Using species distribution models to direct native seed collection efforts

[1]Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, Illinois 60022, USA

Reed Clark Benkendorf[1]*

Abstract

Developing many locally adapted germplasms for habitat restoration has been a challenging process. In particular, finding a large enough number of populations which can serve as source material for agricultural increase across the range of species is proving difficult. The population scouting efforts of native seed collection crews can be informed using Species Distribution Models, allowing them to prioritize areas with higher probabilities of successfully finding the target species. However, SDM's themselves do not have a mechanism for addressing the discontuity between suitable habitat and dispersal limitation, requiring that some of connectivity analysis be performed as a post-processing step to aid in their interpretation. Herein we detail the high-throughput and semi-automated development of SDM's for over 350 common taxa native to Western North America. We further employ simple connectivity analysis, based solely on drainage basins and neighborhood analysis, on these hypothesis to try and ensure that areas with higher probabilities of propagule dispersal and population establishment are visited. Using the hypothesis of suitable habitat generated by these hypothesis, with ground verification data from 13 crews in 5 Western states, we show that SDMs are useful for crews to find populations of native species, and that they are most useful when combined with connectivity analysis. Our workflow is readily replicable by users on a variety of low-cost computing systems, and our implementation strategies for field crews are briefly discussed.

Implications for Practice

- Finding enough large populations across seed collection zones (e.g. seed transfer zones) for economically viable agricultural increase has proven to be difficult.
- Heuristic tools which can be generated cheaply and computationally such as Species Distribution Models,

*Author for Correspondence: rbenkendorf@chicagobotanic.org

and connectivity analysis, offer a possibility to increase the number of populations eligible for wild land seed collection.

- Our code and protocol for generating and utilizing these tools are available and widely applicable to other users.

Introduction

With the United Nations 'Decade of Restoration' well underway limitations in our capacity to carry out active restoration, which normally requires large volumes of seed, are becoming apparent (National Academies of Sciences et al. 2023). Given the scale of our seed-based restoration needs, in most scenarios, the only sustainable source of this quantity of seed are from the amplification ('grow-out') of wild harvested seed in agricultural settings (Pedrini et al. 2020; Broadhurst et al. 2015; National Academies of Sciences et al. 2023). Enormous efforts are now underway to increase the number of species, the number of populations within these species, and the genetic diversity of the seed available from agricultural amplification for restoration projects Merritt & Dixon (2011). However, numerous difficulties exist in both the wild harvest of seed and it's subsequent amplification which are limiting our ability to develop adequate amounts of germplasm.

While most species desired in restorations have historically had relatively expansive geographic ranges, multitudes of populations - many of which had enormous census sizes, the development of native germplasm remains behind targets (National Academies of Sciences et al. 2023). We posit that in part this is due to the difficulty of finding populations with the appropriate number of individual plants, which are experiencing climatic conditions conducive to producing enough viable seed which can be sustainably harvested to begin economically feasible agricultural increase. Essentially the same reasons we need to enact active restoration in the first place, widespread habitat degradation and unnatural wildfires spurred by climate change and ill-informed historic land management practices, are the same hindrances to developing a locally-adapted and affordable germplasm *(Abatzoglou et al. 2021)).* Heuristic tools which are capable of predicting a species geographic range, the presence of populations across the range, and which can predict them across seed collection target units (e.g. seed transfer zones), such as Species Distribution Models offer promise to increase the rate at which native germplasm can be developed.

A Species Distribution Model (alternatively referred to as 'Environmental niche models' etc.) uses presence, and optionally absence, records of a species and correlates these records to environmental variables (e.g. climate, land cover, and soil), using various families of statistical models (e.g. GLMs, MaxEnt, decision trees), which allows for the fit model to be predicted onto a gridded (raster) surface. While Species Distribution Models

(SDMs) only generate hypothesis of whether areas have conditions similar to the observed environmental niche of the species based on the training records these hypothesis are rarely tested, and the link between their predictions and ecological reality have been credibly questioned as being tenuous (A. Lee-Yaw et al. 2022). In the few published instances where SDMs have been tested, they have oftentimes been for rare species which oftentimes have rather obvious and consequential drivers of their distributions, e.g. they proliferate on uncommon soils and are outcompeted on more typical soil types (Johnson et al. 2023). Whether the hypotheses generated by these models may be useful for detecting populations or common species, whereby their distributions often appear more governed by more subtle cues, appears unexplored. Testing the applicability of SDMs to modeling the range of common species is further valuable as much biodiversity planning incorporates common species which are oftentimes viewed as the fabric holding together ecosystems, and as they may assist with planning restoration plantings under climate change.

