Method development to extract spatial association structure from soil polygon maps

István SISÁK[1], Mihály KOCSIS[1], András BENŐ[1] and Gábor VÁRSZEGI[2]

Abstract

Existing soil information systems contain mainly qualitative data on soilscapes, however, quantitative data would be necessary to more effectively guide digital soil mapping efforts. Detailed analysis of small scale overview maps offers the most appropriate way to delineate soilscapes where they are available. In our study, the genetic soil map of Hungary have been used which displays the most complete representation of the Hungarian Soil Classification System. Our goal was to analyse spatial association structure based on the boundary segments between soil polygons. We transformed the polygons into lines. The features of each line segment were the names (or codes) of the soil polygons on both sides. After omission soils with low representation (less than three polygons) and boundaries beside state border, forests and cities, 69 soil units were retained. We calculated a similarity matrix among soil types based on logarithm of ratios between existing segment lengths and theoretical segment lengths. The theoretical lengths were calculated with a Chi-squared calculation by using sums of lengths in rows and columns in the 69 × 69 matrix. The similarity matrix was converted into dissimilarity matrix to distinguish between complete dissimilarity (missing values) and complete similarity (main diagonal). Dissimilarity matrix was clustered and represented in a form of dendrogram both in original form and after dimension reduction with multidimensional scaling method. Our method has resulted a promising approach for delineating soilscapes in presence of overview soil maps. The study resulted fuzzy soilscapes with broad transition zones. The method could be refined by using variable sized moving window method and by combining boundary data with terrain, geology etc.

Keywords soilscape quantification, genetic soil map of Hungary, boundary segment based Chi-squared calculation, hierarchical clustering, multidimensional scaling

Introduction

Since the work of DOKUCHAEV, the axiom of the soil science is that soil forming factors (climate, geology, hydrology, biota, elevation, time and humans) and their specific interaction determine soil formation and soil properties. JENNY, H. (1941) suggested that these complex relationships should be described with mathematical formulas thus, qualitative and quantitative

[1] Department of Plant Production and Soil Science, Georgikon Faculty, University of Pannonia, H-8360 Keszthely, Deák F. u. 16. E-mails: talajtan@georgikon.hu, kocsis.mihaly@2010.georgikon.hu, beno.andras@gmail.com
[2] Department of Agro-environment Coordination, Directorate of Plant and Soil Protection and Agro-environmental issues, National Food Chain Safety Office H-1024 Budapest, Keleti K. u. 24.
E-mail: varszegig@nebih.gov.hu

soil properties will be predictable. McBratney, A.B. *et al.* (2003) gave an overview on digital soil mapping (DSM) which is Jenny's idea put into practice with help of GIS software and geostatistical analysis.

There is a tremendous complexity of soil associations in some landscapes and this requires segmentation of landscapes into soilscapes as a basis for digital soil mapping (McBratney, A.B. *et al.* 1991; Lagarcherie, P. *et al.* 2001; Schmidt, K. *et al.* 2010). Soilscape is a term introduced by Buol, S.W. *et al.* (1973) and conceptually extended by Hole, F.D. (1978) in the context of pedology. According to Lagarcherie, P. *et al.* (2001) soilscape is a landscape unit including a limited number of soil classes that are geographically distributed according to an identifiable pattern. Very often, mapping soilscapes from soil forming factor maps is more realistic than mapping soil classes. The primary task in mapping larger areas should be to account for these spatial soil-association patterns as a basis to segment landscapes (Schmidt, K. *et al.* 2010).

McSweeny, K. *et al.* (1994) proposed to set up a hierarchical multistage strategy to explain the variability of soils and soil properties in space. The second stage of the proposed method was a geomorphometric characterization of the landscape from digital terrain models, which provides (i) a land surface representation to which other data are referenced and (ii) a division of the land surface into areas that correspond with soil patterns. The recently adapted hierarchical approach to define soilscapes follows the World Soils and Terrain Digital Database (SOTER) methodology (ISRIC, 1993). SOTER has become widely evaluated in European and broader context (Dobos, E. *et al.* 2001, 2005, 2010). However, these terrain-based approaches are more appropriate for finer scales as they mainly focus on deriving terrain facets instead of deriving larger homogeneous geomorphological or pedological regions (Schmidt, K. *et al.* 2010).

Existing soil information systems store data on association of soil bodies within soilscapes or soil series in relation tables. This description is strictly qualitative (Finke, P. *et al.* 2001). Efforts have been made to better define the objects resulting from these groupings (Hewitt, A.E. 1993) and to define the criteria used in their construction (Hudson, B.D. 1990). Recent findings provide more and more quantitative results on how soil bodies are associated (Behrens, T. *et al.* 2009; Hewitt, A.E. *et al.* 2010; Schmidt, K. *et al.* 2010). The latest nationwide digital soil mapping projects in New Zealand (Hewitt, A.E. *et al.* 2010) or Ireland (Creamer, R. *et al.* 2014) adapt strong soilscape-based approach.

In spite of the recent trend (Scull, P. *et al.* 2005) that predictive soil models shift from research to operational phase, Grinand, C. *et al.* (2008) observed that soil class prediction accuracy can only be approximated correctly if test samples are collected at a certain distance from the training samples when predicting unvisited areas.

However, digital soil mapping approaches which utilize soil information from existing (usually small or medium scale) soil maps and field observations perform much better than pure theoretical constructions (Mendonça-Santos, M.D.L. *et al.* 2008). Soil maps are physical representations of the mental models of the mappers on how soil forming factors interact (Bui, E. 2004). They provide us a path through the almost infinite number of theoretically possible combinations to the most probable outcome. In countries where small or medium scale soil maps exist their statistical analysis may help to define homogenous soil regions or soilscapes and representative areas for detailed soil surveys (Behrens, T. *et al.* 2009; Schmidt, K. *et al.* 2010).

The aim of our study was to evaluate an existing nationwide soil map of Hungary and to define soil association rules which then can be used to delineate soil regions or soilscapes. We evaluated boundary line segments of neighbouring polygons and we were using Chi-squared method, hierarchical classification and multidimensional scaling in the analysis.

