`MBHdesign`: an `R`-package for efficient spatial survey designs

Scott D. Foster*

26 Nov 2020

Abstract

1. A considered survey design will generate data that is representative of the population that the sample is taken from. All good design takes is a little thought, some information and some good software tools.
2. Spatially-balanced randomisation with unequal inclusion probabilities is a modern and efficient design method. These designs are embedded within sampling theory and should be easy to generate.
3. The `R`-package `MBHdesign` allows field researchers easy access to these designs. It implements point-based and transect-based methods and allows for the incorporation of legacy sites.
4. The functionality of the package is illustrated with example designs in an environmentally heterogeneous region.

1 Introduction

Robust science can only be achieved using a rigorous and carefully planned scientific process (e.g. Leek & Peng 2015; Hayes *et al.* 2019). A pivotal link in this process is survey design, which transforms the research questions into a formal plan about how data is to be collected. It is important that the survey plan is designed so that the resulting data are: (i) representative of the population under investigation so that inference is valid; and (ii) information rich so that uncertainty, about inferences for the research question, is reduced. Without these attributes, surveys are less likely to deliver fit-for-purpose data (Hayes *et al.* 2019). To aid good survey design, the `MBHdesign` `R`-package is presented.

*Data61 CSIRO, Hobart, TAS, Australia, scott.foster@data61.csiro.au

The `MBHdesign` package leverages off the modern design strategy of spatial-balance (see Stevens & Olsen 2004) using Balanced Acceptance Sampling (BAS Robertson *et al.* 2013) with unequal inclusion probabilities (see Thompson 2012). Spatially-balanced designs are robust because they are based on randomisation, which guards against unintentional bias (see Altman 1991). The BAS method uses quasi-random numbers, which in the context of survey design can be treated as random. Spatial-balance increases efficiency in two ways, compared to simple randomisation. Firstly, it approximately balances over all spatially-smooth covariates that are not considered or even measured (Grafström & Lundström 2013). Secondly, when considering model-based analysis methods spatial-balance reduces the spatial autocorrelation between observations. Unequal inclusion probabilities increase efficiency by allowing researchers to increase sampling effort for certain environmental conditions, such as those that are likely to have higher variance (e.g. Thompson 2012). In ecology, where variance often increases with the mean, sites with higher abundance should be sampled more often. One consequence of unequal inclusion probabilities is that unweighted means will no longer be an unbiased description of the sample.

In addition to providing efficient BAS designs, `MBHdesign` makes accessible two extensions. Firstly, `MBHdesign` can accommodate *legacy sites*, which are often sites that have a long historical time-series and continuation of this time-series is beneficial. The legacy sites are incorporated using the methods described in Foster *et al.* (2017) and incorporates the legacy sites' locations into the spatially-balanced sample. Secondly, `MBHdesign` can sample units that are *transects* rather than a *points*. The method described in Foster *et al.* (2020) is implemented, which attempts to spatially-balance the centres of the transects whilst simultaneously respecting the inclusion probabilities for each point. Transect sampling is only available in `MBHdesign`.

Several packages exist for generating spatially-balanced designs (Kermorvant *et al.* 2019). These include: `spsurvey` (Kincaid *et al.* 2019) that implements the generalized random-tessellation stratified (GRTS) algorithm (Stevens & Olsen 2004); `SDraw` (McDonald & McDonald 2020) that implements a range of spatially-balanced methods including BAS, and; `BalancedSampling` (Grafström & Lisic 2019) that implements the local pivotal method (LPM Grafström 2012) and spatially correlated Poisson sampling (SCPS Grafström *et al.* 2012). All these packages provide a good platform for generating spatially-balanced designs, but they have different foci in terms of algorithms, functionality, scope, computation requirements and user-interface. `MBHdesign` is distinguished from these packages in that it provides field scientists with a tool that has a simplified yet flexible interface to generating designs that are: 1) superior spatial-balanced (unlike `spsurvey`), 2) based on unequal inclusion probabilities (unlike `SDraw`), and 3) based on computationally efficient methods that scale well to large problems (unlike `BalancedSampling`). In addition `MBHdesign` provides functionality for incorporating legacy sites (Foster *et al.* 2017) into the new survey design, as well as generating designs for transect-based

sampling platforms (Foster *et al.* 2020, but with extra computational demand). The package is freely available from the Comprehensive R Archive Network (CRAN), at https://cran.r-project.org/package=MBHdesign with a GNU GPL-3 license. This document was created with `MBHdesign` version 2.1.8.

