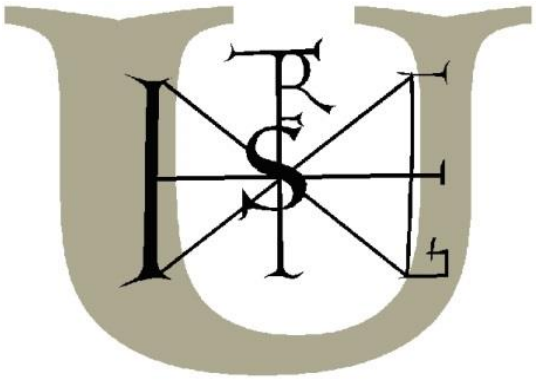
Szent István University
Doctoral School of Environmental Sciences



Development of harmonization, correlation methods and data storage system to support soil conservation and international correlation

Thesis of PhD dissertation

Vince Láng

Gödöllő
2013

Name of doctoral school: Environmental Sciences

discipline: Soil science, agrochemistry, environmental chemistry

Head of School:

Csákiné Dr. Michéli Erika, D.Sc.
professor
Szent István University
Faculty of Agricultural and Environmental Sciences
Institute of Environmental Sciences
Department of Soil Science and Agrochemistry

Scientific Supervisor:

Csákiné Dr. Michéli Erika, D.Sc.
professor
Szent István University
Faculty of Agricultural and Environmental Sciences
Institute of Environmental Sciences
Department of Soil Science and Agrochemistry

……...
Approval of Head of School

……....
Approval of Scientific Supervisor

Introduction and objectives

The rapid growth in world population, related increasing need of food and energy resulted in data hunger to feed environmental and socio-economical models to predict the future changes, issues and needs, but their uncertainty strongly relies on the reliability of the input data. Spatial soil information supports these models, while serve as input for many others. The quality and spatial distribution of the available data sources are limited in many cases. New surveys are not fore seen, hence the harmonization of soil data collected, analyzed and processed with different methods at different times is one of the greatest challenge for data users and applications. The lack of harmonization and correlation of these datasets is another serious issue. Many international and national programs were brought on to serve data harmonization on an international level: The European Union's INSPIRE directive (European Commission, 2007), the EU FP7th Framework's eSOTER project (eSOTER, 2008a,b), the GSSOil eContentplus program (Feiden, 2012) or the Universal Soil Classification System Working Group of the International Union of Soil Science.

According to the international trends it is necessary to collect, and store the valuable information. Not only to serve the European Commission's data needs, but to serve the present technological needs of modeling, other scientific disciplines and the soil science community. Hungary is rich in soil data, but the lack of harmonization, standardization, synthesis and free, easy access of these datasets, makes interpretation, utilization difficult. The development of a data structure and harmonization procedure to store the wide variety of soil data is still required. Such a system would help to save ant interpret the different datasets stored at local scientific workshops, in soil protection plans, and datasets developed by national or international funded projects.

Objectives:

Based on the previously defined objectives and the acquired experiences the following objectives were drawn up:

1. Develop a system to filter, qualify, harmonize and convert archive soil data in accordance with international needs.
2. Develop a data structure model, which is able to store data from different sources and can serve international harmonization needs.

3. Improve the lower classification level of the proposed new Hungarian soil classification system - developed at the Szent Istvan University's, Department of Soil Science and Agrochemistry – with the use of mathematical and statistical methods based on legacy data.
4. Test the developed dataset and the stored data for soil mapping purposes

Materials

The research was based on different national and international classification systems, soil description guidelines and datasets, which are briefly discussed below.

The Hungarian National Soil Classification System

The Hungarian National Soil Classification System (Szabolcs, 1966; Stefanovits, 1981) was the base of the soil correlation tasks. The soil types of the system were correlated to the World Reference Base Reference Soil Groups.

The Proposed Hungarian National Soil Classification System

The proposed classification system played a major role at the data converting task. A database was also developed with an automated classification system to serve the data need for the improvement of the lower level of the classification system (Michéli és mtsai, 2013a).

World Reference Base for Soil Resources (WRB)

The WRB was the basis for the soil classification data harmonization on an international level. Hungarian soil types were correlated to the system's units (IUSS Working group, 2006).

FAO Guidelines for Soil Description

Legacy data harmonization was based on the Proposed Guideline for Field Soil Description (Szabóné Kele, 2013a), in case relevant information was not available the FAO Guidelines for Soil Description was used as a harmonization platform (FAO, 2006).

Hungarian Soil Information and Monitoring System (SIMS)

The most comprehensive (due to standard field and laboratory methods) and detailed available national dataset was the bases of the harmonization process and the analysis of the lower classification levels of the Proposed Hungarian National Soil Classification System (TIM, 1995).

World Inventory of Soil Emission Potentials dataset (v3.1) (WISE)

The most comprehensive international soil profile dataset contains more than 11,000 profiles, all of them classified according the WRB. International data to study the correlation possibilities between different soil classification systems was derived from this dataset (Batjes, 2008).

USDA NRCS National Soil Information System (NASIS)

The database structure of the NASIS database was studied along with other national and international databases, to develop a new data structure, which can efficiently store data from different sources and serves interpretations.

Python programming language

The data filtering, qualifying and converting and classification algorithms were written in Python programming language, due to its efficient memory handling and capability to handle large datasets.

Microsoft Office Access

The data structure was built in MS Access format to support easy handling for novice users and the ability to use with open source software.

Methods

Methods of data quality check

The SIMS served as a basis for the task. Different quality check algorithms were defined and written in Python environment. Algorithms included limits checks, internal consistency checks, and invalid relationship checks. The development of the algorithms was based on functions used in other datasets and based on the idiosyncrasy of SIMS.

Harmonization of morphological and other descriptive data

Beside the data quality checks the harmonization of morphological and other descriptive data is important, to store the different sourced data in one platform and to serve further studies. Accepted international soil correlation platforms like the FAO Guidelines for Soil Description and the WRB can serve as a platform for international correlation. The Proposed Guideline for Field Soil Description and the Proposed Hungarian National Soil Classification System strongly correlates with these systems. During the harmonization process priority was given to the proposed Hungarian system and in case of lacking definitions international platforms were used as a guideline.

Harmonization of laboratory measured data

Harmonization of laboratory measured data is primarily part of the database structure development method the details are discussed later in that section.

Conversion and correlation of soil classification related data

One of the most problematic part of data harmonization is the correlation between soil classification units. Is most cases one to one correlation is not possible between units of two system. The reclassification of soil pedons is a very time consuming task. Automated classification algorithms were developed by Michéli et al. (2011) and Waltner et al. (2012) to derive WRB related classification information from several Hungarian datasets. A similar classification algorithm was developed in this study, for the Proposed Hungarian National Soil Classification System.

For the WRB such a system is almost inconceivable. Eberhardt and Waltner (2010) attempted to define an algorithm system to derive WRB Reference Soil Groups (RSG) from German

legacy datasets. The system has huge data need, which is not available in any Hungarian dataset.

