"Collection Bias" and the Importance of Natural History Collections in Species Habitat Modeling: A Case Study Using *Thoracophorus costalis* Erichson (Coleoptera: Staphylinidae: Osoriinae), with a Critique of GBIF.org

Author(s): Michael L. Ferro and Andrew J. Flick
Source: The Coleopterists Bulletin, 69(3):415-425.
Published By: The Coleopterists Society
DOI: http://dx.doi.org/10.1649/0010-065X-69.3.415
URL: http://www.bioone.org/doi/full/10.1649/0010-065X-69.3.415

The Coleopterists Bulletin, 69(3): 415–425. 2015.

"Collection Bias" and the Importance of Natural History Collections in Species Habitat Modeling: A Case Study using *Thoracophorus costalis* Erichson (Coleoptera: Staphylinidae: Osoriinae), with a Critique of GBIF.org

Michael L. Ferro
Louisiana State Arthropod Museum, Department of Entomology
Louisiana State University Agricultural Center, 404 Life Sciences Building
Baton Rouge, LA 70803, U.S.A.
spongymesophyll@gmail.com

and

Andrew J. Flick
Department of Biological Sciences, Louisiana State University
232 Life Sciences Building, Baton Rouge, LA 70803, U.S.A.
aflick1@lsu.edu

Abstract

When attempting to understand a species' distribution, knowing how many collections should be surveyed to achieve an adequate sample (exhaustiveness) is important. A test for exhaustiveness using species distribution models created with Diva-GIS was performed on county level locality information recorded from more than 4,900 specimens of *Thoracophorus costalis* Erichson (Staphylinidae: Osoriinae) borrowed from 38 collections. Size and location of distribution models based on specimens from single collections varied greatly, indicating "collection bias." At least 15 collections needed to be combined before the resultant model averaged 90% of the area of a reference model created from all available specimens. By themselves, alternative distribution data from literature, Bugguide.net, and GBIF.org performed poorly, resulting in models with less than 15% the area of the reference model. Comments on the use of online data, the importance of maintaining and growing regional collections, and the future of natural history collections are included.

Key Words: biogeography, predictive modeling, niche modeling, rove beetles, furrowed rove beetle, landscape ecology, gamma distributions

---

We've learned from experience that the truth will out. Other experimenters will repeat your experiment and find out whether you were wrong or right. Nature's phenomena will agree or they'll disagree with your theory. And, although you may gain some temporary fame and excitement, you will not gain a good reputation as a scientist if you haven't tried to be very careful in this kind of work. And it's this type of integrity, this kind of care not to fool yourself, that is missing to a large extent in much of the research in Cargo Cult Science.
-Richard Feynman (1974)

Natural history collections of today evolved from "wunderkammer" of the 16th century. Modern collections are still "cabinets of curiosities" but are also considered "Mission-Critical Infrastructure" (NSTC 2009) and are the subject of a plethora of publications describing their importance and maintenance, including a 2400+ page "handbook" by the US National Park Service (NPS 2006). Whether to continue growing natural history collections or declare the exercise good and done continues to be an unending topic of gleeful debate (for recent examples, see Lujan and Page 2015 and Rocha *et al*. 2014 and associated papers).

Specimens housed in natural history collections are important for, among other things, understanding the distribution of a particular species. The importance of any given specimen is difficult to determine because records represent spatial as well as biotic information. For example, a forest dwelling species collected in a forest 50 km farther west than the species has previously been reported represents a spatial range extension, while a specimen from a nearby prairie indicates a biotic range extension and may be more important to the overall understanding of the species' distribution. Therefore, number of specimens or number of localities are, by themselves, poor indicators of whether a given data set will provide an accurate representation of a species' overall distribution.

When determining a species' distribution, researchers should examine as many specimens

as possible to confirm identifications and locality data. When trying to deduce if specimens of a particular species will be present before requesting loans or personally visiting collections, researchers judge natural history collections informally based on the collection's location, age, quantity of specimens, taxonomic emphasis, *etc*. However, within the holdings of a collection that does contain the desired species, potential variation of distribution (or phenology, morphology, genetics, *etc*.) of that species is difficult to surmise.

Natural history collections represent a functional "natural unit" of specimens, because: 1) curators tend to database or loan all specimens of a particular species they are aware they have; 2) specimens are generally exclusive to a single collection; and 3) researchers use all the data they get from a collection. From a practical standpoint, an investigator conducting research on a particular species does not gather data on an arbitrary number of specimens but rather assembles data from the holdings of discrete collections. Little guidance, other than intuition, is available when attempting to judge the exhaustiveness (*i.e*., proportion of information compiled, Hortal *et al*. 2007) of data accumulation for any given species. When attempting to understand a given attribute of a species, such as distribution, phenology, genetic variation, *etc*., knowing how many collections should be surveyed is more important than the predetermined number of specimens examined to achieve an adequate sample but not waste resources by going beyond a point of diminishing returns.

Species distribution modeling, also referred to as species habitat modeling, environmental niche modeling, *etc*., overlays specimen locality information on environmental variables to create a prediction of a species' full distribution (Elith and Leathwick 2009). Models can vary based on technique, such as regression or machine learning, and variables used, such as various aspects of climatic data, spatial scales, species interactions, availability of food/resources, dispersal ability, and use of presence/absence or presence only data (Newbold 2010). When variables are held constant, species distribution models should be able to provide standardized comparisons of distributional data among natural history collections.

During compilation of an updated distribution of the furrowed rove beetle, *Thoracophorus costalis* Erichson (Staphylinidae: Osoriinae), the first author reviewed 4,926 specimens from 38 collections and compiled the most comprehensive collection of distributional data for the species to date (Ferro 2015). The resultant data set allowed for a case study evaluating how the model of a species' distribution was affected by the number of collections surveyed, and how it compared to models created solely from literature records and online databases.

Material and Methods

Adult specimens of *T. costalis* were examined from the following institutions. Collections and their acronyms are from Evenhuis (2014). Collection managers and curators are indicated.

BYUC Monte L. Bean Life Science Museum, Brigham Young University (Provo, UT, USA; Shawn Clark).