2 The `MBHdesign` Package

The `MBHdesign` R-package is purposefully simple, with only three main functions and three 'helper' functions (see Table 1 for an overview). All functions have only a small number of arguments that are mandatory, but finer control can be achieved by changing default values.

Table 1: Functions available in `MBHdesign`. The first three functions are the primary functions in the package.

Function	Description
`quasiSamp`	Generate a BAS sample in arbitrary dimensions.
`alterInclProbs`	Adjusts inclusion probabilities to respect the locations of legacy sites.
`transectSamp`	Generates a spatially-balanced design for transect-based sampling.
`findDescendingTrans`	Finds transects within the survey area that run down gradients.
`findTransFromPoint`	Finds transects within survey area that originate from a given set of points.
`modEsti`	Simple model estimation method introduced in Foster *et al.* (2017).

`MBHdesign` takes a consistent and simple approach to spatial data: all functions can be called using a dense grid of points stored in a data frame (a non-spatial object). The data frame can be constructed using the `raster` package (Hijmans 2018), see supporting R-scripts for example code. It is recommended that the grid contains all locations within the extent of the area to be surveyed, and that an equal-area projection is used to maximise the efficacy of spatial balance. That is a (hyper-)rectangle that contains all the possible sampling locations. Those locations that are not to be sampled should be encoded as `NA`. For point-based designs, these locations can be deleted from the data set, but they are required for transect-based designs.

2.1 Hippolyte Rocks Examples

The functions within `MBHdesign` will be illustrated by way of creating designs for the marine environment surrounding the islands of the Hippolyte Rocks, Australia. For illustrative purposes, the bathymetric data

used here are a spatially-degraded version of those described in Nichol *et al.* (2009) and Spinoccia (2018), see Figure 1A. The exact degraded version of the data is also available (see Data and Code Availability). Bathymetry is a well-known delineator of marine biodiversity, and so inclusion probabilities are chosen to vary with depth. To achieve this, the inclusion probabilities are specified by: 1) defining 4 depth bins (Figure 1B), 2) stipulating that within each bin there is the expectation of the same number of samples, and 3) specifying the inclusion probabilities within each depth bin so sums within bins are equal across bins.

