Effects of Ignoring Survey Design Information for Data Reuse

Foster, Scott D.

2021-09

Foster , S D , Vanhatalo , J , Trenkel , V M , Schulz , T , Lawrence , E , Przeslawski , R & Hosack , G 2021 , ' Effects of Ignoring Survey Design Information for Data Reuse ' , Ecological Applications , vol. 31 , no. 6 , 02360 . https://doi.org/10.1002/eap.2360

http://hdl.handle.net/10138/334029
https://doi.org/10.1002/eap.2360

acceptedVersion

Running Head: Effects of Ignoring Survey Design

Title: Effects of Ignoring Survey Design Information for Data Reuse

Scott D. Foster[1*], Jarno Vanhatalo[2,3], Verena M. Trenkel[4], Torsti Schulz[3], Emma Lawrence[5], Rachel Przeslawski[6], and Geoffrey R. Hosack[1]

[1]Data61 CSIRO, Hobart, Tasmania, Australia

[2]Department of Mathematics and Statistics, University of Helsinki, Finland

[3]Organismal and Evolutionary Biology Research Program, University of Helsinki, Finland

[4]Ifremer, Nantes, France

[5]Data61 CSIRO, Brisbane, Queensland, Australia

[6]Geoscience Australia, Canberra ACT, Australia

[*]Corresponding Author. email: scott.foster@data61.csiro.au

November 5, 2020

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/EAP.2360

Abstract

Data are currently being used, and reused, in ecological research at an unprecedented rate. To ensure appropriate reuse however, we need to ask the question: "Are aggregated databases currently providing the right information to enable effective and unbiased reuse?" We investigate this question, with a focus on designs that purposefully favour the selection of sampling locations (upweighting the probability of selection of some locations). These designs are common and examples are those designs that have uneven inclusion probabilities or are stratified. We perform a simulation experiment by creating datasets with progressively more uneven inclusion probabilities, and examine the resulting estimates of the average number of individuals per unit area (density). The effect of ignoring the survey design can be profound, with biases of up to 250% in density estimates when naive analytical methods are used. This density estimation bias is not reduced by adding more data. Fortunately, the estimation bias can be mitigated by using an appropriate estimator or an appropriate model that incorporates the design information. These are only available however, when essential information about the survey design is available: the sample location selection process (e.g. inclusion probabilities), and/or covariates used in their specification. The results suggest that such information must be stored and served with the data to support meaningful inference and data reuse.

Key Words: Bias, Survey Design, Database, Population Density Estimate, Model, Horvitz-Thompson, FAIR, Reuse, Data, Inclusion Probability

1 Introduction

Ecology and other environmental sciences, like most scientific disciplines, are currently utilising an unprecedented volume of data (e.g. LaDeau et al., 2017) and are poised to make use of even more (e.g. Culina et al., 2018). In our opinion, this trend is due to two parts: the increase in publicly available databases, and the realisation that incorporating data from many sources increases the information available for any particular study (Fletcher Jr. et al., 2019). The intended and desirable outcomes from this trend are that individual ecological studies are now broadening their ecological scale (e.g. global studies: Phillips et al., 2019; McKenzie et al., 2020; Gagné et al., 2020), or are shedding brighter lights on smaller scales so that data-poor systems can be quantitatively studied (e.g. Kindsvater et al., 2018; Fletcher Jr. et al., 2019).

The quality of the inferences from these analyses is only as good as the data that goes into them (e.g. Dobson et al., 2020). For aggregated data this means the quality of the contributing datasets and how well they can relate to each other. This is well recognised, and endeavours have been undertaken to improve data quality, with primary focus on two aspects: FAIR (Findable, Accessible, Interoperable, Reusable; Wilkinson et al., 2016; Stall et al., 2019), and standardisation of collection methods (e.g. Przeslawski et al., 2019). Undoubtedly, these will increase data reusability. However, are there any other hitherto overlooked aspects that will impede the reusability of ecological data?

All ecological data are the result of some sort of sampling process, and this process is based on a survey plan that describes where and how to collect samples. Many surveys do not consider these aspects in sufficient detail before implementation (Legg and Nagy, 2006). Recent modelling efforts with data aggregated from multiple surveys have suggested that survey information, such as the survey plan and sampling gear, should be taken into account to help data ‘speak’ to one another (Fletcher Jr. et al., 2019, and references therein). Without this information, it is hard to understand the meaning of the data and further (potentially wrong) assumptions are required for analysis and interpretation. Indeed, the survey information, or survey metadata, is sometimes not even available to users as the data themselves are. The importance of this omission may be under-appreciated, and it is yet unknown how much of an effect this has on subsequent analyses.

