Non-stationary drivers on fish sampling efforts in Brazilian freshwaters

Thatyane Caetano de Almeida | Geiziane Tessarolo | João Carlos Nabout | Fabrício Barreto Teresa

Abstract
Aim: Closing knowledge gaps in species taxonomy and distribution (i.e. Linnean and Wallacean shortfalls, respectively) require spatially distributed high-quality data. However, studies on terrestrial taxa have shown that occurrence data are biased owing to higher sampling efforts towards areas with greater accessibility, research infrastructure and attractiveness. Here, we tested whether these biasing factors are also important drivers of freshwater fish species research by assessing fishes’ sampling efforts in Brazil, a continental-sized and megadiverse country. We hypothesized that the influence of biasing factors is scale-dependent, with differential effects across regions, that is that they present non-stationarity.

Location: Brazil.

Methods: We modelled sampling events of fish (gathered from online databases) as a function of accessibility (population density, distance from access routes and density of access routes), availability of research infrastructure (distance from research centres) and attractiveness of protected areas (distance from protected areas), by using a traditional stationary model (ordinary least square models—OLS) and a model taking into account non-stationarity (geographically weighted regression—GWR).

Results: We recorded 43,538 sampling events distributed across Brazil. Freshwater fish samplings were spatially biased with a tendency for major efforts to be concentrated in sites with a higher density of access routes, population and nearness to research centres and protected areas. However, GWR models performed better than the OLS model, revealing the non-stationarity of the effects of predictors in explaining the number of sampling events.

Main conclusions: Overcoming Linnean and Wallacean shortfalls for freshwater fish species in megadiverse regions such as Brazil can be aided by financially supporting the sampling efforts in less accessible and less populated sites farther from research centres and well-studied protected areas. However, the source of biasing has strong scale dependency, so that reducing a specific source of bias is not expected to have the same effectiveness in all areas.
1 | INTRODUCTION

Information regarding species occurrence is essential for ecological and evolutionary research to understand processes that affect their geographical distribution (Meyer et al., 2015; Troia & McManamay, 2017). Additionally, high-quality species distribution data are required to define conservation strategies and for the prioritization of resource allocation (Hortal et al., 2007; Ladle & Whittaker, 2011; Tessarolo et al., 2017). The increased concern relating to data availability has stimulated initiatives for online aggregation of distributional data which are now publicly available (Correia et al., 2019; Meyer et al., 2015). Initiatives such as GBIF (Global Biodiversity Information Facility) and SpeciesLink manage billions of species records from museums, herbaria and research centres which have been extensively used for various ecological and conservation studies (e.g. Canhos et al., 2015). Despite these online databases, which have improved the quality and accuracy of their data over time (Gaiji et al., 2013), high-quality data regarding species occurrences are missing for most known species and regions, a gap known as Wallacean shortfall (Bini et al., 2006; Hortal et al., 2015; Lomolino, 2004; Whittaker et al., 2005).

The knowledge shortfall related to species distribution is a result of the complex dynamic of species distribution itself (Tessarolo et al., 2017), but also of the uneven collection of data (Hortal et al., 2015; Meyer et al., 2015). Ideally, sampling efforts should be distributed in the geographical space, in a way that all real natural variation is detected (Brooks et al., 2004). However, even though increasing concern about intensification and spatialization of samplings has emerged (Girardello et al., 2018; Hortal et al., 2007; Oliveira et al., 2016; Reddy & Dávalos, 2003; Romo et al., 2006), spatial biases in biodiversity records are still considerable (Brooks et al., 2004).

Sampling bias results partially from the small spatial scale of the studies developed, which generates a spatial concentration of records restricting the knowledge and inferences of distribution patterns at larger scales (Dias et al., 2016; Hecnar & M’Closkey, 1997), consequently affecting conservation actions (Brooks et al., 2004). Unequal distribution of species sampling efforts may be driven by several factors such as ease of access (Correia et al., 2019; Oliveira et al., 2016; Romo et al., 2006; Sánchez-Fernández et al., 2008), availability of financial resources (Anderson, 2012; Yang et al., 2013), attractiveness for specific areas or species (Reddy & Dávalos, 2003) and security (Amano & Sutherland, 2013). In fact, researchers tend to concentrate their samplings on sites close to their work, generally with the presence of roads near large cities seeking to facilitate travel (Sánchez-Fernández et al., 2008; Sastre & Lobo, 2009) and minimize logistics and financial expenses generated by fieldwork (Botts et al., 2011; Meyer et al., 2015). In addition, some sites also exert a certain attractiveness on some researchers, including protected areas, representing more conserved sites where there is a greater probability of finding high species richness or even rare species (Oliveira et al., 2016; Sánchez-Fernández et al., 2008).

Sampling bias studies often examine relationships between sample bias and factors causing it (e.g. presence of roads) in a unique global model applied to the entire studied region, not considering the spatial variability of sources bias (Oliveira et al., 2016). However, the relationship among spatial sample bias and the factors causing it can vary between regions, indicating the occurrence of non-stationary geographical patterns. For example, the known positive relationship between terrestrial access routes and sampling efforts (Oliveira et al., 2016) may be weaker or negative in regions where the navigable rivers are the main access routes, such as in the Amazon basin (Jézéquel et al., 2019). Therefore, if spatial non-stationarity is present, other statistical approaches such as GWR (geographically weighted regression; Fotheringham et al., 2003) could be used to detect the spatial variability of the predictors and improve them when compared to global models (see examples using GWR in broad-scale studies—Caetano et al., 2018; Eme et al., 2015; Foody, 2004; Terrible & Diniz-Filho, 2009).

