Mapping Soils Using the Fuzzy Approach and Regression-kriging – Case Study from the Považský Inovec Mountains, Slovakia

Juraj BALKOVIČ[1,3], *Gabriela ČEMANOVÁ*[1], *Jozef KOLLÁR*[2], *Miroslav KROMKA*[1] and *Katarína HARNOVÁ*[1]

[1]*Department of Soil Science, Faculty of Natural Science, Comenius University in Bratislava, Bratislava, Slovak Republic;* [2]*Institute of Landscape Ecology, Slovak Academy of Sciences, Bratislava, Slovak Republic;* [3]*Soil Science and Conservation Research Institute, Bratislava, Slovak Republic*

Abstract: The paper introduces a method of digital mapping of spatial distribution of soil typological units. It implements fuzzy *k*-means to classify the soil profile data (study area from the Považský Inovec Mountains, Slovakia) and regression-kriging with the selected digital terrain and remote sensing data to draw membership maps of soil typological units. Totally three soil typological units were identified: Haplic Cambisols (Skeletic, Dystric), Albic Stagnic Luvisols, and Haplic Stagnosols (Albic, Dystric). We analysed the membership values to these units with respect to terrain and remote sensing data. The membership values appeared as spatially smoothly dependant on the terrain gradients (linearly or exponentially) whereas the residua showed spatial autocorrelation. Based on regression and kriging analyses, the regression-kriging model was successfully deployed to draw raster membership maps. These maps yield coefficients of determination between R^2 = 56% (Albic Stagnic Luvisols) to R^2 = 79% (Haplic Cambisols (Skeletic, Dystric)) when evaluated by cross validation. The grid-based continuous soil map represents an alternative to the classical polygon soil maps and can offer a wide range of interpretations for landscape studies.

Keywords: fuzzy *k*-means; regression-kriging; digital landscape data; grid interpretation; spatial distribution; soil classification

Soil cover represents a continuous body that respects several natural gradients in the landscape, and which continuously changes along these gradients. The ambiguity in soils can be a problem for the classification issues and thematic soil maps because one has to decide for explicit soil typological unit or the soil mapping unit. Soils usually show a diffuse spatial distribution that is hard to address in chorochromatic polygon soil maps. Burrough *et al.* (1997) describe the polygon soil

Supported by the Scientific Grant Agency of the Ministry of Education of Slovak Republic and the Slovak Academy of Sciences, Projects No. VEGA 1/2375/05 and VEGA 2/5001/25.

map as double crisp for it creates discontinuity in both taxonomical and geographical space. These authors suggest the continuous raster maps as a better alternative to mapping soils and soil properties. In this contribution, we test the method for mapping the soil cover with continuous raster maps in the study area, which occurs in the eastern part of the Považský Inovec Mountains, Slovakia. The study is based on a systematic profile sampling of forest soils. To address soil fuzziness in the classification and to implement it into the soil maps, the model includes (*i*) fuzzy *k*-means classification (Bezdek *et al.* 1984) to identify soil typological units (STUs), and (*ii*) regression-kriging (Odeh *et al.* 1994) to spatially interpret the fuzzy *k*-partition to STUs using predictor GIS variables. Fuzzy techniques are known to provide taxonomically interpretable data with the floating numeric format (e.g. McBratney & Moore 1985; De Gruijter & McBratney 1988; Burrough *et al.* 1997; De Gruijter *et al.* 1997; Hengl *et al.* 2004; Lagacherie 2005), which can be treated as numeric indices of the spatial variability of soils. Fuzzy *k*-means method implements the theory of fuzzy sets (Zadeh 1965) and it partitions soil profiles into an explicit number of classes through the fuzzy *k*-partition, i.e. the set of membership values (MV). The method is used here to classify the soil-profile data into *k* STUs – each unit is characterised by a set of MVs (weights of belonging to STU) for *n* soil profiles. The set of MVs to a particular STU represents the target variable for the model in this paper, which is mapped to continuous soil maps. The regression-kriging technique is used as the mapping agent to interpolate these MVs along GIS predictor variables into grid coverage. The approach presented uses terrain and remote sensing (RS) data as predictors to support spatial interpretation of the fuzzy *k*-partition. Two basic requirements exist to use successfully GIS predictors – (i) they must closely determine, or copy, the distribution of the soil cover, and (ii) they must exist as high-resolution GIS information so that they can be used to predict the target variable at un-sampled locations. The spatial distribution of STU is then expressed as a membership map, i.e. the continuous coverage of MVs where the cell values range between 0 and 1 (0 – no similarity between cell and centroid of particular STU, 1 – very great similarity between cell and centroid of particular STU).

MATERIAL AND METHODS

Study area

The study area is situated in the eastern part of the Považský Inovec Mountains (Slovakia) and covers approximately 700 ha of forests (Figure 1). Soils on the summits and hillsides developed from weatherings of granitic rock and deluvium (Haplic Cambisols (Skeletic, Dystric)) or from the mixture of coarse and fine-earth materials at the foot of the hills (Haplic Stagnosols (Albic, Dystric) or

