DOI: 10.15201/hungeobull.64.1.4 Hungarian Geographical Bulletin **64** 2015 (1) 35–48.

An application of a spatial simulated annealing sampling optimization algorithm to support digital soil mapping

Gábor SZATMÁRI[1], Károly BARTA[1] and László PÁSZTOR[2]

Abstract

Spatial simulated annealing (SSA) was applied to optimize the sampling configuration for soil organic matter mapping through various sampling scenarios in a Hungarian study site. Prediction-error variance of regression kriging was applied as quality measure in the optimization procedures. Requisites of SSA come from a legacy soil dataset and from spatial auxiliary information. Four scenarios were set to represent the major capabilities of SSA. Scenario 1 and 2 represented completely new sampling designs to optimize with predefined constraints. In scenario 1, number of new observations was the constraint, whilst in scenario 2, it was the value of the quality measure. In both scenarios, areas inaccessible for sampling (roads, farms etc.) were also taken into account. Scenario 3 and 4 represented complementary sampling configurations to optimize taking the previously collected samples into consideration. In scenario 3, the constraint was the number of new observations, whilst in scenario 4, it was the value of the quality measure. In both cases, two types of previously collected sampling design were simulated, a regular and a clustered configuration. The resulted designs were evaluated by Kolmogorov–Smirnov test, nearest neighbour distribution function and empty space function. In cases of scenario 1 and 3, the results showed that, all of the optimized sampling configurations cover properly both geographic and feature space, respectively. In cases of scenario 2 and 4, the resulted calibration curves can be used to determine the sample size for a given quality measure value. Furthermore, we could determine the minimal sample size for a given scenario, which has to be collected to represent properly both geographic and feature space. In conclusion, SSA is a valuable tool to optimize the sampling design considering a lot of constraints.

Keywords: spatial simulated annealing, sampling optimization, geostatistics, regression kriging prediction-error variance, digital soil mapping

Introduction

Digital soil mapping (DSM) aims at spatial prediction of soil properties by combining soil observation at points with auxiliary information, such as contained in digital elevation models, remote sensing images and climate data records (McBratney, A.B. *et al.* 2003; Heuvelink, G.B.M. *et al.* 2007). Hence, the direct observations of the soil are im-

[1] Department of Physical Geography and Geoinfor-matics, Faculty of Science and Informatics, University of Szeged. H-6722 Szeged, Egyetem u. 2., E-mails: szatmari.gabor.88@gmail.com, barta@geo.u-szeged.hu
[2] Institute for Soil Science and Agricultural Chemistry, Centre for Agricultural Research, Hungarian Academy of Sciences. H-1022 Budapest, Herman Ottó u. 15. E-mail: pasztor@rissac.hu

portant for two main reasons (Heuvelink, G.B.M. *et al.* 2007):
– they are used to characterize the relationship between the soil property of interest and the auxiliary information,
– they are used to improve the predictions based on the auxiliary information, by spatial interpolation of the differences between the observations and predictions.

Regression kriging (RK) (also termed universal kriging or kriging with external drift, see Hengl, T. *et al.* 2007) illustrates well that twofold application of the soil observations. Spatial prediction method of RK combines a regression of the target pedological variable on covariates with kriging of the regression residuals. Nevertheless RK assumes that, the sampling points represent properly both geographic and feature space (Hengl, T. 2009), where the latter is defined by the covariates.

Extensive work has been done on sampling strategy optimization for DSM over the past decades to satisfy the topical demands, which were suggested by soil surveyors, pedometricans, end-users, and so forth. These demands can be e.g. the expectation of the accuracy and/or uncertainty of the prediction(s), taking auxiliary information into account, optimization of the sampling design for more than one soil variable, taking previously collected samples into account, consideration of any kind of constraints, such as the number of the new observations, inaccessible areas for sampling, budget and/or accuracy constraints. One of the optimization algorithms is spatial simulated annealing (SSA) (van Groenigen, J.W. and Stein, A. 1998) that has been frequently applied in soil surveys to optimize the sampling design using the RK prediction-error variance (RKV) as optimization criterion (Brus, D.J. and Heuvelink, G.B.M. 2007; Heuvelink, G.B.M. *et al.* 2007; Baume, O.P. *et al.* 2011; Melles, S.J. *et al.* 2011; Szatmári, G. 2014). SSA with RKV is sporadically able to satisfy the above mentioned demands.

The main aim of this paper is to present and test the SSA sampling optimization algorithm through various sampling scenarios in a Hungarian study site. The scenarios were set to represent the major capabilities of SSA and to cover a major part of soil sampling issues. In all scenarios, the goal was to optimize the sampling design for soil organic matter (SOM) mapping considering some constraints (e.g. number of new observations, inaccessible areas for sampling, previously collected samples). The resulted sampling configurations were evaluated by various statistical and point pattern analysis tools, in order to examine how they cover both the geographic and feature space.

Theoretical backgrounds

Some thoughts on (spatial) soil sampling for digital mapping

Sampling concerns selection of a subset of individuals from a population to estimate the characteristics of the whole population; where these characteristics could be the total or mean parameter value for a random field, values at unvisited sites or location of target(s) (Wang, J.-F. *et al.* 2012).

In case of DSM, the main aim for a given pedological variable is to estimate its values at unsampled locations. For this purpose, various statistical models (i.e. spatial prediction methods) have been widely used, where we assume that the models and the "real world" are compatible. Furthermore, this implies that the sampling is representative for the whole population. According to Bárdossy, Gy. (1997), the sampling is said to be representative (from a statistical viewpoint) for a population, if it reflects the characteristics of the population the best.

On other hand, we do not know exhaustively the whole population, just only a small part of it (provided by the samples). How can we decide that, the sampling is representative for the whole population? If we know the components of the given statistical model, we can set a "quasi optimal state" through the sampling strategy, where we can assume that, the collected samples are representative for the whole population.

Therefore, the statistical inferences are compatible with the "real world". The setting of the sampling strategy can be regarded as an optimization problem.

