Ciencia e Investigación AGRARIA

Cien. Inv. Agr. 41(1):33-48. 2014
www.rcia.uc.cl

CROP PRODUCTION

DOI: 10.4067/S0718-16202014000100004

RESEARCH PAPER

Optimization of sample design sizes and shapes for regionalized variables using simulated annealing

Luciana P.C. Guedes[1], Miguel A. Uribe-Opazo[1], and Paulo J. Ribeiro Junior[2]

[1]Universidade Estadual do Oeste do Paraná (UNIOESTE), Cascavel/ Centro de Ciências Exatas e Tecnológicas (CCET)/ Programa de Pós-Graduação em Engenharia Agrícola (PGEAGRI), Universitária Street 2069 – 85819-110, Cascavel, Paraná, Brazil.

[2]Universidade Federal do Paraná (UFPR), Curitiba/Laboratório de Estatística e Geoinformação (LEG), Coronel Francisco Heráclito dos Santos Street 210 – 85531-990, Curitiba, Brazil.

Abstract

L.P.C. Guedes, M.A. Uribe-Opazo, and P.J. Ribeiro Junior. 2014. Optimization of sample design sizes and shapes for regionalized variables using simulated annealing. Cien. Inv. Agr. 41(1): 33-48. The spatial variability of structures in regionalized variables are defined with the aid of geostatistical techniques, which facilitate the estimation of values for these variables in unsampled localizations and generate thematic maps to be used in decision making for localized treatments in the area under study. The quality of these maps depends on the trustworthiness of these estimates that can be modified with the choice for the sample design. The objective of this work was to establish an optimal size and shape of the sample designs in order to enhance the efficiency of sampling plans for the prediction of space dependent variables. These designs were obtained with the use of a stochastic search method called Simulated Annealing. This method is based on a sampling grid with a large number of points. Here, it is initially used to consider simulated data sets with distinct spatial dependence structures and is then used to consider real data on soy productivity. The simulated results are used as reference for the achievement of the best sample design with the lowest number of sample points that can efficiently represent the spatial dependence structure of soy productivity in a commercial area harvested by the harvester monitor. The results reported for the simulations and soy productivity data show that the optimization process was efficient in determining sample designs with reduced size, especially when using the Global Accuracy as the measurement to be maximized.

Key words: Geostatistics, interpolation, precision agriculture, spatial variability.

Introduction

The need for optimal solutions to most agricultural sector problems is a priority because the current focus on concepts, such as performance, efficiency and costs, demand that the agricultural sector look for a higher sustainability in competitive markets. In this context, information gathered by studies on the spatial variability analysis of culture productivity with geostatistics techniques has increased the scope of this area of study, with the identification of high and low productivity zones, and the total produced and possible relations with

Received May 14, 2013. Accepted March 6, 2014.
Corresponding author: luciana_pagliosa@hotmail.com

spatial variability in soil properties (Reichert *et al.,* 2008; Johann, *et al.,* 2010).

However, the description of the spatial variability of any attribute that is georeferenced in thematic maps depends substantially on the accuracy of the geostatistics analysis results and the experimental viability in terms of the available resources for data collection and the measurement of variables. Therefore, it is necessary to devise a reduced sampling scheme that can minimize the operational costs needed while maximizing the quality of results obtained in spatial predictions for unsampled localizations.

Studies in the present literature have assessed the use of classical spatial sampling schemes, including random, centered systematic, triangular, hexagonal and stratified schemes, for experimental planning in order to find information that may support an efficient description of spatial variability in the georeferred data (McBratney *et al.,* 1981; Yfantis *et al.,* 1987, Oda-Souza *et al.,* 2010). Nevertheless, with the acceptance of variations in the parameters and characteristics that support the spatial process, such as the nugget, sill, range and anisotropy effects, the sampling schemes described above have generated very unreliable results (Yfantis *et al*., 1987; Dunn and Harisson, 1993; Diggle and Ribeiro Junior, 2007).

However, failing to use these classical sampling schemes in an experimental area where the sample population of points is infinite and uncountable makes the determination of a sample design based on this population of points an extremely complex task. A helpful methodology that can be used in this case involves the discretization of the area under study, which is represented by an initial sampling grid with a large number of points. This provides a limited area of sampling points. Thus, choosing the most efficient sampling design for spatial prediction can be defined as an optimization problem that consists of choosing, from an initial sampling grid, the best small-sized sampling design that can minimize both losses in the accuracy of results regarding the spatial prediction and costs in the field data collection and laboratory analyses.

Studies to find optimized sampling designs in variables displaying spatial dependence are very recent. This optimization methodology is used, for example, in problems of rationalization of environmental monitoring network stations (Ruiz-Cárdenas *et al*., 2010).

A few studies have investigated optimized sampling designs in relation to maximizing prediction efficiency for the analysis of spatial variability in assessing soil properties and crop yields, specifically (Van Groenigen, 2000; Ferreyra *et al.,* 2002). However, these studies have not compared thematic maps generated by the initial sampling grid or by the optimized sampling design using similarity measures (Guedes *et al*., 2011). They also did not use the smallest possible sampling size for this optimized design to minimize the loss of spatial prediction accuracy.

There are many search procedures to tackle the problem of finding the best sample design. For instance, sequential search methods (Bôer *et al.,* 2002), such as the global reduction enumeration method, can be used, which consist of analyzing all possible solution combinations (Le *et al.,* 2003). The methods previously mentioned are only efficient and computer-friendly when given a small number of possible groups.

However, metaheuristic strategies are used in artificial intelligence, such as the simulated annealing algorithm (Costa Filho *et al.,* 2010; Guedes *et al*., 2011) and the genetic algorithm (Chakrapani and Rajan, 2008; Ruiz-Cárdenas, 2010, Guedes *et al.,* 2011), which use an iterative search method to determine the optimal solution.

The present work has developed optimized sample designs with the smallest possible size relative to their efficiency in predicting spatial locations not

sampled by the simulated annealing search method. This methodology was applied to simulated data sets while considering two objective functions: the sum quadratic error for the spatial prediction, which should be minimized, and an accuracy measure called overall accuracy, which should be maximized. Nonetheless, simulated data sets, with different structures of spatial variability, have been used and evaluated to discover how much this structure may influence the determination of optimized sampling designs.

