The power, potential, and pitfalls of open access biodiversity data in range size assessments: Lessons from the fishes

Jennifer A. Smitha,⁎, Abigail L. Bensonb, Ye Chenb, Steffany A. Yamadac, Meryl C. Mimsa

a Department of Biological Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
b U.S. Geological Survey, Science Analytics and Synthesis, Denver, CO, USA

ABSTRACT

Geographic rarity is a driver of a species’ intrinsic risk of extinction. It encompasses multiple key components including range size, which is one of the most commonly measured estimates of geographic rarity. Range size estimates are often used to prioritize conservation efforts when there are multiple candidate species, because data for other components of rarity such as population size are sparse, or do not exist for species of interest. Range size estimates can provide rankings of species vulnerability to changing environments or threats, identifying rare species for future study or conservation initiatives. However, range sizes can be estimated by several different metrics, and the degree of overlap in the identification of the rarest or most common species across methodologies is not well understood. This knowledge gap compromises our ability to prioritize correctly rare species, and presents a particularly difficult challenge for stream-dwelling organisms with distributions constrained to river networks. We evaluated the relationship of multiple range size estimates of a subset of freshwater fishes native to the United States to determine the degree of overlap in rarity rankings using different data sources and grain sizes. We used publicly available, open access data from the Global Biodiversity Information Facility (GBIF) to calculate extent of occurrence (minimum convex polygons) and area of occupancy (total area occupied, measured across various grain sizes). We compared range sizes estimated using GBIF data with the best available estimates of current distributions described by publicly available digital maps (NatureServe) to evaluate the efficacy of GBIF data in assessments of range size. We found strong correlations between range size estimates across analytical approaches and data sources with no detectable bias of taxonomy. We found that variation among rarity rankings was highest for species with intermediate range sizes indicating that the approaches considered here generally converge when used to identify the rarest or the most common species. Importantly, our results show that the rarest, and perhaps the most vulnerable, species are consistently identified across common methodological approaches. More broadly, our results support the use of open access biodiversity data that include opportunistically collated and collected point occurrence records as a complement to coarse-grain (e.g., whole range map) approaches, as we observed no systematic bias or deviation across data sources in our analyses. This indicates databases such as the GBIF may help fill important fundamental and applied knowledge gaps for many poorly understood species, particularly in a broad-scale, multispecies framework.

1. Introduction

Vulnerability of a species to extinction can be described by three key attributes: exposure to a threat or stressor, intrinsic sensitivity to the stressor, and adaptive capacity (e.g., the ability to move, or adapt in place) (Foden et al., 2013). Intrinsic sensitivity underlies the degree to which a species is at risk due to global change. Geographic rarity (i.e., spatial extent of occupancy) can influence a species’ intrinsic and extrinsic risk to extinction, particularly in the context of other intrinsic factors (e.g., low population density, diet, reproductive rates; Gaston, 1994, Purvis et al., 2000; Pritt and Frimpong, 2010). Geographic rarity is informed by range size (i.e., the geographical area occupied by a species) with wide ranging species being less geographically rare than those with smaller ranges. Range size estimates are often more readily available than estimates of other factors underlying extinction risk (e.g., demographic trends through time, adaptive capacity), and thus are...
frequently used to prioritize conservation efforts (Carter et al., 2000; Ceballos et al., 2005). However, range size is sensitive to the spatial scale (grain and extent) at which it is measured (Hartley and Kunin, 2003; Mims et al., 2018), and thus the approach used to assess range size may affect conservation decisions (Jotz et al., 2008). What is not known is whether range size estimates derived from different descriptors of rarity (e.g., different range size metrics) are comparable and accurate. Furthermore, how do data source, taxonomy, or geographic rarity (e.g., common versus rare species) influence the comparability of range size estimates across approaches? As range size metrics continue to be applied to conservation criteria and decision-making worldwide, these questions represent a significant knowledge gap in our ability to confidently compare range size estimates derived from different data sources and approaches, and thus make informed conservation decisions.

