MASTER THESIS

Regional biodiversity patterns of marine and terrestrial animal fauna in French Polynesia.

Author: Kilian GRIHAULT–BARREIRO

Internship tutor: **Dr. Cristián MONACO**

Academic tutors : Florence HUBERT & Laurence RODER

October 10, 2023

Acknowledgments

I would like to thank my internship tutor, Cristián MONACO, for his commitment and for his demand.
I would like to thank Laura BENESTAN, postdoc for her precious pieces of advice and her support.
I would like to thank Charlotte MORITZ, for her ideas and the time devoted to this project.
Thanks also to my academic tutors Florence HUBERT and Laurence RODER for their support and flexibility.
A special thought for M2 interns: Grégoire, Clémentine, and Clara who have colored this internship of tears and laughs.
Thanks also to the staff of the Pacific Ifremer Center for the reception.
Finally, I would like to express my eternal gratitude to my mother, without whom none of my study projects would have been possible, as my first and last supporter.

Abstract

Despite the importance of biodiversity in French Polynesia, there was a glaring lack of reliable and comprehensive data on the distribution of marine and terrestrial species in the region. The raw datasets from global aggregators had several shortcomings, including erroneous, incomplete data and frequent duplication. In addition, these data were affected by sampling biases linked to the accessibility of the study sites, such as the presence of roads, ports and airports. This study on regional biodiversity patterns encompasses both marine and terrestrial animal fauna within the unique ecological context of French Polynesia. We provide a cleaned database consisting of 156,727 occurrences, 7,101 species, and 97 and 98% of taxonomic reliability for marine and terrestrial data, respectively. The Gastropods and insects are the largest represented classes of species in marine and terrestrial data, respectively. There are five times more marine than terrestrial species, and most of them (71.1%) are located in the Society archipelago. In terms of spatial distribution, the Tuamotu Islands had the lowest mean marine species richness. A positive correlation was observed between marine and terrestrial species richness across the archipelagos. Dissimilarity matrices revealed two main clusters in the composition of species for terrestrial data, one that links Marquesas and Society islands and the other one links Tuamotu island together. It also revealed a relatively more diluted cluster in the marine community between Tuamotu islands. Multivariate community analysis highlights the difference in the Marquesas islands clustering and the difference in marine species composition in relation to other archipelagos. Tuamotu terrestrial species composition is clearly different than the other archipelagos. Most of the Tuamotu and Gambier islands are poorly surveyed. With regard to biases, accessibility biases were quantified, notably due to the presence of roads, ports, and airports.

All in all, these biodiversity data will contribute to a better understanding of the unique ecosystem of French Polynesia, and to more effective conservation..

1 Introduction

Anthropocene crisis Humans are currently modifying the structure and functioning of natural ecosystems globally Gorman et al. (2023), leading to an unprecedented erosion of biodiversity and the 6th mass extinction recorded on Earth (Ceballos and Ehrlich (2023)). This rapid biodiversity crisis is destabilizing marine and terrestrial ecosystems worldwide, thus threatening the essential services they provide (Díaz et al. (2019)).

To cope with the rapid change and the preservation of ecosystems, we need to set up conservation strategies that rely on baseline biodiversity information, which is unfortunately not evenly available across the planet (Singh (2002)).

Big data era To support responses to global change, the integration of biodiversity knowledge is a must, and the era of big data in ecological sciences is welcome. (Heberling et al. (2021)).

For instance, Farley et al. documented an exponential increase in biodiversity records and other museum databases that should enable us to document biodiversity as well as track changes in biodiversity over time (Farley et al. (2018), Kays et al. (2020))

Thankfully, numerous initiatives aimed at centralizing and capitalizing available data on biodiversity have emerged over the last decades, including online data repositories with an emphasis on free accessibility and geospatial information providing, such as the Global Biodiversity Information Facility (**GBIF**, `https://www.gbif.org/`), and the Ocean Biodiversity Information System (**OBIS**, `https://obis.org/`). These intergovernmental research infrastructures provide gathered occurrences data from museums, surveys, and citizen science observations through the FAIR principles ("Findability, Accessibility, Interoperability, and Reusability") and the Darwin Core (DwC) data standard (Wieczorek et al. (2012)), which is an international standard to facilitate metadata sharing (*metadata refers to format, date, publisher, and overall data description*). OBIS and GBIF are the largest occurrences data portals covering marine and all data, respectively, and they are both routinely used for spatial planning and conservation issues (Takashina and Kusumoto (2023), Underwood et al. (2018), Amano et al. (2016), Levin et al. (2014), Frishkoff et al. (2016), Lin et al. (2022)).

Limits and biases of open databases Nevertheless, the rapid increase in the availability of open geospatial biodiversity databases needs to take into account various types of quality issues, such as incomplete, incorrect information (García-Roselló et al. (2023), Zizka et al. (2020b)), and bias (Qian et al. (2022)). Spatiotemporal biases includes sampling bias, i.e. when certain areas are oversampled and others are undersampled, and observer species identification bias when data is collected either by researchers or by non-specialists.

While cleaning and filtering methods improve the dataset's quality to some extent, spatial and temporal biases still require further attention. This involves (i) estimating the extent of the of influence of such bias on the data collected from open databases (Delgado-Rodríguez and Llorca (2004) which provides a non-exhaustive list of biases and their impact), and (ii) applying correction methods to ensure accuracy and reliability of the dataset in order to explore hypothesis in biogeographical studies (Schiesari et al. (2007)) or formulate conservation recommendations. The difficulty for identifying the different biases in a dataset lies in their overlapping. For example, spatial and taxonomical biases are intrinsically linked as some species in a given location are more studied than others existing elsewhere, suggesting that some species can be considered as more present when they are simply more studied. Similarly, spatial and temporal biases are not independent when taxonomic groups are recorded at different frequencies, which can lead to incorrect estimates of species relative abundance. Another parameter potentially skewing the observations derived from geospatial biodiversity databases is the scale dependency, i.e. spatial patterns can significantly vary within the same area when considered at different scales. Considering large scales can also introduce spatial autocorrelation issues (i.e. nearby locations that should tend to have similar values have not).

Accessibility limits biodiversity survey Frequently, spatial biases can be partially explained by socio-economical context, i.e. areas are more likely to be sampled if they present economic advantages (Meyer et al. (2015)), and by their physical accessibility (Kadmon et al. (2004), Engemann et al. (2015)). The geographical isolation of remote areas, such as islands, implies logistical challenges for the scientific community to survey, which may explain the relatively low sampling effort compared to easily accessible zones. Data collection constraints mean that less spatial, temporal, and taxonomic data can be sampled and exploited. While the conservation status (e.g., level of endemism and vulnerability) tends to be well documented in regions of easy access and in zones of high species richness (refs), biodiversity surveys often neglect remote and small areas, including islands. The paucity of information for islands and atolls is particularly worrying because (1) they are a priori highly vulnerable ecosystems due to reduced species diversity, and (2) they potentially harbor high levels of endemism (Russell and Kueffer (2019), Envorn Resour, Simberloff (2000)).

French Polynesia With 118 islands and atolls spread over five archipelagos and scattered over 4.8 mi km2 (Andréfouët and Adjeroud (2019)), French Polynesia represents a unique model of fragmented territory with isolated components within. The number of islands and the distance between them imply a strong presumption of heterogeneity in the sampling effort. French Polynesia biogeography has been only partially studied, mainly through specific taxonomic groups over its entire surface. An incomplete sample of marine species has been investigated such as marine molluscs (Salvat and Tröndlé (2017)) or brown seaweeds (*Phaeophyceae*) (Vieira et al. (2021)), and reef fishes in all South

Pacific (Kulbicki (2007)). Terrestrial and freshwater arthropods have also been surveyed (Ramage (2017)). An overview of terrestrial biota and biogeography has been provided by (Gillespie et al. (2008)) and (Hembry et al. (2018)) but was primarily based on phylogenetic studies and for a series of hot spots archipelagos (Society and Marquesas Island).

Due to the lack of centralized, complete, and evenly distributed data, it is difficult to exhaustively gauge the biogeographical status across French Polynesia.

Despite these challenges, studying those geographical structures can offer unique scientific opportunities and insights to understand and respond to the current biodiversity crisis since insularity has long been recognized as a living laboratory tracing the evolution of species, with several renowned examples of endemism and extinction of marine and terrestrial island species (Baldacchino (2006)).

Establishing spatial patterns of marine and terrestrial animal fauna prior to assessing endemism rate, geographical isolation, or dispersal for each taxonomic group across the region is, therefore, crucial for assessing ecological scenarios (Sarzo et al. (2023)) and examining the role of animal species in an island ecosystem.

Here, using data originally downloaded from open-access databases (GBIF, OBIS), we compiled and curated the first marine and terrestrial animal occurrence records for French Polynesia. We then used the dataset to (i) draw up an initial characterization of those species, (ii) identify poorly surveyed islands, and (iii) quantify island-specific accessibility biases leading to heterogeneous sampling efforts. This study lays the foundations for metaecology (in the sense of Schiesari et al. (2007)) in French Polynesia with a reliable set of empirical geospatial data, with the aim of facilitating future work in biogeography or any other ecological branch requiring such data, and thus responding to environmental conservation and prediction issues.

