Julia Druce
NEAS Time Series
Winter 2015
Julia.sedona@gmail.com

Independent Project:

Modeling the Number of Downy Woodpeckers Counted in NJ from 1966 to 2014 as an ARIMA Process

Introduction

The Downy Woodpecker (*Picoides pubescens*) is a year round resident of New Jersey. This black and white bird is a common sight in suburban backyards. It is personally one of my favorite birds to watch, so I decided to try to model its population size in NJ as an ARIMA process. In every year since 1966, the USGS has conducted standardized surveys to count the number of individuals of each bird species found along specific routes across the United States I obtained the number of Downy Woodpeckers counted in New Jersey in each year from 1966 to 2014 (49 years) during this bird count (data and more info at https://www.pwrc.usgs.gov/bbs/).

Modeling the Downy Woodpecker Population Size as ARIMA(0,1,1)

After plotting time versus the number of Downy Woodpeckers counted, it was clear that this time series was non-stationary, as the mean was increasing over time (Figure 1). Plotting the sample autocorrelation function also confirmed that the time series was non-stationary, as the sample ACF did not decay quickly and remained significantly large for the first 9 lags (Figure 1).

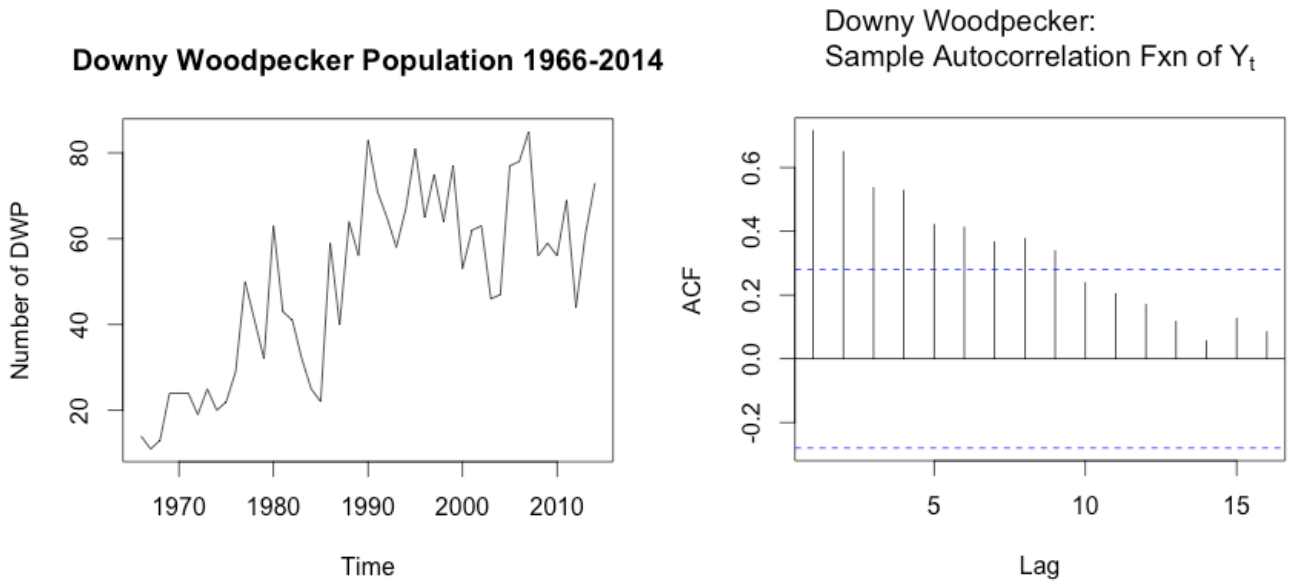

Figure 1. Evidence of non-stationarity

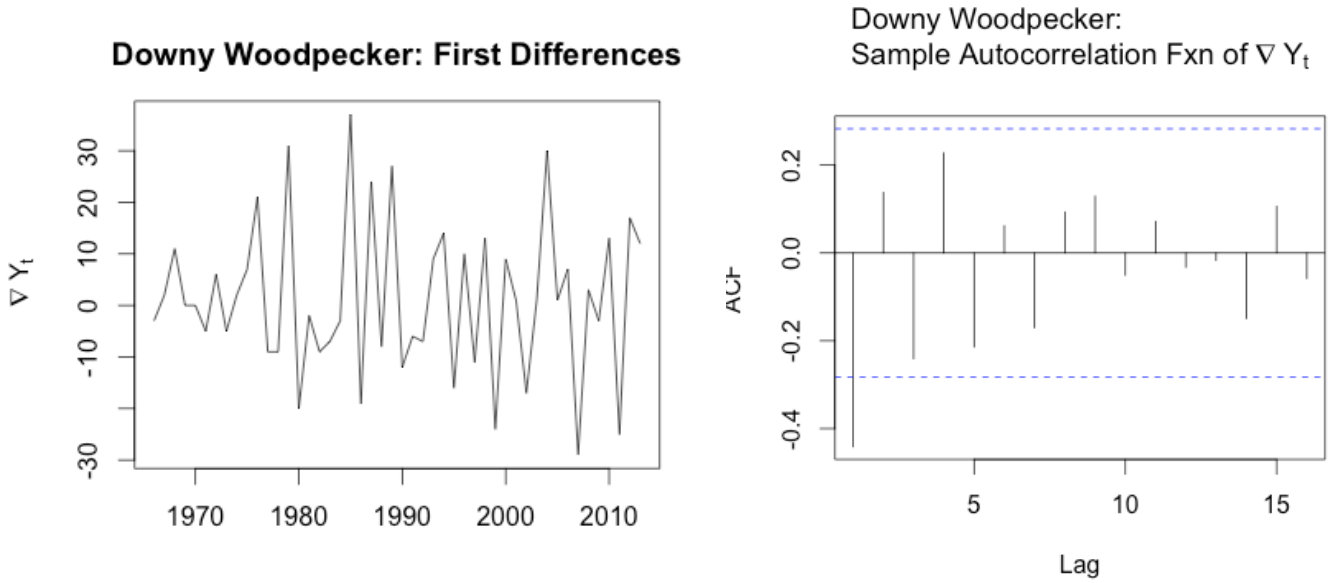

Figure 2. Obtaining stationarity by differencing

To obtain a stationary time series, I took the first difference of the yearly count data, which greatly improved the situation. Afterwards, the mean of the time series appeared to be nearly constant, and only the sample autocorrelation at lag 1 was statistically significantly different from zero (Figure 2). There does not appear to be any significant seasonality among the ∇Y_t. The pattern of autocorrelation seen in the correlogram in Figure 2 was reaffirmed when I plotted ∇Y_t vs. ∇Y_{t-1} and ∇Y_t vs. ∇Y_{t-2} (Figure 3). Again, there was a moderate strong negative correlation between ∇Y_t's at lag 1, but no correlation was present when the lag increased to 2

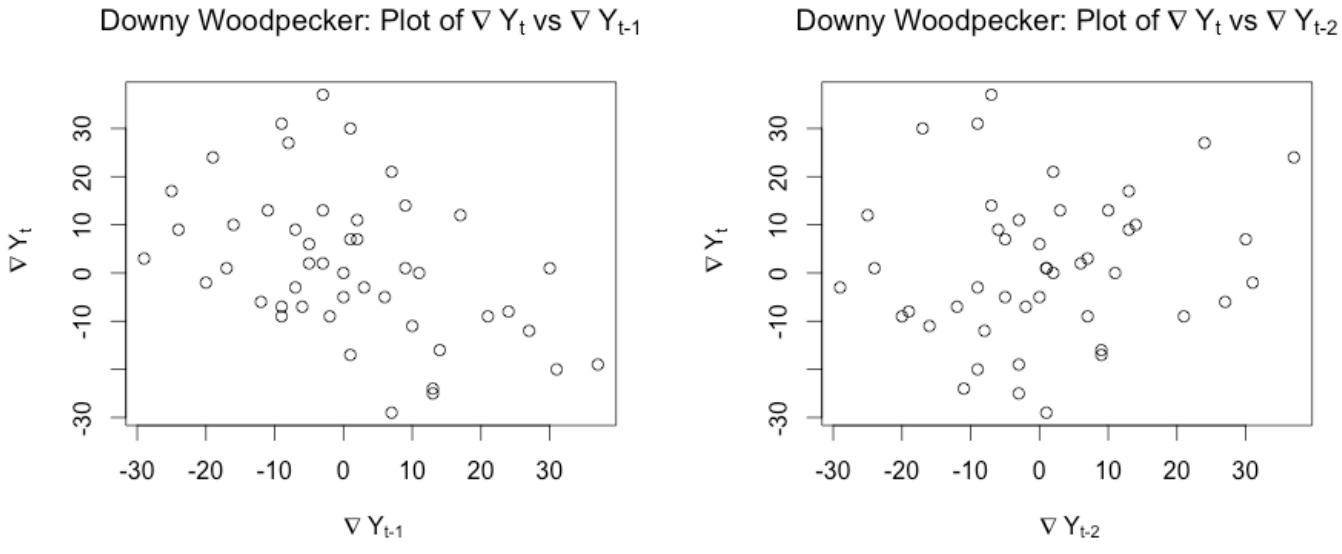

Figure 3. ARIMA (0,1,1) process support (r_1 significant only)