Two of the most common range size metrics used to estimate geographic rarity of species are extent of occurrence (EOO) and area of occupancy (AOO; Hartley and Kunin 2003). EOOs are among the most coarse-grain approaches and are often defined using minimum convex polygons (MCP), considered as the smallest polygon that encloses all occurrence points for a given species (i.e., the locations at which a species has been recorded). They are simple to construct and allow for quick estimation of range size. However, MCPs are heavily influenced by outlying occurrence points and are likely to contain large unoccupied areas within their bounds. Therefore, this coarse metric can overestimate a species range (Burgman and Fox, 2003), particularly if it is characterized by a patchy distribution (Rabinowitz, 1981). AOOs are typically calculated by summing the total occupied area on the landscape using a gridded or tessellated grid approach, in which areas of grid cells are summed if the target species has been observed as present (Kunin, 1998; Hartley and Kunin, 2003). A modification of this approach is to create buffered occurrence points of a given radius (or width), merge buffered areas for the target species, and calculate the total area (sensu Mims et al., 2018). AOOs reduce the likelihood of including unoccupied areas, but may underestimate range sizes by excluding unsampled, but occupied areas (Rondinini et al., 2006), or excluding occupied areas where detectability may be problematic. The grain size of grids or buffered-occurrence points influences the range size estimates (Hartley and Kunin, 2003). Smaller grain sizes more closely approximate the number of occurrences and may be more sensitive to uneven sampling efforts; large grain sizes are less precise, but are likely less sensitive to uneven sampling efforts.

Point occurrence data, often used for AOO range size metrics, are collected either systematically or opportunistically. Systematic sampling (here defined as sampling based on a study design) supports more rigorous statistical analysis and hypothesis testing. However, systematic sampling may be prohibitively expensive, and data collected systematically are often available for select locations, resulting in biased inferences of species distribution (Rondinini et al., 2006). As an alternative, opportunistically collected point occurrence data can be used to provide reliable range size estimates (e.g., Sullivan et al., 2009; Clark, 2017). However, opportunistic data may be geographically or temporally biased, or particularly sensitive to imperfect detection (Rondinini et al., 2006; Dickinson et al., 2010). Both opportunistic and systematic data are increasingly integrated into publicly available and open access biodiversity databases (e.g., eBird, iNaturalist, the Global Biodiversity Information Facility) with concomitant use to address ecological questions (Sullivan et al., 2009; Clark, 2017). Thus, understanding how opportunistic and systematic data affect range size estimates is becoming increasingly important (Ficetola et al., 2014).

Certain taxa with specific habitat requirements or patchy distributions, such as many aquatic species, may exhibit high variation among range size metrics. For example, range sizes of fish estimated using MCPs may incorporate large areas of terrestrial habitat, resulting in inflated range size estimates (McGrath and Austin, 2009). Defining range sizes using watersheds may offer a more refined approach (Bertuzzo et al., 2009; Matthews and Marsh-Matthews, 2015; Januchowski-Hartley et al., 2016), but may still result in overestimated range sizes if the resolution of the watershed is coarse (e.g., watersheds categorized at the sub-basin level defined by US Geological Survey (USGS) 8-digit Hydrologic Unit Codes [HUC-8]) (Frimpong et al., 2016), or if distributions are patchy or discontinuous. Alternative solutions include describing range sizes using stream-reaches (Paul and Post, 2001; DeWeber and Wagner, 2015), biologically relevant patches (Dunham et al., 2002), or using the summed area within buffers centered on occurrence points (Mims et al., 2018), all of which likely reflect more accurately the areas occupied by fish (i.e., the river networks) than ranges described by watersheds. Given the potential for high variability across range sizes estimated using different metrics for fish, evaluations of range size comparability are especially pertinent.

Here, we address these knowledge gaps by evaluating the performance of different range size metrics for a subset of native freshwater fishes in the contiguous United States (US) representing taxonomic and geographic diversity. Specifically, our first goal was to assess the comparability of range sizes constructed using a suite of range size metrics, and the effects of geographic rarity (e.g., common versus rare species) and taxonomy on such relationships. Our second goal was to evaluate the comparability and accuracy of range sizes constructed using publicly available point occurrence records from the Global Biodiversity Information Facility (GBIF), and range sizes reflecting the best available estimates of current distributions described by publicly available digital distribution maps (NatureServe 2010).

2. Materials and methods

2.1. Selection of study species

Our overall aim was to select > 150 species to form an initial species selection that represented taxonomic and geographic diversity of freshwater fishes native to the contiguous US (lower 48 states) (Fig. 1A, Fig. A1). First, we identified candidate species by considering all species in the native freshwater fish database from Mims et al. (2010) (J.D. Olden, University of Washington, unpublished data). We excluded most species identified in the USGS Nonindigenous Aquatic Species (NAS) database to minimize the influence of artificially expanded range sizes due to invasions or stocking events outside of their native ranges. We retained some species represented in NAS to maintain taxonomic or geographic representation. Second, we attempted to maximize the number of families and genera included in the study, with proportional representation of large families and genera. Third, we aimed to include species with small to large range sizes, as estimated initially by IchthyMaps (Frimpong et al., 2015). IchthyMaps is a database containing approximately 600,000 point occurrence records from 1083 species in the US that were collected between 1950 and 1980, and it includes the number of watersheds occupied by each species. We used the count of HUC-8 watersheds (hereafter ‘HUC-8 counts’) as a rough estimate of range size for species with ≥ 10 occurrence point records in IchthyMaps (Appendix A, Fig. A2). Finally, we adjusted the initial species selection to ensure representation of regional species richness for native freshwater fish species selected.