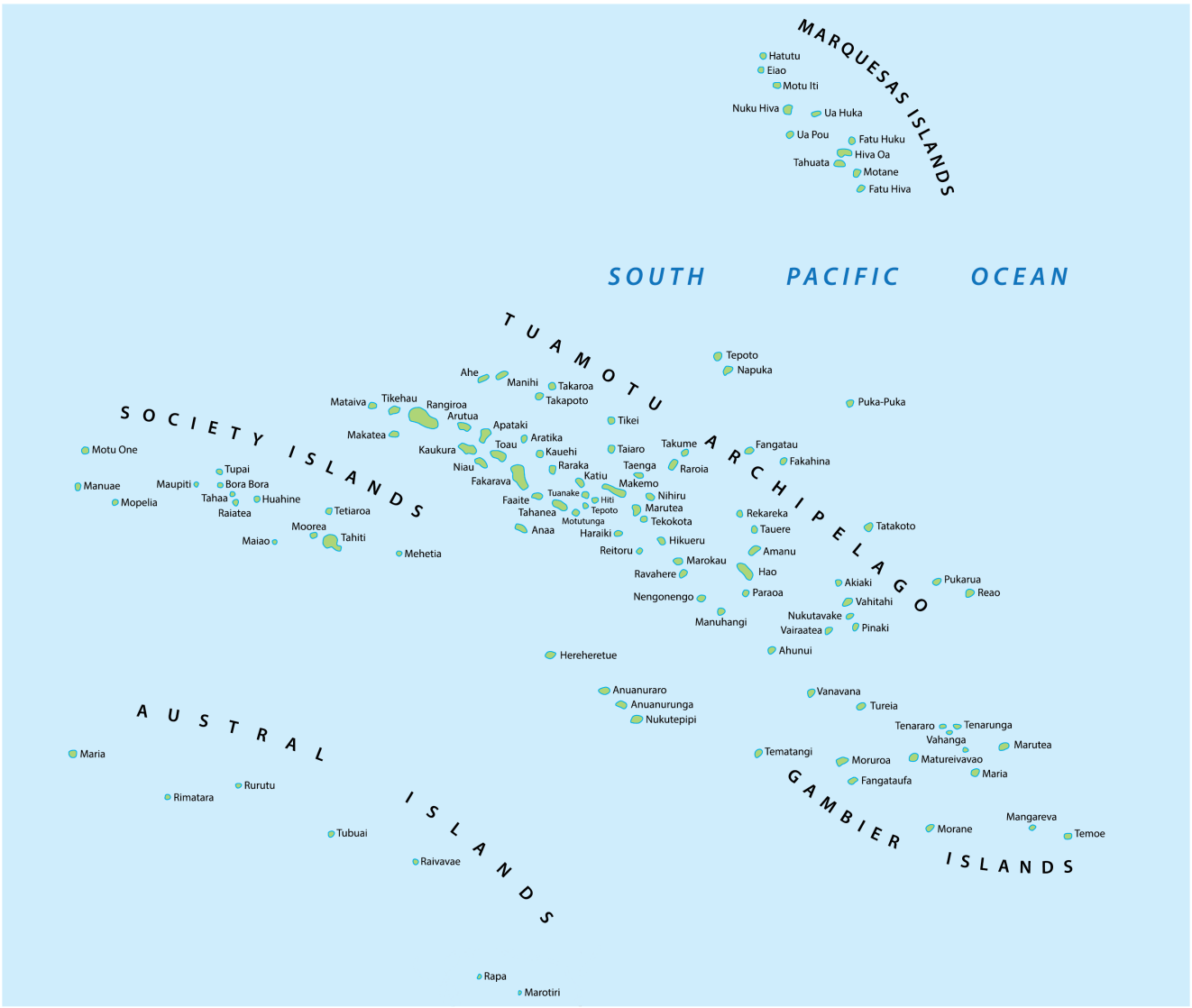

Figure 1: Map of French Polynesia, divided into five administrative subdivisions (Source : *WorldAtlas*)

2 Materials and methods

2.1 Data access

Occurrences datasets were downloaded through the GBIF portal, i.e. sets of species occurrences (or another taxonomic level) at a particular place on a specified date. The raw dataset was then downloaded following the Darwin Core standards, on May 24, 2023 (GBIF.Org User (2023)). Animal occurrences in French Polynesia were selected, yielding a total of 297,789 occurrences. The area delimiting our region was defined by the polygon formed by the intersection of latitudes 5°S and 30°S with longitudes 134°W and 155°W. GBIF and OBIS signed a data-sharing agreement on August 31, 2020, which was effective at the time we downloaded the data, so all data extracted from OBIS was also present in GBIF. We considered 112 islands, six shoals, and two reefs. The nearest geographical structure (i.e., island, seamount, or reef) was attributed to each occurrence.

2.2 Data reliability

2.2.1 Filtering

Cleaning the dataset was done by filtering the occurrence records based on ad-hoc criteria implemented sequentially (fig. 2). Geographic and taxonomic reasons (Zizka A et al. (2020)), are notoriously prevalent in old species occurrence records (Maldonado et al. (2015)). Because the standards of data collection have improved since 1950, the dataset was restricted to records produced after that year. (**??**). Absence data often contain many biases, including false negatives (i.e., the species is present, but has not been detected or recorded) and sampling biases, (i.e. confusing no-observation data with false positives). To rule out potential biases with false positives and no-observation data, we removed all absence data (Bonnet-Lebrun et al. (2023)). We restricted occurrences to those described as: "Human observation", "machine observation", "material sample", "material citation" and "preserved specimen" based on (Smith et al. (2018)) methods. Real duplicates were removed according to the simultaneous combination of similar values in *decimalLatitude, decimalLongitude, ScientificName, Year, Month* and *Day* categories.
Occurrences with no information in *Scientificname* were removed from the dataset. Records missing a date stamp or habitat information were automatically removed from the dataset.

2.2.2 Taxonomic reliability

To get the taxonomic score of the dataset we followed the taxonomic reliability nomenclature used by WoRMS (*World Register of Marine Species* https://www.marinespecies.org/), ITIS (*Integrated Taxonomic Information System*, https://www.itis.gov/), *CoL* (Catalogue of Life, https:

//www.catalogueoflife.org/), and TaxRef (MNHN taxonomic referential https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref). We also assumed that if there was a taxonomic misidentification, it would occur at the species level (Appendix fig. 9), so we selected only occurrences at species or subspecies levels in the dataset. First, we used the WoRMS match taxon tool (https://www.marinespecies.org/aphia.php?p=match) and selected the following options to display: *ScientificName*, *Accepted name*, *Taxon status*, and *Environment* (i.e. habitat). Matched records were kept in the dataset while ambiguous or unmatched records were verified by hand with the above-mentioned taxonomic repositories. According to the validation of scientific names, we attributed a binary code to each species, which was 1 for a match and 0 for a non-match. The taxonomic score was calculated by dividing the number of corresponding species by the total number of species.

2.2.3 Habitat

The habitat occupied by species (i.e. terrestrial, marine) was extracted mainly from the WoRMS repository using the R package **worrms**. This was completed using TaxRef, and sometimes, manually imputed. Species described as exclusively terrestrial were extracted to form our terrestrial subsets and species described as exclusively marine were selected from the marine subset. Species that occupy several habitats (especially terrestrial) were classified as "others". Biogeographical status was attributed to species based on Taxref (exhaustive list of considered biogeographic status provided by the NHMN available in table 1).

2.2.4 Geospatial reliability

The information on habitat was split into 4 categories based on WoRMS criteria, i.e. *Marine, Brackish, Fresh* and *Terrestrial*. To estimate an index of geographical coordinate precision, the latitude or longitude for each occurrence with the fewest decimals was selected, and then divided by six, which is the maximum decimal precision.

2.2.5 Barcoding issues

In occurrences with a material sample method of *basisOfRecord*, some species are identified using DNA environmental DNA methods. These records do not necessarily have a taxon name but a BOLD (Barcode of Life Data). To determine the correct taxon name of each entry, the *bold_identify* function of the **bold** package in R (R Core Team (2021)).

2.3 Evaluation of biases

2.3.1 Taxonomic depth

The taxonomic depth score was assessed by computing occurrences recorded up to the phylum, class, order, family, genus, and species. We assume that fewer errors can occur in well-detailed taxonomic information (Bonnet-Lebrun et al. (2023)).

2.4 Biodiversity analyses

2.4.1 Species richness

α-diversity The species richness is the number of distinct species in the considered location, it is a widely common measure for assessing α-diversity. Species richness was visualized on a reference square grid, using WGS 84 (World Geodetic System 1984) projection and a 0.5°*0.5° resolution, by plotting the number of species recorded in each grid cell. A scatterplot across the different islands was plotted to identify potential differences between marine and terrestrial richness. We used a linear regression curve to model the relationship between terrestrial species richness and marine species richness.

β-diversity The community composition of each island was established at the species levels. Matrices of species community dissimilarity based on the Jaccard index have been displayed to see the potential correlation between geographical structures. The minimum species threshold selected to use the *avgdist* function of the R package **vegan** was 10 for terrestrial data and 100 for marine data. The variation in the composition of species between islands for marine and terrestrial data was performed through principal coordinate analysis (PCoA).

2.5 Accessibility bias

The heterogeneous distribution of occurrences is not only caused by the absence of the species but also because of insufficient sampling effort in the area, preventing comparison of biodiversity across areas.

