UDACITY

Udacity Data Analyst Nanodegree

Create a Tableau Story:

Data Cleaning and Visualization Report

Written by:

Anastasia Makarevich

Date: March, 11, 2019

Introduction

For this project I took a dataset that contains the data on distribution and host range of Chytridiomycosis in Australia and covers the period from 1956 to 2007. The full description of the dataset can be found here:

http://www.esapubs.org/archive/ecol/E091/108/metadata.htm

The goal of the project is to explore what species and what categories are most affected by the fungus, how widespread is the disease and what is the general trend if there is one.

Link to Tableau Story

Initial version:

https://public.tableau.com/profile/anastasia7889#!/vizhome/Chytridiomycosis/Chytridiomycosis

Revised version:

https://public.tableau.com/profile/anastasia7889#!/vizhome/ChytridiomycosisinAustraliaFinal/Chytridiomycosis

Data Gathering and Cleaning

For this project I had to gather additional data so that we can get more insights. In first place, I extracted the genus for each species and then family for each genus. Next, I checked each of the species in the dataset and found it's conservations status on Wikipedia. I also extracted full names of the states to make it more readable. Lastly, I extracted individuals's life stage and gender and split it into two columns. Gathering and cleaning notebook is available here:

https://github.com/AnaMakarevich/DAND/blob/master/Projects/

Project_8_Chytridiomycosis/Chytridiomycosis_Data_Cleaning.ipynb

Summary

The story explores how the data about the diseases was gathered and what methods were used. It shows what species are the most endangered – species from Myobatrachidae family have the status of critically endangered and there were 949 species examined during the survey period (1956-2007). We also discovered that most of the infected species are adult males and that most of the infected frogs were detected in Queensland.

We concluded that the general trend is inconsistent based on the moving average plot.However, there were much fewer field investigations in years 2006 and 2007. One interesting finding is the growing percentage of infected frogs among endangered and critically endangered species:

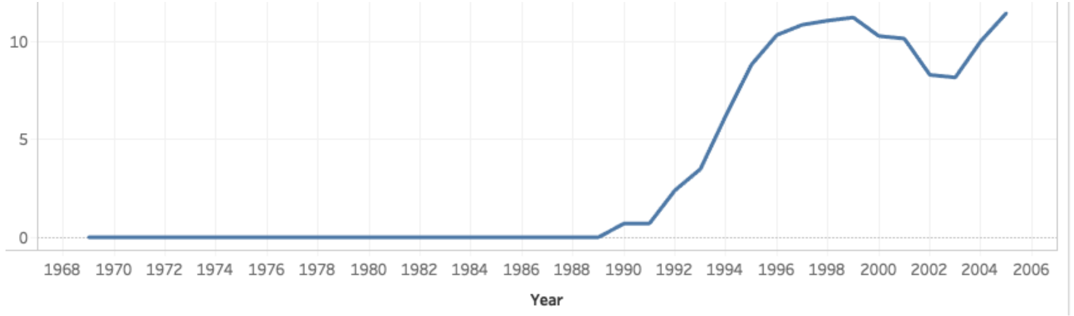

Fig. 1 – Moving average over 5-year period for % of infected individuals among endangered and critically endangered species.

Also, what we've seen suggests that future research is required to confirm any findings. The data is unbalanced in terms of the diagnosis methods, number of samples per year and the level of accuracy.

Design

Overview

The story was designed in a way that the reader's attention is focused on one thing at a time only. Each dashboard is focused on only one aspect of the story. I also added multiple text boxes to help the reader navigate through the story. The topic is very new to me and I had to dig a lot to understand the details. I assumed that it would be hard for the reader to understand the core ideas and provided the context to enable him to reason about the issue.