Taxonomic distance calculation can be an alternative tool to correlate the units of the different systems. This method was studied for such purposes. The method is not unknown in soil science. The first application for soil classification was made by Hole and Hironaka (1960), later Bidwell and Hole (1964a) calculated numerical indices of similarity for 29 Kansas soils. In the early stages of numerical classification many studies were completed for soil classification (Bidwell and Hole 1964b, Sarkar et al. 1966, McBratney et al. 2000). These studies were mainly based on local data with limited scope. The idea has been revisited in the 20th century. Minasny and McBratney (2007) introduced taxonomic distance as criteria for supervised classification of soil groups. For the WRB Reference Soil Groups (RSGs) Minasny et al. (2009) derived taxonomic distances based on the presence (coded with 1) and absence (coded with 0) of key properties

Correlation of Brown Forest Soil types of the Hungarian National Soil Classification System to WRB RSGs with the use of taxonomic distance calculations

Correlation possibilities were studied on two bases:

1. A concept-based approach, where the method of Minasny et al. (2009) was further developed to derive taxonomic distances between the WRB RSGs and the HSCS BFS types based on dominant identifiers according to the concepts and definitions of the soil units.

2. A centroid-based approach, where legacy laboratory data were used to calculate centroids for the WRB RSGs and the HSCS BFS types.

The previously discussed methods were studied for correlation purposes and discussed on the example of the Brown Forest Soils and relevant Reference Soil Groups of the WRB. Results were compared to a former study (Michéli et al., 2006).

The concept-based approach

Taxonomic distance calculations were based on a property matrix, which contains the BFS soil classes (7) and possibly related RSGs (12), coded against selected dominant identifiers.

The concept of using „dominant identifiers" in order to characterize certain soil groups was introduced in the 2006 edition of WRB (IUSS Working Group WRB, 2006), to support better understanding of the logic for the sequence and grouping of RSGs in the WRB classification key. Dominant identifiers are soil-forming factors or processes that most clearly condition the soil formation, and in cases when the prevailing pedogenetic process or processes are not sufficient to characterize, and to distinguish certain RSGs, the results of soil formation, morphological, physical and chemical soil characteristics are used, single or in combination. The dominant identifiers in our study were determined as sets of soil properties developed due to the dominant soil forming factors and processes, and define the most important characteristics of the certain RSG or soil type. The dominant soil forming factors and processes are defined in the "Rationalized Key to the WRB RSGs" for the selected RSGs (IUSS Working Group WRB, 2006), and in the descriptions of HSCS for the BFS classification units (Stefanovits, 1999; Michéli, 2006).

The identifier properties were matched with the 19 soil groups, and were coded based on the probability of the presence of the attributes. Minasny et al. (2009) introduced 2 codes: 0 when

the condition is not present and 1, when the condition is likely to be present for the RSG. In this study an additional code was applied for better characterization of the units (Table 1). In the case of BFS expert judgment was often required during the coding, because of the lack of definitions and quantitative criteria (Table 1).

Table 1. Description of codes applied to characterize soil groups based on selected identifiers

Code	Definition
0	Condition cannot be present for the unit
0.5	Condition may be present for the unit
1	Condition is criteria for the unit

The centroid-based approach

The dominant identifiers were attempted to be converted to calculated centroid values. 9 different centroids have been defined, taking the available data (WISE and SIMS) and the representation of every selected group into consideration. The calculations of the centroid-based approach were affecting profiles from the first mineral horizon, counting as top, respectively to the studied depth of each attribute. Presence of pedogenesis in the horizon was not taken into account, resulting in the inclusion of parent material (C horizons) into the calculations. Mean (centroid) values were calculated weighted on the thickness of the horizon for each examined WRB RSGs and Hungarian BFS soil classes. When the centroid is referred to a depth of occurrence of a certain property, and the defined criteria was not fulfilled, the maximum value of 200 cm was given. The depth of 200 cm was chosen based on the maximum depth criteria occurring in the WRB key.

The calculation of distances

On the basis of the matrices, the taxonomic distances between the selected WRB and Hungarian BFS groups were calculated via R software (R Development Core Team, 2009) using Mahalanobis distance metrics to take the covariance into account:

$$d_{ij}=((x_i-x_j)^t S^{-1}(x_i-x_j))^{1/2}$$

where: d_{ij} is the element of distance matrix D with size (c×c), c is the number of soil groups, S represents the covariance matrixThe value of d_{ij} represents the taxonomic distance between soil group i and group j, and x refers to a vector of indicators of the soil identifiers.

Converting soil pedon data into the units of the Proposed Hungarian National Soil Classification System

Based on the classification key of the system an automated series of algorithms was developed in Python environment. Priority is given to measured laboratory properties over the field described data. Thanks to the well defined and simple definitions, the key is easily programmable.

Development of a modern soil database structure

Available soil database structures were studied to identify weaknesses and strength in accommodating data from different sources. The NASIS database's structure and the European soil data structure was further studied and with the fusion of the two a new structure was developed to serve efficient data handling and international needs. The developed structure can also be handled by novice users.

Improvement of the lower classification levels of the Proposed Hungarian Soil Classification System with the use of mathematical and statistical methods

The SIMS database was reclassified with the previously discussed automated classification algorithm. The derived soil classes were essential to use mathematical and statistical methods to improve the lower classification levels. Methods were tested on the proposed "Soils with clay accumulation" type. The following methods were used to study the propose soil types:

Silhouette analysis

Silhouette analysis (Rousseeuw, 1987) refers to a method of cluster validation. With a graphical interpretation the method provides information, how well an object lies within its cluster. The definition is as follows:

$$SW_i = (b_i - a_i) / \max(a_i, b_i);$$

where a is the average distance of i to other individuals in the same cluster, b is the average distance of i to individuals other cluster, according to this: $-1 < SW_i < 1$.

SW_i can be interpreted as follows:

0,71 – 1,00 strong relationship (good clustering)

0,51 – 0,70 moderate relationship

0,26 – 0,50 weak relationship

≤0,25 no relationship (no real clustering)

Principal component analysis (PCA)

PCA (Pearson, 1901) is an orthogonal transformation procedure, which converts observations and possibly correlated variables into a set of values, called principal components. The first principal component has the largest variance in the dataset, each succeeding component has the highest variance and orthogonal to the preceding component. The number of principal components is equal or less than the original variables.

k-Mean clustering

Simple k-Mean clustering (MacQueen, 1967) was used as a clustering algorithm. The method partitions the observations into *k* user defined clusters, based on their Euclidean distance to the cluster means. Several computational methods were developed. In this study the Hartigan and Wong (1979) computation was used:

$$SS(k) = \sum_{i=1}^{n} \sum_{i=0}^{p} (x_{ij} - \bar{x}_{kj})^2$$

where k is the cluster, x_{ij} is the value of j variable at i observation and x_{kj} is the mean of variable j in cluster k.