To assess the gap of information due to accessibility biases in French Polynesia, we used the R package *sampbias* (Zizka et al. (2020a)), which provides a means to quantify the importance (relative weight) of different anthropic accessibility factors (cities, roads, airports, rivers) driving the sampling effort. Estimates of relative weights are done using a Bayesian approach, which compute the geographic distance between occurrences records and each anthropic accessibility factor. The estimated sampling rate (i.e. the expected number of occurrences) are then calculated for each pixel according to the cumulative effects of each bias. For the geospatial data required in the analyses, we

extracted the river and road vectorial cartographic data from datagouv data and cities (larger than 1000 inhabitats) from the open data (open data soft cities data). Finally airports and ports were taken from airports data humdata and ports data humdata. The shapefiles were merged on QGIS 3.30.3 before being used to prevent issues related to map projection. The pixel resolution was set at 0.05 degrees (ca 5*5km). After extracting the information for each pixel and each habitat of the grid, we gather pixels around islands centroids. Ninety-nine percent of distances between centroids were superior to 58.8km, which is why we limited the buffer size to five, which is equivalent to a ten-pixel distance between centroids and is equivalent to 50km. The choice of buffer sizes for each habitat was made following an analysis of variance (ANOVA) that compared five buffer values from one to five pixels. The final buffer size was selected considering the p-value associated with each comparison. Specifically, we chose the buffer size that had the highest p-value, indicating that it had the least significant differences compared to the other buffer sizes, here 3.

3 Results

3.1 Data description

Raw description The raw GBIF dataset contained 396,381 occurrences concerning 22987 species. The occurrences per species ranged from 1 to 10476, with an average of 17 records per species. No scientific names were found for 63822 (21.10%) records. 5.7% of records did not have date information at all. Years, months, and days were missing for 5.1, 5.7 and 6.3% of occurrences, respectively. Years 2011, 2006, and 2009 had the highest number of records (94288, 27074, and 21374, respectively).

Filtering Animals in the raw dataset accounted for 297789 occurrences (75.1%). 20967 occurrences were dated before 1950 or do not have information on the year. Only 144 absence records were present and discarded to avoid false negatives. Invalid recording methods, i.e. those that have been discarded such as fossil materials, included 0.3% of occurrences. 33.3% of data have an abundance of species information (i.e. *individualCount* column), and the abundance of species goes from 0 to 40000 (1 species: *Onychoprion fuscatus*), and the overall mean for French Polynesia is approximately 277 individuals per species. 22% of records had an identified dataset name among the 127 known (based on variable *datasetName*). The largest data collection comes from the Pelagis observatory with 22014 occurrences and 43 species belonging to Chondrichthyes, Aves, and Scyphozoa classes. Contributions of the National Museum of Natural History and of the Pelagis laboratory account for 79.6% of identifiable data. 107187 records were found to be duplicated data (35.99%) and duplicates came from 24.68% of "Citizen science" publishers. At the end of our filtering process, 71 species (12893 occurrences) without habitat were discarded from the dataset.

The cleaned dataset was composed of 156727 occurrences accounting for 7101 species. Human observation records are the most frequently used recording method with 71% (111349) of total occurrences and 48.67% (3458) of total species. *Gygis Alba* is the most represented species. the marine habitat is the most represented in this recording method. The cleaned dataset has a taxonomic score of 0.874 for species. The three most represented classes through occurrences are "Teleostei", "Aves", and "Gastropoda" with respectively 48.43% (22%), 23.05% (1.77%), and 10.40% (26.46%) of total occurrences (species). In terms of species, the second most represented class is "Malacostraca" with 1117 species (15.72%). 5738 marine and 980 terrestrial species were extracted from the cleaned dataset with 110890 and 9765 occurrences, respectively (fig. 3c).

Biogeographic status has been provided for 70.5% of marine occurrences and 40.7% of terrestrial data. 91.9% of occurrences and 85.4% of species are marine data.

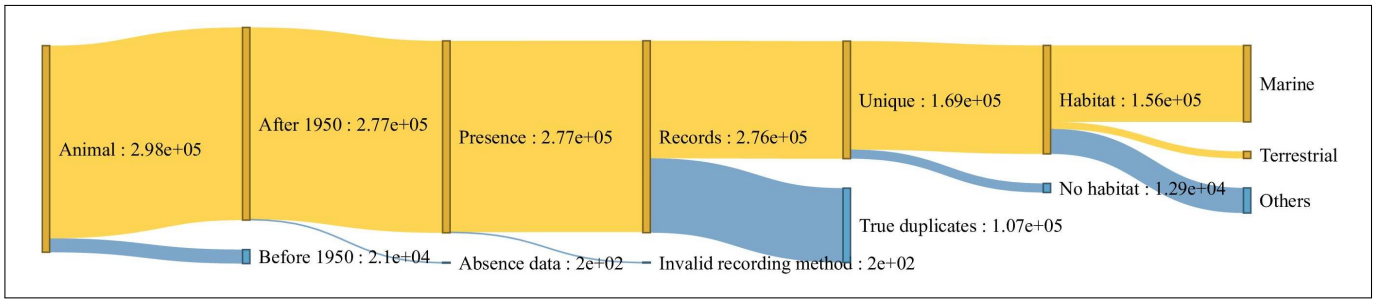

Figure 2: Sankey diagram of filtering and quality controlling steps: OBIS records already present in GBIF, removing data before 1950, absence data, invalid recording methods, true duplicates, species with no habitat information.

Taxonomical assessment The taxon score is 88% for species (815 species have an incomplete or incorrect taxon name). Marine and terrestrial original data had a taxonomic score for species of 97% and 98%, respectively.

3.1.1 Spatial distribution

Marine 116 geographical structures over the 120 considered (see 2.1 on page 7) have information on marine species and 63.3% of them have less than 100 different marine species. Citizen science observation accounts for 15.3% of occurrences and 1419 species (including 147 unique species). The Society archipelago contains 65.5 % of total marine occurrences, of which 92% are found in the high islands of Moorea, Tahiti, and Raiatea-Tahaa. The Society archipelago hosts 71% (4079) of marine species (fig. 3a), and has a mean species richness mean of 628 per island (fig. 3f). The Gambier archipelago is the least represented with 12.5% of species and 3.33% of all marine occurrences. The Tuamotu archipelago had the lowest mean marine richness with 70.8 species per island.

Terrestrial 50.4% of geographical structures contain terrestrial information, and 95% of islands have less than 100 different terrestrial species. For marine data, the Society archipelago had the largest number of occurrences and species referenced with 63.5% and 58.5% (573), respectively (fig. 3b). Moorea, Raiatea-Tahaa, and Tahiti islands represent 83.9% and 53.3% of Society archipelago and French Polynesia occurrences. The Gambier archipelago included only 46 species referenced as terrestrial (4.7%), with 1.5% of total terrestrial occurrences, and a mean of 3.31 species per island.

