Ecological Informatics

Spatial and temporal representation of marine fish occurrences available online

--Manuscript Draft--

Manuscript Number:	ECOINF-D-23-00167R3
Article Type:	Research Paper
Keywords:	Ecoinformatics; Ecological Information Biases; Marine Fish; Spatial and Temporal Representativeness; Species Richness
Corresponding Author:	Horacio Samaniego, PhD Austral University of Chile - Campus Isla Teja CHILE
First Author:	Vanessa Pizarro
Order of Authors:	Vanessa Pizarro
	Andrea G Castillo
	Andrea Piñones, PhD
	Horacio Samaniego, PhD
Abstract:	Despite the 243,000 marine species described by 2022, our knowledge about the oceanic biodiversity is still incomplete. This knowledge gap carries potentially adverse and far-reaching consequences for the preservation of marine ecosystems, particularly in the context of the ongoing human-induced alterations to our biosphere and the rapid progression of climate change and global environmental shifts. Recently, however, a large number of online repositories have emerged, which catalogue, store and distribute biodiversity information, including taxonomic and species occurrence data. FishBase, the Global Biodiversity Information Facility (GBIF) and the Ocean Biodiversity Information System (OBIS) are part of these publicly available repositories representing a variety of sources that have exploded in number. However, despite the incredible accumulation of biodiversity records, not all the information is actually useful, nor does it represent any new knowledge regarding global species richness patterns. In this study, we assessed the spatial and temporal representativeness of marine fish records (order Actinopterygii) found in the GBIF and OBIS global repositories. The methodological framework that we developed relies on a series of non-parametric estimators for computing species richness from incidence data. This methodology employs hexagonal grids as sampling units that overlay marine bioregions across the globe. Using standard ecological and spatial analysis tools, we identify regions that are adequately represented in terms of available records and therefore have more reliable data, as well as regions with few records that do not represent current species richness. We overlap these results with the location of marine protected areas and fishing exploitation zones to understand the anthropogenic effect on marine ichthyofauna. We additionally evaluate hypotheses regarding the taxonomic, geographic, and temporal distribution of information biases to deepen our current understanding of public records of species occurrences worldwide. Considering that more than 40 years of information was analyzed, the results showed that, on a global scale, the primary data on marine fish available on GBIF and OBIS platforms are still far from being representative and complete. Only 1.14\% of the records were useful for our analyses. In addition, we found that the information seems to be biased towards coastal areas, regions close to developed countries, and areas where there is a large fishing activity. Finally, the best represented species and families are those with a small body size, which use shallow habitats and are usually recognized as having commercial or cultural value.
Suggested Reviewers:	Camilo Mora professor, University of Hawai'i System moracamilo@hotmail.com
	Derek P Tittensor Dalhousie University

	derekt@mathstat.dal.ca
	Derek Corcoran Aarhus University derek.corcoran@bio.au.dk
	Signe Normand Aarhus University signe.normand@bio.au.dk
	Duccio Rocchini University of Bologna duccio.rocchini@unibo.it
Response to Reviewers:	

Valdivia, November 30th 2023

Dear Editor,

I am submitting this rebuttal in response to the reviewers' comments on our manuscript titled "Spatial and Temporal Representation of Marine Fish Occurrences," which is currently accessible online.

We express our gratitude for the opportunity to address the minor revisions brought forth by the reviewers regarding our contribution. We have conscientiously integrated their feedback, a process that has bolstered the quality and robustness of our manuscript. It is noteworthy that we sought additional professional assessment of the manuscript from a native English speaker.

Sincerely,

Horacio Samaniego

Universidad Austral de Chile

horacio@ecoinformatica.cl

Reviewer #1: The authors have again incrementally improved the manuscript. I remain unconvinced of the utility of the SRI index given its fundamental issues when the denominator is a low value.

The classification scheme remains problematic. "Adequate" is a synonym of "sufficient" so different terms need to be chosen. If a classification is strictly needed then why not something like high, medium and low?
Thanks for this suggestion, we have now changed the classification to high, medium and low representativeness all through the text.

It remains unclear to me why the authors are ignoring the confidence intervals that are provided by the richness estimation indices. Chao2, for example, has well defined formulae that are fairly easily calculated, and which are already provided by most implementations.
While this is certainly a possibility, we have chosen to follow the literature that is using the mean of several richness estimators. All of these have a different statistical method to quantify richness which makes them unsuitable to include a CI. As discussed in the ms, Mora et all 2008 provides in-depth discussion on how different richness estimators behave and proposes the mean as a good compromise to obtain a smoothed richness curve. While we understand and recognize the concern to include a CI estimation, it would require evaluating how the mean of richness indices compare to the CI resulted from the Chao2

estimator, for example. This is beyond the scope of our particular analysis, as we are simply attempting to evaluate the spatial distribution of species representativeness gathered from online global repositories.

There remain many issues with the English expression. I do appreciate the challenges of writing in a secondary language. However, this journal has an international audience so the text needs to be readily comprehended by readers from other regions for whom English is also a secondary language.
The ms was now professionally reviewed for its English form.

AUTHOR RESPONSES

1. "Nonetheless, we maintain our classification, as we believe it offers a more accurate representation of the relative knowledge of species occurrence numbers and contributes to maintaining the manuscript's focus. It is important to underscore that while the ratios of 12/24th and ½ do indeed signify a considerable disparity in the number of missing species, the primary objective of this study is not to provide an exhaustive enumeration of species absent from the database records. Rather, our aim is to furnish a macro-level description of the most substantial differences between observed and estimated species richness, facilitated by the Species Richness Index (SRI)."

Perhaps this is an issue of word choice, but there is nothing accurate about this classification. This response also does not address the fundamental issue of ratios with small denominators. Why not apply a secondary filter to exclude cells where there are few observed species? Or perhaps check the data to see how many cells have low denominators and where they are? If there are few such cells then the issue does not greatly affect the conclusions.

This was addressed by adding a category "IR" (insufficient records), precisely to identify those cells that only have a single record and therefore generate this low denominator problem. As indicated in the manuscript, these cells marked with IR were excluded from the analysis.

And what do the results look like if one uses the difference of expected from observed? Sometime the absolute difference is the topic of interest.

To address this, we have included a new figure in the appendix showing the log10 of the absolute difference, as suggested.

Colour scheme: Log scaling does not always work, but there are many other alternatives. For example one could use quantiles to classify the data. The main objective is that a small number of very large values do not dominate the colour range, obscuring potentially interesting patterns.

The unique purpose of Fig. 1 is to provide visual support to a) the location bioregions, and b) a sense of the magnitude of richness within them. The detailed, and raw, richness data for each bioregion is available in Table 1. Log transform is here used to minimize the orders of magnitude difference between the richness, family and Shannon diversity in one panel.

2. Averaging richness estimators.
Why not show the bounds? Knowing where the metrics are consistent and where there is a high disparity is useful information in itself.

See the comments below

4. Classes.
This response does not address the problem.

6. OBIS and GBIF.
The response does not address the issue. I have since confirmed through other sources that OBIS is not a complete subset of GBIF, but the overlap remains. Please clarify in the MS text how this was dealt with.

We have clarified this in the ms. and included new references providing support for our choice. It now reads as follow**:** “Both repositories have collaborated since 2001, sharing data on the co-occurrence of marine life.(OBIS, 2021). However, Recent studies have shown unequal contributions between the two repositories, with very low percentages of shared data (Moudrý & Devillers, 2020; Chollett & Robertson, 2020). Also, both platforms include different data sources and methodologies, as well as important differences in temporal and spatial scales at which data were collected (Ziska *et al.*, 2020). It is for these reasons that some authors suggest carefully examining and purifying these data repositories (Bonnet-Lebrun *et al.*, 2023). We have applied a series of filters to this information, first …

References:

Bonnet-Lebrun, A. S., Sweetlove, M., Griffiths, H. J., Sumner, M., Provoost, P., Raymond, B., ... & Van de Putte, A. P. (2023). Opportunities and limitations of large open biodiversity occurrence databases in the context of a Marine Ecosystem Assessment of the Southern Ocean. *Frontiers in Marine Science*, *10*, 1150603.
https://doi.org/10.3389/fmars.2023.1150603

Moudrý, V., & Devillers, R. (2020). Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. *Ecological Informatics*, *56*, 101051. DOI: https://doi.org/10.1016/j.ecoinf.2020.101051

Chollett, I., & Robertson, D. R. (2020). Comparing biodiversity databases: Greater Caribbean reef fishes as a case study. *Fish and Fisheries*, *21*(6), 1195-1212. DOI: https://doi.org/10.1111/faf.12497

19. This does not address the issue that the equal area cylindrical projection is not ideal.
While your statement is accurate, it's important to note that all projections come with certain limitations."

23. Use the actual values not the normalised ones. Make it easy for the reader to understand the data represented in the figure.

Please see the previous response above, regarding Fig 1. and Table1

SPECIFIC COMMENTS

Page and line numbers are for the version with tracked changes.

24. P2, L27. What is "commercial cultural"?
Thanks, we have reworded this last sentence.

25. P3, L11. Species richness does not subsume the overall variety of life. It is one metric amongst many, and in many ways is simplistic given it is subject to taxonomic biases such as lumping and splitting. It also does not account for phylogenetic or functional similarity amongst a sample. Same for P4, L3.

Thank you for bringing this to our attention. While it might appear to be more of a semantic discussion, some may contend that richness, distinct from other diversity metrics, offers a specific and raw quantification of the outcomes of evolutionary processes. Despite not accounting for the splitting and lumping in the tree of life, it inherently reflects the consequences of these processes. This adjustment in wording has now been made to encapsulate this perspective.

26. P7, L2. "Any record with NA values was removed". Do you mean an NA value for any field?
Thank you, we have rephrased the sentence to clarify that any data with NA in the mentioned columns was removed.

27. P7, L6. singular-plural disagreement: "a marine bioregions". (There are many such issues in the MS that need to be corrected).
Thanks for your comment. We corrected it.

28. P7, L16. One does not create a subset. One can, however, extract it.
Thanks for your comment. We corrected it.

29. P7, L23. No need to note WGS1984 here as it is part of the coordinate system definition.
Thanks for your comment. We corrected it.

30. P7, L3. The formula is not needed. Just refer to the mean instead of average on the preceding line.
Thanks for your comment. We corrected it.

31. P7, L5. Averaging will lessen the effect of outliers but not minimise them given there are only three indices and they are each equally weighted. If one of the indices is highly biased then the average will still be biased.
Please see response to the 2nd general comment above.

32. L7, L11. The SRI does not measure the actual species richness. It is a ratio of the observed to estimated richness and is an index of undersampling, one that is affected by the low denominator problem.
Thanks, we have included a small paragraph addressing this in the discussion, which read: "*Striving for simplicity, we employ the ratio of observed to expected species richness (SRI) as a means to indicate the spatial distribution of undersampled regions. While acknowledging the potential for misrepresentation, particularly in cases of extremely low observed richness, we mitigate this concern by confining our analysis to locations with more than one observed species record. This approach offers a straightforward method for identifying areas that warrant additional sampling.*"

33. P10, L6. "Intervals of 30 bins"? Please rephrase. And if the bin sizes are equal then state the sizes.

Thank you, we have now clarified that these were 30 equal sized bins.

34. P10, L20. "off"
Thanks for your comment. We corrected it.

35. P12, L15. This area based summary is an example of the modifiable areal unit problem. If one reconfigured the regions then the results would potentially be very different.

This statement is accurate, our analysis primarily aims to depict the spatial distribution of the representativeness of online repositories. It's important to note that this representation is subject to change, and this dynamic aspect is a key focus of the Modifiable Areal Unit Problem (MAUP). For more in-depth discussions on this topic, we refer the reader to the contributions of Tittensor et al. (2010), Garcia-Rosello et al. (2015), Meyer et al. (2015), and Troia and McManamay (2016, 2017). However, delving into the effect of scale on these online repositories goes beyond the scope of this manuscript.

36. Fig 2 caption. As I have noted in previous reviews, this is not a 1 degree hexagonal lattice. It is approximately one degree at the equator. However this description becomes increasingly inaccurate as one approaches the poles.
Thanks for your comment. We corrected the caption.

37. Fig 2. It is impossible to see the patterns. Why not make the land areas grey and the marine region background white? But now I zoom in the issue is the hexagon boundaries. These need to be removed, as requested in my previous review.
We did our best to increase the transparency of hexagon borders.

38. P13, L6. This is the same issue as for the Fig 2 caption.
Thanks for your comment. We corrected it.

39. Table 2 caption. "percentaje"
Thanks for your comment. We corrected it.

40. Fig 5. Give the units for each histogram x-axis: "species length (cm)" and "habitat depth (m)". There is no need to say they are log10 scaled as this is obvious from the axis labels.
Thanks for your comment. We corrected it.

41. P21, L4. "contrar"
Thanks for your comment. We corrected it.

42. P22, L4. "preference" is not really the correct term here. The databases collate available data and more is collected in some regions than others. It is uncommon for these data to be collected for the purposes of understanding biodiversity.
Thanks for your comment. We corrected it.

43. P23, L30. This could be an artefact of data availability. The North West Pacific region includes the coastline of the Peoples Republic of China. Many PRC records exist that are not part of the databases analysed here.

Indeed, the Republic of China uses the database "China species 2000" (http://www.sp2000.org.cn/). However, our gathering shows that during 2001 this repository was merged with other local repositories into the Catalogue of Life, which in fact is currently included in GBIF. See:
https://www.gbif.org/dataset/b8e085cf-5f2c-47f8-9244-fd501202e475#registration

Spatial and ~~temporal representation~~ Temporal Representation of ~~marine fish occurrences available online~~ Marine Fish Occurrences Available Online

Vanessa Pizarro[a], Andrea G. Castillo[a,b], Andrea Piñones[c,d,e,f], Horacio Samaniego[a,g,*]

[a] *Laboratorio de Ecoinformática, Instituto de Conservación, Biodiversidad y Territorio, Universidad Austral de Chile, Valdivia, Chile*

[b] *Programa de Doctorado en Ciencias mención Ecología y Evolución, Escuela de Graduados, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile*

[c] *Instituto de Ciencias Marinas y Limnológicas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile*

[d] *Centro FONDAP de Investigación en Dinámica de Ecosistemas Marinos de Altas Latitudes (IDEAL), Valdivia, Chile*

[e] *Centro de Investigación Oceanográfica COPAS-COASTAL, Universidad de Concepción, Chile*

[f] *Millenium Institute Biodiversity of Antarctic and Subantarctic Ecosystems - BASE, Chile*

[g] *Instituto de Sistemas Complejos de Valparaíso, Subida Artillería 470, Valparaíso, Chile*

Abstract

Despite the 243,000 ~~species of~~ marine species described by 2022, our knowledge about the oceanic biodiversity is still incomplete. This knowledge gap carries potentially adverse and far-reaching consequences for the preservation of marine ecosystems, particularly in the context of the ongoing human-induced alterations to our biosphere and the rapid progression of climate change and global environmental shifts.

~~However~~Recently, however, a large number of online repositories ~~cataloging, storing and distributing biodiversity information~~hosting taxonomic information have emerged, which catalogue, store and distribute biodiversity information, including taxonomic and species occurrence data~~have emerged recently~~. FishBase, the Global Biodiversity Information Facility (GBIF) and the Ocean Biodiversity Information System (OBIS) are part of these publicly available repositories representing a variety of sources that have exploded in number. However, despite the incredible accumulation of biodiversity records, not all

*Corresponding author

Email address: horacio@ecoinformatica.cl (Horacio Samaniego)

the information is ~~really~~ actually useful, nor does it represent any new knowledge regarding global species richness patterns.

In this study, we assessed the spatial and temporal representativeness of marine fish records (order Actinopterygii) found in the GBIF and OBIS global repositories. ~~We have developed a~~ The methodological framework that we developed relies on a series of non-parametric estimators for computing species richness from incidence data. This methodology employs hexagonal grids as sampling units that overlay marine bioregions across the globe.

Using standard ecological and spatial analysis tools, we identify regions that are adequately represented in terms of available records and therefore have more reliable data, as well as regions with few records that do not represent current species richness. We overlap these results with the location of marine protected areas and fishing exploitation zones to understand the anthropogenic effect on marine ichthyofauna. We additionally evaluate hypotheses regarding the taxonomic, geographic, and temporal distribution of information biases to deepen our current understanding of public records of species occurrences worldwide.

Considering that more than 40 years of information was analyzed, the results showed that, on a global scale, the primary data on marine fish available on GBIF and OBIS platforms are still far from being representative and complete. Only 1.14% of the records were useful for our analyses. In addition, we found that the information seems to be biased towards coastal areas, regions close to developed countries, and areas where there is a large fishing activity. Finally, the best represented species and families are those with a small body size, which use shallow habitats and ~~have commercial~~ are usually recognized as having commercial or cultural value.

Keywords: Ecoinformatics, Ecological Information Biases, Marine Fish, Spatial and Temporal Representativeness, Species Richness

1. Introduction

Currently, the more than 243,000 species included in the World Register of Marine Species database (WORMS, 2022) suggests that only 11% to 78% of all marine species have been discovered, revealing a striking picture of vastly incomplete knowledge that may have serious implications for marine conservation (Luypaert et al., 2020). Moreover, ongoing climate change

represents one of the greatest threats to biodiversity (Malhi et al., 2020; Turner et al., 2020) and has already been documented to modify the distribution of marine species (Lenoir et al., 2020). Some of the ~~described effects~~ effects described includes the invasion of non-native species leading to massive species turnover that may ~~lead to~~ result in the local extinction of large ~~proportions~~ share of species (Cheung et al., 2009).

It is crucial to recognize that species richness, while being a diversity metric among many, is, in itself, an aggregate variable ~~subsuming the overall variety of life~~ quantifying the end result of the splitting and lumping of the tree of life as a product of evolutionary processes (Marquet et al., 2004). Consequently, numerous endeavors have been directed towards the development of more comprehensive diversity indices, giving rise to significant scientific literature, aimed at describing ecological heterogeneity (Tuomisto, 2011; Moreno and Rodríguez, 2011; Daly et al., 2018). However, within this literature, there appears to be a shifting focus towards examining the ramifications of biodiversity loss. This shift involves the adoption of new terminology designed to provide pragmatic concepts, such as "species inventory", "taxonomic inventory", or "inventory completeness", which are intended to convey more precise messages to policymakers, summarizing the richness of biodiversity (Pereira et al., 2013; Butchart et al., 2010). Nevertheless, while the scientific community engages in debates over the use of biodiversity terminology, it is important to note that species richness continues to offer a concise and easily manageable description of variability across various other parameters characterizing the biota in both spatial and temporal dimensions (Appeltans et al., 2012). Species richness remains an essential feature for comprehending how diversity evolves in response to natural and anthropogenic influences within biomes, regions, and ecosystems (Troia and McManamay, 2017; Magurran and McGill, 2011).

Likewise, biodiversity can also be assessed through life history traits, which are modulated by both evolutionary factors and ~~the variation in habitats~~

~~and ecosystems~~ habitat ecosystem variations (Neigel, 1997; Hutchings and Baum, 2005). We now know that biodiversity is more likely an expression of the heterogeneity of such life history traits. Alò et al. 2021, for example, ~~shows~~ show that while some of the fish diversity is certainly due to environmental processes, a large fraction of such richness variance is also determined by evolved life history traits related, for example, to migratory habits. Therefore, evaluating how life history traits impact richness metrics should deepen our understanding of fish diversity patterns.

While still short of having a robust and standardized biodiversity infrastructure (Heberling et al., 2021), ~~a~~ there is great diversity of online repositories with taxonomic information and species occurrences data~~exist~~. Among the most important databases hosting marine information are FishBase, a platform that hosts information on the taxonomy of fish, their ecology, trophic information, habitat, and history of uses dating back to more than 250 years (Froese and Pauly, 2000); and the Global Biodiversity Information Facility (GBIF), a platform that stores and allows for the free access to species occurrence records from around the world. GBIF is currently one of the repositories hosting the largest amount of such data in the world (Telenius, 2011; GBIF: The Global Biodiversity Information Facility , 2021); and finally Ocean Biodiversity Information System (OBIS), which houses data on the occurrence and abundance of species from exclusively marine environments (OBIS: Ocean Biodiversity Information System, 2021). Records entered in these repositories are often used for research related to biodiversity assessment, taxonomic reviews, red listing of threatened species, species distribution, and generation of ecological niche models, among others (Yesson et al., 2007). GBIF currently offers more than 1.62 billion occurrence records and OBIS more than 63 million, which increase considerably each year (GBIF: The Global Biodiversity Information Facility , 2021; OBIS: Ocean Biodiversity Information System, 2021).

The records of both platforms come from a wide variety of sources col-

lected following different methodologies at different temporal and spatial scalesintroducing , which introduces a great variety of biases (Beck et al., 2014; Zizka et al., 2020). Among these, three main types of biases have been described: (i) taxonomic, this occurs when some species and/or families are better sampled than other rarer species (Chandler et al., 2017); (ii) geographic, when data input is unevenly distributed across geographic regions and may prove to obscure inter-region interregional comparisons (Yang et al., 2013; Yesson et al., 2007); and (iii) temporal, which may be prevalent when comparing different time periods as data coverage is unevenly distributed over time (Chandler et al., 2017; Yang et al., 2013). While these biases introduce some uncertainty regarding reliability of species richness descriptions obtained from online platforms (Beck et al., 2014; García-Roselló et al., 2015), they have largely been used to provide an extensive overview of macroecological patterns of distribution not available otherwise (Mora et al., 2008; Troia and McManamay, 2017).

Still, identifying how sampling effort is efforts are distributed across space and time is a necessary required step to interpret biodiversity patterns and reduce biasesas understanding the distribution of our biota is essential to design , as understanding our biota distribution is critical for well-designed protection efforts. This may be achieved through different weighting schemes for records in areas with sufficient sampling that provide a more reliable contribution compared to underrepresented regions (Phillips et al., 2009; Hortal et al., 2008; Yang et al., 2013).

We here assessed the spatial and temporal representativeness of marine fish records available in the global GBIF and OBIS repositories at the level of marine bioregionsmarine bioregions' level in order to pinpoint the location of records that best quantify the diversity of marine fishesmarine fish diversity. The result is a spatial representativeness analysis that we then overlay on marine conservation areas (UNEP-WCMC and IUCN, 2022) and fisheries exploitation areas (FAO, 2014) to learn whether marine conservation efforts,

as well as large fisheries, are located in areas of high species richness or ~~in areas of~~ areas which insufficient data coverage.

Finally, we also analyzed the potential effect ~~that some attributes could have~~ of some attributes on the incidence of more records in global database repositories. Specifically, we evaluated three research questions related to how body size, habitat depth, and commercial use ~~relates~~ relate to the representation of marine fish occurrences. We ask whether: (i) a better representation in ~~the~~ online platforms may be due to the ~~over sampling~~ oversampling of larger fish, ~~caused by its easy identification; that~~ resulting from an easier identification; (ii) shallow areas provide easy access to sampling; and (iii) economic and commercial ~~interest have elicit~~ interests have elicited a larger representation of culturally relevant species ~~among~~ in online biodiversity repositories.

2. Methods

2.1. Species data

We use all ~~recorded ocurrences from the order Actinopterygii hosted in~~ the recorded occurrences of the Actinopterygii order hosted in the GBIF and OBIS repositories (GBIF.org, 2021; OBIS.org, 2021). Following Alò et al. (2021), ~~evolutionarily~~ evolutionary older taxa, such as Cephalaspidomorphi, were excluded from this analysis. ~~The libraries~~ Libraries *rgbif* and *robis* of the statistical package R were used for data extraction (Chamberlain, 2017; Provoost and Bosch, 2020; R Core Team, 2018). Both repositories have collaborated since 2001, sharing data on the co-occurrence of marine life (OBIS.org, 2021). Nevertheless, recent investigations have shown significant disparities in data contributions, revealing remarkable low shares of shared data (Chollett and Robertson, 2020; Moudrý and Devillers, 2020). Noteworthy distinctions exist between the two platforms, encompassing diverse data sources and methodologies, along with substantial variations in temporal and spatial scales associated with data collection (Zizka et al., 2020). Due to

these disparities, scholars recommend a thorough examination and refinement of these data repositories (Bonnet-Lebrun et al., 2023). To enhance the quality and reliability of the information, a comprehensive series of filters has been systematically applied to our analysis. To minimize errors associated with the public usage of GBIF and OBIS repositories, we curated the dataset following Zizka et al. (2020) and filtered the dataset by the columns labeled "scientific name", "family", "year", "~~Longitude~~longitude" and "~~Latitude~~latitude". We retained all taxonomic information down to the species level ~~. Any record~~ and removed records with `NA` ~~values was removed~~in these columns. We also removed ~~any duplicated record~~ all duplicate records with identical latitude and longitude data, as well as ~~any record~~ records collected before 1980 (see Alò et al., 2021; García-Roselló et al., 2015). Each record was further assigned to a marine ~~bioregions~~ bioregion following Costello et al. (2017). Spatial data manipulation and plotting was performed with the aid of the following libraries: *sf*, *dplyr*, and *cartography* (Giraud and Lambert, 2016; Pebesma, 2018; Wickham et al., 2021). We finally labeled and removed ~~any~~ all exotic species record using the `distribution()` function provided by the *rfishbase* library (Boettiger et al., 2012; Froese and Pauly, 2021). To limit our analysis to species occurring within their native range, each record was checked against the ~~classification of~~ FAO fisheries area classification for consistency (FAO, 2014). A summary of the number of records is provided in Appendix A.

2.2. Data Analysis by Bioregion

Once the database was cleaned, a ~~subset of the data was created~~ data subset was extracted for each of the 30 bioregions. For each bioregion, ~~a count of~~ records, species~~and families was made~~, and families were counted, and the Shannon diversity index was calculated using the *vegan* library in R (Oksanen et al., 2020).

2.2.1. Spatial Representativeness Analysis

To assess the ~~spatial representativeness of the data~~data's spatial representativeness, bioregions were gridded into hexagonal cells of ~~equal~~ the same surface area to maximize ~~fit to bioregions~~the fit whit the bioregions' areas using a cylindrical equal area projection ~~and the World Geodetic System of 1984~~ (i.e. EPSG Code:54034). We approximated a 1°~~$\times$ 1°~~ $^{\circ}$ hexagonal lattice by computing cells of 10^4 square-kilometers, resulting in a total of 57,067 cells. In the appendix, we evaluated two additional spatial resolutions: ~~5° and 10°~~ $\sim 5^{\circ}$ and $\sim 10^{\circ}$ lattice with a total of 3,029 and 953 cells~~respectively using a gridcell of~~ , respectively, using a 2.5×10^5, and 10^7 ~~square-kilometers in order to analyze~~ square-kilometer gridcell to assess different biodiversity macropatterns (Tittensor et al., 2010).

The expected species richness (S_{exp}) was computed as the ~~average~~ mean between three non-parametric richness estimators~~so that $S_{exp} = \frac{1}{3}\sum_i^3 S_i$, where S_i is~~ : Chao2 (S_{chao}), Bootstrap ($S_{bootstrap}$) and Jackknife 1 ($S_{jackknife1}$) ~~(see Magurran and McGill, 2011, for individual definition of indices). Such averaging seeks~~ (see Magurran and McGill, 2011, for individual index definitions) . The purpose of this averaging is to minimize biases and potential ~~errors of under- or over-estimation~~ under or overestimation errors by using a single richness estimator following the work ~~of (Mora et al., 2008; Troia and McManamay, 2017)~~ by Mora et al. (2008) and Troia and McManamay (2017).