Freshwater systems encompass a variety of habitats, harbouring around 10% of all known biodiversity, with high levels of endemism (Collen et al., 2013; Strayer & Dudgeon, 2010). Additionally, freshwater systems play an important role in providing resources for humans, such as water filtration, food provision and air quality regulation (Collen et al., 2013; Vigerstol & Aukema, 2011). Although the number of studies focusing on the identification of spatial biases regarding biodiversity data has been increasing (Boakes et al., 2010; Hortal et al., 2007; Oliveira et al., 2016; Reddy & Dávalos, 2003; Romo et al., 2006; Sánchez-Fernández et al., 2008; Stropp et al., 2016; Yang et al., 2013), the majority encompass only terrestrial vertebrate species. Freshwater biodiversity generally has been overlooked (Lévêque et al., 2005; Sánchez-Fernández et al., 2008), with few relevant studies conducted in this field (Pelayo-Villamil et al., 2018; Sánchez-Fernández et al., 2008; Troia & McManamay, 2017). Species in freshwater systems face severe anthropogenic threats, suffering a greater extinction risk than terrestrial species and significant biodiversity loss (Collen et al., 2013; Gallardo et al., 2018). In this context, filling the knowledge gaps in freshwater species distribution will be useful for informing freshwater conservation management (Howard et al., 2015). Thus, it has become imperative to understand the factors causing biases in freshwater fish data distribution and assess its limitations (Boyero et al., 2012; Troia & McManamay, 2017). As the influence of these biasing factors may vary spatially, predictions of biodiversity data would be required to go beyond stationary models (Oliveira et al., 2016) and to consider context-dependent processes more explicitly.

In this study, we evaluate the spatial distribution of the sampling efforts of Brazilian freshwater fish and determine the influence of...
commonly assessed bias factors on such data. Brazil is a continental-sized country with the most diverse freshwater ichthyofauna in the world (Reis et al., 2003; Tisseuil et al., 2013), which has attracted the attention of researchers and resulted in significant cumulated sampling efforts over decades (Dias et al., 2016; Rapp Py-Daniel et al., 2015). It encompasses different biomes, exhibiting a broad range of environmental, infrastructural and socio-economic contexts. This combination of high biological diversity and spatial variability in potential biasing factors makes it a suitable model to analyse drivers of sampling efforts (Oliveira et al., 2016). We tested the hypothesis that the sampling efforts directed towards freshwater fish become biased owing to the accessibility, research infrastructure and attractiveness of protected areas. Furthermore, considering that these relationships are expected to be scale-dependent, with the differential effect of biasing factors across regions, we predicted that non-stationary models would provide better predictions than global models.

2 | MATERIALS AND METHODS

2.1 | Data collection

We listed 3,358 species of freshwater fish for Brazil supported on the database “Fishbase” (www.fishbase.org) by applying the following search filters: “Brazil” and “Freshwater.” Marine species that were erroneously included within the database and non-native species (sensu Azevedo-Santos et al., 2015, Latini et al., 2016) were removed so that the list contained exclusively native freshwater species. We obtained all occurrence records per species available from four online databases (search was made one species at a time in each database): Global Biodiversity Information Facility (GBIF; www.gbif.org), SpeciesLink (www.splink.org.br), Portal da Biodiversidade (portaladabiodiversidade.icmbio.gov.br/portal/) and Sistema de Informação sobre a Biodiversidade Brasileira (SBBR; www.sibbr.gov.br/).

To guarantee that unreliable and non-relevant records were removed, we carried out a data filtering process. Records lacking information regarding geographical coordinates, year of sampling, complete species names and those falling outside Brazilian borders were removed. Additionally, duplicated records—same species name, geographical coordinates and sampling date—were removed. The final database was aggregated into a grid of 0.5° × 0.5° resolution. We used the software QGis v2.18 and “vegan” and “rgbif” R packages (Oksanen et al., 2018; Quantum GIS, 2017; R Core Team, 2018).

2.2 | Sampling efforts

We used the number of sampling events as a surrogate for sampling efforts (Menegotto & Rangel, 2018; Righetti et al., 2020). We regarded as sampling events the unique combinations of longitude, latitude and year of sampling based on species records (Menegotto & Rangel, 2018). We used the year, since it is commonly the only information available about a sampling date. Therefore, records of different species with the same coordinates in a given year were considered a single sampling event (e.g. a group of species collected in a stream reach), while records with equal coordinates but carried out in different years were considered as different sampling events (e.g. temporal sampling in which the same stream is sampled repeatedly over time). By quantifying the sampling events and not the occurrence records of the species, our effort measure is not directly influenced by the differences in geographical patterns of species richness (Vieira et al., 2018).

2.3 | Drivers of spatial biases in sampling efforts

To assess the spatial variation in sampling efforts, we selected factors related to accessibility (population density, distance from terrestrial and fluvial access routes), availability of research infrastructure (distance from research centres) and the attractiveness of protected areas (distance from protected areas).

2.3.1 | Population density (people by km²)

Data for Brazilian municipalities were obtained from the online platform of the Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics), demographic census of 2010 (IBGE: www.ibge.gov.br). The values were resampled to the same resolution of occurrence data (0.5° × 0.5°). For cells containing more than one municipality, we used the mean value for the site in question.

2.3.2 | Terrestrial access routes

Maps of the access routes (highways, streets, and paved and unpaved roads) were obtained from different platforms: Ministry of Environment (mapas.mma.gov.br/), Sistema Estadual de Geoinformação (www.sieg.go.gov.br) and Instituto Brasileiro de Geografia e Estatística (www.ibge.gov.br). The influence of access routes in sampling efforts was evaluated with two measures: the terrestrial access route density in each cell and the distance from each cell to the closest terrestrial access route.

