Second-phase spatial sampling: local and global objectives to optimize sampling patterns

Dr. Eric Delmelle
Department of Geography and Earth Sciences
University of North Carolina - Charlotte
Charlotte, U.S.A.
eric.delmelle@uncc.edu

***Abstract*—In geographic sampling, once initial samples of the primary variable have been collected, it is possible to take additional measurements, an approach known as second-phase sampling. It is generally desirable to collect such additional samples in areas far away from existing observations to reduce redundancy, which coincide with regions where the kriging variance is maximum. However, the kriging variance is independent of data values and computed under the assumption of stationary spatial process, which is often violated in practice. Weighting the kriging variance with another criterion, giving greater sampling importance to locations exhibiting significant spatial roughness, can serve as an alternative objective (Delmelle & Goovaerts 2009). This roughness is computed by a spatial moving average window. Another objective function consists of locally determined variogram models to obtain local kriging variances, reflecting non-stationarity (Haas 1990). The benefits and drawbacks of these three approaches are illustrated in a case study using an exhaustive remote sensing image. Combinations of first-phase systematic and nested sampling designs (or patterns) are generated, while the location of additional observations is guided in a way which optimizes each objective function. Augmented sampling sets minimizing the weighted kriging variance or minimizing the kriging variance computed by local variograms lead to better reconstruction of the true image, while patterns minimizing the kriging variance computed by a global variogram lead to reconstruction similar to a random addition. This indicates that accounting for spatial roughness in second-phase sampling improves the overall accuracy of the prediction.**

Keywords: sampling pattern, local variance, spatial roughness, weighted kriging variance

I. INTRODUCTION

When surveying a phenomenon characterized by spatial variation, it is necessary to find optimal sample locations in the study area *D*. This problem is referred to spatial or two-dimensional sampling (Haining 2003, de Gruijter et al. 2006, Muller 2007, Delmelle 2009). Sampling efforts are usually concentrated in areas deemed to be critical. Once initial samples of the primary variable have been collected and the variogram estimated, the primary variable is interpolated throughout a study region using kriging for instance. Second-phase sampling is necessary when interpolation results from the initial set are judged inaccurate due to a lack of samples or poor sampling design. The lack of accuracy is generally measured by the kriging variance, which is a function of sampling patterns, sampling density and covariance structure. The kriging variance is minimal at existing sample points, and increase away from them. The kriging variance is unfortunately misused as a measure of reliability of the kriging estimate (Deutsch and Journel, 1997).

One objective in second-phase sampling is to collect new samples to reduce the kriging variance or uncertainty by as much as possible (Van Groenigen, Siderius and Stein, 1999). The addition of new samples to minimize the kriging variance allocates new measurements at intermediate positions between existing samples while ignoring the underlying spatial variation (Delmelle and Goovaerts 2009). However, the magnitude of the spatial variation can be expressed in the form of weights combined with the kriging variance. A weighted objective aims at collecting new samples to maximize the change in weighted kriging variance (Cressie 1991, Rogerson et al. 2004).

This paper compares the benefits of allocating new samples according to the optimization of three objectives: (1) the reduction of the kriging variance, (2) the change in weighted kriging variance where the weights reflect spatial roughness and (3) the decrease in kriging variance computed from locally determined variograms. The quality of each objective is measured by comparing how far the predictions are from the true image.

II. SECOND-PHASE SAMPLING

In a first sampling phase, a variable of interest *Y* is collected at *m* locations, $y(\boldsymbol{s}_i)\ \forall \mathrm{i} = 1 \ldots \mathrm{m}$. For notation purposes, the initial sampling set is denoted *M*. Differences in data values can be plotted using a variogram cloud, which graphs the difference in value between two points *i* and *k* separated by distance h.

$$\gamma(h) = \left(y(\boldsymbol{s}_i) - y(\boldsymbol{s}_k)\right)^2, \forall i, k \in M,\ i \neq k, d(\boldsymbol{s}_i, \boldsymbol{s}_k) = h \quad (1)$$

The variogram cloud is sensitive to outliers and it may be computationally time-consuming to find an appropriate fitting model. However, it is not subject to additional parameters which must be tuned, such as lag size, number of lags and tolerance. Given three parameters (nugget effect, sill, range), an exponential variogram model is fitted on the variogram cloud using a non-linear least square strategy. When the number of pairs of data points is high, this fitting

procedure is rather time-consuming and it is desirable to rely on optimization routine. In this paper, a Nelder-Mead algorithm is used to fit the cloud.

