Urban Tree Canopy Classification Using Supervised Learning

Patrick Page
patpage@iu.edu

Introduction

As climate change brings rising temperatures, urban vegetation renewal has been a major topic of concern for communities everywhere. Downtown neighborhoods that lack urban tree canopies may experience temperatures that are 10°F hotter than the surrounding rural areas. This effect is known as the urban heat island effect and one of the ways to combat it is to create green spaces in urban areas [1].

Trees help the urban microclimate in several ways. The lighter color of tree leaves allows for more light to be reflected than the typical grey and dark grey colored urban buildings and infrastructure. Trees transpire water which provides a cooling effect to their surrounding area [1]. Trees serve as quality filters by helping remove pollutants from air and soil [2]. Trees also provide shade to sidewalks and streets.

Thankfully, there are many local and regional NGOs focused on restoring urban tree canopies. In Indianapolis, the organization *Keep Indianapolis Beautiful, Inc.*, or KIBI for short, is focused on planting 100,000 trees within the Indianapolis municipality.

Local and regional NGOs, like KIBI, are in need of recent urban tree canopy cover data in order to inform policy and prioritize locations of tree canopy restoration projects. The aim of this project is to utilize supervised learning algorithms within Google Earth Engine (GEE) to create an updated canopy map of Indianapolis.

Supervised learning is a method of machine learning where the machine is trained with "labeled" data [3]. For this project, a training set of classified land cover types was provided to the model. The type of supervised model used was the random forest model. Random forest is a powerful decision tree model that crowd-sources the prediction from a set of numerous other decision trees [4].

Methodology

Step 1: Filtering Satellite Imagery & Band Selection

The earth remote sensing data used by the random forest model came from the Sentinel-2 satellite. Sentinel-2 provides high resolution, multispectral imagery that is ideal for monitoring vegetation, soil, and water cover [5]. Imagery from Sentinel-2 was filtered between the dates of April 1st, 2022 and October 30th, 2022 so that the changing leave colors could be accounted for in the model. The bands selected for the model are as follows: B2, B3, B4, B5, B8, B9, B11,

NDVI, NDWI, and SAVI. These bands were selected to best capture the different spectral signatures of the land cover types the model would be trained to classify.