As we will see in the next subsection, the RK spatial prediction method assumes that, the variation of the soil property of interest can be modelled as a sum of a deterministic (which is based on the covariates) and a stochastic (which is based on the variogram or covariance function) components. Therefore, if we describe properly, through the sampling design, both the feature (which is defined by the covariates) and geographic space, we can assume that, the statistical inference (i.e. map of the soil property of interest) represent the real situation. It can be regarded as an optimization problem, where we need an optimization algorithm and an optimization criterion. As we will see in the next subsections, SSA will be this algorithm and RKV will be this criterion.

Regression Kriging (RK) spatial prediction method

In the last decade, RK has been more and more popular in DSM (Hengl, T. *et al.* 2004; Dobos, E. *et al.* 2007; Hengl, T. *et al.* 2007; Minasny, B. and McBratney, A.B. 2007; Illés, G. *et al.* 2011; Szatmári, G. and Barta, K. 2013; Pásztor, L. *et al.* 2014), as well as in SSA sampling optimization procedure using its prediction-error variance as optimization criterion (Brus, D.J. and Heuvelink, G.B.M. 2007; Heuvelink, G.B.M. *et al.* 2007; Baume, O.P. *et al.* 2011; Melles, S.J. *et al.* 2011; Szatmári, G. 2014). RK assumes that, the deterministic component of the target soil variable is accounted for by the regression model, whilst the model residuals represent the spatially varying but dependent stochastic component, as well as both components can be modelled separately and simultaneously. The estimation for Z variable at an unvisited location s_0 is given by

$$Z(_{s0}) = q_0^T \cdot \beta + \lambda_0^T \cdot (z - q \cdot \beta), \quad (1)$$

where β is the vector of the regression coefficients, q_0 is the vector of the covariates at the unvisited location, λ_0 is the vector of the kriging weights, z is the vector of the observations and q is the matrix of covariates at the sampling locations. Its prediction-error variance at s_0 is given by

$$\sigma^2(s_0) = c(0) - c_0^T \cdot C^{-1} \cdot c_0 + (q_0 - q^T \cdot C^{-1} \cdot c_0)^T \cdot \cdot (q^T \cdot C^{-1} \cdot q)^{-1} \cdot (q_0 - q^T \cdot C^{-1} \cdot c_0), \quad (2)$$

where $c(0)$ is the variance of the residuals, c_0 is the vector of covariances between the residuals at the observed and unvisited locations and C is the variance-covariance matrix of the residuals. RKV is independent from the observed values (see Eq. [2]), so it can be calculated before the actual sampling takes place, which can be considered as a beneficial property in point of costs and time. Furthermore, it incorporates both the prediction error variance of the residuals (first two terms on the right-hand side of Eq. [2]) and the estimation error variance of the trend (third term on the right-hand side of Eq. [2]), which endeavour SSA algorithm to optimize the sampling design both in geographic and feature space (Heuvelink, G.B.M. *et al.* 2007). However, it mainly depends on, how the two types of error variance contribute to RKV.

Spatial simulated annealing (SSA) sampling optimization algorithm

In brief, SSA is an iterative, combinatorial, model-based sampling optimization algorithm in which a sequence of combinations is generated by deriving a new combination from slightly and randomly changing the previous combination (van Groenigen, J.W. *et al.* 1999). When a new combination is generated, the quality measure (in this study the spatially averaged RKV) is calculated and compared with the quality measure value of the previous combination (van Groenigen, J.W. *et al.* 1999; Brus, D.J. and Heuvelink, G.B.M. 2007). The Metropolis criterion defines the probability that, either accepts the new combination as a basis for the further computation, or rejects it and the previous combination stays as a basis further (van Groenigen, J.W. *et al.* 1999):

$$P(C_i \to C_{i+1}) = 1, \quad \text{if } \Phi(C_{i+1}) \le \Phi(C_i)$$
$$P(C_i \to C_{i+1}) = \exp\left(\frac{\Phi(C_i) - \Phi(C_{i+1})}{c}\right), \quad \text{if } \Phi(C_{i+1}) > \Phi(C_i) \tag{3}$$

where C_i and C_{i+1} are the previous and the new combination, c is the positive control parameter (so-called "system temperature", which is lowered as optimization progresses) and $\Phi(\cdot)$ is the quality measure (so-called "fitness or objective function").

For a given soil variable, SSA (using RKV as optimization criterion) requires that the structure of the regression model and the variogram or covariance function of the residuals are known (Heuvelink, G.B.M. *et al.* 2007), which is one of the main drawbacks of the method. On other hand, the algorithm is able to take inaccessible areas and/or previously collected samples into account.

Material and methods

Study site and legacy soil data

The study site (approx. 17 km^2) is located in the central part of Hungary, in the Mezőföld region, near village Előszállás (*Figure 1*). The area of interest is mainly covered by Haplic Chernozems and Kastanozems with significant secondary carbonates. Calcisols and Regosols are found on the eroded steeper slopes, where the top-horizon is too thin for Mollic or it is completely missing. Colluvic material can be found at the bottom of the slopes, where Phaeozems or Regosols were formed. The study site can be characterized mainly by arable lands sown with winter wheat, maize and sunflower.

The available legacy soil data was collected at the end of the 1980s in the framework of the National Land Evaluation Programme. The dataset incorporates 117 topsoil (0–30 cm) observations from the area of interest. Various pedological variables were quantified during the fieldwork and laboratory analyses. In this study, the soil organic matter (SOM) was chosen as target pedological variable to optimize the sampling design for various scenarios. Exploratory data analysis was performed on SOM data to remove the outliers, to calculate summary statistics and to test the normality of the SOM probability distribution. The analysis has shown that, the probability distribution of SOM is close to normal. The summary statistics of SOM are presented in *Table 1.*

