Check for updates

esa

ECOSPHERE

FRESHWATER ECOLOGY

Assessing temporal biases across aggregated historical spatial data: a case study of North Carolina's freshwater fishes

Kyra Sigler,[1,2] Dan Warren,[3] Bryn Tracy,[1] Elisabeth Forrestel,[4] Gabriela Hogue,[1] and Alex Dornburg[5,†]

[1]*North Carolina Museum of Natural Sciences, Raleigh, North Carolina 27601 USA*
[2]*Department of Biological and Agricultural Engineering, North Carolina State University, Raleigh, North Carolina 27695 USA*
[3]*Biodiversity and Biocomplexity Unit, Okinawa Institute of Science and Technology, Okinawa, Japan*
[4]*Department of Viticulture and Enology, University of California, Davis, California 95616 USA*
[5]*Department of Bioinformatics and Genomics, University of North Carolina Charlotte, Charlotte, North Carolina 28223 USA*

Citation: Sigler, K., D. Warren, B. Tracy, E. Forrestel, G. Hogue, and A. Dornburg. 2021. Assessing temporal biases across aggregated historical spatial data: a case study of North Carolina's freshwater fishes. Ecosphere 12(12):e03878. 10.1002/ecs2.3878

Abstract. Historical records from museums or government agencies are of tremendous utility for illuminating the factors that shape the spatial distribution of the planet's biodiversity. However, these data were often collected under heterogeneous and opportunistic sampling designs and therefore likely contain significant sampling biases that change over time. Understanding historical biases is particularly important for aquatic vertebrates, where no studies of changes in sampling effort have yet been conducted. Here, we use a dataset of 276,138 records that span all freshwater fishes known to occur in North Carolina as a case study from which to highlight major shifts in collection trends that cause sample biases in datasets that aggregate historical records. We found evidence for three distinct phases of sampling over the last two centuries: (1) early sampling in the late 19th and early 20th century that was largely dominated by the research interests of a few "mega-collectors"; (2) a mid-20th century shift toward more widespread sampling; and (3) a major surge of sampling that corresponds to the rise of major environmental movements. We find each period possesses distinct phylogenetic and spatial biases. Moreover, these phases mirror trends in other spatial datasets that aggregate historical records spanning plants to terrestrial vertebrates, thereby suggesting that historical contingency and a temporal bias toward recent records are likely hallmarks of compiled historical datasets. Given that the pace of spatial data sampling continues to grow, our results strongly caution that the continued development of new models and methods to mitigate against bias-driven statistical artifacts will be critical to effectively harnessing the power of historical data.

Key words: biodiversity survey; camera trap surveillance; environmental DNA; environmental niche modeling; museum records.

Received 1 April 2021; revised 30 July 2021; accepted 2 August 2021. Corresponding Editor: Debra P. C. Peters.

Copyright: © 2021 The Authors. *Ecosphere* published by Wiley Periodicals LLC on behalf of The Ecological Society of America. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

† **E-mail** : adornbur@uncc.edu

Introduction

The 21st century has given rise to historically unparalleled access to data on the distribution of the planet's biodiversity (La Salle et al. 2016; Soltis and Soltis 2016; Soltis et al. 2018). Such data often represent the culmination of hundreds of years of sampling efforts (Griffiths et al. 1999,

Blagoderov and Smith 2012, Stropp et al. 2016). In turn, analysis of this spatial data forms the basis for numerous conservation and management decisions (Tolley et al. 2016), as well as predictions for how species or communities will respond to ecological changes (Daufresne and Boët 2007, Jetz et al. 2012, Franklin et al. 2017). While these datasets have incredible statistical value, the information contained within is not necessarily uniformly or systematically sampled over time (Hortal et al. 2007, Dornburg et al. 2017*b*). Heterogeneity in the timing of sample collection (Bulluck et al. 2006, Boakes et al. 2010, Yang et al. 2013, Tessarolo et al. 2017), collection techniques (Patton et al. 1998), target species (Dallas et al. 2017), areas sampled (Botts et al. 2011, Hickisch et al. 2019), and other aspects of data collection remain a specter that haunts meta-analyses. Such heterogeneity in sampling can lead to spurious or inaccurate interpretations of shifts in ecological communities that would compromise species distribution maps (Hortal et al. 2008), and/or mislead analyses of suitable habitat and predicted responses to environmental change (Bulluck et al. 2006, Engemann et al. 2015, Daru et al. 2018, Monsarrat et al. 2019*b*). Therefore, assessing aggregated spatial data for underlying changes in bias through time is a critical step in the conservation and management of species (Romo et al. 2006).

With a level of species richness that exceeds half of all living vertebrates (Near et al. 2012*b*), understanding the distribution of the world's fishes has long been a major challenge. Over the past century, numerous logistical problems associated with sampling, including accessibility (Willis and Babcock 2000, Koenig and Stallings 2015), as well as biases in the efficiency and detection rates of different monitoring equipment (Kjelson and Colby 1977, Beamesderfer and Rieman 1988, Dornburg et al. 2017*b*), have had to be overcome. These efforts have resulted in an unprecedented picture of the spatial distribution of fishes that has been fundamental for studies of niche modeling (Bond et al. 2011, Ruaro et al. 2019, McMahan et al. 2020), biogeography (Dornburg et al. 2015, 2017*a*, Cowman et al. 2017, Siqueira et al. 2019), and population genetics (Rocha 2003, Echelle et al. 2015, Healey et al. 2018, Warren et al. 2021) to name but a few. However, the potential that underlying biases persist in this aggregated data remains. In particular, advances in infrastructure and technology, as well as increased investment in the monitoring of natural resources, have not been uniform over time. This raises the question of what time frame should be considered a baseline for each unique ecological study. Given the growing wealth of accessible data on the distributions of fishes, we now have the opportunity to assess the extent to which such sampling biases have accrued in our collective knowledge. In turn, this provides an important baseline for future sampling and the management of our aquatic resources.

The freshwater fishes of North Carolina provide an exemplar case study for considering biases in spatial and temporal distributional data of fishes. Spanning 21 river basins across 4 physiographic regions, North Carolina's 100 counties have been continuously sampled since the late 1800s (Tracy et al. 2020). Tracy et al. (2020) recently compiled over 276 thousand records stemming from sampling conducted by multiple state agencies and museum-vouchered specimens and used them to redefine the distribution of the state's freshwater fishes. While such a dataset is a tremendous resource that can result in new discoveries, the extent to which the amalgamation of these records reflects shifting historical sampling biases remains unclear. Early efforts to survey fish did so without the aid of a robust transportation infrastructure. This lack of accessibility posed logistical constraints and limitations on sampling effort that remains unquantified, yet previous work suggests that this would produce bias (Monsarrat et al. 2019*a*). Additionally, sampling focus likely varied through time, from early naturalists interested in specific taxa and/or areas to more recent attention on game and select non-game species. Such changes in sampling could yield cryptic phylogenetic sampling biases that in turn could mislead estimates of functional diversity that are crucial to conservation and management. Therefore, assessing whether there is significant temporal or phylogenetic heterogeneity in the species representing these data points can be of high value to future ecological modeling efforts in the region and provide a framework for assessing bias in other collections.

Here, we analyze the distributional records of North Carolina's freshwater fishes to identify, quantify, and visualize biases in the temporal

and phylogenetic distribution of spatial samples. We begin by analyzing the temporal changes in sampling records to assess differences in sampling frequency through time, thereby determining the timescale of baseline historical data. Next, we evaluate taxonomic patterns of sampling to ascertain how sampling efforts have been distributed across major clades of fishes. Finally, we use generalized linear models (GLMs) to test if sampling efforts reflect time spent in areas with higher concentrations of people, a commonly invoked source of sampling bias (Millar et al. 2019), or simply areas with more waterways. We assess whether these relationships with sampling effort are constant or change through time. These analyses provide essential baseline quantifications of bias necessary to more accurately forecast the impacts of rapid urbanization predicted over the next several decades in this region of the southeastern United States (Van Metre et al. 2019). More broadly, our results also provide expectations of similar biases that can exist in datasets composed of aggregated historical records and an analytical framework for assessing bias in biological collection and occurrence data.

Methods

Data acquisition

We used the data from Tracy et al. (2020) that contain 276,139 records across all species of freshwater fishes known to occur in North Carolina, including all obligatory freshwater species, as well as records of species that that occur in both freshwater and saltwater (e.g., anadromous, catadromous, and freshwater tolerant brackish species). This dataset is comprised of distributional data from the North Carolina Division of Water Resources, the North Carolina Wildlife Resources Commission (NCWRC) Portal Access to Wildlife Systems (www.ncpaws.org), the Tennessee Valley Authority, the North Carolina Division of Marine Fisheries Programs (Nos. 100, 115, 123, 127, 146, 150, and 915), and data from vouchered lots housed at the Academy of Natural Sciences of Drexel University, American Museum of Natural History, Auburn University Museum, California Academy of Sciences, Cornell University Museum of Vertebrates, Field Museum of Chicago, Florida Museum of Natural History, Harvard Museum of Comparative Zoology, Illinois Natural History Survey, National Museum of Natural History, North Carolina Division of Parks and Recreation, North Carolina Museum of Natural Sciences, Ohio State University Museum, Roanoke College Ichthyological Collection, Royal Ontario Museum, Tulane University, University of Alabama Ichthyological Collection, University of Kansas, University of Michigan Museum of Zoology, University of North Carolina at Wilmington, University of Tennessee, Virginia Institute of Marine Sciences, and the Yale Peabody Museum of Ichthyology.

