A Julia package for farm-scale soil carbon auditing. User's Manual – Version 1.01

J.J. de Gruijter*, I. Wheeler*, B.P. Malone[a,*]

[a] *Faculty of Agriculture and Environment, Biomedical Building C81, The University of Sydney, NSW 2006, Australia*

1. Introduction

The methodology of optimal spatial stratification by package *ospats+* is documented in de Gruijter et al. (2016). Users are encouraged to read this article before using the software.

It is assumed here that the user has installed Julia, freely available from https://julialang.org/.

Package *ospats+* can be downloaded from https://github.com/jjdegruijter/ospats+. It is ready to be used, assuming that Julia has been installed. No other package dependencies are needed, except for the Julia packages CSV and DataFrames. For the latter, simply do Julia > Pkg.add("CSV") and Julia > Pkg.add("DataFrames).

The package contains four script files: "main", "readdata", "ospats" and "ospall".

Script "main" first serves to fill in all process parameters by the user (see below), it then invokes the other three scripts. Script "readdata" reads the datafile mentioned in "main". Scripts "ospats" and "ospall" produce both an optimal design using the datafile and the process parameters. The difference is that "ospats" optimises by iterative re-allocation of all N grid points, while "ospall" re-allocates only a sample of the grid points, to avoid working with an $N \times N$ matrix of generalised distances in case of very large grids. After a sample of grid points has been stratified, "ospall" continues by (once and

*Corresponding author

Email addresses: `jjdegruijter468@gmail.com` (J.J. de Gruijter), `ichsani.wheeler@gmail.com` (I. Wheeler), `brendan.malone@sydney.edu.au` (B.P. Malone)

definitively) allocating the remaining grid points to the sample strata, using the same optimisation criterion as "ospats".

The requested function is called by setting the parameter *in* (from sampling INterval) in the 'main' script. If $in = 1$, then *ospats*() is called. If $in \geq 2$, then *ospall*() is called. For instance, if $in = 3$, then every third grid point gets included in the sample.

2. Initialisation

Before running the script 'main' the following data must be filled in.

Fill in $filename$: name of the file with the grid data.

> *Explanation*: A text file with 5 comma-separated columns without headers, and N (the number of grid points) rows. Default order of the columns is: x-coordinate, y-coordinate, grid point identification, prediction, prediction error variance. In case of a different order, adapt the column numbers in the function *readdata*. If the data file is incomplete, i.e. not all columns have the same length, then Julia issues a LoadError.

Fill in the following process parameters:
(Note: If an invalid number has been filled in for a parameter (out of range, or not an integer when an integer is expected), then 'main' issues an error message and stops Julia.)

$H_{\min}$: smallest acceptable number of strata.

$H_{\max}$: largest number of strata still assumed to be possibly optimal.

nh_{minim}: smallest sample size allowed within the strata.

CP: carbon offset price, in currency unit (e.g. Aus \$) per Mg.

f: predicted average cost of obtaining data per grid point, in currency unit.

$Area$: surface area of the farm (ha).

Z_γ: quantile of the standard normal distribution (1.645 for the 95% quantile).

$R2$: squared multiple correlation coefficient from the regression model used to generate the predictions.

range: parameter of an exponential auto-covariance function fitted to the prediction residuals.

maxcycle: the maximum number of iteration cycles allowed for iterative re-allocation.

> *Explanation*: This is intended as a safe-guard against unforeseen endless looping. In our experiments the number of iteration cycles needed to fully complete the re-allocation process not yet exceeded 100. The setting $maxcycle = 0$ forces the system to skip the iterative re-allocation, and to proceed with the final calculations on the initial stratification.

in: interval used to draw a systematic sample from the grid.

> *Explanation*: if $in = 1$ then function "ospats" will be called, which optimises a stratification for the entire grid. If $in > 1$ then function "ospall" will be called, which optimises a stratification for a sample from the grid, i.e. after coarse-gridding. The size of the sample is determined by in. For instance, if $in = 10$ then every 10th point is included in the sample, starting with a randomly chosen first point. In principle, the sample size should be taken as large as computer capacity allows for calculating the $N \times N$ matrix of generalised distances. Without recourse to super-computing, that will be in the order of some thousands for a computing size of one 2.5 GHz IntelCore i5 processor and 4 RAM.

seed: seed for the pseudo-random number generator, used to randomise sequences of grid points and to draw a stratified random sample.