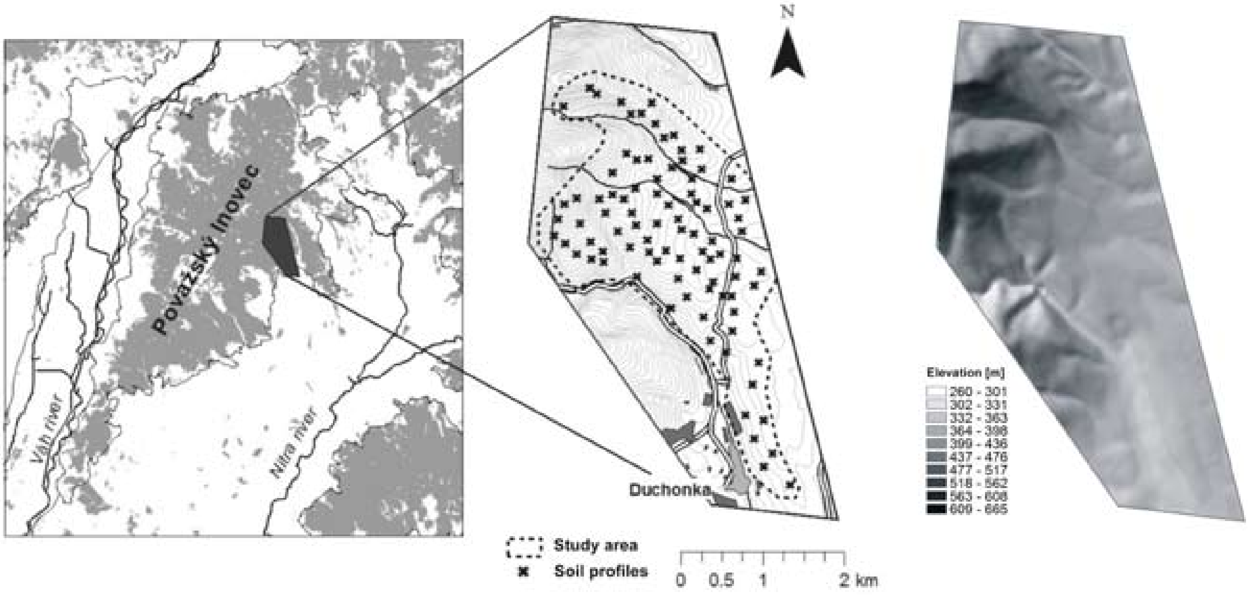

Figure 1. The situation map of the study area with locations of soil profiles

Stagnic Cambisols; nomenclature according to WRB by FAO 2006). The south-eastern part of the study area is built by loess of the Nitrianská pahorkatina hilly-country where acidic Albic Luvisols occur. The area belongs to the warm and moderately dry climatic region with mild winters and moderate humidity. Summer days count for less than 50, average July temperature is ≥ 16°C, and average January temperature above –3°C. The annual rainfall exceeds 700 mm in average (Lapin *et al.* 2002; SHMÚ – meteorological observatory in Závada). The vegetation cover is created by acidophilous oak and oak-hornbeam forests (Čemanová *et al.* 2005).

Soil sampling and coding of soil profile properties

The pattern of soil sampling was designed as more or less systematic; only the extreme erosion grooves and creek alluvia were omitted from the sampling. The field sampling was designed so as to get the between-plot distances to approximately 250 m, i.e. the diversity at shorter distances is not addressed in this study. Totally 90 soil profiles were sampled by the authors of this paper during the period of 2004 and 2005. The individual plots were located by GPS in WGS 84 geographical coordinate system. The following attributes of soils were sampled: genetic features of horizons, horizon depth in cm, colour by Munsell charts for homogenised soil samples, percentage of oxidation and reduction features, sand and clay contents, and stoniness (for all genetic horizons; see Table 1). The attributes were quantified by different criteria, such as directly measured, estimated by percentage, or by interval scales. The soil attributes were encoded in a numerical matrix (90 profile × 73 attributes) as an input for the numerical classification. The numerical scheme respects two main principles: (i) depth of diagnostic horizons and (ii) vertically explicit stratification of soil attributes. All variants and sub-horizons of diagnostic horizons, which were identified by the field research, were aggregated and cross-indexed by signatures of diagnostic horizons as shown in Table 1. The stratification of soil horizons was idealised to the sequence E-EB-B1-B2-B3-BC, whereas the numerical soil properties were assigned to each horizon of the idealised sequence. If a horizon does not exist in the soil profile, each property is set to zero. A-horizon was omitted from the fuzzy *k*-means classification as it is very homogenous in its properties throughout the study area. No more than three B-horizons were ever noticed in each soil profile. The identified soil typological units were classified with respect to WRB system (FAO 2006). "Dystric" suffix was used for STUs in accordance with the information published by Čemanová *et al.* (2005).

Predictor GIS data

The following GIS terrain and RS data were tested to be included into mapping: digital elevation model (DEM), slope in degrees (SLOPE), topography wetness index (TWI), length-slope factor (LS), and normalised difference vegetation index (NDVI). All data were available in 10-m cell resolution. DEM, which represents a raster model of the elevation values, was interpolated by RBF from the elevation contour-line vertexes of topography maps 1:10 000 using Geostatistical analyst for ArcGIS (Johnston *et al.* 2001). SLOPE and LS were calculated from DEM as described by Wilson and Gallant (2000). TWI, which reflects the tendency of water to accumulate at any point of the landscape, was calculated from DEM (see Wilson & Gallant 2000), and it is assumed that soil transmitivity is constant throughout the catchment area. Landsat TM satellite images (source SSCRI Bratislava) from May were used to calculate NDVI as the normalised quotient (Band 4 – Band 3)/(Band 4 + Band 3) (Maselli *et al.* 1998). With NDVI, we aim to gather the main differences in the vegetation cover, which usually indirectly reflects the soil properties. All topography analyses but DEM were calculated in the R open software (http://www.r-project.org). The basic statistics for the predictor GIS data are summarised in Table 2.

Model description

The model presented in the paper consists of two main components: (*i*) fuzzy *k*-means classifier and (*ii*) regression-kriging with predictor GIS variables. Fuzzy *k*-means classifier partitions multivariate soil objects into the given *k* classes, where the centroids of the classes are calculated by minimising the fuzzy partition error as proposed by Bezdek (1981). The fuzzy *k*-partition of MV(m_{ij}) follows the criteria given by statement (1):

$$\left\{m_{ij} = \in [0,1]; \sum_{j=1}^{k} m_{ij} = 1, i = 1 \dots n; \sum_{i=1}^{n} m_{ij} > 0, j = 1 \dots k\right\} \quad (1)$$

where:
n – total number of plots,
k – number of classes.