In this work, we investigate what effect ignoring survey design information can have on analysis outputs. We make our inference from a simulation experiment based on a 2018 survey of deep-water corals, which was formally and purposefully designed to increase information content by modifying the selection process for sample locations (Foster et al., 2020). The specific questions we ask are: 1) If these data were contributed to databases that aggregate multiple surveys, would naive reuse generate a false picture of the ecology or provide misleading information for management? and 2) How much, if any, modification of the sample location selection process (away from complete randomisation) is tolerable before data reuse needs to incorporate survey design information? We discuss what survey design information is needed to be stored within aggregated databases.

2 Methods

2.1 Deep-Water Corals

A population of the deep-water stony coral *Solenosmilia variabilis* is located in the Huon Australia Marine Park, which contains geomorphological features known as the Tasmanian seamounts, located south of Tasmania, Australia. The distribution of *S. variabilis* in this region is not well understood, except in vague terms – it prefers outcropping locations within a partially-known depth ranges (Thresher et al., 2011). To rectify this knowledge gap, a scientific survey was undertaken in late 2018 (Williams et al., 2018, 2020), which follows a 2010 survey in a comparable region (Williams et al., 2010). The design for the 2018 survey is outlined in Foster et al. (2020) and consisted of favouring sample locations where *S. variabilis* presence/abundance is thought to be uncertain.

The method used to create the survey was to sample potential sampling locations with specified uneven *inclusion probabilities* (e.g. Thompson, 2012). For the 2018 seamount survey, these probabilities were expert derived and up-weight the locations that: 1) are within the broad species bathymetric range; and 2) are locally elevated in relation to neighbouring locations, measured by the topographic position index (TPI; Weiss, 2001) – see Fig. 1. Only those locations within 485 m and 2015 m deep were considered for sampling.

[Figure 1 about here.]

In this work, we utilise the 2018 survey's uneven inclusion probabilities defined in Foster et al. (2020, Table 2), which links our simulation to procedures used in practice. These inclusion probabilities are highly skewed as the area covered by seamounts is comparatively small (See Fig. 1). The distribution of inclusion probabilities is given in Appendix S1: Fig. S2. To simplify computation, we only use the survey area within the Huon Marine Park, which also contains many of the seamounts in the broader region.

We also utilise data on *S. variablilis* from a 2010 survey described in Williams et al. (2010). The survey design for the 2010 survey was less formal but did target the coral's depth range and sites with higher TPI. For modelling purposes, we assume that the 2010 design is *ignorable* once the depth and TPI are included as covariates (Gelman et al., 2013). These data were generated from a camera towed along the seafloor, and later quantified by counting the number of live *S. variabilis* coral heads within regularly spaced images. The size of the seafloor covered by the quantification area, within each image is also recorded. Overall, in the Huon park there are 1517 images spaced along 19 transects with the longest transect having 212 images and the shortest 12. Images from the 2018 survey were not used in this work as, at the time of writing, the images are not yet quantified.

2.2 A Model for Coral Distribution

To analyse the 2010 image data, we use a geostatistical model. In particular, we use the 'SPDE' approach, which is implemented using the 'INLA' approximation (Rue et al., 2009; Lindgren and Rue, 2015) implemented for R (R Core Team, 2019). This approach to computing is relatively fast, so that many models can be fitted. We notate each of the ($i = 1 \ldots 1517$) observed *S. variabilis* coral abundance data as y_i, and model all observations as a function of geographical position (s_i), bathymetry, and TPI. That is:

$$\log\left[\mathrm{E}\left(y_i|\boldsymbol{\theta}, b(s_i), t(s_i)\right)\right] = \beta_0 + \beta_1 b(s_i) + \beta_2 b(s_i)^2 + \beta_3 t(s_i) + u(s_i) + \log\left(A_i\right), \quad (1)$$

where β_j is a regression parameter, $b(s_i)$ and $t(s_i)$ are bathymetry and TPI covariates respectively, $u(s_i)$ is a spatial random variable, A_i is the area that the ith image sampled, and all effects are gathered into the parameter vector $\boldsymbol{\theta}$. A quadratic effect for depth was assumed to reflect the belief that the *S. variabilis* depth-niche was covered by the data, whereas it is thought that there is no upper limit to TPI preference. We assume that the conditional distribution of $\boldsymbol{y}_i|\boldsymbol{\theta}, b(s_i), t(s_i)$ is Poisson and that the spatial random variable, $u(s_i)$, is assumed to follow a Matérn Gaussian process with mean zero and smoothness $\nu = 1$. This model gives the spatial covariance of the random effect as