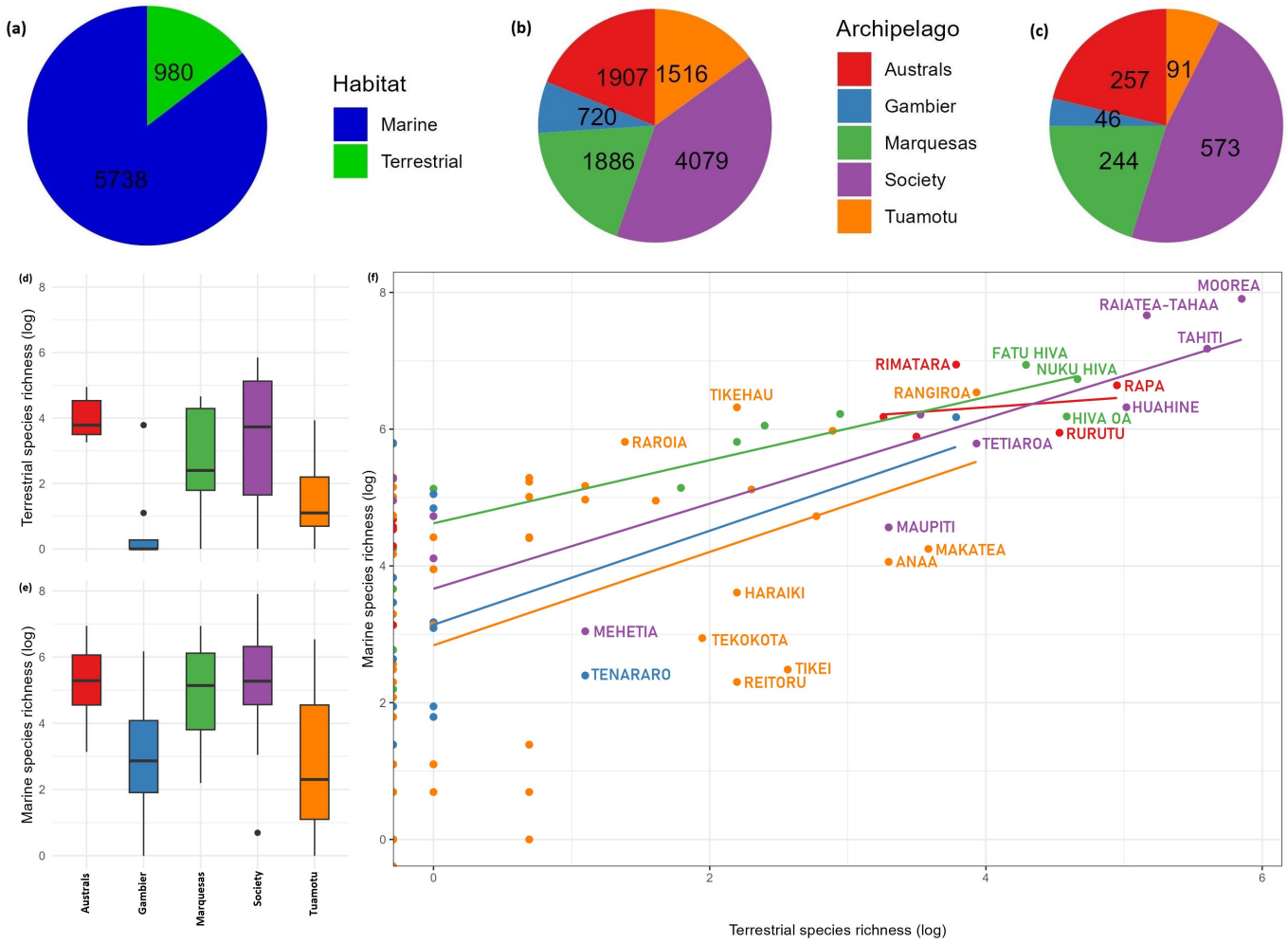

Figure 3: (a) Proportion of marine and terrestrial occurrences in the cleaned dataset. (b) Marine and (c) terrestrial occurrences distribution across the five archipelagos of French Polynesia (the numbers inside the pie charts are the number of species corresponding to the archipelago). (d) Terrestrial and (e) marine species richness boxplots across archipelagos.(f) Relationship between marine and terrestrial species richness.

3.1.2 Species richness analysis

Considering all archipelagoes, there was a positive correlation between the species richness of marine and terrestrial ecosystems (fig. 3f). Tuamotu and Society islands have the highest variability of marine and terrestrial mean species richness, respectively (fig. 3d,e). .

Marine The vast majority of marine species were recorded in the islands of the Society archipelago (fig. 4). The east part of Tuamotu and the west part of Gambier have the lowest species richness. Marine data were composed of 18 different phyla and is concentrated near terrestrial structures (e.g. in lagoons (e.g. Appendix, **??**). Among them, Chordata, Mollusca, and Arthropoda represent 94.5% of records, of which 79.6% concern the classes Teleostei and Gastropoda. *Tridacna maxima*, *Acridotheres tristis*, and *Megaptera novaeangliae* are the most represented species with more than 1000 occurrences each across French Polynesia. 88.9% of *Tridacna maxima* have been recorded by institutions, found in all archipelagos except the Marquesas, of which 58.6 % have been provided by the "UMS PatriNat (OFB-CNRS-MNHN), Paris".

Terrestrial 99.8% of records are referenced with the same trio of phyla as marine data: Chordata, Mollusca, and Arthropoda. The *Aves, Insecta*, and *Gastropoda* represent 92.9% of occurrences. 82.6% of terrestrial species have less than 10 occurrences. *Geopelia striata* is the most represented species in terrestrial records with 1174 occurrences, of which 92.3% were provided by the "Cornell Lab of Ornithology".

3.1.3 Multivariate community analysis

The terrestrial dissimilarity heatmap has a mean Jaccard index of 0.94 across 32 geographical structures containing at least 10 species each. The marine dissimilarity heatmap has a mean Jaccard index of 0.96 across 63 geographical structures containing at least 100 species each. Two main clusters were found, in the terrestrial dissimilarity heatmap (fig. 5 b). One of them shows a similarity between some Marquesas and Society islands (e.g.Bora Bora and Ua Huka), while the other one appears to connect the Tuamotu islands together (e.g. Tikehau and Takapoto). Three island clusters from marine data were distinct (relatively more diluted than terrestrial ones), linking some Tuamotu islands together, such as some Austral geographical structures (fig. 5). Tuamotu islands have the widest spread on the two main axes of the PCoA (e.g. spread for marine: $\Delta_{PCo1} = 0.38$ and $\Delta_{PCo2} = 0.58$) for both habitats (fig. 5 c, d). For marine data, Marquesas species community structure is significantly different from that of the other archipelagos. For terrestrial data, the islands of the Tuamotu archipelago have an orthogonal direction to the main tendency driven by Marquesas and Society along the PCo2 axis.

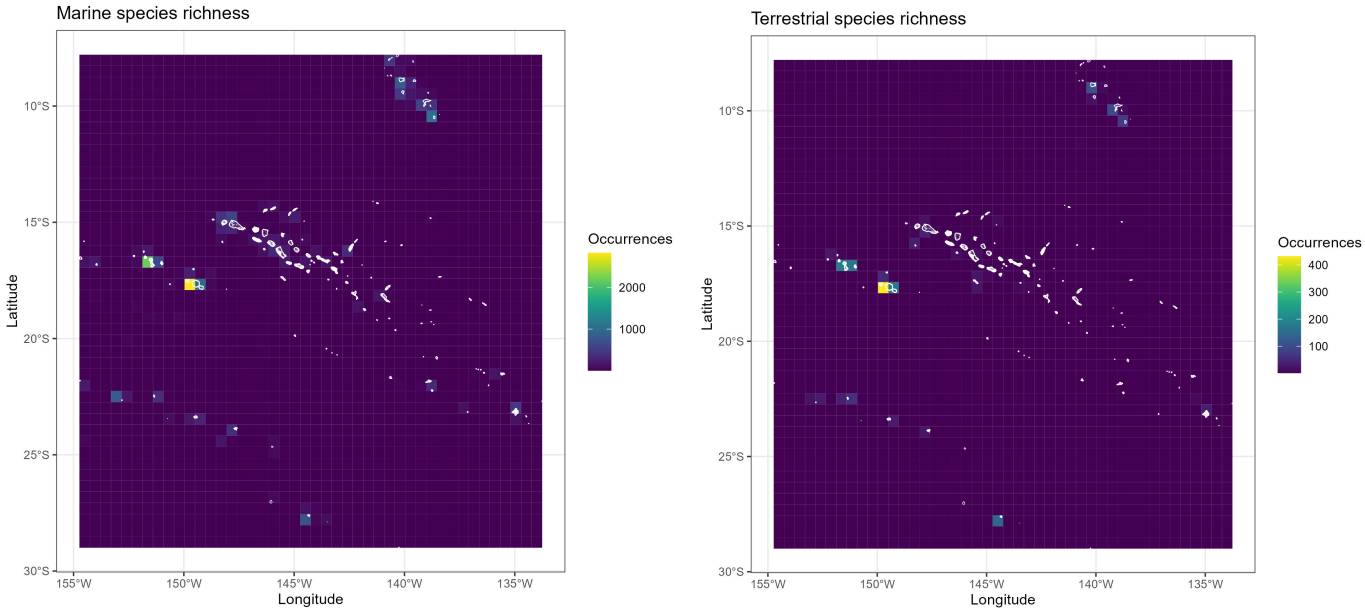

Figure 4: Observed species richness across French Polynesia for marine (left) and terrestrial (right) (i.e, number of species per 0.5*0.5° cell). Legend displays the cube root of actual species number

3.2 Bias assessments

Taxonomic depth Cleaning of the dataset increased the taxonomic depth score by approximately 157.7% (Appendix fig. 9). In the resulting dataset, marine and terrestrial data had a taxonomic depth score of 97.0% and 92%, respectively (fig. 11).

Sampling bias For the marine data, a strong effect of roads, a moderate effect of ports and airports, and a negligible effect of cities and rivers (waterbodies) on the sampling effort have been found at the 0.05 degree study scale (fig. 7). Similar results were found for the terrestrial data, where airports, ports, and roads were mainly responsible for the accessibility bias. All models provided by the *sampbias* package estimated a low number of collected records in the Tuamotu and Gambier archipelagos for both datasets, except for the Mangareva, Hao, and Arutua islands. Most islands of the Society archipelago are estimated to be oversampled compared to others (see *Figure* **??**).