Quantifying temporal and taxonomic heterogeneity

We first heuristically determined which species are most represented in the composite data by assessing the relative proportion of records belonging to single species. We classified species as "game," "nongame," and "introduced" based on the classifications of the NCWRC (http://www.eregulations.com/northcarolina/hunting-fishing/inland-fishing-regulations/). It should be noted that this classification includes Redbreast Sunfish (*Lepomis auritus*) as both introduced and native, as this species has been widely introduced outside of its native range throughout the western region of the state. As such, we follow this classification with representation of this species in both categories. These data were visualized using a Treemap for each classification category, displaying proportionally scaled nested boxes that represent the number of samples within a category (e.g., "game fish"). Treemaps were generated using Tableau Public 2020.2 (Tableau Software and LLC 2020).

To assess patterns of temporal and taxonomic heterogeneity, we extracted year and month data and used these data entries to standardize date formatting within the Tracy et al. (2020) dataset. We then assessed the slope of relationship between records retained and time by plotting the accumulation of species-level records through time. This is akin to "lineage through time" plots commonly used in evolutionary biology (Nee et al. 1994, Near et al. 2012*a*) and allowed us to determine if records accumulated gradually or if there were pulses of accelerated or decelerated sampling efforts at specific time points. Breakpoints in the slope of the species

records through time were computed using the r-package breakpointR (Porubsky et al. 2020) using Bayesian information criterion (Neath and Cavanaugh 2012) to assess the fit of breakpoints. We next assigned species to the taxonomic level of "family" and used CollVizDashboards (Mailhot 2020) in conjunction with Tableau Public 2020.2 (Tableau Software and LLC 2020) to visualize the scope of family-level collections over time. As heterogeneity in family-level sampling may mask phylogenetic bias at deeper scales, we included representation of the ordinal level of each family in this visualization to ensure congruence between taxonomic levels. This visualization complements the records through time plot, facilitating assessments of how uniform sampling efforts were between major clades of fishes.

To assess phylogenetic bias, we downloaded the species-level time-calibrated phylogeny of all ray-finned fishes from Rabosky et al. (2018) from the fish Tree of Life online resource (Chang et al. 2019). This tree topology was then pruned using Geiger 2.0 (Pennell et al. 2014) in R to include only species present in North Carolina and subjected to a clade significance test of node-based phylogenetic clustering (Forrestel et al. 2014, 2015). This test is similar to the nodesig test in PhyloCom that determines whether taxa are over-represented in an assemblage by means of a randomization test. However, the clade significance test also incorporates the density of sampling within clades and their descendents to determine clades with disproportionate sampling through a randomization of tip states and a one-tailed test of significance ($P < 0.05$). Tests were conducted in R and visualized using the packages APE (Paradis et al. 2004) and phytools (Revell 2012).

Testing for potential spatial biases through time

We used the coordinate data in the Tracy et al. (2020) dataset to assign each record to a county. To accomplish this, we used the county boundaries shape file (https://www.nconemap.gov/datasets/NCDOT::ncdot-county-boundaries) and custom R scripts to determine the location of each point. Although Tracy et al. (2020) did include data on the county of collection for some records, this step was necessary as the county-level designations within this dataset were incomplete. If collections were made on a county line, such as a river that divides two counties, these entries were split into two identical entries, one for each county. This mirrors the designation of records in multiple counties used by Tracy et al. (2020) whose data defined these as "countyA–countyB." Likewise, if a collection was made along a state line (i.e., NC-TN, NC-SC, NC-VA), the entry was split into separate entries and the one for the non-NC state record was excluded. Sampling efforts were visualized using a choropleth county map of NC using Tableau Public 2020.2 (Tableau Software and LLC 2020) and CollVizDashboards (Mailhot 2020) to provide an overview of spatial sampling efforts.

To test if any hypothesized factors bias spatial sampling efforts, we tested the fit of several candidate parameters using generalized linear models with county-level samples as the dependent variable: average population growth, county size, the proportion of water, and coastal vs. non-coastal counties. County size was quantified in R using the North Carolina Department of Transportation County Boundaries shape file (https://www.nconemap.gov/datasets/NCDOT::ncdot-county-boundaries). Census data by county per decade (1970–2010) were acquired from the North Carolina Budget and Management, Population and Housing Information (https://linc.osbm.nc.gov/pages/population-housing/). As preliminary analyses indicated that population density was highly correlated with population growth ($r^2 = 0.977$), we only included average population growth in our analyses. To quantify the proportion of water per county, we obtained shapefiles of major water bodies and rivers from NC Onemap (https://www.nconemap.gov/datasets/major-hydrography-waterbodies/data?page=) and the North Carolina Department of Environmental Quality (http://data-ncdenr.opendata.arcgis.com/datasets/surface-water-classifications). We extracted the area of each water body by county using custom R scripts. As many rivers fluctuate in width, these did not contain area data. As such, we used 5 m to represent average stream and river widths and used this to quantify area. To assess how much uncertainty in river area impacted analyses, we repeated all analyses using 1 and 10 m as average stream and river widths. We compared the fit of each parameter using Akaike information criterion (AIC) (Akaike 1998), allowing us to

assess model fit while penalizing for extra parameters. AIC differences of >4 were used to determine improvement of model fit (Burnham and Anderson 2007). AIC differences were additionally used to quantify evidence ratios, providing an alternative quantification of the explanatory power of the best-fit model over other models in the candidate pool (Wagenmakers and Farrell 2004). This approach allowed us to test whether sampling efforts are more proportionate to water area than potential sources of bias such as population or county size or location. Analyses were repeated between time major temporal periods identified using the approaches above.

Results

Taxonomic and temporal heterogeneity

A few species dominate the number of spatial records. Among game species, Striped Bass (*Morone saxatilis*) and White Perch (*Morone america*) account for a combined total of 44.5% of all records (Fig. 1A). This is followed by White Catfish (*Ameiurus catus*; 7.6%) and Bluegill (*Lepomis macrochirus*; 7%) and a few other species that collectively represent 78.8% of gamefish records (Fig. 1A). This pattern of uneven representation between species is also reflected in records of nongame fish, with common estuarine/lower river-dwelling species accounting for over one third of all records (Fig. 1B). Records of introduced species are likewise dominated by few species (Fig. 1C). Placing these records into a temporal context reveals an acceleration of record accumulation through time that is evident across nearly all ordinal or family-level groupings (Fig. 2A) and also reflected in the long fuse and subsequent steep rise of the species-level records through time plot (Fig. 2B). The best-fit breakpoint model provides strong support (ΔBIC 6.9407) for two breakpoints at 1949 and 1976, suggesting three phases of different sampling intensity to underlie the accumulation of these spatial records. Prior to 1949, large temporal gaps between sampling events that span a decade or more are common (Fig. 2A), and sampling efforts are generally characterized by small numbers of records within taxonomic families. Following 1949, temporal sampling intervals became more evenly distributed within taxonomic families, with a general increase of sampling intensity beginning in the late 1970s. Sampling intensity also varied between taxonomic families, such as Leuciscidae, Clupeidae, and Centrarchidae being particularly prominent in sampling efforts towards the end of the 20th century and into the 21st century (Fig. 2A).

Placing the uneven density of sampling between families into a phylogenetic context revealed significant clade-specific clustering of sampling that varied through time (Fig. 3). Prior to 1949, sampling efforts were most concentrated across leuciscine minnows, esoscids (Pikes), four subclades of darters (Percidae), and a clade of campostids (Fig. 3A). During the middle of the 20th century and prior to 1976, sampling shifted notably toward centrarchids (Sunfishes), clupeids, Striped Bass, and White Perch (*Morone*; Fig. 3B), trending away from high-intensity sampling across all leuciscine minnows. Through the present day, significantly higher intensity sampling remains supported for eight clades, including clupeids, *Morone*, Chain Pickerel, *Esox niger,* and Redfin Pickerel, *E. americanus* (Fig. 3C), that were also found significantly sampled at higher rates in the prior time slice (Fig. 3B). Sampling among *Ictalurus* spp. (Channel Catfish and Blue Catfish), a clade of *Nocomis* (chubs), and Central Stoneroller, *Campostoma anomalum* also significantly increased while sampling in centrarchids, a subclade of *Etheostoma* (darters), and *Rhinichthys* (daces) decreased relative to these other clades (Fig. 3C).