$$\operatorname{cov}\left[u(s_i), u(s_{i'})\right] = \sigma^2 \kappa ||s_i - s_i'|| K_1(\kappa ||s_i - s_i'||),$$

which has standard deviation (σ) and scaling parameter (κ). The function $K_1(\cdot)$ is the modified Bessel function of the second kind and order 1. The Matérn process has *effective range* of $\sqrt{8}/\kappa$, which is the empirically derived spatial distance where correlation is $\gamma \approx 0.1$ (Lindgren et al., 2011; Lindgren and Rue, 2015). We specify a penalised complexity prior (Simpson et al., 2017) where there is $\Pr(\sigma > 5) = 0.1$, which penalises overly flexible spatial processes. The effective range (γ) of the process has a prior such that $\Pr(\gamma < 50m) = 0.05$ so that the spatial dependence is unlikely to be very short. Priors for the regression coefficients are chosen to penalise extreme values.We define these to be normal distributions with zero mean and variance equal to 5. Both covariates were standardised to have mean zero and variance 1 before analysis.

2.3 Simulation Experiment

The base form of the simulation experiment is: 1) vary inclusion probabilities to be more and less severe than the 2018 inclusion probabilities, 2) generate a survey design from these inclusion probabilities, 3) simulate data at the sampling locations generated (using the model fitted to the 2010 image data), 4) analyse the simulated data with naive (ignoring sampling probabilities) and more sophisticated methods that account for the survey design, and 5) summarise the simulations' analyses as a response to variation in the unevenness of inclusion probabilities. This approach will inform if the survey data can be naively reused in the analysis of aggregated data.

To vary the inclusion probabilities for the $N = 8840$ sites that define the sampling area, we start with the inclusion probabilities used to design the 2018 survey, and we arrange these probabilities into an $N \times 1$ vector $\boldsymbol{p}$. The N sites are arranged on a 300m $\times$ 300m grid and match the grid of the covariates (see Fig. 1). This was chosen to match that used in Foster et al. (2020), who used this as a compromise between accuracy and computational expense. The inclusion probabilities for the simulation experiment are defined as

$$\boldsymbol{p}_\alpha = \max\left(\boldsymbol{p}_\alpha^*, \boldsymbol{0}\right)/K, \qquad \text{where}$$
$$\boldsymbol{p}_\alpha^* \triangleq \left[\boldsymbol{1}\bar{p} + \alpha\left(\boldsymbol{p} - \boldsymbol{1}\bar{p}\right)\right],$$

$\bar{p} = {(\boldsymbol{1}^\top \boldsymbol{p})}/{N}$ is the mean of $\boldsymbol{p}$, $K = \boldsymbol{1}^\top \boldsymbol{p}_\alpha$ is a normalising constant, and the maximum function is applied element-wise. If an inclusion probability is zero, then that site will not be chosen in the sample. The parameter α indexes the severity of the unevenness in the inclusion probabilities, with $\alpha = 0$ corresponding to even inclusion probabilities (and completely randomised sampling), $\alpha = 1$ corresponding to the 2018 survey's inclusion probabilities and $\alpha > 1$ giving inclusion probabilities more extreme. We allow α to vary from 0 to 2 in increments of 0.1. For each α, $J = 1000$ surveys were simulated, each consisting of $n = 50, 100, 200$ observations from the N sites withing the sampling area. The locations of the observations where chosen at random using $\boldsymbol{p}_\alpha$.

For each simulated survey, data were simulated at the n selected locations using parameters drawn from the posterior distribution of the model in Section 2.2, fitted to the 2010 data. This ensures that all modelled aspects of the 2010 data, including variability, are incorporated into the simulation study. The marginal posterior distribution of the covariate effects is presented in Appendix S1: Fig. S3.

