Predictive Analytics Final Project

The Future of U.S. Honey Bee Populations:
A Time Series Analysis

Team 13
Brendan Hogan, Erin Ospeck, Casey Samagalsky, Celeste Sowell
November 3, 2020

I. Abstract

The purpose of this analysis is to forecast the U.S. Honey Bee population through 2020 using historical annual population data from 1987 to 2017. By determining the most effective model for prediction, we will be able to determine the general direction of the Honey Bee population. The models that will be used for the analysis will include three types of simple models (Mean Forecasts, Naive Forecasts, and Drift Forecasts), an ARIMA analysis, as well as a Regression analysis. Comparing the accuracy of each of these five models will determine the most fitting model to the dataset as well as help predict the future of Honey Bees in the United States.

II. Introduction

Having enormous impacts in agriculture, horticulture, and the general ecosystem of the United States, the Honey Bees' pollination keeps the delicate balance of the environment steady. According to the Food and Drug Administration of the U.S., this pollination is vital to over 250,000 species of flowering plants, which without it would struggle to reproduce. This pollination in crops adds over $15 billion in added crop value and sustains the American agricultural sector (FDA.org).

Image 1.1 - The Honey Bee: Our Friend in Danger by Zachary Huang (fllt.org)

Nearly three decades ago, bee populations across the United States experienced a drastic decline, which has since been attributed to a combination of two factors -- a phenomenon known as Colony Collapse Disorder (CDC), where bees leave their lives and never return, and the introduction of neonicotinides, neuro-active insecticides that started to be used in the late 1980s. Neonicotides have since become one of the most widely-used pesticides in the world. Efforts, such as the use of other pesticides, have been made to avoid such bee colony loss, but the causes of this CDC phenomena are still unclear.

In more recent years, American foulbrood, a bacterial disease, is destroying entire colonies of honey bees across the United States, putting the environmental balance held together by the Honey Bee and the crops that they pollinate at risk. After the initial outbreak of foulbrood in 2006, measures were taken to prevent the destruction of this important species, such as the development of antibiotics and the "Save the Bee" movement.

III. Motivation

Since the development of this movement and the response by environmentalists, it is critical to determine what has been happening to the bee population. From here, it can be seen if

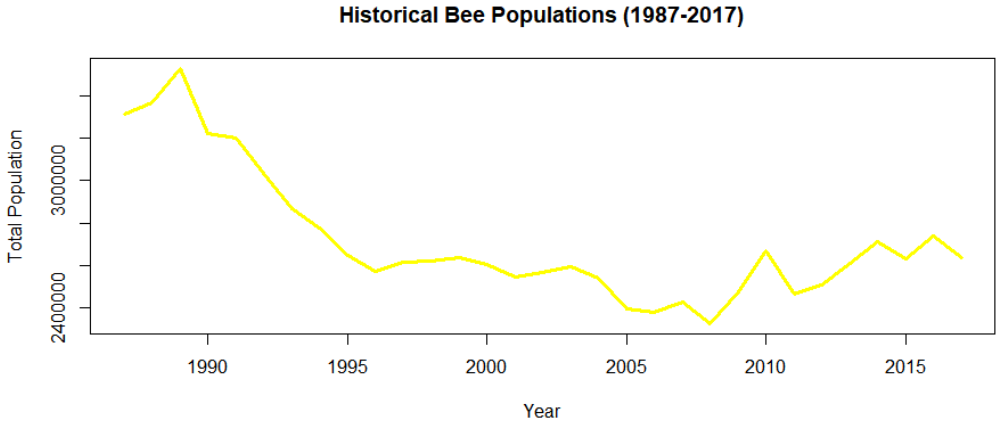
this response was sufficient enough to save the bees, or if further measures would be needed. Regarding the Honey Bees, what can be expected of their population in the upcoming years? Which forecasting method would be the most accurate at the prediction of this? Is the outcome consistent with unofficial population determinations by experts? Decreasing bee populations would have devastating impacts on the environment and many industries in the United States. We hope to find a forecasting method that can estimate the U.S. bee population for 2018 and 2019, and figure out if it is increasing, decreasing, or staying stagnant. With this information, the United States could then see if the bee population is something worth paying more attention to in the future.

IV. The Data

The Honey Bee population data that is being used for the predictive analysis is from the U.S. Department of Agriculture (via data.world). This is a breakdown of total honey bee populations by state for each year from 1987-2017, so it contains a modest amount of historical data to build the model off of. However, because this dataset is annual, it can be a bit more restrictive than a dataset that contains seasonal or monthly data.

As can be seen from Figure 4.1, there was a strong decreasing trend in Bee populations starting in 1987 until the mid-1990s. As mentioned above, this was likely attributed to the Colony Collapse Disorder phenomena and introduction of new insecticides. The Save the Bee movement and the antibiotics that were developed to fight American foulbrood have further aided in the efforts to save bee populations from extinction. Figure 4.2 describes the bee population across the midwest divided by region of the United States. For predictive analysis, factoring the location of the bee population was an important factor, especially if a regression-based forecasting model was to be used.

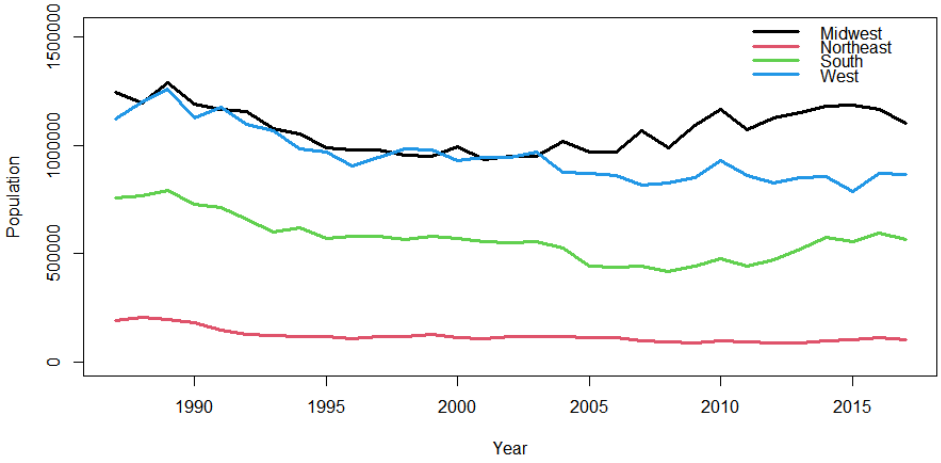
Figure 4.1: Total Bee Populations for United States (1987-2017)



It was decided in the analysis to perform four separate predictions based on the four main regions of the U.S. (Northeast, South, Midwest, and West) in order to get a true perspective on

which populations have changed over time and into the next upcoming years. These regions are determined by the U.S. Department of Agriculture and contain whole states, so aggregating the dataset into each region was simple. The data for each region was broken down into training and validation sets, to have one to build the model off of, and the other to compare with to determine accuracy. The training data consisted of 27 points and the validation data consisted of 4 data points. This allows for a large enough dataset to build the model off of, with a proportionate amount of the data being used for validation purposes. For the model testing, the Southern region was used to determine the model accuracy. After the most effective model was found, that model was applied to all four regions.

