A geostatistical approach to optimize sampling designs for local forest inventories

Jaime Hernández and Xavier Emery

Abstract: In forest management, it is of interest to obtain detailed inventories such that the local prediction errors on forest attributes are less than a prespecified threshold, while keeping the number of ground samples as low as possible. Given an initial sampling design, we propose an algorithm to determine the additional sample locations. The algorithm relies on two tools: geostatistical simulation, which allows measuring the uncertainty in the values of the attribute of interest, and simulated annealing, which allows finding an infill design that minimizes a given objective function. The proposed approach is applied to a data set from a *Prosopis* spp. plantation located in the Atacama Desert, in which the measured attribute is the rate of tree survival.

Résumé : En gestion des forêts, il est intéressant d'obtenir des inventaires détaillés tels que les erreurs de prédiction locale sur les variables forestières soient inférieures à un seuil prédéfini, tout en réduisant le plus possible le nombre d'échantillons de terrain. Étant donné un plan d'échantillonnage initial, nous proposons un algorithme pour déterminer l'emplacement d'échantillons supplémentaires. L'algorithme repose sur deux outils : la simulation géostatistique, qui permet de mesurer l'incertitude sur les valeurs de la variable étudiée, et le recuit simulé, qui permet de trouver un plan d'échantillonnage minimisant une fonction objectif donnée. L'approche proposée est appliquée à un jeu de données d'une plantation de *Prosopis* située dans le désert d'Atacama, où la variable mesurée est le taux de survie des arbres.

Introduction

Forest inventories usually are done by predicting relevant forest attributes on the basis of information from ground sampling and remote sensing (aerial photographs or satellite images) by means of statistical techniques (Loetsch et al. 1973; Burkhart et al. 1984; De Vries 1986; Husch et al. 1993; Philip 1994). Within this framework, the choice of the sampling design is a long-standing problem that involves budget and time constraints, measurement errors, as well as the expected prediction errors on the attributes under consideration (Cochran 1977; De Gruijter et al. 2006; Gilabert 2007; Mandallaz 2007).

Because global inventories are of little practical use in precision forest management, it is of interest to obtain detailed (local) inventories with prediction errors and confidence levels similar to those required for the entire area. Such local inventories are needed to efficiently assign bucking pattern schemes, silvicultural treatments, and harvesting machinery. In this context, our objective is to present a geostatistical approach to defining a cost-effective sampling design such that, at any local (within-stand) area, prediction errors remain bounded. Such a problem is common in all forest stand attributes for any kind of forest type, so in global terms, the proposed approach can be extremely useful in forestry.

The idea for this work originated during the planning phase of a forest inventory to assess the condition of a new plantation in Chile. According to Chilean laws, when a new plantation is established, an incentive bonus of 75% of the plantation cost is offered to the owners. To obtain this bonus, the most important requirement is that, 1 year after initial plantation, the rate of plant survival is >75%. The Chilean Forest Service (CONAF) verifies this requirement by means of traditional inventory techniques and tolerates a maximum error of 10% (with a 95% confidence) over the mean survival rate. Because plantations have a natural spatial variation and to maximize bonus recovery, owners can divide the total area into smaller stands and discriminate between areas with less than the threshold value of 75% survival rate and areas with more than this threshold.

Data and model

Presentation of the data

The data under study consist of 738 measurements from a *Prosopis* spp. plantation located in the Tamarugal pampa, a region of the Atacama Desert, northern Chile (Fig. 1). The measured attribute is the number of surviving trees per hectare.

The trees were planted on a quasiregular grid with a mesh of approximately 10 m × 10 m, and an irrigation system was installed to water them by drip irrigation using underground water. The data were collected later by sampling plots of 60 m × 40 m in a regular array oriented north–south and east–west with a separation of 200 m between adjacent plots (Fig. 2). The coordinates of the northeastern vertex of all plots were uploaded into simple frequency global positioning system (GPS) devices to navigate to every plot location. Because of the flatness of the Chilean altiplano and the very favorable atmospheric conditions, the errors on

Received 26 August 2008. Accepted 26 March 2009. Published on the NRC Research Press Web site at cjfr.nrc.ca on 6 August 2009.

J. Hernández.[1] Department of Forest Resources Management, University of Chile, Santiago, Chile.
X. Emery. Department of Mining Engineering, University of Chile, Santiago, Chile.

[1] Corresponding author (e-mail: jhernand@uchile.cl).

doi:10.1139/X09-048

Fig. 1. Location of the study area in the Atacama Desert, northern Chile.

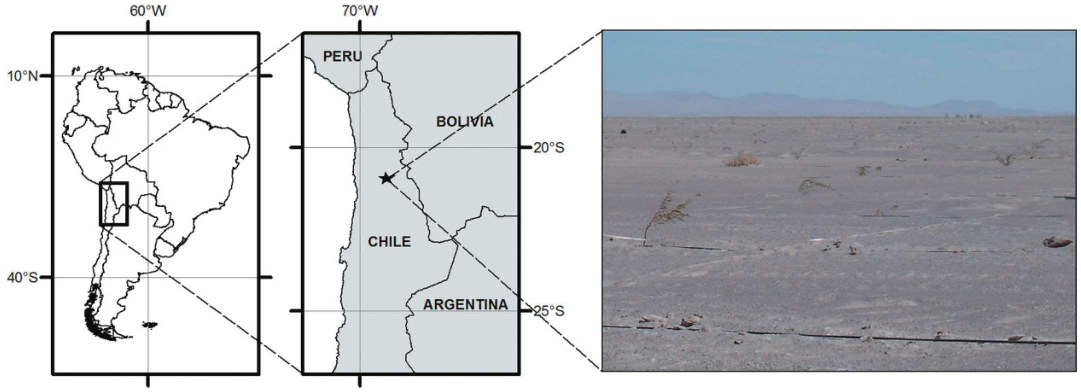
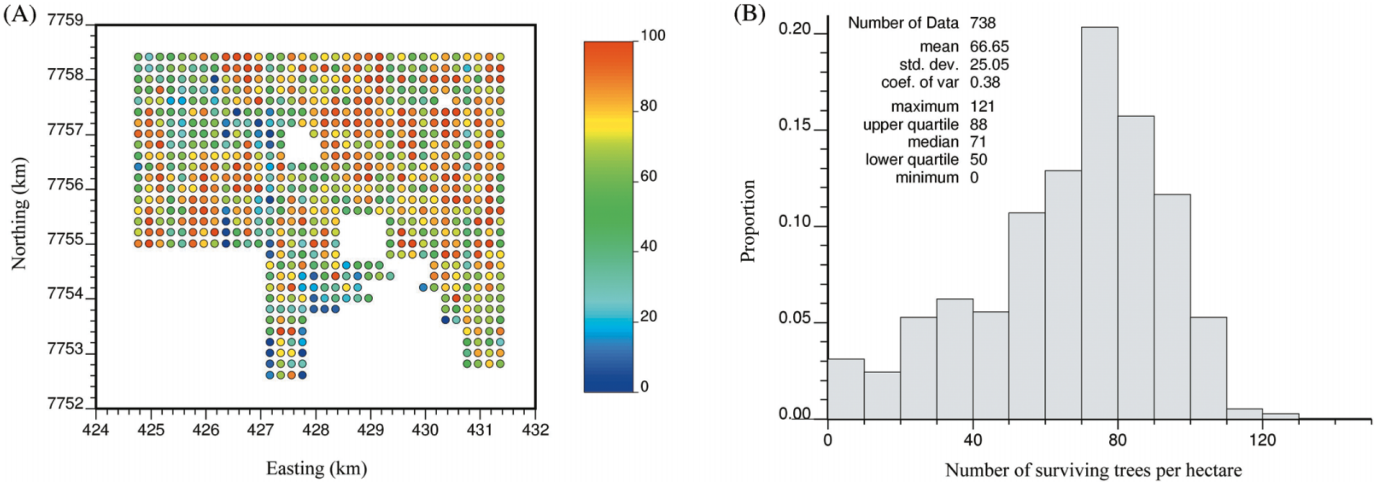