Materials and methods

The genetic soil map of Hungary and the conversion of its categories into WRB categories

There was a nationwide campaign in Hungary in the 1970's and 1980's to renew the old land evaluation system based on detailed new soil maps. The genetic soil map (MÉM-NAK, 1983) was released as a part of the preparation phase for the fine-scale soil mapping. The purpose of the 1:200,000 scale map was to gather all the available information and to give orientation for the field work before the detailed soil surveys. The latest field guide for soil mapping and an official version of the Hungarian Soil Classification System (HSCS) was published (Horváth, B. *et al.* 1987) as part of the project and it served as a compulsory tool for field surveyors. Soil classification system did not change much between 1983 (release of the genetic map) and 1989 (release of the field guide). Slight changes were introduced but basic concepts and categories stayed intact. The genetic soil map is the most complete display of the HSCS and also contains data on parent material, texture and chemical reaction but does not show soil data for the area of forests and larger towns. We completed and improved the digital version (AIR, 2013) of the genetic soil map of Hungary. We used only soil classes of HSCS (soil types, sub-types) in our analysis and did not use other data.

In *Table 1* we provide an approximate conversion between HSCS soil units of the genetic soil map (MÉM-NAK, 1983) based on the work of Horváth, B. *et al.* (1989) and the IUSS Working Group WRB (2007). We should state that clear one-to-one conversion is not possible at all because of the different soil investigation methods, different limit values of the individual properties and partly because of the different concepts. We still decided to use this conversion since one of the declared primary objectives of the WRB is to serve as "common language" between national soil classification systems.

Despite limitations, approximate conversion is possible (Michéli, E. *et al.* 2006; Krasilnikov, P. *et al.* 2009). We applied the following procedure:

1. We considered the basic concepts of the Reference Soil Groups (RSGs) and their qualifiers and specifiers and we used them to express similar concepts in the HSCS without strict investigations of the detailed definitions and limits.

2. Whenever the HSCS expressed properties which were not part of the specifier set of the given RSG, we used similar specifiers from other RSGs but we added them in *italics*.

3. If the Hungarian concept was not included in the WRB concepts, we added a short explanation in *italics*.

Codes are also an easy way to identify soil units in the figures and tables. We decided to provide approximate categories of an earlier version of the WRB (IUSS Working Group WRB, 2007) because this has been well known in the soil science community. Newly introduced changes (IUSS Working Group WRB, 2014) may not be well established beyond experts in soil classification.

The HSCS contains 99 individual units either as soil types (e.g. 10 Lithic Leptosol) or sub-types (e.g. 31 Haplic Regosol, Calcaric). The code of the soil types can be divided by ten without remainder (see *Table 1*). The codes of the sub-types contain numbers in the place of the last digit other than zero. The MÉM-NAK (1983) soil map displays 81 different soil units. However, some of them are represented only by three or less polygons and those were excluded from our analysis. On this way, 69 soil units were retained and converted into approximate WRB units *(Table 1)*.

Data analysis

In the first step we determined the length of each line segment between the soil category polygons (soil types or sub-types). The boundary lines at the state border or in the neighbourhood of forests, lakes or towns were not considered since only one of the

Table 1. Approximate conversion of the units in the Hungarian Soil Classification System into WRB categories

Soil unit codes[1]	Soil unit names in the Hungarian Soil Classification System[1] (in Hungarian in the brackets)	Approximate equivalent in the WRB classification[2]
10	Stony, rocky skeletal soil (Köves sziklás váztalaj)	(Nudi-)Lithic Leptosol
20	Gravelly skeletal soil (Kavicsos váztalaj)	Hyperskeletic Leptosol
31	Calcareous earthy barren (Karbonátos földes kopár)	Haplic Regosol, Calcaric
41	Calcareous blown sand (Karbonátos futóhomok)	Protic Arenosol, Aridic, Calcaric
42	Non-calcareous blown sand (Nem karbonátos futóhomok)	Protic Arenosol, *not calcareous*
45	„Kovárvány" blown sand (Kovárványos futóhomok)	Lamellic Arenosol
51	Calcareous humic sand (Karbonátos humuszos homok)	Haplic Arenosol, Calcaric
52	Non-calcareous humic sand (Nem karbonátos humuszos homok)	Haplic Arenosol, *not calcareous*
53	Calcareous multilayered humic sand (Karbonátos több rétegű humuszos homok)	Haplic Arenosol, Calcaric *with buried A horizon(s)*
54	Non-calcareous multilayered humic sand (Nem karbonátos több rétegű humuszos homok)	Haplic Arenosol *not calcareous with buried A horizon(s)*
60	Humus-carbonate soil (Humuszkarbonát talaj)	Haplic Regosol, Humic, Calcaric
71	Black rendzina (Fekete rendzina)	Rendzic Leptosol *black usually on limestone*
72	Brown rendzina (Barna rendzina)	Rendzic Leptosol *brown usually on dolomite*
112	Nonpodzolic brown forest soil with clay illuviation (Nem podzolos agyag-bemosódásos barna erdőtalaj – BET)	Haplic Luvisol
121	Podzolic pseudogley brown forest soil (Podzolos pszeudoglejes BET)	Albic Stagnic Luvisol, Manganiferric
122	Pseudogley brown forest soil with clay illuviation (Agyagbemosódásos pszeudo-glejes BET)	Stagnic Luvisol
131	Typical Ramann's brown forest soil (Típusos Ramann-féle BET)	Haplic Cambisol, Eutric, Siltic
132	Rust brown Ramann's brown forest soil (Rozsdabarna erdőtalaj)	Brunic Arenosol, Eutric, Chromic
141	Typical brown forest soil with alternating thin layers of clay substance („kovárvány") (Típusos kovárványos BET)	Brunic Lamellic Arenosol
143	„Kovárvány" brown forest soil with clay illuviation (Agyagbemosódásos kovárványos BET)	Lamellic Luvisol
161	Calcareous chernozem brown forest soil (Karbonátos csernozjom BET)	Haplic Cambisol, Humic, Calcaric
162	Non-calcareous chernozem brown forest soil (Nem karbonátos csernozjom BET)	Haplic Cambisol, Humic
171	Calcareous chernozem soils with forest residues (Karbonátos erdőmaradványos csernozjom)	Luvic Phaeozem, Calcaric
172	Non-calcareous chernozem soils with forest residues (Nem karbonátos erdőmaradványos csernozjom)	Luvic Phaeozem, *not calcareous*