2.2. Species point occurrence records and digital distribution maps

The GBIF is an international network and research infrastructure supported by governmental organizations worldwide that provides open access to a biodiversity database (Edwards, 2004). It integrates data collected through both systematic and opportunistic efforts from multiple sources including citizen science, museum collections, and state and federal level agencies (Wheeler 2004). The digital distribution maps (hereafter ‘NatureServe maps’) considered in our study reflect the best available estimates of current distributions of native freshwater fishes in the US by HUC-8 watersheds (i.e., watersheds categorized at
taking a multi-step process to ensure that we removed point occurrence records with missing attributes (i.e., year, longitude, and latitude), those with spurious dates and/or mismatches between coordinates and country of origin, and those located outside of the US. Species for which there were < 50 point occurrence records were then removed from the dataset. Next, we clipped point occurrence records with a spatially explicit polygon of the lower 48 states to remove records located outside of the contiguous US. We removed point occurrence records located within estuaries by clipping the data with spatially explicit estuary shapefiles obtained via the Environment Protection Agency’s (EPA) Estuary Data Mapper (EPA, 2017). At this stage, we mapped occurrence point records and distributed the resulting maps to expert reviewers selected based on their regional expertise pertaining to native fish distributions. The aim of the expert review was not to complete a rigorous quantitative assessment. Rather, we used it to identify if at least some of the occurrence point records fell outside of native ranges, and thus if we needed an additional filtering step. This review process identified occurrence points outside of native ranges. Subsequently, we removed non-native occurrences by filtering for point occurrence records that fell outside of native ranges (defined by NatureServe maps). All spatially explicit data were projected using Albers Equal Conic Projection and analysis completed in Program R (R Core Team, 2016).

In some cases, we relaxed selection criteria to include focal species that represented taxonomically unique groups, or those that occurred in regions for which these selection criteria resulted in disproportionately low species representation. This was largely a concern for the southwestern US, in which many native species either had < 50 point occurrence records, or have been introduced outside their native HUCs, and thus did not meet the selection criteria for being included in the final species dataset.

2.3. Taxonomic and geographic representation of the final species dataset

We visually compared range sizes (i.e., small to large) calculated using IcthyMaps point occurrence records (Frimpong et al., 2016) for species in our final dataset with those for species represented in Mims et al. (2010) (n = 708) using scatterplots. This allowed us to determine if the range sizes represented in our final species dataset were representative of native freshwater fishes in the lower 48 states as a whole. We then compared the geographic distribution of species represented in our final species dataset with those of a broad suite of native freshwater species for which NatureServe maps were available (n = 867; NatureServe 2010) taking a multi-step process to ensure that our final dataset provided geographic representation of species found in the lower 48 states. First, we delineated six geographic regions based on information gleaned from the literature regarding species richness of freshwater fishes (Sheldon, 1988; Warren and Burr, 1994) as follows: 1) the region west of the Continental Divide (hereafter ‘West Region’), 2) the Mississippi River drainage (hereafter ‘Mississippi Region’), 3) the southeastern region encompassing the watersheds of the Alabama and Chattahoochee Rivers (hereafter ‘SE Region’), 4) the region east of the Eastern Continental Divide excluding the Mississippi River drainage (hereafter ‘Atlantic Region’), 5) the area encompassing the watersheds for the Rio Grande and Brazos Rivers (hereafter ‘Texas-Gulf Region’), and 6) the area encompassing the Great Lakes (hereafter ‘Great Lakes Region’). Second, for each of the six geographic regions, we summarized the number of species in our final dataset and NatureServe maps that were present. We then determined the proportional representation of species for each geographic region. Finally, we compared the proportional representation of family groups (i.e., number of species per family) in our final dataset with that in the dataset from Mims et al. (2010). All spatially explicit data were projected using Albers Equal Conic Projection and analysis completed in ArcMap v.10.6.1 (ESRI, California, US).

