Capstone Project Modeling

Introduction to Data Science

Kevin Tajkowski

Library Calls

Load the necessary packages.

```
library(readr)
library(dplyr)
library(ggplot2)
library(lubridate)
```

Data Preparation

Load the data into a data frame. Remove all non-fish species by Family_Name. Non-fish species observed in the data set include sharks, rays, eels, turtles and a dolphin.

```
Little_Cayman <- read_csv("Little_Cayman.csv")
non_fish_family <- c("Requiem Shark", "Conger", "Dolphin", "Stingray",
"Manta", "Moray", "Eagle Ray", "Snake Eel", "Carpet Shark", "Hammerhead
Shark", "Electric Ray", "Round Stingrays", "Sea Turtles", "#N/A")
LC <- Little_Cayman %>% filter(!(Family_Name %in% non_fish_family))
```

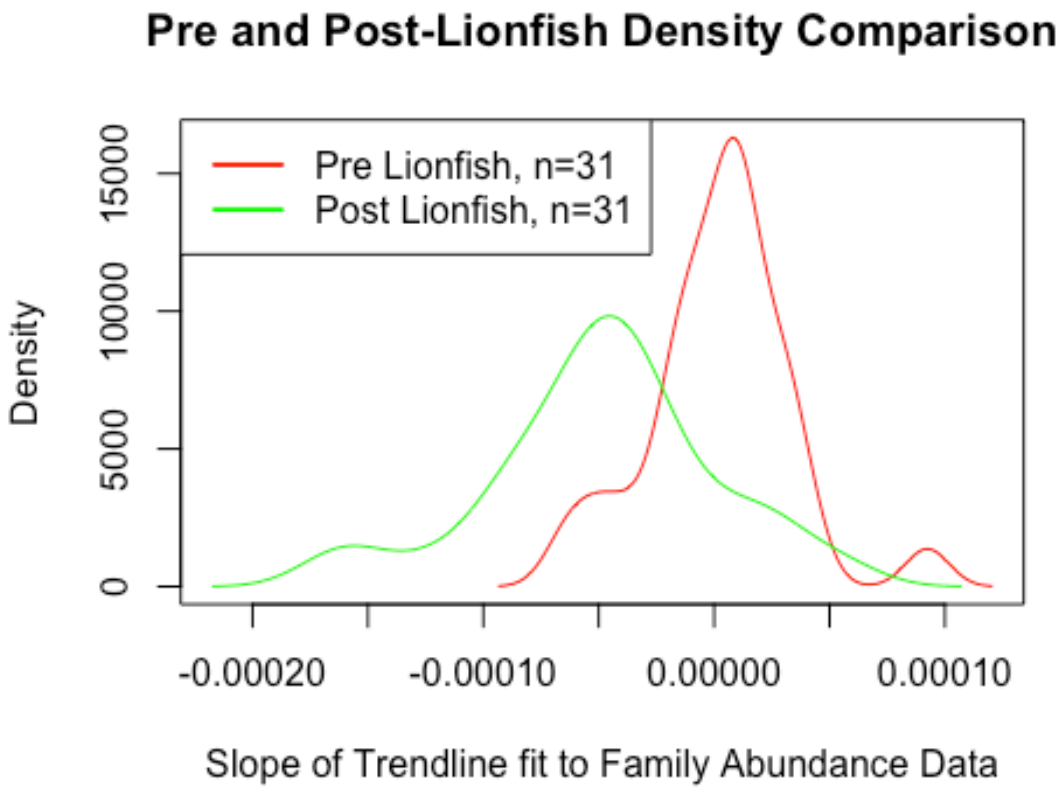
Data Frame Manipulation

Red lionfish are removed from the data set. Fish families with less than a prescribed number of surveys for all species in the family are identified and removed from the data set.

Density Comparison

'LC_Mod' excludes non-fish species, lionfish and fish families with < a certain number of surveys 'family_list' is a vector of family names This section of code generates two vectors of numerical values. The values are the slopes of trendlines that fit the abundance versus date plots. Two plots were generated for each family: before the first known lionfish sighting and after the first sighting. The vector 'pre_slope' corresponds to the plots of abundance versus date before the lionfish and the vector 'post_slope' corresponds to the plots after the lionfish sighting.

This section of code is used to generate a density plot of the pre_slope and post_slope vectors created above.



Moving Window Average Slope

Use a moving window to analyze the data from specific time frames progressing from the first survey date (March 27, 1994) to the last survey date available (April 17, 2018). The size of the window (number of years) and the time to shift the window (number of years) are variables.

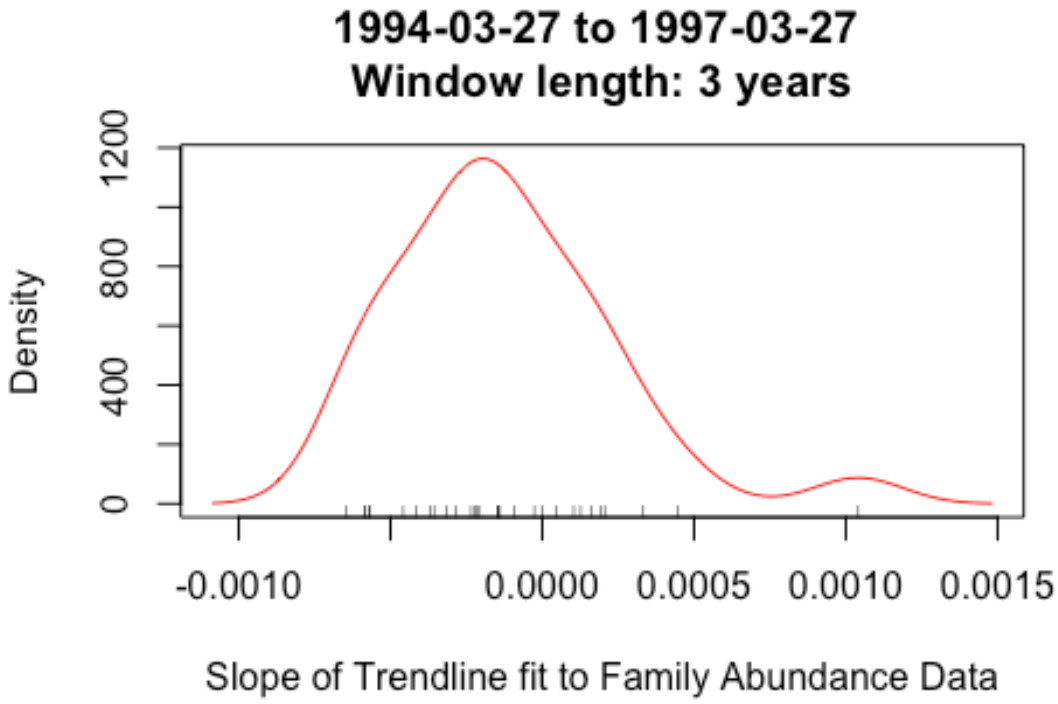
The following code is used to identify important dates and initialize several variables.

