Testing a sequential stochastic simulation method based on regression kriging in a catchment area in Southern Hungary

Gábor Szatmári[1], Károly Barta[1], Andrea Farsang[1] and László Pásztor[2]

[1] Department of Physical Geography and Geoinformatics, Faculty of Science and Informatics, University of Szeged, Egyetem u. 2., H-6722 Szeged, Hungary; (szatmari.gabor.88@gmail.com; barta@geo.u-szeged.hu; farsang@geo.u-szeged.hu)

[2] Department of Environmental Informatics, Institute for Soil Sciences and Agricultural Chemistry, Hungarian Academy of Sciences, Herman Ottó út 15., H-1022 Budapest, Hungary; (pasztor@rissac.hu)

Geologia Croatica

ABSTRACT

Modelling spatial variability and uncertainty is a highly challenging subject in soil- and geosciences. Regression kriging (RK) has several advantages; nevertheless it is not able to model the spatial uncertainty of the target variable. The main aim of this study is to present and test a sequential stochastic simulation approach based on regression kriging (SSSRK), which can be used to generate alternative and equally probable realizations in order to model the spatial variability and uncertainty of the target variable; meanwhile the advantages of the RK technique are retained. The SSSRK method was tested in a sub-catchment area of the Lajvér stream, in Southern Hungary for the high resolution modelling (i.e. 10 metre grid spacing) of the spatial distribution of soil organic matter (SOM). In the first step, secondary information was derived according to the soil-forming factors; then the RK system was built up, which provides the base of SSSRK. 100 realizations were generated, which reproduced the model statistics and honoured the input dataset. These realizations provide 100 simulated values for each grid node, which is an appropriate number for calculating the cumulative distributions for each grid node. Using these cumulative distributions the following maps were derived: the map of the E-type estimation, the corresponding 95% confidence interval width's map and the map of the probability of the event of {SOM < 1.5%}. The latter map is highly informative in soil protection and management planning. The resulting model and maps showed that, SSSRK is a valuable technique to model and assess the spatial variability and uncertainty of the target variable. Furthermore, the comparison of RK and SSSRK showed that the SSSRK's E-type estimation and the RK estimation gave almost the same results due to the fairly high R^2 value of the regression model (R^2=0.809), which decreased the smoothing effect.

Keywords: geostatistics, regression kriging, sequential stochastic simulation, spatial uncertainty, soil organic matter

1. INTRODUCTION

Modelling the spatial distribution, variability and uncertainty of soil related attribute(s) (e.g. soil organic matter content, rooting depth, pH, particle size distribution, bulk density) is a challenging subject in the soil- and geosciences, as well as in general environmental research. The resulting model(s) can be applied to support various soil and environmental related decisions, such as delineation of contaminated or endangered zones, estimation of remediation costs, identification areas for fertilization or crop growth and so forth. Geostatistics, which can be regarded as a subset of sta-

tistics specialized in analysis and interpretation of geographically referenced data (GOOVAERTS, 1997), provide a huge amount of tools to support these decisions (WEBSTER & OLIVER, 2007).

A majority of geographically referenced data (e.g. soil attributes) is related to discrete (sampling) points in the geographic space, which means that we do not have any information about these variables at the unvisited locations. However, an increasing amount of spatially exhaustive secondary information (e.g. digital elevation models, satellite images, geological maps and land cover maps) is available today with increasing spatial and temporal resolution, which can be used in combination with geostatistical tools to satisfy certain requirements (e.g. intrinsic hypothesis, second-order or weak stationarity), and to improve the estimation or simulation models (MINASNY & MCBRATNEY, 2007). **Regression kriging (RK)** is a representative and widely used technique, which combines the regression of the target variable on spatially exhaustive secondary information with simple kriging of the regression residuals to estimate the value of the target variable at an unvisited location (HENGL et al., 2004). RK has several advantages in contrast to other kriging methods, e.g. it can take spatially exhaustive secondary information into account in the estimation process, it can handle the trend (or drift), where the trend term means that the local mean systematically varies from place to place. Furthermore, RK is more flexible than kriging with external drift or cokriging methods (SIMBAHAN et al., 2006; ELDEIRY & GARCIA, 2010; HENGL et al., 2003), which also can take secondary information into account in the estimation process. One of the main drawbacks of RK is that it is unable to model the spatial uncertainty or to provide, for example, the 95% confidence interval for the estimates. According to GOOVAERTS (1997), if the intrinsic hypothesis holds, the kriging variance could be used to derive the 95% confidence interval for the estimates. Unfortunately, the intrinsic hypothesis is not reasonable to hold in many cases (there is a trend, which makes it unacceptable). The main reason why RK is widely used for spatial modeling is that it can handle the trend (SZATMÁRI & BARTA, 2013).

During past decades, stochastic simulations became widespread in the soil- and geosciences to model and assess the spatial variability and uncertainty of the target variable(s) (GOOVAERTS, 1997; DEUTSCH & JOURNEL, 1998; GEIGER, 2006; MALVIĆ, 2008; NOVAK ZELENIKA & MALVIĆ, 2011; GEIGER, 2012; MALVIĆ et al., 2012; NOVAK ZELENIKA et al., 2012). As opposed to any kind of kriging techniques, the main aim of simulation methods is to generate alternative and equally probable realizations, which reproduce the model statistics (e.g. histogram and variogram model), rather than to minimize the local error variance Var{Z*(**u**) – Z(**u**)}. Hence, simulation methods model the "reality" in a certain global (and not local!) sense, which give an opportunity to model the spatial uncertainty (GOOVAERTS, 1997; DEUTSCH & JOURNEL, 1998; GEIGER, 2006; MALVIĆ, 2008). Based on this, we can classify the geostatistical techniques into two classes: estimation and simulation methods. MALVIĆ (2008) also used this classi-

fication but they were referred to as "deterministic" and "stochastic" approaches.

The main aim of this study is to present and test a **sequential stochastic simulation method** based on **regression kriging (SSSRK)**, which is able to generate alternative and equally probable realizations (in order to model the spatial uncertainty) with the constraint that they have to reproduce the model statistics; meanwhile the advantages of RK are retained. SSSRK is presented and tested in a sub-catchment area of the Lajvér stream in southern Hungary, where former soil (water) erosion research has resulted in particular soil sampling and laboratory analysis. The study site is a good example from the point of view of soil science because of the heterogeneous landscapes, where the various effects of soil-forming factors and soil erosion, as well as the various land cover types diversify the sub-catchment soil pattern. Soil organic matter (SOM) content was chosen as a sample variable, being a soil attribute, which has an important role due to its multipurpose functionality in the soil- and geosciences, as well as in environmental research. The goal is to build up the high resolution model (i.e. 10 metre grid spacing) of the SOM spatial distribution based on SSSRK. The resulting model and the derived "maps" are of interest for precision agriculture, water erosion and soil protection research, small scale landscape planning and evaluation, integrated catchment management and climate change research (sources and sinks for atmospheric carbon dioxide).

2. STUDY SITE

The sub-catchment (area is approximately 1.32 km^2) which drains into the Lajvér stream is located in the southern part of Hungary, in the Szekszárd Hills, near the village of Szálka (Fig. 1). The area of interest is covered by loess-like sediments (DÖVÉNYI, 2010). The annual precipitation is 650 mm in the study site. The original soil types are Cambisols and Luvisols with a loamy soil texture, but there are several eroded types of them because of the high relief and the long-term agricultural land use. Even Regosols can be observed in small areas. However, the eroded soil material forms Fluvisols on the valley bottoms. Land use more or less conforms to the relief conditions: approximately 50% of the area (65.2 hectares) is used as arable land (see Fig. 1), but the steeper slopes are covered by meadows and forests. These latter ones mean mainly acacia but we can see oak forests on the northern part of the sub-catchment. There are new vineyards on the southeastern parts (see Fig. 1).

