Leaf Color Shift: Biostats Final Project Proposal

Kevin Buck, Kyle Elliott, Sydney Higgins

2023-04-10

Question 1

What data source(s) are you using? Harvard Forest Data Archive HF003-08

(a) Provide a website link and/or citation for the data source with a link to the data

link: Harvard Forest Mean Fall Phenology By Species

O'Keefe J. 2023. Phenology of Woody Species at Harvard Forest since 1990. Harvard Forest Data Archive: HF003 (v.35). Environmental Data Initiative: https://doi.org/10.6073/pasta/eb0dd36c6ec62a918340b6bda38be832.

(b) How were the data collected?

Data was collected annually since 1991 at a forest plot in Massachusetts. Weekly observations of percent leaf coloration and percent leaf fall started in September each year. Researchers observed when 50% of the leaves of trees of different species had changed color,and started to fall, and recorded those dates.

(c) Describe the data set (e.g. number of observations, number of variables, concerns with missing data)

The original dataset has 6 variables. We are only interested in year, species, and the Julian date of the year that 50% of the leaves have changed color. There is some missing in 1992, which is slightly concerning. There are 256 observation. Once concern is beyond 2002 only 14 species were studied (started with 33 spp). As such, we plan to only look at three species that were continued throughout the study (from 1991 to 2022: paper birch, red maple and red oak.

(d) Is there sufficient metadata (i.e. data about the data) for you to understand what the rows and columns of your data mean? Yes, the dataset is very well documented by the researchers with descriptions of protocol, possible sources of error, etc.

Question 2

What is your interesting scientific question using those data?

Question: Does species or year (or the interaction between the two) affect when leaf color change occurs?

Question 3

What kind of graphics would be most effective in displaying your data?

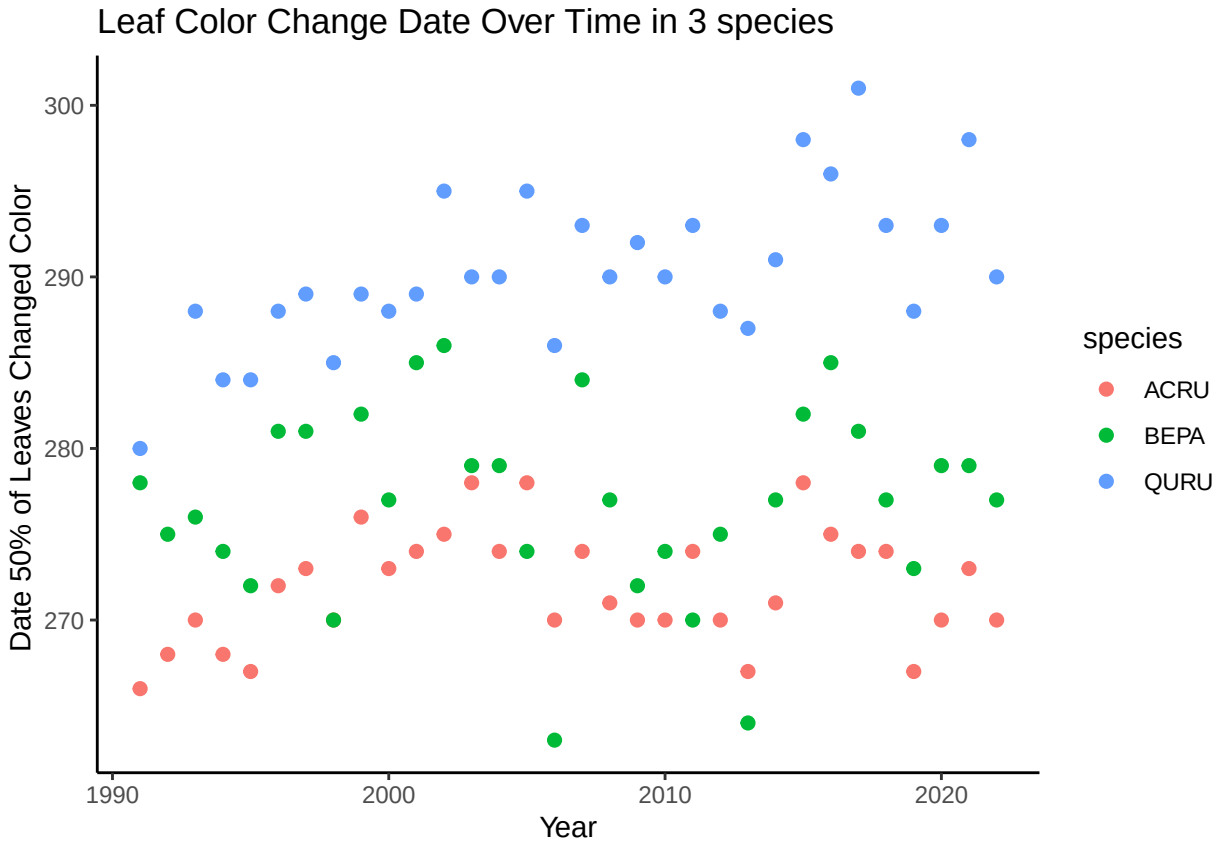
We want to start with a comparison between tree species and the mean date 50% of the leaves have turned, which we can use a mean box plot to demonstrate trends grouped by species. We can also do a scatter plot to look at the year on the x axis, day on the y axis, and species as the color. This will clearly show any differences between species and any long-standing trends over time.