We used FuzME programme (Minasny & McBratney 2002) to execute fuzzy k-means with our data. Fuzzy k-means was calculated with the diagonal distance, with which the input data are transformed into equal variance. Although mathematical methods were suggested to optimise the parameters in the fuzzy k-means (e.g. McBratney & Moore 1985), we manually set the parameters as follows: fuzzy exponent to 1.8 and the number of classes to 3. Such parameterisation seems to yield spatially autocorrelated MVs for each of the output classes (STUs). Hereafter the fuzzy partition MV(m_{ij}) to three STUs is considered as a target soil variable that inputs further procedures of the model.

Regression-kriging (Odeh *et al.* 1994) represents the spatial interpreter that combines multiple regression and kriging, and is used to interpolate MVs into membership maps. The modification that uses multiple linear regression (MLR) and punctual kriging is expressed by Eq. (2). In one case, the modification with multiple exponential regression (MER) was used in this paper.

$$\hat{m}_j(s_0) = \sum_{l=1}^{p} \beta_l \times q_l(s_0) + \sum_{i=1}^{n} w_i(s_0) \times \varepsilon_j(s_i) \quad (2)$$

where:
$\hat{m}_j(s_0)$ – membership value of j^{th} STU at unsampled location s_0 (located by X and Y coordinates in nodes of regular grid),
$q_l(s_0)$ – the l^{th} predictor GIS variable at location s_0,

Table 1. List of diagnostic horizons, soil properties and their coding

Genetic soil horizons (their depth was sampled in cm)			
A	topsoil A horizon	Bw	cambic B-horizon
A_diff	transitive A/x-horizon	Bt	argic B-horizon
Eg	albic horizon with stagnic properties	Bg	B-horizon with stagnic properties
E	albic horizon	BC	transitive B/C-horizon
EB	transitive E/B-horizon		
Scheme of idealised soil horizon stratification			
B2	middle B-horizon*	E	Eluvial horizon
B3	bottom B-horizon**	EB	E/B-horizon
BC	B/C-horizon	B1	upper B-horizon
The list of numeric soil properties and their coding (x stands in for each horizon of idealised stratification):			
Sg_x	Stagnic features {0,1,2,3}[1]	Ox_x, Red_x	oxidation and reduction features (%)
Lv_x	Luvic features {0,1,2,3}[2]	Snd_x, Clay_x	sand, clay (%)
Cb_x	Cambic features {0,1,2,3}[3]	Stn_x	Stoniness {1,2,3,4,5}[5]
X_x, Y_x, Z_x	colour {X, Y, Z}[4]		

*if soil has only one B-horizon, all parameters for B2 are identical to B1
**if soil has only one B-horizon, all parameters for B3 are identical to B1
if soil has only two B-horizons, all parameters for B3 are identical to B2
[1]0 – without stagnic features, 1 – weakly developed stagnic features (e.g. Bw(g)-horizon), 2 – moderately developed stagnic features (e.g. Bwg-horizon), 3 – strongly developed stagnic features (Bg-horizons)
[2]0 – without luvic features, 1 – weakly developed luvic features (e.g. Bw(t)-horizon), 2 – moderately developed luvic features (e.g. Bwt-horizon), 3 – strongly developed luvic features (Bt-horizons)
[3]0 – without cambic features, 1 – weakly developed cambic features (e.g. Bg(w)-horizon), 2 – moderately developed cambic features (e.g. Bgw-horizon), 3 – strongly developed cambic features (Bw-horizons)
[4]Munsell hue, chroma and value data were transferred into CIELab coordinates as introduced by Melville & Atkinson (1985)
[5]1 – without stones, 2 – less than 10%, 3 – 10 to 25%, 4 – 25 to 50%, 5 – more than 50%

β_l – parameter of l^{th} predictor in MLR,
p – number of predictors used for j^{th} STU,
$\varepsilon_j(s_i)$ – MLR residuum at sampled location s_i,
w_i – weight of punctual kriging operator (for more details on punctual kriging see e.g. Burgess & Webster 1980).

Following the theory of kriging (e.g. Burgess & Webster 1980), the weights (w_i) depend on the distances between the observations and the predicted location s_0 and the spatial relations between the sampled data around the predicted location. Whereas geographic distances are determined by X and Y coordinates as Euclidean distance, spatial relations are described by the experimental semivariogram (Burgess & Webster 1980):

$$\hat{\gamma}(h) = \frac{1}{2d(h)} \sum_{i=1}^{d}[\varepsilon(s_i) - \varepsilon(s_{i+h})]^2 \tag{3}$$

where:
$\hat{\gamma}(h)$ – semi-variance,
h – separation lag-distance between locations s_i and s_{i+h},
$\varepsilon(s_i)$, $\varepsilon(s_{i+h})$ – model residua at locations s_i and s_{i+h},
$d(h)$ – number of pairs at any separation distance h.

The semivariogram is a quantitative measure of how the variance between the sampled points is reduced as the separation distance decreases, and it can be modelled by some of the authorised semivariogram equations, such as gaussian or exponential (Webster & Oliver 2006). Finally, the weighting factors of Eq. (2) are estimated by solving the kriging equations (e.g. Webster & Oliver 2006). Both regression-kriging model (2) and semivariogram model (3) were solved in *R* open software (http://www.r-project.org; OLS for linear and exponential regression; punctual kriging). Target fuzzy *k*-partition to *k* STUs was interpolated to a grid with 10-m cell resolution. Alluvia of creeks were omitted from the model (refer to Figure 4) as they are not covered by the sampling, and the model is not calibrated for such areas.

The confusion index (*CI*) was used as the measure of ambiguity for the fuzzy partition (Burrough & McDonell 1998) in each cell of the GRID interpretation. It is described as $CI = 1 - (m_{max} - m_{max-1})$, where m_{max} is the maximum membership value and m_{max-1} is the next highest membership value in the cell. It is used to draw geographical boundaries (Burrough *et al.* 1997) between the analysed STUs as zones of confusion.

