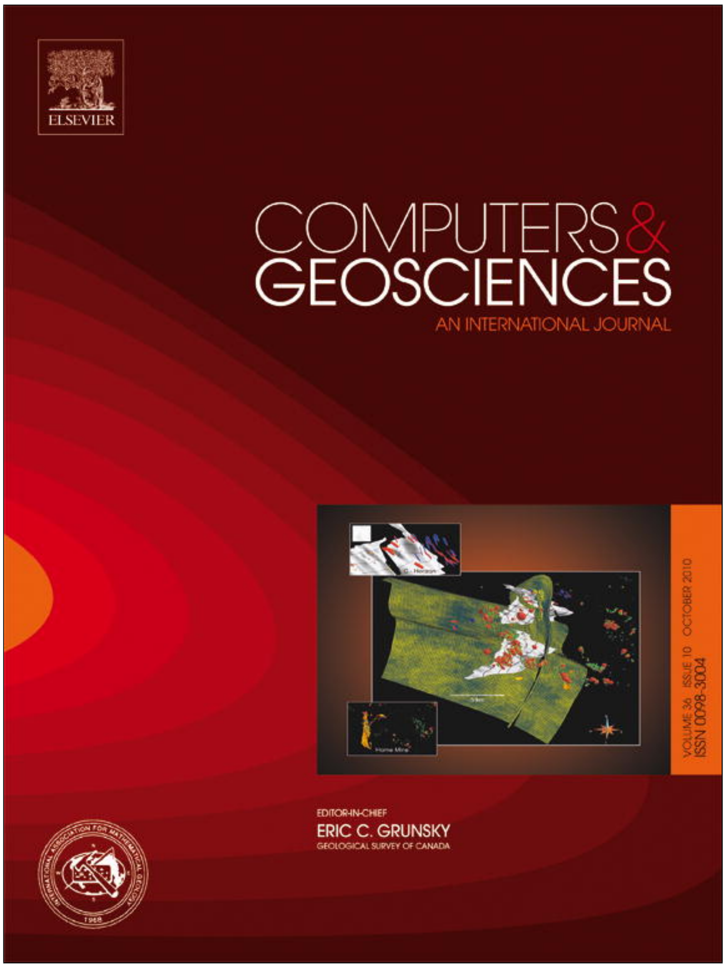
ELSEVIER
COMPUTERS &
GEOSCIENCES
AN INTERNATIONAL JOURNAL
VOLUME 36 ISSUE 10 OCTOBER 2010
ISSN 0098-3004
EDITOR-IN-CHIEF
ERIC C. GRUNSKY
GEOLOGICAL SURVEY OF CANADA

Computers & Geosciences 36 (2010) 1261–1267

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

An R package for spatial coverage sampling and random sampling from compact geographical strata by *k*-means[☆]

D.J.J. Walvoort *, D.J. Brus, J.J. de Gruijter

Alterra, Wageningen University and Research Centre, PO Box 32, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 8 June 2009
Received in revised form
16 April 2010
Accepted 19 April 2010

Keywords:
Design-based
Model-based
Stratified random sampling
Kriging

ABSTRACT

Both for mapping and for estimating spatial means of an environmental variable, the accuracy of the result will usually be increased by dispersing the sample locations so that they cover the study area as uniformly as possible. We developed a new R package for designing spatial coverage samples for mapping, and for random sampling from compact geographical strata for estimating spatial means. The mean squared shortest distance (MSSD) was chosen as objective function, which can be minimized by *k*-means clustering. Two k-means algorithms are described, one for unequal area and one for equal area partitioning. The R package is illustrated with three examples: (1) subsampling of square and circular sampling plots commonly used in surveys of soil, vegetation, forest, etc.; (2) sampling of agricultural fields for soil testing; and (3) infill sampling of climate stations for mainland Australia and Tasmania. The algorithms give satisfactory results within reasonable computing time.

1. Introduction

It is well known that, both for mapping and for estimating spatial means of an environmental variable, the accuracy of the result will usually be increased by dispersing the sample locations so that they cover the study area as uniformly as possible (Cochran, 1977). A simple method to achieve this is sampling on a regular grid. For mapping a variable sampled on a grid, the maximum prediction error variances occur at the centers of the grid cells, and are approximately equal. However, at the border of the study area the prediction error variance (kriging variance) increases considerably when there are no measurements outside the study region that can be used to predict the values near the border. Of course this border effect can be reduced by shifting the sample locations towards the edges, but this goes hand in hand with an increase of the kriging variance at the centers of the grid cells. This raises the question, should we relax the constraint of a regular pattern of sample locations to obtain the best result? This question emerges with greater concern when the area is irregularly shaped, or has enclosures that cannot be sampled (built-up areas) or need not be mapped.

