# A Julia package for farm-scale soil carbon auditing

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*

## Abstract

We introduce Julia package *ospats+* for optimal sampling design in the context of farm-scale soil carbon auditing. The main difference with package *ospats* is that *ospats+* maximises the expected profit for the farmer, rather than the statistical criterion of estimation precision. The package is written in Julia for speed of computation.

Our methodology has been discussed in general terms by de Gruijter et al. (2016), here we go into the computational aspects. Using a grid of predicted carbon content with associated uncertainty, we optimise a stratified random sampling design: number of strata, stratification of the grid, total sample size and sample sizes within strata. The expected profit is maximised on the basis of sequestered carbon price, sampling costs, and a trading parameter that balances farmer's and buyer's risks due to uncertainty of the estimated amount of sequestered carbon.

The core of the methodology is optimisation of the stratification by the *Ospats* method (de Gruijter et al., 2015), an iterative procedure that reallocates grid points to strata on the basis of pairwise generalised distances between grid points. The distances are a function of the locations, the predictions and the covariances of the prediction errors. We illustrate the use of *ospats+* with an application to an Australian farm.

*Keywords:* soil carbon auditing, stratified random sampling, spatial stratification, prediction error, map uncertainty, value of information, Julia

*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)

## 1. Introduction

The purpose of this paper is to introduce a software package, *ospats+* supporting farm-scale soil carbon auditing. The statistical methodology has been discussed in detail by de Gruijter et al. (2016). Using a grid of predicted carbon content with associated uncertainty, it optimises a stratified random sampling design, i.e. number of strata, stratification of the grid, total sample size and sample sizes within strata. The optimisation criterion is the expected financial profit for the farmer, whom we assume to have a contract for soil carbon sequestration.

The expected profit is maximised on the basis of the sequestered carbon price, the sampling costs, and a trading parameter $\gamma$ that balances farmer's and buyer's risks due to uncertainty of the estimated amount of sequestered carbon.

The process of data acquisition and analysis for soil carbon auditing is schematically presented in Table 1. Package *ospats+* covers step 2 of the scheme: design optimisation for the first sampling round, also referred to as the 'baseline'. The actual optimisation takes place in step 2c and 2d, which combines stratification by the *Ospats* method introduced by de Gruijter et al. (2015), the Value Of Information (VOI) approach explained in de Gruijter et al. (2016), and Neyman allocation of optimal sample sizes to each stratum.

The main difference between packages *ospats* and *ospats+* lies in the optimisation criterion. Package *ospats* (https://github.com//jjdegruijter/ospats) minimises the expected sampling error of the estimated mean or total of any target variable for which a grid of predictions with associated error is available. Package *ospats+* on the other hand maximises the expected profit to the farmer from carbon sequestration by the VOI approach. It is therefore a more specialised application, however it is also a more rationalised approach in the sense that it directly optimises for the final goal of maximising profits from soil carbon sequestration efforts.

## 2. Method

Our method is discussed in detail by de Gruijter et al. (2016). Here we only re-iterate the essentials as far as computation is concerned. The process of optimisation is summarised as step 2c and 2d in Table 1 and is further detailed in Table 2. In short, the optimal design is found by subsequently optimising the stratification, total sample size and Neyman



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

Table 1: Schematic overview of the auditing procedure

| Step | Action |
|---|---|
| 1 | PREPARATION: |
| 1a | Delineate the area. |
| 1b | Superimpose a grid with predictions and error variances. |
| 1c | Determine cost per grid point and carbon offset price. |
| 2 | OPTIMIZE DESIGN FOR THE FIRST SAMPLING ROUND: |
| 2a | Choose allowed minimum sample size within strata, $nh_{min}$ (e.g. 3). |
| 2b | Choose a feasible range of strata numbers, $[H_{min}, H_{max}]$. |
| 2c | For each number of strata in the range, calculate stratification (*Ospats*), total sample size (Eq. 3) and sample sizes within strata (Eq. 4). |
| 2d | Select the design with the largest strata number that still fulfils the condition of step 2a. |
| 2e | Draw a stratified random sample according to the design from step 2d. |
| 3 | EXECUTE THE FIRST SAMPLING ROUND: |
| 3a | Collect samples at the locations from step 2e, and take laboratory measurements to determine the carbon stock for each location. |
| 3b | Estimate the total carbon stock and its variance. |
| 4 | OPTIMIZE DESIGN FOR THE SECOND SAMPLING ROUND: |
| 4a | Update the predictions and error variances using the sample data from the first round. |
| 4b | Repeat step 2. |
| 5 | EXECUTE THE SECOND SAMPLING ROUND: repeat step 3. |
| 6 | FINISH: calculate the confidence interval for the total amount of sequestered carbon. |

