Workshop on Visualisation in Environmental Sciences (EnvirVis) (2021)
S. Dutta and K. Feige and K. Rink and D. Zeckzer (Editors)

Assessing the Geographical Structure of Species Richness Data with Interactive Graphics

P. Morgades[†], A. Slingsby[‡] & J. Moat[§2]

Abstract

Understanding species richness is an important aspect of biodiversity studies and conservation planning, but varying collection effort often results in insufficient data to have a complete picture of species richness. Species accumulation curves can help assess collection completeness of species richness data, but these are usually considered by discrete area and do not consider the geographical structure of collection. We consider how these can be adapted to assess the geographical structure of species richness over geographical space. We design and implement two interactive visualisation approaches to help assess how species richness data varies over continuous geographical space. We propose these designs, critique them, report on the reactions of four ecologists and provide perspectives on their use for assessing geographical incompleteness in species richness.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation]: User Interfaces—User-centered design

1. Introduction

Understanding species richness is fundamentally important for biodiversity studies and conservation planning [DC04, Sto16]. There have been major efforts to collate species distribution data collected over the past decades, but these are of varying quality with different species targeted, changing nomenclature, differences in geographical and temporal precision, and varying collection effort [MMZ*15].

The Global Biodiversity Information Facility (GBIF; `https://www.gbif.org/`) is one of a number of initiatives that usefully collate fieldwork-based observations, structuring them in a common format, to help ecologists quantify species richness and biodiversity. For the reasons given above, bias and incompleteness are inherent in this patchwork of datasets [BBES14], but metadata can help determine which ones are more or less reliable for different purposes. However, the purpose of this short paper is to consider the geographical structure of species richness data, reporting on the outcome of a short MSc project by the first author.

There is already a large body of work and good practice that considers these issues in terms of data collection (e.g. incorporating species distribution models [Fra10]), quantification metrics and analysis techniques (e.g. [BBES14]), but many of these focus on habitat-, administrative region- and/or biozone-defined discrete areas, failing to capture other spatial or scale-dependent variation [San15, WWF01] that is important for species richness. We will focus on considering continuous space that considers both the scale and extent of collections. **Our first contribution is to consider ways to assess the geographical species richness in geographical space.**

The geographical structure of species richness has a number of aspects that are worth considering, including scale and extent. Interactive visualisation provides means for the exploratory analysis of data with multiple aspects [DMK05], particularly for spatial data which can be considered at different extents, scales [GDST15] and levels of aggregation [BDW*08]. **Our second contribution is to suggest interactive visualisation designs to help assess species richness data in geographical space.** We report on the reactions of four ecologists who tried these out.

2. Related work

Studies of species richness usually have a strong reliance on maps for indicating geographical variation and the impact of scale on species richness [KBB*11, RG01]. Interactive maps which enable scale and extent to be varied through interactions such as zooming/panning, interactively changing the level spatial aggregation/smoothing and filtering on different species and other characteristics are widespread for spatial phenomena (e.g. [BDW*08]) and for biodiversity (e.g. [SvL13]). We use these interactive approaches, including zooming and planning to select geographical extents for comparison.

† Department of Computer Science, City, University of London, UK
‡ Department of Computer Science, City, University of London, UK
§ Biodiversity Informatics and Spatial Analysis, Royal Botanic Gardens, Kew.

DOI: 10.2312/envirvis.20211085

Completeness of species richness can be assessed with a species accumulation curve representing the species richness [UGE03, R*95] (e.g. number of unique species) as a function of a measure of the sampling effort (e.g. number of observations). With sampling effort on the *x*-axis and species richness on the *y*-axis, if the resulting species accumulation curve flattens off (as in Fig. 2, bottom), we can identify the collection effort (*x*-axis) at which we appear to have a relatively complete picture of species richness (*y*-axis). Where curves do not flatten off (as in Fig. 2, top), we likely have an incomplete picture of species richness. Lobo *et al.* developed an application to visualise the survey effort of different regions at different scales, inferring level of completeness from species accumulation curves [LHY*18]. Edler *et al.* use an interactive map and a multi-scale approach to help identify biozones based on species distribution data [EGZ*17]. **We adapt species accumulation curves** by putting distance or scale on the *x*-axis. The shape of the curve then reflects the geographical heterogeneity of species. Where there are discontinuities in the curve's gradient, there is heterogeneity in the area, perhaps indicating that the area spans difference habitat types. Curves that do not flatten off may indicate that areas are incompletely collected areas (as for species accumulation curves), however this may also be caused by insufficient data samples.

