Potential journals:

Ecological Applications (first target)

NO MORE COMMENTS! (will send through word version or next round of comments shortly)

Quantifying effort needed to estimate species richness from citizen science data

Corey T. Callaghan1, 2, Diana E. Bowler1, 3, 4, Shane Blowes1, 5, 6, Jonathan M. Chase1, 5, 6, Mitchell B. Lyons7, Henrique M. Pereira1, 2, 8

1 German Centre for Integrative Biodiversity Research (iDiv) Halle - Jena - Leipzig, Puschstraße 4, 04103 Leipzig, Germany

2 Institute of Biology, Martin Luther University Halle - Wittenberg, Halle (Saale), Germany

3 Institute of Biodiversity, Friedrich Schiller University Jena, Dornburger Straße 159, 07743 Jena, 4Germany Helmholtz Center for Environmental Research - UFZ, 4Department of Ecosystem Services, Permoserstraße 15, 04318 Leipzig, Germany

5XXXXX

6XXXXX

7Centre for Ecosystem Science, School of Biological, Earth and Environmental Sciences, UNSW Sydney, Sydney, NSW, Australia

8CIBIO (Research Centre in Biodiversity and Genetic Resources) – InBIO (Research Network in Biodiversity and Evolutionary Biology), Universidade do Porto, Vairão, Portugal

# Abstract

<https://www.youtube.com/watch?v=9Gc4QTqslN4&ab_channel=VDJMikeyMike>

# 

# Introduction

With the development of open science, biodiversity databases have enabled better estimates of the distribution and abundance of species and understanding of biodiversity patterns in time and space. For example, there are currently more than 1.6 billion species observations in the Global Biodiversity Information Facility (GBIF; <https://www.gbif.org/>) and the amount of data in GBIF has increased 12-fold since 2007 ([Heberling et al. 2021](https://www.pnas.org/content/118/6/e2018093118)). But despite this rapid growth of global biodiversity data, estimates of biodiversity in many parts of the world remain at best imprecise, and at worst non-existent (Stork 1993; Boakes et al. 2010; Scheffers et al. 2012; Essl et al. 2013; Cornwell et al. 2019). These geographic differences largely stem from variation in sampling effort. Monitoring biodiversity is increasingly important in the face of anthropogenic changes and associated ongoing biodiversity losses (Butchart et al. 2010; IPBES 2019) and plays a core role in indicator development in conservation decision-making, such as the COP’s Post 2020 global biodiversity framework. But biodiversity monitoring is expensive and time consuming, such that funding is generally insufficient ([Bakker et al. 2010](https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1755-263X.2010.00125.x)).

Future biodiversity monitoring will likely rely, at least in part, on broad-scale citizen science (also called community science) data (Bonney et al., 2009; Chandler et al., 2017; McKinley et al., 2017). Already, citizen science data are useful for a wide variety of applications, including monitoring marine protected areas ([Friewald et al. 2018](https://doi.org/10.1111/maec.12470)), developing conservation strategies for migratory species of birds ([Schuster et al. 2019](https://www.nature.com/articles/s41467-019-09723-8)), highlighting the negative impacts of roads on amphibians ([Cosentino et al. 2014](https://doi.org/10.1016/j.biocon.2014.09.027)), enabling continental-scale population trend estimation of birds ([Brlík et al. 2021](https://www.nature.com/articles/s41597-021-00804-2#auth-Vojt_ch-Brl_k)), rapidly mapping the effects of bushfires on biodiversity ([Kirchoff et al. 2020](https://doi.org/10.1016/j.scitotenv.2020.142348)), and understanding the effects of COVID-19 on biodiversity ([Vardi et al. 2021](https://doi.org/10.1016/j.biocon.2021.108953)). Clearly, the scope and extent of citizen science projects and generated data is rapidly increasing (Pocock et al., 2017). In 2020, for example, the web app iNaturalist ([www.inaturalist.org](http://www.inaturalist.org)), which allows citizen science observers to take pictures and upload georeferenced observations, averaged about 63,000 observations per day. As the quantity of data continues to increase, the potential of these data for biodiversity research may also increase.

Biodiversity-focused citizen science projects cover a broad range of biodiversity data collection activities ([Welvaert and Caley 2016](https://link.springer.com/article/10.1186/s40064-016-3583-5); [Kelling et al. 2019](https://doi.org/10.1093/biosci/biz010); [Isaac and Pocock 2015](https://doi.org/10.1111/bij.12532)). Generally, projects range from unstructured (e.g., no formal training and observations can take place wherever and whenever; iNaturalist) to semistructured (e.g., minimal training and directions, but important metadata are collected as part of the observation process; eBird) to structured (e.g., formally-trained, experienced, participants who conduct repeated surveys in time and space; Reef Life Survey). For those projects which share in common that they allow participants to select their own sites for submitting observations, sites are unsurprisingly not selected at random. There are often hotspots of activity, for instance around urban areas or within nature reserves ([Hugo and Altwegg 2017](https://doi.org/10.1002/ece3.3228); [Husby et al. 2021](https://doi.org/10.1111/ibi.12896)). As the number of participants has grown, there is increased potential for redundancy of effort as new participants sample the same set of sites already being sampled by others ([Callaghan et al. 2019a](https://doi.org/10.1371/journal.pbio.3000357)). There is a need to better understand the data collection patterns of citizen science projects, allowing for the development of guidelines for how biodiversity sampling might be improved for estimation of both local and regional biodiversity metrics ([Callaghan et al. 2019b](https://doi.org/10.1098/rspb.2019.1487); [Callaghan et al. 2021](https://doi.org/10.1016/j.biocon.2020.108912)). Importantly, however, the most valuable biodiversity observations will vary depending on the biodiversity metric of interest.

Species richness plays an important role in conservation spatial planning, for instance in identifying priority areas for protection with conservation planning decisions made across spatial scales ranging from local (e.g., patch-specific) to landscape (e.g., municipality) levels ([Boyd et al. 2008](https://conbio.onlinelibrary.wiley.com/doi/full/10.1111/j.1755-263X.2008.00002.x)). Species richness is also intuitive, easily calculated, and probably the most commonly used metric in studies of biodiversity patterns ([Gotelli and Colwell 2001](https://onlinelibrary.wiley.com/doi/full/10.1046/j.1461-0248.2001.00230.x); [Magurran and McGill 2011](https://books.google.de/books/about/Biological_Diversity.html?id=7mwVDAAAQBAJ&redir_esc=y)). Moreover, there are a number of techniques that allow for the extrapolation and comparison of species richness (e.g., [Colwell and Coddington 1994](https://doi.org/10.1098/rstb.1994.0091); [Chao et al. 2009](https://doi.org/10.1890/07-2147.1); [McGlinn et al. 2019](https://doi.org/10.1111/2041-210X.13102)). Therefore, despite potential drawbacks ([Fleishman et al. 2006](https://doi.org/10.1016/j.ecolind.2005.07.005); [Hillebrand et al. 2017](https://doi.org/10.1111/1365-2664.12959)), species richness remains a useful metric that is possible to estimate with citizen science data. Indeed, species richness measured from citizen science data has been shown to be comparable with professionally-collected data ([Van der Wal et al. 2015](https://link.springer.com/article/10.1007/s13280-015-0709-x); [Roman et al. 2017](https://doi.org/10.1016/j.ufug.2017.02.001)). But in order to (a) use citizen science data to estimate species richness and (b) encourage more informed sampling of species richness by citizen science participants, the spatial variation in the sampling effort of participants and how this variation affects the ability to make local estimates of species richness needs to be quantified.

Here, we develop a generalized workflow to estimate the optimal distribution of sampling effort for inference of species richness patterns. This workflow can be continuously updated as data are collected, and thus can be used to guide the site selection of participants to improve data collection. To achieve this goal, we had three specific objectives: (1) test relationships between species richness, landscape attributes (e.g., habitat heterogeneity, land cover), and citizen science sampling effort; (2) quantify the relationship between sample completeness (i.e., a statistical estimate of the likelihood a new species would be found if sampling continued; [Chao and Jost 2012](https://doi.org/10.1890/11-1952.1)) and landscape attributes; and (3) predict ideal effort (i.e., number of citizen science samples) needed to infer species richness. Because estimates of species richness are themselves scale-dependent ([Chase et al. 2019](https://doi.org/10.1111/oik.05968)) we examined how our results vary with spatial grain, testing the ability of citizen science data to assess species richness at different spatial grains. Together, these analyses can provide a useful template for optimizing biodiversity sampling across heterogeneous landscapes.