A. Kriging variance

Kriging is performed over a set of grid nodes $\boldsymbol{s_g}(\mathrm{g} = 1,2 \ldots \mathrm{G})$. The associated kriging variance $(\sigma_K(\boldsymbol{s_g}))^2$ measures the uncertainty of the prediction:

$$(\sigma_K(\boldsymbol{s_g}))^2 = \sigma^2 - \mathbf{c}^{\mathrm{T}}(\boldsymbol{s_g})\mathbf{C}^{-1}\mathbf{c}(\boldsymbol{s_g}), \qquad (1)$$

where $\mathbf{C}^{-1}$ is the inverse of the covariance matrix **C** obtained the covariogram function. Integrating equation (1) over the study region yields the total kriging variance, but it is computationally easier to calculate an average kriging variance (AKV) over a fine set of grid node *G*:

$$\mathrm{AKV} = \frac{1}{|\mathrm{G}|}\sum_{\mathrm{g}\in\mathrm{G}} (\sigma_K(\boldsymbol{s_g}))^2, \qquad (2)$$

A common second-phase sampling objective is to find an augmented sampling pattern S which will maximize the change in kriging variance by as much as possible over the study region (Van Groenigen, Siderius and Stein, 1999):

$$\mathrm{Q[S]} = \frac{1}{|\mathrm{G}|}\sum_{\mathrm{g}\in\mathrm{G}} \left| \left(\sigma_K^{old}(\boldsymbol{s_g})\right)^2 - \left(\sigma_K^{new}(\boldsymbol{s_g})\right)^2 \right| \qquad (3)$$

$$\max_{\{\boldsymbol{s}_{m+1\ldots}\boldsymbol{s}_{m+n}\},n\in\boldsymbol{P}} \mathrm{Q[S]}, \qquad (4)$$

where Q stands for the objective function. The *n* additional samples are selected from a set of *p* potential locations $\boldsymbol{P}$ which, for simplicity, coincide with the set of grid nodes.

B. Weighting the kriging variance

The kriging variance does not account for the variation of the kriging estimates reflected by differences in data value between nearby grid nodes. Let $\hat{y}(\boldsymbol{s_g})$ be the interpolated value of the primary variable *Y* at node $\boldsymbol{s_g}$.

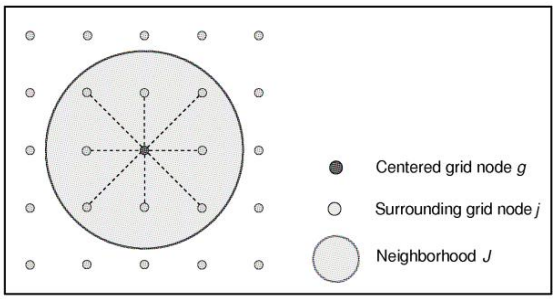

Figure 1. A 3x3 filter facilitates the detection of spatial variation between a grid node from neighboring nodes. In computing the grid node specific weight, greater importance is sgiven to nearby nodes.

The objective consists of estimating by how much that grid node is different in value from its surrounding nodes $\boldsymbol{s_j}$ defined by a neighborhood *J* (Delmelle and Goovaerts 2009). From Fig. 1, a window is constructed around each grid node $\boldsymbol{s_g}$ encompassing its neighbors. The squared difference in interpolated values between the central grid node $\hat{y}(\boldsymbol{s_g})$ and the surrounding ones $\hat{y}(\boldsymbol{s_j})$ is computed. The process moves from one node to another and is repeated for each grid point. The squared difference is then summed over the set *G*. To regulate the importance of nearby nodes, a distance factor $d(\boldsymbol{s_j}, \boldsymbol{s_g})$ and a parameter β are introduced in the weight:

$$\lambda(\boldsymbol{s_g}) = \sum_{\mathrm{j}\in\mathrm{J},\mathrm{j}\neq\mathrm{g}} \frac{d(\boldsymbol{s_j},\boldsymbol{s_g})^{-\beta} * (\hat{y}(\boldsymbol{s_g}) - \hat{y}(\boldsymbol{s_j}))^2}{\sum_{\mathrm{j}=1,\mathrm{j}\neq\mathrm{g}}^{\mathrm{J}} d(\boldsymbol{s_j},\boldsymbol{s_g})^{-\beta}}, \qquad (5)$$

To account for the spatial roughness of the primary variable, equation [3] is modified by introducing a location-specific weighting factor, defined in equation [5]. The second-phase sampling problem is then formulated as a single-weighted objective (Cressie 1991):

$$\max_{\{\boldsymbol{s}_{m+1\ldots}\boldsymbol{s}_{m+n}\},n\in\boldsymbol{P}} \frac{1}{|\mathrm{G}|}\sum_{\mathrm{g}\in\mathrm{G}} \left(\frac{\lambda(\boldsymbol{s_g})}{\mathrm{argmax}_{\boldsymbol{s_g}\in\mathrm{G}}(\lambda(\boldsymbol{s_g}))}\right)^{\alpha} \left| \left(\sigma_K^{old}(\boldsymbol{s_g})\right)^2 - \left(\sigma_K^{new}(\boldsymbol{s_g})\right)^2 \right|, \qquad (6)$$

where α is a parameter controlling the importance given to the weights. A value of α= 1 is used in this paper, but when α= 0, equation [6] reduces to equation [3].