2.4. Species range size calculations

We estimated range sizes using seven range size metrics applied to point occurrence records from GBIF. For each species, we calculated an MCP by creating a polygon that joined the outer most point occurrence records and that encapsulated the remaining point occurrence records, and six AOO range size metrics. To estimate AOOs, we used a modified version of a grid-based approach following Mims et al. (2018). In brief, we summed the area within circular buffers centered on point occurrence records. Overlapping buffers were merged such that heavily sampled areas did not artificially inflate AOO estimates. Given that grain size, or in this case buffer radius, can influence estimates of AOO (Hartley and Kunin, 2003), we evaluated multiple buffer sizes (buffer radii: 1 km, 5 km, 10 km, and 20 km) to calculate four estimates of AOO based on this modified gridded approach. We also considered AOO by watershed, summing the total area within occupied watersheds at both the HUC-8 and HUC-12 scale (i.e., sub-basin and sub-watershed scale, respectively) using available watershed boundary datasets (USGS, 2015). Additionally, we estimated range sizes for each species using NatureServe maps (NatureServe 2010). Because NatureServe maps reflect the best available estimates of current species distributions, these estimates likely provided the most accurate description of range sizes of those species considered in our study. All spatial data were projected using Albers Equal Conic Projection and analysis conducted using the packages ‘rgeos’ v.0.3–28 (Bivand et al., 2018a) and ‘adehabitatHR’ v.0.4.15 (Calenge, 2017) in R v.1.1.453 (R Core Team, 2016).

2.5. Comparing range size estimates

We compared range size estimates in two key ways. First, we evaluated collinearity between range sizes estimates using Spearman’s Rank Correlation in package ‘psych’ v.1.8.4 (Revelle, 2018) in R v.1.1.453 (R Core Team, 2016); p-values were adjusted using a Bonferroni correction to account for multiple tests and we considered relationships significant if p < 0.002. Second, we ranked species according to range size (1 = rarest), allowing a direct comparison across range size estimates. We then calculated an average rank (hereafter ‘geographic rarity ranking’) and standard deviation (SD) across the ranks for each species.

We evaluated the relationship between geographic rarity ranking and SD to determine whether discordance between rarity rankings is correlated with range size (e.g., is discordance higher for the rarest or the most common species). To do this, we fitted a suite of linear regression models using the lm() function in R v.1.1.453 (R Core Team, 2016). We also evaluated whether geographic rarity ranking or SD varied significantly by taxonomy (family group) by using a Kruskal-Wallis test and considered there to be a significant difference between families when p < 0.05. We considered family groups with > 5 species and conducted analyses in R v.1.1.453 (R Core Team, 2016).

3. Results

Our final dataset following our initial data selection process and data filtering process (Fig. 1) consisted of 128 species representing a broad range of taxonomic and geographic diversity (Appendix D) and included species from 29 families, of which Cyprinidae accounted for the most species (36), followed by Percidae (14), and Catostomidae (13) (Fig. 2A). Taxonomic representation was similar between the final dataset and the dataset in Mims et al. (2010) (Fig. 2A). Our final dataset also included species with a wide range of geographic extents, as described by IcthyMaps point occurrence records, which were
representative of geographic extents explained by IcthyMap point occurrence records of species represented in Mims et al. (2010) with small to large range sizes (Appendix A, Fig. A2). In contrast, species represented in Mims et al. (2010) that had larger range sizes (as described by IcthyMaps point occurrence records) were not represented in our final dataset. This is likely because our study focused on species not included in the NAS dataset (with some exceptions) to assess putative native ranges and avoid species with extensive introductions, whereas those in Mims et al. (2010) included many species within the NAS dataset that had generally larger ranges, in some cases due to, or confounded by introductions outside of their native ranges. All geographic regions in the US were reasonably well represented among species in our final dataset (Fig. 2B).

4. Range size estimates and variability among and within species

Range sizes estimated using GBIF point occurrence records varied across the seven range size metrics considered (Table 1). On average, range sizes were largest for those estimated using MCPs (718,820 km²) and smallest for those estimated using 1 km buffered points (1062 km²). Range sizes estimated using GBIF point occurrence records were significantly correlated across all seven metrics considered (all pairwise comparisons: p < 0.0001, Table 2). EOO range sizes defined using MCPs tended to be larger than those estimated by other metrics and GBIF point occurrence records (EOOs represented the largest range size for 91 of 128 species). However, the strength of this relationship depended upon the average range size of a species; range size estimates explained by EOOs were smaller than those explained by other metrics for 28 of the 30 species with the smallest average range sizes. For example, the EOO for Gambusia heterochir, the species with the most restricted range size in our dataset, was smaller than the AOO from the 1 km buffer. In comparison, EOO range size estimates were consistently larger than all other metrics for the 30 species with the largest average range sizes.

