AUTHOR1

AUTHOR2

The Power of Storms

Research Questions

1. How has the average storm intensity changed over time?

   For this question we will be computing highest wind speed for each storm, then calculate the average wind speed for each year, which we'll call the intensity. We will then plot the intensity against the year. We want to look for trends in tropical storms over time.

   Storm intensity appears to have increased from 1950 to present. We believe that the data from 1850 to 1950 is incomplete.
2. How has the frequency of large storms changed over time?

   We will calculate how many storms there have been per year, and how many of those storms are category three of higher (>95 knot wind speeds).

   The number of large storms per year has increased from around 10 in 1950 to 25 recently. The ratio of large storms to small storms has also increased from around 10% in 1950 to around 25% recently.
3. Which basin has the highest intensity storms?

   We will calculate the highest wind speed for each storm, then group by basin and take the average for each year. We will then graph intensity against time for each basin and compare basins. We want to find out which basins have the worst storms, and if the basin with the worst storms has changed over time.

   There is not a basin that clearly has more intense storms. The basins that generally have more intense storms are the North Atlantic, the East Pacific, and the West Pacific.
4. Can we predict the maximum category the storm will reach using the first 16 data points (2 days) from the storm?

   We will manipulate the data into numpy arrays that work in a neural network machine learning algorithm. We will split the data into a training set and a test set and run a machine learning program, then check how well we were able to predict maximum storm category.

We do not get a very good estimate of the maximum storm category with our machine learning algorithm. The accuracy score on the test set was about 40%, which is better than guessing randomly, but still not very good.

Background

We want to compute the average storm intensity over time because we want to see if storm intensity has increased over time. There are a couple reasons to calculate this. First, we want to see if climate change has affected storm intensity in any noticeable way. Second, we want to see if storm intensity has increased so we can raise more money for storm relief.

Understanding which areas in the world are most affected by tropical storms and hurricanes can help us focus disaster relief efforts and decide which countries/areas should be the focus of protection and mitigation efforts.

If we can look at the first two days of data on the storm and accurately predict the SSHS category the hurricane will reach, we can decide what hurricanes will be the most dangerous.

Dataset

Our dataset is the IBTrACS version 4 dataset. It's a dataset from NOAA providing information about tropical storms from 1842 to 2019. Each row of the data represents a measurement of the storm, so it specifies time, wind speed, storm type, basin, etc. To access the data and a description of each column, go to https://www.ncdc.noaa.gov/ibtracs/index.php?name=ib-v4-access. The columns we care most about are columns 1 to 12. We may have to use data from other columns since storm data from separate agencies is put into different columns.

Methodology

Our first step will be to compile all the wind speed columns into one column called wind speed. To do this we will look at every column where there could be a wind speed value, and we will take the first value that exists since all the values should agree. We have to do this because in our examination of the data we discovered that the wind speed is spread out over many different columns, and often some columns will have no value, while others will.

The next step will be to filter our data down to only the columns we will actually be using. These columns include SID (1), Season (2), Basin (4), Subbasin (5), Name (6), ISO Time (7), Latitude (9), Longitude (10), USA_SSHS (26), and the wind speed column we created.

For research question one we will then drop rows with no reported wind speed. Then we group by SID (storm ID) and use the aggregating function max to find each storm's maximum wind speed. Next we will group by Season and compute the average for each season. We can then create a graph with year along the x-axis and average wind speed along the y-axis.

