UNIVERSITY OF SZEGED
Faculty of Science and Informatics
Doctoral School of Environmental Sciences
Department of Physical Geography and Geoinformatics

ELABORATION OF A DIGITAL SOIL MAPPING AND SAMPLING OPTIMIZATION METHODOLOGY BASED ON GEOSTATISTICAL APPROACH AND ITS APPLICATION IN DIFFERENT SCALES

Theses of Ph.D. Dissertation

GÁBOR SZATMÁRI

Supervisor:
Dr. Károly Barta
associate professor

Advisor:
Dr. László Pásztor
senior research fellow

Szeged, 2017

„The soil varies from place to place,
and many of its properties vary in time too.
This is what makes the soil so fascinating"

HEUVELINK & WEBSTER
(2001)[*]

INTRODUCTION AND AIMS OF THE RESEARCH

Demands on soil related information have been significant worldwide and are still increasing. However, the recognition of soil multifunctionality and the decreasing of the available resources for soil surveys have occurred at the same time. Hence, soil surveyors strain after the exploitations of legacy soil information, as well as minimization/optimization of field surveys. In Hungary, demands on soil related information is satisfied by the available spatial soil information systems. Soil data of these spatial information systems are related to point observations, which means that the specific demands of end-users on soil information can be satisfied by the regionalization of these point observations. This regionalization can be carried out the most effectively by geostatistics.

Nowadays, digital soil mapping and sampling optimization highly rely on geostatistical approaches, which were implemented successfully by soil surveyors in Hungary as well. However, the nature of demands on soil related information has changed too, which meets rarely with the available soil information and the expected results of the conventional geostatistical methods. Common request is, for example, to provide the error or the spatial uncertainty of the soil property map(s), which are indispensable to make a valid decision. Nevertheless, the recently applied, conventional geostatistical methods are not (or just partly) able to fulfill these requirements. Hence, the adaptation of those geostatistical methods, which are able to fulfill the above mentioned requirements, is highly important. The first step of the adaptation is to summarize the theoretical considerations related to geostatistics in a coherent and consequent methodology. Accordingly, the aims of my doctoral research are as follows:

- Elaboration of a coherent, digital soil mapping and sampling optimization methodology based on geostatistical approaches, which:
  - provides theoretical as well as practical framework for soil sampling optimization, digital soil mapping, spatial uncertainty modelling, and the assessment of the resulted map(s),

[*] Heuvelink, G.B.M., Webster, R., 2001. Modelling soil variation: past, present and future. Geoderma 100, 269–301.

  - has consequent and compatible components,
  - can be applied to map soil properties in different scales,
  - can take the available secondary information into account in the course of spatial prediction, spatial uncertainty modelling, as well as sampling optimization,
  - has an essential part to determine the error of the map using different measures, as well as to model the spatial uncertainty.
  - Furthermore, the components of the methodology is exchangeable in accordance with the adopted random function model.
- The elaborated methodology would be implemented with free and open source software.
- Application of regression kriging in different scales. Assessment of the error of the created maps.
- Application of sequential stochastic simulation method based on regression kriging to model the spatial uncertainty.
- Application of spatial simulated annealing (SSA) sampling optimization algorithm.
- Multivariate generalization of the univariate SSA algorithm and its application to multivariate soil sampling optimization.

The elaborated methodology has been applied and evaluated on three Hungarian pilot areas. The selected study areas show significant differences in their extensions (scales), as well as the affecting soil forming processes. The study areas are: (1) sub-catchment of Szálka, (2) arable lands of Előszállás and (3) Zala County. I have selected the soil organic matter content to test the methodology, which soil property has a great importance not only for soil science, but for earth and environmental sciences as well. It is important to remark that the methodology is suited for the treatment of other soil related properties.

THEORETICAL CONSIDERATIONS

The soil properties and the related phenomena are complex and complicated results of physical, chemical and biological processes in combination. Nevertheless, our knowledge about the spatial and temporal variability of these processes is fragmentary. As a consequence, the soil appears to us as if it were an outcome of a random process. Hence, the soil can be approached and modelled as a stochastic process. According to this, locally the point values of the soil property are considered as random variables, which variables are not independent but are related by a correlation expressing the spatial structure of the phenomenon. These (soil) properties are referred to as regionalized variables by geostatistics.

Based on this theoretical consideration, the elaborated digital soil mapping and sampling optimization methodology regards explicitly the soil properties as regionalized variables, which means the basis of sampling optimization, spatial prediction, as well as spatial uncertainty modelling.