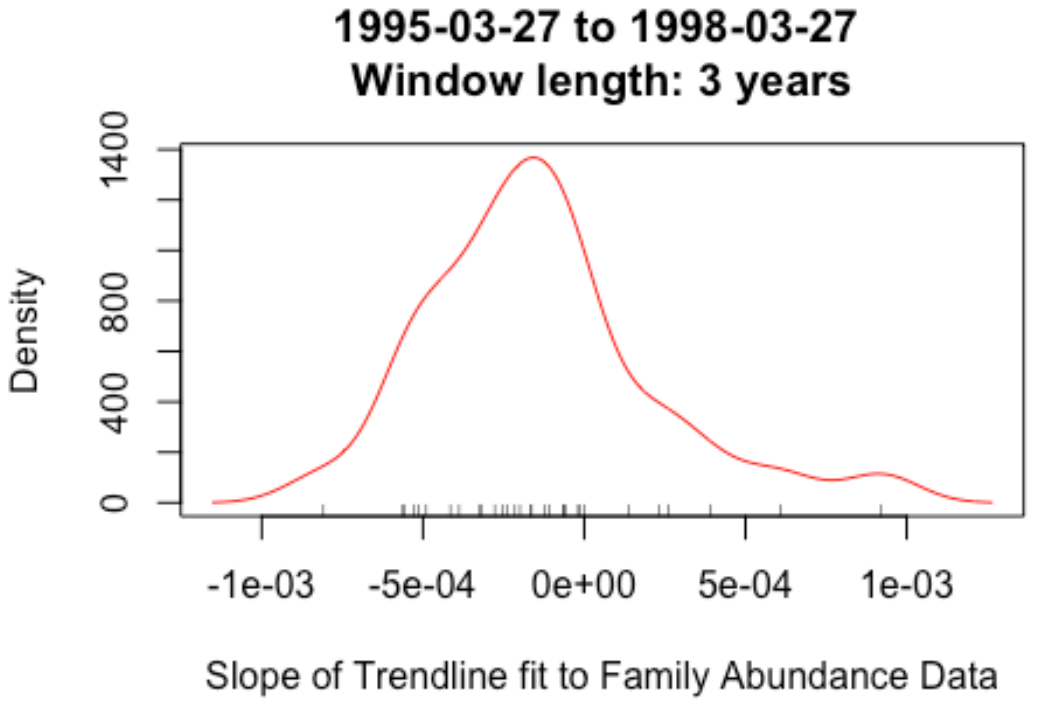
```
initial_date <- as.Date("1994-03-27") # Date that the first survey was
taken
final_date <- as.Date("2018-04-17") # Date of the last survey taken
shift <- 1 # Duration from beginning of window to the beginning of next
window (in years)
window <- 3 # Size (length of time) of window (in years) - Should be eq
ual to or larger than 'shift'

begin_date <- initial_date # initialize beginning of window to date of
first survey
end_date <- begin_date %m+% years(window) # initialize end of window

slope_DF <- data.frame(date = as.Date(NA), mean_slope = NA)

i <- 1
j <- 1
```

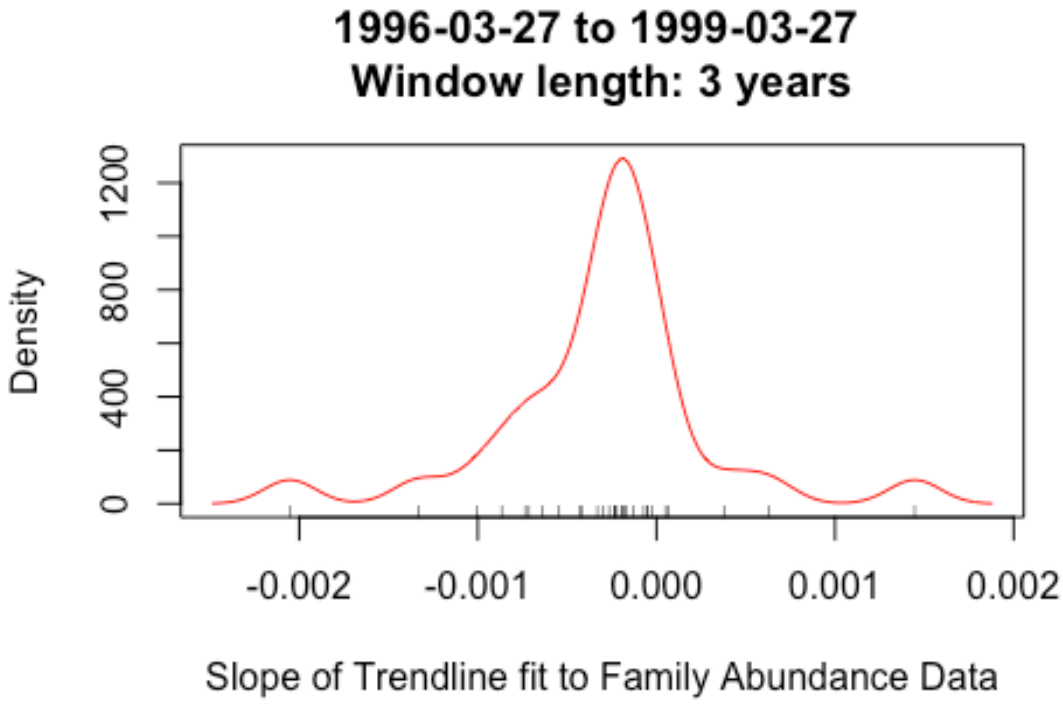
The following code is used to determine the average slope of all of the slopes of the trendlines for every family during a specific time period (the width of the window).

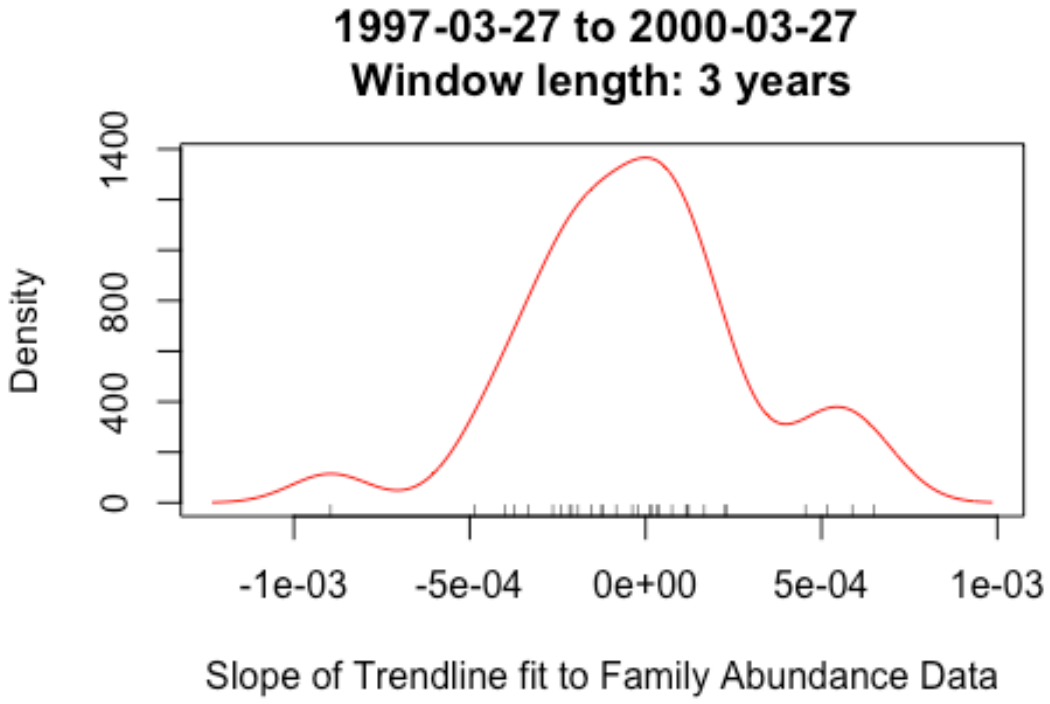
```
while(end_date <= final_date){
  current_window <- LC_Mod %>% filter(Date >= begin_date & Date < end_d
ate)
  slopes <- NA

  for(i in seq_along(family_list)){
    c_wind <- current_window %>% filter(Family_Name == family_list[i])
    TL_slope <- lm(c_wind$Abundance ~ c_wind$Date)
    slopes[i] <- TL_slope$coefficients["c_wind$Date"]
  }

  slope_DF[[j,1]] <- begin_date
  slope_DF[[j,2]] <- mean(slopes)
  j <- j+1

  d1 <- density(slopes)
  xlim <- range(d1$x)
  ylim <- range(d1$y)
  cols <- "red"
```