Figure 4.2: Bee Populations by Region 1987-2017



V. Simple Model Analysis

We ran three simple forecasting analyses on the bee populations in the South region -- Mean, Naive, and Drift. The South region is made up of the highest number of states, so therefore this would render the most accurate results. From this, we plan to find which model is best for forecasting our data and then replicating the analysis on the remainder of the regions (Northeast, West, and Midwest), using the years 1987-2014 as training data and the years 2015-2017 as validation data. Looking at the calculated RMSE (root mean squared error) for each model, we are able to easily compute the solution using a differentiable, symmetric, simple loss function in order to approximate the conditional expected value of the next observation (to be predicted) given the explanatory variables (historical data).

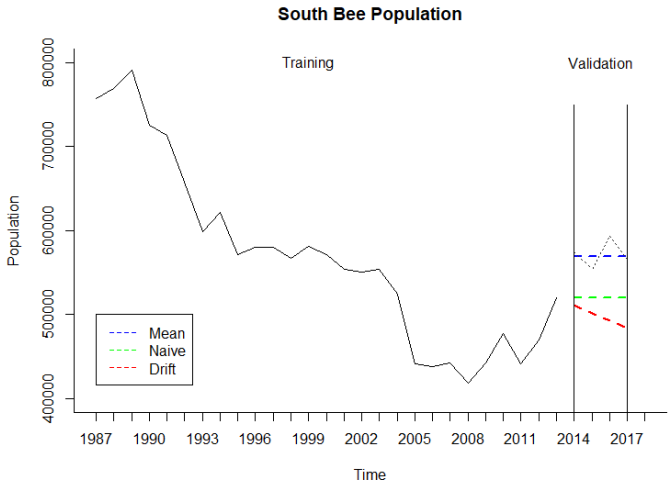


Figure 5.1: Simple Method Models

Mean (Figure 1.4 in Appendix)

For the Mean forecast, where all of the future forecasts are equal to the average of historical data (1987-2017), the graph simply continues the mean of the bee populations around 570,000. This shows that bee populations are projected to remain constant in the next few years. The RMSE found for this model is 14,574.32.

Naive (Figure 1.5 in Appendix)

For the Naive forecast, which sets all forecasts to the value of the last observation (in this case, 2017), we can see that the graph shows a slight decrease in the mean of the bee population to approximately 520,000. The RMSE found for this model is 53,911.97, which is over two times larger than the RMSE of the Mean model.

Drift (Figure 1.6 in Appendix)

For the Drift forecast analysis, which is simply a variation on the naive method which allows the forecasts to increase or decrease based on the historical average amount of change over time, we can see that the mean of the bee populations is forecasted to slightly decrease in the coming years to between 490,000 and 520,000. The RMSE found for this model is 76,987.16, which is much higher than both the Mean and the Naive models.

Forecast Results

After analyzing the data from all three forecasting models, we conclude that the Mean forecasting model is the most effective for looking at the future of bee populations. Utilizing the Mean method will convey more accurate results for the forecasted future of bee populations, as historical data is beneficial in future forecasts. Similarly, this forecasting method rendered the lowest RMSE, MAE, and MAPE, therefore proving to be the most accurate model.

VI. ARIMA Model

The AR Model, or Autoregressive Model, takes advantage of autocorrelation by including it in the regression model itself. A type of AR Model known as the ARIMA model (Autoregressive Integrated Moving Average) does this similarly to a linear regression model, but is mostly more accurate. After performing the ARIMA analysis on the Southern Region data using auto.arima(), the next step was to verify the autocorrelation of the model residuals to make sure that there is no significant autocorrelation. This also verifies that the model was performed correctly. According to Figure 6.1, there is little to no autocorrelation, and it can be determined that the model performed correctly.

Figure 6.1: Acf of Arima Model Residuals

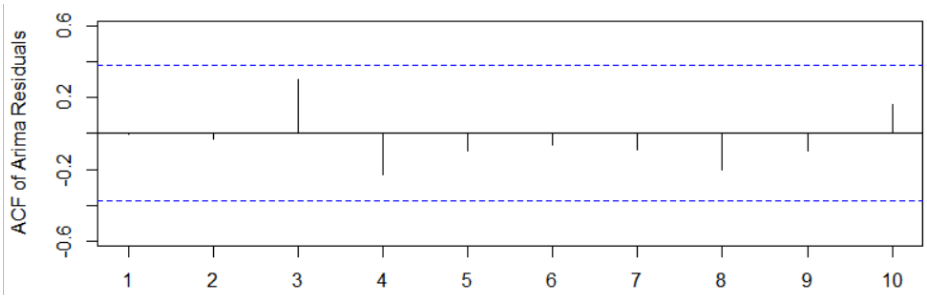
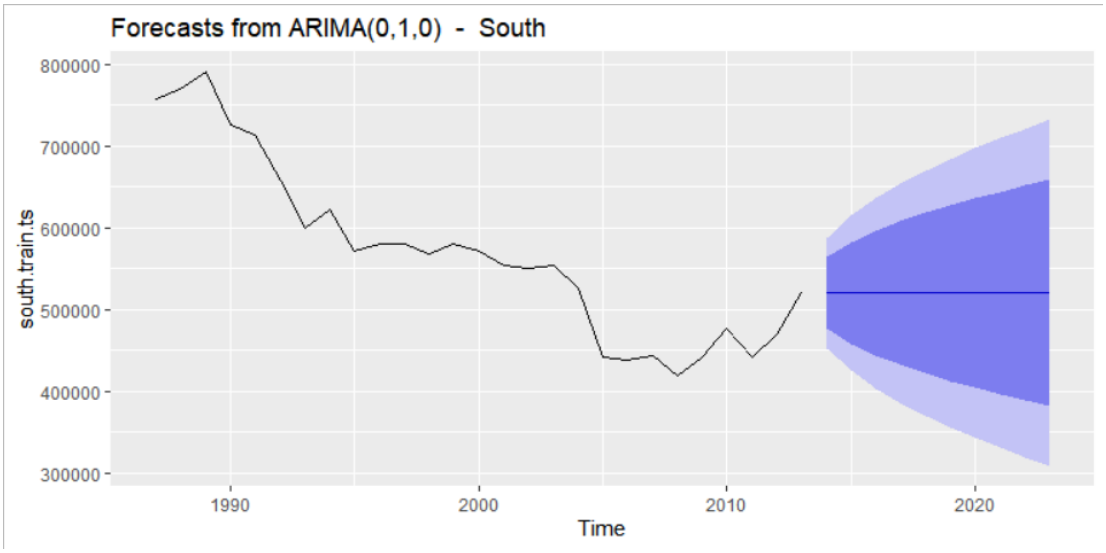