Utteridge *et al.* use two species accumulation curves to compare the species richness and collection patterns for two different habitats of New Guinea [UdK07]. By comparing the shapes of the curves, they concluded that the assessment of species richness was better in one habitat compared to the other with higher species richness. **We use interactive graphics** to help compare likely completeness of species richness between areas, but also a gridded approach that facilitates comparison over a large area.

3. Case study area: New Guinea

This work uses New Guinea as a case study. The island holds a diverse range of habitats with the world's richest flora and is recognised as a biodiversity and conservation hot-spot [CLFA*20]. However, the island is far from being uniformly collected and large areas remain under-explored.

Species occurrence data is from the Global Biodiversity Information Facility (GBIF) from the past century which contains data from a range of data collections activities at different times and precisions. Each row contains information on one observed specimen. We only consider occurrences that are georeferenced, representing about 400,000 occurrences, including more than 35,000 different plant species.

The island of New Guinea comprises the comparatively well-collected Papua New Guinea and Western New Guinea, which is part of Indonesia. Fig. 1 indicates that the number of observations has a similar pattern to species richness (i.e. number of unique species). The question is: is this increase species richness simply because there has been greater collection effort there?

4. Designs and critique

Our prototype designs help make our two stated contributions: *(a)* to consider how to assess the spatial structure of species richness and completeness in geographical space and to *(b)* suggest interactive visualisation designs to help to this. We will use the term "well collected" to refer to apparent high completeness of the species richness picture, which is where the species accumulation curves flattens off, if indeed it does (assuming enough samples have been collected, as explained in section 2). Both our designs use species accumulation curves, but use spatial equivalents of "measurement effort", and take two approaches. Design 2 specifically addresses some limitations we identified with Design 1 as described below.

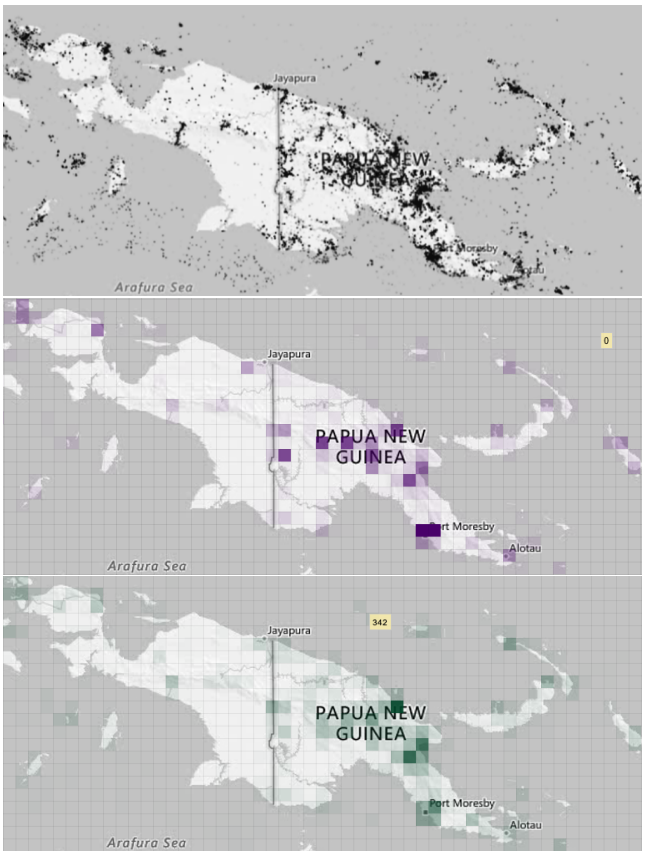

Figure 1: *New Guinea: Top: Observations. Middle: Density of observations. Bottom: Density of species. There are more observations and more species diversity for Papua New Guinea (Eastern half) that Western New Guinea, but how does this relate to collection effort?*

Note that for the purposes of this study, we are assuming that the species accumulation curve is a "well collected", however, as already noted, the inherent data bias and other issues mentioned earlier are likely to compromise this.