3. MATERIALS AND METHODS

3.1. Sampling, laboratory measurements and preliminary data analysis

Former soil (water) erosion research has resulted in a particular pattern of topsoil (0-10 cm depth) sampling and laboratory analyses including SOM. The database contains 47 records on SOM content originating from 2 soil profiles and 45 boreholes (Fig. 1), sampled in 2008-2009.
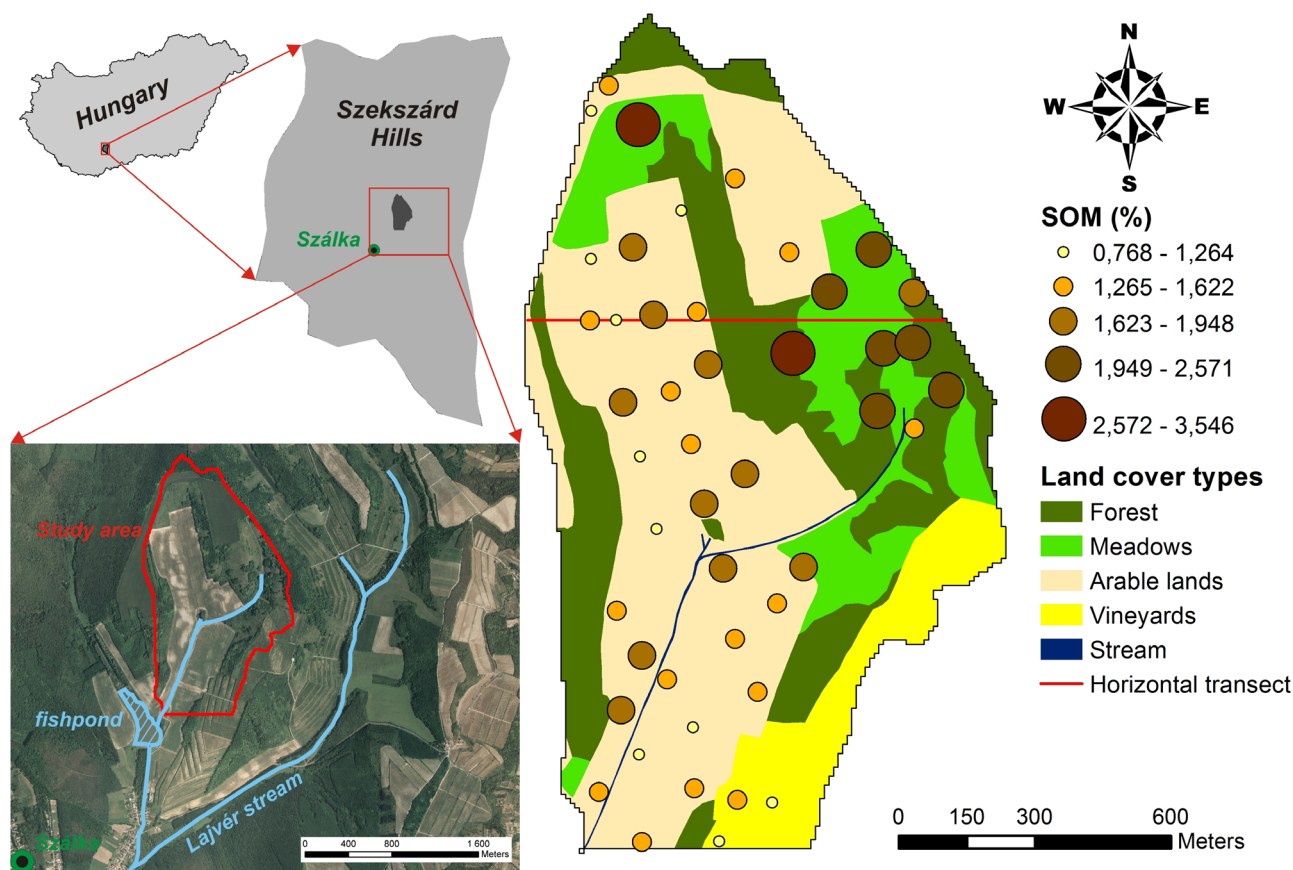
Szatmári et al.: Testing a sequential stochastic simulation method based on regression kriging in a catchment area in Southern Hungary

275

Geologia Croatica

Figure 1: The location of the study site in Hungary and its land cover presented with the measured soil organic matter (SOM) data at the sampling points.

The topsoil sampling design was planned for modelling the soil erosion process (BORCSIK et al., 2011). A stratified sampling strategy (supported with geostatistical considerations) was used for this purpose to determine the spatial variability of the eroded and accumulated soil patterns. The land cover (LC) map and the steepness (slope from the digital elevation model; see. Fig. 3) of the study area was the basis for the stratification. Arable lands are the most seriously affected LC type by soil erosion. Hence, this LC type (including the eroded and accumulated soil patterns) had more weights in the sampling strategy than forests and meadows, because the previous LC types are not affected by this soil degradation process (BORCSIK et al., 2011). As a consequence, arable lands (including the eroded and accumulated soil patterns, as well as the flat valley bottom) are relatively overrepresented; whilst forest and meadows are relatively underrepresented. According to WEBSTER & OLIVER (2007), geostatistical considerations were applied in the sampling strategy to make sure that the sampling design covered the study site uniformly, as far as possible. The soil profiles were excavated; their total depth was 140 cm, where the parent material (loess-like sediments) was reached. The topsoil layer (0-10 cm depth) of the profiles was sampled. A gouge auger was used to excavate the boreholes with an average total depth of 120 cm. The total depths of the boreholes were determined by the depth of the parent material. The samples were collected from the topsoil layer (0-10 cm depth) of the

boreholes. The soil profiles and the boreholes were used to characterize the main soil types, as well as the soil erosion process.

The laboratory analyses included the determination of SOM, particle size distribution, pH, bulk density, as well as the carbonate content (BORCSIK et al., 2011). This study has used the SOM measurement data according to the Hungarian Standard (MSZ 21470-52:1983), which means that SOM content was determined after sulfuric acid digestion in the presence of 0.33 mol/dm^3 potassium dichromate by spectrophotometer (type: Helios-gamma).

Table 1: Summary statistics of the soil organic matter (SOM) data excluding outliers.

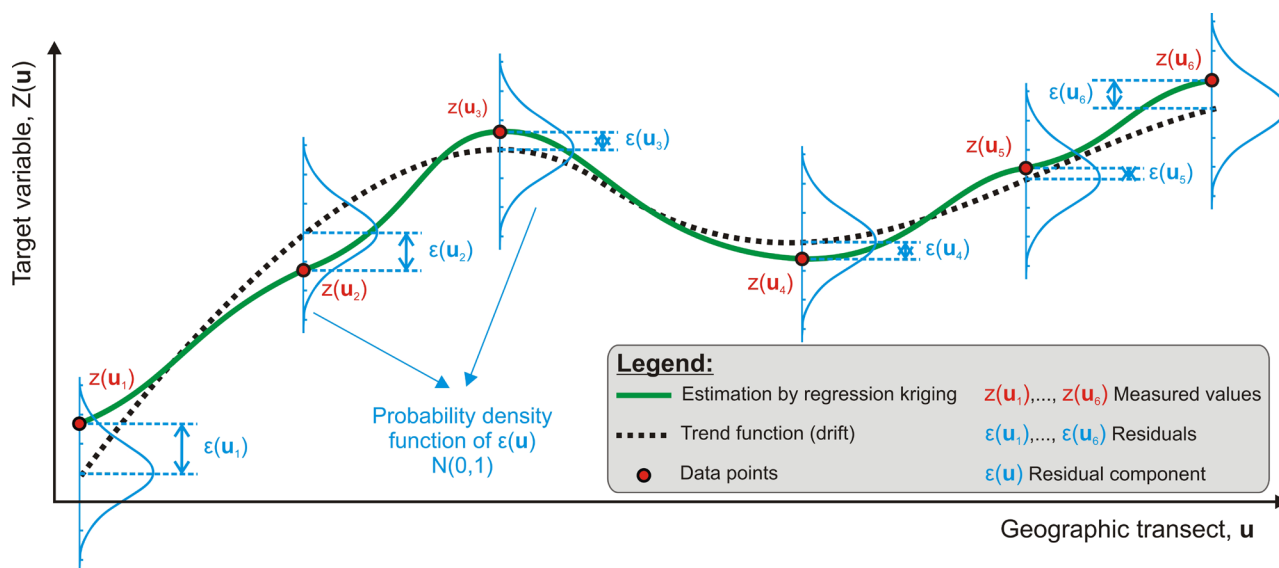| Statistics	Value
Mean	1.600
Median	1.531
Std. dev.	0.502
Minimum	0.768
Maximum	3.546
Skewness	0.603
Kurtosis	0.141

Figure 2: The schema of regression kriging.