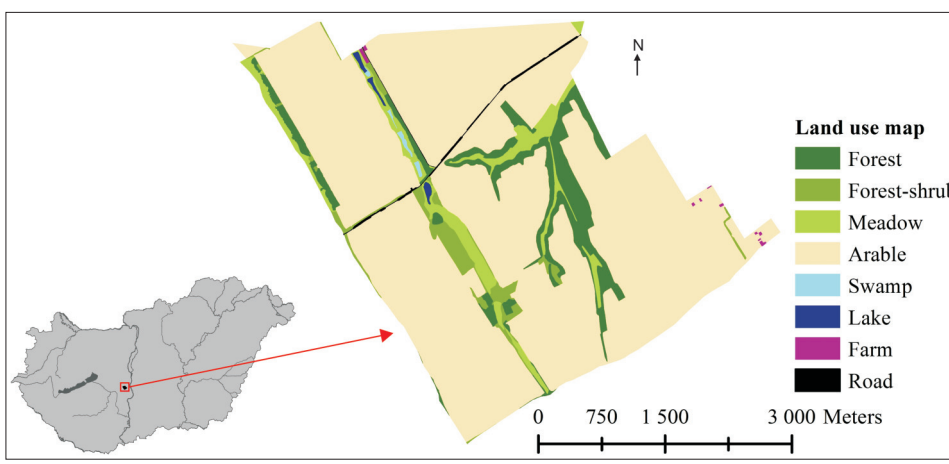

Fig. 1. The location of the study area in Hungary and its land use map

Table 1. Summary statistics of soil organic matter (SOM) computed from the legacy soil dataset without outliers

Variable	Mean	Median	Minimum	Maximum	Standard deviation	Skewness
	value					
SOM, %	2.90	2.95	1.51	4.44	0.56	–0.28

Auxiliary information from the study site

Spatially exhaustive auxiliary information were derived from digital elevation model (DEM) (with 20 meters resolution) and from land use (LU) map of the study area, since Szatmári, G. and Barta, K. (2012, 2013) pointed out that the spatial distribution/variability of SOM mainly depends on the topography and the LU at the area of interest. The following morphometric parameters were derived from DEM: altitude, slope (in percent), slope length, aspect, profile and planar curvature, LS factor (Wischmeier, W.H. and Smith, D.D. 1978), topographic wetness index, vertical distance to channel network and potential incoming solar radiation (direct and diffuse). LU map was derived from the products of the official aerial photography campaign of Hungary, taken in 2005.

In contrast with the morphometric parameters, LU type is a categorical variable. For the sake of the application of RK each LU type was converted into indicator variables. Raster maps were generated for each LU types with value domain showing 1 at the locations of the given LU type and showing 0 for all other locations. These raster maps were resampled for 20 meters.

Principal component (PC) analysis was performed on the auxiliary data and the resulted PCs were used as covariates in the further analysis. It is a crucial step, since the PCs are orthogonal and independent; hence they satisfy the requirements of the multiple linear regression analysis and their application decreases the multi-collinearity effect.

Settings of spatial simulated annealing and sampling scenarios

The requirements of SSA (using RKV as optimization criterion) are the structure of the regression model and the variogram or covariance function of residuals of the model. These requisites were generated from the legacy soil dataset and from the covariates, respectively. Multiple linear regression analysis was performed to characterize the relationship between SOM and covariate data, using a "stepwise" selection method and a significance level of 0.05. In the next step, the residuals were derived from the resulted regression model and exploratory variography was performed on them. The experimental variograms were calculated and the spatial structure was modelled with a theoretical variogram model. The fitted variogram and regression model were used along the optimization process provided by SSA to calculate (using Eq. [2]) the quality measure (i.e. spatially averaged RKV).

There are some land use types (swamp, lake, farm and road), which are out of the scope of soil mapping, so we excluded them from the optimization process as inaccessible areas for sampling.

The initial "system temperature" for SSA was chosen such that the average increase acceptance probability was 0.8 and the "system cooling" was exponentially. Furthermore, a stopping criterion was defined to rein up the simulation when the quality measure did not improve in many tries. The stopping criterion value was set 200.

The sampling scenarios were set to represent the major capabilities of SSA and to cover a major part of soil sampling issues. The following four scenarios were set to optimize the sampling design for SOM mapping:

– Scenario 1 (Sc1): Completely new sampling strategy with fixed number of new observations,
– Scenario 2 (Sc2): Completely new sampling strategy to achieve a predefined quality measure value,

– Scenario 3 (Sc3): Complementary sampling with fixed number of new observations to supplement the previously collected samples,
– Scenario 4 (Sc4): Complementary sampling to supplement the previously collected samples and to achieve a predefined quality measure value.

Two types of previously collected sampling configuration were applied as complementary sampling scenarios (Sc3 and Sc4):
– Regular design, where the sampling points located at the nodes of a square grid,
– Clustered design, where the sampling points showed a clustered pattern in the geographic space.

In case of Sc1, the number of new observations was set 120, which is commensurable with the sample size of the legacy soil dataset. In Sc3 and Sc4, the previously collected sample size was set 35, which were following regular and clustered design, respectively. In case of Sc3, the fixed number of new observation was set 50. In cases of Sc2 and Sc4, the main aim was to create a so-called calibration curve. This calibration curve can be used to determine the sample size for a given quality measure value and vice versa. To calculate this curve, the sample size was systematically increased and the quality measure value of the optimized configuration was calculated. In next step, the quality measure values were plotted as a function of the sample size.

Evaluation of the optimized sampling designs

The optimized sampling designs were evaluated by various statistical and point pattern analysis tools. Kolmogorov–Smirnov (K–S) test was applied to examine for a given covariate, if its distribution from the optimized design is equal to the distribution from the complete area of interest. Based on the test results we can examine how the sampling configurations cover the feature space created by the covariates.

The nearest neighbour distances distribution functions *G(r)* and the empty space functions *F(r)* were calculated, based on the sampling designs, to explore the type of interaction between the sampling points and to examine how they cover the geographic space. The *G(r)* function measures the distribution of the distances from an arbitrary sampling point to its nearest sampling point, while the *F(r)* function measures the distribution of all distances from an arbitrary point of the plane to its nearest sampling point (Bivand, R.S. *et al.* 2008). In case of *F(r)*, the grid nodes of the planned prediction locations were applied to measure the so-called empty space distances. It gives direct information on the kriging neighbourhood.

Results and discussion

Regression and variogram models

The determination coefficient of the resulted regression model was 0.41, which means that the model explains more than 40 percent of the total variability of SOM and the remaining approx. 60 percent have to be modelled stochastically. Five covariates were selected into the model by the "stepwise" method. The observed significance level, which was calculated for the model, was practically zero.

