SJ McNeill, N Odgers, L Lilburne, S Carrick, S Hainsworth, S Fraser, C Hedley, P Roudier, G Grealish 2018. Spatial sampling tools for representative soil sampling. In: *Farm environmental planning – Science, policy and practice*. (Eds L. D. Currie and C. L. Christensen). http://flrc.massey.ac.nz/publications.html. Occasional Report No. 31. Fertilizer and Lime Research Centre, Massey University, Palmerston North, New Zealand. 8 pages.

SPATIAL SAMPLING TOOLS FOR REPRESENTATIVE SOIL SAMPLING

S J McNeill[1], N Odgers[1], L Lilburne[1], S Carrick[1], S Hainsworth[2], S Fraser[3], C Hedley[2], P Roudier[2], G Grealish[2]

[1] Manaaki Whenua Landcare Research, PO Box 69040, Lincoln
[2] Manaaki Whenua Landcare Research, Private Bag 11052, Palmerston North
Email: mcneills@landcareresearch.co.nz

Abstract

Soil surveys provide information for farm environment plans, regional surveys, and national inventories. For large scale surveys of soil properties, field sampling is typically carried out on a selection of sites that are intended to be representative of the soil information on the landscape. There are many considerations involved in the decision where representative samples should be sited, dictated by the resource cost involved, site access restrictions, practical aspects of the soil sampling protocol, and the statistical efficiency of the samples that are gathered. While assessment of the "best" sample design for a survey can be undertaken by a statistician, it is often useful for the field scientist to understand how different sampling approaches affect the distribution of samples, and how compromises in sample designs affect the sampling result. Software packages for spatial sampling design have improved over recent years, but they can still be complicated for many who are not specialists in sampling design.

This paper reports on the development of a graphical user interface (GUI) application tool to assist the field scientist understand some of the compromises involved in using different spatial sampling strategies. Using a shapefile of the study area as input, different spatial sampling strategies can be applied to select one or more versions of the required samples, and display them in the study area. Simple summary statistics are provided to provide some comparison between the methods. The completed design can be saved directly to an Excel spreadsheet with map coordinates for further analysis. In simple cases, the saved results can be used for direct use in the field. The intention is that the application can be used to improve understanding of the compromises involved in selecting sampling designs, and allows the field scientist more direct involvement in site selection, to allow sampling design options to be tested in terms of practical achievement and implications on statistical robustness.

Introduction

Soil surveys provide critical information for farm environment plans delineating soil patterns, as well as regional surveys, and national inventories detailing broad-scale soil properties. The analysis from these efforts typically requires a set of samples designed to be representative of the soil information on the landscape. In this sense, a representative sample can be defined as a group of points whose soil characteristics closely match the characteristics of the population of all soils in the study area. For example, if what is required is an estimate of some physical property (e.g. pH, carbon, Cd concentration) from soils in the high country grasslands, then the field study points should all be from high country grasslands, with characteristics that

should broadly match those in the whole population. In other words, the samples should be an accurate reflection of the population from which the sample is drawn.

The historical assembly of soil information has in some cases involved expert assessment of suitable locations for field study; these are judgement or purposive samples. While acceptable as a means of characterising the soil resource, purposive sample designs can be subject to unintentional bias, and the overall variability of the soil properties may be under-represented as a result (Kuehl 2000; de Gruitjer et al 2006). Random sample designs, by contrast, guarantee an unbiased assessment of the soil properties.

Since soil sampling can be expensive due to the time involved in the field effort and subsequent analysis in the laboratory and computer, there is considerable incentive to try and reduce the overall resources required to carry out the survey. The determination of what is the "best" sample design for a survey depends on how one frames the problem, but as a rough rule of thumb the cost scales with the number of field points involved, hence the desire to reduce the number of samples involved to as few as can be tolerated. The precise calculations involved depend on the nature of the study, and the type of inference required. For instance, one might wish to know whether a soil trace element concentration in a certain land use class is increasing over time (or decreasing, or remaining unchanged) and the number of samples involved in that case depends on information concerning the spatial variability of the trace element concentration as well as its variability over time. Complex designs for which a certain inference power are required (Stroup 2002) may require detailed analysis by a statistician, but simple designs can often be executed by the field scientist, provided the tools are available to perform the task. Even if the task is quite complex, it is useful for the field scientist to understand how different spatial sampling approaches affect the distribution of samples, and how compromises in sample designs could affect the sampling result. In that way, the final sampling design can be made as efficient as possible, whether executed by the field scientist, or as a result of collaboration between the field scientist and statistician.