2.3.3 | Fluvial access routes

We considered main rivers of sixth or greater order as fluvial access routes (sensu Strahler, 1957). Usually, rivers of these hierarchies are big enough to permit navigation. Maps of the rivers classified according to Strahler’s hierarchy were obtained from HydroRIVERS database (data available at www.hydrosheds.org; Lehner & Grill, 2013). We used two variables to describe local availability of fluvial routes:
the fluvial access route density in each cell and the minimum distance from each cell to the closest fluvial access route.

2.3.4 | Research centres

Brazilian research centres with the potential to bias freshwater fish sampling efforts were defined as those with research linked to the study of freshwater fish biodiversity. To this end, we surveyed all of the Brazilian graduate programmes of four major areas likely to involve studies on freshwater fish (Environmental Sciences, General Biology, Biodiversity and Fisheries Resources), numbering 262 graduate courses in Brazil. These areas are defined by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES—Coordination for the Improvement of Higher Education Personnel), the governmental body which manages graduate programmes in Brazil. For the programmes included in each major area, the curricula of all the participating professors were assessed regarding their representation in studies on freshwater fish. Finally, we calculated the minimum distance from each cell to the closest research centre containing research on fish.

2.3.5 | Protected areas

We obtained the map of the protected areas from the online platform of the Ministry of Environment (http://mapas.mma.gov.br/), considering all the Brazilian federal and state units of conservation. Subsequently, we calculated the percentage of the protected area. Only the cells containing more than 50% of its area covered by a protected area were regarded for this analysis. Therefore, we calculated the minimum distance of the centroid of each cell, for the centroid of the closest cell containing a protected area.

2.4 | Data analysis

The geographical pattern of sampling efforts in Brazilian freshwater studies was estimated using the spatial correlogram of Moran’s I (Legendre & Legendre, 2012), estimated at 21 equal distance classes. The Moran’s I usually takes values in the interval -1 to +1 and in the correlogram, this coefficient is estimated to each class of distance. Therefore, the visual inspection of the correlogram (i.e. values of Moran’s I to each distance class) allows the evaluation of occurrence of the spatial structure of sampling efforts, in which positive and significative values in first classes and a decrease to last classes (with zero or negative values) indicate spatial gradient in sampling efforts. The significance of Moran’s I was tested using the Monte Carlo procedure with 999 permutations, and the significance of each class was established using Bonferroni’s criterion (see Legendre & Legendre, 2012).

The effect of predictor variables on sampling efforts of fish data was investigated using two approaches: ordinary least square (OLS) and geographically weighted regression (GWR) (Fotheringham et al., 2003). The OLS indicates the global and static angular coefficient along with space, while the GWR takes into account the non-stationarity of predictor variables and generates angular coefficients (and other parameters) for each cell in the grid. The GWR and OLS results were compared using the Akaiake information criterion (AIC).

Initially, the two statistical analyses with all variables were log-transformed (log X + 1) and the predictor variables were standardized using z-score. The z-score allows the comparing of the coefficients of predictor variables. The collinearity of predictor variables was evaluated using variance inflation factor (VIF), and only variables with a VIF < 2.5 were retained in the models. Thus, two variables (distance to terrestrial routes and distance to fluvial routes) were excluded from the models by the VIF criterion.

In GWR analysis, the non-stationarity of the predictor variables was checked using the Monte Carlo procedure and all variables presented non-stationarity (p < .05). The influence of neighbouring areas in GWR was defined by an adaptive Kernel Bi-square function, with bandwidths determined interactively using AIC (see details in Fotheringham et al., 2003). The number of nearest neighbours selected was equal to 101. Local angular coefficients obtained in the GWR were mapped in the grid cell.

The Moran’s I correlogram was performed using the SAM program (Rangel et al., 2010), and the OLS and GWR were performed using software R. The GWR was calculated using the package “GWmodel” (Gollini et al., 2015) and the functions bw.gwr to estimate the bandwidth, montecarlo.gwr to investigate the non-stationarity of predictor variables and gwr.basic to perform the statistics of GWR. The VIF was estimated using the function vif.cor in the package “usdm” (Naimi et al., 2014). Data and R scripts used in our analyses are available in an open-access repository (https://zenodo.org/recorde/3934832#YDUXpGhKilM).

3 | RESULTS

3.1 | Spatial patterns of fish sampling efforts

Our search of the online databases resulted in 339,246 records, numbering 243,901 records after the data filtering process. The GBIF and SpeciesLink databases contributed the largest number of records (Table S1)—from 1900 to 1964, respectively. From these data, we identified a total of 43,538 sampling events distributed across Brazil (Figure 1a), with a higher number of records in southern and south-eastern Brazil, even though all of the regions had sites with high sampling efforts. The sampling efforts presented a clear spatial structure, with geographically near sites having a similar number of sampling events. The spatial correlogram indicated that the sites with a distance of <215 km (first distance class in correlogram) presented the highest similarity among the number of sampling events (Figure S1). Distribution of protected areas, research centres, population density, terrestrial and fluvial access routes is also spatially uneven (Figure 1b–f).
3.2 | Modelling fish sampling efforts

The GWR presented better performance than OLS according to statistics of $R^2$ and AIC (Table 1), showing the occurrence of non-stationarity in the effect of predictor variables in explaining the number of sampling events. Considering the median values of GWR results, the population density and terrestrial route access density were the variables with a higher effect on the number of sampling events (Table 1). Regions with higher density of terrestrial access routes and population density attracted a higher number of sampling events. Moreover, the higher the density of fluvial access routes and the closer a cell is to protected areas and to centres of research, the higher its number of sampling events (Table 1).