We then produced a species representativeness index (SRI) by comparing the observed richness (S_{obs}) per cell to S_{exp} (Troia and McManamay, 2017), $SRI_i = \frac{S_{obs}}{S_{exp}}$. This ~~index indicates the degree of representativeness of records~~is an undersampling index that points to the records' representativeness to quantify the actual species richness in each cell (i). Its value ranges from 0 to 1, where 0 represents an unsampled cell and 1 represents a fully sampled ~~one~~cell.

~~Since the~~ The Species Richness Index (SRI) ~~serves as an indicator of how databases depict~~ is used as a metric to assess the databases accuracy in

depicting actual species richness~~, it is reasonable to categorize cells arbitrarily~~. Consequently, we propose a systematic categorization of cells into three classes~~: "Few," "Sufficient," and "Adequately representative" of estimated species richness. We establish these classifications~~, *"low"*, *"medium"* and *"high"*, based on the frequency distribution of SRI~~(as depicted~~, as illustrated in Fig. A.1~~). Some cells contain only one species record and are labeled~~. Cells with only one record are identified as having insufficient records (IR) ~~to estimate~~ for estimating S_{est}. ~~Certain cells may exhibit limited knowledge with SRI falling within~~ Those with an SRI in the range (0, 0.60) ~~, while others may demonstrate a sufficient level of species diversity knowledge for a comprehensive representation if SRI falls within~~ are categorized as *low*, while those falling within the interval (0.60, 0.85) ~~. Additionally, some cells will possess an adequate representativeness level if SRI falls within~~ may be characterized as having a *medium* level of representativeness. Furthermore, cells within the range (0.85, 1.00) ~~. Cells with one or no records are treated as distinct classes, as are cells with a single record, in order to identify those cells with insufficient records for SRI estimation. Maps displaying~~ are identified as *high*, meaning an adequate representation of species diversity. Fig. A.2 show maps illustrating the raw values for observed species richness (S_{obs}~~,~~), expected species richness (S_{exp}), and SRI ~~can be found in Fig. A.2~~values.

2.2.2. Temporal Representativeness Analysis

We constructed species accumulation curves, ~~employing~~ using years as the ~~units of sampling~~ sampling unit, to examine the temporal distribution of data records within each bioregion. To assess the ~~adequacy of the sample~~sample's adequacy, we focused on data from the last four years ~~of data~~ (2016-2020), representing the final 10% of each accumulation curve. We ~~employed~~ used a linear fit following the rescaling of the SRI to facilitate statistically comparable slope measurements. Slopes approaching zero suggest bioregions that have been adequately sampled, whereas slopes deviating from zero indicate insufficient sampling efforts over time.

2.2.3. Gap Analysis

We overlaid the spatial representativeness map (§2.2.1) with ~~shapefiles of~~ Marine Protected Areas (~~MPA)~~ MPAs) shapefiles (UNEP-WCMC and IUCN, 2022) and fishing exploitation areas reported by (FAO, 2014). The superposition of these layers allowed us to calculate the extent of protection offered by ~~MPA~~ MPAs for each bioregions on a cell basis, and the extent of cells in designated fishing zones. ~~This exerciseallows to jointly assess the relationship between~~ Based on this exercise, the relationship among two opposing human impacts and current uncertainties about marine fish diversity can be assessed.

2.2.4. Bias Assessment

~~The evaluation of potential biases generated by~~ Potential biases resulting from body size, habitat depth, and cultural value of species (§2.1) ~~was assessed from the fishbase database~~ were assessed using Fishbase repository information (Froese and Pauly, 2021). We ~~generated~~ developed a frequency distribution plot for ~~the reported length of each species~~each species' length reported in the database, employing ~~intervals of~~ equal 30 ~~bins~~bin intervals. Habitat depth ~~were~~ was determined according to the classification of oceanic layers used in Costello et al. 2010 ~~,~~(i.e. epipelagic = 0 - 200 m, mesopelagic = 200 - 1,000 m, and bathypelagic= 1,000 - 4,000 m). A pie chart is used to show how cultural values are represented in the database.

All data and scripts are available (~~see~~ Appendix A).

3. Results

3.1. Records by Bioregions

Approximately 1.14% of the total ~~published occurrences in~~ reported occurrences of the order Actinopterygii were retained in our analysis. That is, from the 71,670,596 ~~downloaded records off~~ records downloaded from the GBIF and OBIS repositories, 820,004 were considered useful (see Appendix A). This

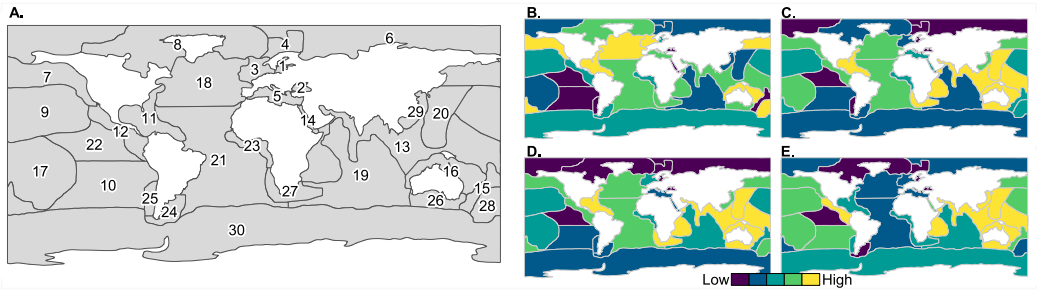


Fig. 1: Marine bioregions and spatial diversity distribution used in this study. **A**. The 30 marine bioregions from Costello et al. (2017) used in this study. Number are identification labels in Table 1. **B**. Records by bioregion; **C**. Overall species richness across bioregions; **D**. Family richness; and **E**. Shannon diversity index. Note that values in **C-E** have been ~~normalized~~ standardized for ~~display~~ illustration purposes. See Table 1 for actual values and a detailed map of observed and expected richness in Fig. A.2.

subset consisted of 10,371 species in 361 families. The most represented families in our dataset are Scombridae, Pleuronectidae, and Gadidae with 103,762, 57,018, and 52,079 records, respectively. The species with the largest representation frequency are *Hippoglossoides platessoides*, *Mola mola*, and *Coryphaena hippurus* with 30,885, 21,042 and 21,089 records, respectively.

The analysis at ~~bioregion~~ bioregions' level (Table 1) shows ~~a~~ large variability. ~~The count of records varies~~ Record counts vary across three orders of magnitudes, ~~that is~~ i.e., from 2.68×10^5 ~~records~~ in the Caribbean Sea and the Gulf of Mexico (11), down to 1.02×10^2 in the Black Sea (2). The ~~bioregion~~ bioregions with the largest species richness and diversity index ~~is the~~ are the Indo-Pacific Seas and the Indian Ocean (13), with 2.95×10^3 recorded species and a Shannon index of 6.93, followed by the Coral Sea bioregion (16), with 2.93×10^3 species and a Shannon index of 6.75. Likewise, the Coral Sea also presents the largest number of families. ~~It is interesting to note that, while being~~ Notably the Southern Ocean (30) is the largest bioregion ~~(i.e. in km^2), the Southern Ocean show the fewest~~ in square kilometers, it has the smallest number of records and the lowest number of species and families across all bioregions. The Black Sea (2) and ~~Norwegian~~ the Norwegian Sea (4) ~~are the bioregions with~~ bioregions have the lowest number of ~~record~~ records and

Table 1: Area (1,000 km^2) and ~~counts of~~ records, species richness, family richness, and Shannon diversity counts for each bioregion. The ~~largest values for~~ highest value in each column is highlighted.

ID	Bioregion	Area	Records	Species	Families	Shannon
1	Inner Baltic Sea	415	8,902	72	30	2.46
2	Black Sea	537	102	37	22	3.21
3	NE Atlantic	2,053	87,377	310	104	3.90
4	Norwegian Sea	1,132	3,046	93	35	2.16
5	Mediterranean	2,859	12,532	372	101	3.39
6	Arctic Seas	10,276	2,506	114	23	3.90
7	North Pacific	12,974	78,070	839	156	4.50
8	North American Boreal	8,001	9,709	162	48	2.99
9	Mid-Tropical N Pacific Ocean	32,685	9,310	615	127	4.59
10	South-East Pacific	21,952	386	190	89	4.97
11	The Caribbean and the Gulf of Mexico	8,427	**268,066**	1,703	209	4.49
12	Gulf of California	6,184	7,639	885	148	5.93
13	Indo-Pacific Seas and Indian Ocean	37,090	16,967	**2,947**	215	**6.93**
14	Gulfs of Aqaba, Aden, Suez, Red Sea	830	926	352	72	5.51
15	Tasman Sea	3,592	1,003	380	120	5.36
16	Coral Sea	7,658	40,107	2,929	**249**	6.75
17	Mid South Tropical Pacific	23,418	6,083	811	123	5.18
18	Offshore and NW North Atlantic	16,012	130,994	897	190	3.46
19	Offshore Indian Ocean	31,076	1,263	337	116	4.06
20	Offshore W Pacific	10,291	6,363	1,839	232	6.81
21	Offshore S Atlantic	41,435	11,960	990	188	3.79
22	Offshore Mid-E Pacific	13,815	687	79	37	3.04
23	Gulf of Guinea	3,325	6,816	384	138	3.95
24	Argentina	2,665	8,701	115	52	2.83
25	Chile	1,739	250	100	54	4.36
26	Southern Australia	3,824	15,643	1,011	201	5.75
27	Southern Africa	4,371	19,954	1,142	210	4.16
28	New Zealand	6,293	53,879	558	154	3.66
29	North West Pacific	2,457	1,767	869	182	6.46
30	Southern Ocean	**62,161**	8,996	294	57	3.98

Shannon index value, respectively. Fig. 1 illustrates the location of the 30 marine bioregions and their respective richness and diversity values.

3.2. Geographic Analysis

Fig.2 shows ~~the~~ cell classification according to SRI (§2.2.1). As expected, no bioregion is completely sampled at the ∼1° resolution. In fact, at this resolution, large empty regions with no records are observed. The bioregions with the largest area classified as ~~*Adequate*~~*high* representativeness are the Northeast Atlantic (3) (37.53%), the Caribbean and the Gulf of Mexico (11) (29.26%), and the Inland Baltic Sea (1) (24.37%). It should be noted that such cells ~~are mostly correspond from~~ mostly correspond to coastal areas in the northern hemisphere. On the other hand, the bioregions that present ~~a greater surface without records correspond to~~ the largest surface area without records are the Southeast Pacific (10) (96.3%), the Arctic Sea (6) (94.9%), and the Southern Ocean (30) (93.7%). While the bioregions with the ~~larger surface with sufficient~~ largest surface area and *medium* representativeness of records are the Gulf of Guinea (23) (32%), the Norwegian Sea (4) (22.3%), and the Gulf of California (12) (21.6%). Additional results for ~~5° ×~~ ∼5° and ∼10~~° × 10°~~ ° spatial resolution grids are ~~available~~ shown in Appendix C.

3.3. Temporal Analysis

Bioregions show similar ~~trends of data accumulation~~ data accumulation trends across the four decades analyzed here (Fig. 3). While a significant increase is apparent in the time period between 2005 and 2010, such increase is not significant for 14 ~~out~~ of the 30 bioregions. The Caribbean and the Gulf of Mexico (11) is the bioregion with the largest ~~increases~~ increase in data contribution to the dataset, while the Black Sea (2) ~~is the bioregion with the lowest rate of data contribution~~ shows the lowest data contribution rate in the 40 years span between 1980 and 2020. (See Appendix D for further analysis).

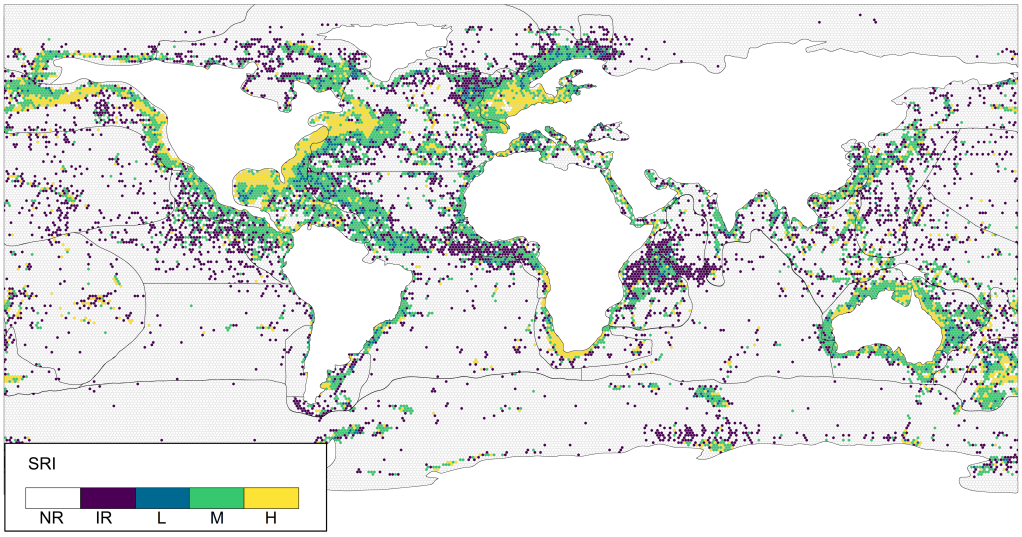


Fig. 2: Spatial Representativeness Index (SRI) in ∼1° hexagonal lattice. **IR** shows cells with insufficient records to evaluate S_{est}. ~~**A**~~ **H** are cells with an ~~adequate~~ *high* representativeness of species richness, i.e. SRI > 0.85. ~~**S**~~ **M** are cells considered as having a ~~sufficient~~ *medium* representativeness, i.e. SRI $\in (0.60, 0.85)$. ~~**F**~~ **L** cells are cells with ~~few~~ *low* representativeness of species records and are thus not considered to be representative of actual species richness, i.e. SRI $\in (0, 0.6)$. **NR** are cells with no records (SRI ~~= *NA*~~= *NA*). Raw values for SRI, S_{obs} and S_{est} are shown in the appendix (Fig. A.2).

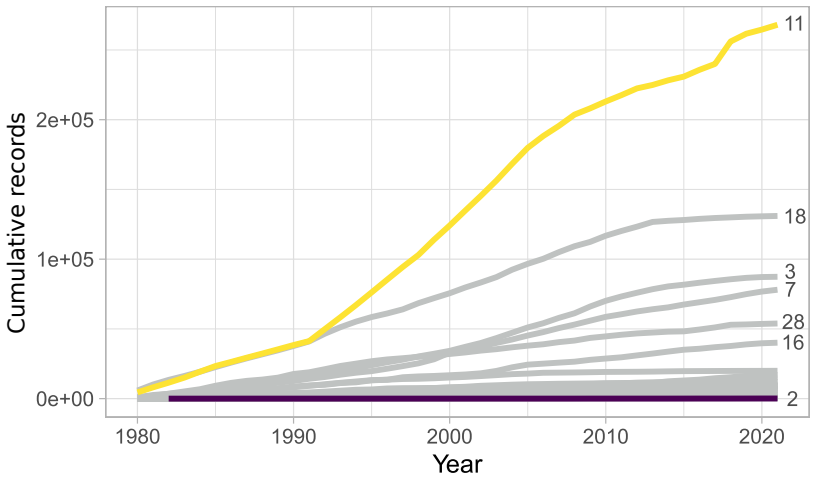


Fig. 3: Records of accumulation rate for each bioregion across the four decades analyzed. The ~~blue~~ yellow line is the accumulation of fish records in the Caribbean and the Gulf of Mexico bioregion (11)~~and~~ , while the ~~red~~ purple line ~~shows~~ is the accumulation rate in the Black Sea (2). ~~Numbers as~~ The numbers at the end of each ~~timeseries~~ time series correspond to the bioregion ID in Table 1.

We ~~categorize~~ classified the slopes of the final 10% of each accumulation ~~curves~~ curve in Fig. 4. Fourteen bioregions show a slope less than 1. The Mediterranean Sea (5) stands out with the lowest slope value (0.47), while the Black Sea (2) is the bioregion with the steepest final slope (3.13).

3.4. Gap ~~analysis~~ Analysis and ~~fishing exploitation areas~~Fishing Exploitation Areas

The bioregions with the largest area covered by protected areas are the Coral Sea (16), the ~~northeast~~ Northeast Atlantic (3) and New Zealand (28)~~covering a~~, covering 37.3%, 17.4%, and 16% of their respective surface areas. Regarding ~~the sampling level of these bioregions~~these bioregions' sampling level, the Offshore Indian Ocean (19)~~;~~, the Gulf of Aqaba, Aden, Suez, Red Sea (14) ~~;~~and Coral Sea (16), are the bioregions with the highest ~~percentages of cell sampled as *Adequate* inside of their~~ share of cells with *high* representativeness, hence well sampled within protected areas (83%, 63.8%, and 59.8% respec-

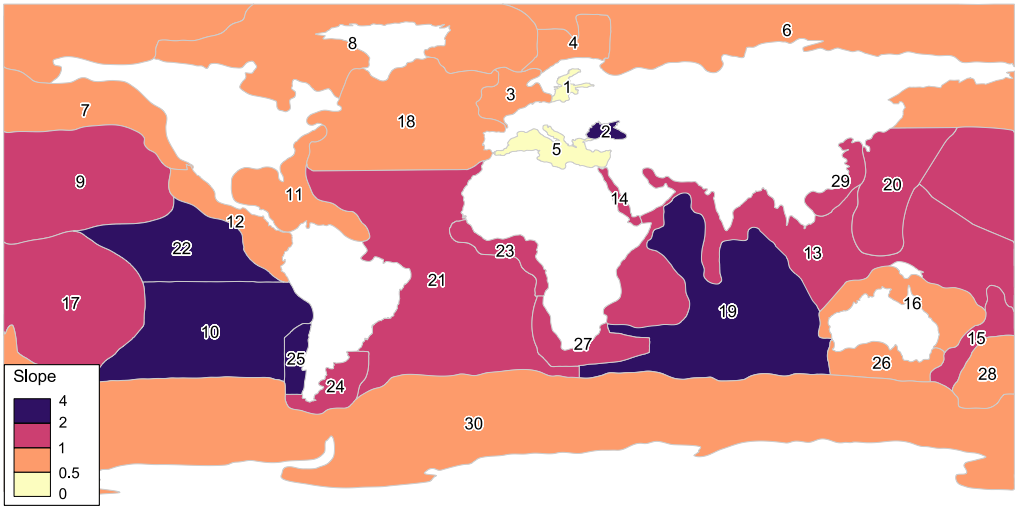


Fig. 4: ~~Graphical representation~~ Illustration of the slope values of the species accumulation curve for each bioregion. The slope corresponds to the final 10% of the species accumulation curve. See §2.2.2 for details regarding the analysis.

tively). ~~While~~ In turn, the Arctic Seas (6), ~~North American boreal~~ the North American Boreal (8) and ~~Mid South~~ Mid-South Tropical Pacific are the bioregions with protected areas ~~with the highest percentage of cell with no~~ showing the highest share of cells without records (86.2%, 83.8%, and 81.2% ~~).~~ respectively). (See Appendix E).

The FAO areas with the largest ~~area categorized as~~ surface area classified as ~~*Adequate*~~*high* representativeness correspond to the ~~northwest~~ Northwest Atlantic (22.1 %), Northeastern part of the Pacific Ocean (14.6 %), and Western part of the Atlantic Ocean (12.6 %) (Table 3). These FAO areas correspond to regions ~~of~~ in the Pacific Ocean (North Pacific, North West Pacific, ~~Mid-tropical~~ Mid-Tropical N Pacific Ocean and Indo-Pacific ~~seas~~ Seas and Indian Ocean, as well as the Gulf of California and ~~Caribbean and~~ the Caribbean and the Gulf of Mexico). Largest FAO areas with *NR* cells correspond to the Antarctic part of the Pacific Ocean, the Antarctic part of the Atlantic Ocean and the Southeastern part of the Atlantic Ocean in the Southern Ocean, Offshore S Atlantic, and Southern Africa.

Table 2: Results of overlapping ~~Marine Protected Areas~~ MPAs and SRI grid. ID is the identification number given to each bioregion (see Table 1 for ~~bioregion~~ bioregions' names). Area corresponds to the ~~percentage~~ share of surface area covered by ~~marine protected areas~~MPAs. ~~NR~~ **NR** is the ~~percentage~~ share of cells with *No Records*; ~~IR~~ **IR** is the ~~percentaje~~ share of cells with ~~*Insufficient records*~~*Insufficient Records*; ~~F~~ **L** is the ~~percentage~~ share of classified cells with ~~*Few*~~ *low* number of records; ~~S~~**M**, the ~~percentage~~ share of classified cells with ~~*Sufficient*~~ *medium* number of records, and ~~A~~**H**, the ~~percentage~~ share of classified cells with ~~*Adequate*~~ *high* number of records. The highest values for each column ~~is~~ are highlighted.

ID	Area	NR	IR	~~F~~L	~~S~~M	~~A~~H
	km^2			%		
1	0.03	2.38	4.30	5.22	49.08	39.01
2	12.89	26.71	**27.61**	10.49	35.19	0.00
3	9.74	3.23	1.35	0.94	40.86	53.62
4	0.15	11.16	19.63	5.18	56.69	7.34
5	0.09	5.71	6.77	**11.20**	47.95	28.37
6	5.01	**86.16**	5.97	0.02	4.65	3.20
7	0.00	26.48	4.24	1.26	23.77	44.25
8	1.23	83.82	7.77	0.62	6.70	1.11
9	0.69	69.58	15.77	0.00	6.40	8.25
10	17.36	73.51	24.58	0.00	0.80	1.11
11	0.28	20.13	6.25	3.44	29.98	40.21
12	0.83	0.33	1.23	8.85	**61.35**	28.25
13	0.45	50.52	11.94	0.88	25.17	11.50
14	0.25	8.87	0.00	1.59	25.65	63.88
15	4.06	57.18	15.27	0.00	7.29	20.26
16	16.00	3.10	0.49	1.10	35.52	59.79
17	0.20	81.19	11.43	0.00	3.07	4.31
18	4.91	35.87	16.63	0.77	25.64	21.09
19	2.78	11.88	0.74	0.00	4.34	**83.04**
20	2.56	34.06	9.68	6.85	35.31	14.10
21	3.79	51.06	19.35	0.38	21.35	7.86
22	0.16	41.98	27.54	0.00	23.22	7.26
23	13.83	9.92	4.92	7.98	61.33	15.85
24	0.21	40.86	20.64	0.24	23.97	14.29
25	0.07	65.27	0.50	0.05	1.29	32.89
26	0.28	30.04	12.83	6.89	38.62	11.63
27	1.45	16.78	5.75	0.15	18.71	58.62
28	0.00	45.36	2.25	0.00	13.71	38.67
29	**37.29**	20.49	17.05	2.68	42.50	17.29
30	1.66	64.05	7.12	4.89	19.86	4.08

Table 3: Results ~~of~~ from overlapping FAO fishery ~~exploitation~~ areas and SRI grid. The surface area corresponding to each bioregion, and the ~~percentage~~ share of surface area of each classification. ~~Area is~~ Areas are in thousands ~~of~~ square km^2; ~~NR~~ **NR** is the ~~percentage~~ share of cells with *No Records*; ~~IR~~ **IR** is the ~~percentage~~ share of ~~cell~~ cells with ~~*Insufficient Record*~~*Insufficient Records*; ~~F~~ **L** is the ~~percentage~~ share of classified cells with ~~*Few*~~ *low* number of records; ~~S,~~ **M** is the ~~percentage~~ share of classified cells with ~~*Sufficient*~~ *medium* number of records, and ~~A,~~ **H** is the ~~percentage~~ share of ~~classified~~ cells with ~~*Adequate*~~ a *high* number of records. The ~~highest~~ largest values for each column ~~is~~ are highlighted.

FAO Area Name	**Area** km^2	**NR**	**IR**	**L** %	**M**	**H**
Arctic Sea	4,086	93.22	3.13	0.29	2.61	0.75
Northwestern part of the Atlantic Ocean	874	31.19	11.66	5.69	29.37	**22.08**
Northeastern part of the Atlantic Ocean	3,223	66.29	12.54	2.55	13.63	4.99
Western part of the Atlantic Ocean	1,285	30.84	13.09	**7.91**	**35.60**	12.55
Eastern Central part of the Atlantic Ocean	1,208	52.61	**24.09**	3.44	18.37	1.19
Mediterranean Sea and the Black Sea	309	46.39	15.43	5.24	24.77	8.17
Southwestern part of the Atlantic Ocean	1,731	82.49	5.85	1.69	8.55	1.42
Southeastern part of the Atlantic Ocean	1,765	89.92	4.19	0.15	2.13	3.61
Antarctic part of the Atlantic Ocean	2,310	93.31	2.80	0.20	2.93	0.76
Western part of the Indian Ocean	2,621	72.45	16.11	1.03	8.51	1.89
Eastern part of the Indian Ocean	3,029	85.40	4.69	0.82	7.39	1.70
Antarctic and South of the Indian Ocean	1,977,29	85.71	7.76	0.56	4.33	1.64
Northwestern part of the Pacific Ocean	2,259	73.55	12.40	0.94	10.32	2.79
Northeastern part of the Pacific Ocean	968	55.13	12.65	1.34	16.26	14.62
Western Central part of the Pacific Ocean	2,963	70.45	12.56	0.43	11.58	4.98
Eastern Central part of the Pacific Ocean	**4,141**	79.36	11.30	0.31	6.94	2.09
Southwestern part of the Pacific Ocean	3,097	85.04	4.42	0.97	6.40	3.17
Southeastern part of the Pacific Ocean	2,997	91.16	6.00	0.10	2.30	0.44
Antarctic part of the Pacific Ocean	2,361	**93.47**	4.57	0.21	1.42	0.33

3.5. Evaluation of Biases

We evaluated biases for body size, habitat depth, and cultural value for 10,371 marine fish species identified in our database (§3.1).

3.5.1. Body ~~size~~Size

The ~~range~~ 10-40 cm range is the most frequently occurring size length, corresponding to the interval between the 1^{st} and 3^{rd} quartile (Fig. 5A). Three species stand out with the highest numbers of records, *Scomber scombrus*, *Lagodon rhomboides* and *Mallotus villosus* with 20,995, 19,563 and 13,609 records respectively. These species are distributed mainly in the Northeast

Atlantic (3) and Offshore and Northwest North Atlantic (18) bioregions. ~~While~~ In turn, the families that accumulate the greatest number of records ~~correspond to~~ are Sparidae , Scombridae and Labridae, with 24,837, 21,719, and 21,035 records, respectively. These families are mainly distributed in the Caribbean Sea and the Gulf of Mexico ~~and~~ (11), and in the Northeast Atlantic (3).