Table 1. Continued

Soil unit codes[1]	Soil unit names in the Hungarian Soil Classification System[1] (in Hungarian in the brackets)	Approximate equivalent in the WRB classification[2]
180	Leached chernozem soil (Kilúgzott csernozjom)	Haplic Chernozem, Pachic
191	Typical calcareous chernozem soil (Típusos meszes vagy mészlepedékes csernozjom)	Calcic Chernozem, Pachic, Siltic
192	Lowland calcareous chernozem soils (Alföldi meszes vagy mészlepedékes csernozjom)	(Endosalic) Calcic Chernozem, Pachic
201	Calcareous meadow chernozem soils (Karbonátos réti csernozjom)	*Bathygleyic* Calcic Chernozem, Pachic
202	Non-calcareous meadow chernozem soil (Nem karbonátos réti csernozjom)	*Bathygleyic* Chernozem, Pachic, *not calcareous*
203	Meadow chernozem soil, salty in deeper horizons (Mélyben sós réti csernozjom)	Endosalic *Bathygleyic* Chernozem, Pachic
204	Meadow chernozem soils, solonetz-like in deeper horizons (Mélyben szolonyeces réti csernozjom)	*Bathygleyic* Chernozem, Pachic *with sodicity in the parent material*
205	Solonetz-like meadow chernozem soil (Szolonyeces réti csernozjom)	*Bathygleyic* Chernozem, Pachic *with sodicity in the subsurface soil horizon*
211	Calcareous terrace chernozem soil (Karbonátos terasz csernozjom)	Calcic Endofluvic Chernozem
221	Calcareous solonchak soil (Karbonátos szoloncsák)	Calcic Solonchak, Carbonatic
231	Calcareous solonchak-solonetz soil (Karbonátos szoloncsák-szolonyec)	Calcic Salic Solonetz *(Carbonatic)*
232	Calcareous and sulphate-containing solonchak-solonetz soil (Karbonátszulfátos szoloncsák-szolonyec)	Calcic Salic Solonetz *(Carbonatic Sulphatic)*
241	Shallow meadow solonetz soil (Kérges réti szolonyec)	Calcic Solonetz *with an A horizon shallower than 7 cm*
242	Medium meadow solonetz soil (Közepesen mély réti szolonyec)	Calcic Solonetz *with an A horizon between 7 and 20 cm*
243	Deep meadow solonetz soil (Mély réti szolonyec)	Calcic Solonetz *with an A horizon deeper than 20 cm*
251	Medium meadow solonetz turning into steppe formation (Közepesen mély sztyepesedő réti szolonyec)	Mollic Calcic Solonetz
281	Sulphate- or chloride-containing solonchak-like meadow soils (Szulfátos vagy kloridos szoloncsákos réti talaj)	Mollic Gleysol *with salt accumulation in the surface horizon (Sulphatic or Chloridic)*
282	Calcareous solonchak-like meadow soils (Karbonátos szoloncsákos réti talaj)	Mollic Gleysol, Calcaric *with salt accumulation in the surface horizon (Carbonatic)*
291	Solonetz-like meadow soils (Szolonyeces réti talaj)	Mollic Gleysol, (Hypo-)sodic
292	Strongly solonetzized solonetz-like meadow soils (Erősen szolonyeces réti talaj)	Mollic Gleysol, Sodic
301	Calcareous meadow soils (Karbonátos réti talaj)	Calcic Mollic Gleysol
302	Non-calcareous meadow soils (Nem karbonátos réti talaj)	Mollic Gleysol, *not calcareous*

Table 1. Continued

Soil unit codes[1]	Soil unit names in the Hungarian Soil Classification System[1] (in Hungarian in the brackets)	Approximate equivalent in the WRB classification[2]
303	Meadow soils, salty in deeper horizons (Mélyben sós réti talaj)	Endosalic Mollic Gleysol
311	Calcareous alluvial meadow soils (Karbonátos öntés réti talaj)	Fluvic Mollic Gleysol, Calcaric
312	Non-calcareous alluvial meadow soils (Nem karbonátos öntés réti talaj)	Fluvic Mollic Gleysol, *not calcareous*
321	Typical marshy meadow soils (Típusos lápos réti talaj)	Histic Gleysol
331	Calcareous chernozem meadow soils (Karbonátos csernozjom réti talaj)	Calcic Gleyic Chernozem, (Pachic)
332	Non-calcareous chernozem meadow soils (Típusos csernozjom réti talaj)	Gleyic Chernozem, (Pachic)
333	Chernozem meadow soils,salty-like in deep-er layers (Mélyben sós csernozjom réti talaj)	Endosalic Gleyic Chernozem
334	Chernozem meadow soils, solonetz-like in deeper layers (Mélyben szolonyeces csernozjom réti talaj)	Gleyic Chernozem *with sodicity in the subsurface soil horizon*
335	Solonetz-like chernozem meadow soils (Szolonyeces csernozjom réti talaj)	Gleyic Chernozem *with sodicity in the parent material*
350	Peat-bog soils (Rétláp talaj)	Fibric Histosol
360	Drained and cultivated lowmoor fen soils (Lecsapolt és telkesített rétláp talaj)	Histosol, Drainic *in general*
361	Drained peat-bog soil (Lecsapolt tőzegláp)	Hemic Histosol Drainic
362	Drained peaty fen soil (Lecsapolt tőzeges láp)	Hemic Histosol, Drainic *with less than 50 cm deep Histic horizon*
363	Drained fen soil with highly decomposed peaty substance 'kotu' (Lecsapolt kotus láp)	Sapric Histosol, Drainic
364	Cultivated low moor fen soil (Telkesített rétláp)	Hemic Histosol, Drainic *with regulated water level*
370	Soils of marsh and alluvial forests (Mocsári erdőtalaj)	Haplic Gleysol, Dystric (alternative: Acric Gleysol)
381	Calcareous recent alluvial soils (Karbonátos nyers öntés talaj)	Haplic Fluvisol, Calcaric
391	Calcareous humic alluvial soils (Karbonátos humuszos öntés talaj)	Haplic Fluvisol, Humic, Calcaric
392	Non-calcareous humic alluvial soils (Nem karbonátos humuszos öntés talaj)	Haplic Fluvisol, Humic, *not calcareous*
393	Calcareous multilayered humic alluvial soils (Karbonátos, több rétegű humuszos öntés talaj)	Haplic Fluvisol, Humic, Calcaric *with buried A horizon in the upper 150 cm*
394	Non-calcareous multilayered humic alluvial soils (Nem karbonátos több rétegű humuszos öntés talaj)	Haplic Fluvisol, Humic *with buried A hori-zon in the upper 150 cm, not calcareous*
395	Meadow-like humic alluvial soils (Réti öntés talaj)	Mollic Gleyic Fluvisol
402	Slope deposits of forest soils (Erdőtalaj eredetű lejtőhordalék talajok)	Colluvic Regosol *derived mainly from Luvisols and Cambisols*

[1]According to Horváth, B. *et al.* 1989, [2]According to IUSS Working Group WRB 2007.

neighbouring polygons had soil data. Then we calculated the sum of lengths for each soil category combinations and thus, we got a square matrix with dimensions of 69 by 69.