Local GWR results showed different predictive powers of the model across sites (Figure 2). The map of $R^2$ revealed that some regions of Brazil presented the highest values ($R^2 > .75$), indicating, for example, that in these regions the predictors explained 75% (or more) of the variability of sampling efforts (Figure 2a). Worth mentioning is the fact that the importance of each predictor needs to be evaluated considering the importance of the model in the same region (local $R^2$). Despite median values to reveal the dominant direction of the driver-sampling efforts relationship, local GWR coefficients showed strong explanatory influence of drivers in some
regions, but not in others (e.g. strongest influence of fluvial access route density in the Amazon). Furthermore, some regions had opposite trends (Figure 2). While terrestrial and fluvial access route density and population density increased sampling efforts in the majority of sites, some regions exhibited the opposite relationship (e.g. in the south-east of Brazil for terrestrial access route density; some sites in the south for fluvial access route density and in the north for population density) (Figure 2). For the distance of protected areas and centres of research, a positive relationship was also observed in some sites in the central-western and north-western regions, contrasting with the overall tendency of negative relationships for these variables (Figure 2).

4 | DISCUSSION

Sampling efforts have enhanced over time for almost all the biological groups and regions, resulting in an expressive cumulated number of species records (Girardello et al., 2018; Oliveira et al., 2016). This is also true for Brazilian freshwater fishes, as shown by the large number of species recorded on databases and sampling efforts employed (Dias et al., 2016). However, our results show that freshwater fish samplings are spatially biased with a tendency for major efforts to be concentrated in sites with a higher density of fluvial and terrestrial access routes, larger populations and near to research centres and protected areas. These results suggest that the well-known factors causing sampling bias for terrestrial organisms (Daru et al., 2017; Oliveira et al., 2016; Romo et al., 2006) are also influential for sampling of freshwater organisms. However, the importance of these drivers varied across sites, revealing the spatial non-stationarity of these effects. Drivers were strongly predictive of fish sampling events in regions with high sampling efforts (southern and south-eastern Brazil) and in areas with low sampling efforts (e.g. part of the north of Brazil).

In line with our expectations, the GWR model performed better than OLS in predicting fish sampling efforts. The former assumes the differential effect of explanatory variables on sampling events across space (i.e. non-stationarity), while the OLS model assumes a consistent relationship of variables along with space. The OLS model appears to have highlighted the influence of predictors at the broad scale which are illustrated by the contrast between the highest sampling efforts in southern and south-eastern Brazil, which presents high population and road density and contains the highest concentration of research centres in Brazil. However, stationary models such as OLS do not highlight variations in local relationships. In fact, the GWR model revealed that the predictive power of the variables in explaining sampling efforts varied spatially, as shown by local $R^2$ (GWR model). While some variables, such as population density and distance of protected areas, seem to show consistent relationship with sampling efforts, others, such as access route density (fluvial and terrestrial) and distance of research centre, show a clear non-stationary relationship. Therefore, the variability of sampling efforts in freshwaters may result from complex and scale-dependent processes.

The importance of terrestrial access routes to predict sampling effort when taking into account the non-stationarity was remarkable. In fact, this variable had the greater range of local coefficients of the GWR model, indicating contrasting relationships of this variable with the sampling effort across the studied area. The positive association between terrestrial access route density and sampling effort occurred especially in regions with low or moderate density of access routes and where there is a broader gradient of this variable. In those regions, the terrestrial access routes represent a key driver influencing sampling. In regions where most areas have a high density of access routes (e.g. south-eastern Brazil), accessibility seems to be a constraint, which is overcome. Carrying out samplings close to access routes generates spatial bias and causes a failure in the representativeness of environmental and biological diversity (Kadmon et al., 2004; Oliveira et al., 2016). Furthermore, access routes have been associated with degradation of the riparian zone, loss of physical habitat integrity, an increase in fishing and the introduction of exotic species (Trombulak & Frissell, 2000). These modifications favour the occurrence of less sensitive species that may not represent regional biodiversity.
Fluvial access routes may favour fish sampling especially in regions with low terrestrial accessibility, as seems to be the case in Occidental Amazonia (northern Brazil), for which the highest positive coefficients of the GWR model were found. In this region, travel is easiest via waterway transport. Most of the samplings are distributed in or near navigable rivers and the high sampling effort is the result of the interest of researchers in sampling the richest fish diversity in the world, which is found in this region (Oberdorff et al., 2019; Tisseuil et al., 2013). The negative relationships of sampling effort with fluvial access route density are probably associated with the idiosyncratic influence of other factors attracting researchers farthest from the main rivers.

The population density was the second variable most strongly associated with sampling effort in both models. Based on the stationary model, it is valid to assume that most areas with greater population density contain more sampled sites, contrasting with sparsely populated areas which receive few sampling efforts. In fact, most of the state capitals and other heavily populated sites received high sampling efforts. These major urban centres provide greater infrastructure and human resources, which could favour the
sampling efforts (Lessa et al., 2019). Furthermore, urbanization has aroused researchers’ interest in concentrating their studies on these sites since such urbanized ecosystems can present high biodiversity (Luck, 2007) and habitat heterogeneity (Araújo, 2003; Pautasso & McKinney, 2007). However, the population density effect is also relevant in less populated regions, as shown by our non-stationary model in some regions (e.g. Amazon), where minor differences in population density are also relevant to predict sampling effort. Exceptions to the overall pattern (i.e. negative association between population density and sampling effort) may be observed when researchers are attracted to less populated regions, usually when other drivers exert a stronger effect. It occurs, for example, when there is the presence of research centres in less populated areas or when large projects fund biodiversity collection in remote regions (Jezéquel et al., 2019).