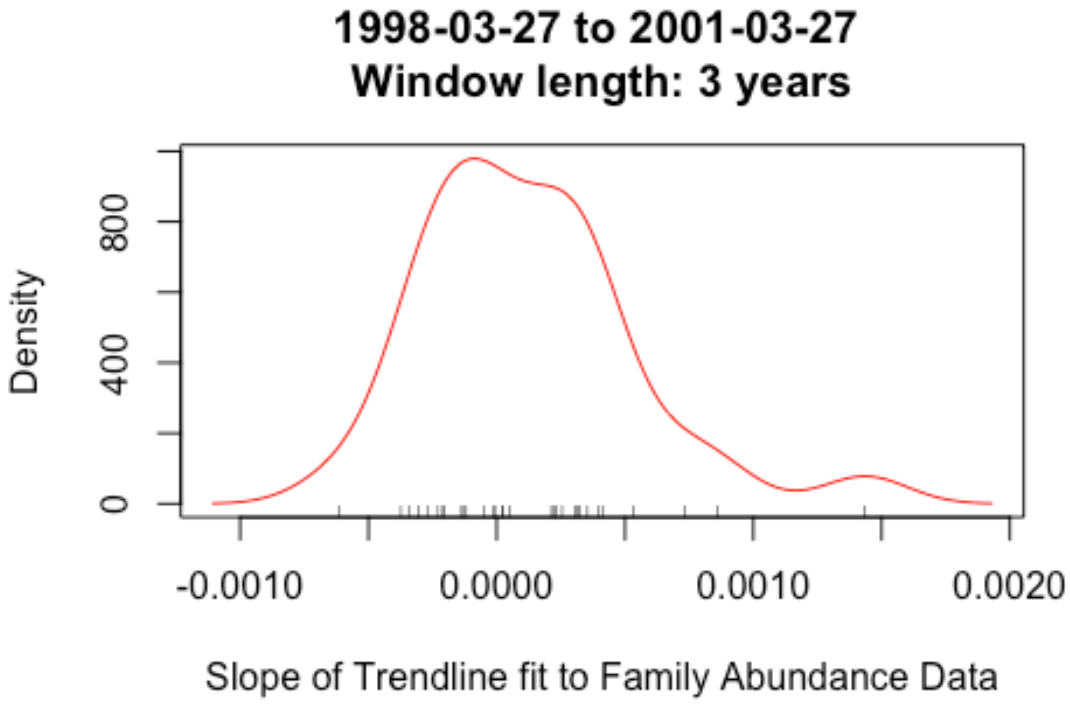
```
  plot(1, xlim=xlim, ylim=ylim, type='n', main=paste(begin_date, "to",
end_date, "\nWindow length:", window, "years"), xlab='Slope of Trendlin
e fit to Family Abundance Data', ylab='Density')
  rug(slopes)
  density(slopes, adjust = 0.5)
  lines(d1$x, d1$y, col=cols)

  begin_date <- begin_date %m+% years(shift)
  end_date <- begin_date %m+% years(window)
}
```

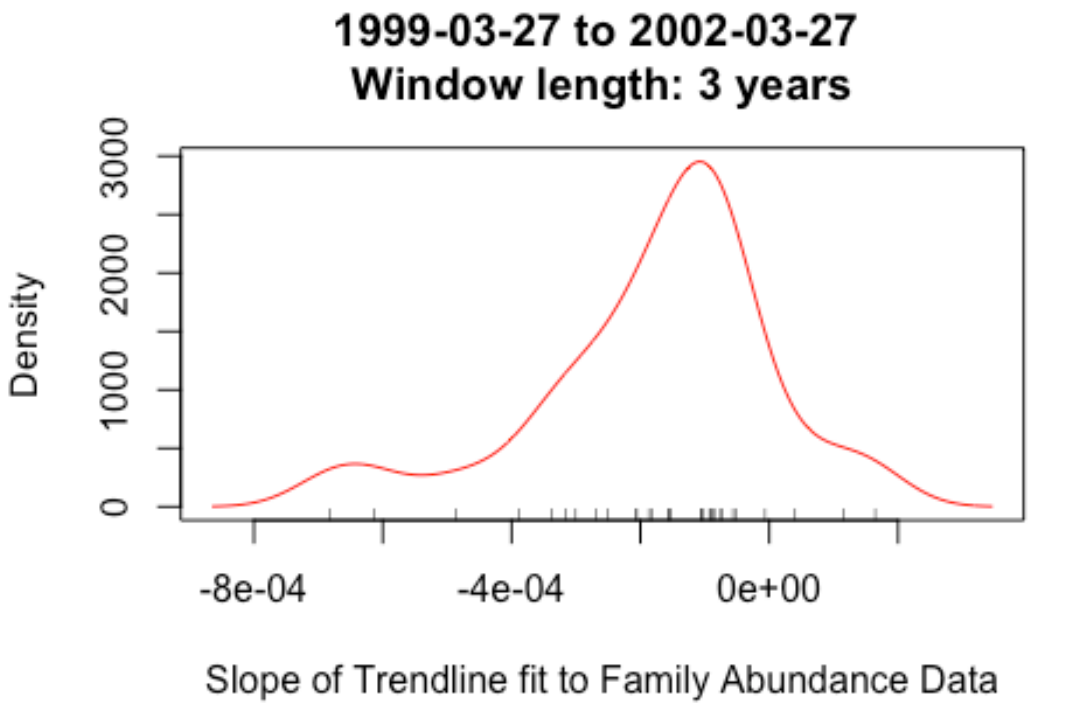

1994-03-27 to 1997-03-27
Window length: 3 years
Density
0
400
800
1200
-0.0010
0.0000
0.0005
0.0010
0.0015
Slope of Trendline fit to Family Abundance Data

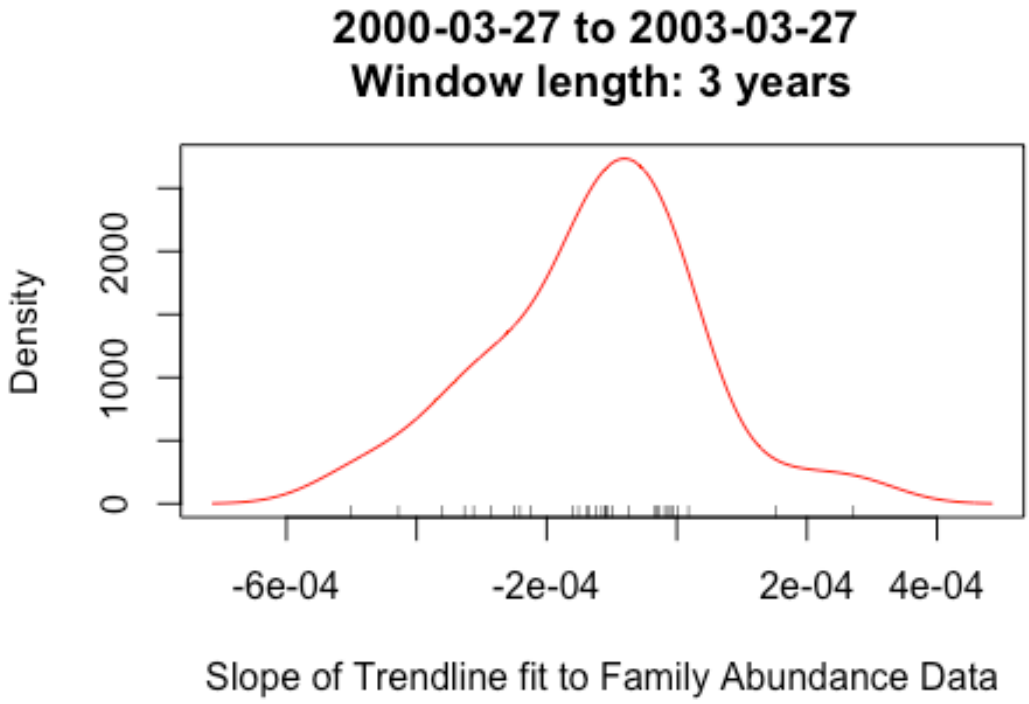

1995-03-27 to 1998-03-27
Window length: 3 years
Density
0
400
800
1400
-1e-03
-5e-04
0e+00
5e-04
1e-03
Slope of Trendline fit to Family Abundance Data