3.5.2. Habitat ~~depth~~Depth

The depth range most commonly observed among records is ~~centered around~~ about 50 meters and decreases as depth increases, particularly from the epipelagic to the mesopelagic ~~zone~~ zones, as illustrated in Fig.5B. Among the species with the highest number of recorded occurrences, *Mola mola*, *Coryphaena hippurus*, and ~~*Lagodon*~~ *L. rhomboides* stand out, with 21,089, 21,042, and 19,563 occurrences in the databases, respectively. These species are distributed mainly around the Caribbean Sea and the Gulf of Mexico (11) bioregions, as well as the following bioregions: Offshore and NW North Atlantic (18) and the South Atlantic Coast (21). The families that accumulate a ~~greater~~ larger number of records correspond to Scombridae, Gadidae, Sparidae with 63,572, 38,876, and 30,041 records, respectively. These are mostly distributed in the northern hemisphere~~. That~~ ; that is, the Caribbean and the Gulf of Mexico (11), Offshore and NW of the North Atlantic (18), and part of the South Atlantic Ocean Coast (21) bioregions.

3.5.3. Cultural ~~value~~Value

Finally, when analyzing the most frequent cultural value represented across our dataset (Fig. 6), "Commercial" use of the species emerges as the most important with ~~a~~ 73.4% among records, followed by the category "No interest" (5.03%), and "Subsistence fishing" (3.08%).

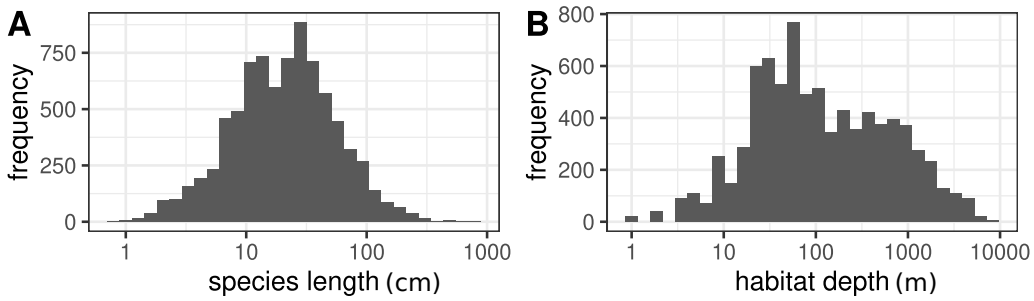


Fig. 5: Distribution of marine fish records in GBIF and OBIS ~~categorized~~ classified by body length and habitat depth. **A**~~,~~. Relationship between record number and species length ($\log_{10}$); and **B**. Relationship between record number and habitat depth ($\log_{10}$) .

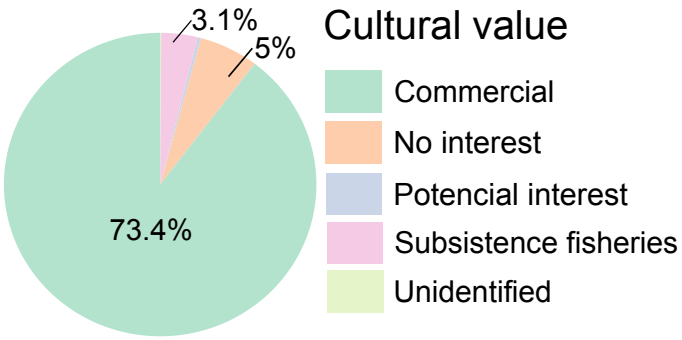


Fig. 6: Frequency of marine fish representation in GBIF and OBIS repositories according to ~~importance of~~ cultural ~~use~~value.

4. Discussion

Our work provides a methodological framework based on a set of non-parametric estimators to quantify the potential number of species from incidence data (Chao et al., 2009). We ~~employed~~ used hexagons due to their suitability as a tessellation that conforms more effectively to the shape of a spheroid, compared to square grid cells. We also placed special emphasis on cleaning the occurrence data in their taxonomy (Jin and Yang, 2020), and any potential input errors associated with large and massive datasets (Zizka et al., 2020). ~~This led us to focus on only~~ Hence, we only focused on evaluating marine species in the order Actinopterygii (Alò et al., 2021).

Publicly accessible occurrence records are growing rapidly, partly due to ~~the significant advances~~ significant progresses in ecoinformatics (Lenoir et al., 2020; Oliver et al., 2021). These databases harbor a growing variety of sources, including museum specimens, field observations, acoustic and visual sensors, and citizen science efforts (Amano et al., 2016). However, despite the incredible accumulation of biodiversity records, not all the data is really useful, nor does it represent new insights into the distribution of species (Bayraktarov et al., 2019; Zizka et al., 2020). That is why a systematic evaluation of the integrity and coverage of this information is required (Troia and McManamay, 2017).

There is an extensive bibliography that evaluates the record quality available for different taxonomic groups. Some examples are: legumes on a global scale (Yesson et al., 2007), lepidoptera from Great Britain, and woody plants in Panama (Chao et al., 2009), global marine biodiversity (Tittensor et al., 2010), vascular plants in China (Yang et al., 2013), marine fish on a global scale (Mora et al., 2008; García-Roselló et al., 2015), freshwater fish in the USA (Troia and McManamay, 2017; Pelayo-Villamil et al., 2018), and terrestrial mammals on a global scale (Oliver et al., 2021), among many others. Assuming that that not all data available in these repositories ~~are~~ is useful for biodiversity analyses, several efforts have proposed parametric and non-parametric estimators for data cleaning and species richness analysis, including ModestR (García-Roselló et al., 2013), KnowBr (Lobo et al., 2018), and RWizard (Guisande and Lobo, 2019)~~among these~~.

Striving for simplicity, we employ the ratio of observed to expected species richness (SRI) as a means to indicate the spatial distribution of undersampled regions. While acknowledging the potential for misrepresentation, particularly in cases of extremely low observed richness, we mitigate this concern by confining our analysis to locations with more than one observed species record. This approach offers a straightforward method for identifying areas that warrant additional sampling.

We evaluated two additional grid sizes (i.e. 2.5×10^4 and 10^7 km^2), and like other studies, our results show that the coarser the resolution used, the greater the overestimation is ~~,~~ in terms of area. That is, the richness index will indicate that a large area is, indeed, well sampled when in reality, ~~the~~ occurrence records could in fact be localized in a very small area. On the ~~contrar~~contrary, the finer the scale of analysis, the more localized and deficient the sampling is ~~(Tittensor et al., 2010; García-Roselló et al., 2015; Meyer et al., 2015; Troia a~~ ~~.~~ (Tittensor et al., 2010; García-Roselló et al., 2015; Meyer et al., 2015; Troia and McManamay, 2 .

Considering that more than 40 years of data were ~~analysed~~analyzed, our results demonstrated that on a global scale ~~,~~ the primary marine fish data available on the GBIF and OBIS platforms are still far from being representative and complete. Compared with other studies evaluating the same taxonomic group (Mora et al., 2008; García-Roselló et al., 2015), although we obtained similar macroecological patterns, only 1.14% of the records extracted from both repositories were useful for our analyses. ~~The large percentage of the~~ A large share of occurrences presented input errors or ~~did not have the necessary data to generate a reliable analysis~~ lacked the data required to develop reliable analyses (Yesson et al., 2007; García-Roselló et al., 2014).

We also found evidence of strong information biases in the records explored. On the one hand, when analyzing the families and species with the greatest representation, they ~~coincide with groups of fish of commercial interest , demonstrating~~ match groups of commercial interest fish, pointing to the existence of ~~**taxonomic bias** of the data~~ data taxonomic biases (Melo-Merino et al., 2020). This is the case of the families Scombridae, Pleuronectidae and Gadidae, which include ~~species of nutritional importance~~ nutritionally-relevant species such as tuna, cod, haddock, among others (Cohen et al., 1990). The same is true for the species with the largest number of records, *H. platessoides* (Pleuronectidae), *C. hippurus*(Coryphaenidae), and *M. mola* (Molidae)~~,~~; while the first two are species exploited by the

fishing industry, ~~with the exception of~~ sunfish (*M. mola*) ~~which~~ has a wide distribution and is mostly associated with scientific and recreational ~~interest~~ interests (Pope et al., 2010).

The unequal contribution of data at the spatial level is another factor that must be considered ~~to work~~ when dealing with data available ~~on~~ in ecoinformatic platforms. ~~There is a clear preference for~~ We show a clear **geoghapic bias** in the sampling of certain regions and/or ecosystems~~as a result of geographical bias~~. The literature indicates that the ~~highest data contribution rates correspond to~~ largest data contributions come from developed countries (Yesson et al., 2007; Chandler et al., 2017), and ~~those~~ coastal regions with ~~better~~ high road connectivity (Chandler et al., 2017; Melo-Merino et al., 2020). This ~~information uncertainty~~ is also particularly prevalent in under-sampled marine habitats, such as the deep sea (Webb et al., 2010). Our results ~~coincide with what is~~ match what has been described in the literature, regardless of the ~~size of the grid that was used to generate the analysis, the~~ grid size used for the analysis. The bioregions that include the ~~Atlantic~~Northeast Atlantic (3), the Caribbean and the Gulf of Mexico (11), and the ~~Baltic Sea are the regions with the highest number of area sampled as *Adequate* associated mainly with coastal areas~~Inland Baltic Sea (1) are regions classified with a high representativeness. However, the number of cells with insufficient ~~data to generate a~~ records to generate an unbiased diversity analysis ~~, is also worrisome. For instance~~is also of concern. For example, our results show that these cells are distributed in more internal areas of the bioregions, zones where sampling is likely to be more difficult. ~~While~~

While, on the other hand, the bioregions that include the South and Southeast Pacific (including the southern coast of South America), the Southern Ocean, and the Arctic ~~Seas~~ Sea are the regions ~~with the least spatial representativeness of records, the proportion~~ where the share of cells without records (*NR*) exceeds 90%. The ~~absence~~ lack of data samples over this

extensive area renders any endeavor to depict species richness and distribution highly unreliable ~~within these bioregions~~ (as noted by Yang et al., 2013; Troia and McManamay, 2017). These marine regions encompassing both the water column and the seabed beyond ~~the territorial jurisdiction of countries constitute~~ national jurisdictions make up nearly half of the Earth's surface and sustain ~~a~~ substantial abundance and diversity of life, as highlighted by (Visalli et al., 2020). Nonetheless, when scrutinizing the occurrence data for marine ichthyofauna, these regions remain the least sampled areas.

Finally, the ~~time bias of the data~~ datas's **time bias** is also present in our study. Differences in species identification and sampling methodologies over the decades have resulted in ~~the production of~~ databases of variable quality. However, the current era is characterized by more accurate data thanks to improvements in individual capture and identification tools (Costello et al., 2015; Jin and Yang, 2020). For these reasons, our approach considers occurrence records since 1980~~, however, ;~~ the coverage of occurrence data, however, is uneven over time when comparing ~~between~~ marine bioregions. Despite ~~evaluating more than~~ assesing four decades of data, ~~still~~ sampling efforts are still insufficient in 46% of marine bioregions~~have insufficient sampling efforts~~. Not surprisingly, the Caribbean and the Gulf of Mexico (11) ~~bioregion is the region~~ is the bioregion with the largest ~~input of data , demonstrating once again that~~ data input, once again showing that the geographic sampling bias has strong ~~effects~~ impacts on spatial predictions of species richness (Yang et al., 2013). Future sampling efforts should focus on bioregions at low or equatorial latitudes, areas where ~~biogeographic studies show that~~ marine biodiversity is concentrated according to biogeographic studies (Costello et al., 2017).

All the biases that we have described, added to ~~the inherent problems in data capture , foster and deepen various~~ typical data capture issues, promote and deepen several information gaps that ~~affect~~ thwart the effective spatio-temporal ~~quantification of biodiversity~~ biodiversity quantification (Magur-

ran and McGill, 2011). In this study, we have overlapped our ~~estimates of species richness~~ species richness estimates with the global ~~marine protected areas declared up to~~ MPAs declared up until the beginning of ~~the year~~ 2022 (UNEP-WCMC and IUCN, 2022), and the ~~areas of fishing exploitation reported by the~~ fishing exploitation areas reported by FAO (FAO, 2014). This exercise demonstrates the importance of public databases that can faithfully reflect the taxonomic and biogeographical knowledge available for each region ~~of the world (Pelayo-Villamil et al., 2018). Our resultsindicate that~~ (Pelayo-Villamil et al., 2018). According to our results, the North West Pacific bioregion (19) has the largest area covered by ~~marine protected areas~~MPAs. However, its ~~percentage of adequately sampled cells~~ share of cells with *high* representativeness is low compared to other bioregions. This ~~latter~~ result is of certain concern as this bioregion is considered a conservation hotspot among other bioregions, such as the Coral Sea (16), a ~~bioregion~~ region with a relatively large ~~percentage of adequately~~ share of highly sampled cells (Ramírez et al., 2017). However, we found a low ~~proportion~~ share of well-sampled cells in both regions, ~~demonstrating~~ pointing to the existence of important information gaps, at least for fish of the order Actinopterygii. We emphasize the need to correct these information gaps so that conservation efforts ~~that seek the implementation of new marine protection areas can have reliable dataso as not to underestimate the biodiversity of species~~ can rely on dependable data, including the design and implementation of new MPAs (Sala et al., 2021).

~~In the same way~~Along these lines, by overlapping the bioregions with ~~the~~ fishing exploitation zones, we determined that the North Pacific (7), the North West Pacific (29), ~~Mid-tropical~~ the Mid-Tropical N Pacific Ocean (9)~~and~~ , and the Indo-Pacific ~~seas~~ Seas and Indian Ocean (13) bioregions, as well the Gulf of California (21)~~and Caribbean and~~ , and the Caribbean and the Gulf of Mexico (11), are the regions with the highest ~~representation of the data~~ data representation and where fishing activity is concentrated.

According to (Kroodsma et al., 2018), the area corresponding to the central Atlantic and Northeast Pacific present little intense fishing ~~effort~~efforts, while the regions associated with the Northeast Atlantic, the Northeast Atlantic (Europe) regions, and the Northwest Pacific are known to have ~~a~~ huge fishing development~~and that is~~ , where fishing efforts are concentrated worldwide. The ~~southeastern~~ Southeastern Atlantic Ocean (FAO area 47 and 88), part of the Pacific Ocean (FAO area 88) and Antarctica (FAO area 48 and 88) are the regions with the highest ~~percentage~~ share of cells without records (NR= >93 %). When compared with the findings ~~of (Kroodsma et al., 2018)~~ by Kroodsma et al. 2018, these areas ~~agree with~~ match the "holes" without fishing effort data, which is explained by the geographical remoteness and the lack of technological development ~~necessary for the~~ required for fisheries to extend to new domains (Visalli et al., 2020). This ~~limits both the exploitation~~ issue restricts both the extraction of marine resources ~~and the collection of data~~ as well as data collection.

The research questions addressed in this study were essential ~~for comprehending the prevailing trends in data collection and laying~~ to understand the prevailing data collection trends and to lay the groundwork for potential corrective measures ~~to~~ than can mitigate the described biases. Our initial inquiry regarding fish body size does not imply a straightforward association between larger records and larger body lengths. Instead, we observe a distinct hump-shaped distribution in ~~the frequency distribution~~frequency distributions, akin to well-documented macroecological patterns observed in various taxa (Smith et al., 2014; Allen et al., 2006). It is worth noting that mid-sized fish species account for the highest number of records. Among these, species such as *S. scombrus* (Scombridae), *L. rhomboides* (Sparidae), and *M. villosus* (Osmeridae) stand out for their numerous records~~, and~~ ; they are predominantly distributed in well-sampled regions such as the Mediterranean Sea ~~,~~ (5), the Caribbean and the Gulf of Mexico ~~and the Caribbean~~(11), and the Atlantic Ocean (e.g. bioregion 3). Furthermore, the inverse relationship between fish

size and abundance, and consequently, the frequency of human ~~utilization~~use, whether for scientific research or commercial purposes, is a well-established concept (Pauly and Palomares, 2005).

This variation in sampling ~~effort~~ efforts results in a noticeable overrepresentation of these species, exacerbating the existing ~~taxonomic bias~~**taxonomic bias**. Conversely, the correlation between the number of records and habitat depth indicates that the pelagic zone ~~exhibits a significant concentration of data~~ shows a significant data concentration, which appears to align with areas more readily accessible for data collection (Melo-Merino et al., 2020). It has been pointed out that ~~the concentration of species~~ species concentration decreases as the ~~depth of the ocean increases ,~~ ocean increases its depth; however, it is precisely these areas that have been the least sampled and where there is ~~the greatest probability~~ a larger chance of discovering new species (Costello et al., 2017). This demonstrates the need to concentrate efforts on the deeper regions of the water column (mesopelagic, bathyal, and abyssal) for a more equitable representation of marine ecosystems. Finally, a straightforward examination of cultural value ~~among~~ within marine records unmistakably reveals that ~~species of marine fish~~ marine fish species with more favorable or ~~economically advantageous utility to~~ economic advantages for humans tend to have stronger ~~representation within the analyzed databases~~ representations within the databases discussed. This observation is likely connected to the significant role of the fishing industry as one of the primary sources of information contributing to platforms such as OBIS, as previously discussed ~~OBIS~~ (Zhang and Grassle, 2002).

Today, marine ecosystems and their biodiversity face the ~~great challenges of climate change and the impact~~ major climate change challenge as well as the impacts of human activity, especially ~~those~~ on species considered key food resources for survival (Hollowed et al., 2013; Ramírez et al., 2017; O'Hara et al., 2021). It is ~~necessary to focus and strengthen~~ important to focus on and further the study of ~~those areas with very~~ areas with few or no records,

since ~~the descriptions of the geographic ranges of the species~~ describing the species geographic ranges and their temporal dynamics ~~are fundamental measures~~ is a key measure for the evaluation of the ~~real state of biodiversity (Lenoir et al., 2020; Oliver et al., 2021). Having~~ actual biodiversity state (Lenoir et al., 2020; Olive . Counting on more reliable data will allow for the implementation of effective conservation actions~~to be implemented~~.

Acknlowledgements

Funding for this research was provided by ~~the National Agency of~~ Chile's National Research and Development ~~of Chile~~ Agency (ANID) through project FONDECYT Regular #11211490 to HS and a doctoral fellowship to AGC (ANID #2022-21220124). We ~~finally~~ thank professor Ricardo Giesecke for his valuable comments on an ~~early~~ earlier version of this manuscript.

References

Allen, C.R., Garmestani, A.S., Havlicek, T.D., Marquet, P.A., Peterson, G.D., Restrepo, C., Stow, C.A., Weeks, B.E., 2006. Patterns in body mass distributions: sifting among alternative hypotheses. Ecology Letters 9, 630–643. doi:10.1111/j.1461-0248.2006.00902.x.

Alò, D., Lacy, S.N., Castillo, A., Samaniego, H.A., Marquet, P.A., 2021. The macroecology of fish migration. Global Ecology and Biogeography 30, 99–116. doi:10.1111/geb.13199.

Amano, T., Lamming, J.D., Sutherland, W.J., 2016. Spatial gaps in global biodiversity information and the role of citizen science. Bioscience 66, 393–400. doi:10.1093/biosci/biw022.

Appeltans, W., Ahyong, S., Anderson, G., Angel, M., Artois, T., Bailly, N., Bamber, R., Barber, A., Bartsch, I., Berta, A., Błażewicz-Paszkowycz, M., Bock, P., Boxshall, G., Boyko, C., Brandão, S., Bray, R., Bruce, N., Cairns, S., Chan, T.Y., Cheng, L., Collins, A., Cribb, T., Curini-Galletti,

M., Dahdouh-Guebas, F., Davie, P., Dawson, M., De Clerck, O., Decock, W., De Grave, S., de Voogd, N., Domning, D., Emig, C., Erséus, C., Eschmeyer, W., Fauchald, K., Fautin, D., Feist, S., Fransen, C., Furuya, H., Garcia-Alvarez, O., Gerken, S., Gibson, D., Gittenberger, A., Gofas, S., Gómez-Daglio, L., Gordon, D., Guiry, M., Hernandez, F., Hoeksema, B., Hopcroft, R., Jaume, D., Kirk, P., Koedam, N., Koenemann, S., Kolb, J., Kristensen, R., Kroh, A., Lambert, G., Lazarus, D., Lemaitre, R., Longshaw, M., Lowry, J., Macpherson, E., Madin, L., Mah, C., Mapstone, G., McLaughlin, P., Mees, J., Meland, K., Messing, C., Mills, C., Molodtsova, T., Mooi, R., Neuhaus, B., Ng, P., Nielsen, C., Norenburg, J., Opresko, D., Osawa, M., Paulay, G., Perrin, W., Pilger, J., Poore, G., Pugh, P., Read, G., Reimer, J., Rius, M., Rocha, R., Saiz-Salinas, J., Scarabino, V., Schierwater, B., Schmidt-Rhaesa, A., Schnabel, K., Schotte, M., Schuchert, P., Schwabe, E., Segers, H., Self-Sullivan, C., Shenkar, N., Siegel, V., Sterrer, W., Stöhr, S., Swalla, B., Tasker, M., Thuesen, E., Timm, T., Todaro, M., Turon, X., Tyler, S., Uetz, P., van der Land, J., Vanhoorne, B., van Ofwegen, L., van Soest, R., Vanaverbeke, J., Walker-Smith, G., Walter, T., Warren, A., Williams, G., Wilson, S., Costello, M., 2012. The magnitude of global marine species diversity. Current Biology 22, 2189–2202. doi:10.1016/j.cub.2012.09.036.

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E.L., Nguyen, H.A., McRae, L., Possingham, H.P., Lindenmayer, D.B., 2019. Do big unstructured biodiversity data mean more knowledge? Frontiers in Ecology and Evolution , 239. doi:10.3389/fevo.2018.00239.

Beck, J., Böller, M., Erhardt, A., Schwanghart, W., 2014. Spatial bias in the gbif database and its effect on modeling species' geographic distributions. Ecological Informatics 19, 10–15. doi:10.1016/j.ecoinf.2013.11.002.

Boettiger, C., Lang, D.T., Wainwright, P., 2012. Rfishbase: exploring, ma-

nipulating and visualizing fishbase data from r. Journal of Fish Biology 81, 2030–2039. doi:10.1111/j.1095-8649.2012.03464.x.

Bonnet-Lebrunm A.S., Sweetlove, A., Griffiths, H.J., Sumner, M., Provoost, P., Raymond, B., Ropert-Coudert, Y., Van de Putte, A.P., 2023. Opportunities and limitations of large open biodiversity occurrence databases in the context of Frontiers in Marine Science 10, 1–13. doi:10.3389/fmars.2023.1150603.

Butchart, S.H.M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J.P.W., Almond, R.E.A., Baillie, J.E.M., Bomhard, B., Brown, C., Bruno, J., Carpenter, K.E., Carr, G.M., Chanson, J., Chenery, A.M., Csirke, J., Davidson, N.C., Dentener, F., Foster, M., Galli, A., Galloway, J.N., Genovesi, P., Gregory, R.D., Hockings, M., Kapos, V., Lamarque, J.F., Leverington, F., Loh, J., McGeoch, M.A., McRae, L., Minasyan, A., Morcillo, M.H., Oldfield, T.E.E., Pauly, D., Quader, S., Revenga, C., Sauer, J.R., Skolnik, B., Spear, D., Stanwell-Smith, D., Stuart, S.N., Symes, A., Tierney, M., Tyrrell, T.D., Vié, J.C., Watson, R., 2010. Global biodiversity: Indicators of recent declines. Science 328, 1164–1168. doi:10.1126/science.1187512.

Chamberlain, S., 2017. rgbif: Interface to the global "biodiversity" information facility "api". r package version 0.9.8. URL: https://CRAN.R-project.org/package=rgbif.

Chandler, M., See, L., Copas, K., Bonde, A.M., López, B.C., Danielsen, F., Legind, J.K., Masinde, S., Miller-Rushing, A.J., Newman, G., et al., 2017. Contribution of citizen science towards international biodiversity monitoring. Biological conservation 213, 280–294. doi:10.1016/j.biocon.2016.09.004.

Chao, A., Colwell, R.K., Lin, C.W., Gotelli, N.J., 2009. Sufficient sampling

for asymptotic minimum species richness estimators. Ecology 90, 1125–1133. doi:10.1890/07-2147.1.

Cheung, W.W., Lam, V.W., Sarmiento, J.L., Kearney, K., Watson, R., Pauly, D., 2009. Projecting global marine biodiversity impacts under climate change scenarios. Fish and fisheries 10, 235–251. doi:10.1111/j.1467-2979.2008.00315.x.

Chollett, I., Robertson, D.R., 2020. Comparing biodiversity databases: Greater Caribbean reef fish Fish and Fisheries 21, 1195–1212. doi:10.1111/faf.12497.

Cohen, D.M., Inada, T., Iwamoto, T., Scialabba, N., 1990. Gadiform fishes of the world. FAO Fisheries Synopsis 10, I.

Costello, M.J., Tsai, P., Wong, P.S., Cheung, A.K.L., Basher, Z., Chaudhary, C., 2017. Marine biogeographic realms and species endemicity. Nature Communications 8, 1057. doi:10.1038/s41467-017-01121-2.

Costello, M.J., Cheung, A., De Hauwere, N., 2010. Surface area and the seabed area, volume, depth, slope, and topographic variation for the world's seas, oceans, and countries. Environmental Science & Technology 44, 8821–8828. doi:10.1021/es1012752.

Costello, M.J., Vanhoorne, B., Appeltans, W., 2015. Conservation of biodiversity through taxonomy, data publication, and collaborative infrastructures. Conservation Biology 29, 1094–1099. doi:10.1111/cobi.12496.

Daly, A.J., Baetens, J.M., De Baets, B., 2018. Ecological diversity: Measuring the unmeasurable. Mathematics 6, 119. doi:10.3390/math6070119.

FAO, 2014. Fao statistical areas for fishery purposes. fao fisheries and aquaculture department [online] URL: http://www.fao.org/fishery/area/search/en.

Froese, R., Pauly, D., 2000. FishBase 2000: concepts designs and data sources. volume 1594. The WorldFish Center]. URL: http://hdl.handle.net/20.500.12348/2428.

Froese, R., Pauly, D.E., 2021. Fishbase. URL: https://www.fishbase.org.

García-Roselló, E., Guisande, C., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., Manjarrás-Hernández, A., Vaamonde, A., Granado-Lorencio, C., 2013. Modestr: a software tool for managing and analyzing species distribution map databases. Ecography 36, 1202–1207. doi:10.1111/j.1600-0587.2013.00374.x.

García-Roselló, E., Guisande, C., Heine, J., Pelayo-Villamil, P., Manjarrés-Hernández, A., González Vilas, L., González-Dacosta, J., Vaamonde, A., Granado-Lorencio, C., 2014. Using modestr to download, import and clean species distribution records. Methods in ecology and evolution 5, 708–713. doi:10.1111/2041-210X.12209.

García-Roselló, E., Guisande, C., Manjarrés-Hernández, A., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., González-Vilas, L., Vari, R.P., Vaamonde, A., Granado-Lorencio, C., et al., 2015. Can we derive macroecological patterns from primary global biodiversity information facility data? Global Ecology and Biogeography 24, 335–347. doi:10.1111/geb.12260.

GBIF: The Global Biodiversity Information Facility , 2021. What is gbif? URL: https://www.gbif.org/what-is-gbif.

GBIF.org, 2021. Occurrence download. URL: https://www.gbif.org/occurrence/download/0039590-210914110416597, doi:10.15468/DL.V2PFS3. last accessed 29 October 2021.

Giraud, T., Lambert, N., 2016. cartography: Create and integrate maps in

your r workflow. Journal of Open Source Software 1, 54. doi:10.21105/joss.00054.