lags or greater. These findings lead me to conclude that an ARIMA (0,1,1) process could be the most appropriate model for modeling the number of Downy Woodpeckers counted yearly in NJ. I selected d = 1 because taking the first difference created a stationary time series and I selected q = 1 because only the sample autocorrelation at lag 1 was significantly different from zero, just as expected in an MA(1) process.

I then used R to estimate the coefficients of the ARIMA (0,1,1) model, obtaining $\nabla \mathbf{Y_t} = \mathbf{1.0814} + \mathbf{e_t} - \mathbf{0.6107}\ \mathbf{e_{t-1}}$ **.** The drift of the model is 1.0814 (*i.e.* the underlying MA(1) process has a mean of 1.0814) and the error term from the previous time point (e_{t-1}) moves the process in the opposite direction, as $\theta = 0.6107$. The standard error of the estimate of θ is 0.1608, so an approximate 95% confidence interval on θ is (.296, .926), which does not contain 0.

To examine the appropriateness of this model, I plotted the standardized residuals and the autocorrelation of the residuals (Figure 4). The residuals do appear to approximately reflect a white noise process: the residuals are randomly scattered around zero, with no significant correlation with each other. There is a slight deviation from the normal distribution expectation, as the QQ plot shows that the distribution of residuals is slightly skewed to the right, but this is not too troubling. In addition, the Ljung-Box statistic is not significant ($\chi^2 = 33.597$ on 24 df), so I do not reject the assumption that the residuals form a white noise process.

