Analyzing Local Land Cover Using Surrounding Data

Yuvraj Sahu

NASA STEM Enhancement in Earth Science (SEES) Internship

Abstract

Land cover can be predicted where it is not feasible to make observations. Using the Land Cover tool within the GLOBE Observer app, data was collected and classified based on percent coverage. Four factors were selected for testing - deciduous trees, evergreen trees, grass, and urban. Using a weighted average formula, the predicted land cover at each possible location within each possible rectangle was calculated. For all 4 corners, it used the 2 directions facing the point being predicted. As the distance or angle difference increased, the weight decreased. The mean absolute error, in percentage points, was then calculated for each of the factors. The overall error was 16.6 for deciduous trees, 13.4 for evergreen trees, 17.5 for grass, and 38.5 for urban. In general, the small rectangles had lower error than the large rectangles. This confirms that, for certain factors, surrounding data can be used to predict land cover at a certain point.

Introduction

Using the GLOBE Observer app, citizen scientists can collect data on the features of Earth's surface, which researchers can then use. While GLOBE Observer has multiple data points for land cover across the world, it does not have a continuous map - there are certain locations where there is no data. This can apply to other data that is derived from citizen science.

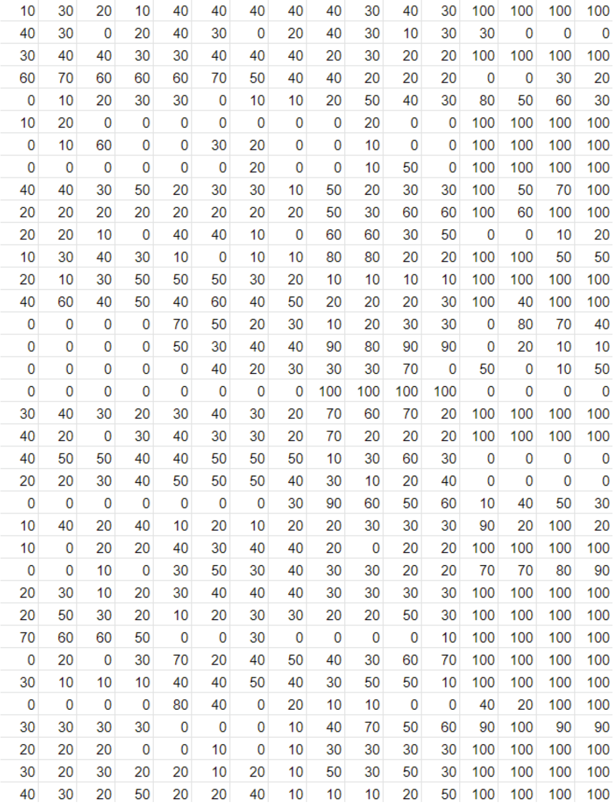
This research paper aims to determine if it is possible to use data science to predict the gaps in land cover data using the surrounding data points. A sufficiently accurate prediction method would allow researchers to utilize discrete land cover data, since the predictions would fill in gaps in data. This could be especially useful when such discrete data is used in conjunction with other data, since researchers would be able to draw conclusions between land cover and other factors.

Throughout this research process, I was mentored by scientist mentors, including Dr. Rusty Low, Dr. Cassie Soeffing, and Dr. Peder Nelson. In addition, I received guidance from peer mentors, including Matteo Kimura. Their mentorship was especially helpful during the data collection and research question phases, and they inspired me to pursue research in the future.

Methods

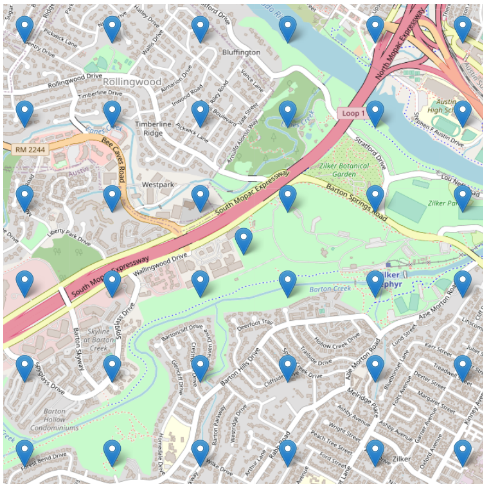
Data Collection

GLOBE Observer is an app that allows users to collect data about the Earth. Under the guidance of the scientist mentors, using this app, land cover data was collected in a 6 by 6 grid where adjacent points were 500 meters apart. This data was then categorized and the percent coverage for each category was determined. For each point, all four cardinal directions were classified separately. Since the observations only included the 100 meter by 100 meter space centered around a point, there were gaps between these points where land cover was not classified. The following table represents the data.