MATERIALS AND METHODS

Soil databases and secondary information

Legacy soil databases were available from the selected, three Hungarian study sites. I have selected the soil organic matter content as target variable to test the elaborated methodology. My doctoral research did not involve new, supplementary soil survey and sampling. The applied secondary information for the selected study areas were derived from their digital elevation models. Furthermore, the land cover maps of Előszállás and Szálka were applied as categorical secondary information, which were compiled interpreting the products of the official aerial photography campaign of Hungary, taken in 2005. In case of Zala County, the Digital Kreybig Soil Information System's mapping units of soil water management were applied as categorical secondary information.

Mapping methodology

First of all, exploratory data analysis was carried out for the raw soil organic matter data. I applied methods, which are able to take explicitly the geographical positions of data points into account. Furthermore, I assessed the relationship between data points and the available secondary information.

I have modelled the spatial distribution of soil organic matter content with regression kriging for each study areas, which spatial prediction method combines the regression of the target soil variable on spatially exhaustive secondary information with simple kriging of the regression residuals to predict the value of the target soil variable at unvisited location:

$$Z^*_{\mathrm{RK}}(\mathbf{u}) = \mathbf{q}^{\mathrm{T}}(\mathbf{u}) \cdot \boldsymbol{\beta}_{\mathrm{GLS}} + \boldsymbol{\lambda}^{\mathrm{T}}_{SK}(\mathbf{u}) \cdot (\mathbf{z} - \mathbf{Q} \cdot \boldsymbol{\beta}_{\mathrm{GLS}}), \quad (1)$$

where $\mathbf{q}(\mathbf{u})$ is the vector of secondary information at $\mathbf{u}$ location, $\boldsymbol{\beta}_{\mathrm{GLS}}$ is the vector pf regression coefficients, $\boldsymbol{\lambda}_{\mathrm{SK}}(\mathbf{u})$ is the vector of simple kriging weights (assigned to the regression's residuals) at $\mathbf{u}$ location, $\mathbf{z}$ is the vector of data points and $\mathbf{Q}$ is the matrix of the secondary information at the sampling locations. I have estimated the regression coefficients using the generalized least squares (GLS) method because it is able to take the covariance matrix of the residuals into account. Furthermore, I applied principal component analysis in order to decrease the multicollinearity. I have created soil organic

matter map for each study sites using the spatial predictions provided by regression kriging.

The soil organic matter maps were validated using independent control points, except in case of Szálka, where the leave-one-out cross validation (LOOCV) was used because of the limited number of data points. The following error measures were calculated: mean error, mean absolute error and the value of root mean square error standardized by kriging variance.

I have applied the sequential stochastic simulation method based on regression kriging to model the spatial uncertainty in Előszállás. As opposed to any kind of kriging techniques, the main aim of stochastic simulation is to generate alternative, but equally probable realizations of the regionalized variable, which realizations reproduce the model statistics. Hence, stochastic simulations model the "reality" in a certain global (and not local) sense, which gives the opportunity to model the spatial uncertainty. Using these uncertainty models, we can support such kind of decision-making, where the kriging techniques would be inadequate.

Model based soil sampling optimization methodology

The elaborated, model based sampling optimization methodology is based on the spatial simulated annealing (SSA) algorithm, which 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. In consideration of the requirements of the mapping methodology, I have selected the regression kriging variance as the quality measure of optimization, which is:

$$\sigma^2_{\mathrm{RK}}(\mathbf{u}) = C(0) - \mathbf{c}^{\mathrm{T}} \cdot \boldsymbol{\lambda}_{\mathrm{SK}}(\mathbf{u}) + [\mathbf{q}(\mathbf{u}) - \mathbf{Q}^{\mathrm{T}} \cdot \boldsymbol{\lambda}_{\mathrm{SK}}(\mathbf{u})]^{\mathrm{T}} \cdot (\mathbf{Q}^{\mathrm{T}} \cdot \mathbf{C}_{\mathrm{R}}^{-1} \cdot \mathbf{Q})^{-1} \cdot \cdot [\mathbf{q}(\mathbf{u}) - \mathbf{Q}^{\mathrm{T}} \cdot \boldsymbol{\lambda}_{\mathrm{SK}}(\mathbf{u})] \quad (2)$$

where, *C(0)* is the variance of the residuals, **c** is the vector of covariances between the residuals at the observed and unvisited locations and $\mathbf{C}_{\mathrm{R}}$ is covariance matrix of the residuals. It incorporates both the kriging variance of the residuals (first two terms on the right-hand side of Eq.2.) and the error variance of the trend (third term on the right-hand side of Eq.2.). I have applied the spatially averaged regression kriging variance as the objective function of optimization.