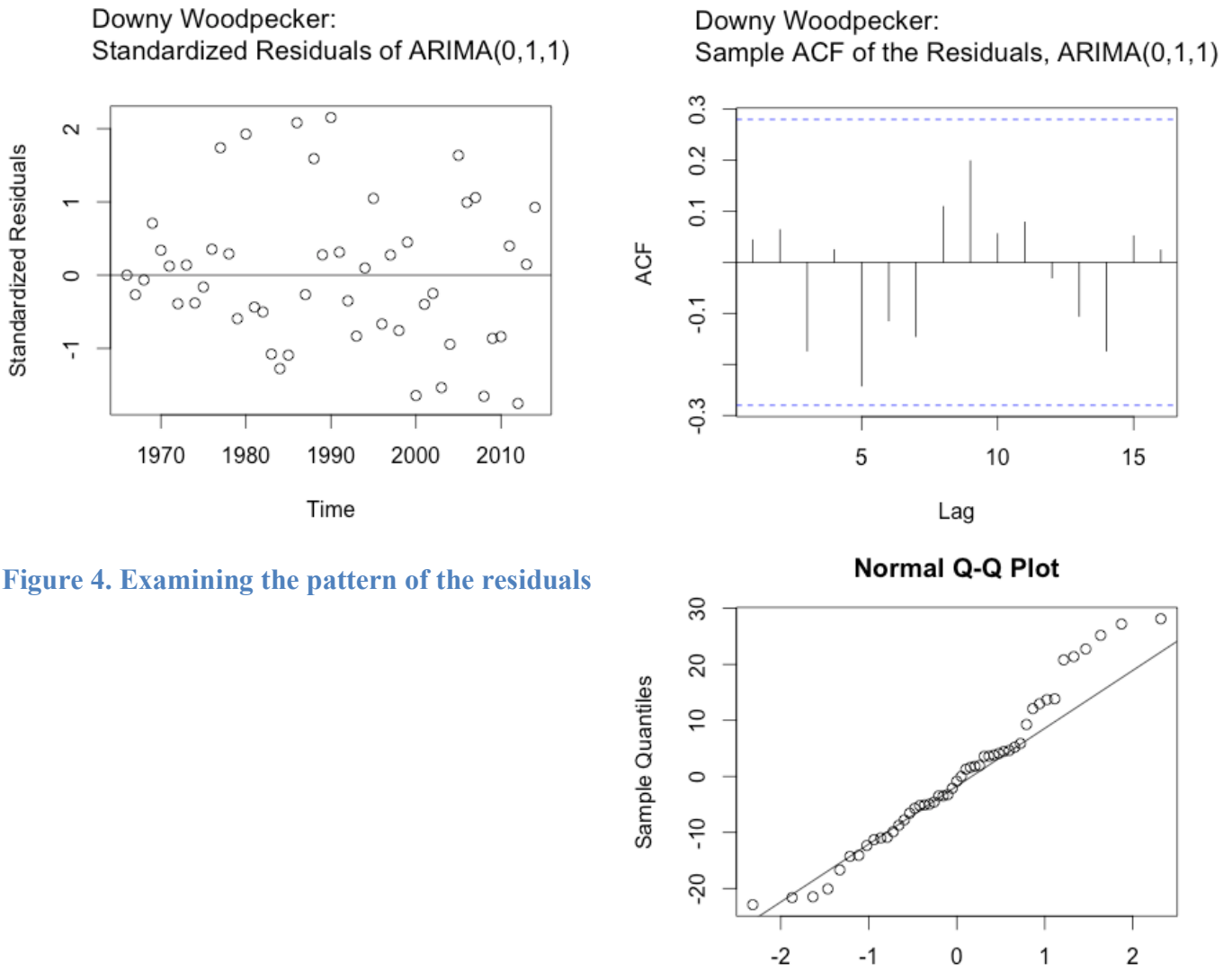

Figure 4. Examining the pattern of the residuals

Modeling the Logarithm of the Number of Downy Woodpeckers as ARIMA(0,1,1)

Next, I tried to see if the model could be improved by performing a transformation on the data. Taking the log of the original time series was not enough to make the time series stationary, so I then took the first difference of the log of the data. As before, after differencing, the mean appeared to be constant and only the sample autocorrelation at lag 1 was significantly different from zero. I used R to estimate the coefficients for an ARIMA (0,1,1) model of the log of the data, getting $\nabla \log(Y_t) = 0.0335 + e_t - 0.4736\, e_{t-1}$

The drift is not significantly different from zero. $\theta = 0.4736$ with a standard error of 0.1627. An approximate confidence interval of θ is (0.155, 0.792), which does not contain zero. Again, the residuals were plotted to assess the suitability of the model (Figure 6). Again, the residuals approximately replicate a white noise process, showing random scatter around zero, no correlation with each other, and approximate normal distribution, although like the previous model, a slight skewness to the right is evident in the QQ plot. Finally, the Ljung-Box statistic is not significant ($\chi^2 = 30.055$ on 24 df), so I do not reject the assumption that the residuals form a white noise process.