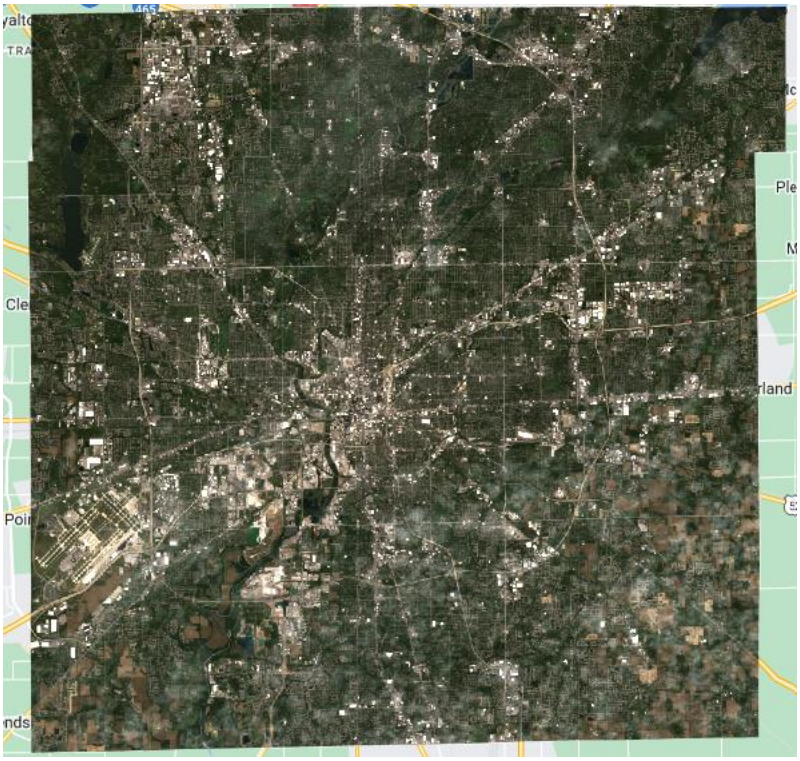

Figure 1: RGB display of Marion County's Sentinel-2 satellite imagery

Step 2: Creating the Labeled Dataset

The four land cover labels used in the model are as follows: tree canopy (tree_canopy), vegetation that is not a tree canopy (veget_non_forest), urban (urban), and water (water). Geometries for each landcover type were carefully drawn on the GEE map. Each landcover type has 16 to 20 different labeled geometries which constitutes a feature collection. The labeled dataset was created by merging the different land cover feature collections. Figures 1 & 2 show how the geometries for the different land covers were constructed on the GEE map.

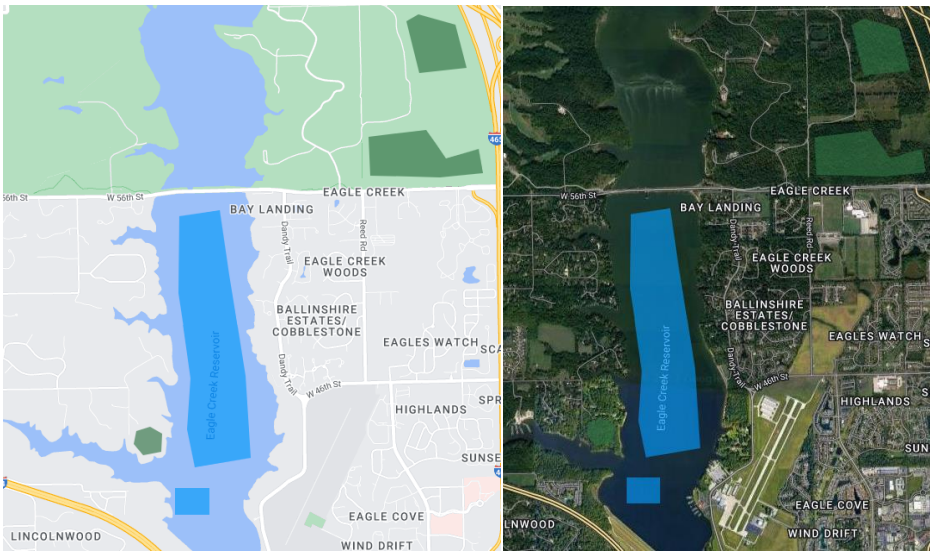

Figure 2: Geometries of water and tree canopy cover in Eagle Creek (left map view & right satellite view)

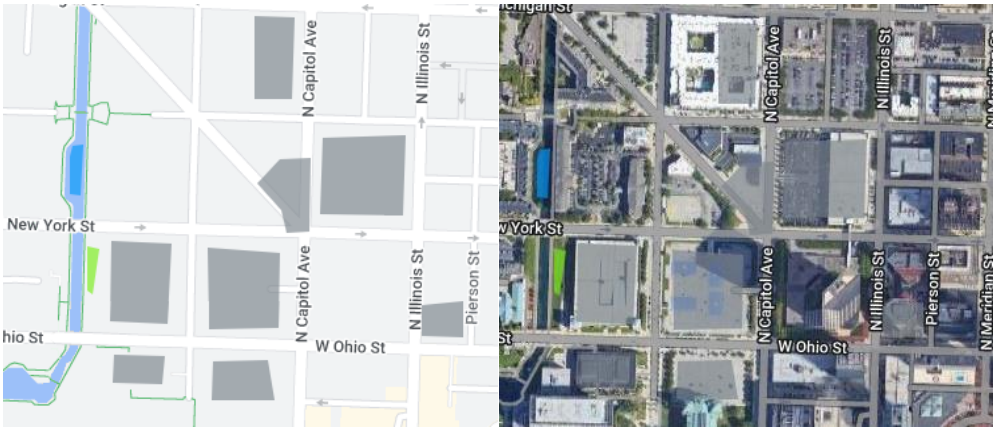

Figure 3: Geometries of urban and vegetation land covers in downtown Indianapolis (left map view & right satellite view)

Step 3: Partitioning Marion County by Zip Code

To create action from the model, the city of Indianapolis needed to be divided out within GEE to better compare tree canopies and to help focus efforts on areas that lack trees. The division of Indianapolis was created by uploading a shapefile asset of Marion County zip codes. In figure 4, the cyan colored outlines represent the different zip code outlines within Marion County.