C. Kriging variance from locally determined variograms

The kriging variance at each grid node is generally calculated according to a similar variogram model. It is, however, feasible to compute a locally-varying covariance model for each grid node, and obtain a kriging variance reflecting local variation (Haas 1990). This directly addresses the problem of covariance non-stationarity. Practically, a search window containing a minimum number of initial sampling points is imposed around each grid node $\boldsymbol{s_g}$. These points are used to compute a locally determined variogram cloud. Locally dependent variogram parameters are determined for each grid node, and the kriging variance locally computed. In this context, the second-phase sampling objective allocates samples in these regions characterized by strong local variance. These points are used to compute a locally determined variogram cloud. Locally dependent variogram parameters are determined for each grid node, and the kriging variance locally computed. In this context, the second-phase sampling objective allocates samples in these regions characterized by strong local variance.

III. SIMULTANEOUS GREEDY ADDITION

A simultaneous addition consists of supplementing the initial set *M* with a set *N* of *n* additional measurements in one time. The major concern lies in the selection of those points. A total enumeration is not recommended because of combinatorial explosion The goal of the greedy algorithm in simultaneous addition is to supplement the initial sampling set by adding *n* points exhibiting a high kriging variance (or weighted kriging variance) value. The major problem consists of locating these new samples in a way that they are

not too close to each other. A solution consists of introducing a minimum separating distance $D_{min} = \zeta$ among selected points. Figure 2 illustrates how the selection of 10 new samples is implemented, depending on the distance constraint. The curve represents the weighted kriging variance, and is interpolated over 50 candidate locations. These candidate points, separated by 100m-interval, range from 0 to 5000m. From the graph on the left, the point with the highest function value is selected by default, and the closest second point is added if $d(\boldsymbol{s}_{m+1}, \boldsymbol{s}_{m+2}, \boldsymbol{s}_i\ (\forall i = 1 \ldots m)) \geq 100m$. This continues in a sequential fashion until all *n* points have been found.

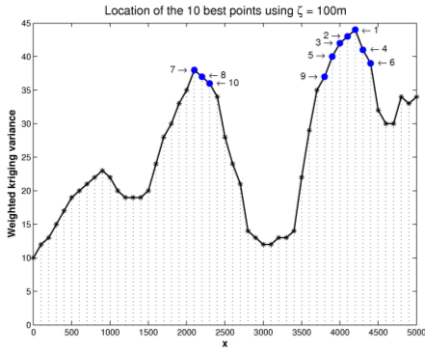

Figure 2. One-dimensional illustration of the simultaneous greedy addition under distance constraints $D_{min} \geq$ 100m. Black dots are potential samples, while blue dots are the resulting locations of the new samples, after the greedy algorithm has been applied.

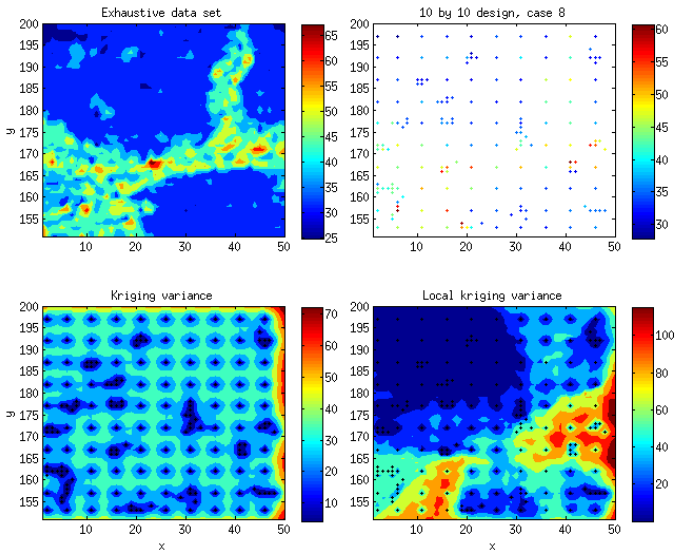

Figure 3. Top left figure: contour of the exhaustive dataset. The 10by10 design consists of 100 gridded points while the remaining 72 are clustered around 18 samples. The bottom figures depict the kriging variance and local kriging variance, respectively.