CAS California Academy of Sciences (San Francisco, CA, USA; Norman Penny).

CNC Canadian National Collection of Insects (Ottawa, ON, Canada; Patrice Bouchard).

CSCA California State Collection of Arthropods (Sacramento, CA, USA; Jacqueline Kishmirian-Airoso).

CSUC Colorado State University (Fort Collins, CO, USA; Boris Kondratieff).

CUAC Clemson University (Clemson, SC, USA; Michael Caterino).

CUIC Cornell University (Ithaca, NY, USA; Jason Dombroskie).

EMEC Essig Museum of Entomology, University of California (Berkeley, CA, USA; Cheryl Barr and Peter Oboyski).

FMNH Field Museum of Natural History (Chicago, IL, USA; James Boone).

FSCA Florida State Collection of Arthropods, Division of Plant Industry (Gainesville, FL, USA; Paul Skelley).

ICUI The University of Iowa Museum of Natural History (Iowa City, IA, USA; Elizabeth Fouts, Cindy Opitz).

INHS Illinois Natural History Survey (Champaign, IL, USA; Jamie Zahniser).

KSPC Kyle Schnepp personal collection (Gainesville, FL, USA; Kyle Schnepp).

LSAM Louisiana State Arthropod Museum, Louisiana State University (Baton Rouge, LA, USA; Victoria Bayless).

MCZ Museum of Comparative Zoology, Harvard University (Cambridge, MA, USA; Rachel Hawkins).

MEM Mississippi State University (Starkville, MS, USA; Terence Schiefer).

MSUC Michigan State University (East Lansing, MI, USA; Anthony Cognato, Gary Parsons).

MTEC Montana State University (Bozeman, MT, USA; Michael Ivie).

NCSU North Carolina State University Insect Collection (Raleigh, NC, USA; Bob Blinn).

OMNH Oklahoma Museum of Natural History, University of Oklahoma (Norman, OK, USA; Katrina Menard).

OSUC C. A. Triplehorn Insect Collection, Ohio State University (Columbus, OH, USA; Luciana Musetti).
PMNH Peabody Museum of Natural History, Yale University (New Haven, CT, USA; Lawrence F. Gall).
SEMC Snow Entomological Museum, University of Kansas (Lawrence, KS, USA; Zachary Falin).
TAMU Texas A & M University (College Station, TX, USA; Ed Riley).
UAAM The Arthropod Museum, Department of Entomology, University of Arkansas (Fayetteville, AR, USA; Jeffrey K. Barnes).
UASM E. H. Strickland Entomological Museum, University of Alberta (Edmonton, AB, Canada; Danny Shpeley).
UCDC R. M. Bohart Museum of Entomology, University of California (Davis, CA, USA; Lynn Kimsey).
UCFC University of Central Florida (Orlando, FL, USA; Sandor Kelly).
UCMS University of Connecticut (Storrs, CT, USA; Jane O'Donnell).
UCRC Entomology Research Museum, Department of Entomology, University of California (Riverside, CA, USA; Doug Yanega).
UGCA University of Georgia (Athens, GA, USA; E. Richard Hoebeke).
UMRM W. R. Enns Entomology Museum, University of Missouri (Columbia, MO, USA; Kristin Simpson).
UMSP University of Minnesota (St. Paul, MN, USA; Robin Thomson).
UNHC University of New Hampshire (Durham, NH, USA; Donald Chandler).
VMNH Virginia Museum of Natural History (Martinsville, VA, USA; Nancy Moncrief).
WFBM W. F. Barr Entomological Collection, University of Idaho (Moscow, ID, USA; Frank Merickel).
WIRC University of Wisconsin Insect Research Center, Department of Entomology, University of Wisconsin (Madison, WI, USA; Steven Krauth).
WSU Maurice T. James Entomological Collection, Washington State University (Pullman, WA, USA; Richard Zack).

For specimens examined, county or county equivalents (hereafter referred to as county) were recorded when provided, or when the county could be reasonably inferred from other locality information such as city and state. Specimens with inadequate locality information were excluded from analysis, *e.g*. "Ill.". A single point at the center of each county was mapped based on coordinates provided on respective pages available at www.wikipedia.org. *Thoracophorus costalis* is known only from the US and Canada, except for a single specimen from Mexico (MCZ), the label reading in full: "Mex.", which is not included in this analysis. Therefore, the data used represent the species' entire range as is currently known.

Alternative distribution data for *T. costalis* were taken from three sources. Occurrence data were downloaded from Global Biodiversity Information Facility (GBIF 2015) on 25 January 2015. Data from GBIF consisted of 142 records, all from SEMC, and were used as-is. County level distribution was recorded for images on BugGuide (2015) on 25 January 2015 and consisted of 14 specimens representing 13 counties. BugGuide data were mapped as above. State level distribution data (five localities: CT, FL, IN, MD, NJ) were taken from Downie and Arnett (1996) and mapped by selecting coordinates at a single location in the center of each state.

Distribution modeling was performed using the program DIVA-GIS (Hijmans *et al.* 2012), version 7.5.0.0. Global climate layers came from www.worldclim.org: generic grids; 30 seconds resolution; Bioclim 1–18 (Hijmans *et al*. 2005). Models were created using the Ecological Niche Modeling function in DIVA-GIS with the following parameters: Output Grid Dimensions MinX −130°, MaxX −50°, MinY 20°, MaxY 60°; Bioclim variables selected 1 (annual mean temperature), 3 (isothermality), 4 (temperature seasonality), 7 (temperature annual range), 12 (annual precipitation), 15 (precipitation seasonality); Bioclim as output variable; all other options default. The model was evaluated in DIVA-GIS by creating a receiver operating characteristic (ROC) curve using 75% of the localities as "training" data. The quality of the model's prediction was qualified by calculating the area under the curve (AUC) of the ROC, where values range from 0.5, indicating the model is no better than random, to 1 where the model has maximum accuracy; values greater than 0.9 represent 'high accuracy' (Swets 1988).

