

# Building global range maps for terrestrial insects with species range kernels and modeled aggregations of georeferenced records

Brice B. Hanberry 

USDA Forest Service, Rocky Mountain Research Station, Rapid City, SD, 57702, USA

## ARTICLE INFO

### Keywords:

Biogeography  
Distributions  
IUCN  
Odonates  
Pollinators  
Macroecology

## ABSTRACT

Range maps for most terrestrial invertebrates generally are not accessible as spatial data. My objective was to build global range maps for 11,844 terrestrial insect species, specifically pollinator species with at least 400 georeferenced records each, by applying species range kernels, which contained about 99% of the records, and polygons of aggregated georeferenced records. Depending on aggregation distance, point aggregation generates a range of polygons, in terms of shape and area. To illustrate, for 225 odonate species that had range maps developed by experts, I selected one of six aggregation distances ranging from 200 km to 5000 km based on resemblance to range maps and fidelity to the georeferenced records. Then to predict aggregation distances for insects without accessible ranges, I modeled variables related to georeferenced records, kernels, and aggregated polygons for six different aggregation distances, based on supervised classification of aggregation distances for 4405 species with ranges. Lastly, I examined the species richness gradients of the species ranges, recognizing incompleteness of sampling. Predicted aggregation distances were predominantly middle distances of 500 km or 1000 km, for 86% of the insect species. Given a six class, complex problem with no true answer of species ranges, model accuracy for aggregation distance was 0.66. Generally, misclassification was only off by one distance, resulting in either better fit in shape or greater connection and coverage. For odonates, the species richness patterns were similar ( $r = 0.91$ ) between expert range maps and data range maps, which helped to correct sampling deficiencies in georeferenced records. Range maps for invertebrates, without previously spatially available range maps, expand underlying sampling gradients and offer at least provisional range maps, filling the gap of missing spatial data.

## 1. Introduction

Invertebrate species overall are unknown, proportionate to vertebrates and vascular plants (Scheffers et al., 2012). While nearly all mammal and bird species have been catalogued and assessed, encompassing documentation of species ranges and population statuses, only about 30% of arthropods have been identified (Scheffers et al., 2012; IUCN Standards and Petitions Committee, 2024). To be fair, species richness for vertebrates probably encompasses less than 81,000 species, with plants at less than 400,000 species, whereas estimates for species richness of tropical arthropods alone range from 3.6 to 11.4 million species (Hamilton et al., 2013). Insects may include 5 million species, with 1 million catalogued (Scheffers et al., 2012). In part due to the sheer number of invertebrate species, more information is available for vertebrates and plants than invertebrates (Ballesteros-Mejia et al., 2017; Vergara-Asenjo et al., 2023). Relatively inaccessible static range maps abound in guidebooks but generally are not available as accessible

spatial layers. The IUCN (2025) has provided range maps for taxa including vertebrates, vascular plants, and freshwater and marine invertebrates, including odonates (i.e., dragonflies and damselflies; Fig. 1).

Range maps are used for multiple purposes, including knowledge of the potential suite of species at any locality for identification, monitoring, management, and research from field studies to wider macroecological studies for conservation planning (Gaston and Fuller, 2009). Ranges are generally suitable areas for species occupancy, by definition, although errors may occur in range map locations (IUCN Standards and Petitions Committee, 2024). Certainly, species do not occur everywhere at all times within ranges, due to factors ranging from simply lack of space to fit all species in one location, and lack of resources, to avoidance of and/or removal by humans (Hurlbert and Jetz, 2007). Furthermore, sampling does not occur completely throughout most species ranges, resulting in unsampled areas within species ranges, similar to taxonomic shortfalls in sampling of invertebrates (Vergara-Asenjo et al., 2023).

E-mail address: [brice.hanberry@usda.gov](mailto:brice.hanberry@usda.gov).

<https://doi.org/10.1016/j.actao.2026.104192>

Received 15 March 2026; Received in revised form 31 July 2026; Accepted 5 August 2026

Available online 7 August 2026

1146-609X/Published by Elsevier Masson SAS. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

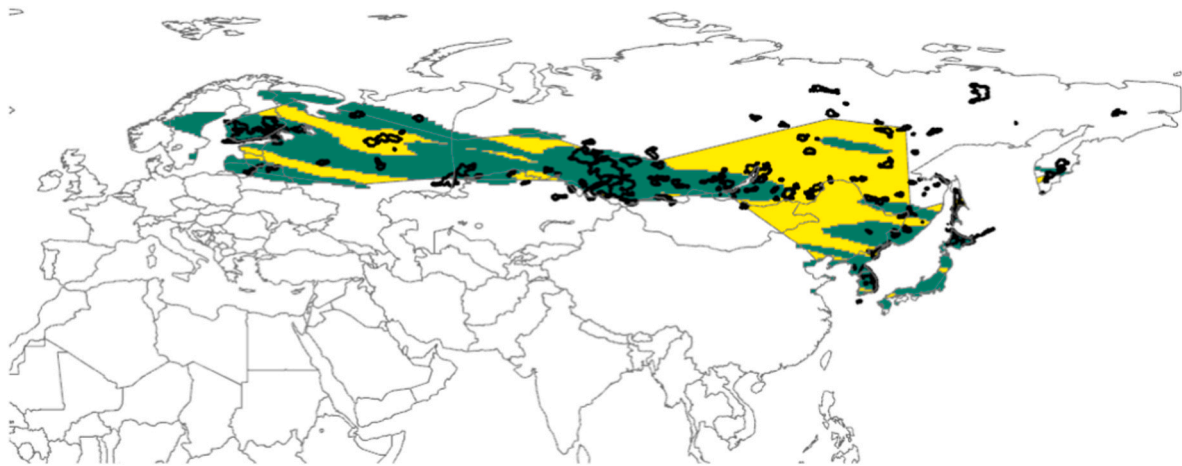


**Fig. 1.** The expert range (IUCN, 2025; dark black outline) for *Aeshna crenata*, an odonate species, contained 51% of georeferenced records, compared to the data range, comprised of a species range kernel (green, with 99% of georeferenced records) and georeferenced records aggregated to polygons at an aggregation distance of 2000 km (yellow). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Range maps typically are time-intensive to produce, due to required development by species experts. Experts develop and modify range outlines by applying expert knowledge, sometimes pooled through consensus in workshops, which is elusive to capture in a standardized workflow (Aronsson et al., 2024). Routinely, the starting point for development of species range maps is to encircle all known species records (Cantú-Salazar and Gaston, 2013; Aronsson et al., 2024). The initial range shape may be drawn manually or through application of a function to compute the shape of the records. The characteristic output is a minimum convex hull, which consistently is too large, with limited shape, relative to species ranges (Burgman and Fox, 2003). While alpha hulls can add curved lines, and georeferenced records can be removed to reduce polygon areas, all hulls need additional modification to replicate ranges developed by experts (Aronsson et al., 2024). Experts know information that is absent in georeferenced records and more customized than one polygon encircling georeferenced records from a database.