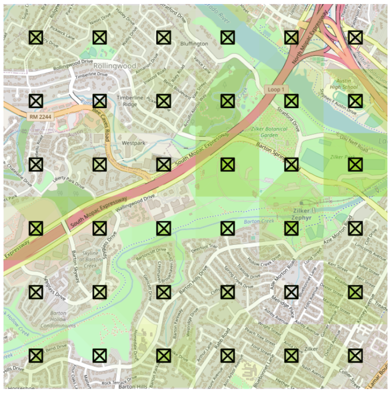


10	30	20	10	40	40	40	40	40	30	40	30	100	100	100	100
40	30	0	20	40	30	0	20	40	30	10	30	30	0	0	0
30	40	40	30	30	40	40	40	20	30	20	20	100	100	100	100
60	70	60	60	60	70	50	40	40	20	20	20	0	0	30	20
0	10	20	30	30	0	10	10	20	50	40	30	80	50	60	30
10	20	0	0	0	0	0	0	0	20	0	0	100	100	100	100
0	10	60	0	0	30	20	0	0	10	0	0	100	100	100	100
0	0	0	0	0	0	20	0	0	10	50	0	100	100	100	100
40	40	30	50	20	30	30	10	50	20	30	30	100	50	70	100
20	20	20	20	20	20	20	20	50	30	60	60	100	60	100	100
20	20	10	0	40	40	10	0	60	60	30	50	0	0	10	20
10	30	40	30	10	0	10	10	80	80	20	20	100	100	50	50
20	10	30	50	50	50	30	20	10	10	10	10	100	100	100	100
40	60	40	50	40	60	40	50	20	20	20	30	100	40	100	100
0	0	0	0	70	50	20	30	10	20	30	30	0	80	70	40
0	0	0	0	50	30	40	40	90	80	90	90	0	20	10	10
0	0	0	0	0	40	20	30	30	30	70	0	50	0	10	50
0	0	0	0	0	0	0	0	100	100	100	100	0	0	0	0
30	40	30	20	30	40	30	20	70	60	70	20	100	100	100	100
40	20	0	30	40	30	30	20	70	20	20	20	100	100	100	100
40	50	50	40	40	50	50	50	10	30	60	30	0	0	0	0
20	20	30	40	50	50	50	40	30	10	20	40	0	0	0	0
0	0	0	0	0	0	0	30	90	60	50	60	10	40	50	30
10	40	20	40	10	20	10	20	20	30	30	30	90	20	100	20
10	0	20	20	40	30	40	40	20	0	20	20	100	100	100	100
0	0	10	0	30	50	30	40	30	30	20	20	70	70	80	90
20	30	10	20	30	40	40	40	30	30	30	30	100	100	100	100
20	50	30	20	10	20	30	30	20	20	50	30	100	100	100	100
70	60	60	50	0	0	30	0	0	0	0	10	100	100	100	100
0	20	0	30	70	20	40	50	40	30	60	70	100	100	100	100
30	10	10	10	40	40	50	40	30	50	50	10	100	100	100	100
0	0	0	0	80	40	0	20	10	10	0	0	40	20	100	100
30	30	30	30	0	0	0	10	40	70	50	60	90	100	90	90
20	20	20	0	0	10	0	10	30	30	30	30	100	100	100	100
30	20	30	20	20	10	20	10	50	30	50	30	100	100	100	100
40	30	20	50	20	20	40	10	10	10	20	50	100	100	100	100

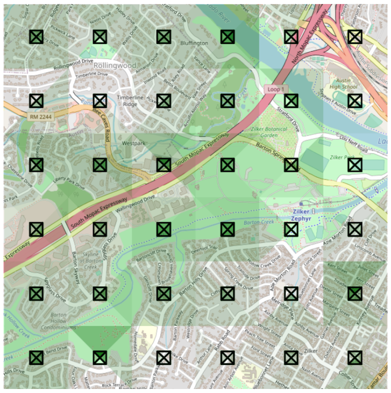
The following diagrams (below) represent the data. The topmost diagram represents the 37 locations where data was collected. While data was collected for the center point, it was not used in this research because it did not align with the grid formed by the other 36 points. The subsequent four diagrams visualize the percent coverage for the four selected factors - deciduous trees, evergreen trees, grass, and urban land cover. The black boxes represent the 100x100 classified region. The coloring within the black boxes was extended to the 500x500 meter space surrounding the location.