One of the most relevant limitations of the SSA algorithm is that it is able to optimize the sampling design only for one soil variable. Hence, I have elaborated a multivariate generalization of the univariate SSA algorithm. In this case, I have selected also the regression kriging variance as the quality measure of optimization, which can be derived from the structures of the

regression models and the variogram of the dominant soil property. I have applied the spatially averaged regression kriging variance as the objective function of optimization.

I have set 8 soil sampling scenarios in Előszállás in order to test and evaluate the elaborated, model based sampling optimization methodology. The scenarios were set to represent the capabilities of the methodology. I set four univariate scenarios for hypothetical mapping of soil organic matter content. I set two multivariate scenarios in order to test the extended SSA method. In this case the soil survey's hypothetical aim was to map soil organic matter content together with rooting depth. I set also two scenarios for that cases, when we adopted different models to the random function. In this scenarios, ordinary kriging and multiple linear regression was adopted as spatial prediction techniques, respectively. The goal of the scenarios was to optimize the soil sampling to these prediction techniques. I have selected the ordinary kriging variance and the regression error variance as the quality measure of optimization, respectively. In the course of scenarios, I have considered numerous sampling constraints, such as the number of new observations, previously collected sampling points, inaccessible or irrelevant areas for sampling, available secondary information and irregular shape of the study area.

I have evaluated the optimized sampling configurations by the empty space function, the maps of kriging neighbourhood and the Kolmogorov-Smirnov test. I examined that the optimized sampling designs cover properly the geographic and/or feature spaces.

SUMMARY OF RESULTS, THESES

1. **The spatial variability of soil properties can be modelled effectively, if we regard them as regionalized variables.**

   In the course of my research, I have pointed out that soil properties are complicated and complex resultants of physical, chemical and biological processes in combination. Nevertheless, our knowledge about the spatial and temporal variability of these processes is fragmentary. As a consequence, the soil appears to us as if it were an outcome of a random process. Hence, the soil can be approached and modelled as a stochastic process. According to these, locally the point values of the soil property are considered as random variables, which variables are not independent but are related by a correlation expressing the spatial structure of the phenomenon. These (soil) properties are referred to as regionalized variables by geostatistics.

2. **The elaborated, digital soil mapping and sampling optimization methodology can be implemented and improved with free and open source software (FOSS).**
    a. I have applied SAGA GIS to prepare the available, continuous and categorical secondary information. Furthermore, I have applied R software environment to geomathematical, as well as geostatistical analyses and modelling. I have presented the results of my research (e.g. figures, histograms, scatterplots, indicator maps, realizations, sampling configurations, calibration curves) with R, too. Only in the case of cartographic design of soil organic matter maps, commercial software (i.e. ArcGIS) was applied.
    b. I have improved and elaborated several methods using R software environment, which methods act on the elaborated methodology. These methods are the following: (1) modified empty space function, (2) mapping of kriging neighbourhoods, (3) sequential stochastic simulation based on regression kriging and (4) multivariate sampling design optimization method.
    c. Furthermore, I have pointed out that FOSS environment provides a good possibility to improve further on the elaborated, digital soil mapping and sampling optimization methodology.

3. **Regression kriging is a suitable spatial prediction method for modelling the spatial distribution of continuous soil properties in different scales. However, one has to keep in mind that regression kriging technique supposes the linear relationship between the target soil properties and the available secondary information.**
   I have evaluated the resulted soil organic matter maps by expert knowledge. I was led to the conclusion the resulted maps represented properly the spatial distribution of the soil organic matter, respectively. The calculated error measures (i.e. mean error, mean absolute error and the value of root mean square error standardized by kriging variance) approached their expectations, which means also that the created soil organic matter maps represented properly the spatial variability of the target soil property. In case of Zala County, I have revealed that in areas with extremely high organic matter content (>10%) regression kriging underestimated the soil organic matter content. This can be attributed to the fact soil organic matter and its content is highly affected by various, complex, nonlinear soil forming processes.

