





# 1 elapid: Species distribution modeling tools for Python

2 **Christopher B. Anderson**  <sup>1,2</sup>

3 <sup>1</sup> Salo Sciences, San Francisco, CA, USA <sup>2</sup> Center for Conservation Biology, Stanford University,  
4 Stanford, CA, USA

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

## Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

## License

Authors of papers retain copyright  
and release the work under a  
Creative Commons Attribution 4.0  
International License ([CC BY 4.0](#)).

## 5 Summary

6 Species distribution modeling (SDM) is based on the Grinnellian niche concept: the environ-  
7 mental conditions that allow individuals of a species to survive and reproduce will constrain the  
8 distributions of those species over space and time ([Grinnell, 1917](#); [Wiens et al., 2009](#)). The  
9 inputs to these models are typically spatially-explicit species occurrence records and a series of  
10 environmental covariates, which might include information on climate, topography, land cover  
11 or hydrology ([Booth et al., 2014](#)). While many modeling methods have been developed to  
12 quantify and map these species-environment interactions, few software systems include both  
13 a) the appropriate statistical modeling routines and b) support for handling the full suite of  
14 geospatial analysis required to prepare data to fit, apply, and summarize these models.

15 elapid is both a geospatial analysis and a species distribution modeling package. It provides an  
16 interface between vector and raster data for selecting random point samples, annotating point  
17 locations with coincident raster data, and summarizing raster values inside a polygon with  
18 zonal statistics. It provides a series of covariate transformation routines for increasing feature  
19 dimensionality, quantifying interaction terms and normalizing unit scales. It provides a Python  
20 implementation of the popular Maxent SDM ([Phillips et al., 2017](#)) using infinitely weighted  
21 logistic regression ([Fithian & Hastie, 2013](#)). It also includes a standard Niche Envelope Model  
22 ([Nix, 1986](#)), both of which were written to match the software design patterns of modern  
23 machine learning packages like sklearn ([Grisel et al., 2022](#)). It also allows users to add spatial  
24 context to any model by providing methods for spatially splitting train/test data and computing  
25 geographically-explicit sample weights. elapid was designed as a contemporary SDM package,  
26 built on best practices from the past and aspiring to support the next generation of biodiversity  
27 modeling workflows.

## 28 Statement of need

29 Species occurrence data—georeferenced point locations where a species has been observed  
30 and identified—are an important resource for understanding the environmental conditions  
31 that predict habitat suitability for that species. These data are now abundant thanks to  
32 the proliferation of institutional open data policies, large-scale collaborations among research  
33 groups, and advances in the quality and popularity of citizen science applications ([GBIF, 2022](#);  
34 [iNaturalist, 2022](#)). Tools for working with these data haven't necessarily kept pace, however,  
35 especially ones that support modern geospatial data formats and machine learning workflows.

36 elapid builds on a suite of well-known statistical modeling tools commonly used by biogeogra-  
37 phers, extending them to add novel features, to work with cloud-hosted data, and to save and  
38 share models. It provides methods for managing the full lifecycle of modeling data: generating  
39 background point data, extracting raster values for each point (i.e. point annotation), splitting  
40 train/test data, fitting models, and applying predictions to rasters. It provides a very high  
41 degree of control for model design, which is important for several reasons.

42 First is to provide simple and flexible methods for working with spatial data. Point data are  
43 managed as GeoSeries and GeoDataFrame objects (Jordahl et al., 2022), which can be easily  
44 merged and split using traditional indexing method as well as with geographic methods. They  
45 can also be reprojected on-the-fly. elapid reads and writes raster data with rasterio, which  
46 provides a similarly convenient set of methods for indexing and reading point locations from  
47 rasters (Gillies, 2013). These features are wrapped to handle many of the routine tasks and  
48 gotchas of working with geospatial data. It doesn't require data to be rigorously pre-processed  
49 so that all rasters are perfectly aligned, nor does it require that all datasets are in matching  
50 projections. elapid can extract pixel-level raster data from datasets at different resolutions,  
51 from multi-band files, and harmonize projections on-the-fly, for both model fitting and for  
52 inference.

53 Another advantage of elapid's flexible design is that it can be used to extend traditional species  
54 distribution models in ways that are difficult to implement in other software systems. Working  
55 with multi-temporal data, for example—fitting SDMs to occurrence records and environmental  
56 data from multiple time periods—is also supported. Each time period's occurrence data can  
57 be annotated using the coincident environmental data. Random background samples can  
58 likewise be generated for each time period, which ensures the background represents a broad  
59 distribution of conditions across the full temporal extent. These presence and background  
60 samples then concatenated into a single GeoDataFrame for model fitting. Fitted models can  
61 be applied to multi-temporal environmental data to map changes in habitat suitability over  
62 time, and can also be saved and restored later for future inference.