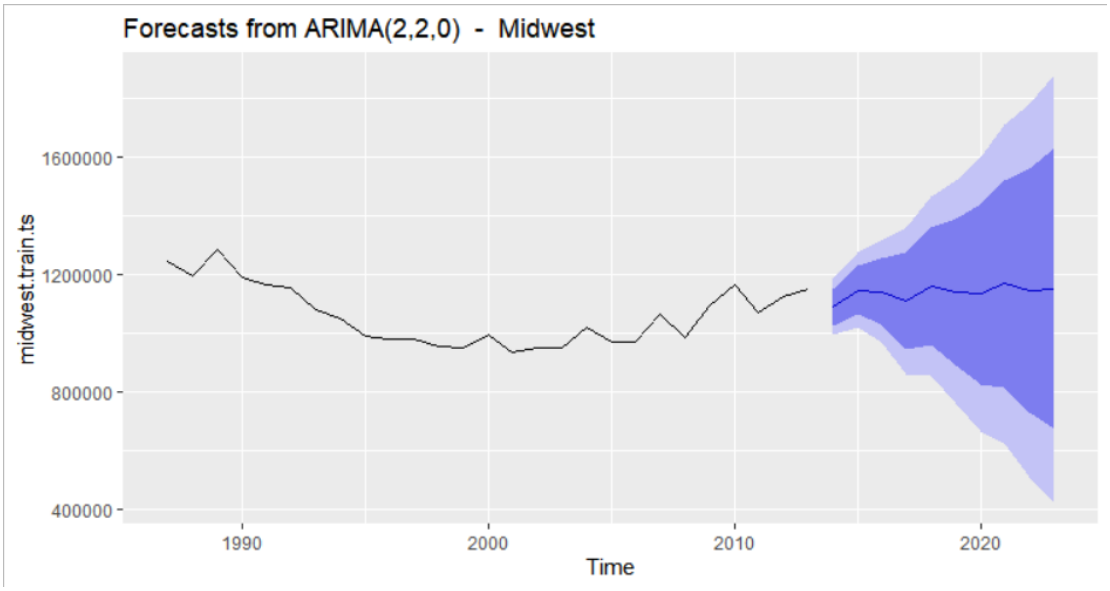
Figure 6.2: South ARIMA



Forecasting the auto.arima() allows for the true trend and results of the ARIMA model to be seen. Very interesting results occurred. According to ARIMA(0, 1, 0) in the results of the model. The Southern Region's data is a random walk. A random walk indicates that the model is not predictable. It takes the previous value in the time series and will add a coefficient that is merely a step away. In this case, the step is equal to zero, so the results of this model are equal to that of the Naive Model (RMSE = 53,911.97). Additionally, the error values are exactly equal to that of the Naive Model. Due to these interesting results, it was decided to perform the ARIMA model on all four regions to see if it was consistent.

The results from the other region analyses determined that the South, West, and Northeast also had random walk results, except that the West had a coefficient that wasn't equal to 0, meaning that it is equal to Drift and not Naive. The Midwest, however, did not have a random walk. The Midwest region had a low error (RMSE = 50,276.43) and a slightly increasing trend towards the future. While this RMSE is very low, if it were to be the winning model, it would be a debate as to whether it would be appropriate to apply ARIMA to every region, as the regions lead to very different results.

Figure 6.3: Midwest ARIMA



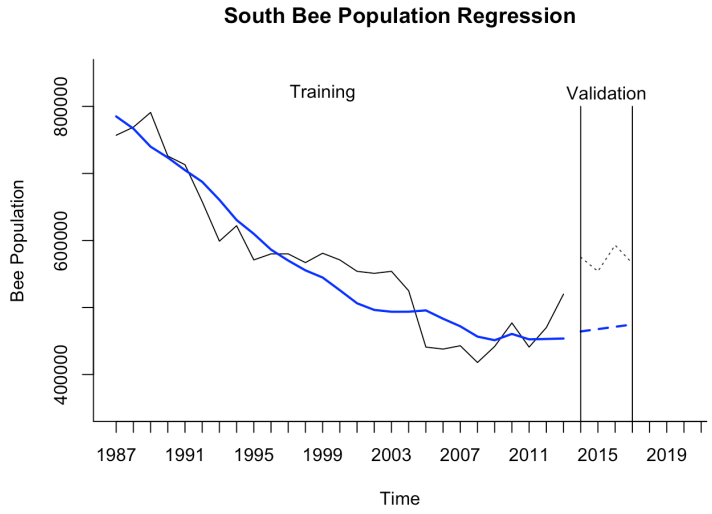
VII. Regression Model

Since bee population trends didn't appear to be strongly predictable by the simple and ARIMA models, it seemed likely that some external trends were driving the large swings in population numbers. Global warming and vegetation health appear to be strong candidates as drivers, as research has shown both to impact bee population. To proxy for warming and vegetation, this model uses *average temperatures* and *annual precipitation* numbers, collected by state and averaged across states in each region.

The regression incorporates three components: a quadratic trend, average yearly temperature across states in a region, and average annual precipitation across states in a region. A quadratic trend makes sense here, not only because the data appears to be U-shaped, but because the quadratic model has a grossly lower error than the linear one.

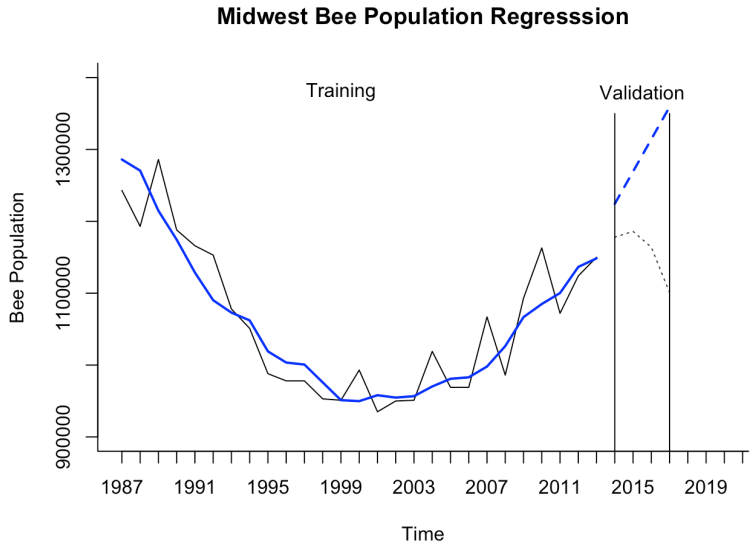
The Southern model correctly suggested a growth in bee population over the next four years, but it grossly underestimated the extent of this growth, with a mean percentage error of 17.87%. This suggests that a quadratic trend in the population, along with rainfall and temperature, probably does not account for most of the change in bee populations over this period.

Figure 7.1: Regression Model for South



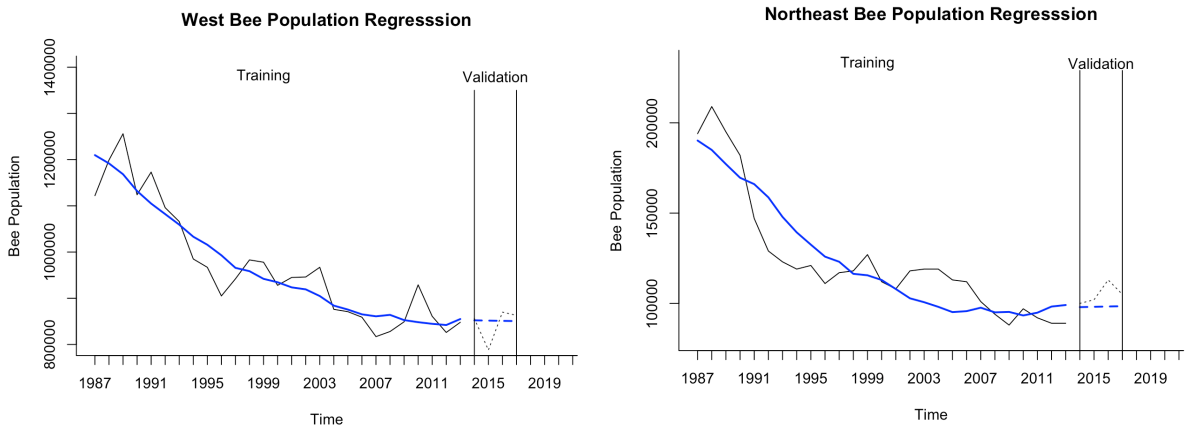
The model performed even worse in the Midwest, at least visually, predicting a steep increase in the validation data that actually saw a sharp decline. The mean percentage error for this region was better, though, at -11.83%. This is probably due to the relatively static trends in Midwest bee populations, relative to the South.