These simulations have oriented the execution of the main objective of this work: to optimize the size and sampling designs for analyzing soy productivity georeferenced data sets inside a commercial area, collected with a harvester machine that represents the discretization of the area studied. The sampling design will be optimized by the simulated annealing method. Based on this optimized sampling design, it will be possible to use geostatistic techniques to estimate, for the reference year and oncoming crop years, the amount to be harvested from each of the regions of the study area and the total production, thus minimizing the costs of the experiment while preserving its spatial prediction trustworthiness and quality.

Materials and methods

First, 10 simulated data sets were generated and distributed over a regular initial 20 × 20 sampling grid with 400 sampling points and a maximum coordinates limit of 100 m, which represented a discretization in the area of study. For each simulated data set, the values used as the regionalized variables in these localizations were simulated by a Monte Carlo experiment, which represented realizations of multivariate stochastic processes and displayed stationary Gaussian variables, with no directional trend, isotropic or exponential model, as described in Equation 1, with a nugget effect (C_0), sill ($C_0 + C_1$) and range (a) of 2, 10 and 60, respectively.

$$\begin{cases} \gamma(h) = C_0 + C_1[1 - e^{\left(\frac{-3h}{a}\right)}], & \text{se } 0 < h < a \\ \gamma(h) = C_0 + C_1, & \text{se } h > a \end{cases} \quad (1)$$

The soy productivity data (t ha-1) for this study was collected during the 2004/2005 crop year, in a commercial area of 57.16 ha located in the city of Cascavel, West of Paraná, Brazil. The geographic coordinates for the area are approximately 24.95º South and 53.57º West, datum SAD-69, at an average height of 650 m above the sea level. In the rhodic hapludox area, which had a clay-like texture, the topographical scope was performed by GPS receivers with fixed positioning and preprocessed differential corrections.

The data set for the productivity corresponds to 5000 sampling points registered in the productivity reading of the harvest monitor. The reading was performed by sensors installed in the harvester machine, which measure the instantaneous grain yield during harvesting using an impact plate in the flowing grain. This data set, which has a large number of sampling points, represents a partitioning of the productivity distribution in the area under study.

The selection of the best set of sampling points and the smallest sampling size that should fulfill a minimally efficient criterion for the spatial prediction consisted of a reduction in the sampling size of the initial grid of 400 points in the simulated data sets and of 5000 points in the crop production data. This was accomplished with an optimization process called Simulated Annealing, consisting of an iterative method to search for optimized solutions to complex problems with a large space for possible solutions and no need for any information about derivatives, nonlinearity and discontinuity.

The method consists of a set of systematized random actions that simulate the natural thermodynamic phenomenon of metallurgy known as annealing. In this process, metal or glass is heated to high temperatures, allowing the atoms to move freely

and rapidly in a disorderly state. The material is then slowly and gradually cooled down until it solidifies and becomes stronger (Benvenga, 2011).

The algorithm implementation, which defined the sampling size and optimal sampling design, consisted of the following stages:

Stage 0. From $i = 0$ it was predetermined that the initial sampling size d_0 should be 15% of the number of points in the initial grid. Some measures for the simulated annealing algorithm were also previously defined based on initial tests: the stopping criterion for the algorithm was equal to 1200 interactions (defined by observing the algorithm stationary trend), the value of the initial temperature was set to 1 and the geometric cooling schedule was equal to $t_{i+1}=0.9 \times t_i$. The last two results guaranteed that the process escaped optimal places and searched for more promising regions in the solution range.

Stage 1. In a simple and random way, a sampling design S_i S_i of the reduced size d_0 was selected from the initial grid.

Stage 2. For this sampling design, an exponential model was adjusted using the maximum likelihood method, and the spatial prediction of the variable values in localizations at the original grid were performed using a geostatistical interpolation method called kriging. An objective function for S_i was then calculated. Two optimal designs were generated independently, with distinct objective functions. In the first process (TS1), the objective was to maximize an accuracy measure called the Overall Accuracy (*OA*), which is used to measure the similarity between two maps and is present in Equation 2 (Anderson *et al*., 2001; De Bastiani *et al*., 2012).

$$OA = \sum_{j=1}^{r} x_{jj} \Big/ N, \qquad (2)$$

where, N is the value for the total area and x_{jj} are the elements in the diagonal error matrix. Each x_{jk} element in this matrix represents the total area of each part that belongs to the class j (j *=1,..., r*) of values in the model map (a map that expresses the spatial variability of values predicted in the initial grid localizations obtained through the sampling points resulting from the optimization process) and to the class k *(k =1,..., r)* of values in the reference map (a map that expresses the spatial variability of values displayed in the initial grid).

Therefore, the main diagonal line represents the amount of area with the same classification in the two maps, while values outside the main diagonal line represent the amount of area with classifications without any match. Ten class intervals (*r=10*) with the same amplitude were used in this work, guaranteeing that if in a certain area of the two maps there is the same classification, the values predicted will be very similar.

In the second optimization process (TS2), the objective function to be minimized was the sum quadratic error of the spatial prediction.

Stage 3. A new sampling design S_{i+1} was registered when a point in the former design S_{i+1} was chosen randomly and replaced by another point chosen randomly in the surroundings.

Stage 4. The objective function was calculated for the sampling design S_{i+1} in the same was as described for stage 2.

Stage 5. The objective function variation between two sampling schemes was calculated by the expression $\Delta_i = f(S_{i+1}) - f(S_i)$. If the objective is to minimize the sum quadratic error of the spatial prediction, the new solution S_{i+1} will be accepted with the probability described in Equation 3.

$$P[\text{aceitar } S_{i+1}] = \begin{cases} 1, & \text{se } \Delta_i \le 0 \\ \exp\left(\dfrac{-\Delta_i}{t_i}\right), & \text{se } \Delta_i > 0 \end{cases}. \qquad (3)$$

However, if the objective is to maximize overall accuracy, the new solution S_{i+1} will be accepted with the probability described in Equation 4.

$$P[\text{aceitar } S_{i+1}] = \begin{cases} 1, & \text{se } \Delta_i \geq 0 \\ \exp\left(\frac{\Delta_i}{t_i}\right), & \text{se } \Delta_i < 0 \end{cases} \quad (4)$$

Stage 6. For the current sampling size d_m, the optimization process will be finished if the stopping criterion has been achieved. Otherwise, the cooling schedule described in stage 0 should be applied to the current temperature value followed by $i=i+1$, and a return to stage 3.