All available data from specimens examined (38 collections, 464 counties, 610 collection-county records) were combined to create a single reference model of *T. costalis* distribution against which all other models were compared. Comparisons among models were based on area of projected distribution only, regardless of the geographic coverage of the model—greater area of individual models equaled greater similarity to the reference model. Distribution area included all suitable prediction levels, from Low to Excellent, and was calculated using the Analyze: Measure function in the program ImageJ (Rasband 2014) for all models.

Rarefaction was used to study the effect of number of collections on the distribution model.

For each possible non-total number of collections (1–37) 20 sets (without resampling) were randomly created using the program R version 3.3 (R Core Team 2014), e.g., 20 sets of 17 randomly selected collections, *etc*. Distribution models were generated for each set and total area of the distribution of each model was calculated (740 total). The average and 95% confidence interval of the mean of model area for each set were calculated using R (R Core Team 2014). A logistic regression was fit to the logistic curve data and used to predict mean number of collections necessary to achieve 90% of the reference distribution.

Rarefaction was used to study the effect of number of localities on the distribution model to test for bias of model area within individual collections. From the set of combined data (unique localities only), 20 sets (without resampling) of 3, 5, 10, 25, 50, 100, 150, 200, 250, and 300 random specimen localities were selected. Distribution models were generated for each set and total area of the distribution of each model was calculated (200 total). Results with 95% confidence intervals of the mean were graphed. If models generated by individual collection holdings behaved the same as the entire data set, they should fall on or near the line created by the random model area distribution. Collections that fall below the line have specimens that result in a smaller than expected model (clumped) in relation to the number of localities, while collections that fall above the line have specimens that result in a larger than expected model (dispersed).

Results

The reference model matched the known distribution of *T. costalis* closely, with the exception of extreme northeastern North America, western Texas, and the mountainous regions in the western US and Canada (Fig. 1). Lack of specimens from those areas may be the result of poor collecting effort, suitable but inaccessible habitat for *T. costalis* (especially in the mountainous west), or errors in the model. The model does project *T. costalis* distribution into Mexico, a prediction supported by the MCZ specimen. The AUC value for the reference model (75% data used) was 0.910, which indicated the model was highly accurate (Swets 1988). Overall, the model parameters appeared to be satisfactory to produce comparative distributions for this research.

Three collections had specimens from only a single county and failed to produce a distribution model (Table 1). Area of distribution models of other collections were generally small compared to the reference model: 18 individual collections covered less than 10%; 13 collections covered 10–50%; and only three covered more than 50% (Table 1; Figs. 2–4, 9). Alternative distribution data resulted in models that ranged 8.5–14.1% of the area of the reference model (Table 1; Figs. 6–8, 9). An argument could be made that most collections surveyed could have been ignored because of poor contribution to the overall model. However, this *ex post facto* reasoning fails to take into account that the contribution of any given collection is unknown before it is "observed", and once observed, the data may as well be used.

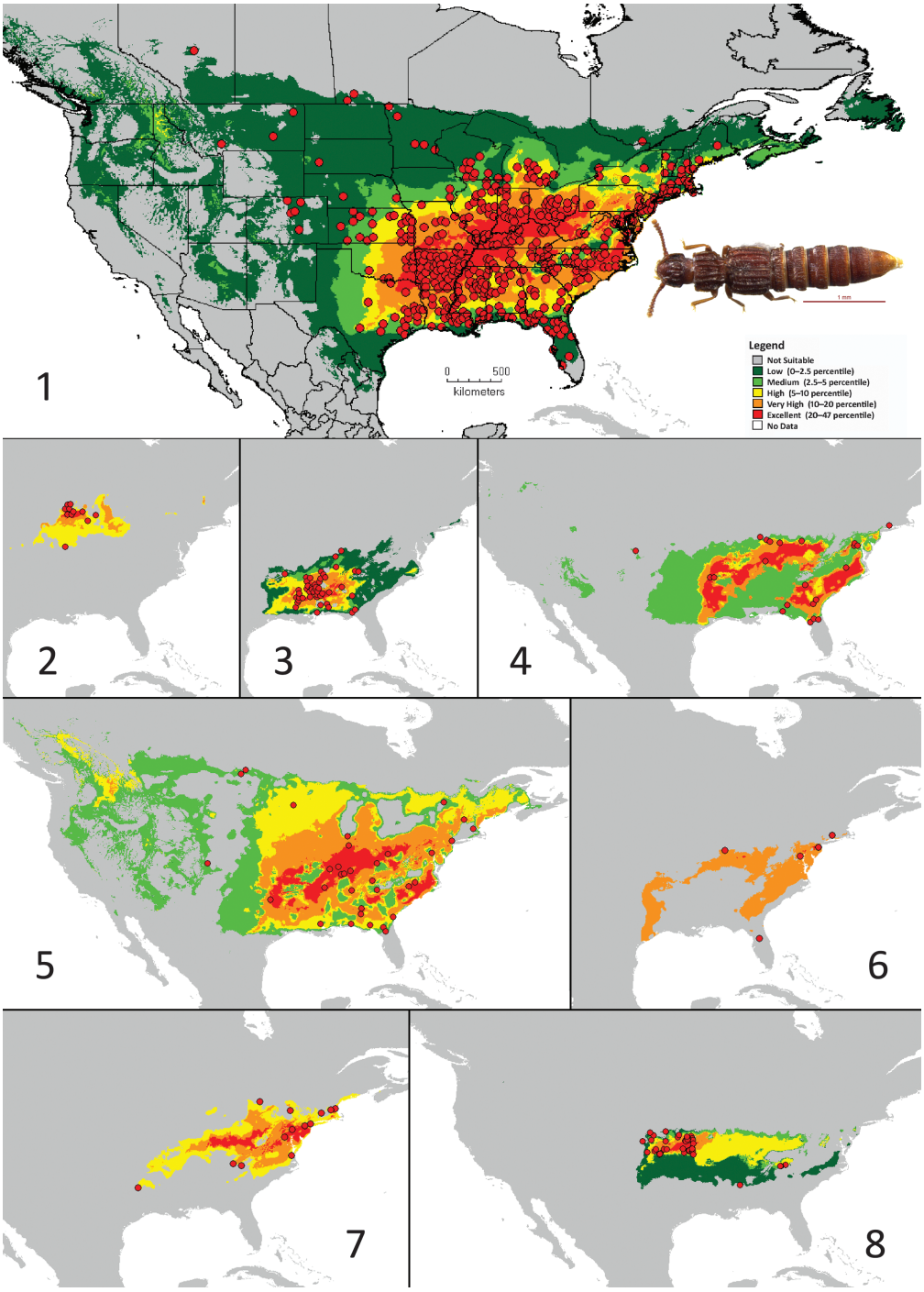
The average area of distribution models from single collections was significantly lower than models resultant from two collections combined (Fig. 10). Alternative distribution data resulted in models with areas indistinguishable from single collections: Downie and Arnett (1996), 8.53%; BugGuide, 11.88%; and GBIF, 14.05% (Table 1; Fig. 10). At least five combined collections were needed to produce a model that averaged greater than 50% the reference model area, and 15 collections were necessary to create a model with 90% of the reference model area. The curve reached an asymptote beginning at 15 and stabilized at 25 collections (Fig. 10).