Figure 7.2: Regression Model for Midwest



The model performed better in the West, with a -1.01 mean percentage error, and in the Northeast, with a 6.29 mean percentage error. Much like the Midwest, though, 2014-2017 was relatively stable for bee populations in these regions, which may suggest that the model itself was still not particularly predictive. Yet again, it would seem that trends in temperature and precipitation do not adequately explain much of the variability in bee populations, even if the error in some regions are relatively low.

Figure 7.3: West and Northeast Regression



VIII. Final Model and Conclusions

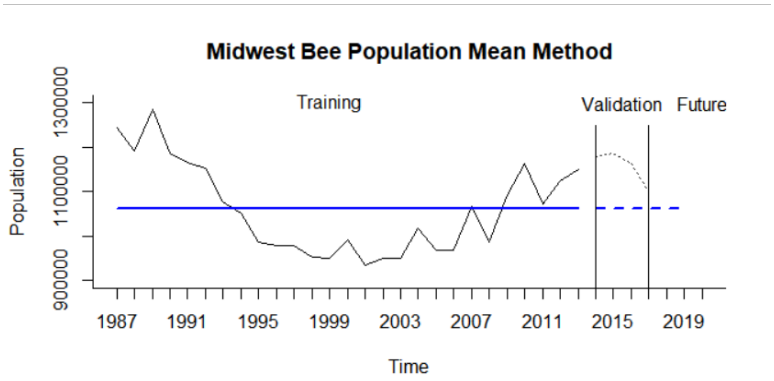
Based on the RMSE values that are featured in Figure ?.?, the mean method is the most accurate model. While it might seem disappointing that a more complex model couldn't predict any more accurately, it demonstrates that the bee data in the U.S. is so complicated and that the best way to predict is with an average approach. The U.S. Honey Bee population is extremely difficult to model, with the ARIMA model even indicating that certain regions may even be unpredictable. With such uncertainty surrounding the future of the bee population, it goes to show that the bees are still extremely vulnerable and need continuing protections.

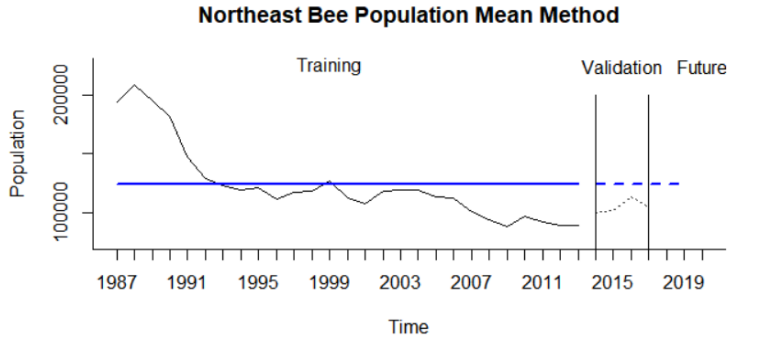
Figure 8.1: RMSE Values of the Models

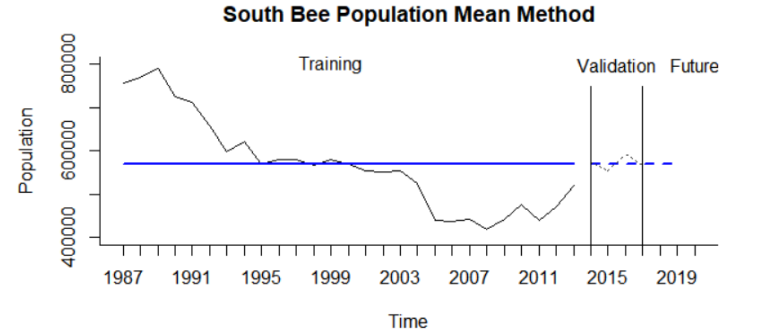
	MEAN	NAIVE	DRIFT	ARIMA (South)	ARIMA (Midwest)	Regression
RMSE	**14,574.32**	53,911.97	76,987.16	53,911.97	50,276.43	103,493.93
MAPE	**2.08**	9.03	13.02	9.03	3.41	17.87
MAE	**12,000.00**	52,000.00	74,788.46	52,000.00	40,045.85	102,488.58

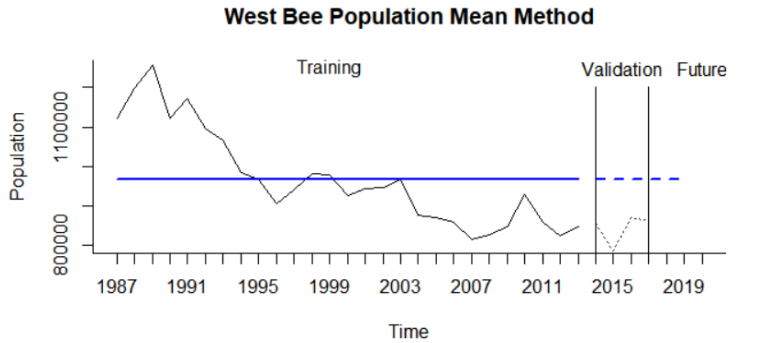
The final model predicts that the U.S. Honey Bee Population will remain somewhere around 2,671,000 colonies through 2018 and 2019. This is a bit lower than numbers recently released by the USDA, showing colony totals slightly over 2.8 million. Whether these overshot numbers continue to climb, or whether they fall again as they did through the 90s, will be vitally important to the future of the country's ecosystem, vegetation, and food supply. Perhaps one of the biggest takeaways from this analysis is that this trajectory--and its sweeping impacts on American life--is incredibly uncertain. It will rely on proper colony management, equitable regulation, and sustainable farming practices to create a healthier ecosystem, a reliable food supply, and a more predictable trajectory for bee colonies.