Software packages for spatial sampling design have improved in recent years, and there are a variety of tools available to help in power analysis and sample design. Examples include tools for spatial power analysis (Stroup 2002, Barry & Maxwell 2018), and for generating randomised spatial designs (Walvoort et al 2010; McDonald 2018). However, these methods may be too complicated for users who simply wish to visualise their simple spatial designs. This paper reports on the development of a graphical application written in the R language (R Core Team 2017) that allows the field scientist to experiment with a number of different spatial sampling strategies in a study area.

Method

Overview of spatial sampling designs

There are many methods for spatial sampling designs available in the literature (Green 1979) and it is beyond the scope of this paper to describe all of them or even to provide a broad range of those available. In this paper, we restrict ourselves to area-based sampling, so that samples can be placed anywhere within some study area. It is, however, possible to infer spatial characteristics by sampling along a curve (e.g. off a roadside), but for simplicity we do not address those methods here.

It is also well known that the accuracy of the measured soil property will usually be improved by dispersing the sample locations so that they cover the study area as uniformly as possible (Cochran 1977). The simplest and most effective way to achieve this is to select samples on a

grid with equal spacing between points (i.e. grid sampling), but this results in high prediction error near the study area border (Walvoort et al 2010), an effect that becomes exacerbated for certain types of study areas. This difficulty can be reduced by placing some samples near the border, which of course disturbs the regular grid pattern. Another way to spread samples is to randomly sample over space (i.e. simple random sampling), and accept samples that happen to fall inside the study area (this is a "hit or miss" approach), which can unfortunately result in some selected points that are too close together. Grid and simple random sampling are well-known, but tend to perform quite poorly for prediction of spatial soil properties (Walvoort et al 2010).

Another more recent approach generates samples that are purposely designed to be well-spread over the study area, giving a class of methods known as spatially balanced designs. There are many spatially balanced design methods (Wang et al 2013), but two of the most popular are Generalised Random Tessellation Stratified (GRTS) sampling (Stevens & Olsen 2004) and Balanced Acceptance Sampling (BAS) (Robertson et al 2013). The general concept of these approaches is to transform the two-dimensional space into a one-dimensional line, from which a systematic sample is drawn. The locations of the balanced samples can then be obtained by inverse-mapping from one-dimensional sampling space back to two-dimensional space. The resultant set of samples is balanced, with all sub-regions being represented with equal probability.

Requirements and implementation

A survey of sampling requirements from recent projects within Manaaki Whenua Landcare Research (McNeill et al 2017) suggested that field scientists preferred tools that enabled them to understand and visualise sampling decisions that were made for their area of interest. While the final design of a complex study might require a detailed analysis by a statistician, it was considered helpful that simple designs could be trialled by the field scientist.

To make the implementation of the tool simple, it was decided to implement these requirements in the R-Shiny GUI framework (Chang 2018). This approach results in an application that runs in a standard web browser (see Figure 1), and if the GUI is implemented in a remote R server, the approach does not require the user to even have the R software installed, and the application can even run on the web browser on a simple mobile phone.

The R-Shiny approach has both advantages and disadvantages from the viewpoint of development. The principle advantage is that all of the packages that are available in the R programming environment are available, and the application makes strong use of the leaflet (Cheng et al 2017) and Sdraw (McDonald 2018) packages for visualisation and sample selection. Another advantage is that Javascript and HTML code can be included in the GUI, providing support for web links and certain helpful graphical features (e.g. a "busy" indicator). From a user's perspective, the application is run using the web address https://mcneills.shinyapps.io/spatialsampler. This method of invoking the application involves the restriction from the providers of R-Shiny that the total amount of processing time is limited, so designs using very complex spatial shapefiles might not complete. However, no limit has been found during testing to date.