Fig. 2. (A) Location map and (B) histogram of available data.



the coordinates remained less than 2 m, therefore negligible. In each plot (an area of 0.24 ha), the total numbers of living and dead plants were recorded and rescaled to 1 ha. A plant was considered living if it had at least one green leaf; otherwise it was classified as dead, even if it had a green stem.

Geostatistical modeling

Geostatistics provides a set of tools and methods for modeling the spatial distribution and variability of forest attributes and has been used in forest inventory over the past decades (Matérn 1960; Marbeau 1976; Hock et al. 1993; Mandallaz 2000; Nanos et al. 2004; Sales et al. 2007).

In the following, D is the domain under consideration, $\mathbf{s}$ is a vector of spatial coordinates (easting and northing) in D, and $z(\mathbf{s})$ is the value of the attribute (number of surviving trees per hectare) on a plot centered at $\mathbf{s}$. Following the classical geostatistical formalism, $z = \{z(\mathbf{s}), \mathbf{s} \in D\}$ is a regionalized variable (i.e., a variable that exhibits spatial continuity) and can be interpreted as a realization of a parent random field $Z = \{Z(\mathbf{s}), \mathbf{s} \in D\}$. We will characterize Z by assuming that it can be transformed into a stationary Gaussian random field $Y = \{Y(\mathbf{s}), \mathbf{s} \in D\}$, i.e., a random field whose finite dimensional distributions are multivariate normal and are invariant under spatial translation:

[1] $$\forall \mathbf{s} \in D, Z(\mathbf{s}) = \phi[Y(\mathbf{s})]$$

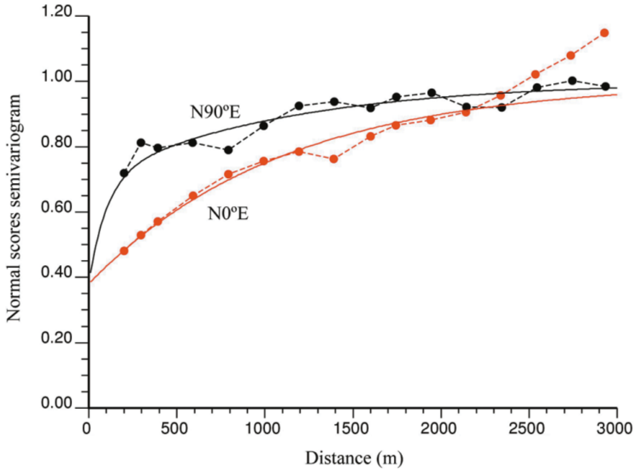
where ϕ is a nondecreasing function called Gaussian anamorphosis (Chilès and Delfiner 1999). In practice, such a function can be inferred empirically by constructing a quantile–quantile transformation between the data histogram (Fig. 2B) and the normal distribution. Details about the inference and modeling of the Gaussian anamorphosis can be found in standard geostatistical textbooks (Rivoirard 1994; Chilès and Delfiner 1999).

The model is completed by fitting the semivariogram of the transformed data (normal scores). In the present case, the semivariogram model consists of a nugget effect plus two spherical structures (Fig. 3, Table 1). One observes that the spatial continuity (spatial correlation) is greater along the north–south direction, as indicated by the larger range in this direction.

Conditional simulation

Conditional simulation consists in constructing a set of realizations of the random field Z that reproduce the values of the attribute measured at the data locations (Chilès and Delfiner 1999). Each realization mimics the spatial continuity of the attribute and provides a scenario of how the actual (unknown) values can be distributed in space, given the ob-

Fig. 3. Sample and modeled semivariogram of transformed data.



served data values. Conditional realizations are helpful to assess the uncertainty prevailing at a given location or the joint uncertainty over several locations.

A conditional simulation of the Gaussian random field Y is obtained as follows (Journel 1974; Chilès and Delfiner 1999):

$$[2] \qquad \forall \boldsymbol{s} \in D, Y_{\mathrm{CS}}(\boldsymbol{s}) = Y_{\mathrm{NCS}}(\boldsymbol{s}) + \sum_{\alpha=1}^{n} \lambda_{\alpha|n}(\boldsymbol{s})[Y(\boldsymbol{s}_\alpha) - Y_{\mathrm{NCS}}(\boldsymbol{s}_\alpha)]$$

where $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, …, $\boldsymbol{s}_n$ are the conditioning data locations, $\lambda_{1|n}(\boldsymbol{s})$, $\lambda_{2|n}(\boldsymbol{s})$, …, $\lambda_{n|n}(\boldsymbol{s})$ are the weights assigned to $Y(\boldsymbol{s}_1)$, $Y(\boldsymbol{s}_2)$, …, $Y(\boldsymbol{s}_n)$ when predicting $Y(\boldsymbol{s})$ by simple kriging (i.e., kriging with a known mean), and Y_{NCS} is a nonconditional simulation of Y (i.e., a simulation that is not constrained to reproduce the values measured at the data locations). There exist a wide variety of algorithms to construct nonconditional realizations of Y (Lantuéjoul 1994). If the mean value of Y is poorly estimated (for example, because of too few data), one can substitute ordinary kriging (i.e., kriging with an unknown mean) for simple kriging to obtain conditional realizations (Journel and Huijbregts 1978; Emery 2007); in this paper, only simple kriging will be used.

In practice, the random field Y is back-transformed to the original unit (Z) using the anamorphosis function (eq. 1). An illustration is given in Fig. 4.

Leave-one-out cross validation

Before addressing the sampling design problem, we will validate the capability of the fitted geostatistical model to measure the uncertainty in the attribute at nonsampled locations. The idea of cross validation is to model the uncertainty at each data location via a probability distribution by temporarily removing the datum at this location. The performance of the uncertainty model can be assessed by examining the accuracy of probability intervals derived from the modeled distributions (Goovaerts 2001).