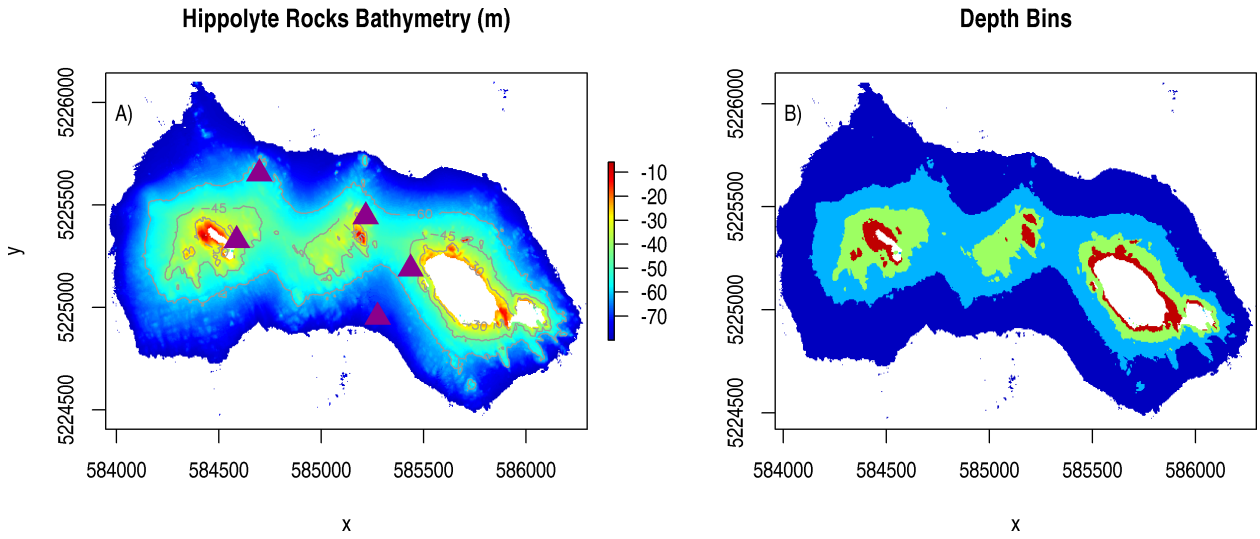

Figure 1: A) Bathymetry of the Hippolyte Rocks survey area. Purple triangles represent the locations of legacy sites that could be incorporated into a survey design. B) The depth bins within which the inclusion probabilities are constant.

2.1.1 Generating Point-Based Designs

The first design that will be generated is an equal probability spatially-balanced design for nSamp=100 survey sites. This design is appropriate when there is no information about the survey area, including which locations may have higher variance. With `bathy.df` being a data frame containing the grid of locations defining the survey area, the call is below, the locations are illustrated in Figure 2A and the first six rows of the return value is in Table 2.

```
#### A spatially-balanced sample within the study area (not depth-related)
evenSample <- quasiSamp( n=nSamp, potential.sites=bathy.df[,c("x","y")])
```

The second design is with unequal inclusion probabilities. To generate this design, the inclusion probabilities are passed to `quasiSamp()`. The resulting design is plotted in Figure 2B and is the result of the call:

Table 2: Four survey sites selected by the function quasiSamp for the even inclusion probability design. In order, the columns are: 'x' coordinate (e.g. longitude), 'y' coordinate (e.g. latitude), the inclusion probability that the site was selected with, and the row number from the potential.sites input argument. Inclusion probabilities, over all potential sites, will sum to 1 by construction.

x	y	inclusion.probabilities	ID
585996.7	5225038	9.1e-06	78146
584027.6	5225018	9.1e-06	79783
585791.7	5225523	9.1e-06	25518
585590.4	5225588	9.1e-06	18839

```
#### A spatially-balanced sample with shallow locations preferred
unevenSample <- quasiSamp( n=nSamp, potential.sites=bathy.df[,c("x","y")],
                                  inclusion.probs=bathy.df[,"inclusion.prob"])
```

The third design additionally incorporates the spatial locations of existing legacy sites (Foster *et al.* 2020). This involves a two-step procedure: 1) alter the inclusion probabilities to down-weight locations near the legacy sites, and 2) take a sample with the altered inclusion probabilities.

```
#adjust the inclusion probabilities for the locations of the legacy sites
bathy.df$altered.inclusion.prob <- alterInclProbs( as.matrix( legacySites),
                potential.sites=bathy.df[,c("x","y")],
                inclusion.probs=(nSamp-nrow(legacySites))*bathy.df[,"inclusion.prob"],
                mc.cores=8)
#a spatially-balanced sample with legacy sites and a preference for shallow locations
unevenLegacySample <- quasiSamp( n=nSamp-nrow(legacySites),
                potential.sites=bathy.df[,c("x","y")],
                inclusion.probs=bathy.df[,"altered.inclusion.prob"])
```

2.1.2 Generating Transect-Based Designs

Generating transect-based designs requires specieal methods as the task of respecting (point-based) inclusion probabilities using a transect sample is non-trivial. `MBHdesign` employes the method of Foster *et al.* (2020), who choose transects based on inclusion probabilities of the cells that the transects intersect. Transects are then chosen based on these derived inclusion probabilities. Complex survey areas and spatially-complex inclusion probabilities maps will mean that finding transect designs is a difficult and computationally demanding task. A design for `nSamp=12` transects is shown in Figure 3A, and is generated using:

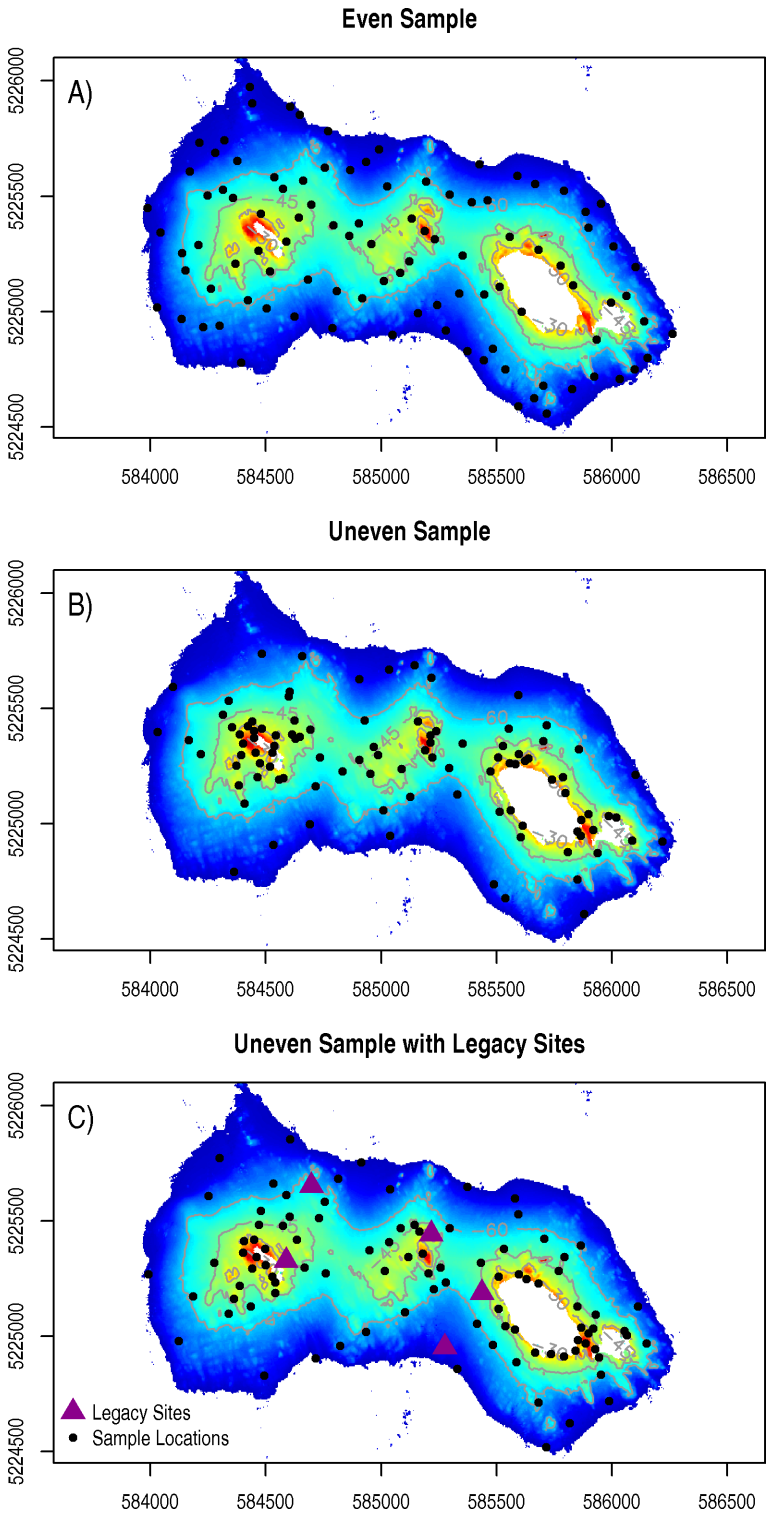

Figure 2: Example point-based designs. A) a spatially-balanced design with equal inclusion probabilities. B) a spatially-balanced design with unequal inclusion probabilities. C) a spatially-balanced design with unequal inclusion probabilities that incorporates 5 legacy sites.