A regular pattern of sample locations can also be too restrictive when we have measurements in the study region collected in a former survey, which we want to use in the geostatistical interpolation and which do not fit into a regular grid. The previous measurements may have left large spaces unsampled, which we would like to fill in because there the greatest gain in accuracy can be achieved. When we expect regular grid sampling to be suboptimal under such practical constraints, we must design in some way an irregular pattern that will lead to more precise spatial predictions than the regular grid. Several methods for optimization of the pattern of sample locations have been described in the literature. The methods differ with respect to the objective function, and in the way the method searches for the optimal pattern (optimization algorithm). In geostatistical sampling, an objective function explicitly defined in terms of the prediction error variance is minimized, usually the average or maximum kriging variance (Sacks and Schiller, 1988; van Groenigen et al., 1999). This requires knowledge of the variogram, and in many situations this variogram is unknown, or at least characterized by uncertainty. In spatial coverage sampling, an objective function is defined in terms of the distance between the sample locations and the nodes of a fine interpolation grid (Royle and Nychka, 1998), and a variogram is not needed.

In a design-based sampling strategy for estimating the spatial mean, spreading of the sample locations can be achieved by sampling on a randomly placed regular grid. There are two disadvantages of random grid sampling. First, estimation of the sampling variance is cumbersome (D'Orazio, 2003). This is because we do not have independent replicates of the sample: the grid can be considered as one 'cluster' of sample locations. Second, in general the number of sample locations with random grid sampling is not fixed, but varies between randomly drawn samples. We may choose the grid spacing such that on average

[☆] Code available from server at http://cran.r-project.org or at http://www.iamg.org/CGEditor/index.htm.

[*] Corresponding author.

E-mail addresses: dennis.walvoort@wur.nl (D.J.J. Walvoort), dick.brus@wur.nl (D.J. Brus), jaap.degruijter@wur.nl (J.J. de Gruijter).

0098-3004/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2010.04.005

the number of sample locations equals the required (allowed) number of sample locations, but for the actually drawn sample, this number can be a few locations smaller or larger. A random number of sample locations may be undesirable, for instance, when this size is prescribed in regulations. An alternative is stratified random sampling, using geographically compact sub-areas as strata. By using these compact sub-areas as strata, spatial clustering of the sample locations can be avoided, which usually increases the accuracy of the estimated spatial mean. The central question then is how to split the area into sub-areas that are geographically as compact as possible.

Although methods and software exist for designing spatial coverage samples (Royle and Nychka, 1998), we decided to develop a new R package (R Development Core Team, 2010). The main reason is that there is a clear need for a simple, straightforward and generally available method that can be used both for designing spatial coverage samples for mapping, and for constructing compact geographical strata for estimating spatial means.

We have chosen the mean squared shortest distance (MSSD) as objective function. It has been shown before that minimizing the MSSD leads to spatial coverage samples with a mean ordinary kriging variance (MOKV) only marginally larger than that of geostatistical samples obtained by directly minimizing the MOKV (Brus et al., 2007). Another attractive property of the MSSD as objective function is that it can be minimized by *k*-means clustering, which is a well-developed branch of cluster analysis, both theoretically and computationally. It can be shown that, in the case where all cluster centers coincide with the cluster centroids, minimizing the trace of the pooled within-cluster variance (tr(**W**)) is equivalent to minimizing the MSSD (see Section 2). Note that we distinguish cluster centers from cluster centroids. A cluster center is the location to which the distances of the objects are calculated, whereas a cluster centroid is the multivariate average of the objects allocated, at a given stage in the clustering process, to that cluster. Existing software for *k*-means clustering is not fully satisfactory for sampling purposes. First, this software has not all the functionality we need, such as the possibility of using prior sample data (infill sampling), and forming clusters of equal size. Clusters of equal size are attractive because, when used as strata in random sampling, the sampling design is self-weighting, i.e. the unweighted sample mean is an unbiased estimator of the spatial mean. The freeware program FuzMe (Minasny and McBratney, 2002) uses a modified fuzzy *k* means algorithm to obtain clusters of equal size. Although the FuzMe program offers many interesting features for multivariate fuzzy cluster analysis, it does not guarantee clusters of equal size and is therefore not suitable for our needs.

Second, existing software is generally not directly linked with sampling, and several data processing activities related to sampling, such as the discretization of the study area, random selection of sampling locations from the final clusters, and design-based or model-based inference are not supported.

The aim of this paper is to present and illustrate a new R package called **spcosa**, that can be used for designing spatial coverage samples and for partitioning the study area into geographically compact blocks to be used as strata in random sampling. R is a programming environment for data analysis and graphics which has become extremely popular during the last decade. Since it is freely available and offers many add-on packages for spatial data analysis and visualization, it seems the natural language of choice for implementing our spatial coverage sampling algorithms.