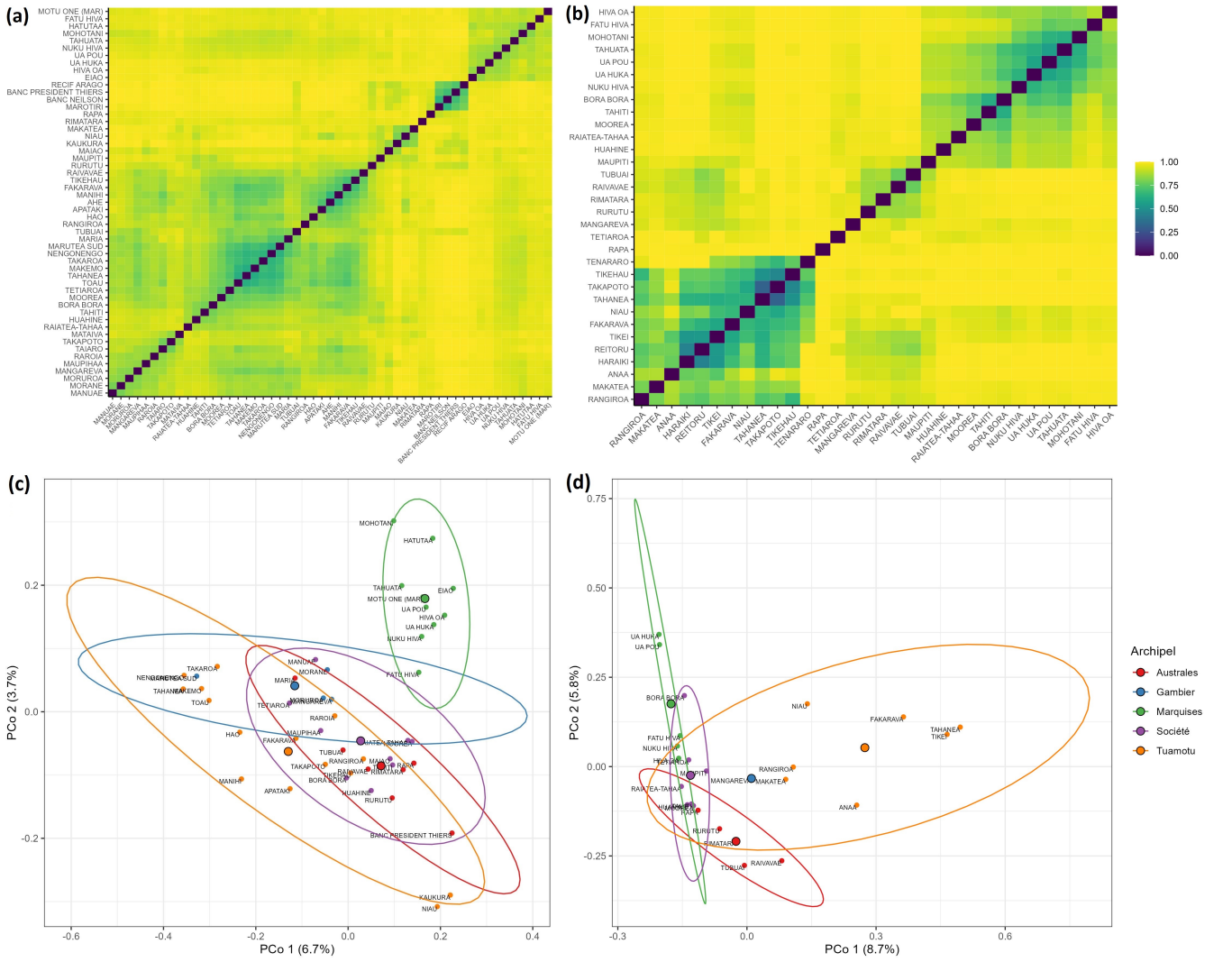

Figure 5: Dissimilarity matrices based on Jaccard distances for (a) marine and (b) terrestrial data. The dimension of dissimilarity matrices is based on the number of islands filtered by the threshold sample (here designed by parameter *sample* of *avgdist* function in **vegan** R-package). Principal coordinate analysis (PCoA) of species community of islands across French Polynesia, for (c) marine and (d) terrestrial fauna.

4 Discussion

General Our dataset had a taxonomic reliability of 98% and 99% for marine and terrestrial datasets, respectively. Despite the size of French Polynesia, the relatively small proportion of land area (4167 km^2) among marine area ($2.5 * 10^6$ km^2) is reflected in the difference of data between these two habitats, indeed, our results show that the number of marine occurrences (110890) and species (5738) is much higher than the number of terrestrial occurrences (9765) and species (980). Gastropoda and Insecta classes are the more represented in the data for marine and terrestrial species,

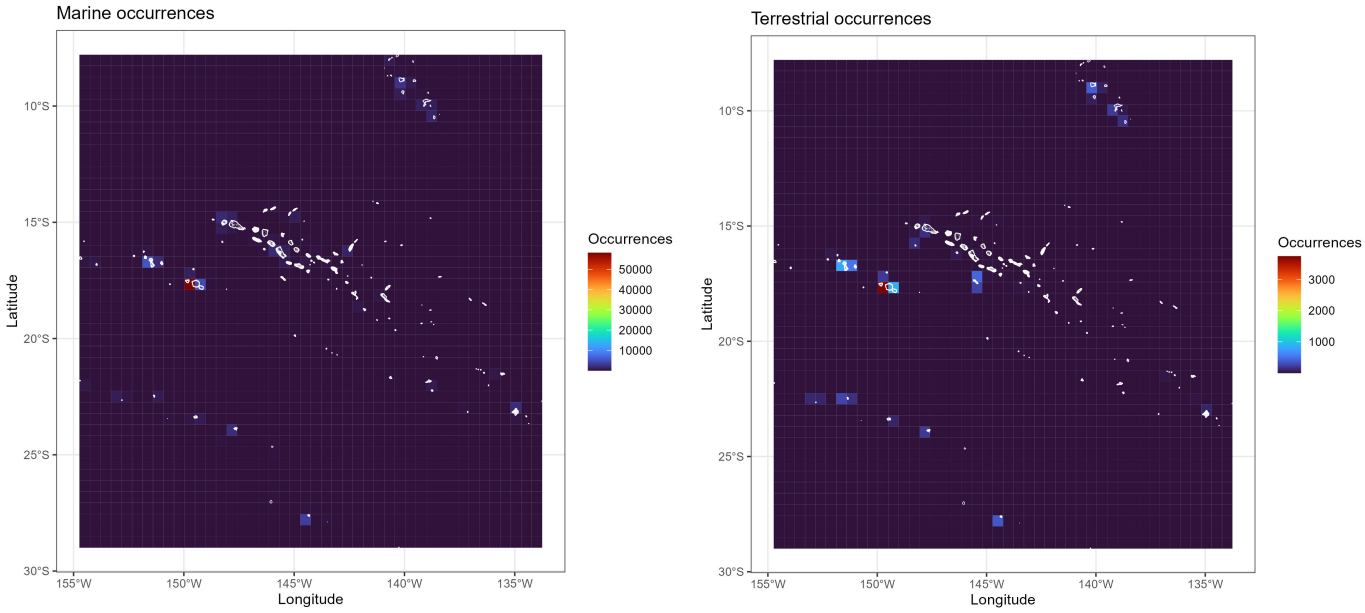

Figure 6: Occurrences across French Polynesia for marine (left) and terrestrial (right) (i.e, number of occurrences per 0.5*0.5° cell).

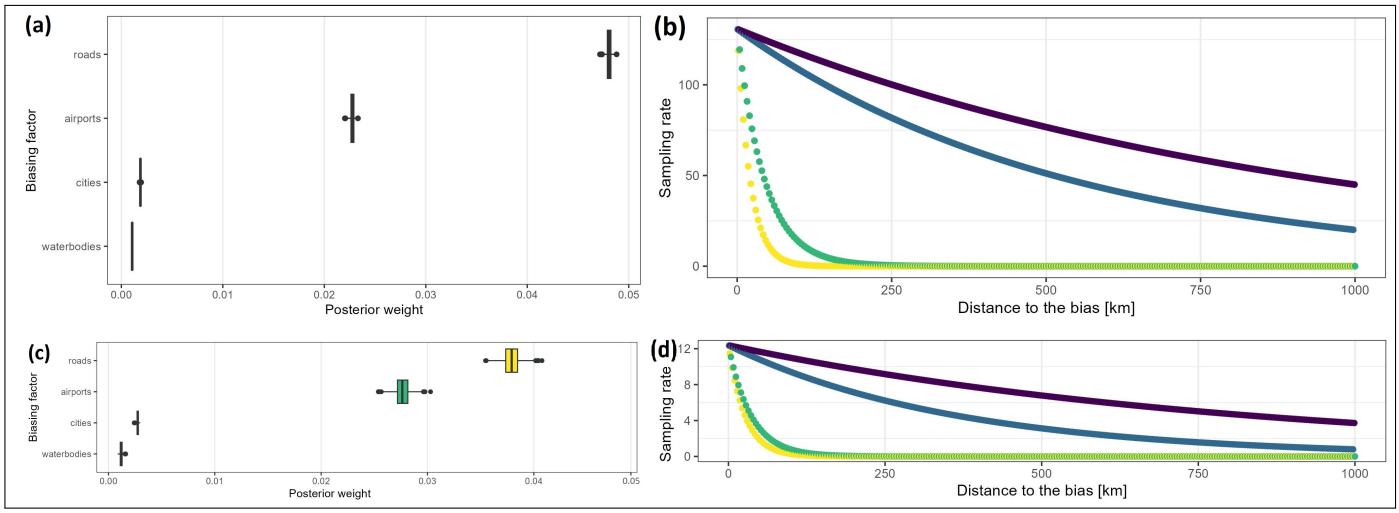

Figure 7: Results of the *calculate_bias* function (*Sampbias* package), estimating accessibility bias in animal occurrences across French Polynesia for marine (a,b) and terrestrial (c,d) species. (a,c) Bias weights of each bias factor (here airports section includes ports). (b,d) Sampling rate as a function of the shortest distance to each bias factor, roads (yellow), airports (green), cities (blue), and waterbodies (purple) (Occurrence expected number).

respectively. There was a positive correlation between terrestrial and marine richness. The composition of marine species is significantly different between the Marquesas archipelago and the others, while the terrestrial species composition was significantly different between the Tuamotu archipelago

and the others. The Society archipelago contains the highest number of occurrences and species for both marine and terrestrial habitats, whereas Tuamotu (especially the western part) and islands of the Gambier archipelago appear to be particularly under-surveyed. The strongest accessibility bias factors in both habitats were the roads, ports, and airports.

