/2$. Since our sample size n is over 30,000, we were attempting to display over 30,000 periodograms! A frequency of 0.02 might seem small, but that frequency covers almost two years. We reduced the frequency window to 0.01, shown below.



Nothing stood out, so we again reduced the frequency window to 0.005.



Even zoomed in to a frequency window of 0.0005, no peaks stood out.

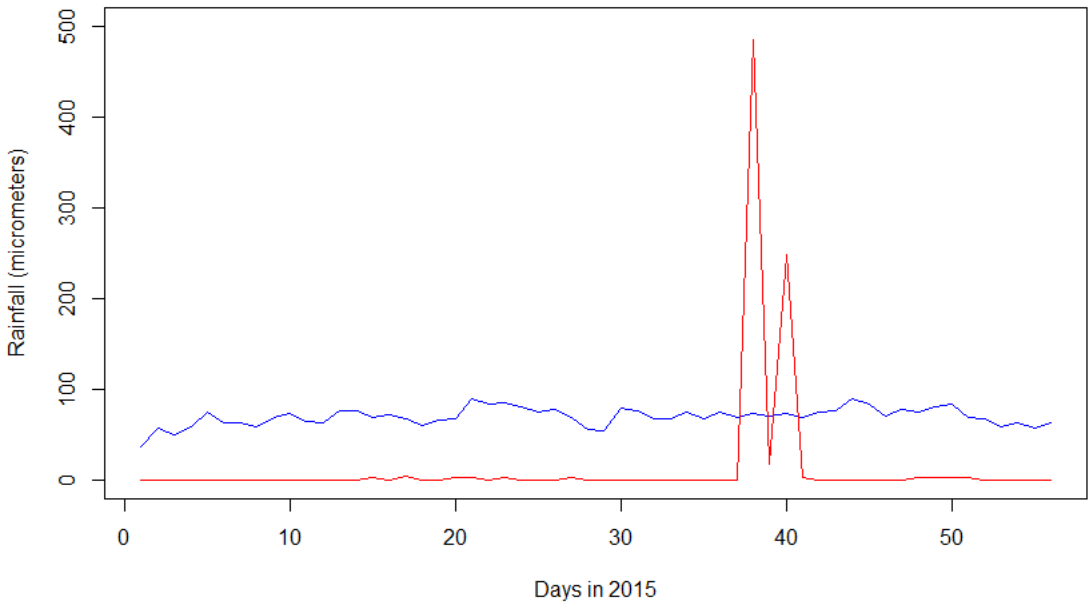


Frequency-Domain Discussion

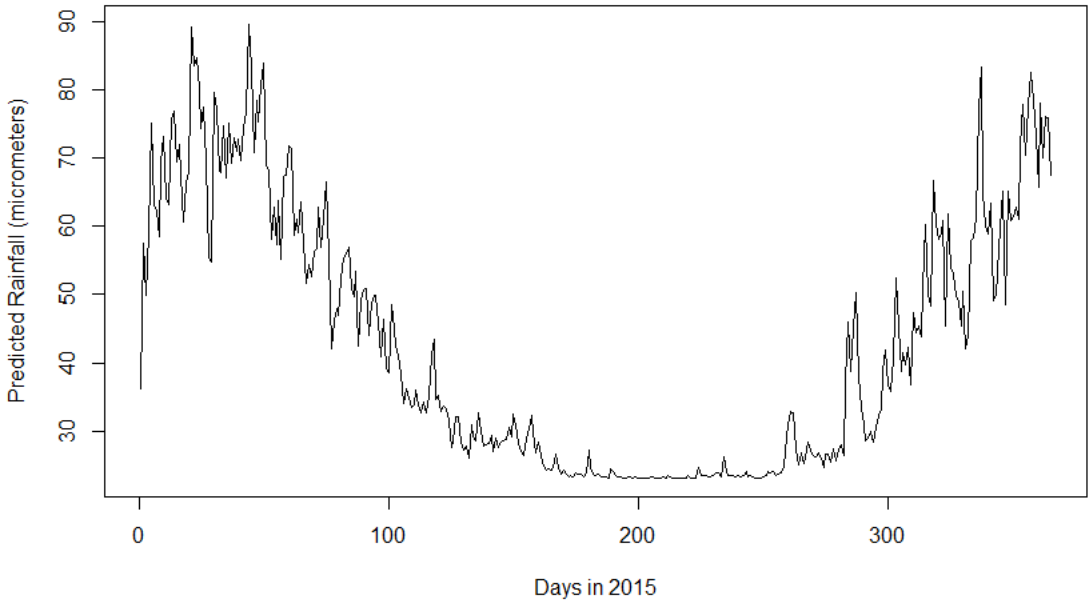
If the residuals had strong periodic components, we would expect to see peaks in the periodogram at the frequency of the oscillation. Since the periodogram appeared to contain nothing by white noise, we stopped our inquiry here. Given a further study, we would have tested the statistical significance of the small peaks.

