A novel approach for mapping WRB soil units – A methodology for a global SOTER coverage

Endre DOBOS[1], Péter VADNAI[1], Károly KOVÁCS[1], Vince LÁNG[1], Márta FUCHS[2] and Erika MICHÉLI[2]

Abstract

Traditional soil maps present soil information in the form of categorical classes of soil types classified on the appropriate level of the applied classification system corresponding to the scale. Soil complexes and associations have been used to describe polygons. This kind of data structure is useful to characterise an area by explaining its soil resources. However, it is difficult to convert these complex categorical units into a simple digital variable, the usage of this kind of data in a digital environment is limited. Users often need single properties instead of the complex classes. Additionally, the problem becomes more complicated when soil information of different origin, based on different classification systems has to be integrated into a common, harmonised database. The presented methodology is part of the efforts to develop a global SOTER (World Soil and Terrain database) coverage and contribute to the global soil observing s as part of the Global Earth Observing System of Systems (GEOSS). The aim is to determine and map the relevant soil properties, horizons and materials following the diagnostic concepts of the World Reference Base (WRB) for soil resources and derive the occurrence probability of soil classes (WRB reference soil groups) of certain spots with the application of remote sensing and digital soil mapping tools. The developed method is referred as the e-SOTER approach and is capable of producing a stack of soil diagnostic element layers with the likelihood of their occurrence within each pixel and a layer of WRB reference soil groups (RSG). This new approach may provide better input for modellers and predict the spatial continuum of the soil cover in a much better resolution than the traditional polygon based approaches. At the same time the diagnostic elements, as building blocks of the classification systems, help the correlation of the national soil classes into integrated databases and maps.

Keywords: soil types and classification systems, soil classification methodology, World Soil and Terrain database, global soil observing, WRB reference soil groups

Introduction

Soil is a continuous cover on landscapes. Appropriate use of this natural resource requires knowledge of its spatial heterogeneity. The physical, chemical and biological properties are distributed differently in the land surface. These properties are often linked together to specific combinations as a response to the soil forming environment (Glinka, K.D. 1927; Jenny, H. 1941). The commonly occurring soil forming process associations have been used to define soil classification categories. Guy Smith was the first who recognised the results of these processes regarding diagnostic features having measurable properties and quantitative characteristics to classify soils (Eswaran, H. 1999). The early and later editions of Soil Taxonomy (Soil Survey Staff, 1999), the legend of the FAO UNESCO Soil Map of the World (1974–1981), the WRB (IUSS Working Group WRB, 2007a, b) and several modern national systems are based on these diagnostic principles. This approach

[1] Institute of Geography and Geoinformatics, University of Miskolc. H-3515 Miskolc, Egyetemváros. Correspondent's e-mail: ecodobos@uni-miskolc.hu

[2] Department of Soil Science and Agro-chemistry. Szent István University, H-2100 Gödöllő, Páter Károly u. 1. E-mails: fuchs.marta@mkk.szie.hu, micheli.erika@mkk.szie.hu

is still valid and represents the state of the art knowledge of soil science.

On the other hand, soil mapping requires several generalisation steps and compromises to dissolve the within-class heterogeneity and create classes that are valid, homogeneous and meaningful, of which these classes have been used for soil mapping. Mapping and interpreting soil classes need soil experts to extract certain soil properties or property associations to stipulate the proper use of soils. It is especially true in case of small scale datasets and maps, when not only the soil property association (soil types) but associations of the soil types (soil associations or complexes) have been used to characterise the area.

Any use of these maps requires the disaggregation of the information and the allocation of the soil properties to specific environmental conditions within the polygon. It used to be done by soil experts using mental soil-landscape models. However, users from other fields of science have difficulties in interpreting the content.

The SOTER approach introduced the application of physiographic and – when it was available – lithological information to delineate the spatial units of small scale maps and datasets (ISRIC 1993; van Engelen, V.W.P. and Wen, T.T. 1995; European Soil Bureau Scientific Committee, 1998; King, D. *et al.* 2002). The soil information is assigned to these terrain units. These geographic units are difficult to be used for any modelling activity, because no spatial disaggregation of the soil type association can be done to geo-locate, spatially define the information.

The past three decades has changed the world of cartography and database development methodology. GIS tools have been developed to analyse and present spatial information more efficiently. In the meantime, a tremendous amount of digital spatial datasets has been collected and compiled, such as digital elevation models and satellite images providing high-resolution environmental covariates for soil mapping. Digital soil mapping has become a very effective tool in soil science, and several applications have been published (Dobos, E. *et al.* 2000, 2006, 2007, 2010, 2013b; Worstell, B. 2000; McBratney, A.B. *et al.* 2003; Lagacherie, P. *et al.* 2006; Szatmári, G. *et al.* 2013; Pásztor, L. and Takács, K. 2014; Sisák, I. and Benő, A. 2014; Szatmári, G. and Pásztor, L. 2016). Soil data is needed for several applications, but only in a format that can be integrated into the existing models. The majority of soil data users require data in raster format with values of certain properties, like pH, clay content or soil organic matter content. Some of these variables are used as it is, as direct inputs into the model, while others are used to estimate complex soil features and properties, like diagnostics, features and horizons. These sophisticated features, like the WRB diagnostics, present valuable information for several applications. Taking the groundwater impact as an example: information on the presence of temporal water saturation in the soil occurs as very important for several environmental, agricultural or civil engineering applications. Water saturation is the function of several soil and environmental properties, such as climate and terrain conditions and soil properties like compaction, total and differential porosity, bulk density, depth to the groundwater, dynamics of its fluctuation etc. We need to know all soil and environmental properties to predict their collective impact on soil. In contrast, soils showing gleyic properties prove that all the required conditions are present at the same time to develop hydromorphic impacts on the soil.

The WRB diagnostic horizons and other diagnostic elements represent a set of well-defined characteristics of a soil horizon. Each of them is important to describe the soil system and can be interpreted to provide information on the proper use and functions of soil. The established (predicted) presence of these horizons and features gives direct answers to the most common questions with no need for further processing, calculating or interpreting several properties to derive the information needed. Therefore, WRB diagnostics have been already used in some applications (Dobos, E. *et al.* 2010; Liess, M. *et al.* 2012; Pásztor, L. *et al.* 2013).

The goal of this study was to develop a novel approach to present complex soil properties in the form of WRB diagnostics that can be used efficiently for modelling using georeferenced soil data and auxiliary digital data sources, like remote sensing and digital elevation model (DEM) data. The developed method may support the completion of the global coverage of the SOTER database by providing a digital soil mapping (DSM) toolset to create harmonised soil information for the SOTER polygons. This method is referred as the e-SOTER approach and produces a stack of soil diagnostic property layers showing the likelihood of their occurrence within each pixel and a layer of the Reference Soil Groups (RSG) of the WRB.

Methods

The overall framework

The e-SOTER approach is based on the major building units of the WRB classification system such as the diagnostic properties and horizons (DPDH). It attempts to estimate the spatial occurrence probability of DPDHs using remote sensing, digital terrain data and pre-processed legacy data - as training dataset. As the WRB includes numerous diagnostics, a limited set of significant units has to be defined by an expert group based on the existence and significance of horizons, properties and materials of the target area. Training datasets for this group of diagnostics are derived from legacy data. Each training dataset consists of points or areas with known existence or absence of the property in question. Therefore, using these training datasets for classifying a complex MODIS/SRTM based image results in numerous continuous layers for each property having the probabilities of the existence of the diagnostic property. The major advantage of this approach is that it provides the needed thematic information on essential soil properties such as; texture, organic matter, salt content etc. Additionally, using these DPDH layers, a WRB-based simplified classification scheme is developed to identify the WRB soil types for each pixel. The success and detail of the approach depend primarily on the quantity and quality of the input training dataset.

The workflow of data development has the following major steps:

– Development of important input physiographic and parent material layers for the classification – landform, bare rock, the texture of the unconsolidated parent material (*Table 1.* lines 1–4);
– Definition of the significant WRB diagnostics (properties and horizons – DPDH) needed to characterise the major soil properties and features of the mapped area *(Table 1.* lines 5–16);
– The collection of legacy data (soil profiles or large scale soil maps) and the development of the training datasets;
– DSM procedure to develop the layers of the WRB diagnostics (properties and horizons);
– The definition of the classification rules to define the WRB RSGs.

