Harnessing Citizen Science to Enhance Land Surface Temperature Prediction

Interns: Kevin Diaz, Kandyce Diep, Suhani Dondapati, Maako Fangajei, Anna Felten, Kei Fry, Conor Furey, Aaren George

Mentors: Andrew Clark, Russanne Low, Peder Nelson, Ollie Snow, Cassie Soeffing, Riya Tyagi

Introduction

Urban heat islands (UHI) are areas indicated by comparatively higher land surface temperatures than surrounding areas, which pose a risk to vulnerable populations and can lead to health issues. Monitoring conditions in these areas is crucial to protect the health of people and the environment.

Figure 1. Simple graphic illustrating an urban heat island. Credit: NASA Climate Kids.

These urban heat islands are strongly related to land cover. Impervious surfaces in cities absorb heat due to their high thermal storage capacity, resulting in higher temperatures, and reduced vegetation compounds this effect.

Our team utilized citizen science to train machine learning models (Random Forest and XGBoost) to predict land surface temperatures. We hypothesized that incorporating vegetation and other land cover data into these models would increase their predictive accuracy.

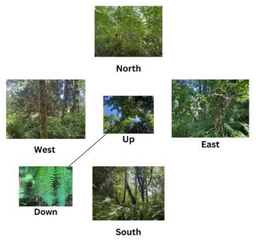

Figure 2. .An example of six directional images taken at a GLOBE Observer sample site. Our research focused on the downward ("down") photo. Credit: GLOBE database.

Literature Review

Land surface temperature has been predicted through Multifractal Detrended Fluctuation Analysis and Spatio-Temporal Semantic Kriging. Random Forest and XGBoost models have promising results for land surface temperature prediction. Moreover, citizen science can be found in biodiversity monitoring, land cover assessment, and climate change studies, but its direct impacts on land cover modeling have yet to be investigated.

Strengths and Limitations of our Study

Several sources of error limit our models' predictive accuracy. Some images had lower quality compared to the rest, resulting in less accurate labels. Additionally, our sample size was relatively small, containing only 437 ground images out of the theoretical maximum of 1480. Our ground images overrepresented a single moderate climate during one season. Finally, the labels did not contain precise quantitative information, as participants were simply asked to select all the land cover types that applied to each ground image.

Figure 8. .An example of a blurry down photo included in the Zooniverse dataset, which may have reduced accuracy of labeled data. Credit: GLOBE database.

Citizen science also introduces both limitations and advantages to the study. An inherent challenge with any program relying on the built-in GPS receiver on a user's device, including the GLOBE Observer app, is spatial accuracy: GPS sensor readings can be inaccurate, especially with older devices. Moreover, using citizen science for data labeling may inherently contribute to lower data quality than could be garnered by researchers.

Nonetheless, citizen science also democratizes research by facilitating a larger volume of data collection across diverse locations. It could serve as a powerful tool in fostering community engagement related to challenges like UHI.

Methods

Directional photos (North, South, East, West) along with zenith (upward) and nadir (downward) were captured using the GLOBE Observer mobile application by the NASA SEES Earth System Explorers intern group.

The Zooniverse platform was utilized to label downward ("down") photos based on land cover with the help of citizen scientists.

Figure 3. Map illustrating the geographic distribution of GLOBE Observer sample sites utilized by the 2024 SEES Earth System Explorer team. Credit: Peder Nelson.

Figure 4. The menu of options presented to Zooniverse volunteers for classifying land cover.

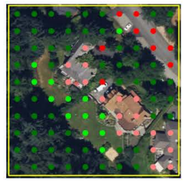

Figure 5. An example 100 x 100 m grid of 100 labeled points on Collect Earth Online. Different colors represent different types of land cover. Credit: Peder Nelson.

Sentinel-2 satellite imagery was labeled using Collect Earth Online and paired to GLOBE down photos based on geographic proximity.

Daily mean LST measurements in Kelvin for June 2024 were found with each site, retrieved from the 30 m multispectral bands on NASA's Landsat-8 satellite using a method developed by Ermida et. al.

Following data preprocessing, which included imputation to fill in missing values, 437 images and their associated LST and land cover use data were deemed fit for analysis.

Random Forest and XGBoost models from sci-kit learn were trained using a 90:10 train/test split and K-fold cross-validation with 10 splits. A Bayesian Search was conducted to find the optimal hyperparameters for each model.