Figure 8.2: Final Models of US Bee Population by Mean









IX. Appendix:

Figure: Training and Validation data split for the South

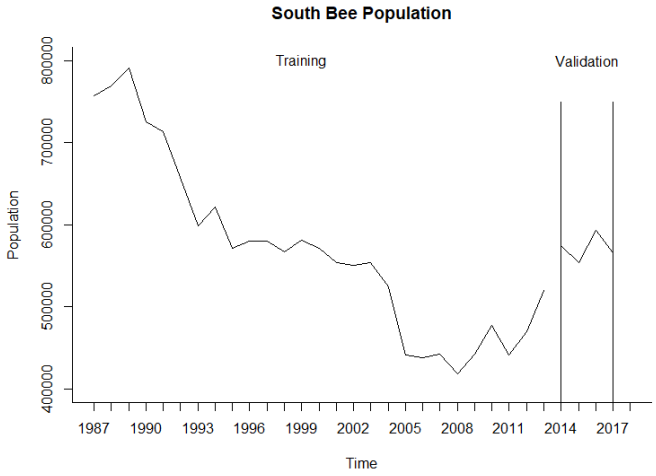
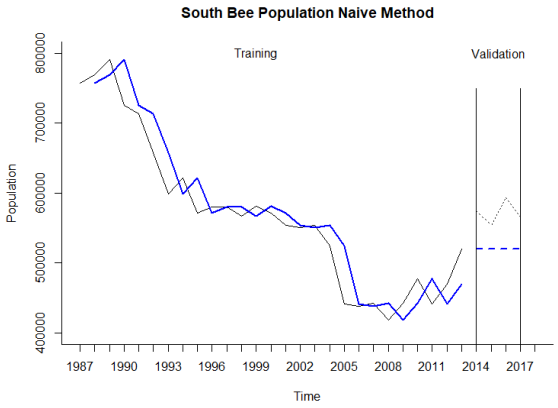
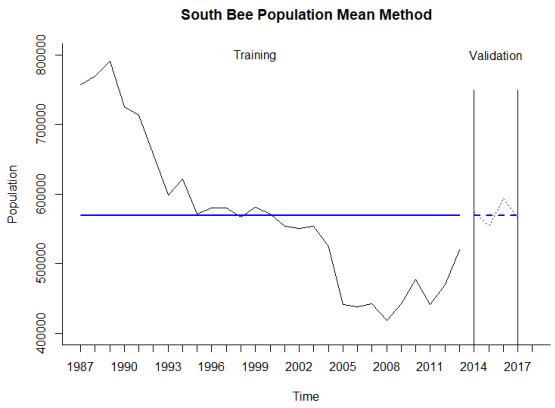
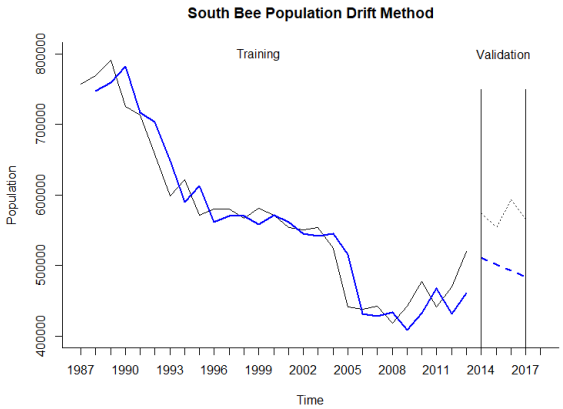


Figure 1.4-1.6: Results for Naive, Drift, and Mean







Links to Data:
https://www.ncdc.noaa.gov/cag/statewide/time-series/24/tavg/ann/12/1987-2020?base_prd=true&begbaseyear=1987&endbaseyear=2017

https://data.world/finley/bee-colony-statistical-data-from-1987-2017/workspace/file?filename=Bee+Colony+Survey+Data+by+State.csv

Images:
https://www.cnn.com/2020/02/06/us/bumble-bee-climate-change-extinction-study-scn/index.html
http://www.fao.org/news/story/en/item/1194910/icode/
https://www.washingtonpost.com/

Bibliography:

Andrei, Mihai. "The World Bee Populations Are Dwindling, and We Still Can't Make It Stop." *ZME Science*, 4 Apr. 2013, www.zmescience.com/ecology/animals-ecology/bee-population-ccd-04042013/.

CMB.Contact@noaa.gov. "Climate at a Glance." *National Climatic Data Center*, NOAA National Centers for Environmental Information, Oct. 2020, www.ncdc.noaa.gov/cag/.

Huang, Zachary. *The Honey Bee: Our Friend in Danger*. Ithaca, New York.

Medicine, Center for Veterinary. "Helping Agriculture's Helpful Honey Bees." *U.S. Food and Drug Administration*, FDA, 30 July 2018, www.fda.gov/animal-veterinary/animal-health-literacy/helping-agricultures-helpful-honey-bees.

Medicine, Center for Veterinary. "KIDDING." *U.S. Food and Drug Administration*, FDA, www.fda.gov/animal-veterinary/animal-health-literacy/helping-agricultures-helpful-hone

y-bees.

"Why Bees Are Important." *Sustain*, Sustain 2020, 2020,

www.sustainweb.org/foodfacts/bees_are_important/.