Quality The first step to develop a regional biogeographical inventory in French Polynesia is to have reliable, usable data. The raw data I downloaded from GBIF was free from several types of recurring commonly-reported errors (e.g., occurrences with inverted latitude and longitude coordinates or occurrences with 0° longitude and 0° latitude). However, the high proportion of true duplicates emphasizes the importance of data-quality checks before use (2).

Specific richness Comparing the biodiversity inventories is essential to assess the wildlife representativeness in an ecologic spatial database. 2327 marine molluscs are present in our database over 3022 that are referenced in the published atlas ([Marine Molluscs of French Polynesia by M. Boutet, R. Gourguet, J. Letourneux]). For arthropods 1990 species are present over the 3025 referenced in Ramage (2017).

Our results showed an uneven geographical distribution of marine and terrestrial species in French Polynesia. The Society archipelago stands out for its rich observed biodiversity, hosting the majority of marine and terrestrial occurrences and species. In contrast, some regions, such as the Tuamotu and Gambier archipelagos, are less well surveyed, which is mainly reflected in the limited number of terrestrial species and occurrences recorded.

Analysis of species richness reveals positive correlations between marine and terrestrial species across the five archipelagos. The Tuamotu and Gambier archipelagos show greater variability in species richness across islands, suggesting greater heterogeneity within these archipelagos, and in particular in the sampling effort.

As far as marine species are concerned, the Society archipelago stands out as a local hot-spot of biodiversity among other archipelagos, home to a large number of species, notably Teleostans and Gastropods. Species such as *Tridacna maxima*, *Acridotheres tristis* and *Megaptera novaeangliae* are abundant in the dataset, with significant contributions from research institutions, which reflects the growing interest of the scientific community in these species.

For terrestrial species, "Aves", "Insecta" and "Gastropoda" dominate the records. The species *Geopelia striata* stands out as the most frequently observed terrestrial species.

Multivariate community analysis A multivariate community analysis was performed to assess the similarities among archipelagos in French Polynesia for marine and terrestrial species assemblages. This analysis revealed crucial information about community structure and the factors influencing

their composition. The dissimilarity analysis based on Jaccard's index showed that both terrestrial and marine communities have high index values (0.94 and 0.96, respectively), suggesting strong dissimilarity between species communities in each habitat. The dissimilarity analysis revealed two main island clusters, based on their terrestrial communities. One cluster shows a marked similarity between certain Marquesas islands and Society Islands, such as Bora Bora and Ua Huka. The other group appears to link the Tuamotu islands, notably Tikehau and Takapoto.

For the marine data, three distinct clusters were observed, although relatively more dispersed than the terrestrial clusters. Some Tuamotu island groups are related, as are some southern geographic structures. The Tuamotu islands show a greater dispersion on the two main axes of the principal component analysis (PCoA) for both habitats.

Community analysis also revealed differences between the archipelagos. For marine data, the species community structure of the Marquesas Islands is significantly different from that of the other archipelagos. For terrestrial data, the islands of the Tuamotu archipelago are distinguished by a direction orthogonal to the main trend, mainly dictated by Marquesas and Society, along the PCo2 axis.

These results highlight the complexity of species community structure in French Polynesia, with observed similarities and differences between terrestrial and marine communities, as well as between different archipelagos. These variations suggest the influence of various biogeographical, ecological, and evolutionary factors on species distribution. Additionally, it's worth noting that the ease of physically accessing these specific zones has been widely recognized as a critical bias (Amano et al. (2016)). This acknowledgment underscores the importance of considering not only natural factors but also potential anthropogenic influences when studying species distributions in the region.

4.1 Sampling bias

To determine which islands are under-sampled and would require a greater effort, an estimated sampling rate, and an accessibility biases assessment were computed. It has shown that in both marine and terrestrial occurrences distribution, the roads had the strongest accessibility bias. The relative estimated sampling rate follows a positive correlation between terrestrial and marine data (figure 10), which means that the geographical structures with terrestrial and marine information have an equivalent relative proportion. Quantitatively, the estimated sampling rate for terrestrial data is significantly lower than for marine data. The sampling rate is based on the observed number of occurrences and needs sufficient and even data to capture spatial variations. If there are no occurrences in a cell, it is impossible to estimate the sampling rate for that cell. Similarly, if there are very few occurrences in a cell, the estimated sampling rate may be unreliable because it will be based on a small sample size. Thus, the fact that 59 geographical structures have no terrestrial information and

the small overall number of occurrences for the terrestrial dataset across such vast space partially explains the relative underestimation.

Assessing the significance of specific biases individually is a complex process that needs to understand the correlations between the different factors. The taxonomic bias linked to sampling bias relies on the homogeneity of taxonomic groups sampling, some taxonomic groups may be more easily sampled than others due to their size, behavior, or habitat preferences, which can lead to biases in the observed number of occurrences and the estimated sampling rate. This is why usually patterns of sampling efforts are studied on homogeneously sampled taxonomic groups (Sanchez-Fernandez 2021 patterns of sampling effort), thus, beyond assessing the taxonomical depth, to minimize this taxonomic bias, it is critical to know the distributions of all taxonomic groups.

Additionally to sharpen this tool for a sparsely distributed occurrences dataset would be to look for areas with a high relative concentration of species richness that are not affected by accessibility biases, e.g. Rapa for terrestrial species and Nuku Hiva for both marine and terrestrial (fig. 8).

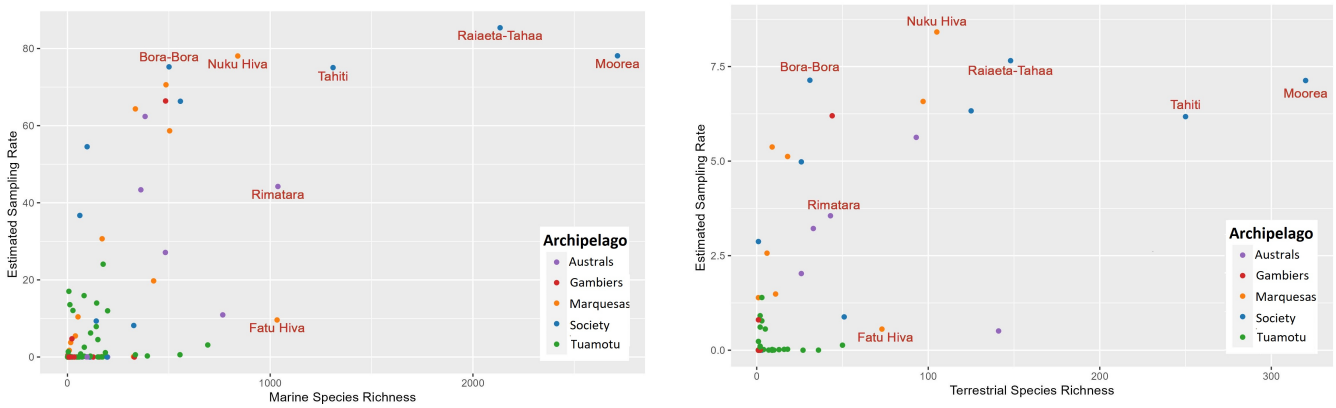

Figure 8: Estimated sampling rate function of species richness for (left) marine and (right) terrestrial data.

List of Figures

List of Tables

References

Amano, T., Lamming, J. D. L., and Sutherland, W. J. (2016). Spatial Gaps in Global Biodiversity Information and the Role of Citizen Science. *BioScience*, 66(5):393–400.

Andréfouët, S. and Adjeroud, M. (2019). Chapter 38 - french polynesia. In Sheppard, C., editor, *World Seas: an Environmental Evaluation (Second Edition)*, pages 827–854. Academic Press, second edition edition.

Baldacchino, G. (2006). Islands, island studies, island studies journal.

Bonnet-Lebrun, A.-S., Sweetlove, M., Griffiths, H. J., Sumner, M., Provoost, P., Raymond, B., Ropert-Coudert, Y., and 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. Publisher: Frontiers Media.

Ceballos, G. and Ehrlich, P. R. (2023). Mutilation of the tree of life via mass extinction of animal genera. *Proceedings of the National Academy of Sciences*, 120(39):e2306987120. Publisher: Proceedings of the National Academy of Sciences.

Delgado-Rodríguez, M. and Llorca, J. (2004). Bias. *Journal of Epidemiology & Community Health*, 58(8):635–641. Publisher: BMJ Publishing Group Ltd Section: Continuing professional education.

Díaz, S. M., Settele, J., Brondízio, E., Ngo, H., Guèze, M., Agard, J., Arneth, A., Balvanera, P., Brauman, K., Butchart, S., Chan, K. M. A., Garibaldi, L. A., Ichii, K., Liu, J., Subramanian, S., Midgley, G., Miloslavich, P., Molnár, Z., Obura, D., Pfaff, A., Polasky, S., Purvis, A., Razzaque, J., Reyers, B., Roy Chowdhury, R., Shin, Y.-J., Visseren-Hamakers, I., Willis, K., and Zayas, C. (2019). *The global assessment report on biodiversity and ecosystem services: Summary for policy makers*. Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Accepted: 2020-10-20T16:12:43Z.

