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 (**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, **u** from *A* to *B*, and **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: $\overline{\mathbf{x}}_1 \cdots \overline{\mathbf{x}}_k$ (*k* two-dimensional vectors).

Step 2: Re-allocation of the first cell. Determine if the first cell (with co-ordinate vector **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 **u** to $\overline{\mathbf{x}}_A$ and to $\overline{\mathbf{x}}_B$, 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 **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 **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