Only one outlier was identified by Box and Whisker plot and then it was removed from the raw SOM data. Then summary statistics were calculated for the filtered SOM data (Table 1). Figure 1 presents the spatial distribution of the measured SOM content values. As it was anticipated, the SOM content is much lower in arable lands and vineyards than in forests and meadows, due to the long-term and intensive agricultural activity, as well as the soil erosion effects, which cause a higher amount of organic matter mineralization, as well as the erosion of the SOM rich topsoil. As a consequence, the study site shows a diversified picture of the SOM content's spatial distribution. These also imply that, the intrinsic hypothesis is not reasonable to hold, because there is an obtrusive trend (i.e. the local mean systematically varies from place to place), which makes several kriging techniques (e.g. ordinary kriging, simple kriging) inappropriate. Moreover, the large number of factors (e.g. land cover types, topography, morphometric parameters) and their abrupt changes in the geographic space make cokriging and kriging with external drift techniques inadequate too, according to GOOVAERTS (1997). Alternatively, RK is able to handle the trend, as well as it can take numerous secondary information into account considering their abrupt changes in geographic space. Hence, the RK technique is a reasonable choice for modeling the SOM content's spatial distribution in the area of interest.

3.2. Theory of regression kriging (RK)

In the last ten years, regression kriging (RK) has been more and more popular to estimate the value(s) of the target variable(s) at unvisited locations taking spatially exhaustive secondary information into account (HENGL et al., 2004; SIMBAHAN et al., 2006; GOOVAERTS, 2010, 2011; KERRY et al., 2012; SZATMÁRI & BARTA, 2013; SZATMÁRI et al., 2013; PÁSZTOR et al., 2014a, b). RK assumes that the random function $Z(u)$ can be deconstructed into a

trend component (also referred to as drift or mean, see GOO-VAERTS [1997]) $m(u)$ and a residual component $\varepsilon(u)$:

$$Z(\mathbf{u}) = m(\mathbf{u}) + \varepsilon(\mathbf{u}) , \qquad (Eq. 1)$$

where the trend component is accounted for by a multiple linear regression model, whilst the model residuals represent the spatially varying but dependent stochastic component with zero mean, normal distribution (see Fig. 2) and covariance structure, which can be modeled with a simple kriging technique. Figure 2 presents the schema of RK. Based on this, the estimation for Z at an unvisited location u_0 is given by

$$\hat{Z}(\mathbf{u}_0) = \mathbf{q}_0^{\mathrm{T}} \times \boldsymbol{\beta} + \lambda_0^{\mathrm{T}} (\mathbf{z} - \mathbf{q} \times \boldsymbol{\beta}) \qquad (Eq. 2)$$

where $\boldsymbol{\beta}$ is the vector of the regression coefficients, $\boldsymbol{q}_0$ is the vector of the secondary information at the unvisited location, λ_0 is the vector of the simple kriging weights (assigned to the regression's residuals), $\boldsymbol{z}$ is the vector of the observations and $\boldsymbol{q}$ is the matrix of the secondary information at the sampling locations. The regression coefficients are estimated by the generalized least squares (GLS) method because it is able to take the covariance matrix of the residuals into account along the estimation process.

3.3. Spatially exhaustive secondary information for RK

In case of soil related attribute(s), secondary information can be compiled according to the soil-forming factors because there is a significant relationship between these factors and the soil attribute(s) (PHILLIPS, 1998; MCBRATNEY et al., 2003; BOCKHEIM & GENNADIYEV, 2010; SZATMÁRI et al., 2013; BOCKHEIM et al., 2014). According to BOCK-HEIM et al. (2014), the soil-forming factors are the following: topography, climate, parent material, organisms (i.e. vegetation and fauna), the age of the soil and the human intervention, as an anthropogenic factor.
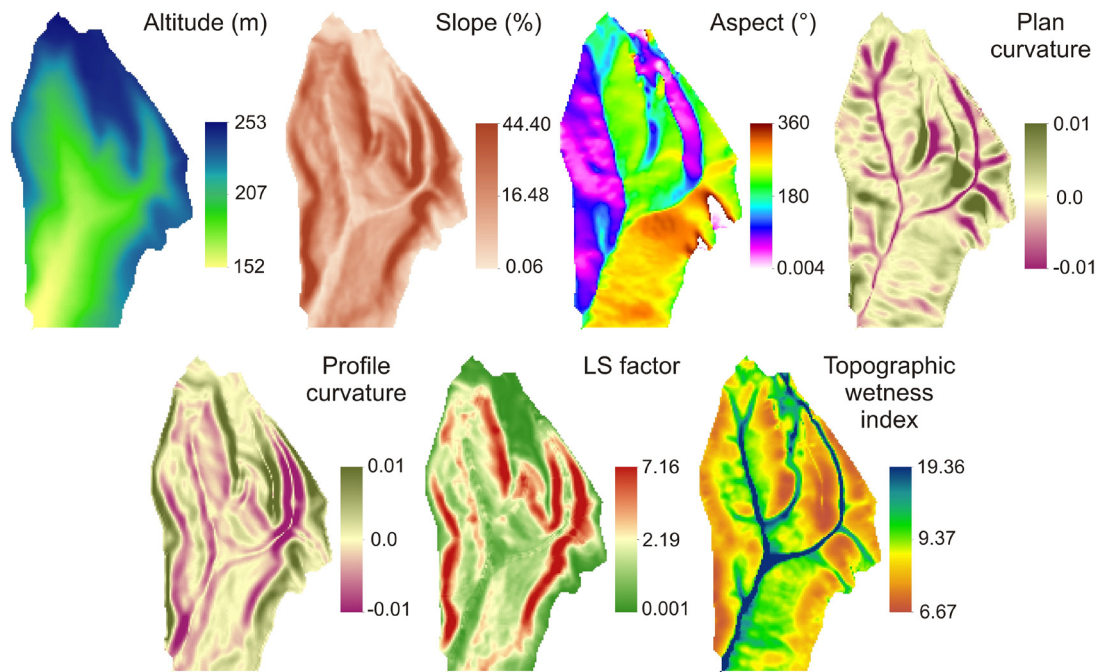
Szatmári et al.: Testing a sequential stochastic simulation method based on regression kriging in a catchment area in Southern Hungary

277
Geologia Croatica

Figure 3: The maps of the derived morphometric parameters.

Climate and parent material can be regarded as homogeneous because of the local aspect of the sub-catchment (DÖVÉNYI, 2010). Unfortunately, we do not have any information about the age of the soils. However, this factor has been frequently omitted in soil related spatial modelling, because it is difficult to characterize well (MCBRATNEY et al., 2003).

Spatially exhaustive information on topography can be derived from the digital elevation model (DEM) of the study area, which was built up with 10 metre resolution. The first step in using this secondary information is morphometric analysis of DEM. The derived, so called "morphometric" parameters' grids are aimed to characterize the geomorphometry of the surface (MCBRATNEY et al., 2003). The grids of the morphometric parameters have the same resolution as DEM. Table 2 and Figure 3 summarize the derived parameters and their characteristics. Note that there is a significant relationship between the derived morphometric parameters, thus the "raw" grids of these parameters cannot be used in further multiple linear regression analysis because of multicollinearity. To avoid this, principal component (PC) analysis was performed to transform the grids of the morphometric parameters, because the resultant PC grids are orthogonal and independent. Hence, their application decreases the effect of multicollinearity; moreover the resulting PC grids preserve the total variation of the morphometric parameters (GEIGER, 2007). The PC analysis was carried out in SAGA GIS software with its "Spatial and Geostatistics / Principal Components" module, which calculates the PC grids from the input grids. The resulting PC grids were used in further multiple linear regression analysis.

