Predicting land cover using GEDI-L2A tree canopy data for New York State: Executive Summary

Frank Seidl, Luke Kiernan, Keavin Moore, Nicholas Barvinok, Noah Rahman
GitHub: https://github.com/fcseidl/erdos-gedi

Overview

The tree canopy height and vertical structure of a given region of Earth's surface can provide important implications for researchers and developers concerned with climatological trends over time. This can provide information to describe a link between important human factors, such as urban population and wildfire persistence, through an area's land cover: developed vs. vegetation. We present a classification model that can **predict land cover (e.g., developed, forested, wetland, open water) using the Global Ecosystem Dynamics Investigation (GEDI)-L2A dataset of tree canopy height and vertical structure** from orbital laser measurements onboard the International Space Station (ISS). We hope to use our best-fit model to predict land cover classifications for another region of interest, such as wildfire-prone California, or the dense forests of northern Europe.

Stakeholders

- Wildfire prevention researchers: determine areas at highest risk of wildfires, based on land cover within a given region (developed areas vs. undeveloped areas of vegetation), since there is a strong link between human activity and wildfires.
- Real estate developers: determine the best region for new housing developments, depending on surrounding land cover classification.
- Solar panel installation companies: find optimal, low-shadow areas based on tree canopy coverage of a region.
- Vegetation and groundwater researchers: use manageable reduced, cleaned, and correctly classified dataset with the most relevant information, instead of working with cumbersome many-GB-sized datasets.

Data Collection & Methodology

We used the **GEDI-L2A Vector Canopy Top Height dataset** which provides orbital laser readings covering a 25-m beam footprint following the orbit of the ISS. We used an API to download the relevant data files for our New York state, selecting the best data based on various quality flags; we chose New York as a test case because it contains forests, plains, lakes/ponds, and urban developments. We performed initial exploratory data analysis on the downloaded GEDI-L2A dataset but were unable to find strong correlations between pairs of features. We hence combine this dataset with the **Multi-Resolution Land Characteristic Consortium (MRLC) land cover data for the contiguous United States**, specified to New York. This land cover data, colored by land cover (i.e., developed, forest, wetland, water), was linked with the GEDI-L2A dataset using latitude-longitude coordinates. We then predict land cover using classifiers indicating whether land is developed or vegetation --- including distinctions between different levels of development and organic regions, such as forests, shrublands, and wetlands --- using some of the 101 relative height '*rh_n*' metrics. These metrics provide the *n*-th percentile of returned beam energy, mapping the tree canopy height as well as its vertical structure. We investigate multiple classification models, including *k*-nearest neighbors, linear/quadratic discriminant analysis, decision trees, and neural networks, evaluating their balanced accuracy score to find the best-fit model. We must reduce the dimensionality of the relative height metrics to a subset to ease model training; this will be iteratively done by selecting different combinations of the metrics and assessing their predictive ability.

Performance Evaluation & Results

We must be careful assessing our model performance: since over 50% of the shots contain forest, random-guessing of forest vs. other would already be 50% accurate. Initial investigations provide a roughly 70% accuracy using a *k*-nearest neighbors approach and around 30-40 neighbors, but we seek to improve this further by removing data with irrelevant columns to land cover prediction, such as 'not classified'. Based on our model testing, we find it easier to predict the land cover 'Group', which contains eight unique classifiers, than 'Class', which is more precise but includes 19 class labels. If the training data is divided using stratified sampling, such that each 'Group' is represented equally, the bias on predicting 'Forest' decreases significantly, while 'Water' becomes the easiest feature to identify. Finally, investigation of the symmetrized confusion matrix for the class prediction problem with spectral clustering automatically identifies sets of commonly confused classes; with two groups, the classes appear to be separated based on tree cover.

We compare our *k*-NN model with another: a neural network approach. By simplifying the problem to a binary classification on the two largest Classes, 'Deciduous Forest' and 'Pasture/Hay', the neural network approach yields an 88% training accuracy, compared to the baseline of 75%. Unfortunately, the model accuracy plateaus after a single epoch. We then ran the multi-class classification on 'Group' instead of on 'Class', which instead plateaus at 70% compared to a baseline of 50%. Even with slight tweaking to the weights and architecture of the neural network, the model mainly settled into the accuracy rates above when training on 'Class' and 'Group'. Assessing the performance of our models, then, our best fit model is *k*-nearest neighbors. The decision tree performance varied strongly (and calibration may vary) with the performance of separation/dimensionality reduction by LDA in advance – a peak accuracy of 81% was reached fold-wise in 5-fold CV with a mean balanced accuracy of 77.8%. This number may improve with more data but one should be wary of consequent over-fitting. Further work could include using AdaBoost and tuning the number of leaves to minimize misclassification error.

Future Directions

- Further optimize tested models, especially decision tree & neural network, and our best-fit *k*-nearest neighbors model.
- Incorporate time-of-year (which correlates with canopy thickness) to inform predictions.
- Compare with historical real estate data for the state of New York. For example, land tax data was recently used in a multiresolution geospatial PM2.5 (particulate air pollution) model for NYC.
- Compare with energy usage data for New York and other regions, to identify ideal regions that will benefit from solar panel installation.
- Apply models to other regions, e.g. connecting to historical wildfire data in California.
- Improve the UI of our model, making a website or app to easily communicate findings to interested stakeholders.