Stage 7. If the overall accuracy value for TS1, with size d_m, is equal to 0.85, which corresponds to a high level of similarity between the two maps (Anderson *et al*., 2001; De Bastiani *et al*., 2012) the optimization process is finished. Otherwise, we have to go back to stage 1 with $d_{m+1} = d_m + 20$ (for the simulations) and $d_{m+1} = d_m + 200$ (for the productivity data). For the optimization process TS2, the process was finished when the sum quadratic error of the spatial prediction corresponded to 15% of the same calculation for the smallest sampling design.

In the simulations, the size-reduced optimized sampling grids were compared between themselves and two simple random sampling schemes generated in a sampling size that satisfied the criteria described in stage 7 of the optimization process.

The measures associated with the spatial prediction of the initial grid, with its 400 sampling points, were as follows: estimates of the parameters in the exponential model adjusted to the semivariance function, mean variance of the spatial prediction, percentage and total values predicted for the variable, which are above the third quartile, and the sum quadratic error of the percentage and total of values predicted, which are above the third quartile.

The following similarity measures were used for comparison: an overall accuracy, described in Equation 2, with a minimum accuracy level of 0.85, a concordance index Tau (T), expressed in the Equation 5- (a), and a Kappa index (K), expressed in Equation 5-(b), both with a low accuracy if T(or K) < 0.67, an average accuracy if 0.67≤T(or K) < 0.80 and a high accuracy if T(or$K$) T(or K)≥ 0.80 (Ma and Redmond, 1995; Anderson *et al*., 2001; De Bastiani *et al*., 2012).

$$T = \frac{\frac{\sum_{i=1}^{r} x_{ii}}{N} - \frac{1}{r}}{1 - \frac{1}{r}} \quad (5\text{-a})$$

$$K = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}\left(x_{i+} \times x_{+i}\right)}{N^2 - \sum_{i=1}^{r}\left(x_{i+} \times x_{+i}\right)} \quad (5\text{-b})$$

In addition to the spatial dependence model already cited, the best methodology to determine sampling designs with reduced sizes was carried out for different sizes of the initial grid (with 625, 900 and 1600 sampling points) and different models of spatial dependence in the sets of simulated data. The vector of the parameters (a, C_0, C_0+C_1) that define the structure of spatial dependence for the spatial models used for the construction of simulated data sets are: (75,2,10), (90,2,10) and (60,5,10). For each size of a regular initial sampling grid and for each model, ten data sets were simulated and the best optimization methodology was applied. The main objective for variations in the sampling sizes of the initial grid, and in range and nugget effect values, was to study the efficiency of the optimization methodology applied in a more refined partitioning of the area under study and to determine how far the parameters that define the spatial dependence structure of the regionalized variable influence the choice for smaller sampling designs.

The geostatistical analyses were performed using the software R (R Development Core Team, 2010) and the module geoR (Ribeiro Jr. and Diggle, 2001).

Results and discussion

Analysis of the simulation results: a comparison between optimization processes

The simulated data sets, which have initial sampling grids of 400 sampling points with a nugget effect (C_0), a sill ($C_0 + C_1$) and a range (a) equal to 2, 10 and 60 meters, respectively, are presented in Figure 1 with one of the simulations and sampling sizes that fit the previously defined criteria for stage 7 of the optimization process. Scatter plots displaying the order of interactions versus the value for the objective function to be optimized and the arrangement of points obtained in the optimization process are shown. These scatter plots show that simulated annealing provides an efficient search of sampling designs that can maximize the overall accuracy (Figure 1-a) and minimize the sum quadratic error of the spatial prediction (Figure 1-b). The two optimization processes may also define a sampling design that will provide a better scope of the area studied (Figures 1-c and 1-d). Similar results can be found in Guedes *et al*. (2011), who considered the spatial prediction average variance as an objective function to be minimized, and Lark (2002), who has proved through simulations that the regularity of points distributed in the area can be found in the generation of optimized designs for a regionalized variable with lower range values.

The exponential model estimates (Equation 1) of the semivariance function (Table 1) are displayed for these simulations. It is clear that there was a moderate dispersion of estimates

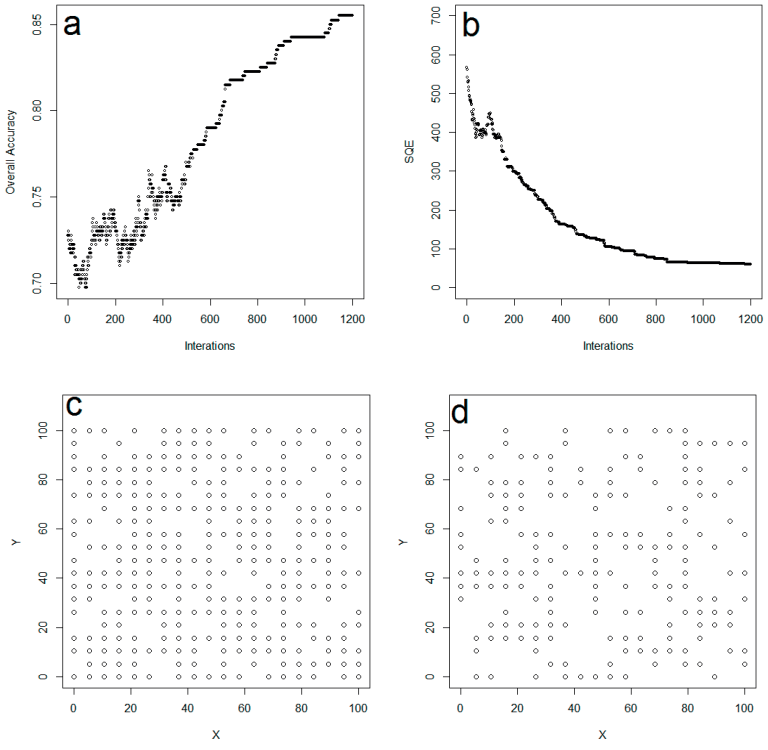

Figure 1. A dispersion chart of the interaction order versus the objective function values for sampling designs, which have improved (a) the overall accuracy and (b) the sum quadratic error (SQE) for the spatial prediction. An arrangement of points for sampling designs, which have improved (c) the overall accuracy and (d) the SQE (sum quadratic error) for spatial prediction, considering sampling sizes that satisfied the minimum values for these measures in one of the simulations.

and an underestimation of the range values for all reduced samplings. The mean values for the nugget effect and sill are closer to real values and are also registered in the samplings optimized by the overall accuracy maximization.