allocation (explained below) for each of the number of strata ($H$) in a pre-chosen range, $[H_{\min}, H_{\max}]$. The optimal $H$ is then the largest one, subject to the condition that the sample sizes allocated across its strata are each at least equal to a pre-chosen minimum $nh_{\min}$. The method works from an input file with four values for each of $N$ grid points: X-coordinate $x$, Y-coordinate $y$, predicted SOC content $\tilde{C}$ and error-variance of the prediction $s^2$.

The stratification for a given $H$, assuming Neyman allocation, is optimised by *Ospats*, the iterative re-allocation method described by de Gruijter et al. (2015). This method starts with a random stratification and improves



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

it by re-allocating the grid points to different strata on the basis of their pair-wise generalised distances (see below). This process is continued as long as it diminishes the objective function $O$, defined as:

$$O = \sum_{h=1}^{H} \left\{ \sum_{i=1}^{N_h - 1} \sum_{j=i+1}^{N_h} D_{ij}^2 \right\}^{1/2} \tag{1}$$

with generalised distance:

$$D_{ij}^2 = \frac{(\tilde{C}_i - \tilde{C}_j)^2}{R^2} + (s_i^2 + s_j^2) \cdot e^{-3 \cdot \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}/range} \tag{2}$$

where $R^2$ denotes the squared correlation coefficient resulting from a regression analysis underlying the SOC prediction, and the *range* is the parameter of an exponential co-variance function fitted to the prediction residuals.

To save computer time, package *ospats+* calculates the $N \times N$ matrix of pairwise generalised distances beforehand, prior to the iterative re-allocation. In case of large grids this would be impractical, so the optimisation process is then split into two phases. In the first phase, a stratification is calculated only for a sample of the grid points, then the remaining grid points are allocated to the sample strata whilst minimising $O$.

As shown by de Gruijter et al. (2016), a stratification that results from this process is optimal for any total sample size. Therefore the total sample size which maximises the expected profit for the farmer can be derived as:

$$n' = \left( \frac{CP \cdot A \cdot Z_\gamma \cdot \overline{O}}{f\sqrt{2}} \right)^{2/3}, \tag{3}$$

where

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

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

$Z_\gamma$ : quantile of the standard normal distribution (1.645 for the 95% quantile).

$\overline{O} = O/N$ : value of the optimisation criterion for the calculated stratification.

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

As discussed in de Gruijter et al. (2016) the data value of the sample data that is going to be collected depends on the precision of the estimated amount of sequestration. The precision of an estimate is usually calculated



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

from the sample data. In our case, however, we can predict the precision of the estimate and indeed the data value beforehand, when we use the SOC predictions and their error variances. To that end we define the tradeable amount of sequestration $tp$ such that there is a sufficiently large probability $\gamma$ (say 95%) that the future sequestration will be equal to or much greater than $tp$, thus minimising chances of a false positive sequestration. This is formalised by taking for $tp$ the lower boundary of the one-sided prediction interval around the predicted amount of sequestration. This boundary depends linearly on $Z_\gamma$. If the average sequestration were selected as $tp$, there would be no value in increasing the certainty of the sequestration estimate.

Given the stratification and the total sample size $n'$, optimal allocation of sample sizes to the strata, in the sense of minimal sampling variance of the mean or total, can be realised by so-called Neyman allocation (Dalenius and Hodges, 1959; Cochran, 1977). The optimal sample size for stratum $h$ is then given by:

$$n'_h = n' \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} \,. \tag{4}$$

where
$N_h$ is the size (number of grid points) of stratum $h$,
$S_h$ is the standard deviation of the SOC predictions in stratum $h$, which is predicted by

$$\tilde{S}_h = \left\{ \sum_{i=1}^{N_h - 1} \sum_{j=i+1}^{N_h} D_{ij}^2 \right\}^{1/2} \tag{5}$$

The total sample size and the sample sizes per stratum are rounded off to the nearest integer. To avoid possible inconsistency between both, the total sample size is adjusted to equal the sum of the sample sizes per stratum.