4.1. Design 1: Exploration and comparison of species richness for two locations

Design 1 was designed to *(a)* explore species richness and its variations across different scales, (b) estimate to which extent the perceived species richness is an accurate reflection of the reality

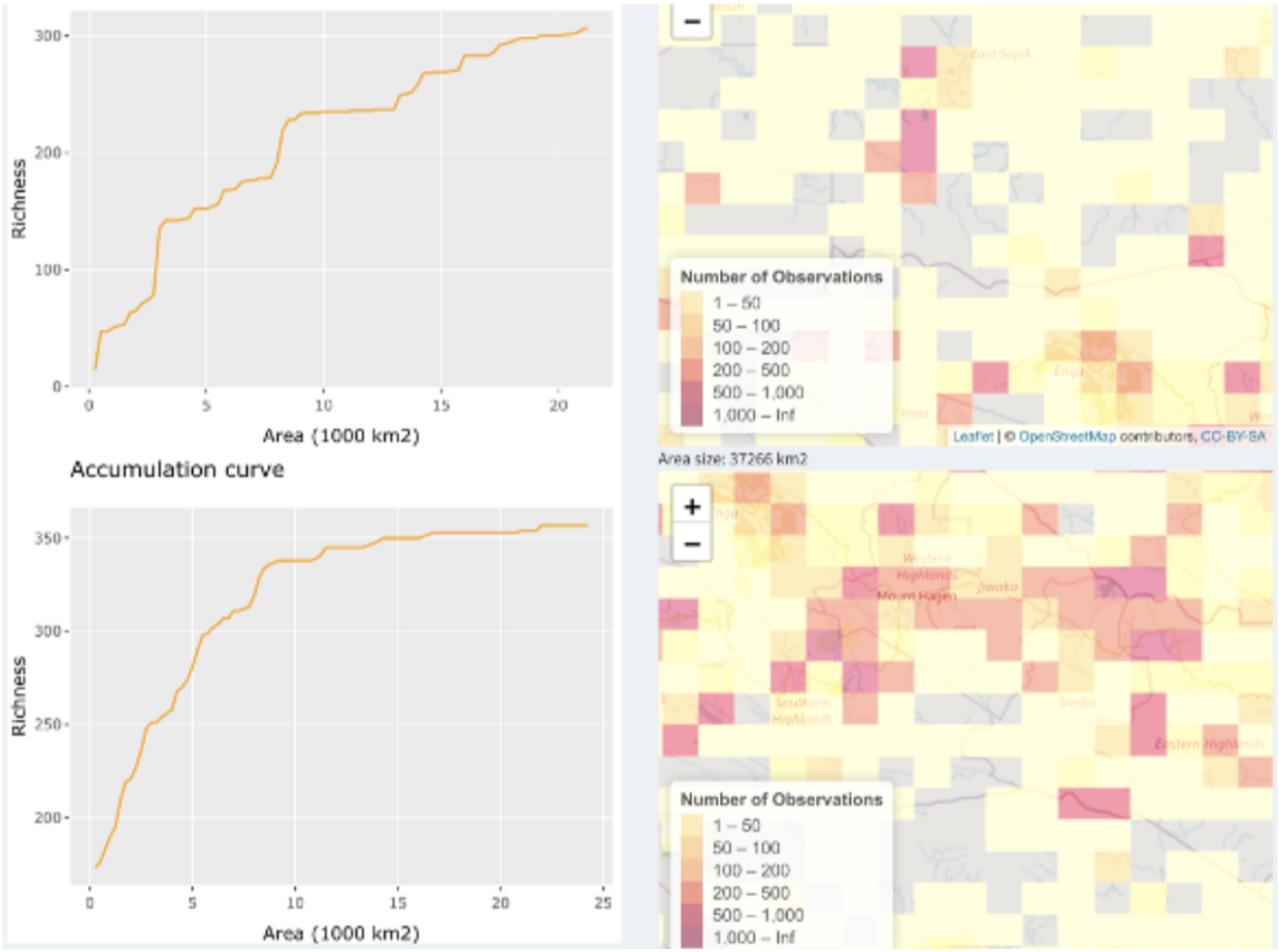

Figure 2: *Design 1 has a species accumulation curve that is constrained to the map area indicated to its right, with species accumulated from the centre outwards. Maps can be interactively zoomed and panned, resulting in the curve being updated dynamically. The lower curve (for the area indicated on the lower right) is more "well-collected" (at a 10 000km^2 collection area) because it flattens off. This is not the case for the upper curve (for the area indicated on the upper right).*

on the ground, and *(c)* enable its comparison to another area. It achieves these goals by facilitating comparison of species accumulation curves for two interactively-defined areas. The prototype was implemented as an RShiny application, using the Vegan, Tidyverse, Leaflet, Leafgl and Plotly libraries. Fig. 2 shows the prototype for Design 1. It has two accumulation curves with corresponding maps. Curves are automatically updated as map areas are interactively defined with zoom/pan interactions. They can be zoom/panned independently of each other. This means that they can be set to have the same or different areas (through panning) and/or scales (through zooming).