Geographic biases

Visualization of the number of records by county reveals that the eastern counties have higher numbers of records than most other counties (Fig. 4A) and that concentrations of records are generally not in the same counties as concentrations of people (Fig. 4B). This mismatch between human population density and spatial records was also supported by our GLM analysis. We found no strong correlation between population and records ($P = 0.06$). Instead, records were higher in coastal counties ($P < 0.004$), a result that is highly correlated with surface water area. This result is consistent whether assuming an average river width of 3 m ($P < 0.0077$), 15 m ($P < 0.0076$), or 30 m ($P < 0.0076$). Comparing the AIC score of a model based on population vs. water area reveals an AIC difference of over 62,

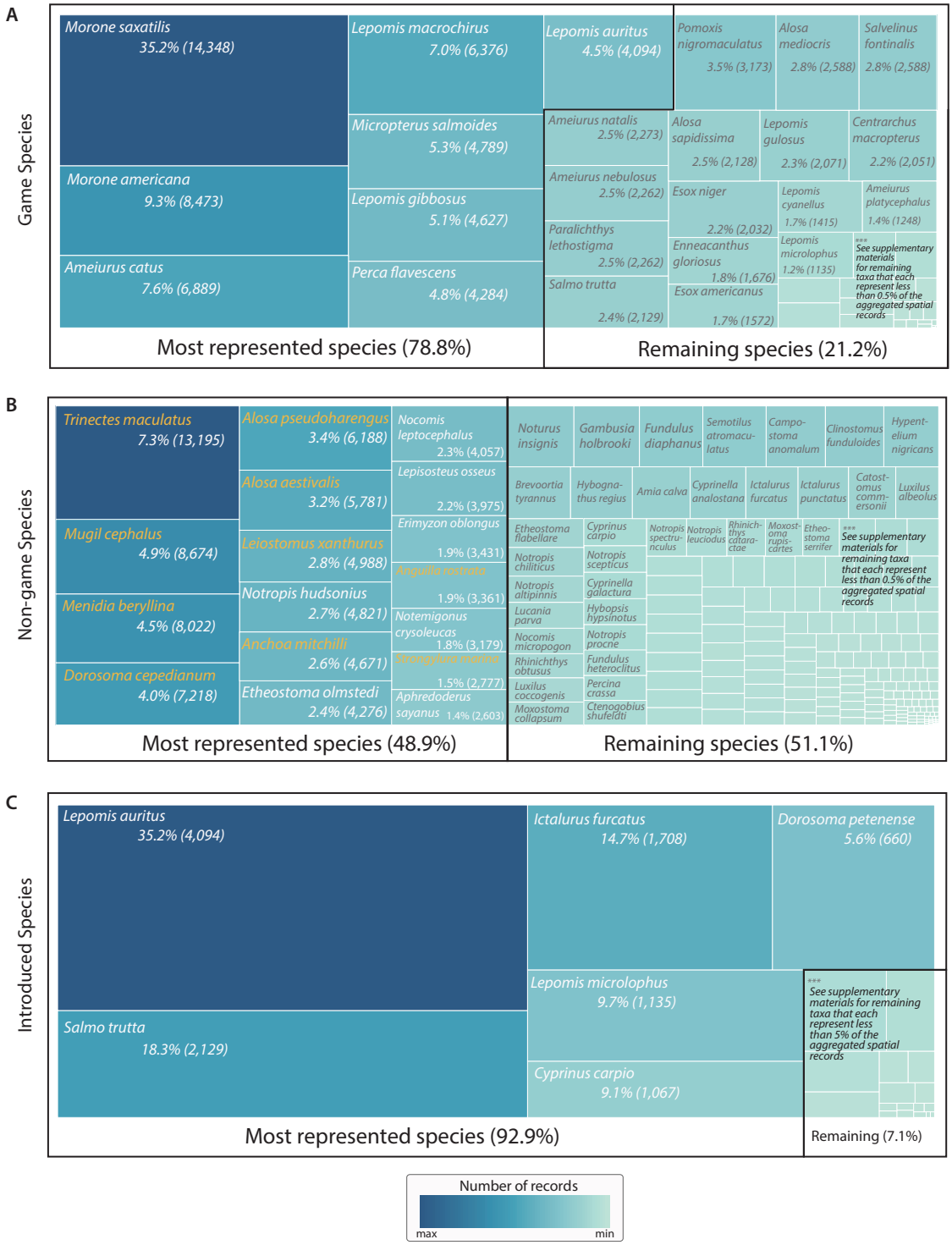

Fig. 1. Number of records per species of game (A), non-game (B), and introduced species (C). Insets indicate predominant (White text) species and all other remaining species (gray) for each category. Background shading and box size corresponds to the number of records. Orange text in B indicates coastal species that are largely estuarine mentioned in the text.

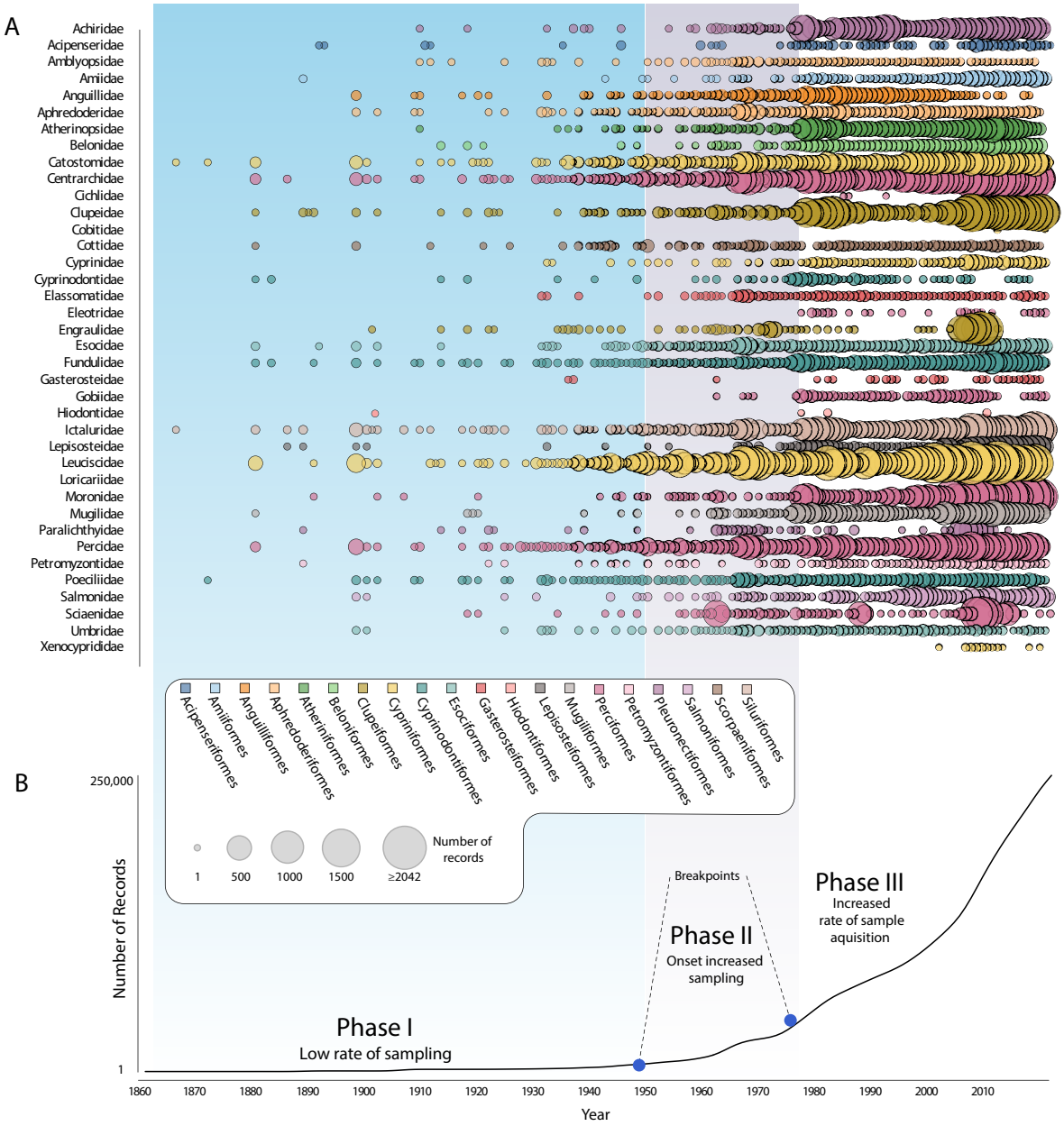

Fig. 2. Changes in rate and intensity of family-level sampling (A) and species-level records through time (B). Size of the circles in A is proportional to the number of samples and colors correspond to the ordinal-level designations indicated in the insert. Background shading corresponds to the identified phases of sampling identified by the breakpoint analysis.

which corresponds to an evidence ratio of surface water area being over 72 billion times more likely to explain spatial records in North Carolina than population density. However, placing the geography of sampling into a temporal context reveals that the dominance of coastal sampling was not a hallmark of early sampling in the state (Fig. 5A). Instead early records prior to