# Methods

## General overview of methods

Our methodological framework can be broken down into three parts, aligning to the objectives mentioned above: (1) test relationships between species richness, landscape attributes (e.g., habitat heterogeneity, land cover), and citizen science sampling effort; (2) quantify the relationship between sample completeness and landscape attributes; and (3) predict ideal effort needed to infer species richness. Using this general approach, we predicted effort needed at a site even when there was currently no citizen science data, using the predictive power of the land-use data.

For objective 1, we first explore the spatial patterns in the sampling effort of citizen science participants. Sampling effort was quantified as the number of submitted checklists. We used piecewise structural equation models (SEMs) to test the effects of land-use on real-world effort at a given site and how these effects were mediated by species richness.

For objective 2, we use rarefaction (Chao et al. 2010) to quantify sample completeness at a site and then use multiple regression to understand the relationship between sample completeness and landscape attributes. Sampling completeness represents a statistical estimate of the likelihood a new species would be found if sampling continued at that site. At a site, we calculated two types of sampling completeness: (1) standardized sampling completeness achieved by bootstrapping random subsamples of samples at a site, thereby standardizing sampling effort among sites, and (2) total sampling completeness achieved by using all samples at a site. In addition, we used two types of the sampling completeness function (Chao et al. 2020): (1) where common species are the focus of species richness estimation and (2) where rare species are disproportionately more important in species richness estimation.

For objective 3, we use the relationships between land-use, sampling completeness, and the number of samples to predict the effort needed to infer species richness across the whole landscape. Since our aim here was prediction, we used random forest models but we note that other forms of models could be used in this step. We also studied how effort needed varied in relation to land-use.

To illustrate this general framework, we used data from eBird across the U.S. state of Florida. Background on eBird data, followed by specifics of how data and analyses were conducted for each objective are treated in turn below.

## eBird citizen science data

We used data from eBird because it is one of the most successful citizen science projects to date with >900 million global observations since 2002 (Sullivan et al. 2014; 2017). eBird is a semistructured project where volunteer birdwatchers submit their observations in the form of ‘checklists’, and indicate whether or not they recorded every species they were able to identify. In addition, useful information regarding sampling effort, such as the distance travelled, the duration of observation, and the spatiotemporal coordinates are recorded. Regional filters are set on suspect bird observations whereby if an observer sees a species or records a count of species outside of those filters, then the data are thoroughly reviewed by regional reviewers before being added to the database (Gilfedder et al. 2018).

As a test region, we used eBird data from peninsular Florida in the United States (Bird Conservation Region 31). This region is relatively well-sampled, but includes very diverse habitats, including a large urban area and large swaths of protected areas mixed with agriculture, forested, and wetland areas. We downloaded the eBird basic dataset (version eBd\_May\_2020) and subsetted the data to include all of 2019 (between January 1st, 2019 and December 31st, 2019). Because the distance and time spent on a given eBird checklist can drastically vary (Kelling et al. 2015), we further subset the checklists by: (1) including only checklists which were complete; (2) including checklists which were <120 minutes and >15 minutes in duration; and, (3) including checklists that travelled <1 km or 150 Ha. Although we focus on 2019 because it had the greatest amount of data, we performed analyses for years 2014-2018 and found qualitatively and quantitatively similar patterns.

Because sites are not predefined in eBird, and different locations can correspond to drastically different sized habitat patches, we aggregated eBird checklists into grids. We did this for grids with six different spatial resolutions corresponding with 5, 10, 15, 20, 25, and 30 km2 grain sizes. Only sites (i.e., grids) that had a minimum of 25 eBird checklists within a given grid were considered for further analyses.

## Landscape habitat attributes

At each site, across our different grain sizes, we quantified four landscape attributes: habitat heterogeneity, percent cover of urban habitat, percent cover of trees, and percent cover of water. To calculate the habitat metrics, we use the Copernicus Global Land Cover Layers ([Buchhorn et al. 2020](https://research.wur.nl/en/publications/copernicus-global-land-cover-layers-collection-2)), which is a 100m pixel resolution product that maps discrete land cover classes, as well as fractional cover of major structural types (bare ground, crops, grass, shrub, tree, urban, permanent water, seasonal water). This provided sufficient detail for our case study site, and additionally allows for transferability to other regions in future studies. Habitat heterogeneity was calculated as the variance of the discrete land cover classes within the 5, 10, 15, 20, 25, and 30 km2 grain size buffers, and the fractional cover of the structural types were calculated as the mean within the buffers. All processing for these metrics was done using Google Earth Engine - a publicly available earth observation platform (Gorelick et al. 2017).

## Understanding patterns of real-world effort

We used the total observed species richness, landscape habitat attributes, and sampling effort (i.e., number of eBird checklists) in a site to quantify the direct and indirect relationships between these variables in a structural equation modelling framework (SEM). Our SEM consisted of three a priori linear models with a Gaussian error distribution: (1) log-transformed number of eBird checklists as the response variable and observed species richness and as predictor variables; (2) log-transformed number of eBird checklists as the response variable and observed species richness, urban cover, and habitat heterogeneity as predictor variables; and (3) observed species richness as the response variable and habitat heterogeneity, urban cover, water cover, and tree cover as predictor variables. The SEM was fitted using the R package piecewiseSEM v2.1.2 ([Lefcheck 2016](https://doi.org/10.1111/2041-210X.12512)) and we present the results for the standardized coefficient estimates representing the strength of the various relationships.

## Relationship between sample completeness and landscape attributes

To quantify sample completeness, we used the iNEXT R package ([Chao et al. 2014](https://doi.org/10.1890/13-0133.1); [Hsieh et al. 2016](https://doi.org/10.1111/2041-210X.12613); [Chao et al. 2020](https://doi.org/10.1111/1440-1703.12102)). As the majority of citizen science data documents the presence (and sometimes absence) of species (e.g., iNaturalist), we calculated sample completeness for incidence data ([Chao et al. 2020](https://doi.org/10.1111/1440-1703.12102)), by converting eBird data to presence/absence data (though we note that a similar approach could be taken using abundance data; Chao and Jost 2012). We focused on two points along the Hill number continuum: q=0 (i.e., species richness) and q=2 (i.e., Simpson’s index). When q=0, sample completeness is the ratio of observed species richness to true species richness, and most sensitive to rare species. Whereas when q=2, sample completeness is disproportionately sensitive to species that occur in high frequencies, or common species ([Chao et al. 2020](https://doi.org/10.1111/1440-1703.12102)). Accordingly, throughout our results we refer to our use of q=0 and q=2 as “Rare species sensitive”, and “Common species sensitive”, respectively.

Similar to species richness estimates, estimates of sample completeness are sensitive to sampling effort. Therefore, to quantify the relationship between sample completeness and landscape characteristics, we standardised the number of checklists per site in our sample completeness calculation. To standardise sampling effort across sites (i.e., grid cells), an equal number of checklists (5, 10, 15, and 20) were resampled (N = 50) randomly without replacement from each site, and the sampling completeness was then estimated as described above and averaged across the resamples. The analysis was robust to the number of checklists chosen each time (Figure S1), and thus, we only present the results using 10 randomly chosen checklists for each sample. This bootstrapped measure of completeness indirectly represents the size of the species pool, where lower values indicate a larger species pool and lead to, on average, lower levels of sampling completeness. Conversely, higher values indirectly indicate a smaller species pool, leading to an increased probability of high sampling completeness.

We used linear models with a gaussian distribution to assess the relationship between sampling completeness and landscape habitat predictors, where the mean sampling completeness was the response variable and the predictor variables were the landscape attributes (i.e., habitat heterogeneity, tree cover, water cover, and urban cover).

## Predicting effort needed to infer species richness

To predict how many checklists are required to achieve a target completeness, we used a random forest model, where the response variable was the total number of observed eBird checklists, log10-transformed, and the predictor variables were the total completeness estimate, habitat heterogeneity, tree cover, water cover, and urban cover. For this objective, we used the total observed sampling completeness, repeating the analysis mentioned above, but calculating only one measure of sampling completeness where all available eBird checklists were included for each site. This differs from the above objective that used a bootstrapping approach to standardize sampling completeness (Figure S2).

We chose a random forest model because our aim here was prediction rather than in understanding the effects of the covariates. Our goal was to make sites comparable for the purpose of comparing species richness in space, and not necessarily estimate the total diversity (Chao et al. 2020) and therefore we predicted the number of eBird checklists necessary to reach a completeness of 0.95 using our fitted random forest models. Exploratory analyses showed that there were marginal differences between completeness values of 0.95, 0.9, 0.85, and 0.8. This analysis was repeated twice, once for completeness sensitive to rare species (q=0) and once for completeness sensitive to common species (q=2). To assess the predictive performance of our random forest models, we report the mean pseudo R2 value, across all decision trees, extracted from the model fit. This pseudo R2 value represents the mean squared error divided by the variance of our response variable. Random forest analysis was done using the R package randomForest v4.6-14 (Liaw and Wiener 2002). Models were fit without replacement and using the default settings, including 500 trees and a node size of 5.