However, data-driven species ranges can approximate species ranges developed by experts. Georeferenced records of species occurrences are becoming available, due to contributions of community naturalists combined with online platforms for sharing observations and identification experience (Tosa et al., 2021). Georeferenced records are ideal building blocks for generating species range maps. Species ranges can be built similarly to home range kernels, with kernel density estimation (Worton, 1987) from georeferenced records, and the species range kernels can be connected to fill in missing spaces with georeferenced records aggregated to polygons (Fig. 1). Aggregation can occur at any distance; therefore, selection of one polygon of the different distance options is necessary.

For species without accessible range maps, the next phase in data range map development is to model classified samples of aggregation distances to select aggregation distances by model prediction. Here, I extended previous selection for aggregated polygons of six distances from 200 km to 5000 km, and evaluation of data ranges relative to expert ranges, for about 3960 vertebrate and vascular plants (Hanberry, in review) to assess and apply supervised classification modeling of aggregation distances for species ranges, with addition of about 450 species, including invertebrates (Supplementary Table 1; <https://zenodo.org/records/17195899>). To incorporate invertebrate range representation into the modeling and to illustrate the process, I determined the best fit for 225 odonates, relative to information provided in species kernels and IUCN range maps, after merging species range kernels with six different options for aggregated polygons. Next, I developed an array of predictor variables related to number of georeferenced records, species range kernels, and polygon aggregations for each aggregation

distance. Then, I applied four random forests algorithms to model the aggregation distance class and predict to 5692 butterfly and moth species (Lepidoptera), 582 bee species (Hymenoptera: Anthophila), 586 wasp species (Hymenoptera: Apocrita), 1613 fly species (Diptera), and 3371 beetle species (Coleoptera; Supplementary Table 2 at <https://zenodo.org/records/17195899>). These taxa are characterized as strongly to moderately efficient pollinators, although any insect species that visits flowers may be pollinators, albeit with weak transport of pollen (Hanberry, 2025a). I examined species richness for these species, including comparisons of odonate data ranges to IUCN expert ranges. I produced 11,844 range maps as spatial layers and static image files, contributing to assessment of invertebrate species.

## 2. Methods

### 2.1. Data

For invertebrate species of odonates, butterflies and moths, bees, wasps, beetles, and flies, georeferenced records occurred during 1990 to 2024, and into the first quarter of 2025, with unique terrestrial coordinates and location distance uncertainties < 36 km or unknown (Global Biodiversity Information Facility [GBIF] 2025; <https://doi.org/10.15468/dl.f4xa7r>; <https://doi.org/10.15468/dl.mnzzxe>; <https://www.gbif.org/occurrence/download/0009004-250325103851331>; <https://www.gbif.org/occurrence/download/0009319-250325103851331>; <https://www.gbif.org/occurrence/download/0009304-250325103851331>; <https://www.gbif.org/occurrence/download/0009004-250325103851331>). For robust range delineation, I retained species that had at least 400 georeferenced records. Regarding odonates, I included only species that had IUCN (2025) ranges  $\geq 10,000$  square kilometers (i.e., likely to be less well-known than small ranges) for comparisons, resulting in 225 species. From previous work, I gathered similar records for 3960 species of mammals, anurans, scaled reptiles, birds, and plants that also had least 400 unique georeferenced records and IUCN range areas of at least 10,000 square kilometers (documented in Hanberry, in review). I also added 220 vertebrate and vascular plant species to represent smaller range areas and less than 400 records, to incorporate more sample densities (i.e., number of records relative to area) into the model.

### 2.2. Generation of kernel densities and aggregated polygons from georeferenced records

Package ks, with standard defaults (Duong, 2007; R Core Team,

2025), generated species range kernels, for which I extracted 99% contour lines to enclose about 99% of georeferenced records, which will filter out outliers outside of ranges that may contain potential error due to species misidentification or vagrant individuals (Hanberry, 2025b). Extracting contour lines to raster layers at a resolution of 0.3° (i.e., about 33 km at the equator, approximately matching the distance uncertainty of the coordinates and within the range of values for macroecological studies; Hurlbert and Jetz, 2007; Hanberry, 2026), instead of to vector files, maximized coverage and represented isolated areas, with fast processing time. Other parameter or input changes also create more generalized or specific shapes. For example, smaller sample sizes typically result in larger, smoother, more generalized kernels due to increased bandwidths automatically selected by the ks package to avoid modeling random patterns (Fieberg, 2007). For rare issues, package MASS (Venables and Ripley, 2002), which applies a coarser default bandwidth for greater smoothing, was a replacement. For species that occur on more than one continent, I produced species range kernels for each continent and then combined kernels.

As for aggregating points into a range of polygon shapes and sizes, ArcGIS Pro (ESRI, Redlands, CA, USA), has an aggregate points function in the cartography tools to create polygons around clusters of three or more points (i.e., necessary to create a polygon) within the specified aggregation distance (ESRI, 2025). Because ArcGIS Pro customarily works better with projected data layers, I projected the georeferenced records into Eckert IV, a global area-preserving projection (Snyder, 1987). Six specified distances of 200, 250, 500, 1000, 2000, and 5000 km overall produced both separation between polygons and typically at least two adjacent distances generated suitable ranges, with one more fitted in shape and one offering greater coverage and connectivity.