This paper is organized as follows. In Section 2 we describe two *k*-means algorithms, one for unequal area partitioning and one for equal area partitioning. Section 3 describes three applications of the proposed sampling method, with a spatial extent ranging from a sampling plot of tens of square meters to a whole continent. The sampling method is discussed in Section 4, and several conclusions are drawn.

2. K-means algorithms

As stated in Section 1, spatial coverage samples can be designed by minimizing tr(**W**). This can be achieved by *k*-means cluster analysis (Hartigan, 1975), originally developed in the context of multivariate analysis. In our spatial application of this method, the objects are the cells of a fine grid, and the classification variables are the geographical co-ordinates of the midpoints of these cells, as explained in more detail by Brus et al. (1999). In *k*-means clustering, starting from an initial solution, the cells are iteratively re-allocated to clusters, and their centroids re-computed, until some stopping criterion is satisfied. The result of this procedure consists of a partition of the grid and the associated cluster centers. The clusters can be used as strata in stratified random sampling, whereas the cluster centers can directly be used as sample locations in a model-based sampling strategy.

Several *k*-means algorithms exist, see for instance MacQueen (1967), Lloyd (1982), Hartigan and Wong (1979) and Ding and He (2004). We defined and implemented two algorithms. In algorithm 1 only 'transfers' take place. By transfer we mean a re-allocation of a cell ($\mathbf{u}$) from its present cluster (A) to another cluster (B). This algorithm is suitable for unequal area partitioning, possibly in the presence of prior points. The algorithm is described in Section 2.1. In algorithm 2 only 'swops' take place. A swop is a simultaneous transition of two cells, $\mathbf{u}$ from A to B, and $\mathbf{v}$ from B to A. This algorithm is suitable for equal area partitioning. Algorithm 2 is described in Section 2.2.

2.1. K-means algorithm 1 for unequal area partitioning

Step 1a: Initial partition. If there are n prior sample points, then these points act as n fixed cluster centers in the following. If k additional sample points are required, then select at random k cells from the grid. Their midpoints act as k variable cluster centers. Create an initial solution in the form of a partition by allocating the unselected cells to the nearest (fixed or variable) of the $n+k$ cluster centers.

Step 1b: Initial cluster centers. Replace each of the k variable cluster centers by the centroid of the cluster around it: $\bar{\mathbf{x}}_1 \cdots \bar{\mathbf{x}}_k$ (k two-dimensional vectors).

Step 2: Re-allocation of the first cell. Determine if the first cell (with co-ordinate vector $\mathbf{u}$) should be transferred from its initial cluster (say A) to the first of the other $n+k-1$ clusters (say B), as follows.

Calculate the squared distances from $\mathbf{u}$ to $\bar{\mathbf{x}}_A$ and to $\bar{\mathbf{x}}_B$, respectively, $d^2(A,\mathbf{u})$ and $d^2(B,\mathbf{u})$. If $d^2(A,\mathbf{u}) > d^2(B,\mathbf{u})$, then the transfer is carried out and, as far as they are not fixed, the two cluster centers are replaced by the centroids of the surrounding clusters. If not, then the transfer is not carried out.

Step 3: Iteration. If the transfer in step 2 was not carried out, then determine in the same way if $\mathbf{u}$ should be transferred to the second of the other $n+k-1$ clusters. If not, then do the same for the third cluster, and so on. When $\mathbf{u}$ has been transferred or when it has been determined that it should not be transferred to any cluster, then go to the second cell and do the same as with the first one. Thereafter continue with the third cell, and so on, until all cells have been addressed. After that, start again from the beginning (a new cycle), and continue until none of the cells are being transferred any more. The resulting partition and cluster centers constitute the final solution.

2.2. K-means algorithm 2 for equal area partitioning

Step 1a: Initial partition. Create an initial solution in the form of a random partition, with k clusters of equal size, say N. This can be accomplished by visiting all cells in random order and assigning cell i to cluster $(i-1)\bmod k+1$. For instance, if $k=10$, then cell 12 will be assigned to cluster 2. Note that if the total number of grid cells is not a multiple of N, not all clusters are exactly of size N. The clusters differ in size by only one grid cell at most.

Step 1b: Initial centroids. Calculate the centroids of the clusters, $\bar{\mathbf{x}}_1 \cdots \bar{\mathbf{x}}_k$ (k two-dimensional vectors).

Step 2: Re-allocation of the first cell. Determine if the first cell (with co-ordinate vector $\mathbf{u}$) should be swopped from its initial cluster (A) with the first cell ($\mathbf{v}$) of the first one of the other $k-1$ clusters (B), as follows.