Each simulated data set is analysed using design-based and model-based estimators. The target metric in each of these analyses is the average number of corals per 20m^2 image (coral density). Theoretically, it is useful to consider the bias in the average density for both design-based and model-based analyses: design-based estimates are intended to be unbiased for the average, and the average is also the Bayes estimate under quadratic loss for model-based methods. We note that other summaries could be of interest, like the maximum coral density, but the average is a very common summary, almost ubiquitously so. The design-based analyses were a naive mean ($^1/_n \sum y_i$), and the Horvitz-Thompson (HT) estimator (see Thompson, 2012) of the form $\sum {}^{y_i}/_{np_{\alpha i}}$ where the sum is over the n samples. The HT estimator is only available when the inclusion probabilities for the samples are known, and it should (theoretically) produce unbiased estimates, even when inclusion probabilities are unequal. The naive mean should (theoretically) only be unbiased when the inclusion probabilities are equal (Thompson, 2012).

The model in Section 2.2 was used to analyse each simulated data set along with three simplifications. These models are used to investigate the effect of only making part of the design information available to the analysis process. The models are:

Covariates + Spatial The full model in Section 2.2.

Spatial Covariates unavailable or neglected and only the spatial effects are included.

Covariates Spatial effects are omitted. The analyst assumes that the observations are independent given the covariates.

Bathymetry/TPI The third simplification is to drop each of the covariates (bathymetry and TPI) in turn, with no spatial effect.

For all models, the 'true' average density of the jth simulation, μ_j, was calculated by taking the mean of the set of predictions formed at a grid of N locations throughout the study region. The same set of draws of the parameters (from the posterior that conditions on the 2010 data, Section 2.2) were used to calculate the set of μ_j. For a given value of α, the average density estimate of the kth estimation method was assessed by calculating a percentage difference between the estimated average density ($\hat{\mu}_{jk}$) and the quantity it is estimating (μ_j). Formally, for the jth simulation replicate and the kth estimation method, the percentage difference is

$$d_p(j,k) = 100\frac{\hat{\mu}_{jk} - \mu_j}{\mu_j}.$$

For each value of α and for each estimation method, there are J estimates of average coral density. We summarise this information using the median and mean absolute deviation (MAD; see Venables and Ripley, 2002). These are relatively robust measures of location and scale that are not unduly affected by extreme values (outliers). We take the median of the naive mean estimates, when the inclusion probabilities were even ($\alpha = 0$), as the reference value for comparison against all other estimators and all other values of α. The naive mean has well known and desirable properties when sampling is even ($\alpha = 0$).

3 Results

Fitting the model to the 2010 image data, see Section 2.2, suggested that coral density peaked around 1350 m deep, and had a much reduced expectation outside of the range (-1700 to -1000 m). Increasing TPI increased the density of corals (about 12 times increase from flat areas to the extremely elevated). The spatial dependence was short with $\mathrm{E}(\gamma|\boldsymbol{y}) = 333$ m ($\mathrm{SD}(\gamma|\boldsymbol{y}) = 72.3$ m), and the spatial standard deviation was $\mathrm{E}(\sigma|\boldsymbol{y}) = 2.8$ ($\mathrm{SD}(\sigma|\boldsymbol{y}) = 0.4$). Posterior distributions for all parameters defined in (1) are presented in Appendix S1: Fig. S3. Posterior predictions from this model are presented in Fig. 1 and show the effect of depth, which is smooth over the survey area, and the relatively patchy effects of TPI and spatial noise.

Results for the simulation experiment, described in Section 2.3, are presented in Fig. 2. Overall, it is clear that ignoring the inclusion probability information can induce substantial bias in average coral density estimates. It is evident though, that even those estimation methods that do incorporate inclusion probabilities can perform badly but in general they work as intended (Fig. 2).

The naive mean is an increasing function of α, implying that the mean increases as more favourable environments are sampled with greater inclusion probabilities. The naive mean also has very high variation, presumably due to not taking the appropriate weighting of each observation. The HT estimator, which does account for unequal inclusion probabilities, *decreased* with α and did so sharply just past $\alpha = 1$ after agreeing with the reference well for all sample sizes for $\alpha < 1$.

[Figure 2 about here.]

The simulation illustrated that model-based analyses can produce unbiased estimates of the average density (Fig. 2). The form of the model appears to be important though. The model with no covariates (just a spatial term) and the model with only the bathymetry covariate had undesirable performance, with a trend similar to, but not as extreme as, the naive mean estimate (Fig. 2). When the full model (covariates and spatial) and the TPI-only model were used to analyse the simulated data sets, Section 2.2, the median of the estimates for average density were comparatively unbiased albeit after having high values for very small α with $n = 50$ (Fig. 2). A similar pattern was observed for the model with both covariates, but this exhibited a slight positive bias.