Table 3: Four points (of 120) on transects selected by transSamp. Stored in the second element of the return object. In order, the columns are: transect number, the coordinates of the transects' midpoints, the compass bearing of the transect, the points on the transect, the user-defined inclusion probability and the (internal) probability of selection to maintain the user-defined probabilities.

transect	mid_x	mid_y	direction	x	y	specifiedInclProb	AdjustedInclProb
12	584573.0	5225438	51.42857	584579.5	5225443	0.0005543	0.0004217
12	584573.0	5225438	51.42857	584553.4	5225423	0.0006148	0.0000000
11	584298.5	5225288	85.71429	584273.5	5225286	0.0004108	0.0001009
7	585634.4	5224713	51.42857	585614.8	5224698	0.0001113	0.0000662

```
#The representation of transects and other algorithm controls
control <- list( transect.pattern='line', transect.nPts=10, nRotate=21,
                                                line.length=150, mc.cores=8)
#a spatially-balanced sample with shallow locations preferred
transSample <- transectSamp( n=nSamp, potential.sites=incl.prob[,c("x","y")],
                             inclusion.probs=incl.prob[,"inclusion.prob"], control=control)
```

The control argument contains information about the type and representation of the transects, as well as information about the underlying algorithm. In this example, the 150 m linear transects are represented by `transect.nPts=10` discrete points and `nRotate=21` different rotations (directions) are considered from each cell within the survey area. Increasing `transect.nPts` and `nRotate` will increase the ability of the algorithm to respect the specified inclusion probabilities, but at the expense of computation. The value of `line` for transect.pattern is a special case, in general its value should be a matrix that represents the (centered) shape of the transect. For a non-linear example, see the supporting information file 'VisualIllustration6.R') of Foster *et al.* (2020).

The remaining arguments are similar to those in `quasiSamp()`. The return value of transectSamp is a list of two elements, each of which is structured similarly to the return object of quasiSamp (Table 3). The first element of the `transectSamp()` return list are the nSamp central locations of the selected transects. The columns are the same as in Table 2 but contain additional information about the transect midpoints: the ID number and the direction from that point. The second element of the return list is for the points on each of the `nSamp` transects, see Table 3. The two types of inclusion probabilities returned are: those specified by the user, and those which the algorithm has used to try and obtain transect samples with the user-supplied probabilities. See Foster *et al.* (2020) for details on the altering process, and see the package vignette for an illustration of its effect.

It may be important to know that a transect's inclusion probability is just the sum of the inclusion probabilities

for the points that represent it. This means that a transect's inclusion probability may be high even though one, or more, of the inclusion probabilities for its constituent points is very low or even zero. This raises a design decision: whether areas just outside the survey area should be excluded completely (inclusion probability of `NA`) or should they be included if the neighbouring sites are sufficiently important (inclusion probability of 0).

It is possible to place constraints on the transects, to remove the chance of obtaining a transect that cannot be executed in the field. As an example, when sampling with towed underwater video (ToV), it is standard to only perform transects that are downhill from the start location. Another example is sampling seamounts, again using ToV, where transects are sometimes taken from a seamount's peak (e.g. Williams *et al.* 2020). There are specific functions within `MBHdesign` to perform both these specific tasks (`findDescendingTrans()` and `findTransFromPoint()`), which return a matrix of text values that describe the behaviour of the transect, which can be then used within `transectSamp()`. Note that with more constraints the inclusion probabilities become less and less well-respected / approximated. Example designs that incorporate these two constraints are given in Figures 3B and 3C.

3 Summary and Discussion

An ubiquitous and fundamental aspect of field ecology is survey design. To aid generation of efficient designs, the `MBHdesign` `R`-package is introduced. The designs generated are spatially-balanced and allow for unequal inclusion probabilities. The package offers a simple interface whilst remaining flexible enough to provide a rich suite of designs.

Whilst *design* is the primary consideration of `MBHdesign`, *inference* is only briefly considered (using the supplied function `modEsti` for inference with legacy sites). It is noted however, that there are many different, valid, design-based and model-based inferential methods for the data that point-based designs from `MBHdesign` will generate. All approaches will benefit from an efficient design. Foster *et al.* (2020) suggested that model-based analyses for transect-based surveys were necessary, because design-based analyses were currently unknown for this type of randomisation. Such analyses should take into account spatial autocorrelation.