## Data analysis and availability

All data analysis was conducted in R statistical software and relied heavily on the tidyverse ([Wickham et al. 2019](https://joss.theoj.org/papers/10.21105/joss.01686)). Statistical significance, in the case of multiple linear regressions and SEMs, was concluded at alpha <0.05. Code and data to reproduce these analyses are currently available here (<https://github.com/coreytcallaghan/spatial_cs_sampling>) and will be permanently archived in a Zenodo repository upon acceptance of this article.

# Results

We used a total of 138,703 eBird checklists comprising 450 species’ observations for the year 2019 throughout peninsular Florida. The number of sites (i.e., grids) included in analysis was 933, 537, 343, 243, 175, and 132 for 5 to 30 km2 grain sizes, respectively.

Our structural equation modelling showed the strong influence of habitat attributes in predicting species richness, with urban cover the strongest supported, followed by habitat heterogeneity, tree cover, and water cover (Figure 1; Figure S3). There was also some support that the number of checklists at a site was predicted by the percentage of urban cover and habitat heterogeneity, suggesting that these two traits influence where people submit eBird checklists. These patterns showed some variation with grain size (Figure S4).

Standardised sampling completeness was negatively correlated with habitat heterogeneity, water cover, and urban cover across grain sizes, indicating consistent patterns. In other words, completeness tended to be higher at sites with lower heterogeneity, water cover, and urban cover. However, at the smaller grain sizes, tree cover was negatively correlated with bootstrapped sampling completeness, but the strength of this relationship weakened as grain size increased (Figure 2). Among these predictors, habitat heterogeneity and urban cover were consistently the strongest predictors of standardised sampling completeness for both common and rare species sensitive analyses, and across grain sizes. However, urban cover seemed to be a stronger predictor for rare species sensitive analyses than common species sensitive analyses (Figure S5).

We found that on average, the number of checklists necessary to meet 95% sample completeness when calculating completeness at q=2 (common species sensitive) was 43, 64, 96, 123, 172, and 176 for 5, 10, 15, 20, 25, and 30 km2 grain sizes, respectively. But when calculating completeness at q=0 (rare species sensitive), the average number of eBird checklists was 129, 214, 362, 498, 712, and 946 at 5, 10, 15, 20, 25, and 30 km2 grain sizes. (Figure 3A). The variability in these average estimates was greater when calculating sampling completeness at q=0 (Figure 3; Table S1). And the relationship between grain size and mean number of samples necessary to meet 95% sample completeness increased much quicker for rare species compared with common species, where it appeared to level off at larger grain sizes (Figure 3B).

Our models to predict the number of checklists necessary to sample in space performed relatively strongly (mean R2 = 0.83; range= 0.79 - 0.86; Table S2) for common species sensitive, but less well for rare species (mean R2 = 0.39; range=0.07 - 0.58; Table S2). Our unsampled sites generally had a narrower distribution of values than our sampled sites, where our sampled sites covered some of the sites that needed the most number of checklists to meet 95% completeness, suggesting that data contributed to eBird are already originating from the most diverse sites (Figure S6). And we found that there was a strong relationship between the predicted number of eBird checklists and the predicted species richness, confirming that our separate random forest models correlated well, as expected. We found that the highest number of necessary sampling was generally along the coast, and the lowest number was inland in large homogenous areas such as throughout the Florida Everglades (Figure 4).

# Discussion

We used more than 100,000 citizen science samples to quantify the relationships between sampling effort, species richness, and sampling completeness. We found that species richness is strongly modified by land cover, and that sampling completeness of richness is strongly tied to sampling effort (Figure 2), . Moreover, we show how it is possible to predict the sampling effort needed to quantify species richness in space. When the species richness estimation is focused on common species (i.e., q=2), a relatively small number of samples, on average, are needed to meet 95% sampling completeness, with an average of 44 samples in a 5 km2 grid and 203 samples in a 30 km2 grid. But if the species richness estimation is focused on rare species (i.e., q=0), substantially more samples are needed, from 322% more to 530% more samples at 5 km2 and 30 km2 grids, respectively. Ultimately, our results highlight the potential of citizen science data to monitor species richness and provide a framework for future adaptive sampling by citizen science participants.

(optimising design)

Biodiversity monitoring frequently relies on stratified sampling designs, where samples are spread across the landscape in an ‘equal’ fashion ([Bibby 2004](https://books.google.de/books?hl=en&lr=&id=wYk6AwAAQBAJ&oi=fnd&pg=PA1&dq=stratified+sampling+design+biodiversity+birds&ots=ouk5yfIk4S&sig=0kcrhiB1ObOebVfFfkA7XIEtA_c&redir_esc=y#v=onepage&q=stratified%20sampling%20design%20biodiversity%20birds&f=false)), and sometimes dependent on the size of a region of interest with larger geographic regions receiving proportionately more sampling effort ([Pavlacky et al. 2017](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0185924); [Wilgenburg et al. 2020](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0234494)). Although funding and accessibility (REF) may be an issue for professionally designed biodiversity monitoring schemes, biodiversity is not spread equally throughout the landscape (REF). Indeed, our results showed that species richness varies considerably in space (Figure 4C) and consequently, the effort required to sample species richness in space should match that of the species diversity (Figure 4). Importantly, however, we found that participants are already pre-selecting sites with high diversity and high habitat heterogeneity. However, sample completeness was still lower in the most species-rich sites? This suggests that the pattern of citizen science sampling is already in a direction to best sample richness. Repeat sampling of citizen scientists at the same sites is not necessarily redundant sampling if these sites are the most species richness.

(comparison to other monitoring design)

Monitoring schemes often differ in terms of whether they target common/widespread species, or rather specific groups of rare species. The challenge is often aligning these efforts for whole community sampling. Previous studies show how the optimal strategy varies among these species etc… See papers in the Github - random sampling for common species and conditional sampling for rare species Specht et al. These findings are consistent with our results since more spatially uniform patterns of effort were effective for a metric that was most sensitive to common species, while more heterogeneity effort was needed for that more sensitive to rare species.

Our analysis demonstrates the unequal sampling that takes place across the landscape, either because of biases in site selection, but also potentially because birders are pre-selecting sites with the highest biodiversity. By relying on citizen science data, the gaps and redundancies in biodiversity sampling ([Boakes et al. 2010](https://doi.org/10.1371/journal.pbio.1000385); [Courter et al. 2013](https://link.springer.com/article/10.1007/s00484-012-0598-7)) can potentially be minimized. A future goal of citizen science projects should be to encourage sampling, for example through gamification (Callaghan et al. 2019), in locations where the samples are most valuable for a given statistical objective - e.g., species richness estimation. For an illustrative example, using the 10 km2 grain size (N=1113 grid cells) and focusing on species richness estimation with a focus on common species, on average, a given grid cell needs 68 samples to meet 95% completeness. And with a total of 138,703 eBird checklists submitted throughout 2019, each grid cell could receive 125 eBird samples in a year, with current effort. Yet only 537 grids had at least 25 eBird samples, illustrating the unequal effort in space, but also highlighting the potential of citizen science data if this effort was spread throughout the landscape. By understanding and quantifying the necessary amount of sampling, we can begin to reduce redundancies in the data, for example, by ‘directing’ participants. Of course, implementation of directing participants will necessarily involve a social science approach that should inform the best strategies to interact with and engage with the participants of a project (REF - Pocock probably?). Nevertheless, some participants’ motivations involve ‘contributing’ to conservation, suggesting that a subset of users would be willing to alter their sampling efforts (REFS) to inform a specific goal.

Most conservation practise takes place at relatively small spatial scales, but citizen science is often seen as a coarse source of information on species occurrences. There is a need to understand the limits of citizen science for practical conservation decision-making.

Beyond our general patterns of citizen science effort needed to quantify species richness, we found significant differences among grain sizes and depending on whether species richness estimation was weighted towards common or rare species. Unsurprisingly we found a general increase in the mean number of samples needed to estimate species richness and the grain size, but this relationship was much stronger when species richness was focused on rare species (Figure 3B). XXXXXXXXXXXXXXXXX.