Specifically, let $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, …, $\boldsymbol{s}_n$ denote the available data locations. The validation consists of the following steps:

Table 1. Parameters of semivariogram model for the Gaussian random field Y.

		Range (m)	
Type	Sill	N0°E	N90°E
Nugget	0.38	—	—
Spherical	0.32	3300	300
Spherical	0.30	3300	3300

1. For $\alpha \in \{1,\ldots n\}$, (*i*) perform simple kriging of the Gaussian random field Y at location $\boldsymbol{s}_\alpha$ by using the normal scores data known at the remaining locations $\{\boldsymbol{s}_\beta\colon \beta \neq \alpha\}$ and obtain the kriging prediction $Y^*(\boldsymbol{s}_\alpha)$ and the standard deviation $\sigma^*(\boldsymbol{s}_\alpha)$ of the kriging error and (*ii*) construct a set of symmetric intervals with probabilities ranging from 0.1 to 0.9. The interval with probability p is bounded by the $(1 - p)/2$ and $(1 + p)/2$ percentiles of the normal distribution with mean $Y^*(\boldsymbol{s}_\alpha)$ and standard deviation $\sigma^*(\boldsymbol{s}_\alpha)$. (This is a consequence of the hypothesis that Y is a stationary Gaussian random field.)
2. For each p between 0.1 and 0.9, calculate the proportion, p^*, of normal scores data that belong to the corresponding p intervals. This proportion is expected to match the underlying probability p, up to reasonable statistical fluctuations.

To get an idea of the acceptable deviation between p^* and p, let us assume (for the sake of simplicity) that there is no spatial dependence. In such a case, np^* is a binomial random variable with size n and parameter p. Because n is large (n = 738), this can be identified with a normal distribution, with mean np and variance $np(1 - p)$. Hence, the deviation $p^* - p$ is expected to be (in absolute value) less than $2[p(1 - p) / n]^{0.5}$ with a 95% probability.

In all cases, the deviations between p and p^* (Table 2) are low, which corroborates the goodness of the fitted model.

Sampling design problem

Statement of the problem

Given an initial sampling design, which in the following will be taken as a subset of the available 738 samples, it is of interest to define a design for additional samples so that the attribute can be predicted locally with an error (in absolute value) less than a prespecified threshold ε. To optimize the sampling cost, the proposed design should contain as few samples as possible.

Two questions will be addressed: (*i*) How should the initial design be defined (should it be on a regular mesh or not)? and (*ii*) How many additional samples are necessary, and where should they be placed?

Choice of prediction support and possible locations for extra samples

Before answering the previous questions, it is of utmost importance to define the area or support of the block (forest stand or a portion of it) on which the prediction errors will be evaluated. Indeed, because of the so-called support effect, the larger that the block support is, the less the error (Chilès and Delfiner 1999). Such an effect has long been recognized in the geostatistical evaluation of natural resources and of polluted sites (Journel and Huijbregts 1978; Isaaks and Srivastava 1989).

Fig. 4. Two realizations of the attribute (number of surviving trees per hectare) conditioned to the available data (Fig. 2A).

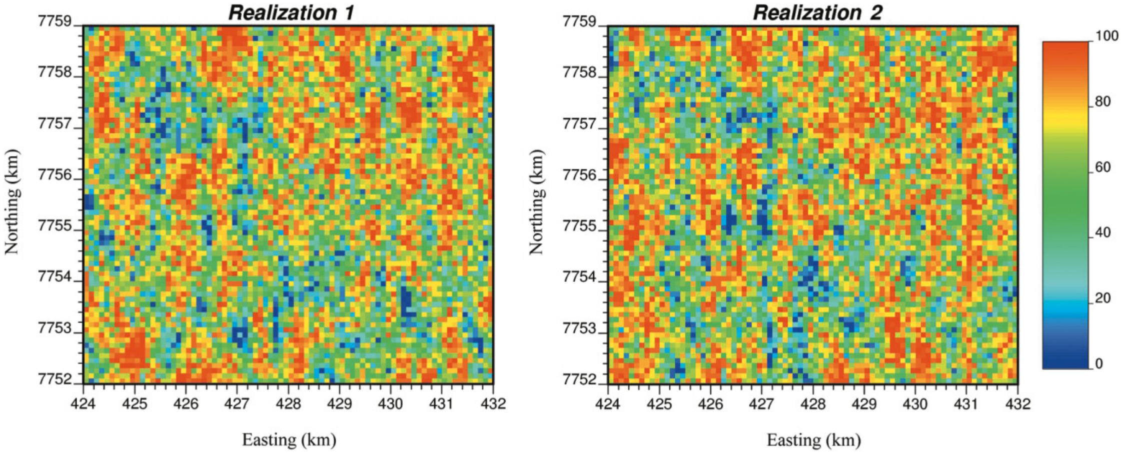


Table 2. Validation of the local uncertainty model (symmetric probability intervals).

Probability (p)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Data proportion (p*)	0.108	0.202	0.301	0.390	0.504	0.617	0.725	0.825	0.911
Allowable deviation between p and p*	0.022	0.029	0.034	0.036	0.037	0.036	0.034	0.029	0.022

To be as conservative as possible, we will choose a support of 600 m × 1200 m, considered as the minimal stand support targeted for prediction; the larger block dimension along the north direction takes account of the greater continuity in this direction. One obtains a total of 40 blocks in the domain of interest (Fig. 5). Besides, we consider a grid with mesh 100 m × 100 m that discretizes each block into 6 × 12 points and corresponds to the set of possible locations for the additional samples.

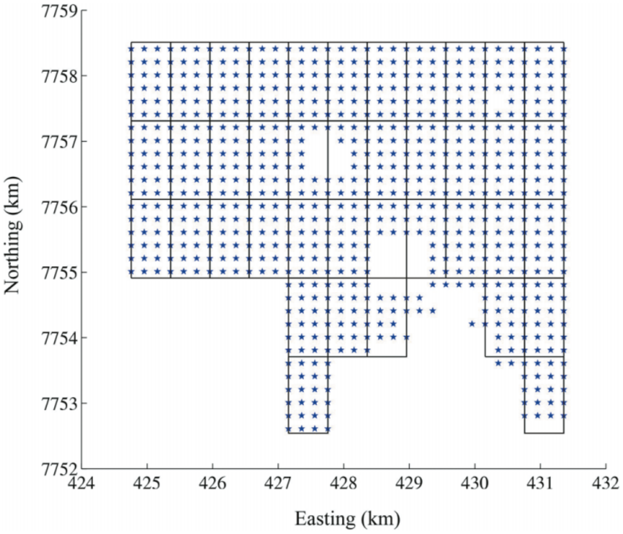
Choice of a measure for the prediction error

A well-known measure of local prediction errors is the kriging standard deviation, which accounts for the spatial configuration of data locations and for the spatial correlation (semivariogram) of the data. This measure has been widely used in geostatistical applications to define where to place additional samples (Delhomme 1978; McBratney and Webster 1981; McBratney et al. 1981; Olea 1984; Barnes 1989; Gao et al. 1996; Brus and Heuvelink 2007; Lin et al. 2008). However, because it does not depend on the data values, the kriging standard deviation does not account for the local variability of the attribute (for instance, a higher variability in subareas with lower local means) and, therefore, does not fully reflect the uncertainty in its values.