Calculate the squared distances from $\mathbf{u}$ to $\bar{\mathbf{x}}_A$ and to $\bar{\mathbf{x}}_B$ ($d^2(A,\mathbf{u})$ and $d^2(B,\mathbf{u})$) and from $\mathbf{v}$ to $\bar{\mathbf{x}}_A$ and to $\bar{\mathbf{x}}_B$ ($d^2(A,\mathbf{v})$ and $d^2(B,\mathbf{v})$). If $d^2(A,\mathbf{u})+d^2(B,\mathbf{v})>d^2(A,\mathbf{v})+d^2(B,\mathbf{u})$, then the swop is carried out and the centroids are updated. If not, then the swop is not carried out.

Step 3: Iteration. If the swop in step 2 was not carried out, then determine in the same way if $\mathbf{u}$ should be swopped with the second cell of the first one of the other clusters. If not, then do the same for the third cell, and so on, until all cells of all other clusters have been checked. When $\mathbf{u}$ has been swopped or when it has been calculated that it should not be swopped with any cell, then go to the second cell and do the same as with the first one. Thereafter the third cell, and so on, until all cells have been treated. After that, start again from the beginning (a new cycle), and continue until none of the cells are being swopped any more. The resulting partition and centroids constitute the final solution.

The squared distances in the algorithms described above can be either defined as Euclidean distances or great circle distances. The latter is relevant for designing spatial coverage samples for the globe as a whole, or at the continental scale.

In basic k-means algorithms, re-calculation of the centroids is postponed until the end of each iteration cycle. By doing so one or more clusters may become empty, with no centroids to be re-calculated. To avoid this problem we adopted algorithms that update the centroids immediately after a transfer or swop.

To enhance performance, the activity of each cluster is tracked during the optimization process. A cluster is said to be active if it was involved in a transfer or swop during the current or previous cycle. During a cycle only pairs of clusters are addressed that have at least one cluster active. Pairs of two inactive clusters are disregarded as long as both clusters stay inactive. This leads to a significant performance gain of about 5–20% for algorithm 2 (swop), and up to 30% for algorithm 1 (transfer).

K-means is a deterministic search technique which means that, given the discretization grid and the stopping rule, all intermediate clusterings and the final clustering are determined by the initial clustering and the order of the cells. This is because in each iteration the algorithm calculates the best re-allocation of grid cells. The final clustering can be a local instead of the global minimum, and therefore a number of initial clusterings should be tried.

3. Examples

We now illustrate the proposed method by three examples. The spatial extent of these examples differs widely. The first example is designing a spatial coverage sample for plots of tens to hundreds of square meters. The second case study is stratified random sampling and spatial coverage sampling of agricultural fields, while the third example is spatial infill sampling at a continental scale.

3.1. Case 1: sub-sampling of sampling plots

In survey and monitoring of natural resources such as soil, vegetation and forest, the sampling units often are not point locations but small areas of tens to hundreds of square meters. Commonly used shapes of the sampling plots are squares and circles. To estimate the spatial mean of a property that varies within a plot, often the plot must be sub-sampled, i.e., several locations must be selected within the plot at which the property is measured. To estimate the plot mean as accurately as possible, the sampling locations should cover the sampling plot as uniformly as possible. Often, for budgetary reasons, only one composite sample is collected per sampling plot. The value of the target property as measured on the composite sample is used as an estimate of the plot mean. In this case we recommend *purposive* spatial coverage sampling, i.e., sampling at the centroids of the k-means clusters formed by algorithm 2. Fig. 1 shows the optimized sample pattern for circular and square sampling plots for a range of sample sizes. The circular sampling plot consists of 648 cells, the square sampling plot of 900 cells. For circular plots and sample sizes $n=2$, 3, 4 and 5, the optimized patterns consist of the centroids of circular sectors. For $n=6$, 7 and 8, one sample location is located at the center. All square plots show line symmetry with one or two lines of symmetry, and some of them point symmetry with the origin at the center of the square. In case an estimate of the accuracy of the estimated plot mean is required, we recommend stratified random sampling, using the subplots (clusters) as strata (see next case) (de Gruijter et al., 2006).

3.2. Case 2: sampling of agricultural fields

All over the world agricultural fields are repeatedly sampled to monitor the nutrient status of the soil. In practice the fields are often sampled in a herringbone pattern (Ferrari and Vermeulen, 1955). Recently, we proposed to sample the fields by means of stratified random sampling, where the strata are formed by k-means algorithm 2 for equal area partitioning (Brus et al., 1999). Fig. 2 (left hand side) shows an example. In this example the field of 2.8 ha (2500 grid cells), is partitioned into 20 strata, and from each stratum two locations are selected by simple random sampling. The order of selection is recorded. In order to save laboratory costs, the soil material collected at the locations firstly drawn is bulked into a composite sample, and similarly the material collected at the secondly drawn locations. As the strata have equal surface area, the nutrient concentration in the composite sample is an unbiased estimate of the mean concentration for the field. A second composite sample is collected, so that the sampling variance of the estimated field mean can be quantified. If a model of the spatial variation can be postulated and an estimate of the laboratory measurement error is available, as well as a model for the costs of fieldwork and laboratory analysis, then the optimal combination of number of composite samples and number of aliquots (increments) per composite can be calculated for a given budget (Brus et al., 1999; Brus and Noij, 2008).