After merging the kernel, which contained the high certainty sampled range areas with aggregated polygons at six different distance, I selected the best fitting range map in a process structurally analogous to supervised classification, which is common for images that often are too complex for unsupervised classification using scores, simple thresholds, or other automatic approaches such as clustering (Cheng et al., 2020). Out of six different distance options ranging from 200 km to 5000 km, I selected the distance for the aggregated polygons based on visual comparison to the kernels of species occurrences, which provided support from georeferenced records, and International Union for Conservation of Nature (IUCN Standards and Petitions Committee, 2024) ranges, generally by connecting the high certainty kernels and with a balance of maximizing overlap and minimizing areas beyond IUCN ranges (see Fig. 1). To extend the process for invertebrates, for 225 odonate species, after merging species range kernels with six different distances of aggregated polygons and comparing with IUCN ranges, I chose the distance that best embodied the species range kernels, while maximizing overlap with IUCN ranges and minimizing areas outside of the IUCN ranges. Fit encompassed fidelity to the merged kernel and aggregation that represented the known georeferenced records, which are accurate species occurrences yet incomplete throughout species ranges. These selections were added to 3960 species of vertebrate and plant species (documented in Hanberry, in review) and 220 vertebrate and vascular plant species that did not meet thresholds in number of georeferenced records or range areas, to make sure that the classification algorithms had exposure to an assortment of variable values.

2.3. Classification modeling of the aggregation distances

With the samples from classified supervision (i.e., selection of the aggregation distance), I iterated variables to determine the most important ones for classification modeling of the aggregation distance, from six options of 200, 250, 500, 1000, 2000, and 5000 km, while maintaining the same accuracy as the full model with all variables. Initially, I calculated 75 predictor variables derived from number of georeferenced records, the species range kernels, and six aggregation

**Table 1**  
Final 32 variables, distance for aggregation or if kernel, and importance values for four slightly different classification algorithms (rf = random forests, par = parallel random forests).

| Variable                                                               | Distance or kernel | Rf  | Par | Rborist | Ranger |
|------------------------------------------------------------------------|--------------------|-----|-----|---------|--------|
| Aggregated polygon area                                                | 200                | 16  | 19  | 0       | 37     |
|                                                                        | 250                | 20  | 21  | 5       | 46     |
|                                                                        | 500                | 27  | 31  | 10      | 55     |
|                                                                        | 1000               | 50  | 49  | 32      | 76     |
|                                                                        | 2000               | 56  | 50  | 53      | 71     |
|                                                                        | 5000               | 47  | 46  | 32      | 62     |
| Intersection between aggregated polygons and kernel to kernel area     | 200                | 71  | 78  | 100     | 75     |
|                                                                        | 250                | 58  | 53  | 30      | 64     |
|                                                                        | 500                | 31  | 25  | 6       | 37     |
|                                                                        | 1000               | 8   | 10  | 1       | 12     |
|                                                                        | 2000               | 3   | 1   | 1       | 2      |
|                                                                        | 5000               | 0   | 0   | 2       | 0      |
| Area of aggregated polygons and kernel to kernel area                  | 200                | 26  | 25  | 3       | 35     |
|                                                                        | 250                | 23  | 22  | 3       | 24     |
|                                                                        | 500                | 6   | 13  | 7       | 12     |
|                                                                        | 1000               | 11  | 7   | 11      | 14     |
|                                                                        | 2000               | 6   | 8   | 4       | 14     |
|                                                                        | 5000               | 8   | 10  | 7       | 13     |
| Sum of intersection to aggregated polygons and intersection to kernels | 200                | 24  | 23  | 13      | 17     |
| Perimeter length of the aggregated polygons to aggregated polygon area | 200                | 7   | 10  | 9       | 13     |
|                                                                        | 250                | 3   | 6   | 2       | 8      |
|                                                                        | 500                | 1   | 1   | 4       | 6      |
|                                                                        | 1000               | 9   | 10  | 10      | 20     |
|                                                                        | 2000               | 9   | 10  | 8       | 14     |
|                                                                        | 5000               | 5   | 9   | 9       | 14     |
| Point densities of records to aggregated polygon area                  | 200                | 81  | 64  | 23      | 72     |
|                                                                        | 250                | 78  | 79  | 38      | 72     |
|                                                                        | 500                | 100 | 94  | 76      | 98     |
|                                                                        | 1000               | 91  | 100 | 84      | 100    |
|                                                                        | 2000               | 84  | 85  | 38      | 82     |
|                                                                        | 5000               | 85  | 96  | 80      | 80     |
| Point densities of georeferenced records to kernel area                | kernel             | 73  | 76  | 29      | 80     |

distances of polygons. These variables encompassed areas, intersected areas (i.e., shared areas), union of areas (i.e., total combined areas), point densities of records relative to areas, ratios of areas, and perimeter lengths relative to areas (Table 1; Supplementary Table 1).

Algorithms were four random forests variants (random forests, parallel random forests, Rborist, and ranger; caret package; Kuhn, 2008; R Core Team, 2025). The random forest algorithm is an ensemble of non-linear decision trees (Breiman, 2001a, 2001b). Each decision tree is built using a random subset of species occurrence samples, selected with replacement (i.e., bootstrapping), and subset of predictor variables, fit independently of any other tree (i.e., parallel tree construction is called bagging as opposed to boosting; Breiman, 2001a, 2001b). I used the caret package in R (Kuhn, 2008; R Core Team, 2025) with 10-fold cross-validation, repeated three times, to determine hyperparameter settings with maximum accuracy. In the caret package, the only automatic tunable hyperparameter is mtry, the number of randomly selected predictors considered at each split in a tree, with robust default values of 500 trees and classification trees completed with one classification in each terminal node (Breiman, 2001a, 2001b; Probst and Boulesteix, 2018). Typically, tuning of the mtry value is unnecessary because the tuned value is similar the default value of the square root of the number of variables (Breiman, 2001a, 2001b; Probst et al., 2019). I modeled a second time with 25% of samples reserved from modeling for accuracy metrics, but in the random forest modeling process, withheld samples, known as out-of-bag samples, for each tree providing internal unbiased accuracy without a need for a separate testing dataset; that is, training and testing occur as part of the ensemble modeling process so that an additional step of withheld testing data is unlikely to identify variation