R Code:

```
# Bee Population Final Project - Brendan Hogan, Erin Ospeck, Casey Samagalsky, Celeste Sowell


# Loading Required Libraries

library(forecast)

library(ggplot2)

library(dplyr)

library(MASS)

library(tidyr)


### PREPROCESSING

##DATA MANIPULATION


# Loading Bee Dataset

bees_state <- read.csv("Bee Population by State 1987-2017.csv", stringsAsFactors=TRUE)

bees_state$Region <- NA

bees_state$Population <- as.numeric(gsub(",", "", bees_state$Population))


## Loading Weather Datasets
```

```
# Precipitation

precip <- read.csv('State_Precip.csv')

names(precip)[1] <- "YEAR"

precip$Region <- NA

precip <- gather(precip, State, Precipitation, ALABAMA:WYOMING)

precip <- precip[ -c(2:3) ]


# Temperature

temp <- read.csv('State_Avg_Temp.csv')

names(temp)[1] <- "YEAR"

temp$Region <- NA

temp <- gather(temp, State, Temperature, ALABAMA:WYOMING)

temp <- temp[ -c(2:3) ]


# Creating Region Vectors

Region <- list("Northeast", "Midwest", "South", "West")

Northeast <- c("CONNECTICUT", "MAINE", "MASSACHUSETTS", "NEW HAMPSHIRE",
"RHODE ISLAND", "VERMONT", "NEW JERSEY", "NEW YORK", "PENNSYLVANIA")

Midwest <- c("ILLINOIS", "INDIANA", "MICHIGAN", "OHIO", "WISCONSIN", "IOWA",
"KANSAS", "MINNESOTA", "MISSOURI", "NEBRASKA", "NORTH DAKOTA", "SOUTH
DAKOTA")

South <- c("DELAWARE", "FLORIDA", "GEORGIA", "MARYLAND", "NORTH
CAROLINA", "SOUTH CAROLINA", "VIRGINIA", "DISTRICT OF COLUMBIA", "WEST
VIRGINIA", "ALABAMA", "KENTUCKY", "MISSISSIPPI", "TENNESSEE", "ARKANSAS",
"LOUISIANA", "OKLAHOMA", "TEXAS")

West <- c("ARIZONA", "COLORADO", "IDAHO", "MONTANA", "NEVADA", "NEW
MEXICO", "UTAH", "WYOMING", "ALASKA", "CALIFORNIA", "HAWAII", "OREGON",
"WASHINGTON")
```

```
# Aggregating bee data across regions

for(i in 1:nrow(bees_state)){

  if(bees_state$State[i] %in% Northeast){

    bees_state$Region[i] <- "Northeast"

  }else if(bees_state$State[i] %in% Midwest){

    bees_state$Region[i] <- "Midwest"

  }else if(bees_state$State[i] %in% South){

    bees_state$Region[i] <- "South"

  }else{

    bees_state$Region[i] <- "West"

  }

}

bees_region <- aggregate(bees_state$Population, by=list(Year=bees_state$Year,
Region=bees_state$Region), FUN=sum)

bees_region

bees_total <- aggregate(bees_region$x, by=list(Year=bees_region$Year), FUN=sum)


# Aggregating Weather data across regions


# Precipitation

for(i in 1:nrow(precip)){

  if(precip$State[i] %in% Northeast){

    precip$Region[i] <- "Northeast"

  }else if(precip$State[i] %in% Midwest){

    precip$Region[i] <- "Midwest"

  }else if(precip$State[i] %in% South){
```

```
    precip$Region[i] <- "South"
  }else{
    precip$Region[i] <- "West"
  }
}
precip <- aggregate(precip$Precipitation, by=list(Year=precip$YEAR, Region=precip$Region), FUN=sum)
precip <- rename(precip, c("Precipitation"="x"))
Precip


# Temperature


for(i in 1:nrow(temp)){
  if(temp$State[i] %in% Northeast){
    temp$Region[i] <- "Northeast"
  }else if(temp$State[i] %in% Midwest){
    temp$Region[i] <- "Midwest"
  }else if(temp$State[i] %in% South){
    temp$Region[i] <- "South"
  }else{
    temp$Region[i] <- "West"
  }
}
temp <- aggregate(temp$Temperature, by=list(Year=temp$YEAR, Region=temp$Region), FUN=mean)
temp <- rename(temp, c("Temperature"="x"))
```

```
temp


# Joining together weather data and bee data & creating a time series for the entire US.

weather <- inner_join(temp, precip, by = c('Year','Region'))

bees_weather <- inner_join(bees_region, weather, by = c('Year','Region'))

bees_weather.ts <- ts(bees_weather, start = c(1987), end = c(2017), freq = 1)


# Subsetting each region out of main data set

bees_midwest <- subset(bees_region, Region == "Midwest", select = c("Year", "x"))

bees_northeast <- subset(bees_region, Region == "Northeast", select = c("Year", "x"))

bees_south <- subset(bees_region, Region == "South", select = c("Year", "x"))

bees_west <- subset(bees_region, Region == "West", select = c("Year", "x"))


# Subsetting regions from weather data set

bees_weather_midwest <- subset(bees_weather, Region == "Midwest", select = c("Year", "x",
"Temperature", "Precipitation"))

bees_weather_northeast <- subset(bees_weather, Region == "Northeast", select = c("Year", "x",
"Temperature", "Precipitation"))

bees_weather_south <- subset(bees_weather, Region == "South", select = c("Year", "x",
"Temperature", "Precipitation"))

bees_weather_west <- subset(bees_weather, Region == "West", select = c("Year", "x",
"Temperature", "Precipitation"))


## Creating time series for each region

# Bee Data

bees.ts <- ts(bees_total$x, start = c(1987, 1), end = c(2017, 1), freq = 1)

midwest.ts <- ts(bees_midwest$x, start = c(1987, 1), end = c(2017, 1), freq = 1)
```

```
northeast.ts <- ts(bees_northeast$x, start = c(1987, 1), end = c(2017, 1), freq = 1)
south.ts <- ts(bees_south$x, start = c(1987, 1), end = c(2017, 1), freq = 1)
west.ts <- ts(bees_west$x, start = c(1987, 1), end = c(2017, 1), freq = 1)

# Weather Data
midwest_weather.ts <- ts(bees_weather_midwest, start = c(1987), end = c(2017), freq = 1)
northeast_weather.ts <- ts(bees_weather_northeast, start = c(1987), end = c(2017), freq = 1)
south_weather.ts <- ts(bees_weather_south, start = c(1987), end = c(2017), freq = 1)
west_weather.ts <- ts(bees_weather_west, start = c(1987), end = c(2017), freq = 1)

# Visualizing Past Populations
options(scipen=999)
x11()
ts.plot(bees.ts, gpars=list(xlab="Year", ylab="Total Population",
                            main="Historical Bee Populations (1987-2017)",
                            col="Yellow", lwd=3))

x11()
ts.plot(midwest.ts, northeast.ts, south.ts, west.ts,
        gpars=list(xlab = "Year", ylab = "Population",
                   main = "Bee Populations 1987-2017 by Region",
                   ylim=c(0, 1525000), col=1:4, lwd=3))
legend("topright", lty=1, col=1:4, box.lty=0, lwd=3,
       legend=c("Midwest", "Northeast", "South", "West"))
options(scipen=0)
```

```
# Creating Valid and Training data for each Region

nValid <- 4

nTrain <- length(south.ts) - nValid


## South

south.train.ts <- window(south.ts, start = c(1987, 1), end = c(1987, nTrain))

south.valid.ts <- window(south.ts, start = c(1987, nTrain + 1), end = c(1987, nTrain + nValid))


south.weather.train.ts <- window(south_weather.ts, start = c(1987, 1), end = c(1987, nTrain))

south.weather.valid.ts <- window(south_weather.ts, start = c(1987, nTrain + 1), end = c(1987,
nTrain + nValid))


# Northeast

northeast.train.ts <- window(northeast.ts, start = c(1987, 1), end = c(1987, nTrain))

northeast.valid.ts <- window(northeast.ts, start = c(1987, nTrain + 1), end = c(1987, nTrain +
nValid))


northeast.weather.train.ts <- window(northeast_weather.ts, start = c(1987, 1), end = c(1987,
nTrain))

northeast.weather.valid.ts <- window(northeast_weather.ts, start = c(1987, nTrain + 1), end =