The values in the main diagonal were dismissed (set to zero) since they represented the same category with slightly different properties (texture or pH). Then we calculated the following theoretical length for each matrix element:

$$L_{ij\text{-}est} = \Sigma L_i \times \Sigma L_j / L_{tot},$$

where $L_{ij\text{-}est}$ = the estimated length for an individual category combination, L_i = the total length of the *i*-th category in the rows of the matrix, L_j = the total length of the *j*-th category in the columns of the matrix, L_{tot} = the total length of all categories (grand total of the matrix).

Then we have calculated the following *P* similarity (neighbourhood) matrix:

$$P_{ij} = \log [(L_{ij} / L_{ij\text{-}est}) \times 100],$$

where L_{ij} = the actual length for an individual category combination. This is the logarithm of the percent ratio between actual and theoretical lengths. Zero values in the main diagonal and missing combinations have no logarithm thus, in this similarity matrix we cannot distinguish between complete similarity (main diagonal) and complete dissimilarity (non-existent combinations). To alleviate this problem, we converted the similarity matrix into $P`_{ij}$ dissimilarity (distance) matrix. All length ratios were less than 100,000 thus, we selected 5 (= log 100,000) as the maximum dissimilarity.

$$P'_{ij} = \begin{cases} 0 & \text{if } P_{ij} = 0 \\ 5 - P_{ij} & \text{if } P_{ij} > 0 \\ 5 & \text{if } P_{ij} = \text{missing} \end{cases}$$

We performed hierarchical cluster analysis with $P`_{ij}$ matrix and presented the results in form of a dendogram. The dimensionality of this matrix is 69 with regard to the soil categories as variables. However, the dissimilarity matrix had several missing combinations and we assumed that the dimensionality can be significantly reduced without much loss of information. We applied the multidimen-

sional scaling procedure to find a simpler and more general structure. Then we applied the hierarchical clustering to the new matrix again and represented the results with another dendogram. We used ArcGIS 10.0 for map data handling and interpretation and SPSS 13.0 for data analysis.

Results and discussion

The frequency distribution of the $P`_{ij}$ distance (dissimilarity) matrix has been shown in *Figure 1* without the values of 5 and 0. The histogram was calculated from the full matrix which means that all values are in duplicate. The distribution is close to the normal. For the half matrix when each combination is considered only once, there are 2,346 possible combinations between 69 soil categories but only 779 of them (33.2%) really exist which means that soil categories can be neighbours of only a subset of other categories which is trivial.

Chi-square statistics are often used for overlaid categorical maps in land use change studies (Pontius, Jr. R.G. 2002). However, the appropriateness of method drew also criticism because mapped area has no clear, statistically independent "case" thus, its error model is flawed (Chrisman, N.R. 1989) and the pixel size or the area of measurement unit will determine the "degree of freedom"
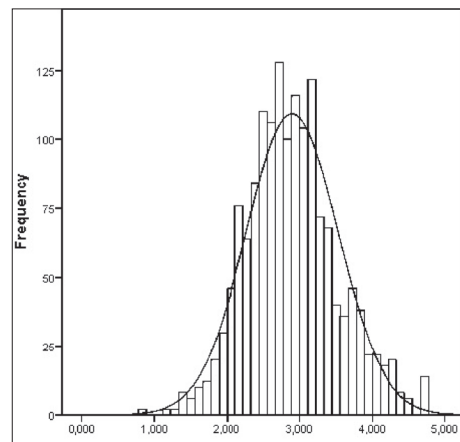

Fig. 1. Data distribution in the dissimilarity matrix

in the test. Similar objections are true for Chi-squared statistics with line segments. However, we did not use the Chi-square calculation in our study to test any significance; we just calculated the P_{ij} matrix elements from segment lengths in a similar way as in Chi-square method without entering into the questioned test calculation.

The resulting dendogram calculated from the first, not simplified distance matrix can be seen in *Figure 2*. After reducing the dimensionality with the PROXSCALE procedure, we got 5 dimensions instead of the previous 69 whereby 7 percent of the information was lost as indicated by the stress-test of the procedure. The second hierarchical clustering with the reduced, five-dimensional matrix has resulted the dendogram shown in *Figure 3*.

There are numerous differences between the two dendograms but generally, the second one has a much more separated structure between the branches than the first one.

The following two soil types are loosely associated with each other and they are rather separated from other categories in the first dendogram (*Figure 2*):

202: Bathygleyic Chernozem, Pachic, *not calcareous,*

301: Calcic Mollic Gleysol.

They lost their separation from other branches, but retained some degree of their association as members of the same group (cluster 3c in *Figure 3*) after dimensionality reduction, however, they were directly associated with other soil categories:

202: Bathygleyic Chernozem, Pachic, *not calcareous,*

363: Sapric Histosol, Drainic,

364: Hemic Histosol, Drainic *with regulated water level* and

301: Calcic Mollic Gleysol,

172: Luvic Phaeozem, *not calcareous.*

The dimensionality reduction may bring forward relationships which explain soil formation processes such as Stagnic Luvisol (112) became associated with Colluvic Regosol *derived mainly from Luvisols and Cambisols* (402) in cluster 5b (*Figure 3*) which association was not so close in the first dendogram (*Figure 2*).

There are very closely related soil categories which, in theory, should express different degree of groundwater influence coupled with strong organic matter accumulation such as Bathygleyic Chernozems (201–205) and Gleyic Chernozems (331–335) as seen in *Figure 3* (clusters 1 and 2c). However, even the latest official field guide (HORVÁTH, B. *et al.* 1989) does not provide enough support to tell them apart in the field. Our analysis points out specific weaknesses in the HSCS which need more precise definitions as part of the necessary future development of the HSCS according to the diagnostic principles (MICHÉLI, E. *et al.* 2006; KRASILNIKOV, P. *et al.* 2009). *Figure 4* shows the map of soil clusters indicated in *Figure 3*.

There is a clear regional distribution of clusters within the area of the country. The clusters marked with "A" are situated on the Great Plain (South-East part of Hungary) and to lesser extent on the Small Hungarian Plain (North-West part). Most of the clusters marked with "D" are situated on the hilly regions with some remarkable exceptions (D_3b and D_4b) which are associated with sandy regions and large rivers on the Great Hungarian Plain. The lead soil types within the clusters are provided in *Table 2*. At that, we followed the method of SCHMIDT, *et al.* (2010) instead of trying to characterize the complete soil associations. Further investigation of the association rules and their regional differences can be the objective of future studies. The major soil type gives more than 2/3 of the area within the cluster in five clusters, this ratio is between 1/3 and 2/3 in four clusters and it is below 1/3 in two clusters. The latter two are on lowland where the genetic soil map shows larger pedodiversity.