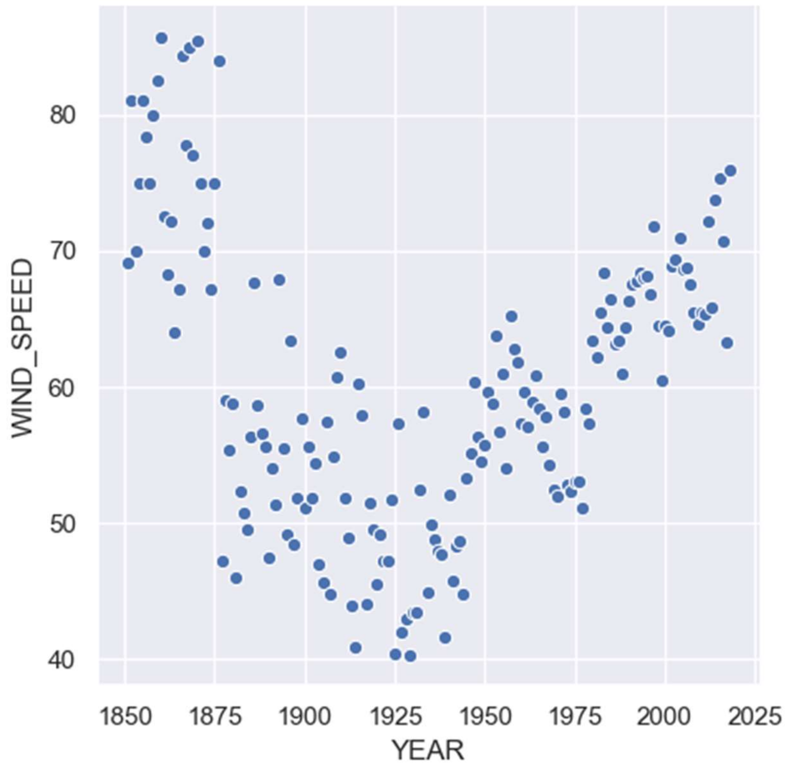
For research question two we will group by Season and use the unique function to find the number of unique storm IDs. After that we can filter the original data to find storms that reached category 3 or higher (USASSHS), and we can once again do a group by and count the number of unique storms that reached category 3 or higher in a season.

For research question three we will group by the storm ID and find the maximum wind speed. Next we will index into the data for each year we are looking at (using a loop), and then we can group by basin and find the average wind speed for the year, and drop all rows that have value MM as the value in the basin column. We can then save the average wind speed for that basin in that year in a list or a dictionary. Once we have done this for each year we're looking at we can graph the average wind speed for each basin against the year and compare trends and values.

For research question four we need to group the data into a numpy array in such a way that each storm is one row in the array and we can run it through the machine learning algorithm. We also need to make sure we only look at storms that have at least two days' worth (16 points) of data. We will use a neural network algorithm from sklearn, and check how well our model predicts the SSHS category of new storms. We tested a few differently sized neural networks to see how well each worked.

Results

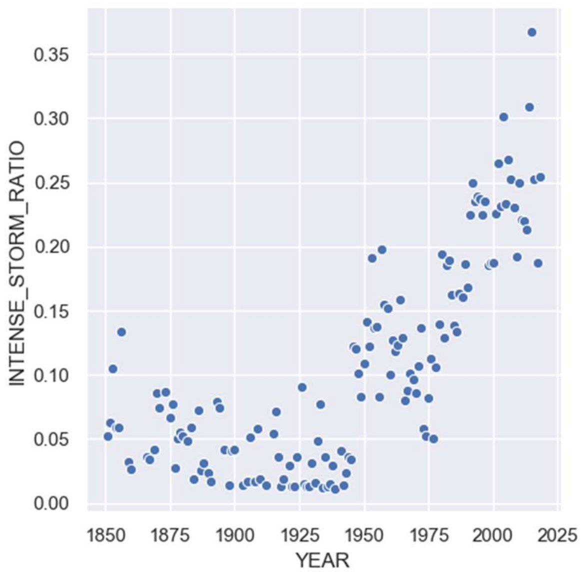
1. How has the average storm intensity changed over time?