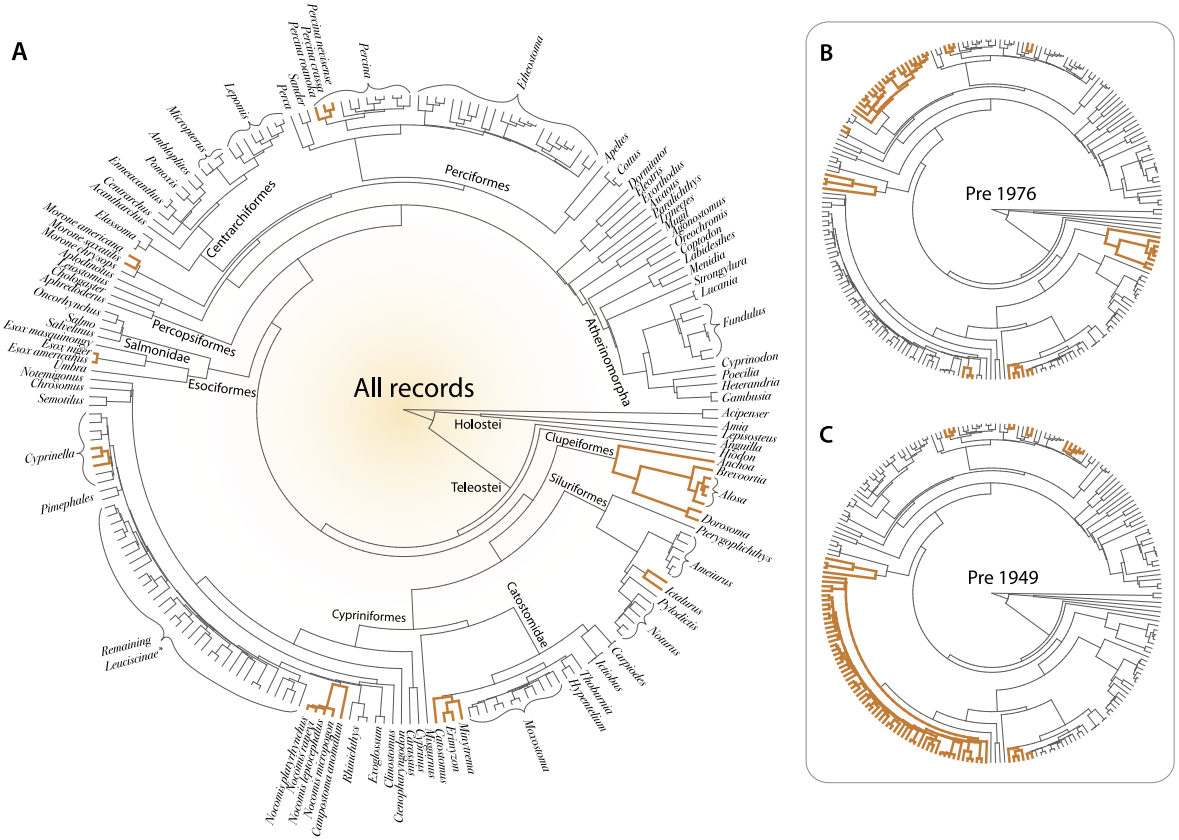

Fig. 3. Comparison of phylogenetic clustering of cumulative records prior to 1949 (A), prior to 1976 (B), and prior to 2020 (C). Orange shading indicates clades with significant clustering of records. Major clades are labeled along internal branches following Dornburg and Near (2021). *Remaining Leuciscinae: *Notropis, Hybognathus, Erimonax, Luxilus, Hybopsis, Lythrurus, Erimystax, Phenacobius*.

1950 are largely concentrated in the western mountainous regions (Fig. 5A), with a dramatic shift in sampling in eastern counties beginning in the 1950s (Fig. 5B) that continues to the present day (Fig. 5C).

Discussion

Our quantification of the temporal distribution of historical records demonstrates several significant patterns of taxonomic, phylogenetic, and spatial bias that have changed over time. Although a handful of taxa that do not represent the widest ranging species account for a disproportionate share of the total records (e.g., Striped Bass, Hogchoker, and Inland Silverside) this taxonomic bias is not static through time, instead shifting through three temporal intervals that are categorized by different rates of sample acquisition. These three intervals also correspond with changes in phylogenetic bias, with the prominence of clades such as leuciscine minnows prior to the 1950s giving way to an increased focus on other taxa such as Clupeiformes (families Clupeidae and Engraulidae) and *Ictalurus* catfish toward the present day. These temporal changes in sampling are also associated with changes in geographic sampling with a distinct shift in emphasis from an early concentration on the Western mountainous regions of the state to higher levels of samples from coastal counties in the latter portion of the 20th century. By the 21st century, there is a significant reduction in overall biases, reflecting a global trend of attempting to reduce sampling biases in wildlife monitoring. A similar heterogeneous history of sampling

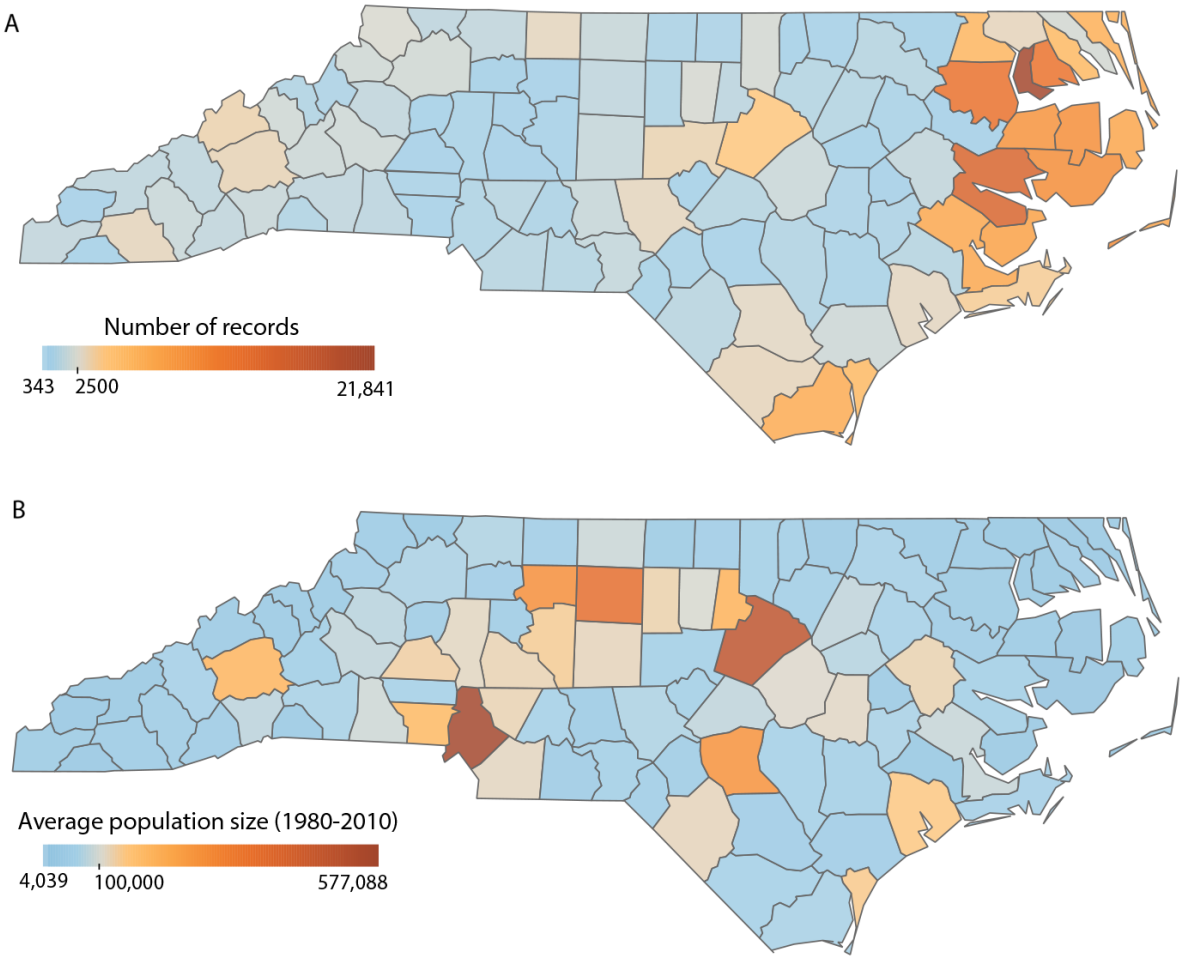

Fig. 4. Total number of records (A) and average population size (B) per county.

designs and shifting scientific interests is also prevalent in aggregate datasets of terrestrial species (Hortal et al. 2008, Escribano et al. 2016, Haque et al. 2018). As such, our results caution that the patterns we observed may represent not only general biases in aggregated historical spatial data of the freshwater fishes in North Carolina, but likely many other composite historical datasets.