While the goal of an SDM is to develop a testable hypothesis of where a self sustaining population of a species *can* grow, they generally lack an explicit link to estimating *where* a population will actually grow (Guisan & Thuiller 2005)). We argue that where populations will occur is driven largely by dispersal limitations, rather than simply whether habitat is suitable (Franklin 2010). A possible tool to associate a probability of occurrence of a species in a suitable habitat patch, is connectivity analysis. Connectivity analysis, as generally implemented in land management, often requires a variety of computationally intensive algorithms stemming from a variety of disciplines (e.g. circuit theory). However we posit that in the case of widespread species useful results can be obtained by simply detailing the occupancy of drainage catchments, and the number of catchments between these occupied sites, and drainages with hypothesized potential habitat.

In this article we use empirical field verified data to explore two main concepts developed from a large number of SDMs across an expansive geographic range. The first is that SDM's are useful for finding populations of species, and that the number of populations found increases with habitat suitability. The second is that SDM's are most useful when combined with knowledge of distance from the predicted to the observed cell.

2 MATERIALS AND METHODS

2.1 Study system

The 353 taxa modeled were selected by land managers working for the Bureau of Land Management in arid & semi-arid areas Western North America and are species they already use, or which they intend to develop for use, in restoration. The spatial extent of modelling broadly encompasses the Western United States being bounded by the Pacific Ocean, the 50th parallel at north, -100 degrees at East and Mexico to the south. This domain has tremendous variation in amounts of elevation, temperature, and precipitation.

Species Occurrence Records

Species presence records were collected from herbaria, citizen science initiatives (e.g. iNaturalist), and standardized ecological monitoring programmes (e.g. VegBank, Forest Inventory and Analysis) using R (Chamberlain et al. 2024; Maitner 2023; Michonneau & Collins 2023). These records were filtered to only those collected after 1950, with coordinate uncertainty less than 250 meters, and only the most recent record per 90m cell was retained. All of these records were then manually reviewed by species, where records with more 2 or more variables in the 1.5% quantiles of several environmental variables (BIO1, BIO4, BIO10, BIO12, & BIO19; see TABLE X) and distance were flagged. During subjective manual review all digitized herbarium specimens which were suspect were reviewed, while other records were dropped based on the analysts review of other occurrence records.

Species absences were generated using three processes, and we sought to have a 1:1 ratio of presences to absences. True absences were acquired from an ecological monitoring program Assess, Inventory, and Monitor (AIM), these absences accounted for a percent equivalent to land managed by BLM within the species range (e.g. if 20% of land ownership across a species range was BLM, than the number of presences * 0.2 defined the number of these absences). Likely absences, representing 15% of records, were generated outside the known range of the species, but bounded to be within 50 miles of the extent extent of occurrences. Pseudo-absences (PA) were randomly selected from areas in, or within XX km, of the species range but greater than 10km from an occurrence, these records accounted for $((1 - (\% \text{ BLM Land} + 15\% \text{ LA records})) * 1.25)$. Because most of these species are highly abundant in order to reduce the probability that a PA was drawn within a species the environmental niche linear discriminant analysis (LDA) was used. The LDA utilized the presences, true absences and likely absences as the classes and several environmental variables (Venables & Ripley 2002), identified by vifstep with theta = 10 as displaying at most moderate collinearity, as independent variables (Naimi et al. 2014). As many records, classified by the LDA as originating from the presence data set, from the pseudo-absences could be removed as to achieve a class balance between presences and absences.

Independent Variable Processing

Up to 44 variables were used in generating the species distribution models (Appendix 1). These variables were selected based on authors previous work, and represent variables commonly associated with the empirical distributions of taxa in arid and semi-arid regions. The thematic contents of these variables largely relate to climate (Karger et al. 2017), landcover (Tuanmu & Jetz 2014), topography and landform (Amatulli et al. 2020 ; Yamazaki et al. 2017), soil (Hengl et al. 2017; Ivushkin et al. 2019), and representations of anthropogenic impacts (Sanderson et al. 2002). All variables were downloaded from source and re-sampled to 90m resolution using terra (Hijmans 2024).