A more informative measure of local uncertainty is obtained by considering the width of an interval in which the attribute has a given probability p to lie (Goovaerts 2001). Such an interval is estimated by generating a large set of conditional realizations: the interval bounds are then defined by the $(1 - p)/2$ and $(1 + p)/2$ percentiles of the distribution of simulated values. If one chooses the interval midpoint as the prediction, then the absolute value of the prediction error will be less than ε (with probability p) if the interval width is less than 2ε. In the following, we shall take $p = 0.95$.

Sampling design algorithm: simulated annealing

Fig. 5. Definition of blocks targeted for local prediction (initial samples are marked with a star).



Let S_0 denote the set of sample locations of the initial design, S_1 the set of locations for the additional samples, and S_2 the remaining locations on the grid of possible samples. Because of the combination of all the potential sampling configurations (choices of S_1 and S_2), an exhaustive search of the optimal design is precluded, so that one has to look for a suboptimal approach to solve the sampling design problem. Possible approaches include forward-selection algorithms, in which sampling locations are added sequentially until the required constraints are fulfilled (Lu et al. 2000), sequential exchange algorithms (Aspie and Barnes 1990), or Bayesian search theory (Freeze et al. 1992; James and Gorelick 1994).

Another approach to determine S_1 and S_2 is the recourse to a simulated annealing (SA) algorithm. SA is an iterative device introduced by Kirkpatrick et al. (1983) to solve combinatorial problems and to find an approximate solution to the optimum of a given objective function. At each step, a candidate solution is generated by randomly perturbing the solution obtained at the previous step and is accepted or rejected on the basis of the Metropolis criterion. Several applications of SA have been proposed to solve sampling design problems with objective functions that account for a variety of criteria (Christakos and Killam 1993; Van Groenigen et al. 1999, 2000; Simbahan and Dobermann 2006; Brus and Heuvelink 2007).

In the following, an SA algorithm will be proposed to minimize an objective function, which is defined as

$$[3] \qquad O = \begin{cases} \infty \text{ if } w_{\max} \geq 2\varepsilon \\ \text{cardinality}(\boldsymbol{S}_1) + \dfrac{w_{\text{mean}}}{2\varepsilon} \text{ otherwise} \end{cases}$$

where $w_{\max}$ and w_{mean} are the maximal and mean widths of the 95% probability intervals over the 40 blocks of interest (Fig. 5). If $w_{\max}$ is less than 2ε, so is w_{mean}; hence, the number of additional samples (cardinality of $\boldsymbol{S}_1$) is the integer part of the objective function, which implicitly incorporates the sampling costs. The optimal design will be the design that contains the minimal number of samples and, at the same time, minimizes the mean prediction error (measured by the fraction $w_{\text{mean}}/2\varepsilon$).

The proposed algorithm consists of the following initialization (step 1) and iteration (steps 2–9) steps (Fig. 6).

1. Find a design such that the objective function (eq. 3) is finite. For instance, one can select one sample in $\boldsymbol{S}_2$ and incorporate it into $\boldsymbol{S}_1$ and repeat the procedure until $w_{\max} < 2\varepsilon$.
2. Draw a random number U uniformly distributed on (0,1).
3. If $3U < 1$, select one sample in $\boldsymbol{S}_2$ and incorporate this sample into $\boldsymbol{S}_1$. If $1 \leq 3U < 2$, select one sample in $\boldsymbol{S}_1$ and incorporate it into $\boldsymbol{S}_2$. Otherwise, exchange one sample from $\boldsymbol{S}_1$ for another sample from $\boldsymbol{S}_2$.
4. Simulate L realizations of the attribute at the locations in $\boldsymbol{S}_1$ conditionally to the data at the locations in $\boldsymbol{S}_0$. The simulated values will supply the lack of data at the locations in $\boldsymbol{S}_1$, to compute the subsequent conditional distributions and 95% probability intervals.
5. For each realization (i = 1, 2, …, L): (*i*) simulate M realizations of the attribute over the grid discretizing the blocks of interest (Fig. 5), conditionally to the data at locations in $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$ and regularize the realizations to the block support by averaging the values simulated within each block; (*ii*) for each block, use the M realizations to compute an interval in which the attribute has a 95% probability to lie; and (*iii*) calculate the maximal width ($w^i_{\max}$) and the mean width (w^i_{mean}) over all the blocks.
6. Calculate the overall interval widths over the L realizations:

$$[4a] \qquad w_{\max} = \max_{i \in \{1 \ldots L\}} (w^i_{\max})$$

and

$$[4b] \qquad w_{\text{mean}} = \underset{i \in \{1 \ldots L\}}{\text{mean}} (w^i_{\text{mean}})$$

7. If $w_{\max} \geq 2\varepsilon$, reject the design proposed in step 3 and revert to the former design.
8. Otherwise, calculate the objective function (eq. 3) for the former design (O) and for the new design (O'). Draw a random number V uniformly distributed on (0,1) and accept the new design if the following inequality is fulfilled:

$$[5] \qquad V < \exp\left(\frac{O - O'}{t}\right)$$

where t is a positive real number called temperature. This way, a design that decreases the objective function is always accepted, which corresponds to a design with fewer samples or a design with the same number of samples but a smaller value for w_{mean}, indicating that the mean prediction error decreases. In contrast, the designs that increase the objective function are accepted with a probability that depends on t. When the simulation progresses, this parameter decreases according to a logarithmic cooling schedule (Hajek 1988).
9. Go back to step 2 until a large number of iterations are done.

Besides the objective function under consideration (eq. 3) and the fact that the number of additional samples is unknown a priori, a major difference with existing approaches for sampling design is the coupling between simulated annealing (to search for additional sample locations) and geostatistical simulation (to model the uncertainty in the attribute at these locations). The authors are not aware of other proposals with such a coupling, probably because of the computational difficulties it implies (Watson and Barnes 1995; see also Implementation below). Most contributions using geostatistical simulation are limited to the cases when the locations of additional samples are known a priori (a circumstance that occurs in mineral resources evaluation with blast hole samples, to assess future ore and waste misclassifications, see Journel and Kyriakidis 2004), or when one looks for a prespecified type of design, e.g., a regular or random stratified design (Englund and Heravi 1993).