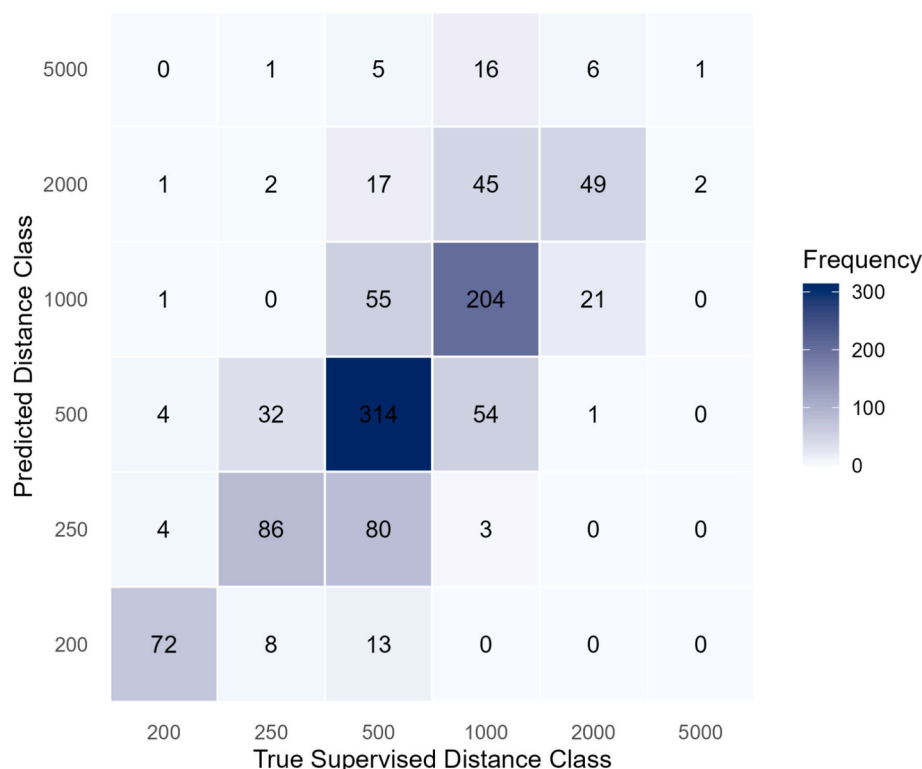


**Fig. 2.** Classification matrix of predicted aggregation distances relative to classified aggregation distances, for withheld samples from 4400 species, with random forests classification.

in model performance (Breiman, 2001a, 2001b).

The algorithms then predicted aggregation distances for 11,844 pollinator species. For any discrepancies in predictions, I selected the mode value and for ties, selected the minimum value. I merged the aggregated polygons of the predicted aggregation distances to the species range kernels to complete the data ranges.

#### 2.4. Intersections of odonate ranges and georeferenced records

For the 225 odonate species, I calculated areas of intersection between the data ranges and IUCN ranges, relative to areas of the IUCN ranges and the union (i.e., the shared intersection area combined with area outside of the intersection; intersection over union score is also known as the Jaccard Index; Gilbert, 1884). I repeated these calculations with a default aggregation distance of 500 km for data ranges to determine if this simple solution was sufficient. Furthermore, I calculated percentage intersection between the odonate georeferenced records and IUCN ranges. Next, I compared these percentages with areas of the data ranges, which encompassed about 99% of georeferenced records.

#### 2.5. Species richness patterns

For macroecological patterns, I converted species ranges by taxa into species richness grids (1-degree resolution; letsR package; Vilela and Villalobos, 2015). For 225 odonate species, I compared species richness grids of IUCN ranges, data ranges with supervised classification, data ranges with a default aggregation distance of 500 km, and also the species range kernels, which represented a slight expansion of the georeferenced records. For other taxa, I compared species richness grids of data ranges and species range kernels. To be clear, species were abundant, filtered based on a minimum of 400 records, and species samples can reflect sampling intensity.

### 3. Results

#### 3.1. Classification

For the full model of 75 variables and four algorithms, accuracies were 0.66 for this six class classification problem, comprised of six possible aggregation distances of 200, 250, 500, 1000, 2000, and 5000 km. Reduction to 32 variables upheld accuracies of 0.66 for four algorithms (Table 1). For withheld samples (25% of all samples), accuracies reduced to 0.65 for parallel random forests (0.6536), Rborist (0.6545), and ranger (0.6481). The four algorithms determined that the most important variables were the point (i.e., georeferenced record) densities relative to areas of the aggregated polygons, along with intersected areas relative to the kernel areas for aggregation distances of 200 and 250 km, followed by areas of the aggregated polygons (Table 1). Additional variables were important, but it appeared that different variables could be retained to maintain accuracies at 0.66 (Supplementary Fig. 1).

For modeling of withheld samples, at aggregated polygon distances of 250, 500, and 1000 km, only 9 out of 859 species (for the random forests algorithm) had predicted classifications that fell outside of one distance above or below the assigned polygon distance (Fig. 2). For the distances of 2000 and 5000 km, predictions became more conservative, with lesser predictions, preventing commission error (Fig. 2). Even though 62% of the modeled species, including 62% of odonate species, were classified to have aggregation distances of 500 and 1000 km, most predictions (86% of species) for the 11,844 pollinator species were aggregation distances of 500 and 1000 km, with another 12% of species predicted to have aggregation distances of 2000 km. The four modeling algorithms concurred for predictions of 11,920 species (93%). As an aside, the random forests and parallel random forests algorithms were the fastest at modeling, with Rborist and ranger relatively slow.