Our results, and those found by Sánchez-Fernández et al., (2008) for aquatic beetles, reinforce the tendency to concentrate samplings close to research centres. The search for sampling sites close to the worksite is a convenience that makes the research cheaper and the logistics easier for the researchers (Moerman & Estabrook, 2006). In fact, almost all sites surrounding research centres showed significant local coefficients in the GWR model. An example is the intensely sampled area surrounding the capital of the state of Amazonas (lowest coefficients in Figure 2c), where traditional ichthyological research centres are located (e.g. National Institute of Amazonian Research, INPA) (Freitas et al., 2021; Zuanon et al., 2016). These results highlight the importance of research centres as vectors for biodiversity samplings and scientific development (e.g. Dias et al., 2016; Nabout et al., 2015). It is worth mentioning that the strength of the research centre sampling relationship may still have been underestimated since our criterion for the definition of a research centre was based on the presence of a graduate programme. Thus, sampling efforts carried out by researchers working in universities or research agencies that are not part of graduate programmes may not have been credited to the research centre effect, which may be the case for part of the sites with positive coefficients.

The association of protected areas with sampling effort found here reinforces the attractiveness effect of such protected areas (Sánchez-Fernández et al., 2008). Predictions of both models (OLS and GWR) were similar, reflecting the high consistency of this effect. The attractiveness of protected areas may be explained by the tendency of researchers to seek less disturbed sites, areas with high species richness and endemism, features usually associated with protected areas (Reddy & Dávalos, 2003; Sánchez-Fernández et al., 2008). In regions with higher degrees of human influence, such as the south and south-east, just a few protected areas with native vegetation remain. These areas may represent the last refuge for many species, including threatened species (Abell et al., 2007), attracting researchers’ efforts to conduct sampling in these areas. Additionally, as a legal requirement of Brazilian federal legislation, the planning, creation, expansion and management of conservation units demand studies and inventories of biodiversity, usually involving fish sampling.

Fish sampling efforts were weakly predicted by our models in some regions, principally in some sites with low sampling efforts. While this unpredictability may be the result of evenness in the sampling efforts, other factors may be considered: (i) in some cases, as in north-eastern Brazil, there is a spatial decoupling between the greater portion of the hydrographic network that drains the interior of the continent and the coastal region where major cities with most of the research centres, infrastructure and human resources are located (Lessa et al., 2019); (ii) occurrence of sampling efforts not recorded in online databases and hence not considered by us. While the online availability of data is a potential issue (Blair et al., 2020), the availability of high-quality biodiversity data within Brazil is aided by the initiative of the SpeciesLink network, which brings together collections nationwide; and (iii) sampling efforts are predicted by other biasing factors not considered by us, such as intermittency of rivers and streams which could make sampling difficult.

Spatial bias is responsible not only for species distribution gaps (Wallacean shortfall) but also for hampering the knowledge of existing species, as sites with no or low sample efforts can harbour a high number of undescribed species (Diniz-Filho et al., 2005; Dutra et al., 2020). This can be especially true for remote regions with singular environmental conditions, where the lack of sampling may lead to missing rare and endangered species (Nogueira et al., 2010; da Silva et al., 2020). Langeani et al. (2007) highlighted that even for the most sampled ecoregion of Brazil (Upper Paraná, south-eastern region), the number of described fish species had increased considerably in the last decade. Thus, we may anticipate that many more species are yet to be discovered in other regions, especially those with few sampling efforts. Therefore, overcoming Linnean and Wallacean knowledge gaps for freshwater fish species in megadiverse regions such as Brazil depends on investing in sampling efforts in less accessible and populated sites, farther from research centres and well-studied protected areas. However, it should be highlighted that the influence of most drivers varies considerably in space and thus such variability should be taken into account when defining new areas for sampling, as the targeting of reducing a specific source of bias cannot have the same effectiveness in all areas.

ACKNOWLEDGEMENTS

We thank Murilo S. Dias, Rodrigo A. Carvalho and Vitor Hugo M. Prado for helpful suggestions in the early version of the manuscript and to Matthijs Strietman for revision of the English of this document. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001 (master’s fellowship to TC and postdoctoral fellowship to GT). GT was also supported by UEG/CAPES no. 817164/2015—PROAP. JCN and FBT receive CNPq fellowship (305798/2019-7, 306912/2018-0, respectively). This study was supported by INCT in Ecology.
Evolution and Biodiversity Conservation, funded by MCTIC/CNPq (proc. 465610/2014-5).

PEER REVIEW
The peer review history for this article is available at https://publon ns.com/publon/10.1111/ddi.13269.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available in Zenodo at https://zenodo.org/record/3934832#.YDUXpGhKiILM. These data were derived from the following resources available in the public domain: Global Biodiversity Information Facility (GBIF; www.gbif.org), SpeciesLink (www.splink.org.br), Portal da Biodiversidade (portaldabiodiversidade.icmbio.gov.br/portal/), Sistema de Informação sobre a Biodiversidade Brasileira (SBBR; www.sibbr.gov.br/), Brazilian Institute of Geography and Statistics (IBGE; www.ibge.gov.br), Ministry of Environment (mapas.mma.gov.br/), Sistema Estadual de Geoinformação (www.sieg.go.gov.br) and HydroRIVERS database (www.hydrosheds.org).