Guisande, C., Lobo, J., 2019. Discriminating well surveyed spatial units from exhaustive biodiversity databases. r package version. 2.0. URL: http://cran.r-project.org/web/packages/KnowBR.

Heberling, J.M., Miller, J.T., Noesgaard, D., Weingart, S.B., Schigel, D., 2021. Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences 118, e2018093118. doi:10.1073/pnas.2018093118.

Hollowed, A.B., Barange, M., Beamish, R.J., Brander, K., Cochrane, K., Drinkwater, K., Foreman, M.G., Hare, J.A., Holt, J., Ito, S.i., et al., 2013. Projected impacts of climate change on marine fish and fisheries. ICES Journal of Marine Science 70, 1023–1037. doi:10.1093/icesjms/fst081.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M., Baselga, A., 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117, 847–858. doi:10.1111/j.0030-1299.2008.16434.x.

Hutchings, J.A., Baum, J.K., 2005. Measuring marine fish biodiversity: temporal changes in abundance, life history and demography. Philosophical Transactions of the Royal Society B: Biological Sciences 360, 315–338. doi:10.1098/rstb.2004.1586.

Jin, J., Yang, J., 2020. Bdcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. Global Ecology and Conservation 21, e00852. doi:10.1016/j.gecco.2019.e00852.

Kroodsma, D.A., Mayorga, J., Hochberg, T., Miller, N.A., Boerder, K., Ferretti, F., Wilson, A., Bergman, B., White, T.D., Block, B.A., et al.,

2018. Tracking the global footprint of fisheries. Science 359, 904–908. doi:10.1126/science.aao5646.

Lenoir, J., Bertrand, R., Comte, L., Bourgeaud, L., Hattab, T., Murienne, J., Grenouillet, G., 2020. Species better track climate warming in the oceans than on land. Nature Ecology & Evolution 4, 1044–1059. doi:10.1038/s41559-020-1198-2.

Lobo, J.M., Hortal, J., Yela, J.L., Millán, A., Sánchez-Fernández, D., García-Roselló, E., González-Dacosta, J., Heine, J., González-Vilas, L., Guisande, C., 2018. Knowbr: An application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. Ecological Indicators 91, 241–248. doi:10.1016/j.ecolind.2018.03.077.

Luypaert, T., Hagan, J.G., McCarthy, M.L., Poti, M., 2020. Status of marine biodiversity in the anthropocene, in: YOUMARES 9-The Oceans: Our research, our future. Springer, Cham, pp. 57–82. doi:10.1007/978-3-030-20389-4_4.

Magurran, A.E., McGill, B.J., 2011. Biological diversity: frontiers in measurement and assessment. Oxford University Press. doi:10.1086/666756.

Malhi, Y., Franklin, J., Seddon, N., Solan, M., Turner, M.G., Field, C.B., Knowlton, N., 2020. Climate change and ecosystems: Threats, opportunities and solutions. doi:10.1098/rstb.2019.0104.

Marquet, P.A., Fernández, M., Navarrete, S.A., Valdovinos, C., 2004. Diversity emerging: towards a deconstruction of biodiversity patterns, in: Lomolino, M., Heaney, L. (Eds.), Frontiers of Biogeography: New directions in the Geography of Nature. Cambridge University Press, pp. 191–209.

Melo-Merino, S.M., Reyes-Bonilla, H., Lira-Noriega, A., 2020. Ecological niche models and species distribution models in marine environments: A

literature review and spatial analysis of evidence. Ecological Modelling 415, 108837. doi:10.1016/j.ecolmodel.2019.108837.

Meyer, C., Kreft, H., Guralnick, R., Jetz, W., 2015. Global priorities for an effective information basis of biodiversity distributions. Nature communications 6, 1–8. doi:10.1038/ncomms9221.

Mora, C., Tittensor, D.P., Myers, R.A., 2008. The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. Proceedings of the Royal Society B: Biological Sciences 275, 149–155. doi:10.1098/rspb.2007.1315.

Moudrý, V., Devillers, R., 2020. Quality and usability challenges of global marine biodiversity data Ecological Informatics 56, 101051. doi:10.1016/j.ecoinf.2020.101051.

Moreno, C.E., Rodríguez, P., 2011. Do we have a consistent terminology for species diversity? back to basics and toward a unifying framework. Oecologia 167, 889–892. doi:10.1007/s00442-011-2125-7.

Neigel, J., 1997. Marine Biodiversity: Patterns and Processes. Cambridge, Cambridge University Press. chapter Population genetics and demography of marine species. URL: http://www.cambridge.org/9780521552226.

OBIS: Ocean Biodiversity Information System, 2021. About obis URL: https://obis.org/.

OBIS.org, 2021. Occurrence download. URL: https://datasets.obis.org/downloads/9fd73b2a-cf6f-4ef9-a0e3-2d1f653520d3.zip. last accessed 29 October 2021.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., , Wagner, H., 2020. The vegan package URL: https://github.com/vegandevs/vegan.

Oliver, R.Y., Meyer, C., Ranipeta, A., Winner, K., Jetz, W., 2021. Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. PLoS Biology 19, e3001336. doi:10.1371/journal.pbio.3001336.

O'Hara, C.C., Frazier, M., Halpern, B.S., 2021. At-risk marine biodiversity faces extensive, expanding, and intensifying human impacts. Science 372, 84–87. doi:10.1126/science.abe6731.

Pauly, D., Palomares, M.L., 2005. Fishing down marine food web: it is far more pervasive than we thought. Bulletin of marine science 76, 197–212.

Pebesma, E.J., 2018. Simple features for r: standardized support for spatial vector data. R J. 10, 439. doi:10.32614/RJ-2018-009.

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. doi:10.1007/s10531-018-1630-y.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013. Essential biodiversity variables. Science 339, 277–278. doi:10.1126/science.1229931.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological applications 19, 181–197. doi:10.1890/07-2153.1.

Pope, E.C., Hays, G.C., Thys, T.M., Doyle, T.K., Sims, D.W., Queiroz, N., Hobson, V.J., Kubicek, L., Houghton, J.D., 2010. The biology and ecology of the ocean sunfish mola mola: a review of current knowledge and future research perspectives. Reviews in Fish Biology and Fisheries 20, 471–487. doi:10.1007/s11160-009-9155-9.

Provoost, P., Bosch, S., 2020. robis: R client to access data from the obis api. ocean biogeographic information system, intergovernmental oceanographic commission of unesco URL: https://cran.r-project.org/package=robis.

R Core Team, 2018. R: A language and environment for statistical computing. vienna, austria: R foundation for statistical computing URL: https://www.r-project.org/.

Ramírez, F., Afán, I., Davis, L.S., Chiaradia, A., 2017. Climate impacts on global hot spots of marine biodiversity. Science Advances 3, e1601198. doi:10.1126/sciadv.1601198.

Sala, E., Mayorga, J., Bradley, D., Cabral, R.B., Atwood, T.B., Auber, A., Cheung, W., Costello, C., Ferretti, F., Friedlander, A.M., et al., 2021. Protecting the global ocean for biodiversity, food and climate. Nature 592, 397–402. doi:10.1038/s41586-021-03371-z.

Smith, F.A., Gittlemann, J.L., Brown, J.H., 2014. Foundations of macroecology: classic papers with commentaries. University of Chicago Press.

Telenius, A., 2011. Biodiversity information goes public: Gbif at your service. Nordic Journal of Botany 29, 378–381. doi:10.1111/j.1756-1051.2011.01167.x.

Tittensor, D.P., Mora, C., Jetz, W., Lotze, H.K., Ricard, D., Berghe, E.V., Worm, B., 2010. Global patterns and predictors of marine biodiversity across taxa. Nature 466, 1098–1101. doi:10.1038/nature09329.

Troia, M.J., McManamay, R.A., 2016. Filling in the gaps: evaluating completeness and coverage of open-access biodiversity databases in the united states. Ecology and evolution 6, 4654–4669. doi:10.1002/ece3.2225.

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. doi:10.1111/ddi.12637.

Tuomisto, H., 2011. Do we have a consistent terminology for species diversity? yes, if we choose to use it. Oecologia 167, 903–911. doi:10.1007/s00442-011-2128-4.

Turner, M.G., Calder, W.J., Cumming, G.S., Hughes, T.P., Jentsch, A., LaDeau, S.L., Lenton, T.M., Shuman, B.N., Turetsky, M.R., Ratajczak, Z., et al., 2020. Climate change, ecosystems and abrupt change: science priorities. Philosophical Transactions of the Royal Society B 375, 20190105. doi:10.1098/rstb.2019.0105.

UNEP-WCMC, IUCN, 2022. Protected Planet: The World Database on Protected Areas (WDPA) [Online], January 2022, Cambridge, UK. Technical Report. URL: https://www.protectedplanet.net.

Visalli, M.E., Best, B.D., Cabral, R.B., Cheung, W.W., Clark, N.A., Garilao, C., Kaschner, K., Kesner-Reyes, K., Lam, V.W., Maxwell, S.M., et al., 2020. Data-driven approach for highlighting priority areas for protection in marine areas beyond national jurisdiction. Marine Policy 122, 103927. doi:10.1016/j.marpol.2020.103927.

Webb, T.J., Vanden Berghe, E., O'Dor, R., 2010. Biodiversity's big wet secret: the global distribution of marine biological records reveals chronic under-exploration of the deep pelagic ocean. PloS one 5, e10223. doi:10.1371/journal.pone.0010223.

Wickham, H., Francois, R., Henry, L., Müller, K., 2021. dplyr: A grammar of data manipulation. r package version 1.0.3. R Found. Stat. Comput., Vienna URL: https://CRAN.R-project.org/package=dplyr.

WORMS, 2022. World register of marine species database: Statistics. number of records in worms 11th april 2022 [online] URL: http://www.marinespecies.org/.

Yang, W., Ma, K., 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. doi:10.1111/jbi.12108.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A., et al., 2007. How global is the global biodiversity information facility? PloS one 2, e1124. doi:10.1371/journal.pone.0001124.

Zhang, Y., Grassle, J.F., 2002. A portal for the ocean biogeographic information system. Oceanologica Acta 25, 193–197. doi:10.1016/S0399-1784(02)01204-5.

Zizka, A., Carvalho, F.A., Calvente, A., Baez-Lizarazo, M.R., Cabral, A., Coelho, J.F.R., Colli-Silva, M., Fantinati, M.R., Fernandes, M.F., Ferreira-Araújo, T., et al., 2020. No one-size-fits-all solution to clean gbif. PeerJ 8, e9916. doi:10.7717/peerj.9916.

Appendix A. The database

Table A.1 below shows the data loss for each criterion that we have used to clean our database. We downloaded 71,670,596 records from GBIF and OBIS. Only 820,004 records were useful for our analyses.

Database state	Number of records
Origial records from GBIF and OBIS	71,670,596
Data curation (following Zizka et al. (2020))	5,380,439
Taxonomically filtered data	5,007,322
Deletion of data outside the native range	820,004

Table A.1: Criteria for filtering occurrence data from GBIF and OBIS using bioregions.

Files of the 10,371 marine fish species and their attributes (body size, habitat depth, and cultural value) from FishBase may be found in the GitHub project page of this manuscript: http://github.com/vapizarro/stp_fishes

Appendix B. Species Representativeness Analysis (SRI)

For each cell (i), the SRI is the simple ratio between the observed number of species S_{obs} and the expected number of species (S_{obs}): $SRI_i = S_{obs}/S_{exp}$. Maps for the smallers resolution analyzed (~~$\sim 1° \times 1°$~~ $\sim 1°$) are in Fig. A.2.

Appendix C. Grids resolutions

For spatial representation analysis we evaluated two additional spatial resolutions (~~$5° \times 5°$~~ $\sim 5°$ = 3,021 cells, and ~~$10° \times 10°$~~ $\sim 10°$ = 958 cells). Table C.2 contains the results of this analysis for these grids. We have also mapped these results (see Figure A.3), to understand how the effect of spatial resolution on the evaluation of biodiversity macropatterns. Finally, we also plot the frequency of cells for each SRI category for the three grid sizes (R1= ~~$1° \times 1°$~~ $\sim 1°$; R5= ~~$5° \times 5°$~~ $\sim 5°$; R10= ~~$10° \times 10°$~~ $\sim 10°$) to understand how the data is distributed in our analyses (see Figure A.4)

	R1 (∼ 1°)					R5 (∼ 5°)					R10 (∼ 10°)				
ID	**NR**	**IR**	**L**	**M**	**H**	**NR**	**IR**	**L**	**M**	**H**	**NR**	**IR**	**L**	**M**	**H**
1	18.49	15.13	5.04	36.97	24.37	0.00	16.67	0.00	33.33	50.00	16.67	16.67	0.00	33.33	33.33
2	68.75	19.79	1.04	10.42	0.00	10.00	40.00	10.00	30.00	10.00	40.00	0.00	0.00	40.00	20.00
3	15.74	6.54	3.39	36.80	37.53	3.57	3.37	0.00	17.86	75.00	0.00	0.00	0.00	10.00	90.00
4	46.35	22.34	7.93	22.13	1.25	28.13	9.38	6.25	40.63	15.63	30.77	15.38	0.00	23.08	30.77
5	42.39	14.75	4.92	27.87	10.07	14.29	3.57	0.00	32.14	50.00	16.67	8.33	0.00	8.33	66.67
6	94.96	2.21	0.13	1.87	0.83	82.13	5.64	0.31	6.58	5.33	62.65	13.25	1.20	10.84	12.05
7	63.24	11.24	0.87	14.46	10.19	17.09	9.40	3.42	29.06	41.03	7.69	7.69	2.56	35.90	46.15
8	79.52	11.27	0.89	7.17	1.15	43.93	11.56	4.05	32.37	8.09	32.69	11.54	3.85	40.38	11.54
9	88.74	8.71	0.00	1.57	0.99	28.74	22.99	2.87	38.51	6.90	7.69	9.62	21.15	38.08	13.46
10	96.31	2.41	0.04	0.88	0.36	70.87	15.75	0.79	7.09	5.51	51.28	15.38	2.56	20.51	10.26
11	23.82	8.42	5.65	32.85	29.26	8.62	0.00	0.00	18.97	72.41	0.00	10.53	0.00	5.26	84.21
12	35.59	21.61	2.45	35.59	4.76	14.29	4.76	2.38	47.62	30.95	5.88	11.76	0.00	17.65	64.71
13	67.52	15.80	1.01	12.00	3.67	13.76	12.84	7.34	44.95	21.10	9.46	6.76	2.70	44.59	36.49
14	45.83	10.83	2.50	30.83	10.00	46.15	0.00	0.00	7.69	46.15	25.00	0.00	0.00	0.00	75.00
15	74.52	13.06	0.00	7.07	5.35	20.00	6.67	6.67	40.00	26.67	37.50	12.50	0.00	37.50	12.50
16	36.68	10.95	3.84	34.65	13.88	5.77	7.69	3.85	28.85	53.85	10.53	0.00	0.00	21.05	68.42
17	91.36	4.90	0.00	1.57	2.17	47.93	19.01	0.00	20.66	12.40	25.00	8.33	0.00	36.11	30.56
18	48.29	16.27	3.78	22.06	9.61	6.50	7.32	7.32	43.09	35.77	10.26	5.13	0.00	28.21	56.41
19	90.40	6.93	0.06	2.27	0.35	53.45	18.39	3.45	17.82	6.90	31.48	12.96	1.85	35.19	18.52
20	63.61	17.35	1.43	13.56	4.04	8.20	8.20	9.84	44.26	29.51	15.00	5.00	0.00	45.00	35.00
21	74.78	9.63	2.84	11.48	1.27	34.68	13.51	3.15	28.38	20.27	21.21	9.09	0.00	27.27	42.42
22	76.12	18.00	0.00	5.10	0.78	33.33	6.17	16.05	43.21	1.23	9.09	4.55	9.09	59.09	18.18
23	34.65	32.02	2.89	24.41	6.04	25.93	0.00	14.81	51.85	7.41	0.00	18.18	0.00	36.36	45.45
24	63.07	17.89	1.38	13.53	4.13	25.93	7.41	3.70	51.85	11.11	20.00	10.00	0.00	50.00	20.00
25	88.02	4.96	0.83	5.37	0.83	52.63	10.53	0.00	21.05	15.79	42.86	0.00	0.00	28.57	28.57
26	60.93	7.04	2.41	22.41	7.22	27.27	12.12	0.00	30.30	30.30	16.67	0.00	0.00	33.33	50.00
27	66.84	10.35	0.70	8.07	14.04	27.78	16.67	0.00	22.22	33.33	7.69	7.69	0.00	23.08	61.54
28	59.84	9.17	2.13	17.67	11.19	30.19	7.55	1.89	30.19	30.19	30.00	10.00	0.00	25.00	35.00
29	41.96	19.87	2.84	26.81	8.52	15.00	10.00	5.00	40.00	30.00	12.50	12.50	0.00	37.50	37.50
30	93.74	3.49	0.20	1.97	0.59	69.45	11.02	1.00	10.52	8.01	42.29	16.57	2.86	22.29	16.00

Table C.2: Surface area as a ~~percentage~~ share of each bioregion (ID) for every SRI category for each of the three grid sizes (R1= ~~1°× 1°~~∼ 1°; R5= ~~5°× 5°~~∼ 5°; R10= ~~10°× 10°~~∼ 10°). Values show the surface area as a ~~percentage~~ share of each bioregion for every SRI category (see §2.2.1). ID is the identification number given to each bioregion (Table 1). ~~A~~ **H** are cells with ~~an adequate~~ a *high* representativeness of species richness (i.e. SRI > 0.85). ~~S~~ **M** are cells considered as having a ~~sufficient~~ *medium* representativeness (i.e. SRI ∈ (0.60, 0.85)). ~~F~~ **L** cells are cells with ~~few~~ a *low* number of records and are thus not considered to be representative of actual species richness (i.e. SRI ∈ (0, 0.6)). ~~NR~~ **NR** as cells with ~~no records~~ *no records* (SRI= ~~NA~~*NA*), and ~~IR~~ **IR** as cell with ~~insufficient records~~ *insufficient records* to apply SRI.

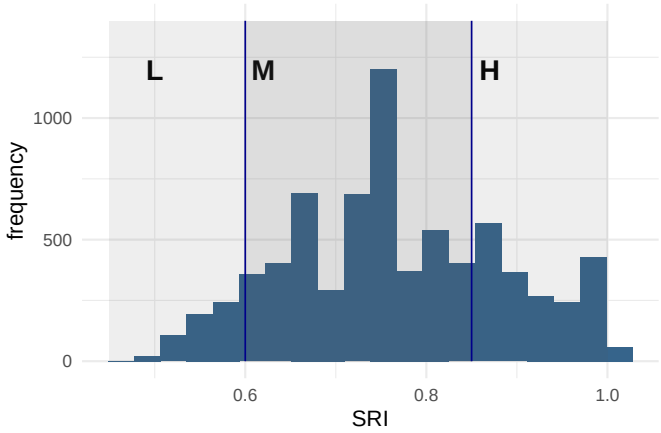


Fig. A.1: Classification of SRI values based on its frequency distribution. This histogram displays the frequency distribution of SRI (Species Richness Index) values and the corresponding class selection thresholds. Cells are categorized as follows: SRI < 0.6 are classified as "~~Few~~ *low* representativeness (**L**), "~~SRI falling in the range (0.6, 0.85) as "**S**~~ufficient~~ *medium* representativeness (**M**), "~~and SRI > 0~~ 0.85 as "~~**A**dequate"~~ *high* representativeness (**H**).

Appendix D. Bioregions slopes

We evaluated the slopes of the last 10% of the accumulation curves of each bioregion in our temporal representation analysis. Table D.3 shows the result for each bioregion.

Appendix E. GAP Analysis

We plotted the ~~percentage of surface with marine protected areas of~~ share of surface areas with MPAs in each bioregion (Fig A.5), and the ~~percentage~~ share of cells of each FAO ~~Area~~ area for each category of SRI value (Fig A.6).

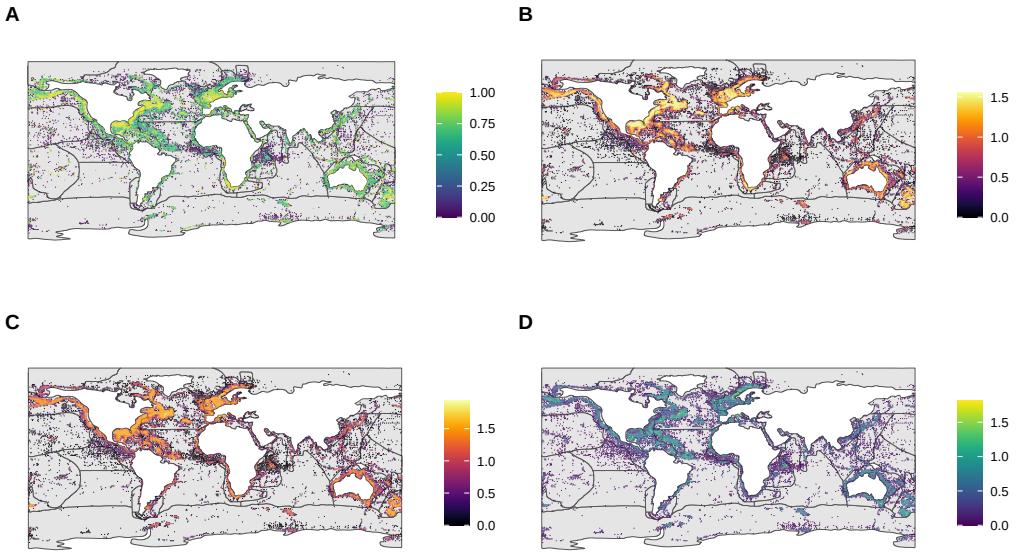


Fig. A.2: SRI and Species richness S depicted from GBIF and OBIS databases. **A**. Species representativeness index; **B**. Observed species richness (S_{obs}); **C**. Expected species richness (S_{exp}); **D**. Difference between raw estimated and observed richness. The difference has been log_{10} transformed after subtraction.

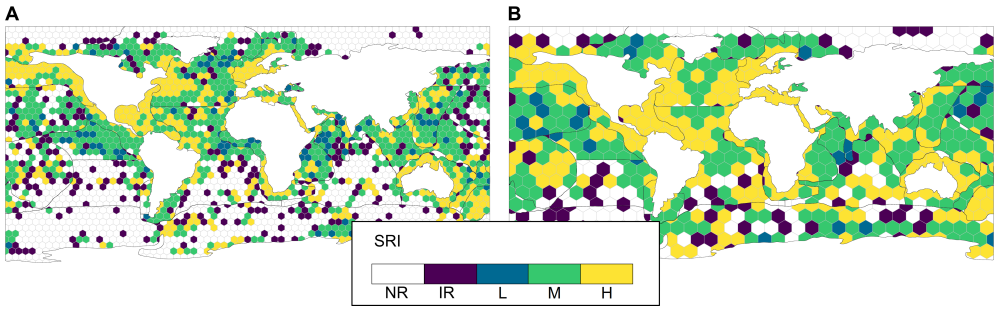


Fig. A.3: Spatial representativeness index (SRI) mapping of cells of size: A=~~5° × 5°~~ ∼ 5°; B= ~~10° × 10°~~ ∼ 10°. The categorization of the cells corresponds to the level reached by the SRI, where SRI > 0.85: Amount of data ~~*Adequate*~~ is *high* for the representation of species richness (~~"A"~~ **H**); SRI=0.60-0.85: Amount of data can be considered ~~*Sufficient*~~ of *medium* representativeness (~~"S"~~ **M**); SRI=0-0.60: Amount of records ~~*Few*~~ is *low* (~~"F"~~ **L**); and SRI = NA: cells with no records (~~"NR"~~ **NR**). **IR** are cells with *insufficient records* to evaluate species diversity representativeness.

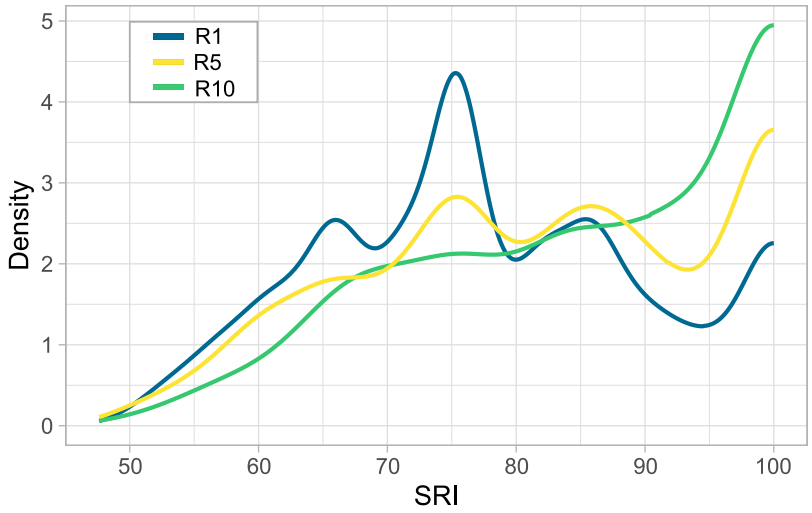


Fig. A.4: Density probability distribution of SRI in three grids of different sizes: R1= ~~1° × 1°~~ ∼ 1°(blue line); R5= ~~5° × 5°~~ ∼ 5° (red line); and R10= ~~10° × 10°~~ ∼ 10° (yellow line).

Bioregion	Slope
1	0.35
2	1.16
3	1.79
4	0.91
5	1.76
6	0.65
7	4.44
8	1.37
9	6.18
10	4.87
11	10.37
12	7.57
13	32.86
14	4.90
15	6.78
16	21.62
17	10.10
18	6.59
19	12.44
20	23.21
21	11.70
22	1.85
23	4.42
24	3.49
25	2.12
26	7.74
27	12.29
28	4.82
29	14.08
30	2.74

Table D.3: Final slope (10%) of the accumulation curves for each bioregion

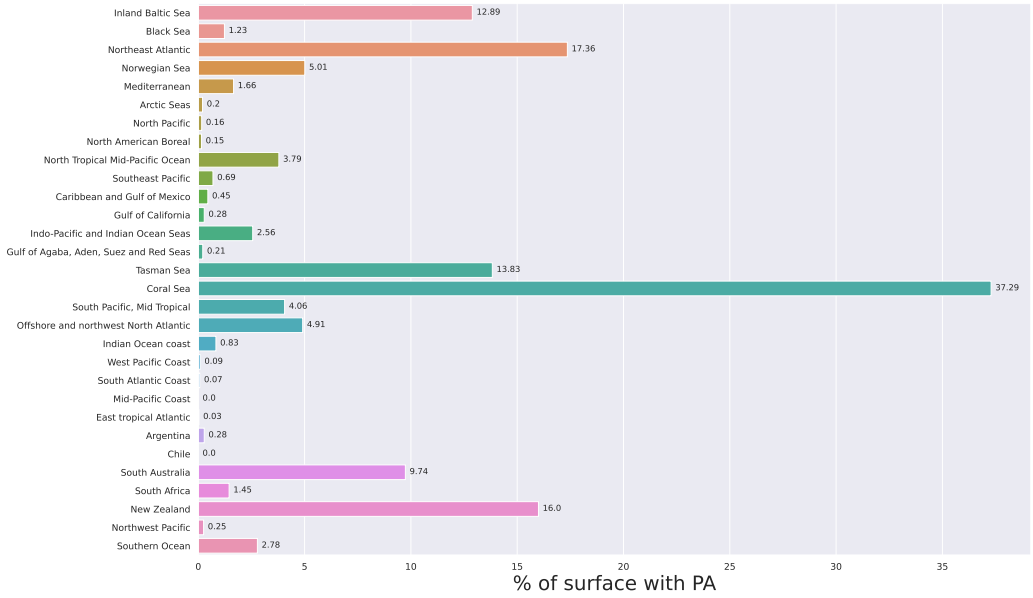


Fig. A.5: ~~Percentage~~ Share of surface area with ~~marine protected areas~~ MPAs by bioregions.