Soil type information derivation of the Proposed Hungarian Soil Classification System from layer based thematic soil information

The focus of digital soil mapping has turned from soil type mapping to property mapping in the last decade. The major soil mapping projects (African Soil Information Service (AfSIS, 2013); GlobalSoilMap (2013)) are focusing on soil properties and no classification related mapping is among the aims. The information content of soil classification units is larger than soil property maps, combined. Mathematical methods were studied based on the GlobalSoilMap specifications to derive soil classification information from layer based, property datasets:

Taxonomic distance based calculation was performed on the SIMS dataset with centroid values. Centroids were calculated for the proposed soil types of the Proposed Hungarian Soil Classification System. A validation dataset of 250 profiles (22% of total) were partitioned to test the method.

A Random Forest based method was also studied, where 30% of the pedons were used for validation. Random Forest method is based on classification tree algorithms, where the user defined number of trees are grown and each individual is classified according to each tree. The assigned class is the one, which was the result for the largest number of trees. The method also assigns a reliability value for each result.

Results

Results of data quality check

The data quality check algorithm series analyses the laboratory measured parameters and their relation to morphological data. Each profiles, each horizon is analyzed based on the created algorithm. Limits checks were performed for each variable with limits defined by the unit (eg. percentage) or limits defined by theoretical minimum, maximum values (eg pH for soil between 2 and 12). The program also generates an extra file, which includes the deletions, modifications performed by the program. This is a necessary to track the modifications.

Harmonization of morphological and other descriptive data

The harmonization process was based on the previously discussed materials. A series of algorithms were developed. Beside the simple harmonization algorithms additional functions were developed to derive additional parameters (e.g. horizon indexes from morphological data). All records were coded for easier storage purposes. The description of the codes can be found in separate tables, specially designed for metadata storage.
In many cases the lack of definition in the Proposed Guideline for Field Soil Description was an issue, in these cases the harmonization was performed according to the FAO Guidelines for Soil Description. These resulted several issues. The FAO guideline describes several properties in a hierarchical system. These systems can differ from the description system in the archive datasets, and definitions can be correlated to different levels of the international hierarchical system. These problems occur for land use, landscape, soil structure etc classification.

Harmonization of laboratory measured data

The created classification program was defined based on the classification key of the Proposed Hungarian National Soil Classification system. The program was built in the same order. The *Anthropogenic soils* type was excluded from the program, due to lack of information stored in legacy soil datasets. The program was written in Python environment and is part of the previously discussed function series. The input data is fed from the output of those programs.
The program was tested on the SIMS database (Figure 1.) The database contains over 1200 soil profiles and the classification can be performed within 1 minute (min. 4 Gb memory, Intel Core i5 processor). The output of the program can be selected. Either the whole database can be saved, or the individual identification numbers with the soil type can be saved to a file. The program is also capable for modification, to support SQL based import into MS Access formats. The program is also capable to create a file with the reliability of the classification, based on the necessary information for the classification.

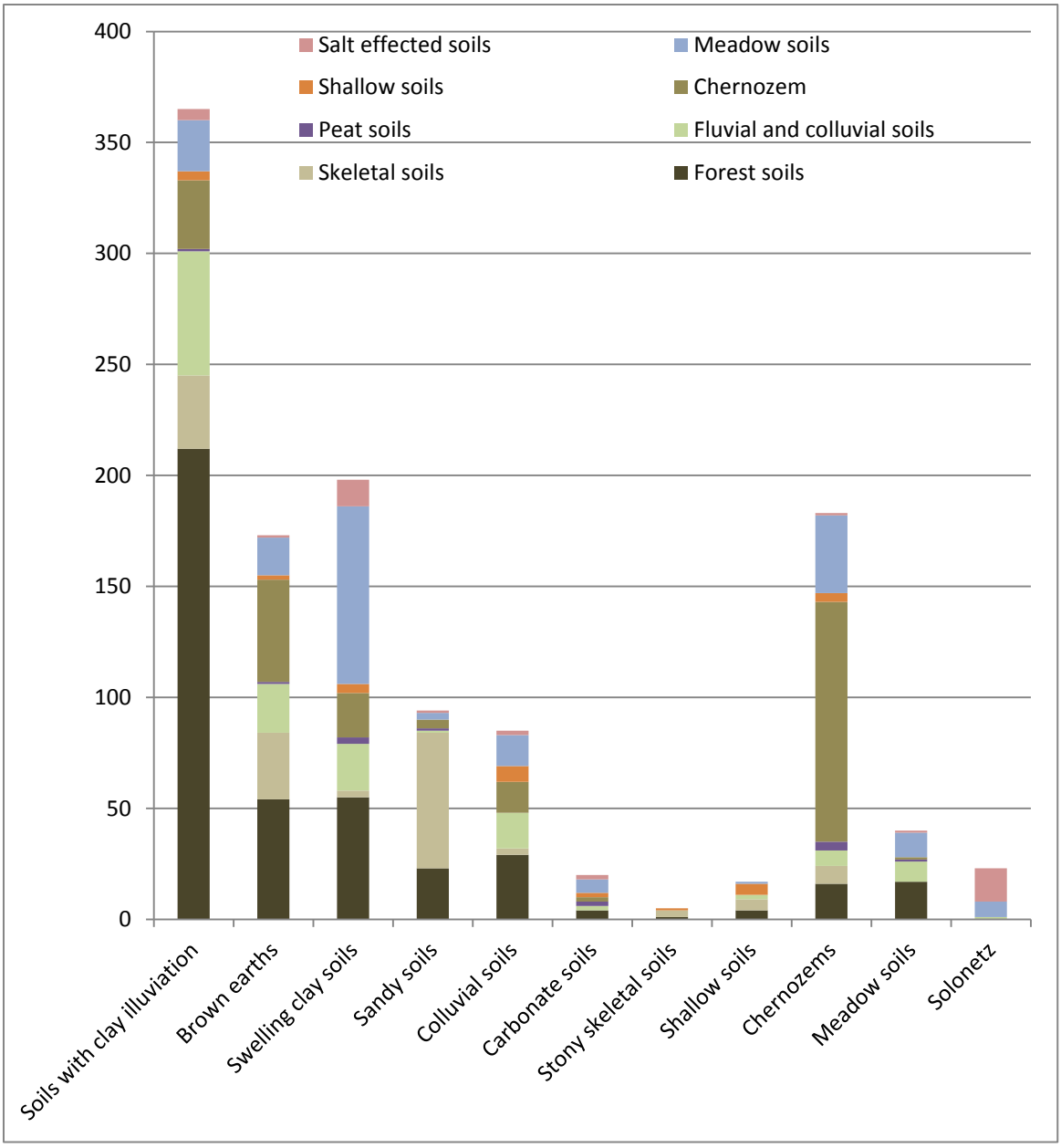


Figure 1. Distribution of soil types derived based on the automated classification algorithm and the actual soil types of the SIMS dataset

