Annual Report: "Identifying Shifts in Forest Communities Using Machine Learning Techniques"

Jonathan A. Knott, Trenton W. Ford, Chathurangi H. Pathiravasan

August 30, 2019

Project overview:

Since the turn of the 20th century, ecologists have debated what defines an ecological community of frequently co-occurring species. Advances in statistics and the collection of large-scale datasets over the last century have provided further evidence of these relationships. Recently, machine learning techniques have presented the opportunity to explore ecological communities in a new, data-driven way, but comparisons between techniques are needed to understand which methods best align with current ecological knowledge. Here, we used two machine learning models, Latent Dirichlet Allocation (LDA), and Node2Vec with USDA Forest Service Forest Inventory and Analysis (FIA) data, spanning 135 tree species and over 75,000 samples across the eastern U.S. to compare and contrast the models and the communities they detect.

Project updates:

Originally, we intended to use three models and to assess changes in forest communities over time using a historic dataset from the 1980s. However, the data from the 1980s contained fewer plots, more sampling bias, and fewer species with consistent records. Similarly, one of the original three models, BigCLAM, did not produce results comparable to other methods. By shifting our analysis to the most recently available FIA data (2015-2017) and reducing our analysis to two machine learning models representing two different ways of utilizing the data (an abundance-based vs. a network-based approach), we are now able to provide a more robust analysis of the communities themselves and highlight the nuances of each method used. For example, in our original dataset, we were using 85 species of interest. Now, we are able to include 135 species since new FIA protocols include species-level identification across the entire range (as opposed to genus-level identification in some states, especially during earlier sampling). Similarly, we encountered an issue with one of our models (Node2Vec) where samples locations without historic data were clustering together, leaving some species missing geographic information and some regions missing community information. Additionally, we intended to use a simple measure of goodness-of-fit (AIC) for our models, but we are now using a suite of metrics to determine which model contains the best-fit number of

communities. Given the lack of machine learning models used in ecological studies, this research remains a novel approach to solving a classical ecological problem even if we are no longer focusing on changes to forest community distributions over time and are instead focusing on the methods themselves.

In our last report (Mar 1, 2019), we had compiled all of the data from the FIA database and had begun some of the modeling. We were waiting for two team members to gain access to Purdue's Research Computing data workbench cluster. We have now secured accounts so all team members can run models and create data visualizations. Although we are still currently running models, much of our modeling framework has been established so that it is easy to make modifications between runs as we further analyze our model output. As such, we have begun the writing process, with most of the introduction and methods completed, and the other sections currently being written. We intend to submit our manuscript to an open-access journal later this fall.

Team Interactions and Meetings:

We have had multiple online team meetings (Zoom video calls on July 17, Sept. 3, and Nov. 19, 2018, and Jan. 7, March 11, and July 9, 2019). Our meetings have been focused on outlining our analysis, discussing steps moving forward, and talking through the implications of our study in preparation for writing. We have an active Slack channel that we use on a daily to weekly basis to provide updates on the analyses, data visualization, and writing. We also met in person at the Symposium on Data Science and Statistics (SDSS 2019) on May 28-June 2, 2019, where we presented our work and discussed more details of our results and future directions (more info below in the "SDSS conference report" section).

SDSS conference report:

We attended the Symposium on Data Science and Statistics (SDSS) in Bellevue, WA at the beginning of summer (May 29-June 1, 2019). At the conference, we presented our findings as a group e-poster presentation. During the poster session, we were able to present our results to many researchers working on a variety of data science topics, including a few who were also working on ecological and natural resources projects. The conference abstract is attached below.

Throughout the week at the conference, we met as a team and discussed the future steps for our

project. We made significant progress on data visualization of the forest communities (which was included in the e-poster presentation and will be adapted for a future publication). We also were able to discuss the project moving forward and how we were planning to finish the analysis and write up the manuscript.

In addition to the work on our specific project, we were able to attend other talks and events and network with fellow data scientists. Jon generally attended statistics education talks as he is hoping to go into academia and teach ecological statistics and GIS to undergraduate and graduate students. Chathu attended sessions focusing on biomedical data science because she is planning to work on a mobile health (mHealth) project for her postdoctoral research. She also participated in a hackathon competition, and her team won best data visualization. Trenton went to sessions about network analysis, aligning with his PhD research. Overall, the team had an excellent experience at SDSS and were able to learn, network, and present our research.

SDSS conference abstract:

Since the turn of the 20th century, ecologists have debated what defines an ecological community of co-occurring species. Advances in statistics and the collection of large-scale datasets over the last century have provided further evidence of these relationships but identifying communities in a data-driven manner has been difficult. Recent machine learning techniques and network analytics tools present opportunities to explore ecological communities in new, data-driven ways, but comparisons between techniques are needed to understand which methods best align with current ecological knowledge. During this research, we applied three machine learning models, Latent Dirichlet Allocation (LDA), Cluster Affiliation Model for Big Networks (BigCLAM), and Metapath2vec, to three decades of U.S. Forest Service Forest Inventory and Analysis (FIA) data, spanning 85 tree species and 70,000 plots across the eastern United States. The models showed that the best-fit number of communities, k, varied between the model input (relative vs. absolute measures of species abundance and sapling vs. adult stems) and the method used (LDA usually found more communities than BigCLAM and Metapath2vec). However, the community composition (the mixture of species within a community) when k was kept constant, was consistent between methods. These methods were also able to identify changes in both the geographic distributions of communities over time and the overlap between communities within a sampling unit, which have close links to ecological processes. For example, observed reductions in communities associated with the Fraxinus (ash)