

Center for the Science of Information Interdisciplinary Research Team Project Proposal

Title:

Identifying Shifts in Forest Communities Using Machine Learning Techniques

Project Type:

Interdisciplinary team project with intent to present results at an international conference

Project Date:

September 2018 - April 2019

Total Funds Requested:


\$6000

Student PIs:

Jonathan Knott¹, Trenton Ford², and Chathurangi Pathiravasan³

Faculty Advisors:

Songlin Fei, PhD¹  _____

Nitesh Chawla, PhD²  _____

Bhaskar Bhattacharya, PhD³  _____

¹Department of Forestry and Natural Resources, Purdue University
195 Marsteller Street, West Lafayette, IN 47907

²Department of Computer Science and Engineering, University of Notre Dame
384E Nieuwland Science Hall, Notre Dame, IN 46556

³Department of Mathematics, Southern Illinois University
1245 Lincoln Drive, Carbondale, IL 62901

1 Problem Statement

The definition of forest communities has been a topic of interest for over a century, and recent advancements in computational capacity and data availability have allowed the analysis of these communities at much larger scales and with more statistical power than in the past. Using the Forest Inventory and Analysis (FIA) dataset from the U.S. Forest Service and recently developed machine learning techniques, we aim to (1) identify the communities of forest species in the eastern U.S., (2) assess their changes over time, and (3) address the possible causes of community change over time, such as climate change, invasive species, and land use.

2 Intellectual Merit and Broader Impacts

As ecology enters a critical era, with vast quantities and varieties of data and the associated computational resources to analyze these data, it is necessary to revisit and expand upon previous understanding of how ecological systems function. For over a century, ecologists have struggled to identify and describe communities of species and their dynamics due to conflicting theories of how to define a “community.” From early works based on localized field data (Clements, 1916; Gleason, 1926), to more recent efforts that take into account regional influences (Ricklefs, 1987; Dyer, 2006), community ecology has flourished with many different viewpoints of what makes a community. Now in a new era of data-driven ecology, it is possible to incorporate concepts of big data (such as the use of FIA) and machine learning into our understanding of ecological communities.

However, although data availability has improved, until recently statistical methods and computational capabilities have caused a bottleneck in the process of properly applying concepts of community ecology to big-data. For example, the “double-zero problem” (Legendre and Legendre, 2012) can occur at a large spatial scale when several species never or rarely co-occur in a sampling unit. The resulting communities can be defined by the lack of a certain species or group of species, rather than be defined by actual species interactions. Given the recent advancements in computational power and machine learning, we are now able to identify these communities based on their presence and co-occurrence. Much like how machine learning algorithms can identify which users in a social network are talking to one another to form an online community, these algorithms can also identify which species co-occur together to form a forest community. In addition, the availability of nation-wide high-quality climate, invasive species, and socio-economic data allows the assessment of major threats to the forest communities at a regional scale. In particular, our project aims to provide a case study for using machine learning in community ecology research.

The outcome of this project will also provide data-driven analyses that can be utilized for forest managers and decision makers. The results of this study will broaden our understanding of the interconnectedness of species across the eastern U.S. Forest management requires careful assessment of the local species pool, as forest communities respond differently to various management practices. Forest managers and decision makers can properly manage the valuable forest resources when properly utilizing information about the dynamics of their forest community. In addition, the results of this analysis can help to protect threatened forest communities by identifying those that are most impacted by invasive species, climate change, and human disturbance.

3 Proposed Activity

3.1 Community Detection

The use of robust community detection techniques is a fundamental part of this project’s success. To support the identification of changes within forest communities and similarly, the dynamic communities that exist across all data time-slices, we must first establish credible communities at the individual time-slices for which data exists. For this project, the data available is being separated into two distinct

time-slices (1980s and 2010s) for which static community structures will be discovered. Using these two static community structures we will create a single dynamic community structure from which we can make comparisons and potentially find, or strengthen, higher-order community structures.

Choosing which machine learning techniques to apply requires both an understanding of the data, and an understanding of what constitutes appropriate results. Given the data, we can extract communities using a wide variety of techniques such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Non-Negative Matrix Factorization (NMF) (Yang and Leskovec, 2013), or a mixture of Node2vec (Grover and Leskovec, 2016) and K-nearest neighbors (KNN) (Keller et al., 1985). Consideration for the appropriateness of community results is the more powerful discriminant. These methods allow us to identify the community associations for each species which allow comparisons of the communities across time.

3.2 Inference

After identifying communities in our data, we will utilize tools for comparing communities to detect changes over time. Species that are associated with fewer communities in recent years than in the past may signify a need for further scrutiny. We intend to address three main questions for these species of interest: (1) Where has this species declined? Are there particular regions that might be associated with loss of this species? (2) What are possible causes of this decline? Are declines in this species or its community associations related to climate, management practices (such as plantation efforts or deforestation), or disease and exotic species invasion? (3) Which other species are impacted by this loss? Are species in the same community as the species of interest increasing or decreasing in abundance?

In addition to the species-level changes over time, community-level changes are also important, as they can indicate losses in ecosystem functioning. One change we expect to find are shifts in the spatial distribution of the communities. Fei et al. (2017) found that many tree species in the eastern U.S. are shifting their distributions westward and northward; therefore, it is reasonable to expect that the distribution of the communities are also shifting geographically. In order to track these changes, we propose the use of the abundance weighted-centroid (i.e. center-of-mass) as a proxy for the communities' distribution, as the weighted-centroid can identify within-distribution shifts that are not apparent when analyzing the leading or trailing edge of the communities (Fei et al., 2017).