Conversion and correlation of soil classification related data

The taxonomic distances between the selected Hungarian soil types and the WRB RSGs were plotted on heat maps for better visual interpretation (Figure 2)

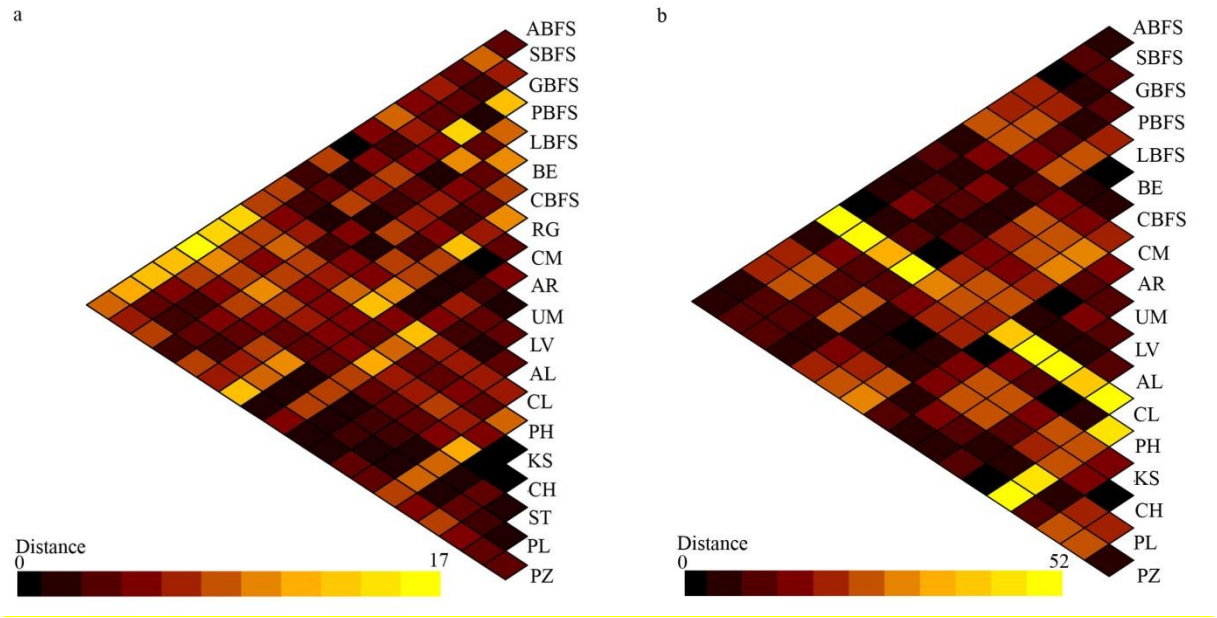


Figure 2 Calculated taxonomic distances plotted on shaded distance matrices based on the concept-based (a) and the centroid-based (b) approach

For easier understanding an extraction of the possible correlation of HBFS types to related WRB RSGs according to the results of the applied methods compared with expert based previous studies (Michéli et al. 2006) can be seen in Table 2.

Table 2. Possible correlation of HBFS types to related WRB RSGs according to the results of the applied methods compared with expert based previous studies (Michéli et al. 2006)

HBFS types	**The 3 closest WRB RSG according to the concept based approach [a]**	**The 3 closest WRB RSG according to the centroid based approach [a]**	**Expert based correlation (no ranking)**
Chernozem BFS (CBFS)	Chernozems, Phaeozems, Kastanozems	Kastanozems, Chernozems, Luvisols	Chernozems, Kastanozems, Phaeozems
Brown earth (BE)	Umbrisols, Cambisols, Chernozems	Kastanozems, Chernozems, Luvisols	Cambisols
Lesivated BFS (LBFS)	Luvisols, Planosols, Stagnosols	Kastanozems, Chernozems, Cambisols	Luvisols, Alisols
Pseudogley BFS (GBFS)	Luvisols, Stagnosols, Planosols	Cambisols, Luvisols, Planosols	Luvisols, Stagnosols
Lamellic BFS (SBFS)	Arenosols, Luvisols, Regosols	Arenosols, Alisols, Umbrisols	Luvisols, Arenosols

Podzolised BFS (PBFS)	Umbrisols, Regosols, Podzols	Alisols, Podzols, Arenosols	Luvisols, Umbrisols, Alisols
Acidic, non-podsolised, BFS (ABFS)	Cambisols, Umbrisols, Regosols	Alisols (0,43), Planosols, Umbrisols	Cambisols, Umbrisols, Alisols

Chernozem BFS (CBFS): both methods resulted the WRB steppe soils taxonomically the closest ones to the CBFS. Luvisols were calculated as the third closest RSG in the centroid-based approach. This result can be explained by the fact that the definition of CBSF lacks the specific criteria for the presence and depth of clay or secondary carbonate accumulations.

Brown earths (BE): has a very general morphological description and lacks numerical criteria. At the same time the TIM database includes more BE profiles than any other BFS type, and from this diverse profile set almost any RSG might be matched. Logically the concept-based mean taxonomic distance is the closest to the Cambisols. In the centroid-based calculation the lack of the numerical definitions of the BE is well reflected.

Lessivated BFS (LBFS): for the concept-based calculations Luvisols are the closest ones, as expected. On the other hand the steppe soils has occurred as close RSGs for the centroid-based method, repeatedly indicating that the presence and depth of accumulation of clay and secondary carbonates should be important differentiating criteria for the taxonomic composition of the classification units.

Pseudogley BFS (GBFS): the expert judgment and the concept-based taxonomic distances are in line; the closest RSG is the Luvisols, followed by the Stagnosols. The centroid-based concept resulted in different order of closest RSGs. The absence of Stagnosols is due to the lack of input data, hence the Stagnosols have been excluded from the calculations.

Lamellic BFS (SBFS): both the concept-, and the centroid-based methods correlate well with the expert judgment. Umbrisols resulted as third closest for the centroid-based approach, this is a result of the low base status and low pH for both units, even though the organic carbon content, of the SBFS is much lower than that of the Umbrisols.

Podzolic BFS (PBFS): most PBFSs lack the spodic horizon, so the WRB Podzols were not expected to be the closest ones, as results show from the centroid-based calculations. In case of the concept-based approach the possibility of a spodic horizon brings them close to the Podzols. Umbrisols, Alisols and the Arenosols could logically get close to the PBFS based on their low base saturation and pH values.

Acidic, non-podzolised BFS (ABFS): The concept-based method resulted the Cambisols as the closest RSG that is in good agreement with the expert based correlation. Umbrisols are the second closest RSG, even though the calculated centroids show that the organic carbon content of the Umbrisols is much higher than that of the ABFS. The centroid-based method resulted Alisols and Umbrisols as close RSGs to ABFS, which is in a good agreement with the expert based correlation results. Planosols occurred as second closest for the centroid-based methods, which was not expected, although the centroid calculations showed that the two units are very similar to each other except for the clay content change centroid. Some of the ABFS profiles may fulfill the clay content change criteria of the Planosols based on the databases, but the abrupt increasing would not be fulfilled.