Range size estimates constructed using GBIF point occurrence records were significantly correlated with range sizes explained by NatureServe maps (all pairwise comparisons: p < 0.0001, Table 2, Fig. 3). On average, range sizes estimated using NatureServe maps were larger than those estimated using GBIF point occurrence records, except when MCPs were used (NatureServe: 402,188 km²; MCPs: 718,820 km², Table 1). However, range sizes calculated using MCPs were not larger than those estimated with NatureServe maps for all species. Rather, range sizes calculated from NatureServe maps were larger for 43/128 species represented in our final dataset.

Range sizes explained by HUC-8 watersheds and GBIF point occurrence records were between 14.87% and 120.81% (mean = 65.25%) the size of those estimated using NatureServe maps (described by HUC-8 watersheds) (Fig. 3B). This can be explained by
poor alignment of the HUC-8 watersheds used to estimate range sizes using GBIF point occurrence records and those used to construct NatureServe maps within the spatially explicit layers. Further, NatureServe maps are truncated by the US border, and thus HUC-8 watersheds shown in NatureServe maps that overlap the US border are reduced in size. In contrast, our range size estimates described by HUC-8 watersheds and GBIF point occurrence records considered the total area of all HUC-8 watersheds occupied within the contiguous US, even if they crossed the US border. We decided not to crop HUC-8 watersheds by the US border because we wanted HUC-8 based range sizes to accurately reflect the true areas delineated by watersheds. In so doing, our HUC-8 range sizes provide a relevant measure to others (e.g., natural resource managers) considering conservation at the HUC-8 scale and a more ecologically driven AOO measure rather than one determined by geopolitical boundaries. Where range sizes estimated by HUC-8 watersheds exceeded those described by NatureServe maps, this departure from the typical relationship was often explained by HUC-8 watersheds transcending the US border (e.g., *Astyanax mexicanus*, *Pteronotropis signipinnis*). Where range sizes explained by HUC-8 watersheds were substantially smaller than those explained by NatureServe maps, inspection of raw GBIF point occurrence records (i.e., records pre-data filtering) suggested that GBIF data underrepresented the distributions for at least some species considered in this study (e.g., *Atractosteus spatula*, *Ictalurus lupus*, *Lota lota*).

5. Range size variability within species and associations with geographic rarity and taxonomy

The relationship between variation among range sizes, described by SD of range size rankings and geographic rarity rankings was explained
by a significant quadratic function ($p < 0.001$, $r^2 = 0.30$); SD of range size rankings was highest for species with intermediate geographic rarity rankings, with geographic rarity rankings converging for the rarest and most common species (Fig. 4). Species with relatively high geographic rarity rankings and low SD tended to be characterized by large, relatively contiguous and non-patchy distributions with high point occurrence record counts (e.g., *Aphrophodon sayanus* [n = 4671], *Esox americanus* [n = 3752], *Pimephales vigilax* [n = 5237]). In contrast, species with relatively low geographic rarity rankings and low SD tended to be characterized by small, linear distributions with low point occurrence record counts (e.g., *Catostomus fumeiventris* [n = 50], *Creiichthys baileyi* [n = 176], *Notropis cahabae* [n = 59]). Species with intermediate geographic rarity rankings and high SD tended to be characterized by medium range sizes with relatively low point occurrence records (e.g., *Cyclopeus elongatus* [n = 94], *Esox masquinongy* [n = 63], *Percaius omiscomaycus* [n = 70]). Four notable outliers were characterized by intermediate geographic rarities and very high SDs (Fig. 4), three of which resulted from large, but patchy distributions, described by low point occurrence record counts (e.g., *Agonostomus monticola* [n = 34], *Lota lota* [n = 40], and *Pungitius pungitius* [n = 28]). In contrast, the very high SD associated with *Lythrurus bellus* was driven by a small, but contiguous distribution described by high point occurrence record counts (n = 2280). Under this scenario, range size rankings were relatively high for very small grain size metrics, with relative range decreasing as grain size increases (1 km buffer = 115, 20 km buffer = 87, MCP = 47). Geographic rarity rankings and SD of rarity rankings were not significantly different across family groups (Catostomidae, Centrarchidae, Cyprinidae, Ictaluridae, Percidae; mean: χ^2 = 7.59, p = 0.11, df = 4; SD: χ^2 = 4.44, p = 0.35, df = 4, Fig. 5).