RESULTS AND DISCUSSION

Fuzzy *k*-means classification

The fuzzy *k*-mean classifier was parameterised to obtain the partition to totally three STUs. Following the output diagnostic features of centroids, the soils can be classified as Haplic Stagnosols (Albic, Dystric), Albic Stagnic Luvisols, and Haplic Cambisols (Skeletic, Dystric). Each STU is defined by the centroid profile (Table 3), which was built from the centroid values of the attributes in the classified attribute space. Albic Stagnic Luvisols are chiefly determined by argic Bt horizon developed from loess material. Some luvic features (luvic features of 1 or 2, see Table 1) occur also in deeper horizons of Stagnosols and Cambisols. This phenomenon relates to the relict processes that occurred in fragipan subsoil layer. Stagnosols are determined by hydromorphic Eg and Bg horizons with strongly developed stagnic features. Both oxidation and reduction signs occur also in some

Table 2. Basic statistics for predictor GIS data

Predictor	Mean	Median	Min.	Max.	Var.	*SD*	Skew.	Kurt.
DEM	384.6	371.7	291.6	584.1	4030.2	63.5	1.03	0.82
SLOPE	7.7	6.1	1.0	21.4	25.7	5.1	0.93	0.01
TWI	6.9	7.0	5.7	8.6	0.5	0.7	0.21	–0.53
NDVI	0.155	0.152	0.107	0.263	0.001	0.025	1.08	3.38
LS	4.5	2.9	0.2	20.4	16.5	4.1	1.56	2.63

DEM – digital elevation model, SLOPE – slope in degrees, TWI – topography wetness index, NDVI – normalised difference vegetation index, LS – length-slope factor

Table 3. Statistically central profiles of soil typological units; the result of fuzzy *k*-means classification

Haplic Stagnosols (Albic, Dystric)	
A	0–2 cm, topsoil A horizon
A/E	2–8 cm, transitive A/E-horizon
Eg	8–38 cm, albic horizon with stagnic properties (stagnic features Sg_E ~ 3, cambic features Cb_E ~ 0, luvic features Lv_E ~ 0), colour (X_E, Y_E, Z_E): 0.8049, 1.1097, 5.9355; 5% of oxidation features (Ox_E), 74% of reduction features (Red_E), 17% of clay (Clay_E), 45% of sand (Snd_E), < 10% of skelet (Stn_E ~ 2),
E/B	38–60 cm, transitive E/B-horizon (stagnic features Sg_EB ~ 3, cambic features Cb_EB ~ 0, luvic features Lv_EB ~ 0), colour (X_EB, Y_EB, Z_EB): 1.0816, 1.4899, 5.3624; 14% of oxidation features (Ox_EB), 62% of reduction features (Red_EB), 25% of clay (Clay_EB), 36% of sand (Snd_EB), 10–25% of skelet (Stn_EB ~ 3),
Bg	60–95 cm, B-horizon with stagnic features (stagnic features Sg_B1 ~ 3, cambic features Cb_B1 ~ 0, luvic features Lv_B1 ~ 0), colour (X_B1, Y_B1, Z_B1): 1.4174, 1.9502, 4.2124; 32% of oxidation features (Ox_B1), 66% of reduction features (Red_B1), 32% of clay (Clay_B1), 37% of sand (Snd_B1), 25–50% of skelet (Stn_B1 ~ 4)
Albic Stagnic Luvisols	
A	0–5 cm, topsoil A horizon
E	5–36 cm, albic E-horizon (stagnic features Sg_E ~ 0, cambic features Cb_E ~ 0, luvic features Lv_E ~ 1), colour (X_E, Y_E, Z_E): 2.0157, 2.8621, 5.6579; 1% of oxidation features (Ox_E), 1% of reduction features (Red_E), 16% of clay (Clay_E), 39% of sand (Snd_E), < 10% of skelet (Stn_E ~ 2),
E/B	36–47 cm, transitive E/B-horizon (stagnic features Sg_EB ~ 0, cambic features Cb_EB ~ 0, luvic features Lv_EB ~ 1), colour (X_EB, Y_EB, Z_EB): 1.5502, 2.2232, 4.0739; 7% of oxidation features (Ox_EB), 12% of reduction features (Red_EB), 15% of clay (Clay_EB), 24% of sand (Snd_EB), < 10% of skelet (Stn_EB ~ 2),
Btg	47–90 cm, argic Bt-horizon with moderate stagnic features (stagnic features Sg_B1 ~ 2, cambic features Cb_B1 ~ 0, luvic features Lv_B1 ~ 3), colour (X_B1, Y_B1, Z_B1): 2.9172, 3.9962, 5.2670; 51% of oxidation features (Ox_B1), 29% of reduction features (Red_B1), 28% of clay (Clay_B1), 33% of sand (Snd_B1), < 10 % of skelet (Stn_B1 ~ 2)
Haplic Cambisols (Skeletic, Dystric)	
A	0–8 cm, topsoil A horizon
Bw	8–96 cm, cambic B-horizon
B1	upper Bw-horizon (stagnic features Sg_B1 ~ 0, cambic features Cb_B1 ~ 3, luvic features Lv_B1 ~ 0), colour (X_B1, Y_B1, Z_B1): 2.2906, 3.3372, 5.4582; 23% of clay (Clay_B1), 39% of sand (Snd_B1), 10–25% of skelet (Stn_B1 ~ 3),
B2	bottom Bw-horizon with signs of agric horizon (stagnic features Sg_B2 ~ 0, cambic features Cb_B2 ~ 3, luvic features Lv_B2 ~ 1), colour (X_B2, Y_B2, Z_B2): 3.0738, 4.1229, 5.1540; 3% of oxidation features (Ox_B2), 1% of reduction features (Red_B2), 26% of clay (Clay_B2), 41% of sand (Snd_B2), 25–50% of skelet (Stn_B2 ~ 4)

Luvisol and Cambisol profiles, but they are not as dominant as in Stagnosols. Cambisol creates variable STU in the study area, for which cambic Bw horizon seems diagnostic. It has developed mostly from recent deluvium strata and shows high stoniness. The whole information on the centroids is readable in Table 3. The set of MVs demonstrates the fuzzy-like boundaries between the individual STUs (Figure 2). Diffuse and broad taxonomical boundaries occur especially between Cambisols and Stagnosols.