IV. Case Study

A small case study illustrates the spatial sampling problem structure. The objective is to allocate a set of new samples using a greedy addition algorithm in order to maximize the change in either three objectives. An exhaustive remote sensing image is used to illustrate the procedure but also to test which augmented sampling patterns lead to better predictions. The sampling strategy is tested using a SPOT High Resolution Visible (HRV) scene of a 4km^2 area covered by tropical forests and savannah (Goovaerts 2002). Fig. 3 illustrates the dataset. The image is divided into 50 rows and columns, yielding a set of 2500 pixels. All computational results are obtained using Matlab v. 7.6. running on a Linux desktop, while the Nelder-Mead function *fminsearch* was used to fit the variogram cloud.

A. Initial sampling patterns

To guarantee coverage over the study region and to estimate the variation of the variogram at small distances, initial patterns are designed using a combination of systematic and nested sampling. A 10x10 pattern divides the study region into 100 cells or intervals (10 rows and columns), generating 1 systematic sample of 100 sample points. The coordinates of the first sample are purposively chosen within the first interval and correspond to a location close to the origin. Depending on the location of the first sample, the remaining 99 samples are aligned regularly by the size of the cell. Nested sampling is carried out next by adding 4 more clustered observations around 18 randomly chosen systematic samples, yielding a total sampling size of 172 observations. Next, a fixed neighborhood is drawn around each of the 18 randomly selected samples and 4 samples are selected at random within this neighborhood. Twenty five total different sets of gridded and clustered data are selected to attenuate the impact of sampling fluctuations. This is implemented by shifting the origin of the systematic sampling pattern, while new clustered data are selected at random. The pixel value at each sampling location $y(\boldsymbol{s}_i)$ is extracted. Exponential models with varying variogram parameters are tested, and the model exhibiting the lowest sum of squares is kept. The variable *Y* is interpolated over the set of grid points *G* (2500 nodes) using ordinary kriging. The average absolute error of prediction (MAE) is reported:

$$MAE = \frac{1}{|G|}\sum_{g \in G}\left|\hat{y}(\boldsymbol{s}_g) - y(\boldsymbol{s}_g)\right| \qquad (7)$$

where $\hat{y}(\boldsymbol{s}_g)$ is the interpolated value of the primary variable at a grid node, and $y(\boldsymbol{s}_g)$ its true value. Ultimately, the goal is to find an augmented sampling pattern reducing equation [7] by as much as possible, leading to a better reconstruction of the "true" image.

B. Spatial roughness

For each of the 25 sampling realizations, local variations in interpolated values are computed with equation [5] for each grid point using its 4 nearest neighbors (|J| = 4), and parameter β= 1.5. Fig. 4 illustrates the weights reflecting the spatial roughness of sampling realization case 8, which are then multiplied with the kriging variance map. The weighted kriging variance map contrasts from the map of the weights in that more importance is given to location away from initial points.

C. Second-phase sampling

Given an initial sampling set M, the objective is to select an additional set *N* (|N| = 10, 20, 30, 40 and 50) at which the

exhaustive dataset will be sampled. The underlying goal is to gather information in those strategic locations, (a) where the kriging variance is maximum, (b) where the weighted kriging variance is the highest and (c) where the local kriging variance is the greatest. Although some regions on the edge of a study region may exhibit high kriging variance, those are not considered as potential locations as they will have a minimal impact on the Eq. [3]. To test whether these augmented designs lead to better predictions (i.e. reconstruction of the "true" image), we compare the merits of these 3 approaches to the average of 100 random simulations (100simulations for each of the 25 sampling realizations).

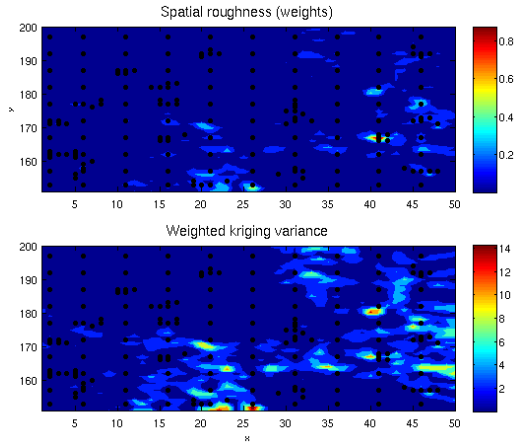

Figure 4. The weighted kriging variance is a combination of weights (top figure) with the kriging variance (figure 3).