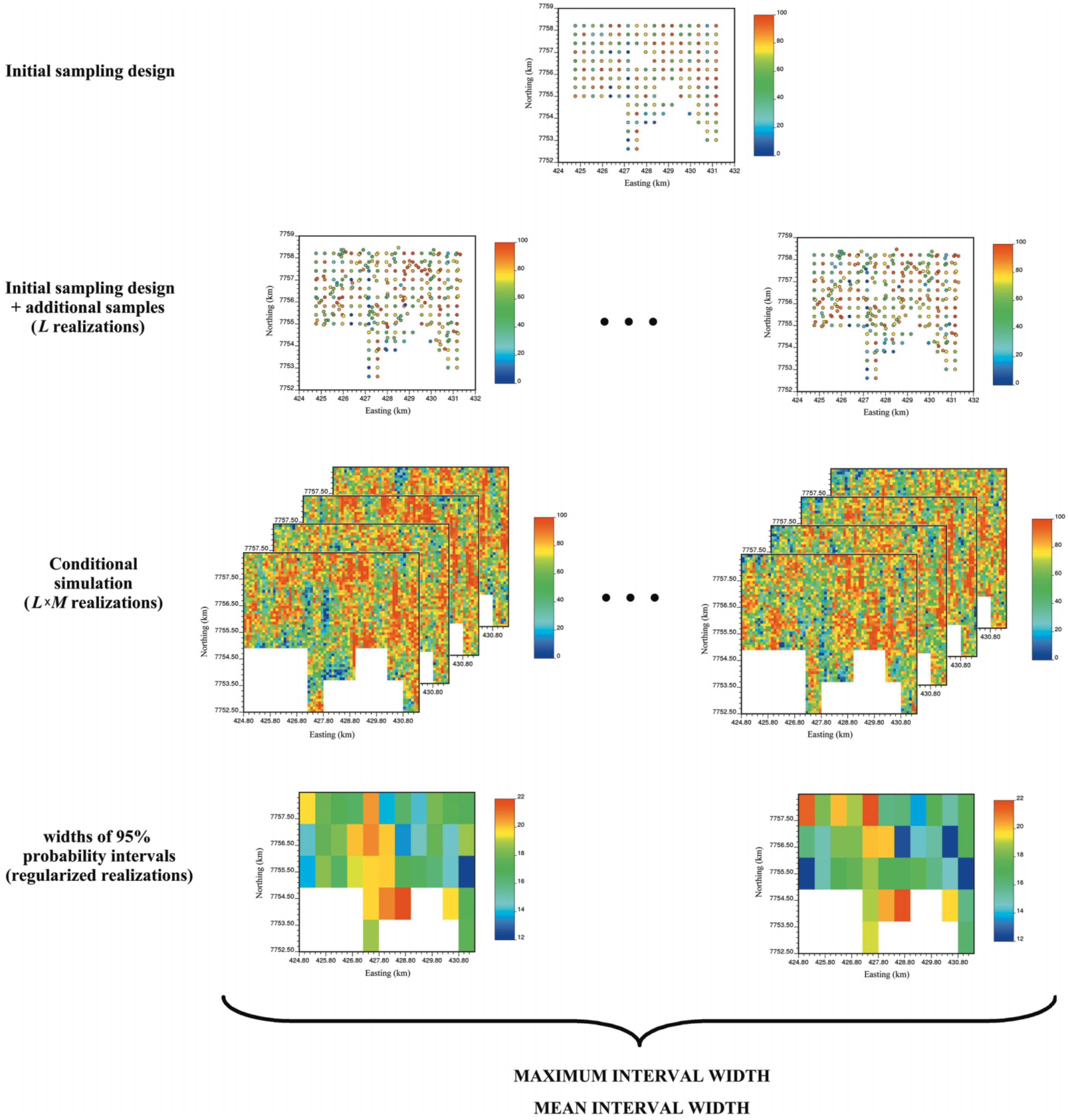
Here, our ambition is to deal with two sources of combinatorial problems at the same time: (*i*) the search of additional sample locations and (*ii*) the uncertainty in the values of the attribute at these locations. The recourse to simulation is necessary for our purposes, insofar as the probability intervals calculated at step 5 explicitly depend on the values at the locations in $\boldsymbol{S}_1$. Because these values are unknown, the intervals cannot be calculated by nonlinear kriging methods such as indicator, disjunctive, or lognormal kriging. Moreover, we need to determine the intervals jointly at all the locations in $\boldsymbol{S}_1$, not separately, which again argues in favor of conditional simulation instead of kriging techniques (Goovaerts 2001).

Implementation

Each iteration of the proposed SA algorithm calls for $L \times M$ conditional realizations of the attribute over the grid discretizing the block shown in Fig. 5 (2880 nodes), which is likely to be prohibitive in terms of CPU time. To speed up the algorithm, the following stratagems are considered.

1. The L realizations needed in step 4 are calculated once,

Fig. 6. Workflow of proposed algorithm.



by simulating the attribute over the grid of possible sample locations ($S_1 \cup S_2$) conditionally to the original data located at S_0.

2. The M realizations needed in step 5(i) are obtained by generating M nonconditional realizations and adding the simple kriging of the residuals between the conditioning data at $S_0 \cup S_1$ and the values of the nonconditional realizations at the same locations (eq. 2). Accordingly, for each iteration and each of the L realizations generated at step 4, the same set of M nonconditional realizations are used, i.e., only $L + M$ realizations are required in total.
3. Because the simple kriging weights only depend on the semivariogram model and on the spatial configuration of the data locations, a single kriging system has to be solved to condition the $L \times M$ realizations at step 5.
4. Between one iteration and the next one, the data config-

uration is almost the same (only one sample is added, removed, or moved). To avoid solving a full kriging system, the simple kriging weights at a given iteration are determined by updating the weights calculated at the former iteration (see Appendix A).

Applications

In the following, we will consider three initial sampling designs (Fig. 7):

1. A regular design with a mesh of 400 m × 400 m, which is obtained by retaining one-quarter of the available 738 samples (every other one along each coordinate axis).
2. A stratified random design in which the available samples are grouped into subsets of 2 × 2 contiguous samples, then one sample is selected at random in each subset.
3. A uniform random design in which each available sample is selected with probability 0.25, independently of the other samples.

Our idea is to apply the proposed SA algorithm to complete these initial sampling design. This will give us an insight into whether or not a systematic initial design is preferable to a random design, and how many additional samples are needed to fulfill the desired accuracy criterion (local prediction errors $<\varepsilon$).

For the particular application, the following parameters are used: (*i*) $L = 200$ realizations, (*ii*) $M = 20$ realizations, (*iii*) $\varepsilon = 10\%$ (maximum allowable error for local prediction), (*iv*) number of iterations = 100 000, (*v*) t_0 (initial temperature) = 2.90, and (*vi*) t_f (final temperature) = 0.17. The numbers of realizations L and M have been chosen so as to accurately determine the probability intervals and to get a reasonable number of possible scenarios for assessing uncertainty. The initial temperature has been taken so that the probability of accepting a design with one more sample is equal to 0.7 at the beginning of the iterative process (Brus and Heuvelink 2007); this probability decreases to 0.003 at the end of the process.