In total, 250 localities randomly selected from the combined data were needed to create a model that averaged 90% the area of the reference model (464 localities) (Fig. 11). Note that localities, not specimens, were used, and the localities came from the entire data set. Most tests of sample size on model accuracy pull samples from an entire data set (Carroll and Pearson 1998; Cumming 2000; Stockwell and Peterson 2002; Hernandez *et al.* 2006). However, in this example most individual collections fell below the curve and exhibited a "clumped" bias, and only two collections were appreciably above it, showing a "dispersed" bias. Therefore, tests of sample size on random subsets of individual collections would have perpetuated the bias of that collection.

Absolute number of localities (or specimens) (Fig. 11) was a poor measure when predicting accuracy of the resultant model. The GBIF data exhibited a clumped bias and had a high number of "localities" compared to other sources, because data were used as-is and "localities" represented individual specimens. The CNC data included 42 localities and created a model with 75% of the reference area, while FMNH had nearly three times the number of localities (123) but produced a model only two-thirds the size (52%) (Fig. 11).

Discussion

Collection Bias. Variation in model area among collections (Table 1) shows that "collection bias" can be large and may eclipse smaller scale biases.



Figs. 1–8. Exemplar distribution models. Red circles = specimen localities (county-level). **1)** Reference model created using all specimen records. Insert: *Thoracophorus costalis*; **2)** University of Wisconsin Insect Research Center, 4.4% reference area; **3)** Mississippi State University, 11.9% reference area; **4)** Florida State Collection of Arthropods, 28.9% reference area; **5)** Canadian National Collection of Insects, 75.5% reference area, greatest of any single collection; **6)** Downie and Arnett (1996), 8.5% reference area; **7)** Bugguide.net, 11.9% reference area; **8)** GBIF.org, 14.1% reference area.

Table 1. Number of specimens (# spec.), counties (# Co.), and percent model area (Model %) for each collection (see text for collections designated by codens). * = alternative distribution data.

Collection	# spec.	# Co.	Model %	Collection	# spec.	# Co.	Model %
BYU	1	1	0	GMNH	52	8	7.80
OMNH	5	1	0	*DA1996	NA	5 states	8.53
UIMNH	2	1	0	OSUC	27	7	11.23
PMNH	6	2	<0.01	TAMU	159	16	11.71
UCDC	4	3	<0.01	*BugGuide	14	13	11.88
UCFC	4	2	0.01	MEM	171	47	11.92
UCMS	2	2	0.03	UASM	8	4	13.33
WSU	27	3	0.03	*GBIF.org	142	NA	14.05
CSUC	3	2	0.09	NCSU	223	14	14.36
EMEC	5	2	0.15	LSAM	1,553	40	19.42
WFBM	4	2	0.16	INHS	175	28	19.97
CSCA	45	6	0.28	UAAM	323	50	21.73
VMNH	15	7	0.52	UNHC	65	23	25.32
UMRM	22	6	0.53	FSCA	119	22	28.93
UMSP	23	3	0.69	CUIC	114	11	30.53
MTEC	13	3	2.21	SEMC	198	45	30.58
UCR	13	7	2.83	CAS	59	15	33.93
KSPC	9	6	3.47	FMNH	1,023	123	52.49
WIRC	28	11	4.40	MCZ	172	28	54.24
CUAC	10	6	4.56	CNC	192	42	75.48
MSUC	52	11	6.70	**All Collections**	**4,926**	**464**	**100.00**

Within species distribution modeling literature, examples of sample bias include biases at the scale of the specimen: spatial, temporal, environmental, and taxonomic (Graham *et al*. 2004; Newbold 2010). For example, Kadmon *et al*. (2004) found that collection localities tended to be near roads. Data users that are not familiar with the practical patchiness of specimens and data from natural history collections may fail to recognize larger scale bias inherent to holdings of a particular collection, such as locality, age, size, and history.

A wide variety of reasonable models could have been created with the data—species distribution modeling is a growing, changing subject. The scale of variation among models due to "museum bias" was, in some cases, almost certainly larger than variation due to other possible model creation protocols. Future researchers should incorporate number of collections (in additional to number of specimens) into tests of model accuracy.

Exhaustiveness. The relationship between the number of collections surveyed (exhaustiveness) and the area of the distribution model is interesting because 1) *at least*… and 2) *only* 15 collections had to be surveyed before a reasonable model (90% reference model) of the distribution of *T. costalis* could be created. The high number of collections needed indicates the importance of maintaining many regional natural history collections—each collection added unique information to the model (although some information became redundant as more collections were added) and reduced overall museum bias.

The relatively low number of collections needed, 15 out of 38 that contributed data, indicates that, *for some species*, adequate large scale distributional data may already exist. However, locality data are available for nearly all specimens in collections. When surveying other attributes, such as pollen on specimens, gut content, or genetic variability, fewer specimens may contribute data. In that case, more collections may need to be surveyed before an adequate amount of information is obtained.