Spatially exhaustive information on organisms (i.e. vegetation) and human intervention can be derived from the LC

Table 2: The derived morphometric parameters and their characteristics.

Morphometric parameters	Characteristics
Altitude (meter)	The vertical distance between the surface and the Baltic Sea level
Slope (%)	The rise or fall of the surface in percent
Aspect or Slope exposure (degree)	The compass direction that a slope faces
Plan curvature (dimensionless)	The curvature of the surface perpendicular to the slope direction
Profile curvature (dimensionless)	The curvature of the surface in the direction of the slope
LS (or topographic) factor (dimensionless)	Quantified effects of slope and slope length on water erosion (WISCHMEIER & SMITH, 1978)
Topographic Wetness Index (dimensionless)	Quantified control of local topography on hydrological processes, and indicator of the spatial distribution of soil moisture and surface saturation

map of the study area (Fig. 1). In the present case, the LC map was compiled interpreting the products of the official aerial photography campaign of Hungary, taken in 2005. As opposed to the morphometric parameters, the LC type is a categorical variable. For the sake of the application of RK, each LC type was converted into an indicator variable (IV). Therefore a grid map was generated (with 10 metre resolution) for each LC type with a value domain showing 1 at the locations of the given LC type and showing 0 for all other locations. The resulting IVs were used in further multiple linear regression analysis.

3.4. Theory of sequential stochastic simulation based on regression kriging (SSSRK)

During previous decades, stochastic simulations became widespread (GOOVAERTS, 1997; DEUTSCH & JOURNEL, 1998; GEIGER, 2006; MALVIĆ, 2008; NOVAK ZELENIKA & MALVIĆ, 2011; GEIGER, 2012; MALVIĆ et al., 2012; NOVAK ZELENIKA et al., 2012). These simulations are methods in which alternative and equally probable high resolution models of spatial distribution of $Z(\mathbf{u})$ are generated (DEUTSCH & JOURNEL, 1998). If the realizations (also referred to as stochastic images) from them honour the input data, then the simulation is called "conditional" (DEUTSCH & JOURNEL, 1998).

According to GOOVAERTS (1997), let $\{Z(\mathbf{u}_j), j=1,\dots,N\}$ be a set of random variables defined at N locations $\mathbf{u}_j$ within the study area. The objective is to generate several joint realizations of these N random variables conditional to the dataset. The corresponding N-point (or N-variate) conditional cumulative distribution function (ccdf) is:

$$F(\mathbf{u}_1,\dots,\mathbf{u}_N;z_1,\dots,z_N\mid(n))=P\{Z(\mathbf{u}_1)\le z_1,\dots,Z(\mathbf{u}_N)\le z_N\mid(n)\}$$

(Eq. 3)

However, this N-point ccdf can be written as the product of N one-point ccdf:

$$F(\mathbf{u}_1,\dots,\mathbf{u}_N;z_1,\dots,z_N\mid(n))=F(u_N;z_N\mid(n+N-1))\cdot$$
$$F(u_{N-1};z_{N-1}\mid(n+N-2))\cdot\dots\cdot$$
$$F(u_2;z_2\mid(n+1))\cdot$$
$$F(u_1;z_1\mid(n))$$

(Eq. 4)

where the one-point ccdfs are conditional to the dataset and all previously simulated values. Based on this, a stochastic image can be generated in N successive steps (GOOVAERTS, 1997; GEIGER, 2006):

1. Model the ccdf at the first location $\mathbf{u}_1$ conditional to the dataset.
2. Draw a realization from this ccdf, which realization becomes a conditioning datum for all subsequent drawings.
3. Model the ccdf of $Z(\mathbf{u}_i)$ at the ith location $(\mathbf{u}_i)$ conditional to the dataset and all (i-1) previously simulated values.
4. Draw a realization from this ccdf, which realization becomes a conditioning datum for all subsequent drawings.
5. Repeat the two previous steps until all N locations are visited and each has been given a simulated value.

The resulting set of simulated values represents one realization of the random function over the N locations. Other
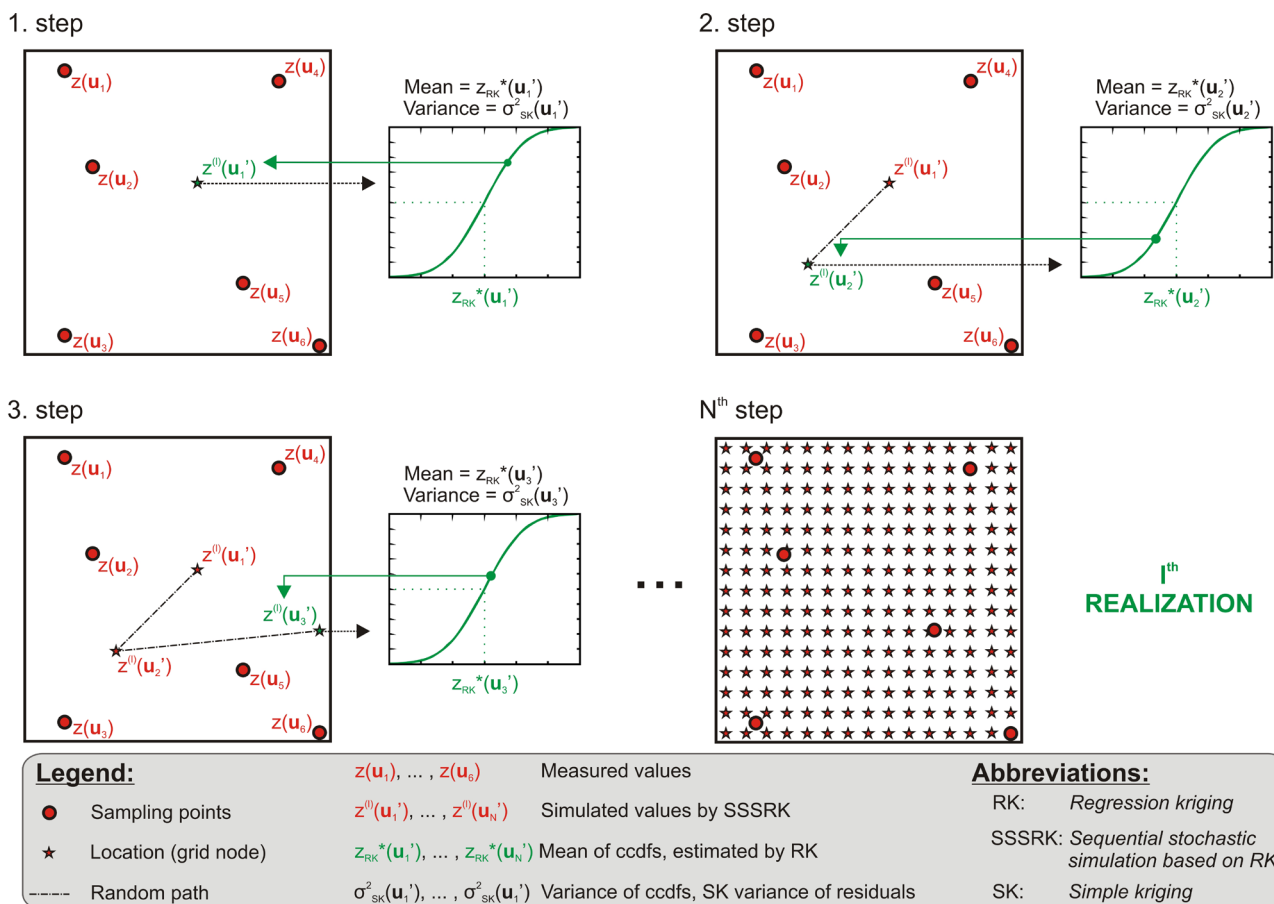

Figure 4: A schema of sequential stochastic simulation based on regression kriging.