By analyzing the results displayed in Table 2, which refer to measurements related to the spatial prediction quality, it becomes clear that all reduced samplings report high accuracy measure values that are above or very close to levels that indicate a high similarity among the maps generated by the values simulated and obtained through the spatial prediction (De Bastiani *et al*., 2012). Moreover, the worst results for the spatial prediction variance mean ($\bar{S}_0^2$) and for the sum quadratic error of the spatial prediction (SQE) could be observed for the sampling design that improved the overall accuracy, as seen in Figures (2-a) and (2-b). This provides evidence that the lowest values of the spatial prediction variance mean in the simple random sampling schemes (Ale1 and Ale2) and the lowest values of the sum quadratic error of the spatial prediction in the sampling schemes have optimized this measure (Ale2 and TS2).

However, the sampling design obtained by the optimization of the overall accuracy generated, in all simulations, optimized sampling patterns containing the lowest number of points, as displayed in Figure (2-c), with a mean of approximately 168 sampling points corresponding to 42% of the initial grid (Table 2). In contrast, other sampling schemes use approximately 71% to 79% of the initial grid in the formation of an efficient sampling design that considers spatial predictions. Therefore, the high values of the accuracy measurements and the low values of the variance mean and the sum quadratic error of spatial prediction that were accomplished by using other sampling schemes were mainly due to the high number of sampling points used to identify a sampling design that could satisfy the optimization criteria.

Analysis of the simulation results: a comparison of the global accuracy optimization in models with distinct parameters or distinct sampling sizes

Considering simulations with 400 sampling points, Table 3 and Figure 3 display results of

Table 1. Descriptive statistics of the parameter estimates for the exponential model.

Statistics	Sampling Scheme	Mean ($\bar{x}$)	Standard Error (SE)	$\bar{x}$±2SE	Coefficient of variation
CE (real = 20%)	TS1	18.68	3.19	[12.30, 25.06]	54.04
	TS2	39.83	2.01	[35.81, 43.85]	15.95
	Ale1	24.49	2.51	[19.47, 29.51]	32.46
	Ale2	22.22	2.31	[17.60, 26.84]	32.81
Sill ($C_0 + C_1$)	TS1	10.41	1.21	[7.99, 12.83]	36.90
	TS2	9.36	1.04	[7.28, 11.44]	35.10
	Ale1	8.89	1.04	[6.81, 10.97]	36.89
	Ale2	9.02	0.93	[7.16, 10.88]	32.47
Range (a)	TS1	43.73	3.94	[35.85, 51,61]	28.74
	TS2	49.56	5.35	[38.86, 60.26]	34.13
	Ale1	52.90	5.78	[41.34, 64.46]	34.53
	Ale2	50.50	5.62	[39.26, 61.74]	35.21

$CE = 100 \times C_0/(C_0 + C_1)$.

Table 2. Descriptive statistics of the measures related to the spatial prediction quality and minimum necessary numbers according to the optimization criteria.

Statistics	Sampling Scheme	Mean ($\bar{x}$)	Standard Error (SE)	$\bar{x}$±2SE	Coefficient of variation
$\bar{\delta}_0^2$	TS1	3.67	0.19	[3.29, 4.05]	16.35
	TS2	2.44	0.13	[2.18, 2.70]	17.11
	Ale1	1.94	0.09	[1.76, 2.12]	14.85
	Ale2	1.03	0.04	[0.95, 1.11]	12.68
SQE	TS1	943.88	43.14	[857.60, 1030.16]	14.45
	TS2	95.58	10.75	[74.08, 117.08]	35.51
	Ale1	668.22	25.88	[616.46, 719.98]	12.25
	Ale2	313.49	11.93	[289.63, 337.35]	12.04
Overall Accuracy(*OA*)	TS1	0.87	0.003	[0.864, 0.876]	1.33
	TS2	0.94	0.006	[0.928, 0.952]	1.78
	Ale1	0.86	0.003	[0.854, 0.866]	0.69
	Ale2	0.92	0.003	[0.914, 0.926]	1.08
Kappa (*K*)	TS1	0.80	0.006	[0.788, 0.812]	2.27
	TS2	0.91	0.006	[0.898, 0.926]	2.69
	Ale1	0.79	0.003	[0.784, 0.796]	1.67
	Ale2	0.88	0.006	[0.868, 0.892]	1.75
Tau (*T*)	TS1	0.83	0.003	[0.824, 0.836]	1.72
	TS2	0.93	0.006	[0.918, 0.942]	1.94
	Ale1	0.82	0.003	[0.814, 0.826]	0.90
	Ale2	0.90	0.003	[0.894, 0.906]	1.37
Minimum Number of Sampling Points (*n*)	TS1	168.20	7.09	[154.02, 182.38]	13.03
	TS2	283.90	4.55	[274.80, 293.00]	5.07
	Ale1	250.40	4.83	[240.74, 260.06]	6.10
	Ale2	316.30	2.61	[311.08, 321.52]	2.61

SQE: sum quadratic error of spatial prediction.