### 63 Why Maxent still matters

64 The main scientific contribution of elapid is extending and modifying the Maxent SDM, a  
65 model and software system as popular as it is maligned (Fourcade et al., 2018; Phillips &  
66 Dudík, 2008). First published in 2006, Maxent remains relevant because it's a presence-only  
67 model designed to work with the kinds of species occurrence data that have proliferated  
68 lately.

69 Presence-only models formulate binary classification models as presence/background (1/0)  
70 instead of presence/absence, which changes how models are fit and interpreted (Fithian &  
71 Hastie, 2013; Merow et al., 2013). Background points are a spatially-random sample of the  
72 landscapes where a species might be found, which should be sampled with the same level  
73 of effort and bias as the species occurrence data. Presence/background models posit the  
74 null expectation is that a species is equally likely to be found anywhere within its range.  
75 Differences in environmental conditions between where a species occurs and in the conditions  
76 across the full landscape should indicate niche preferences. Relative habitat suitability is  
77 then determined based on differences in the relative frequency distributions of conditions in  
78 these regions. Presence-only models reduce the burden of finding absence data, which are  
79 problematic to boot, but they increase the burden of precisely selecting background points.  
80 These define what relative habitat suitability is defined as relative to (Barbet-Massin et al.,  
81 2012; Elith et al., 2011).

82 elapid includes several methods for sampling the background. Points can be sampled uniformly  
83 within a polygon, like a range map or an ecoregion extent. Sampling points from rasters can be  
84 done uniformly across the full extent or only from pixels with valid, unmasked data. Working  
85 with bias rasters is also supported. Any raster with monotonically increasing values can be used  
86 as a sample probability map, increasing the probability that a sample is drawn in locations  
87 with higher pixel values. One important role for the niche envelope model is to create bias  
88 maps to ensure background points are only sampled within the broad climatic envelope where  
89 a species occurs. The target-group bias sampling method has also been shown to effectively  
90 correct for sample bias (Barber et al., 2022).

91 A common criticism of Maxent is that, though it depends on spatially-explicit data, it's not

92 really a spatial model. Covariate data are indexed and extracted spatially, but there are no  
93 model terms based on location, distance, or point density, and all samples are treated as  
94 independent measurements. While I generally maintain the perspective that many of the ails of  
95 spatial autocorrelation are typically overstated (Hawkins, 2012), spatial data have unique and  
96 very interesting properties that should be handled carefully. Non-independence is inherent to  
97 spatial data, driven both by underlying ecological patterns and processes (e.g. dispersal, species  
98 interactions, climatic covariance) as well as by data collection biases (e.g. occurrence records  
99 are common near roads or trails despite many species typically preferring less fragmented  
100 habitats).

101 Spatial models should include methods for handling spatially-specific modeling paradigms,  
102 particularly the lack of independence of nearby samples or spatial biases in sample density.  
103 Quantifying and understanding model skill requires accounting for these spatial autocorre-  
104 lations, and `elapid` includes several methods for doing so. Checkerboard cross-validation  
105 can mitigate bias introduced by spatially clustered points. Creating spatially-explicit  $k$ -fold  
106 splits—-independent clusters based on  $x/y$  locations—can quantify how well model predictions  
107 generalize to new areas. And tuning sample weights based on the density of nearby points  
108 decreases the risk of overfitting to autocorellated environmental features from areas with high  
109 sample density. This is particularly important for mitigating the effects of density-dependent  
110 non-independence.

111 These methods are not solely restricted to the SDMs implemented in `elapid`. They can add  
112 spatial context to other machine learning models, too. Geographic sample weights can be  
113 used to fit random forests, boosted regression trees, generalized linear models, and other  
114 approaches commonly used to predict spatial distributions. `elapid` also includes a series of  
115 feature transformers, including the transformations used in Maxent, which can extend covariate  
116 feature space to improve model skill.

117 `elapid` was designed to provide a series of modern tools for quantifying biodiversity change.  
118 The target audience for the package includes ecologists, biodiversity scientists, spatial analysts  
119 and machine learning scientists. Working with software to understand the rapid changes  
120 reshaping our biosphere should be easy and enjoyable. Because thinking about the ongoing  
121 annihilation of nature that's driving our current extinction crisis is decidedly less so.

## 122 Acknowledgments

123 Many thanks to Jeffrey R. Smith for many long and thought-provoking discussions on species  
124 distribution modeling. Thanks also to David C. Marvin for helping me think creatively about  
125 novel applications for Maxent. And many thanks to Gretchen C. Daily for promoting and  
126 supporting access to open source software for biodiversity and ecosystem services modeling.