4. **In digital soil mapping, the sequential stochastic simulation method based on regression is suitable for modelling spatial uncertainty. The**

generated realizations reproduce the applied variogram model of the regression residuals. The realizations provide a visual and quantitative measure of spatial uncertainty.

a. I have generated 100 alternative and equally probable realizations by the sequential stochastic simulation method based on regression kriging in Előszállás. In the course of stochastic simulation, the simple kriging variance was applied to identify the variance of the conditional cumulative distribution function for each grid point. According to the results, the realizations reproduced the applied variogram model of the regression residuals.
b. The difference between the generated 100 realizations represents the uncertainty of soil organic matter content mapping. Based on the generated realizations, I can stated that the differences between the realizations in steeper slopes are more pronounced, than in plain areas.
c. I have applied the generated realizations to calculate the cumulative distribution around an infinitesimally small neighbourhood of each grid point. Using these distributions, I have derived the following maps: (1) E-type estimation of soil organic matter content and (2) the corresponding 95% confidence interval. I have demonstrated that confidence interval's width provides direct information about the uncertainty of the E-tpye estimation of soil organic matter content.

5. **The calculated cumulative distribution for each grid point can be applied to derive such goal oriented digital soil maps, which can be used effectively in decision-making. For this purposes, the conventional kriging techniques are inadequate.**

In Előszállás, the calculated cumulative distribution for each grid point was applied to create four probability maps. The created maps represent spatially the probability of several statistical events, such as what is the probability that soil organic matter content is higher than 2%, but lower than or equal to 3%. I have pointed out that the conventional kriging techniques cannot answer this goal oriented "task" because of their designs. Based on the resulted probability maps, I was led to the conclusion, probability maps – which can be considered as goal oriented digital soil maps – can be applied to satisfy certain demands on soil related information. Furthermore, they can be used to support effectively decision-making.

6. **In general, the regression kriging variance is not a measure of local prediction accuracy. However, it is fully suitable for ranking alternative, geometric data configuration, which can be applied to soil sampling optimization. Since it is independent from the observed values, it can be calculated before the actual sampling takes place, which is advantageous in point of costs and time.**
    a. Regression kriging is the best linear unbiased estimator (BLUE), which is "best" in that sense that it minimizes the error variance of the prediction. Nevertheless, the regression kriging variance is independent from the observed values. Hence, it can be applied to measure the local prediction accuracy only if (1) the errors can be modelled as a realization of a Gaussian random variable and (2) the error variance is independent from the data points. As a consequence, the regression kriging variance cannot be considered as universal measure of local prediction accuracy.
    b. I pointed out that regression kriging variance is suitable for ranking alternative, geometric data configuration, which follows from its definition. Furthermore, it is able to take simultaneously the coverage of geographic and feature spaces into account, which is advantageous in point of sampling optimization.
    c. As it follows from its definition, the regression kriging variance is independent from the observed values. This means that it can be determined/estimated before the actual sampling takes place. This can be considered as advantageous property in point of costs and time. This beneficial property was used in the course of sampling scenarios.

7. **The elaborated, model based sampling optimization methodology using the regression kriging variance as quality measure is suitable for optimizing soil sampling designs. The methodology is suitable for considering numerous sampling constraints and demands. Furthermore, it supports the best the digital soil mapping methodology, which is based on regression kriging.**
    a. According to the results of modified empty space functions, kriging neighbourhoods' maps and the Kolmogorov-Smirnov test, the optimized sampling configurations for the scenarios (1^{st}, 2^{nd}, 3^{rd} and 4^{th}) covered properly both geographic and feature space. Furthermore, the calculated calibration curves (2^{nd} and 4^{th} scenarios) can be used to determine those minimal sample sizes, which are necessity to cover properly both geographic and feature space (for given sampling constraints).
    b. The sampling scenarios have revealed that elaborated soil sampling methodology is able to consider numerous sampling

constraints and demands in the course of optimization, such as the number of new observations, previously collected sampling points, inaccessible or irrelevant areas for sampling, the available secondary information, the predefined quality measure value, as well as the irregular shape of the study area.

c. Moreover, the optimized scenarios have revealed that the regression kriging variance (as quality measure) endeavors SSA to optimize simultaneously the sampling design in both geographic and feature space considering the predefined sampling constraints. This satisfies the requirements of the digital soil mapping methodology, which is based on regression kriging. Consequently, the elaborated, model based sampling optimization methodology supports the digital mapping methodology.

