Predicting, habitat suitability, occupancy, and census sizes of a rare plant species using iterative adaptive niche based sampling

Reed Clark Benkendorf[1]*, Jeremie B. Fant[1,2], Sophie Taddeo[3]

[1] Chicago Botanic Garden, Glencoe, Illinois 60022, USA

[2] Plant Biology and Conservation, Northwestern University, Evanston, Illinois 60208, USA

[3] Department of Environmental and Ocean Sciences, University of San Diego, San Diego, California 921

Abstract

1)

2)

3)

4)

?? use habitation rather than occupancy to avoid confusion with wildlife people??

1 | INTRODUCTION

The effects of anthropogenic stressors, e.g. land use and climate change, have lead to a global extinction crisis with estimates of the number of plant species facing extinction ranging from 20-40% ((**brummitt2015green?**), (**pimm2015many?**), (**nic2020extinction?**)). Determining which plant species to focus our conservation efforts (e.g. active restoration, preserve creation, *ex situ* collections) on requires an array of data which seldom exist for decision makers ((**heywood2017plant?**)). These data generally outline simple biological and ecological parameters of species useful for detailing there rarity and how it's distribution relates to current and future anthropogenic stressors. Chief amongst these parameters, are the geographic extent of occurrence (range), the distribution of suitable habitat - the occupancy of this habitat as well as the spacing of occupied patches, and the census size of individual (sub-)populations ((**natural2001iucn?**), (**faber2012natureserve?**), (**usfws2016ssa?**)). While these parameters are seemingly simple, characterizing

*Author for Correspondence: rbenkendorf@chicagobotanic.org

them can be time consuming, generally requiring extensive field work - hence they oftentimes require proxies or heuristics for estimation in conservation assessments ((**juffe2016assessing?**), (**bland2015cost?**), (**pelletier2018predicting?**)). Environmental niche models have made enormous headway in resolving the former two problems (extent, and occupancy), however the historic mismatch between the resolution of variables governing species distributions and the data available to serve as predictors of environmental have restricted the interpretation and implementation of these models in highly heterogeneous environments (taddeo2024grimes, (**guisan2013predicting?**)). Recent advances in remote-sensing technologies have allowed for the generation of useful models in these generally biodiverse systems, further these data offer promise to model additional population parameters (e.g. census size, population extents), however the utility and usage of high resolution have rarely been ground verified or ostensibly reported upon (Chiffard *et al.* (2020)).

Recently, considerable headway has been made in generating statistically robust environmental niche models (ENM's) spirited by: recent advances in collecting high-resolution environmental data, compute power, digitization of natural history museum records and the acquisition of citizen science records, and statistical methods especially downsampling ((**markham2023review?**), (**feldman2021trends?**)).

However, ENM's are rarely ground verified, and even more seldom at landscape scales. Hence, most of our knowledge about producing ENM's rely on simulated species and data, especially at spatial resolutions considered relatively coarse to those desired by many analysts. An historic complication with the implementation and interpretation of ENM's is a mismatch between the spatial resolution of the independent variables available to model the species fundamental niche and the factors governing the true distribution of populations - the realized niche ((**carscadden2020niche?**), (**chauvier2022resolution?**), (**lembrechts2019incorporating?**)). Recent papers have had mixed results regarding the effects of imprecisely mapped occurrence data on ENM predictions, with indications that models generated in more heterogenous environments, and at finer resolutions (e.g. ca. 3 arc-second relate to 10 arc-minutes (~14.5 km) at 38*)) suffer minor decreases in model performance ((**graham2008influence?**), (**smith2023including?**)) with real species, while increasing error in mapping has drastic effects on model predictions with virtually simulated species ((**gabor2022positional?**)).

A further mismatch of resolution is the year in which data on geographic localities were obtained and current conditions which allow for positive population growth ((**bracken2022maximizing?**)). Indeed, historic occurrence data may now represent conditions which are inhospitable to the maintenance of populations, or may even represent populations which even then were simply sinks from more robust populations ((**bracken2022maximizing?**)). Collecting data on whether areas are favorable to continued recruitment of individuals from the soil seed bank, are perhaps more astute than whether long-lived individuals persist.