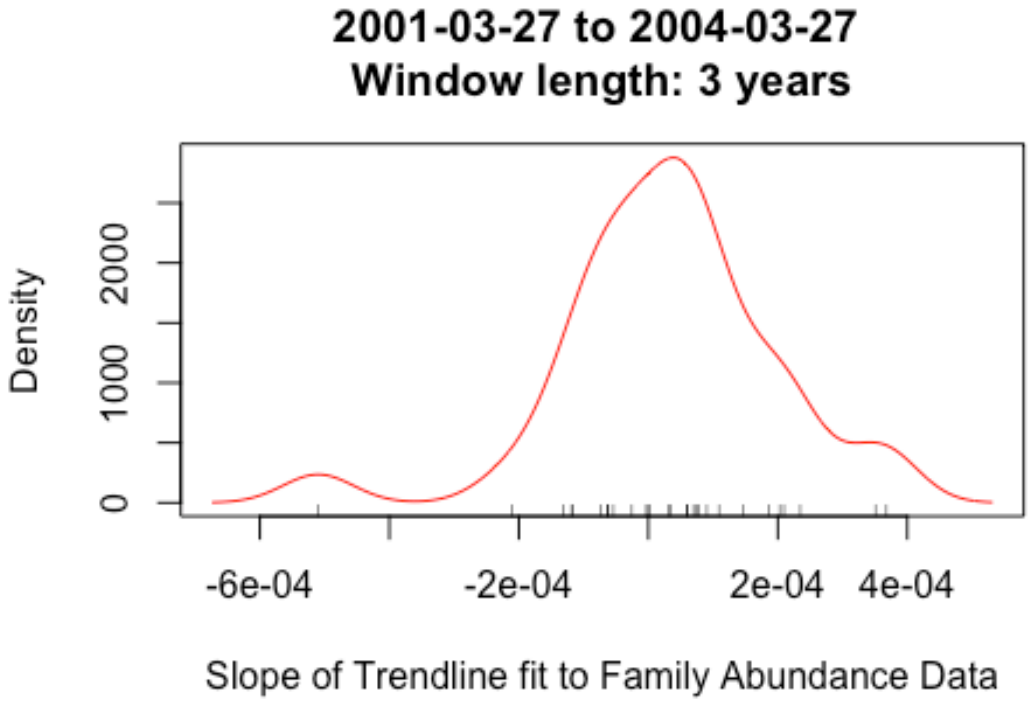

1996-03-27 to 1999-03-27
Window length: 3 years
Density
0
400
800
1200
-0.002
-0.001
0.000
0.001
0.002
Slope of Trendline fit to Family Abundance Data

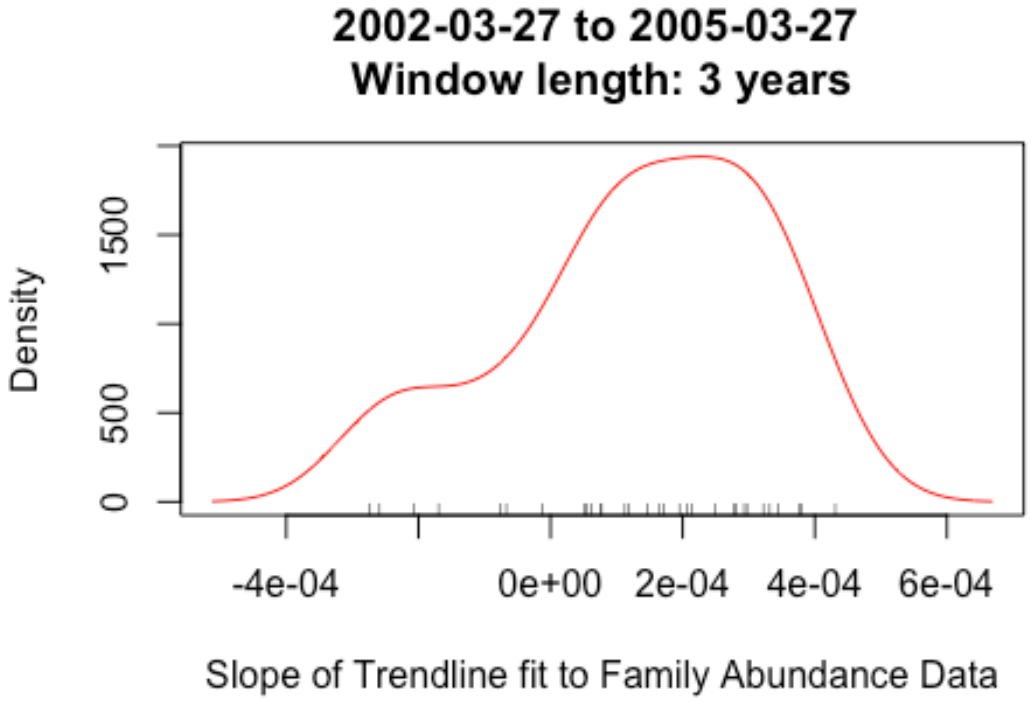

1997-03-27 to 2000-03-27
Window length: 3 years
Density
0
400
800
1400
-1e-03
-5e-04
0e+00
5e-04
1e-03
Slope of Trendline fit to Family Abundance Data

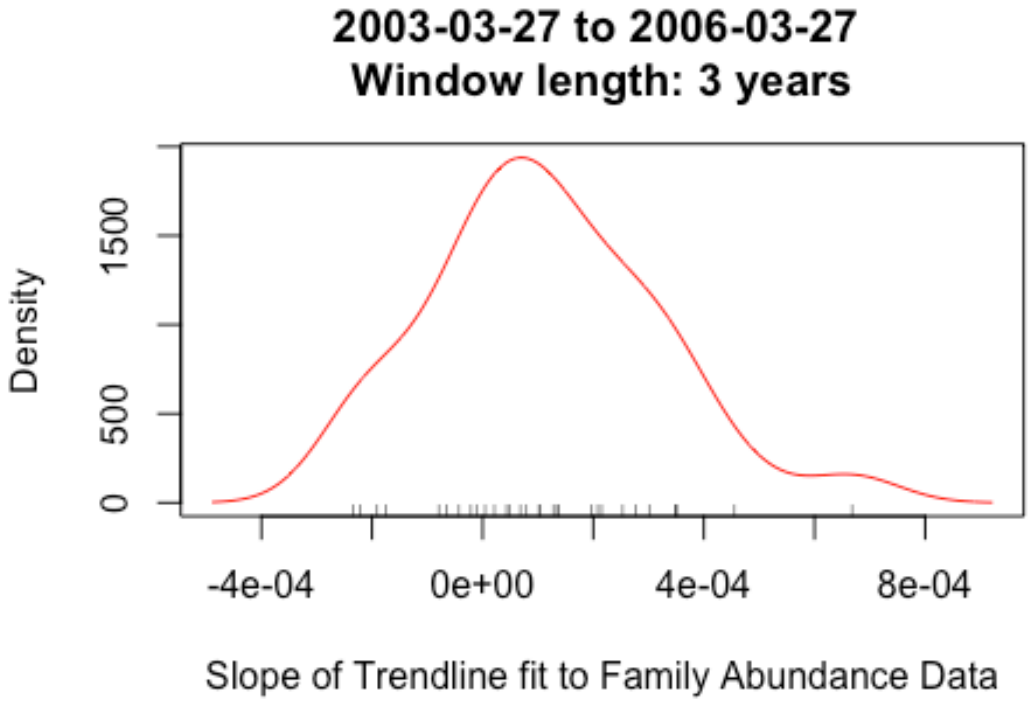

1998-03-27 to 2001-03-27
Window length: 3 years
Density
0
400
800
-0.0010
0.0000
0.0010
0.0020
Slope of Trendline fit to Family Abundance Data