Close proximity in the dendogram may originate from strong association in one region but in other region this relatedness does not exist sometimes simply because one of the soil categories is not present in the other region. This observation is most striking for the cluster 5a (Mollic Gleysol, *not calcareous* and associated soils). Stagnic Luvisols are included in this cluster (code 121 and 122) and they are
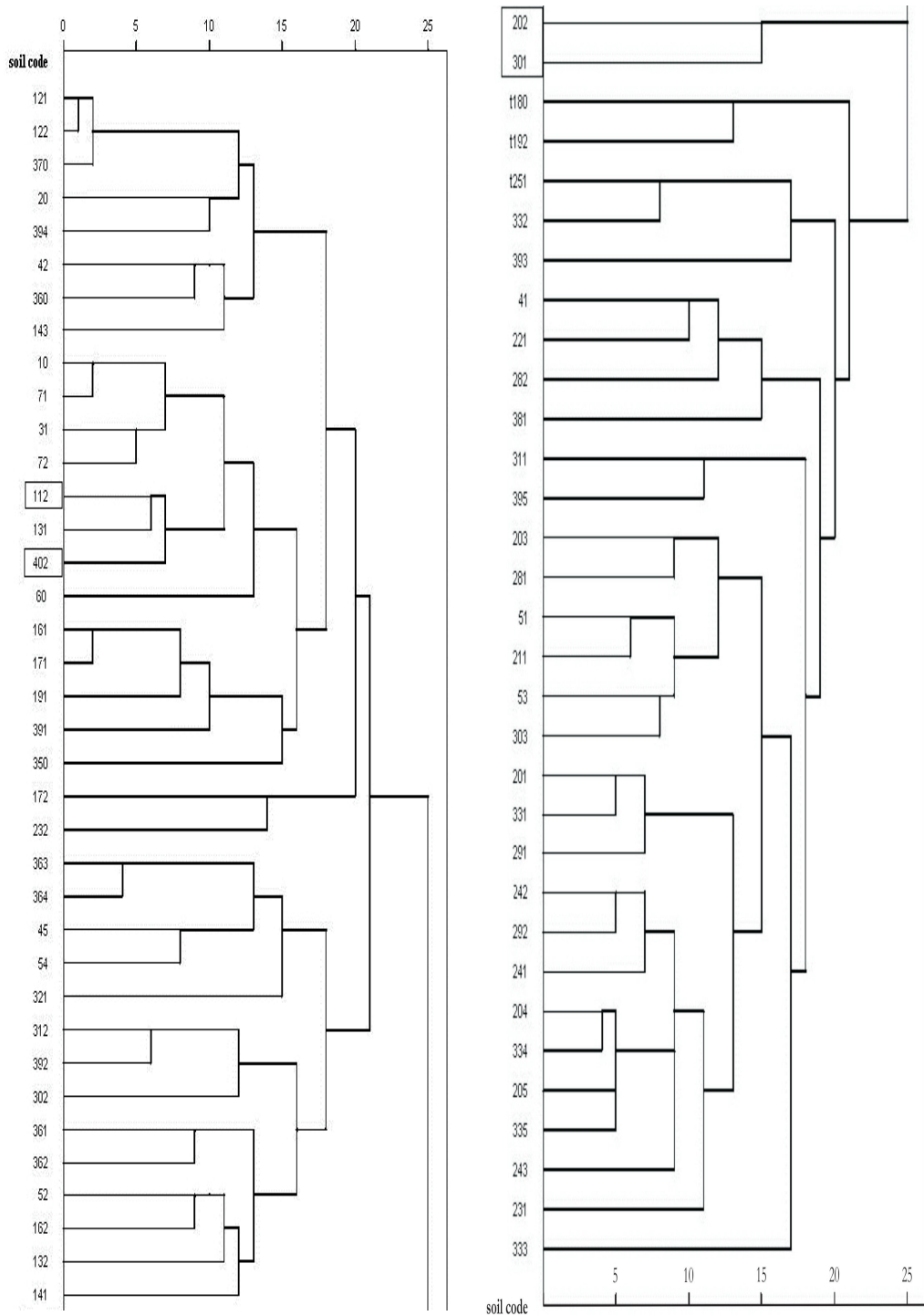
Fig. 2. Dendogram derived by hierarchical clustering from the original dissimilarity matrix
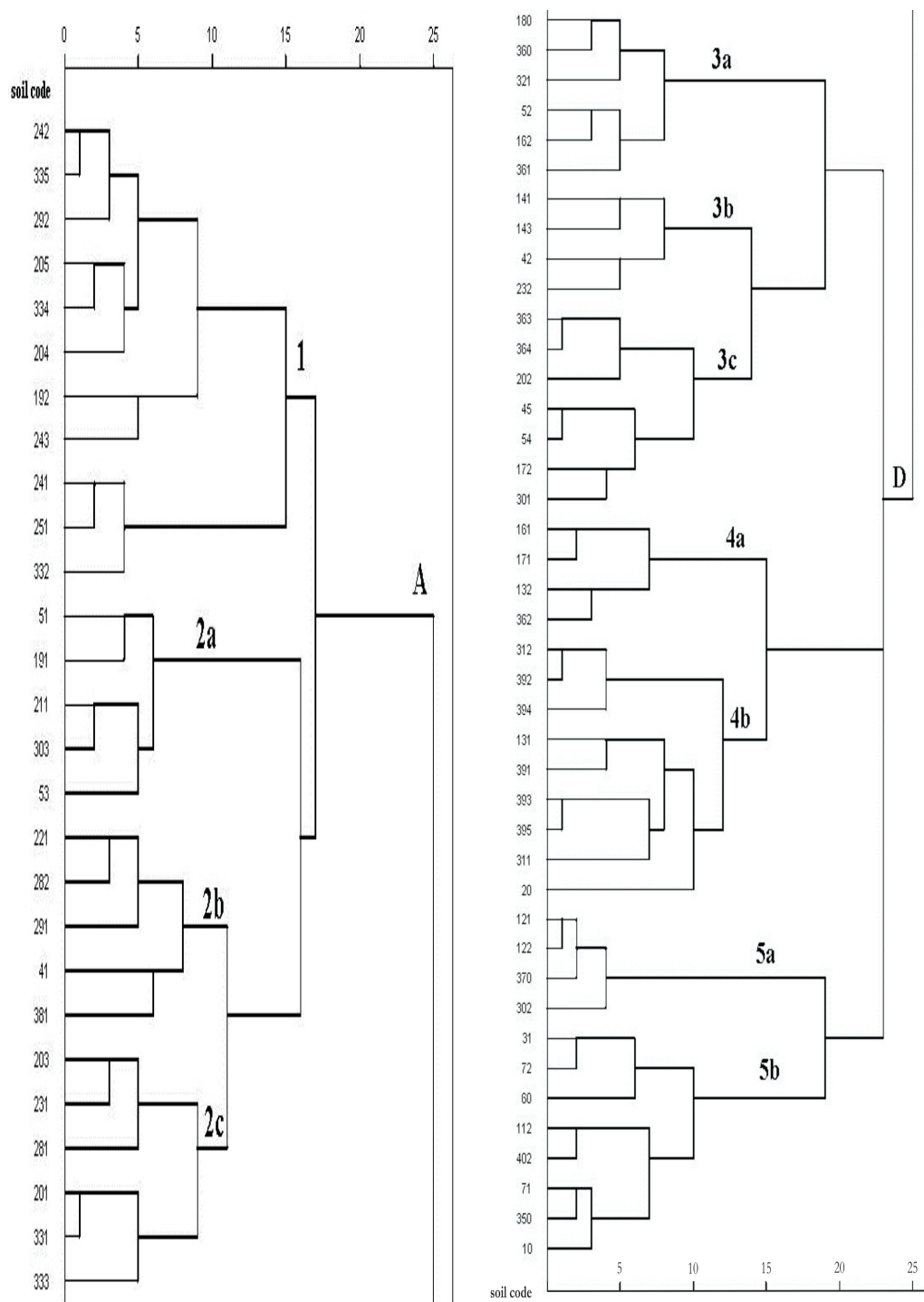
Fig. 3. Dendogram derived by hierarchical clustering from the dissimilarity matrix after dimension reduction
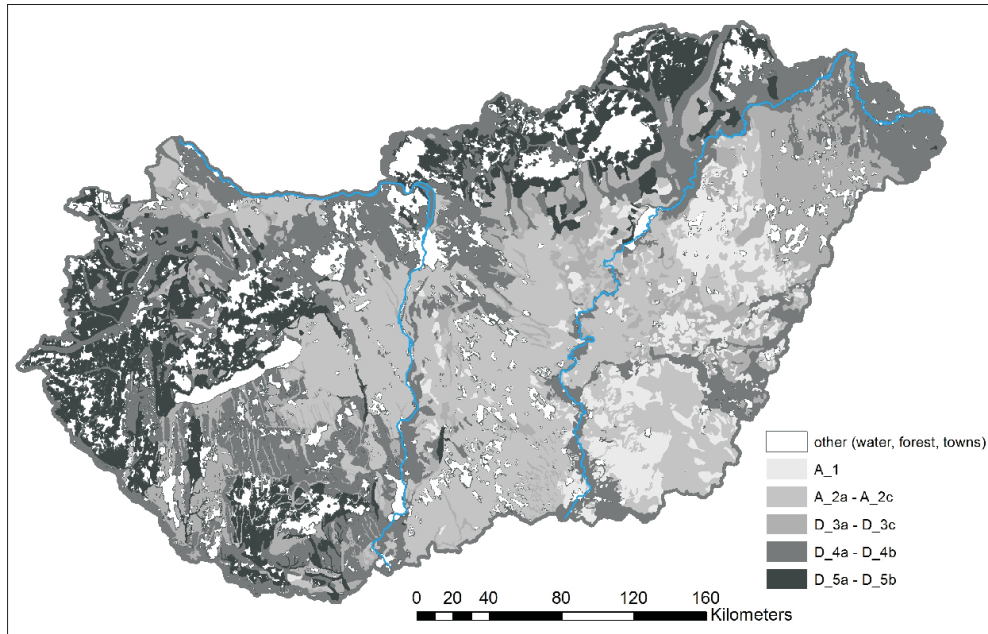
Fig. 4. Map of the soil type clusters