Ray's Rule of Precision: Measure with a micrometer. Mark with chalk. Cut with an axe. Global Biodiversity Information Facility (GBIF) is currently a major compilation of biodiversity data. The compilers claim to host the "biggest biodiversity database on the Internet," having compiled more than 500 million records over 1.5 million species and contributed to over 1,000 peer-reviewed publications (www.gbif.org/whatisgbif). Data from GBIF have been used in publications on endangered species conservation (*e.g*., Mota-Vargas and Rojas-Soto 2012) and the impact of climate change (*e.g*., Hof *et al*. 2012), both controversial subjects.

Despite criticism over data quality (Graham *et al*. 2004; Yesson *et al*. 2007; Beck *et al*. 2013), only one cautionary statement on *only one page* concerning data quality is provided by GBIF: "The quality and completeness of data cannot be guaranteed. Users

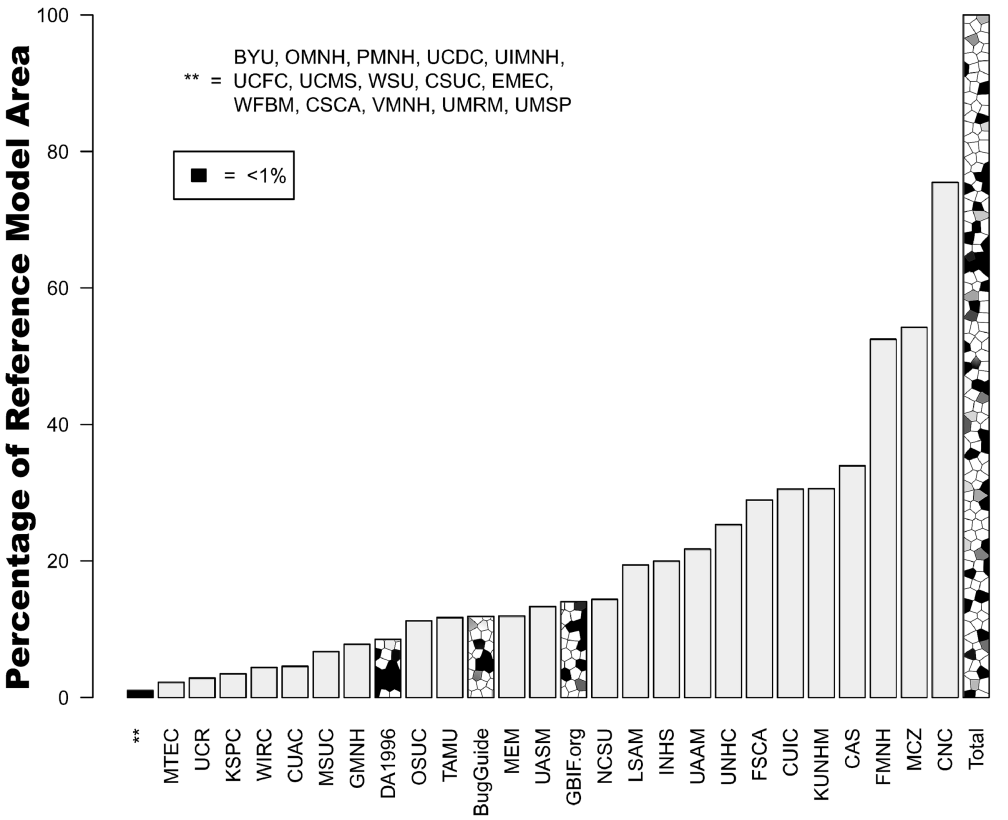


Fig. 9. Percentage of the reference area for each model created from individual collection data (see text for collections designated by codens). Bars for alternative data sources and total have tile fill. ** = collections with individually less than 1% of the reference model area.

employ these data at their own risk." (www.gbif.org/terms/licences/data-use). Two excellent resources (Chapman 2005a, b) concerning specimen-level data quality are hosted by GBIF, but both are more appropriate for data contributors than data users (www.gbif.org/resources/for-users). Otherwise, the site is largely self-promoting and lacks the skepticism and caution expected from a scientific resource. Other aggregator sites provide cautionary statements and even entire sections on their reliability. Bugguide.net states the following above *every* range map: "The information below is based on images submitted and identified by contributors. Range and date information may be incomplete, overinclusive, or just plain wrong." Wikipedia maintains a page on its reliability (en.wikipedia.org/wiki/Reliability_of_Wikipedia).

Online databases offer an opportunity for naïve or lethargic researchers to quickly produce poor quality research with little effort. Some researchers even justify immediate use of available data with the sentiment, "We cannot wait indefinitely for better information, but must use the knowledge that we already have." (Newbold 2010). For this research, it took less time to download GBIF data for *T. costalis* than it did to write an email requesting a loan from a museum, and it took less time to create the species distribution model than it took to confirm the identification of the requested specimens. Once a model is created, many measures exist to test its accuracy (*e.g.*, Liu *et al.* 2011 offer 30 accuracy measures), which, to naïve researchers or inattentive reviewers, may provide a false sense of quality control.

Computers allow for incredibly fast data manipulation, and many researchers may feel that they can, or are expected to, perform research at a proportional speed. Ecological studies involving tens, hundreds, or thousands of species would never be completed if every species were addressed at a critical level. For example, researchers estimating average change in distribution of North American insects due to climate change need never look at