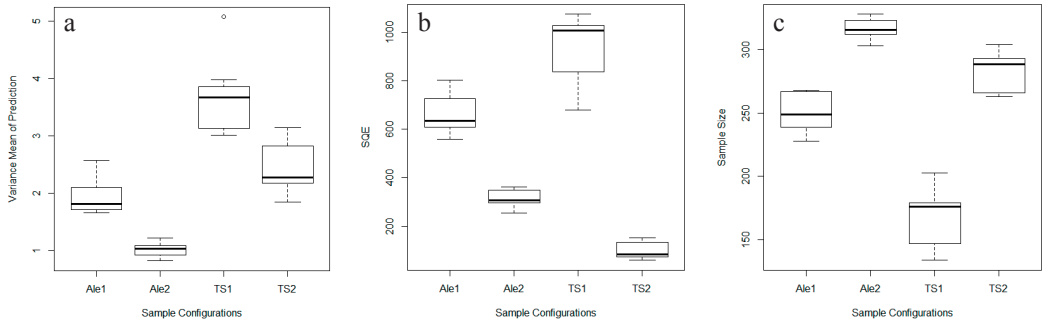

Figure 2. A box plot of (a) the mean spatial prediction variances, (b) the SQE (sum quadratic error) for spatial prediction and (c) the sampling size necessary in optimization for each reduced sampling scheme.

the quality measures for the spatial prediction and the number of necessary points required to maximize the overall accuracy for the sampling designs, considering the models proposed initially. These results suggest that, on average, with an increase in the range values, there is a similar increase in the optimized sampling designs in relation to the values for the spatial prediction variance mean (Figure 3-d) and for the accuracy measures (Figures 3-a, 3-b, 3-c), and a decrease in the sum quadratic error of the spatial prediction (Figure 3-e) and in the number of necessary points required by the optimization criteria.

An increase in the nugget effect also resulted in an increase in the variance mean of the spatial prediction (Figure 3-i) and in the sum quadratic error of the prediction (Figure 3-j). An increase in the nugget effect value in the simulated model also decreased the accuracy measure values (Figures 3-f, 3-g, 3-h) obtained for the optimized designs. The reduction in accuracy measures for the optimized designs, generated by an increase in the nugget effect, caused an increase in the number of necessary points for the optimization process. According to Lark (2002, 2011), the conception of a sampling design optimized by simulated annealing depends on the spatial structures underlying the regionalized variable.

Figure 4 displays box plots for some measures obtained for the best designs arrived at through the simulated annealing process in many simulations with different initial grid sizes (400, 625, 900 and

Table 3. Descriptive statistics of measures of spatial prediction accuracy and the minimum number of points necessary to reach the lowest level of high similarity between maps among the optimized sampling schemes, in relation to the distinct models simulated.

Statistics	Model	Mean ($\bar{x}$)	Standard Error (SE)	$\bar{x}$±2SE	Coefficient of variation
$\bar{S}_0^2$	Range = 60	3.67	0.19	[3.29, 4.05]	16.35
	Range = 75	3.67	0.11	[3.45, 3.89]	9.65
	Range = 90	3.57	0.13	[3.31, 3.83]	11.78
	Nugget Effect = 5	5.71	0.16	[5.39, 6.03]	8.87
SQE	Range = 60	943.88	43.14	[857.60, 1030.16]	14.45
	Range = 75	857.77	37.14	[783.49, 932.05]	13.69
	Range = 90	765.87	33.17	[699.53, 832.21]	13.70
	Nugget Effect = 5	1377.76	59.46	[1258.84, 1496.68]	13.65
Overall Accuracy (*GA*)	Range = 60	0.87	0.003	[0.864, 0.876]	1.33
	Range = 75	0.86	0.003	[0.854, 0.866]	0.92
	Range = 90	0.87	0.003	[0.864, 0.876]	1.61
	Nugget Effect = 5	0.86	0.003	[0.854, 0.866]	1.24
Kappa (*K*)	Range = 60	0.80	0.006	[0.788, 0.812]	2.27
	Range = 75	0.79	0.003	[0.784, 0.796]	1.71
	Range = 90	0.80	0.006	[0.794, 0.806]	2.74
	Nugget Effect = 5	0.79	0.006	[0.778, 0.802]	2.22
Tau (*T*)	Range = 60	0.83	0.003	[0.824, 0.836]	1.72
	Range = 75	0.83	0.003	[0.824, 0.836]	1.20
	Range = 90	0.84	0.006	[0.828, 0.852]	2.10
	Nugget Effect = 5	0.82	0.003	[0.814, 0.826]	1.61
Minimum Number of Sampling Points (*n*)	Range = 60	168.20	7.09	[154.02, 182.38]	13.03
	Range = 75	161.60	5.16	[151.28, 171.92]	10.10
	Range = 90	158.91	5.92	[147.07, 170.75]	11.78
	Nugget Effect = 5	185.45	5.24	[174.97, 195.93]	8.94

SQE: the sum quadratic error of the spatial prediction.

