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

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

March 1, 2019

Problem Background :

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 are planning to apply 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 86 tree species and 70,000 plots across the eastern U.S. to compare and contrast the three models and the communities they detect.

Findings :

These methods were also able to identify changes in both the geographic distributions of the communities over time and the overlap in communities within a sampling unit, which have close links to ecological processes. For example, a decline in communities associated with Fraxinus (ash) species detected with the BigCLAM model might be a result of the invasive emerald ash borer beetle. Conversely, some communities have also shown little change over recent years, which may indicate a lag between anthropogenic impacts (such as climate change) and forest community responses. The relationship between these stressors and the shifts in forest communities can provide insight into the future sustainability of forest ecosystems across the eastern U.S.

Team Interactions/Meetings :

We had multiple online team meetings (Zoom conference calls), and our meetings have been focused on outlining our project and creating a plan for the analyses we intend to carry out. We have an active Slack channel that we keep in touch through on a daily to weekly basis. We are also planning to meet at the Symposium on Data Science and Statistics (SDSS 2019) on May 28 and discuss more details of our results and future directions (more info below in the "Remaining project timeline" section).

Remaining project timeline :

Currently, we have compiled all of the data from the FIA database to be used in the models. We have also begun some of the modeling, primarily the LDA model. Over the next weeks, we will continue to run models on the Purdue RCAC data workbench server. We expect to gain full access to the server for Trenton and Chathurangi within the next two weeks, so that they can begin to run additional models, summarize and compare results, and create output figures for upcoming presentations.

We are planning to attend the Symposium of Data Science and Statistics (SDSS) at the end of May, so we will begin working on preparing our presentation in late April/early May. This presentation will be an e-poster, so we plan to use an online platform such as Overleaf to prepare our presentation. We also plan to arrive early to the conference to finalize our presentation in person.

After the SDSS conference, we intend to begin drafting a manuscript on our project. We have started an outline for this manuscript, but will focus on writing after the SDSS conference. This manuscript will likely be submitted to the journal Bioscience due to the mixture of application, theory, and methods in this project.

List of conferences planning to attend :

- E-poster presentation at the Symposium on Data Science & Statistics in Bellevue, Washington. It will be held on May 29 - June 1, 2019 (Abstract submitted).