We use the circular area (1000km^2) that is centred on the map centre, for the *x*-axis as "measurement effort". Records are accumulated from the centre (thus the square of the distance from the map centre) helping estimate the collection level in different areas. To assist with defining appropriate map areas, the heatmap indicates density of observations.

A couple of design limitations became apparent for this design:

- It is sensitive to the specific location centred on the map area. Panning the map may result in quite a different curve, because it affects the way in which observations are accumulated. Additionally, the reliance on point locations makes broader comparisons across space more difficult to do
- It is limited to two locations and – as such – difficult to compare across space systematically.

4.2. Design 2: Comparison across geographical space

Fig. 3 shows Design 2, which helps address the limitations identified in Design 1. Visualisation can be effective for summarising data across multiple scales [GDST15] and this TileMap [Sli18]

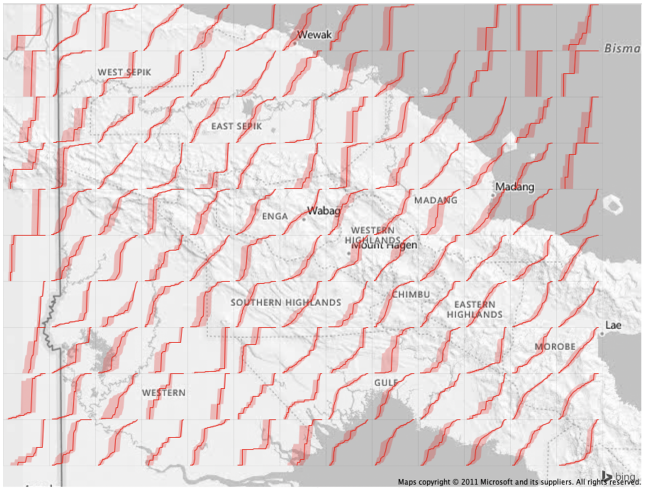

Figure 3: *Design 2 is a regular array of spatially-local species accumulation curves indicating variation in species richness at different scales and across geographical space.*

enables comparison of the required collection scale for well-collected data across a systematic set of multiple areas. It does this by tiling a set of local species accumulation curves on a map, in which we use scale on the *x*-axis as the "measurement effort". Thus these spatially-local species accumulation curves, indicate variation in species richness at **different scales** and **across geographical space**.