Species Distribution Modelling

Random Forests models were generated for each species following several steps to reduce the number of features. While superfluous features do not notably decrease the performance of random forests, they increase the amount of time required to predict models onto gridded surfaces. In our experience, predicting models onto raster surfaces is the most time consuming step of SDM generation. The Boruta algorithm was used on the full stack of 44 independent variables to remove un-informative variables (Kursa & Rudnicki 2010), and the variable importance factor (VIF) scores were then used to subset the most informative features in several auto-correlated pairs, reducing the independent variables to at most 33 (Naimi et al. 2014). Recursive Feature Elimination (rfe) was then used to identify the fewest number of independent variables which could either increase model performance (measured using accuracy), or only decrease it by 1.5% relative to the full set of variables, these variables were subset and used as features for random forest modelling, using the 'randomForest' package with all default values except for optimized mtrys (Kuhn 2008; Liaw & Wiener 2002). Models were then predicted onto raster surfaces which exceeded the species known range, based on the training and test data, in all directions by 50 miles (Hijmans 2024).

Patch identification

To identify putative metapopulations raster cells predicted as having less than 0.8 probability of suitable habitat were masked as NA's, and then areas which were crossed by streams with Strahler orders of three of more (2004), and the divides of HU10 watersheds (2023) were 'burnt' away from the raster. The resulting rasters were aggregated by a factor of 2 to 5, depending of their sizes (< 300 MiB 2, < 500 MiB 3, < 700 MiB 4, > 700 MiB 5) to accommodate, a rooks case search using terra in a scalable time period. Resultant patches < 5 acres were then discarded, and the rasters was resampled to it's input resolution.

Putative populations were identified using the patches generated above. These patches followed the same processing, except that HU12 watersheds were used to delineate populations.

Because the > 0.8 threshold used above did not capture all colonized patches, all patches with predicted suitability scores of >0.55 which contained species observations were then identified using Terra's patches to create a data set of known populations.

Predicting Species Occurrence in Patches

All patches identified above were used to identify patches within 5 contiguous neighbors, or 5 kilometers, of a patch known to be occupied. Occupied patches were determined using both the training and test occurrences data set used to generate the SDMs. To identify each patch within 5 orders of contiguous neighbors `nblag`, and to determine all patches within 5k of a populated patch `dnearneigh`, were used (Bivand & Wong 2018). For each of these non-occupied patches the number of occupied patches at different lags were counted.

Each non-occupied patch was assigned an arbitrary rank based upon whether they were contiguous with an occupied patch, and if contiguous than their lag number to the nearest occupied patch, and the number of occupied patches connected (TABLE XX). The arbitrarily assigned numbers increase from '1', for an occupied patch, to '7' for a patch which has fewer than 3 second-order contiguous neighbors to an occupied patch(es).

Connection	No. Connections	Rank
1st	>= 2	2
1st	1	3
2nd	>= 3	4
2nd	<= 2	5
3rd	>= 4	6
3rd	<= 3	7

Patches which have no contiguous neighbors, but which have neighbors within 5km were also assigned rank values based in this system (TABLE XX).

No. Occupied Patches	Rank
>= 3	8
<= 2	9

Results

Of the 9048 records which were flagged, and manually investigated by an analyst a total of 3581 were removed. Of these 332 manually reviewed taxa, 66 had no records removed, the mean proportion of records which were removed (based on the entire number of records present) was 0.011, median = 0.003, with a maximum of

0.388. These values are skewed upwards due to a few species which serve as common 'dumping' grounds for difficult to identify taxa in species rich clades.

While the training accuracy scores for the linear-discriminant analysis were high all across the board (mean = 0.9997, min = 0.987), the realized range on testing values were lower, and reflected values more typical of complex applications (mean = 0.77, median = 0.789, min = 0.317). Of the randomly generated psuedo-absences which underwent linear discriminant analysis for 335 taxa, in only 13 instances did the number of records flagged for removal not exceed the number of records flagged. On average 0.175 records were flagged per species (median = 0.165).

The final models utilized 33 variables, the variables which showed up in the most models, also tended to be the most important in driving decisions (FIGURE X). In general climate variables, especially the Bioclim variables, tended to be the most important in decreasing mean Gini, with variables of other types also contributing to models in lesser capacities (FIGURE X). On the partitioned data, most models performed very well across a variety of diagnostic metrics (FIGURE XX). Area under the receiver operator curve (AUC) values were generally the highest with a mean = 0.945, median = 0.953, kappa scores were generally lowest with a mean = 0.772, median = 0.778, and balanced accuracy scores were intermediate with mean = 0.885, median = 0.888

How were model evaluation statistics?