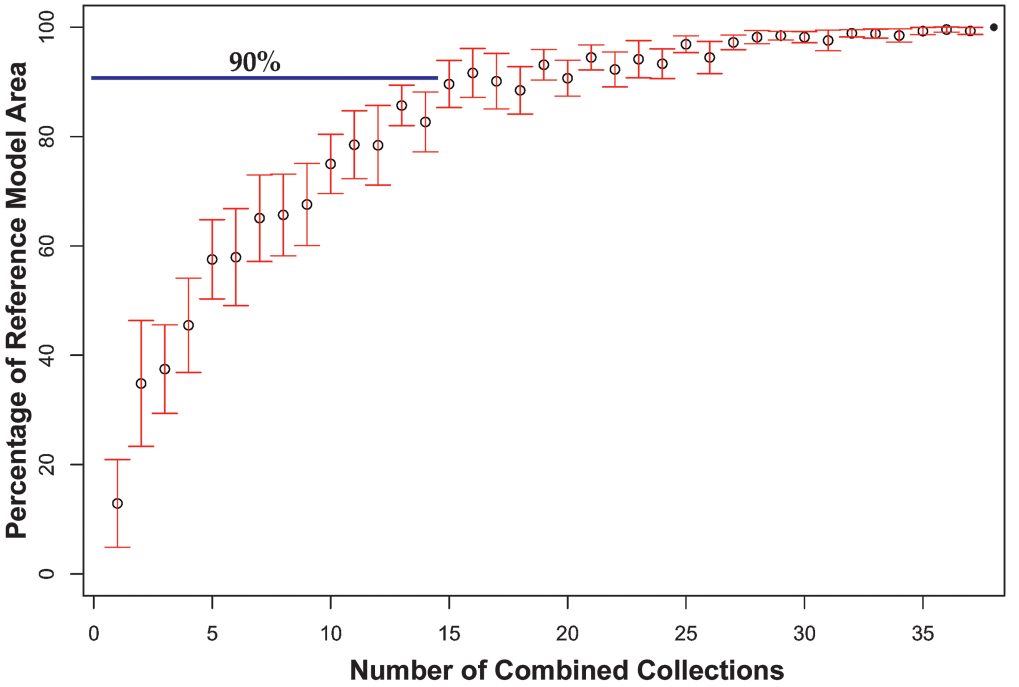


Fig. 10. Average and 95% confidence intervals of model areas for each set of combined collections. The reference model is represented by a black dot at 38 collections = 100% area.

the species list, just simply download data and run models. To be fair, checking data quality and exhaustiveness should not rest entirely on the shoulders of data users. To quote Soberón and Peterson (2004), "without a strong and active taxonomic community, BI [Biodiversity Informatics] will never be more than a clever set of software tools lacking a substantial factual basis."

The expectation of model creators and users that the data they are using will meet a minimum level of quality is not inappropriate. Our understanding of where a species is found will certainly change, and hopefully improve, over time. However, some minimum level of quality should be expected, otherwise research results will be worse than non-existent, they will be wrong. Research commenting on endangered species, conservation strategies, and climate change is especially important and especially controversial. Every effort should be made to assure the primary data used are of the utmost quality. If data quality and cleaning are the responsibility of the data custodian (Chapman 2005a, b), then providing a *measure* of data quality and exhaustiveness should be the responsibility of the data purveyor.

The "Digital" Museum: Beyond GBIF.org. Technological advances, especially the Internet, have spurred institutions to "digitize" holdings of natural history collections. In this sense, "digitize" refers to the creation of specimen-level databases that include traditional "label data" such as date, locality, and collector, and a taxonomic identifier, such as family, genus, or species. Sadly, parallel funding for confirmation of specimen identification and/or continued identification of specimens to lower taxonomic levels, particularly species, has not been included in the digitization zeitgeist. The culmination of "digitization" is websites such as GBIF.org that aggregate specimen data for use by researchers. Creation of these databases is time-consuming and costly and will not be completed any time soon (~1,500 years, Blagoderov *et al.* 2012).

In his famous essay, *The Tragedy of the Commons*, Garrett Hardin (1968) suggested the issue of overpopulation fell into a special class of human related dilemmas he called "no technical solution problems." His contention was that some problems faced by humanity could not be solved by technology, but only by changes in human behavior. Natural history collections face two dilemmas for which there is "no technical solution." Both problems are known but have not been properly classified, making discovery of a solution particularly difficult.

Problem 1. Recently, data, *sensu lato,* have become digital. Originally, data were married to a physical medium (a cave wall, a scroll, a building,

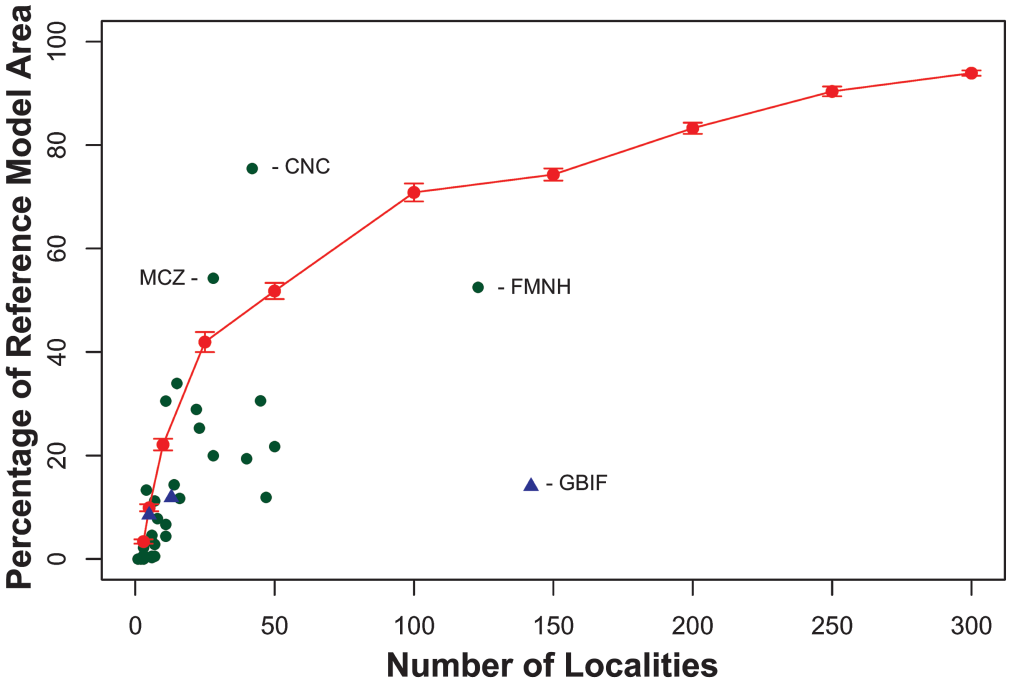


Fig. 11. Model area versus number of localities. Red circles = average and 95% confidence intervals of model areas for each set of random localities. Green circles = individual collections. Blue triangles = alternative distribution data.

a book, magnetic tape, film, *etc*.). Today, technological innovations have allowed information to be independent of a specific physical scaffolding (paper, film, *etc*.) and exist in a digital format that can be realized using general equipment (computers, monitors, speakers, *etc*.). Cost of resources (time, space, money, manpower, *etc*.) to store, reproduce, transport, acquire, *etc*. these materials have been reduced dramatically.