Table 2. Clusters in the dendogram and the major soil type in the cluster

Cluster No. in *Figure 3*	Legend in *Figure 4*	Approximate WRB equivalent of the major soil type in the cluster and its code No.	Area % within the cluster
1	A_1	205: *Bathy*gleyic Chernozem, Pachic *with sodicity in the subsurface soil horizon*	20.3
2a	A_2a–A_2c	51: Haplic Arenosol, Calcaric	45.3
2b		291: Mollic Gleysol, (Hypo-)sodic	68.7
2c		201: *Bathy*gleyic Calcic Chernozem, Pachic	72.1
3a	D_3a–D_3c	162: Haplic Cambisol, Humic	37.7
3b		141: Brunic Lamellic Arenosol	76.5
3c		301: Calcic Mollic Gleysol	71.5
4a	D_4a–D_4b	132: Brunic Arenosol, Eutric, Chromic	55.3
4b		131: Haplic Cambisol, Eutric, Siltic	25.5
5a	D_5a–D_5b	302: Mollic Gleysol, *not calcareous*	56.8
5b		112: Haplic Luvisol	84.7

common near the Western border of Hungary. The cluster presents itself in other parts of the country but Stagnic Luvisols do not.

Soils in a landscape are associated spatially as well as taxonomically (Hole, F.D. 1978). However, spatially associated soils might not be associated taxonomically (Campbell, J.B. and Edmons, W.J. 1984). Thus, a spatial approach seems appropriate to derive soilscapes as a basis for subsequent digital soil-mapping purposes (Schmidt, K. *et al.* 2010).

According to the summarizing works by McBratney, A.B. *et al.* (2003) and Scull, P. *et al.* (2003) tree-based methods are rapidly gaining popularity as means to develop prediction rules that can be rapidly and repeatedly

evaluated. Because of the clear advantages, several authors applied tree-based methods for soil mapping problems (Hengl, T. *et al.* 2007; Grinand, C. *et al.* 2008; Cambule, A.H. *et al.* 2013; Sun, X.L. *et al.* 2011; Häring, T. *et al.* 2012; Pásztor, L. *et al.* 2013). Complex similarity (relatedness) or dissimilarity (distance) matrices and their analysis in tree form are routine procedures in several disciplines such as in psychology (Pecora, L.M. *et al.* 1995) genetics (Yu, J. *et al.* 2005) or in scientometrics (Boyack, K.W. *et al.* 2005). One of the early publications is on representing demographic data (Hartigan, J.A. 1967).