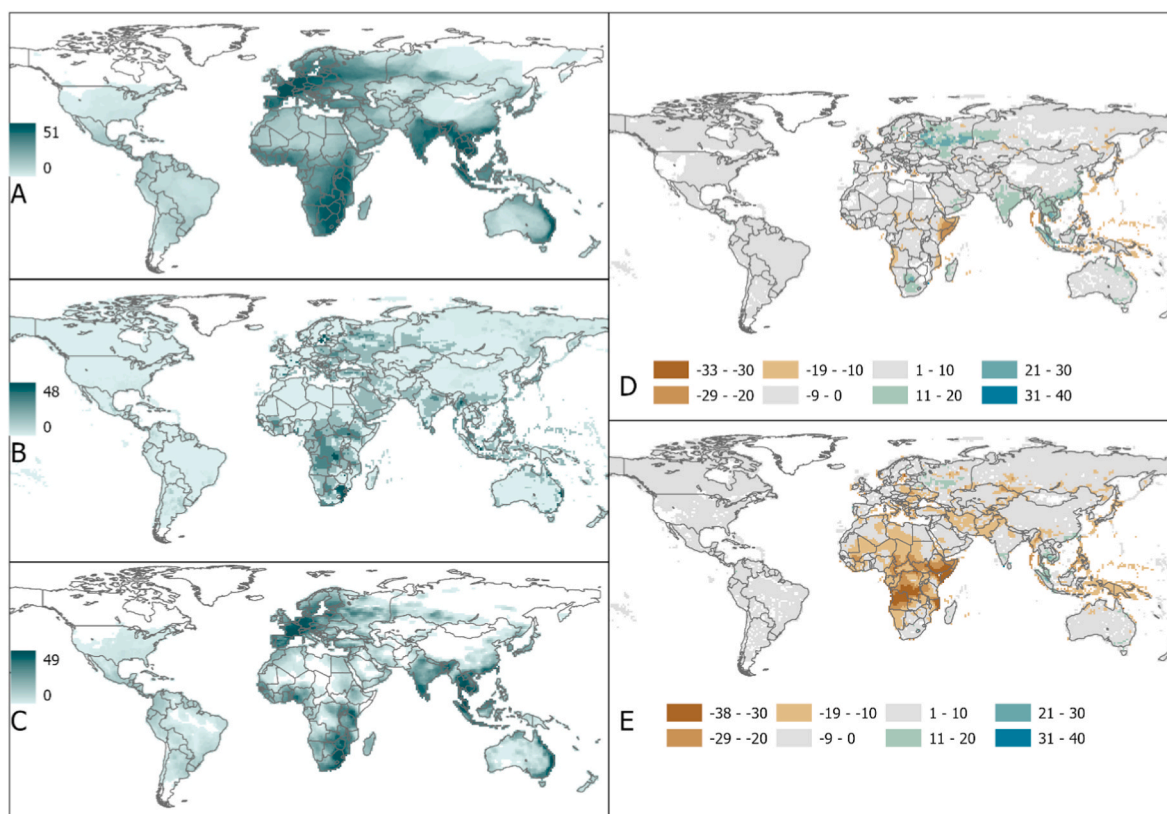


**Fig. 3.** Species richness grids for expert ranges (A), data ranges (B), and species range kernels (C), and differences between data ranges and expert ranges (D) and species range kernels and expert ranges (E), for 225 odonate species.

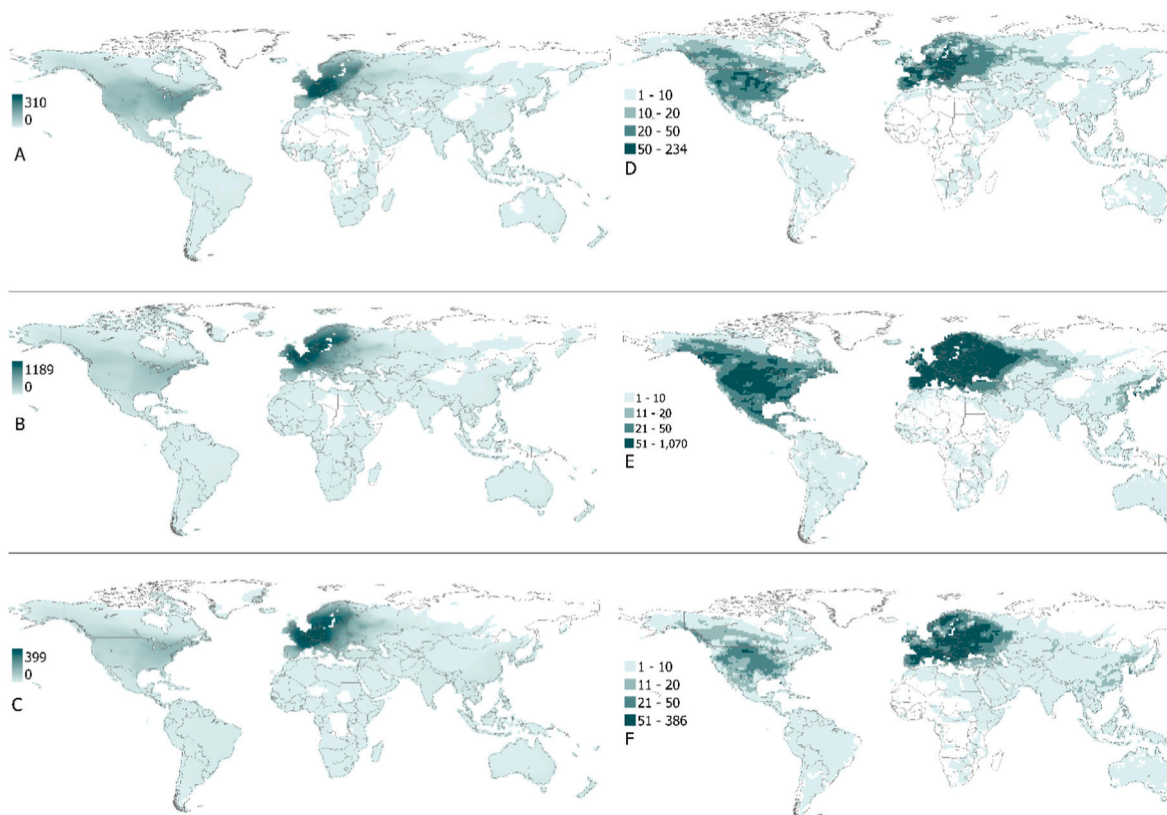


**Fig. 4.** Species richness grids for data ranges of bees (A), flies (B), and wasps (C) and differences between data ranges and species range kernels of bees (D), flies (E), and wasps (F).

### 3.2. Odonate intersections between records and ranges

To illustrate information about expert ranges relative to georeferenced records and data ranges, I summarized intersections for 225 odonate species. For 25% of species, IUCN expert ranges intersected at least 96.2% of georeferenced records. The median value, for 50% of species, was IUCN range intersection with 86.0% of records. But the first quartile was IUCN range intersection with 50.9% of records, meaning that 25% of species ranges contained less than 50.9% of records and another 25% of species ranges contained between 50.9% and 86.0% of records.