3. Starting the optimisation process

The optimization process is started by running 'main', which calls first "readdata" and then either "ospats" or "ospall", depending on the setting of parameter in.

```
File name :  Nowley.txt
Grid size : 4381
Seed for random number generator : 1234
Sampling interval :  3
R2 :  0.36
Range :  582
Maximum number of iteration cycles :  150
Minimum sample size allowed in the strata : 3
Minimum number of strata : 3
Maximum number of strata : 7
Carbon offset price (Aus dollar per Mg) : 10
Cost of obtaining data per grid point (Aus dollar) : 120
Surface area of the farm (ha) : 2336
Calling function ospall
-------------------- START FUNCTION OSPALL -----------------------
sample size = 1460
______________________________ Number of strata : 7
Total number of transfers = 2908
Number of iteration cycles = 21
Intermediate results :
Sample size :   55
Sample sizes in strata : [1.0 15.0 16.0 3.0 11.0 7.0 2.0]
Smallest sample size allocated to a stratum :  1.0
______________________________ Number of strata : 6
Total number of transfers = 2931
Number of iteration cycles = 14
Intermediate results :
Sample size :   58
Sample sizes in strata : [15.0 3.0 9.0 19.0 4.0 8.0]
Smallest sample size allocated to a stratum :  3.0

FINAL RESULTS:
Number of strata : 6
Total sample size : 58
Sample sizes in strata : [15.0 3.0 9.0 19.0 4.0 8.0]
Smallest sample size allocated to a stratum :  3.0
Size of grid-strata : [1085 220 701 1219 454 702]
---- END OF FUNCTION OSPALL ----
```

Figure 1: Process parameters and results of design optimisation with grid data Nowley.txt.

```
SampleNr,StratNr,PointNr,X,Y
1,1,2285,1.51390554010775e6,-3.63485161267572e6
2,1,2976,1.51480554010775e6,-3.63537661267572e6
3,1,1909,1.51240554010775e6,-3.63455161267572e6
4,1,2703,1.51675554010775e6,-3.63515161267572e6
5,1,2417,1.51690554010775e6,-3.63492661267572e6
.
.
.
54,6,2861,1.51375554010775e6,-3.63530161267572e6
55,6,2811,1.51750554010775e6,-3.63522661267572e6
56,6,3185,1.51510554010775e6,-3.63552661267572e6
57,6,3067,1.51398054010775e6,-3.63545161267572e6
58,6,3065,1.51383054010775e6,-3.63545161267572e6
```

Figure 2: First and last 5 sample points of the sample generated by *ospats+* with grid data Nowley.txt and process parameters as in Fig. 1

4. Output

The resulting stratification is written in file "Stratification", with three columns: x-coordinate, y-coordinate and stratum number for each of the grid points. The present version of *ospats+* does not provide a map of the stratification. The stratified random sample is written in file "Sample" with five columns: sample number, stratum number, grid point number, x-coordinate and y-coordinate.

As shown in Figure 1, along with the settings of the process parameters and intermediate optimisation results, the following final results are displayed on screen:
1) Number of strata;
2) Total sample size;
3) Sample sizes allocated to the strata;
4) Smallest sample size allocated to a stratum;
5) Sizes of the grid strata.

5. Replication test

To check if the package as installed works correctly, run *ospats+* on the test data Nowley.txt at https://github.com/jjdegruijter/ospats+, with process parameters as given in Figure 1. The resulting sample should be as specified in Figure 2

References

de Gruijter, J.J., McBratney, A.B., Minasny, B., Wheeler, I., Malone, B.P., Stockmann, U., 2016. Farm-scale soil carbon auditing. Geoderma 265, 120–130.

de Gruijter, J.J., Minasny, B., McBratney, A.B., 2015. Optimizing stratification and allocation for design-based estimation of spatial means using predictions with error. Journal of Survey Statistics and Methodology 3, 19–42.