Improving Model Performance - Iterative adaptive-nice based sampling after expert mapping and sampling. ENM models generally suffer from having few, generally spatially biased, occurrence records to serve as dependent variables, which generally fail to characterize the ecological breadth of the species (CITE). While many ENM's have high-performance metrics while tested on small subsets of hold-out data, they are unlikely to detect many new populations during ground verification ((**a2022species?**)). To increase the number of presences, and absences in sites which are relatively similar to those harboring presences, which can be used for training models iterative adaptive-niche based sampling (ANBS) has become increasingly employed ((**guisan2006using?**)). In ANBS cells with high probabilities of occurrence are preferentially visited, and after each bout of field visits, a new model is fit incorporating the original data, plus the recently collected data.

However, we posit that evaluating the effects of ANBS is complicated as the species are oftentimes initially 'under surveyed'; in these instances the true distribution of the population is generally poorly defined - despite it being easy to do so by a naturalist. Here we showcase the usage of a robust first-stage expert subjective sampling effort to generate a large pool of random variables for elaboration after a second modelling bout. Using this process not only allows for the acquisition of a larger number of presences and absences, but also allows for verifying coordinate placement, that historic points are still extent, and for the acquisition of additional data such as census estimates and life stages ((**stockwell2002effects?**), (**wisz2008effects?**)). For the purposes of this paper, we believe that it elevates the challenge of future sampling bouts to be more representative of typical rare species, i.e. that they have on occasion been thoroughly surveyed for.

Prioritizing Survey Efforts - Occupied Sites An ENM predicts a single outcome; the probability of suitable habitat for the species. Although generally an analysts true feature of interest is the species realized distribution.

However, the bridge between an ENM and a plant populations presence is related to the dispersal of propagules and the establishment of the population, rather than the fundamental niche alone.

To assist practitioners in detecting new populations, or extending the range of currently known populations, we propose modelling plant occupancy as a random variable dependent on distance from sites known to be occupied, and landscape metrics of the target site. 'Distance' may be defined as euclidean (or Haversine for large distances), or as a least-cost distance reflecting a generalized surface which conveys the difficulty for typical seeds to travel between the nearest occupied sites and the site of interest.

Landscape metrics postulated to relate to the occupancy at a site include the patch metrics Core Area Index, roughly reflecting the probability of a propagule arriving to a larger patch, and the class metrics proximity and contagion both of which reflect the aggregation between occupied sites.

Estimating deme borders and census sizes Obtaining reliable estimates of plant population census sizes can be a time consuming task, requiring two major pieces of information 1) measurements of density, and 2) population extent. Generally these values are obtained via the use of many long (50m) transects, which may require multiple reads and data recorders.

In addition to these personnel requirements, many plant species are endemic to steep, and often loose, slopes, prohibiting the use of transects. Given these complications, many now prefer to obtain these data via genomic methods (e.g. linkage disequilibrium), a similarly expensive task. Here we estimate plant density gleaned during preliminary surveys by a naturalist using simple statistical techniques, and use these to feed into ENM's ((**oliver2012population?**)). Precisely mapping the boundaries of a population is another time consuming task, but essential for the effective estimates of population census size. Currently, population boundaries are generally delineated by have practitioners walk distances (e.g. 1km) in several directions searching for more plants. However, as the distance from a central location increases the amount of area to survey increases IN A FASHION… This is essentially problematic for rare species with many small clusters of individuals, which require considerable survey effort to chance upon.

2 | METHODS

Study Species & System

Eriogonum coloradense Small (Polygonaceae) is a synoecious mat forming perennial herb endemic to the Central Southern Rocky Mountains in Colorado, U.S.A. It's known geographic range covers XX km^2, has 26 formally described populations, and is thought to occur across a range of elevation, slopes, aspects, soil types and habitats. The elevation range from which it is xx - xx, and the broad habitat types it's known from include: high elevation sagebrush steppe, sub-alpine grasslands, and alpine slopes. It is treated as an S3/G3 species by NatureServe, and a Tier 2 species by State Wildlife Action Plan by the Colorado Parks and Wildlife Service.

Data Acqusition

Dependent data were gathered from iNaturalist (XX records) and the Consortium of Southern Rockies herbaria (XX records) ((**soro2024?**)). These data were manually reviewed and 16 herbarium records which had low geolocation quality, or which were place in localities which did not match their herbarium labels were removed.

Digital Elevation Models at 3arc (type), and 1arc (type), and Digital Elevation Products at 1/3 arc, and 1m

(), resolution were acquired from the United States Geological Survey and clipped to the domain of analysis (a rectangle buffered 10 mi. beyond any known population). The 3m data set was created by bilinearly re-sampling the 1m DEP. These elevation products were used to create all geomorphology data sets using whiteboxTools ((**wu2022whitebox?**)). Vegetation cover data were made by combining the raster data into continuous covers: Forested, Shrub, and Herbaceous vegetation (**tuanmu2014global?**).