Prediction

Perhaps the most important component of time series analysis is prediction. Since our dataset only extends through February 20th, we decided to predict rainfall through February 20th using the aforementioned AR(6) process, as well as the mean and seasonal estimated components. Our prediction appear below, with blue representing predicted rainfall and red representing actual rainfall:



As one can see, our prediction error is very high, but this is to be expected. We generally expect January to be wetter than June, since Davis rain tends to pour instead of slowly drizzle, predicting on which precise day rain will pour is very more difficult. Our prediction really only speaks to the mean behavior of the residuals, seeing as the variance of the residuals is not only large (2331.912), but substantially larger than the mean (0). In other words, given a day in January, we expect the day to have a large amount of rain with a very low probability. This results in poor predictions based on the residuals. We also plotted our predictions for the entirety of 2015, displayed below.



Conclusion

A rather surprising result of our data analysis was that we found no significant long term precipitation trend in Davis. In fact, the slope coefficient of the trend was positive, so Davis gardeners and musical aficionados have ever-so-slightly wetter weather to look forward to, despite evidence of the state as a whole getting dryer. Less surprising was the strong seasonal trend in precipitation. Any resident of California knows that the summer months are dry, especially in the Central Valley. It is interesting to note that there are some days for which the mean rainfall over the last 85 years has been 0. In other words, there are days (about 45) in which it has not rained a single day since the Davis weather station began recording precipitation data.

When examining the residuals it was apparent that an AR model was most appropriate. This coincides with the intuition that the past few days would be a good indication of how much rain there might be today. An AR(6) model makes sense, as it seems reasonable that the longest stretch of time a single day might have predictive power for rainfall would be about a week. As it is so infrequently the case, it is rather comforting to have one's intuition supported by statistical analysis, rather than unceremoniously torn down.

This report has been enlightening to us in many respects. It has provided insight into long term precipitation trends in our locale. It has also given us an opportunity to practice our freshly acquired, though somewhat limited, knowledge of Time Series Analysis. While we may not be able to fool an expert in our field into thinking we have achieved mastery, like our friend Eliza Doolittle so cunningly did to Hungarian phonetician, Zoltan Karpathy, we will be "dancing all night" if we have attained, at the very least, some level of proficiency such as to keep us off the streets selling flowers.

References

Thanks to Aaron Nip for his code to construct the AIC matrix.

City of Davis. Public Works > Water > Water Conservation > Drought. http://water.cityofdavis.org/water-conservation/drought

Richard Howitt, Jay Lund, Josu Medelln-Azuara, Kat Kerlin. Dateline. "Drought impact study: California agriculture faces greatest water loss ever seen." July 15th, 2014.

Dana Nuccitelli. The Guardian. "California just had its worst drought in over 1200 years, as temperatures and risks rise." December 8th, 2014.