The first problem natural history collections face is the *expectation* that all types of data can be digitized. Currently, select aspects of natural history collections are being digitized: databases with specimen-level collection information, photographs with morphological information, micro-CT scans with internal and external morphology, audio recordings of vocalizations, genetic code, *etc*. Museums and researchers are profiting from the increased efficiency associated with digitizing this information. But can a natural history specimen, and therefore an entire collection, be completely digitized? Stated another way, can all there is to know about a given specimen be captured, can that information be stored in an electronic format, and the specimen (and eventually the entire collection) be discarded without discarding any information?

The answer is probably yes—no laws of physics would have to be broken to "record" all the information a specimen contains, but that is certainly well outside of our current or even near-future technological sophistication. The expectation that natural history collections can be digitized—the specimens discarded like old journals with no meaningful information lost—cannot be met. As such, natural history collections engage in at least three activities that must continue to physically take place (*i.e*., cannot be fully digitized): acquisition of specimens, retention of specimens, and lending of specimens (including collection visitation, which can provide valuable results, see Chatzimanolis 2014).

Non-digitizable data, other than natural history collections, include original artwork, original and historic documents, anthropological artifacts, architecture, and live cultures. If maintained, all of the above represent long-term, multigenerational resources that accrue in value over time. Interestingly, with the exception of natural history collections, few critics seriously advocate discarding the above items, even when suitable reproductions can be made, despite the fact that maintenance of originals can be costly.

Problem 2. Creation, maintenance, and growth of natural history collections is nearly always justified for utilitarian (often economic) reasons (*e.g*., Suarez and Tsutsui 2004). The argument is that

information from the collection: 1) is currently helping to solve human-related problems; or 2) may help solve unknown problems in the future. However, those arguments fail to recognize that discovery and exploration of the natural world is a part of the human condition, whether it has a utilitarian use or not. Historically, "explorers" and "discoverers" visited new geographic regions, but as the map was filled in, modern discoverers shifted their focus and began to explorer the universe at different scales of space and time or began to explore emergent properties of combined systems. Exploration, along with sports, music, and art, belongs to a class of human activities that represent "endeavors without end."

The second problem natural history collections face is the impression that building a natural history collection will come to an end, that the collection can be "finished." Natural history collections are the product of exploration at a smaller scale—discovering mites instead of mountains—and provide base material for the study of systems, such as ecology or mineralogy. The notion that growth of natural history collections will some day be "finished" is as nonsensical as the notion that the need to have children, make music, be entertained, or conduct scientific inquiry will be "finished." Policies such as strategic planning for the future and institutional goal-setting should recognize the "endeavor without end" qualities of natural history collections.

Conclusion

Thoracophorus costalis is not rare or difficult to collect. Rather, it is a widespread generalist that is commonly collected using a dozen techniques, is available throughout the year, is easily recognized by researchers, and is distributed throughout eastern North America (Ferro 2015), a well-surveyed region. Despite or because of this, specimens from at least 15 collections had to be surveyed before a reasonable distribution model (90% total reference) could be created. "Collection bias" greatly affected area of distribution of models, and in this case study number of collections was a better measure of exhaustiveness than number of specimens or localities. Therefore, maintenance and growth of numerous, regional natural history collections is important. Online databases are important resources but currently offer opportunities for researchers to unexpectedly obtain poor or incomplete data. Quality control and adequate warnings should be initiated or else inappropriate data usage by some could call into question research conducted by all. Growth and maintenance of natural history collections is an essential and enduring aspect of human endeavors.

Acknowledgments

We thank the curators, collection managers, and others that assisted in the loan of specimens. We thank Rachel Hawkins and Phil Perkins for help with "rediscovery" of the "lost" Mexican specimen of *T. costalis*. We thank Christopher Carlton and two anonymous reviewers for reviewing this manuscript. Partial funding for this research was made possible by donations from backers to the project "Lucid Key to Staphylinidae Subfamilies" posted at Experiment.com (DOI: 10.18258/0674).

References Cited

Beck, J., L. Ballesteros-Mejia, P. Nagel, and I. J. Kitching. 2013. Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? Diversity and Distributions 19: 1043–1050.

Blagoderov, V., I. J. Kitching, L. Livermore, T. J. Simonsen, and V. S. Smith. 2012. No specimen left behind: industrial scale digitization of natural history collections. Zookeys 209: 133–146.

BugGuide. 2015. Species *Thoracophorus costalis*. bugguide.net/node/view/52017/bgimage (accessed 25 January 2015).

Carroll, S. S., and D. L. Pearson. 1998. The effects of scale and sample size on the accuracy of spatial predictions of tiger beetle (Cicindelidae) species richness. Ecography 21: 401–414.

Chapman, A. D. 2005a. Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen, Denmark.

Chapman, A. D. 2005b. Principles and Methods of Data Cleaning — Primary Species and Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen, Denmark.

Chatzimanolis, S. 2014. Darwin's legacy to rove beetles (Coleoptera, Staphylinidae): A new genus and a new species, including materials collected on the Beagle's voyage. Zookeys 379: 29–41.

Cumming, G. S. 2000. Using between-model comparisons to fine-tune linear models of species ranges. Journal of Biogeography 27: 441–455.

Downie, N. M., and R. H. Arnett, Jr. 1996. The Beetles of Northeastern North America, 2 vols. The Sandhill Crane Press, Gainesville, FL.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40: 677–697.

Evenhuis, N. L. 2014. Abbreviations for insect and spider collections of the world. hbs.bishopmuseum.org/codens/codens-inst.html (accessed 30 September 2014).