For large fields with significant spatial variation of basic soil properties such as clay, pH, organic matter content, or the nutrient status, an estimate of the overall field mean is often not satisfactory. To adapt soil management to local conditions as

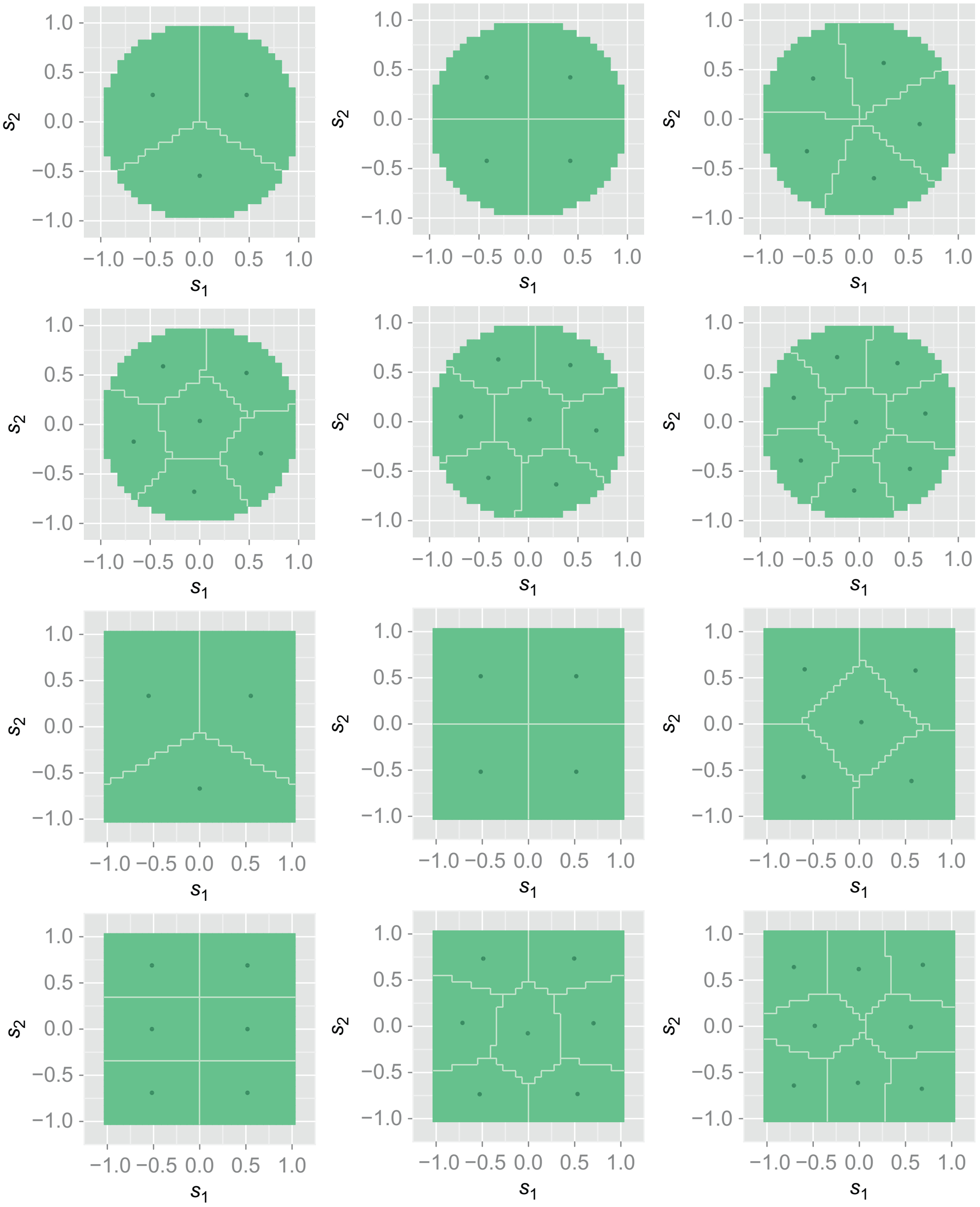

Fig. 1. Examples of spatial coverage samples to estimate spatial means of circular (top rows) and square (bottom rows) sampling plots.

done in precision agriculture, maps of soil properties are required. For mapping, the soil material collected at the sample locations must be analyzed separately. For this aim a purposive spatial coverage sample is more appropriate. The *k*-means algorithm 1 for unequal area partitioning generally leads to a slightly better spatial coverage than algorithm 2 for equal area partitioning. Fig. 2 (right hand side) shows the spatial pattern of a spatial coverage sample of 40 locations for the field, obtained with algorithm 1.