ClimateNA was used to create a data set at 3 arc-second resolution which then underwent simple bilinear interpolation to generate products at the finer resolutions ((**wang2016locally?**)). Gray co-occurrence level matrices were produced using the glcm r package using 2023 NAIP aerial imagery, which underwent bilinear resampling to resolution, using default settings, but with windows of 5 in both directions ((**zvoleff2020glcm?**)).

Ground Verification

The first round of ground verification was carried out from June-September 2024. All pre-existing occurrence points were considered candidates for revisits and all X trails leading to them were marked as SAMPLE UNITS. Each trail was manually mapped, and buffered 45m in each direction and XXX random points were drawn, thinned to distances > XXXm, leaving XXX random plots for assessment. XX trails were visited, allowing for the assessment of XX random points and XX occurrences. When conducting field work, all presences of *E. coloradense* were opportunistically noted, and to better describe the spread of the population points were subjectively placed ca. 30-50m from the previous one until passing out of the population (n =). Additionally subjectively placed absences were also collected in areas which seemed favorable, or were in close proximity, to *E. coloradense* individuals; this occurred both in field (n =) and through use of aerial imagery on a computer afterwards (n =).

Adaptive Niche-Based Sampling was carried out in July of 2025. To determine whether adaptive sampling performed better than alternative sampling regimes (e.g. stratified, random), and if it could be improved by occupancy modelling 30 points were selected for each of the aforementioned sampling schema plus 30 occupancy points.

Comparision of Different Spatial Resolutions

Environmental Niche Models were generated at five different spatial resolutions, 3 (~72m), 1 (~24m), and 1/3 (~8m) arc-seconds, and 3 and 1m resolution. The details of modelling were similar for each resolution.

Records were thinned to the distance of an hypotenuse of a cell to avoid replicates ((**aeillo2015spthin?**)). XXX Absence records were generated using the background function with method environmental distance from sdm ((**naimi2016sdm?**)), these records were then manually reviewed and six records which were deemed in areas which may be possible presences were removed, after this the records were randomly sampled to reduce the data set size to XX records.

After the first iteration of modelling all additional presences and absences were thinned via a similar manner and combined with the original absence records. 'Presences' which had greater coordinate uncertainty than the resolution of modelling were removed.

compare the results of each spatial resolution at each of the 3 stages (naive, expert, adaptive)

Adaptive and Occupancy Based Field Sampling

500 stratified points ranging from 1-100% probability of suitable habitat were generated using sample (terra). Occupancy scores for each of these points were then calculated using (whiteboxtools). 200 points which maximized the spread of values along both dimensions (habitat, occupancy) were then chosen for ground verification.

Plant Density

Plant density was modelled using five methods, a generalized linear model (GLM) with a poisson error distribution and spatial autocorrelation structures, spAbundance, Random Forest and XgBoost regression.

Species Occupancy

Comparision of Juvenile and Mature Plant Models

Using the top performing model resolution () models were refit using only either juvenile or mature plant presences, the number of mature presences were limited to the number of juvenile occurrences. Models were fit at 6 sample sizes (n = 15, 30, 50, 75, 125, 200) 15 times each using a randomly sampled 60%-40% train-test split of data.

Simulations of Sample Size

The effect of sample size on model performance was simulated at 8 sample sizes (n = 15, 30, 50, 75, 125, 200, 300, 400) each 25 times, using a randomly sampled 60%-40% train-test split of data.

Simulations of Coordinate Errors

The effect of sample size on model performance was simulated at 6 sample sizes (n = 15, 30, 50, 75, 125, 200), with three proportions of records in error (0%, 5%, 10%, 20%), and 3 levels of coordinate uncertainty (0m, 10m, 100m, 1000m), at each of the 5 resolutions 25 times.

3 | RESULTS

Comparision of Different Spatial Resolutions

Ground Verification

Plant Density

Species Occupancy

Comparision of Juvenile and Mature Plant Models

Simulations of Sample Size

Simulations of Coordinate Errors

4 | DISCUSSION

5 | CONCLUSIONS

6 | ACKNOWLEDGMENTS

REFERENCES

Chiffard, J., Marciau, C., Yoccoz, N.G., Mouillot, F., Duchateau, S., Nadeau, I., Fontanilles, P. & Besnard, A. (2020). Adaptive niche-based sampling to improve ability to find rare and elusive species: Simulations and field tests. *Methods in Ecology and Evolution*, **11**, 899–909.