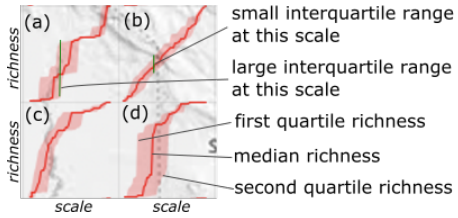

Figure 4: *These spatially local species accumulation curves show species richness as a function of* ***scale****, with the red area indicating the interquartile range of all the areas within the grid square.*

Fig. 4 illustrates how the species accumulation curves show variation in species richness across different geographical scales and across geographical space. These spatially-local curves (constrained for the land area defined by the grid cell boundaries) report the distribution of species richness at a variety of scales, from the scale represented by 1 pixel to the whole grid square. The red line indicates the median species richness. The shaded area indicated the interquartile range which represents heterogeneity in recorded species richness, converging to zero on the right as only one species richness measurement exists for the whole grid cell. There are updated in response to zoom/pan interactions; since the grid cells remain the same size in screen-space, it results in a new set of local curves for a different spatial extent (though panning) and spatial resolution (through zooming). This design enable comparison across scales (through *x*-axis of the curve) across space (through the regular grid sampling). The shape of the curves gives indication on the species richness and geographical completeness. In Fig. 3 a variety of shapes of curve, indicate different likely completeness of species richness scenarios, with area at the top-right in which there is more heterogeneity in species richness within the grid cells.

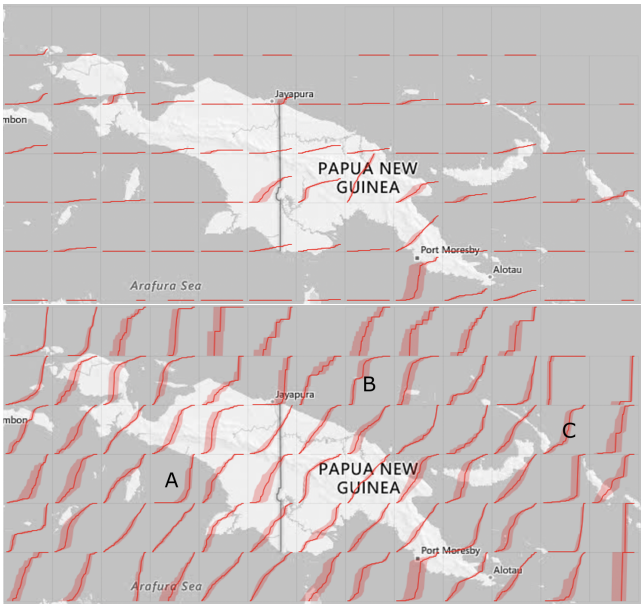

Figure 5: *Local species accumulation curves, with a common y-axis (top) and a locally-scaled y-axis (bottom), indicating large differences in species richness.*

Fig. 5 shows the effect of absolute and relative species-richness scaling; the former indicating species richness and the latter giving a better representation of the local species accumulation curves. Switching between both is useful for interpretation, but symbolising the absolute species richness on the relatively-scaled curves might be more effective.

Some observations in the upper map (Fig. 5; absolute species richness scaling):

- In the bottom row, a species accumulation curve with high species richness indicates "well collected" data at a scale of about half the grid-cell size scale, but has a lot of heterogeneity within, probably due to it being a grid cell containing both land and sea.
- Under the "PAPUA NEW GUINEA" label is a species accumulation curve, again, indicating high species richness, but is not "well collected" because it doesn't flatten off. This suggests the species richness is likely to be higher than recorded.

The lower map has some really interesting curves. Some observations (Fig. 5; relative species richness scaling):

- The curve at A has in unusual inverse shape and may indicate where species are very locally-specific and are only picked up when considered at larger scales.
- The curve at B is has a large discontinuity in it, perhaps indicating that it straddles different habitat types.

- The curve at C has an 'S-shape', indicating that it is well-collected at one scale, but poorly collected at larger scales. This may be because we have two very different habitats: marine and terrestrial.

