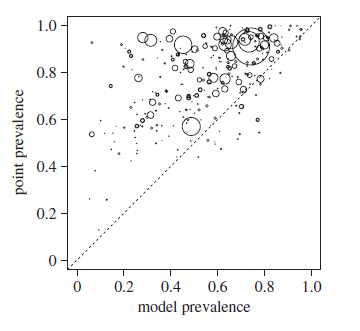
**Assessing habitat maps using presence-only point occurrence locations and their uncertainty**

Rondinini et al. (2011) created deductive habitat suitability models (high, medium, unsuitable) for >5000 terrestrial mammals globally at a spatial grain of 300 meters to examine patterns of richness and protection. They conducted model evaluation on a subset of only 263 species because detailed occurrence data are lacking for most mammal taxa. The occurrence data used contains information on species presence only, limiting assessment methodologies. Therefore, Rondinini et al. chose to compare what they term ‘model prevalence’ and ‘point prevalence’ to species’ landscapes of randomly distributed habitat at a spatial grain of 1 km2 (the assumed accuracy of their point locality data). Point prevalence is defined as the proportion of 1 km2 cells containing one or more occurrences and one or more 300 m cells with high and/or medium suitability (i.e. proportion of correctly predicted occurrences). Model prevalence is defined as the proportion of 1 km2 cells with high or medium suitability within the species’ range. The results of their evaluation produced the following chart:

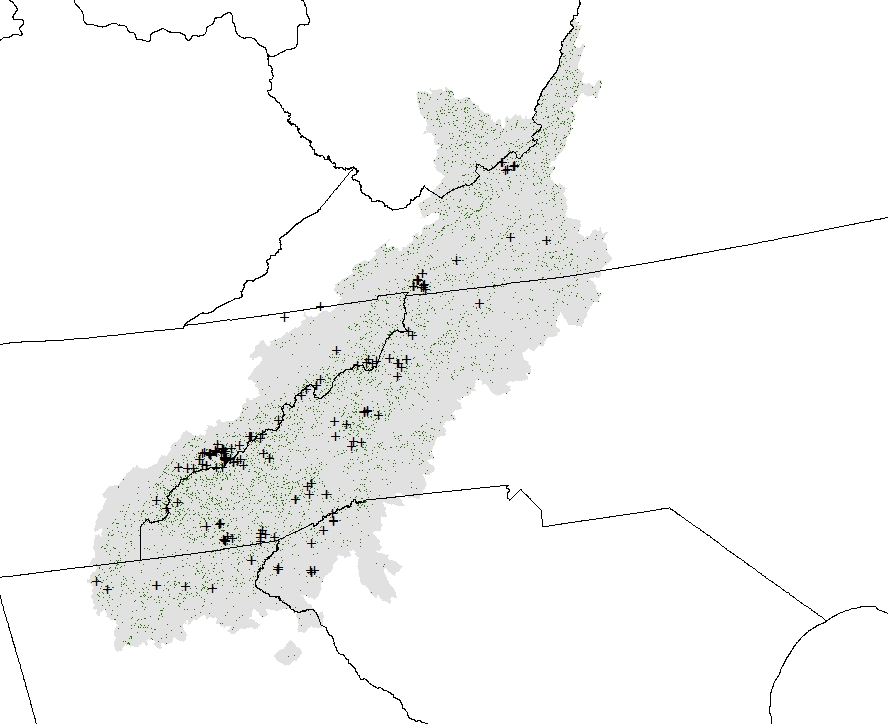


As Rondinin et al. describe: *“If predicted suitable cells were distributed randomly with respect to species occurrences, the expectation would be that the point prevalence equals the model prevalence.”* Hence, the comparison assumes a random habitat landscape is one in which any given location contains habitat at proportions equal to model predicted habitat across the entire species’ range. When point prevalence is greater than model prevalence, the model predicts habitat at species locations better than the range *“because it tends to correctly flag as unsuitable the cells that do not contain occurrences.”*