The data ranges, which encompassed about 99% of records, had a median value of 83.6% intersection with IUCN expert ranges, with a third quartile value of 91.1% intersection and first quartile value of 70.5% intersection with the expert ranges. However, the median value of the area of data range was twice as large as IUCN ranges. Data ranges included greater area to incorporate 99% of records. For IUCN ranges that overlapped with at least 75% of georeferenced records, data range areas were slightly greater (median value of 1.1) than IUCN ranges. For IUCN ranges that overlapped with less than 50% of georeferenced records, data range areas were almost 34-fold greater than IUCN range areas. Therefore, the intersection area to union area score median value was 35.3%. Instead of supervised classification of the aggregation distance, using a default aggregation distance of 500 km reduced the median values to 74.0% intersection (from 83.6%) with IUCN expert ranges and 26.4% intersection area to union area score (from 35.3%).

### 3.3. Species richness of ranges and species range kernels

Due to inclusion of only species that had at minimum 400 georeferenced records, species richness reflected well-sampled locations, which encompassed Europe for most taxa, and typically parts of the United States, eastern Australia, and some smaller countries or nature preserves. For 225 odonate species, species richness was more similar (correlation of  $r = 0.91$ ) between data ranges and IUCN ranges than between the species range kernels (i.e., a slight expansion of georeferenced records) and IUCN ranges ( $r = 0.80$ ; Fig. 3). Data ranges and species range kernels were similar as well ( $r = 0.90$ ). As a direct approach, species richness of data ranges based on species range kernels and aggregation distances of 500 km only, with no specification of distance, was not an improvement upon species range kernels alone, according to correlation with IUCN ranges ( $r = 0.81$ ).

Data ranges helped reduce sampling deficiencies as reflected in species range kernels, specifically in Africa for odonates. These odonate species were abundantly sampled in Europe, and followed a less typical sampling pattern of greater richness in the Afrotropics south of the Sahara Desert and the Indo-Malayan realm than in North America, as displayed in species richness for data ranges of bees, flies, and wasps, for example (Fig. 4). In comparison to species richness of species range kernels, species richness of data ranges increased with number of species, that is, with greatest increases in Europe and North America for bees, flies, and wasps.

## 4. Discussion

Although species ranges of invertebrates typically reside as static images, georeferenced records are available to build species ranges as spatial data. One approach is to combine species range kernels that contain about 99% of georeferenced records, but are limited by the sampling extent, with polygons created by point aggregation of georeferenced records, which extend the extent. However, without a guide from spatially available ranges developed by experts (IUCN, 2025), the distance for aggregating points to polygons is uninformed. Here, I modeled aggregation distances based on supervised classification of about 4400 species, determined based on fit to expert range maps and faithfulness to georeferenced records. Classification accuracy was 0.66,

for a six-class problem with numerous possible representations of species ranges. There is no absolute right solution, as objective reality is realized through approximate representations (Denzin, 2012; Bans-Akutey and Tiimub, 2021). Expert ranges overall are useful representations of species ranges but include omission and commission errors (Jetz et al., 2012). As illustrated with odonate species, some omission errors are possible to detect through comparison to georeferenced records (Ficetola et al., 2014; Alhajeri and Fourcade, 2019), although global sampling coverage is incomplete (Hughes et al., 2021).

A shorter aggregation distance will produce better fit and a longer aggregation distance will improve coverage. While 0.66 may not appear to be the most accurate classification, in practice for classification of six classes, this meant that most species had predicted distances of either the aggregated polygon distance of supervised classification (i.e., accurate classification of approximation similarity), or one distance above or below the aggregated polygon distance, based on results from withheld samples. For aggregated polygon distances of 250, 500, and 1000 km, only about 1% of species had predicted classifications outside of one distance above or below the assigned polygon distance. The shortest distance of 200 km had a greater tendency to be classed as a longer distance (i.e., 13 species had predicted classifications of 500 km, or two distance classes above 200). As an outcome of an intentional classification decision to be conservative and avoid commission error, the aggregated polygon distances of 2000 and 5000 were predicted primarily within the 500, 1000, and 2000 distance classes. Indeed, for the 11,844 pollinator species, 86% of species were predicted to have aggregation distances of 500 or 1000, as opposed to 62% of the modeled odonate, vertebrate, and plant species.

This result of convergence to the middle aggregation distances suggests that for fast generation of terrestrial insect ranges, simply generating polygons from aggregated distances of 500 km may be serviceable for most terrestrial insect species, encompassing similar ranges at the 250 and 1000 km aggregation distances. Unfortunately, at least for odonate species, this more direct approach generated less accurate ranges than with supervised classification, according to lesser values of intersection with IUCN range maps and correlation of species richness grids with species richness grids from IUCN range maps. Along similar lines, distances less than 200 km may be valuable for a small subset of species, although then perhaps the species range kernels would adequately cover the ranges. Supervised classification of thousands of species did result in improved outcomes in matching IUCN ranges, as opposed to using a default aggregation distance when generating ranges. Similarly, simply widening expert ranges to reduce omission error (Jetz et al., 2012), or using a default distance of at least 1000 km, is likely to increase commission error unbalanced by decreased omission error.

By design, the objective in developing ranges is to build ranges as much as possible without risking inclusion of large extents of unoccupied areas (IUCN Standards and Petitions Committee, 2024). The aggregated polygons extended the range conservatively overall from the species range kernels, with few aggregated distances of 2000 and 5000 km. Apparently for odonate species, experts were particularly cautious by targeting numerous, known small locations as species ranges (Fig. 1; IUCN Standards and Petitions Committee, 2024). However, this meant that 25% of species ranges contained less than 51% of georeferenced records, and another 25% of species ranges had less 86% of georeferenced records. Less conservative range delineation may be more desirable to incorporate greater intersection with georeferenced records, corroborating range locations through research triangulation (Denzin, 2012; Bans-Akutey and Tiimub, 2021). Data ranges accounted for 99% of georeferenced records and were greater in areas than IUCN ranges. Certainly, omission errors remained, even though corrected by the locational information of georeferenced records. For example, species with Asian distributions demonstrated omissions (Figs. 3 and 4), but support for greater distributions was not present in the georeferenced records (Hughes et al., 2021). Nonetheless, data ranges were produced with a consistent and objective approach, which then could be tailored