The regression residuals were derived and the experimental variograms (directional and omnidirectional) were calculated to model their spatial continuity. The directional variograms showed an isotropic spatial structure, which structure was approached by a spherical variogram model type. *Table 2* summarizes the parameters of the fitted isotropic variogram model.

Table 2. Parameters of the fitted isotropic variogram model for soil organic matter (SOM) residuals

Variable	Model type	Nugget	Partial sill	Sill	Nugget/Sill, %	Range, m
SOM residuals	Spherical	0.04	0.12	0.16	25.00	1,420

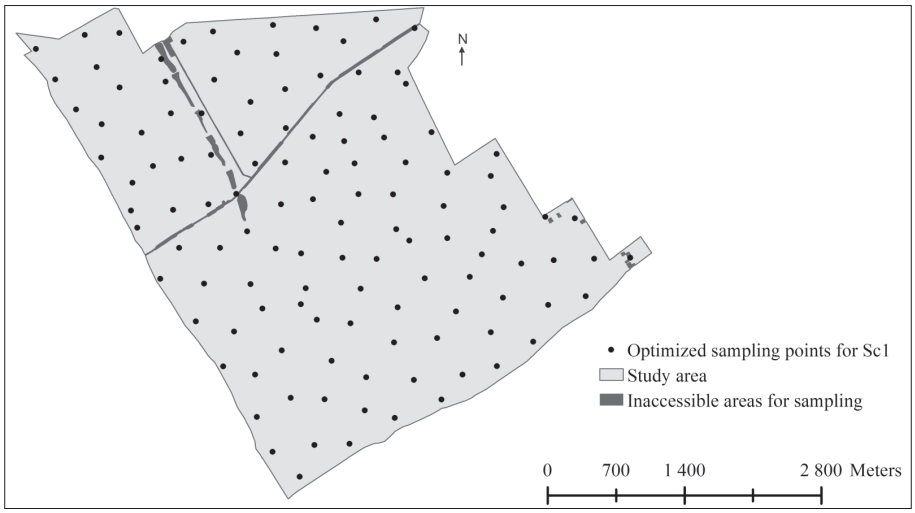

Fig. 2. The optimized sampling design for scenario 1

Optimized sampling designs for Sc1–Sc2 and their performance

The optimized sampling configuration for Sc1 is presented in *Figure 2*, which sampling design shows a "quasi" regular point pattern. *Figure 3* presents the calibration curve for Sc2 (denoted with solid line), as well as the nugget variance of SOM residuals (denoted with dashed line), where the latter is constant, because this part of variance cannot be modelled (Webster, R. and Oliver, M.A. 2007). The so-called "nugget effect" arises from measurement errors and/or small-scale heterogeneity (Goovaerts, P. 1999; Geiger, J. 2006; Webster, R. and Oliver, M.A. 2007). It also means that the value of the spatially averaged RKV cannot be less than this nugget variance (see Eq. [2]). Hence, the calibration curve converges to the nugget variance, if the sample size is infinitely large (see *Figure 3*).

The calculated calibration curve for Sc2 can be used to determine the sample size for a given spatially averaged RKV value expected to be achieved for the SOM map. In a practical point of view, this kind of calibration curve is a useful tool to estimate the sample size considering the predefined RKV value (expected to be achieved for the map) and/or the sampling budget's constraints. For example, if the soil surveyors want to achieve 0.08 [%]^2 value of spatially averaged RKV for the SOM map, then the sample size, using this calibration curve (*Figure 3*) is 98. On other hand, if the budget allows to collect 42 number of soil samples and the question is "What is the expectation of the spatially averaged RKV for the SOM map?", then, using the calibration curve (*Figure 3*), the expectation is 0.1 [%]^2.

The observed significance levels of K–S test for Sc1 and Sc2 are presented by *Table 3*. The null hypothesis was that, the two distributions were drawn from the same distribution.

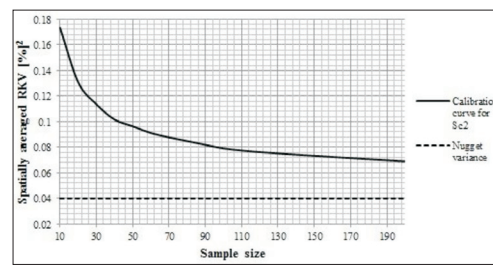

Fig. 3. The calibration curve for scenario 2 and the nugget variance. RKV = regression kriging prediction-error variance

The applied significance level was 0.05. In *Table 3* the values of the observed significance level *were bolded*, where the null hypothesis was accepted. In case of Sc1, the null hypothesis was accepted for all covariates, which means that, the optimized sampling design for Sc1 covers properly the feature space. In case of Sc2, we examined for a given sample size that, how the optimized sampling configuration covers the feature space. As we can see, 60 is the minimal sample size, which is needed to cover properly the feature space (see *Table 3*). Based on this, samples with less than 60 observations are not suitable to describe the trend function, as well as the spatial distribution of SOM.

The observed *F(r)* and *G(r)* functions gave almost the same results for Sc1 and Sc2, thanks to the relatively large range of the variogram (see *Table 2*). We can, however, state that, the optimized sampling configurations covered properly the geographic space, because the *r* value for *F(r) = 1* was lower than the variogram range, respectively. As a consequence, there was no any planned prediction location, which did not have any kriging neighbours. Furthermore, there is an inhibition (i.e. competition) between the sampling points, which follows from that, the $G_{obs}(r)$ function is below the theoretical distribution of complete spatial randomness (e.g. in *Figure 4*, a), whilst the $F_{obs}(r)$ function is above the theoretical distribution of complete spatial randomness (e.g. in *Figure 4*, b).

As a consequence, it causes a quasi-regular point pattern, respectively (as we can also see in *Figure 2*). *Figure 4* presents the observed *G(r)* and *F(r)* functions of the optimized sampling design for Sc1 (the calculated *G(r)* and *F(r)* functions for Sc2 were omitted, because they gave a similar results as in case of Sc1, due to the large range of the variogram).