Table 1. The list of important terrain, texture and WRB properties, diagnostics and horizons (DPDH) in the gridstack

1	Terrain type with 5 classes (stratification map): 1. fine plain 2. coarse plain 3. hill 4. mountain 5. water
2	Consolidated-unconsolidated image
3	Texture image
4	Bare rock image
5	Spodic Horizon Class Probability
6	Argic Horizon Class Probability
7	Cambic Horizon Class Probability
8	Vertisol Class Probability (only Vertisol vertic horizons)
9	Salic Horizon Class Probability
10	Natric Horizon Class Probability
11	Gleyic-Stagnic-Reducting conditions Class Probability
12	Mollic Horizon Class Probability
13	Calcic Horizon Class Probability
14	Calcisol Class Probability (only Calcisol calcic horizons)
15	Dystric Class Probability
16	Eutric Class Probability

The study area

The pilot area is located in Central Europe and covers the territory of Austria, Hungary, Slovakia, Czech Republic, Southern Poland, and a small part of Germany and Romania. This window has been chosen to cover the Central European pilot area of the e-SOTER project *(Figure 1).* The final area is much larger than the pilot; it follows the tile borders of the SRTM (Farr, T.G. and Kolbrick, M. 2000) and MODIS images that fully includes the e-SOTER pilot. Training data has been available only for the pilot window and the territory of Hungary.

The terrain and the soils of the area are quite variable (Pásztor, L. *et al.* 2018). It includes parts of the Alps, the Carpathian mountain range, the Czech-Moravian Mountains, the Pannonian Basin and the southern, hilly and flat region of Poland. The parent material varies as all kinds of consolidated siliceous and carbonaceous rocks occur the area, together with Holocene alluvial and aeolian sediments, and Pleistocene glacial and periglacial materials. The soils on the lowland are mainly Chernozems, Vertisols, Arenosols, Gleysols and Calcisols, while on the hilly and mountainous areas Luvisols, Cambisols, Stagnosols, Regosols and Leptosols are the dominant ones.

Covariates used to derive the thematic layers

To strengthen the performance of the classification, multi-temporal images of MODIS bands were compiled into a 55 layer image representing the visible, Near Infra-Red (NIR), Mid-infrared (MIR) and thermal bands to capture the temporal environmental conditions and changes that reveal to surface conditions. Multi-temporal 8 days MODIS composites were used, five dates evenly distributed over the vegetation period:

– MOD09A1: Band 1-2 (250 m resolution)-7 (Layers 3–7), 500 m resolution;
– MOD11A2: Band 31-32 (Layers 9–10), LST (Land Surface Temperature) Day (Layer 1) and LST Night (Layer 5), 500 m resolution.

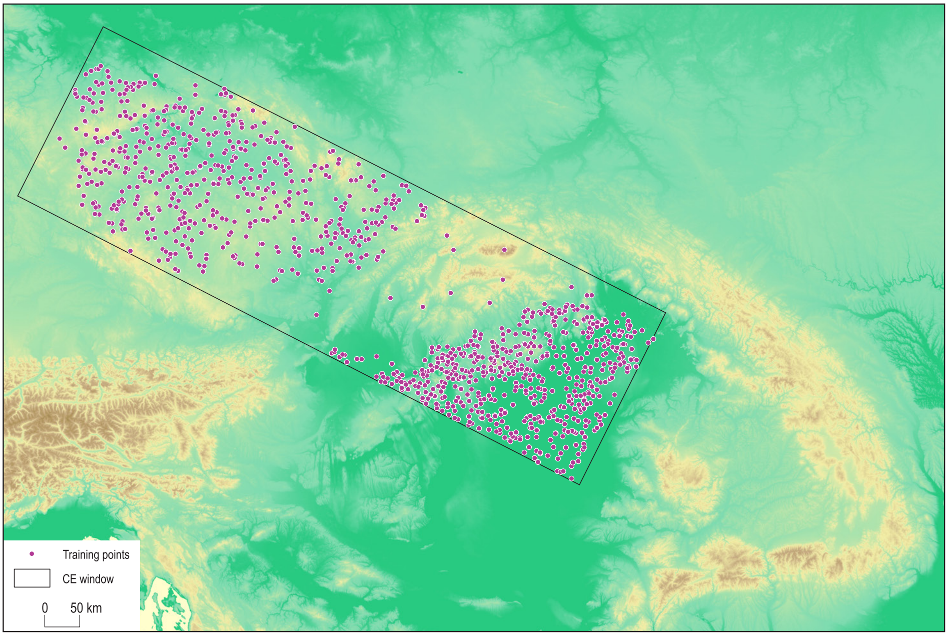

Fig. 1. The pilot window and the distribution of the profile dataset for the Central European window

However, the 55 layers have a significant portion of overlapping information, redundant information in the images, hence a Principal Component Analysis (PCA) was used to decrease the number of input images and de-correlate the information of the bands. The first 15 PCA components were maintained and incorporated into the final image.

Previous studies also suggested using surface temperature information, like the thermal bands of the MODIS (Bands 31, 32) and the LST (Land Surface Temperature) products (night and day) that have been derived from them. The daily temperate fluctuation is a function of the thermal capacity of the surface material, which is the function of the kind of material, texture, colour and water content; primarily the factors of key interest to the study. Therefore, a new normalised band combination was developed and added to the PCA image set for each date. The daily temperature difference was calculated by simply subtracting the LST night from the LST day, and the values were multiplied with the ratio of the LST_{max} / LST_{day} to reduce the effect of the climatic variation due to the difference in potential energy intake from the sun. (Note: LST_{max} = LST maximum value for the whole area.)

There were many attempts recorded in the literature to use band ratios to identify certain lithology classes or to highlight lithology differences in Landsat images. These band ratios were adopted to MODIS and were derived for each of the five dates, resulting in 15 other images, that were added to the final image. Three band ratios adopted after Drury, S. (1987) and Segal, D. (1982) of 6/2 (ferrous minerals ratio), 1/3 (iron-oxide ratio) and 7/6 (clay mineral ratio) were created to represent lithological variations better.

SRTM (Farr, T.G. and Kolbrick, M. 2000) data were used in combination with the MODIS derived layers as well. The basic parameters were the following:

- Elevation (sinks are filled up to a certain level);
- Slope per cent;
- Relief Intensity;
- Potential Drainage Density (Dobos, E. and Daroussin, J. 2007);
- Groundwater level (developed via the interpolation of the SRTM derived drainage network points heights and subtracted from the original elevation values);
- Topographic Wetness Index;
- Upland/Lowland: (elevation range/2 + elevation min – elevation) for a 10 pixels radius circle;
- Convexity (not added to the basic image, used only for the colluvial image derivation).

The listed derivatives are either used in the SOTER methodology already or believed to add significant information for differentiating between the classified parameters. The SRTM images were spatially degraded to the level of MODIS resolution, the final resolution of the image was 456 m, partly to stay close to the original resolution of the MODIS and partly to be multipliable by the SRTM resolution.

Besides of the 43 layers (15 PCA layers, 8 SRTM derivatives, 5 normalised LST difference images and 15 band ratios), three further layers were added to the image to represent the climatic variability. These were the images of Easting and Northing, which defines the geographic location, and the distance from the sea. With these extra 3 layers, 46 layers image was developed and used for the classification.

The development of input layers for the final classification

The SOTER physiography layer development

The SOTER approach is based on the assumption that the landform and the parent material are the most critical factors of the soil formation when working within a relatively small land surface area, like a SOTER polygon, in which the natural macroclimatic variability is negligible. Therefore, the major portion of the climatic variability is due to the terrain that defines the meso and microclimate as well. The vegetation develops in the function of terrain, climate and soil, so the majority of the vegetation variability is already explained by them. This assumption is the basis for the pre-stratification of the area, the physiography and simplified parent material classes that defined the homogeneous units. Five classes have been distinguished: water, mountain, hill, and two plain classes, namely the fine and coarse plains. The terrain classification used the SRTM-based, modified physiographic classification of the SOTER developed by Dobos, E. *et al.* (2007, 2013b) and is shown in *Figure 2.* Elevation and relief intensity variables were used to separate the mountain, hill and plain classes. The plain class was further divided into fine and coarse plains using the MODIS and LUCAS based texture classification image. The fine plain area has clay or loam texture. The coarse plain is the sandy and gravelly textured plain area. The texture class database development used the method defined by Dobos, E. *et al.* (2013b). The terrain/parent material based pre-stratification of the European window is shown in *Figure 3.*

Parent material image-set development

Parent material is vital to define the soil associations within the SOTER units. However, it is often difficult to compile harmonised parent material datasets from legacy data.