Taxonomic distance based correlation of selected proposed soil types with WRB RSGs

Centroid based taxonomic distance calculation was performed on some selected soil types of the Proposed Hungarian National Soil Classification System, derived from the SIMS dataset. The previously used centroids of the WRB RSGs were used and centroids were calculated for 5 proposed soil types. The centroids were defined to correlate the Hungarian Brown Forest Soil type and due to this they may not represent all the important properties of the selected types. The extracted result of the calculations can be seen, in Table 3, where the selected proposed soil types are matched with the first 3 closest WRB RSGs.

Table 3. Proposed Hungarian soil types and the 3 closest WRB RSGs according to a centroid based distance calculation

Proposed type	**Closest RSG**	**2nd closest RSG**	**3rd closest RSG**
Soils with clay accumulation	Luvisol	Cambisol	Chernozem
Brown earths	Chernozem	Kastanozem	Luvisol
Sandy soils	Arenosol	Luvisol	Cambisol
Carbonate soils	Calcisol	Kastanozem	Chernozem
Chernozems	Calcisol	Kastanozem	Chernozem

Soils with clay accumulation: The type shows strong relation to the Luvisol RSG, which is an expected result due to the similar definition in the classification systems. Cambisols occur as second closest, due to the unixpectidly high clay content change in the profile.
Brown earths: The concept of the Brown earth represent soils with weak profile development, hence the type should be close to Cambisols. In contrast with this the WRB steppe soils occur as closest RSGs. Based on the calculated centroid values the Cambisols, represented in the WISE 3.1 dataset are much more acid, than the Brown earths of Hungary. According to this the type can not be close to the Cambisol RSG. This could be improved, if the profiles of the RSG centroid calculations would be derived from similar environmental conditions.
Sandy soils: Based on the calculations the Arenosol RSG is the closest to the type, which can be expected due to the similar definitions.
Carbonate soils: The Calcisol RSG occured as the closest one to the type, which is expected due to the similar definitions of the to units.
Chernozem soils: The occurance of Calcisols as the closest RSG is unexpected. Based on the calculated centroid values the Hungarian steppes soils are less leached (the depth of calcic horizon is at 47 cms) than their international counterparts. The previously discussed environmental conditions play a huge role in the results. The fact, that the definition of the centroids were designed to study the Brown Forest Soils should also be noted. Probably the better definition of the centroids would improve the results.

Development of a modern soil database structure

The creation of the new database structure was based on the previous harmonization and quality check experience, the analysis of the NASIS and other datasets, and the recommendations of the Proposed Guideline for Field Soil Description and the FAO Guidelines for Soil Description.

Entry of soil morphological properties

The creation of the morphological data tables was based on the analysis of the NASIS dataset, where each morphological type has been studied. The number of described morphological properties by genetic horizons has been revealed as exemplified in Figure 3.

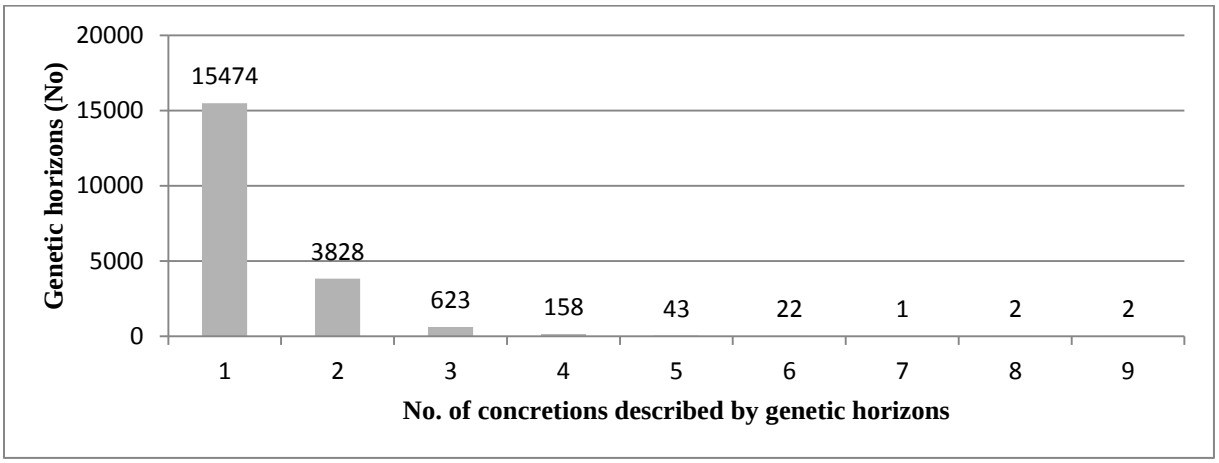


Figure 3. Number of concretions described by genetic horizons based on the analysis of NASIS database

Harmonization of laboratory measured properties based on standard methods.

The data structure also accommodates a table, where internationally accepted correlation factors are stored, to convert the results of non-standard methods into the accepted platform. The table is connected, to the horizon properties through the *Profile* table and the result of the queries can contain only harmonized values.

The developed database structure

The final data structure contains 38 data tables in a relational database. Both morphological and laboratory measured data can be stored in the data structure. For testing purposes the database has been filled with SIMS data.

Improvement of the lower classification levels of the Proposed Hungarian Soil Classification System with the use of mathematical and statistical methods

Analysis on the depth occurrence of redox features at the *Soils with clay illuviation* soil type

The depth occurrence of redox features can be important for classification purposes. A Silhouette analysis has been performed on the dataset. After a k-Mean clustering (Figure 4.) 3 clusters can be identified. Some of the individuals have equal membership in 2 clusters, but most of the individuals have valid membership in their cluster.