### Table 1

<table>
<thead>
<tr>
<th>Range Size Metric</th>
<th>Mean (km²)</th>
<th>SE (km²)</th>
<th>Smallest range size (km²)</th>
<th>Largest range size (km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCP 1km</td>
<td>718,820</td>
<td>91,092</td>
<td>7</td>
<td>5,873,911</td>
</tr>
<tr>
<td>5km</td>
<td>1062</td>
<td>152</td>
<td>11</td>
<td>10,258</td>
</tr>
<tr>
<td>10km</td>
<td>18,433</td>
<td>2562</td>
<td>146</td>
<td>167,496</td>
</tr>
<tr>
<td>20km</td>
<td>51,029</td>
<td>6703</td>
<td>443</td>
<td>400,728</td>
</tr>
<tr>
<td>HUC-8</td>
<td>119,378</td>
<td>18,433</td>
<td>146</td>
<td>167,496</td>
</tr>
<tr>
<td>MCP 10km</td>
<td>402,188</td>
<td>44,051</td>
<td>6,014</td>
<td>2,996,608</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Range Size Metric</th>
<th>Range Size Metric</th>
<th>MCP 1km</th>
<th>5km</th>
<th>10km</th>
<th>20km</th>
<th>HUC-8</th>
<th>HUC-12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1km</td>
<td>0.58</td>
<td>&lt; 0.0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5km</td>
<td>0.67</td>
<td>&lt; 0.0001</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10km</td>
<td>0.75</td>
<td>&lt; 0.0001</td>
<td>0.96</td>
<td></td>
<td>0.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20km</td>
<td>0.83</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUC-8</td>
<td>0.92</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>HUC-12</td>
<td>0.96</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>NatureServe</td>
<td>0.96</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

We found strong correlations between range size estimates across analytical approaches and data sources with no detectable bias of taxonomy. We also found that variation (SD) among rankings of range sizes estimated using publicly available point occurrence records was greatest for species with intermediate range sizes and lowest for species with the smallest and largest range sizes, indicating that range size rankings for metrics considered here are more similar (i.e., they converge in size) for the geographically rarest or the most common species. Specifically, our results show that the rarest, and perhaps the most vulnerable species are consistently identified across common analytical approaches. More broadly, we found evidence that the use of publicly available databases containing both opportunistically and systematically collated and collected point occurrence records complement coarse-grain (e.g., whole range map) approaches, as we observed strong correlations between, and thus no systematic bias across range sizes estimated using different data sources (i.e., GBIF data and NatureServe maps). While our results demonstrate that point occurrence records from publicly available databases often underestimate, and on occasion over estimate absolute areas occupied by focal species, our method
provides evidence that the use of rankings offers a robust approach in comparative assessments of range sizes. Importantly, this indicates databases such as the GBIF may help fill important fundamental and applied knowledge gaps for many poorly understood species, particularly in a broad-scale, multispecies framework.

Our results demonstrate that range sizes estimated using GBIF point occurrence records correlate with range sizes described by NatureServe maps (i.e., best available estimates of current distributions), highlighting the efficacy of publicly available databases to provide insight into the distribution of native freshwater fishes in the contiguous US. Therefore, given the fine-grained nature of GBIF data, our results suggest that publicly available point occurrence records have the potential for use as an alternative to coarse-grained range maps in ecological assessments of species, and/or to complement existing efforts incorporating coarse-grained approaches. For example, point occurrence records from publicly available databases could be used to elucidate species-habitat relationships and predict species distribution models at fine-scale resolutions (Torres et al., 2012; Abolafya et al., 2013; Smith et al., 2017), or to assess temporal changes in species distributions (Jiguet et al., 2012; Ferrer-Paris et al., 2014; Clark, 2017). However, before taking such approaches we recommend that potential biases associated with the use of publicly available data should be considered and explored (Beck et al., 2013, 2014). Given the increasing concerns of global change on biodiversity (e.g., due to climate change and habitat loss), we also encourage exploration of the efficacy of these data in species sensitivity assessments (Mims et al., 2018).

