NASA GLOBE CLOUD GAZE

Audrey Cabrera
Data Science Intern
Summer 2022

Scan for my LinkedIn!

About me:
My name is Audrey Cabrera and I am a rising senior at UCLA. I am majoring in statistics and primarily focus on working with data. I intend to pursue a career in the data science field after I graduate.

My Project Objective

- Use three datasets:
  - CLOUD GAZE Cloud Cover
  - CLOUD GAZE Cloud Type
  - GLOBE Observations
- Work with these datasets to:
  - Examine usability of the data to catch and fix errors before other scientists use the data
  - Explore different averaging methods to find the best way to average cloud cover values
  - Compare averages from CLOUD GAZE, GLOBE, and the satellites

Example observation:
- One image for each direction:
  - N, E, S, W, Up

What is NASA GLOBE CLOUD GAZE?

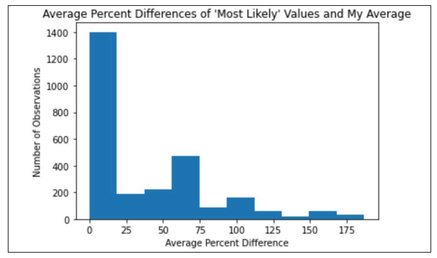
- Combines GLOBE and Zooniverse to form GAZE
- **GLOBE**
  - Over 100 countries, 34,000 schools and 142,000 citizen scientists
  - Citizen scientists take and submit photographs of the sky, and note what they see
- **CLOUD GAZE**
  - Citizen scientists characterize photographs and add cloud cover and cloud type tags
  - Set of citizen scientists classify the GLOBE observer images on cloud cover and cloud type
- Third source of data is from satellites:
  - GEO (GOES geostationary satellite): include METEOSAT and HIMAWARI (part of the suite of GEO satellites)
    - About 35,800 kilometers high
  - Terra: About 705 kilometers high
  - Aqua: About 705 kilometers high
- Since satellites can't catch every detail from their position, citizen scientists help fill in the missing data and details with their ground observations
- Satellites provide a perspective from above and citizen scientists complete the other half, with the ground perspective
  - When a GLOBE observation is captured within 15 minutes of a satellite observation, the two perspectives are matched
  - Satellite matches provide augmented data for research
- Overall mission: understand how clouds affect the climate on Earth
  - Compare cloud observations collected by citizen scientists and satellites

Measuring precision of averages

- Method: percent difference
  - Measures difference between two related values
- 43% of the data has 0% average percent difference
- 11% of the data has over 100% average percent difference
- Once I found the averages, I wanted to examine how precise and reliable my average was so I used percent difference to measure the difference between two values.
- I found the percent difference between my average and each direction. I averaged them together for a single percent difference for each observation.
- Almost half of the observations (43%) had a percent difference of zero, which means that all the values were the same, which indicates a precise average that I found. Looking at the plot to the right, you can see that its distribution is mostly concentrated around lower average percent difference.
- Only about 11% of the data had 100 or more percent difference, meaning that the values were very different.

Table of Recalculated Averages:

	N	E	S	W	Up	My Avg	Avg % Diff
0	70.0	70.0	70.0	70.0	70.0	70.0	0.00
1	5.0	5.0	5.0	5.0	5.0	5.0	0.00
2	5.0	5.0	5.0	17.5	5.0	7.5	48.00
3	37.5	70.0	37.5	37.5	37.5	44.0	21.88
4	70.0	70.0	70.0	70.0	70.0	70.0	0.00
5	70.0	70.0	37.5	37.5	70.0	57.0	28.79
6	95.0	95.0	70.0	95.0	95.0	90.0	9.32
7	95.0	95.0	95.0	95.0	95.0	95.0	0.00
8	95.0	95.0	95.0	95.0	95.0	95.0	0.00
9	95.0	95.0	95.0	95.0	95.0	95.0	0.00
10	70.0	95.0	95.0	70.0	95.0	85.0	14.41
11	95.0	95.0	70.0	95.0	70.0	85.0	14.41
12	95.0	95.0	95.0	95.0	95.0	95.0	0.00
13	5.0	5.0	5.0	5.0	5.0	5.0	0.00
14	70.0	70.0	0.0	37.5	70.0	49.5	66.10



Exploring the data

- I focused on averaging cloud cover: 'Most Likely' column
  - 'Most Likely' represents the most likely answer for total cloud cover from GAZE classifications
- 'Most Likely' cloud cover data:
  - Few (less than 10%)
  - Isolated (10-20%)
  - Scattered (25-50%)
  - Broken (50-90%)
  - Overcast (more than 90%)
  - Other (clouds and sky more than 25% blocked, not clouds or sky)
- 'Most Likely' data is represented by integers ranging from 0-5, however these numbers are codes for the categorical values
- Limitations of the data
  - Unique structure: Each row is a separate observation with five images representing it. Since each row had to be treated as its own population in a sense, I was limited on how to find averages and compare values. The population of each observation was too small to use any type of statistical test or complex statistical method, and the data was not continuous which prevented it from meeting the assumptions of many tests and methods in statistics.
  - Categorical: The 'Most Likely' column of data is categorical which limits the possible averaging methods to just qualitative averages, which is equivalent to finding the mode.

Averaging methods

- Qualitative average vs numerical average
  - Finding the mode was a problem here because many of the observations had multimodal data, which means that there were multiple modes because there were equally common values.
- My steps taken in order to be able to properly analyze categorical average of the data:
  - Convert integer to a meaningful total cloud cover numeric value
    - Take the midpoint of each category's interval
    - Assign this value as the new 'Most Likely' value
  - Take average of all five directions using the meaningful numeric values
  - Convert numeric average back to one of the given categories based on the cloud cover intervals

North Numeric Most Likely	East Numeric Most Likely	South Numeric Most Likely	West Numeric Most Likely	Up Numeric Most Likely	Numeric Most Likely Average	Average Most Likely (Categorical)
70.0	70.0	70.0	70.0	70.0	70.0	Broken
5.0	5.0	5.0	5.0	5.0	5.0	Few
5.0	5.0	5.0	17.5	5.0	7.5	Few
37.5	70.0	37.5	37.5	37.5	44.0	Scattered
70.0	70.0	70.0	70.0	70.0	70.0	Broken

Percent Difference and Agreement

- When analyzing the values, I was able to verify that agreement corresponds to percent difference.
- When there is a higher percent difference, there is lower agreement and the cloud cover data tends to have higher standard deviation. This means that the data is more spread out and there is more variation in the responses of the classifications.
- As a result, a high percent difference is inevitable since there is less agreement in the values.
- The left plot shows the average agreement of observations with 0% difference and the distribution is concentrated around higher agreement.
- Whereas in the middle plot, we can see that there is lower agreement on average for the data with 100% or more difference.