Optimized sampling designs for Sc3–Sc4 and their performance

The optimized sampling configurations for Sc3 regular and Sc3 clustered are presented in *Figure 5. Figure 6* presents the calculated calibration curves for Sc4 regular (denoted with solid line) and Sc4 clustered (denoted with dashed line), as well as the nugget variance of the fitted variogram model (denoted with dotted line). Both calibration curves converge to the nugget variance, if the sample size is infinitely large (see *Figure 6).*

The calculated calibration curves for Sc4 regular and Sc4 clustered can be used to determine the sample size for a given spatially averaged RKV value and vice versa. For example, if the soil surveyors want to achieve

Table 3. The values of the observed significance level of Kolmogorov-Smirnov test calculated for Scenario 1 and 2.

Sample size	Covariates*				
	SPC1	SPC2	SPC3	SPC4	SPC5
10	0.017	0.000	*0.060*	0.006	0.035
20	*0.240*	0.013	0.021	0.042	*0.173*
30	*0.240*	0.035	*0.153*	0.006	0.013
40	*0.454*	*0.172*	*0.617*	0.017	*0.173*
50	*0.454*	0.035	*0.617*	*0.095*	*0.172*
60	*0.454*	*0.082*	*0.617*	*0.194*	*0.082*
70	*0.734*	*0.329*	*0.617*	*0.194*	*0.173*
80	*0.734*	*0.173*	*0.905*	*0.358*	*0.082*
90	*0.734*	*0.329*	*0.617*	*0.194*	*0.560*
100	*0.454*	*0.173*	*0.617*	*0.358*	*0.329*
110	*0.954*	*0.329*	*0.905*	*0.194*	*0.329*
120	*0.954*	*0.560*	*0.905*	*0.358*	*0.082*
150	*0.734*	*0.560*	*0.617*	*0.193*	*0.173*
200	*0.734*	*0.560*	*0.617*	*0.841*	*0.173*

*The observed significance levels are in italics, where the null hypothesis was accepted at 0.05 significance level

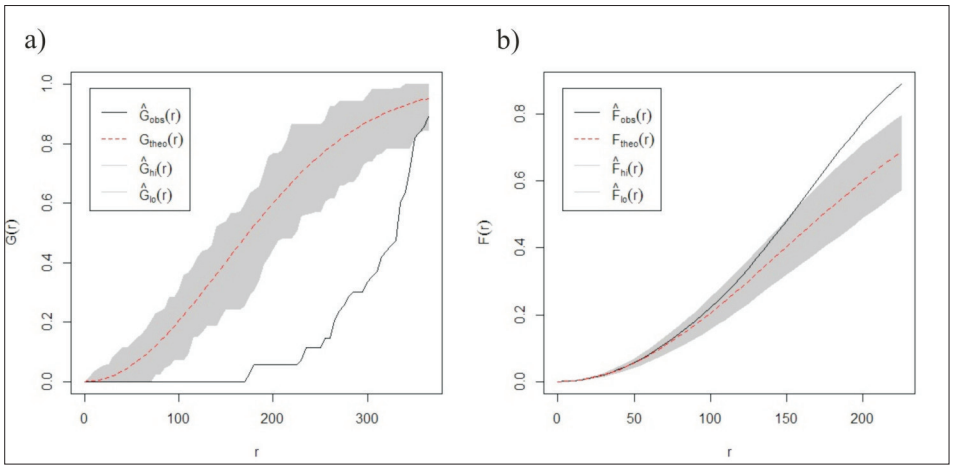

Fig. 4. The observed $G_{obs}(r)$ nearest neighbour distances distribution (a) and $F_{obs}(r)$ empty space function (b) for scenario 1. Abbreviations inside the legend: theo = theoretical distribution of complete spatial randomness; hi = upper envelope of theo; lo = lower envelope of theo

0.08 [%]² value of spatially averaged RKV for the SOM map, when the previously collected sampling design is regular, then the number of new observations, using the calibration curve (*Figure 6*), is 64. On other hand, when the previously collected sampling design is clustered the number of new observations, using the corresponding calibration curve (*Figure 6*) is 84. The large difference between them can be attributed to the follows: when the previously collected sampling design was clustered, the existing samples concentrated only on a small part of the complete area of interest (see *Figure 5*, b), which yielded higher RKV values, as well as caused a poor coverage both in geographic and feature space. On other hand, the existing regular sampling design covered more properly the geographic space (see *Figure 5*, a).

The observed significance levels of K–S tests for Sc3 regular and Sc4 regular are presented in *Table 4*, whilst the observed significance levels for Sc3 clustered and Sc4 clustered are presented in *Table 5*. The null hypothesis and the applied significance level were the same as in case of Sc1 and Sc2. In *Table 4* and *5*, the values of the observed significance level *were bolded*, where the null hypothesis was accepted.

In both cases of Sc3 regular and Sc3 clustered, the null hypothesis was accepted for all covariates, which means that, the optimized sampling designs for Sc3 regular and Sc3 clustered cover properly the feature space. In cases of Sc4 regular and Sc4 clustered, we examined for a given sample size, how the optimized sampling configuration covers the feature space. As we can see in *Table 4* and *5*, 40 is the minimal sample size, which is needed to cover properly the feature space. Based on this, samples with less than 40 observations are not suitable to describe the trend function, as well as the spatial distribution of SOM.

The observed *F(r)* and *G(r)* functions for the previously collected sampling designs are presented in *Figure 7*. In case of clustered design, the *r* value for *F(r) = 1* is higher than the variogram range (which means that, there are some planned prediction locations, which do not have any kriging neighbours), whilst in case of regular design, this *r* value is lower than the variogram range. They support the ascertainment, the clustered design does not cover properly the geographic space, whilst the regular design does (see *Figure 7*).