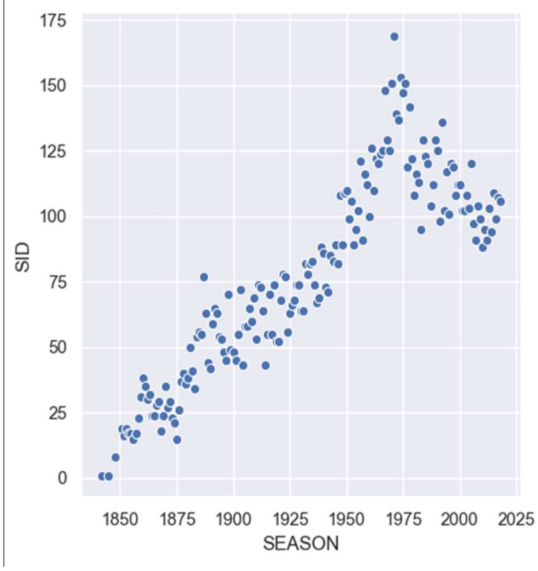


[Year vs. Wind Speed (knots)]

The trends in this graph is very interesting and unexpected; we didn't expect this V shape. We suspect that before about 1950 storm data was not kept very well, and this causes our points from before 1950 to be unrealistically high. You can certainly see a sudden shift in the data, which suggests that something changed around that time, either a change in actual storm intensity, or the method of data collection changed. In the last fifty years there has been a steady increase in storm intensity. The positive trend in the last fifty years supports the idea that storms have been getting worse, at least recently.

2. How has the frequency of large storms changed over time?

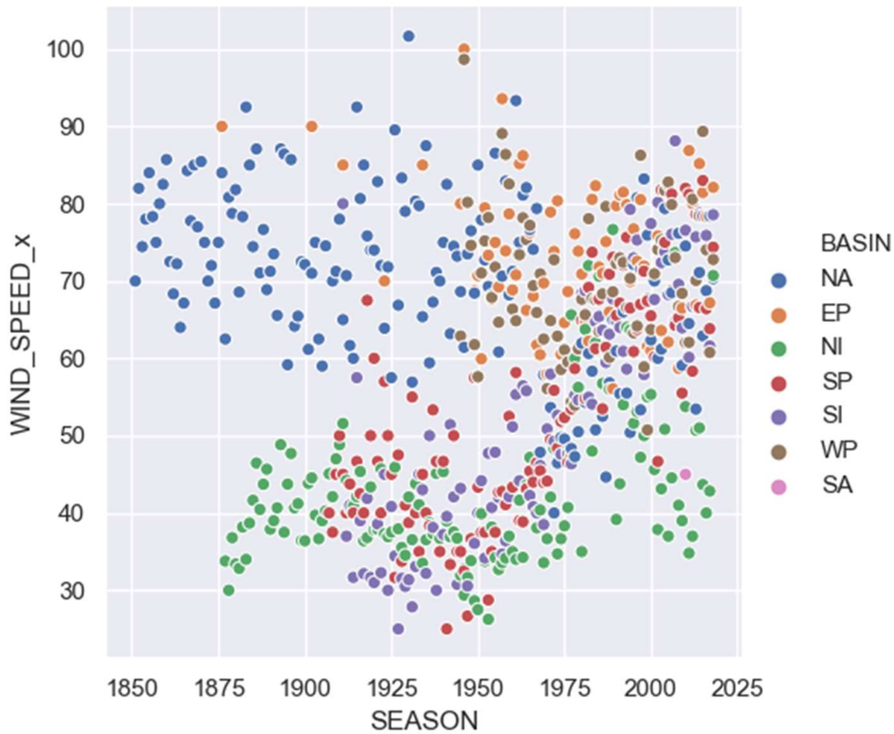




[Year vs. Storm Ratio] [Year vs. Number of Storms]

Since 1950 there has been a marked trend upwards in the fraction of total storms that are category three or higher. This strongly supports the idea that storms are getting worse, and means that more should be invested to support safety measures and strong infrastructure in endangered areas. This evidence supports the climate change theory. We expected the total number of storms to consistently increase. The sudden drop in number of storms around 1975 surprised us, and we're not sure what happened. The reason we chose to use scatter plots is because there is natural variability between each year, so a line plot or bar plot would look messy compared to a scatter plot.