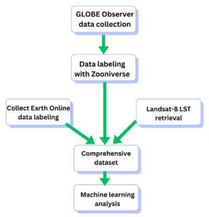

Figure 6. A streamlined overview of the methods used to prepare data for machine learning analysis.

Conclusions

Although the incorporation of land cover data failed to improve predictive accuracy, our team concluded that this was due to a lack of quality data rather than the usefulness of the data itself. Improving data collection methods and ensuring higher quality data could potentially reveal the true value of citizen-sourced land cover data.

Applications:

Through this project, our team supports individuals from various regions to engage in citizen science and develop a better understanding of the complex relationship between land cover and land surface temperature. Our research may equally be useful for urban planners seeking to avoid the onset of the UHI in their urban environment.

Future Work:

Our team recommends investigating the use of pre-trained Convolutional Neural Networks (CNN) to extract relevant features from down photos instead of relying on manual labeling. Moreover, to enhance dataset quality, we suggest collecting image samples from a more diverse array of geographical locations, including colder climates and various elevations.

Results and Discussion

Table 1. Evaluation metrics for three Random Forest models after tuning.

Dataset	R^2	RMSE (K)	MAE (K)
Model 1	0.84	2.49	1.89
Model 2	0.78	3.00	2.30
Model 3	0.79	2.97	2.25

Table 2. Evaluation metrics for three XGBoost models after tuning.

Dataset	R^2	RMSE (K)	MAE (K)
Model 1	0.82	2.86	2.26
Model 2	0.80	3.00	2.36
Model 3	0.82	2.87	2.24

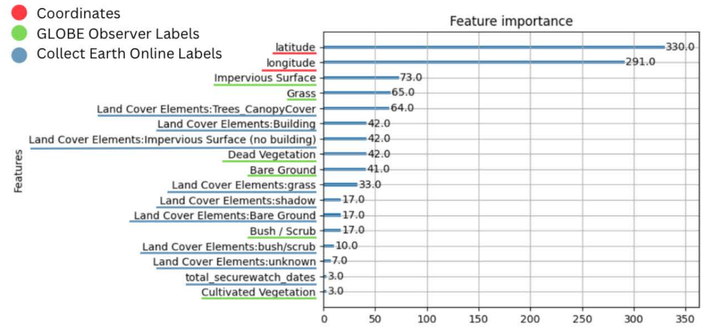

Figure 7. A bar graph showing the features our models used to predict LST, measured using their F-score. Each feature comes from either coordinate data (red), GLOBE Observer Labels (green), or Collect Earth Online Labels (blue).

For both models, the addition of GLOBE Observer image labels resulted in a decrease in predictive performance. Moreover, the addition of the CEO data resulted in a marked improvement in predictive performance, despite only 41% of samples having complete CEO data. This is suggestive that CEO land cover data is a much better predictor of LST than GLOBE Observer nadir image labels.

We deduce from the feature importance plot that the most important features (following longitude and latitude) are from two GLOBE Observer labels, and then two CEO labels. Also, Random Forest models performed slightly better than their XGBoost counterparts across all dataset sizes. This is potentially due to Random Forest performing better on noisy data; exploring the predictive capabilities of a Random Forest model with an expanded dataset would be valuable.

References

ARSET - Fundamentals of Machine Learning for Earth Science | *NASA Applied Sciences*. (2023, April 20). https://appliedsciences.nasa.gov/get-involved/training/english/arset-fundamentals-machine-learning-earth-science

McCartney, S. (2023). Overview and Access of Land Surface Temperature (LST). In *NASA's Applied Remote Sensing Training Program* [Report].

Low, R. D., Nelson, P. V., Soeffing, C., & Clark, A. (2021). Adopt a Pixel 3 km: A Multiscale Data Set Linking Remotely Sensed Land Cover Imagery With Field Based Citizen Science Observation. *Frontiers in Climate, 3*. https://doi.org/10.3389/fclim.2021.658063

NASA/JPL-Caltech. (n.d.). *An illustration of an urban heat island* [Image]. Climate Kids. Retrieved July 19, 2024, from https://climatekids.nasa.gov/heat-islands/

Link to Complete List of Sources:

The material contained in this poster is based upon work supported by the National Aeronautics and Space Administration (NASA) cooperative agreements NNX16AE28A to the Institute for Global Environmental Strategies (IGES) for the NASA Earth Science Education Collaborative (NESEC) and NNX16AB89A to the University of Texas Austin for the STEM Enhancement in Earth Science (SEES). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NASA.