In cases of Sc3 regular and Sc4 regular, the *F(r)* and *G(r)* functions gave "quasi" the same

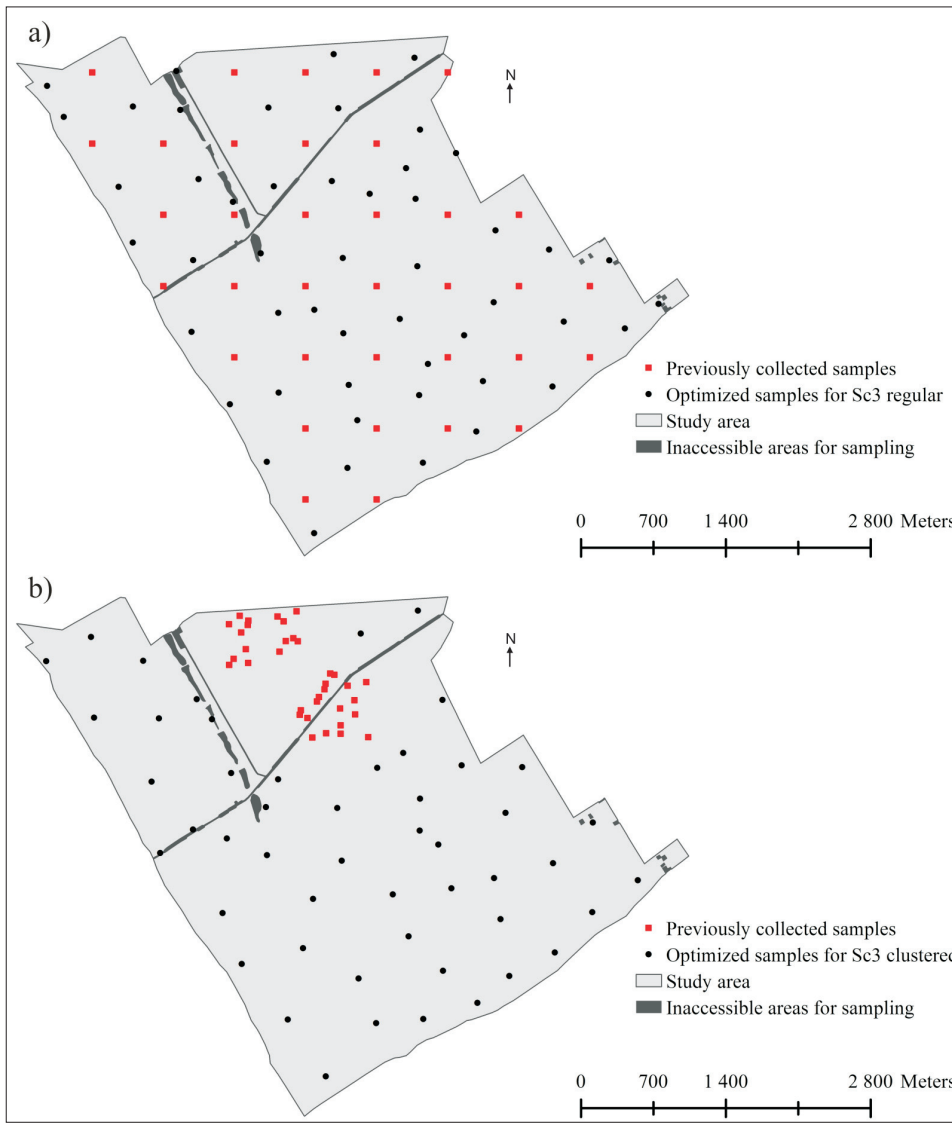

Fig. 5. The optimized sampling designs for scenario 3 regular (a), and scenario 3 clustered (b)

results, thanks to the relatively large range of the variogram model. However, we can state that, the optimized sampling configurations for Sc3 regular and Sc4 regular covered properly the geographic space, so there was no any planned prediction location, which did not have any kriging neighbours. There is an inhibition (i.e. competition) between the sampling points, which causes a quasi-regular point pattern. In case of Sc3 clustered, the optimized sampling design covers properly the geographic space. On other hand, the calculated *F(r)* and *G(r)* functions show a transition between the regular and clustered point pattern types.

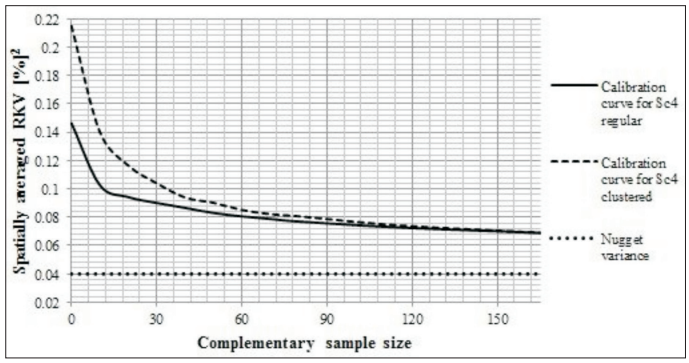

Fig. 6. The calibration curves for scenario 4 regular, scenario 4 clustered and the nugget variance. RKV = regression kriging prediction-error variance

Table 4. The values of the observed significance level of Kolmogorov-Smirnov test calculated for Scenario 3 regular and Scenario 4 regular

Complementary sample size	Covariates*				
	SPC1	SPC2	SPC3	SPC4	SPC5
0	< 0.05				
10	*0.454*	0.035	*0.617*	*0.095*	0.013
20	*0.734*	0.035	*0.617*	*0.358*	*0.082*
30	*0.954*	0.035	*0.617*	*0.194*	*0.329*
40	*0.734*	*0.173*	*0.617*	*0.095*	*0.173*
50	*0.734*	*0.173*	*0.334*	*0.358*	*0.173*
60	*0.954*	*0.329*	*0.617*	*0.095*	*0.329*
70	*0.734*	*0.329*	*0.617*	*0.358*	*0.560*
80	*0.954*	*0.329*	*0.905*	*0.591*	*0.329*
115	*0.954*	*0.560*	*0.905*	*0.358*	*0.560*
165	*0.734*	*0.560*	*0.617*	*0.841*	*0.173*

*The observed significance levels are in italics, where the null hypothesis was accepted at 0.05 significance level

Table 5. The values of the observed significance level of Kolmogorov-Smirnov test calculated for Scenario 3 clustered and Scenario 4 clustered