Where were the populations crews ended up finding? weird elbow plot) How did presences and absences interact? (logistic regression)

Discussion

Acknowledgments

The development of the species distribution models, and the connectivity data products, were funded on a contract awarded to the Chicago Botanic Garden from the Bureau of Land Management. Ground verification was conducted in the routine operations of 13 contract crews funded by the BLM on the same set of agreements. Kristy Snyder is thanked for sharing Seeds of Success field data which were required for the analyses in this manuscript.

Literature Cited

Supporting Information

(2004) National hydrography dataset.

(2023) Watershed boundary dataset.

A. Lee-Yaw J, L. McCune J, Pironon S, N. Sheth S (2022) Species distribution models rarely predict the biology of real populations. Ecography 2022:e05877

Abatzoglou JT, Battisti DS, Williams AP, Hansen WD, Harvey BJ, Kolden CA (2021) Projected increases in western US forest fire despite growing fuel constraints. Communications Earth & Environment 2:1–8

Amatulli G, McInerney D, Sethi T, Strobl P, Domisch S (2020) Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. Scientific Data 7:162

Bivand R, Wong DWS (2018) Comparing implementations of global and local indicators of spatial association. TEST 27:716–748

Broadhurst L, Driver M, Guja L, North T, Vanzella B, Fifield G, Bruce S, Taylor D, Bush D (2015) Seeding the future–the issues of supply and demand in restoration in australia. Ecological Management & Restoration 16:29–32

Chamberlain S, Barve V, Mcglinn D, Oldoni D, Desmet P, Geffert L, Ram K (2024) Rgbif: Interface to the global biodiversity information facility API.

Franklin J (2010) Moving beyond static species distribution models in support of conservation biogeography. Diversity and Distributions 16:321–330

Guisan A, Thuiller W (2005) Predicting species distribution: Offering more than simple habitat models. Ecology letters 8:993–1009

Hengl T, Mendes de Jesus J, Heuvelink GB, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B, others (2017) SoilGrids250m: Global gridded soil information based on machine learning. PLoS one 12:e0169748

Hijmans RJ (2024) Terra: Spatial data analysis.

Ivushkin K, Bartholomeus H, Bregt AK, Pulatov A, Kempen B, De Sousa L (2019) Global mapping of soil salinity change. Remote sensing of environment 231:111260

Johnson S, Molano-Flores B, Zaya D (2023) Field validation as a tool for mitigating uncertainty in species distribution modeling for conservation planning. Conservation Science and Practice 5:e12978

Karger DN, Conrad O, Böhner J, Kawohl T, Kreft H, Soria-Auza RW, Zimmermann NE, Linder HP, Kessler M (2017) Climatologies at high resolution for the earth's land surface areas. Scientific data 4:1–20

Kuhn M (2008) Building predictive models in r using the caret package. Journal of Statistical Software 28:1–26

Kursa MB, Rudnicki WR (2010) Feature selection with the boruta package. Journal of Statistical Software 36:1–13

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2:18–22

Maitner B (2023) BIEN: Tools for accessing the botanical information and ecology network database.

Merritt DJ, Dixon KW (2011) Restoration seed banks—a matter of scale. Science 332:424–425

Michonneau F, Collins M (2023) Ridigbio: Interface to the iDigBio data API.

Naimi B, Hamm N a.s., Groen TA, Skidmore AK, Toxopeus AG (2014) Where is positional uncertainty a problem for species distribution modelling. Ecography 37:191–203

National Academies of Sciences Engineering, Medicine, others (2023) An assessment of native seed needs and the capacity for their supply.

Pedrini S, Gibson-Roy P, Trivedi C, Gálvez-Ramìrez C, Hardwick K, Shaw N, Frischie S, Laverack G, Dixon K (2020) Collection and production of native seeds for ecological restoration. Restoration Ecology 28:S228–S238

Sanderson EW, Jaiteh M, Levy MA, Redford KH, Wannebo AV, Woolmer G (2002) The human footprint and the last of the wild: The human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not. BioScience 52:891–904

Tuanmu M-N, Jetz W (2014) A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. Global Ecology and Biogeography 23:1031–1045

Venables WN, Ripley BD (2002) Modern applied statistics with s. Fourth. Springer, New York

Yamazaki D, Ikeshima D, Neal JC, O'Loughlin F, Sampson CC, Kanae S, Bates PD (2017) MERIT DEM: A new high-accuracy global digital elevation model and its merit to global hydrodynamic modeling. In: AGU fall meeting abstracts.Vol. 2017 pp. H12C–04.