However, there is no evidence in the scientific literature that boundary line segments between soil polygons would have ever been analyzed and spatial association rules would have been extracted as trees from legacy soil maps. Compared to other regionalization studies (Schmidt, K. *et al.* 2010; Lilburne, L.R. *et al.* 2012), we used only boundary segments and soil classes on both sides of the line instead of complex data sets on soil, terrain, geology and other surface properties and analyzed the whole data set instead of subsetting by moving window method with rasterized data (Behrens, T. *et al.* 2009; Schmidt, K. *et al.* 2010). The consequence of our approach is that the region boundaries are rather fuzzy with large mosaicked transition zones around the more homogenous core zones *(Figure 4).* Variable sized moving window method (Behrens, T. *et al.* 2009; Schmidt, K. *et al.* 2010) combined with our boundary line approach may result more homogenous soilscapes. This combination of methods may alleviate the problem of Stagnic Luvisols mentioned above where existing associations in one region were false in another region in spite of the presence of the same cluster simply because one soil class was missing.

Conclusions

There are three nationwide legacy soil maps in Hungary. The first one was published in 1953 at a scale of 1:200,000 (Mattyasovszky, J.

et al. 1953), the second one (popularly called AGROTOPO) was published between 1983 and 1988 on 1:100,000 sheets (Várallyay, Gy. *et al.* 1979, 1980; MÉM 1983–1988) and the third one (genetic soil map) was compiled by the experts of the agricultural extension agency of the agricultural ministry in 1983 at scale of 1:200,000 (MÉM-NAK 1983). The genetic soil map provides the most complete display of the HSCS thus it is the most appropriate basis for soilscape analysis. Despite its relative completeness, it does not contain all the soil types and sub-types of the HSCS. Further digital soil mapping works are needed since spatial resolution of existing maps are insufficient to the requirements of the policy making (Pásztor, L. *et al.* 2013; Sisák, I. and Benő, A. 2012, 2014).

In conclusion, our method has resulted a promising approach for delineating soilscapes in presence of overview soil maps. We used the method for whole area of Hungary but it has resulted fuzzy soilscapes with broad transition zones. The method could be refined by using variable-sized moving window method and by combining boundary data with terrain, geology etc.

REFERENCES

AIR 2013. *Agrár-környezetgazdálkodási Indormációs rendszer* (Information System for the Agri-environmental schemes). Nyilvános térképek (Open access maps). (in Hungarian) http://terkep.air.gov.hu/terkep/nyilvanos/nyilvanos.htm

Behrens, T., Schneider, O., Lösel, G., Scholten, T., Hennings, V., Felix-Henningsen, P. and Hartwich, R. 2009. Analysis on pedodiversity and spatial subset representativity – the German soil map 1:1 000 000. *Journal of Plant Nutrion and Soil Science* 172. 91–100.

Boyack, K.W., Klavans, R. and Börner, K. 2005. Mapping the backbone of science. *Scientometrics* 64. (3): 351–374.

Bui, E. 2004. Soil survey as a knowledge system. *Geoderma* 120. (1–2): 17–26.

Buol, S.W., Hole, F.D. and McCracken, R.J. 1973. *Soil genesis and classification*. Ames, University Press, IA, USA. 404 p.

Cambule, A.H., Rossiter, D.G. and Stoorvogel, J.J. 2013. A methodology for digital soil mapping in poorly-accessible areas. *Geoderma* 192. 341–353.

Campbell, J.B. and Edmonds, W.J. 1984. The missing geographic dimension to soil taxonomy. *Annals of the Association of American Geographers* 74. 83–97.

Chrisman, N.R. 1989. Modelling error in overlaid categorical maps. In *Accuracy of spatial databases.* Eds.: Goodchild, M. and Gopal S. London, Taylor & Francis, 21–34.

Creamer, R., Simo, I., Reidy, B., Carvalho, J., Fealy, R., Hallett, S., Jones, R., Holden, A., Holden, N., Hannam, J., Massey, P., Mayr, T., McDonald, E., O'Rourke, S., Sills, P., Truckell, I., Zawadzka, J. and Schulte R. 2010. *Irish Soil Information SystemSynthesis Report.* No. 130, Johnstown Castle, Environmental Protection Agency, Ireland.

Dobos, E., Daroussin, J. and Montanarella, L. 2005. *An SRTM-based procedure to delineate SOTER Terrain Units on 1:1 and 1:5 million scales.* EUR 21571 EN, Luxembourg, Office for Official Publications of the European Communities. 55 p.

Dobos, E., Daroussin, J. and Montanarella, L. 2010. A quantitative procedure for building physiographic units supporting a global SOTER database. *Hungarian Geographical Bulletin* 59. (2): 181–205.

Dobos, E., Montanarella, L., Nègre, T. and Michéli, E. 2001. A regional scale soil mapping approach using integrated AVHRR and DEM data. *International Journal of Applied Earth Observation and Geoinformation* 3. (1): 30–42.

Finke, P., Hartwich, R., Dudal, R., Ibàñez, J., Jamagne, M., King, D., Montanarella, L. and Yassoglou, N. 2001. *Georeferenced Soil Database for Europe. Manual of procedures. Version 1.1.* European Soil Bureau, Research Report No. 5. EUR 18092 EN. European Communities

Grinand, C., Arrouays, D., Laroche, B. and Martin, M.P. 2008. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma* 143. (1–2): 180–190.

Häring, T., Dietz, E., Osenstetter, S., Koschitzki, Th. and Schröder, B. 2012. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. *Geoderma* 185–186. 37–47.

Hartigan, J.A. 1967. Representation of similarity matrices by trees. *Journal of the American Statistical Association* 62. (320.) 1140–1158.

Hengl, T., Toomanian, N., Reuter, H.I. and Malakouti, M.J. 2007. Methods to interpolate soil categorical variables from profile observations: lessons from Iran. *Geoderma* 140. 417–427.

Hewitt, A.E., 1993. Predictive modelling in soil survey. *Soil and Fertilizers* 56. (3): 305–314.

Hewitt, A.E., Barringer, J.R.F., Forrester, G.J. and McNeill, S.J. 2010. Soilscapes basis for digital soil mapping in New Zealand. In *Digital Soil Mapping Bridging Research, Environmental Application, and Operation.* Eds.: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E. and Kienast-Brown, S. Dortdrecht, Springer, The Netherlands, 297–308.

Hole, F.D. 1978. An approach to landscape analysis with emphasis on soils. *Geoderma* 21. 1–23.

Horváth, B., Izsó I., Jassó, F., Király, L., Parászka, L. and Várallyay, Gy. 1987. *Guide to conduct nationwide fine-scale soil mapping. Melioráció-öntözés és talajvédelem.* Budapest, Agroinform.