The full model (with random spatial effects) consistently exhibits small variation in the distribution of estimates, except for $n = 50$ and for small α (Fig. 2, right column). This result is linked to the extrapolation/leverage issues (see Discussion). The covariates model and the TPI model also suffer from this behaviour, at $n = 50$ and $\alpha = 0$, but do not have the low variability in the distribution of estimates, which is exhibited by the full model.

4 Summary and Discussion

For data to be FAIR it must be reusable (Wilkinson et al., 2016; Stall et al., 2019). For it to be reusable the relevant information must be made available about *how* to reuse it. Without this information assumptions must be made, with the naive assumption (equal probability random sample) often being wrong.

In this study we investigated the effect of ignoring survey-design information using a simulation experiment based on a 2018 survey design, and 2010 image data, for a chain of seamounts in southern Australia. We found that ignoring survey design information can induce a substantial bias in estimates of average population density when a naive or an inappropriate analysis method is used; the median of the simulations' average density estimates can be up to $\sim 250\%$ biased and estimates for individual data sets even worse. The potentially large bias has the potential to make seemingly straightforward inferences wrong and misleading. We note that the density bias does not disappear with increased sample sizes (Fig. 2), so 'big-data' are no panacea. Even worse, big-data may lead to confident, but biased, inferences.

The simulation experiment showed that some analysis methods performed better than others with uneven inclusion probabilities. The naive mean estimate for population density was the worst performer and some model-based estimators also produced consistently poor results (Fig. 2). The bias was alleviated by incorporating survey design information into the analysis, either through inclusion probabilities for the Horvitz-Thompson (HT) estimator, or through inclusion of the appropriate covariates in a model-based analysis. The sudden appearance of bias in the HT estimator at $\alpha = 1$ is suspected to be caused by the introduction of sites with inclusion probabilities of zero at $\alpha = 1$ (see Methods Section) and the associated severe right skew in the distribution of inclusion probabilities (Appendix S1: Fig. S2). We stress that obtaining bias by ignoring design information is not a new result, see Gelman et al. (2013, Chapter 8), Diggle et al. (2010) and Pati et al. (2011). However, this is perhaps under-appreciated by those who deal with ecological data (but see Pennino et al., 2018; Dobson et al., 2020). In fisheries, the problem is receiving recent attention for commercial catch data (e.g. Trenkel et al., 2013).

The poor performance of the models with covariates for smaller sample sizes is likely to be due to insufficient sampling of covariate space (top panel of Appendix S1: Fig. S1, $\alpha \lesssim 0.2$). The insufficient sampling of covariates potentially leads to survey data that must be extrapolated, in covariate space, to predict to all locations (to calculate the average density). This extrapolation in covariates may be erratic and of low-quality. The poor sampling of covariates potentially also leads to samples that have undue leverage, which can distort the model estimates. The supplementary study in Appendix S1: Section S1 indicates that small sample sizes underestimate the range of both the bathymetry and TPI covariates. A second reason for poor performance is poor sampling of the spatial extent and hence poor prediction of the spatial random effect throughout the entire region. However, the spatial effect has a relatively small effective range so it is likely that only the largest sample sizes will cover the area sufficiently.

Survey designs are often based on covariates. To account for the influence of the survey design on the model's predictions, these covariates should be included in any model utilising the survey data (Gelman et al., 2013). If there is no information about how the survey was designed, then it may be most appropriate to include the covariates that the analysts *assumes* to be important in the design, or to use a preferential sampling model (Diggle et al., 2010). We stress that not including any covariates makes the assumption that there were no design-covariates – corresponding to the naive mean in our simulation study – which may be a very inappropriate assumption. We are also aware that this simple advice may be hard to implement in certain situations; an example is when all covariates are not available for all surveys utilised in a particular reuse. In these situations, careful and skilful analyses must be undertaken, which will rest on assumptions that are necessary to describe *both* the sampling process *and* ecological processes (Diggle et al., 2010; Pati et al., 2011; Liu and Vanhatalo, 2020). We note that including a spatial random effect in the southern seamount simulation is not an effective replacement for covariates and that all the design covariates need to be included (Fig. 2). Both these results are likely to be due to the relatively noisy, patchy and spatially non-smooth geographical distribution of TPI.

