An Image is Worth a Thousand Species: Scaling high-resolution plant biodiversity prediction to biome-level using citizen science data and remote sensing imagery

Lauren Gillespie[‡,§], Megan Ruffley[§], Moisés Expósito-Alonso[§,|]

‡ Department of Computer Science, Stanford University, Stanford, United States of America
§ Department of Plant Biology, Carnegie Institution for Science, Stanford, United States of America
| Department of Biology, Stanford University, Stanford, United States of America

Corresponding author: Lauren Gillespie (gillespl@cs.stanford.edu)

Abstract

Accurately mapping biodiversity at high resolution across ecosystems has been a historically difficult task. One major hurdle to accurate biodiversity modeling is that there is a power law relationship between the abundance of different types of species in an environment, with few species being relatively abundant while many species are more rare. This "commonness of rarity," confounded with differential detectability of species, can lead to misestimations of where a species lives. To overcome these confounding factors, many biodiversity models employ species distribution models (SDMs) to predict the full extent of where a species lives, using observations of where a species has been found, correlated with environmental variables. Most SDMs use bioclimatic environmental variables as the dependent variable to predict a species' range, but these approaches often rely on biased pseudo-absence generation methods and model species using coarse-grained bioclimatic variables with a useful resolution floor of 1 km-pixel.

Here, we pair iNaturalist citizen science plant observations from the Global Biodiversity Information Facility with RGB-Infrared aerial imagery from the National Aerial Imagery Program to develop a deep convolutional neural network model that can predict the presence of nearly 2,500 plant species across California. We utilize a state-of-the-art multilabel image recognition model from the computer vision community, paired with a cutting-edge multilabel classification loss, which leads to comparable or better accuracy to traditional SDM models, but at a resolution of 250m (Ben-Baruch et al. 2020, Ridnik et al. 2020). Furthermore, this deep convolutional model is able to accurately predict species presence across multiple biomes of California with good accuracy and can be used to build a plant biodiversity map across California with unparalleled accuracy. Given the widespread availability of citizen science observations and remote sensing

imagery across the globe, this deep learning-enabled method could be deployed to automatically map biodiversity at large scales.

Keywords

biodiversity mapping, machine learning, species distribution models

Presenting author

Lauren Gillespie

Presented at

TDWG 2021

Funding program

Conflicts of interest

The author reports no outstanding conflicts of interest

References

- Ben-Baruch E, Ridnik T, Zamir N, Noy A, Friedman I, Protter M, Zelnik-Manor L (2020) Asymmetric Loss For Multi-Label Classification. arXiv:2009.14119 [cs] URL: http://arxiv.org/abs/2009.14119
- Ridnik T, Lawen H, Noy A, Baruch EB, Sharir G, Friedman I (2020) TResNet: High Performance GPU-Dedicated Architecture. arXiv:2003.13630 [cs, eess] URL: http://arxiv.org/abs/2003.13630