Horváth, B., Izsó, I., Jassó, F., Király, L., Parászka, L. and Szabóné Kele, G. 1989. *Guide to conduct nationwide fine-scale soil mapping. Melioráció-öntözés és tápanyaggazdálkodás.* Budapest, Agroinform.

http://napok.georgikon.hu/cikkadatbazis-2012/cat_view/3-cikkadatbazis/4-2012/10-vii-szekcio-vizgazdalkodas

Hudson, B.D. 1990. Concepts of soil mapping and interpretation. *Soil Survey Horizons* 31. (3): 63–72.

ISRIC 1993. *Global and National Soils and Terrain Digital Databases (SOTER)*. Procedures Manual. UNEP-ISSS-ISRIC-FAO. ISRIC. Wageningen, Netherlands. 115 p.

IUSS Working Group WRB 2007. World Reference Base for Soil Resources 2006, first update 2007. *World Soil Resources Reports* No. 103. Rome, FAO.

IUSS Working Group WRB. 2014. World Reference Base for Soil Resources 2014. International soil classification system for naming soils and creating legends for soil maps. *World Soil Resources Reports* No. 106. Rome, FAO.

Jenny, H. 1941. *Factors of Soil Formation, A System of Quantitative Pedology.* New York, McGraw-Hill, 191 p.

Krasilnikov, P., Arnold, R. and Michéli, E. 2009. Soil Classification of Hungary. In *A handbook of soil terminology, correlation and classification.* Eds.:

Arnold, R., Shoba, S., Krasilnikov, P. and Marti, J.J.I. London, Earthscan, Sterling, VA.

Lagacherie, P., Robbez-Masson, J.M., Nguyen-The, N. and Barthes, J.P. 2001. Mapping of reference area representativity using a mathematical soilscape distance. *Geoderma* 101. 105–118.

Lilburne, L.R., Hewitt, A.E. and Webb, T.W. 2012. Soil and informatics science combine to develop S-map: A new generation soil information system for New Zealand. *Geoderma* 170. 232–238.

Mattyasovszky, J., Görög, L. and Stefanovits, P. 1953. *1:200 000 m.a. mezőgazdasági talajtérkép* (Agricultural soil map at a scale of 1:200 000). Budapest, Tervgazdasági Könyvkiadó.

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

McBratney, A.B., Hart, G.A. and McGarry, D. 1991. The use of region partitioning to improve the representation of geostatistically mapped soil attributes. *Journal of Soil Science* 42. 513–532.

McSweeney, K., Slater, B. K., Hammer, R.D., Bell, J.C., Gessler, P.E., Petersen, G.W. 1994. Towards a new framework for modelling the soil-landscape continuum. In *Factors of Soil Formation: A 50th Anniversary Retrospective.* Eds.: Amundson, R., Harden, J. and Singer, M. SSSA Special Publication 33. 127–145.

MÉM (1983–1988). *Magyarország agrotopográfiai térképe 1:100 000 m.a.* (Agrotopographic map of Hungary 1:100 000). Budapest, MÉM Országos Földügyi és Térképészeti Hivatal, 84 térképlap /mapsheet, 51 x 66 cm.

MÉM-NAK 1983. *Magyarország genetikai talajtérképe 1:200 000 m.a.* (Genetic soil map of Hungary 1:200,000. Budapest, Mezőgazdasági és Élelmezésügyi Minisztérium.

Mendonça-Santos, M.D.L., Santos, H.G., Dart, R.O. and Pares, J.G. 2008. Digital mapping of soil classes in Rio de Janeiro State, Brazil: data, modelling and prediction. In *Digital Soil Mapping with Limited Data.* Eds.: Hartemink, A.E., Mcbratney, A. and Mendonça-Santos, M.L. Dordrecht, Springer, 381–396.

Michéli, E., Fuchs, M., Hegymegi, P. and Stefanovits, P. 2006. Classification of the major soils of Hungary and their correlation with the World Reference Base for Soil Resources (WRB). *Agrokémia és Talajtan* 55. (1): 19–28.

Pásztor, L., Bakacsi, Zs., Laborczi, A. and Szabó, J. 2013. Downscaling of categorical soil maps with the aid of auxiliary spatial soil information and data mining methods. *Agrokémia és Talajtan* 62. 205–218.

Pecora, L.M., Carroll, T.L. and Heagy, J.F. 1995. Statistics for mathematical properties of maps between time series embeddings. *Physical Review E* 52. (4): 3420. p.

Pontius Jr, R.G. 2002. Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. *Photogrammetric Engineering and Remote Sensing* 68. (10): 1041–1050.

Schmidt, K., Behrens, T., Friedrich, K. and Scholten, T. 2010. A method to generate soilscapes from soil maps. *Journal of Plant Nutrition and Soil Science* 173. (2): 163–172.

Scull, P., Franklin, J. and Chadwick, O.A. 2005. The application of classification tree analysis to soil type prediction in a desert landscape. *Ecological Modelling* 181. 1–15.

Scull, P., Franklin, J., Chadwick, O.A. and McArthur, D. 2003. Predictive soil mapping: a review. *Progress in Physical Geography* 27. (2): 171–197.

Sisák, I. and Benő, A. 2012. *Digital publication of the 1:200 000 scale agricultural soil map on the Georgikon Mapserver.* LIV. Georgikon Napok, Keszthely, 2012. okt. 11–12. Papers, 431–436.

Sisák, I. and Benő, A. 2014. Probability-based harmonization of digital maps to produce conceptual soil maps. *Agrokémia és Talajtan* 63. (1): 89–98.

Sun, X.L., Zhao, Y.G., Zhang, G.L., Wu, S.C., Man, Y.B. and Wong, M.H. 2011. Application of a digital soil mapping method in producing soil orders on mountain areas of Hong Kong based on legacy soil data. *Pedosphere* 21. (3): 339–350.

Várallyay, Gy., Szűcs, L., Murányi, A., Rajkai, K. and Zilahy, P. 1979. Magyarország termőhelyi adottságait meghatározó talajtani tényezők 1:100 000 méretarányú térképe I. (1:100,000 map of the properties determining soil productivity in Hungary I.) *Agrokémia és Talajtan* 28. 363–384.

Várallyay, Gy., Szűcs, L., Murányi, A., Rajkai, K. and Zilahy, P. 1980. Magyarország termőhelyi adottságait meghatározó talajtani tényezők 1:100 000 méretarányú térképe II. (1:100,000 map of the properties determining soil productivity in Hungary II.) *Agrokémia és Talajtan* 29. 35–76.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. and Buckler, E.S. 2005. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38. (2): 203–208.