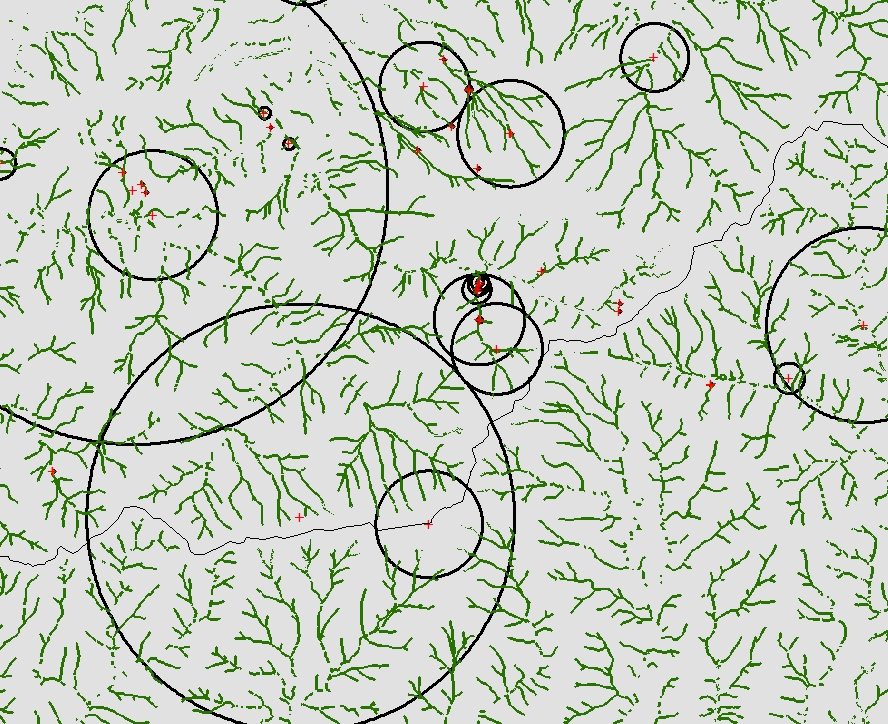
**Buffer Proportion Assessment (BPA)**

GAP species maps include geographic range along with distribution of potential habitat generated at a spatial raster resolution of 30 m. Therefore, each species has a measurement of model prevalence, which is simply the areal proportion of predicted habitat within the delineated range. However, in order to calculate point prevalence at a 30 m spatial grain, point occurrences would need to have an equivalent spatial accuracy. Point prevalence calculations at a 30 m spatial grain will be available only for a very limited number of species and with very few data points because most point occurrence datasets (e.g. GBIF, BiSON, eBird) have either no information regarding individual location accuracy or vary widely in coordinate uncertainty (e.g. <100 m to province-level centroids). However, buffering point occurrences by the degree of uncertainty of presence locations and calculating the proportion of habitat within those buffers would allow for comparisons to random habitat landscapes using range alone. As with point prevalence vs. model prevalence, models that predict habitat better than random would have a mean proportion of habitat in point buffers greater than the proportion of habitat throughout the species’ range.

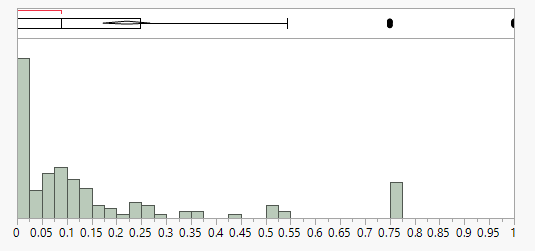
The following is an illustration of this assessment method for the Black-bellied salamander (*Desmognathus quadramaculatus*). The range for this species extends throughout the southern Appalachian Mountains from West Virginia to northern Georgia. GBIF records were obtained that have latitude and longitude coordinates with a measure of accuracy (uncertainty). The GAP mapped range and habitat along with the 174 GBIF records are pictured below where the gray polygon is the GAP range, green pixels are predicted habitat, and the black cross symbols are GBIF records. Model prevalence (proportion of GAP predicted habitat in range) for this species is 0.0401.



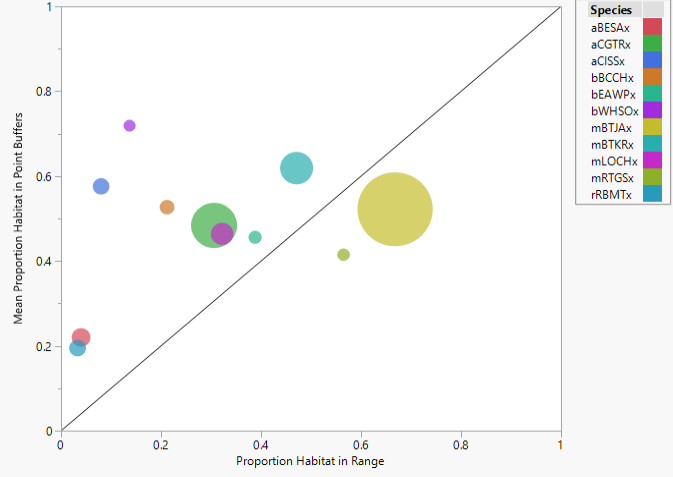
Each occurrence location has a coordinate uncertainty measured in meters, which was used to generate circular buffers around each point. Occurrence locations with coordinate uncertainty less than 30 meters were buffered at 30 meters and only points with a coordinate uncertainty of 10000 meters or less were used in the assessment.