## 3. Architecture of package *ospats+*

The package consists of 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 functions of 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



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

Table 2: Schematic overview of the optimization algorithm

| Step | Action |
|---|---|
| 2c | **Design optimisation for maximum number of strata $H_{\max}$:** |
| 2c-1 | Calculate the optimal stratification with $H_{\max}$ as number of strata. |
| 2c-2 | Calculate the optimal sample size, using Eq. 3. |
| 2c-3 | Calculate the optimal (Neyman) allocation of sample sizes to the strata, using Eq. 4. |
| 2c-4 | Determine the smallest sample size within a stratum: $nh_{\mathrm{l}}$. |
| 2c-5 | If $nh_{\mathrm{l}} < nh_{\min}$, then lower $H$ by 1. |
| 2d | **Select the optimal number of strata:** |
| 2d-1 | Repeat steps 2c-1 through 2c-5 until $nh_{\mathrm{l}} \geq nh_{\min}$. |
| 2d-2 | Keep the last design resulting from step 2d-1 as the optimal design. |

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 definitively) allocating the remaining grid points to the sample strata, using the same optimisation criterion described above.

The process parameters to be set by the user in "main" are:

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

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

$nh_{\mathrm{minim}}$ : 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).

$Z_\gamma$ : quantile of the standard normal distribution (1.645 for the 95% quantile).

$R^2$ : squared multiple correlation coefficient from the regression model used to generate the predictions.

$range$ : estimated parameter of the exponential auto-covariance of the prediction errors.

$maxcycle$ : maximum number of iteration cycles allowed for iterative re-allocation. 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 has not yet exceeded 100. The setting $maxcycle = 0$ forces the system to skip the iterative re-allocation, and to



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

proceed with calculating statistics of the random initial stratification.
*in* : interval used to draw a systematic sample from the grid. 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 random number generator.

See Figure 1 for a broad overview of the optimisation process as implemented in *ospats+*.

The following general comments on alternative solutions in the algorithm are to be made.

1) The random starting solution.

The process of iterative re-allocations starts from a random initial stratification, i.e. one where the strata consist of a random collection of grid points. Initial solutions that are closer to the eventual optimum than a random draw are possible, e.g. by the cum-root-f rule (Dalenius and Hodges, 1959). We decided not to implement a closer starting solution, because preliminary experiments (not reported here) showed that the computation time needed to generate a closer start can easily outweigh any saving from fewer iteration cycles. This is primarily due to the first few iteration cycles covering the majority of the distance between a random draw and convergence to the optimal solution.

2) The option of skipping unchanged pairs of strata.

If any two strata are not changed during a cycle, then it is known beforehand that in the next cycle there can be no improving transfers of points between these two strata, hence it is an unnecessary computation step. This could in principle be skipped to save computation time. However, preliminary experiments (not reported here) show that the search functions required to enable such a skipping device is more computationally expensive than the possible savings. Thus the 'inefficiency' remains conceptional when employing conditional functions (e.g. if-else constructions) within loops.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

3) The option of swapping.

If the iteration process get trapped in a local minimum, then it could be possible to escape from it via a swap, i.e. a simultaneous transfer of two grid points to and from their current strata. An inbuilt swapping device would therefore reduce the risk of a local minimum. However, preliminary experiments (not reported here) show that only very few improving swaps are found after a complete run using sequential transfers. These swaps had a negligible effect on $O$. In addition, the swapping device proved to be relatively time consuming. Therefore our provisional conclusion is that multiple runs are more efficient than swapping.

## 4. Use of package *ospats+*

We selected Julia as programming language primarily due its speed. R was not a suitable candidate as it tends to be slower when used for large scale optimisation problems. Initially Matlab was used by de Gruijter et al. (2015) and de Gruijter et al. (2016). However, speed comparisons in the literature suggest that Julia is usually faster than Matlab, and Julia is a free and open-source language.