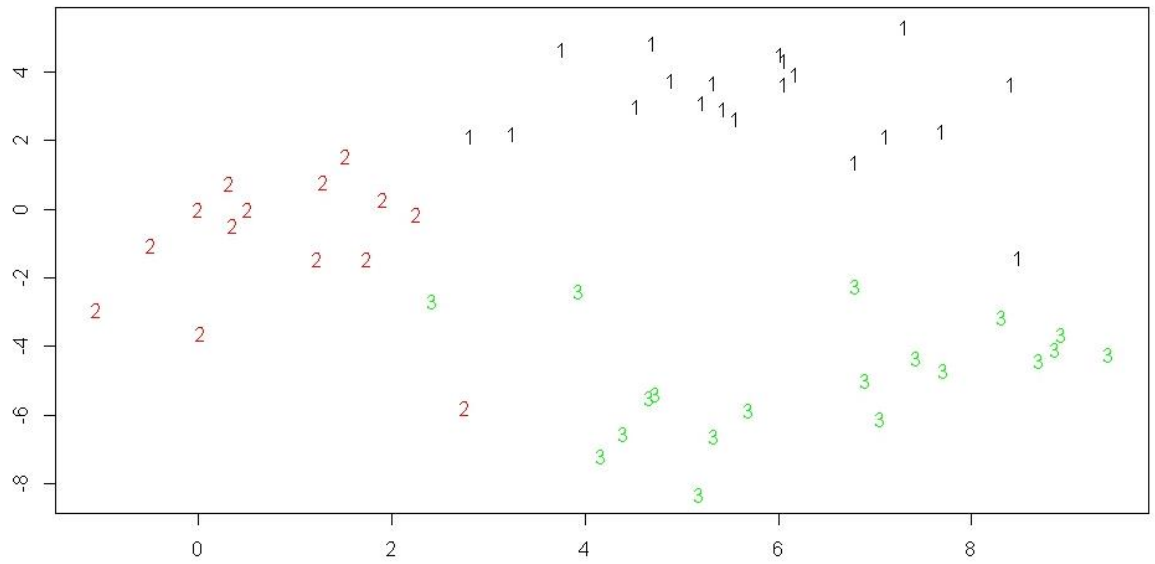


Figure 4. k-Mean clustering of the Soils with clay illuviation into 3 clusters (1-3)based on the occurance of redox features, plotted against 2 principal components

With the plotting of the probability (binary value ranged from 0 to1) of the evidence of redox features by depth for the 3 clusters of the *Soils with clay illuviation* type the 3 different depth occurrence can be identified (Figure 5.). The first cluster (Klaszter1) contains soil pedons, having redox features only in deep layers (below 1 meter the occurrence is almost sure). The second cluster (Kalszter2), has no evidence of redox features, or it is minimal. The third cluster (Klaszter3) has a maximum in evidence from 35-75 cms. This group can be identified as soils with stagnatic water close to the surface, due to intense clay illuviation.
Based on the results the *Soils with clay illuviation* can be classified into 3 lower level classes based on the occurrence of redox features.

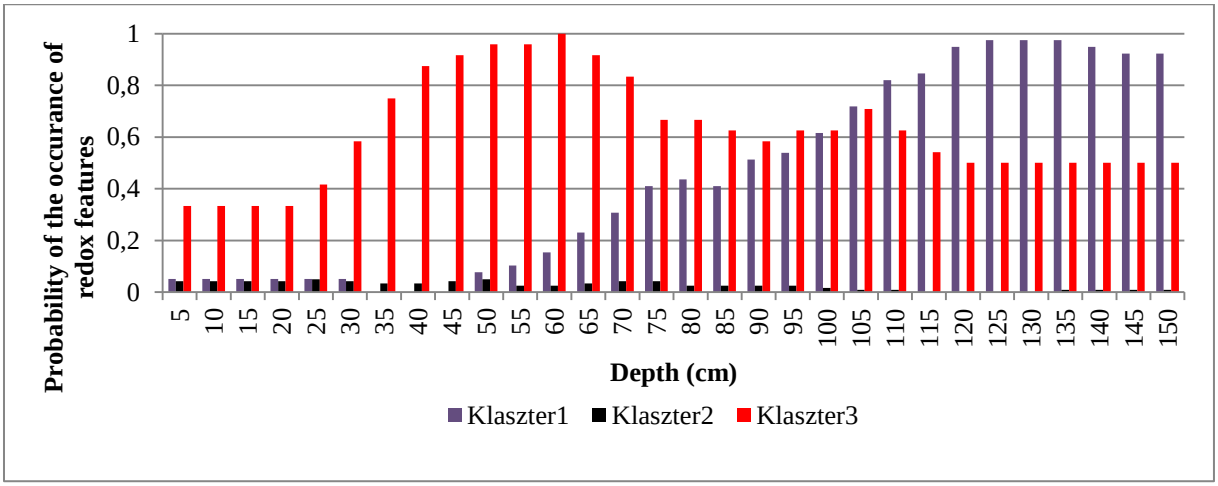


Figure 5. The probability of the evidence of redox features by depth for the 3 clusters of the Soils with clay illuviation type

Analysis on the coarse fragment content at the *Soils with clay illuviation* soil type

Based on the Silhouette analysis (Figure 6.) 2 clusters can be identifies. The second cluster occurs with a low silhouette width.

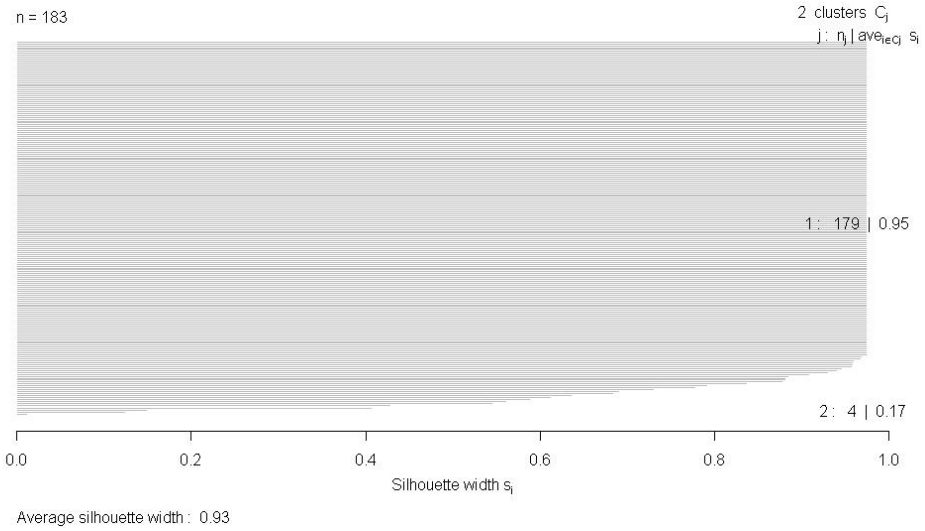


Figure 6. Silhouette analysis of the Soils with clay illuviation based on their coarse fragment content

After a k-Mean clustering (Figure 7.) 2 clusters can be identified. The first cluster has a high number of individuals, grouped together (no coarse fragment in the profile). The second cluster contains profiles with 30% or more coarse fragment somewhere in the profile. Based on the results 30% coarse fragment is a mathematical threshold to cluster Soils with clay illuviation.