The implemented algorithm for generating point-based designs (BAS; Robertson *et al.* 2013) is computationally thrifty. However, the algorithm for transect-based designs (Foster *et al.* 2020) is computationally more expensive. This is partly due to the relative infancy of transect-based approaches, especially compared to points.

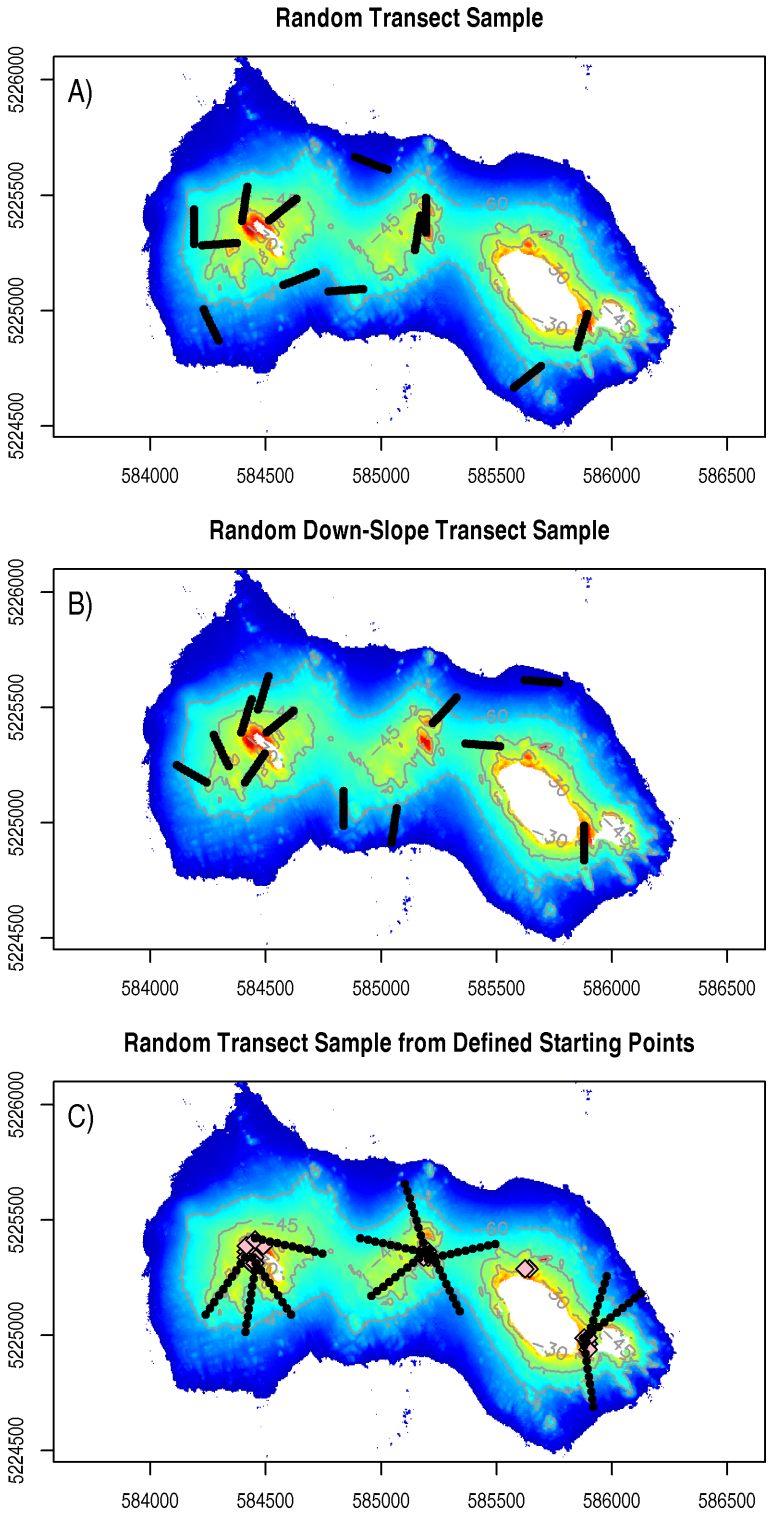

Figure 3: Example transect-based designs. A) A spatially-balanced design with unequal inclusion probabilities. B) Like A) but with transects that are downhill. C) Like A) but only with transects that start in the shallowest 20 grid cells. In all panels, black dots represent way-points on the transect and in C) pink diamonds represent the set of potential start locations.