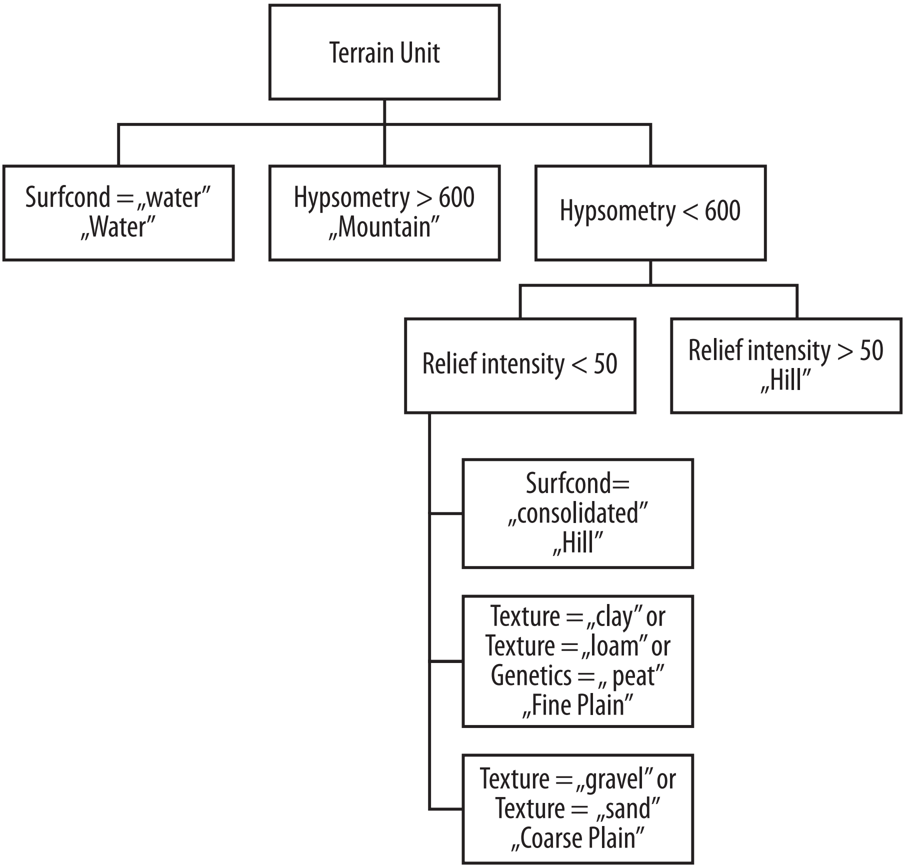
Terrain Unit
Surfcond = „water" „Water"
Hypsometry > 600 „Mountain"
Hypsometry < 600
Relief intensity < 50
Relief intensity > 50 „Hill"
Surfcond= „consolidated" „Hill"
Texture = „clay" or Texture = „loam" or Genetics = „ peat" „Fine Plain"
Texture = „gravel" or Texture = „sand" „Coarse Plain"

Fig. 2. The decision tree for the pre-stratification of the study area. The relief intensity value is calculated for a 1 km diameter circle area.

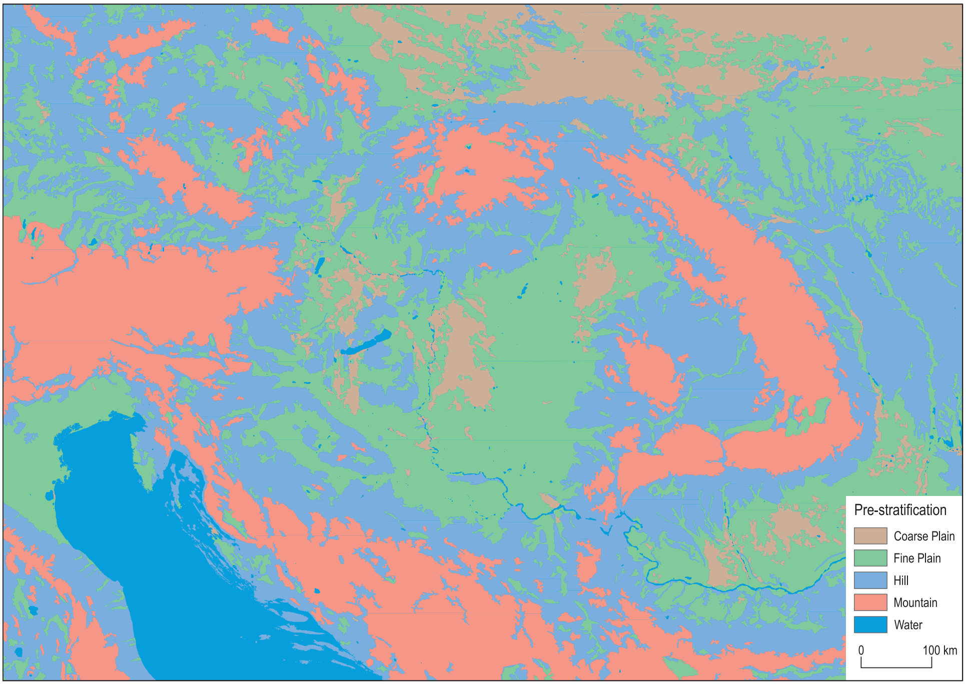

Fig. 3. The terrain/parent material based pre-stratification of the Central European window

Our approach differs from the traditional, legacy data-based approach, where existing surface geology data of different origin is used to define the soil forming units. Our approach comes from the geomorphological approach. The assumption is that different parent materials (PM) have different surface morphology, therefore part of the properties can be estimated by geomorphometric tools. The term "parent material" is used herein a simplified way. The approach is a digital soil mapping procedure based, modified SOTER approach developed within the e-SOTER project by Dobos, E. *et al.* (2013b).

This study is using only the first two levels of the classification tree, namely the consolidated-unconsolidated layer – complemented with the bare rock layer – and the texture, including the diagnostic organic soil material as a separated class. The calcaric nature of the material is handled later in the DPDH classification by their nature.

The final parent material image is a combination of the consolidated/unconsolidated image (a), the bare rock surfaces for the consolidated parent material areas (b) and the texture classes for the unconsolidated areas (c).

a) Consolidated/unconsolidated areas – Unconsolidated material is defined here as a loose inorganic/organic material, that is by nature, accumulated/deposited in a deeper stratum by wind, water or ice (fluvial, estuarine, lacustrine, marine, glacial, aeolian) or by mass movements (like the colluvial materials). The consolidated material – as it is defined here – is the solid rock and its shallow weathering residuum, having mainly the typical mountain soil associations like bare rock/Leptosol/Cambisol, and by genetics, it can be eluvial (locally weathered residuum), colluvial or bare rock. The widening of the content with the weathering residuum is an unavoidable compromise because the existing soil maps with parent material information for this kind of areas describe only the underlying rocks and gives no information on the properties of the weathered material. This statement was concluded by the authors; it is still not commonly agreed.

Maximum likelihood supervised classification algorithm using the combined image of 46 layers was applied to derive the consolidated/unconsolidated image (Dobos, E. *et al.* 2013b). There are only stochastic relationships between specific terrain parameters and the consolidatedness of the PM. It is also true for the RS images, especially in the temperate and tropical zones, when the vegetation masks out the PM signal of the images.