Once sample points have been generated, simple statistics are displayed on screen, but most users would most likely wish to process the data separately. The data can be downloaded directly to an Excel spreadsheet, with the sample locations returned in both geographic

(WGS84 latitude and longitude) as well as New Zealand Transverse Mercator (NZTM) coordinates.

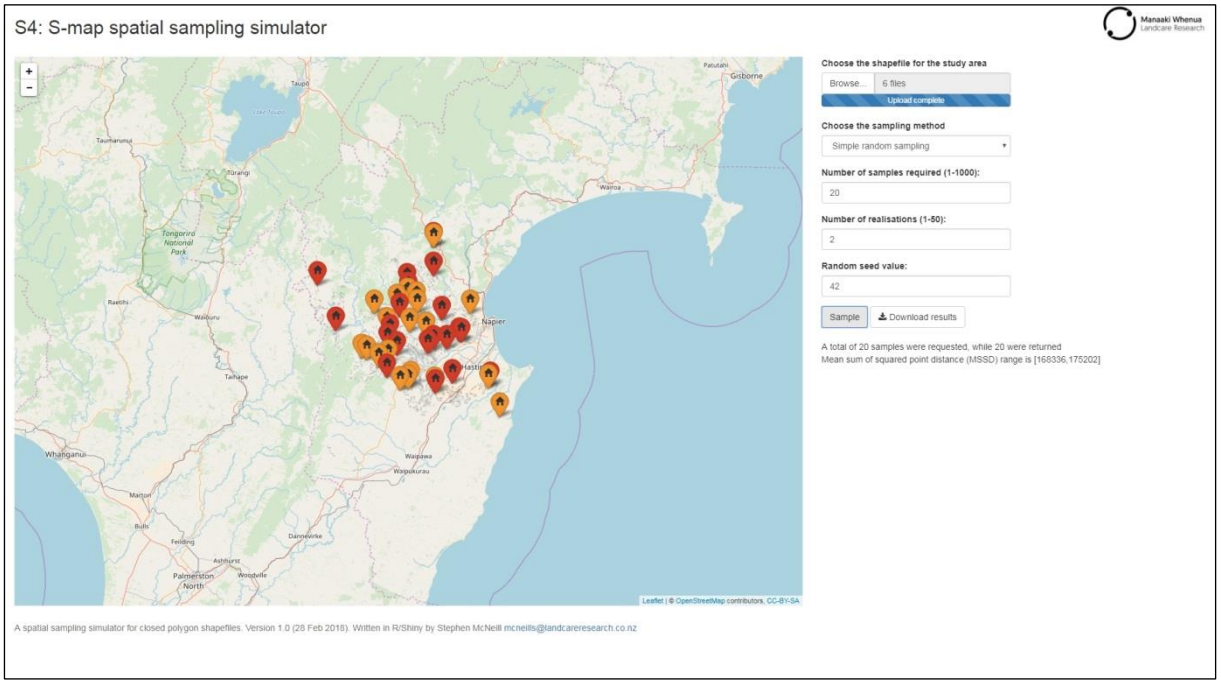

Figure 1: Screenshot of the GUI. The sample locations for two realisations of 20 samples using simple random sampling are shown with red or orange markers for the different realisations.

Results

A case study

To illustrate how the application can be used to investigate different sampling strategies, we use a shapefile from a study area slightly adapted from a recent study by one of us. The shapefile, which covers an area in the Hawkes Bay region, is derived from the road network, and the valid area for sampling is between 50 and 400m from roads and farm tracks, derived using LINZ Topo50 road centrelines, LINZ Topo50 farm tracks, and district council road centrelines. The study area excludes dense vegetative cover from the Land Cover Data Base, as well as land above 35 degrees slope. The study area is shown shaded in grey in Figure 2. One complication in this study area is that there are several small "islands" within the study area some distance away from the study area bulk (e.g. south east of Havelock North), which must also have equal probability of sample coverage.