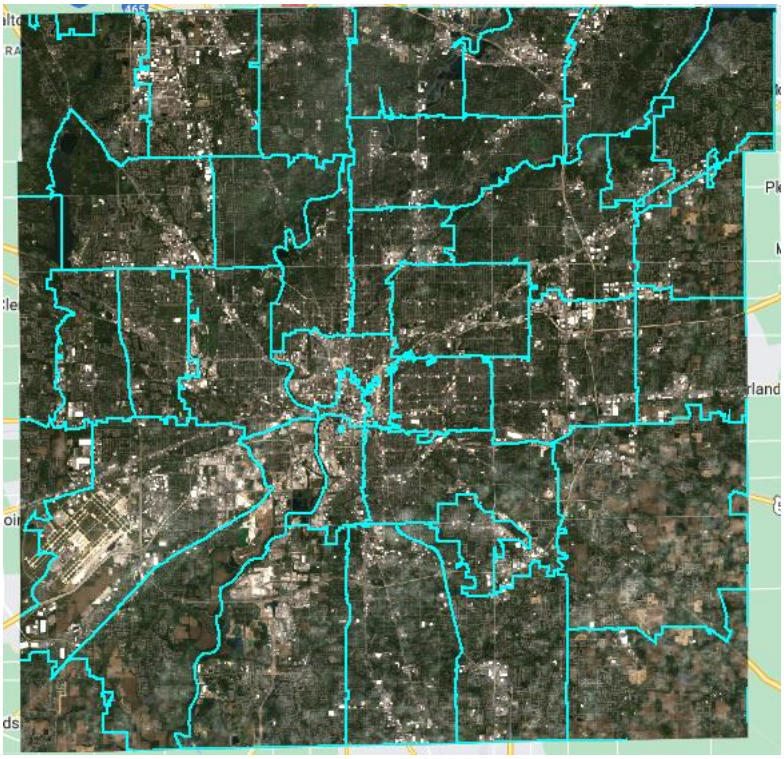

Figure 4: Zip code outlines within Marion County

Step 4: Training and Testing the Model

The labeled data set was split into a training and test dataset in 80% - 20% fashion (80% of the labeled data was used for training the model while 20% of the labeled data was for testing the model). A confusion matrix was created to observe the accuracy of the model. The random forest model scored an accuracy level of 99.8% from the test dataset.

Results

As stated earlier, the model did score an accuracy of 99.8% using the test dataset. This value should be taken with a grain of salt, however, as confusion matrices are not the best method for assessing model accuracy. The limitation here is that it is difficult to acquire sufficient labeled data to serve as a test dataset. It is not surprising the model did extremely well with data that closely resembles the same data that it was trained on.