The code-base of `MBHdesign` is freely available from CRAN (https://cran.r-project.org/package=MBHdesign) and this article was created using version 2.1.8.

4 Acknowledgements

I would like to thank Emma Lawrence, Jac Monk, Geoff Hosack, Keith Hayes and Skip Woolley for advice and support. This work was undertaken for the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Science Program.

5 Data and Code Availability

The Zenodo repository (Foster 2020) contains the Hippolyte Rocks bathymetry data, and the R-code used to create the examples in this work. The data are a degraded version of that obtained from Spinoccia (2018). The MBHdesign package is available from https://cran.r-project.org/package=MBHdesign.

6 Author Contributions

S Foster is responsible for all parts of this work.

References

Altman, D. (1991). Randomisation. *BMJ*, **302**, 1481.

Foster, S.D. (2020). Code and data required to run examples in MBHdesign description paper. Retrieved from https://doi.org/10.5281/zenodo.4291227

Foster, S.D., Hosack, G.R., Lawrence, E., Przeslawski, R., Hedge, P., Caley, M.J., Barrett, N.S., Williams, A., Li, J., Lynch, T., Dambacher, J.M., Sweatman, H.P. & Hayes, K.R. (2017). Spatially balanced designs that incorporate legacy sites. *Methods in Ecology and Evolution*, **8**, 1433–1442.

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

Grafström, A. (2012). Spatially correlated poisson sampling. *Journal of Statistical Planning and Inference*,

142, 139–147.

Grafström, A. & Lisic, J. (2019). *BalancedSampling: Balanced and spatially balanced sampling*.

Grafström, A. & Lundström, N.L.P. (2013). Why well spread probability samples are balanced. *Open Journal of Statistics*, **3**, 36–41.

Grafström, A., Lundström, N.L.P. & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, **68**, 514–520.

Hayes, K.R., Hosack, G.R., Lawrence, E., Hedge, P., Barrett, N.S., Przeslawski, R., Caley, M.J. & Foster, S.D. (2019). Designing monitoring programs for marine protected areas within an evidence based decision making paradigm. *Frontiers in Marine Science*, **6**, 746.

Hijmans, R.J. (2018). *Raster: Geographic data analysis and modeling*.

Kermorvant, C., D'Amico, F., Bru, N., Caill-Milly, N. & Robertson, B. (2019). Spatially balanced sampling designs for environmental surveys. *Environmental Monitoring and Assessment*, **191**.

Kincaid, T.M., Olsen, A.R. & Weber, M.H. (2019). *Spsurvey: Spatial survey design and analysis*.

Leek, J.T. & Peng, R.D. (2015). What is the question? *Science*.

McDonald, T. & McDonald, A. (2020). *SDraw: Spatially balanced samples of spatial objects*.

Nichol, S., Anderson, T., McArthur, M., Heap, A., Siwabessy, P. & Brooke, B. (2009). *Southeast Tasmania temperate reef survey. Post survey report record 2009/43.* Geoscience Australia, Canberra.

Robertson, B.L., Brown, J.A., McDonald, T. & Jaksons, P. (2013). BAS: Balanced acceptance sampling of natural resources. *Biometrics*, **69**, 776–784.

Spinoccia, M. (2018). AusSeabed bathymetry holdings. Retrieved July 14, 2020, from http://pid.geoscience.gov.au/dataset/ga/116321

Stevens, D. & Olsen, A. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, **99**, 262–278.

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

Williams, A., Althaus, F., Green, M., Maguire, K., Untiedt, C., Mortimer, N., Jackett, C.J., Clark, M., Bax, N., Pitcher, R. & Schlacher, T. (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.