The study area is a good example where it is quite difficult to spread samples spatially since the boundary length is long compared with the total area covered (i.e. the area is non-compact). This contrasts with compact shapes (e.g. a rectangular field), where the candidate sites for sampling can be uniformly spaced within the shape and most points will be far from the boundary. In such cases, the poor performance of grid-based methods is minimised.

To illustrate some features of some of the sampling methods, we selected 20 samples from regular grid sampling and BAS. In practise, more samples would be required to give a realistic account of the soil properties in the region, but selecting a small number exacerbates the relative advantages and disadvantages of the two methods.

Figure 2: Study area, shaded grey, within the Hawkes Bay region.

Figure 3 shows the result of requesting 20 samples from regular grid sampling and balanced acceptance sampling. Figure 3(a) shows that a total of 23 samples were generated, and this feature of obtaining more (or fewer) samples than requested is a feature of grid sampling, due to the random selection of the starting point in the grid (Walvoort 2010). In practise, one would probably continue generating new versions of the regular grid samples until a set was obtained with exactly 20 samples. Another feature of grid sampling is that the local density of points depends critically on how the study area aligns with the grid itself. The BAS sample in Figure 3(b), by contrast, returns exactly 20 samples, and the points tend to be well spread across the study area.

Comparing sampling schemes on the basis of the visual spread of the samples is difficult if the number of samples becomes large, since in that case the sample positions appear to be random and appear to fill the study area. More insight can be gained by calculating diagnostic measures from the selected samples. Figure 4(a) shows the average of the median distance from points in each of three sampling schemes when 50 realisations of 100 samples are sought from the study area in Figure 2. Figure 4(b) shows the average of the mean distance from points in the same study area, again over 50 different random realisations.

Since points must be spaced a multiple of a fixed distance in grid sampling (see Figure 3(a)), the average median distance from a point to its nearest neighbour will be one of those fixed distances, as evidenced in Figure 4(a). For simple random sampling, there is a chance that a randomly-selected point is close to a point that has been previously chosen, since no account is taken of previous points selected, so the average value of the median distance from a point to its neighbour will be smaller when compared with grid or BAS. Thus, the points are poorly distributed over the study area as a result. From the viewpoint of the average distance from a point to its nearest neighbour alone (whether using the median or the mean distance), there is very little difference between grid and BAS.

(a) (b)

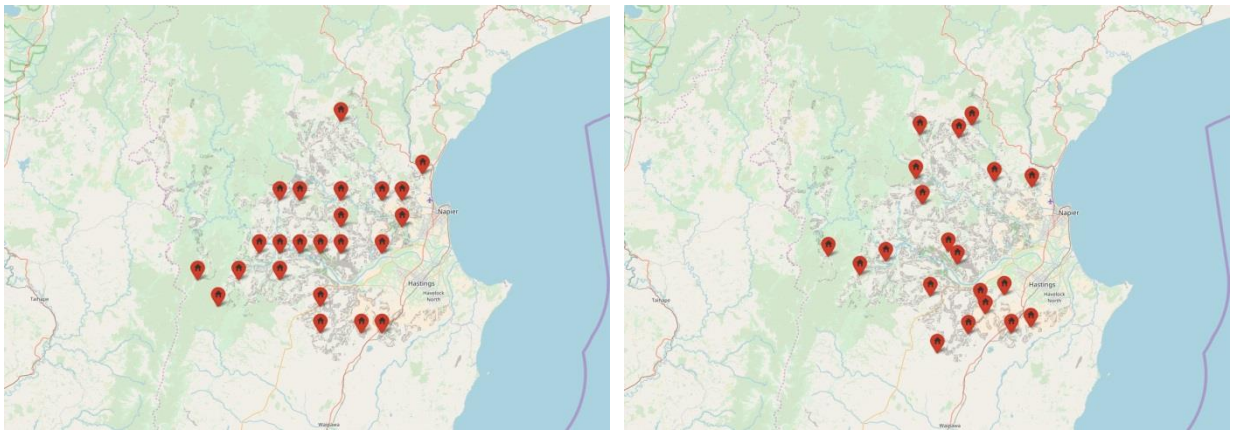

Figure 3: Selection of 20 samples in the study area from two different methods, (a) Regular grid sampling, and (b) Balanced acceptance sampling.