Engemann, K., Enquist, B. J., Sandel, B., Boyle, B., Jørgensen, P. M., Morueta-Holme, N., Peet, R. K., Violle, C., and Svenning, J.-C. (2015). Limited sampling hampers "big data" estimation of species richness in a tropical biodiversity hotspot. *Ecology and Evolution*, 5(3):807–820. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.1405.

Farley, S. S., Dawson, A., Goring, S. J., and Williams, J. W. (2018). Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. *BioScience*, 68(8):563–576.

Frishkoff, L. O., Karp, D. S., Flanders, J. R., Zook, J., Hadly, E. A., Daily, G. C., and M'Gonigle, L. K. (2016). Climate change and habitat conversion favour the same species. *Ecology Letters*, 19(9):1081–1090.

García-Roselló, E., González-Dacosta, J., and Lobo, J. M. (2023). The biased distribution of existing information on biodiversity hinders its use in conservation, and we need an integrative approach to act urgently. *Biological Conservation*, 283:110118.

GBIF.Org User (2023). Occurrence Download.

Gillespie, R. G., Claridge, E. M., and Goodacre, S. L. (2008). Biogeography of the fauna of French Polynesia: diversification within and between a series of hot spot archipelagos. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1508):3335–3346. _eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2008.0124.

Gorman, C. E., Torsney, A., Gaughran, A., McKeon, C. M., Farrell, C. A., White, C., Donohue, I., Stout, J. C., and Buckley, Y. M. (2023). Reconciling climate action with the need for biodiversity protection, restoration and rehabilitation. *The Science of the Total Environment*, 857(Pt 1):159316.

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., and Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, 118(6):e2018093118. Publisher: Proceedings of the National Academy of Sciences.

Hembry, D. H., Raimundo, R. L. G., Newman, E. A., Atkinson, L., Guo, C., Guimarães Jr., P. R., and Gillespie, R. G. (2018). Does biological intimacy shape ecological network structure? A test using a brood pollination mutualism on continental and oceanic islands. *Journal of Animal Ecology*, 87(4):1160–1171. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1365-2656.12841.

Kadmon, R., Farber, O., and Danin, A. (2004). Effect of Roadside Bias on the Accuracy of Predictive Maps Produced by Bioclimatic Models. *Ecological Applications*, 14(2):401–413. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/02-5364.

Kays, R., McShea, W. J., and Wikelski, M. (2020). Born-digital biodiversity data: Millions and billions. *Diversity and Distributions*, 26(5):644–648.

Kulbicki, M. (2007). Biogeography of reef fishes of the French territories in the south pacific.

Levin, N., Coll, M., Fraschetti, S., Gal, G., Giakoumi, S., Göke, C., Heymans, J., Katsanevakis, S., Mazor, T., Öztürk, B., Rilov, G., Gajewski, J., Steenbeek, J., and Kark, S. (2014). Biodiversity data requirements for systematic conservation planning in the Mediterranean Sea. *Marine Ecology Progress Series*, 508:261–281.

Lin, H., Caley, M. J., and Sisson, S. A. (2022). Estimating global species richness using symbolic data meta-analysis. *Ecography*, 2022(3):e05617. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.05617.

Maldonado, C., Molina, C. I., Zizka, A., Persson, C., Taylor, C. M., Albán, J., Chilquillo, E., Rønsted, N., and Antonelli, A. (2015). Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? *Global Ecology and Biogeography*, 24(8):973–984. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/geb.12326.

Meyer, C., Kreft, H., Guralnick, R., and Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, 6(1):8221. Number: 1 Publisher: Nature Publishing Group.

Qian, H., Zhang, J., and Jiang, M.-C. (2022). Global patterns of fern species diversity: An evaluation of fern data in gbif. *Plant Diversity*, 44(2):135–140.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramage, T. (2017). Checklist of the terrestrial and freshwater arthropods of French Polynesia (Chelicerata; Myriapoda; Crustacea; Hexapoda). *Zoosystema*, 39(2):213.

Russell, J. C. and Kueffer, C. (2019). Island Biodiversity in the Anthropocene. *Annual Review of Environment and Resources*, 44(1):31–60. _eprint: https://doi.org/10.1146/annurev-environ-101718-033245.

Salvat, B. and Tröndlé, J. (2017). Biogéographie des mollusques marins de Polynésie française. *Revue d'Écologie*, 72(3):215–257. Publisher: Société nationale de protection de la nature (SNPN).

Sarzo, B., Martínez-Minaya, J., Pennino, M. G., Conesa, D., and Coll, M. (2023). Modelling seabirds biodiversity through Bayesian Spatial Beta regression models: A proxy to inform marine protected areas in the Mediterranean Sea. *Marine Environmental Research*, 185:105860.

Schiesari, L., Britta, G., and Grillitsch, H. (2007). Biogeographic Biases in Research and Their Consequences for Linking Amphibian Declines to Pollution. *Conservation biology : the journal of the Society for Conservation Biology*, 21:465–71.

Simberloff, D. (2000). Extinction-promeness of island species - causes and management implications. *THE RAFFLES BULLETIN OF ZOOLOGY 2000*.

Singh, J. S. (2002). The biodiversity crisis: A multifaceted review. *Current Science*, 82.

Smith, J. R., Letten, A. D., Ke, P.-J., Anderson, C. B., Hendershot, J. N., Dhami, M. K., Dlott, G. A., Grainger, T. N., Howard, M. E., Morrison, B. M. L., Routh, D., San Juan, P. A., Mooney, H. A., Mordecai, E. A., Crowther, T. W., and Daily, G. C. (2018). A global test of ecoregions. *Nature Ecology & Evolution*, 2(12):1889–1896. Number: 12 Publisher: Nature Publishing Group.

Takashina, N. and Kusumoto, B. (2023). A perspective on biodiversity data and applications for spatio-temporally robust spatial planning for area-based conservation. *Discover Sustainability*, 4(1):1.

Underwood, E., Taylor, K., and Tucker, G. (2018). The use of biodiversity data in spatial planning and impact assessment in Europe. *Research Ideas and Outcomes*, 4:e28045. Publisher: Pensoft Publishers.

Vieira, C., Steen, F., D'Hondt, S., Bafort, Q., Tyberghein, L., Fernández-García, C., Wysor, B., Tronholm, A., Mattio, L., Payri, C., Kawai, H., Saunders, G., Leliaert, F., Verbruggen, H., and Clerck, O. (2021). Global biogeography and diversification of a group of brown seaweeds (Phaeophyceae) driven by clade-specific evolutionary processes. *Journal of Biogeography*.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., and Vieglais, D. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLOS ONE*, 7(1):e29715. Publisher: Public Library of Science.

Zizka, A., Antonelli, A., and Silvestro, D. (2020a). sampbias , a method for quantifying geographic sampling biases in species distribution data. *Ecography*, 44.

Zizka, A., Azevedo, J., Leme, E., Neves, B., da Costa, A. F., Caceres, D., and Zizka, G. (2020b). Biogeography and conservation status of the pineapple family (Bromeliaceae). *Diversity and Distributions*, 26(2):183–195. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ddi.13004.

Zizka A, A. C. F., Calvente A, R. B.-L. M., Cabral A, C. J., M, C.-S., MR, F., MF, F., T, F.-A., F, G. L. M., Santos NMC, S. T., dos Santos-Costa RC, S. F., Alves da Silva AP, d. S. S. A., Cavalcante de Souza PG, C. T. E., and Vale VF, Vieira TL, A. A. (2020). No one-size-fits-all solution to clean GBIF [PeerJ].

Appendix

A List bio status

STATUS	DESCRIPTION
P	Present (native or uncertain)
E	Endemic
S	Subendemic
C	Cryptogenic
I	Introduced
J	Introduced invasive
M	Introduced non-established (including cultivated/domesticated)
B	Occasional
D	Doubtful
Q	Mentioned in error
A	Absent
W	Extinct
X	Extinct
Z	Endemic extinct
Y	Introduced extinct

Table 1: Biogeography status list (NHMN Provider)

P: Taxon present in a broad sense in the considered geographical area, meaning either a native taxon or a taxon whose status is uncertain. Lack of knowledge benefits native status. By native, we mean a taxon that originated from the considered geographical area and naturally developed there without human contribution, or a taxon that arrived there without human intervention (intentional or unintentional) from a zone where it is native. **E**: Taxon is naturally restricted to the considered geographical area.

S: Taxon naturally restricted to an area not entirely included in the considered geographical area, but with the main populations located in it (e.g., distribution spanning the French and Spanish Pyrenees, the French and Swiss Jura, Corsican-Sardinian endemism, etc.). For overseas regions, this status applies to regional endemism: for Guyana = endemic to the Guyana Shield.

C: Taxon whose original range is unknown, and therefore, it cannot be determined if it is native or introduced.

I: Taxon introduced (established or possibly established) in the considered geographical area. By introduced, we mean a taxon whose presence in the considered geographical area is the result of human intervention, intentional or unintentional, or a taxon that arrived in the area without human intervention but from a zone where it is introduced.