Appendix

```
#Necessary for AIC Corrected
library(forecast)

#Import data
data = read.csv("Precipitation Data.csv")

#PURIFY DATA
#NOAA's data is based on five collection sites in Davis.  Four began relatively recently.
#We remove all data points not from the longest-running source (DAVIS 2 WSW EXPERIMENTAL FARM CA US)
data = data[data$'STATION_NAME' == 'DAVIS 2 WSW EXPERIMENTAL FARM CA US',]

#Some observations are missing and are stored as -9999.  Replace these values with NA
data[data$PRCP < -100,]$PRCP = NA

#The date values in the imported file are actually integers.  We need to convert them to Date objects.
date = 1:nrow(data)
class(date) <- "Date"
for (i in 1:nrow(data)){
  date[i] = as.Date(toString(data$DATE[i]),"%Y%m%d")
}

#Create a data frame consisting of two columns: date and precipitation on that day
dataset = data.frame(date, data$PRCP)

#Rename column data.PRCP to PRCP
names(dataset)[names(dataset)=="data.PRCP"] <- "prcp"


#Data is missing for some days. We find which dates are missing and insert NA values on those dates.
all_dates = seq(as.Date(dataset$date[1]), as.Date(as.Date("20151231","%Y%m%d")), by="1 day")
missing_dates = structure(setdiff(all_dates,dataset$date),class="Date")
missing_rows = data.frame(missing_dates,NA)
names(missing_rows)[names(missing_rows)=="missing_dates"] <- "date"
names(missing_rows)[names(missing_rows)=="NA."] <- "prcp"
dataset = rbind(missing_rows,dataset)

#Sort the data
dataset = dataset[order(dataset$date),]

#Remove leap years
leap_years = dataset[format(dataset$date, "%m") == "02" & format(dataset$date, "%d") == "29",]
for(i in 1:nrow(dataset)){
  if(nrow(leap_years) == 0){
    # do nothing
  }
```

```
  else if(dataset$date[i] == leap_years$date[1]){
    dataset = dataset[-i,]
    leap_years = leap_years[-1,]
  }
}

#Check if variance changes over time
var_by_year = c(1931:2014)
years = 1931:2014
for (i in 1931:2014){
  var_by_year[i-1930] = var(na.omit(dataset[format(dataset$date,"%Y") == i,]$prcp))
}
fit = lm(var_by_year ~ years)
summary(fit)
plot(1931:2014,var_by_year,xlab = "Year",ylab = "Annual Variance")

#Check if mean changes over time
mean_by_year = c(1931:2014)
for (i in 1931:2014){
  mean_by_year[i-1930] = mean(na.omit(dataset[format(dataset$date,"%Y")==i,]$prcp))
}
fit = lm(mean_by_year ~ years)
summary(fit)
plot(1931:2014,mean_by_year,xlab="Year",ylab="Annual Mean")

#Plot precipitation by month
plot(format(dataset$date,"%m"),dataset$prcp,xlab="Months",ylab="Precipitation")

#Store dataset as timeseries
myts <- ts(dataset$prcp)

#Construct 2d matrix
datamatrix = matrix(dataset$prcp,ncol = 365, byrow = TRUE)

#View data
plot(myts,xlab="Index",ylab="Precipitation (micrometers)")
summary(myts)

#Create matrix for residuals
datamatrix2 = datamatrix

#Calculate row means and subtract
row_means = c()
for (i in 1:nrow(datamatrix)){
  row_means[i] = mean(datamatrix[i,],na.rm = TRUE)
  datamatrix2[i,] = datamatrix[i,] - row_means[i]
}

#Calculate column means and subtract
column_means = c()
for(i in 1:ncol(datamatrix)){
  column_means[i] = mean(datamatrix[,i], na.rm = TRUE)
  datamatrix2[,i] = datamatrix[,i] - column_means[i]
}
```

```
#Plot the detrended, deseasonalized data
plot(1:(85*365),t(datamatrix2),xlab = "Days since January 1, 1931",ylab="Residuals",type="l")
mean(na.omit(as.vector(datamatrix2)))
var(na.omit(as.vector(datamatrix2)))

#Plot ACF
ACF = acf(as.vector(t(datamatrix2)), na.action = na.pass, type = "correlation",main="ACF by Lag")
#qqplot of ACF values
mean(ACF$acf)
var(ACF$acf)
OneOverN = 1 / length(dataset$prcp)
qqnorm(ACF$acf)
#Plot PACF
PACF = pacf(as.vector(t(datamatrix2)), na.action = na.pass,main="PACF by Lag")

#Calculate fit using AIC corrected.
#The rows are q values, the columns are p values
fitmatrix = matrix(nrow=8,ncol = 8,byrow=TRUE)
for (p in 0:7){
  for (q in 0:7){
    model = Arima(as.vector(t(datamatrix2)), order = c(p,0,q))
    fitmatrix[p+1,q+1] = model$aicc
  }
}

#See which ARMA process has minimum AICC
optimalp = which(fitmatrix == min(fitmatrix), arr.ind=TRUE)[1]-1
optimalq = which(fitmatrix == min(fitmatrix), arr.ind=TRUE)[2]-1
optimalAIC = Arima(as.vector(t(datamatrix2)), order = c(optimalp,0,optimalq))

#Find parameters for best model i.e. AR(6)
model = Arima(as.vector(t(datamatrix2)), order = c(optimalp,0,optimalq))
model
plot(as.vector(t(datamatrix2)),col="red",type="l",xlab = "Days since January 1, 1931",
 ylab="Residuals")
lines(fitted(model),col="blue",type="l")

#Time Domain Prediction
previous_year_mean = row_means[length(row_means) -1]

#Clone dataset and make predictions for the following year
predicted_rainfall = datamatrix2
for (i in 1:365){
  predicted_rainfall[,i] = predicted_rainfall[,i]+column_means[i]
}
for (i in 1:84){
  predicted_rainfall[i,] = predicted_rainfall[,i]+row_means[i]
}
predicted_rainfall = as.vector(t(predicted_rainfall))
for (i in 1:365){
  predicted_rainfall[365*84+i] = previous_year_mean + column_means[i] +
  0.2605*predicted_rainfall[365*84+i-1] + 0.0340*predicted_rainfall[365*84+i-2] +
```

```
  0.0409*predicted_rainfall[365*84+i-3]+0.0116*predicted_rainfall[365*84+i-4] +
  0.0105*predicted_rainfall[365*84+i-5] + 0.0192*predicted_rainfall[365*84+i-6]
}
plot(1:365,predicted_rainfall[(365*84+1):length(predicted_rainfall)],
     xlab = "Days in 2015", ylab = "Rainfall (micrometers)", type="l")

plot(1:56,predicted_rainfall[(365*84+1):(365*84+56)],
     xlab = "Days in 2015", ylab = "Rainfall (micrometers)", type="l",
     ylim=c(0,500), col="blue")

#Compare to observed values
lines(1:56,dataset[(365*84+1):(365*84+56),]$prcp,type="l",col="red")

#Spectral analysis
freqts = spectrum(as.vector(t(datamatrix2)),demean = TRUE,plot=TRUE,na.action=na.exclude)
plot(freqts,main="Daily Precipitation Raw Periodogram")

#Limit periodogram to two years
plot(freqts,xlim=c(0,0.01),main="Daily Precipitation Raw Periodogram")
plot(freqts,xlim=c(0,0.005),main="Daily Precipitation Raw Periodogram")
plot(freqts,xlim=c(0,0.0005),main="Daily Precipitation Raw Periodogram")
```