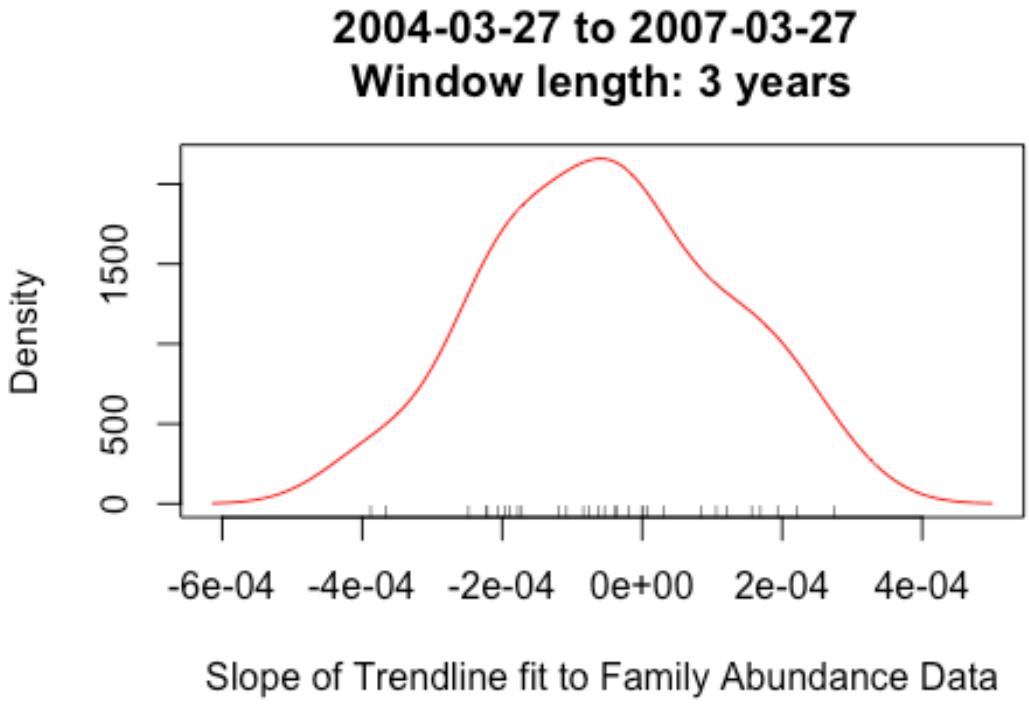

1999-03-27 to 2002-03-27
Window length: 3 years
Density
0
1000
2000
3000
-8e-04
-4e-04
0e+00
Slope of Trendline fit to Family Abundance Data

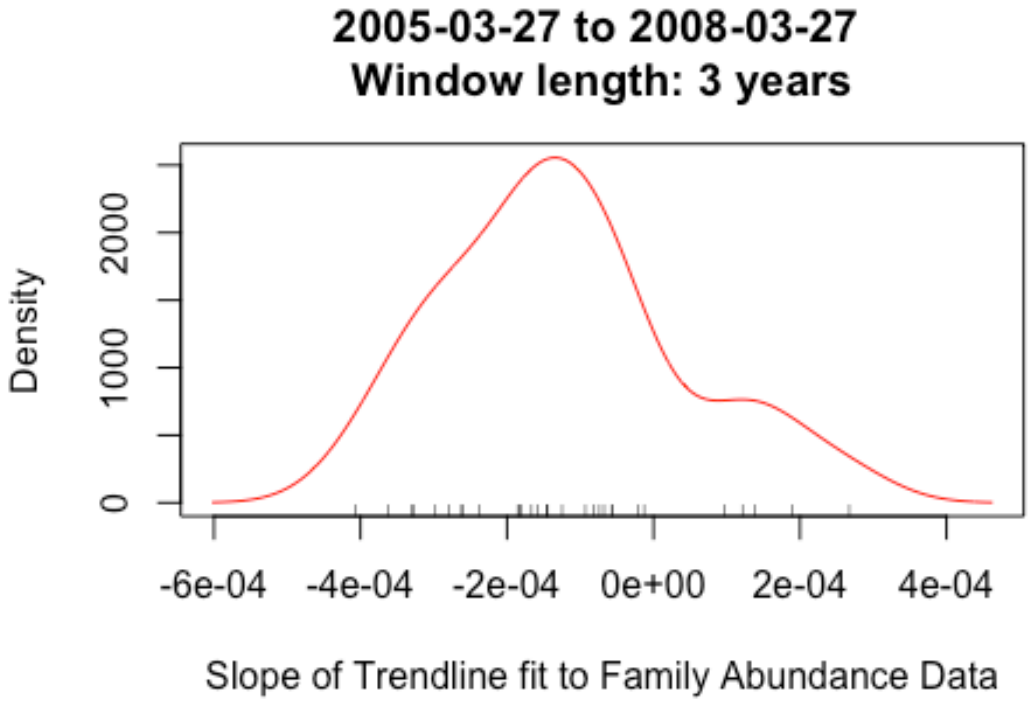

2000-03-27 to 2003-03-27
Window length: 3 years
Density
0
1000
2000
-6e-04
-2e-04
2e-04
4e-04
Slope of Trendline fit to Family Abundance Data

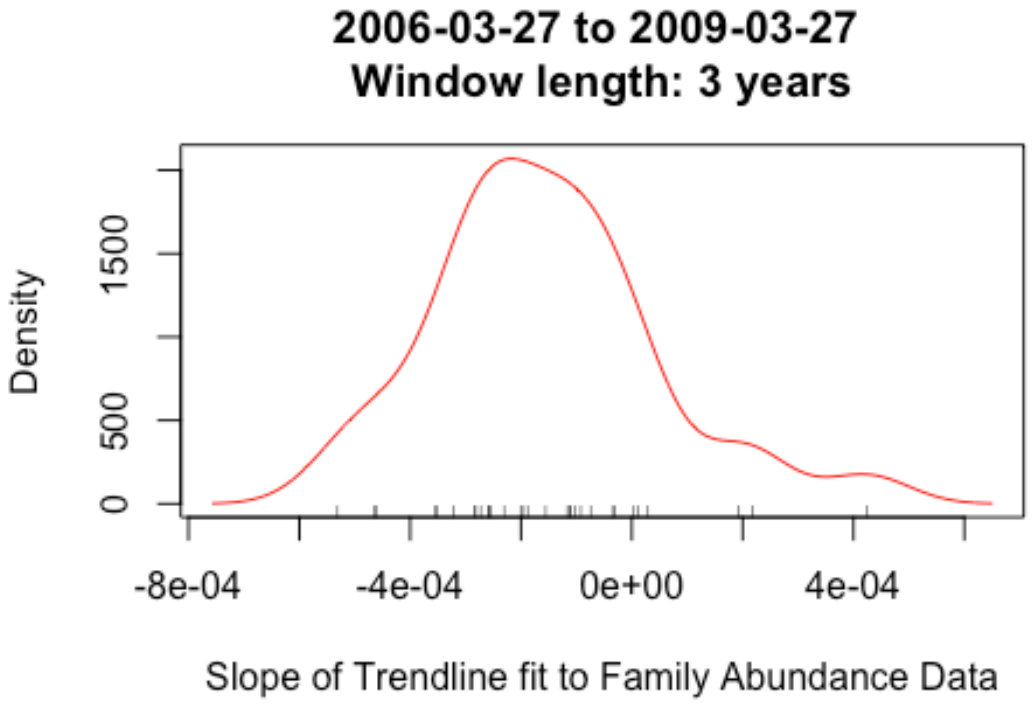

2001-03-27 to 2004-03-27
Window length: 3 years
Density
0
1000
2000
-6e-04
-2e-04
2e-04
4e-04
Slope of Trendline fit to Family Abundance Data