We intend to use distance metrics (such as Sørensen's distance) to measure the compositional change across time. Using Sørensen's distance, Thompson et al. (2013) found that compositional change in the northeast U.S. is most strongly associated with the historical extent of agricultural clearing. To build on Thompson et al. (2013), we can use regression tree analysis and Random Forest models with the Sørensen's distance between time periods as the response variable and a variety of predictor variables (such as climate change, temperature, rate of land use change, etc.).

3.3 Close Examination

Some species have known declines over the study period, and we plan to assess how these declines impact forest communities. For example, at the CSoI workshop in May 2018, we found one species, the black ash (*Fraxinus nigra*) was associated with three communities in the 1980s, but is currently only associated with two communities. This decline in the number of community associations may reflect the decrease in black ash abundance due to the invasion by the emerald ash borer beetle (*Agrilus planipennis*) during the last two decades. For this species and other species of interest identified by the community detection and inference stages (see sections 3.1 and 3.2), we plan to investigate the causes of shifts in community association, including invasive pests, changes in climate, and human disturbance.

We hypothesize that changes at the species level are manifested at the community level. If the species within a community are impacted by a variety of threats, then the community as a whole is impacted; however, due to the complexity of communities versus individual species, we hypothesize that these changes are not as large for communities as they are for species. It is likely that as one species in a community

declines, others in the community compensate for this loss, rather than the community as a whole declining. Our planned analyses are robust enough to validate our hypothesis.

4 Goals and Outcomes

This project has four main goals, two related to the ecological question, and two related to the presentation of results at an international conference and drafting of a manuscript. The goals are:

1. Identify and describe the forest communities using machine learning techniques.
2. Infer ecological implications and potential drivers of community change over time.
3. Present results at the 2019 International Conference on Forest Ecology, Conservation, and Management (ICFECM) in Rome, Italy.
4. Draft a manuscript for this project for potential publication in a high-tier journal such as *Global Change Biology* or *Diversity and Distributions*.

5 Proposed Work Statement

This is an interdisciplinary project that involves combining big-data and machine learning with on-the-ground ecological applications. Team member JK will gain experience using machine learning, an up-and-coming topic in ecology. Team members TF and CP will be able to utilize their expertise in machine learning and statistics with a real-world problem, providing them with experience working outside of the theoretical realm of computer science and mathematics.

In addition, the team members will contribute to the project via:

- Team meetings that will consist of video conference sessions at least once a month, more frequently if necessary.
- An in-person meeting at Purdue University in early 2019 to plan for the conference presentation
- Communication through a Slack workspace for intermediate conversation regarding progress on the project.
- Data and code shared on a BitBucket repository so that each team member can update and track progress.
- Unique contributions to this project:
 - **JK:** He is interested in the ecology of the study system. He will focus on outlining the project in an ecological context and provide ecological interpretations of results. In addition, he will perform some of the analyses using the Purdue Rosen Center for Advanced Computing (RCAC) DataWorkbench cluster.
 - **TF:** He is interested in the machine learning techniques used for community detection. He will focus on performing analyses using the various machine learning techniques and will provide knowledge about the implementation of these techniques.
 - **CP:** She is interested in investigating and analyzing the causes for forest decline. She is planning to use Random Forest models and Regression Trees to evaluate the relationships between compositional change and the suite of predictor variables (land use, climate change, and bio-physical predictor variables)

6 Diversity Statement

Our team brings specialties in ecology, mathematics, and machine learning as well as vastly different personal and cultural backgrounds. This diversity informs not only the means by which we answer the research question but also the language and manner that we use to present the question and our findings to an audience. Given the confluence of disciplines employed to answer this research question, it would be easy to lose a reader well-versed in a single discipline in the minutia of another. To this end, the resulting research findings will be made as simple as possible, but no simpler.

As it relates to academic diversity, each of our team members come from different disciplines and research foci. This eclecticism has allowed for the generation of wide ranges of questions to be answered, and brings together a large set of tools to answer them. Our research team consists of three PhD students from three different universities (see names and affiliations above). In addition to academic diversity, two team members are US citizens and one team member is a woman. All three team members are from racial minorities.

7 Budget and Justification

Travel to ICFECM 2019: 21st International Conference on Forest Ecology, Conservation, and Management, April 9-10, 2019 in Rome, Italy for three graduate students.

Item	Cost	Quantity	Total
Registration	\$400	3	\$1200
Flight	\$1400	3	\$4200
Lodging	\$120	9 (3 nights/each)	\$1080
		Total:	\$6480
		Amount requested:	\$6000

References

- David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.
- Frederic E. Clements. *Plant succession: An analysis of the development of vegetation*. Carnegie Institute of Washington Publication, Washington, DC, 1916.
- James M Dyer. Revisiting the deciduous forests of eastern north america. *BioScience*, 56(4):341–352, 2006.
- Songlin Fei, Johanna M Desprez, Kevin M Potter, Insu Jo, Jonathan A Knott, and Christopher M Oswalt. Divergence of species responses to climate change. *Science Advances*, 3(5), 2017.
- H. A. Gleason. The individualistic concept of the plant association. *Bulletin of the Torrey Botanical Club*, 53(1):7–26, 1926.
- Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *KDD*, 2016.
- James M Keller, Michael R Gray, and James A Givens. A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, (4):580–585, 1985.
- Pierre Legendre and Loic FJ Legendre. *Numerical Ecology*, volume 24. Elsevier, 2012.
- Robert E Ricklefs. Community diversity: relative roles of local and regional processes. *Science*, 235(4785): 167–171, 1987.
- Jonathan R Thompson, Dunbar N Carpenter, Charles V Cogbill, and David R Foster. Four centuries of change in northeastern united states forests. *PLoS One*, 8(9):e72540, 2013.
- Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 587–596, New York, NY, USA, 2013. ACM.