In support of previous studies, our results also showcase that range sizes estimated using GBIF point occurrence records correlate with range sizes described by NatureServe maps (i.e., best available estimates of current distributions), highlighting the efficacy of publicly available databases to provide insight into the distribution of native freshwater fishes in the contiguous US. Therefore, given the fine-grained nature of GBIF data, our results suggest that publicly available point occurrence records have the potential for use as an alternative to coarse-grained range maps in ecological assessments of species, and/or to complement existing efforts incorporating coarse-grained approaches. For example, point occurrence records from publicly available databases could be used to elucidate species-habitat relationships and predict species distribution models at fine-scale resolutions (Torres et al., 2012; Abolafya et al., 2013; Smith et al., 2017), or to assess temporal changes in species distributions (Jiguet et al., 2012; Ferrer-Paris et al., 2014; Clark, 2017). However, before taking such approaches we recommend that potential biases associated with the use of publicly available data should be considered and explored (Beck et al., 2013, 2014). Given the increasing concerns of global change on biodiversity (e.g., due to climate change and habitat loss), we also encourage exploration of the efficacy of these data in species sensitivity assessments (Mims et al., 2018).

In support of previous studies, our results also showcase that range sizes (km²) are sensitive to the scale at which they are measured (Hartley and Kunin, 2003). Intuitively, as grain size increased for buffered points (i.e., buffer radius), range sizes increased. Similarly, range sizes described by HUC-8 watersheds were consistently larger than those described by HUC-12 watersheds (i.e., sub-basin and sub-
watershed scale, respectively). Interestingly, MCPs often described the smallest range size of species with relatively small average range sizes, but this pattern deteriorated as average range size increased; for species with relatively large average range sizes, MCPs consistently described the largest range size. Because MCPs can include large areas of unoccupied habitat, especially for species distributed in linear networks (i.e., rivers and streams), our results suggest that MCPs likely overestimated range sizes, especially for wide ranging species considered in our analysis. Our results support those of others who have demonstrated that bias in range sizes explained by MCPs increase as sample size and spatial distribution of occurrence points increase (Burgman and Fox, 2003; Mota-Vargas and Rojas-Soto, 2012).

Range size rankings were also sensitive to the scale considered. For example, variation among range size rankings (SD) was smallest for the rarest and most common species (i.e., those with the smallest and largest geographic rarity ranking, respectively) and largest for species with intermediate geographic rarity rankings. These results suggest that the rarest, and perhaps the most vulnerable species are consistently identified across common methodological approaches, and that consideration of the spatial scale at which range sizes are calculated may be more important for species with intermediate range sizes. We did not detect any systematic bias of taxonomy on geographic rarity ranking, or on variation among range size rankings, suggesting that the sensitivity of range size to the scale at which it is measured is not taxonomically defined. This demonstrates that our approach and results have broad relevance across all taxonomic groups.

We aimed to provide insight into the value of publicly available data in multi-species assessments. Given the demonstrated ability of GBIF data to provide range sizes that are directly comparable to those described by NatureServe maps, we argue that publicly available point occurrence data from GBIF offer a robust approach in such assessments. We also acknowledge alternative publicly available point occurrence databases including Biodiversity Information Serving Our Nation (BISON), Biodata, and Multistate Aquatic Resources Information System (MARIS) (USGS). However, due to the fact that GBIF has global contributions from a broader list of data providers it likely provides the most up to date and complete set of point occurrence data. Regardless of the database considered, we acknowledge the potential for imperfect detection and spatial bias in publicly available point occurrence data due to incomplete and uneven sampling across species’ ranges (Beck et al., 2014) to bias range size estimates.

This study advances our understanding of the relationships between range sizes measured using different grain sizes both within, and across species with different geographic rarities. It also demonstrates the efficacy of using publicly available data in assessments of range sizes of freshwater fishes, indicating the value of using publicly available data for making management decisions and informing conservation strategies. Future work could consider exploring systematic bias in GBIF data that considers spatial biases in sampling efforts, detectability, conservation status, and geographic regions occupied by focal species. In so doing, more confidence may be placed in the ability to make informed management decisions based on future assessments that consider such data in a multispecies framework.

Declaration of Competing Interest

There is no conflict of interest.

Acknowledgements

We thank the U.S. Geological Survey (USGS) for funding. For data acquisition, we are grateful to S. Aulenbach (USGS), D. Ignizio (USGS), D. Wesley, D. Wieferich (USGS), A. McKerrow (USGS), & B. Young (NatureServe). We are also indebted to the GBIF network for making data available for public acquisition, to C. Moore for assistance with graphics, and to A. Carbajal for helpful edits. Finally, we thank J. Dunham, M. Freeman, K. Gido, and T. Hitt for QA/QC of our dataset. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. R scripts for this work can be found on Github-https://github.com/MimsLabVT/National-Fishes-Vulnerability-Assessment-Project.