Szatmári et al.: Testing a sequential stochastic simulation method based on regression kriging in a catchment area in Southern Hungary

279

Geologia Croatica

realizations are obtained by repeating the entire sequential process with possibly different random path for each realization (GOOVAERTS, 1997; GEIGER, 2006).

When secondary information is available, then this information can be used in the sequential stochastic simulation process (GOOVAERTS, 1997; DEUTSCH & JOURNEL, 1998; MALVIĆ, 2008). As mentioned above, the secondary information is related to DEM and the LC map of the pilot area. In this study, the RK estimation was used to identify the mean of ccdf at any grid node and the simple kriging variance of the residuals was used to identify the variance of ccdf at any grid node, according to GOOVAERTS (1997) and DEUTSCH & JOURNEL (1998). Figure 4 presents the schema of SSSRK. The consequence of this practice is that the variogram and the histogram of the residuals are reproduced by the simulation model. Furthermore, the realizations honour the input dataset.

In our work, 100 realizations (according to GOO-VAERTS [2001] and GEIGER [2006]) were generated by the previously detailed SSSRK method. The resulting stochastic images can be used, for example, to map the E-type estimation, as well as the corresponding upper and lower bound of the confidence interval for each grid node, to assess the spatial uncertainty using the differences between the realizations or to solve tasks like "contouring the probability of the event of {SOM < 1.5%}" (GEIGER & MUCSI, 2005; GEIGER, 2006; MUCSI, et al. 2013).

4. RESULTS AND DISCUSSION

4.1. Results of RK

The generalized least squares (GLS) method was used to estimate the regression coefficients for the multiple linear regression model. The response variable was the SOM content, whilst the explanatory variables were the PC grids of the morphometric parameters and the indicator variables (IV) of the LC types. The applied significance level was 0.05 and the "stepwise" method was used to select the explanatory variables into the regression model. Table 3 summarizes the results of the multiple linear regression analysis.

Table 3: Results of the multiple linear regression analysis.

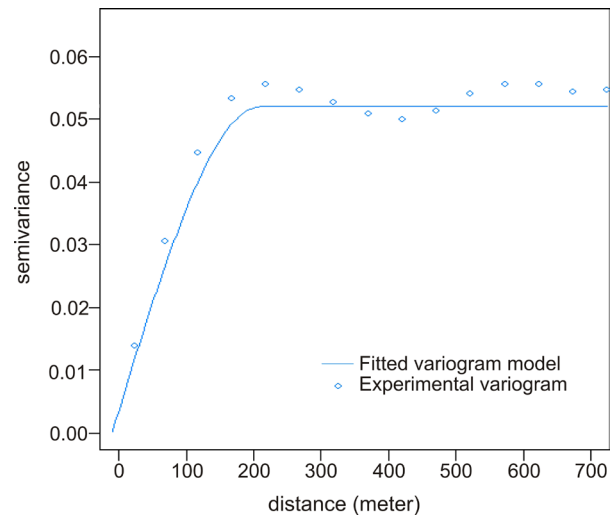|	Regression coefficients	Sig.
Intercept	1.878	$< 2*10^{-16}$
Explanatory variables		
PC-1	-0.004	$3.08*10^{-02}$
PC-2	-0.035	$1.20*10^{-08}$
PC-4	-0.172	$1.29*10^{-05}$
IV-Forests	-0.621	$1.63*10^{-04}$
IV-Vineyards	-0.859	$5.33*10^{-08}$
IV-Eroded Arables	0.624	$3.40*10^{-05}$
PC-6	-0.533	$1.14*10^{-02}$
Determination coefficient (R^2)	0.809	
model p-value	$3.59*10^{-11}$	

Figure 5: An experimental variogram and the fitted variogram model of the residuals.

Seven explanatory variables were selected into the multiple linear regression model by the "stepwise" method (see Table 3), where 4 explanatory variables were related to the PC grids (i.e. PC-1, PC-2, PC-4 and PC-6), whilst 3 explanatory variables were related to the IV grids (i.e. IV-Forests, IV-Vineyards and IV-Eroded Arables). In case of the selected
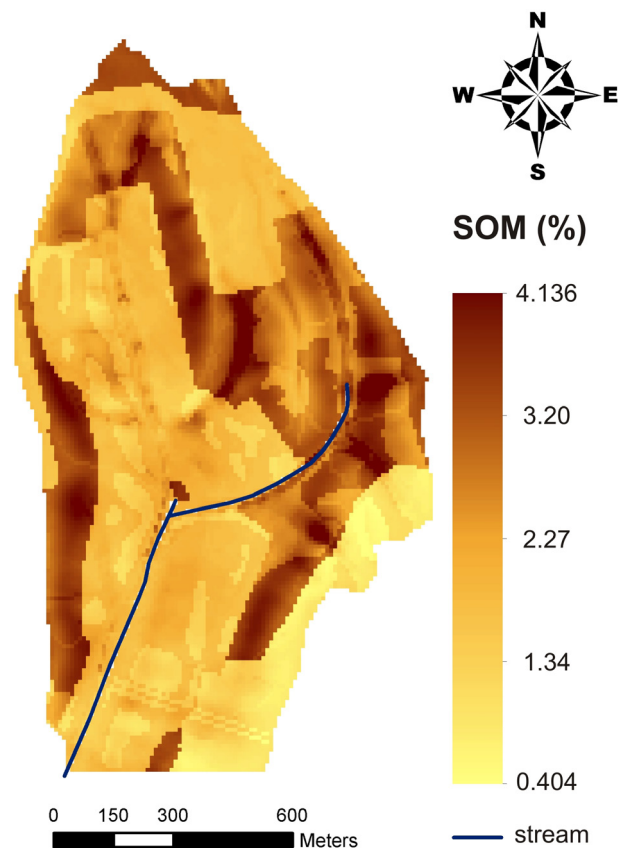

Figure 6: A map of the soil organic matter (SOM) content created by regression kriging.
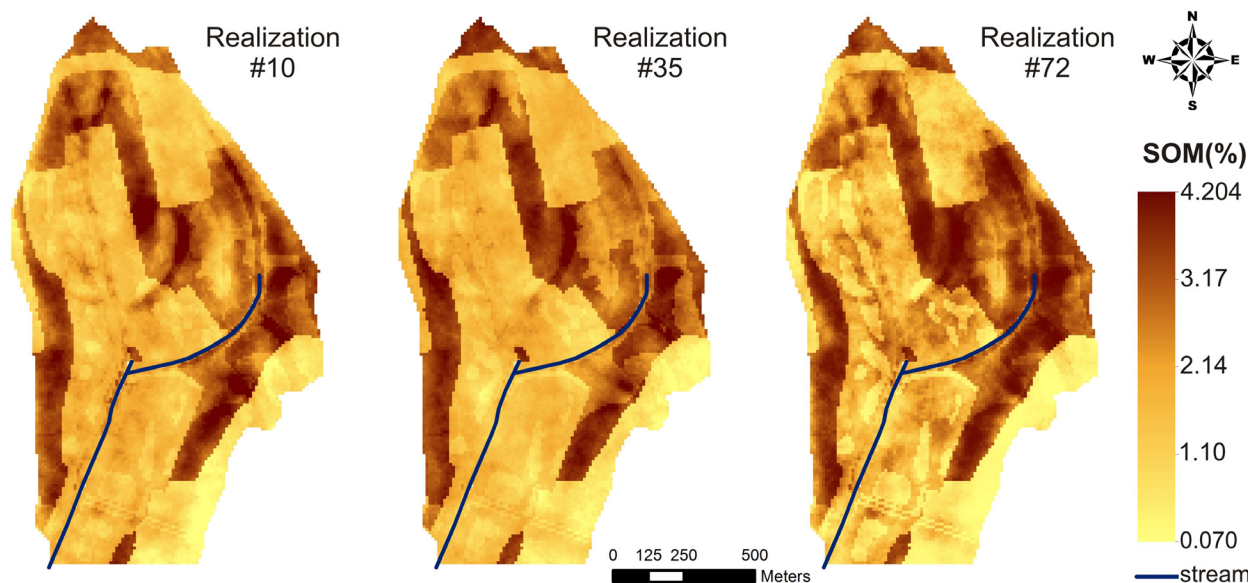
Figure 7: Three realizations of the sequential stochastic simulation based on regression kriging. Abbreviation: SOM: soil organic matter.