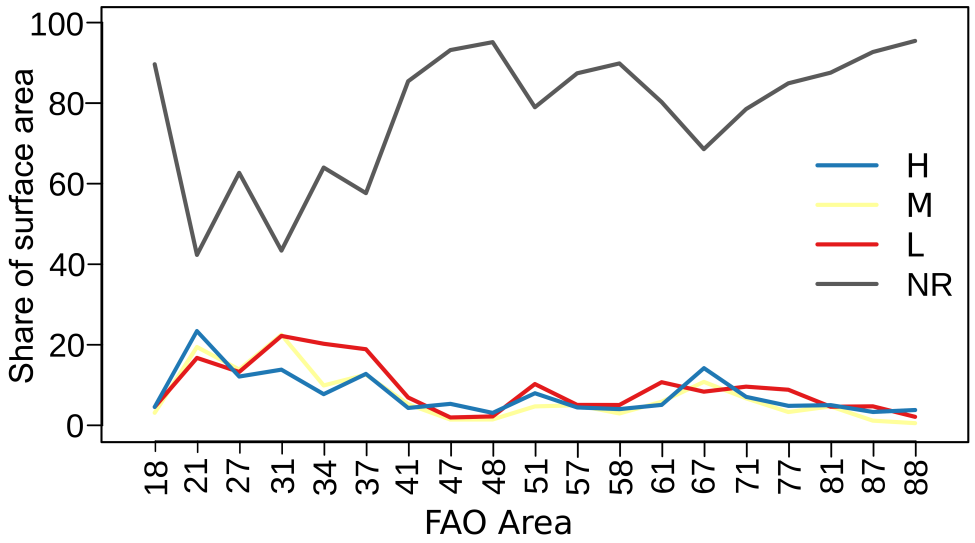


Fig. A.6: ~~Percentage~~ Share of cells ~~of~~ each FAO ~~Area~~ area for each ~~category of~~ SRI value category. ~~Amount of data *Adequate* for the representation~~ SRI= >0.85: *High* representativeness of observed species richness (~~"A"~~**H**); SRI=0.60-0.85: ~~Amount~~ *Medium* representativeness of ~~data can be considered *Sufficient*~~ observed species richness (~~"S"~~**M**); SRI=0-0.60: ~~Amount~~ *Low* representativeness of ~~records *Few*~~ observed species richness (~~"F"~~**L**); and SRI= ~~NA~~*NA*: cells with no records (~~"NR"~~**NR**).

Spatial and Temporal Representation of Marine Fish Occurrences Available Online

Vanessa Pizarro[a], Andrea G. Castillo[a,b], Andrea Piñones[c,d,e,f], Horacio Samaniego[a,g,*]

[a]*Laboratorio de Ecoinformática, Instituto de Conservación, Biodiversidad y Territorio, Universidad Austral de Chile, Valdivia, Chile*
[b]*Programa de Doctorado en Ciencias mención Ecología y Evolución, Escuela de Graduados, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile*
[c]*Instituto de Ciencias Marinas y Limnológicas, Facultad de Ciencias, Universidad Austral de Chile, Valdivia, Chile*
[d]*Centro FONDAP de Investigación en Dinámica de Ecosistemas Marinos de Altas Latitudes (IDEAL), Valdivia, Chile*
[e]*Centro de Investigación Oceanográfica COPAS-COASTAL, Universidad de Concepción, Chile*
[f]*Millenium Institute Biodiversity of Antarctic and Subantarctic Ecosystems - BASE, Chile*
[g]*Instituto de Sistemas Complejos de Valparaíso, Subida Artillería 470, Valparaíso, Chile*

Abstract

Despite the 243,000 marine species described by 2022, our knowledge about the oceanic biodiversity is still incomplete. This knowledge gap carries potentially adverse and far-reaching consequences for the preservation of marine ecosystems, particularly in the context of the ongoing human-induced alterations to our biosphere and the rapid progression of climate change and global environmental shifts.

Recently, however, a large number of online repositories have emerged, which catalogue, store and distribute biodiversity information, including taxonomic and species occurrence data. FishBase, the Global Biodiversity Information Facility (GBIF) and the Ocean Biodiversity Information System (OBIS) are part of these publicly available repositories representing a variety of sources that have exploded in number. However, despite the incredible accumulation of biodiversity records, not all the information is actually useful, nor does it represent any new knowledge regarding global species richness patterns.

In this study, we assessed the spatial and temporal representativeness

*Corresponding author
Email address: horacio@ecoinformatica.cl (Horacio Samaniego)

of marine fish records (order Actinopterygii) found in the GBIF and OBIS global repositories. The methodological framework that we developed relies on a series of non-parametric estimators for computing species richness from incidence data. This methodology employs hexagonal grids as sampling units that overlay marine bioregions across the globe.

Using standard ecological and spatial analysis tools, we identify regions that are adequately represented in terms of available records and therefore have more reliable data, as well as regions with few records that do not represent current species richness. We overlap these results with the location of marine protected areas and fishing exploitation zones to understand the anthropogenic effect on marine ichthyofauna. We additionally evaluate hypotheses regarding the taxonomic, geographic, and temporal distribution of information biases to deepen our current understanding of public records of species occurrences worldwide.

Considering that more than 40 years of information was analyzed, the results showed that, on a global scale, the primary data on marine fish available on GBIF and OBIS platforms are still far from being representative and complete. Only 1.14% of the records were useful for our analyses. In addition, we found that the information seems to be biased towards coastal areas, regions close to developed countries, and areas where there is a large fishing activity. Finally, the best represented species and families are those with a small body size, which use shallow habitats and are usually recognized as having commercial or cultural value.

Keywords: Ecoinformatics, Ecological Information Biases, Marine Fish, Spatial and Temporal Representativeness, Species Richness

1. Introduction

Currently, the more than 243,000 species included in the World Register of Marine Species database (WORMS, 2022) suggests that only 11% to 78% of all marine species have been discovered, revealing a striking picture of vastly incomplete knowledge that may have serious implications for marine conservation (Luypaert et al., 2020). Moreover, ongoing climate change represents one of the greatest threats to biodiversity (Malhi et al., 2020; Turner et al., 2020) and has already been documented to modify the distribution of

marine species (Lenoir et al., 2020). Some of the effects described includes the invasion of non-native species leading to massive species turnover that may result in the local extinction of large share of species (Cheung et al., 2009).

It is crucial to recognize that species richness, while being a diversity metric among many, is, in itself, an aggregate variable quantifying the end result of the splitting and lumping of the tree of life as a product of evolutionary processes (Marquet et al., 2004). Consequently, numerous endeavors have been directed towards the development of more comprehensive diversity indices, giving rise to significant scientific literature, aimed at describing ecological heterogeneity (Tuomisto, 2011; Moreno and Rodríguez, 2011; Daly et al., 2018). However, within this literature, there appears to be a shifting focus towards examining the ramifications of biodiversity loss. This shift involves the adoption of new terminology designed to provide pragmatic concepts, such as "species inventory", "taxonomic inventory", or "inventory completeness", which are intended to convey more precise messages to policymakers, summarizing the richness of biodiversity (Pereira et al., 2013; Butchart et al., 2010). Nevertheless, while the scientific community engages in debates over the use of biodiversity terminology, it is important to note that species richness continues to offer a concise and easily manageable description of variability across various other parameters characterizing the biota in both spatial and temporal dimensions (Appeltans et al., 2012). Species richness remains an essential feature for comprehending how diversity evolves in response to natural and anthropogenic influences within biomes, regions, and ecosystems (Troia and McManamay, 2017; Magurran and McGill, 2011).

Likewise, biodiversity can also be assessed through life history traits, which are modulated by both evolutionary factors and habitat ecosystem variations (Neigel, 1997; Hutchings and Baum, 2005). We now know that biodiversity is more likely an expression of the heterogeneity of such life history traits. Alò et al. 2021, for example, show that while some of the fish

diversity is certainly due to environmental processes, a large fraction of such richness variance is also determined by evolved life history traits related, for example, to migratory habits. Therefore, evaluating how life history traits impact richness metrics should deepen our understanding of fish diversity patterns.

While still short of having a robust and standardized biodiversity infrastructure (Heberling et al., 2021), there is great diversity of online repositories with taxonomic information and species occurrences data. Among the most important databases hosting marine information are FishBase, a platform that hosts information on the taxonomy of fish, their ecology, trophic information, habitat, and history of uses dating back to more than 250 years (Froese and Pauly, 2000); and the Global Biodiversity Information Facility (GBIF), a platform that stores and allows for the free access to species occurrence records from around the world. GBIF is currently one of the repositories hosting the largest amount of such data in the world (Telenius, 2011; GBIF: The Global Biodiversity Information Facility , 2021); and finally Ocean Biodiversity Information System (OBIS), which houses data on the occurrence and abundance of species from exclusively marine environments (OBIS: Ocean Biodiversity Information System, 2021). Records entered in these repositories are often used for research related to biodiversity assessment, taxonomic reviews, red listing of threatened species, species distribution, and generation of ecological niche models, among others (Yesson et al., 2007). GBIF currently offers more than 1.62 billion occurrence records and OBIS more than 63 million, which increase considerably each year (GBIF: The Global Biodiversity Information Facility , 2021; OBIS: Ocean Biodiversity Information System, 2021).

The records of both platforms come from a wide variety of sources collected following different methodologies at different temporal and spatial scales, which introduces a great variety of biases (Beck et al., 2014; Zizka et al., 2020). Among these, three main types of biases have been described:

(i) taxonomic, this occurs when some species and/or families are better sampled than other rarer species (Chandler et al., 2017); (ii) geographic, when data input is unevenly distributed across geographic regions and may prove to obscure interregional comparisons (Yang et al., 2013; Yesson et al., 2007); and (iii) temporal, which may be prevalent when comparing different time periods as data coverage is unevenly distributed over time (Chandler et al., 2017; Yang et al., 2013). While these biases introduce some uncertainty regarding reliability of species richness descriptions obtained from online platforms (Beck et al., 2014; García-Roselló et al., 2015), they have largely been used to provide an extensive overview of macro-ecological patterns of distribution not available otherwise (Mora et al., 2008; Troia and McManamay, 2017).

Still, identifying how sampling efforts are distributed across space and time is a required step to interpret biodiversity patterns and reduce biases, as understanding our biota distribution is critical for well-designed protection efforts. This may be achieved through different weighting schemes for records in areas with sufficient sampling that provide a more reliable contribution compared to underrepresented regions (Phillips et al., 2009; Hortal et al., 2008; Yang et al., 2013).

We here assessed the spatial and temporal representativeness of marine fish records available in the global GBIF and OBIS repositories at the marine bioregions' level in order to pinpoint the location of records that best quantify marine fish diversity. The result is a spatial representativeness analysis that we then overlay on marine conservation areas (UNEP-WCMC and IUCN, 2022) and fisheries exploitation areas (FAO, 2014) to learn whether marine conservation efforts, as well as large fisheries, are located in areas of high species richness or areas which insufficient data coverage.

Finally, we also analyzed the potential effect of some attributes on the incidence of more records in global database repositories. Specifically, we evaluated three research questions related to how body size, habitat depth,

and commercial use relate to the representation of marine fish occurrences. We ask whether: (i) a better representation in online platforms may be due to the oversampling of larger fish, resulting from an easier identification; (ii) shallow areas provide easy access to sampling; and (iii) economic and commercial interests have elicited a larger representation of culturally relevant species in online biodiversity repositories.

2. Methods

2.1. Species data

We use all the recorded occurrences of the Actinopterygii order hosted in the GBIF and OBIS repositories (GBIF.org, 2021; OBIS.org, 2021). Following Alò et al. (2021), evolutionary older taxa, such as Cephalaspidomorphi, were excluded from this analysis. Libraries *rgbif* and *robis* of the statistical package R were used for data extraction (Chamberlain, 2017; Provoost and Bosch, 2020; R Core Team, 2018). Both repositories have collaborated since 2001, sharing data on the co-occurrence of marine life (OBIS.org, 2021). Nevertheless, recent investigations have shown significant disparities in data contributions, revealing remarkable low shares of shared data (Chollett and Robertson, 2020; Moudrý and Devillers, 2020). Noteworthy distinctions exist between the two platforms, encompassing diverse data sources and methodologies, along with substantial variations in temporal and spatial scales associated with data collection (Zizka et al., 2020). Due to these disparities, scholars recommend a thorough examination and refinement of these data repositories (Bonnet-Lebrun et al., 2023). To enhance the quality and reliability of the information, a comprehensive series of filters has been systematically applied to our analysis. To minimize errors associated with the public usage of GBIF and OBIS repositories, we curated the dataset following Zizka et al. (2020) and filtered the dataset by the columns labeled "scientific name", "family", "year", "longitude" and "latitude". We retained all taxonomic information down to the species level and removed records with `NA` in these

columns. We also removed all duplicate records with identical latitude and longitude data, as well as records collected before 1980 (see Alò et al., 2021; García-Roselló et al., 2015). Each record was further assigned to a marine bioregion following Costello et al. (2017). Spatial data manipulation and plotting was performed with the aid of the following libraries: *sf*, *dplyr*, and *cartography* (Giraud and Lambert, 2016; Pebesma, 2018; Wickham et al., 2021). We finally labeled and removed all exotic species record using the `distribution()` function provided by the *rfishbase* library (Boettiger et al., 2012; Froese and Pauly, 2021). To limit our analysis to species occurring within their native range, each record was checked against the FAO fisheries area classification for consistency (FAO, 2014). A summary of the number of records is provided in Appendix A.

2.2. Data Analysis by Bioregion

Once the database was cleaned, a data subset was extracted for each of the 30 bioregions. For each bioregion, records, species, and families were counted, and the Shannon diversity index was calculated using the *vegan* library in R (Oksanen et al., 2020).

2.2.1. Spatial Representativeness Analysis

To assess the data's spatial representativeness, bioregions were gridded into hexagonal cells of the same surface area to maximize the fit whit the bioregions' areas using a cylindrical equal area projection (i.e. EPSG Code:54034). We approximated a 1° hexagonal lattice by computing cells of 10^4 square-kilometers, resulting in a total of 57,067 cells. In the appendix, we evaluated two additional spatial resolutions: $\sim 5°$ and $\sim 10°$ lattice with a total of 3,029 and 953 cells, respectively, using a 2.5×10^5, and 10^7 square-kilometer gridcell to assess different biodiversity macropatterns (Tittensor et al., 2010).

The expected species richness (S_{exp}) was computed as the mean between three non-parametric richness estimators: Chao2 (S_{chao}), Bootstrap ($S_{bootstrap}$) and Jackknife 1 ($S_{jackknife1}$) (see Magurran and McGill, 2011, for

individual index definitions). The purpose of this averaging is to minimize biases and potential under or overestimation errors by using a single richness estimator following the work by Mora et al. (2008) and Troia and McManamay (2017).

We then produced a species representativeness index (SRI) by comparing the observed richness (S_{obs}) per cell to S_{exp} (Troia and McManamay, 2017), $SRI_i = \frac{S_{obs}}{S_{exp}}$. This is an undersampling index that points to the records' representativeness to quantify the actual species richness in each cell (i). Its value ranges from 0 to 1, where 0 represents an unsampled cell and 1 represents a fully sampled cell.

The Species Richness Index (SRI) is used as a metric to assess the databases accuracy in depicting actual species richness. Consequently, we propose a systematic categorization of cells into three classes, *"low"*, *"medium"* and *"high"*, based on the frequency distribution of SRI, as illustrated in Fig. A.1. Cells with only one record are identified as having insufficient records (IR) for estimating S_{est}. Those with an SRI in the range (0, 0.60) are categorized as *low*, while those falling within the interval (0.60, 0.85) may be characterized as having a *medium* level of representativeness. Furthermore, cells within the range (0.85, 1.00) are identified as *high*, meaning an adequate representation of species diversity. Fig. A.2 show maps illustrating the raw values for observed species richness (S_{obs}), expected species richness (S_{exp}), and SRI values.

2.2.2. Temporal Representativeness Analysis

We constructed species accumulation curves, using years as the sampling unit, to examine the temporal distribution of data records within each bioregion. To assess the sample's adequacy, we focused on data from the last four years (2016-2020), representing the final 10% of each accumulation curve. We used a linear fit following the rescaling of the SRI to facilitate statistically comparable slope measurements. Slopes approaching zero suggest bioregions that have been adequately sampled, whereas slopes deviating from zero in-

dicate insufficient sampling efforts over time.

2.2.3. Gap Analysis

We overlaid the spatial representativeness map (§2.2.1) with Marine Protected Areas (MPAs) shapefiles (UNEP-WCMC and IUCN, 2022) and fishing exploitation areas reported by (FAO, 2014). The superposition of these layers allowed us to calculate the extent of protection offered by MPAs for each bioregions on a cell basis, and the extent of cells in designated fishing zones. Based on this exercise, the relationship among two opposing human impacts and current uncertainties about marine fish diversity can be assessed.

2.2.4. Bias Assessment

Potential biases resulting from body size, habitat depth, and cultural value of species (§2.1) were assessed using Fishbase repository information (Froese and Pauly, 2021). We developed a frequency distribution plot for each species' length reported in the database, employing equal 30 bin intervals. Habitat depth was determined according to the classification of oceanic layers used in Costello et al. 2010 (i.e. epipelagic = 0 - 200 m, mesopelagic = 200 - 1,000 m, and bathypelagic= 1,000 - 4,000 m). A pie chart is used to show how cultural values are represented in the database. All data and scripts are available (Appendix A).

3. Results

3.1. Records by Bioregions

Approximately 1.14% of the total reported occurrences of the order Actinopterygii were retained in our analysis. That is, from the 71,670,596 records downloaded from the GBIF and OBIS repositories, 820,004 were considered useful (see Appendix A). This subset consisted of 10,371 species in 361 families. The most represented families in our dataset are Scombridae, Pleuronectidae, and Gadidae with 103,762, 57,018, and 52,079 records, respectively. The

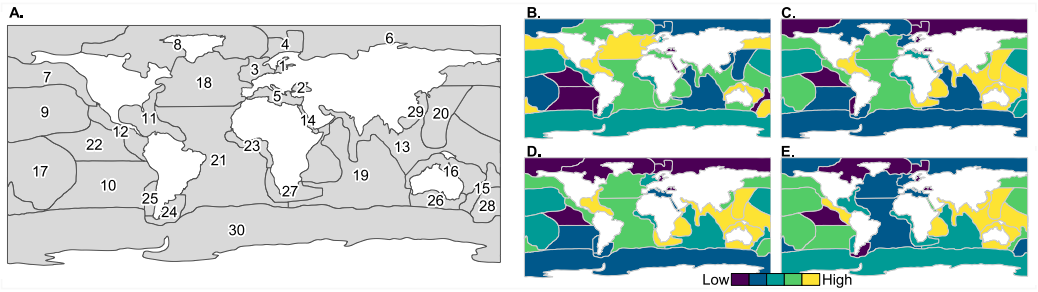


Fig. 1: Marine bioregions and spatial diversity distribution used in this study. **A**. The 30 marine bioregions from Costello et al. (2017) used in this study. Number are identification labels in Table 1. **B**. Records by bioregion; **C**. Overall species richness across bioregions; **D**. Family richness; and **E**. Shannon diversity index. Note that values in **C-E** have been standardized for illustration purposes. See Table 1 for actual values and a detailed map of observed and expected richness in Fig. A.2.

species with the largest representation frequency are *Hippoglossoides platessoides*, *Mola mola*, and *Coryphaena hippurus* with 30,885, 21,042 and 21,089 records, respectively.

The analysis at bioregions' level (Table 1) shows large variability. Record counts vary across three orders of magnitudes, i.e., from 2.68×10^5 in the Caribbean Sea and the Gulf of Mexico (11), down to 1.02×10^2 in the Black Sea (2). The bioregions with the largest species richness and diversity index are the Indo-Pacific Seas and the Indian Ocean (13), with 2.95×10^3 recorded species and a Shannon index of 6.93, followed by the Coral Sea bioregion (16), with 2.93×10^3 species and a Shannon index of 6.75. Likewise, the Coral Sea also presents the largest number of families. Notably the Southern Ocean (30) is the largest bioregion in square kilometers, it has the smallest number of records and the lowest number of species and families across all bioregions. The Black Sea (2) and the Norwegian Sea (4) bioregions have the lowest number of records and Shannon index value, respectively. Fig. 1 illustrates the location of the 30 marine bioregions and their respective richness and diversity values.

Table 1: Area (1,000 km^2) and records, species richness, family richness, and Shannon diversity counts for each bioregion. The highest value in each column is highlighted.

ID	Bioregion	Area	Records	Species	Families	Shannon
1	Inner Baltic Sea	415	8,902	72	30	2.46
2	Black Sea	537	102	37	22	3.21
3	NE Atlantic	2,053	87,377	310	104	3.90
4	Norwegian Sea	1,132	3,046	93	35	2.16
5	Mediterranean	2,859	12,532	372	101	3.39
6	Arctic Seas	10,276	2,506	114	23	3.90
7	North Pacific	12,974	78,070	839	156	4.50
8	North American Boreal	8,001	9,709	162	48	2.99
9	Mid-Tropical N Pacific Ocean	32,685	9,310	615	127	4.59
10	South-East Pacific	21,952	386	190	89	4.97
11	The Caribbean and the Gulf of Mexico	8,427	**268,066**	1,703	209	4.49
12	Gulf of California	6,184	7,639	885	148	5.93
13	Indo-Pacific Seas and Indian Ocean	37,090	16,967	**2,947**	215	**6.93**
14	Gulfs of Aqaba, Aden, Suez, Red Sea	830	926	352	72	5.51
15	Tasman Sea	3,592	1,003	380	120	5.36
16	Coral Sea	7,658	40,107	2,929	**249**	6.75
17	Mid South Tropical Pacific	23,418	6,083	811	123	5.18
18	Offshore and NW North Atlantic	16,012	130,994	897	190	3.46
19	Offshore Indian Ocean	31,076	1,263	337	116	4.06
20	Offshore W Pacific	10,291	6,363	1,839	232	6.81
21	Offshore S Atlantic	41,435	11,960	990	188	3.79
22	Offshore Mid-E Pacific	13,815	687	79	37	3.04
23	Gulf of Guinea	3,325	6,816	384	138	3.95
24	Argentina	2,665	8,701	115	52	2.83
25	Chile	1,739	250	100	54	4.36
26	Southern Australia	3,824	15,643	1,011	201	5.75
27	Southern Africa	4,371	19,954	1,142	210	4.16
28	New Zealand	6,293	53,879	558	154	3.66
29	North West Pacific	2,457	1,767	869	182	6.46
30	Southern Ocean	**62,161**	8,996	294	57	3.98

3.2. Geographic Analysis

Fig.2 shows cell classification according to SRI (§2.2.1). As expected, no bioregion is completely sampled at the ∼1° resolution. In fact, at this resolution, large empty regions with no records are observed. The bioregions with the largest area classified as *high* representativeness are the Northeast Atlantic (3) (37.53%), the Caribbean and the Gulf of Mexico (11) (29.26%), and the Inland Baltic Sea (1) (24.37%). It should be noted that such cells mostly correspond to coastal areas in the northern hemisphere. On the other hand, the bioregions that present the largest surface area without records are the Southeast Pacific (10) (96.3%), the Arctic Sea (6) (94.9%), and the Southern Ocean (30) (93.7%). While the bioregions with the largest surface area and *medium* representativeness of records are the Gulf of Guinea (23) (32%), the Norwegian Sea (4) (22.3%), and the Gulf of California (12) (21.6%). Additional results for ∼5° and ∼10° spatial resolution grids are shown in Appendix C.

3.3. Temporal Analysis

Bioregions show similar data accumulation trends across the four decades analyzed here (Fig. 3). While a significant increase is apparent in the time period between 2005 and 2010, such increase is not significant for 14 of the 30 bioregions. The Caribbean and the Gulf of Mexico (11) is the bioregion with the largest increase in data contribution to the dataset, while the Black Sea (2) shows the lowest data contribution rate in the 40 years span between 1980 and 2020. (See Appendix D for further analysis).

We classified the slopes of the final 10% of each accumulation curve in Fig. 4. Fourteen bioregions show a slope less than 1. The Mediterranean Sea (5) stands out with the lowest slope value (0.47), while the Black Sea (2) is the bioregion with the steepest final slope (3.13).

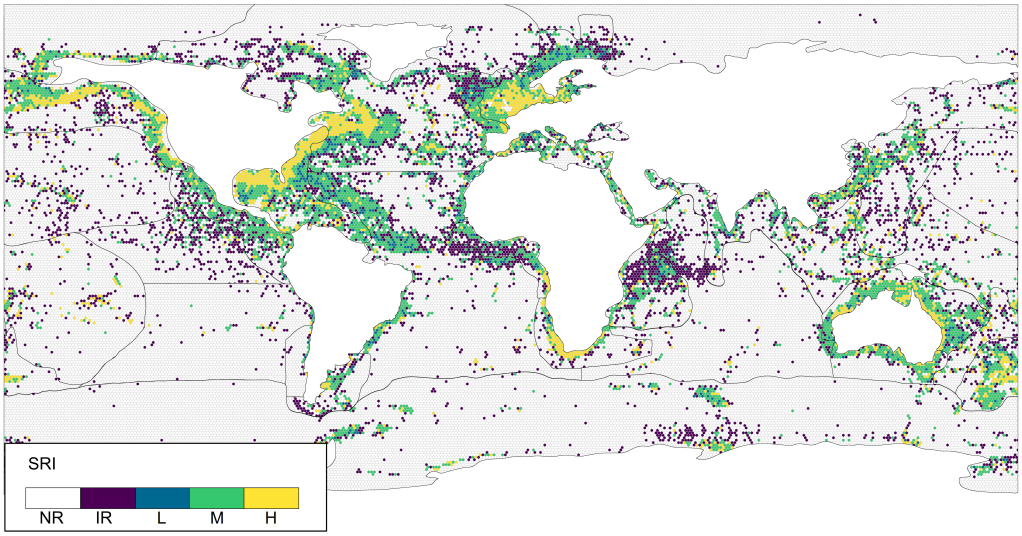


Fig. 2: Spatial Representativeness Index (SRI) in ∼1° hexagonal lattice. **IR** shows cells with insufficient records to evaluate S_{est}. **H** are cells with an *high* representativeness of species richness, i.e. SRI > 0.85. **M** are cells considered as having a *medium* representativeness, i.e. SRI $\in (0.60, 0.85)$. **L** cells are cells with *low* representativeness of species records and are thus not considered to be representative of actual species richness, i.e. SRI $\in (0, 0.6)$. **NR** are cells with no records (SRI $= NA$). Raw values for SRI, S_{obs} and S_{est} are shown in the appendix (Fig. A.2).