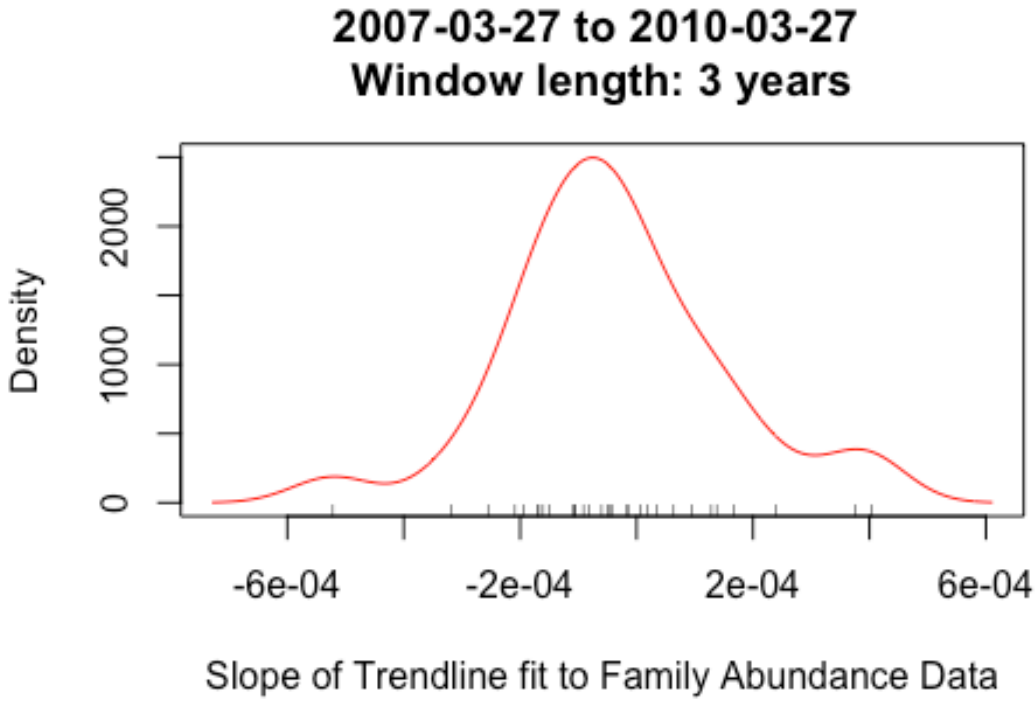

2002-03-27 to 2005-03-27
Window length: 3 years
Density
0
500
1500
-4e-04
0e+00
2e-04
4e-04
6e-04
Slope of Trendline fit to Family Abundance Data

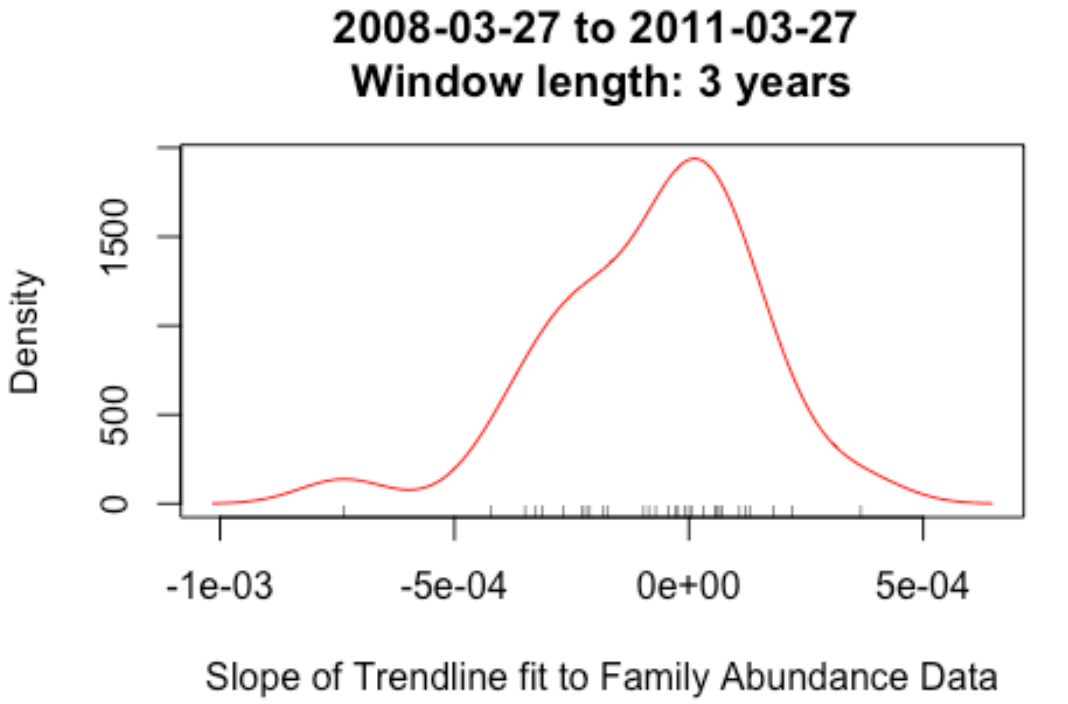

2003-03-27 to 2006-03-27
Window length: 3 years
Density
0
500
1500
-4e-04
0e+00
4e-04
8e-04
Slope of Trendline fit to Family Abundance Data

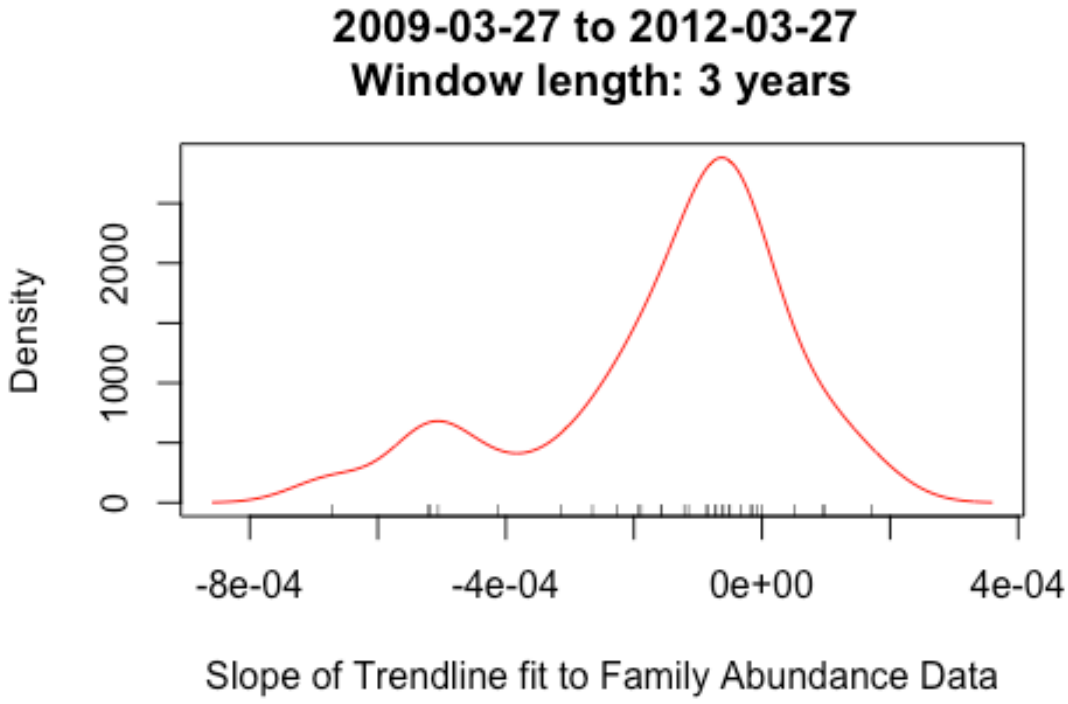

2004-03-27 to 2007-03-27
Window length: 3 years
Density
0
500
1500
-6e-04
-4e-04
-2e-04
0e+00
2e-04
4e-04
Slope of Trendline fit to Family Abundance Data