PC grids, PC-1 relates to the altitude parameter (i.e. the altitude morphometric parameter had the highest PC coefficient in the PC analysis), PC-2 relates to slope, PC-4 represents the Topographic Wetness Index, whilst PC-6 relates to the LS factor. As a consequence, SOM spatial distribution is mainly determined by the soil erosion related morphometric parameters and the LC types, as was anticipated. The determination coefficient of the regression model is 0.809 (see Table 3), which means that the model explains more than 80% of the total variability of the SOM data and the remaining approximately 20% have to be modeled with a simple kriging system. The regression residuals were derived and the corresponding experimental variogram was calculated to model their spatial structure (Fig. 5). The experimental variogram was approached with a spherical variogram model type with zero nugget, 0.0515 sill (which is approximately 20% of the total variance, see Table 1), 204 meter range and isotropic characteristic (see Fig. 5). The spatial estimation by RK was carried out using the multiple linear regression model and the fitted variogram model. The RK estimation is presented in Fig. 6.

4.2. Results of SSSRK

In this study, 100 equally probable realizations were generated based on the regression kriging system presented earlier and using the SSSRK algorithm. Figure 7 shows three of the resulting stochastic images of SSSRK. As we can see in Fig. 7, there are areas where the stochastic images are not so different (e.g. arable lands), whereas we can find some regions (e.g. forests and meadows) where the differences are more pronounced.

The experimental variograms of the resulted realizations can be derived and they can be compared with the applied variogram model, which was used in the SSSRK process (Fig. 8). One of our constraints was that SSSRK algorithm

has to reproduce, through the resulted realizations, the applied variogram model of the residuals. Figure 8 shows that this was achieved.

The 100 realizations provide 100 simulated values for each grid node and this number is quite appropriate to calculate the cumulative distribution around an infinitesimally small neighbourhood of each grid node (GEIGER, 2006; MUCSI et al., 2013). Using these cumulative distributions the E-type estimation and the corresponding upper and lower boundary of the 95% confidence interval can be calculated for each grid node. Moreover, the 95% confidence interval's width also can be derived, which provides a measure of uncertainty of the SOM estimation, i.e. when this interval width is relatively high, then the SOM estimation is more uncertain, according to GEIGER (2006). Figure 9 shows the E-
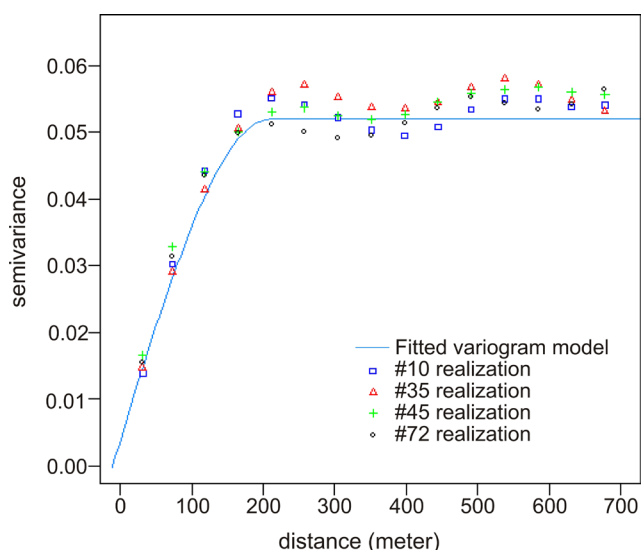

Figure 8: Derived variograms from several realizations and the applied variogram model.

Szatmári et al.: Testing a sequential stochastic simulation method based on regression kriging in a catchment area in Southern Hungary
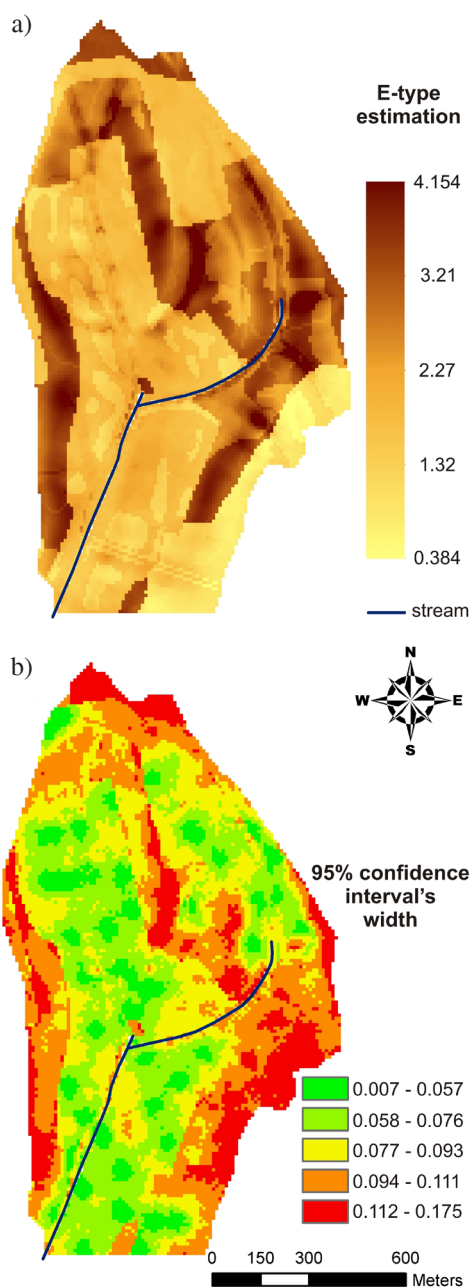
281

Geologia Croatica

a)

b)

Figure 9: The E-type estimation of the soil organic matter content (a) and the corresponding 95% confidence interval width (b) on the basis of 100 realizations.

type estimation and the corresponding confidence interval width. As we can see in Fig. 9b, the E-type estimations are more uncertain in forests, vineyards and meadows, than in arable lands. It can be attributed to the sampling strategy, which underrepresented the former LC types.

A horizontal transect (see Fig. 1) was traced out in the northern part of the study site, which intersects the most frequent LC types (arable, meadows and forests). Along the transect, the E-type estimation, the upper and lower boundary of the 95% confidence interval and every tenth simulated value were plotted (see Fig. 10) in order to analyze how the simulated values vary in different LC types. Figure 10 shows clearly that the simulated values in forests and meadows have

a wider range, than is the case for arable lands. This is in accordance with the aforementioned statements that the E-type estimations are more uncertain in forests and meadows. Furthermore, the transect also indicates that the SOM content is much lower in arable lands, than in forests and meadows, due to the long-term and intensive agricultural activity, as well as the effects of soil erosion, which cause a higher amount of organic matter mineralization, as well as the erosion of the SOM rich topsoil. At 140 metres along the transect, the range of the simulated values is fairly small, which can be attributed to a sampling point, which is pretty near that grid node (see Fig. 1). We can conclude that, the SSSRK algorithm honours the input dataset, which was the other constraint on the SSSRK technique.

In contrast, we can use the calculated cumulative distribution for each grid point to solve tasks like "contouring the probability of the event of {SOM < 1.5%}", according to GEIGER (2006). Figure 11 presents that "probability map", which can be directly used in a soil protection and management plan of the sub-catchment to delineate areas for SOM archiving.