3. Which basin has the highest intensity storms?



There doesn't seem to be any basin that clearly has more intense storms. This stops us from being able to suggest a region that we should focus relief and mitigation efforts on. The North Atlantic, Eastern Pacific, and Western Pacific have fairly constant storm intensities. On the other hand, the South Pacific and South Indian have dramatic increases in storm intensity over time. That is, the increase in overall storm intensity seems to be largely a result of increases in these regions. This means we should probably keep an eye on how storm intensity changes in these regions in the future. We focused on the average strength of storms. This may not be the best way to look at where funding should go, since there are many other factors contributing to damage done by hurricanes. For example, many category two hurricanes can do as much damage as only a few larger hurricanes.

4. Can we predict the maximum category the storm will reach using the first 16 data points (2 days) from the storm?

   We can't really predict the maximum category the storm will reach. If the machine learning model was guessing randomly, we would guess correctly roughly 15% of the time. We are reaching about 40% accuracy, which is significantly better than randomly guessing, but still not a very accurate model. We had to base our model on data from 2000 to present since it turned out that in previous years there was missing data that made it very difficult to use older data effectively in our model.

Reproducing Results

First, download the data. Open the link given above, then click on "CSV (Comma Separated Values)" under the header "links to v04 data". Then download "ibtracs.ALL.list.v04r00.csv". Then rename the csv file "storms.csv". Open our code in Visual Studio Code. Run the code in the terminal, and after a few minutes the graphs will be saved to your computer as png files with the following names: "wind_speed_over_time.png", "high_inten_storm_freq_over_time.png", "intense_storm_ratio_over_time.png", "number_of_storms.png", and "maximum_intensity_by_basin.png". These graphs are the results which let us answer our first three research questions.

The machine learning testing and training scores will be printed in the terminal after running our code. Since it's using a random number generator it won't always get exactly the same results, but it should be reasonably consistent.

Work Plan Evaluation

1. Data cleanup. Create the compiled wind speed column as mentioned above, and then filter down to only the data columns we need. This will likely take the most time to complete (3 hours).
   This step took a while since we had a large dataset, and we were originally using an inefficient for loop, so we wasted a significant amount of time waiting for the program to run. We also had to ensure that our data ended up as the right type, since our data sometimes seemed to end up as a different type than it began as.
   Actual Time: 4 hours
2. Research questions one, two, and three, as detailed above; we will meet in person to work on code so we can bounce ideas off each other and discuss how we want to solve any problems we encounter. We will each test our functions separately to be as sure as we can that they work properly. (approx. 1.5 hours per research question).

Research question one was the most challenging since that was where we ran into most of the issues we had never seen before. After that, the other two were simpler since we had a better idea of some techniques we could use.

Actual Time: 5 hours

3. Creating data visualizations for each research question, using either matplotlib or seaborn. We will use either scatter plots or line plots, depending on which better shows our data trends.

   This part was simple. We used a scatter plot for each since we had enough data points that a line or bar plot would be unnecessarily hard to read. We tried using an lmplot to see trends for research question three, but that didn't work out very well. Most of the time associated with this part was lumped in with answering our research questions.
4. We decided we needed another research question to incorporate something else we had learned in class, and to allow us to create more code. We decided on implementing a machine learning neural network. This took about 5 hours in total, most of which was spent on transforming the data into a workable form.

Testing

Testing was performed as we went along. We felt this was best since it let us catch bugs as we went along. We created test sets that were subsets of our full dataset. Because our data didn't produce any numeric outputs we didn't use the assert equals function at any point. Our results can be trusted since we have a reliable data source (from NOAA), and because we ran our program on a test set each time we changed something significant. We checked our data to ensure that the result made sense, and adjusted accordingly if there were any problems. For example, we removed the data from 2019 since it showed up as an extreme outlier because the year hasn't finished yet, so that data is incomplete.

Presentation Type

Live presentation.

Collaboration

In addition to ourselves and course staff we looked online at documentation and other people's questions (e.g. stackexchange).