The southern seamount survey example is quite extreme in its patchy topography and hence the unevenness of the inclusion probabilities. This is why we chose this survey design – to investigate how bad things could be if ignored. However, altering the amount of unevenness (varying α, Section 2.3) and coupling to the more general theoretical results (e.g. Gelman et al., 2013; Diggle et al., 2010) suggest that our results are generalisable to any survey. Of course, the severity will depend on the amount of variation in the inclusion probabilities, the sample size (Fig. 2), and the survey design (through specification of inclusion probabilities/strata, Fig. 2).

To ensure the ability to reuse data, we suggest that database managers should facilitate the storage and serving of information about survey design, perhaps even incorporated into formal data formats. Reusers of data should be encouraged, perhaps by changing default function settings, to download this information with the data. Data reusers should also be educated about the importance of survey design information. To be clear, this information at minimum should consist of a detailed description of, or accurate reference to, the survey design procedure. Additionally, it is highly desirable to also include: 1) the inclusion probabilities (the H-T estimator only needs these at the *sampled* locations), and 2) the values of the covariates at each location within the well-defined study region. We note that the inclusion probabilities could be stored as a field in the data (architecturally similar to another biological measurement), and that the covariates could be part of a meta-data record (or a link to them).

A corollary to this work is that it is best, and in many ways practically necessary, to have a formal survey design if the data are to be reused. Whilst it is possible to model the data from surveys without formal designs, the process becomes more complex (see the variety of models in Diggle et al. 2010 and Gelman et al. 2013, Chapter 8), and is liable to ambiguity through the necessity of making assumptions that are oftentimes untestable. The data may end up being unusable, produce ambiguous results, and their curation and analysis may create a large, hidden research cost (Dobson et al., 2020).

We recommend that surveys should be formally designed *and importantly:* the survey design should be stored along with the data. This work serves as a cautionary tale for those who wish to use and reuse data: Do not ignore how the data were obtained, unless you are confident that there is no intentional, or unintentional, specification of unequal inclusion probabilities in the survey design. Further, this work demonstrates what is needed to interpret survey data: information about the survey design employed to collect the data.

Acknowledgements

This work was undertaken for the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Science Program. J.V. was additionally funded by the Academy of Finland (grant 317255). R Przeslawski publishes with the permission of the CEO, Geoscience Australia. We would like to thank Franzis Althaus, Alan Williams, Tim Langlois, Zhi Huang, Jasmine Bursic, Nicholas Johannsohn, Amy Nau, Keith Hayes and two anonymous reviewers.

References

Culina, A., T. Crowther, J. Ramakers, P. Gienapp, and M. Visser (2018). How to do meta-analysis of open datasets. *Nature Ecology and Evolution 2*, 1053–1056.

Diggle, P. J., R. Menezes, and T.-l. Su (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(2), 191–232.

Dobson, A., E. Milner-Gulland, N. J. Aebischer, C. M. Beale, R. Brozovic, P. Coals, R. Critchlow, and A. Dancer et al. (2020). Making messy data work for conservation. *One Earth 2*, 455–465.

Fletcher Jr., R. J., T. J. Hefley, E. P. Robertson, B. Zuckerberg, R. A. McCleery, and R. M. Dorazio (2019). A practical guide for combining data to model species distributions. *Ecology 100*(6), e02710.

Foster, S. D., G. R. Hosack, J. Monk, E. Lawrence, N. S. Barrett, A. Williams, and R. Przeslawski (2020). Spatially balanced designs for transect-based surveys. *Methods in Ecology and Evolution 11*(1), 95–105.

Gagné, T. O., G. Reygondeau, C. N. Jenkins, J. O. Sexton, S. J. Bograd, E. L. Hazen, and K. S. Van Houtan (2020, 02). Towards a global understanding of the drivers of marine and terrestrial biodiversity. *PLOS ONE 15*(2), 1–17.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Kindsvater, H. K., N. K. Dulvy, C. Horswill, M.-J. Juan-Jordá, M. Mangel, and J. Matthiopoulos (2018). Overcoming the data crisis in biodiversity conservation. *Trends in Ecology & Evolution 33*(9), 676 – 688.

LaDeau, S., B. Han, E. Rosi-Marshall, and K. Weathers (2017). The next decade of big data in ecosystem science. *Ecosystems 20*, 274–283.

Legg, C. J. and L. Nagy (2006). Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management 78*(2), 194 – 199.

Lindgren, F. and H. Rue (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software, Articles 63*(19), 1–25.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 423–498.