3.3. *Case 3: spatial infill sample of Australia*

The third case concerns the design of a spatial infill sample of climate stations for a continent as a whole, *viz.* Australia. The Australian Reference Climate Station (RCS) network has been establishment for high quality, long-term climate monitoring, particularly with regard to climate change analysis. More details on the RCS-network can be found at the website of the Australian Government Bureau of Meteorology.[1] The locations of the 94 climate stations on mainland Australia and Tasmania have been selected for this case study. As an example we assumed that 11 new stations were to be added, an increase of about 10%. The locations of the new stations were computed with *k*-means algorithm 1 (unequal area partitioning). The new stations are located at the centers of the strata not containing an existing station (Fig. 3). Considering the large spatial extent of the study area, we used squared great circle distances instead of squared Euclidean distances in the *k*-means algorithm.

Fig. 3 shows that the new stations are mainly located in the interior part of Australia. No new stations are added to the coastal zones where the density of existing stations was already large. Note that existing stations do not generally coincide with the

[1] http://www.bom.gov.au/climate/change/reference.shtml, (accessed 14 June, 2010).

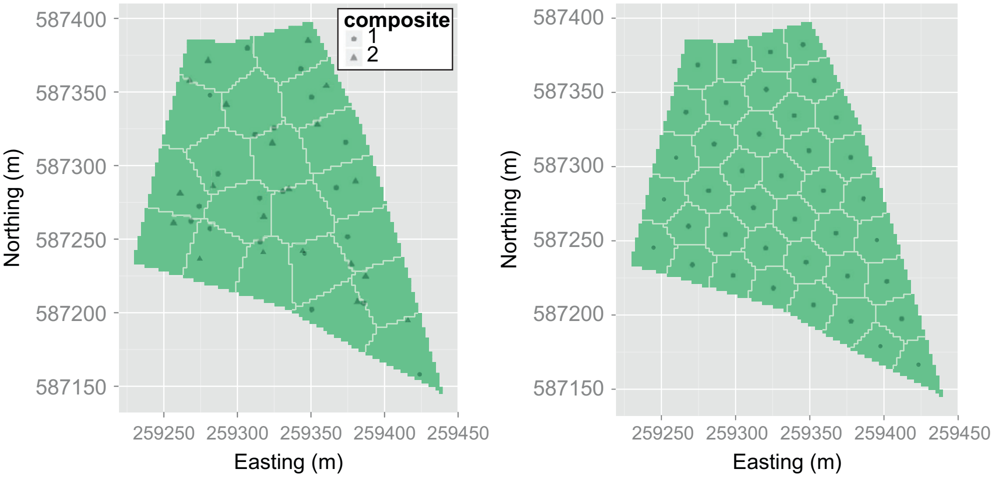

Fig. 2. Left: stratified simple random sample. Strata are formed by *k*-means algorithm 2 for equal area partitioning. Two composite samples are formed, by bulking 20 soil aliquots (one aliquot per stratum). Right: spatial coverage sample. Clusters are formed by *k*-means algorithm 1 for unequal area partitioning.