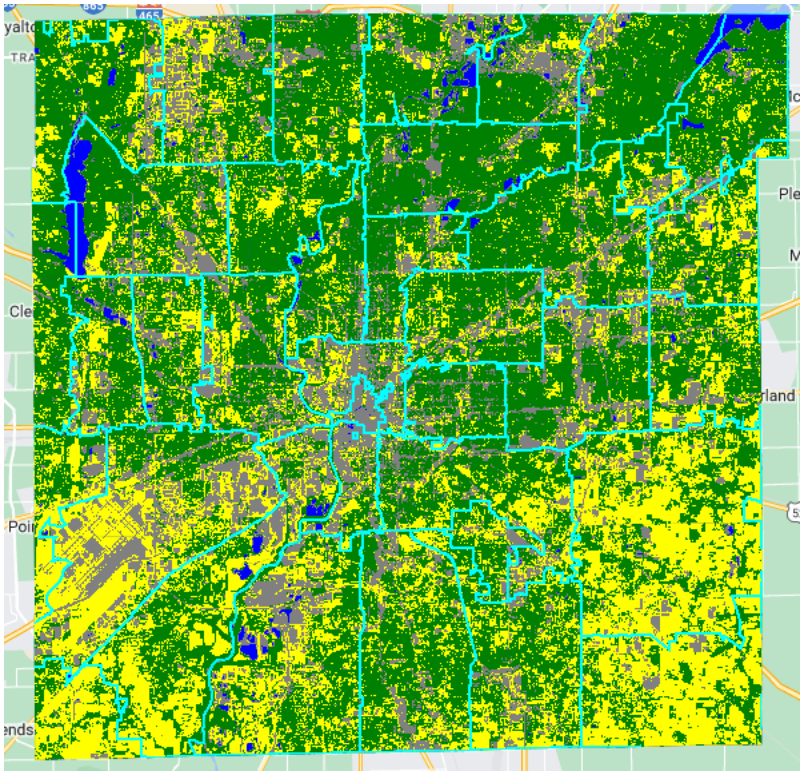

Figure 5: Classified land cover image generated by the random forest model

The model still performed generally well based on eye-tests. Figure 5 shows the classifications generated by the model for all of Marion County. In figure 5, green represents tree canopy, yellow represents non-tree vegetation, grey represents urban cover, and blue represents water.

Discussion

Based on the classifications from the model, a dataset of each zip code tree canopy ratio was obtained. This dataset was then joined with datasets from the 2021 American Community survey to provide information of population, demographics, and income for each of the zip codes in Marion County. Mean surface temperature data for each zip code was also found in GEE and merged with the final dataset.

Table 1 shows the zip codes (that contain more than 10,000 people) with the lowest tree canopy ratios. The tree canopy ratio is the division of the tree canopy area by the total area of the zip code. According to the model, these are the zip codes that need the most attention for planting trees.

	zipcode	tree canopy ratio	Total pop
21	46204	0.0470	10246.0
11	46163	0.1847	13948.0
46	46202	0.1855	21039.0
34	46241	0.2108	33756.0
18	46259	0.2266	14955.0
9	46112	0.2983	40574.0
24	46239	0.3023	33456.0
28	46168	0.3517	37058.0
32	46231	0.3575	10824.0
45	46113	0.3717	15798.0

Table 1: Top 10 Marion County zip codes with lowest tree canopy ratios

A quick study was performed to identify any correlations between tree canopy ratios and communities dominated by certain races and ethnicities as well as household income.

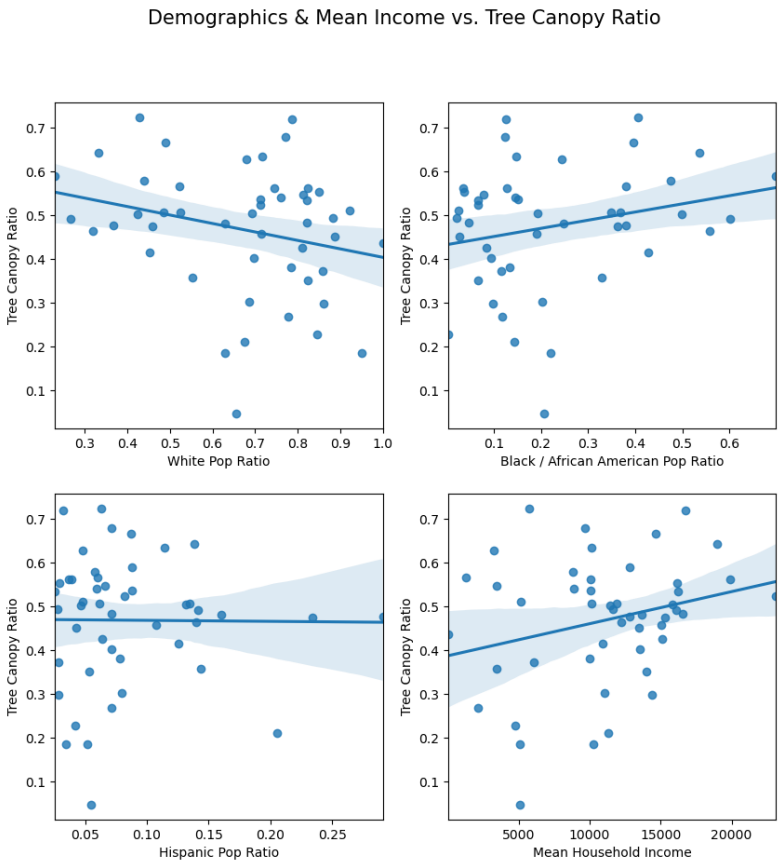

Figure 6: Scatterplots of Demographics and Household Income vs. Tree Canopy Ratio

Based on the study, tree canopy ratios slightly decline when the white population ratio increases. Tree canopy ratios increase when black population ratios increase. No correlation was found between tree canopy ratios and Hispanic population ratios. Finally, tree canopy ratios increase when mean household income increases.