The results (Fig. 8, Table 3) call for the following comments:

1. The final number of additional samples is less when the initial design is random than when it is regular. (This is not coincidental; similar results have been obtained by repeating the exercise with other initial designs.)
2. As shown by the evolution of the objective function with the number of iterations, the proposed algorithm converges faster in the case of the stratified random initial design. However, after 100 000 iterations, the convergence is reached in all the cases, and the mean width of the 95% probability intervals is very close to the maximum allowable width (20) (Table 3).
3. The domain of interest is not evenly covered by the final design. This is due to the fact that the uncertainty in the values at nonsampled locations (measured by the widths of the 95% probability intervals) depends not only on the number of neighboring data and on their spatial configuration, but also on the data values (Isaaks and Srivastava 1989). In particular, the areas where the measured values of the attribute are similar require fewer ad-

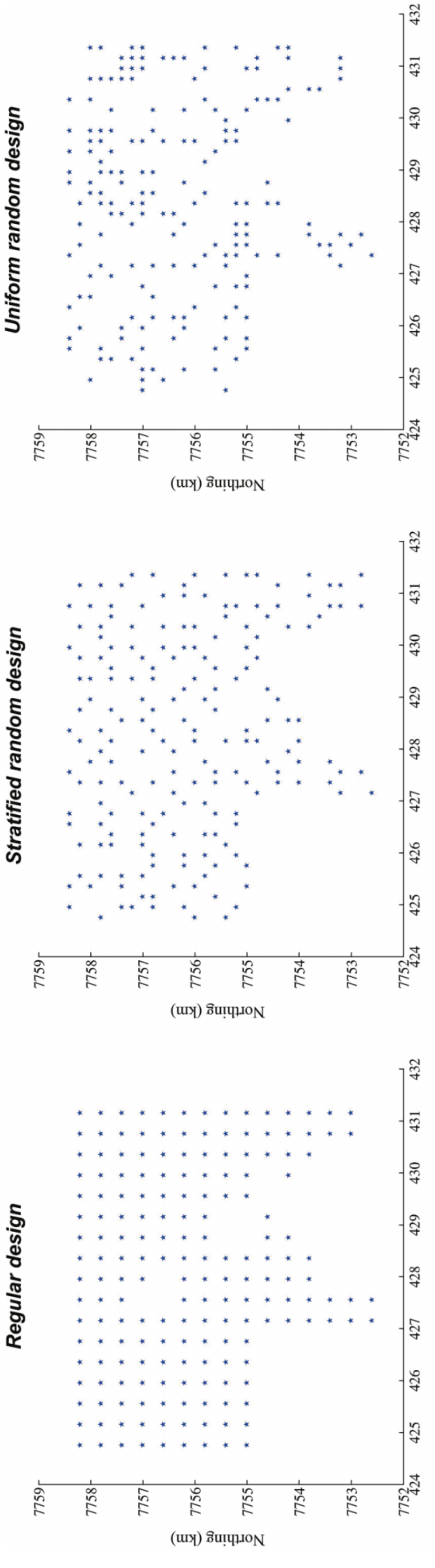
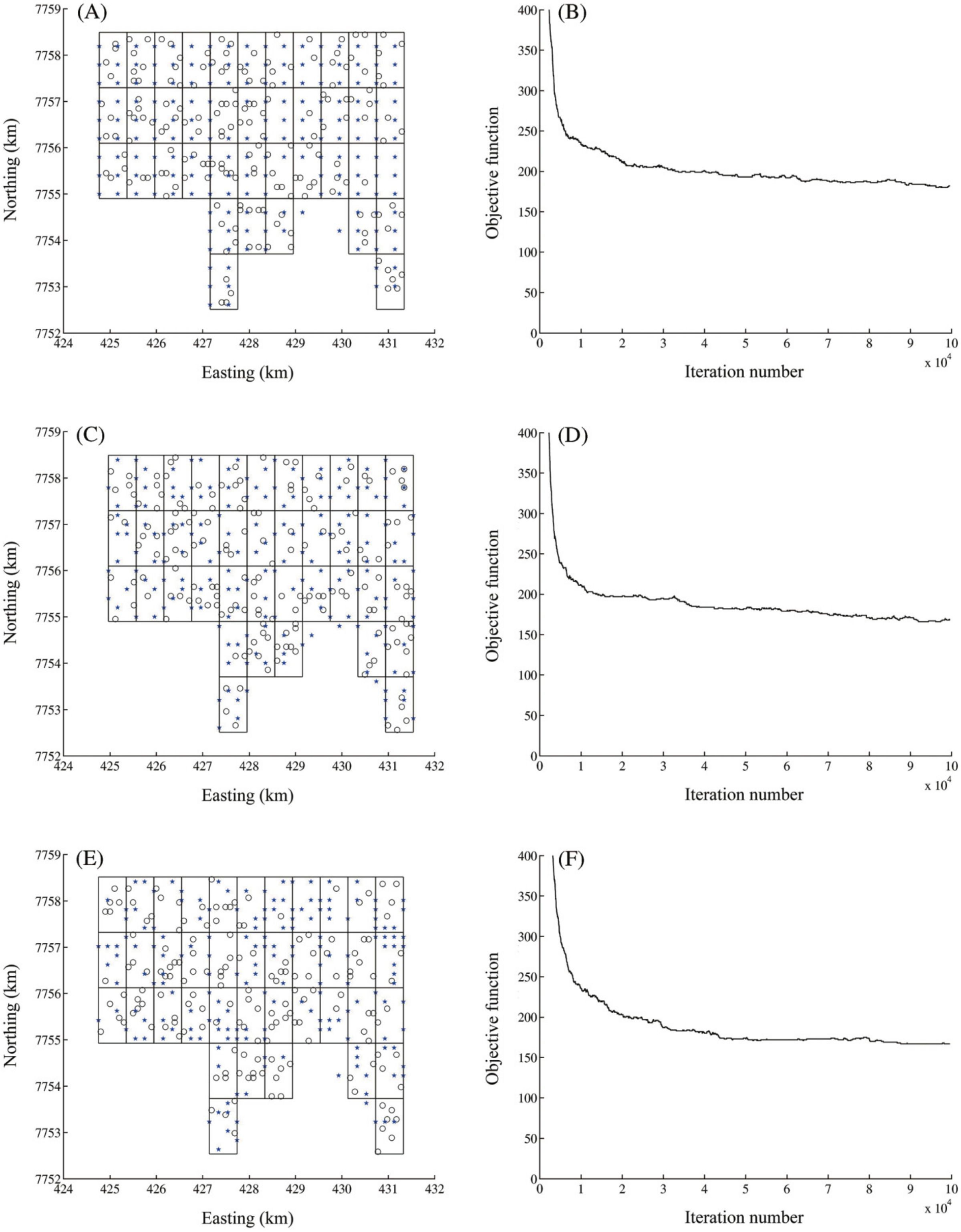


Fig. 7. Three initial sampling designs (regular, stratified random, and uniform random). Sample locations are indicated with stars.