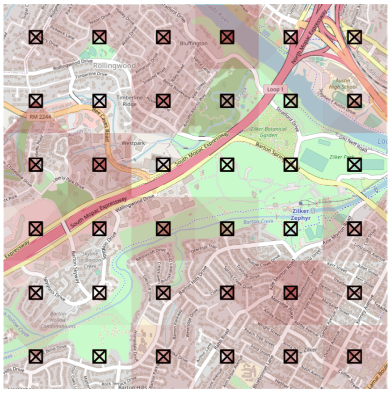
Map of the Areas of Interest (AOI) Points



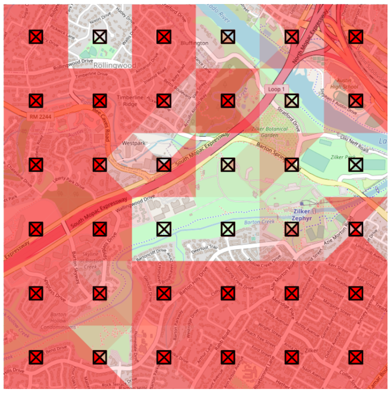
Map of Grass



Map of Evergreen Trees



Map of Deciduous Trees



Map of Urban Spaces

Research Method

For every possible rectangle that can be made from the grid, and for every point within each rectangle, it predicts the land cover based on the four corners in order to simulate predicting the areas between each of the points. To do this, it uses a weighted average. For each of the corners, only the two directions that are facing the point to be predicted are used. This means that 8 factors are used in the weighted average. Distance and weight have an inverse relationship - as the distance increases, the weight in the weighted average decreases. Also, angle difference and weight have an inverse relationship - as the difference in angle increases, the weight decreases.

Specifically, this formula was used:

$$\frac{\sum_{corners} \frac{p_{NS} * sin^{2}(\theta) + p_{EW} * cos^{2}(\theta)}{\sqrt{(\Delta x)^{2}+(\Delta y)^{2}}}}{8}$$

Where:

p_{NS} is the proportion of land cover that was covered by a specific feature at a given corner in either the north direction (for the two corners south of the target) or the south direction (for the two corners north of the target).

p_{EW} is the proportion of land cover that was covered by a specific feature at a given corner in either the east direction (for the two corners west of the target) or the west direction (for the two corners east of the target).

θ is the angle from the corner to the target, using standard Polar angles.

Δx is the horizontal (east-west) distance between the corner and the target.

Δy is the vertical (north-south) distance between the corner and the target.

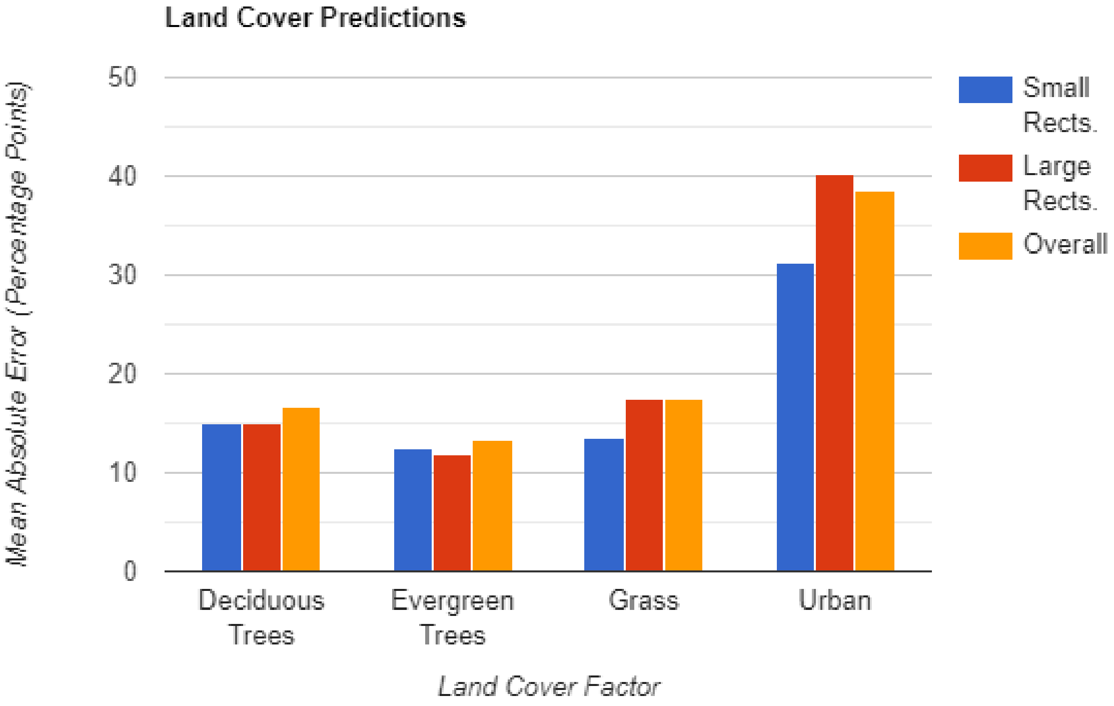
Implementation

Using the programming language Python, I developed a program to create visualizations and test the formula for each possible target and rectangle. The code can be found at: https://colab.research.google.com/drive/1p7Rg9taYoThZKLJfBaoyc3MqZ4Ge2EOW

Results

For each prediction, the mean absolute error - the average of the positive differences between the predicted values and the actual values - was calculated. The units for the errors are percentage points. The errors were calculated for small rectangles (1km by 1km), for large rectangles (3km by 3km), and for all rectangles.