Liu, J. and J. Vanhatalo (2020). Bayesian model based spatiotemporal survey designs and partially observed log gaussian cox process. *Spatial Statistics 35*, 100392.

McKenzie, L., L. M. Nordlund, B. L. Jones, L. C. Cullen-Unsworth, C. M. Roelfsema, and R. Unsworth (2020). The global distribution of seagrass meadows. *Environmental Research Letters*.

Pati, D., B. J. Reich, and D. B. Dunson (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika 98*(1), 35–48.

Pennino, M., I. Paradinas, J. Illian, F. Muñoz, J. Bellido, A. López-Quílez, and D. Conesa (2018). Accounting for preferential sampling in species distribution models. *Ecology and evolution 9*(1), 653–663.

Phillips, H. R. P., C. A. Guerra, M. L. C. Bartz, M. J. I. Briones, G. Brown, T. W. Crowther, O. Ferlian, K. B. Gongalsky, J. van den Hoogen, and J. Krebs et al. (2019). Global distribution of earthworm diversity. *Science 366*(6464), 480–485.

Przeslawski, R., S. Foster, J. Monk, N. Barrett, P. Bouchet, A. Carroll, T. Langlois, V. Lucieer, J. Williams, and N. Bax (2019). A suite of field manuals for marine sampling to monitor australian waters. *Frontiers in Marine Science 6*, 177.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rue, H., S. Martino, and N. Chopin (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

Simpson, D. P., H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science 32*(1), 1–28.

Stall, S., L. Yarmey, J. Cutcher-Gershenfeld, B. Hanson, K. Lehnert, B. Nosek, M. Parsons, E. Robinson, and W. Lesley (2019). Make all scientific data fair. *Nature 570*, 27–29.

Thompson, S. (2012). *Sampling*. Wiley.

Thresher, R. E., J. Adkins, S. J. Fallon, K. Gowlett-Holmes, F. Althaus, and A. Williams (2011). Extraordinarily high biomass benthic community on southern ocean seamounts. *Scientific Reports 1*, 119.

Trenkel, V. M., J. A. Beecham, J. L. Blanchard, C. T. T. Edwards, and P. Lorance (2013). Testing cpue-derived spatial occupancy as an indicator for stock abundance: application to deep-sea stocks. *Aquat. Living Resour. 26*(4), 319–332.

Venables, W. and B. Ripley (2002). *Modern Applied Statistics with S. Fourth Edition*. Springer.

Weiss, A. (2001). Topographic positions and landforms analysis (poster). ESRI International User Conference, July 2001. San Diego, CA: ESRI.

Wilkinson, M., M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, L. da Silva Santos, and P. Bourne et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data 3*.

Williams, A., F. Althaus, M. Green, K. Maguire, C. Untiedt, N. Mortimer, C. J. Jackett, M. Clark, N. Bax, R. Pitcher, and T. Schlacher (2020). True size matters for conservation: A robust method to determine the size of deep-sea coral reefs shows they are typically small on seamounts in the southwest pacific ocean. *Frontiers in Marine Science 7*, 187.

Williams, A., N. Bax, M. Clark, and T. Schlacher (2018). RV Investigator voyage summary IN2018_v08: Status and recovery of deep-sea coral communities on seamounts in iconic Australian marine reserves. Australian Marine National Facility Report. Retrieved from `https://www.marine.csiro.au/data/reporting/get_file.cfm?eov_pub_id=187`.

Williams, A., T. A. Schlacher, A. A. Rowden, F. Althaus, M. R. Clark, D. A. Bowden, R. Stewart, N. J. Bax, M. Consalvey, and R. J. Kloser (2010). Seamount megabenthic assemblages fail to recover from trawling impacts. *Marine Ecology 31*(1), 183–199.

List of Figures

Figure 1: Detail of the sampling locations within the Huon Australian Marine Park, located south of Tasmania, Australia. These locations are those that are within the depth range of 485 m and 2015 m. Bathymetry is water depth (m), and TPI is 'topographic position index' and gives an indication of how elevated each cell is with respect to its neighbours (units of TPI are metres). The inclusion probabilities are those used to draw the sampling locations for the survey. The predicted values are from the model defined in Section 2.2, fitted to the original survey data whose locations are grey '+' on the bathymetry map. The image-frame size for prediction ($20m^2$) is arbitrary. The coordinate reference system used is WGS 84 / UTM zone 55S, with units of metres east and north.