Fig. 8. Sample locations (left) and objective functions (right) for the three initial sample designs: regular (A and B), stratified random (C and D), and uniform random (E and F). Stars are the initial samples, and circles are additional samples.



ditional samples than the areas where the measured values are highly dispersed.

4. Additional samples are sometimes located close together or close to samples of the initial design, which can be explained because of the nugget effect present in the fitted semivariogram model (Fig. 3, Table 1). The nugget effect indicates a lack of spatial correlation, so that neighboring samples do not necessarily convey redundant information. For this reason also, there may be many solutions to the sample design problem.

Table 3. Statistics on cost-effective sample designs.

Initial design	No. of initial samples	No. of additional samples	Total no. of samples	Mean width of 95% probability intervals
Regular	181	181	362	19.90
Stratified random	187	167	354	19.83
Uniform random	181	166	347	19.90

5. If the number of additional samples found were too large, this would indicate that the accuracy criterion is too demanding. In such a case, an option would be to increase the size of the blocks targeted for prediction (recall the discussion on the support effect in Choice of prediction support and possible locations for extra samples).

Although the uniform random design yields the best results in terms of additional samples, the stratified random design offers a more homogeneous distribution of the initial samples over the entire area (Fig. 7), which facilitates the inference of the model parameters (Gaussian anamorphosis and semivariogram of transformed data). Besides, the numbers of additional sampling units for both strategies are similar (167 vs. 166); thus, in practice, the stratified random design may be preferable as an initial design.

Conclusions

The objective of this work was to define a cost-effective infill sampling design by using geostatistics. The proposed algorithm rests upon two important methods, conditional simulation to quantify spatial uncertainty and simulated annealing to minimize an objective function, and is able to ensure local accuracy while keeping sampling costs as low as possible. Compared with the original systematic sampling consisting of 738 samples, our method reduces the sampling costs by >50%. The algorithm is also versatile insofar as any criterion for assessing the quality of a sampling design can be incorporated into the objective function, e.g., criteria that depend on economic parameters, such as sampling costs, remediation costs or misclassification costs (Aspie and Barnes 1990; Englund and Heravi 1993; Christakos and Killam 1993; James and Gorelick 1994); on available secondary information (Van Groenigen et al. 2000); or on extreme value occurrences (Watson and Barnes 1995).

The sampling design problem arises with attributes that describe natural resources, as long as an underlying spatial dependence law exists. In forestry, because of the relationship between vegetation productivity and site condition, such a spatial dependence is always present. Therefore, the proposed methodology can introduce a new paradigm in forest and natural resource sampling and has a great potential for the spatial prediction of descriptive variables of both natural and introduced flora and fauna.

Future studies should address the incorporation of covariates, such as environmental attributes, to improve local predictions and define alternative strategies for initial sampling design to better assess the semivariogram at short distances.

Acknowledgements

The authors are grateful to two anonymous reviewers for their constructive comments that helped to improve the manuscript.

References

Aspie, D., and Barnes, R.J. 1990. Infill-sampling design and the cost of classification errors. Math. Geol. **22**(8): 915–932. doi:10.1007/BF00890117.

Barnes, R.J. 1989. Sample design for geologic site characterization. *In* Geostatistics. *Edited by* M. Armstrong. Kluwer Academic Publishers, Dordrecht, the Netherlands. pp. 809–822.

Brus, D.J., and Heuvelink, G.B.M. 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma, **138**(1–2): 86–95. doi:10.1016/j.geoderma.2006.10.016.

Burkhart, H.E., Barrett, J.P., and Lund, H.G. 1984. Timber inventory. *In* Forestry handbook. *Edited by* K.F. Wenger. John Wiley & Sons, New York. pp. 361–412.

Chilès, J.P., and Delfiner, P. 1999. Geostatistics: modeling spatial uncertainty. Wiley, New York.

Christakos, G., and Killam, B.R. 1993. Sampling design for classifying contaminant level using annealing search algorithms. Water Resour. Res. **29**(12): 4063–4076. doi:10.1029/93WR02301.

Cochran, W.G. 1977. Sampling techniques. Wiley, New York.

De Gruijter, J., Brus, D., Bierkens, M., and Knotters, M. 2006. Sampling for natural resource monitoring. Springer, Berlin.

De Vries, P.G. 1986. Sampling theory for forest inventory. Springer-Verlag, Berlin.

Delhomme, J.P. 1978. Kriging in the hydrosciences. Adv. Water Resour. **1**(5): 251–266. doi:10.1016/0309-1708(78)90039-8.

Dubrule, O. 1983. Cross validation of kriging in a unique neighborhood. Math. Geol. **15**(6): 687–699. doi:10.1007/BF01033232.

Emery, X. 2007. Conditioning simulations of Gaussian random fields by ordinary kriging. Math. Geol. **39**(6): 607–623. doi:10.1007/s11004-007-9112-x.

Emery, X. 2009. The kriging update equations and their application to the selection of neighboring data. Computat. Geosci., In press. doi:10.1007/s10596-008-9116-8.

Englund, E.J., and Heravi, N. 1993. Conditional simulation: practical application for sampling design optimization. *In* Geostatistics Tróia'92. *Edited by* A. Soares. Kluwer Academic Publishers, Dordrecht, the Netherlands. pp. 613–624.

Freeze, R.A., James, B., Massmann, J., Sperling, T., and Smith, L. 1992. Hydrogeological decision analysis: 4. The concept of data worth and its use in the development of site investigation strategies. Ground Water, **30**(4): 574–588. doi:10.1111/j.1745-6584.1992.tb01534.x.

Gao, H., Wang, J., and Zhao, P. 1996. The updated kriging variance and optimal sample design. Math. Geol. **28**(3): 295–313. doi:10.1007/BF02083202.

Gilabert, H. 2007. Optimizing yield data collection efforts for forest management planning. Ph.D. dissertation, Pennsylvania State University, University Park, Pa.

Goovaerts, P. 2001. Geostatistical modelling of uncertainty in soil science. Geoderma, **103**(1–2): 3–26. doi:10.1016/S0016-7061(01)00067-2.

Hajek, B. 1988. Cooling schedules for optimal annealing. Math. Oper. Res. **13**(2): 311–329. doi:10.1287/moor.13.2.311.

Hock, B., Payn, T., and Shirley, J. 1993. Using a geographic information system and geostatistics to estimate site index of *Pinus radiata* for Kaingaroa forest, New Zealand. N.Z. J. For. Sci. **23**(3): 264–277.

Husch, B., Miller, C., and Beers, T. 1993. Forest mensuration. 3rd ed. Krieger Publishing Co., Malabar, Fla.

Isaaks, E.H., and Srivastava, R.M. 1989. An introduction to applied geostatistics. Oxford University Press, New York.

James, B.R., and Gorelick, S.M. 1994. When enough is enough: the worth of monitoring data in aquifer remediation design. Water Resour. Res. **30**(12): 3499–3513. doi:10.1029/94WR01972.

Journel, A.G. 1974. Geostatistics for conditional simulation of orebodies. Econ. Geol. **69**(5): 673–687. doi:10.2113/gsecongeo.69.5.673.