The supplied data file is assumed to have $N$ rows, i.e. one for each grid point and no headers. The values are comma-separated and presented in the order X-coordinate, Y-coordinate, SOC prediction and variance of the prediction error. The file may also include a column with grid point identifications. In that case the user must specify the order of the columns in script "readdata". If the data file is incomplete, i.e. not all columns have the same length, Julia issues a LoadError.

The output from *ospats+* consists of two files:

"Stratification": a file with x-coordinate, y-coordinate and stratum number for the $N$ grid points. The present version of *ospats+* does not provide a map of the stratification.

"Sample": the stratified random sample is written in this file with five columns, for sample number, stratum number, grid point number, x-coordinate and y-coordinate.

*ospats+* has been developed with Julia Version 0.6.2. Julia can be downloaded from https://julialang.org/downloads/. *ospats+* can be downloaded from https://github.com/jjdegruijter/ospats-plus, together with a user's manual and replication material. It is ready to be used, assuming that Julia has



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

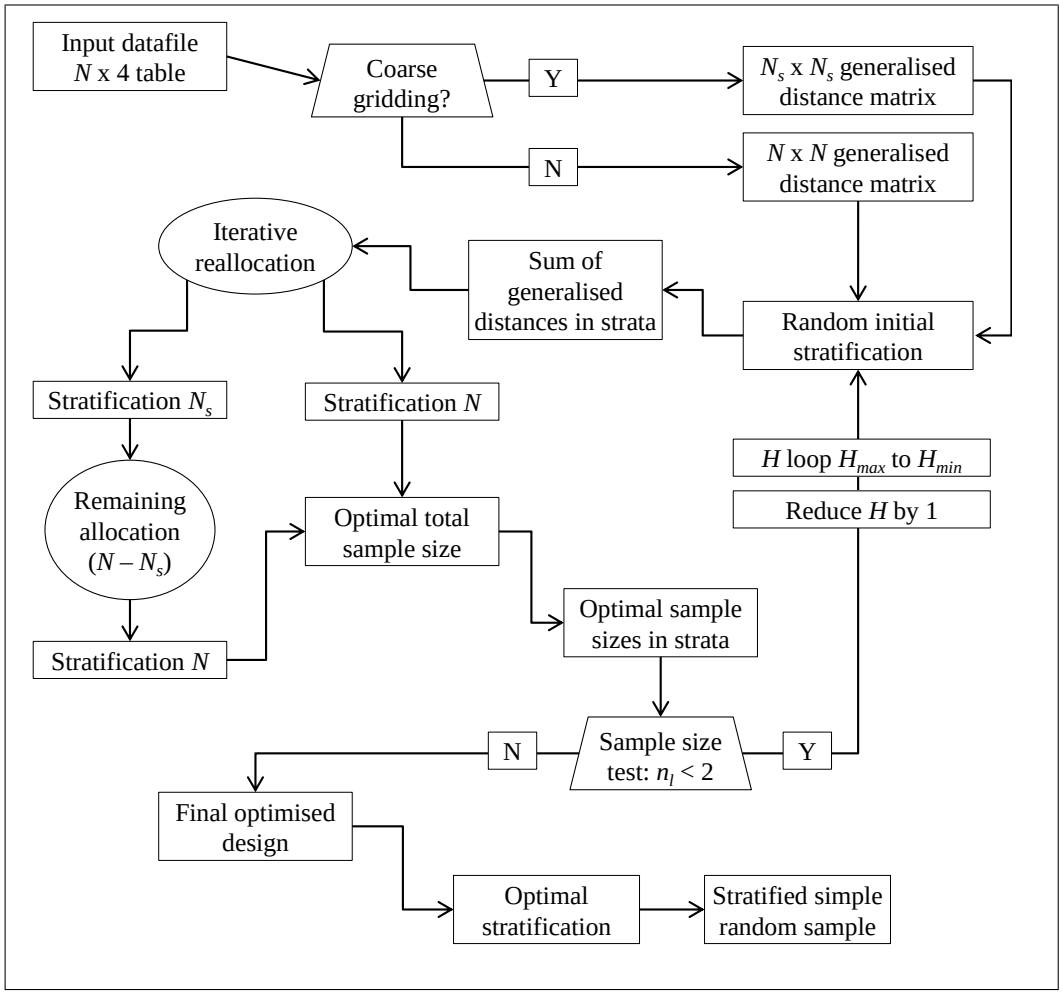


Figure 1: Overview of the optimisation process in *ospats+*.

been installed. No other package dependencies are needed, except for the Julia packages CSV and DataFrames (simply do Julia > Pkg.add("CSV") and Julia > Pkg.add("DataFrames").