The last finding about the correlation between income and tree canopies is particularly fascinating because it hints that lower income, impoverished neighborhoods tend to lack tree canopies while higher income neighborhoods tend to have more trees.

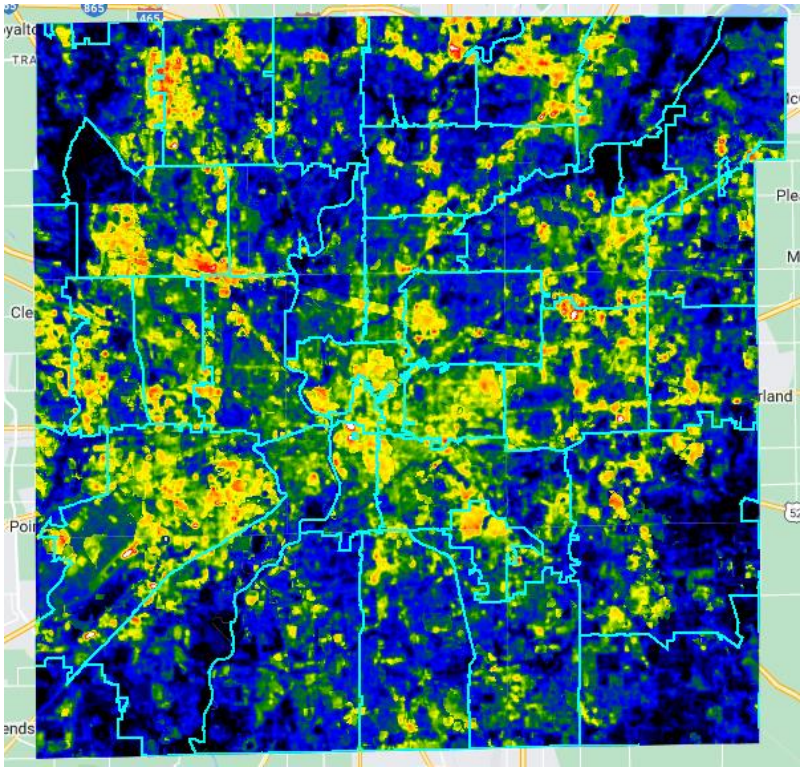

Figure 7: Mean Surface Temperatures of Marion County for Summer 2022

Lastly, a study between the effects of tree canopies and mean surface temperature was conducted. Figure 7 depicts the mean surface temperatures of Marion County during the summer of 2022. It can be seen that central downtown appears the hottest as it appears bright yellow in the image. This same area has the largest collection of grey pixels based on the classification model.

Figure 8 is a scatterplot of tree canopy ratios vs. the mean surface temperatures for each zip code. As the tree canopy ratio increases, the mean surface temperature decreases. This plot emphasizes the importance of tree canopies and that they really do improve the microclimates of urban areas by helping reduce surface temperatures.

