Using Species Distribution Models to inform native seed collection efforts

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

Reed Clark Benkendorf[1]*

Abstract

words

IMPLICATIONS FOR PRACTICE

up to six of these. let's shoot 3 with 2 sentence max per.

Introduction

A primary challenge to restoring Earths terrestrial ecosystems is the lack of available plant germplasm (National Academies of Sciences, Medicine & others (2023), Merritt & Dixon (2011)). Given the scale of our restoration needs, in most scenarios, the only sustainable source of seed is from grow-outs of wild harvested seed in agricultural settings (Pedrini, Gibson-Roy, Trivedi, Gálvez-Ramìrez, Hardwick, Shaw, Frischie, Laverack & Dixon (2020), Broadhurst, Driver, Guja, North, Vanzella, Fifield, Bruce, Taylor & Bush (2015), National Academies of Sciences, Medicine & others (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 for restoration National Academies of Sciences, Medicine & others (2023). However, numerous difficulties exist in both the wild harvest seed and it's increase which are limiting our ability to develop adequate amounts of germplasm.

While most species desired in restorations have historically had relatively large geographic ranges, numbers of populations, and number of individuals per populations, the development of native germplasm remains behind targets (National Academies of Sciences, Medicine & others (2023)). We posit that in part this is due

*Author for Correspondence: rbenkendorf@chicagobotanic.org

to the difficulty of finding populations with the appropriate number of individuals, which are experiencing climatic conditions conducive to producing enough viable seed to being agricultural increase, a complications borne of widespread habitat degradation and unnatural wildfires *(Abatzoglou, Battisti, Williams, Hansen, Harvey & Kolden (2021))*. Tools which are capable of predicting a species geographic range, the presence and size of populations across the range and in seed collection target units (such as empirical or provisional seed transfer zones or ecoregions), such as Species Distribution Models offer promise to increase the rate at which native germplasm can be developed.

However, while SDM's generate hypothesis of whether areas have environmental conditions similar to the observed environmental niche of the species they do not consider the probabilities of colonization, nor the populations census sizes. A possible tool to associate a probability of occurrence of a species in a suitable habitat patch, is connectivity analysis. Utilizing the predicted unsuitability of habitat, with extreme barriers to dispersal, to create cost-resistance surfaces between patches with known populations and predicted populations allows for simulating the probability of occurrence. While previous correlations between habitat suitability and population size have often been low, we posit that patch level parameters offer a more useful prediction of population size (Weber, Stevens, Diniz-Filho & Grelle (2017), Waldock, Stuart-Smith, Albouy, Cheung, Edgar, Mouillot, Tjiputra & Pellissier (2022)). Patch metrics will more appropriately reflect parameters of a potential population, e.g. geographic extent, which we posit is a more informative proxy of population size than environmental niche correlation.
Utilizing this approach will allow field crews to associate probabilities with previously unground-truthed patches.

Here we showcase the utility of SDM's to predict suitable habitat for common species in natural settings and use those data to test two specific hypotheses. Firstly that SDM's can identify more populations than exist in the occurrence based records they were derived from. And secondly patch metrics derived from SDM's correlate with observed population census size. In order to determine whether SDM's are useful in detecting new populations, over $_{10}$ field crews were given 50 putative populations to survey for the presence and absence of the modeled species. To determine if patch metrics can be used to predict population census sizes, these crews noted whether the populations was large enough to support

2 MATERIALS AND METHODS

2.1 Study system

The 353 modelled species were selected by land managers in arid & semi-arid areas Western North America and are species they already or immediately intend to use in restoration. The spatial domain 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 Natural History Museums, citizen science initiatives, and standardized ecological monitoring programmes using R (Chamberlain, Barve, Mcglinn, Oldoni, Desmet, Geffert & Ram (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 2.5% quantiles of several environmental variables (bio1, 4, 10, 12, & 19; 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 a massive 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 - (% BLM Land + 15 % 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, Hamm, Groen, Skidmore & Toxopeus (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.

Environmental Variables

Up to 44 variables were used in generating the species distribution models (table X, Karger, Conrad, Böhner, Kawohl, Kreft, Soria-Auza, Zimmermann, Linder & Kessler (2017), Hengl, Mendes de Jesus, Heuvelink, Ruiperez Gonzalez, Kilibarda, Blagotić, Shangguan, Wright, Geng, Bauer-Marschallinger & others (2017), Ivushkin, Bartholomeus, Bregt, Pulatov, Kempen & De Sousa (2019), Tuanmu & Jetz (2014), Yamazaki, Ikeshima, Neal, O'Loughlin, Sampson, Kanae & Bates (2017), Amatulli, McInerney, Sethi, Strobl & Domisch (2020), Sanderson, Jaiteh, Levy, Redford, Wannebo & Woolmer (2002)). 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. 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 generally do not notably decrease the performance of random forests, they increase the amount of time required to predict models onto gridded surfaces. 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.

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 with all default values except for optimized mtrys (Kuhn (2008), Liaw & Wiener (2002)). Models were then predicted onto gridded surfaces which exceeded the species range 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 practical 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	5
<= 2	6

Results

Discussion

Acknowledgments

Literature Cited

(2004) National hydrography dataset.

(2023) Watershed boundary dataset.

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.

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

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, et al. (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

Waldock C, Stuart-Smith RD, Albouy C, Cheung WW, Edgar GJ, Mouillot D, Tjiputra J, Pellissier L (2022) A quantitative review of abundance-based species distribution models. Ecography 2022

Weber MM, Stevens RD, Diniz-Filho JAF, Grelle CEV (2017) Is there a correlation between abundance and environmental suitability derived from ecological niche modelling? A meta-analysis. Ecography 40:817–828

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.