Training data was limited for the window as only 10 per cent of the whole area was covered with legacy data. Training areas for the Czech Republic and the Hungarian part of the window were used. The data sources were interpreted in the training areas for the classes defined. The consolidated and unconsolidated parts are handled and classified differently from this point.

b) Bare rock image – The bare rock classification was done using an NDVI (Normalized Difference Vegetation Index) image generated from the peak of the vegetative period, like summer in the Central European (CE) window, when strong vegetation cover is expected. Only areas having no soil and thus vegetative cover are expected to have very low NDVI value. A threshold was set by selecting known areas with bare rock and the corresponding; representative NDVI values were identified and used for threshold the NDVI values. This value and the procedure in general works very well in the temperate and tropical zones, it has been tested for France/United Kingdom and Southwest China as well (Dobos, E. *et al.* 2013b).

c) Developing the texture class layer – The most critical part of the procedure is the training data. The optimum case is when relatively high-resolution training data is available with well defined, none overlapping classes. 1:100K to 1:250K data sources are commonly available for the developed part of the World, which contain aggregated but still concrete classes (not associations). These data sources can be used as direct inputs for the supervised classification.

The texture classification was done the same way as the consolidated/unconsolidated layer, using the 43 layer combined image and training data for the supervised classification. The legacy training set for the training area was the same as well, but this dataset was complemented by the LUCAS dataset (Tóth, G. *et al.* 2013; Orgiazzi, A. *et al.* 2018). The sand, silt and clay percentages of the LUCAS, TIM (Várallyay, Gy. *et al.* 1995; Várallyay, Gy. 2012) and the Czech topsoil datasets were converted to texture classes and used for the supervised classification. The texture layer is shown in *Figure 4.*

The definition of the significant WRB diagnostics (properties and horizons) and the classification rules to define the soil classes

Typical soil types for the four terrain/parent material classes and their corresponding diagnostic horizons and properties were defined by expert knowledge and listed as required information layers for the soil characterisation. This list of the selected DPDHs was then compared with the local legacy database to test for missing DPDH or real existence/significance of the selected DPDHs in the database. New DPDH was added to the list when the legacy data proved its importance, or DPDH was removed when the legacy data did not contain information on the feature, or the frequency of occurrence was too low to support the classification algorithm. In some cases, the DPDH was kept, even if the data was not supporting its importance, but the experts flagged it as an important factor for the classification.

The last step of the procedure was the classification of the WRB reference soil groups (RSGs). A WRB classification tree was developed to estimate the most likely RSG for each pixel. A nested conditional function was developed to classify/define the corresponding RSG for each pixel using the variables of the stratification image and the DPDH images – described below. This classification tree depends strongly

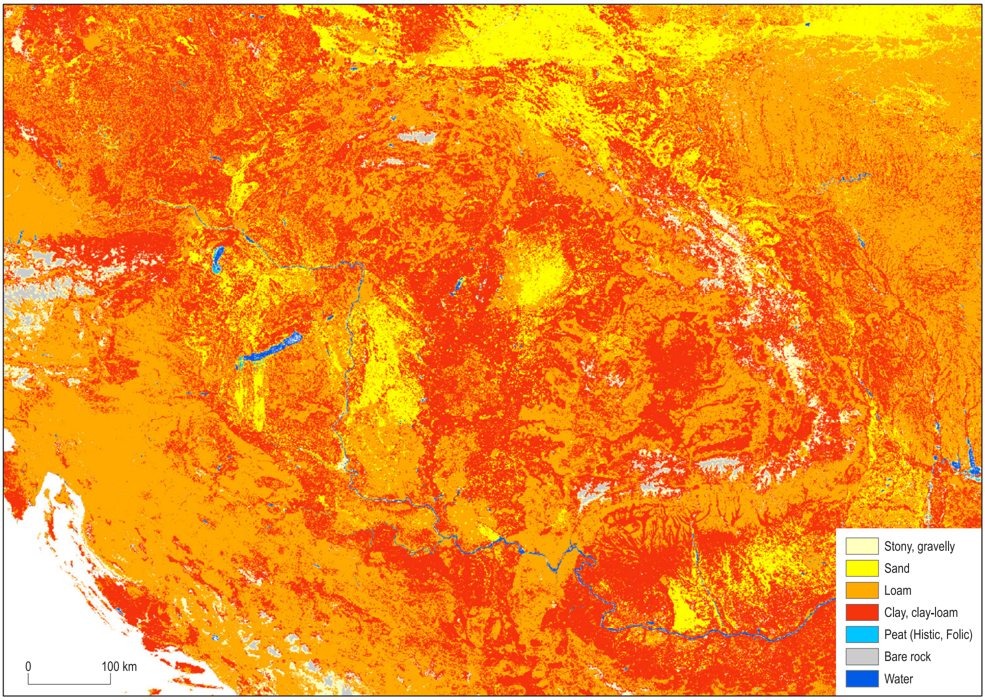

Fig. 4. The classified texture classes for the pilot window (Dobos, E. *et al.* 2010).

on the detail and content of the legacy data. The one shown below was developed for the Central European window and was adapted to the available set of soil information. The more complete soil data, the more detailed classification tree and the more refined RSG classes can be elaborated. On the other side, similar or related soil types may need to be combined into more complex units when input data is limited.

This classification tree was developed specifically to the Central European soil associations and data availability conditions of the region. Other regions may require different trees with different terrain and DPDH elements and different rules explaining the soil formation. Narrowing the potential variability to the ones occurring in the area simplifies the classification tree and makes it more efficient and site specific. However, this methodology requires strong local knowledge and understanding of the geographic distribution of the soil resources and the major driving forces of the genetics of the soil types.

One may recognize that Fluvisol, that should be common in the area, is not included among the RSG classes. The reason is that all major rivers of the region have been chanelized and the natural floodplains have been narrowed by dike systems built along the rivers. The remaining active floodplain has been cut to only a 100 m or narrower strips along the river, which is not wide enough to present on the map.

The classification tree followed the WRB key RSG order to make sure that the final categories match the WRB classification. For example, if a soil occurs on the plain area and has a mollic or chernic horizon and also has a clay, heavy clay texture – which is very common on the Great Hungarian Plain – the soil keyed out as Vertisols, while the remaining ones have classified to Chernozem/Kastanozem RSG.

Development of translation algorithms and correlation tools for the harmonization process

The legacy soil data originated from different sources (i.e. national and international data sets) is usually very variable, because the collection, determination and classification of it are based on various methodologies. For international projects, European and global initiatives using data from diverse sources, preliminary harmonization is necessary to provide a standardized input dataset for further research.

For soil data harmonization international standards are provided. Within the European Union, the harmonized master horizon designation, subordinate characteristics and site descriptions follow the 2006 edition of the FAO Guidelines for soil description (FAO, 2006). The classification of soils and the related diagnostic horizons, properties and materials are described and coded according to the World Reference Base for Soil Resources (IUSS Working Group WRB, 2007a, b). The success of harmonization largely depends on the quantity and quality of the input datasets.

In many cases, the number and completeness of field observation and laboratory data is insufficient for proper correlation. The determination of the WRB diagnostics, Reference Soil Groups and qualifiers often requires morphological, chemical and physical soil data as well. Simplification of the requirements and expert judgment is often needed to overcome the shortage in information.

In this study a computer assisted determination of the diagnostics was applied for all datasets with simplified requirements of the WRB (2007a, b). The simplification was adjusted to the availability of the required information. In many cases, even the simplified requirements were not available and expert judgment was used to determine the presence or absence of the diagnostics. Since many of the diagnostic features require morphological criteria that are not commonly part of the legacy data sets, a significant portion of uncertainty is introduced into the procedure.

In order to generate the training dataset for the image classification 31 simplified algorithms for the WRB diagnostic horizons, properties and materials and 29 simplified algorithms for the WRB qualifiers were performed on the harmonised dataset. The output database contained simple information, such as the presence or absence of the given diagnostic criteria for each profile. Classification was performed only when sufficient information was available for the given criteria to avoid additional uncertainty of the training dataset. Due to the lack of information in the database a large number of profiles were necessary to generate a suitable number of training points for each DPDH.