While our framework here focused on Florida, it can be generalized to other parts of the world, given the global increase in citizen science data (Pocock et al. 2017) and other citizen science datasets (e.g., iNaturalist). We focused on an annual temporal resolution, but smaller temporal windows (e.g., seasonally or monthly) could similarly be implemented to quantify effort needed to estimate species richness (Figure 4). In the case of Florida, for example, the effort would be greatest during spring and fall migration when the maximum levels of bird diversity are found in the region. Of course, the temporal window will ultimately depend on the quantity of data in a given locale. An important component of our framework is relying on the strong relationships we found between landscape metrics and species richness (Figure 2) and using a globally-applicable and openly-available remote sensing product (e.g., Copernicus land cover), ensuring the generalizability of this framework to other parts of the world. Our SEM model suggested that variables with good predictive power of effort needed are likely to be strongly linked to those affecting species richness. Hence, other studies might tailor the land-use predictors to their specific context.

(gap between current effort and effort needed)

Citizen science data will continue to play an important role in biodiversity monitoring in the future (McKinley et al. 2017; Chandler et al. 2017). Despite their promise, there remains reluctance to use these data (Burgess et al. 2017), in large part stemming from the gaps and redundancies in these data. A key goal in reducing the gaps and redundancies in these data, and concomitantly increasing the utility of these data is understanding how biodiversity sampling should be conducted in space and time (REF). We find relatively few eBird samples are necessary to meet 95% completeness, highlighting the potential of citizen science data to make informed comparisons of species richness in space and/or time. But the sampling effort inherently depends on the monitoring goal - c.f., common or rare species. Regardless, our framework, applicable and generalizable, allows for the quantification of sampling effort needed to estimate species richness with citizen science data.

# 

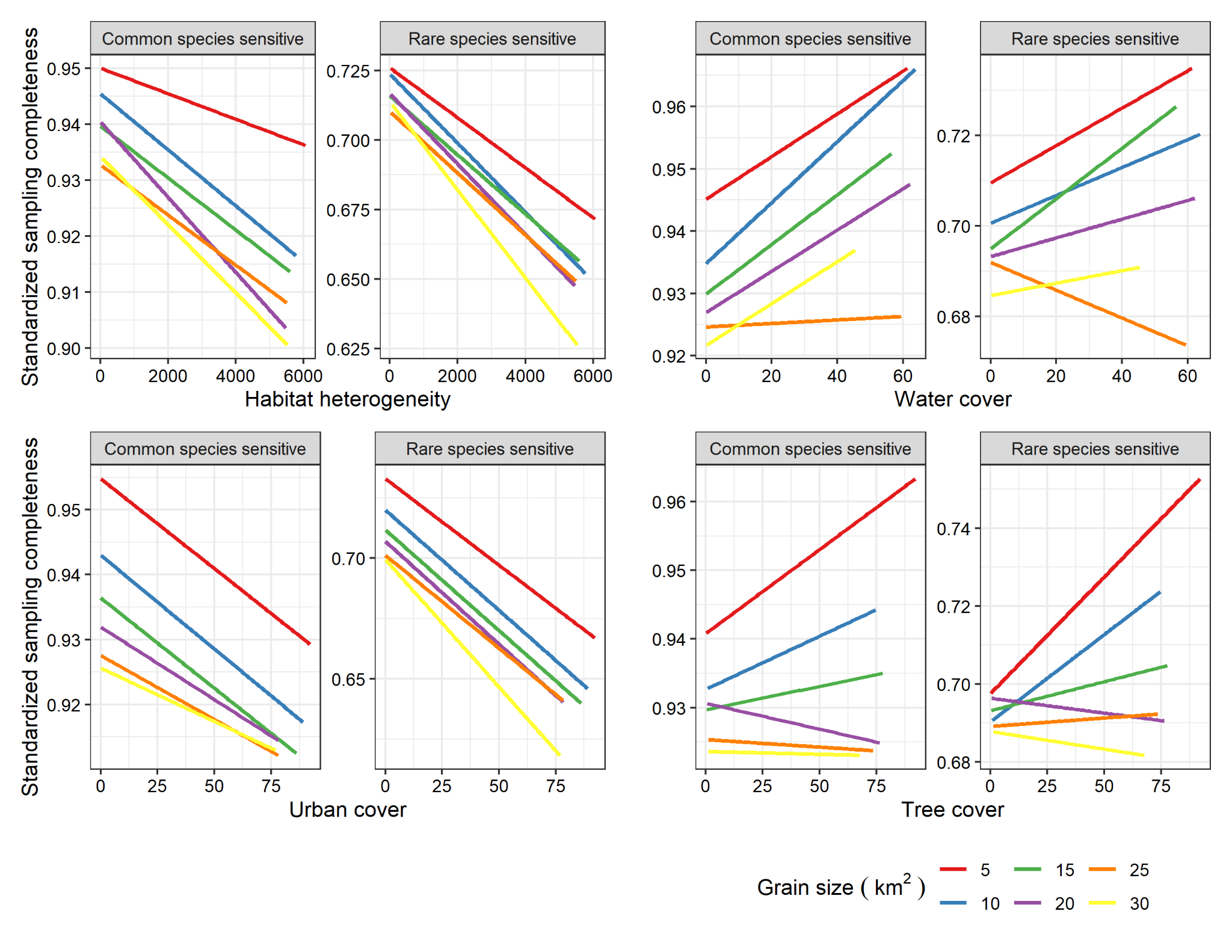
# References

For objective 1, we expected that sampling effort would be positively related to species richness, because high diversity areas can contain the rare and uncommon species that are often of interest to observers. We also expected that habitat heterogeneity and land-use would explain variation in sampling effort, partly mediated by their impacts on species richness. Finally, we expected the greatest disparity between the observed sampling effort and effort needed to quantify species richness at the smallest spatial scales.