Ferro, M. L. 2015. Review of the genus *Thoracophorus* (Coleoptera: Staphylinidae: Osoriinae) in North America north of Mexico, with a key to species. The Coleopterists Bulletin 69(1): 1–10.

GBIF. 2015. Occurrences of *Thoracophorus costalis* (Erichson, 1840). www.gbif.org/occurrence/search?taxon_key=4989463 (accessed 25 January 2015).

Graham, C. H., S. Ferrier, F. Huettman, C. Mortiz, and A. T. Peterson. 2004. New developments in museum-based informatics and applications in biodiversity analysis. TRENDS in Ecology and Evolution 19(9): 497–503.

Hardin, G. 1968. The tragedy of the commons. Science 162: 1243–1248.

Hernandez, P. A., C. H. Graham, L. L. Master, and D. L. Albert. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29: 773–785.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965–1978.

Hijmans, R. J., L. Guarino, and P. Mathur. 2012. DIVA-GIS Version 7.5 manual. www.diva-gis.org (accessed 15 January 2015).

Hof, A. R., R. Jansson, and C. Nilsson. 2012. Future climate change will favour non-specialist mammals in the (sub)arctics. PloS One 7(12): p.e52574. 1–11.

Hortal, J., J. M. Lobo, and A. Jiménez-Valverde. 2007. Limitations of biodiversity databases: case studies on seed-plant diversity in Tenerife, Canary Islands. Conservation Biology 21: 853–863.

Kadmon, R., O. Farber, and A. Danin. 2004. Effects of roadside bias on the accuracy of predictive maps produced by bioclimatic models. Ecological Applications 14(2): 401–413.

Liu, C., M. White, and G. Newell. 2011. Measuring and comparing the accuracy of species distribution models with presence-absence data. Ecography 34: 232–243.

Lujan, N. K., and L. M. Page. 2015. Libraries of life. New York Times 27 February 2015: A25.

Mota-Vargas, C., and O. R. Rojas-Soto. 2012. The importance of defining the geographic distribution of species for conservation: the case of the bearded wood-partridge. Journal for Nature Conservation 20(1): 10–17.

Newbold, T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. Progress in Physical Geography 34(1): 3–22.

NPS. 2006. NPS Museum Handbook. Parts I, II, and III. www.nps.gov/museum/publications/handbook.html (accessed on 21 December 2014).

NSTC. 2009. Scientific Collections: Mission-Critical Infrastructure of Federal Science Agencies. National Science and Technology Council, Committee on Science, Interagency Working Group on Scientific Collections. Office of Science and Technology Policy, Washington, DC.

R Core Team. 2014. R: A Language and Environment for Statistical Computing. www.R-project.org (accessed 21 December 2014).

Rasband, W. S. 2014 [1997–2014]. ImageJ. imagej.nih.gov/ij/ (accessed 15 January 2015).

Rocha, L. A., A. Aleixo, G. Allen, F. Almeda, C. C. Baldwin, M. V. L. Barclay, J. M. Bates, A. M. Bauer, F. Benzoni, C. M. Berns, M. L. Berumen, D. C. Blackburn, S. Blum, F. Bolaños, R. C. K. Bowie, R. Britz, R. M. Brown, C. D. Cadena, K. Carpenter, L. M. Ceríaco, P. Chakrabarty, G. Chaves, J. H. Choat, K. D. Clements, B. B. Collette, A. Collins, J. Coyne, J. Cracraft, T. Daniel, M. R. de Carvalho, K. de Queiroz, F. Di Dario, R. Drewes, J. P. Dumbacher, A. Engilis Jr., M. V. Erdmann, W. Eschmeyer, C. R. Feldman, B. L. Fisher, J. Fjeldså, P. W. Fritsch, J. Fuchs, A. Getahun, A. Gill, M. Gomon, T. Gosliner, G. R. Graves, C. E. Griswold, R. Guralnick, K. Hartel, K. M. Helgen, H. Ho, D. T. Iskandar, T. Iwamoto, Z. Jaafar, H. F. James, D. Johnson, D. Kavanaugh, N. Knowlton, E. Lacey, H. K. Larson, P. Last, J. M. Leis, H. Lessios, J. Liebherr, M. Lowman, D. L. Mahler, V. Mamonekene, K. Matsuura, G. C. Mayer, H. Mays Jr., J. McCosker, R. W. McDiarmid, J. McGuire, M. J. Miller, R. Mooi, R. D. Mooi, C. Moritz, P. Myers, M. W. Nachman, R. A. Nussbaum, D. Ó Foighil, L. R. Parenti, J. F. Parham, E. Paul, G. Paulay, J. Pérez-Emán, A. Pérez-Matus, S. Poe, J. Pogonoski, D. L. Rabosky, J. E. Randall, J. D. Reimer, D. R. Robertson, M.-O. Rödel, M. T. Rodrigues, P. Roopnarine, L. Rüber, M. J. Ryan, F. Sheldon, G. Shinohara, A. Short, W. B. Simison, W. F. Smith-Vaniz, V. G. Springer, M. Stiassny, J. G. Tello, C. W. Thompson, T. Trnski, P. Tucker, T. Valqui, M. Vecchione, E. Verheyen, P. C. Wainwright, T. A. Wheeler, W. T. White, K. Will, J. T. Williams, G. Williams, E. O. Wilson, K. Winker, R. Winterbottom, and C. C. Witt. 2014. Specimen collection: an essential tool. Science 344(6186): 814–815.

Soberón, J., and A. T. Peterson. 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London B 359: 689–698.

Stockwell, D. R. B., and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. Ecological Modeling 148: 1–13.

Suarez, A. V., and N. D. Tsutsui. 2004. The value of museum collections for research and society. BioScience 54(1): 66–74.

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. Science 240(4857): 1285–1293.

Yesson, C., P. W. Brewer, T. Sutton, N. Caithness, J. S. Pahwa, M. Burgess, W. A. Gray, R. J. White, A. C. Jones, F. A. Bisby, and A. Culham. 2007. How global is the Global Biodiversity Information Facility? PLoS One 11: e1124. 1–10.

(Received 14 April 2015; accepted 28 July 2015. Publication date 18 September 2015.)