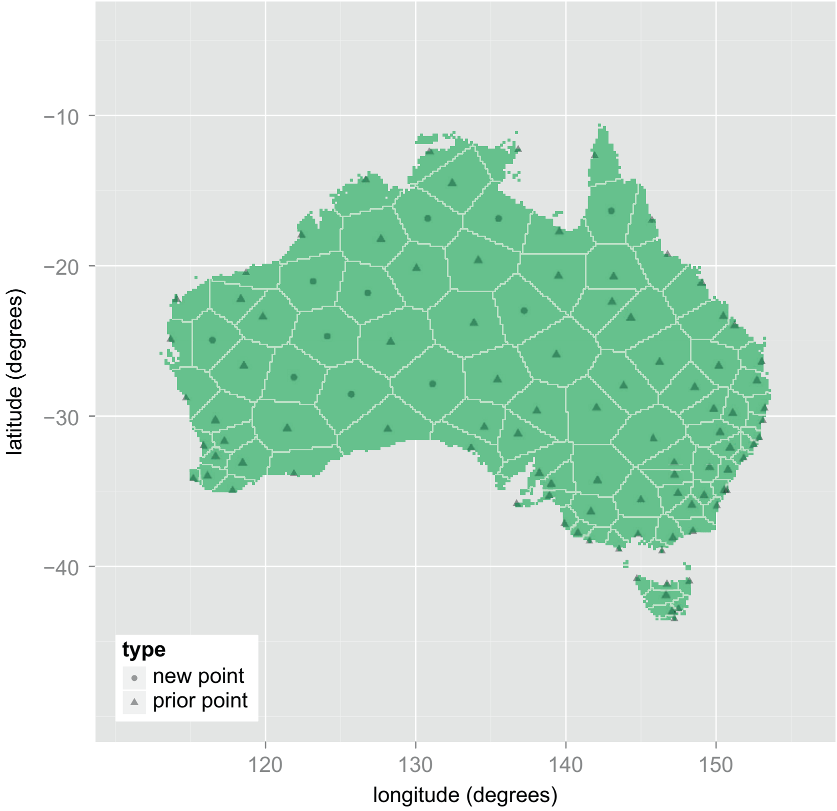

Fig. 3. Spatial infill sample of 11 new climate stations. Existing climate stations (94) are part of the Australian Reference Climate Station (RCS) network.

centroids of the strata. Especially, existing stations near the coastline are distant from the centroid of their associated stratum.

Optionally, spatial constraints can be taken into account in the optimization process. For example, urban centers can be filtered out from the map so that new stations will be placed in areas without urban influence.

The *k*-means algorithm is sensitive for local optima, especially in the case of spatial infill sampling. In this case new centroids can be trapped by surrounding prior points (existing weather stations). Therefore, it is recommended to apply the algorithm to several initial clusterings to reduce the risk of ending up in an unfavorable local optimum.

4. Discussion and conclusions

If the study area is non-convex, due to a concave outer-boundary or non-contiguity of the area or inclusions that are not of interest or that cannot be sampled, then the proposed algorithms may produce a purposive or infill sample with one or more sample locations falling outside the study area. When there are only a few and they are close to the boundary, a pragmatic solution is to shift such locations to the nearest point at the boundary. This solution has been implemented in the current version of the R package. Another simple solution could be to divide the area first into a number of (approximately) convex

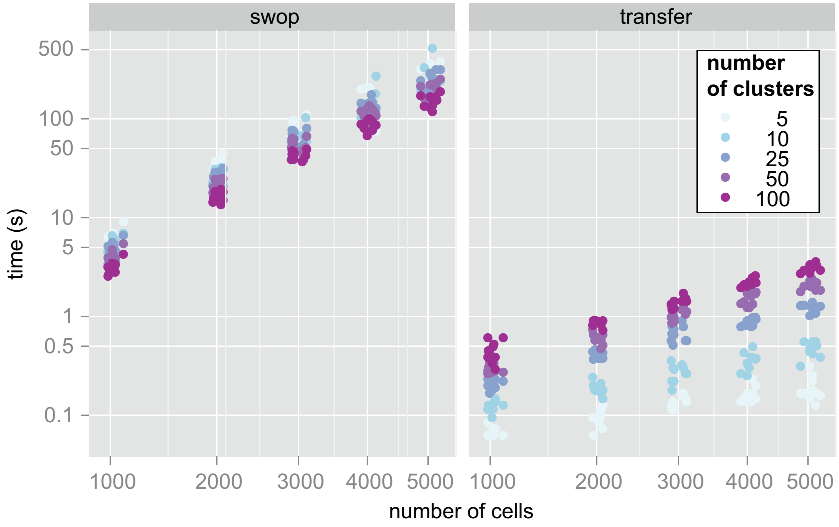

Fig. 4. Performance of algorithm 1 (right) and algorithm 2 (left) as a function of number of clusters and number of grid cells. Each dot represents total computation time given 10 random initial configurations.

sub-areas, which are then sampled separately via the proposed algorithm. This will not work when there are many inclusions, or when the outer boundary is highly irregular, or when the area consists of many geographically disconnected sub-areas. In these situations other methods, such as simulated annealing, may be more appropriate (see for instance van Groenigen and Stein, 1998; van Groenigen et al., 1999 for examples). Another option could be to apply *k*-medoids instead of *k*-means. K-medoids resembles *k*-means in that both partition a data-set under minimization of the mean distance or squared distance between a data point and the center of the cluster to which it is allocated. The difference is that in *k*-means the cluster centers are the centroids of the clusters, while in *k*-medoids the cluster centers are data points (see e.g. Kaufman and Rousseeuw, 1990). K-medoids requires greatly increased computing time than *k*-means.

As mentioned before, the final clustering can be a local instead of the global minimum, and therefore the whole optimization process must be repeated with several initial clusterings. Further research is needed on how many of these initial clusterings are needed to avoid such a local minimum (Brus et al., 2007).