ORCID
Thaityane Caetano de Almeida https://orcid.org/0000-0002-5075-4797
Geiziane Tesserolo https://orcid.org/0000-0003-1361-0062
João Carlos Nabout https://orcid.org/0000-0001-9102-3627
Fabricio Barreto Teresa https://orcid.org/0000-0002-1357-4391

REFERENCES
Abell, R., Allan, J. D., & Lehner, B. (2007). Unlocking the potential of protected areas for freshwater. Biological Conservation, 134, 48–63.
Amano, T., & Sutherland, W. J. (2013). Four barriers to the global understanding of biodiversity conservation: Wealth, language, geographical location and security. Proceedings of the Royal Society B: Biological Sciences, 280, 20122649.
Anderson, R. P. (2012). Harnessing the world’s biodiversity data: Promise and peril in ecological niche modeling of species distributions. Annals of the New York Academy of Sciences, 1260, 66–80. https://doi.org/10.1111/j.1749-6632.2011.06440.x
Araújo, M. (2003). The coincidence of people and biodiversity in Europe. Global Ecology & Biogeography, 12, 5–12. https://doi.org/10.1046/j.1466-822X.2003.00314.x
Azevedo-Santos, V. M., Pelicice, F. M., Lima-Junior, D. P., Magalhães, A. L. B., Orsi, M. L., Vitule, J. R. S., & Agotinho, A. A. (2015). How to avoid fish introductions in Brazil: Education and information as alternatives. Natureza & Conservação, 13, 123–132. https://doi.org/10.1016/j.njcon.2015.06.002
Bini, L. M., Diniz-Filho, J. A. F., Rangel, T. F. L. V. B., Bastos, R. P., & Pinto, M. P. (2006). Challenging Wallacean and Linnean shortfalls: Knowledge gradients and conservation planning in a biodiversity hotspot. Diversity and Distributions, 12, 475–482. https://doi.org/10.1111/j.1366-9516.2006.00286.x
Blair, J., Gwiazdowski, R., Borrelli, A., Hotchkiss, M., Park, C., Perrett, G., & Hanne, R. (2020). Towards a catalogue of biodiversity databases: An ontological case study. Biodiversity Data Journal, 8, e32765. https://doi.org/10.3897/BDJ.8.e32765
Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-qing, D., Clark, N. E., O’Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. Plos Biology, 8, e1000385.– https://doi.org/10.1371/journal.pbio.1000385
Botts, E. A., Barend, F. N., Erasmus, B. F. N., & Alexander, G. J. (2011). Geographic sampling bias in the South African Frog Atlas Project: Implications for conservation planning. Biodiversity and Conservation, 20, 119–139. https://doi.org/10.1007/s10531-010-9950-6
Boyero, L., Pearson, R. G., Dudgeon, D., Ferreira, V., Graça, M. A. S., Gessner, M. O., Boulton, A. J., Chauvet, E., Yule, C. M., Albariño, R. J., Ramírez, A., Nelson, J. E., Callisto, M., Aruachalam, M., Chará, J., Figueroa, R., Mathookie, J. M., Gonçalves Jr, J. F., Moretti, M. S., ... Barmuta, L. A. (2012). Global patterns of stream detritivore distribution: Implications for biodiversity loss in changing climates. Global Ecology and Biogeography, 21, 134–141. https://doi.org/10.1111/j.1466-8238.2011.00673.x
Brooks, T. H., Fonseca, G. A. B., & Rodrigues, A. S. L. (2004). Protected areas and species. Conservation Biology, 18, 616–618. https://doi.org/10.1111/j.1523-1739.2004.00183.x
Caetano, J. M., Tesserolo, G., de Oliveira, G., Souza, K. D. S., Diniz-Filho, J. A. F., & Nabout, J. C. (2018). Geographical patterns in climate and agricultural technology drive soybean productivity in Brazil. PLoS One, 13, e0191273.– https://doi.org/10.1371/journal.pone.0191273
Canhos, D. A. L., Sousa-Baena, S., de Souza, S., Maia, L. C., Stehmann, J. R., Canhos, V. P., De Giovanni, R., Bonacelli, M. B. M., Los, W., & Peterson, A. T. (2015). The importance of biodiversity e-infrastructures for megadiverse countries. PLoS Biology, 13, e1002204.– https://doi.org/10.1371/journal.pbio.1002204
Collen, B., Whitton, F., Dyer, E. E., Baillie, J. E. M., Cumberlidge, N., Darwall, W. R. T., Pollock, C., Richman, N. I., Soulsby, A. M., & Böhm, M. (2013). Global patterns of freshwater species diversity, threat and endemicism. Global Ecology and Biogeography, 23, 40–50. doi:10.1111/glob.12096
Correira, R. A., Ruete, A., Stropp, J., Malhado, A. C., dos Santos, J. W., Lessa, T., & Ladle, R. J. (2019). Using ignorance scores to explore biodiversity recording efforts for multiple taxa in the Caatinga. Ecological Indicators, 106, 105539.
da Silva, A. F., Malhado, A. C. M., Correira, R. A., Ladle, R. J., Vital, M. V. C., & Mott, T. (2020). Taxonomic bias in amphibian research: Are researchers responding to conservation need? Journal for Nature Conservation, 56, 125829.– https:// doi.org/10.1016/j.jnc.2020.125829
Daru, B. H., Park, D. S., Primack, R. B., Willis, C. G., Barrington, D. S., Whitfeld, T. J. S., Seidler, T. G., Sweeney, P. W., Foster, D. R., Ellison, A. M., & Davis, C. (2017). Widespread sampling biases in herbaria revealed from large-scale digitization. New Phytologist, 217, 939–955.
Dias, M. S., Zuanon, J., Couto, T. B. A., Carvalho, M., Carvalho, L. N., Espírito-Santo, H. M. V., Frederico, R., Leitão, R. P., Mortati, A. F., Pires, T. H. S., Torrente-Vilara, G., Vale, J. D., Anjos, M. B. D., Mendonça, F. P., & Tedesco, P. A. (2016). Trends in studies of Brazilian stream fish assemblages. Natureza & Conservação, 14, 106–111. https://doi.org/10.1111/njcon.2016.06.003
Diniz-Filho, J. A. F., Bastos, R. P., Rangel, T. F. L. V. B., Bini, L. M., Carvalho, P., & Silva, R. J. (2005). Macroecological correlates and spatial patterns of anuran description dates in the Brazilian Cerrado. Global Ecology and Biogeography, 14, 469–477.
Dutra, G. M., Freitas, T. M. S., Prudente, B. S., Salvador, G. N., Leão, M. D. V., Peixoto, L. A. W., Mendonça, M. B. N., Netto-ferreira, A. L. Silva, F. R., Montag, L. F. A., & Wosiacki, W. B. (2020). Rapid assessment of the ichthyofauna of the southern Guiana Shield tributaries of the Amazonas river in Pará, Brazil. Acta Amazonica, 50, 24–36. https:// doi.org/10.1590/1809-4392201901402
Eme, D., Zagmajster, M., Fíšer, C., Galassi, D., Marmonier, P., Stoch, F., Corru, J.-F., Oberdorff, T., & Malard, F. (2015). Multi-causality and spatial non-stationarity in the determinants of groundwater crus-tacean diversity in Europe. Ecography, 38, 531–540. https://doi.org/10.1111/ecog.01092
Foody, G. M. (2004). Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. Global Ecology and Biogeography, 13, 315–320. https://doi.org/10.1111/j.1466-822X.2004.00097.x