The use of *ospats+* need not be limited to a farm as a whole. It can also be applied to different parts of a farm, such as management units. Another option is to use it for a group of farms, e.g. a co-operation of carbon farmers. In a research setting *ospats+* can be employed as a tool for what-if studies, to investigate the effects of, for instance, changes in carbon offset price, costs of data collection and accuracy of SOC prediction.

It should be noted that *ospats+* has several limitations. Firstly, the



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

present version supports only the first sampling round in SOC monitoring, i.e. step 2 in Table 1. A future extension may well include optimal design for the second round. The methodology has been worked out by de Gruijter et al. (2016), and coding can largely follow the same lines as in the present version.

Secondly, *ospats+* optimises a sampling design for a single target variable only: soil organic carbon. The resulting design, especially the stratification, may *not be optimal for other soil variables in general*. However, we expect the design to be reasonably efficient for other variables as well, dependant the degree in which they are correlated with SOC.

For instance, among the most common macro and micro nutrients, the most correlated with organic carbon is organic nitrogen (see Figure 4 Hengl et al. (2017)). As most nutrients are inter-correlated, more than 75% of variation in values can be explained by the first 5 principal components: PC1 (48.8%), PC2 (19.4%), PC3 (6.7%), PC4 (5.2%) and PC5 (3.8% variation) (Hengl et al., 2017). Therefore it is reasonable to assume that optimised sampling designs for organic carbon will capture a decent portion of the spatial variation of other common plant nutrients - should they also be measured.

Regardless of efficiency, the unbiasedness of the statistics estimated from the sample data like means, totals and fractions, as well as standard errors and confidence intervals, remains valid for any variables measured using these designs.

Thirdly, but less importantly, iterative re-allocation may get trapped in a local minimum. In other words it does not warrant a global optimum. This is why package *ospats* has the option of multiple runs, retaining the best result. In our experience, however, differences between the results from multiple runs appeared to be practically irrelevant in all our cases, if at all existent. This option was therefore not included in *ospats+*.

## 5. Illustrative Example

### *5.1. Description of area and data*

As an illustration, we applied *ospats+* to soil carbon data from 'Nowley farm', the same farm as in the case study by de Gruijter et al. (2016). For this example we used data from previous sampling campaigns. However, prior data collection on-site is becoming less necessary for optimising sampling designs as carbon mapping with associated uncertainty, at sufficient resolution, is becoming increasingly available. Part of the drive of this



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

increased availability/suitability of carbon prediction maps is based on increasing availability of both covariates (e.g. remote sensing based) and field measurements based on proximal sensing..

Nowley farm covers approximately 2300 ha and is situated in the highly agriculturally productive Liverpool Plains region in north west NSW, Australia. It is run as a mixed farming enterprise centred around cropping of wheat, barley and canola in winter, sorghum and sunflower in summer, and a cattle herd of breeders, replacement heifers and bulls. Nowley has a combination of fertile basaltic soils together with more challenging soil types that are poorly drained, with considerably high amounts of subsoil sodium.

Soil point observations of total soil carbon concentration were collected over two separate soil sampling campaigns during 2014 and 2015 from across Nowley farm. The sampling for each campaign was based on stratified random sampling, where at each site a 7.5 cm depth core of soil (0 - 7.5 cm and with known volume) was collected. A total of 130 samples was collected from these two sampling campaigns.

Soil carbon stocks ($CS$, t ha$^{-1}$ ) to 7.5 cm were calculated from measured carbon concentrations, bulk densities and gravel contents. The mean carbon stock of these samples was 16.06 t ha$^{-1}$, while the minimum and maximum was 6.03 and 43.20 t ha$^{-1}$ respectively.