J: Taxon introduced into the considered geographical area, producing fertile descendants often in large numbers, and having the potential to exponentially expand over a large area, rapidly increasing its distribution range. This often leads to negative ecological, economic, or health consequences (IUCN, 2000). All taxa categorized as "introduced invasive," "exotic invasive," or "invasive" (invasive in English) in a scientific publication are grouped under this status.
M: Taxon introduced that can occasionally reproduce outside of its cultivation or captivity area but cannot maintain itself in the wild, as it cannot form viable populations without human intervention, and therefore depends on repeated introductions to persist in nature. All taxa categorized as "introduced occasional," "subspontaneous," and "escaped from cultivation or captivity" (in English: casual alien) are grouped under this status. This status includes strictly domestic taxa (fauna), cultivated taxa (flora), as well as commercially traded species (e.g., species for aquariums).
B: Occasional taxon, non-breeding, accidental in the considered geographical area (e.g., migratory passerby).
D: Taxon whose presence in the considered geographical area is not confirmed (pending confirmation).
Q: Taxon mentioned in error as present in the considered territory.
A: Taxon not present in the considered geographical area.
W: Taxon no longer present in the wild in the considered geographical area but not globally extinct. Note: in cases of doubt regarding the historical presence or absence of the taxon in the wild, use the absent status (A).
X: Taxon globally extinct (= completely disappeared from the surface of the Earth).
Z: Endemic taxon that is extinct today, and therefore, globally extinct (X).
Y: Taxon introduced in the past but now extinct in the considered geographical area (W) or globally extinct (X).

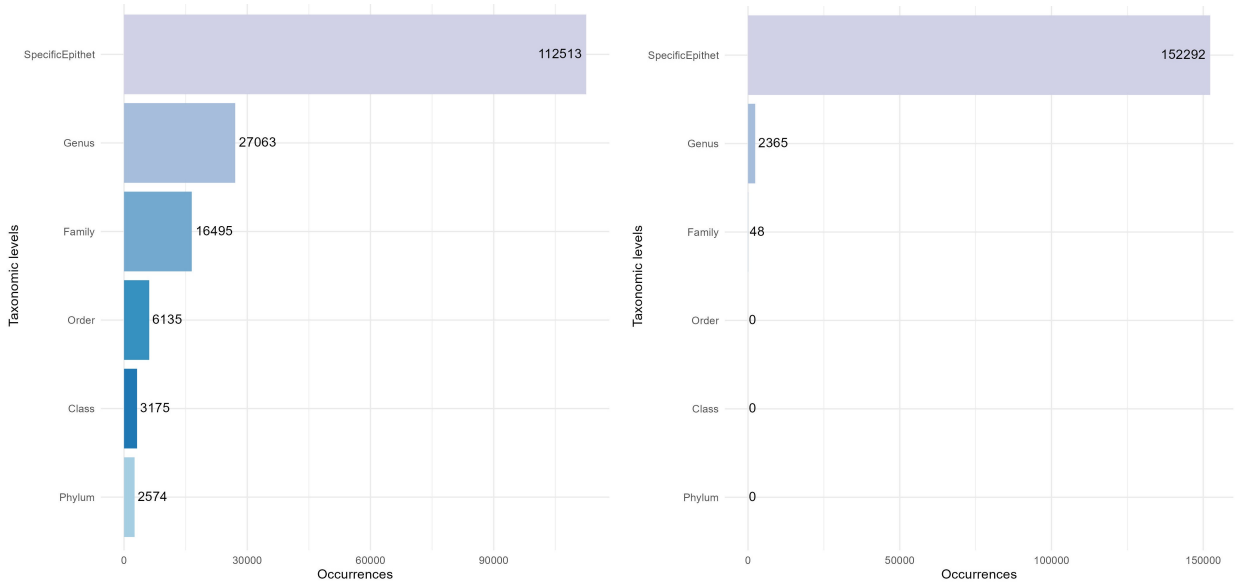

Figure 9: Taxonomic depth for raw (left) and cleaned (right) dataset, with 37.7% of data up to specific epithet and 97.2% respectively.

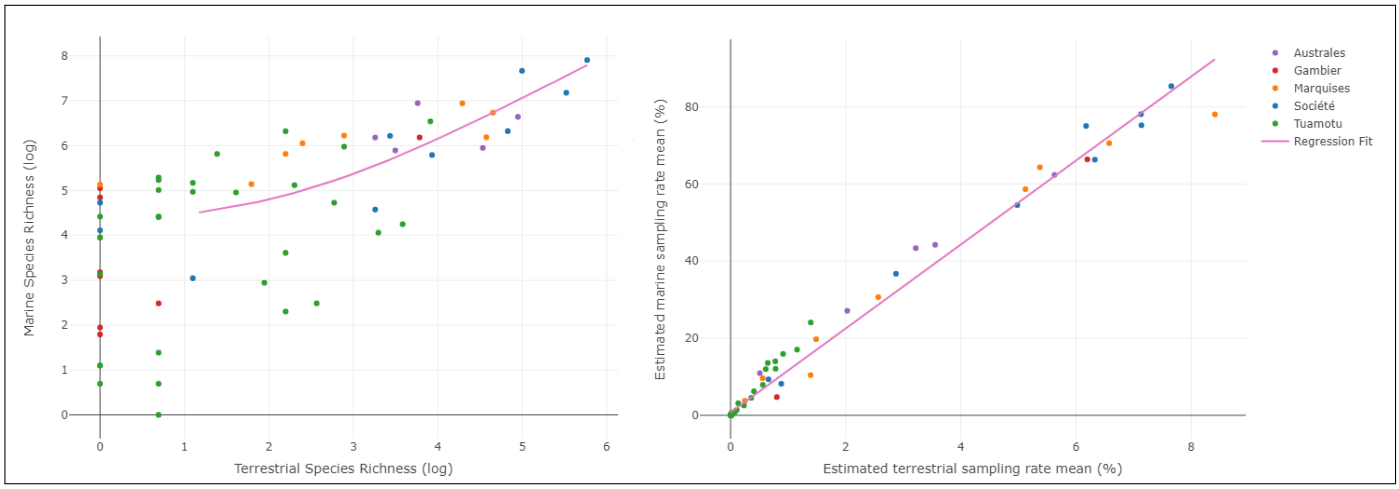

Figure 10: (Left)

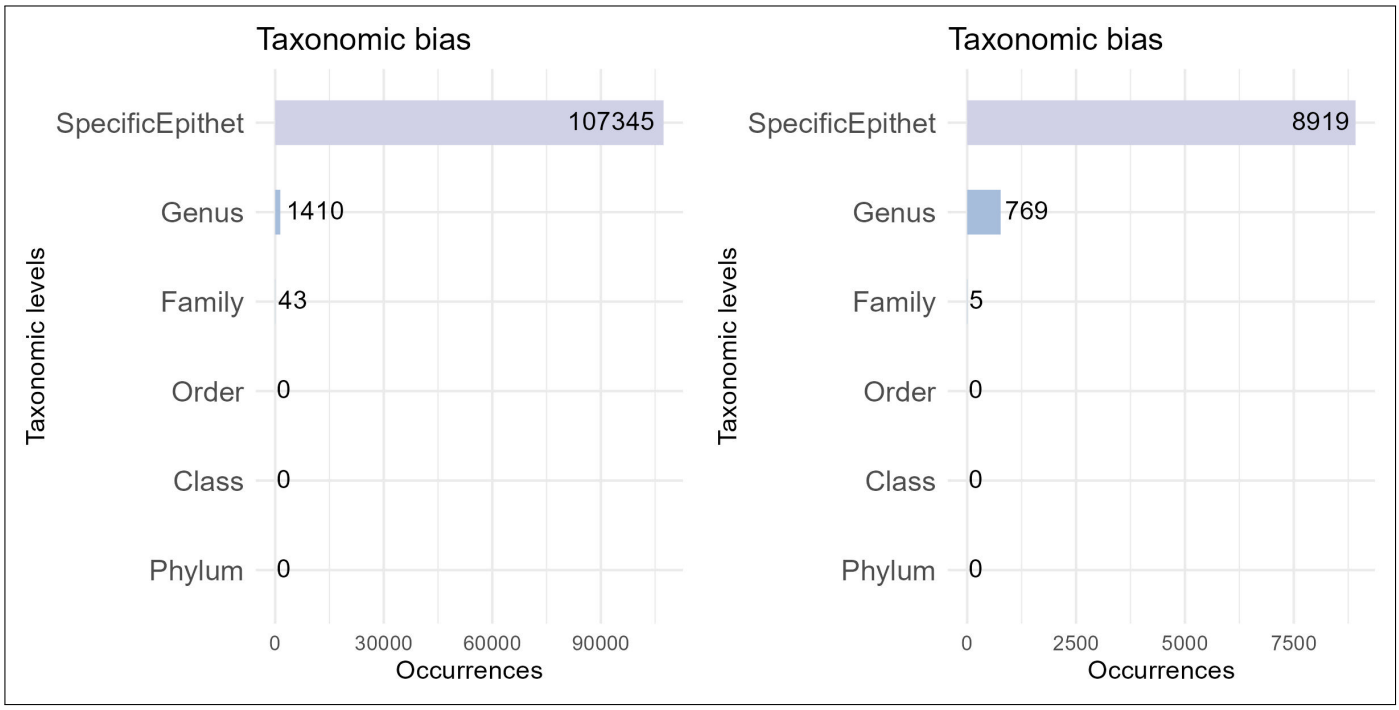

Figure 11: Number of marine (left) and terrestrial (right) occurrences up to the registered taxonomic level for marine and terrestrial data, the species level is named *ScientificEpithet* (e.g. 43 marine occurrences have their Taxonomic information filled from phylum to order).