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). Geographically weighted regression: The analysis of spatially varying relationships. Wiley.

Freitas, T. M. S., Stropp, J., Calegari, B. B., Calatauyud, J., De Marco-Jr, P., Montag, L. F. A., & Hortal, J. (2021). Quantifying shortfalls in the knowledge on Neotropical Auchenipteridae fishes. Fish and Fisheries, 22, 87–104. https://doi.org/10.1111/faf.12507

Gaijii, S., Chavan, V., Aríño, A. H., Otegui, J., Hobern, D., Sood, R., & Robles, E. (2013). Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potential. Biodiversity Informatics, 8, 94–172.

Gallardo, B., Bogan, A. E., Harun, S., Jainih, L., Lopes-Lima, M., Pizarro, M., Rahim, K. A., Sousa, R., Virdis, S. G. P., & Zieritz, A. (2018). Current and future effects of global change on a hotspot’s freshwater diversity. Science of the Total Environment, 635, 750–760. https://doi.org/10.1016/j.scitotenv.2018.04.056

Girardello, M., Martellos, S., Pardo, A., & Bertolino, S. (2018). Gaps in biodiversity occurrence information may hamper the achievement of international biodiversity targets: Insights from a cross-taxon analysis. Environmental Conservation, 45, 370–377. https://doi.org/10.1017/S0376892918000115

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2015). GWmodel: An R Package for exploring spatial heterogeneity using geographically weighted models. Journal of Statistical Software, 63(17), 1–50.

Hecnar, S. J., & M’Closkey, R. T. (1997). Spatial scale and determination of species status of the green frog. Conservation Biology, 11, 670–682.

Hortal, J., Bello, F., Diniz-Filho, J. A., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven shortfalls that beset large-scale knowledge of biodiversity. Annual Review of Ecology, Evolution, and Systematics, 46, 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Hortal, J., Lobo, J. M., & Jimenez-Valverde, A. (2007). Limitations of biodiversity databases: Case study on seed-plant diversity in Tenerife, Canary Islands. Conservation Biology, 21, 853–863. https://doi.org/10.1111/j.1523-1739.2007.00686.x

Howard, J. K., Klausmeyer, K. R., Fesenmyer, K. A., Furnish, J., Gardali, J., Torrente-Vilara, G., Sousa, L. M., Guiarrizzo, T., Valdiviezo, J., Sarmiento, J., Tedesco, P. A., Ramos Neto, M. B., & da Silva, J. M. (2010). Restricted-range fishes and the conservation of Brazilian freshwaters. PLoS One, 5, e11390.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O’Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2018). Vegan: Community ecology package. R package version 2.5-20. 2018.

Paudtasso, M., & McKinney, M. L. (2007). The botanist effect revisited: Plant species richness, county area and human population size in the United States. Conservation Biology, 21, 1333–1340.