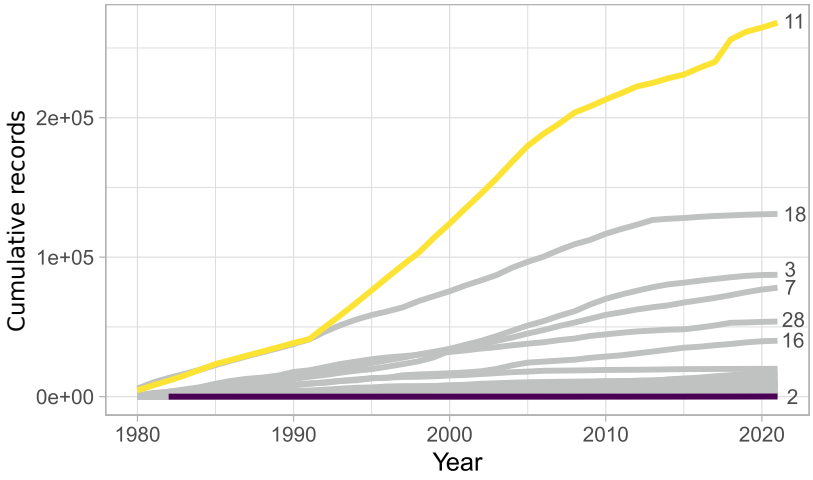


Fig. 3: Records of accumulation rate for each bioregion across the four decades analyzed. The yellow line is the accumulation of fish records in the Caribbean and the Gulf of Mexico bioregion (11), while the purple line is the accumulation rate in the Black Sea (2). The numbers at the end of each time series correspond to the bioregion ID in Table 1.

3.4. Gap Analysis and Fishing Exploitation Areas

The bioregions with the largest area covered by protected areas are the Coral Sea (16), the Northeast Atlantic (3) and New Zealand (28), covering 37.3%, 17.4%, and 16% of their respective surface areas. Regarding these bioregions' sampling level, the Offshore Indian Ocean (19), the Gulf of Aqaba, Aden, Suez, Red Sea (14) and Coral Sea (16), are the bioregions with the highest share of cells with *high* representativeness, hence well sampled within protected areas (83%, 63.8%, and 59.8% respectively). In turn, the Arctic Seas (6), the North American Boreal (8) and Mid-South Tropical Pacific are the bioregions with protected areas showing the highest share of cells without records (86.2%, 83.8%, and 81.2% respectively). (See Appendix E).

The FAO areas with the largest surface area classified as *high* representativeness correspond to the Northwest Atlantic (22.1 %), Northeastern part of the Pacific Ocean (14.6 %), and Western part of the Atlantic Ocean (12.6 %) (Table 3). These FAO areas correspond to regions in the Pacific Ocean (North Pacific, North West Pacific, Mid-Tropical N Pacific Ocean and Indo-

Table 2: Results of overlapping MPAs and SRI grid. ID is the identification number given to each bioregion (see Table 1 for bioregions' names). Area corresponds to the share of surface area covered by MPAs. **NR** is the share of cells with *No Records*; **IR** is the share of cells with *Insufficient Records*; **L** is the share of classified cells with *low* number of records; **M**, the share of classified cells with *medium* number of records, and **H**, the share of classified cells with *high* number of records. The highest values for each column are highlighted.

ID	**Area**	**NR**	**IR**	**L**	**M**	**H**
	km^2			%		
1	0.03	2.38	4.30	5.22	49.08	39.01
2	12.89	26.71	**27.61**	10.49	35.19	0.00
3	9.74	3.23	1.35	0.94	40.86	53.62
4	0.15	11.16	19.63	5.18	56.69	7.34
5	0.09	5.71	6.77	**11.20**	47.95	28.37
6	5.01	**86.16**	5.97	0.02	4.65	3.20
7	0.00	26.48	4.24	1.26	23.77	44.25
8	1.23	83.82	7.77	0.62	6.70	1.11
9	0.69	69.58	15.77	0.00	6.40	8.25
10	17.36	73.51	24.58	0.00	0.80	1.11
11	0.28	20.13	6.25	3.44	29.98	40.21
12	0.83	0.33	1.23	8.85	**61.35**	28.25
13	0.45	50.52	11.94	0.88	25.17	11.50
14	0.25	8.87	0.00	1.59	25.65	63.88
15	4.06	57.18	15.27	0.00	7.29	20.26
16	16.00	3.10	0.49	1.10	35.52	59.79
17	0.20	81.19	11.43	0.00	3.07	4.31
18	4.91	35.87	16.63	0.77	25.64	21.09
19	2.78	11.88	0.74	0.00	4.34	**83.04**
20	2.56	34.06	9.68	6.85	35.31	14.10
21	3.79	51.06	19.35	0.38	21.35	7.86
22	0.16	41.98	27.54	0.00	23.22	7.26
23	13.83	9.92	4.92	7.98	61.33	15.85
24	0.21	40.86	20.64	0.24	23.97	14.29
25	0.07	65.27	0.50	0.05	1.29	32.89
26	0.28	30.04	12.83	6.89	38.62	11.63
27	1.45	16.78	5.75	0.15	18.71	58.62
28	0.00	45.36	2.25	0.00	13.71	38.67
29	**37.29**	20.49	17.05	2.68	42.50	17.29
30	1.66	64.05	7.12	4.89	19.86	4.08

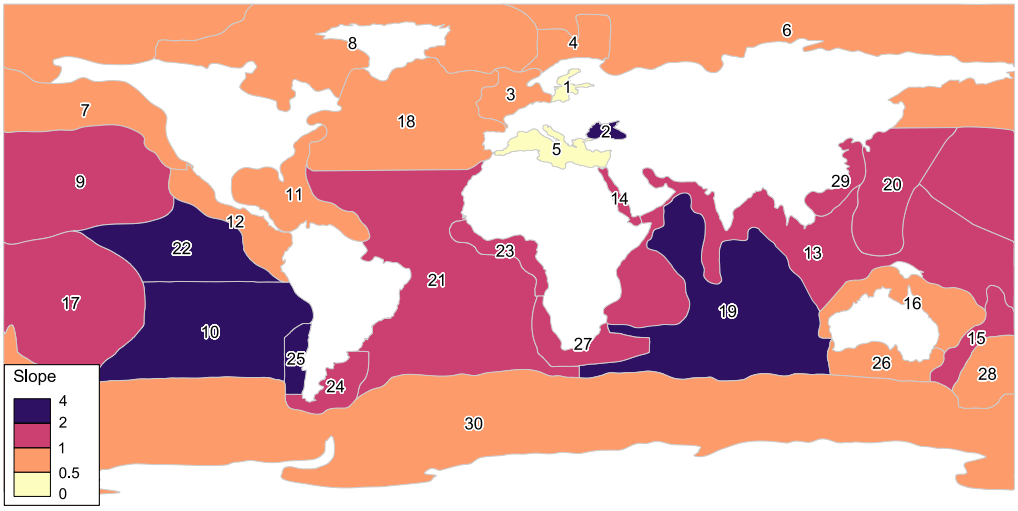


Fig. 4: Illustration of the slope values of the species accumulation curve for each bioregion. The slope corresponds to the final 10% of the species accumulation curve. See §2.2.2 for details regarding the analysis.

Pacific Seas and Indian Ocean, as well as the Gulf of California and the Caribbean and the Gulf of Mexico). Largest FAO areas with *NR* cells correspond to the Antarctic part of the Pacific Ocean, the Antarctic part of the Atlantic Ocean and the Southeastern part of the Atlantic Ocean in the Southern Ocean, Offshore S Atlantic, and Southern Africa.

3.5. Evaluation of Biases

We evaluated biases for body size, habitat depth, and cultural value for 10,371 marine fish species identified in our database (§3.1).

3.5.1. Body Size

The 10-40 cm range is the most frequently occurring size length, corresponding to the interval between the 1^{st} and 3^{rd} quartile (Fig. 5A). Three species stand out with the highest numbers of records, *Scomber scombrus*, *Lagodon rhomboides* and *Mallotus villosus* with 20,995, 19,563 and 13,609 records respectively. These species are distributed mainly in the Northeast

Table 3: Results from overlapping FAO fishery areas and SRI grid. The surface area corresponding to each bioregion, and the share of surface area of each classification. Areas are in thousands square km; **NR** is the share of cells with *No Records*; **IR** is the share of cells with *Insufficient Records*; **L** is the share of classified cells with *low* number of records; **M** is the share of classified cells with *medium* number of records, and **H** is the share of cells with a *high* number of records. The largest values for each column are highlighted.

FAO Area Name	**Area** km^2	**NR**	**IR**	**L** %	**M**	**H**
Arctic Sea	4,086	93.22	3.13	0.29	2.61	0.75
Northwestern part of the Atlantic Ocean	874	31.19	11.66	5.69	29.37	**22.08**
Northeastern part of the Atlantic Ocean	3,223	66.29	12.54	2.55	13.63	4.99
Western part of the Atlantic Ocean	1,285	30.84	13.09	**7.91**	**35.60**	12.55
Eastern Central part of the Atlantic Ocean	1,208	52.61	**24.09**	3.44	18.37	1.19
Mediterranean Sea and the Black Sea	309	46.39	15.43	5.24	24.77	8.17
Southwestern part of the Atlantic Ocean	1,731	82.49	5.85	1.69	8.55	1.42
Southeastern part of the Atlantic Ocean	1,765	89.92	4.19	0.15	2.13	3.61
Antarctic part of the Atlantic Ocean	2,310	93.31	2.80	0.20	2.93	0.76
Western part of the Indian Ocean	2,621	72.45	16.11	1.03	8.51	1.89
Eastern part of the Indian Ocean	3,029	85.40	4.69	0.82	7.39	1.70
Antarctic and South of the Indian Ocean	1,977,29	85.71	7.76	0.56	4.33	1.64
Northwestern part of the Pacific Ocean	2,259	73.55	12.40	0.94	10.32	2.79
Northeastern part of the Pacific Ocean	968	55.13	12.65	1.34	16.26	14.62
Western Central part of the Pacific Ocean	2,963	70.45	12.56	0.43	11.58	4.98
Eastern Central part of the Pacific Ocean	**4,141**	79.36	11.30	0.31	6.94	2.09
Southwestern part of the Pacific Ocean	3,097	85.04	4.42	0.97	6.40	3.17
Southeastern part of the Pacific Ocean	2,997	91.16	6.00	0.10	2.30	0.44
Antarctic part of the Pacific Ocean	2,361	**93.47**	4.57	0.21	1.42	0.33

Atlantic (3) and Offshore and Northwest North Atlantic (18) bioregions. In turn, the families that accumulate the greatest number of records are Sparidae , Scombridae and Labridae, with 24,837, 21,719, and 21,035 records, respectively. These families are mainly distributed in the Caribbean Sea and the Gulf of Mexico (11), and in the Northeast Atlantic (3).

3.5.2. Habitat Depth

The depth range most commonly observed among records is about 50 meters and decreases as depth increases, particularly from the epipelagic to the mesopelagic zones, as illustrated in Fig.5B. Among the species with the highest number of recorded occurrences, *Mola mola*, *Coryphaena hippurus*, and *L. rhomboides* stand out, with 21,089, 21,042, and 19,563 occurrences in

the databases, respectively. These species are distributed mainly around the Caribbean Sea and the Gulf of Mexico (11) bioregions, as well as the following bioregions: Offshore and NW North Atlantic (18) and the South Atlantic Coast (21). The families that accumulate a larger number of records correspond to Scombridae, Gadidae, Sparidae with 63,572, 38,876, and 30,041 records, respectively. These are mostly distributed in the northern hemisphere; that is, the Caribbean and the Gulf of Mexico (11), Offshore and NW of the North Atlantic (18), and part of the South Atlantic Ocean Coast (21) bioregions.

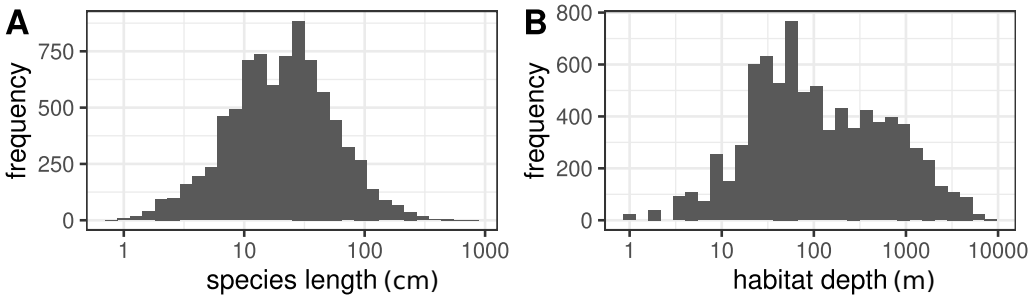


Fig. 5: Distribution of marine fish records in GBIF and OBIS classified by body length and habitat depth. **A**. Relationship between record number and species length ($\log_{10}$); and **B**. Relationship between record number and habitat depth ($\log_{10}$) .

3.5.3. Cultural Value

Finally, when analyzing the most frequent cultural value represented across our dataset (Fig. 6), "Commercial" use of the species emerges as the most important with 73.4% among records, followed by the category "No interest" (5.03%), and "Subsistence fishing" (3.08%).

4. Discussion

Our work provides a methodological framework based on a set of non-parametric estimators to quantify the potential number of species from incidence data (Chao et al., 2009). We used hexagons due to their suitability

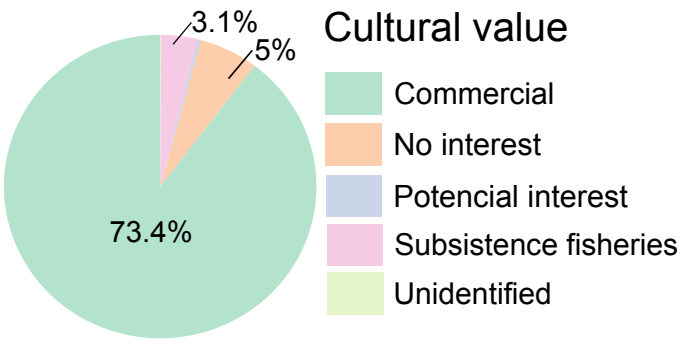


Fig. 6: Frequency of marine fish representation in GBIF and OBIS repositories according to cultural value.

as a tessellation that conforms more effectively to the shape of a spheroid, compared to square grid cells. We also placed special emphasis on cleaning the occurrence data in their taxonomy (Jin and Yang, 2020), and any potential input errors associated with large and massive datasets (Zizka et al., 2020). Hence, we only focused on evaluating marine species in the order Actinopterygii (Alò et al., 2021).

Publicly accessible occurrence records are growing rapidly, partly due to significant progresses in ecoinformatics (Lenoir et al., 2020; Oliver et al., 2021). These databases harbor a growing variety of sources, including museum specimens, field observations, acoustic and visual sensors, and citizen science efforts (Amano et al., 2016). However, despite the incredible accumulation of biodiversity records, not all the data is really useful, nor does it represent new insights into the distribution of species (Bayraktarov et al., 2019; Zizka et al., 2020). That is why a systematic evaluation of the integrity and coverage of this information is required (Troia and McManamay, 2017).

There is an extensive bibliography that evaluates the record quality available for different taxonomic groups. Some examples are: legumes on a global scale (Yesson et al., 2007), lepidoptera from Great Britain, and woody plants in Panama (Chao et al., 2009), global marine biodiversity (Tittensor et al., 2010), vascular plants in China (Yang et al., 2013), marine fish on a global

scale (Mora et al., 2008; García-Roselló et al., 2015), freshwater fish in the USA (Troia and McManamay, 2017; Pelayo-Villamil et al., 2018), and terrestrial mammals on a global scale (Oliver et al., 2021), among many others. Assuming that that not all data available in these repositories is useful for biodiversity analyses, several efforts have proposed parametric and non-parametric estimators for data cleaning and species richness analysis, including ModestR (García-Roselló et al., 2013), KnowBr (Lobo et al., 2018), and RWizard (Guisande and Lobo, 2019).

Striving for simplicity, we employ the ratio of observed to expected species richness (SRI) as a means to indicate the spatial distribution of undersampled regions. While acknowledging the potential for misrepresentation, particularly in cases of extremely low observed richness, we mitigate this concern by confining our analysis to locations with more than one observed species record. This approach offers a straightforward method for identifying areas that warrant additional sampling.

We evaluated two additional grid sizes (i.e. 2.5×10^4 and 10^7 km^2), and like other studies, our results show that the coarser the resolution used, the greater the overestimation is in terms of area. That is, the richness index will indicate that a large area is, indeed, well sampled when in reality, occurrence records could in fact be localized in a very small area. On the contrary, the finer the scale of analysis, the more localized and deficient the sampling is (Tittensor et al., 2010; García-Roselló et al., 2015; Meyer et al., 2015; Troia and McManamay, 2016, 2017).

Considering that more than 40 years of data were analyzed, our results demonstrated that on a global scale the primary marine fish data available on the GBIF and OBIS platforms are still far from being representative and complete. Compared with other studies evaluating the same taxonomic group (Mora et al., 2008; García-Roselló et al., 2015), although we obtained similar macroecological patterns, only 1.14% of the records extracted from both repositories were useful for our analyses. A large share of occurrences

presented input errors or lacked the data required to develop reliable analyses (Yesson et al., 2007; García-Roselló et al., 2014).

We also found evidence of strong information biases in the records explored. On the one hand, when analyzing the families and species with the greatest representation, they match groups of commercial interest fish, pointing to the existence of data taxonomic biases (Melo-Merino et al., 2020). This is the case of the families Scombridae, Pleuronectidae and Gadidae, which include nutritionally-relevant species such as tuna, cod, haddock, among others (Cohen et al., 1990). The same is true for the species with the largest number of records, *H. platessoides* (Pleuronectidae), *C. hippurus*(Coryphaenidae), and *M. mola* (Molidae); while the first two are species exploited by the fishing industry, sunfish (*M. mola*) has a wide distribution and is mostly associated with scientific and recreational interests (Pope et al., 2010).

The unequal contribution of data at the spatial level is another factor that must be considered when dealing with data available in ecoinformatic platforms. We show a clear **geoghapic bias** in the sampling of certain regions and/or ecosystems. The literature indicates that the largest data contributions come from developed countries (Yesson et al., 2007; Chandler et al., 2017), and coastal regions with high road connectivity (Chandler et al., 2017; Melo-Merino et al., 2020). This is also particularly prevalent in undersampled marine habitats, such as the deep sea (Webb et al., 2010). Our results match what has been described in the literature, regardless of the grid size used for the analysis. The bioregions that include the Northeast Atlantic (3), the Caribbean and the Gulf of Mexico (11), and the Inland Baltic Sea (1) are regions classified with a high representativeness. However, the number of cells with insufficient records to generate an unbiased diversity analysis is also of concern. For example, our results show that these cells are distributed in more internal areas of the bioregions, zones where sampling is likely to be more difficult.

While, on the other hand, the bioregions that include the South and

Southeast Pacific (including the southern coast of South America), the Southern Ocean, and the Arctic Sea are the regions where the share of cells without records (NR) exceeds 90%. The lack of data samples over this extensive area renders any endeavor to depict species richness and distribution highly unreliable (as noted by Yang et al., 2013; Troia and McManamay, 2017). These marine regions encompassing both the water column and the seabed beyond national jurisdictions make up nearly half of the Earth's surface and sustain substantial abundance and diversity of life, as highlighted by (Visalli et al., 2020). Nonetheless, when scrutinizing the occurrence data for marine ichthyofauna, these regions remain the least sampled areas.

Finally, the datas's **time bias** is also present in our study. Differences in species identification and sampling methodologies over the decades have resulted in databases of variable quality. However, the current era is characterized by more accurate data thanks to improvements in individual capture and identification tools (Costello et al., 2015; Jin and Yang, 2020). For these reasons, our approach considers occurrence records since 1980; the coverage of occurrence data, however, is uneven over time when comparing marine bioregions. Despite assesing four decades of data, sampling efforts are still insufficient in 46% of marine bioregions. Not surprisingly, the Caribbean and the Gulf of Mexico (11) is the bioregion with the largest data input, once again showing that the geographic sampling bias has strong impacts on spatial predictions of species richness (Yang et al., 2013). Future sampling efforts should focus on bioregions at low or equatorial latitudes, areas where marine biodiversity is concentrated according to biogeographic studies (Costello et al., 2017).

All the biases that we have described, added to typical data capture issues, promote and deepen several information gaps that thwart the effective spatio-temporal biodiversity quantification (Magurran and McGill, 2011). In this study, we have overlapped our species richness estimates with the global MPAs declared up until the beginning of 2022 (UNEP-WCMC and IUCN,

2022), and the fishing exploitation areas reported by FAO (FAO, 2014). This exercise demonstrates the importance of public databases that can faithfully reflect the taxonomic and biogeographical knowledge available for each region (Pelayo-Villamil et al., 2018). According to our results, the North West Pacific bioregion (19) has the largest area covered by MPAs. However, its share of cells with *high* representativeness is low compared to other bioregions. This result is of certain concern as this bioregion is considered a conservation hotspot among other bioregions, such as the Coral Sea (16), a region with a relatively large share of highly sampled cells (Ramírez et al., 2017). However, we found a low share of well-sampled cells in both regions, pointing to the existence of important information gaps, at least for fish of the order Actinopterygii. We emphasize the need to correct these information gaps so that conservation efforts can rely on dependable data, including the design and implementation of new MPAs (Sala et al., 2021).

Along these lines, by overlapping the bioregions with fishing exploitation zones, we determined that the North Pacific (7), the North West Pacific (29), the Mid-Tropical N Pacific Ocean (9), and the Indo-Pacific Seas and Indian Ocean (13) bioregions, as well the Gulf of California (21), and the Caribbean and the Gulf of Mexico (11), are the regions with the highest data representation and where fishing activity is concentrated. According to (Kroodsma et al., 2018), the area corresponding to the central Atlantic and Northeast Pacific present little intense fishing efforts, while the regions associated with the Northeast Atlantic, the Northeast Atlantic (Europe) regions, and the Northwest Pacific are known to have huge fishing development, where fishing efforts are concentrated worldwide. The Southeastern Atlantic Ocean (FAO area 47 and 88), part of the Pacific Ocean (FAO area 88) and Antarctica (FAO area 48 and 88) are the regions with the highest share of cells without records (NR= >93 %). When compared with the findings by Kroodsma et al. 2018, these areas match the "holes" without fishing effort data, which is explained by the geographical remoteness and the lack of technological

development required for fisheries to extend to new domains (Visalli et al., 2020). This issue restricts both the extraction of marine resources as well as data collection.

The research questions addressed in this study were essential to understand the prevailing data collection trends and to lay the groundwork for potential corrective measures than can mitigate the described biases. Our initial inquiry regarding fish body size does not imply a straightforward association between larger records and larger body lengths. Instead, we observe a distinct hump-shaped distribution in frequency distributions, akin to well-documented macroecological patterns observed in various taxa (Smith et al., 2014; Allen et al., 2006). It is worth noting that mid-sized fish species account for the highest number of records. Among these, species such as *S. scombrus* (Scombridae), *L. rhomboides* (Sparidae), and *M. villosus* (Osmeridae) stand out for their numerous records; they are predominantly distributed in well-sampled regions such as the Mediterranean Sea (5), the Caribbean and the Gulf of Mexico (11), and the Atlantic Ocean (e.g. bioregion 3). Furthermore, the inverse relationship between fish size and abundance, and consequently, the frequency of human use, whether for scientific research or commercial purposes, is a well-established concept (Pauly and Palomares, 2005).

This variation in sampling efforts results in a noticeable overrepresentation of these species, exacerbating the existing **taxonomic bias**. Conversely, the correlation between the number of records and habitat depth indicates that the pelagic zone shows a significant data concentration, which appears to align with areas more readily accessible for data collection (Melo-Merino et al., 2020). It has been pointed out that species concentration decreases as the ocean increases its depth; however, it is precisely these areas that have been the least sampled and where there is a larger chance of discovering new species (Costello et al., 2017). This demonstrates the need to concentrate efforts on the deeper regions of the water column (mesopelagic, bathyal, and abyssal) for a more equitable representation of marine ecosystems. Fi-

nally, a straightforward examination of cultural value within marine records unmistakably reveals that marine fish species with more favorable or economic advantages for humans tend to have stronger representations within the databases discussed. This observation is likely connected to the significant role of the fishing industry as one of the primary sources of information contributing to platforms such as OBIS, as previously discussed (Zhang and Grassle, 2002).

Today, marine ecosystems and their biodiversity face the major climate change challenge as well as the impacts of human activity, especially on species considered key food resources for survival (Hollowed et al., 2013; Ramírez et al., 2017; O'Hara et al., 2021). It is important to focus on and further the study of areas with few or no records, since describing the species geographic ranges and their temporal dynamics is a key measure for the evaluation of the actual biodiversity state (Lenoir et al., 2020; Oliver et al., 2021). Counting on more reliable data will allow for the implementation of effective conservation actions.

Acknlowledgements

Funding for this research was provided by Chile's National Research and Development Agency (ANID) through project FONDECYT Regular #11211490 to HS and a doctoral fellowship to AGC (ANID #2022-21220124). We thank professor Ricardo Giesecke for his valuable comments on an earlier version of this manuscript.

References

Allen, C.R., Garmestani, A.S., Havlicek, T.D., Marquet, P.A., Peterson, G.D., Restrepo, C., Stow, C.A., Weeks, B.E., 2006. Patterns in body mass distributions: sifting among alternative hypotheses. Ecology Letters 9, 630–643. doi:10.1111/j.1461-0248.2006.00902.x.

Alò, D., Lacy, S.N., Castillo, A., Samaniego, H.A., Marquet, P.A., 2021. The macroecology of fish migration. Global Ecology and Biogeography 30, 99–116. doi:10.1111/geb.13199.

Amano, T., Lamming, J.D., Sutherland, W.J., 2016. Spatial gaps in global biodiversity information and the role of citizen science. Bioscience 66, 393–400. doi:10.1093/biosci/biw022.

Appeltans, W., Ahyong, S., Anderson, G., Angel, M., Artois, T., Bailly, N., Bamber, R., Barber, A., Bartsch, I., Berta, A., Błażewicz-Paszkowycz, M., Bock, P., Boxshall, G., Boyko, C., Brandão, S., Bray, R., Bruce, N., Cairns, S., Chan, T.Y., Cheng, L., Collins, A., Cribb, T., Curini-Galletti, M., Dahdouh-Guebas, F., Davie, P., Dawson, M., De Clerck, O., Decock, W., De Grave, S., de Voogd, N., Domning, D., Emig, C., Erséus, C., Eschmeyer, W., Fauchald, K., Fautin, D., Feist, S., Fransen, C., Furuya, H., Garcia-Alvarez, O., Gerken, S., Gibson, D., Gittenberger, A., Gofas, S., Gómez-Daglio, L., Gordon, D., Guiry, M., Hernandez, F., Hoeksema, B., Hopcroft, R., Jaume, D., Kirk, P., Koedam, N., Koenemann, S., Kolb, J., Kristensen, R., Kroh, A., Lambert, G., Lazarus, D., Lemaitre, R., Longshaw, M., Lowry, J., Macpherson, E., Madin, L., Mah, C., Mapstone, G., McLaughlin, P., Mees, J., Meland, K., Messing, C., Mills, C., Molodtsova, T., Mooi, R., Neuhaus, B., Ng, P., Nielsen, C., Norenburg, J., Opresko, D., Osawa, M., Paulay, G., Perrin, W., Pilger, J., Poore, G., Pugh, P., Read, G., Reimer, J., Rius, M., Rocha, R., Saiz-Salinas, J., Scarabino, V., Schierwater, B., Schmidt-Rhaesa, A., Schnabel, K., Schotte, M., Schuchert, P., Schwabe, E., Segers, H., Self-Sullivan, C., Shenkar, N., Siegel, V., Sterrer, W., Stöhr, S., Swalla, B., Tasker, M., Thuesen, E., Timm, T., Todaro, M., Turon, X., Tyler, S., Uetz, P., van der Land, J., Vanhoorne, B., van Ofwegen, L., van Soest, R., Vanaverbeke, J., Walker-Smith, G., Walter, T., Warren, A., Williams, G., Wilson, S., Costello, M., 2012. The magni-

tude of global marine species diversity. Current Biology 22, 2189–2202. doi:10.1016/j.cub.2012.09.036.