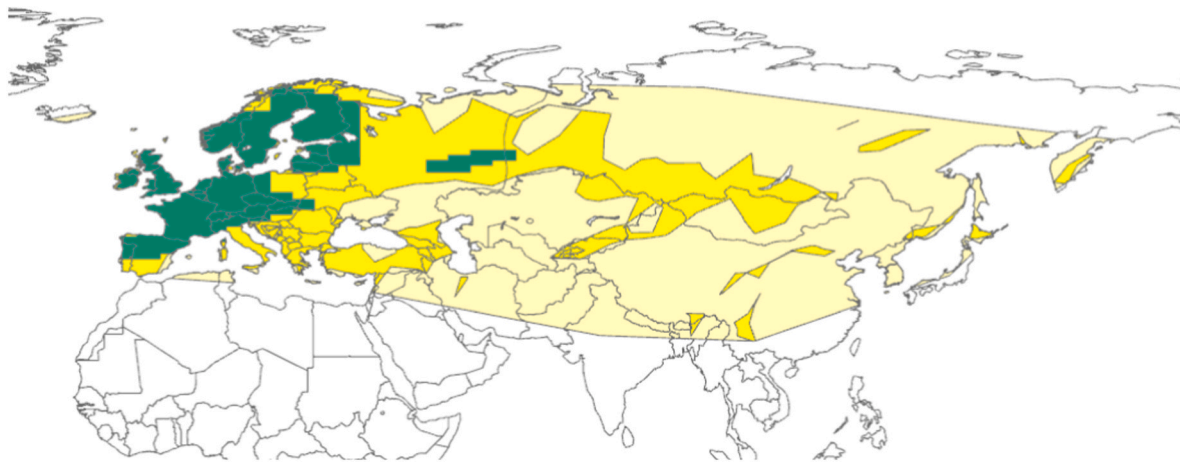


**Fig. 5.** Data range of *Aglais urticae*, a butterfly species, incorporating a species range kernel (green, with 99% of georeferenced records) and georeferenced records aggregated to polygons at an aggregation distance of 500 km, based on model predictions (bright yellow). An aggregation distance of 5000 km (light yellow) provides a more comprehensive Eurasian distribution (Meshcheryakova et al., 2023) but does not have support from the georeferenced records (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

directly by expert opinion or static maps developed by experts.

Data ranges helped reduce sampling deficiencies in georeferenced records (Hughes et al., 2021), while resembling expert ranges. Species range kernels encircled about 99% of georeferenced records, expanding yet primarily representing recorded areas of occupancies, which were extended again by data ranges. Likewise, most of the data ranges overlapped with most of the IUCN ranges (i.e., a first quartile value of 70.5% intersection for 225 odonate species with IUCN ranges). Species richness grids were more similar (correlation of  $r = 0.91$ ) between data ranges and IUCN ranges than between species range kernels and IUCN ranges ( $r = 0.80$ ). Species richness grids were at 1-degree resolution, which is typical for global richness studies, and appropriate for the resolution of ranges, although 2-degree resolution may be preferable, particularly for taxa with information deficits (Hurlbert and Jetz, 2007; Ballesteros-Mejia et al., 2017). Relative to species richness grids of species range kernels, which directly reflected the georeferenced records, species richness grids of data ranges increased with number of species (that is, where sampling intensity resulted in the most species with at least 400 records). Data ranges still represented underlying sampling gradients, as both an advantage in remaining true to known information and a disadvantage, due to incomplete sampling coverage (Figs. 4 and 5, Hughes et al., 2021).

Species occupancies often are complex and dynamic, with species densities exhibiting a variety of patterns (Hurlbert and Jetz, 2007; Hanberry, 2021). Ranges identify reasonable areas of possible species presence during a given time period (Jetz et al., 2012). However, the geographic space may not contain the full suite of necessary environmental conditions available at all locations, including absence of deterrents such as competitors and other barriers (Hurlbert and Jetz, 2007). The areas inside ranges incorporate most individuals during a given time interval, with restraint in delimiting all known individuals if too much area is required to gather vagrants or individuals transported long distances by different vectors (IUCN Standards and Petitions Committee, 2024). The expert range maps particularly appeared to limit recent expansion, whereas the data ranges reflected georeferenced records as faithfully as possible, with circumscription (e.g., 99% and not 100% of georeferenced records). Data ranges, due to aggregated polygons, amplified areas of occupancies (species range kernels) to extents of occurrences (ranges). Data ranges may be helpful for applications that are limited by sampling deficiencies, such as species distribution models, which require untruncated samples (Thuiller et al., 2004). For invertebrate species without range maps, data ranges fill the gap by contributing accessible spatial data, which can be developed over time, similar

to expert ranges.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgements

I thank the Global Biodiversity Information Facility, International Union for Conservation of Nature, and contributors to the range maps and georeferenced observations. This research was supported by the U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, and the U.S. Department of Agriculture, Forest Service, with funding from the Infrastructure Investment and Jobs Act of 2021. The findings and conclusions in this publication are those of the author and should not be construed to represent any official USDA or U.S. Government determination or policy.

### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.actao.2026.104192>.

### Data availability

Statement Insect ranges are available at <https://zenodo.org/records/17195899>.

### References

- Alhajeri, B.H., Fourcade, Y., 2019. High correlation between species-level environmental data estimates extracted from IUCN expert range maps and from GBIF occurrence data. *J. Biogeogr.* 46 (7), 1329–1341.
- Aronsson, H., Zizka, A., Antonelli, A., Faurby, S., 2024. Expert-based range maps cannot be replicated using data-driven methods but macroecological conclusions arising from them can. *J. Biogeogr.* 51 (8), 1549–1559.
- Ballesteros-Mejia, L., Kitching, I.J., Jetz, W., Beck, J., 2017. Putting insects on the map: Near-global variation in sphingid moth richness along spatial and environmental gradients. *Ecography* 40 (6), 698–708.
- Bans-Akutey, A., Tiimub, B.M., 2021. Triangulation in research. *Academia Lett.* 2 (3392), 1–7.
- Breiman, L., 2001a. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16 (3), 199–231.
- Breiman, L., 2001b. Random forests. *Mach. Learn.* 45 (1), 5–32.