Regression-kriging interpolation

The deterministic component of regression-kriging (refer to the right side of Eq. (2) left to + sign) describes the distribution of MVs over the study area as the MLR or MER functions of the predictor GIS variables. As expected, the fuzzy *k*-partition shows deterministic responses to the predictor GIS data (see Table 4). The distribution of MVs to Stagnosols along DEM, SLOPE, TWI, and LS

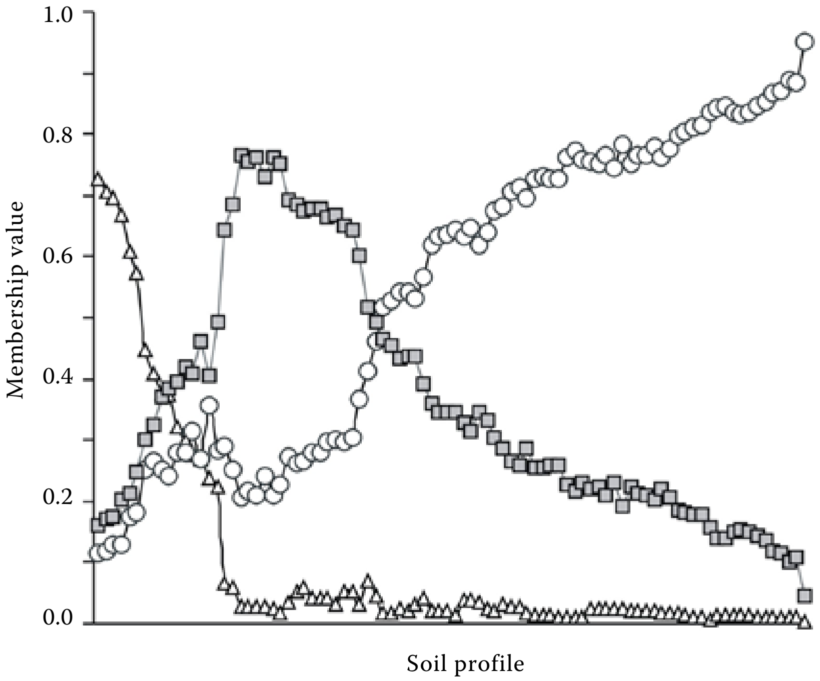

Figure 2. The distribution of membership values for soil typological units in the set of profiles (white triangles – Albic Stagnic Luvisols, dark squares – Haplic Stagnosols (Albic, Dystric), white circles – Haplic Cambisols (Skeletic, Dystric)); soil profiles order at X-axis is sorted to draw an optimal shape of membership data

is significantly described by the linear regression model, where R^2 reaches up to 37% of variance with SLOPE and TWI. Also the fuzzy k-partition to Cambisols is significantly defined by the linear regression model; DEM is the best explanatory variable (R^2 is about 56% of variance). Both Cambisols and Stagnosols seem to copy well the relief-based parameters in the study area. On the other hand, the fuzzy k-partition to Luvisols can not be statistically predicted by any of parameters but DEM, to which it shows exponential response (exponential regression model; R^2 = 44%). The data in Table 4 prove that a kind of deterministic spatial trend occurs in the data (linear or exponential), which is a function of the profile location on the relief gradient. It also shows that NDVI predictor is the weakest one of the GIS data used and is suboptimal to predict the soil cover in this study.

Since DEM-based relief parameters are strongly autocorrelated in the study area, their use in MLR or MER yields no significant improvement compared to simple regressions with the most explanatory GIS parameter (statistical test not shown here). Therefore, the single linear or exponential regression with the most explanatory variable was used instead of MLR/MER in Eq. (2). So SLOPE was used with MVs to Haplic Stagnosols (Albic, Dystric) and DEM with MVs to Haplic Cambisols (Skeletic, Dystric) and Albic Stagnic Luvisols.

The regression yields residua that are positive or negative depending on overestimating or underestimating the fuzzy k-partition at particular locations of the profile in Eq. (2). The estimation of the residua at the unsampled location (refer to the right side of equation (2) right to + sign) involves defining the semivariogram (Eq. (3)) from the located residua that helps to define weights, with which the residua of the individual soil profiles input the kriging estimator of Eq. (2). The spatial dependency in the residua of STUs is depicted by bounded semivariogram models (Figure 3). Here, the semi-variance depends on the distances between the locations of the profiles and not on the absolute geographic position, in contrast to the previous part of the model. It can be seen that the fuzzy k-partition residua demonstrate spatial autocorrelation for all STUs. The error of the spatial distribution model is caused by the nugget variance which is the semivariance at the zero between-sample distance. In the model, it represents the variation within the shortest sampling interval. The portion of scale (c1 parameter of the

Table 4. Regression analysis with predictor GIS data (N = 90)

Soil type	Predictor	Transformation	Model	R^2 (%)	R	P-level	B0	B1	C
Haplic Stagnosols (Albic, Dystric)	DEM	$\log_{10}$	LR	23.8	0.49	+++	4.048	–1.432	–
	SLOPE	$\log_{10}$	LR*	36.9	0.61	+++	0.665	–0.394	–
	TWI	no	LR	36.8	0.61	+++	–0.826	0.169	–
	NDVI	no	LR	8.2	0.29	+++	0.704	–2.255	–
	LS	$\log_{10}(1 + x)$	LR	33.3	0.58	+++	0.599	–0.387	–
Haplic Cambisols (Skeletic, Dystric)	DEM	$\log_{10}$	LR*	56.2	0.75	+++	–6.666	2.796	–
	SLOPE	$\log_{10}$	LR	46.7	0.68	+++	0.101	0.563	–
	TWI	no	LR	47.7	0.69	+++	2.250	–0.245	–
	NDVI	no	LR	14.3	0.38	+++	–0.044	3.798	–
	LS	$\log_{10}(1 + x)$	LR	42.7	0.65	+++	0.192	0.556	–
Albic Stagnic Luvisols	DEM	$\log_{10}$	EXP*	43.8	0.66	+++	44.188	–18.177	–0.019
	SLOPE	$\log_{10}$	EXP	8.3	0.29	–	2.985	–0.009	–19.551
	TWI	no	EXP	8.9	0.30	–	3.333	0.003	–28.450
	NDVI	no	EXP	5.1	0.23	–	0.058	–13.951	–0.027
	LS	$\log_{10}(1 + x)$	EXP	7.8	0.28	–	2.756	–0.011	–15.532