Some of these may be vulnerable to low numbers of observations, but are worthy of more investigation.

Again, we critiqued our design:

- Where the y-axes of curves are scaled to the local (grid) maximum, although curves can be easily compared, it is difficult to identify those for which there is too little data to produce reliable results.
- The gridding is a rather arbitrary discretisation that does not take into account topographic features (including coastlines) and biozones. Differences in species richness that result from grid cells straddling very different habitats may be obscured, and the context to determine this is not provided (though the base-map will provide some clues).
- Although local curves are not so sensitive to location because they consider the whole distribution of all areas at the given through throughout the grid cell, they are still sensitive to the "Modifiable Areal Unit Problem (MAUP)" in which results are dependent on the grid aggregation [Ope81,Sli18,MBFB18]. Panning the map will result in changes the curves. The degree of this change will indicate the degree to which MAUP is responsible for the patterns depicted by the curves

5. Participant study

As demonstrated above, we have critiqued our designs in terms of data quality and the visual depiction of the data. However, we wanted to get some ecologists' views on our approaches. We asked four ecologist scientists from Kew to give their perspectives on these approaches and our designs. We set up one-to-one sessions with all four. Unfortunately, due to COVID restrictions, we were not able to do this in-person and due to associated unexpected technical problems, participants were unable to use the interactions themselves. Instead, we screen-shared and carried out the interactions on their behalf according to their instructions.

For each participant, we *(a)* demonstrated each design; *(b)* asked them to carry out a number of tasks related to our stated contributions (interactive visualisation for assessing species richness and completeness in a continuous geographical context); and *(c)* asked for their perspectives on its apparent success, suggestions of further things to consider, and potential uses of these techniques.

5.1. Suitability of the designs

Participants were asked a set of indicative questions that enabled them to use the designs for interactively assessing species richness and completeness in a continuous geographical context: *(a)* interpreting a curve at a specified location; *(b)* Interpreting comparison of curves for two specified areas; *(c)* identifying on well-collected vs non-well-collected areas and commenting on their reasons; *(d)* making general statements across geographical space. None of the participants had seen or used such interactive tools for helping make such assessments. Perhaps not surprisingly, they were able to carry out these tasks, with Design 1 being more suitable for tasks a-c and Design 2 being more suitable for task d. However, the greatest value in asking them to carry out these tasks was to get their reactions and suggestions.