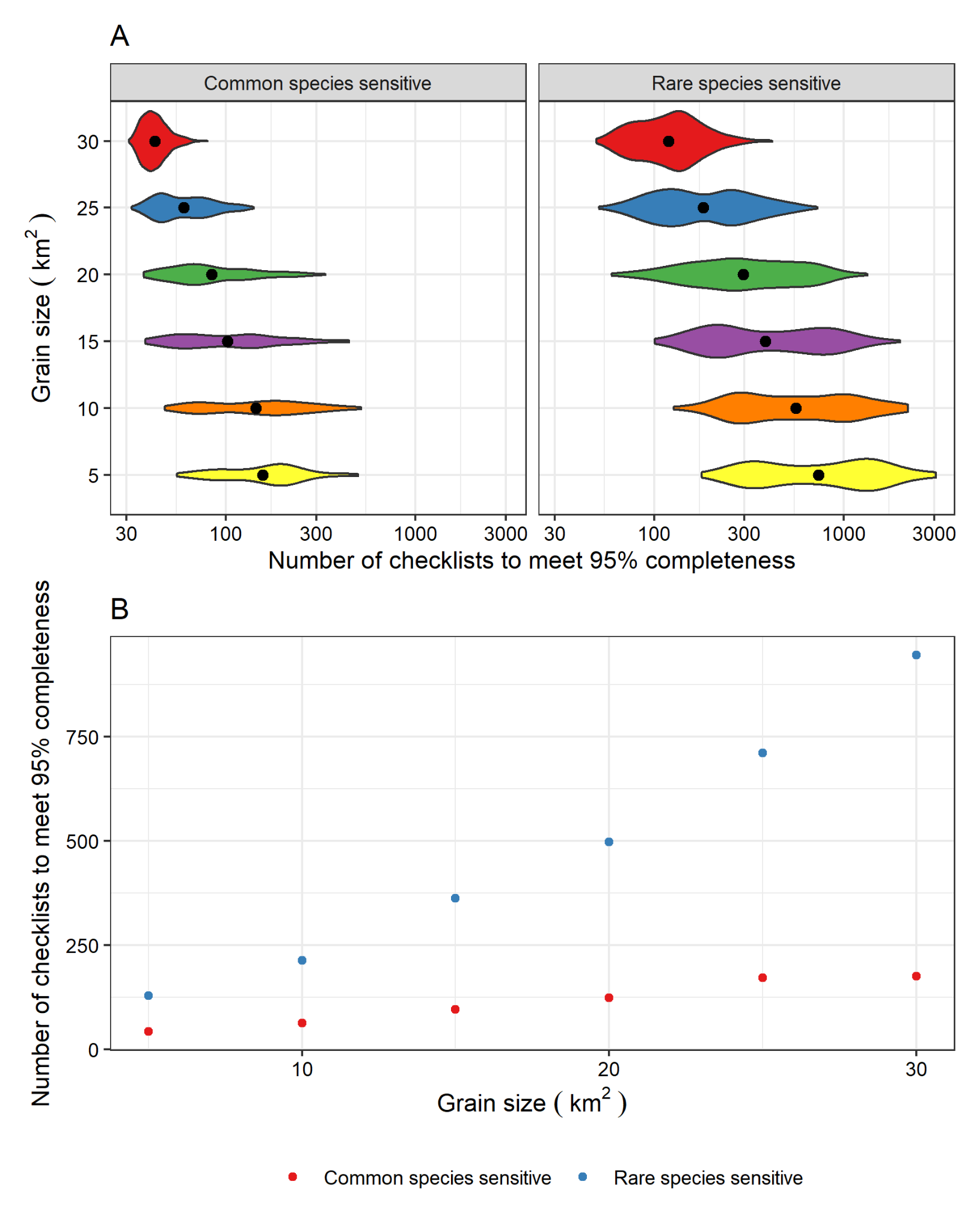
# Figures

# 

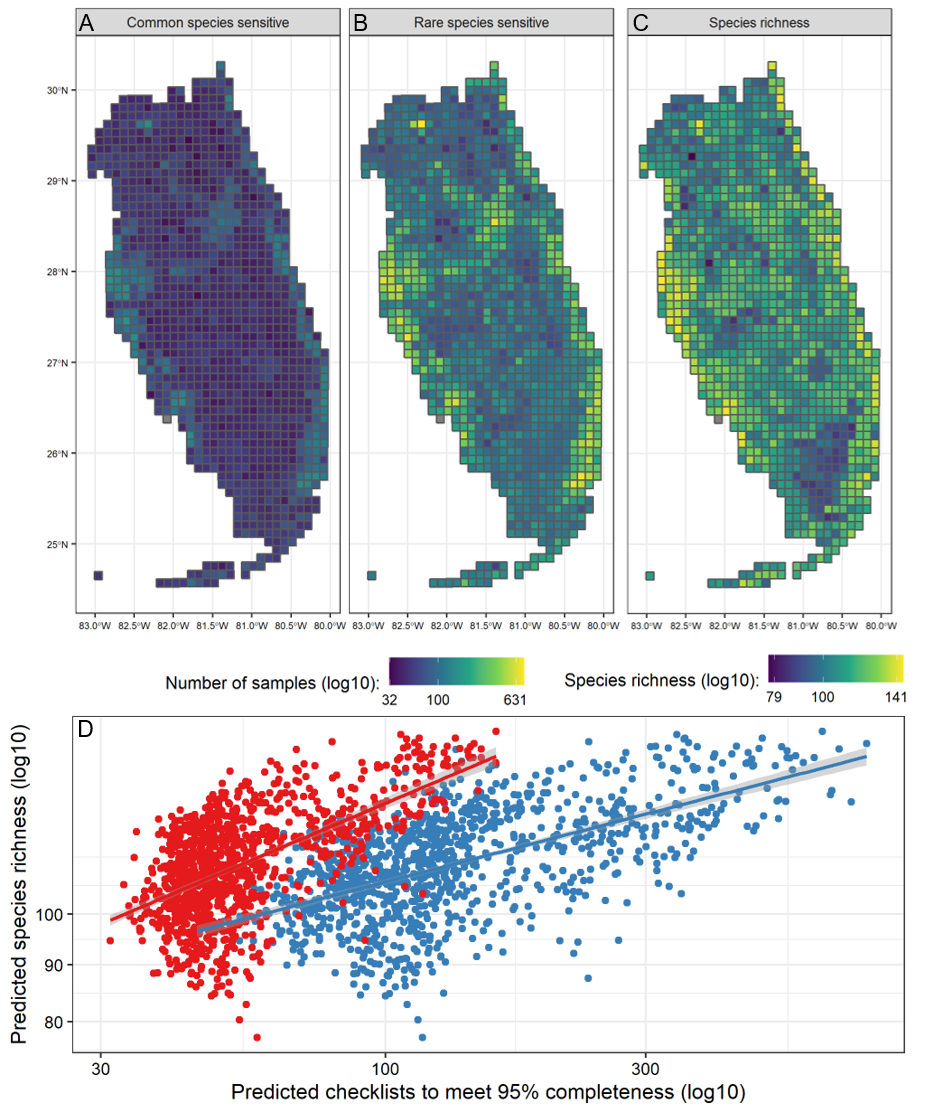
**Figure 1**. The results of our structural equation modelling, where the arrow goes from the predictor to the response variable, and the numbers represent standardized regression coefficients. These results represent the mean standardized regression coefficient across all grain sizes.



**Figure 2**. The relationship between bootstrapped sampling completeness (y-axis) and four landscape habitat attributes (x-axis), with habitat heterogeneity (top left), water cover (top right), urban cover (bottom left) and tree cover (bottom right) for our two measures of completeness. The relationship represents a linear model fit for each grain size from 5-30 km2. Note the different scales for the y-axis.