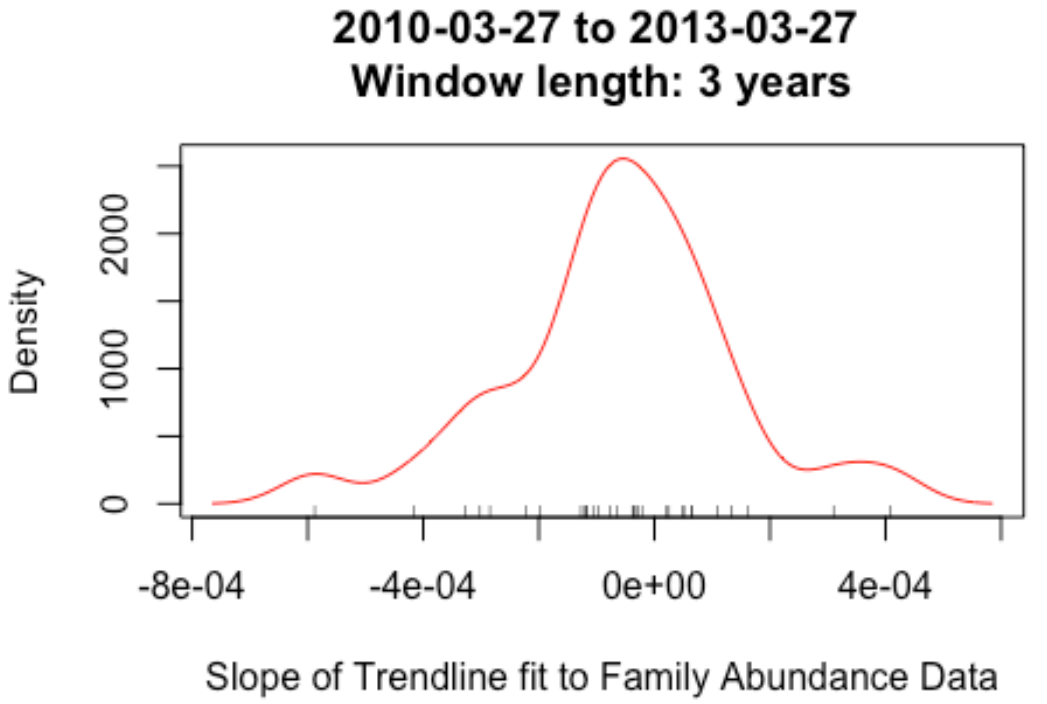

2005-03-27 to 2008-03-27
Window length: 3 years
Density
0
1000
2000
-6e-04
-4e-04
-2e-04
0e+00
2e-04
4e-04
Slope of Trendline fit to Family Abundance Data

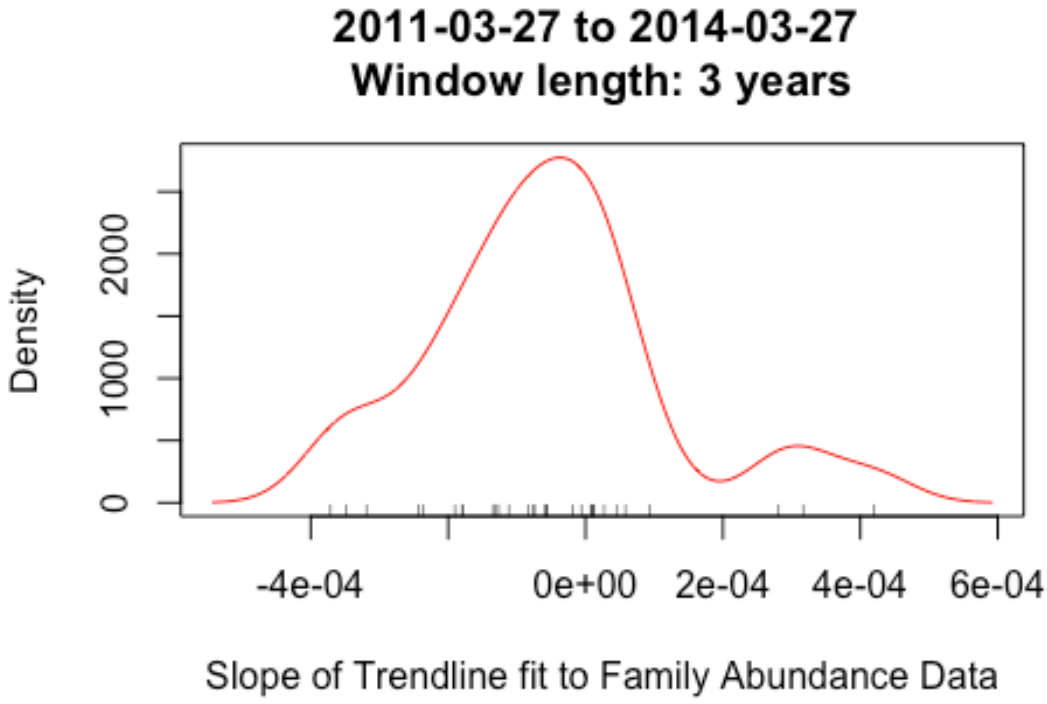

2006-03-27 to 2009-03-27
Window length: 3 years
Density
0
500
1500
-8e-04
-4e-04
0e+00
4e-04
Slope of Trendline fit to Family Abundance Data

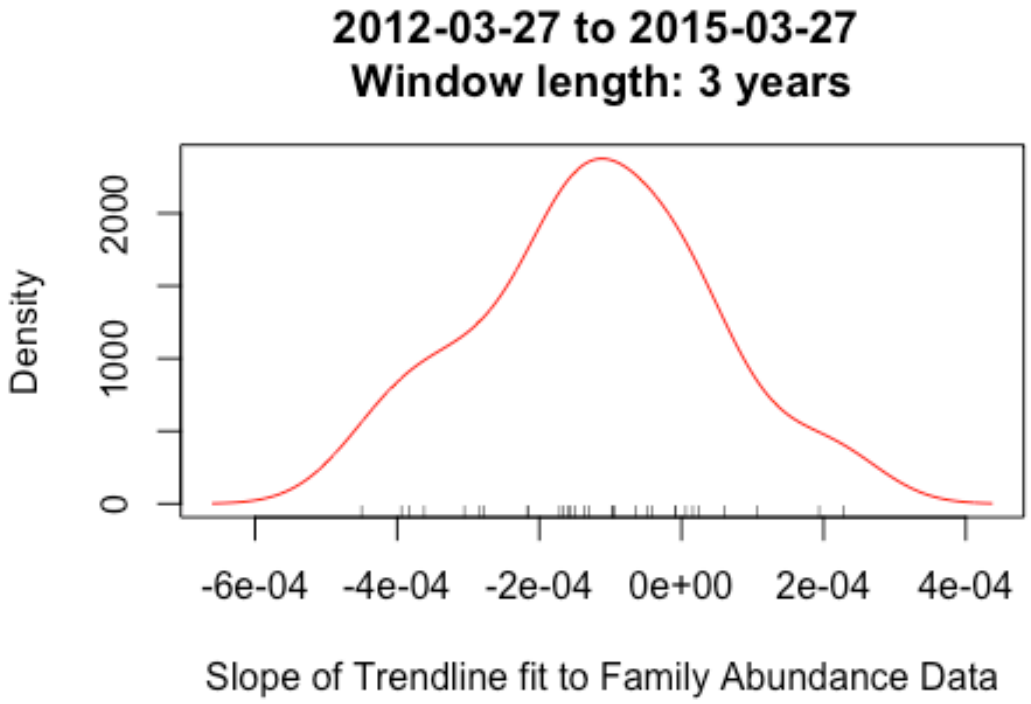

2007-03-27 to 2010-03-27
Window length: 3 years
Density
0
1000
2000
-6e-04
-2e-04
2e-04
6e-04
Slope of Trendline fit to Family Abundance Data

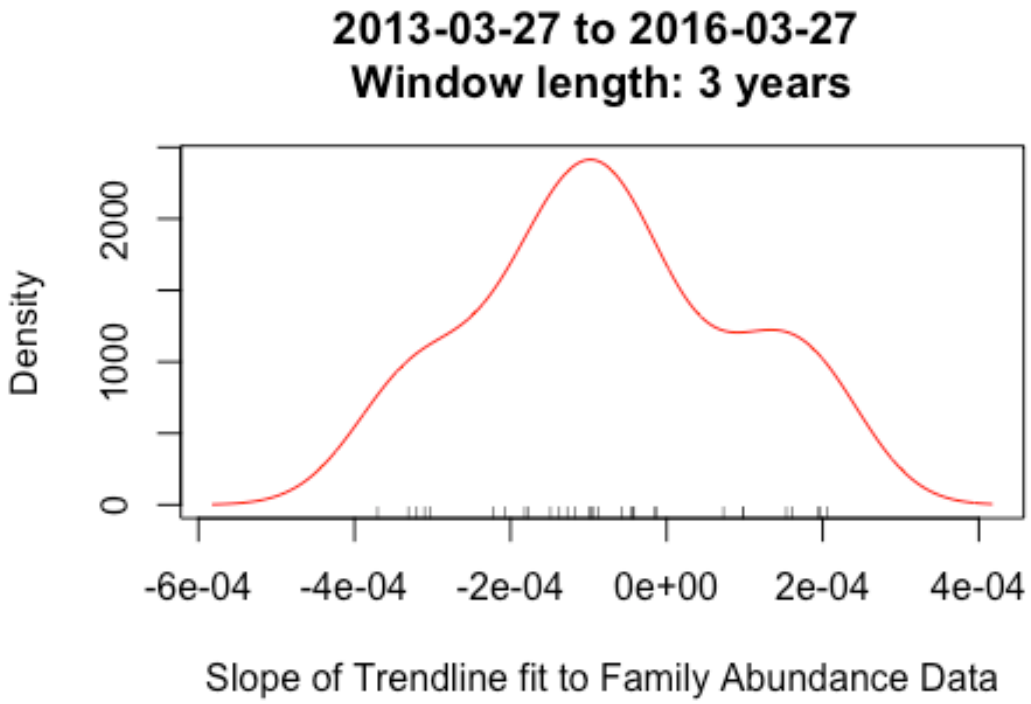

2008-03-27 to 2011-03-27
Window length: 3 years
Density
0
500
1500
-1e-03
-5e-04
0e+00
5e-04
Slope of Trendline fit to Family Abundance Data