Considering bias in historical data

Historical species records are invaluable for understanding changes or trends in contemporary distributional patterns (Shaffer et al. 1998, Monsarrat et al. 2019*b*), as well as for focusing future conservation efforts. However, effectively harnessing the power of historical data requires assessing potential sources of bias that could mislead statistical analyses (Daru et al. 2018, Gippoliti and Groves 2018). Our survey of North Carolina's available freshwater fish records reveals substantial shifts in historical sampling that correspond to major transitions in the history of freshwater ichthyology in the region as well as the United States at large. Early surveys of North Carolina's fishes were most intensive in western regions of the state (Fig. 5A) corresponding largely with surveys and work by naturalists such as Cope (1870*a*, *b*), Jordan (1889), and Joseph Bailey (Menhinick et al. 1974). This bias of records toward specific collectors is not unique to North Carolina. Thousands of historical records from the 1800s across the America's stem from similar naturalist surveys (e.g., Baird and Girard 1853; Cope 1871, 1875,

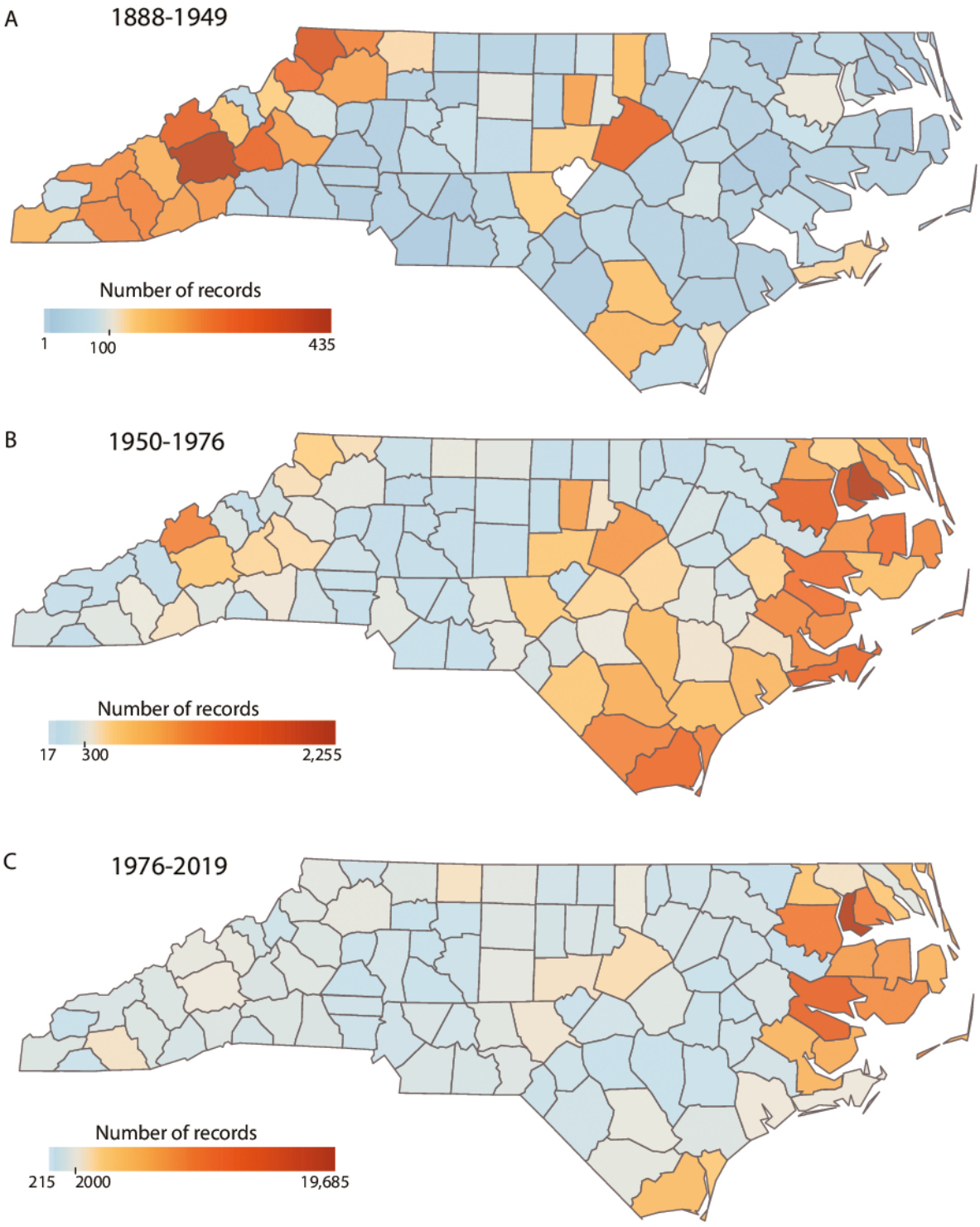

Fig. 5. Number of records per county pre-1949 (A), 1950–1976 (B), and 1977–2019 (C).

1878). Given that early naturalists were often conducting systematic work and faced limitations in collection gear and travel logistics, our results add to the growing evidence that phylogenetic and spatial biases are an intrinsic component of aggregated historical data stemming from early "mega-collectors" (Daru et al. 2018, Gippoliti and Groves 2018).

It is indisputable that the efforts of early naturalists to systematically chronicle the biodiversity of the Western Hemisphere continue to be of extreme value to our understanding of fish distributions (Skelton et al. 1995, Labay et al. 2011). Nevertheless, it is important to recognize that by the mid-1900s, biodiversity inventory efforts themselves began to diversify. Our results support an abrupt shift in sampling focus in the late 1940s (Fig. 2B) with a sudden increase in overall taxonomic coverage (Fig. 2A) and a new emphasis on the sampling of coastal counties (Fig. 5B). In addition to focused surveys by individual researchers (see Tracy et al. 2020), this second phase of historical sampling is associated with a phylogenetic bias toward the sampling of commercially important clupeids (i.e., river herrings), moronids (i.e., Striped Bass), and game fish such as sunfish and bass. This shift was likely the result of concerns about the future of river herring and Striped Bass fisheries as well as a systematic effort by the NCWRC to inventory all of the freshwater lakes, rivers, and streams of the state that support game fish due to general concerns about water quality in the state (Menhinick et al. 1974). During this time, interest among the scientific community concerning fish distributions rapidly expanded from natural history into the study of the effects and accumulation of chemicals on fish populations, often with an emphasis on food and sportfish (Knoll and Fromm 1960, Burdick et al. 1964, Mount and Stephan 1967). In turn, these studies contributed to the foundation of ecotoxicology as a discipline (Jouany 1971) and led to the passage of the Clean Water Act in 1972 (Hines 2012). Additionally, the cultural divide between "laboratory" and "field" science began to rapidly erode (Kohler 2002), coinciding with a strong environmental movement that promoted a surge of interest in ecology (Johnson and Frickel 2011). As such, the shift in sampling and associated aggregate biases we observed in our analyses are part of a larger historical trend occurring throughout many regions of the world and again a very likely a general feature of other historical datasets.

The many technological and computational innovations developed by the end of the 20th century are broadly recognized as empowering rapid growth and dissemination of scientific research (Larsen and von Ins 2010). However, our results suggest that this trend is likely correlated with a significant additional temporal skewing of distributional records (Fig. 2A). We found that over 81% of North Carolina's 276,138 freshwater fish records were collected after 1976. Collectively, these efforts resulted in spatial, temporal, and taxonomic breadth of sampling that are without a historical parallel and reflect the growth of aquatic science in the region. From the late 1970s on, the number of research laboratories working on the taxonomy, ecology, and other aspects of fish biology continued to grow (Davis and Louder 1971, Harrell and Cloutman 1978, Weinstein and Davis 1980, Cloutman and Harrell 1987, Burkholder et al. 1992, Midway et al. 2014, 2015), and more work was conducted in ecologically unique areas such as the Sand Hills (Rohde and Ross 1987, Rohde and Arndt 1991), as well as in areas experiencing changes in urban infrastructure (Baumann and Gillespie 1986) or land use (Harding et al. 1998, Kennen et al. 2005). Fish also became used as indicator species by state water quality regulatory agencies such as the North Carolina Department of Water Quality in point and non-point source pollution and in basin-wide surveys (Hocutt 1981, Karr 1981, Simon 1999, NCDENR 2013), and the NCWRC expanded surveys to include non-game fish throughout the state. This trend of rapid data acquisition is not unique to North Carolina and reflects a larger global effort to monitor biodiversity more effectively. Indeed, the level of temporal bias we observed is nearly identical to frequencies of plants collected in the history of Australian botanical collections (Haque et al. 2018) and also consistent with historical surveys of insects (Hortal et al. 2008) as well as mammals (Escribano et al. 2016). As such, we expect temporal skewing towards the present day to be a hallmark of other aggregate datasets.