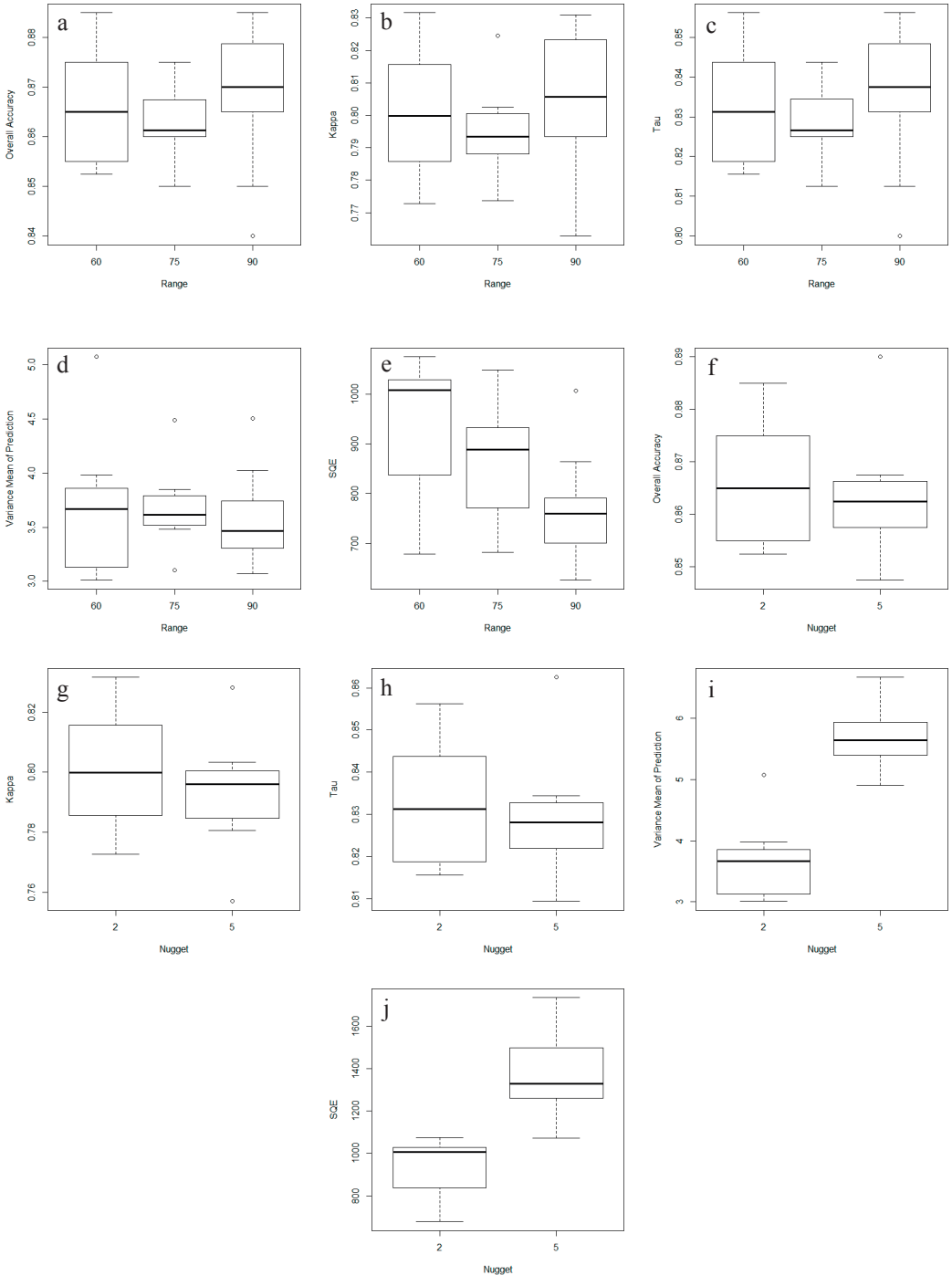

Figure 3. Box plots of the measures for optimized sampling designs: (a) Overall Accuracy, (b) Kappa Index, (c) Tau Index, (d) Variance mean spatial prediction and (e) SQE for models with distinct range values. (f) Overall Accuracy, (g) Kappa Index, (h) Tau Index, (i) Variance mean spatial prediction and (j) SQE for models with distinct nugget values.

1600) that aim to optimize the overall accuracy but that have the same values for the range (a = 60), nugget effect (C_0 = 2) and sill ($C_0 + C_1$ = 10) in the spatial dependence model.

With these results, it is possible to see that as the initial grid size increases, there is a corresponding decrease in the accuracy measures (Figures 4-a, 4-b, 4-c), especially in simulations with an initial sampling size of 1600. There is, however, a decrease in the spatial prediction mean variance in the optimized designs that comes from simulations with higher initial design sizes (Figure 4-d).

An increase in the initial grid also causes an increase in the sum quadratic error of spatial prediction and in the number of points necessary to generate the best sampling design. This increase is generated by the direct relationship between

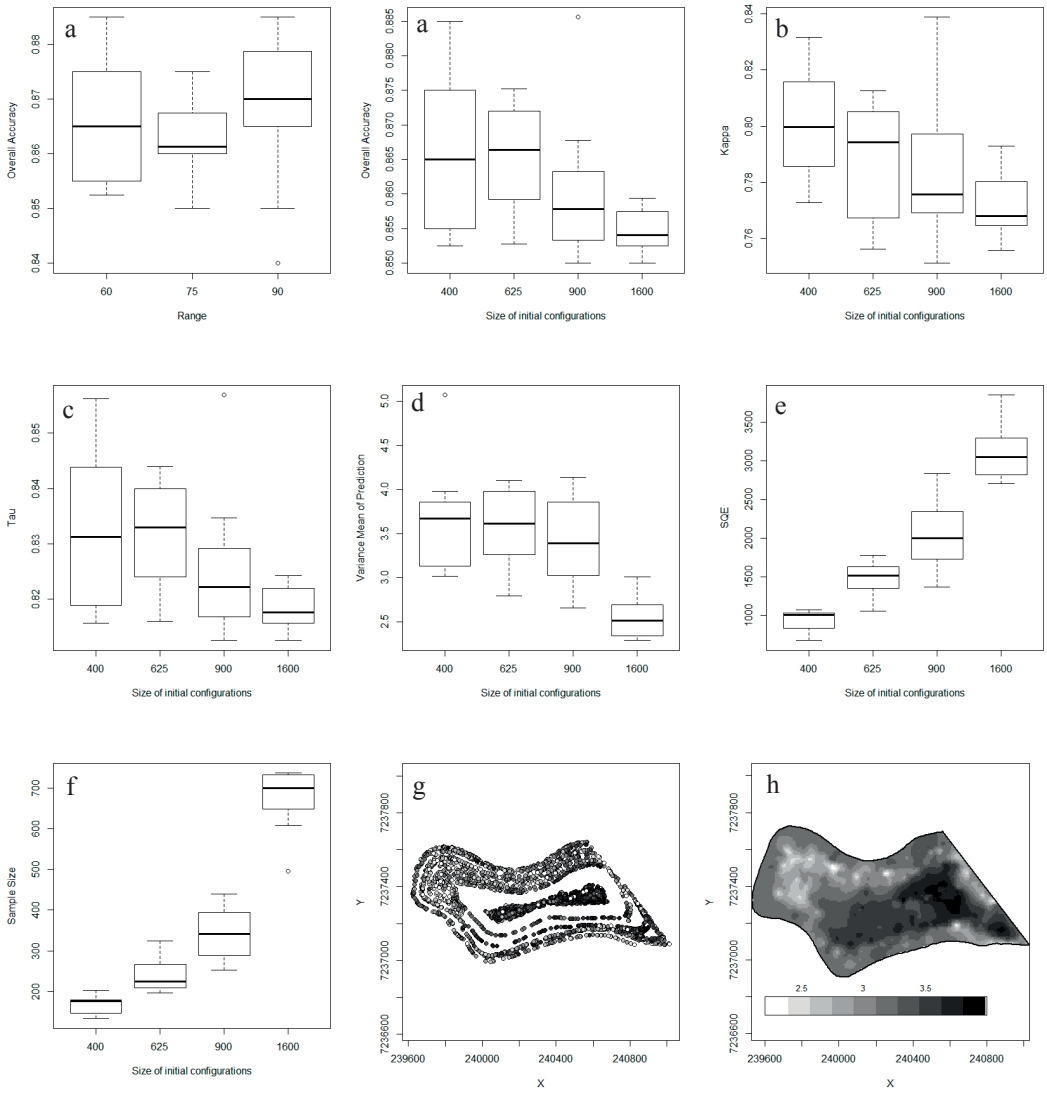

Figure 4. Box plots of the measures: (a) Overall Accuracy, (b) Kappa Index, (c) Tau Index, (d) Variance mean spatial prediction, (e) SQE and (f) number of points necessary to reach a minimum index of overall accuracy equal to 0.85 for optimized sampling designs reported from simulations with distinct initial sampling sizes. (g) Location for soy productivity values, distributed around their respective localizations in the reduced size optimized sampling grid, in which the colors for each point represent intervals delimited by quartiles and (h) a Thematic Map representing the spatial variability of soy production.