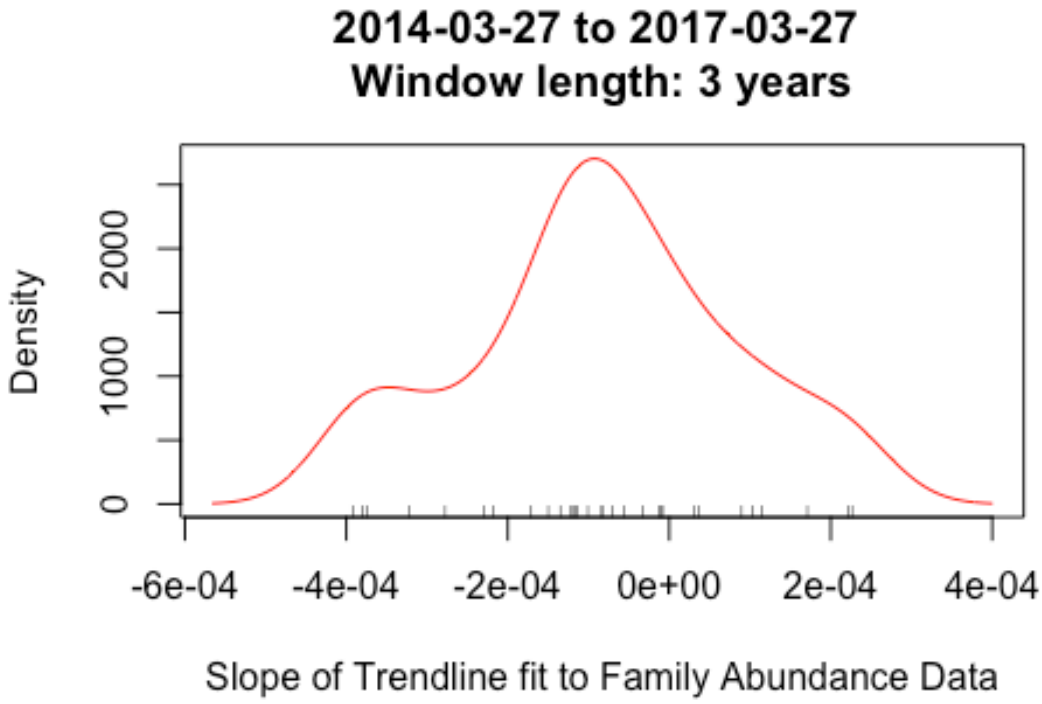

2009-03-27 to 2012-03-27
Window length: 3 years
Density
0
1000
2000
-8e-04
-4e-04
0e+00
4e-04
Slope of Trendline fit to Family Abundance Data

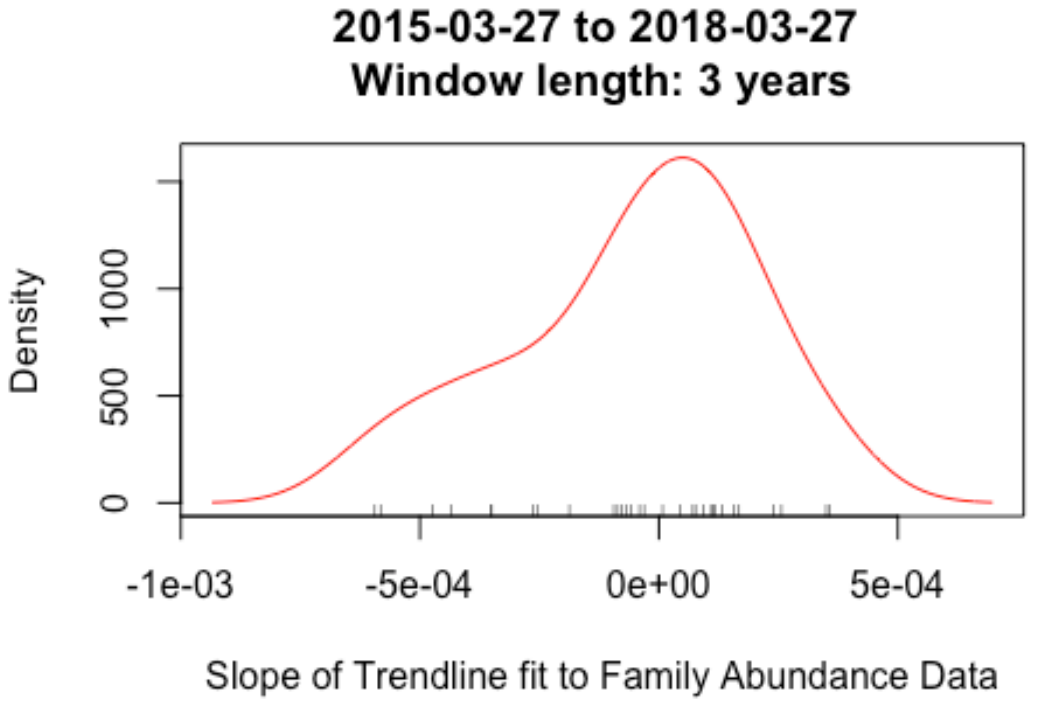

2010-03-27 to 2013-03-27
Window length: 3 years
Density
0
1000
2000
-8e-04
-4e-04
0e+00
4e-04
Slope of Trendline fit to Family Abundance Data


2011-03-27 to 2014-03-27
Window length: 3 years
Density
0
1000
2000
-4e-04
0e+00
2e-04
4e-04
6e-04
Slope of Trendline fit to Family Abundance Data


2012-03-27 to 2015-03-27
Window length: 3 years
Density
0
1000
2000
-6e-04
-4e-04
-2e-04
0e+00
2e-04
4e-04
Slope of Trendline fit to Family Abundance Data


2013-03-27 to 2016-03-27
Window length: 3 years
Density
0
1000
2000
-6e-04
-4e-04
-2e-04
0e+00
2e-04
4e-04
Slope of Trendline fit to Family Abundance Data


2014-03-27 to 2017-03-27
Window length: 3 years
Density
0
1000
2000
-6e-04
-4e-04
-2e-04
0e+00
2e-04
4e-04
Slope of Trendline fit to Family Abundance Data


2015-03-27 to 2018-03-27
Window length: 3 years
Density
0
500
1000
-1e-03
-5e-04
0e+00
5e-04
Slope of Trendline fit to Family Abundance Data

Generate a plot of the average of the slopes for the moving window.

```
first_LF <- as.Date("2009-02-17")

ggplot(slope_DF, aes(x = date, y = mean_slope)) +
  geom_point(size = 0.75) +
  geom_line() +
  ggtitle("Mean Slope of Trendlines for all Families") +
  theme(plot.title = element_text(hjust = 0.5, size = 12.0)) +
  geom_line(aes(slope_DF$date, 0), col = "green") +
  geom_vline(xintercept=first_LF, col = "red") +
  xlab(paste("Start Date of", window, "- year Window")) +
  ylab("Mean Slope") +
  theme_bw() +
  annotate("text", x = as.Date("2012-01-01"), y = 1.25e-04, label = "Fi
rst Lionfish", col = "red") +
  annotate("text", x = as.Date("2012-01-01"), y = 0.9e-04, label = "Sig
hting", col = "red")
```