Balanced sampling methods (GRTS or BAS) have advantages when compared with other sampling methods that are not obvious from Figure 4. Grid sampling has the disadvantage that the information from the sample set can only be obtained to the scale of the grid itself, and relationships at a finer scale are not available (Brown et al 2015). Since all positions within the study area in Figure 2 have a finite probability of inclusion, analysis of samples from BAS can be used to analyse relationships at finer scales than possible with a grid approach, but with greater spread efficiency (i.e. larger point-to-point distance) than is possible with simple random sampling. This is an important consideration if it is known that the soil properties vary at a broad range of scales, from fine- to coarse-scale.

Discussion

Limitations of the application

While the application cannot provide all the relevant diagnostics for the above study area, it provides a mechanism for the study designer to experiment with different sampling schemes and determine if there are problems using one particular approach. Since the application can save the sampling points to an Excel spreadsheet, specialised diagnostics (e.g. Figure 4) can be calculated externally in a spreadsheet or other analysis program.

The application has limitations concerning the complexity of the sampling designs possible, at least at present. It is assumed that the study area is a single stratum, so stratified sampling cannot be accommodated. It is also assumed that the sampling effort is known, so power analysis is not provided. This latter constraint recognises that many soil surveys have a fixed budget, so the field effort is often pre-determined and what is required is to arrange the available spatial samples so that the best geographic spread is obtained.

Finally, one limitation is that the shapefile must lie within New Zealand, simply because exporting of coordinates to NZTM is only valid in this region. This restriction may be eased in the future.