LR (linear regression): $Y = \beta_0 + \beta_1 x$, EXP (exponential regression): $Y = C + \exp(\beta_0 + \beta_1 x)$, R^2: coefficient of determination, R – Pearson's correlation coefficient , (+++) significant at α = 0.01, (–) not significant at α = 0.01; DEM – digital elevation model, SLOPE – slope in degrees, TWI – topography wetness index, NDVI – normalised difference vegetation index, LS – length-slope factor; *variable and regression type used for regression-kriging

semivariogram model, see Figure 3) and nugget variance (c0 parameter of the same model) defines the strength of the semivariogram model and, consequently, the validity of the residuum estimation. Figure 3 indicates that the models used for Cambisol and Luvisol STUs are more explanatory than the model used for Stagnosol residua which has a higher nugget than scale variance. Hence, the estimation of the residua for Stagnosols will be less accurate in this case. The semivariogram models analysed (Figure 3) were used to optimise the weights for kriging estimator of Eq. (2) for each unsampled location of the output map. Model (2) was solved for the grid with 10-m cell size where the extrapolation up to 250 m outside the sampling area (which is an average distance between the locations of soil profiles) was allowed so that the validity of the model might be fulfilled.

Final membership maps of the STUs studied (Figure 4) visualise the results of the model, which is the spatial distribution of STUs. It is obvious from the maps that the STUs studied geographically closely abut and it would be problematic to state other boundary than the diffuse one between these types of soils. The map of confusion index (Figure 4d) identifies where the diffuse boundaries occur (the highest confusion), and where distinct areas of the individual STUs are located. The fuzzy maps presented give a reliable picture of the STU distribution in the study area. Haplic Cambisols (Skeletic, Dystric) occur on convex geomorphologic units, whereas some luvic and stagnic features occur in Cambisols at the feet of hills (diffuse transition to Luvisols and Stagnosols). Stagnosols and Luvisols occur at the bottom parts of hillsides where the mountains meet the hilly country region. Haplic Stagnosols (Albic, Dystric) occur mostly in submontane depressions with seasonal surface water logging.

Cross validation

We used the cross-validation method (e.g. Johnston *et al.* 2001) to validate the model. The membership values of the individual profiles were correlated to the modelled estimations of MVs at the same

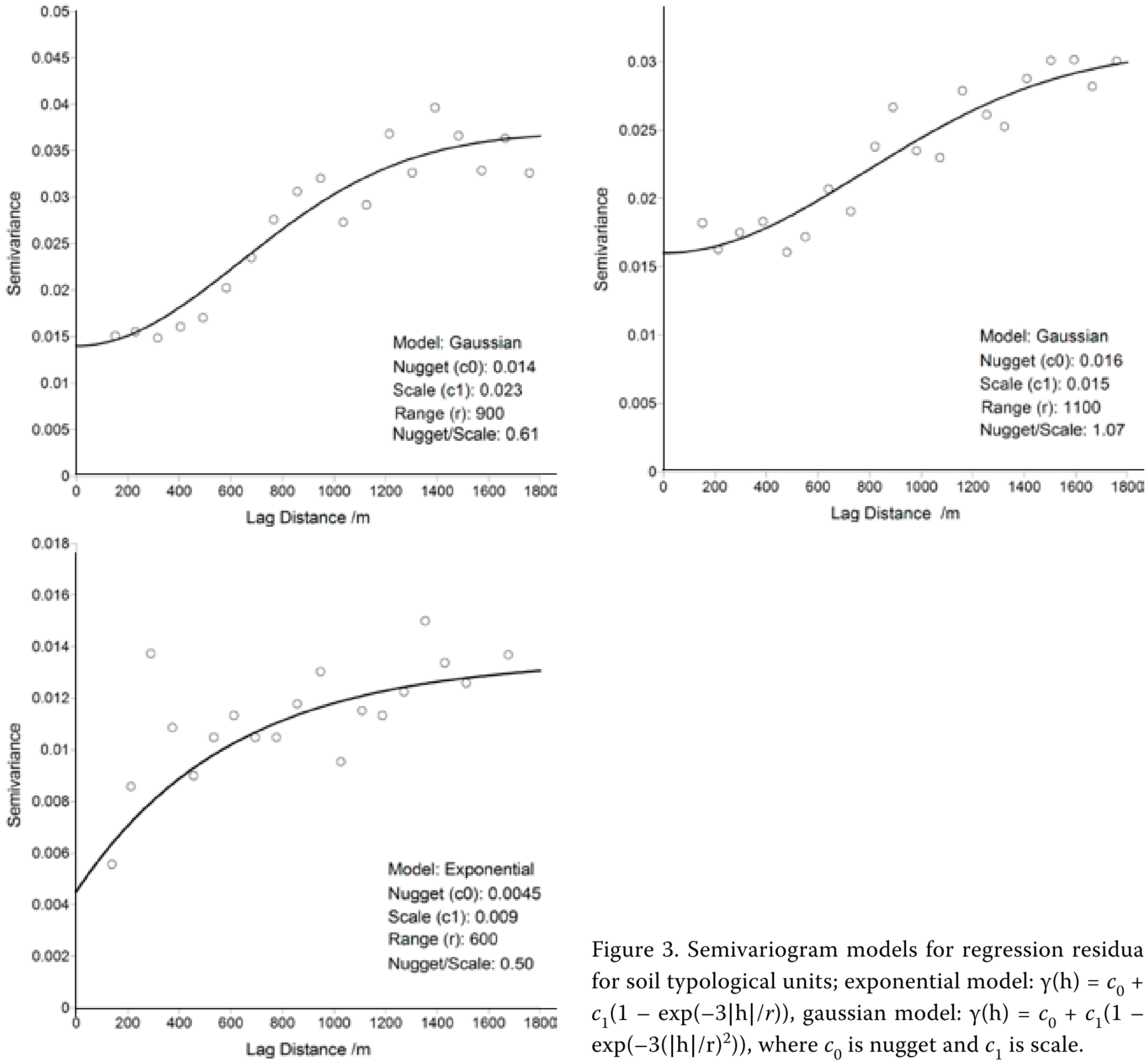

Figure 3. Semivariogram models for regression residua for soil typological units; exponential model: $\gamma(h) = c_0 + c_1(1 - \exp(-3|h|/r))$, gaussian model: $\gamma(h) = c_0 + c_1(1 - \exp(-3(|h|/r)^2))$, where c_0 is nugget and c_1 is scale.