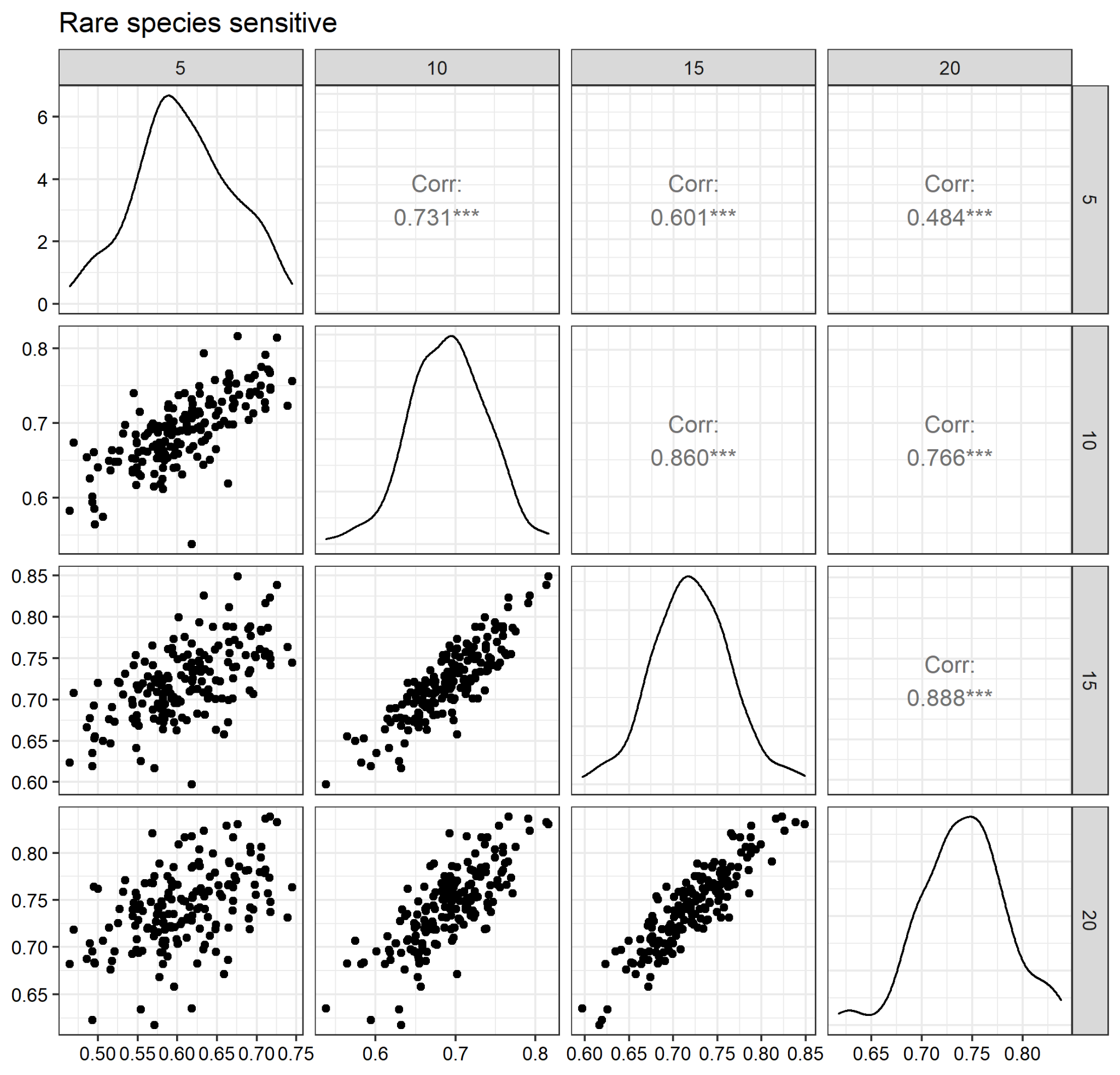
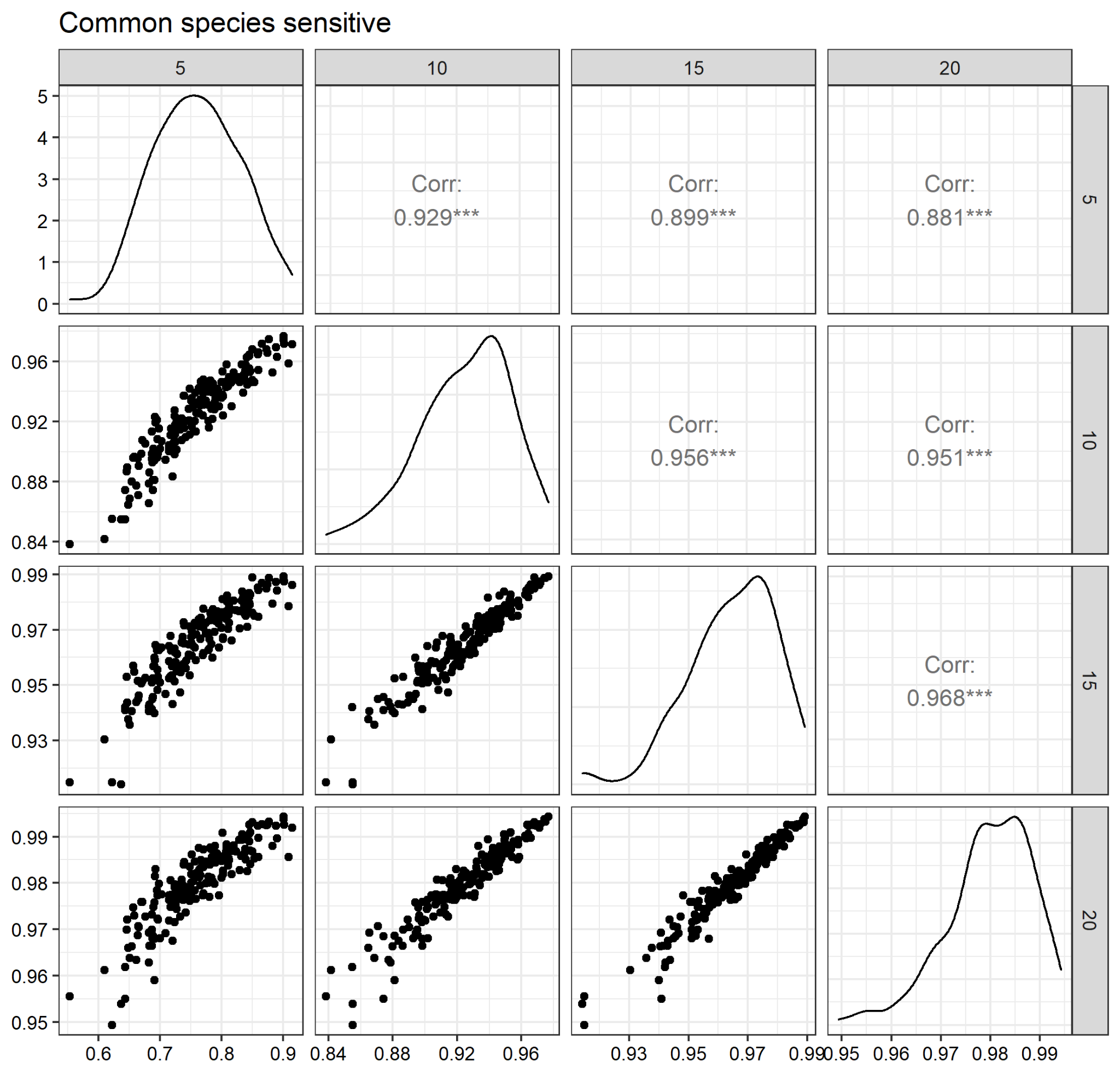
**Figure 3**. (A) The predicted number of checklists (log-10 scale) necessary to meet 95% completeness when completeness is calculated sensitive to common species (left) and sensitive to rare species (right). The black dot represents the mean value. (B) The mean number of checklists necessary to meet 95% completeness (y-axis) against grain size (x-axis).



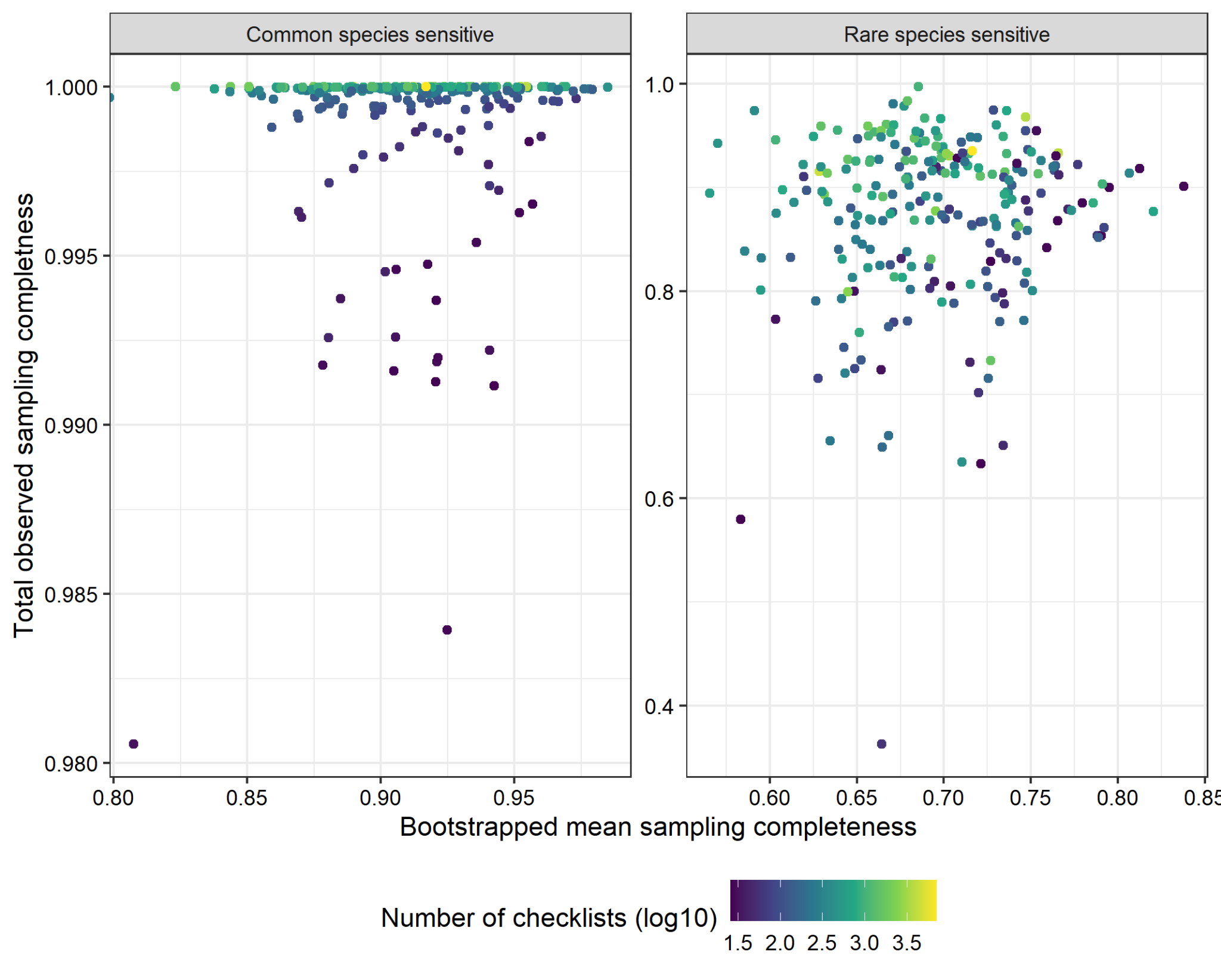
**Figure 4**. Spatial representation of the number of samples necessary to meet 95% completeness for (A) when completeness is calculated towards common species sensitive and (B) towards rare species sensitive. (C) A map of species richness throughout peninsular Florida, and (D) the generally positive relationship between species richness and the necessary number of samples needed to reach 95% completeness for common species (red) and rare species (blue). Results are shown for the 10 km2 grain size only.