The collection of legacy data and the development of training datasets

The representative datasets could be points or polygons having unique identifier for each of the objects. A Microsoft Excel sheet was created with the identifier column and one column for each selected DPDH. Experts derived the existing DPDHs for each object based on the provided classification units and the measured properties of the legacy data. A value of “1” was assigned to each object, when the DPDH in question was existing, and a “0” value for non-existence. The cell was left empty when no decision could be made. Therefore, two classes were created for each DPDH, the existing class and the non-existing one; while the empty ones were not used for the classification of the specific DPDH.

Additional data points were needed when the resulting number of points for the DPDH was insufficient. Due to the matrix inversion steps used in the calculation process of the image classification, the number of training pixels for each class used in a maximum likelihood classification procedure has to be at least one more than the number of image layers used for the classification. In our case, the image had 46 layers, so the minimum number of training pixels had to be at least 47. In case of less than 47 training pixels for the classes, the legacy data points were extended "artificially" to a larger region using statistical thresholds of Euclidian distance for the surrounding pixel values to make sure that similar pixels are involved into the training procedure. A similar procedure to transform the point dataset into a raster with a size of 1 km^2 was used as well, and the whole area of the pixel was used as a training area. This latter approach is simpler, however unavoidably introducing unsupervised uncertainty into the procedure. Therefore, it was used only when large numbers of points were involved. For the Central European window 1091 profiles were available. The distribution of the profiles is shown in *Figure 1.*

The development of layers of DPDH using image classification procedures

Probability classifications for the DPDH were done using the MODIS/SRTM image and legacy data based training dataset. Signature files for each DPDH were created from the training dataset and used for Class probability classification. The classification was performed using the ClassProb command of the ArcGIS software setting the range of values between 0 and 100, where the value of 50 means the equal possibility of the two classes (existing or missing DPDH), the higher values mean higher likelihood for the existence, while the lower for the missing DPDH. The mapped area was pre-stratified into the terrain/PM classes described above and the classification of the DPDHs was done simultaneously and individually for the four regions. In the end, the 4 (5 with the water class) classified images of the same DPDH were mosaicked together to create the final probability image for each DPDH.

Validation methodology

The details of the validation procedure have been described by Dobos, E. *et al.* (2013a). The final dataset has several DPDH probability layers and some categorical ones, like the RSGs and the texture classes. The percentage value of the occurrence probability can be taken as a probability of being correct in the classification, or – having an approximately 500 by 500-metre pixel area that being potentially heterogeneous – the spatial coverage/existence or share of the given feature within the pixel area. In order to validate the percentage values, we needed to know the real share or existence of the DPDH within the cell. Therefore, a new validation dataset has been collected and developed. Four datasets are planned for the Visegrád Countries (Czech Republic, Hungary, Poland and Slovakia) using the same procedure.

The validation was based on randomly selected validation sites. These sites were moved to the pixel centre, where a profile was excavated, described and all observed DPDH were recorded. Four additional auger holes were deepened at 100 m distance to the North, East, South and West directions from the opened profile. These auger holes were described in the same way and DPDHs were recorded as well. In the end, each validation site had five observations within the pixel, and existence percentages of 20, 40, 60, 80 or 100 could be calculated as the likelihood of occurrence within the pixel. These numbers can be used to characterise the homogeneity of the pixel and to validate the results of the probability classification. The validation was done for the RSG and texture data of the area of Hungary, the distribution of the validation sites is shown in *Figure 5.*

Results and discussion

Any DSM exercise and model development requires a deep understanding of the soil resources and the soil forming environment of the target area. Therefore, the first step in this study was the definition of the potential soil classes that occur within the area. The workflow of this step follows the original SOTER framework. It starts with the stratification of the landscape into homogeneous units defined by physiography and parent material. These

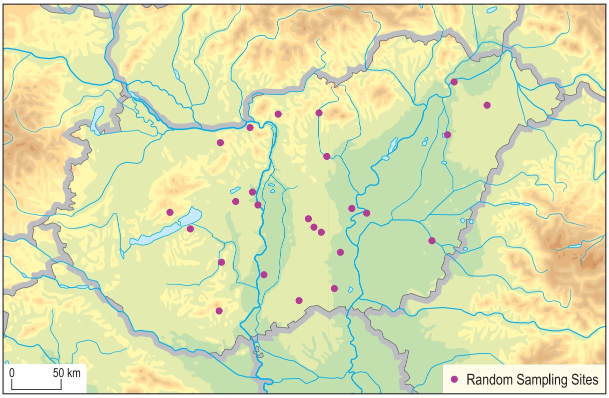

Fig. 5. The randomly selected validation sites around Hungary

two maps were combined to create the final stratification for the area (see *Figure 3*). The texture of the unconsolidated sediments on the plains has a strong correlation with the origin of the parent material in Central Europe. Clayey materials of the plains always have an alluvial origin and occur in the lower lying floodplains. The loam class refers to the deposited in situ or transported loess with a little bit higher elevation and still relatively low relief, while the sands are alluvial and later reworked by the wind. The first two classes often have low relief, while the dune formation of the sand regions results in stronger relief. That is why relief intensity, potential drainage density (PDD) (Dobos, E. and Daroussin, J. 2007) and the groundwater level layers had a significant contribution to the separation of these classes (Dobos, E. *et al.* 2013b).

Any kind of existing texture map can be integrated into the process. However, due to the lack of high resolution, consistence texture datasets, we used LUCAS data within a DSM procedure to derive the texture layer. The LUCAS – combined with the Hungarian, Czech and Romanian monitoring point datasets – has been reclassified into three texture classes, namely sand, loam, and clay, and a maximum likelihood classification using the integrated MODIS/SRTM dataset was performed. The overall class performance calculated with the leave-one-out method was 45.5 per cent with 22.8 per cent Kappa value. The visual check indicated a strong over-classification of the clay class over the loam. Therefore, additional loam areas were identified from different soil maps and added to the training dataset to refine the results. With the addition of the new training data, the overall class performance was increased to 88.7 per cent (with 26.1% Kappa), but the strong over-classification of the clay class was still evident.

These classes make a real and significant difference between the soil forming processes of the plain areas and separates the different soil associations. Therefore, the texture was used to refine the stratification of the plain areas and divide them into two subclasses of fine and coarse textured parent material. The soils of the sand and gravelly-sand regions are different from the ones forming on loamy or clayey material. Further differentiation within the fine texture class to loam and clay classes would have been advantageous, but the input texture image did not make a good separation between these classes. Areas having clay-loam texture – prevalent on the alluvial area of the Great Hungarian Plain – were classified as clay. Therefore, clayey soils – and thus the Vertisols – are artificially overrepresented in the target area. Fortunately, the separation between the fine and coarse textured areas is much more reliable and makes a good input for further classification (see *Figure 4*).

Stratification was followed by the definition of soil associations, the existing WRB reference soil groups for each stratification classes. The list of WRB reference soil groups is given in the last column of *Figure 6*. A minimum set of WRB DPDH was defined that is needed to classify the RSG classes – lines 5–16 in *Table 1*. By using the DPDH set and the stratification classes, a simplified classification tree was developed to classify each pixel according to its most likely RSG class *(Figure 6)*.