the sum quadratic error of spatial prediction and the stopping criterion of the optimization process ($OA \geq 0.85$) and the number of sampling points to be predicted, which has a similar size to the initial grid. However, the ratio of necessary points to the result of the best sampling design, in relationship to the initial grid size, is similar for all the simulations, varying from 38 to 42%.

Analysis of results obtained for the optimized reduction of the soy productivity data

Table 4 shows the descriptive statistics for soy productivity, in an optimized grid with a reduced sampling size, and in the initial grid of results obtained by the harvester monitor. The optimization process has reduced the initial grid size to 2000 points according to the stopping criterion of the algorithm (OA ≥ 0.85). This result represents 40% of the initial grid and is part of the ratio interval of points necessary for the configuration of the best sampling design obtained in the simulations. Even with this reduction, the descriptive statistics of the reduced grid appeared similar to the results presented for the initial grid.

Figure 4-g displays a chart with soy productivity values and their respective localizations, in an optimized reduced sampling grid, where the color for each point represents a value interval that includes the productivity; the intervals used are the same as the ones defined in the Materials and Methods for the calculation of the accuracy measures. There is regularity in the point distribution, with the appearance of regions with closer values and an excellent scope of the selected sample points in the optimization process from the initial grid.

Estimates in the exponential model parameters of the semivariance function were obtained for the optimally reduced sampling design and the random sampling design, using the same sample size as in the previous configuration displayed in Table 4. It is important to highlight that the choice for the best spatial model was made according to cross validation criteria and the Akaike Information Criterion (Faraco et al., 2008). Both models showed a moderate spatial dependence (25% ≤ CE ≤ 75%; Cambardella *et al*., 1994). However, the model that was estimated by considering the random sample configuration showed higher estimates for the parameters that described the structure of spatial dependence.

Table 4 also displays results concerning measures related to spatial prediction. The OA indexes and the concordance indexes of Kappa (K) and Tau (T) have proved that, even considering 40% of

Table 4. Descriptive statistics of soy productivity, as reported in the results of the harvester monitor, and the reduced size optimized design. Parameters estimated for the exponential model and measures associated with the spatial prediction of soy productivity based on the reduced optimal sampling design.

Statistics	Harvester Monitor	Optimized grid
Number of points	5000	2000
Average	3.23	3.23
Median	3.27	3.27
Q1	2.98	2.99
Q3	3.55	3.54
Minimum	0.68	0.83
Maximum	5.44	4.99
DP	0.67	0.48
CV (%)	20.74	14.75
Parameters for the Spatial Model	**Estimates for the optimized grid**	**Estimates for the sample grid**
C_0	0.1069	0.1457
$C_0 + C_1$	0.2173	0.2645
a	197.16	256.01
CE	49.19	55.09
$\bar{S}_0^2$	0.158	0.127
OA	0.851	0.834
K	0.702	0.666
T	0.834	0.815
ME	0.0002	-0.0001
S_{ME}	0.3660	0.4214
MSE	0.0003	-0.0002
S_{MSE}	0.9989	1.0070

$CE = 100 \times C_0/(C_0 + C_1)$, $\bar{s}_0^2$ is the spatial prediction variance mean, *OA* is the Overall Accuracy, and *K* and *T* are the Kappa and Tau indexes. The measurements obtained by cross-validation are *ME* (mean error), S_{ME} (standard deviation of the error), SME (mean standardized error) and S_{MSE} (standard deviation of standardized error).

the points displayed by the harvester monitor, the spatial prediction of the 5000 points presents a high similarity to the results obtained by the harvester monitor using the optimized sample configuration (OA ≥ 0.85, K ≥ 0.80, T ≥ 0.80; De Bastiani *et al*., 2012). This indicates that reductions in the sampling sizes due to the optimization process were efficient for the spatial prediction accuracy. This conclusion is confirmed by results obtained by cross-validation (Table 4), showing that the values of the mean error and mean standardized error were near zero and that the standard deviation of the mean standard error was close to 1. Furthermore, Table 4 also displays results concerning measures related to spatial prediction, using the random sample configuration. These results showed a slight decrease in measures of accuracy and an increase in the standard deviation of the error, which showed a better efficiency in the spatial prediction using the optimized design when compared to a random sampling.

Figure 4-h shows a thematic map for soy productivity for the determination of soybean yield (t ha^{-1}) in the study area, allowing for the zoning of areas with lower and higher productivity. This can facilitate the decision making of farmers, such that they can better standardize and maximize their productivity without affecting the environment.

We conclude that the analyses performed for the simulated data sets and real soy productivity data demonstrate that optimization methodology could define sampling designs that can provide a better scope of the area studied, apart from the objective function related to spatial predictions. We recommend using the objective Overall Accuracy function because it allows for the determination of an optimized sampling design with the smallest sampling size and a comparison between this design and the other methodologies presented.

Acknowledgements

We would like to express our gratitude to CAPES, CNPq, Brasília, Brazil and Fundação Araucária, Paraná, Brazil, for the financial support.