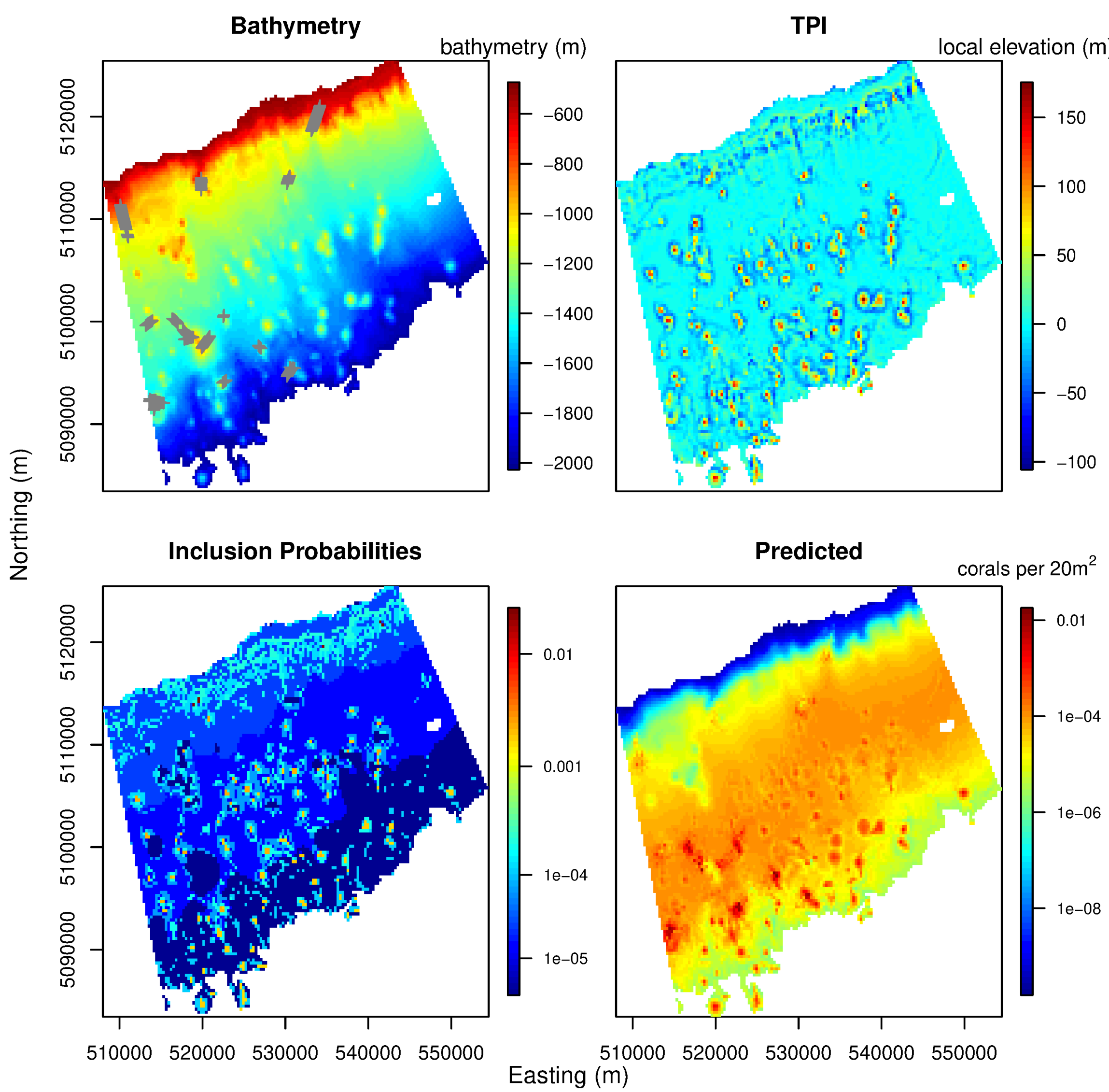

Figure 2: Results of the simulation experiment based on the survey of the Huon Australian Marine Park. Top row is for surveys with $n = 50$ sample locations, middle row with $n = 100$ and bottom row with $n = 200$. Left panels give, for each method and for each α, the median of the estimates from each of the $J = 1000$ simulated data sets. Right panels show the mean absolute deviation (MAD) estimate of variation of the same estimates. See Methods Section for the definition of percent difference and for the choice of reference. Solid grey line is 0% difference and dashed grey line is the median of the naive mean at $\alpha = 0$ (an unbiased estimator). Small values of α give more even inclusion probabilities.