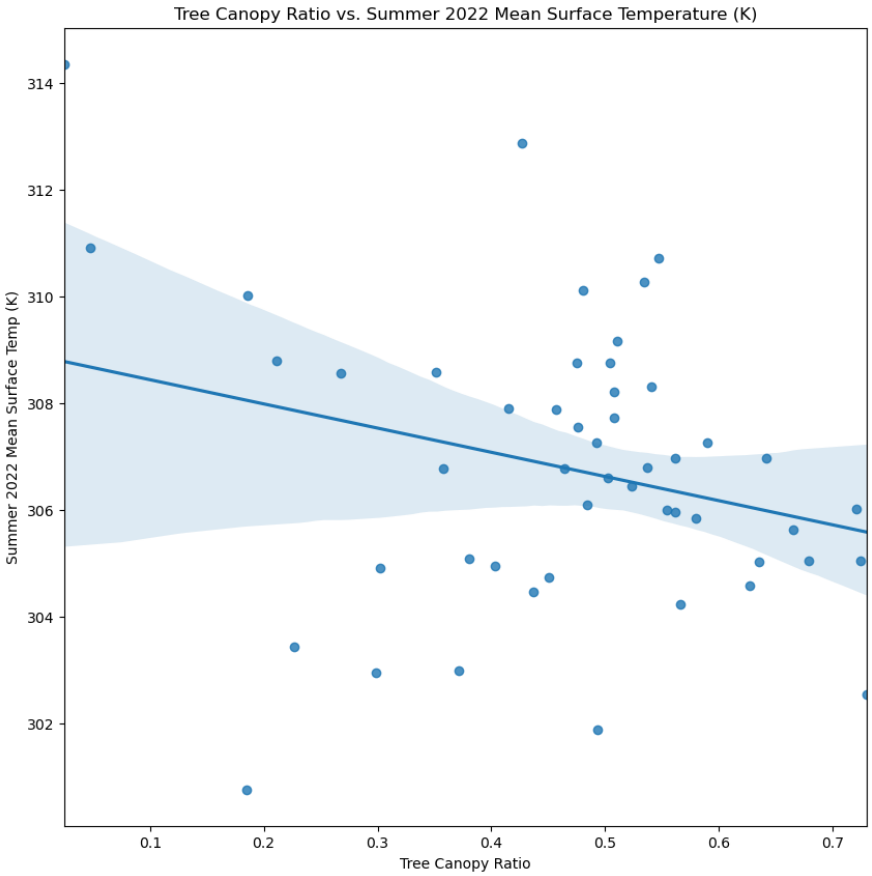

Figure 8: Scatterplot of tree canopy ratio vs. mean surface temperature (K)

Conclusion

The random forest algorithm within GEE yielded a decent tree canopy map of Marion County. Although there were certainly misclassifications, the model still served as a great overview of where tree planting efforts should be focused. For supervised learning methods, like random forest, to be a truly viable method of gathering tree canopy data, more reliable labeled data needs to be acquired for training and testing. Alternatively, unsupervised learning methods may be a better option since these methods do not require labeled data sets for training.

Other ways to improve the robustness of the model would be to supplement the Sentinel-2 data with other data sources. After doing some research, I found that similar machine learning efforts to gather tree canopy data utilized LIDAR data as well as spectral band data. LIDAR data would help differentiate vegetation types, like trees from bushes since there tends to be drastic height differences between the two.

One thing I wish I did differently was to partition Marion County by census tract rather than zip codes. Supposedly zip code areas can be variable depending on the organization classifying them. Partitioning by census tract would also allow the merging of census data which would allow for more studies between urban vegetation and socioeconomic factors.

Overall, I think the project was successful because it provided action by providing zip codes with the lowest tree canopy ratios. I do plan on taking the project further and creating a UI using the built in GEE libraries to create a more polished tree canopy map. I also plan on exploring other machine learning algorithms that GEE supports.

GEE Repository Link

https://code.earthengine.google.com/?scriptPath=users%2FpatpageERS%2Fexer02%3AGradProj

References

[1] Architects, J. (2018, October 14). *How landscape architecture mitigates the urban heat island effect*. Land8. Retrieved December 12, 2022, from https://land8.com/how-landscape-architecture-mitigates-the-urban-heat-island-effect/

[2] *Community forestry*. Keep Indianapolis Beautiful, Inc. (n.d.). Retrieved December 12, 2022, from https://www.kibi.org/community-forestry

[3] *Supervised and unsupervised learning*. GeeksforGeeks. (2022, August 24). Retrieved December 12, 2022, from https://www.geeksforgeeks.org/supervised-unsupervised-learning/

[4] Yiu, T. (2021, September 29). *Understanding random forest*. Medium. Retrieved December 12, 2022, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[5] Google. (n.d.). *Harmonized sentinel-2 MSI: Multispectral Instrument, level-2a | Earth Engine Data catalog | google developers*. Google. Retrieved December 12, 2022, from https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED#description