Apart from the number of initial clusterings, the computing time is strongly determined by the number of cells to be clustered and the number of clusters. Fig. 4 gives the computing time needed for the algorithms to converge, as a function of the number of cells and the number of clusters. Each dot represents the best solution found given 10 random initial clusterings. Computing time increases drastically with the number of cells. In practical sampling applications, in general a fine discretization is not needed, and a cell size of about 1/2500 to 1/5000 of the size of the study area, is satisfactory. Given the number of grid cells, computing time for the swop algorithm is considerably larger than for the transfer algorithm. For the swop algorithm, computing time decreases with the number of clusters, whereas for the transfer algorithm it increases. In general, it only takes a few minutes to complete most designs. This is minor cost of implementing sampling.

In stratified random sampling, the estimator for the spatial mean is linked with the sampling design. For this reason the R package also supports the estimation of the spatial mean and its standard error from stratified simple random samples, when all sampling locations are measured separately, or when soil aliquots collected at the sampling locations are bulked into composite samples. Spatial interpolation (e.g., by kriging) of spatial coverage samples is supported by other R packages, for instance `gstat` (Pebesma, 2004) and `geoR` (Ribeiro and Diggle, 2001).

5. Availability

Spatial coverage sampling has been implemented in the **spcosa**-package (Walvoort et al., 2009) and is freely available at the Comprehensive R Archive Network (CRAN, http://cran.r-project.org/).

References

Brus, D.J., de Gruijter, J.J., van Groenigen, J.W., 2007. Designing spatial coverage samples using the *k*-means clustering algorithm. In: Lagacherie, P., McBratney, A.B., Voltz, M. (Eds.), Digital Soil Mapping. An Introductory Perspective. Elsevier, Amsterdam, The Netherlands, pp. 183–192.

Brus, D.J., Noij, I.G.A.M., 2008. Designing sampling schemes for effect monitoring of nutrient leaching from agricultural soils. European Journal of Soil Science 59, 292–303.

Brus, D.J., Spätjens, L.E.E.M., de Gruijter, J.J., 1999. A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. Geoderma 89, 129–148.

Cochran, W.G., 1977. Sampling Techniques. Wiley, New York 428pp.

de Gruijter, J., Brus, D., Bierkens, M., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer, Berlin 332pp.

Ding, C., He, X., 2004. K-means clustering via principal component analysis. In: Brodley, C.E. (Ed.), Proceedings of the 21st International Conference on Machine Learning (ICML), Banff, Canada, pp. 225–232.

D'Orazio, M., 2003. Estimating the variance of the sample mean in two-dimensional systematic sampling. Journal of Agricultural, Biological, and Environmental Statistics 8, 280–295.

Ferrari, T.J., Vermeulen, F.M.B., 1955. Soil heterogeneity and soil testing. Netherlands Journal of Agricultural Science 3, 265–275.

Hartigan, J.A., 1975. Clustering Algorithms. Wiley, New York 351pp.

Hartigan, J.A., Wong, M., 1979. A *k*-means clustering algorithm. Applied Statistics 28, 100–108.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York 368pp.

Lloyd, S.P., 1982. Least squares quantization in PCM. IEEE Transactions on Information Theory 28, 129–137.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. In: Le cam, L.M., Neyman, J. (Eds.), Proceedings of 5th Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 281–297.

Minasny, B., McBratney, A., 2002. FuzME version 3.5. Australian Centre for Precision Agriculture, The University of Sydney, Australia ⟨http://www.usyd.edu.au/agriculture/acpa/software/fuzme.shtml⟩ (Accessed June 10, 2010).

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences 30, 683–691.

R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 ⟨ http://www.R-project.org⟩ (Accessed June 10, 2010).

Ribeiro, P.J., Diggle, P.J., 2001. geoR: a package for geostatistical analysis. R-NEWS 1 (2), 14–18 ISSN 1609-3631 ⟨http://cran.r-project.org/doc/Rnews⟩ (Accessed June 10, 2010)..

Royle, J.A., Nychka, D., 1998. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Computers & Geosciences 24, 479–488.

Sacks, J., Schiller, S., 1988. Spatial designs. In: Gupta, S.S., Berger, J.O. (Eds.), Statistical Decision Theory and Related Topics IV, vol. 2. Springer Verlag, New York, pp. 385–399.

van Groenigen, J.W., Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality 27, 1078–1086.

van Groenigen, J.W., Siderius, W., Stein, A., 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87, 239–259.

Walvoort, D.J.J., Brus, D.J., de Gruijter, J.J., 2009. spcosa: Spatial coverage sampling and random sampling from compact geographical strata. R package version 0.2-1 ⟨http://cran.r-project.org/package=spcosa⟩ (Accessed June 10, 2010).