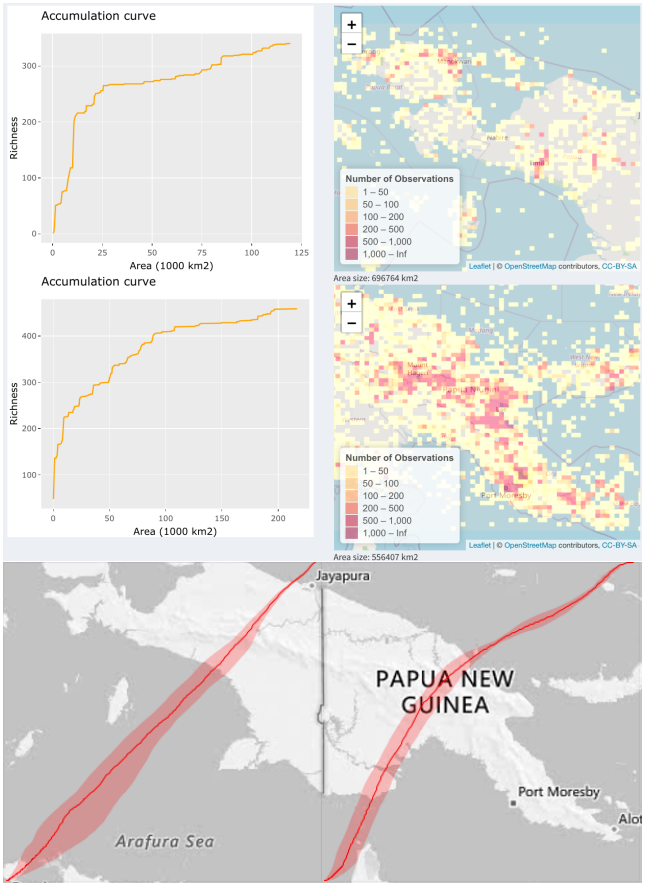

Figure 6: *Investigating whether the higher species richness in Papua New Guinea (East) than in Western New Guinea is due to more collection effort. Top: Design 1, indicating that both sides are relatively "well-collected". Bottom: Design 2, indicating that Western New Guinea is not "well-collected" but Papua New Guinea is slightly better collected.*

5.2. Additional contextual information that would be useful

In a follow-up questionnaire, participants suggested that habitat and land-use boundaries (as an indication of changes in the species assemblage), roads (as an indication of researcher accessibility), elevation (as an indication of conditions) and protected areas (as an indication of human disturbance) would be very useful contextual information for helping with interpretation.

Many of the issues originally mentioned about data quality were unsurprisingly parts of the discussion, with participants confirming that more information about the quality of data (spatial precision, when collected, etc) and the ability to filter on families and taxa would help assess to what degree observed patterns may be artefacts of data deficiencies.

5.3. Possible uses of these designs

In the same follow-up questionnaire, participants suggested possible uses in their work. Suggestions included: (a) using for planning field work and collecting trips; (b) helping understand problems and gaps in the collection; (c) exploring the species turnover between habitats, (d) exploring species loss in certain areas; and (e) helping define protected areas.

6. Papua New Guinea (East) vs Western New Guinea

To turn our attention to the question posed in Fig. 1, Design 1 in Fig. 6 (top) seems to indicate that both parts of New Guinea are reasonably "well-connected". However this if the maps are panned, the curves are very unstable, indicating a sensitivity to the location at the centre of the map. However, Design 2 in Fig. 6 (bottom) indicated that neither part is well collected across the scales, but Papua New Guinea is perhaps more so.

7. Reflections and conclusion

We consider and suggest approaches, visual encodings and interactions for helping assess spatial variation in collection completeness for species richness. We adapted species accumulation curves by using distance and scale measures of sampling effort, and incorporating these into two interactive designs that facilitate their construction and comparison. The interactively enabled geographical extents define and update species accumulation curves.

We critiqued our designs, with Design 2 addressing some of the issues in Design 1. Design 2 summarises the spatial structural of species richness more robustly; however some issues remain. Our observations in Fig. 5 and experiment in section 6 are initial investigations into how well these work; but this needs more work. Our small participant study suggested that both designs are capable of helping different aspects of assessing species richness and completeness in a continuous geographical context.

Feedback from participants has suggested additional geographical contextual information that could be usefully added. We also think that automated approaches for extracting insights from these species accumulation curves could be a useful addition. Participants also suggested a number of other possible uses, each of which could be the basis of a design study in future and interactive visualisation could be part of the solution. The suggested use for assessing helping assess data quality is a good use-case.

Although participants use species accumulation curves they had not used them in an interactive way before. The implementation of Design 1 as an RShiny application using standard open-source packages that are already used by ecologists, makes these methods accessible to the wider community and there is the opportunity to be involved in popularising their use. Design 2 was a custom-built prototype whose techniques are not widely accessible yet, but it demonstrates the possibilities for depicting multi-scale information across geographical space that likely also apply to other aspects of ecology.

References

[BBES14] BECK J., BÖLLER M., ERHARDT A., SCHWANGHART W.: Spatial bias in the gbif database and its effect on modeling species' geographic distributions. *Ecological Informatics 19* (2014), 10–15. 1

[BDW*08] BUTKIEWICZ T., DOU W., WARTELL Z., RIBARSKY W., CHANG R.: Multi-focused geospatial analysis using probes. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1165–1172. doi:10.1109/TVCG.2008.149. 1