Bayraktarov, E., Ehmke, G., O'Connor, J., Burns, E.L., Nguyen, H.A., McRae, L., Possingham, H.P., Lindenmayer, D.B., 2019. Do big unstructured biodiversity data mean more knowledge? Frontiers in Ecology and Evolution , 239. doi:10.3389/fevo.2018.00239.

Beck, J., Böller, M., Erhardt, A., Schwanghart, W., 2014. Spatial bias in the gbif database and its effect on modeling species' geographic distributions. Ecological Informatics 19, 10–15. doi:10.1016/j.ecoinf.2013.11.002.

Boettiger, C., Lang, D.T., Wainwright, P., 2012. Rfishbase: exploring, manipulating and visualizing fishbase data from r. Journal of Fish Biology 81, 2030–2039. doi:10.1111/j.1095-8649.2012.03464.x.

Bonnet-Lebrunm A.S., Sweetlove, A., Griffiths, H.J., Sumner, M., Provoost, P., Raymond, B., Ropert-Coudert, Y., Van de Putte, A.P., 2023. Opportunities and limitations of large open biodiversity occurrence databases in the context of a Marine Ecosystem Assessment of the Southern Ocean. Frontiers in Marine Science 10, 1–13. doi:10.3389/fmars.2023.1150603.

Butchart, S.H.M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J.P.W., Almond, R.E.A., Baillie, J.E.M., Bomhard, B., Brown, C., Bruno, J., Carpenter, K.E., Carr, G.M., Chanson, J., Chenery, A.M., Csirke, J., Davidson, N.C., Dentener, F., Foster, M., Galli, A., Galloway, J.N., Genovesi, P., Gregory, R.D., Hockings, M., Kapos, V., Lamarque, J.F., Leverington, F., Loh, J., McGeoch, M.A., McRae, L., Minasyan, A., Morcillo, M.H., Oldfield, T.E.E., Pauly, D., Quader, S., Revenga, C., Sauer, J.R., Skolnik, B., Spear, D., Stanwell-Smith, D., Stuart, S.N., Symes, A., Tierney, M., Tyrrell, T.D., Vié, J.C., Watson, R., 2010. Global biodiversity: Indicators of recent declines. Science 328, 1164–1168. doi:10.1126/science.1187512.

Chamberlain, S., 2017. rgbif: Interface to the global "biodiversity" information facility "api". r package version 0.9.8. URL: https://CRAN.R-project.org/package=rgbif.

Chandler, M., See, L., Copas, K., Bonde, A.M., López, B.C., Danielsen, F., Legind, J.K., Masinde, S., Miller-Rushing, A.J., Newman, G., et al., 2017. Contribution of citizen science towards international biodiversity monitoring. Biological conservation 213, 280–294. doi:10.1016/j.biocon.2016.09.004.

Chao, A., Colwell, R.K., Lin, C.W., Gotelli, N.J., 2009. Sufficient sampling for asymptotic minimum species richness estimators. Ecology 90, 1125–1133. doi:10.1890/07-2147.1.

Cheung, W.W., Lam, V.W., Sarmiento, J.L., Kearney, K., Watson, R., Pauly, D., 2009. Projecting global marine biodiversity impacts under climate change scenarios. Fish and fisheries 10, 235–251. doi:10.1111/j.1467-2979.2008.00315.x.

Chollett, I., Robertson, D.R., 2020. Comparing biodiversity databases: Greater Caribbean reef fishes as a case study. Fish and Fisheries 21, 1195–1212. doi:10.1111/faf.12497.

Cohen, D.M., Inada, T., Iwamoto, T., Scialabba, N., 1990. Gadiform fishes of the world. FAO Fisheries Synopsis 10, I.

Costello, M.J., Tsai, P., Wong, P.S., Cheung, A.K.L., Basher, Z., Chaudhary, C., 2017. Marine biogeographic realms and species endemicity. Nature Communications 8, 1057. doi:10.1038/s41467-017-01121-2.

Costello, M.J., Cheung, A., De Hauwere, N., 2010. Surface area and the seabed area, volume, depth, slope, and topographic variation for the world's seas, oceans, and countries. Environmental Science & Technology 44, 8821–8828. doi:10.1021/es1012752.

Costello, M.J., Vanhoorne, B., Appeltans, W., 2015. Conservation of biodiversity through taxonomy, data publication, and collaborative infrastructures. Conservation Biology 29, 1094–1099. doi:10.1111/cobi.12496.

Daly, A.J., Baetens, J.M., De Baets, B., 2018. Ecological diversity: Measuring the unmeasurable. Mathematics 6, 119. doi:10.3390/math6070119.

FAO, 2014. Fao statistical areas for fishery purposes. fao fisheries and aquaculture department [online] URL: http://www.fao.org/fishery/area/search/en.

Froese, R., Pauly, D., 2000. FishBase 2000: concepts designs and data sources. volume 1594. The WorldFish Center]. URL: http://hdl.handle.net/20.500.12348/2428.

Froese, R., Pauly, D.E., 2021. Fishbase. URL: https://www.fishbase.org.

García-Roselló, E., Guisande, C., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., Manjarrás-Hernández, A., Vaamonde, A., Granado-Lorencio, C., 2013. Modestr: a software tool for managing and analyzing species distribution map databases. Ecography 36, 1202–1207. doi:10.1111/j.1600-0587.2013.00374.x.

García-Roselló, E., Guisande, C., Heine, J., Pelayo-Villamil, P., Manjarrés-Hernández, A., González Vilas, L., González-Dacosta, J., Vaamonde, A., Granado-Lorencio, C., 2014. Using modestr to download, import and clean species distribution records. Methods in ecology and evolution 5, 708–713. doi:10.1111/2041-210X.12209.

García-Roselló, E., Guisande, C., Manjarrés-Hernández, A., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., González-Vilas, L., Vari, R.P., Vaamonde, A., Granado-Lorencio, C., et al., 2015. Can we derive macroecological patterns from primary global biodiversity informa-

tion facility data? Global Ecology and Biogeography 24, 335–347. doi:10.1111/geb.12260.

GBIF: The Global Biodiversity Information Facility , 2021. What is gbif? URL: https://www.gbif.org/what-is-gbif.

GBIF.org, 2021. Occurrence download. URL: https://www.gbif.org/occurrence/download/0039590-210914110416597, doi:10.15468/DL.V2PFS3. last accessed 29 October 2021.

Giraud, T., Lambert, N., 2016. cartography: Create and integrate maps in your r workflow. Journal of Open Source Software 1, 54. doi:10.21105/joss.00054.

Guisande, C., Lobo, J., 2019. Discriminating well surveyed spatial units from exhaustive biodiversity databases. r package version. 2.0. URL: http://cran.r-project.org/web/packages/KnowBR.

Heberling, J.M., Miller, J.T., Noesgaard, D., Weingart, S.B., Schigel, D., 2021. Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences 118, e2018093118. doi:10.1073/pnas.2018093118.

Hollowed, A.B., Barange, M., Beamish, R.J., Brander, K., Cochrane, K., Drinkwater, K., Foreman, M.G., Hare, J.A., Holt, J., Ito, S.i., et al., 2013. Projected impacts of climate change on marine fish and fisheries. ICES Journal of Marine Science 70, 1023–1037. doi:10.1093/icesjms/fst081.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M., Baselga, A., 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117, 847–858. doi:10.1111/j.0030-1299.2008.16434.x.

Hutchings, J.A., Baum, J.K., 2005. Measuring marine fish biodiversity: temporal changes in abundance, life history and demography. Philosophical

Transactions of the Royal Society B: Biological Sciences 360, 315–338. doi:10.1098/rstb.2004.1586.

Jin, J., Yang, J., 2020. Bdcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. Global Ecology and Conservation 21, e00852. doi:10.1016/j.gecco.2019.e00852.

Kroodsma, D.A., Mayorga, J., Hochberg, T., Miller, N.A., Boerder, K., Ferretti, F., Wilson, A., Bergman, B., White, T.D., Block, B.A., et al., 2018. Tracking the global footprint of fisheries. Science 359, 904–908. doi:10.1126/science.aao5646.

Lenoir, J., Bertrand, R., Comte, L., Bourgeaud, L., Hattab, T., Murienne, J., Grenouillet, G., 2020. Species better track climate warming in the oceans than on land. Nature Ecology & Evolution 4, 1044–1059. doi:10.1038/s41559-020-1198-2.

Lobo, J.M., Hortal, J., Yela, J.L., Millán, A., Sánchez-Fernández, D., García-Roselló, E., González-Dacosta, J., Heine, J., González-Vilas, L., Guisande, C., 2018. Knowbr: An application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. Ecological Indicators 91, 241–248. doi:10.1016/j.ecolind.2018.03.077.

Luypaert, T., Hagan, J.G., McCarthy, M.L., Poti, M., 2020. Status of marine biodiversity in the anthropocene, in: YOUMARES 9-The Oceans: Our research, our future. Springer, Cham, pp. 57–82. doi:10.1007/978-3-030-20389-4_4.

Magurran, A.E., McGill, B.J., 2011. Biological diversity: frontiers in measurement and assessment. Oxford University Press. doi:10.1086/666756.

Malhi, Y., Franklin, J., Seddon, N., Solan, M., Turner, M.G., Field, C.B.,

Knowlton, N., 2020. Climate change and ecosystems: Threats, opportunities and solutions. doi:10.1098/rstb.2019.0104.

Marquet, P.A., Fernández, M., Navarrete, S.A., Valdovinos, C., 2004. Diversity emerging: towards a deconstruction of biodiversity patterns, in: Lomolino, M., Heaney, L. (Eds.), Frontiers of Biogeography: New directions in the Geography of Nature. Cambridge University Press, pp. 191–209.

Melo-Merino, S.M., Reyes-Bonilla, H., Lira-Noriega, A., 2020. Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. Ecological Modelling 415, 108837. doi:10.1016/j.ecolmodel.2019.108837.

Meyer, C., Kreft, H., Guralnick, R., Jetz, W., 2015. Global priorities for an effective information basis of biodiversity distributions. Nature communications 6, 1–8. doi:10.1038/ncomms9221.

Mora, C., Tittensor, D.P., Myers, R.A., 2008. The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. Proceedings of the Royal Society B: Biological Sciences 275, 149–155. doi:10.1098/rspb.2007.1315.

Moudrý, V., Devillers, R., 2020. Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. Ecological Informatics 56, 101051. doi:10.1016/j.ecoinf.2020.101051.

Moreno, C.E., Rodríguez, P., 2011. Do we have a consistent terminology for species diversity? back to basics and toward a unifying framework. Oecologia 167, 889–892. doi:10.1007/s00442-011-2125-7.

Neigel, J., 1997. Marine Biodiversity: Patterns and Processes. Cambridge, Cambridge University Press. chapter Population genetics and demography of marine species. URL: http://www.cambridge.org/9780521552226.

OBIS: Ocean Biodiversity Information System, 2021. About obis URL: https://obis.org/.

OBIS.org, 2021. Occurrence download. URL: https://datasets.obis.org/downloads/9fd73b2a-cf6f-4ef9-a0e3-2d1f653520d3.zip. last accessed 29 October 2021.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., , Wagner, H., 2020. The vegan package URL: https://github.com/vegandevs/vegan.

Oliver, R.Y., Meyer, C., Ranipeta, A., Winner, K., Jetz, W., 2021. Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. PLoS Biology 19, e3001336. doi:10.1371/journal.pbio.3001336.

O'Hara, C.C., Frazier, M., Halpern, B.S., 2021. At-risk marine biodiversity faces extensive, expanding, and intensifying human impacts. Science 372, 84–87. doi:10.1126/science.abe6731.

Pauly, D., Palomares, M.L., 2005. Fishing down marine food web: it is far more pervasive than we thought. Bulletin of marine science 76, 197–212.

Pebesma, E.J., 2018. Simple features for r: standardized support for spatial vector data. R J. 10, 439. doi:10.32614/RJ-2018-009.

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. doi:10.1007/s10531-018-1630-y.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013. Essential biodiversity variables. Science 339, 277–278. doi:10.1126/science.1229931.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological applications 19, 181–197. doi:10.1890/07-2153.1.

Pope, E.C., Hays, G.C., Thys, T.M., Doyle, T.K., Sims, D.W., Queiroz, N., Hobson, V.J., Kubicek, L., Houghton, J.D., 2010. The biology and ecology of the ocean sunfish mola mola: a review of current knowledge and future research perspectives. Reviews in Fish Biology and Fisheries 20, 471–487. doi:10.1007/s11160-009-9155-9.

Provoost, P., Bosch, S., 2020. robis: R client to access data from the obis api. ocean biogeographic information system, intergovernmental oceanographic commission of unesco URL: https://cran.r-project.org/package=robis.

R Core Team, 2018. R: A language and environment for statistical computing. vienna, austria: R foundation for statistical computing URL: https://www.r-project.org/.

Ramírez, F., Afán, I., Davis, L.S., Chiaradia, A., 2017. Climate impacts on global hot spots of marine biodiversity. Science Advances 3, e1601198. doi:10.1126/sciadv.1601198.

Sala, E., Mayorga, J., Bradley, D., Cabral, R.B., Atwood, T.B., Auber, A., Cheung, W., Costello, C., Ferretti, F., Friedlander, A.M., et al., 2021. Protecting the global ocean for biodiversity, food and climate. Nature 592, 397–402. doi:10.1038/s41586-021-03371-z.

Smith, F.A., Gittlemann, J.L., Brown, J.H., 2014. Foundations of macroecology: classic papers with commentaries. University of Chicago Press.

Telenius, A., 2011. Biodiversity information goes public: Gbif at your service. Nordic Journal of Botany 29, 378–381. doi:10.1111/j.1756-1051.2011.01167.x.

Tittensor, D.P., Mora, C., Jetz, W., Lotze, H.K., Ricard, D., Berghe, E.V., Worm, B., 2010. Global patterns and predictors of marine biodiversity across taxa. Nature 466, 1098–1101. doi:10.1038/nature09329.

Troia, M.J., McManamay, R.A., 2016. Filling in the gaps: evaluating completeness and coverage of open-access biodiversity databases in the united states. Ecology and evolution 6, 4654–4669. doi:10.1002/ece3.2225.

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. doi:10.1111/ddi.12637.

Tuomisto, H., 2011. Do we have a consistent terminology for species diversity? yes, if we choose to use it. Oecologia 167, 903–911. doi:10.1007/s00442-011-2128-4.

Turner, M.G., Calder, W.J., Cumming, G.S., Hughes, T.P., Jentsch, A., LaDeau, S.L., Lenton, T.M., Shuman, B.N., Turetsky, M.R., Ratajczak, Z., et al., 2020. Climate change, ecosystems and abrupt change: science priorities. Philosophical Transactions of the Royal Society B 375, 20190105. doi:10.1098/rstb.2019.0105.

UNEP-WCMC, IUCN, 2022. Protected Planet: The World Database on Protected Areas (WDPA) [Online], January 2022, Cambridge, UK. Technical Report. URL: https://www.protectedplanet.net.

Visalli, M.E., Best, B.D., Cabral, R.B., Cheung, W.W., Clark, N.A., Garilao, C., Kaschner, K., Kesner-Reyes, K., Lam, V.W., Maxwell, S.M., et al., 2020. Data-driven approach for highlighting priority areas for protection in marine areas beyond national jurisdiction. Marine Policy 122, 103927. doi:10.1016/j.marpol.2020.103927.

Webb, T.J., Vanden Berghe, E., O'Dor, R., 2010. Biodiversity's big wet secret: the global distribution of marine biological records reveals chronic under-exploration of the deep pelagic ocean. PloS one 5, e10223. doi:10.1371/journal.pone.0010223.

Wickham, H., Francois, R., Henry, L., Müller, K., 2021. dplyr: A grammar of data manipulation. r package version 1.0.3. R Found. Stat. Comput., Vienna URL: https://CRAN.R-project.org/package=dplyr.

WORMS, 2022. World register of marine species database: Statistics. number of records in worms 11th april 2022 [online] URL: http://www.marinespecies.org/.

Yang, W., Ma, K., 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. doi:10.1111/jbi.12108.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A., et al., 2007. How global is the global biodiversity information facility? PloS one 2, e1124. doi:10.1371/journal.pone.0001124.

Zhang, Y., Grassle, J.F., 2002. A portal for the ocean biogeographic

information system. Oceanologica Acta 25, 193–197. doi:10.1016/S0399-1784(02)01204-5.

Zizka, A., Carvalho, F.A., Calvente, A., Baez-Lizarazo, M.R., Cabral, A., Coelho, J.F.R., Colli-Silva, M., Fantinati, M.R., Fernandes, M.F., Ferreira-Araújo, T., et al., 2020. No one-size-fits-all solution to clean gbif. PeerJ 8, e9916. doi:10.7717/peerj.9916.

Appendix A. The database

Table A.1 below shows the data loss for each criterion that we have used to clean our database. We downloaded 71,670,596 records from GBIF and OBIS. Only 820,004 records were useful for our analyses.

Database state	Number of records
Origial records from GBIF and OBIS	71,670,596
Data curation (following Zizka et al. (2020))	5,380,439
Taxonomically filtered data	5,007,322
Deletion of data outside the native range	820,004

Table A.1: Criteria for filtering occurrence data from GBIF and OBIS using bioregions.

Files of the 10,371 marine fish species and their attributes (body size, habitat depth, and cultural value) from FishBase may be found in the GitHub project page of this manuscript: http://github.com/vapizarro/stp_fishes

Appendix B. Species Representativeness Analysis (SRI)

For each cell (i), the SRI is the simple ratio between the observed number of species S_{obs} and the expected number of species (S_{obs}): $SRI_i = S_{obs}/S_{exp}$. Maps for the smallers resolution analyzed ($\sim 1°$) are in Fig. A.2.

Appendix C. Grids resolutions

For spatial representation analysis we evaluated two additional spatial resolutions ($\sim 5° = 3{,}021$ cells, and $\sim 10° = 958$ cells). Table C.2 contains the results of this analysis for these grids. We have also mapped these results (see Figure A.3), to understand how the effect of spatial resolution on the evaluation of biodiversity macropatterns. Finally, we also plot the frequency of cells for each SRI category for the three grid sizes (R1= $\sim 1°$; R5= $\sim 5°$; R10= $\sim 10°$) to understand how the data is distributed in our analyses (see Figure A.4)

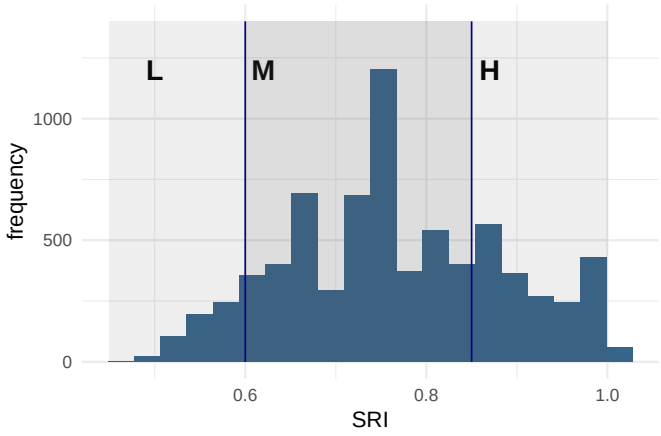


Fig. A.1: Classification of SRI values based on its frequency distribution. This histogram displays the frequency distribution of SRI (Species Richness Index) values and the corresponding class selection thresholds. Cells are categorized as follows: SRI < 0.6 are classified as *low* representativeness (**L**), SRI falling in the range (0.6, 0.85) as *medium* representativeness (**M**), and SRI > 0.85 as *high* representativeness (**H**).

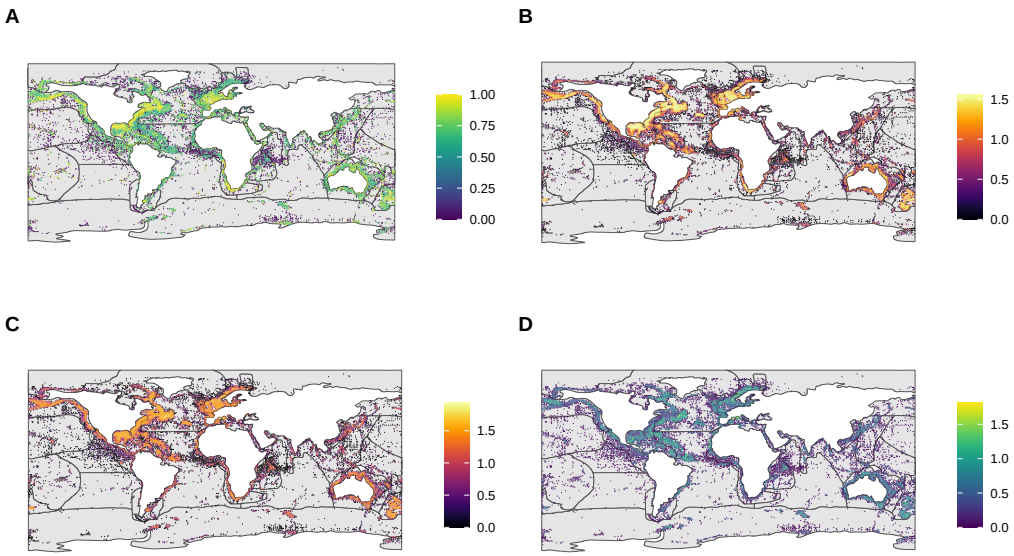


Fig. A.2: SRI and Species richness S depicted from GBIF and OBIS databases. **A**. Species representativeness index; **B**. Observed species richness (S_{obs}); **C**. Expected species richness (S_{exp}); **D**. Difference between raw estimated and observed richness. The difference has been log_{10} transformed after subtraction.

	R1 ($\sim 1°$)					**R5** ($\sim 5°$)					**R10** ($\sim 10°$)				
ID	**NR**	**IR**	**L**	**M**	**H**	**NR**	**IR**	**L**	**M**	**H**	**NR**	**IR**	**L**	**M**	**H**
1	18.49	15.13	5.04	36.97	24.37	0.00	16.67	0.00	33.33	50.00	16.67	16.67	0.00	33.33	33.33
2	68.75	19.79	1.04	10.42	0.00	10.00	40.00	10.00	30.00	10.00	40.00	0.00	0.00	40.00	20.00
3	15.74	6.54	3.39	36.80	37.53	3.57	3.37	0.00	17.86	75.00	0.00	0.00	0.00	10.00	90.00
4	46.35	22.34	7.93	22.13	1.25	28.13	9.38	6.25	40.63	15.63	30.77	15.38	0.00	23.08	30.77
5	42.39	14.75	4.92	27.87	10.07	14.29	3.57	0.00	32.14	50.00	16.67	8.33	0.00	8.33	66.67
6	94.96	2.21	0.13	1.87	0.83	82.13	5.64	0.31	6.58	5.33	62.65	13.25	1.20	10.84	12.05
7	63.24	11.24	0.87	14.46	10.19	17.09	9.40	3.42	29.06	41.03	7.69	7.69	2.56	35.90	46.15
8	79.52	11.27	0.89	7.17	1.15	43.93	11.56	4.05	32.37	8.09	32.69	11.54	3.85	40.38	11.54
9	88.74	8.71	0.00	1.57	0.99	28.74	22.99	2.87	38.51	6.90	7.69	9.62	21.15	38.08	13.46
10	96.31	2.41	0.04	0.88	0.36	70.87	15.75	0.79	7.09	5.51	51.28	15.38	2.56	20.51	10.26
11	23.82	8.42	5.65	32.85	29.26	8.62	0.00	0.00	18.97	72.41	0.00	10.53	0.00	5.26	84.21
12	35.59	21.61	2.45	35.59	4.76	14.29	4.76	2.38	47.62	30.95	5.88	11.76	0.00	17.65	64.71
13	67.52	15.80	1.01	12.00	3.67	13.76	12.84	7.34	44.95	21.10	9.46	6.76	2.70	44.59	36.49
14	45.83	10.83	2.50	30.83	10.00	46.15	0.00	0.00	7.69	46.15	25.00	0.00	0.00	0.00	75.00
15	74.52	13.06	0.00	7.07	5.35	20.00	6.67	6.67	40.00	26.67	37.50	12.50	0.00	37.50	12.50
16	36.68	10.95	3.84	34.65	13.88	5.77	7.69	3.85	28.85	53.85	10.53	0.00	0.00	21.05	68.42
17	91.36	4.90	0.00	1.57	2.17	47.93	19.01	0.00	20.66	12.40	25.00	8.33	0.00	36.11	30.56
18	48.29	16.27	3.78	22.06	9.61	6.50	7.32	7.32	43.09	35.77	10.26	5.13	0.00	28.21	56.41
19	90.40	6.93	0.06	2.27	0.35	53.45	18.39	3.45	17.82	6.90	31.48	12.96	1.85	35.19	18.52
20	63.61	17.35	1.43	13.56	4.04	8.20	8.20	9.84	44.26	29.51	15.00	5.00	0.00	45.00	35.00
21	74.78	9.63	2.84	11.48	1.27	34.68	13.51	3.15	28.38	20.27	21.21	9.09	0.00	27.27	42.42
22	76.12	18.00	0.00	5.10	0.78	33.33	6.17	16.05	43.21	1.23	9.09	4.55	9.09	59.09	18.18
23	34.65	32.02	2.89	24.41	6.04	25.93	0.00	14.81	51.85	7.41	0.00	18.18	0.00	36.36	45.45
24	63.07	17.89	1.38	13.53	4.13	25.93	7.41	3.70	51.85	11.11	20.00	10.00	0.00	50.00	20.00
25	88.02	4.96	0.83	5.37	0.83	52.63	10.53	0.00	21.05	15.79	42.86	0.00	0.00	28.57	28.57
26	60.93	7.04	2.41	22.41	7.22	27.27	12.12	0.00	30.30	30.30	16.67	0.00	0.00	33.33	50.00
27	66.84	10.35	0.70	8.07	14.04	27.78	16.67	0.00	22.22	33.33	7.69	7.69	0.00	23.08	61.54
28	59.84	9.17	2.13	17.67	11.19	30.19	7.55	1.89	30.19	30.19	30.00	10.00	0.00	25.00	35.00
29	41.96	19.87	2.84	26.81	8.52	15.00	10.00	5.00	40.00	30.00	12.50	12.50	0.00	37.50	37.50
30	93.74	3.49	0.20	1.97	0.59	69.45	11.02	1.00	10.52	8.01	42.29	16.57	2.86	22.29	16.00

Table C.2: Surface area as a share of each bioregion (ID) for every SRI category for each of the three grid sizes (R1= $\sim 1°$; R5= $\sim 5°$; R10= $\sim 10°$). Values show the surface area as a share of each bioregion for every SRI category (see §2.2.1). ID is the identification number given to each bioregion (Table 1). **H** are cells with a *high* representativeness of species richness (i.e. SRI > 0.85). **M** are cells considered as having a *medium* representativeness (i.e. SRI $\in (0.60, 0.85)$). **L** cells are cells with a *low* number of records and are thus not considered to be representative of actual species richness (i.e. SRI $\in (0, 0.6)$). **NR** as cells with *no records* (SRI= *NA*), and **IR** as cell with *insufficient records* to apply SRI.