profile locations. When running cross-validation, the profile at the just-analysed location was omitted from the model, so only (N – 1) profiles input the model at each of 90 soil profile locations. The cross-validation yields statistical data as shown in Table 5; Pearson's correlation is higher than 0.75 for all STUs. Addressing the noisy character of the soil data, the model appeared as a good spatial predictor and mapping agent for all STUs studied. It needs to be mentioned that this approach validates the model in its sub-optimal use, where the distance between the estimated cell in the map and the closest neighbouring profile will be about twice as long than in the case of the optimally parameterised model. Therefore, the correlation obtained by cross-validation may be assumed as slightly pessimistic with respect to the digital outputs.

The derived membership maps to STUs can slightly deviate from the assumption that the sum of MVs through STUs in each cell is equal to one (Eq. (1)) because each map is independently smoothed by the mapping algorithm. Further improvements can be therefore expected in the compositional parameterisation of the model (implementing compositional kriging). In order to visualise the results consistently, MV coverages were rescaled to meet the assumption of Eq. (1).

CONCLUSIONS

The paper offers a way to map digitally the ambiguous soil data in terms of their classification, and to analyse the relations between STUs and the digitally-described landscape variables.

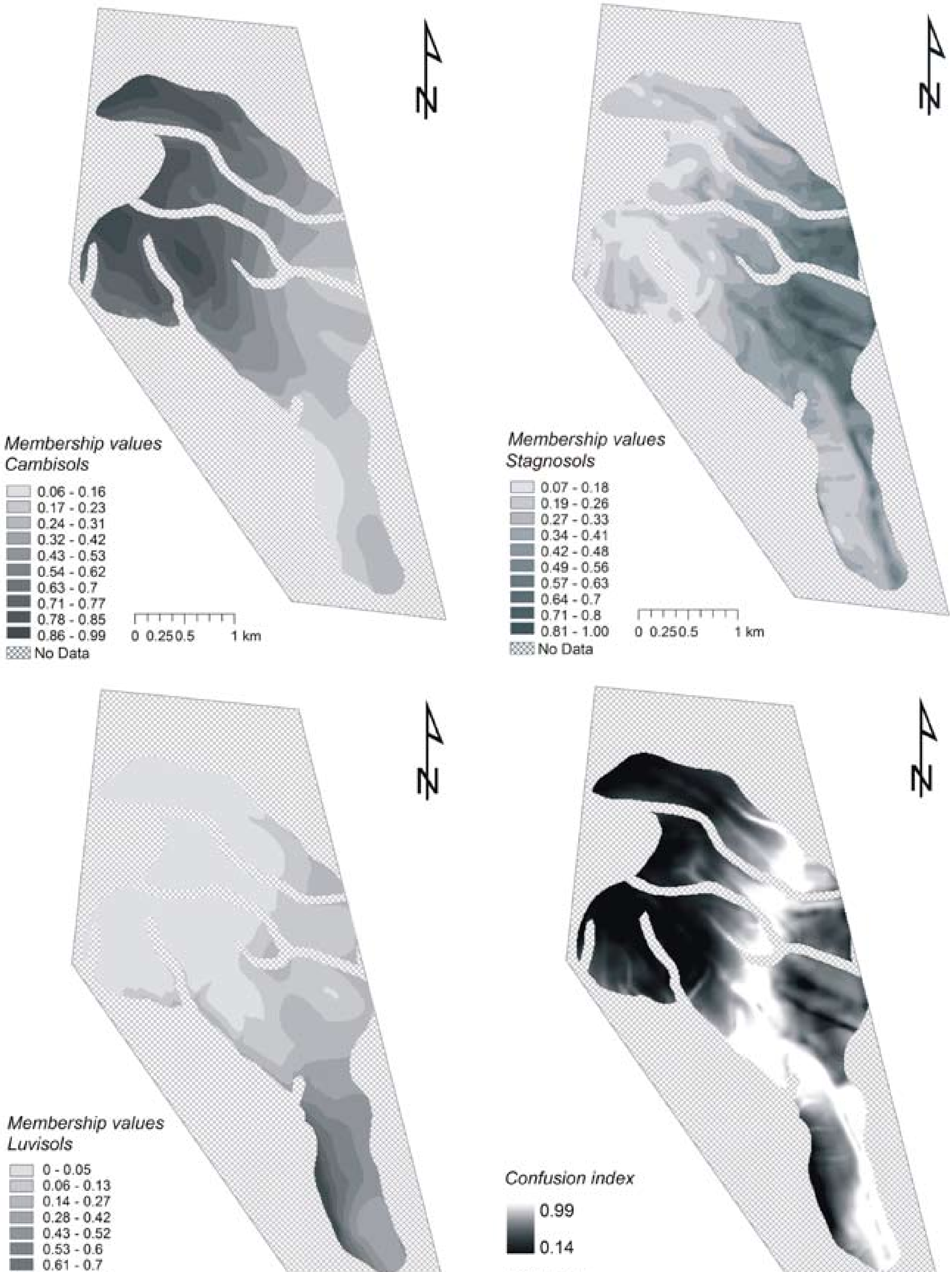

Figure 4. Membership maps of soil typological units

Table 5. Cross-validation statistics with soil typological units

Soil type	Model	N	R^2 (%)	R	P-level
Haplic Stagnosols (Albic, Dystric)	LIN	90	56.8	0.75	+++
Haplic Cambisols (Skeletic, Dystric)	LIN	90	79.4	0.89	+++
Albic Stagnic Luvisols	LIN	90	55.6	0.75	+++