c(1987, nTrain + nValid))


# Midwest

midwest.train.ts <- window(midwest.ts, start = c(1987, 1), end = c(1987, nTrain))
```

```
midwest.valid.ts <- window(midwest.ts, start = c(1987, nTrain + 1), end = c(1987, nTrain + nValid))


midwest.weather.train.ts <- window(midwest_weather.ts, start = c(1987, 1), end = c(1987, nTrain))

midwest.weather.valid.ts <- window(midwest_weather.ts, start = c(1987, nTrain + 1), end = c(1987, nTrain + nValid))


# West

west.train.ts <- window(west.ts, start = c(1987, 1), end = c(1987, nTrain))

west.valid.ts <- window(west.ts, start = c(1987, nTrain + 1), end = c(1987, nTrain + nValid))


west.weather.train.ts <- window(west_weather.ts, start = c(1987, 1), end = c(1987, nTrain))

west.weather.valid.ts <- window(west_weather.ts, start = c(1987, nTrain + 1), end = c(1987, nTrain + nValid))


# Plotting South Population Training and Validation

options(scipen=999)

plot(south.train.ts, ylim = c(400000, 800000), ylab = "Population", xlab = "Time",

    bty = "l", xaxt = "n", xlim = c(1987, 2018), main = "South Bee Population")

axis(1, at = seq(1987, 2018, 1), labels = format(seq(1987, 2018, 1)))


# Add Validation Data

lines(south.valid.ts)


# Section off Validation Data

lines(c(2017 - 3, 2017 - 3), c(0, 750000))
```

```
lines(c(2017, 2017), c(0, 750000))


# Add Text Labels
text(1999, 800000, "Training")
text(2015.5, 800000, "Validation")


### PREDICTIVE ANALYTICS PROJECT
## SIMPLE METHOD ANALYSIS
## Mean Method
south.mean.pred <- meanf(south.train.ts, h = nValid)
# Plot Population
plot(south.train.ts, ylim = c(400000, 800000), ylab = "Population", xlab = "Time",
    bty = "l", xaxt = "n", xlim = c(1987, 2018), main = "South Bee Population Mean Method")
axis(1, at = seq(1987, 2018, 1), labels = format(seq(1987, 2018, 1)))
# Add Mean Forecast
lines(south.mean.pred$mean, lwd = 2, col = "blue", lty = 2)
# Add Mean Fitted Values
lines(south.mean.pred$fitted, lwd = 2, col = "blue")
# Add Validation Data
lines(south.valid.ts, col = "grey20", lty = 3)
# Section off Validation Data
lines(c(2017 - 3, 2017 - 3), c(0, 750000))
lines(c(2017, 2017), c(0, 750000))
# Add Text Labels
text(1999, 800000, "Training")
```

```
text(2015.5, 800000, "Validation")
accuracy(south.mean.pred, south.valid.ts)

## Naive Method
south.naive.pred <- naive(south.train.ts, h = nValid)
# Plot Population
options(scipen=999)
plot(south.train.ts, ylim = c(400000, 800000), ylab = "Population", xlab = "Time",
    bty = "l", xaxt = "n", xlim = c(1987, 2018), main = "South Bee Population Naive Method")
axis(1, at = seq(1987, 2018, 1), labels = format(seq(1987, 2018, 1)))

# Add Naive Mean Forecast
lines(south.naive.pred$mean, lwd = 2, col = "blue", lty = 2)

# Add Naive Fitted Values
lines(south.naive.pred$fitted, lwd = 2, col = "blue")

# Add Validation Data
lines(south.valid.ts, col = "grey20", lty = 3)

# Section off Validation Data
lines(c(2017 - 3, 2017 - 3), c(0, 750000))
lines(c(2017, 2017), c(0, 750000))

# Add Text Labels
```

```
text(1999, 800000, "Training")
text(2015.5, 800000, "Validation")

accuracy(south.naive.pred, south.valid.ts)

## Drift Method

south.drift.pred <- rwf(south.train.ts, drift=TRUE, h= nValid)

# Plot Population
plot(south.train.ts, ylim = c(400000, 800000), ylab = "Population", xlab = "Time",
    bty = "l", xaxt = "n", xlim = c(1987, 2018), main = "South Bee Population Drift Method")
axis(1, at = seq(1987, 2018, 1), labels = format(seq(1987, 2018, 1)))

# Add Drift Mean Forecast
lines(south.drift.pred$mean, lwd = 2, col = "blue", lty = 2)

# Add Drift Fitted Values
lines(south.drift.pred$fitted, lwd = 2, col = "blue")

# Add Validation Data
lines(south.valid.ts, col = "grey20", lty = 3)

# Section off Validation Data
lines(c(2017 - 3, 2017 - 3), c(0, 750000))
```

```
lines(c(2017, 2017), c(0, 750000))

# Add Text Labels
text(1999, 800000, "Training")
text(2015.5, 800000, "Validation")
accuracy(south.drift.pred, south.valid.ts)

# Plot All Forecasts
plot(south.train.ts, ylim = c(400000, 800000), ylab = "Population", xlab = "Time",
    bty = "l", xaxt = "n", xlim = c(1987, 2018), main = "South Bee Population")
axis(1, at = seq(1987, 2018, 1), labels = format(seq(1987, 2018, 1)))

# Add Mean Forecast
lines(south.mean.pred$mean, lwd = 2, col = "blue", lty = 2)
# Add Naive Mean Forecast
lines(south.naive.pred$mean, lwd = 2, col = "green", lty = 2)
# Add Drift Mean Forecast
lines(south.drift.pred$mean, lwd = 2, col = "red", lty = 2)
# Add Validation Data
lines(south.valid.ts, col = "grey20", lty = 3)

# Section off Validation Data
lines(c(2017 - 3, 2017 - 3), c(0, 750000))
lines(c(2017, 2017), c(0, 750000))
```

```
# Add Text Labels

text(1999, 800000, "Training")

text(2015.5, 800000, "Validation")

# Add Legend

legend(1987, 500000, legend=c("Mean", "Naive", "Drift"),

       col=c("blue", "green", "red"), lty = 2)


# Find which model is most accurate

accuracy(south.mean.pred, south.valid.ts)

accuracy(south.naive.pred, south.valid.ts)

accuracy(south.drift.pred, south.valid.ts)


## Mean method is most accurate out of simple methods


## Time Series Regression Analysis

# Create the linear model for the South

south.weather.train.lm <- tslm(x ~ trend  + I(trend^2) + Temperature + Precipitation,
data=south.weather.train.ts)

south.weather.train.lm.pred <- forecast(south.weather.train.lm$fitted.values, h = nValid)


# Plot the Southern model

options(scipen=999)

plot(south.train.ts,  ylab = "Bee Population", xlab = "Time",

    bty = "l", xaxt = "n", ylim = c(350000,850000), xlim = c(1987, 2020), main = "South Bee
Population Regression")

axis(1, at = seq(1987, 2025, 1), labels = format(seq(1987, 2025, 1)))
```

```
lines(south.weather.train.lm.pred$fitted, lwd = 2, col = "blue")

lines(south.weather.train.lm.pred$mean, lwd = 2, col = "blue", lty = 2)

lines(south.valid.ts, col = "grey20", lty = 3)


# Check Southern Accuracy

accuracy(south.weather.train.lm.pred, south.valid.ts)


lines(c(2017 - 3, 2017 - 3), c(0, 800000))

lines(c(2017, 2017), c(0, 800000))


text(1999, 820000, "Training")

text(2015.5, 820000, "Validation")


# Create the linear model for the Midwest

midwest.weather.train.lm <- tslm(x ~ trend + I(trend^2) + Temperature + Precipitation,
data=midwest.weather.train.ts)

midwest.weather.train.lm.pred <- forecast(midwest.weather.train.lm$fitted.values, h = nValid)

# Plot the Midwestern model

plot(midwest.train.ts, ylab = "Bee Population", xlab = "Time",

    bty = "l", xaxt = "n", ylim = c(900000,1400000), xlim = c(1987, 2020), main = "Midwest Bee
Population Regresssion")

axis(1, at = seq(1987, 2025, 1), labels = format(seq(1987, 2025, 1)))

lines(midwest.weather.train.lm.pred$fitted, lwd = 2, col = "blue")

lines(midwest.weather.train.lm.pred$mean, lwd = 2, col = "blue", lty = 2)


# Add Validation Data
```

```
