Network optimization algorithms and scenarios in the context of automatic mapping

Olivier Baume[1], Albrecht Gebhardt[2], Claudia Gebhardt[2],
Gerard Heuvelink[1], Juergen Pilz[2]

[1]Environmental Sciences / Wageningen University
olivier.baume@wur.nl

[2]Institute of Statistics / University of Klagenfurt
albrecht.gebhardt@uni-klu.ac.at

Abstract. Many different algorithms can be used to optimize spatial network designs. For spatial interpolation of environmental variables in routine and emergency situations, computation time and interpolation accuracy are important criteria. The objective of this work is to compare the performance of different optimization algorithms for both criteria. Both adding to and deleting measurements from an existing network are considered. We applied four algorithms to three datasets with known variogram models, in all cases taking the mean universal kriging variance (MUKV) as the interpolation accuracy measure. Preliminary results show that greedy algorithms that minimize the entropy perform best, both in computing time and MUKV.

1 INTRODUCTION

Optimization of measurement locations is a key issue in spatial sampling design. When network running costs limit the number of measurements, optimization methods allow to choose the locations such that the interpolation error is minimized, either locally or globally. In the case of emergency situations, computation time is an important constraint that the optimization procedure must take into account. For instance, in the case of an accidental radioactivity release by a nuclear power plant, the placement of mobile devices to best assess the location of the plume needs a fast tool to compute the optimal locations of additional measurement devices. Heuvelink et al. (2009) [8] used plume simulations to optimize the locations of additional measurements such that the expected costs of wrong decisions – areas of false positive and false negative detection of the plume – but the solution method is time consuming because it uses loops of geostatistical simulation in an iterative numerical optimization algorithm.

For estimation of global quantities in space, sampling design can be done using design-based and model-based approaches [7]. However, model-based approaches are generally preferred when the objective is to estimate local quantities. In such a case one defines and applies a geostatistical model of spatial variation, which may include spatial trends. The quality of the interpolation for a given design can then be diagnosed using the mean universal kriging variance (MUKV) [3].

In this paper we focus on network design optimization in the context of automatic mapping, where a network is already in use but additional measurements may be collected or where the network must be thinned. We consider the situation in which a user needs to map a natural resource or environmental variable under time, cost and quality constraints. The objectives of this paper are to compare several optimization methods applied to the

adding and deleting scenarios and to guide the user in choosing the most appropriate method in a given situation. We base the comparison on two criteria: computation time and interpolation accuracy (i.e., the MUKV). We test four optimization methods using three well-known datasets from the geostatistical literature, all available in the "Gstat" package library in R [12, 11]. We define a geostatistical model for each dataset and assume these valid. In other words, we assume that the random functions are second-order stationary – possibly after removal of a spatial trend – and do not consider uncertainty in variogram structure and variogram parameters.

2 MATERIAL AND METHODS

2.1 Datasets

Three datasets are taken from the gstat library in R [1]. The "Meuse" and "Jura" datasets [11, 6] contain point samples of soil minerals in two different European regions. The Sic2004 dataset – named after the Spatial Inter Comparison exercise of 2004 [4] – contains Gamma dose rate observations from Germany. The measurement locations of the three datasets are given in Figure 1. The Meuse and Sic2004 datasets show a fairly random distribution of the measurement locations (with a somewhat larger sampling density near the western border for the Meuse dataset), whereas the Jura dataset uses a regular grid with extra measurement clusters and relatively few points near the boundary of the study area.

We defined case-specific geostatistical models and associated interpolation modes to each of the three datasets: kriging with external drift for the Meuse dataset; ordinary kriging for the Jura dataset and universal kriging for Sic2004. The model defined for Meuse dataset is the log of Zinc ppm against the distance of measurements to the river; the model for Jura dataset is an ordinary kriging model of Nitrate ppm; for Sic2004 dataset, Gamma dose against a linear trend in the geographic coordinates. All three model residuals have a known variogram structure composed with a spherical model and nonzero nugget component.

2.2 Methods

Four optimization methods were used for the comparison exercise: a greedy algorithm maximizing entropy reduction; a greedy algorithm minimizing the MUKV – so-called A-optimal; a simulated annealing approach minimizing the MUKV; and a spatial coverage approach. All algorithms were implemented in R [12]. It must be noted that the first two methods use a grid as candidate locations to add new measurements, while the two other methods use the entire space of the study areas as candidate locations.