LIN – linear regression model, N – number of observations, R^2 – coefficient of determination, R – Pearson's correlation coefficient, (+++) significant at α = 0.01

It shows that the systematic sampling of the soil profiles provides data that include a high degree of fuzziness in its attribute dataset, and which can be treated as fuzzy sets. It is also demonstrated that fuzzy *k*-partition to classified STUs can be optimised in such way as to produce smoothly geographically distributed data that yield significant correlations with the terrain variables. The predictor GIS-based data can support the spatial interpretation of MVs. The fuzzy partition data allow to include geostatistical methods, such as regression-kriging in this case, for their spatial interpretation. Confusion index spatially allocates the ambiguity in the classification obtained and draws an idea where the soil-profile data overlap the classification used. It seems to be a suitable technique also to mark geographical boundaries between the STUs studied as zones of confusion. We can conclude the following:

(*i*) – It appears that different STUs are mapped with different validity, which is dependant on the calibration of both regression and kriging components. To keep the validity of the maps for STUs, which demonstrate weak relations to the terrain-based variables (and are therefore highly sensitive to kriging component), a high sampling density is required.

(*ii*) – Membership maps slightly deviate from the assumption that the sum of MVs in each cell is equal to 1 as each map is independently smoothed by the mapping algorithm. Further improvements can be therefore expected in parameterisation of the model for all STUs in one run fulfilling the assumption of the compositional nature of MVs.

(*iii*) – The algorithm is not applicable for alluvia in the study area since it was not calibrated for such conditions. To solve this lack, additional sampling and model adjustments will be needed.

The fuzzy approach with the soil data could offer a wide range of interpretations for the soil and landscape studies (see e.g. Grunwald 2006). It provides opportunities to describe digitally soils and could potentially lead to spatial modelling of soils and landscapes. Moreover, the model produces results which can easily meet other landscape data in a GRID-based platform (e.g. various RS data), which are commonly used to analyse the landscape.

References

Bezdek J.C. (1981): Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York.

Bezdek J.C., Ehrlich R., Full W. (1984). FCM: the fuzzy c-means clustering algorithm. Computers and Geosciences, **10**: 191–203.

Burgess T.M., Webster R. (1980): Optimal interpolation and isarithmic mapping of soil properties. I. The semivariogram and punctual kriging. Journal of Soil Science, **31**: 315–331.

Burrough P.A., McDonnell R.A. (1998): Principles of Geographical Information Systems. Oxford University Press, New York.

Burrough P.A., Van Gaans P.F.M., Hootsmans R. (1997): Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma, **77**: 115–135.

Čemanová G., Harnová K., Balkovič J., Kollár J., Kromka M. (2005): Phytocoenological and soil inventory in the Nature Reserve of Čepušky and its surroundings (Považský Inovec Mountains). Phytopedon (Bratislava), **4**: 49–59.

De Gruijter J.J., McBratney A.B. (1988): A modified fuzzy k-means method for predictive classification. In: Bock H.H. (ed.): Classification and Related Methods of Data Analysis. Elsevier Science Publishers B.V., Amsetrdam.

De Gruijter J.J., Walvoort D.J.J., van Gaans P.F.M. (1997): Continuous soil maps – a fuzzy set approach to bridge the gaps between aggregation levels of process and distribution models. Geoderma, **77**: 169–195.

FAO (2006): World Reference Base for Soil Resources 2006 – A Framework for International Classification, Correlation and Communication. FAO, Rome.

Grunwald S. (ed.) (2006): Environmental Soil-landscape Modelling: Geographic Information Technologies and Pedometrics. CRC Press, Boca Raton, London, New York.

Hengl T., Heuvelink G.M.B., Stein A. (2004): A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma, **120**: 75–93.

Johnston K., Ver Hoef J.M., Krivoruchko K., Lucas N. (2001): Using ArcGIS Geostatistical Analyst. ESRI, Redlands.

Lagacherie P. (2005): Using a fuzzy pattern matching algorithm for allocating soil individuals to pre-existing soil classes. Geoderma, **128**: 274–288.

Lapin M., Faško P., Melo M., Šťastný P., Tomlain J. (2002): Climatic Regions. In: Atlas of the Slovak Republic. 1st Ed., MŽP SR, SAŽP, Bratislava, Banská Štiavnica. (in Slovak)

Maselli F., Gilabert A.M., Conese C. (1998): Integration of high and low resolution NDVI data for monitoring vegetation in Mediterranean environments. Remote Sensing Environments, **63**: 208–218.

McBratney A.B., Moore A.W. (1985): Application of fuzzy sets to climatic classification. Agricultural and Forest Meteorology, **35**: 165–185.

Melville M.D., Atkinson G. (1985): Soil color: Its measurements and its designation in models of unifoem color space. Journal of Soil Science, **36**: 495–512.

Minasny B., McBratney A.B. (2002): FuzME Version 3. Australian Centre for Precision Agriculture, The University of Sydney, Sydney.

Odeh I., McBratney A., Chittleborough D. (1994): Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma, **63**: 197–214.

Webster R., Oliver M.A. (2006): Modeling spatial variation of soils as random functions. In: Grunwald S. (ed.): Environmental Soil-landscape Modelling: Geographic Information Technologies and Pedometrics. CRC Press, Boca Raton, London, New York.

Wilson J.P., Gallant J.C. (eds) (2000): Terrain Analysis: Principles and Applications. John Wiley & Sons, New York.

Zadeh L. (1965). Fuzzy sets. Information and Control, **8**: 338–353.

Received for publication July 12, 2007
Accepted after corrections September 6, 2007

Corresponding author:

RNDr. Juraj Balkovič, Ph.D., Univerzita Komenského v Bratislave, Prírodovedecká fakulta, Katedra pedológie, Mlynská dolina, pavilón B2, 842 15 Bratislava, Slovenská republika
tel.: + 421 260 296 662, e-mail: balkovic@fns.uniba.sk