Complementary sample size	Covariates*				
	SPC1	SPC2	SPC3	SPC4	SPC5
0	< 0.05				
10	0.046	0.000	*0.153*	0.017	*0.082*
20	*0.240*	0.005	*0.617*	*0.095*	*0.082*
30	*0.240*	0.035	*0.334*	0.006	*0.082*
40	*0.954*	*0.329*	*0.617*	*0.095*	*0.560*
50	*0.954*	*0.173*	*0.617*	*0.095*	*0.173*
60	*0.734*	*0.329*	*0.905*	*0.194*	*0.560*
70	*0.954*	*0.329*	*0.905*	*0.194*	*0.560*
80	*0.734*	*0.173*	*0.617*	*0.358*	*0.173*
115	*0.954*	*0.329*	*0.905*	*0.591*	*0.329*
165	*0.734*	*0.560*	*0.617*	*0.841*	*0.173*

*The observed significance levels are in italics, where the null hypothesis was accepted at 0.05 significance level

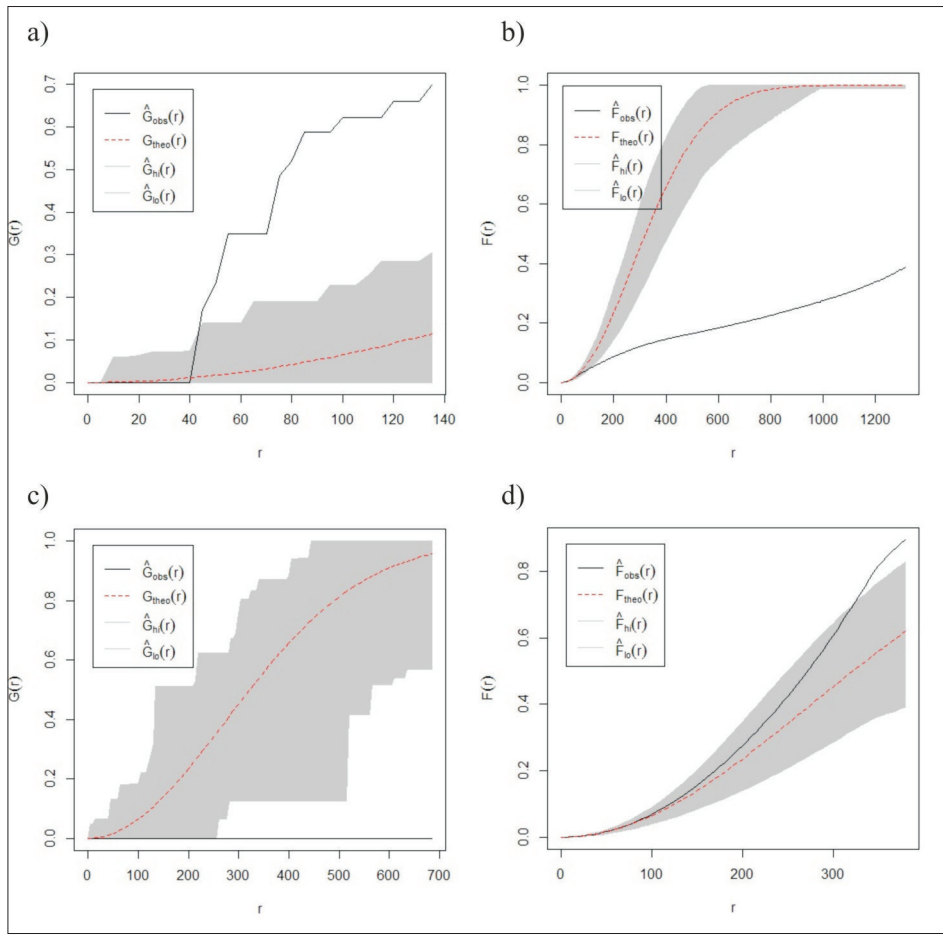

Fig. 7. The observed nearest neighbour distances distribution $G_{obs}(r)$ and empty space $F_{obs}(r)$ function for the previously collected samples in clustered (a–b) and regular (c–d) designs. Abbreviations inside the legend: see *Fig. 4.*

Some thoughts on RKV and SSA

About RKV, we have to notice that, its value(s) mainly characterizes the spatial prediction model rather than the local accuracy of the prediction(s); since it is independent from the observed values (Deutsch, C.V. and Journel, A.G. 1998; Goovaerts, P. 1999; Geiger, J. 2006). We have to consider this fact when we want to use directly its values. However, RKV is a fully suitable measure to compare alternative sampling configuration and to optimize the sampling design for DSM, which follows from its definition (see Eq. [2]).

As we mentioned, the optimized sampling designs for Sc1, Sc2, Sc3 regular and Sc4 regular showed a quasi-regular point pattern. It means that, the variogram model had the dominant influence along these optimization procedures rather than the structure of the regression model, according to Heuvelink, G.B.M. *et al.* (2007). It can be explained by that, the area of interest is fairly homogeneous in point of topography and land use, in other words it has a small "niche" in the feature space, according to Hengl, T. *et al.* (2003). The study site belongs to the Sárbogárd Loess Plateau and only two loess-valleys slice up the area of interest. On

other hand, approx. 85% of the total area is arable (Szatmári, G. and Barta, K. 2012).

There are some limits of SSA using RKV as optimization criterion, e.g. the optimization of sampling design for more than one soil variable. However, it seems to be solved by Szatmári, G. (2014). Another drawback of SSA algorithm is the calculation time, which is lingering. In this study, the elapsed time for a sampling design simulation can take a few hours up to a day. It depends on the settings of the SSA algorithm (initial "system temperature", number of iterations, "cooling" scheme, stopping criterion, etc.), the number of new observations, the size and complexity of the area of interest, the resolution of auxiliary data, the size of matrices for the quality measure calculation (see Eq. [2]), and so forth. We found that, if the maximum of the kriging neighbourhood is restricted to a finite number of observations (according to Webster, R. and Oliver, M.A. 2007, it was set 25, which number of observations is reasonable in point of kriging), then the calculation time decreased significantly.

Conclusions

As it was illustrated by the scenarios, SSA (using RKV as optimization criterion) is a valuable algorithm to optimize soil sampling strategy considering a lot of constraints and demands, which were suggested by soil surveyors, pedometricans and end-users (e.g. the number of new observations, predefined quality measure value (i.e. RKV), as well as taking auxiliary information, previously collected samples and inaccessible areas into account).