As shown in the bar chart and table below, smaller rectangles generally achieved better accuracy, especially for the factors of grass and urban land cover. In addition, the predictions for deciduous trees, evergreen trees, and grass were significantly better than those for urban areas.



Mean Absolute Errors (Percentage Points)

	Small Rectangles	Large Rectangles	Overall (All Rectangles)
Deciduous Trees	14.9	15.1	16.6
Evergreen Trees	12.5	11.8	13.4
Grass	13.6	17.6	17.5
Urban	31.2	40.2	38.5

Discussion

For future research, different formulas and techniques could be experimented with in order to find a more accurate model. In addition, these findings could be applied to different contexts. This could include using other point grids (including those of different sizes) in order to better determine the strength of a model, using public GLOBE data (which would not be organized in a grid) to develop a model that works for scattered discrete data, and developing models for different fields outside of predicting land cover. Finally, these land cover predictions could be paired with other data, such as data on mosquito populations, in order to determine relationships.

Conclusion

Many sources of data pertaining to the Earth's surface are discrete, meaning that they have individual data points in different locations on Earth's surface. This can be problematic for researchers who require data in certain locations that are not part of the discrete data set. The goal of this research paper was to determine if it was possible to use data science to predict land cover data using surrounding points, which would allow researchers to use predicted values at locations where there is no data.

To test this, land cover data was collected in a grid, and the percent coverage for each factor of land cover was determined. Then, for each rectangle in the grid, and for each target point within the rectangle, a Python program attempted to predict the land cover at the target given the data of the corners of the rectangle. To do this, it used a weighted average. A total of eight items were used in the weighted average - for each of the four corners, the data from the two cardinal directions facing the target were used. The weight for each item was directly proportional to the sine or cosine of the angle to the target and inversely proportional to the distance to the target.

The results showed that, to a certain degree of accuracy, for a grid format, surrounding data can be used to predict land cover at a given point within a grid. Smaller rectangles seemed to achieve better accuracy, especially for the factors of grass and urban land cover. In addition, the predictions for deciduous trees, evergreen trees, and grass were significantly better than those for urban areas. This research can continue to be developed by using new sources of data, testing new models, and developing models for other areas outside of land cover.

Citations

Global Learning and Observations to Benefit the Environment (GLOBE) Program, 6/19/2021, globe.gov

IVSS Badges - STEM Professional, Engineer, Data Scientist, and STEM Storyteller

I Am A STEM Professional

Throughout the internship, I collaborated with STEM professionals. This included the scientist mentors, which included Dr. Rusty Low, Dr. Cassie Soeffing, and Dr. Peder Nelson, as well as the peer mentors, which included Matteo Kimura. The scientist mentors guided me through using the GLOBE Observer app, which is how I collected data for my research. Later, conversations with Matteo helped me find my research question. Finally, Dr. Rusty Low and Dr. Cassie Soeffing assisted me through the creation of a research abstract, poster, and paper as well as a video explaining my research. These mentors provided great mentorship and inspired me to continue to undertake research in the future.

I Am An Engineer

This badge focuses on using engineering principles to answer a question. In this case, I used software engineering principles when creating my Python program. I used Python to create the visualizations for the land cover factors. Later, I used the program to iterate through each rectangle and each target point within the rectangle in order to test if the data from the corners of the rectangle could predict the land cover at the target point. In order to address the issue of gaps in data, I used software engineering to create a program that would predict land cover.

I Am A Data Scientist

This badge focuses on the use of data to solve problems. In this case, I collected the data used in this research paper by classifying the land cover at each of the 36 locations. I then used this data to fuel my research, and I studied how the data science model I developed performed on the data. Finally, I discussed the ways that I could further my data science research - by exploring different data sources (both in grid formats and not in grid formats), by exploring different models, and by relating these predictions to new contexts. Throughout the paper, I made sure to discuss how I used data and data science techniques and how more data and model development could improve my research.

I Am A STEM Storyteller

This badge focuses on the use of creativity, which includes artistic renditions. In this case, I created visualizations for the land cover data. I made sure to carefully delineate the boundary between known data and extrapolated unknown data by using black lines. In addition, the extrapolated portions are lighter in color, which allows the viewer to see the background while simultaneously reminding them that this is extrapolated data. Finally, I made sure to show the differences in the data between the north, south, east, and west directions as well as show the differences between locations by changing the intensity of the color. I later used other visualizations, such as the bar chart, which would help the viewer better understand the results.