# 

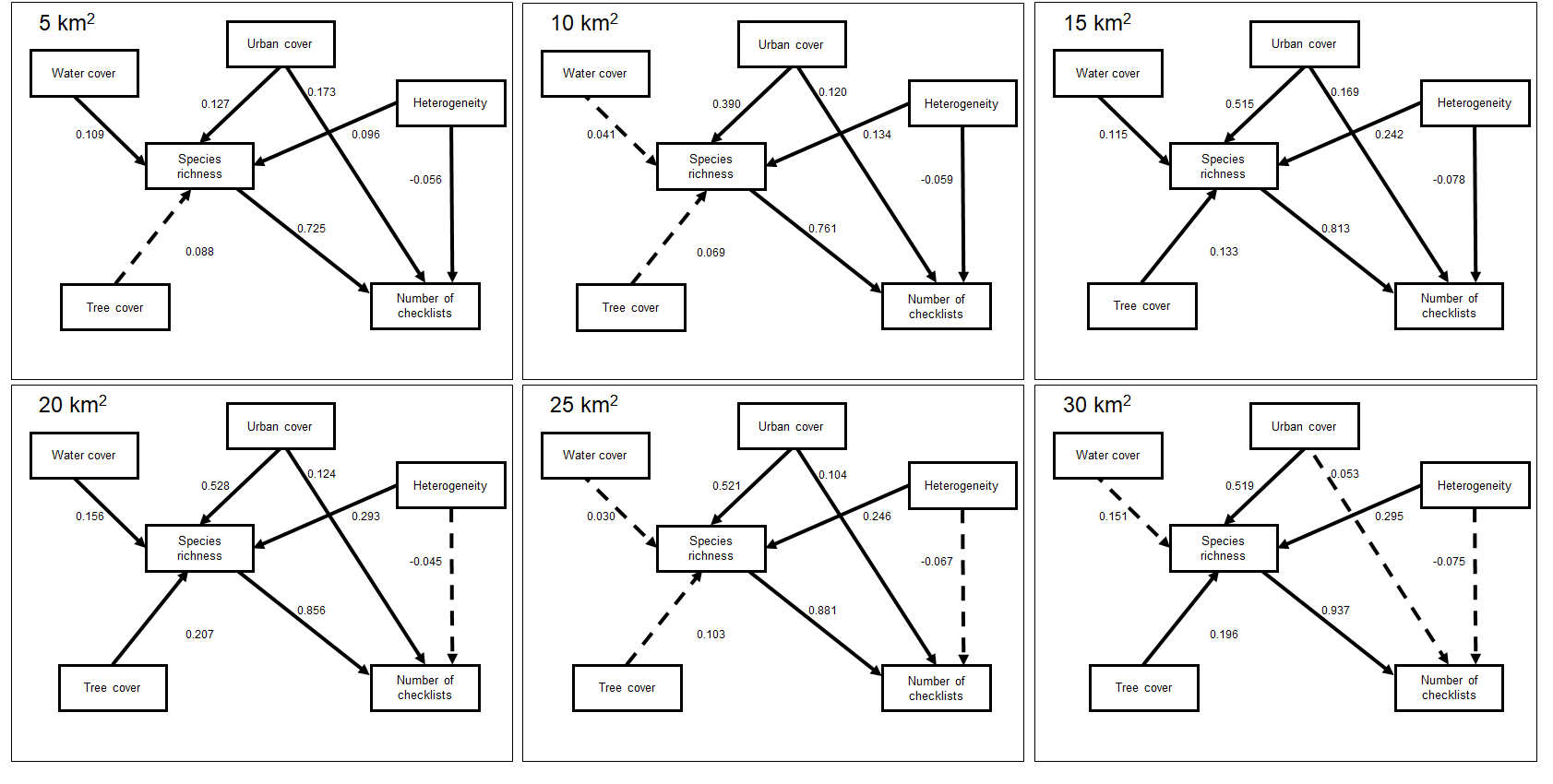
# Supplementary figures



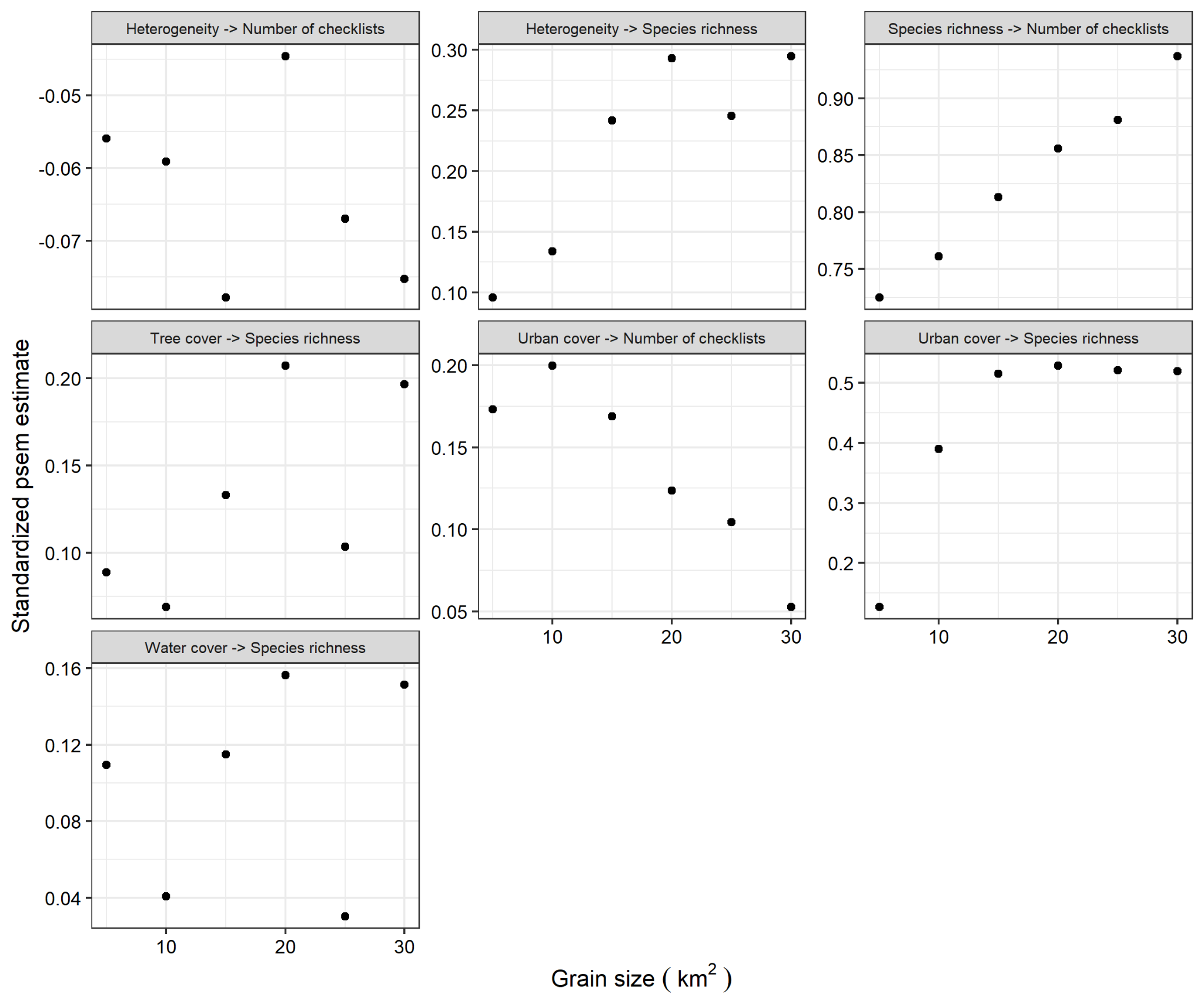
**Figure S1**. The correlations of completeness values for common species (top) and rare species (bottom), when 5, 10, 15, or 20 checklists were used and randomly sampled to calculate our bootstrapped mean completeness values. Because the correlations were high among the number of samples used, we use 10 checklists as our value and present these throughout our results.