- Burgman, M.A., Fox, J.C., 2003. Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. *Animal Conserv. Forum* 6 (1), 19–28.
- Cantú-Salazar, L., Gaston, K.J., 2013. Species richness and representation in protected areas of the Western hemisphere: discrepancies between checklists and range maps. *Divers. Distrib.* 19 (7), 782–793.
- Cheng, G., Xie, X., Han, J., Guo, L., Xia, G.S., 2020. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 13, 3735–3756.
- Denzin, N.K., 2012. Triangulation 2.0. *J. Mix. Methods Res.* 6 (2), 80–88.
- Duong, T., 2007. ks: kernel density estimation and kernel discriminant analysis for multivariate data in R. *J. Stat. Software* 21, 1–16.
- ESRI, 2025. Aggregate points (standard Feature Analysis). <https://pro.arcgis.com/en/pro-app/latest/tool-reference/cartography/aggregate-points.htm>. (Accessed 21 February 2025).
- Ficetola, G.F., Rondinini, C., Bonardi, A., Katariya, V., Padoa-Schioppa, E., Angulo, A., 2014. An evaluation of the robustness of global amphibian range maps. *J. Biogeogr.* 41 (2), 211–221.
- Fieberg, J., 2007. Kernel density estimators of home range: smoothing and the autocorrelation red herring. *Ecology* 88 (4), 1059–1066.
- Gaston, K.J., Fuller, R.A., 2009. The sizes of species' geographic ranges. *J. Appl. Ecol.* 46 (1), 1–9.
- Gilbert, G.K., 1884. Finley's tornado predictions. *American Meteorol. J.* 1, 166–172.
- Global Biodiversity Information Facility [GBIF], 2025. Free and open access to biodiversity data. Available at: [www.gbif.org](http://www.gbif.org). (Accessed 2 January 2025).
- Hamilton, A.J., Novotný, V., Waters, E.K., Basset, Y., Benke, K.K., Grimbacher, P.S., Miller, S.E., Samuelson, G.A., Weiblen, G.D., Yen, J.D., Stork, N.E., 2013. Estimating global arthropod species richness: refining probabilistic models using probability bounds analysis. *Oecologia* 171, 357–365.
- Hanberry, B.B., 2021. Transition from fire-dependent open forests: alternative ecosystem states in the southeastern United States. *Diversity* 13 (9), 411.
- Hanberry, B.B., 2025a. Potential expanded pollinator distributions in North America under future climate. *Ecol. Solutions Evid.* 6 (2), e70025.
- Hanberry, B.B., 2025b. Clarifying influences of sampling bias (concentration) and locational errors (uncertainties) on precision or generality of species distribution models. *Land* 14 (8), 1620.
- Hanberry, B.B., 2026. Beyond thinning: equalized sampling based on global sampling intensity gradients for species distribution models. *Front. Ecol. Evol.* 14, 1831501.
- Hughes, A.C., Orr, M.C., Ma, K., Costello, M.J., Waller, J., Provoost, P., Yang, Q., Zhu, C., Qiao, H., 2021. Sampling biases shape our view of the natural world. *Ecography* 44 (9), 1259–1269.
- Hurlbert, A.H., Jetz, W., 2007. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proc. Natl. Acad. Sci.* 104 (33), 13384–13389.
- IUCN, 2025. The IUCN red list of threatened species. Spatial Data Download. <http://www.iucnredlist.org/resources/spatial-data-download>. (Accessed 25 March 2025).
- IUCN Standards and Petitions Committee, 2024. Guidelines for using the IUCN red list categories and criteria. Version 16. Prepared by the standards and petitions committee. Downloadable from: <https://www.iucnredlist.org/documents/RedListGuidelines.pdf>. (Accessed 27 March 2025).
- Jetz, W., McPherson, J.M., Guralnick, R.P., 2012. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.* 27 (3), 151–159.
- Kuhn, M., 2008. Building predictive models in R using the caret package. *J. Stat. Software* 28, 1–26.
- Meshcheryakova, E.N., Bulakhova, N.A., Zhigulskaya, Z.A., Shekhovtsov, S.V., Berman, D.I., 2023. Wintering and cold hardiness of the small tortoiseshell *Aglais urticae* (Linnaeus, 1758) (Nymphalidae, Lepidoptera) in the West and East of the Northern Palearctic. *Diversity* 15 (1), 72.
- Probst, P., Boulesteix, A.L., 2018. To tune or not to tune the number of trees in random forest. *J. Mach. Learn. Res.* 18 (181), 1–18.
- Probst, P., Wright, M.N., Boulesteix, A.L., 2019. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery* 9 (3), e1301.
- R Core Team, 2025. R: a language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Scheffers, B.R., Joppa, L.N., Pimm, S.L., Laurance, W.F., 2012. What we know and don't know about Earth's missing biodiversity. *Trends Ecol. Evol.* 27 (9), 501–510.
- Snyder, J.P., 1987. Map Projections – a Working Manual. USGS, Denver, Colorado, USA.
- Thuiller, W., Brotons, L., Araújo, M.B., Lavorel, S., 2004. Effects of restricting environmental range of data to project current and future species distributions. *Ecography* 27 (2), 165–172.
- Tosa, M.I., Dziedzic, E.H., Appel, C.L., Urbina, J., Massey, A., Ruprecht, J., Eriksson, C.E., Dolliver, J.E., Lesmeister, D.B., Betts, M.G., Peres, C.A., 2021. The rapid rise of next-generation natural history. *Front. Ecol. Evol.* 9, p698131.
- Venables, W.N., Ripley, B.D., 2002. *Modern Applied Statistics with S*, fourth ed. Springer, New York <https://www.stats.ox.ac.uk/pub/MASS4/>.
- Vergara-Asenjo, G., Alfaro, F.M., Pizarro-Araya, J., 2023. Linnean and Wallacean shortfalls in the knowledge of arthropod species in Chile: challenges and implications for regional conservation. *Biol. Conserv.* 281, 110027.
- Vilela, B., Villalobos, F., 2015. letsR: a new R package for data handling and analysis in macroecology. *Methods Ecol. Evol.* 6 (10), 1229–1234.
- Worton, B.J., 1987. A review of models of home range for animal movement. *Ecol. Model.* 38 (3–4), 277–298.