4.3. Comparison of RK estimates with SSSRK's E-type estimates

Numerous authors have compared the results of the estimation and simulation methods in the last decade, such as GOOVAERTS (2000), GEIGER (2006), MALVIĆ (2008), NOVAK ZELENIKA & MALVIĆ (2011), MALVIĆ et al. (2012), NOVAK ZELENIKA et al. (2012) or MUCSI et al. (2013). Following this practice, the results of RK and SSSRK were compared. If we compare the RK estimation (Fig. 6) with the SSSRK's E-type estimation (Fig. 9.a), then we can state that, they are very similar. To test this impression, a difference map was calculated (Fig. 12), which quantifies the difference between the map of RK estimation and SSSRK's E-type estimation. The range of the difference map is [0.068; -0.091] (see Fig. 12), which is fairly small. Based on this, we can conclude that the two maps, from the practical point of view, present the same result. However, we have to notice that this similarity may can be attributed to the fact that the determination coefficient of the regression model is fairly high (R^2=0.809), which decreased the smoothing effect of the RK technique (SZATMÁRI et al., 2013).

5. CONCLUSIONS

In this paper, a sequential stochastic simulation method based on regression kriging (SSSRK) was presented and tested in a sub-catchment area of the Lajvér stream, in Southern Hungary. For this purpose, the soil organic matter (SOM) content was chosen because this particular soil property has an important role due to its multipurpose applicability.

As it was illustrated in this study, SSSRK (as opposed to RK) is able to model the spatial uncertainty of the target variable using the generated equally probable realizations (which reproduce the model statistics and honour the input dataset). It is able to provide a measure of the uncertainty of
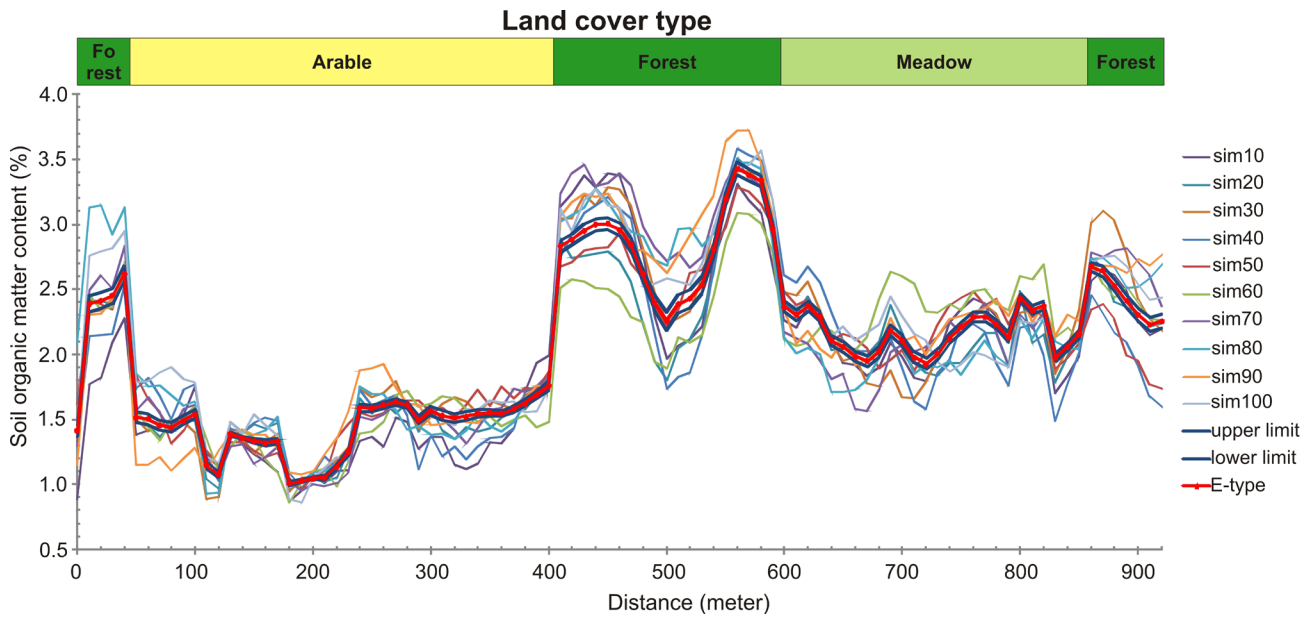
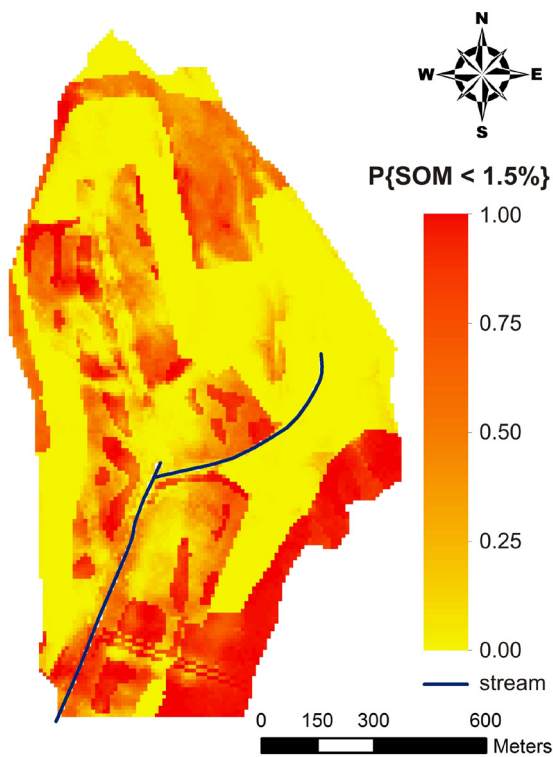**Figure 10:** Horizontal transect of every tenth simulated stochastic image in the northern part of the sub-catchment.

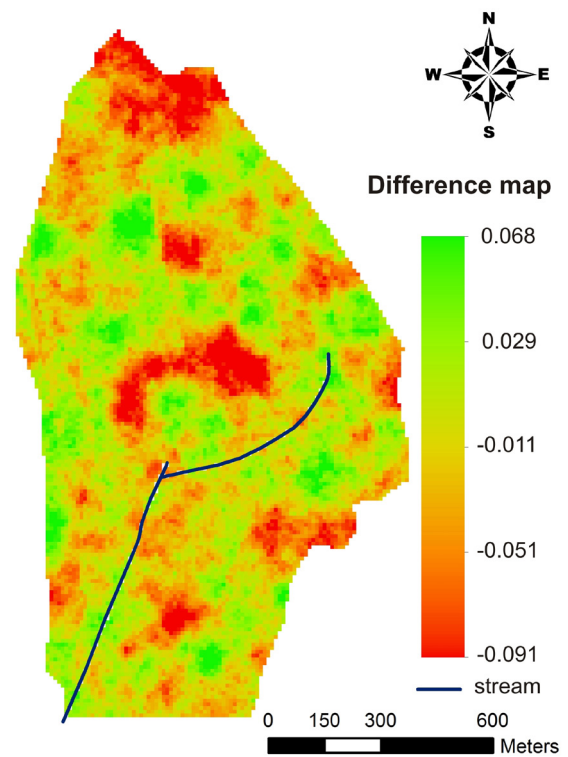**Figure 11:** A map of the probability of the event of {SOM < 1.5%}.

Figure 12: A difference map of the regression kriging estimation and the sequential stochastic simulation based on regression kriging's E-type estimation.

the E-type estimates using the derived confidence interval width. Moreover, the calculated cumulative distribution around an infinitesimally small neighbourhood of each grid node can be used to support various decisions (e.g. identification of SOM rich or SOM poor areas). In addition, SSSRK retained the main advantages of the RK technique such as

its flexibility, but it can also handle the trend (or drift) and it can take spatially exhaustive secondary information into account in the simulation process.

In conclusion, SSSRK is a valuable technique to model the spatial distribution, variability and uncertainty of the target variable and to complete RK's several shortcomings.

283

Szatmári et al.: Testing a sequential stochastic simulation method based on regression kriging in a catchment area in Southern Hungary

Geologia Croatica

ACKNOWLEDGEMENT

REFERENCES

BOCKHEIM, J.G. & GENNADIYEV, A.N. (2010): Soil-factorial models and earth-system science: A review.– Geoderma, 159, 243–251.