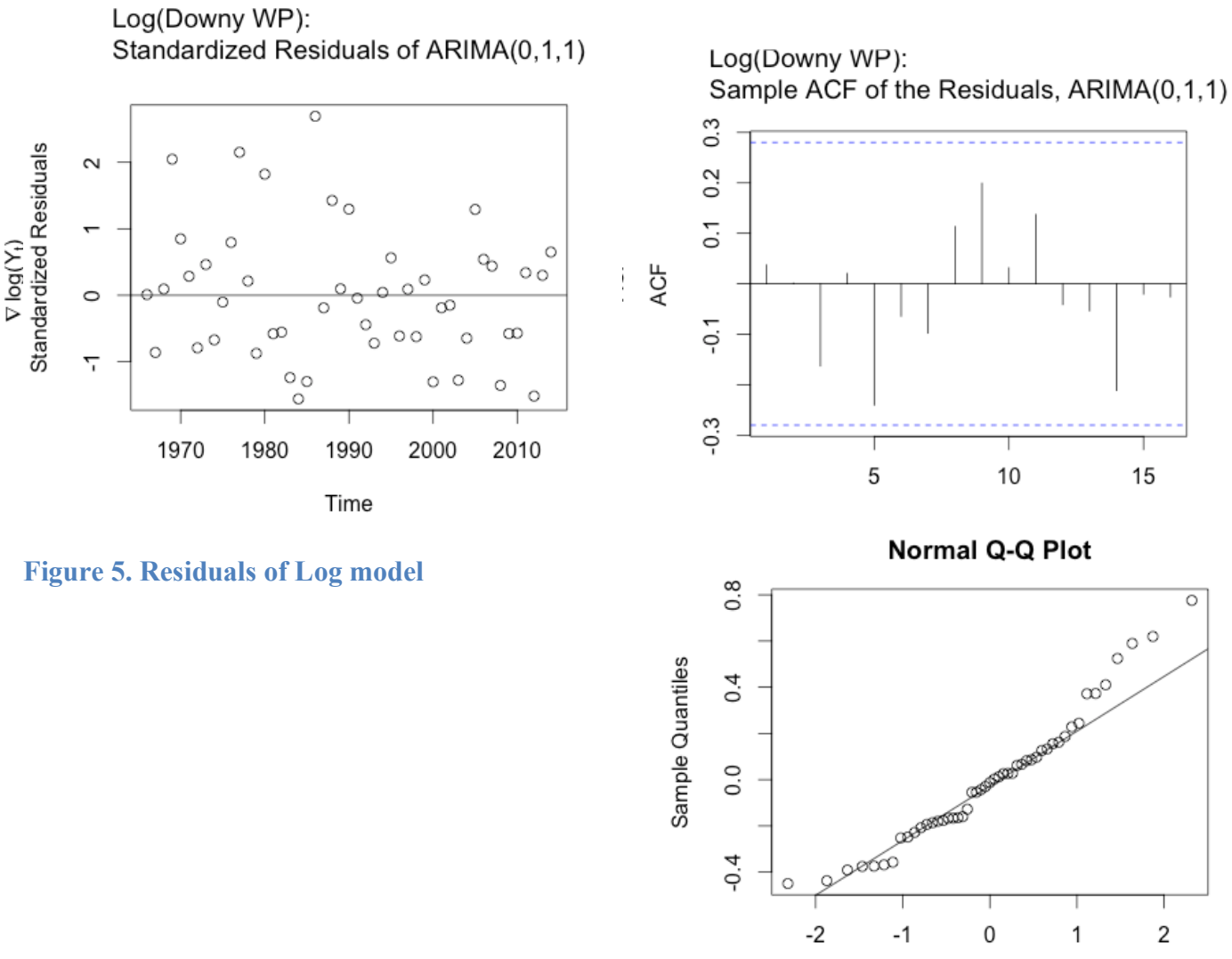

Figure 5. Residuals of Log model

Conclusion:

The ARIMA (0,1,1) process was selected after observation that taking the first difference generates a stationary time series and that only the sample autocorrelation of the Both the ARIMA (0,1,1) model and the log of the data ARIMA (0,1,1) model exhibited similar properties, including approximate white noise behavior of the residuals. However, the log of the data model was superior when comparing AIC and BIC values (AIC=21.65 and BIC=27.26 for the log model vs. AIC=380.45 and BIC=386.07 for the untransformed data model). Figure 7 shows the result of forecasting from both of the models. The forecasts are similar in that forecasting beyond a lead of one, the best forecast is just the linear drift of the mean.

The final model, $\nabla \log(Y_t) = 0.0335 + e_t - 0.4736\, e_{t-1}$, implies that if the difference of $\log(Y_{t-1})$ and $\log(Y_{t-2})$ is larger (smaller) than expected, then the difference between $\log(Y_t)$ and $\log(Y_{t-1})$ will tend to be small (large). In simpler terms, if the change in population for a given year was larger than expected, then the change in population in the following year should be small, and vice versa. There is probably a natural force at work causing this type of fluctuation in the counted numbers. Perhaps overly large population growth in one year causes intense competition for resources such as shelter and food, and in the subsequent year the population can't grow as much as because survival is limited by lack of resources caused by the previous year's boom. Alternatively, if population growth is lower than expected, in the following year the population can rebound due to excess resources.