## 127 References

- 128 Barber, R. A., Ball, S. G., Morris, R. K., & Gilbert, F. (2022). Target-group backgrounds  
129 prove effective at correcting sampling bias in maxent models. *Diversity and Distributions*,  
130 28(1), 128–141. <https://doi.org/10.1111/ddi.13442>
- 131 Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences  
132 for species distribution models: How, where and how many? *Methods in Ecology and*  
133 *Evolution*, 3(2), 327–338. <https://doi.org/10.1111/j.2041-210x.2011.00172.x>
- 134 Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. (2014). BIOCLIM: The first  
135 species distribution modelling package, its early applications and relevance to most current  
136 MAXENT studies. *Diversity and Distributions*, 20(1), 1–9. <https://doi.org/10.1111/ddi.12144>
- 137

- 138 Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A  
139 statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17(1), 43–57.  
140 <https://doi.org/10.1111/j.1472-4642.2010.00725.x>
- 141 Fithian, W., & Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-  
142 only data. *The Annals of Applied Statistics*, 7(4), 1917. [https://doi.org/10.1214/  
143 13-aos667](https://doi.org/10.1214/13-aos667)
- 144 Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of  
145 species, or the challenge of selecting environmental predictors and evaluation statistics.  
146 *Global Ecology and Biogeography*, 27(2), 245–256. <https://doi.org/10.1111/geb.12684>
- 147 GBIF. (2022). *GBIF: The global biodiversity information facility*. Global Biodiversity Informa-  
148 tion Facility. <https://www.gbif.org/what-is-gbif>
- 149 Gillies, S. (2013). *Rasterio: Geospatial raster i/o for Python programmers*. Mapbox. [https:  
150 //github.com/rasterio/rasterio](https://github.com/rasterio/rasterio)
- 151 Grinnell, J. (1917). The niche-relationships of the california thrasher. *The Auk*, 34(4), 427–433.  
152 <https://doi.org/10.2307/4072271>
- 153 Grisel, O., Mueller, A., Lars, Gramfort, A., Louppe, G., Prettenhofer, P., Blondel, M.,  
154 Niculae, V., Nothman, J., Fan, T. J., Joly, A., Lemaitre, G., Vanderplas, J., kumar,  
155 manoj, Estève, L., Qin, H., Hug, N., Varoquaux, N., Layton, R., ... Eren, K. (2022).  
156 *Scikit-learn/scikit-learn: Scikit-learn 1.1.2* (Version 1.1.2) [Computer software]. Zenodo.  
157 <https://doi.org/10.5281/zenodo.6968622>
- 158 Hawkins, B. A. (2012). Eight (and a half) deadly sins of spatial analysis. *Journal of*  
159 *Biogeography*, 39(1), 1–9. <https://doi.org/10.1111/j.1365-2699.2011.02637.x>
- 160 iNaturalist. (2022). *iNaturalist*. California Academy of Sciences. <https://www.inaturalist.org>
- 161 Jordahl, K., Bossche, J. V. den, Fleischmann, M., McBride, J., Wasserman, J., Richards, M.,  
162 Badaracco, A. G., Gerard, J., Snow, A. D., Tratner, J., Perry, M., Farmer, C., Hjelle, G.  
163 A., Ward, B., Cochran, M., Taves, M., Gillies, S., Culbertson, L., Bartos, M., ... Wasser, L.  
164 (2022). *Geopandas/geopandas: v0.11.1* (Version v0.11.1) [Computer software]. Zenodo.  
165 <https://doi.org/10.5281/zenodo.6894736>
- 166 Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling  
167 species' distributions: What it does, and why inputs and settings matter. *Ecography*,  
168 36(10), 1058–1069. <https://doi.org/10.1111/j.1600-0587.2013.07872.x>
- 169 Nix, H. A. (1986). A biogeographic analysis of australian elapid snakes. *Atlas of Elapid Snakes*  
170 *of Australia*, 7, 4–15.
- 171 Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening  
172 the black box: An open-source release of maxent. *Ecography*, 40(7), 887–893. [https:  
173 //doi.org/10.1111/ecog.03049](https://doi.org/10.1111/ecog.03049)
- 174 Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with maxent: New  
175 extensions and a comprehensive evaluation. *Ecography*, 31(2), 161–175. [https://doi.org/  
176 10.1111/j.0906-7590.2008.5203.x](https://doi.org/10.1111/j.0906-7590.2008.5203.x)
- 177 Wiens, J. A., Stralberg, D., Jongsomjit, D., Howell, C. A., & Snyder, M. A. (2009). Niches,  
178 models, and climate change: Assessing the assumptions and uncertainties. *Proceedings of*  
179 *the National Academy of Sciences*, 106(supplement\_2), 19729–19736. [https://doi.org/10.  
180 1073/pnas.0901639106](https://doi.org/10.1073/pnas.0901639106)