The proportion of predicted habitat amount was calculated for each point buffered area. The histogram of point buffer habitat proportions:



The mean proportion of habitat in point buffers was calculated and plotted against the proportion of habitat in the species’ range ( = 0.2204; approximately 5 times greater than proportion of habitat in range). Similar calculations were made for locations and their associated buffer distances for 10 additional species across a continuum of model prevalences (proportion of habitat in range) illustrating most species have models considered ‘better than random’.

~~~~

Note: the size of the circles is proportional to the number of point occurrence records used to make the habitat proportion comparison for each species.

Species codes in the above graph are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Species Code** | **Scientific Name** | **Common Name** | **Prop. Hab. in Range** | **Mean Prop. Hab. in Point Buffers** |
| rRBMTx | Sternotherus carinatus | Razor-backed Musk Turtle | 0.033 | 0.195 |
| aBESAx | Desmognathus quadramaculatus | Black-bellied Salamander | 0.040 | 0.220 |
| aCISSx | Batrachoseps pacificus | Channel Islands Slender Salamander | 0.080 | 0.576 |
| bWHSOx | Megascops trichopsis | Whiskered Screech-owl | 0.137 | 0.719 |
| bBCCHx | Poecile atricapillus | Black-capped Chickadee | 0.212 | 0.527 |
| aCGTRx | Hyla chrysoscelis | Cope's Gray Treefrog | 0.306 | 0.484 |
| mLOCHx | Tamias speciosus | Lodgepole Chipmunk | 0.322 | 0.464 |
| bEAWPx | Contopus virens | Eastern Wood-pewee | 0.388 | 0.456 |
| mBTKRx | Dipodomys spectabilis | Banner-tailed Kangaroo Rat | 0.471 | 0.619 |
| mRTGSx | Spermophilus tereticaudus | Round-tailed Ground Squirrel | 0.565 | 0.415 |
| mBTJAx | Lepus californicus | Black-tailed Jackrabbit | 0.668 | 0.522 |

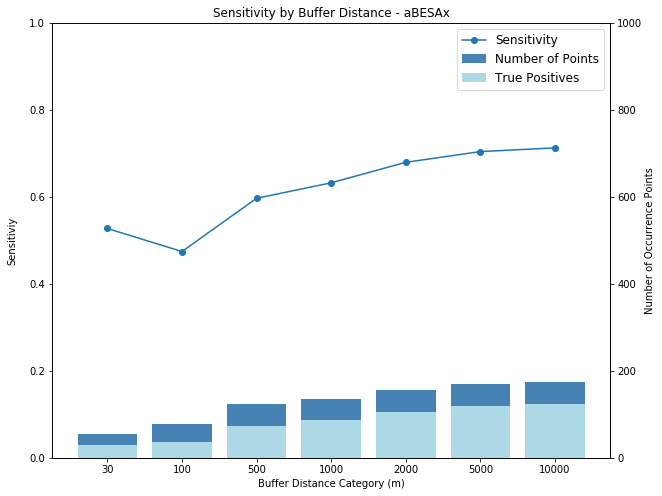
**Buffer Sensitivity Metric (BSM)**

Since occurrence points have varying amounts of uncertainty associated with their geographic location, comparing them with the habitat map will result in a varying degree of assessment. Illustrating this variation across a number of scales can help model users determine which scale is most suitable for their intended use. The accuracy assessment measure ***‘sensitivity’*** is also referred to as the true positive fraction in model validation. For habitat map assessment, it is a measure of locations correctly predicted to be habitat. Sensitivity is calculated as the count of ***true positives*** divided by the sum of ***true positive*** plus ***false negative*** counts:

Using occurrence points in GAP habitat map evaluation, a ***true positive*** is recorded when at least one predicted habitat cell occurs within an occurrence point uncertainty buffer. A ***false negative*** is recorded when no habitat cells occur within an occurrence point buffer because occurrence points only record presence of a species and not absence. A sensitivity measure can be calculated for increasing buffer distances (i.e. scales) where a cumulative count of ***true positive*** (at least one habitat cell) and ***false negative*** (no habitat cells) are made for buffer sizes less than or equal to one of several size categories (for example 30, 100, 500, 1000, 2000, 5000, and 10000 meters). High sensitivity is indicative of a low omission rate or false positive fraction because sensitivity and omission rate sum to one. NOTE: The rule wherein an occurrence location is counted as a true positive for habitat if at least one cell is within the point buffer could be altered to increase stringency. Any arbitrary value such as at least 15% of the buffer area could be used for all species or the threshold could be species specific such as the percentage of habitat predicted within the range.

*The assumption in these measurement analyses is that occurrence points and their uncertainty buffer areas are located within the species’ habitat. This assumption may be violated, that is, a species may be recorded at a given location but the species may not perceive it as suitable habitat (e.g., it may be moving between habitat patches). Additionally, this metric becomes increasingly un-informative as the amount of predicted habitat within the range increases because there are no absences to identify areas that would not be considered suitable. Predicting every cell within the range as habitat will produce an omission rate of 0; that is, no false negatives will be recorded, sensitivity will be 1, and 1-sensitivity = omission rate which will equal 0 .*

Sensitivity was calculated for seven buffer distance categories for the Black-bellied salamander and plotted on a graph illustrating increasing sensitivity with increasing buffer size. Blue bars indicate the number of occurrence points used to calculate sensitivity for each buffer size category; light blue bars indicate the number of true positives:



NOTE: At 10000 meters, all points are used to calculate sensitivity because GBIF occurrences were limited to those with a coordinate uncertainty of 10000 meters or less.

Pearson, R.G. (2007) Species’ distribution modeling for conservation educators and practitioners. Available at: <http://biodiversityinformatics.amnh.org>.

Rondinini, C., Di Marco, M., Chiozza, F., Santulli, G., Baisero, D., Visconti, P., ... Boitani, L. (2011). Global habitat suitability models of terrestrial mammals. Philosophical Transaction of the Royal Society B: Biological Sciences, 366, 2633-2641.

Additional explanation of proportion habitat in buffer area vs. proportion habitat in range (BPA)

If the GAP model predicts that 4% of the black-bellied salamander range is suitable habitat, then a finite number of cells in the range could be considered habitat pixels. If these cells were distributed randomly across the range, the probability of any given cell being identified as habitat would be 0.04. Likewise, if we used point buffers to clip the random habitat landscape, the percentage of habitat cells in those buffers would be very close or equal to 4% (I ran multiple 100+ points of varying buffer distances on 100+ simulated landscapes just to test this and it is indeed the case). Assume that occurrence point locations are indicative of suitable habitat for the species. If a cell predicted by the model as habitat falls within the occurrence point buffer, because the number of habitat cells is finite, that means a cell outside the buffer was correctly not identified as habitat. The higher the proportion of cells identified as habitat within occurrence point buffers, the better the model was at identifying areas of the range not considered habitat. The model is identifying parts of the range where there are occurrences and by extension, habitat. One caveat is that not all of the range extent was sampled so we cannot really know if predicted non-habitat cells truly have no species occurrences and therefore are not habitat. But, those are caveats for all modeling methods and evaluations.

Pearson (2007):

*We cannot test if the model is correct (since we do not know the true potential distribution* [of habitat]*), but rather we test if the model is useful. Our criteria for usefulness are that the model successfully predicts presence* [of habitat] *in a high proportion of test localities (i.e., known occurrences) whilst not predicting that an excessively large proportion of the study area is suitable.* *Thus, a model that successfully predicts whether the species is present at all test localities whilst classifying most of the study area as suitable may be correct (the environment may truly be suitable throughout most of the study area); however, the model is not useful because it is not more informative than a random prediction.*

GAP models are predicting two things: the proportion of habitat in the range (aspatial) and the geographic distribution of that habitat throughout the range (spatial). I believe the buffer sensitivity metric (BSM) evaluates the aspatial aspect of the model at varying scales (buffer sizes), and the buffer proportion assessment (BPA) evaluates the spatial aspect of the model, i.e. the randomness of the habitat pixels on the landscape given the predicted proportion of habitat in range.

NOTES

Here is a proposed workflow to conduct these analyses using scripting. Using Python, the workflow would require the [**pygbif**](https://pygbif.readthedocs.io/en/latest/index.html) and [**python-dwca-reader**](https://python-dwca-reader.readthedocs.io/en/latest/index.html) packages as well as Pandas for dataframe manipulation. Utilizing the **pygbif** package would be exclusive to accessing the GBIF API and would require the user to have an account, along with username and password, to access the functionality of downloading GBIF occurrences programmatically.

* Create a script, function, or module that downloads GBIF occurrences. Note that using the occurrences.search functionality of **pygbif** is limited to 300 records/page and 200,000 records/request. Using the occurrences.download function with a username and password is not limited to 200,000 records/request.
* Create a script(s), function(s), or module that processes downloaded GBIF records to filter records accordingly. Flexibility of these programmatic methods would be necessary to account for various taxa and provide utility to other types of analyses.
* Create a script(s), function(s), or module that creates spatial versions of the occurrence data based on the latitude and longitude and coordinate uncertainty of the occurrences. That is, vector datasets of points and point buffers.
* Create a script, function, or module that performs spatial analyses between point buffers and species’ habitat rasters to calculate and compare for each species the mean proportion of habitat within point buffers and the proportion of habitat within the species’ range (model prevalence).