2.2.1 Greedy algorithm based on Entropy(GE)

Maximum entropy sampling is based on Shannon's information theory. For a continuous random field with probability density $f(z)$, its information is defined as $E(log(f(z)))$, and its entropy as $H(f) = -\int f(z)log(f(z))dz$, i.e. information is negative entropy. In

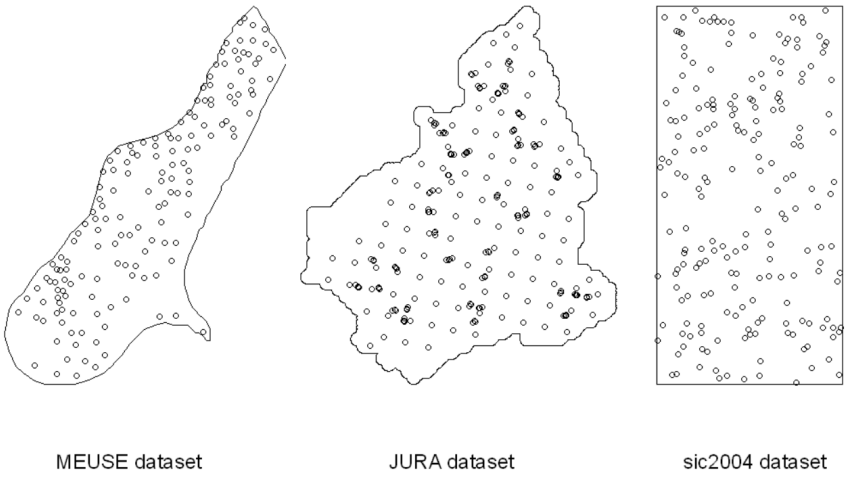

Figure 1: Initial measurement locations of the three datasets.

the spatial sampling context, the goal is to maximize the expected increase in information when changing from the prior to the posterior density. This is equivalent to maximizing $E\{H(f(\theta)) - H(f(\theta|z))\}$, where θ is the vector of model parameters (see Gebhardt (2003) [5]).

We applied this criterion to the case of sampling from a grid of potential sites, which is split into two disjoint subsets: the design points at which the random field will be observed and the complementary set. Shewry and Wynn (1987) [14] proposed an exchange-type algorithm to find the optimal design. Their iterative procedure converges, but does not necessarily lead to an optimum. Ko *et al.* (1995) [9] and Lee and Williams (2000) [10] developed branch-and-bound-methods, which, under certain conditions, lead to the global optimum. The computational complexity of these methods make their practical implementation computationally prohibitive when it comes to choosing several dozens of design points from a grid of several thousands of potential sites.

We propose the implementation of greedy algorithms as suggested in Gebhardt (2003) [5]. At each step, greedy algorithms select the design which leads to the minimum entropy (when adding a new measurement) or to the minimum increase in entropy (when deleting an existing measurement).

2.2.2 Greedy algorithm based on the Kriging Variance (GKV)

As a second optimization method, we propose to apply the greedy algorithms to the mean (universal) kriging variance instead of the entropy. Whereas there is a well-developed theory for optimum regression designs, there is no analogous catalogue of methods for spatial regression prediction and interpolation design. The basic difficulty stems from the fact that classical design functionals are no longer convex, due to correlated errors. When we employ ordinary or (Bayesian) universal kriging for prediction, it is natural to use the corresponding (Bayesian) kriging variance as a measure of prediction accuracy.

2.2.3 Simulated Annealing (SA)

Simulated Annealing (SA) was also applied to optimize the MUKV. Brus and Heuvelink (2007) [3] applied SA to MUKV minimization and we refer to it for a detailed presentation of the method. The basic idea of SA is to derive a new candidate design from the current by perturbing the current design slightly, evaluating the criterion, accepting the new design when the criterion has improved, and accepting it with some probability when the criterion has deteriorated. Simulated annealing requires several parameters to be defined. The initial probability to accept worsening designs, the 'cooling' schedule and a stopping criterion of the optimization procedure have to be chosen in order to avoid to be trapped in local optima and to avoid a too long procedure. The selection of the best value for these parameters is largely dependent on each specific application.

2.2.4 Spatial Coverage (SC)

The final optimization method targets at geometrical criteria. Geometrical criteria are based only on the spatial configuration of the measurements and not on the measurement values or underlying geostatistical model. SC algorithms are more often used in the context of design-based sampling design to estimate global quantities such as the global mean [7].

In this work two SC algorithms were applied. In the scenario where measurement locations are added, we used the algorithm developed by Brus et al. [2]. Their method is based on the mean squared distance criterion, which allows optimization with k-means. In the case of deleting measurements from the initial dataset, we used the definition of coverage as in Royle and Nychka (1998) [13]. The heuristic search is a point swapping algorithm, similar to the one used in the greedy algorithms.

3 RESULTS AND DISCUSSION

Table 1: SIC 2004 dataset: adding and deleting measurements to the initial design – GE:Greedy Entropy; GKV: Greedy Kriging Variance; SA: Simulated Annealing; SC: Spatial Coverage.