Resumen

L.P.C. Guedes, M.A. Uribe-Opazo, and P.J. Ribeiro Junior. 2014. Optimización del tamaño y de la forma de configuraciones muestrales para variables regionalizadas usando lo recocido simulado. Cien. Inv. Agr. 41(1): 33-48. La definición de la estructura de la variabilidad espacial de las variables regionalizadas por medio de técnicas geoestadísticas, permite estimar los valores de estas variables en lugares no incluidos en el muestreo, generando mapas temáticos que serán utilizados en la construcción de sectores agrícolas para aplicación diferenciada de tratamientos del suelo. La calidad de estos mapas depende de las estimativas confiables, que pueden cambiar por la selección la configuración del muestreo. El objetivo de este estudio fue determinar el tamaño de la muestra y la configuración de muestreo óptima, con el fin de maximizar la eficiencia del plano de muestreo, en la predicción de las variables con dependencia espacial. Las configuraciones muestrales optimizadas fueron obtenidas utilizando un método de búsqueda estocástica llamado Recocido simulado, a partir de una cuadrícula de muestreo con un gran número de puntos, inicialmente teniendo en cuenta un conjunto de datos simulados, con diferentes estructuras de dependencias espaciales y posteriormente un conjunto de datos reales de la productividad de la soya. Los resultados obtenidos de la simulación y del conjunto de datos reales de la productividad de la soya mostraron que el proceso de optimización fue eficiente en la determinación de configuración de muestreo óptima con tamaño de muestrea reducido, principalmente usando el índice de Exactitud Global como medida para maximizarla.

Palabras clave: Agricultura de precisión, geoestadística, interpolación, variabilidad espacial.

References

Anderson, J.F., E.E. Hardy, J.T. Roach, and R.E. Witmer. 2001. A land use and land cover classification system for use with remote sensor data. Geological Scope Professional Paper 964, Geologic Scope. Washington, USA. 41 pp.

Benvenga, M.A.C., S.A. Araújo, A.F.H. Librantz, J.C.C. Santana and E.B. Tambourji. 2011. Application of simulated annealing in simulation and optimization of drying process of *Zea mays* malt. Eng. Agríc. [online] 31: 940-953.

Bôer, E.P.J., A.L.M. Dekkers and A. Stein. 2002. Optimization of a monitoring network for sulfur dioxide. Journal of Environmental Quality 31:121-128.

Cambardella, C.A., T.B. Moorman, J.M. Novak, T.B. Parkin, D.L. Karlen, R.F. Turco, and A.E. Konopka. 1994. Field scale variability of soil properties in Central Iowa soils. Soil Sci. Soc. Am. J. 58:1501-1511.

Chakrapani, Y., and K.S. Rajan. 2008. Hybrid Genetic-Simulated Annealing Approach for Fractal Image Compression. International Journal of Computational Intelligence 4: 308-313.

Costa Filho, C.C.F., A.T. Albuquerque, and M.G.F. Costa. 2010. Luminance Optimization in closed environments by simulated annealing. IEEE Latin America Transactions 8:229-23.

De Bastiani, F., M.A. Uribe-Opazo, and G.H. Dalposso. 2012. Comparison of maps of spatial variability of soil resistance to penetration constructed with and without covariables using a spatial linear model. Engenharia Agrícola (Impresso) 32:394-404

Diggle, P.J., and P.J. Ribeiro Junior. 2007. Model-based Geostatistics. 1ª Edição. Springer, New York, USA. 230 pp.

Dunn, R., and A.R. Harrison. 1993. Two dimensional systematic sampling of land use. Applied Statistics 42:585-601.

Faraco, M.A., M.A. Uribe-Opazo, E.A. Silva, J.A. Johann, and J.A. Borssoi. 2008. Seleção de modelos de variabilidade espacial para elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja. Revista Brasileira de Ciência do Solo 32:463-476.

Ferreyra, R.A., H.P. Apezteguía, R. Sereno and J.W. Jones. 2002. Reduction of soil water spatial sampling density using scaled semivariograms and simulated annealing. Geoderma 100:265-289.

Guedes, L.P.C., P.J. Ribeiro Junior, S.M.S. Piedade and M.A. Uribe-Opazo. 2011. Optimization of spatial sample configurations using hibrid genetic algorithm and simulated annealing. Chilean Journal of Statistics 2:39-50.

Johann, J.A., M.C.A. Silva, M.A. Uribe-Opazo, and G.H. Dalposso. 2010. Variabilidade espacial da rentabilidade, perdas na colheita e produtividade do feijoeiro. Eng. Agríc. [online] 30:700-714.

Lark, R.M. 2002. Optimized spatial of soil for estimation of the variogram by maximum likelihood. Geoderma 105:49-80.

Lark, R.M. 2011. Spatially nested sampling schemes for spatial variance components: scope for their optimization. Computers e Geosciences 37:1633-1641.

Le, N.D., L. Sun, and J.V. Zidek. 2003. Designing networks for monitoring multivariate environmental fields using data with monotone pattern. Technical Report, Statistical and Apllied Mathematical Sciences Institute, NC. 37 pp.

Ma, Z., and R.L. Redmond. 1995. Tau coefficients for accuracy assessment of classification of remote sensing data. Photogrammetric Engineering and Remote Sensing 61:4535-439.

Mcbratney, A.B., R. Webster, and T.M. Burgess. 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables. I. Theory and method. Comput. Geosci. 7:331–334.

Oda-Souza, M., J.L.F. Batista, P.J. Ribeiro Junior, and R.R. Rodrigues. 2010. Influência do tamanho e forma da unidade amostral sobre a estrutura de dependência espacial em quatro formações florestais do estado de São Paulo. Floresta 40:849-860.

R Development Core Team. 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 1706 pp.

Ribeiro Jr., P.J., and Diggle, P.J. 2001. geoR: A package for geostatistical analysis. R-News 1/2:15-18.

Reichert, J.M., T.A. Dariva, D.J. Reinert, and V.R. Silva. 2008. Variabilidade espatial de Planossolo e produtividade de soja em várzea sistematizada: análise geoestatística e análise de regressão. Ciência Rural 38:981-988.

Ruiz-Cárdenas, R., M.A.R. Ferreira and A.M. Schmidt, 2010. Stochastic search algorithms for optimal design of monitoring networks. Environmetrics 21:102-112.

Van Groenigen, J.W. 2000. The influence of variogram parameters on optimal sampling schemes for mapping by kriging. Geoderma 97:223-236.

Yfantis, E.A., G.T. Flatman, and J.V. Behar. 1987. Eficiency of kriging estimation for square, triangular and hezagonal grids. Mathematical Geology 19:183-205.