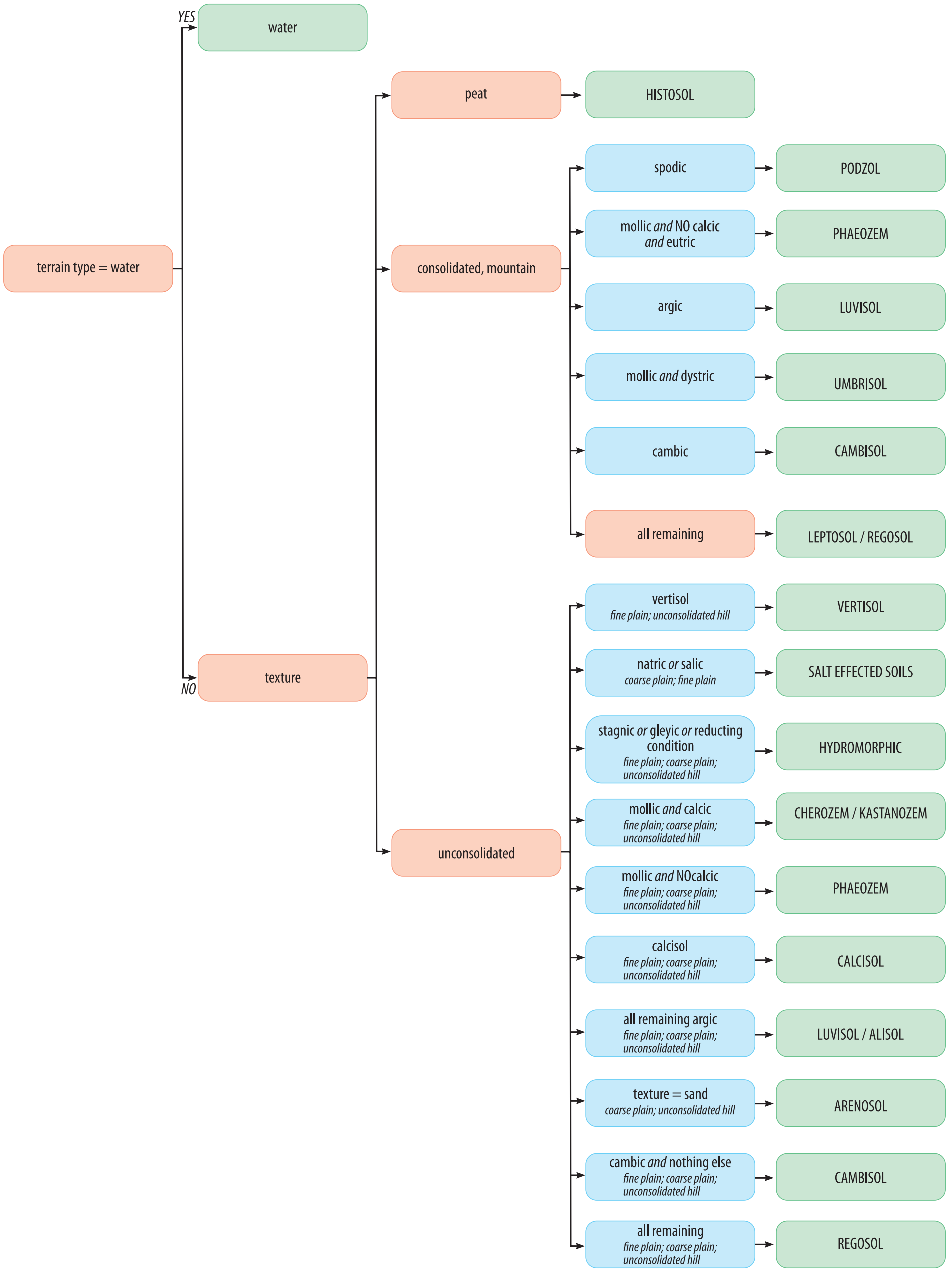
terrain type = water
YES
water
NO
texture
peat
HISTOSOL
consolidated, mountain
spodic
PODZOL
mollic and NO calcic and eutric
PHAEOZEM
argic
LUVISOL
mollic and dystric
UMBRISOL
cambic
CAMBISOL
all remaining
LEPTOSOL / REGOSOL
unconsolidated
vertisol fine plain; unconsolidated hill
VERTISOL
natric or salic coarse plain; fine plain
SALT EFFECTED SOILS
stagnic or gleyic or reducting condition fine plain; coarse plain; unconsolidated hill
HYDROMORPHIC
mollic and calcic fine plain; coarse plain; unconsolidated hill
CHEROZEM / KASTANOZEM
mollic and NOcalcic fine plain; coarse plain; unconsolidated hill
PHAEOZEM
calcisol fine plain; coarse plain; unconsolidated hill
CALCISOL
all remaining argic fine plain; coarse plain; unconsolidated hill
LUVISOL / ALISOL
texture = sand coarse plain; unconsolidated hill
ARENOSOL
cambic and nothing else fine plain; coarse plain; unconsolidated hill
CAMBISOL
all remaining fine plain; coarse plain; unconsolidated hill
REGOSOL

Fig. 6. Simplified classification tree for the WRB RSGs

An ArcGIS module containing the nested conditional function system was developed and made available to the public. This module requires an input gridstack with a predefined structure and standardised layers of DPDH (see *Table 1*). The input layers of the

Table 2. The validation results for the Hungarian territory of the image

Nr.	Classification category		Texture class	
	Field observation	Estimated	Field observation	Estimated
1	Lamellic Arenosol	Regosol	sand	clay
2	Calcic Chernozem	Vertisol	silt	clay
3	Calcic, Endogleyic Chernozem (pachic, siltic)	Vertisol	silt	clay
4	Calcic Chernozem (pachic, siltic)	Vertisol	silty clay	clay
5	Endocalcic, Vertic Chernozem (pachic, epiruptic, episiltic)	Vertisol	loamy clay	clay
6	Calcaric Arenosol	Arenosol	sand	sand
7	Calcic Chernozem	Vertisol	loamy clay	clay
8	Calcaric Arenosol	Arenosol	sand	sand
9	Calcic Endogleyic Chernozem	Chernozem	silt loam	clay
10	Endogleyic Regosol (calcaric)	Hidromorphic	loamy clay	clay
11	Pheozem/Kastanozem	Chernozem	silt loam	loam
12	Kastanozem/Chernozem	Regosol	loamy clay	loam
13	Leptosol	Luvisol	clay	loam
14	Chernozem/Gleysol	Histosol	silt	peat
15	Haplic Calcisol (siltic)	Luvisol	silt	consolidated
16	Luvic Phaeozem (clayic)	Luvisol	clay	consolidated
17	Endogleyic, Cutanic, Luvisol (siltic)	Luvisol/Alisol	loam	loam
18	Calcic Chernozem (pachic)	Chernozem	sandy loam	clay
19	Chernozem	Chernozem	silt loam	loam
20	Chernozem	Chernozem	loam	loam
21	Chernozem/Calcisol	Regosol	loamy sand	clay
22	Regosol	Regosol	loam	loam

gridstack are developed with a standard RS classification procedure using the MODIS/SRTM image described in the 2.3. section of the methods part of the paper. This probability classification approach is based on the signature file/training dataset, which was developed from legacy data using several transformations, translation and correlation algorithms and expert knowledge. Examples of these images are given in *Figure 7.*

Figure 8 shows the final product of the classification, having the WRB RSG classes assigned to each of the pixels. Only the Hungarian part of the window is described here in detail, because the rest of the image has not been validated by unbiased data. However, the general trends of soil class distribution are also recognisable and matches with the legacy soil maps.

A group of experts has interpreted the Hungarian part of the image. This image corresponds well with the known picture of the soil class distribution of the area. There were two major comments on the content. The first one is the overestimation of the Vertisol area, which was due to the texture map. It was stated earlier that the areas having clay-loam texture, the most dominant texture of the alluvial areas of the Hungarian plain, was classified mainly to the clay class. That has increased the potential area of Vertisol occurrence and resulted in some misclassification between the fine textured Chernozems

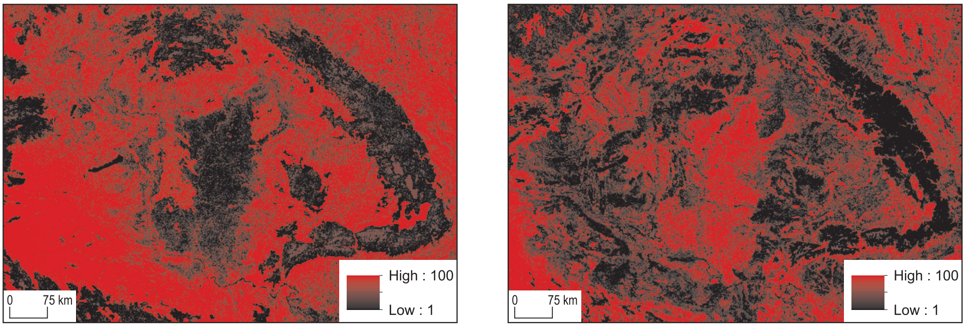

Fig. 7. Examples of the probability layers for the WRB Argic and the Mollic horizons

and the Vertisols. However, even the field differentiation between the two classes is often difficult and makes the training data sometimes unreliable. Especially, because the Hungarian soil classification system does not recognise Vertisols as a separate class, and no diagnostic criteria had been collected at the field for identifying the vertic properties, thus no related input data is available. As there is no one to one correlation between any Hungarian soil classification unit (soil type) and the WRB Vertisols, this information was extracted indirectly from related properties, like clay content, which is not always satisfactory and significant uncertainty may have been introduced this way.