Journel, A.G., and Huijbregts, C.J. 1978. Mining geostatistics. Academic Press, London.

Journel, A.G., and Kyriakidis, P.C. 2004. Evaluation of mineral reserves: a simulation approach. Oxford University Press, New York.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. 1983. Optimization by simulated annealing. Science (Washington, D.C.), **220**(4598): 671–680. doi:10.1126/science.220.4598.671. PMID:17813860.

Lantuéjoul, C. 1994. Non conditional simulation of stationary isotropic multigaussian random functions. *In* Geostatistical simulations. *Edited by* M. Armstrong and P.A. Dowd. Kluwer Academic Publishers, Dordrecht, the Netherlands. pp. 147–177.

Lin, Y., Yeh, M., Deng, D., and Wang, Y. 2008. Geostatistical approaches and optimal additional sampling schemes for spatial patterns and future sampling of bird diversity. Glob. Ecol. Biogeogr. **17**(2): 175–188. doi:10.1111/j.1466-8238.2007.00352.x.

Loetsch, F., Zöhrer, F., and Haller, K.E. 1973. Forest inventory. Vol. 2. BLV Verlagsgesellschaft mbH, München, Germany.

Lu, Z., Berliner, L.M., and Snyder, C. 2000. Experimental design for spatial and adaptive observations. *In* Studies in the atmospheric sciences. *Edited by* L.M. Berliner, D. Nychka, and T. Hoar. Springer-Verlag, Berlin. pp. 65–78.

Mandallaz, D. 2000. Estimation of the spatial covariance in universal kriging: application to forest inventory. Environ. Ecol. Stat. **7**(3): 263–284. doi:10.1023/A:1009619117138.

Mandallaz, D. 2007. Sampling techniques for forest inventories. Chapman & Hall/CRC Press, Boca Raton, Fla.

Marbeau, J.P. 1976. Géostatistique forestière. Ph.D. dissertation, University of Nancy, Nancy, France.

Matérn, B. 1960. Spatial variation. Springer-Verlag, Berlin.

McBratney, A.B., and Webster, R. 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables — II: program and examples. Comput. Geosci. **7**(4): 335–365. doi:10.1016/0098-3004(81)90078-9.

McBratney, A.B., Webster, R., and Burgess, T.M. 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables — I: theory and method. Comput. Geosci. **7**(4): 331–334. doi:10.1016/0098-3004(81)90077-7.

Nanos, N., Calama, R., Montero, G., and Gil, L. 2004. Geostatistical prediction of height/diameter models. For. Ecol. Manage. **195**(1–2): 221–235. doi:10.1016/j.foreco.2004.02.031.

Olea, R.A. 1984. Sampling design optimization for spatial functions. Math. Geol. **16**(4): 369–392. doi:10.1007/BF01029887.

Philip, M.S. 1994. Measuring trees and forests. 2nd ed. CAB International, Wallingford, UK.

Rivoirard, J. 1994. An introduction to disjunctive kriging and nonlinear geostatistics. Oxford University Press, New York.

Sales, M.H., Souza, C.M., Kyriakidis, P.C., Roberts, D.A., and Vidal, E. 2007. Improving spatial distribution estimation of forest biomass with geostatistics: a case study for Rondonia, Brazil. Ecol. Modell. **205**(1–2): 221–230. doi:10.1016/j.ecolmodel.2007.02.033.

Simbahan, G.C., and Dobermann, A. 2006. Sampling optimization based on secondary information and its utilization in soil carbon mapping. Geoderma, **133**(3–4): 345–362. doi:10.1016/j.geoderma.2005.07.020.

Van Groenigen, J.W., Siderius, W., and Stein, A. 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma, **87**(3–4): 239–259. doi:10.1016/S0016-7061(98)00056-1.

Van Groenigen, J.W., Pieters, G., and Stein, A. 2000. Optimizing spatial sampling for multivariate contamination in urban areas. Environmetrics, **11**(2): 227–244. doi:10.1002/(SICI)1099-095X(200003/04)11:2<227::AID-ENV404>3.0.CO;2-#.

Watson, A.G., and Barnes, R.J. 1995. Infill sampling criteria to locate extremes. Math. Geol. **27**(5): 589–608. doi:10.1007/BF02093902.

Appendix A

For $\alpha \in \{1, 2, \ldots, n\}$, let us denote by $\lambda_{\alpha|n}(\boldsymbol{s})$ the simple kriging weight assigned to the datum located at $\boldsymbol{s}_\alpha$ when predicting an attribute at location $\boldsymbol{s}$ by using the data at locations $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, …, $\boldsymbol{s}_n$. Consider the incorporation of a new datum at location $\boldsymbol{s}_{n+1}$. The weights obtained by using n and $n + 1$ data fulfill the following relationships (Emery 2009):

$$[A1] \qquad \forall\alpha \in \{1, 2, ..., n\}, \lambda_{\alpha|n}(\boldsymbol{s}) = \lambda_{\alpha|n+1}(\boldsymbol{s}) + \lambda_{n+1|n+1}(\boldsymbol{s})\lambda_{\alpha|n}(\boldsymbol{s}_{n+1})$$

Let $\mathbf{C}$ be the variance–covariance matrix of the $n + 1$ data, and B the inverse of $\mathbf{C}$. Dubrule (1983) has shown the following identities:

$$[A2] \qquad \forall\alpha \in \left\{1, 2, ..., n\right\}, \lambda_{\alpha|n}(\boldsymbol{s}_{n+1}) = -\frac{B_{\alpha,n+1}}{B_{n+1,n+1}}$$

Accordingly, the set of kriging weights $\{\lambda_{\alpha|n}(\boldsymbol{s}_{n+1}), \alpha = 1, 2, \ldots, n\}$ can be obtained as soon as the last row of $\mathbf{B}$ (i.e., $\boldsymbol{B}_{\bullet,n+1}$) is known, which is done by solving the following equation:

$$[A3] \qquad \mathbf{C}\boldsymbol{B}^t_{\bullet,n+1} = \boldsymbol{D}$$

where $\boldsymbol{D}$ is an $(n + 1) \times 1$ vector whose entries are equal to 0, except for the last one that is equal to 1. Equations A1–A3 allow one to quickly update the simple kriging weights when passing from $n + 1$ data to n data. Reciprocally, when passing from n to $n + 1$ data according to eq. A1, the calculation of $\lambda_{n+1|n+1}(\boldsymbol{s})$ is also required:

$$[A4] \qquad \lambda_{n+1|n+1}(\boldsymbol{s}) = \boldsymbol{B}_{\bullet,n+1}\boldsymbol{C}_0$$

where $\boldsymbol{C}_0$ is the $(n + 1) \times 1$ covariance vector between the attribute at location $\boldsymbol{s}$ and the $n + 1$ data at locations $\boldsymbol{s}_1$, $\boldsymbol{s}_2$, …, $\boldsymbol{s}_{n+1}$.