8. **The elaborated soil sampling optimization methodology is suitable for optimizing sampling designs if other model was adopted to the random function. This means that the methodology is flexible.**
   Two sampling scenarios were set to test the sampling methodology in that case if other models were adopted to the random function. The results have revealed that the methodology is flexible considering the adopted models.
   a. The ordinary kriging variance (as quality measure) endeavored SSA to optimize the sampling configuration in geographic space (7^{th} scenario). The results of the modified empty space function and the kriging neighbourhoods map pointed out that the optimized sampling design shows a regular configuration, which is adequate for ordinary kriging. The optimized sampling points covered properly the geographic space.
   b. The regression error variance (as quality measure) endeavored SSA to optimize the sampling configuration in feature space (8^{th} scenario). The results of the modified empty space function pointed out that the optimized sampling design shows a clustered configuration. According to the results of the Kolomogorov-Smirnov test, the optimized sampling points covered properly the feature space, which supports the multiple linear regression.

9. **The extended SSA algorithm is suitable for optimizing simultaneously the sampling designs for the target soil properties.**
   The results of multivariate sampling scenarios (5^{th} and 6^{th}) have revealed that the extended algorithm was able to optimize the sampling design for the target soil variables (i.e. soil organic matter content and rooting depth). According to the results of modified empty space functions, kriging

neighbourhoods' maps and the Kolmogorov-Smirnov test, I can state that the optimized sampling configurations of the multivariate scenarios covered properly both geographic and feature space. Furthermore, the results of the scenarios pointed out that the extended algorithm retained the advantageous properties of univariate SSA, which means that it is able to consider numerous sampling constraints.

LIST OF PUBLICATIONS RELATED TO THE THESES

Szatmári, G., 2013. High-resolution mapping of soil organic matter content based on Regression Kriging in a study area endangered by erosion in Hungary. In: Horváth, J. et al. (Eds.) XVI. Congress of Hungarian Geomathematics and V. Congress of Croatian and Hungarian Geomathematics. Hungarian Geological Society, Mórahalom. pp. 76–79. (ISBN:978-963-8221-49-0)

Szatmári, G., Barta, K., 2013. Csernozjom talajok szervesanyag-tartalmának digitális térképezése erózióval veszélyeztetett mezőföldi területen. *Agrokémia és Talajtan* 62(1), 47–60.

Szatmári, G., Laborczi, A., Illés, G., Pásztor, L., 2013. A talajok szervesanyag-készletének nagyléptékű térképezése regresszió krigeléssel Zala megye példáján. *Agrokémia és Talajtan* 62(2), 219–234.

Szatmári, G., 2014. Optimization of sampling configuration by spatial simulated annealing for mapping soil variables. In: Cvetkovic, M. et al. (Eds.) XVII. Congress of Hungarian Geomathematics and VI. Congress of Croatian and Hungarian Geomathematics. Croatian Geological Society, Zagreb. pp. 105–111. (ISBN:978-953-95130-8-3)

Szatmári, G., 2015. Using a sequential stochastic simulation approach based on regression kriging to generate functional soil maps. In: Horváth, J. et al. (Eds) XVIII. Congress of Hungarian Geomathematics and VII. Congress of Croatian and Hungarian Geomathematics. Hungarian Geological Society, Mórahalom. pp. 134–140. (ISBN:978-963-8221-58-2)

Szatmári, G., Barta, K., Farsang, A., Pásztor, L., 2015. Testing a sequential stochastic simulation method based on regression kriging in a catchment area in Southern Hungary. *Geologia Croatica* 68(3), 273–283. (**IF: 0,625 [2015]**)

Szatmári, G., Barta, K., Pásztor, L., 2015. An application of a spatial simulated annealing sampling optimization algorithm to support digital soil mapping. *Hungarian Geographical Bulletin* 64(1), 35–48.

Szatmári, G., Barta, K., Pásztor, L., 2016. Multivariate Sampling Design Optimization for Digital Soil Mapping. In: Zhang, G.-L. et al. (Eds.) Digital Soil Mapping Across Paradigms, Scales and Boundaries. Springer-Verlag, Singapore. pp. 77–87. (ISBN:978-981-10-0414-8)

Szatmári, G., Pásztor, L., 2016. Geostatisztika a talajtérképezésben: Szemle. *Agrokémia és Talajtan* 65 (1), 95–114.

Tóth, G., Hermann, T., **Szatmári, G.**, Pásztor, L., 2016. Maps of heavy metals in the soils of the European Union and proposed priority areas for detailed assessment. *Science of the Total Environment* 565, 1054–1062. (**IF: 3,976 [2015/2016]**)