The second comment was on the Calcisols. It was very interesting to see the Calcisol RSG on the Hungarian soil map. A Calcic horizon has three basic criteria to meet, 15 per cent or more total $CaCO_3$, has at least 5 per cent secondary carbonate and has a minimum thickness of 15 cm. There are large sandy and loess areas in Hungary, where huge amount of primary carbonates – over 15–20 per cent – are present in the parent material. This carbonate is partly leached entirely from the upper horizons and accumulated in the deeper horizons as accumulated secondary carbonate. In case of sand having no any significant diagnostic horizon other than the calcic, the soil keys out as Haplic Calcisol (Arenic) according to WRB classification, which is quite a common situation in the Danube–Tisza Interfluve area. A similar situation may occur on loess, where the calcic horizon is formed under a mollic that may have been eroded away due to intensive agriculture and high relief resulting in a Haplic Calcisol (Siltic). These two kinds of conditions are quite common and represent significantly large areas of Hungary, but it has not been recognised in the Hungarian classification yet, and these soils were classified as Arenosols or Regosols.

The concept of Calcisols has been developed for the semiarid regions with strong evaporation and $CaCO_3$ accumulation from the $CaCO_3$ rich groundwater, but diagnostics based classification systems describe the current features and soil genetics has only secondary importance. Not following the diagnostic rules may result in a definite inconsistency in any of these datasets. Subjectivity, or having a preconception in soil classification or correlation process is quite a common problem and is difficult to overcome. One of the main advantages of this approach is the objectivity of the classification rule.

In order to validate the results, the Validat. DSM dataset was used (Dobos, E. *et al.* 2013a, 2014). There were 23 randomly selected validation sites in Hungary distributed all over the country (see *Figure 5*). Each site had five observations, one profile and four auger holes. This procedure was developed to handle the within-pixel heterogeneity of the soils. *Table 2.* shows the comparison results on the RSG and the texture classes.

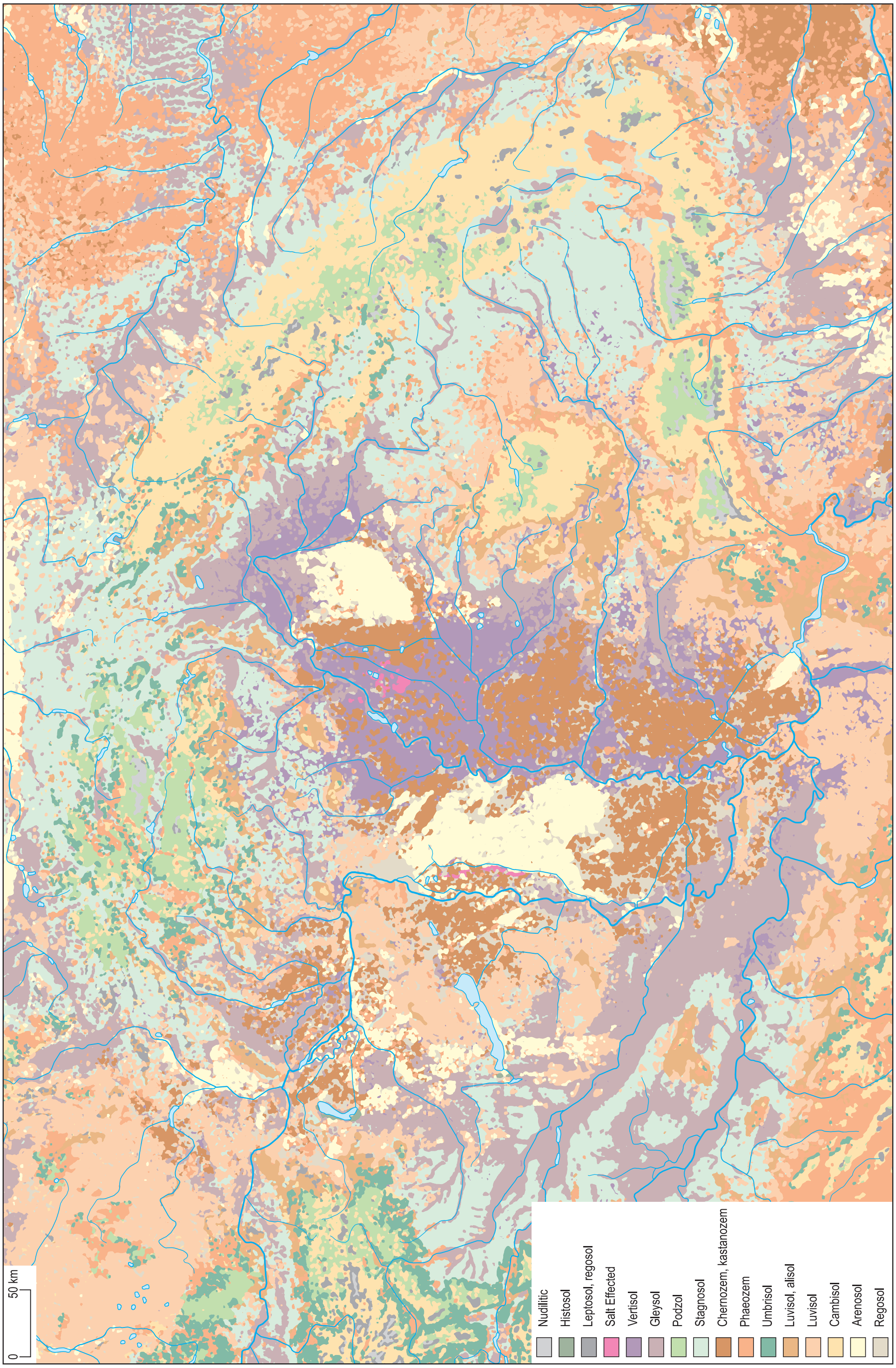

Fig. 8. The output image of the WRB RSG classification modul

Observations 2, 3, 4, 5 and 7 represent the Chernozem/Vertisol problem, that has been identified by the experts as well. These five observations represent the same genetic area and are located in the valleys of the Tisza and Körös rivers, where the silt-loam and clay-loam textured alluvial sediments are the dominant ones, but were partly misclassified as clayey textured soils and as a consequence of this to be Vertisols. Arenosols, Chernozems and Luvisols are mainly classified well. Even in the case of a few misclassifications, the diagnostic feature of the estimated DPDH is partly there, like the argic for the Luvisols, or gleyic, endogleyic for the Hydromorphic soil types (combined class of Gleysols and Stagnosols). A similar trend is evident, in the texture class comparisons as well, the majority of the misclassified classes are located in the Vertisol/Chernozem problem area.

Conclusions

Traditional soil maps are no longer able to present our knowledge on soils in a format that matches the need of interdisciplinary users, have the thematic and spatial resolution comparable with other digital data sources or that fits into a GIS-based modelling environment. Soil Database developers focus more on the property based maps with measured or more commonly estimated values, than on the soil classification category-based ones. These category maps require soil expert knowledge to interpret their content and translate them to specific properties and processes.

There are several important, commonly used properties, which are very difficult to measure and are usually derived from soil classes using pedotransfer formulas. The efficiency and uncertainty depends largely on the input data quality and resolution. This situation is not likely to change in the near future, soil science and its knowledge on soil genesis and processes expressed in the classification categories is still needed for soil data development. A new generation of soil maps that meet the requirements of data users and present our qualitative knowledge on the soil processes, like this RSG map, is needed.

This paper presented a novel way of soil information and soil database development using legacy data and DSM tools. The output is a multi-layer dataset containing several important WRB diagnostic features, reference soil groups, horizons and properties in raster format, capable of modelling the spatial continuum of the complex soil processes and features. These features alone represent complex properties of the soils, which can be easily linked to soil management and soil function related problems, and can be integrated into any specific model, where complex soil classification categories were inappropriate. The SRTM and MODIS supported development of these layers make use of the high spatial resolution of these covariates describing the variability of the soil forming environment in high detail.