Digital soil mapping was used to create a carbon stock map for Nowley using the point observations of carbon stocks and a number of environmental variables derived principally from a digital elevation model, air-borne gamma radiometric data and associated derivatives from each. The map was made using stepwise multiple linear regression which lead to a model containing parameters for 4 variables: Elevation ($E$), Topographic wetness index ($TW$), gamma radiometric potassium ($GK$), and Wilford's weathering index ($WI$). The model took the form:

$$CC = 5.02 + 0.07 \times E - 0.83 \times TW - 1.05 \times GK - 0.81 \times WI \quad (6)$$

Model residuals showed a week spatial autocorrelation. Fitting an exponential variogram with zero nugget (the default in *ospats+*), gave an estimated range of 582 m. We used Leave-one-out cross validation to evaluate the goodness of fit of the model. Here we estimated the RMSE = 5.5 and $R^2$ = 0.36. The prediction variance of the model was also estimated in order to quantify the uncertainty about the map predictions of soil carbon stocks, see Figure 2. Together, these maps were created using a 10 m x 10 m grid cell



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

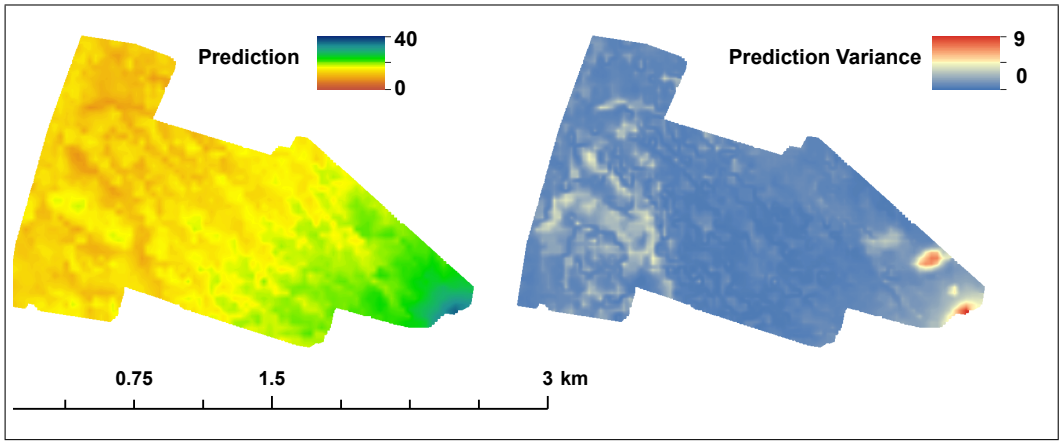


Figure 2: Nowley farm: soil carbon prediction and prediction variance.

Table 3: Process parameters used to run *ospats+* on the Nowley data set

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| $H_{\min}$ | 3 | $H_{\max}$ | 7 |
| $nh_{\text{minim}}$ | 3 | $CP$ | 10 Aus$ |
| $f$ | 120 Aus$ | $Area$ | 2336 ha |
| $Z_\gamma$ | 1.645 | $R^2$ | 0.36 |
| $range$ | 582 m | $maxcycle$ | 150 |
| $in$ | 2 | $seed$ | 1234 |

resolution, as this was the resolution of the environmental covariates used. However, subsequent to this modelling we coarse-gridded the maps to 30 m x 30 m grids to avoid undue computational load for this example. This resulted in 26,079 grid points.

### *5.2. Application of* ospats+

We ran *ospats+* on the data described above, with process parameters given in Table 3.

It turned out that in these circumstances the optimal number of strata is 5, the optimal total sample size is 58, and the optimal sample sizes within the strata are 8, 12, 21, 4 and 13. A map of the optimised stratification and the sample locations is presented in Fig. 3.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