Sampling biases should be expected in aggregate datasets stemming from records that were collected without a unifying sampling protocol

(Boakes et al. 2010), and our results highlight general biases that are likely to be encountered when analyzing historical species distributional records. Some of these biases can be accounted for in more sophisticated statistical models (Dorazio 2014, Daru et al. 2018) or by extending existing methods that account for sample biases. For example, current "target group background" approaches (Phillips and Dudík 2008, Phillips et al. 2009) that use data from related species to correct for sample bias could be extended to also account for temporal sampling bias. While the continued development of models that account for historical biases is vital to the analyses of historical spatial datasets, there is also a trade-off between parameter-rich models and statistical power (Tingley and Beissinger 2009). This trade-off can be amplified by the degradation of statistical power through time as a function of the underlying distribution of temporal records (Tessarolo et al. 2017). As such, the development of methods like those in other fields used to profile predicted statistical utility based on data information content through time (Townsend 2007, Dornburg et al. 2016, 2017*c*) could provide a means to empower dataset scrutiny. Moreover, these types of methods could potentially mitigate model misspecification and aid in determining the credibility of analysis results. The development of such tools alongside new models that capture historical shifts in sampling designs offer an exciting research frontier that will allow us to more effectively harness the wealth of historical information contained within aggregate datasets.

Historical data and the changing face of 21st century data collection

As we continue into the 21st century, continued technological and methodological innovations have enhanced our ability to observe and monitor species to an extent that was previously impossible. Camera traps (Long et al. 2012, Fleming et al. 2014, Kays 2016), video surveys (Andradi-Brown et al. 2016, Fisher et al. 2016, Galaiduk et al. 2017), and citizen scientists (Bodilis et al. 2014, Donnelly et al. 2014, Parsons et al. 2018) can provide tens of thousands of records in a single field season. Likewise, environmental DNA can be used to barcode entire communities (Yamamoto et al. 2016, Sato et al. 2018, Taberlet et al. 2018) and complement visual surveys (Stat et al. 2019). Maximizing the utility of these powerful new resources will not only require consideration of potential biases intrinsic to each collection effort (Isaac et al. 2014, Collins et al. 2019, Hofmeester et al. 2019) but also quantification of shifts in sampling biases that correspond with changing trends in data collection to avoid distorting our view of biodiversity (Boakes et al. 2010). Although the pace of sampling today vastly exceeds that of the 20th century, the problem of inducing sampling biases when aggregating heterogeneous sampling designs targeting specific regions or taxa remains just as relevant today as when working with data collected over a century ago. As such, studies that quantify and report sampling biases will be of continual importance for avoiding bias driven errors (Bystriakova et al. 2012), as will developing and utilizing methods that can account for identified sampling biases if we are to serve as effective stewards of the lineages forecast to face rapid changes in environmental conditions over the next century.

Acknowledgments

Funding for this work was provided in part by the Institute of Museum and Library Services to AD & GH (MA-30-18-0275-18) and the state of North Carolina Internship program to KS. We thank J. Townsend for helpful discussions about aspects of early drafts of this manuscript. G. Hogue and A. Dornburg contributed equally to the work reported here.

Literature Cited

Akaike, H. 1998. Information theory and an extension of the maximum likelihood principle. Selected papers of Hirotugu Akaike. Pages 199–213. Springer, New York, New York, USA.

Andradi-Brown, D. A., C. Macaya-Solis, D. A. Exton, E. Gress, G. Wright, and A. D. Rogers. 2016. Assessing Caribbean shallow and mesophotic reef fish communities using Baited-Remote Underwater Video (BRUV) and Diver-Operated Video (DOV) survey techniques. PLOS ONE 11:e0168235.

Baird, S. F., and C. Girard. 1853. Descriptions of some new fishes from the River Zuni. Proceedings of the Academy of Natural Sciences of Philadelphia 1:368–370.

Baumann, P. C., and R. B. Gillespie. 1986. Selenium bioaccumulation in gonads of largemouth bass and bluegill from three power plant cooling reservoirs. Environmental Toxicology and Chemistry 5:695–701.

Beamesderfer, R. C., and B. E. Rieman. 1988. Size selectivity and bias in estimates of population statistics of smallmouth bass, Walleye, and Northern Squawfish in a Columbia river reservoir. North American Journal of Fisheries Management 8:505–510.

Blagoderov, V., and V. Smith. 2012. Bringing collections out of the dark. ZooKeys 209:1–6.

Boakes, E. H., P. J. K. McGowan, R. A. Fuller, D. Chang-qing, N. E. Clark, K. O'Connor, and G. M. Mace. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biology 8:e1000385.

Bodilis, P., P. Louisy, M. Draman, H. O. Arceo, and P. Francour. 2014. Can citizen science survey non-indigenous fish species in the eastern Mediterranean Sea? Environmental Management 53:172–180.

Bond, N., J. Thomson, P. Reich, and J. Stein. 2011. Using species distribution models to infer potential climate change-induced range shifts of freshwater fish in south-eastern Australia. Marine and Freshwater Research 62:1043.

Botts, E. A., B. F. N. Erasmus, and G. J. Alexander. 2011. Geographic sampling bias in the South African Frog Atlas Project: implications for conservation planning. Biodiversity and Conservation 20:119–139.

Bulluck, L., E. Fleishman, C. Betrus, and R. Blair. 2006. Spatial and temporal variations in species occurrence rate affect the accuracy of occurrence models. Global Ecology and Biogeography 15:27–38.

Burdick, G. E., E. J. Harris, H. J. Dean, T. M. Walker, J. Skea, and D. Colby. 1964. The accumulation of DDT in Lake Trout and the effect on reproduction. Transactions of the American Fisheries Society 93:127–136.

Burkholder, J. M., E. J. Noga, C. H. Hobbs, H. B. Glasgow Jr, and S. A. Smith. 1992. New "phantom" dinoflagellate is the causative agent of major estuarine fish kills. Nature 358:407–410.

Burnham, K. P., and D. R. Anderson. 2007. Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach. Springer Science & Business Media, New York, USA.

Bystriakova, N., M. Peregrym, R. H. J. Erkens, O. Bezsmertna, and H. Schneider. 2012. Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. Systematics and Biodiversity 10:305–315.

Chang, J., D. L. Rabosky, S. A. Smith, and M. E. Alfaro. 2019. An r package and online resource for macroevolutionary studies using the ray-finned fish tree of life. Methods in Ecology and Evolution 10:1118–1124.

Cloutman, D. G., and R. D. Harrell. 1987. Life history notes on the Whitefin Shiner, *Notropis niveus* (Pisces: Cyprinidae), in the Broad River, South Carolina. Copeia 1987:1037–1040.

Collins, R. A., J. Bakker, O. S. Wangensteen, A. Z. Soto, L. Corrigan, D. W. Sims, M. J. Genner, and S. Mariani. 2019. Non-specific amplification compromises environmental DNA metabarcoding with COI. Methods in Ecology and Evolution 10:1985–2001.

Cope, E. D. 1870*a*. A partial synopsis of the fishes of the fresh waters of North Carolina. Proceedings of the American Philosophical Society 11:448–495.

Cope, E. D. 1870*b*. On some Etheostomine perch from Tennessee and North Carolina. Proceedings of the American Philosophical Society 11:261–270.

Cope, E. D. 1871. On the fishes of the Ambyiacu River. Proceedings of the Academy of Natural Sciences of Philadelphia 23:250–294.

Cope, E. D. 1875. Report Upon the Collections of Fishes Made in Portions of Nevada, Utah, California, Colorado, New Mexico, and Arizona: during the Years 1871, 1872, 1873, and 1874.

Cope, E. D. 1878. Synopsis of the Fishes of the Peruvian Amazon, Obtained by Professor Orton During His Expeditions of 1873 and 1877.

Cowman, P. F., V. Parravicini, M. Kulbicki, and S. R. Floeter. 2017. The biogeography of tropical reef fishes: endemism and provinciality through time. Biological Reviews of the Cambridge Philosophical Society 92:2112–2130.

Dallas, T., S. Huang, C. Nunn, A. W. Park, and J. M. Drake. 2017. Estimating parasite host range. Proceedings of the Royal Society B: Biological Sciences 284:20171250.

Daru, B. H., et al. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. New Phytologist 217:939–955.

Daufresne, M., and P. Boët. 2007. Climate change impacts on structure and diversity of fish communities in rivers. Global Change Biology 13:2467–2478.

Davis, J. R., and D. E. Louder. 1971. Life history and ecology of the cyprinid fish *Notropis petersoni* in North Carolina waters. Transactions of the American Fisheries Society 100:726–733.

Donnelly, A., O. Crowe, E. Regan, S. Begley, and A. Caffarra. 2014. The role of citizen science in monitoring biodiversity in Ireland. International Journal of Biometeorology 58:1237–1249.

Dorazio, R. M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecology and Biogeography 23:1472–1484.

Dornburg, A., S. Federman, A. D. Lamb, C. D. Jones, and T. J. Near. 2017*a*. Cradles and museums of Antarctic teleost biodiversity. Nature Ecology & Evolution 1:1379–1384.