RKV is a suitable optimization criterion, because it incorporates the error variance of the trend, as well as the estimation error variance of the residuals, which endeavour SSA to optimize the sampling design both in geographic and feature space. As a consequence, the optimized design absolutely accommodates to the requirements of the RK spatial prediction technique. Therefore, we can assume that the statistical inference (i.e. map of the soil property of interest) is compatible with "the real world". Another beneficial property of RKV is that, it can be calculated before the actual sampling takes place, which can be important in a viewpoint of costs and time. Nevertheless we have to keep in mind that, RKV is independent from the observed values.

The so-called calibration curve can be used to determine the sample size for a given quality measure value and vice versa. As a consequence, this kind of calibration curve is a useful tool to estimate the sample size considering the predefined quality measure value (which is expected to be achieved for the map) and/or the sampling budget's constraints.

***Acknowledgements*:** Our work has been supported by the Hungarian National Scientific Research Foundation (OTKA, Grant No. K105167) and by the TÁMOP-4.1.1.C-12/1/KONV-2012-0012 project (ZENFE).

REREFENCES

Baume, O.P., Gebhardt, A., Gebhardt, C., Heuvelink, G.B.M. and Pilz, J. 2011. Network optimization algorithms and scenarios in the context of automatic mapping. *Computers and Geosciences* 37. 289–294.

Bárdossy, Gy. 1997. Geomatematikai kérdések geológus szemmel (Questions of Geomathematics from the point of view of a geologist). *Magyar Geofizika* 38. (2): 124–141.

Bivand, R.S., Pebesma, E.J. and Gómez-Rubio, V. 2008. *Applied Spatial Data Analysis with R*. New York, Springer, 375 p.

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

Deutsch, C.V. and Journel, A.G. 1998. *GSLIB: Geostatistical Software Library and User's Guide* (2nd Ed.). New York, Oxford University Press, 369 p.

Dobos, E., Michéli, E. and Montanarella, L. 2007. The population of a 500-m resolution soil organic matter spatial information system for Hungary. In *Developments in Soil Science*, Vol. 31. Eds.: Lagacherie, P., McBratney, A.B. and Voltz, M. Amsterdam, Elsevier B.V. 487–495.

Geiger, J. 2006. *Geostatisztika* (Geostatistics). Szeged, University of Szeged, 77 p.

Goovaerts, P. 1999. Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma* 89. 1–45.

Hengl, T. 2009. *A Practical Guide to Geostatistical Mapping*. 2nd Ed. Amsterdam, University of Amsterdam, 291 p.

Hengl, T., Heuvelink, G.B.M. and Rossiter, D.G. 2007. About regression-kriging: from equations to case studies. *Computers and Geosciences* 33. 1301–1315.

Hengl, T., Heuvelink, G.B.M. and Stein, A. 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* 122., 75–93.

Hengl, T., Rossiter, D.G. and Stein, A. 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. *Australian Journal of Soil Research* 41. 1403–1422.

Heuvelink, G.B.M., Brus, D.J. and de Gruijter, J.J. 2007. Optimization of sample configurations for digital mapping of soil properties with universal kriging. In *Developments in Soil Science*, Vol. 31. Eds.: Lagacherie, P., McBratney, A.B. and Voltz, M. Amsterdam, Elsevier B.V. 137–151.

Illés, G., Kovács, G. and Heil, B. 2011. Nagyfelbontású digitális talajtérképezés a Vaskereszt erdőrezervátumban (High resolution digital soil mapping in the Vaskereszt forest reserve). *Erdészettudományi Közlemények* 1. 29–43.

McBratney, A.B., Mendonca Santos, M.L. and Minasny, B. 2003. On digital soil mapping. *Geoderma* 117. 3–52.

Melles, S.J., Heuvelink, G.B.M., Twenhöfel, C.J.W., van Dijk, A., Hiemstra, P.H., Baume, O. and Stöhlker, U. 2011. Optimizing the spatial pattern of networks for monitoring radioactive releases. *Computer and Geosciences* 37. 280–288.

Minasny, B. and McBratney, A.B. 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. *Geoderma* 140. 324–336.

Pásztor, L., Szabó, J., Bakacsi, Zs., Laborczi, A., Dobos, E., Illés, G. and Szatmári, G. 2014. Elaboration of novel, countrywide maps for the satisfaction of recent demands on spatial, soil related information in Hungary. In *Global Soil Map: Basis of the Global Spatial Soil Information System*. Eds.: Arrouays, D. *et al.* London, Taylor &Francis Group, 207–212.

Szatmári, G. 2014. Optimization of sampling configuration by spatial simulated annealing for mapping soil variables. In *6th Croatian–Hungarian and 17th Hungarian Geomathematical Congress: "Geomathematics – from theory to practice"*. Eds.: Cvetković, M., Novak Zelenika, K. and Geiger, J., Zagreb. Croatian Geological Society, 105–111.

Szatmári, G. and Barta, K. 2012. Az erózió, az erózió-veszélyeztetettség és a területhasznosítás kapcsolata mezőföldi területeten (Relationship between water erosion, potential erosion and land use on an area in the Mezőföld region). *Agrokémia és Talajtan* 61. (1): 41–56.

Szatmári, G. and Barta, K. 2013. Csernozjom talajok szervesanyag-tartalmának digitális térképezése erózióval veszélyeztetett mezőföldi területen (Digital mapping of the organic matter content of chernozem soils on an area endangered by erosion in the Mezőföld region). *Agrokémia és Talajtan* 62. (1): 47–60.

van Groenigen, J.W. and Stein, A. 1998. Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality* 27. 1078–1086.

van Groenigen, J.W., Siderius, W. and Stein, A. 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma* 87. 239–259.

Wang, J.-F., Stein, A., Gao, B.-B. and Ge, Y. 2012. A review of spatial sampling. *Spatial Statistics* 2. 1–14.

Webster, R. and Oliver, M.A. 2007. *Geostatistics for Environmental Scientists* 2nd Ed. Chichester, Wiley, 330 p.

Wischmeier, W.H. and Smith, D.D. 1978. *Predicting rainfall erosion losses: A guide to conservation planning*. Washington D.C., U.S. Government Printing Office, 58 p.