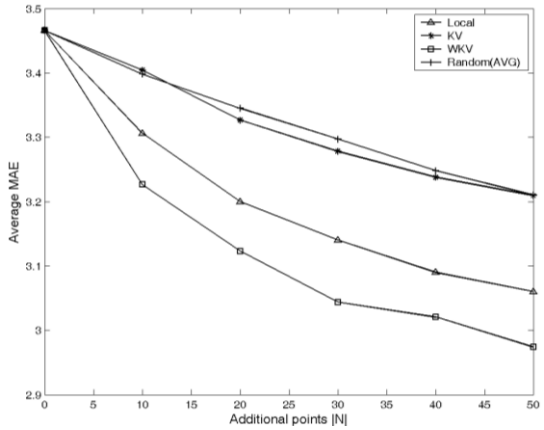

Figure 5. Reduction in the Mean Average Error (MAE) from the true image as a function of the second-phase sampling set size (|N|), and the methodology adopted to allocate these points. Random (AVG) is the average of 100 simulation of augmented sets, for each sampling realization.

Fig. 5 graphs the reduction of MAE for the different sampling strategies as a function of the number of additional points (|N|). Regardless of the strategy adopted to allocate additional samples, the overall improvement in reducing MAE increases when a greater number of additional points are added to the initial set. The relative reduction decreases however as the size of the second-phase sampling becomes larger. Augmented sampling designs maximizing the change in weighted kriging variance (WKV) strongly contrast with augmented designs maximizing the kriging variance (KV) alone. The former leads to better reconstruction of the true image, which is very significant. The same pattern is observed for sampling designs maximizing the change in kriging variance, when the latter is computed from local variograms (Local). Interestingly, designs maximizing the kriging variance alone perform slightly better in reducing the overall error than by random addition alone.

V. Discussion and Future Research

In this paper, three different sampling objectives have been used to augment an initial sampling set. The objectives which maximize the change in weighted kriging variance or minimize the kriging variance computed from local variograms tend to allocate new samples in areas of strong spatial roughness. Additional samples optimizing the change in kriging variance are allocated more geometrically, ignoring the underlying spatial variation. Empirical results show that additional samples added in areas of strong spatial variations have a greater impact in improving the overall interpolation accuracy, rather than measurements added in a geometric fashion, far away from existing points. Future research is necessary in 3 different arenas. First, the results in this paper should be extended to various sampling densities of lower and higher order (e.g. 6by6, 8by8, 12by12, etc…). Secondly, the additional points have been allocated following a greedy approach, which is suboptimal. Other search algorithms such as simulated annealing or tabu search may provide better results. Finally, a variogram model was fitted onto the variogram cloud, while computationally it is easier to work with empirical variograms rather than clouds.

Acknowledgements

The author would like to thank Pierre Goovaerts, Timothy Haas, Budiman Minasny, Werner Muller, Edzer Pebesma for their valuable input and comments.

References

Cressie, N. (1991). *Statistics for spatial data*. Wiley, New York, USA.

De Gruijter, J., Brus, D.J., Bierkens, M.F.P. and Knotters M. (2006). *Sampling for natural resource monitoring*. Springer, 332p

Delmelle, E. (2009). Spatial sampling. In: Rogerson, P. and S. Fotheringham (eds). *The SAGE handbook of spatial analysis*. Sage Publication. 528p.

Delmelle, E., and P. Goovaerts (2009). Second-phase sampling designs for non-stationary variables. *Geoderma*. vol 153: 205-216.

Deutsch, C.V. and Journel, A.G. (1997). *Gslib: Geostatistical software library and user's guide*. Oxford University Press, 2nd edition, 369p.

Goovaerts, P. (2002). Geostatistical modelling of spatial uncertainty using *p*-field simulation with conditional probability fields. *International Journal of Geographical Information Science*, vol. 16(2): 167-178.

Haining, R.P. (2003). *Spatial data analysis : Theory and practice*. Cambridge University Press. 452p.

Muller, W.G. (2007). *Colleting spatial data: Optimum designs of experiments for random fields.* 3rd ed. Springer, Germany. 242p.

Rogerson, P.A., Delmelle, E.M., Batta, R., Akella, M.R., Blatt, A and Wilson, G., (2004). Optimal sampling design for variables with varying spatial importance. *Geographical Analysis*, vol. 36: 177-194.

Van Groenigen, J.W., Siderius, W. and Stein, A. (1999). Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, vol. 87: 239-259.