Dornburg, A., E. Forrestel, J. Moore, T. Iglesias, A. Jones, L. Rao, and D. Warren. 2017*b*. An assessment of sampling biases across studies of diel activity patterns in marine ray-finned fishes (Actinopterygii). Bulletin of Marine Science 93:611–639.

Dornburg, A., J. P. Townsend, and Z. Wang. 2017*c*. Maximizing power in phylogenetics and phylogenomics: a perspective illuminated by fungal big data. Advances in Genetics 100:1–47.

Dornburg, A., J. N. Fisk, J. Tamagnan, and J. P. Townsend. 2016. PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. BMC Evolutionary Biology 16:262.

Dornburg, A., J. Moore, J. M. Beaulieu, R. I. Eytan, and T. J. Near. 2015. The impact of shifts in marine biodiversity hotspots on patterns of range evolution: evidence from the Holocentridae (squirrelfishes and soldierfishes). Evolution 69:146–161.

Dornburg, A., and T. J. Near. 2021. The emerging phylogenetic perspective on the evolution of Actinopterygian fishes. Annual Review of Ecology, Evolution, and Systematics 52:427–452.

Echelle, A. A., M. R. Schwemm, N. J. Lang, J. S. Baker, R. M. Wood, T. J. Near, and W. L. Fisher. 2015. Molecular systematics of the least darter (Percidae: *Etheostoma microperca*): historical biogeography and conservation implications. Copeia 103: 87–98.

Engemann, K., B. J. Enquist, B. Sandel, B. Boyle, P. M. Jørgensen, N. Morueta-Holme, R. K. Peet, C. Violle, and J.-C. Svenning. 2015. Limited sampling hampers "big data" estimation of species richness in a tropical biodiversity hotspot. Ecology and Evolution 5:807–820.

Escribano, N., A. H. Ariño, and D. Galicia. 2016. Biodiversity data obsolescence and land uses changes. PeerJ 4:e2743.

Fisher, R. B., Y.-H. Chen-Burger, D. Giordano, L. Hardman, and F.-P. Lin. 2016. Fish4Knowledge: collecting and analyzing massive coral reef fish video data. Springer International Publishing, New York, USA.

Fleming, P., P. Meek, G. Ballard, P. Banks, A. Claridge, J. Sanderson, and D. Swann. 2014. Camera trapping: wildlife management and research. CSIRO PUBLISHING, Melbourne, Victoria, Australia.

Forrestel, E. J., D. D. Ackerly, and N. C. Emery. 2015. The joint evolution of traits and habitat: ontogenetic shifts in leaf morphology and wetland specialization in Lasthenia. New Phytologist 208:949–959.

Forrestel, E. J., M. J. Donoghue, and M. D. Smith. 2014. Convergent phylogenetic and functional responses to altered fire regimes in mesic savanna grasslands of North America and South Africa. New Phytologist 203:1000–1011.

Franklin, J., J. M. Serra-Diaz, A. D. Syphard, and H. M. Regan. 2017. Big data for forecasting the impacts of global change on plant communities. Global Ecology and Biogeography 26:6–17.

Galaiduk, R., B. T. Radford, S. K. Wilson, and E. S. Harvey. 2017. Comparing two remote video survey methods for spatial predictions of the distribution and environmental niche suitability of demersal fishes. Scientific Reports 7:17633.

Gippoliti, S., and C. P. Groves. 2018. Overlooked mammal diversity and conservation priorities in Italy: impacts of taxonomic neglect on a Biodiversity Hotspot in Europe. Zootaxa 4434:511–528.

Griffiths, G. H., B. C. Eversham, and D. B. Roy. 1999. Integrating species and habitat data for nature conservation in Great Britain: data sources and methods. Global Ecology and Biogeography 8:329–345.

Haque, M. D. M., D. A. Nipperess, J. B. Baumgartner, and L. J. Beaumont. 2018. A journey through time: exploring temporal patterns amongst digitized plant specimens from Australia. Systematics and Biodiversity 16:604–613.

Harding, J. S., E. F. Benfield, P. V. Bolstad, G. S. Helfman, and E. B. 3rd Jones. 1998. Stream biodiversity: the ghost of land use past. Proceedings of the National Academy of Sciences of the United States of America 95:14843–14847.

Harrell, R. D., and D. G. Cloutman. 1978. Distribution and life history of the Sandbar Shiner, *Notropis scepticus* (Pisces: Cyprinidae). Copeia 1978:443–447.

Healey, A. J. E., N. J. McKeown, A. L. Taylor, J. Provan, W. Sauer, G. Gouws, and P. W. Shaw. 2018. Cryptic species and parallel genetic structuring in Lethrinid fish: implications for conservation and management in the southwest Indian Ocean. Ecology and Evolution 8:2182–2195.

Hickisch, R., T. Hodgetts, P. J. Johnson, C. Sillero-Zubiri, K. Tockner, and D. W. Macdonald. 2019. Effects of publication bias on conservation planning. Conservation Biology 33:1151–1163.

Hines, W. M. 2012. History of the 1972 clean water act: the story behind how the 1972 act became the capstone on a decade of extraordinary environmental reform. U Iowa Legal Studies Research Paper No. 12-12, SSRN Electronic Journal.

Hocutt, C. H. 1981. Fish as indicators of biological integrity. Fisheries 6:2831.

Hofmeester, T. R., J. P. G. M. Cromsigt, J. Odden, H. Andrén, J. Kindberg, and J. D. C. Linnell. 2019. Framing pictures: a conceptual framework to identify and correct for biases in detection probability

of camera traps enabling multi-species comparison. Ecology and Evolution 9:2320–2336.

Hortal, J., A. Jiménez-Valverde, J. F. Gómez, J. M. Lobo, and A. Baselga. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117:847–858.

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

Isaac, N. J. B., A. J. van Strien, T. A. August, M. P. de Zeeuw, and D. B. Roy. 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. Methods in Ecology and Evolution 5:1052–1060.

Jetz, W., J. M. McPherson, and R. P. Guralnick. 2012. Integrating biodiversity distribution knowledge: toward a global map of life. Trends in Ecology & Evolution 27:151–159.

Johnson, E. W., and S. Frickel. 2011. Ecological threat and the founding of U.S. National Environmental movement organizations, 1962–1998. Social Problems 58:305–329.

Jordan, D. S. 1889. Report of explorations made during the summer and autumn of 1888, in the Alleghany region of Virginia, North Carolina and Tennessee, and in western Indiana, with an account of the fishes found in each of the river basins in those regions. Bulletin of the United States Fish Commission 8:97–173.

Jouany, J. M. 1971. Nuisances et Ecologie. Actualités Pharmaceutiques 69:11–22.

Karr, J. R. 1981. Assessment of biotic integrity using fish communities. Fisheries 6:21–27.

Kays, R. 2016. Candid creatures: How camera traps reveal the mysteries of nature. JHU Press, Baltimore, Maryland, USA.

Kennen, J. G., M. Chang, and B. H. Tracy. 2005. Effects of landscape change on fish assemblage structure in a rapidly growing metropolitan area in North Carolina, USA. American Fisheries Society Symposium 47:39–52.

Kjelson, M. A., and D. R. Colby. 1977. The evaluation and use of gear efficiencies in the estimation of estuarine fish abundance. Estuarine Processes II:416–424.

Knoll, J., and P. O. Fromm. 1960. Accumulation and elimination of hexavalent chromium in rainbow trout. Physiological Zoology 33:1–8.

Koenig, C. C., and C. D. Stallings. 2015. A new compact rotating video system for rapid survey of reef fish populations. Bulletin of Marine Science 91:365–373.

Kohler, R. E. 2002. Landscapes and labscapes: exploring the lab-field border in biology. University of Chicago Press, Chicago, Illinois, USA.

La Salle, J., K. J. Williams, and C. Moritz. 2016. Biodiversity analysis in the digital era. Philosophical Transactions of the Royal Society B: Biological Sciences 371:20150337.

Labay, B., A. E. Cohen, B. Sissel, D. A. Hendrickson, F. D. Martin, and S. Sarkar. 2011. Assessing historical fish community composition using surveys, historical collection data, and species distribution models. PLOS ONE 6:e25145.

Larsen, P. O., and M. von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 84:575–603.

Long, R. A., P. MacKay, J. Ray, and W. Zielinski. 2012. Noninvasive survey methods for carnivores. Island Press, Washington, DC, USA.

Mailhot, J. 2020. Seeing the story: Making data visualization accessible to natural history collections through the creation of custom-made dashboards and resources. Dissertation. University of Colorado at Boulder, Boulder, Colorado, USA.

McMahan, C. D., C. E. Fuentes-Montejo, L. Ginger, J. C. Carrasco, P. Chakrabarty, and W. A. Matamoros. 2020. Climate change models predict decreases in the range of a microendemic freshwater fish in Honduras. Scientific Reports 10:12693.