Pelayo-Villamil, P., Guisande, C., Manjarrés-Hernández, A., Jiménez, L. F., Granado-Lorencio, C., García-Roselló, E., González-Dacosta, J., Heine, J., González-Vilas, L., & Lobo, J. M. (2018). Completeness of national freshwater fish species inventories around the
world. Biodiversity and Conservation, 27, 3807–3817. https://doi.org/10.1007/s10531-018-1630-y
Quantum GIS. (2017). Development Team. 2015. Quantum GIS Geographic Information System. V. 2.18. Open Source Geospatial Foundation Project.
R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/
Rangel, T. F., Diniz-Filho, J. A. F., & Bini, L. M. (2010). SAM: A comprehensive application for spatial analysis in macroecology. Ecography, 33, 46–50. https://doi.org/10.1111/j.1600-0587.2009.06299.x
Rapp Py-Daniel, L. H., Ito, P. M. M., Ota, R. P., Soares, I. M., Bastos, D. A., & Hashimoto, S. (2015). 30 anos da maior coleção de peixes amazônicos: A coleção de peixes do INPA. Boletim Da Sociedade Brasileira De Ictiologia, 116, 4–16.
Reddy, S., & Dávalos, L. M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography, 30, 1719–1727. https://doi.org/10.1046/j.1365-2699.2003.00946.x
Reis, R. E., Kullander, S. O., & Ferraris-Jr, C. J. (2003). Check list of the freshwater fishes of South and Central America. Edipucrs.
Righetti, D., Vogt, M., Zimmermann, N. E., Guiy, M. D., & Gruber, N. (2020). PhytoBase: A global synthesis of open-ocean phytoplankton occurrences. Earth System Science Data, 12, 907–933. https://doi.org/10.5194/essd-12-907-2020
Romo, H., Barros, E. G., & Lobo, J. M. (2006). Identifying recorder-induced geographic bias in an Iberian butterfly database. Ecography, 29, 873–885. https://doi.org/10.1111/j.2006.0906-7590.04680.x
Sánchez-Fernández, D., Lobo, J. M., Abellán, P., Ribera, I., & Millán, A. (2008). Bias in freshwater biodiversity sampling: The case of Iberian water beetles. Diversity and Distributions, 14, 754–762. https://doi.org/10.1111/j.1472-4642.2008.00474.x
Sastre, P., & Lobo, J. M. (2009). Taxonomist survey biases and the unveiling of biodiversity patterns. Biological Conservation, 142, 462–467. https://doi.org/10.1016/j.biocon.2008.11.002
Strahler, A. N. (1957). Quantitative analysis of watershed geomorphology. Eos Transactions, American Geophysical Union, 38, 913–920. https://doi.org/10.1029/TR038i006p00913
Strayer, D. L., & Dudgeon, D. (2010). Freshwater biodiversity conservation: Recent progress and future challenges. Journal of the North American Benthological Society, 29, 344–358. https://doi.org/10.1899/08-171.1
Stropp, J., Ladle, R. J., M. Malhado, A. C., Hortal, J., Gaffuri, J., H. Temperley, W., Olav Skøien, J., & Mayaux, P. (2016). Mapping ignorance: 300 years of collecting flowering plants in Africa. Global Ecology and Biogeography, 25, 1085–1096. https://doi.org/10.1111/greb.12468
Terlibe, L. C., & Diniz-Filho, J. A. F. (2009). Spatial patterns of species richness in New World coral snakes and the metabolic theory of ecology. Acta Oecologica, 35, 163–173. https://doi.org/10.1016/j.actao.2008.09.006
Tessarolo, G., Ladle, R., Rangel, T., & Hortal, J. (2017). Temporal degradation of data limits biodiversity research. Ecology and Evolution, 7, 6863–6870. https://doi.org/10.1002/ece3.3259
Tisselui, C., Cornu, J.-F., Beauchard, O., Brosse, S., Darwall, W., Holland, R., Hugueny, B., Tedesco, P. A., & Oberdorff, T. (2013). Global diversity patterns and cross-taxa convergence in freshwater systems. Journal of Animal Ecology, 82, 365–376. https://doi.org/10.1111/1365-2656.12018
Troia, M. J., & McManamay, R. A. (2017). Completeness and coverage of open-access freshwater fish distribution data in the United States. Diversity and Distributions, 23, 1482–1498. https://doi.org/10.1111/ddi.12637
Trombulak, S. C., & Frissell, C. A. (2000). Review of ecological effects of roads on terrestrial and aquatic communities. Conservation Biology, 14, 18–30. https://doi.org/10.1046/j.1523-1739.2000.99084.x
Vieira, T. B., Pavaneli, C. S., Casatti, L., Smith, W. S., Benedito, E., Mazzoni, R., Sánchez- Botero, J. I., Garcez, D. S., Lima, S. M. Q., Pompeu, P. S., Agostinho, C. S., Montag, L. F. D. A., Zuanon, J., Aquino, P. D. P. U. D., Cetra, M., Tejerina-Garro, F. L., Duboc, L. F., Corrêa, R. C., Pérez-Mayorga, M. A., ... De Marco, P. (2018). A multiple hypothesis approach to explain species richness patterns in neotropical stream-dweller fish communities. PLoS One, 13, e0204114. https://doi.org/10.1371/journal.pone.0204114
Vigerstol, K. L., & Aukema, J. E. (2011). A comparison of tools for modelling freshwater ecosystem services. Journal of Environmental Management, 92, 2403–2409. https://doi.org/10.1016/j.jenvman.2011.06.040
Whittaker, R. J., Araújo, M. B., Jepson, P., Ladle, R. J., Watson, J. E. M., & Willis, K. J. (2005). Conservation biogeography: Assessment and prospect. Diversity and Distributions, 11, 3–23. https://doi.org/10.1111/j.1365-2972.2005.00143.x
Yang, W., Keping, M. A., & Kreft, H. (2013). Geographical sampling bias in a large distributional database and its effects on species richness–environment models. Journal of Biogeography, 40, 1415–1426. https://doi.org/10.1111/jbi.12108
Zuanon, J., Mendonça, F. P., Espírito-Santo, H. M. V., Dias, M. S., Galuch, A. V., & Akama, A. (2016). Guia de Peixes da Reserva Ducke, Amazônia Central. Instituto Nacional de Pesquisas da Amazonia.