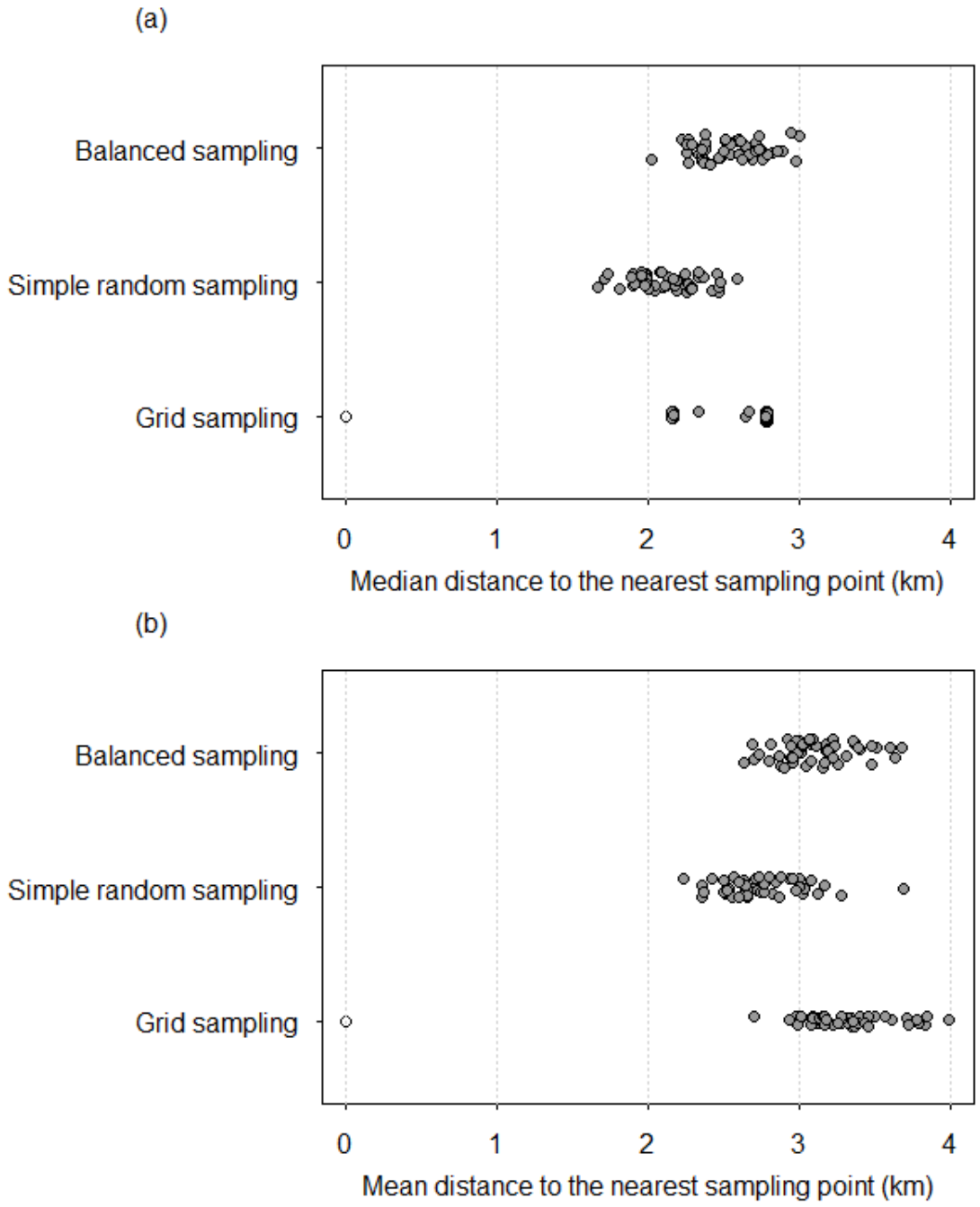

Figure 4: (a) Average median distance between points for three spatial sample designs in the study area, from a total of 50 different realisations. (b) Average mean distance between points for the same study as (a).

Conclusions

A simple application has been written in the R-Shiny GUI framework that allows a user without specialised software expertise to test out various spatial sample designs on a shapefile within New Zealand. While the application does not address all possible requirements that a field scientist might have (e.g. power analysis) exporting the sample design to an Excel spreadsheet permits more complicated diagnostics and analysis to be carried out.

The application has certain limitations as a result of its design, and it cannot handle stratified designs. Additional features may be added in the future, depending on the needs of users.

Acknowledgements
This research was funded by the Ministry of Business, Innovation and Employment's Science and Innovation Group.

References

Barry J, Maxwell D. 2018. emon: Tools for Environmental and Ecological Survey Design. https://cran.r-project.org/package=emon. Accessed February 2018.

Brown JA, Robertson BL, McDonald T. 2015. Spatially balanced sampling: application to environmental surveys. *Procedia Environmental Sciences*, 27:6-9.

Chang W. 2018. shiny: Web Application Framework for R. https://cran.r-project.org/package=shiny. Accessed February 2018.

Cheng J, Karambelkar B, Xie Y. 2017. leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 1.1.0. https://CRAN.R-project.org/package=leaflet. Accessed February 2018.

de Gruitjer J, Brus D, Bierkens M, Knotters M. 2006. Sampling for natural resource monitoring. Springer, Berlin, 332p.

Green RH. 1979. Sampling Design and Statistical Methods for Environmental Biologists. John Wiley & Sons, 257p.

Kuehl RO. 2000. Design of experiments: Statistical principles of research design and analysis. Second edition, Duxbury Press.

McDonald T. 2018. SDraw: Spatially Balanced Sample Draws for Spatial Objects. https://cran.r-project.org/package=SDraw. Accessed February 2018.

McNeill S, Lilburne L, Carrick S, Hedley C, Stevenson B, Mudge P, Hainsworth S. 2017. Improvements in spatial soil sample design efficiency, Poster presentation for Pedometrics 2017, Wageningen, Netherlands.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Robertson BL, Brown JA, McDonald T, Jaksons P. 2013. BAS: Balanced acceptance sampling of natural resources. *Biometrics* 69:776-784.

Walvoort DJJ, Brus DJ, de Gruitjer JJ. 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers and Geosciences*, 36:1261-1267.

Wang JF, Stein A, Gao BB, Ge Y. 2012. A review of spatial sampling. *Spatial Statistics*. 2:1-14.