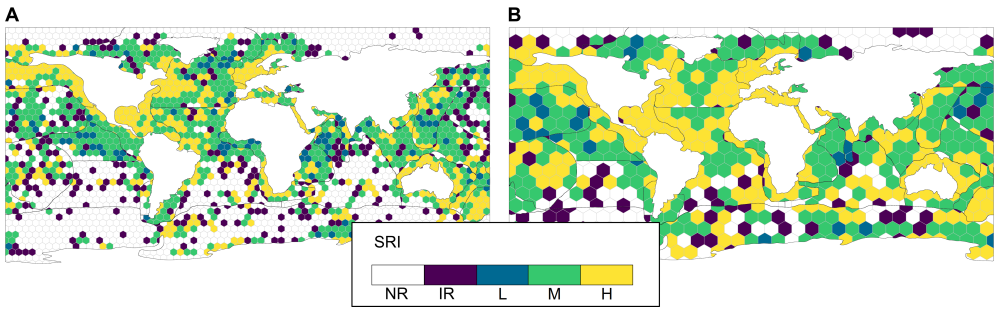


Fig. A.3: Spatial representativeness index (SRI) mapping of cells of size: A=$\sim$ 5°; B= $\sim$ 10°. The categorization of the cells corresponds to the level reached by the SRI, where SRI $>$ 0.85: Amount of data is *high* for the representation of species richness (**H**); SRI=0.60-0.85: Amount of data can be considered of *medium* representativeness (**M**); SRI=0-0.60: Amount of records is *low* (**L**); and SRI = NA: cells with no records (**NR**). **IR** are cells with *insufficient records* to evaluate species diversity representativeness.

Appendix D. Bioregions slopes

We evaluated the slopes of the last 10% of the accumulation curves of each bioregion in our temporal representation analysis. Table D.3 shows the result for each bioregion.

Appendix E. GAP Analysis

We plotted the share of surface areas with MPAs in each bioregion (Fig A.5), and the share of cells of each FAO area for each category of SRI value (Fig A.6).

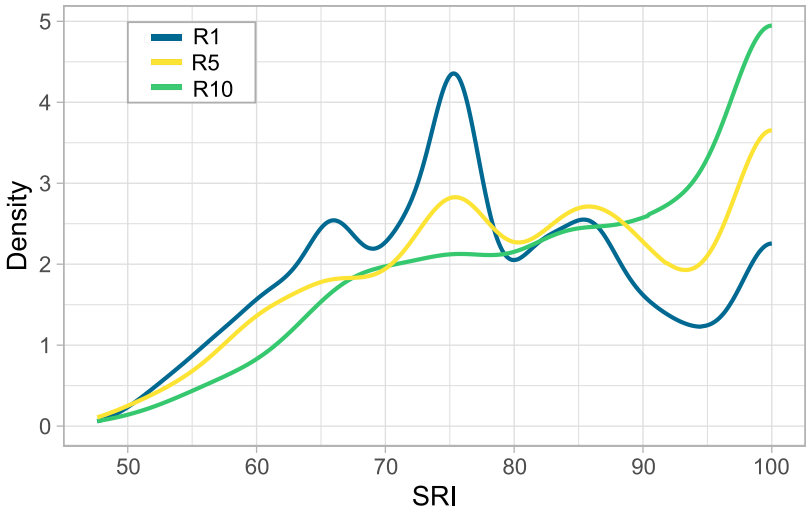


Fig. A.4: Density probability distribution of SRI in three grids of different sizes: R1= $\sim 1°$(blue line); R5= $\sim 5°$ (red line); and R10= $\sim 10°$ (yellow line).

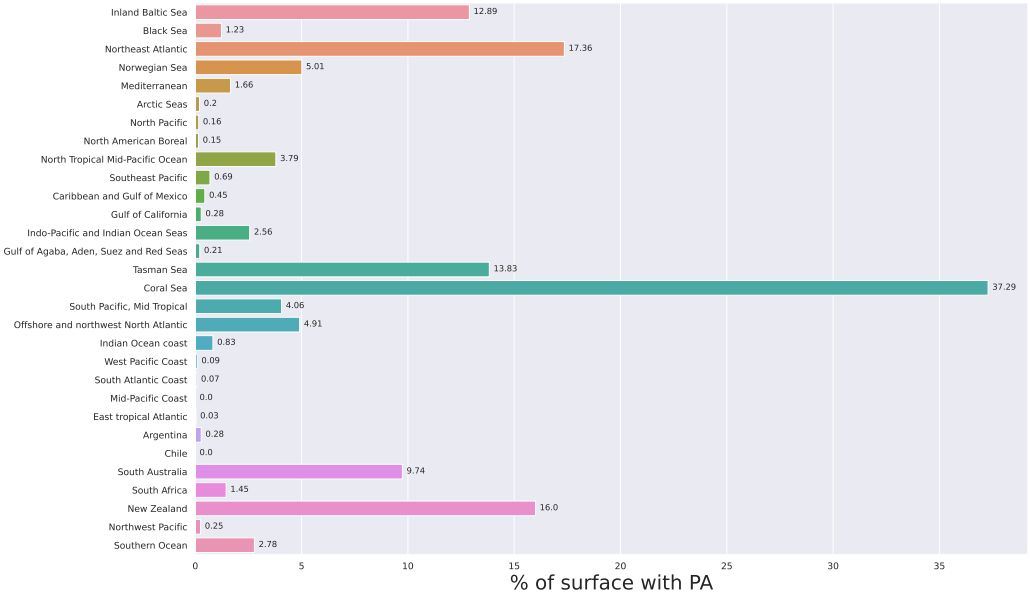


Fig. A.5: Share of surface area with MPAs by bioregions.

Bioregion	Slope
1	0.35
2	1.16
3	1.79
4	0.91
5	1.76
6	0.65
7	4.44
8	1.37
9	6.18
10	4.87
11	10.37
12	7.57
13	32.86
14	4.90
15	6.78
16	21.62
17	10.10
18	6.59
19	12.44
20	23.21
21	11.70
22	1.85
23	4.42
24	3.49
25	2.12
26	7.74
27	12.29
28	4.82
29	14.08
30	2.74

Table D.3: Final slope (10%) of the accumulation curves for each bioregion

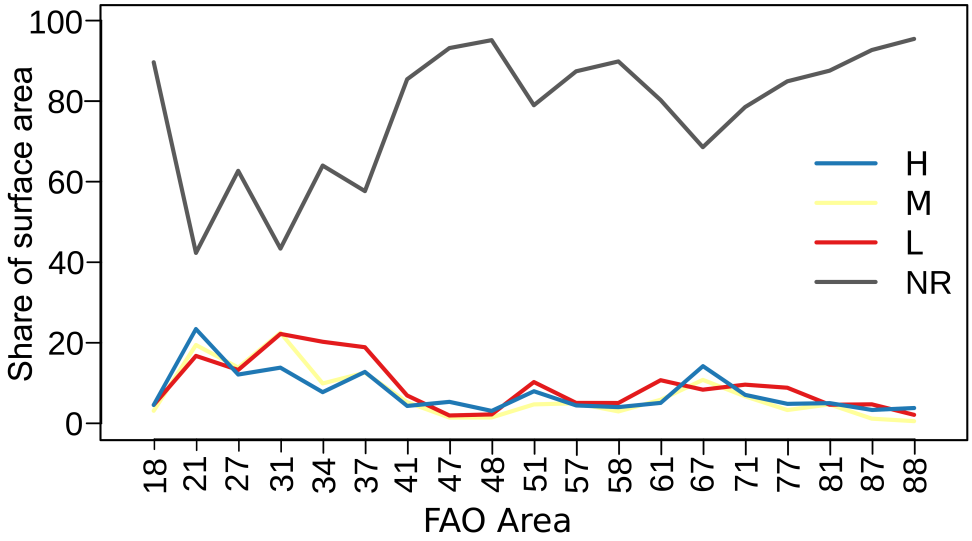
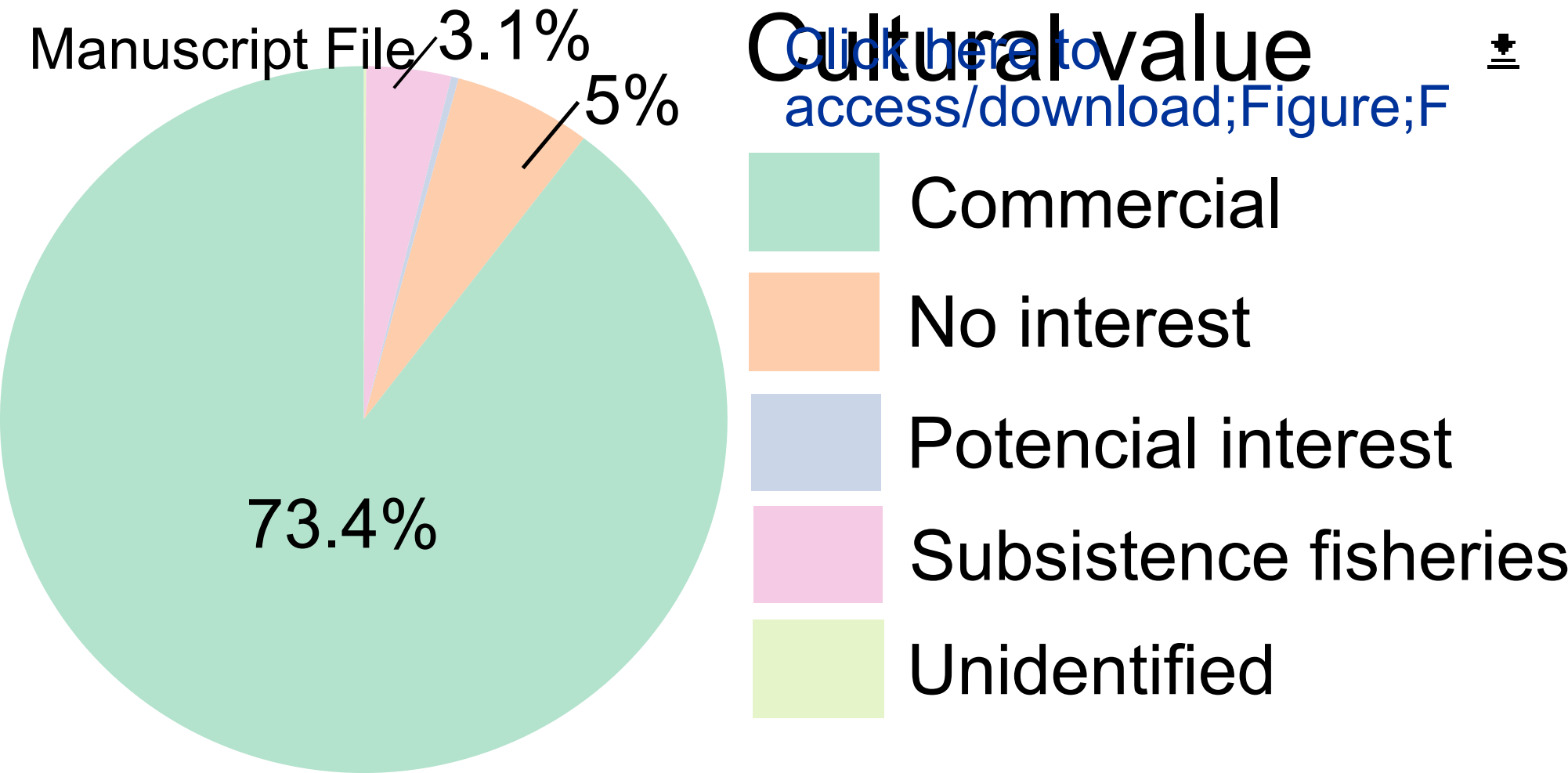
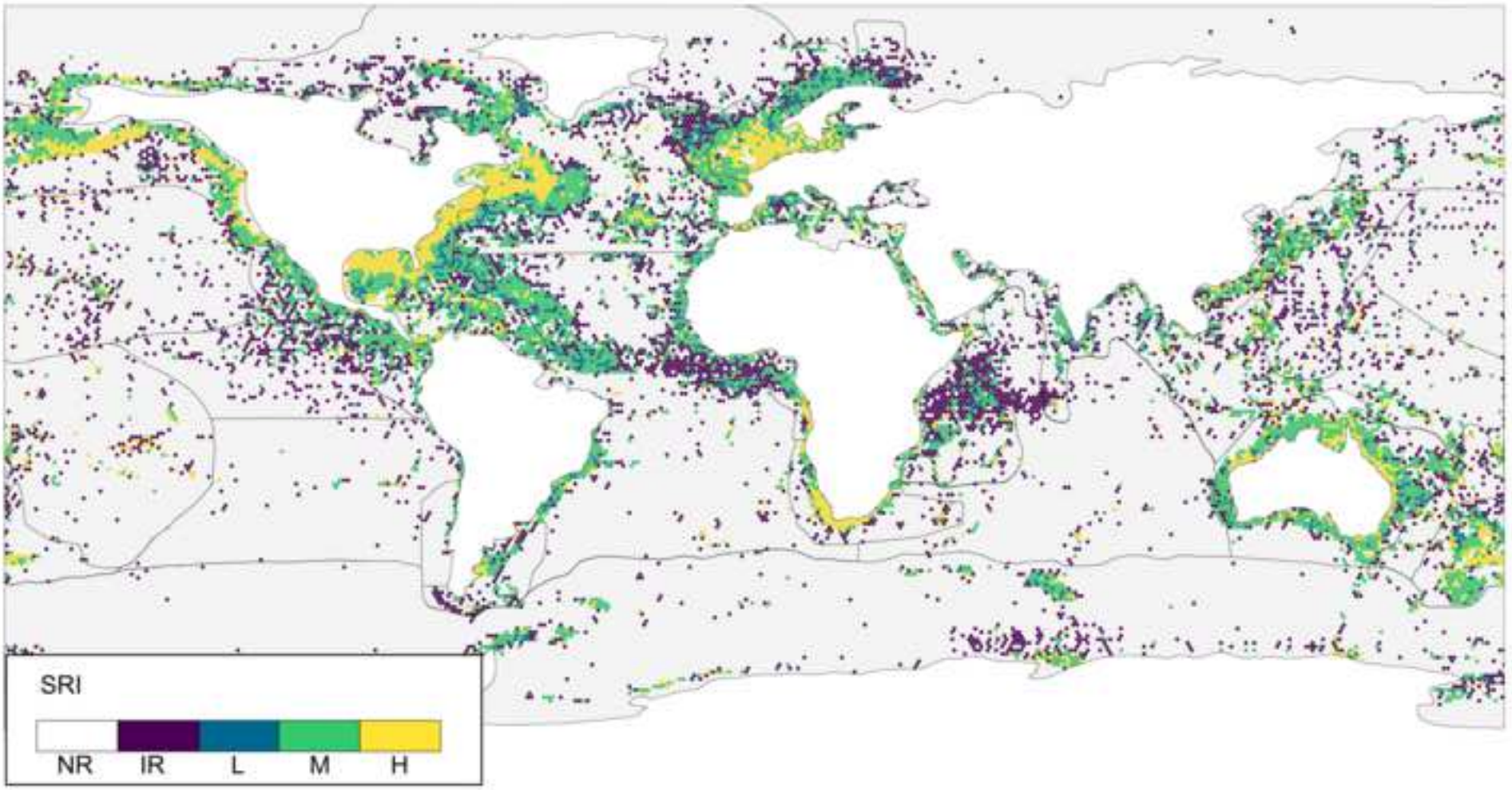
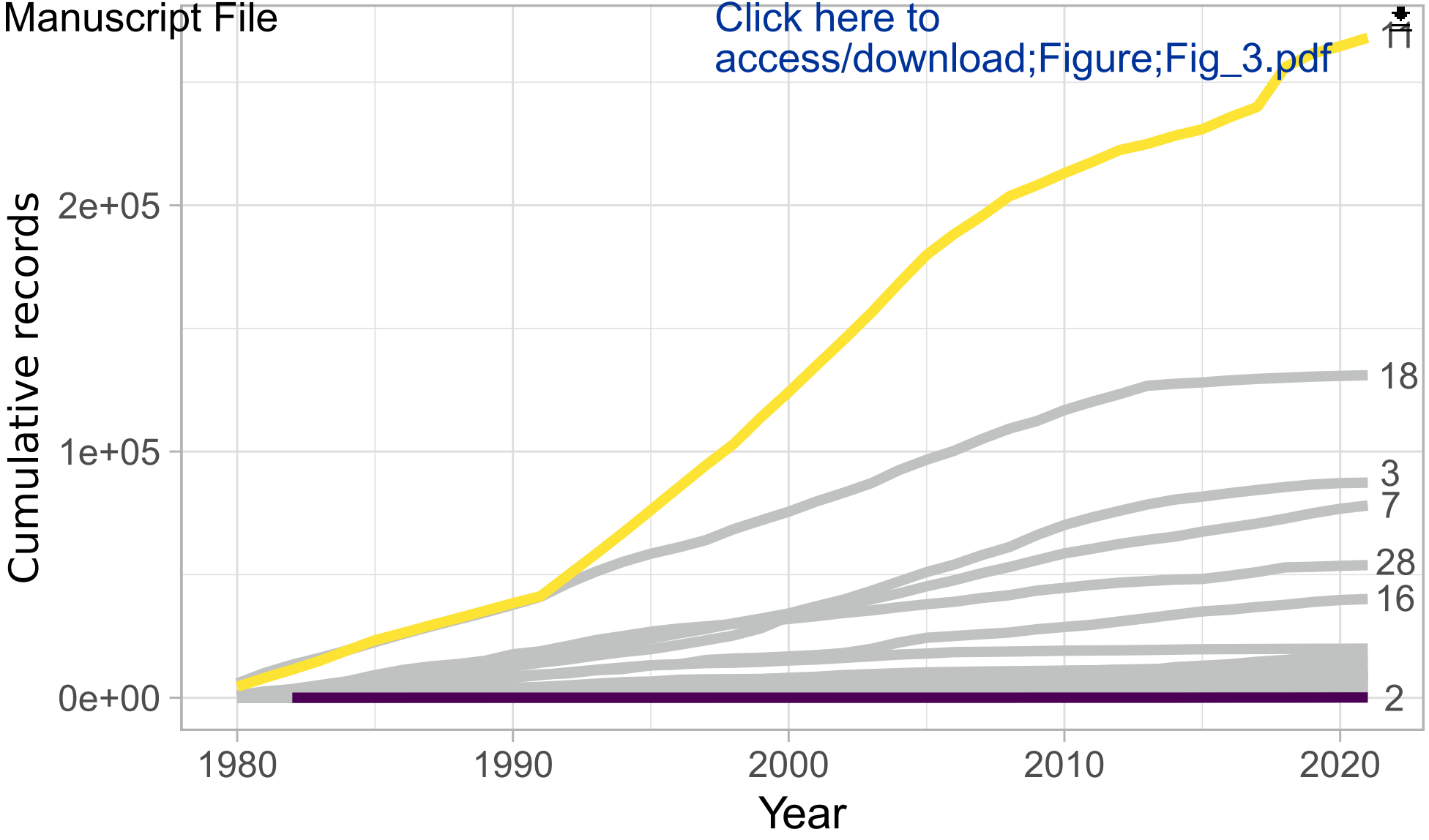
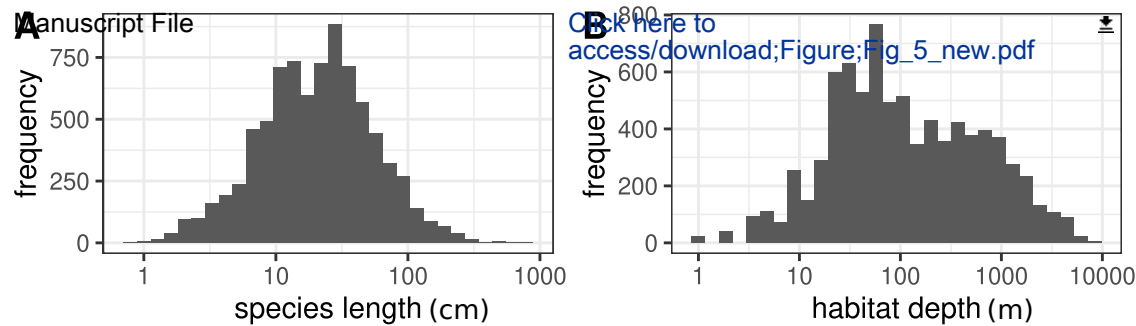


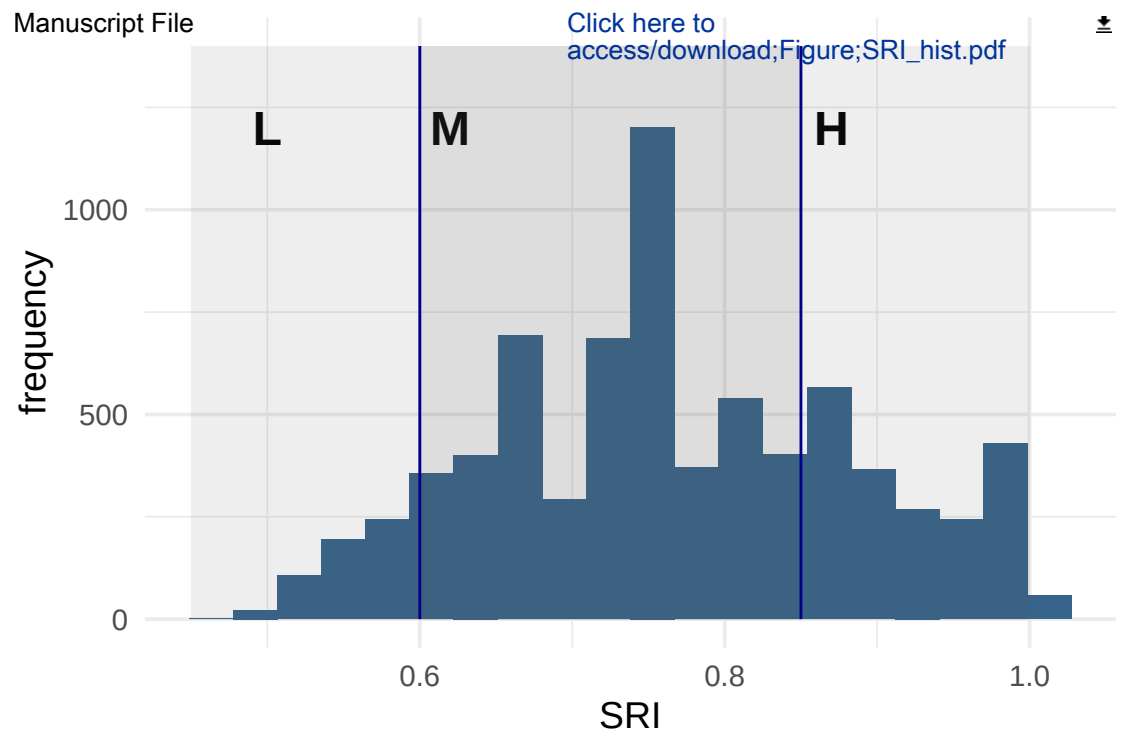
Fig. A.6: Share of cells each FAO area for each SRI value category. SRI= >0.85: *High* representativeness of observed species richness (**H**); SRI=0.60-0.85: *Medium* representativeness of observed species richness (**M**); SRI=0-0.60: *Low* representativeness of observed species richness (**L**); and SRI= *NA*: cells with no records (**NR**).


Manuscript File
3.1%
5%
73.4%
Cultural value
Click here to
access/download;Figure;F
Commercial
No interest
Potencial interest
Subsistence fisheries
Unidentified

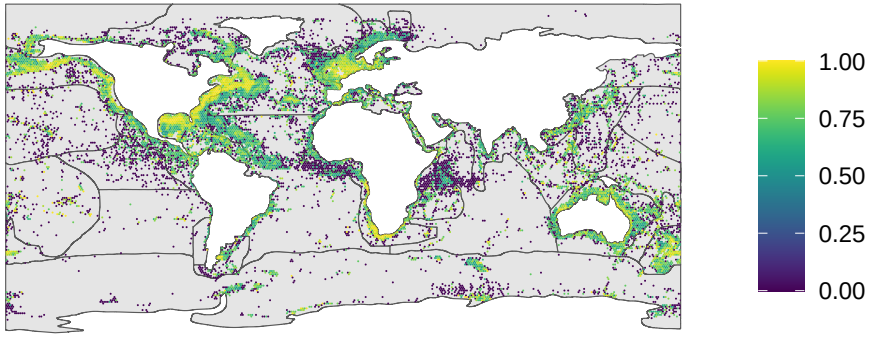




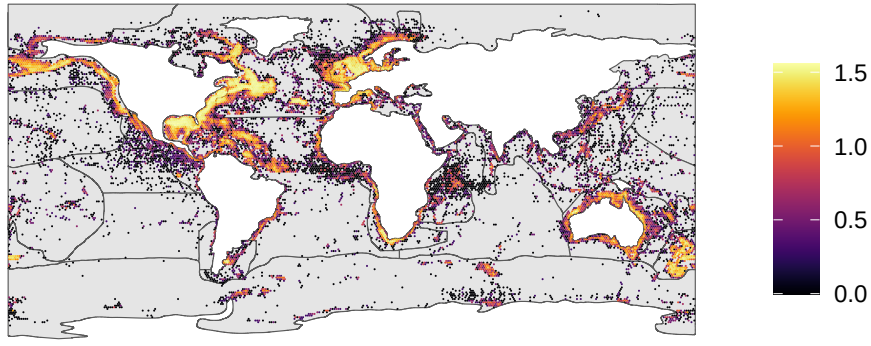




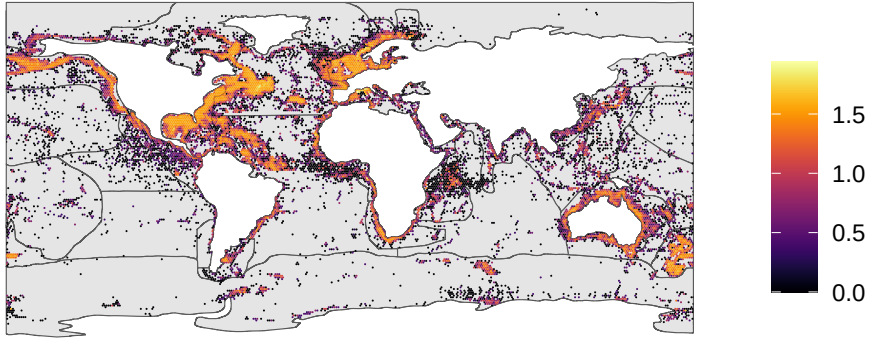
A



B



C



D