BOCKHEIM, J.G., GENNADIYEV, A.N., HARTEMINK, A.E. & BREVIK, E.C. (2014): Soil-forming factors and Soil Taxonomy.– Geoderma, 226–227, 231–237.

BORCSIK, Z., FARSANG, A., BARTA, K. & KITKA, G. (2011): Humuszanyagok mennyiségi és minőségi eróziójának mérése a Tolna megyei Szálka település melletti vízgyűjtőn [Quantitative and qualitative measurements of the erosion affected humic material's distribution in a subcatchment area near village Szálka, in Tolna County – in Hungarian, with an English Abstract].– Talajvédelem (Special Issue), 127–137.

DEUTSCH, C.V & JOURNEL, A.G. (1998): GSLIB: Geostatistical Software Library and User's Guide.– Oxford University Press, New York, 369 p.

DÖVÉNYI, Z. (ed.) (2010): Magyarország kistájainak katasztere [*The cadastre of the Hungarian small regions* – in Hungarian].– MTA FKI, Budapest, 876 p.

ELDEIRY, A. & GARCIA, L. (2010): Comparison of Ordinary Kriging, Regression Kriging, and Cokriging Techniques to Estimate Soil Salinity Using LANDSAT Images.– J. Irrig. Drain Eng., 136/6, 355–364.

GEIGER, J. (2006): Szekvenciális gaussi szimuláció az övzátonytestek kisléptékű heterogenitásának modellezésében [*Sequential Gaussian simulation in the modelling of small-scale heterogeneity of point-bars* – in Hungarian, with an English Abstract].– Földtani közlöny, 136/4, 527–546.

GEIGER, J. (2007): Geomatematika [Geomathematics – in Hungarian]. – JATEPress, Szeged, 117 p.

GEIGER, J. (2012): Some thoughts on the pre- and post-processing in sequential Gaussian simulation and their effects on reservoir characterization.– In: GEIGER, J., PÁL-MOLNÁR, E. & MALVIĆ, T. (eds.): New horizons in central european geomathematics, geostatistics and geoinformatics.– GeoLitera, Szeged, 17–34.

GEIGER, J. & MUCSI, L. (2005): A szekvenciális sztochasztikus szimuláció előnyei a talajvízszint kisléptékű heterogenitásának térképezésében [*The advantages of stochastic simulation in mapping the small-scale heterogeneity of water tables* – in Hungarian, with an English Abstract].– Hidrológiai közlöny, 85/2, 37–47.

GOOVAERTS, P. (1997): Geostatistics for Natural Resources Evaluation.– Oxford University Press, New York, 512 p.

GOOVAERTS, P. (2000): Estimation or simulation of soil properties? An optimization problem with conflicting criteria.– Geoderma, 97, 165–186.

GOOVAERTS, P. (2001): Geostatistical modelling of uncertainty in soil science.– Geoderma, 103, 3–26.

GOOVAERTS, P. (2010): Combining Areal and Point Data in Geostatistical Interpolation: Applications to Soil Science and Medical Geography.– Math. Geosci., 42, 535–554.

GOOVAERTS, P. (2011): A coherent geostatistical approach for combining choropleth map and field data in the spatial interpolation of soil properties.– Eur. J. Soil Sci., 62, 371–380.

HENGL, T., HEUVELINK, G.B.M. & STEIN, A. (2003): Comparison of kriging with external drift and regression-kriging (technical report).– International Institute for Geo-information Science and Earth Observation (ITC), Enschede, 18 p.

HENGL, T., HEUVELINK, G.B.M. & STEIN, A. (2004): A generic framework for spatial prediction of soil variables based on regression-kriging.– Geoderma, 120, 75–93.

KERRY, R., GOOVAERTS, P., RAWLINS, B.G. & MARCHANT, B.P. (2012): Disaggregation of legacy soil data using area to point kriging

for mapping soil organic carbon at the regional scale.– Geoderma, 170, 347–358.

MALVIĆ, T. (2008): Kriging, cokriging or stochastical simulations, and the choice between deterministic or sequential approaches.– Geol. Croat., 61/1, 37–47.

MALVIĆ, T., NOVAK ZELENIKA, K. & CVETKOVIĆ, M. (2012): Indicator vs. gaussian geostatistical methods in sandstone reservoirs – case study from the Sava Depression, Croatia.– In: GEIGER, J., PÁL-MOLNÁR E. & MALVIĆ, T. (eds.): New horizons in central european geomathematics, geostatistics and geoinformatics.– GeoLitera, Szeged, 37–45.

MCBRATNEY, A.B., MENDONCA SANTOS, M.L. & MINASNY, B. (2003): On digital soil mapping.– Geoderma, 117, 3–52.

MINASNY, B. & MCBRATNEY, A.B. (2007): Spatial prediction of soil properties using EBLUP with the Matérn covariance function.– Geoderma, 140, 324–336.

MUCSI, L., GEIGER, J. & MALVIĆ, T. (2013): The advantages of using sequential stochastic simulations when mapping small-scale heterogeneities of the groundwater level.– J. En. Geo., 6/3–4, 39–47. doi: 10.2478/jengeo-2013-0005

MSZ 21470–52:1983 (1983): Környezetvédelmi talajvizsgálatok: Talajok szervesanyag-tartalmának meghatározása [*Environmental protection. Testing of soils. Determination of organic matter* – in Hungarian].– Hungarian Standard Association, Budapest, 4 p.

NOVAK ZELENIKA, K. & MALVIĆ, T. (2011): Stochastic simulations of dependent geological variables in sandstone reservoirs of Neogene age: A case study of the Kloštar Field, Sava Depression.– Geol. Croat., 64/2, 173–183.

NOVAK ZELENIKA, K., VELIĆ, J. & MALVIĆ, T. (2012): Application of geostatistics in the description of turbiditic depositional environments, a case study of the Kloštar Field, Sava Depression, Croatia.– In: GEIGER, J., PÁL-MOLNÁR E. & MALVIĆ, T. (eds.): New horizons in central european geomathematics, geostatistics and geoinformatics.– GeoLitera, Szeged, 61–71.

PÁSZTOR, L., KÖRÖSPARTI, J., BOZÁN, CS., LABORCZI, A. & TAKÁCS, K. (2014a): Spatial risk assessment of hydrological extremities: Inland excess water hazard, Szabolcs-Szatmár-Bereg County, Hungary.– Journal of Maps, DOI: 10.1080/17445647.2014.954647

PÁSZTOR, L., SZABÓ, J., BAKACSI, ZS., LABORCZI, A., DOBOS, E., ILLÉS, G. & SZATMÁRI, G. (2014b): Elaboration of novel, country-wide maps for the satisfaction of recent demands on spatial, soil related information in Hungary.– In: ARROUAYS, D., MCKENZIE, N., HEMPEL, J., RICHER DE FORGES, A. & MCBRATNEY, A.B. (eds.): Global Soil Map: Basis of the global spatial soil information system.– Taylor & Francis, London, 207–212.

PHILLIPS, J.D. (1998): On the relations between complex systems and the factorial model of soil formation (with Discussion).– Geoderma, 86, 1–42.

SIMBAHAN, G.C., DOBERMANN, A., GOOVAERTS, P., PING, J. & HADDIX, M.L. (2006): Fine-resolution mapping of soil organic carbon based on multivariate secondary data.– Geoderma, 132, 471–489.

SZATMÁRI, G. & BARTA, K. (2013): Csernozjom talajok szervesanyag-tartalmának digitalis 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* – in Hungarian, with an English Abstract].– Agrokem. 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 [*Large-scale mapping of soil organic matter content by regression kriging in Zala County* – in Hungarian, with English Abstract].– Agrokem. Talajtan, 62/2, 219–234.

WEBSTER, R. & OLIVER, M.A. (2007): Geostatistics for Environmental Scientists (2nd Ed.).– Wiley, Chichester, 330 p.

WISCHMEIER, W.H. & SMITH, D.D. (1978): Predicting rainfall erosion losses: A guide to conservation planning.– U.S. Government Printing Office, Washington D.C., 58 p.