Bar Plots

The main type of chart that I used was bar plot. This is one of the most convenient plots when it comes to summarizing categorical data. It's typically used to count the number of records per category, so only one column is used. However, for this project, I already had some pre-summarized data (number of individuals per record), so I used this column to calculate the bar heights. I tried various type of visualization for this data – heatmaps, packed bubbles and pie chart, but ended up with simple, but reliable bar plot. Also, I learnt from the lectures that humans better respond to difference in height rather than in size and color. For that same reason, I used stacked bar plots when adding another level of detail (e.g. showing each bar's structure by gender).

Many initial bar plots in the dataset were heavily skewed because of the dominance of certain types of species (e.g. Litoria genus or species with the conservation status 'Least Concern`) - in these cases I provided filters, so that the viewer can exclude the dominating category. Also, in order to make the plots more readable, I used logarithmic scale for heavily skewed bar plots. It made the plots much more appealing. I also noted in the plot title that this is the logarithmic scale to avoid confusion. Examples: see story points: "What is the most vulnerable category", "What percentage of endangered species is infected?", "What is conservation status and why it matters?".

Map

I used map to convey spatial information about the data and also set map as a filter for the bar chart that shows the percentage of infected individuals by year. I used line chart for that as well and added labels to make it easier to compare yearly data.

Line Chart

I used line chart to show the dynamics. Line charts often work well for temporal data and for showing the trend. I also used smoothing technique to show the moving average rather than raw numbers.

Coloring

I used colorblind palette for all colorings. In general, I left the default color assignments – when the color was used only to distinguish between categories. However, for the conservation status I manually edited the colors (while still staying within the colorblind palette), so that the color corresponds to the category. For example, blue is a calm color, so I used it for 'Least Concern' species. Dark orange (close to red) is intense color that often means danger, so I used it encode 'Critically endangered' species. 'Endangered' category has less intense orange color. Also, the colors are consistent throughout the dashboards and mean the same thing for 'Conservation Status' on each plot.

Feedback

I asked my peers that I met during the Challenge course for feedback and they noticed that:

- the colors were inconsistent with the meaning, and I changed the color scheme, so that it's obvious that 'redness' increases with as the conservation status of the species increases.

- Moving average plot: it was showing total number of examined individuals rather than the both the number of examined and number of infected. As a result, I also decided to include moving average for the percentage of infected amphibians as well.

- My last slide suggested that there were species resistant to the disease, so one of my peers asked if there were species that were never diagnosed with Chytridiomycosis. So I created another dashboard that shows most affected species and most unaffected.

- I was told that some plots are hard to interpret due to some categories dominating over the other. My solution was to put them on logarithmic scale which significantly improved the visuals.

- Just a few seconds after submitting the project for the second time, I've got feedback from one of the mentors – it was suggested to change bar plot to line chart for the dashboard 'How was the disease spread by state over the years?'

- Lastly, the project reviews suggested that I should enhance the report by adding rationale for using the types of charts that I used. So the 'Design' section of this report was updated.

Resources

1. Disease Overview: http://wildlife.ohiodnr.gov/portals/wildlife/pdfs/species%20and%20habitats/chytrid.pdf

2. Wikipedia on Chytridiomycosis: https://en.wikipedia.org/wiki/Chytridiomycosis

3. Chytridiomycosis on Amphibiaweb:

https://amphibiaweb.org/chytrid/chytridiomycosis.html

4. Original Dataset Metadata:

http://www.esapubs.org/archive/ecol/E091/108/metadata.htm

5. First Documented Exctinction by Infection

https://www.researchgate.net/publication/29463188_The_Decline_of_the_Sharp-Snouted_Day_Frog_Taudactylus_acutirostris_The_First_Documented_Case_of_Extinction_by_Infection_in_a_Free-Ranging_Wildlife_Species

6. Detailed Report on the Disease

https://www.cabi.org/ISC/datasheet/109124

7. Chytridiomycosis causes catastrophic organism-wide metabolic dysregulation including profound failure of cellular energy pathways:

https://www.researchgate.net/publication/325426490_Chytridiomycosis_causes_catastrophic_organism-wide_metabolic_dysregulation_including_profound_failure_of_cellular_energy_pathways