Scenario	Method	Time(s)	MUKV		Scenario	Method	Time(s)	MUKV
Add 1	GE	2.0	116.05		Delete 1	GE	0.6	116.26
Add 1	GKV	130.1	116.04		Delete 1	GKV	3.7	116.38
Add 1	SA	414.3	116.06		Delete 1	SA	261.5	116.26
Add 1	SC	9.5	116.07		Delete 1	SC	4.4	116.28
Add 10	GE	3.07	114.56		Delete 10	GE	4.3	116.26
Add 10	GKV	1211.3	114.76		Delete 10	GKV	33.5	116.75
Add 10	SA	1144.4	114.97		Delete 10	SA	438.48	116.51
Add 10	SC	10.0	115.08		Delete 10	SC	174.7	117.82
Add 50	GE	9.0	110.89		Delete 50	GE	14.7	118.15
Add 50	GKV	7196.5	111.27		Delete 50	GKV	133.2	120.03
Add 50	SA	2242.4	111.53		Delete 50	SA	566.6	118.51
Add 50	SC	11.2	112.14		Delete 50	SC	2649.3	129.35

The results for Sic2004 are presented in Table 1. On the left side, results of scenarios with adding 1, 10 and 50 measurements are given (to the 200 existing locations of the initial dataset). On the right side, results of scenarios with deleting 1, 10 and 50 measurements are presented. The comparison for the Meuse and Jura datasets give similar results.

In terms of minimizing the mean interpolation error variance, using the Entropy criterion (GE algorithms) gives the best results. In the SA case, different configurations of parameters were tested but none of these performed better than the GE algorithm.

Note that the computation times reported in Table1 may have been influenced by the implementation and choice of the grid with candidate locations. Also, the greedy algorithms used R interfaces with Fortran and C to run the optimization process, which may speed up the algorithms. In the scenarios where measurements were deleted, the greedy and spatial coverage algorithm use the same type of swapping procedure and hence should lead to comparable computation times, which is not the case. It must also be noted that in the greedy algorithms, the grid size of eligible new locations had to be reduced to about a third of the original size to ensure that R could allocate the required memory. This also lead to shorter computation times.

Theoretically, simulated annealing should reach the best MUKV performance because it explores the entire domain of the study area and is not restricted to grid nodes. However, time is a limiting factor for SA to obtain better results than greedy algorithms. Moreover, Gebhardt [5] showed that in many cases the initial heuristic solutions from greedy algorithms were already close to optimal.

4 CONCLUSION

From the preliminary results presented here, we conclude that greedy algorithms outperform the other algorithms, both for the case of adding and deleting measurements. Greedy algorithms yield the best results, both for MUKV and computation time. However, for a fair comparison in terms of computation time, all the computationally expensive parts of the methods should be implemented in the same or similar lower level language, such as Fortran or C.

ACKNOWLEDGMENTS

This work is funded by the European Commission, under the Sixth Framework Program, by the Contract N. 033811 with the DG INFSO, action Line IST-2005-2.5.12 ICT for Environmental Risk Management. The views expressed herein are those of the authors and are not necessarily those of the European Commission. Authors are grateful to two anonymous reviewers for their suggestions.

REFERENCES

[1] R.S. Bivand, E.J. Pebesma, and V. Gómez-Rubio. *Applied spatial data analysis with R*. Springer-Verlag, Berlin Heidelberg, 2008.

[2] D. Brus, J. de Gruijter, and J. van Groenigen. Designing spatial coverage samples using the k-means clustering algorithm. In A. McBratney amd M. Voltz P. Lagacherie, editor, *Digital Soil Mapping: An Introductory Perspective*, Developments in Soil Science, vol. 3., Amsterdam, 2006. Elsevier.

[3] D. Brus and G.B.M. Heuvelink. Optimization of sample patterns for universal kriging of environmental variables. *Geoderma*, 138:86–95, 2007.

[4] G. Dubois. Introduction to the spatial interpolation comparison (sic) 2004 exercise and presentation of the datasets. *Applied GIS*, 1(2), 2005.

[5] C. Gebhardt. *Bayesian Methods for Geostatistical Design*. University of Klagenfurt, 2003.

[6] P. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford Univ. Press, New-York, 1997.

[7] J. de Gruijter, D. Brus, M. Bierkens, and M. Knotters. *Sampling for Natural Resource Monitoring*. Springer-Verlag, Berlin Heidelberg, 2006.

[8] G.B.M. Heuvelink, Z. Jiang, S. de Bruin, and C.J.W. Twenhöfel. Optimization of mobile radioactivity monitoring networks. *International Journal of GIS*, page in press, 2009.

[9] C. Ko, J. Lee, and M. Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43:684–691, 1995.

[10] J. Lee and J. Williams. A linear integer programming bound for maximum entropy sampling. *IBM Research Report*, Sept. 2000.

[11] E.J. Pebesma. Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, 30(6):683–691, 2004.

[12] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[13] J.A. Royle and D. Nychka. An algorithm for the construction of spatial coverage designs with implementation in splus. *Computers & Geosciences*, 24:479–488, 1998.

[14] P. Shewry and H.P. Wynn. Maximum entropy sampling. *J. Applied Statistics*, 14:165–170, 1987.