lines(midwest.valid.ts, col = "grey20", lty = 3)


# Section off Validation Data

lines(c(2017 - 3, 2017 - 3), c(0, 1350000))

lines(c(2017, 2017), c(0, 1350000))

# Add Text Labels

text(1999, 1380000, "Training")

text(2015.5, 1380000, "Validation")


# Compute midwest accuracy

accuracy(midwest.weather.train.lm.pred, midwest.valid.ts)


# Create the linear model for the West

west.weather.train.lm <- tslm(x ~ trend + I(trend^2) + Temperature + Precipitation,
data=west.weather.train.ts)

west.weather.train.lm.pred <- forecast(west.weather.train.lm$fitted.values, h = nValid)


# Plot the Western model

plot(west.train.ts, ylab = "Bee Population", xlab = "Time",

    bty = "l", xaxt = "n", ylim = c(800000,1400000), xlim = c(1987, 2020), main = "West Bee
Population Regresssion")

axis(1, at = seq(1987, 2025, 1), labels = format(seq(1987, 2025, 1)))

lines(west.weather.train.lm.pred$fitted, lwd = 2, col = "blue")

lines(west.weather.train.lm.pred$mean, lwd = 2, col = "blue", lty = 2)


# Add Validation Data
```

```
lines(west.valid.ts, col = "grey20", lty = 3)


# Section off Validation Data

lines(c(2017 - 3, 2017 - 3), c(0, 1350000))

lines(c(2017, 2017), c(0, 1350000))


# Add Text Labels

text(1999, 1380000, "Training")

text(2015.5, 1380000, "Validation")


# Compute West accuracy

accuracy(west.weather.train.lm.pred, west.valid.ts)




# Create the linear model for the Northeast

northeast.weather.train.lm <- tslm(x ~ trend + I(trend^2) + Temperature + Precipitation,
data=northeast.weather.train.ts)

northeast.weather.train.lm.pred <- forecast(northeast.weather.train.lm$fitted.values, h = nValid)


# Plot the Northeastern model

plot(northeast.train.ts, ylab = "Bee Population", xlab = "Time",

    bty = "l", xaxt = "n", ylim = c(80000,234000), xlim = c(1987, 2020), main = "Northeast Bee
Population Regresssion")

axis(1, at = seq(1987, 2025, 1), labels = format(seq(1987, 2025, 1)))

lines(northeast.weather.train.lm.pred$fitted, lwd = 2, col = "blue")

lines(northeast.weather.train.lm.pred$mean, lwd = 2, col = "blue", lty = 2)
```

```
# Add Validation Data
lines(northeast.valid.ts, col = "grey20", lty = 3)

# Section off Validation Data
lines(c(2017 - 3, 2017 - 3), c(0, 229000))
lines(c(2017, 2017), c(0, 229000))

# Add Text Labels
text(1999, 233000, "Training")
text(2015.5, 233000, "Validation")

# Compute West accuracy
accuracy(northeast.weather.train.lm.pred, northeast.valid.ts)

## ARIMA
# Creating Model for South
south.arima <- auto.arima(south.train.ts)
south.arima.forecast <- forecast(south.arima)

Acf(south.arima.forecast$residuals, lag.max = 10, ylab="ACF of Arima Residuals")

#Plotting the Model
options(scipen=999)
autoplot(south.arima.forecast, main="Forecasts from ARIMA(0,1,0)  -  South")
```

```
#Determining Accuracy
accuracy(south.arima.forecast, south.valid.ts)

# Checking the other Regions
northeast.arima <-auto.arima(northeast.train.ts)
autoplot(forecast(northeast.arima))

west.arima <- auto.arima(west.train.ts)
autoplot(forecast(west.arima))

midwest.arima <- auto.arima(midwest.train.ts)
midwest.arima.forecast <- forecast(midwest.arima)
autoplot(midwest.arima.forecast, main="Forecasts from ARIMA(2,2,0)  -  Midwest")
midwest.arima

Acf(midwest.arima.forecast$residuals, lag.max = 10, ylab="ACF of Arima Residuals")

accuracy(midwest.arima.forecast, midwest.valid.ts)


## CONCLUSION

# Use mean method to forecast bee population in 2018 and 2019 for all regions
```

```
## Midwest

midwest.mean.pred <- meanf(midwest.train.ts, h = 6)

# Plot Population
plot(midwest.train.ts, ylim = c(900000, 1300000), ylab = "Population", xlab = "Time",
    bty = "l", xaxt = "n", xlim = c(1987, 2020), main = "Midwest Bee Population Mean Method")
axis(1, at = seq(1987, 2020, 1), labels = format(seq(1987, 2020, 1)))

# Add Mean Forecast
lines(midwest.mean.pred$mean, lwd = 2, col = "blue", lty = 2)

# Add Mean Fitted Values
lines(midwest.mean.pred$fitted, lwd = 2, col = "blue")

# Add Validation Data
lines(midwest.valid.ts, col = "grey20", lty = 3)

# Section off Validation Data
lines(c(2017 - 3, 2017 - 3), c(0, 1250000))
lines(c(2017, 2017), c(0, 1250000))

# Add Text Labels
text(1999, 1300000, "Training")
text(2015.5, 1300000, "Validation")
```

```
text(2020, 1300000, "Future")


# Find Value

mean(midwest.valid.ts)



## Northeast


northeast.mean.pred <- meanf(northeast.train.ts, h = 6)


# Plot Population

plot(northeast.train.ts, ylim = c(75000, 225000), ylab = "Population", xlab = "Time",

    bty = "l", xaxt = "n", xlim = c(1987, 2020), main = "Northeast Bee Population Mean
Method")

axis(1, at = seq(1987, 2020, 1), labels = format(seq(1987, 2020, 1)))


# Add Mean Forecast

lines(northeast.mean.pred$mean, lwd = 2, col = "blue", lty = 2)


# Add Mean Fitted Values

lines(northeast.mean.pred$fitted, lwd = 2, col = "blue")


# Add Validation Data

lines(northeast.valid.ts, col = "grey20", lty = 3)


# Section off Validation Data
```

```
lines(c(2017 - 3, 2017 - 3), c(0, 200000))
lines(c(2017, 2017), c(0, 200000))

# Add Text Labels
text(1999, 225000, "Training")
text(2015.5, 225000, "Validation")
text(2020, 225000, "Future")

# Find Value
mean(northeast.valid.ts)



## South

south.mean.pred <- meanf(south.train.ts, h = 6)

# Plot Population
plot(south.train.ts, ylim = c(400000, 800000), ylab = "Population", xlab = "Time",
    bty = "l", xaxt = "n", xlim = c(1987, 2020), main = "South Bee Population Mean Method")
axis(1, at = seq(1987, 2020, 1), labels = format(seq(1987, 2020, 1)))

# Add Mean Forecast
lines(south.mean.pred$mean, lwd = 2, col = "blue", lty = 2)

# Add Mean Fitted Values
```

```
lines(south.mean.pred$fitted, lwd = 2, col = "blue")

# Add Validation Data
lines(south.valid.ts, col = "grey20", lty = 3)

# Section off Validation Data
lines(c(2017 - 3, 2017 - 3), c(0, 750000))
lines(c(2017, 2017), c(0, 750000))

# Add Text Labels
text(1999, 800000, "Training")
text(2015.5, 800000, "Validation")
text(2020, 800000, "Future")

# Find Value
mean(south.valid.ts)



## West

west.mean.pred <- meanf(west.train.ts, h = 6)

# Plot Population
plot(west.train.ts, ylim = c(800000, 1250000), ylab = "Population", xlab = "Time",
    bty = "l", xaxt = "n", xlim = c(1987, 2020), main = "West Bee Population Mean Method")
```

```
axis(1, at = seq(1987, 2020, 1), labels = format(seq(1987, 2020, 1)))

# Add Mean Forecast
lines(west.mean.pred$mean, lwd = 2, col = "blue", lty = 2)

# Add Mean Fitted Values
lines(west.mean.pred$fitted, lwd = 2, col = "blue")

# Add Validation Data
lines(west.valid.ts, col = "grey20", lty = 3)

# Section off Validation Data
lines(c(2017 - 3, 2017 - 3), c(0, 1200000))
lines(c(2017, 2017), c(0, 1200000))

# Add Text Labels
text(1999, 1250000, "Training")
text(2015.5, 1250000, "Validation")
text(2020, 1250000, "Future")

# Find Value
mean(west.valid.ts)
```