[CLFA*20] CÁMARA-LERET R., FRODIN D. G., ADEMA F., ANDERSON C., APPELHANS M. S., ARGENT G., GUERRERO S. A., ASHTON P., BAKER W. J., BARFOD A. S., ET AL.: New guinea has the world's richest island flora. *Nature 584*, 7822 (2020), 579–583. 2

[DC04] DESMET P., COWLING R.: Using the species–area relationship to set baseline targets for conservation. *Ecology and Society 9*, 2 (2004). 1

[DMK05] DYKES J., MACEACHREN A. M., KRAAK M.-J.: *Exploring geovisualization*. Elsevier, 2005. 1

[EGZ*17] EDLER D., GUEDES T., ZIZKA A., ROSVALL M., ANTONELLI A.: Infomap bioregions: interactive mapping of biogeographical regions from species distributions. *Systematic biology 66*, 2 (2017), 197–204. 2

[Fra10] FRANKLIN J.: *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, 2010. 1

[GDST15] GOODWIN S., DYKES J., SLINGSBY A., TURKAY C.: Visualizing multiple variables across scale and geography. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 599–608. 1, 3

[KBB*11] KEIL P., BIESMEIJER J. C., BARENDREGT A., REEMER M., KUNIN W. E.: Biodiversity change is scale-dependent: an example from dutch and uk hoverflies (diptera, syrphidae). *Ecography 34*, 3 (2011), 392–401. 1

[LHY*18] 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.: Knowbr: An application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. *Ecological Indicators 91* (2018), 241–248. 2

[MBFB18] MOAT J., BACHMAN S. P., FIELD R., BOYD D. S.: Refining area of occupancy to address the modifiable areal unit problem in ecology and conservation. *Conservation biology 32*, 6 (2018), 1278–1289. 5

[MMZ*15] MALDONADO C., MOLINA C. I., ZIZKA A., PERSSON C., TAYLOR C. M., ALBÁN J., CHILQUILLO E., RØNSTED N., ANTONELLI A.: Estimating species diversity and distribution in the era of b ig d ata: to what extent can we trust public databases? *Global Ecology and Biogeography 24*, 8 (2015), 973–984. 1

[Ope81] OPENSHAW S.: The modifiable areal unit problem. *Quantitative geography: A British view* (1981), 60–69. 5

[R*95] ROSENZWEIG M. L., ET AL.: *Species diversity in space and time*. Cambridge University Press, 1995. 2

[RG01] RAHBEK C., GRAVES G. R.: Multiscale assessment of patterns of avian species richness. *Proceedings of the National Academy of Sciences 98*, 8 (2001), 4534–4539. 1

[San15] SANDEL B.: Towards a taxonomy of spatial scale-dependence. *Ecography 38*, 4 (2015), 358–369. 1

[Sli18] SLINGSBY A.: Tilemaps for summarising multivariate geographical variation. In *Summarization workshop at VIS 2018* (2018). URL: https://openaccess.city.ac.uk/id/eprint/20884/. 3, 5

[Sto16] STORCH D.: The theory of the nested species–area relationship: geometric foundations of biodiversity scaling. *Journal of Vegetation Science 27*, 5 (2016), 880–891. 1

[SvL13] SLINGSBY A., VAN LOON E.: Visual analytics for exploring changes in biodiversity. In *Workshop on Visualisation in Environmental Sciences (EnvirVis)* (2013). URL: https://openaccess.city.ac.uk/id/eprint/2385/. 1

[UdK07] UTTERIDGE T., DE KOK R.: Collecting strategies for large and taxonomically challenging taxa: Where do we go from here, and how often? *SYSTEMATICS ASSOCIATION SPECIAL VOLUME 72* (2007), 297. 2

[UGE03] UGLAND K. I., GRAY J. S., ELLINGSEN K. E.: The species–accumulation curve and estimation of species richness. *Journal of Animal Ecology 72*, 5 (2003), 888–897. 2

[WWF01] WHITTAKER R. J., WILLIS K. J., FIELD R.: Scale and species richness: towards a general, hierarchical theory of species diversity. *Journal of biogeography 28*, 4 (2001), 453–470. 1