Menhinick, E. F., T. M. Burton, and J. R. Bailey. 1974. An annotated checklist of the freshwater fishes of North Carolina. Journal of the Elisha Mitchell Scientific Society 1:24–50.

Midway, S. R., T. Wagner, and B. H. Tracy. 2014. A hierarchical community occurrence model for North Carolina Stream Fish. Transactions of the American Fisheries Society 143:1348–1357.

Midway, S. R., T. Wagner, B. H. Tracy, G. M. Hogue, and W. C. Starnes. 2015. Evaluating changes in stream fish species richness over a 50-year time-period within a landscape context. Environmental Biology of Fishes 98:1295–1309.

Millar, E. E., E. C. Hazell, and S. J. Melles. 2019. The "cottage effect" in citizen science? Spatial bias in aquatic monitoring programs. International Journal of Geographical Information Science 33:1612–1632.

Monsarrat, S., A. F. Boshoff, and G. I. H. Kerley. 2019*a*. Accessibility maps as a tool to predict sampling bias in historical biodiversity occurrence records. Ecography 42:125–136.

Monsarrat, S., P. Novellie, I. Rushworth, and G. Kerley. 2019*b*. Shifted distribution baselines: neglecting long-term biodiversity records risks overlooking potentially suitable habitat for conservation management. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 374:20190215.

Mount, D. I., and C. E. Stephan. 1967. A method for detecting cadmium poisoning in fish. Journal of Wildlife Management 31:168.

NCDENR. 2013. Standard operating procedure, biological monitoring, stream fish community assessment program. Page 52. North Carolina Department of Environment and Natural Resources, Division of Water Resources, Raleigh, North Carolina, USA.

Near, T. J., A. Dornburg, K. L. Kuhn, J. T. Eastman, J. N. Pennington, T. Patarnello, L. Zane, D. A. Fernández, and C. D. Jones. 2012*a*. Ancient climate change, antifreeze, and the evolutionary diversification of Antarctic fishes. Proceedings of the National Academy of Sciences of the United States of America 109:3434–3439.

Near, T. J., R. I. Eytan, A. Dornburg, K. L. Kuhn, J. A. Moore, M. P. Davis, P. C. Wainwright, M. Friedman, and W. L. Smith. 2012*b*. Resolution of ray-finned fish phylogeny and timing of diversification. Proceedings of the National Academy of Sciences of the United States of America 109:13698–13703.

Neath, A. A., and J. E. Cavanaugh. 2012. The Bayesian information criterion: background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics 4:199–203.

Nee, S., E. C. Holmes, R. M. May, and P. H. Harvey. 1994. Extinction rates can be estimated from molecular phylogenies. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 344:77–82.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Parsons, A. W., C. Goforth, R. Costello, and R. Kays. 2018. The value of citizen science for ecological monitoring of mammals. PeerJ 6:e4536.

Patton, T. M., F. J. Rahel, and W. A. Hubert. 1998. Using historical data to assess changes in Wyoming's fish fauna. Conservation Biology 12:1120–1128.

Pennell, M. W., J. M. Eastman, G. J. Slater, J. W. Brown, J. C. Uyeda, R. G. FitzJohn, M. E. Alfaro, and L. J. Harmon. 2014. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics 30:2216–2218.

Phillips, S. J., and M. Dudík. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.

Phillips, S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19:181–197.

Porubsky, D., A. D. Sanders, A. Taudt, M. Colomé-Tatché, P. M. Lansdorp, and V. Guryev. 2020. breakpointR: an R/Bioconductor package to localize strand state changes in Strand-seq data. Bioinformatics 36:1260–1261.

Rabosky, D. L., et al. 2018. An inverse latitudinal gradient in speciation rate for marine fishes. Nature 559:392–395.

Revell, L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3:217–223.

Rocha, L. A. 2003. Patterns of distribution and processes of speciation in Brazilian reef fishes. Journal of Biogeography 30:1161–1171.

Rohde, F. C., and R. G. Arndt. 1991. Distribution and status of the sandhills chub, *Semotilus lumbee*, and the pinewoods darter, *Etheostoma mariae*. Journal of the Elisha Mitchell Scientific Society 1:61–70.

Rohde, F. C., and S. W. Ross. 1987. Life-history of the pinewoods darter, *Etheostoma maraie* (Osteichthyes, Percidae), a fish endemic to the Carolina sandhills. Brimleyana 13:1–20.

Romo, H., E. García-Barros, and J. M. Lobo. 2006. Identifying recorder-induced geographic bias in an Iberian butterfly database. Ecography 29:873–885.

Ruaro, R., et al. 2019. Climate change will decrease the range of a keystone fish species in La Plata River Basin, South America. Hydrobiologia 836:1–19.

Sato, Y., M. Miya, T. Fukunaga, T. Sado, and W. Iwasaki. 2018. MitoFish and MiFish Pipeline: a mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding. Molecular Biology and Evolution 35:1553–1555.

Shaffer, H. B., R. N. Fisher, and C. Davidson. 1998. The role of natural history collections in documenting species declines. Trends in Ecology & Evolution 13:27–30.

Simon, T. P. 1999. Assessing the sustainability and biological integrity of water resources using fish communities. Pages 671. CRC Press, Boca Raton, Florida, USA.

Siqueira, A. C., D. R. Bellwood, and P. F. Cowman. 2019. Historical biogeography of herbivorous coral reef fishes: the formation of an Atlantic fauna. Journal of Biogeography 46:1611–1624.

Skelton, P. H., J. A. Cambray, A. Lombard, and G. A. Benn. 1995. Patterns of distribution and conservation status of freshwater fishes in South Africa. South African Journal of Zoology 30:71–81.

Soltis, D. E., and P. S. Soltis. 2016. Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. Plant Divers 38:264–270.

Soltis, P. S., G. Nelson, and S. A. James. 2018. Green digitization: online botanical collections data answering real-world questions. Applications in Plant Sciences 6:e1028.

Stat, M., J. John, J. D. DiBattista, S. J. Newman, M. Bunce, and E. S. Harvey. 2019. Combined use of eDNA metabarcoding and video surveillance for the assessment of fish biodiversity. Conservation Biology 33:196–205.

Stropp, J., R. J. Ladle, A. C. M. Malhado, J. Hortal, J. Gaffuri, W. H. Temperley, J. O. Skøien, and P. Mayaux. 2016. Mapping ignorance: 300 years of collecting flowering plants in Africa. Global Ecology and Biogeography 25:1085–1096.

Taberlet, P., A. Bonin, L. Zinger, and E. Coissac. 2018. Environmental DNA: for biodiversity research and monitoring. Oxford University Press, Oxford, UK.

Tableau Software, LLC. 2020. Tableau Public. Tableau Software, LLC, Seattle, Washington, USA.

Tessarolo, G., R. Ladle, T. Rangel, and J. Hortal. 2017. Temporal degradation of data limits biodiversity research. Ecology and Evolution 7:6863–6870.

Tingley, M. W., and S. R. Beissinger. 2009. Detecting range shifts from historical species occurrences: new perspectives on old data. Trends in Ecology & Evolution 24:625–633.

Tolley, K. A., G. J. Alexander, W. R. Branch, P. Bowles, and B. Maritz. 2016. Conservation status and threats for African reptiles. Biological Conservation 204:63–71.

Townsend, J. P. 2007. Profiling phylogenetic informativeness. Systematic Biology 56:222–231.

Tracy, B. H., F. C. Rohde, and G. M. Hogue. 2020. An Annotated Atlas of the Freshwater Fishes of North Carolina. Southeastern Fishes Council Proceedings: No. 60. https://trace.tennessee.edu/sfcproceedings/vol1/iss60/1

Van Metre, P. C., et al. 2019. Projected urban growth in the southeastern USA puts small streams at risk. PLOS ONE 14:e0222714.

Wagenmakers, E.-J., and S. Farrell. 2004. AIC model selection using Akaike weights. Psychonomic Bulletin & Review 11:192–196.

Warren, D. L., R. I. Eytan, A. Dornburg, T. L. Iglesias, M. C. Brandley, and P. C. Wainwright. 2021. Reevaluating claims of ecological speciation in *Halichoeres bivittatus*. Ecology and Evolution 11:11449–11456.

Weinstein, M. P., and R. W. Davis. 1980. Collection efficiency of seine and rotenone samples from Tidal Creeks, Cape Fear River, North Carolina. Estuaries 3:98.

Willis, T. J., and R. C. Babcock. 2000. A baited underwater video system for the determination of relative density of carnivorous reef fish. Marine and Freshwater Research 51:755.

Yamamoto, S., et al. 2016. Environmental DNA as a "Snapshot" of fish distribution: a case study of Japanese Jack Mackerel in Maizuru Bay, Sea of Japan. PLOS ONE 11:e0149786.

Yang, W., K. Ma, and H. Kreft. 2013. Geographical sampling bias in a large distributional database and its effects on species richness-environment models. Journal of Biogeography 40:1415–1426.

Data Availability

Data and R script are available from Zenodo: https://doi.org/10.5281/zenodo.4081842