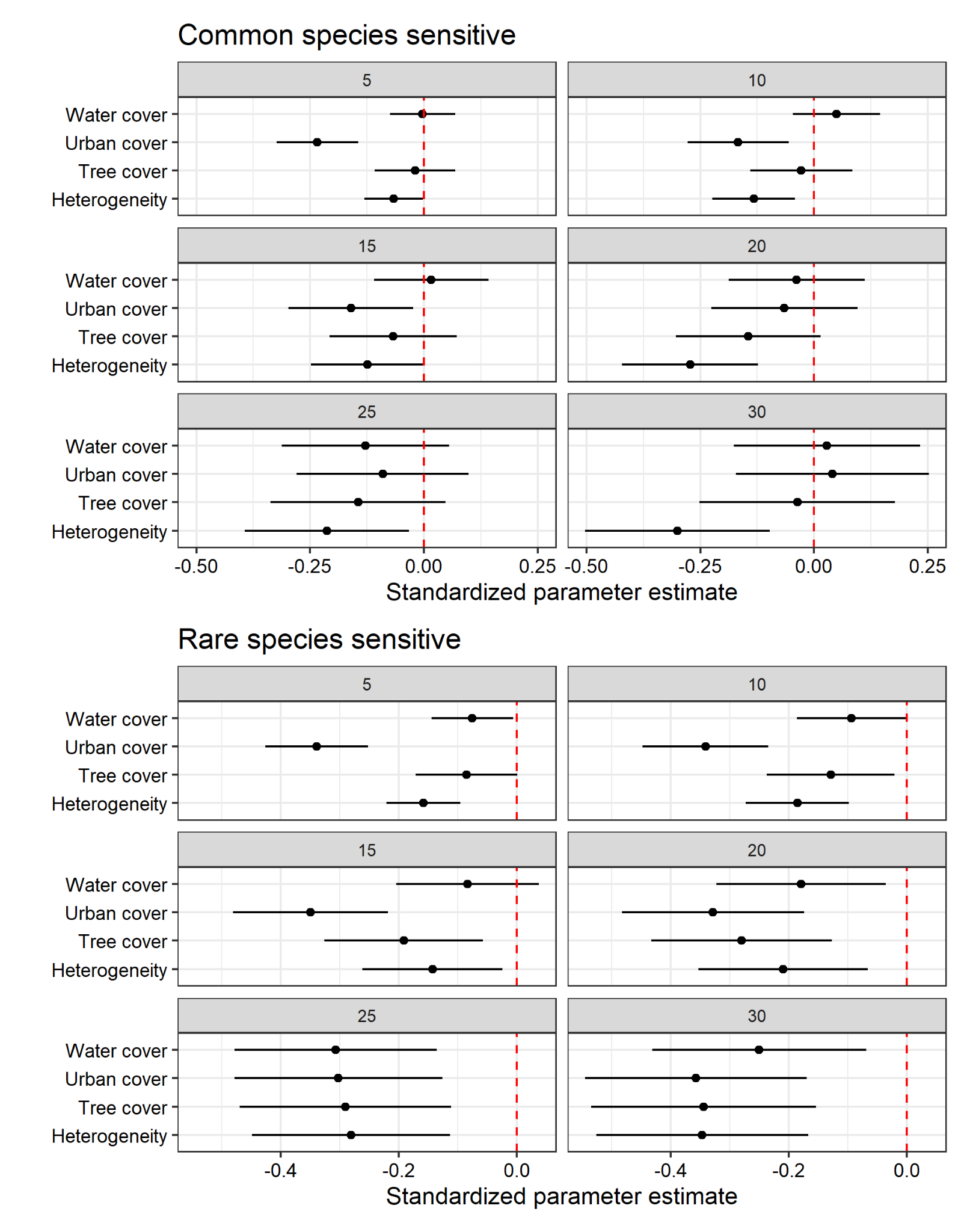
**Figure S2**. The relationship between bootstrapped mean sampling completeness and the total observed sampling completeness, calculated when all checklists are used at a site. The color represents the total number of eBird checklists at a site.



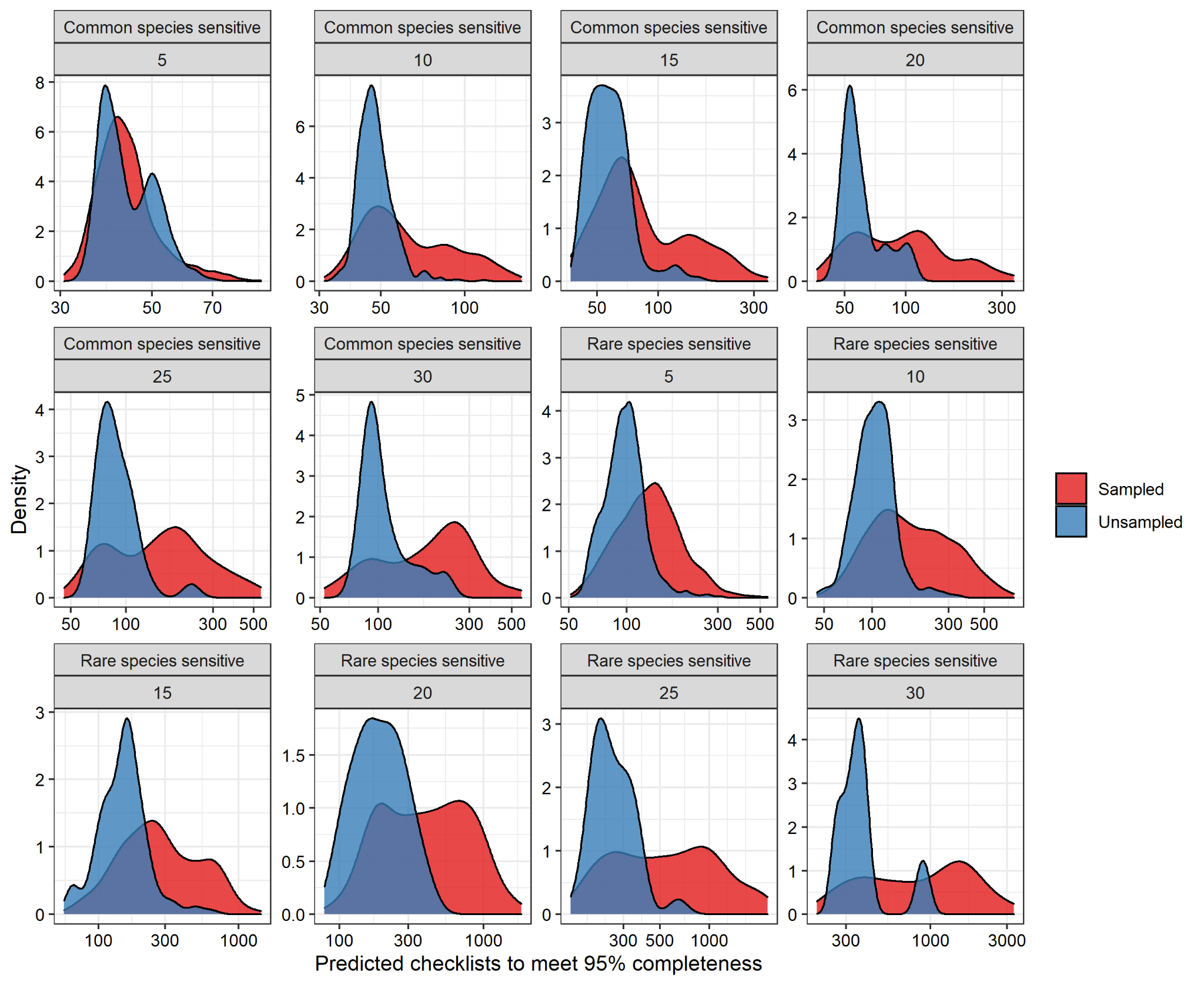
**Figure S3**. The results of our structural equation modelling analysis, applied across grain sizes - Figure 1 in the main text represents the mean values across grain sizes. For each instance, the dashed lines represent an insignificant link between variables.



**Figure S4**. The SEM results (see Figure S4) visualized as a function of grain size (x-axis), where the y-axis represents the standardized coefficient estimate from our psem analysis, and each panel represents the predictor - response relationship.



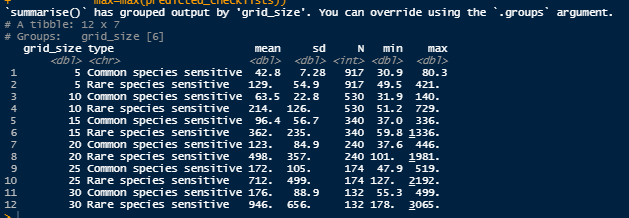
**Figure S5**. Results of multiple linear regressions where the response variable was bootstrapping sampling completeness (corresponding to Figure 1) considered for common species sensitive (top) and rare species sensitive (bottom). The red line is 0, the black dot is the standardized parameter estimates, and the black line represents the 95% confidence interval.



**Figure S6**. The distribution of predicted number of checklists to meet 95% completeness for unsampled sites (blue) and sampled sites (red) showing that the random forest models predicted generally less samples needed for unsampled sites.

# Supplementary Tables

**Table S1**. The number of checklists needed to meet 95% sampling completeness among sites (i.e., grid cells), and the variability associated with these estimates.



**Table S2**. The R2 of our random forest models.