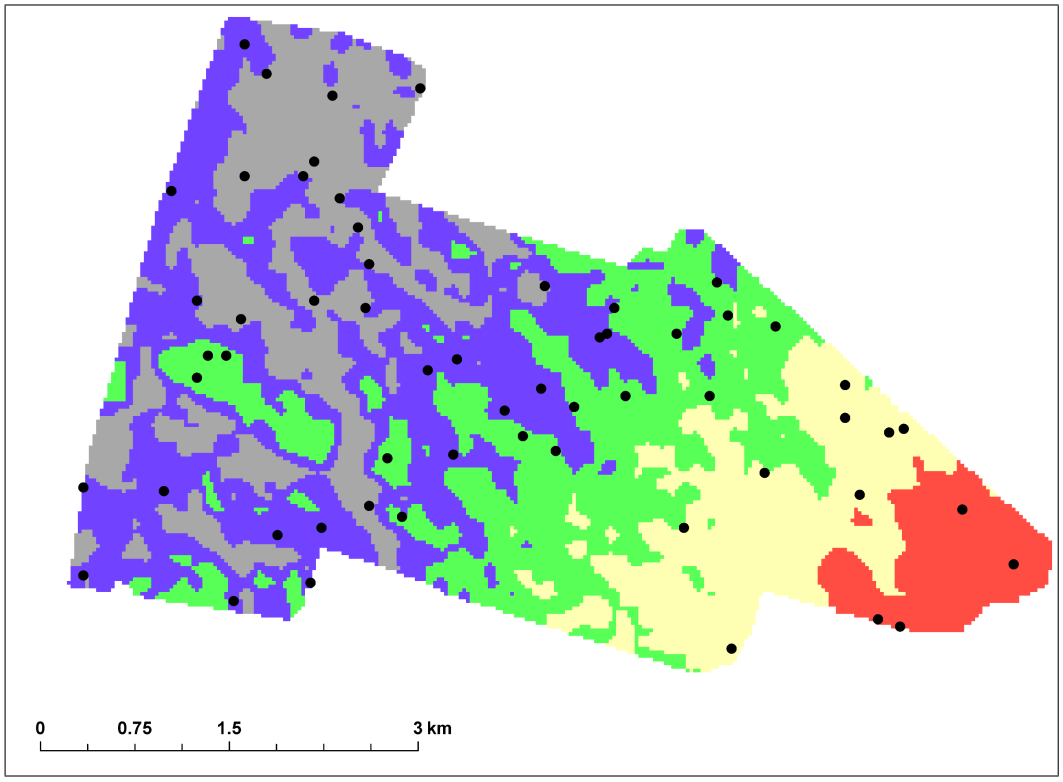


Figure 3: Ospats+ stratification and stratified sample based on data in Fig. 2.

## 6. Discussion

When using *ospats+* one should realise that the following assumptions underly the methodology as implemented.

1. *The second round sampling is independent from the first round.*
Revisiting the sampling sites from the first round again in the second round would usually lead to a higher precision of the estimated change. However, to avoid possible fraudulent practices we adopted full independence between both rounds. Additionally, differing sample points each time allows a more complete picture of the spatial variation of SOC to emerge.

2. *The variable cost of collecting the data is linearly related to the number of sample points.*
The present version of *ospats+* uses a linear cost function. If that does not predict the real costs well enough, then a non-linear function could replace the linear one. In that case Eq. 3 should be adapted, or replaced by a discrete optimisation algorithm to determine the optimal sample size..



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018

3. *The variances of the prediction errors are correctly quantified.*
Over-estimated and under-estimated variances of the prediction errors will expectedly lead to a less efficient sampling design. The same applies to over- and under-estimation of the auto-covariance range and $R^2$. However, regardless of efficiency, unbiasedness remains warranted for statistics estimated from the sample data like means, totals and fractions, as well as standard errors and confidence intervals.

4. *Measurement errors in determining SOC stocks of samples are negligible compared to prediction errors.*
If measurement errors are not negligible, such as with proximal sensing of SOC stocks, then the sample size should be increased to achieve the same data value. This is not accounted for in the present version of *ospats+*.

## References

Cochran, W., 1977. Sampling Techniques. Wiley, New York.

Dalenius, T., Hodges, J.L., 1959. Minimum variance stratification. Journal of the American Statistical Association 54, 88–101.

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.

Hengl, T., Leenaars, J.G.B., Shepherd, K.D., Walsh, M.G., Heuvelink, G.B.M., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E., Wheeler, I., Kwabena, N.A., 2017. Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m resolution using machine learning. Nutrient Cycling in Agroecosystems 109, 77–102.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26964v1 | CC BY 4.0 Open Access | rec: 28 May 2018, publ: 28 May 2018