(a) Provide at least one example using code, ideally something that supports points 1.d and 1.e, above. This is just an example of our regular admonition to "Always graph your data

Below is the code for a scatter plot which shows how the day that 50% of the leaves for a given species have changed color varies over the years of 1991 to 2022. We will also add a trend line to the data in our analysis

```
pheno <- read.csv(file="hf003-08-fall-mean-spp.csv") #read in the data
pheno %>% #pipe in data
  #filter in correct species
  filter(species=="ACRU"|species=="BEPA"|species=="QURU") %>%
  ggplot(aes(x = year, y = lc_doy)) + #set year as x axis and julian day of change as the y
  geom_point(aes(color=species, fill=species), size=2) +
  ggtitle("Leaf Color Change Date Over Time in 3 species")+ #add title
  xlab("Year") + #label x axis
  ylab("Date 50% of Leaves Changed Color ") + #label y axis
  theme_classic() #changing background
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



Question 4

What kind of statistical model do you think would be appropriate to analyze your data in order to inform your scientific question? Note that we expect that the final analyses in many full proposals will deviate from the planned analyses in the proposal. At this point, we're only asking you to propose plausible suggestions of analyses that could address your scientific questions.

Our group will use a multi-factor ANOVA to analyze the effect of species and year on leaf color change day. This is best because both of our variables are categorical.

For post-hoc analysis, we will use Tukey-Tests to make pairwise comparisons.

(a) What is your response variable and what are your explanatory variables (at least two)? State whether you are each of your explanatory variables as numerical or Categorical. Response Variable: leaf color turn date (numerical) Explanatory Variables: year (categorical), species (categorical)

(b) You should follow the guidance in the 2016 American Statistical Association Statement on p-values (Make sure to read to the end of the text on that link). Okay we will!

(c) Do you have concerns about meeting the assumptions of the statistical model you are choosing (e.g. linearity, independence, homogeneity of variance, normality in residuals, etc.)? If so, how do you plan to address those concerns? At the moment, we do not have concerns. Linearity seems to be met, as any trends in change of color date over time would be unidirectional. Variance should be equal, as all the individual trees are experiencing the same conditions bc they are in the same plots. We are assuming that the residuals are normal, and can transform the data logistically if necessary to deal with violations of that assumption. The measurements are certainly independent, as they are collected on different species on different dates in each year.

Question 5

If your dataset comes from a previously published dataset, describe the original analysis and how your proposed analysis differs from the original work.

The original dataset doesn't have analysis attached to it, it was simply collected and published. Other researchers have done analysis with this data in order to predict the effects of climate change on phenological shift, but there is not much emphasis on interspecific variation, which is where our analysis is unique.

Question 6

Since you will be working collaboratively, and we expect every team member to contribute, you should include a short "anticipated contribution statement," which identifies the tasks different team members expect to make to the project. Since this is a statistics class, we encourage each team member to find a way to contribute to the analysis. This contribution statement will be revised for the final project submission to reflect the actual contribution of each team member

Our Project will have three phases, and each group member will do one of them

Kevin - Data tidying and Statistical Analysis ~in code~ Kyle - Graph Production (what graphs make this clear) Sydney - Results (answers in text!!)

If there is a hiccup, we can all help the person who is struggling! solidarity

Question 7

(7) As in all academic assignments, it is important to properly reference the sources of ideas, data, & analysis. This practice (a) helps strengthen your argument by providing outside references, and (b) guards against plagiarism. You can use any standard bibliographic format. If you include any references apart from the original study, there should be few of them, however. This goal of this analysis is a demonstration of your analytical creativity, not a literature review. We are not expecting you to do any extra reading for this assignment, but if you do take ideas or text from somewhere, you need to make the appropriate attribution. The entire proposal pdf, including figures, citations, etc, should be 3 pages or less, single spaced, with 11 point font. Note: there are ways to adjust the size of your figures in your Rmarkdown PDF that you may want to consider to efficiently use space (see link).

Data Citation:

O'Keefe J. 2023. Phenology of Woody Species at Harvard Forest since 1990. Harvard Forest Data Archive: HF003 (v.35). Environmental Data Initiative: https://doi.org/10.6073/pasta/eb0dd36c6ec62a918340b6bda38be832.