Funding

This work was supported by the United States Geological Survey Cooperative Agreement G17AC00235 with Virginia Polytechnic Institute and State University.
Appendix B.: References for point occurrence data downloaded from the Global Biodiversity Information Facility between 4th and 18th October 2017

GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.9dsa5r
GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.minus7
GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.3ezruu
GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.3s8qym
GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.uituvb
GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.yebehc
GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.1zhzsh
GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.xriybs
GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.zvrcil
GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.lng6xw
GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.l1xvny
GBIF.org (4th October 2017) GBIF Occurrence Download https://doi.org/10.15468/dl.haiaau

Fig. A1. IcthyMap record count (x-axis, log-transformed) and GBIF record count (post-filtering; y-axis, log-transformed) for n = 128 final species. Adjusted r-squared = 0.37 (p < 0.001).

Fig. A2. IcthyMap record count (x-axis, log-transformed) and IcthyMap HUC-8 counts (y-axis, log-transformed) for n = 128 species in this study (red) and all species with > 10 records in IcthyMaps (grey). Adjusted r-squared are 0.81 (p < 0.001) and 0.79 (p < 0.001) for species included and not included, respectively.

Appendix B.: References for point occurrence data downloaded from the Global Biodiversity Information Facility between 4th and 18th October 2017
Appendix C. Step-by-step procedure for filtering GBIF point occurrence data

Point occurrence data were downloaded for species represented in our initial species list via the Global Biodiversity Information Facility (GBIF) online database (https://www.gbif.org/) on the 4th October 2017 (Appendix B). Each species was represented by a .csv file named with a unique string of numbers. Data consisted of 44 attributes; Each .csv file was renamed to reflect the species name in attribute column 'species';

For each species, we deleted records that were missing data for latitude, longitude, or year. Specifically, we removed records if data was missing for one or more of the following attribute columns: ‘decimallatitude’, ‘decimallongitude’, and ‘year’;

We removed records with spurious dates and/or mismatches between coordinates and country of origin. Specifically, we removed records if the following terms were represented in the attribute column ‘issue’: ‘COUNTRY_COORDINATE_MISMATCH’, ‘IDENTIFIED_DATE_UNLIKELY’, ‘RECORDED_DATE_MISMATCH’, ‘ZERO_COORDINATE’;

We removed records with data missing from the following attributes: ‘gbifid’, ‘countrycode’, i.e. records for which ‘gbifid’ and/or ‘countrycode’ were populated with ‘na’;

We removed all records that fell outside of the US. Specially, we retained records for which the attribute ‘countrycode’ was populated by ‘US’; Species with < 50 records were removed, apart from Catostomus insignis which was retained to maintain geographic representation of the southwestern US;

We removed records located outside of the lower 48 states by clipping the data with a spatially explicit shapefile of the lower 48 states obtained through package ‘maps’ version 3.3.0 (Becker et al., 2018) in Program R (R Core Team, 2016);

We removed records located within estuaries by clipping the data with a spatially explicit shapefile of estuaries in the US. Estuary shapefiles were obtained via the Environment Protection Agency’s (EPA) Estuary Data Mapper, downloaded on 19th January 2018 (EPA, 2017);

Simultaneously, we removed records that fell outside of native ranges by clipping the data with publicly available digital distribution maps describing species native distributions at the HUC-8 watershed level. Digital distribution maps were obtained from NatureServe on the 17th July 2018 (http://www.natureserve.org/conservation-tools/data-maps-tools/digital-distribution-native-us-fishes-watershed);

All spatially explicit data were projected using Albers Equal Conic Projection and analyses were completed using the packages ‘sf’ (Pebesma et al., 2018), ‘rgdal’ version 1.3-6 (Bivand et al., 2018), ‘mapdata’ version 2.3.0 (Becker et al. 2018a), ‘maps’ version 3.30 (Becker et al. 2018b), ‘maptools’ version 0.9-4 (Bivand et al., 2018), and ‘mapproj’ version 1.2.6 (McIlroy et al. 2018) in Program R (R Core Team, 2016). R script explaining the filtering process of GBIF data can be found at the following URL: https://github.com/MimsLabVT/National-Fishes-Vulnerability-Assessment-Project/blob/master/GBIF%20data%20filtering%20and%20AOO%20estimation_11_13_2018

References

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2019.105896.