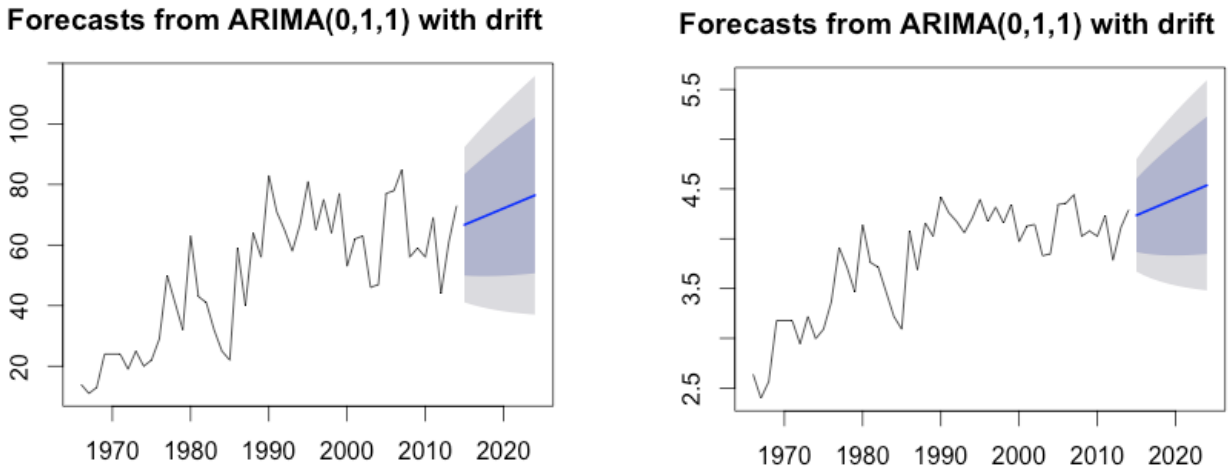

Figure 7. Forecasting from the models. Left: original model. Right: log transformed model.

R code:

** Extracting the relevant data from a larger data set**

```
library(plyr)
library(TSA)
library(forecast)

nj<-read.csv("NJersey.csv", header=T)
df<-aggregate(nj, by=list(nj$Aou, nj$Year), FUN=sum)
dfs<-split(df,df$Group.1)
extractYears <- function(df)
{
    r <- t(data.frame(df$SpeciesTotal, row.names=df$Group.2))
    r<-as.data.frame(r)
    row.names(r) <- df$Group.1[1]
    return(r)
}
flattened<-lapply(dfs, extractYears)
birds<-rbind.fill(flattened)
rownames(birds) <- names(flattened)
birds <- birds[,order(colnames(birds))]
birds[is.na(birds)] <- 0
```

****Model creation and testing****

```
downy<-t(birds["3940",])

downyts<-ts(downy, start=1966)

downy1<-diff(downy)

downy1ts<-ts(downy1, start=1966) logdowny<-log(downy)

plot(downyts, ylab="Number of DWP", main="Downy Woodpecker
Population 1966-2014")

plot(downy1ts, ylab=expression("Y"[t]), main="Downy Woodpecker:
First Differences")

plot(y=downy1, x=zlag(downy1), ylab=expression(~ nabla ~"Y"[t]),
xlab=expression(~ nabla ~"Y"['t-1']),
     main=expression("Downy Woodpecker: Plot of" ~ nabla
~"Y"[t]* " vs" ~ nabla ~"Y"['t-1']))

plot(y=downy1, x=zlag(downy1,2), ylab=expression(~ nabla
~"Y"[t]), xlab=expression(~ nabla ~"Y"['t-2']),
     main=expression("Downy Woodpecker: Plot of" ~ nabla
~"Y"[t]* " vs" ~ nabla ~"Y"['t-2']))
```

```
acf(downyts, main=expression("Downy Woodpecker:\nSample
Autocorrelation Fxn of Y"[t]))

acf(downy1ts, main=expression("Downy Woodpecker:\nSample
Autocorrelation Fxn of" ~nabla~ "Y"[t]))

dwp1<-Arima(downyts, c(0,1,1), include.constant=T)
plot(dwp1$residuals/sqrt(dwp1$sigma2), ylab="Standardized
Residuals", main=expression("Downy Woodpecker:\nStandardized
Residuals of ARIMA(0,1,1)"), type="p")

abline(h=0)

acf(residuals(dwp1), main=expression("Downy Woodpecker:\nSample
ACF of the Residuals, ARIMA(0,1,1)"))

qqnorm(residuals(dwp1))

qqline(residuals(dwp1))

Box.test(residuals(dwp1), lag=25, type="Ljung-Box", fitdf=1)
```

****Modeling the Log transformed data****

```
logdowny1<-diff(log(downy))

logdowny1ts<-ts(logdowny1, start=1966)

plot(logdowny1ts,  ylab=expression(~ nabla ~"log(Y"[t]*")"),
main="Downy WP: First Differences of Log")

logdwp1<-Arima(logdownyts, c(0,1,1), include.constant=T)

plot(logdwp1$residuals/sqrt(logdwp1$sigma2), ylab="Standardized
Residuals", main=expression("Log(Downy WP):\nStandardized
Residuals of ARIMA(0,1,1)"), type="p")

abline(h=0)

acf(residuals(logdwp1), main=expression("Log(Downy WP):\nSample
ACF of the Residuals, ARIMA(0,1,1)"))

qqnorm(residuals(logdwp1))

qqline(residuals(logdwp1))

Box.test(residuals(logdwp1), lag=25, type="Ljung-Box", fitdf=1)
```