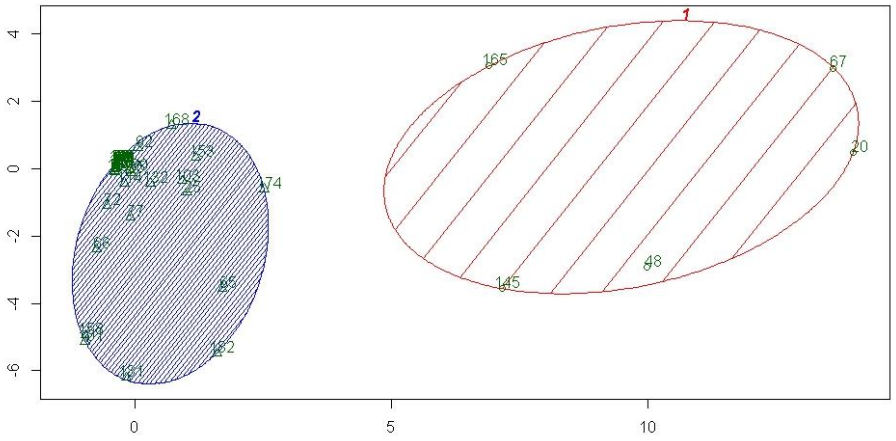


Figure 7. k-Mean clustering of the Soils with clay illuviation into 2 clusters (1-3) based on their coarse fragment content, plotted against 2 principal components

Based on the results the use of mathematical, statistical methods to identify significant properties for classification on the lower levels of a system is recommended. The results show correlation with the soil type level of the Hungarian National Soil Classification System,

which is a more detailed level than the proposed soil type level. Beside these properties further analysis can identify additional significant properties.

Soil type information derivation of the Proposed Hungarian Soil Classification System from layer based thematic soil information

Classification based on GSM specification with taxonomic distance calculations

Based on the results of the 250 validation profiles the mean accuracy of the method is below 30%. The Soils with sandy texture type was the only exception, where a greater accuracy has occurred. The method is not recommended for such purposes.

Classification based on GSM specification with random forest method

Mean accuracy is around 75% with the use of random forest (Table 4). For some soil types the accuracy is only around 50% (Solonetz), this may be the result of the lack of main identifying properties (Exchangeable Na). Some of the soil types (eg.: Shallow soils) had limited number of profiles in the database, this may also cause misclassification. On the whole, the random forest method could be used to derive soil classification related information from layer based soil property datasets. With the addition of environmental covariates, these results could be improved.

Table 4. Classification accuracy results based ont he GSM specifications, with the use of random forest

Soil Type	Probability
Soil with clay accumulation	79%
Brown earths	71%
Swelling clay soils	88%
Sandy soils	82%
Colluvial soils	62%
Carbonate soils	54%
Stony skeletal soils	58%
Chernozems	69%
Meadow soils	76%
Solonetz soils	60%
Mean	75%

Conclusions

The aim of the research was to develop a system for harmonization, correlation and storage of different national datasets, and also to provide tools for international correlation, and mapping purposes. A series of automated algorithms were developed in a programming environment to provide a tool for fast and reliable data correlation. The program is capable to deliver data in a format suitable to import into a database system, developed to provide a platform for data of different sources on a harmonized way.

Correlation of taxonomic information was performed with different methods. The soil types of the Hungarian soil classification system were correlated with World Reference Base for Soil Resources Reference Soil Groups with taxonomic distance calculation, a mathematical method that is capable to define similarities and dissimilarities of taxa. The soil types of the recently proposed, improved Hungarian soil classification system were also derived from datasets, through a developed series of classification algorithms, defined according to the classification key.

Correlated datasets, made mathematical studies possible, to define the lower classification levels of the system. Silhouette, cluster and principal component analysis were performed to derive important characteristics.

Layer based datasets, like the specification of the Global Soil Map project, was also successfully tested to derive soil classification information. These methods can provide a tool to (re)map the country according to the Proposed Hungarian National Soil Classification system.

The study revealed fast and reliable methods for soil data harmonization and correlation, and a feasible database structure for storing diverse data. Different techniques were proposed and used for the further improvement of the proposed Hungarian soil classification system, and also to derive the units of the systems from datasets developed for thematic and not for soil class mapping.

Summary of new scientific results

1. I have developed a data harmonization, a data quality check and a classification methods and algorithms, which are able to handle SIMS nomenclature based soil data to serve international harmonization, and are able to convert data to the Proposed Hungarian National Soil Classification System

2. I have successfully applied the taxonomic distance calculations - on a conceptual and centroid bases - as a tool for harmonization between differet classification systems.

3. I have succesfully applied and recommended mathematical methods to study and improve the lower classification levels of the Proposed Hungarian Soil Classification System.

4. With the use of mathematical methods I have successfully derived soil types of the Proposed Hungarian Soil Classification System form layer based thematic soil information.

Related publications

ARTICLES

1. Peer-reviewed research articles

1.1. With impact factor (according to WEB OF SCIENCE), in English

1.1.2. International publisher

LÁNG, V., FUCHS, M., WATNER, I. & MICHÉLI, E., 2013. Taxonomic distance metrics, a tool for soil correlation. *Geoderma* **192**, 269-276.
http://dx.doi.org/10.1016/j.geoderma.2012.07.023
Impact factor: 2,318

1.2. Without impact factor, in English

1.2.1. Hungarian publisher

LÁNG, V., FUCHS, M., WALTNER, I. & MICHÉLI, E., 2010. Taxonomic distance measurements applied for soil correlation. *Agrokémia és Talajtan*, **59**, 1. 57-64.

Független idézettség:
Zádorová, T. & Penížek, V., 2011. Problems in correlation of Czech national soil classification and World Reference Base 2006. *Geoderma* **167-168:** 54-60.

MICHÉLI, E., SZABARI, SZ., LÁNG, V., WALTNER, I., DOBOS, E. 2009: Applying diagnostic categories of the World Reference Base for Soil Resources (WRB) for identifying and delineating risk areas of salinization and sodification, Proceedings of the VIII. Alps-Adria Scientific Workshop, Neum, Bosnia-Herzegovina, 27 April – 2 May, 2009. *Cereal Research Communications*, Vol. **37** No. 3. 399-402. pp

1.3. Without impact factor, in Hungarian

FUCHS M., WALTNER I., SZEGI T., LÁNG V. & MICHÉLI E., 2011. A hazai talajtípusok taxonómiai távolsága a képződésüket meghatározó folyamattársulások alapján. *Agrokémia és Talajtan*, **60**, 1. 33-44.

WALTNER I., FUCHS M., MICHÉLI E., LÁNG V., 2012. Hazai archív talajadatok beillesztésének lehetőségei nemzetközi adatbázisokba. *Agrokémia és Talajtan*, **61**, 2. 263-76.

CONFERENCE PROCEEDINGS

4. Conference proceedings with ISBN, ISSN or other certification

4.1. Full text, peer-reviewed, in English:

LÁNG, V., FUCHS, M., WALTNER, I. & MICHÉLI, E., 2010. Pedometrics applications for correlation of Hungarian soil types with WRB. In: Gilkes R.J., Prakongkep N. (eds.), Proceedings of the 19th World Congress of Soil Science; Soil Solutions for a Changing World; ISBN 978-0-646-53783-2; Published on DVD; http://www.iuss.org; Symposium WG 1.1; The WRB ®evolution; 2010 Aug 1-6. Brisbane, Australia, IUSS; 2010, pp. 21-24.

4.2. Full text, peer reviewed, in Hungarian

FUCHS, M., SZŐCS, A., LÁSZLÓ, P., LÁNG, V., & MICHÉLI, E., 2008. A Bodrogköz vízhatás alatt álló talajainak osztályozási problémái. Talajvédelem Különszám. Talajtani vándorgyűlés, Nyíregyháza, 2008. május 28-29. Talajvédelmi Alapítvány, Bessenyei György Könyvkiadó, Nyíregyháza, 595-601. p. ISBN 978-963-9909-03-8

References

AFSIS (2013): African Soil Information Service projekt honlap , 2013, November, (http://www.africasoils.net/about/rationale)

BATJES N.H.(2008): ISRIC-WISE Harmonized Global Soil Profile Dataset (Ver. 3.1). Report 2008/2, ISRIC – World Soil Information, Wageningen, The Netherlands.

BIDWELL O. W., HOLE F. D. (1964a): Numerical taxonomy and soil classification. *Soil Science* 97**.** 58–62.

BIDWELL O. W., HOLE F. D., (1964b): An experiment in the numerical classification of some Kansas soils. *Soil Sci. Soc. Amer. Proc.* 26. 263–268.

BREIMAN L. (2001): Random Forests, *Machine Learning* 45 (1): 5–32. doi:10.1023/A: 1010933404324.

EBERHARDT E., WALTNER I. (2010): Finding a way through the maze – WRB classification with descriptive data. [5–8. p.] In: GILKES, J. R. & PRAKONGKEP, N (szerk.):Soil Solutions for a Changing World: Proc. 19th World Congress of Soil Science, 1–6 August 2010, Brisbane, Australia 5–8. International Union of Soil Sciences. Brisbane.

EKLUND A. (2011): Color-based plots for multivariate visualization (squash) package for R version 1.0.1

EURÓPAI TANÁCS (2007): Az Európai Parlament és a Tanács 2007/2/EK irányelve (2007. március 14.) az Európai Közösségen belüli térinformációs infrastruktúra (INSPIRE) kialakításáról. Az Európai Unió Hivatalos Lapja, L108/1

e-SOTER (2008a): e-SOTER projekt honlap. http://www.esoter.net/?q=category/home-page/welcome (2013.11.02)

e-SOTER (2008b): Description of Work (DOW), Propsal No. 211578, 19-Jul-2008. http://www.esoter.net/sites/default/files/files/DoW_e-SOTER_19jul.pdf

FAO (2006): Guidelines for soil description. 4th edition. FAO, Rome.

FEIDEN K. ÉS MUNKACSOPORT VEZETŐK (2012): ECP-2008-GEO-318004 GS Soil, „Assessment and strategic development if INSPIRE compliant geodata services for European soil data" Final report, 03. June 2012 (http://www.gssoil-portal.eu/Best_Practice/GS_SOIL_D1.5.3_3_annual_public_report.pdf)

GLOBALSOILMAP.NET (2011): Specifications GlobalSoilMap.net products Version 2.1, Report 1. July, 2011 (http://www.globalsoilmap.net/system/files/GlobalSoilMap_net_specifications_v2_0_edited_draft_Sept_2011_RAM_V12.pdf)

HARTIGAN J. A., WONG M. A. (1979): A K-means clustering algorithm. *Applied Statistics* 28, 100–108.

HOLE F. D., HIRONAKA M. (1960): An experiment in ordination of some soil profiles. *Soil Sci. Soc. Amer. Proc.* **24.** 309–312.

http1: http://ncsslabdatamart.sc.egov.usda.gov/
http2: http://en.wikipedia.org/wiki/Python_(programming_language)
IUSS WORKING GROUP WRB (2006): World Reference Base for Soil Resources 2006. World Soil Resources Reports, No. 103. FAO. Rome.
MACQUEEN J. B. (1967): Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press. pp.281–297.MR 0214227.Zbl 0214.46201. Retrieved 2009-04-07.
MCBRATNEY A. B., ODEH I. O. A., BISHOP T. F. A., DUNBAR M. S., SHATAR T. M. (2000): An overview of pedometric techniques for use in soil survey. *Geoderma*. **97.** 293–327.
MCFERRIN L. (2013): Statistical Analysis Tools for High Dimension Molecular Data (HDMD) package for R version 1.2
MICHÉLI E., LÁNG V., FUCHS M., WALTNER I., SZEGI T., DOBOS E., SERES A., VADNAI P., VAN ENGELEN V., DIJKSHOORN K., DAROUSSIN J., EBERHARDT E., SCHULER U., ZADOROVA T., KOZAK J., HANNAM J., HALLETT S., ZHANG G., YUGUO Z., BALAGHI R., MOUSSADEK R. (2011): Deliverable D5 – A soil data base for the 1:1 million scale windows. WP1 and WP2 report of the „e-SOTER – Regional pilot platform
MICHÉLI E., FUCHS M., HEGYMEGI P., 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. p.
MICHÉLI E., FUCHS M., LÁNG V., SZABÓNÉ KELE G. (2013a): Alapelvek, osztályozó kulcs. Vitaanyag a Magyar Talajtani társaság 2013. június 20-i ülésére. www.talaj.hu/magyar/szakosztalyok/Talajgenetika
MINASNY B., MCBRATNEY A. B. (2007): Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. *Geoderma* **142.** 285–293.
MINASNY B., MCBRATNEY A., HARTEMINK A.E. (2010): Global pedodiversity, taxonomic distance, and the World Reference Base. *Geoderma*, 155(3-4), 132-139. p.
PEARSON K. (1901): On Lines and Planes of Closest Fit to Systems of Points in Space (PDF). *Philosophical Magazine* 2 (11): 559–572.
R DEVELOPMENT CORE TEAM (2009): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
ROUSSEEUW P. J. (1987): Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* 20: 53–65. doi:10.1016/0377-0427 (87) 90125-7.
SARKAR P. K., BIDWELL O. W., MARCUS L. F. (1966): Selection of characteristics for numerical classification of soils. *Soil Sci. Soc. Amer. Proc.* **30.** 269–272.
STEFANOVITS P. (1981): Talajtan. Mezőgazdasági Kiadó. Budapest
SZABOLCS I. (szerk.) (1966): A genetikus üzemi talajtérképezés módszerkönyve. OMMI. Budapest.
SZABÓNÉ KELE, G. (2013a): Javaslat helyszíni talajfelvételezés általános módszertanára. Vitaanyag a Magyar Talajtani társaság 2013. június 20-i ülésére. www.talaj.hu/magyar/szakosztalyok/Talajgenetika
TIM (TALAJVÉDELMI INFORMÁCIÓS ÉS MONITORING RENDSZER) (1995): Módszertan. Földművelésügyi Minisztérium Növényvédelmi és Agrár-környezetgazdálkodási Főosztály, Budapest.
WALTNER I., FUCHS M., MICHÉLI E., LÁNG V., (2012): Hazai archív talajadatok beillesztésének lehetőségei nemzetközi adatbázisokba. *Agrokémia és Talajtan*, 61, 2. 263-76.