It was concluded, that these images have a lot more detail than any of the previous, national scale maps. Despite its rough thematic resolution – only RSG without any prefix or suffix qualifiers is given – it shows the soil regions and soil associations clearly and also the transitions between them. A huge amount of spatial detail was introduced by the SRTM and MODIS data, which makes this dataset more applicable, even at a regional level. This spatial detail is further strengthened by the additional DPDH layers, which are ready to serve specific requirements without any intermediate interpretation need.

The original database idea and structure follows the SOTER approach, and in its sense, it can be correlated with several legacy datasets, but at the same time, a lot of extra knowledge, spatial details were integrated through the introduction of the high resolution digital data sources, like digital elevation data, or satellite images. While the resulted structure is similar to the legacy datasets, the development procedure is novel. Then legacy datasets were developed by compiling and harmonising soil maps having already generalised information. The traditional harmonisation procedure is often based on only rough estimations and correlation algorithms due to the lack of detailed, specific soil information. This procedure leaves off the use of existing map geometrics and applies point and specific soil property data based DSM tools and approaches. The input variables include the commonly accepted terrain and parent material features agreed on by the traditional soil science community, but in a high resolution, quantitative environment. Besides the resulting soil datasets and attribute data, the most important value of this dataset is the regionalization, the derived spatial patterns defined by the soil forming factors – described by the input datasets. The integration of any of these images into a quantitative data estimation algorithm as the pre-stratification image may significantly improve the model performance. The integration of one RSG layer can replace several covariates that describe the soil forming environment.

The model performance and accuracy are difficult to measure, but the general performance is always the function of the quality and quantity of input calibration, training datasets and the expert knowledge of the modellers. The Hungarian window shows the highest detail, due to the highest amount of data and understanding of the local soil resources. Subjectivity is still involved in a certain sense because some of the datasets are based on field morphology, which is impossible to overcome.

Besides, one of the major improvements has been the application of the updated standards for soil descriptions and soil classification. The master horizon designation, subordinate characteristics and site descriptions follow the 2006 edition of the FAO Guidelines for soil description (FAO, 2006). The classification of soils and the related diagnostic horizons, properties and materials are described and coded according to the World Reference Base for soil recourses (IUSS Working Group WRB, 2007a, b). The new soil map of the Carpathian Basin – where several soil datasets of the different countries occupying the area have to be harmonised – has applied this methodology to produce the soil map of the region (Pásztor, L. *et al.* 2018).

***Acknowledgements*:** Our work has been supported by the FP7 project "Regional pilot platform as EU contribution to a Global Soil Observing System" Grant agreement no.: 211578, financed by the European Commission. The research was also supported by the GINOP-2.3.2-15-2016-00031 "Innovative solutions for sustainable groundwater resource management" project of the Faculty of Earth Science and Engineering of the University of Miskolc in the framework of the Széchenyi 2020 Plan, funded by the European Union, co-financed by the European Structural and Investment Funds. This research was supported by the Higher Education Institutional Excellence Program (1783-3/2018/FEKUTSTRAT) awarded by the Ministry of Human Capacities and by Hungarian National Fund (OTKA) 113171 Program.

REFERENCES

Dobos, E. and Daroussin, J. 2007. Calculation of potential drainage density index (PDD). In *Digital Terrain Modelling. Development and Applications in a Policy Support Environment.* Eds.: Peckham, R.J. and Jordan, G., Berlin, Springer Verlag, 283–295.

Dobos, E., Bialkó, T., Michéli, E. and Kobza, J. 2010. Legacy data harmonization and database development. In *Digital Soil Mapping. Bridging Research, Environmental Application, and Operation. Progress in Soil Science.* Eds.: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E. and Kienast-Brown, S., Dordrecht, Springer, 309–323.

Dobos, E., Carré, F., Hengl, T., Reuter, H.I. and Tóth, G. 2006. *Digital Soil Mapping as a support to production of functional maps*. Luxemburg, Office for Official Publications of the European Communities, EUR 22123 EN.

Dobos, E., Daroussin, J. and Montanarella, L. 2007. A quantitative procedure for building physiographic units for the European SOTER database. In *Digital Terrain Modelling. Development and Applications in a Policy Support Environment.* Eds.: Peckham, R. and Jordan, G., Berlin, Springer, 227–259.

Dobos, E., Michéli, E., Baumgardner, M.F., Biehl, L. and Helt, T. 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. *Geoderma* 97. 367–391.

Dobos, E., Michéli, E., Fulajtár, E., Penížek, V. and Świtoniak, M. 2013a. ValiDat.DSM , a new soil data validation dataset for Central Europe. *Hungarian Geographical Bulletin* 62. (3). 313–320.

Dobos, E., Seres, A., Vadnai, P., Michéli, E., Fuchs, M., Láng, V., Bertóti, R.D. and Kovács, K. 2013b. Soil parent material delineation using MODIS and SRTM data. *Hungarian Geographical Bulletin* 62. (2): 133–156.

Dobos, E., Vadnai, P., Bertóti, D., Kovács, K., Michéli, E., Láng, V. and Fuchs, M. 2014. A novel approach for validating raster datasets with categorical data. In *GlobalSoilMap. A basis of the global spatial soil information system.* Eds.: Arrouays, D., McKenzie, N., Hempel, J., Richer de Forges, A. and McBratney, A.B., London UK, CRC Press, Taylor & Francis Group, 347–353.

Drury, S. 1987. *Image Interpretation in Geology.* London, Allen and Unwin.

Eswaran, H. 1999. Time zero of modern soil classification. *Soil Survey Horizon* 40. (3): 104–105.

European Soil Bureau Scientific Committee 1998. *Georeferenced Soil Database for Europe. Manual of Procedures. Version 1.0.* EUR 18092 EN European Community.

FAO 2006. *Guidelines for soil description*. Rome, FAO.

FAO IUSS Working Group WRB 2007. *World Reference Base for Soil Resources*. Rome, ISRIC.

Farr, T.G. and Kolbrick, M. 2000. Shuttle Radar Topography Missions produces a wealth of data. *American Geophysical Union EOS* 81. 583–585.

Glinka, K.D. 1927. *Dokuchaiev's ideas in the development of pedology and cognate sciences*. Russian pedology. Leningrad, Invest. I. Acad. Sci. USSR.

ISRIC 1993. *Global and National Soils and Terrain Databases (SOTER): Procedures Manual.* UNEP-ISSS-ISRIC-FAO. Wageningen, International Soil Reference and Information Centre.

IUSS Working Group WRB 2007. *'World Reference Base for Soil Resources 2006, update 2007.* 2nd edition. World Soil Resources Reports 103. Rome, FAO.

Jenny, H. 1941. *Factors of Soil Formation. A System of Quantitative Pedology.* New York, Dover Publications.

King, D., Saby, N., Le Bas, C., Nachtergaele, F., van Engelen, V., Eimberck, M., Jamagne, M., Lambert, J.J., Bridges, M., Reinhard, R. and Montanarella, L. 2002. *A method for generalization of a soil geographical database: the example of a transfer of the European database EUSIS at 1:1M to the world SOTER program at 1:5M.* 17th. Bangkok, Thailand, World Congress of Soil Science.

Lagacherie, P., McBratney, A.B. and Voltz, M. (eds.) 2006. *Digital soil mapping: an introductory perspective*. Amsterdam, Elsevier.

Liess, M., Glaser, B. and Huwe, B. 2012. Making use of the World Reference Base diagnostic horizons for the systematic description of the soil continuum – Application to the tropical mountain soil-landscape of southern Ecuador. *Catena* 97. 20–30.

McBratney, A.B., Mendonça Santos, M.L. and Minasny, B. 2003. On digital soil mapping. *Geoderma* 117. (1–2): 3–52.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A. and Fernández-Ugalde, O. 2018. LUCAS Soil, the largest expandable soil dataset for Europe: A review. *European Journal of Soil Science* 69. (1): 140–153, Doi: 10.1111/ejss.12499

Pásztor L., Dobos, E., Michéli, E. and Várallyay, Gy. 2018. Soils. In *National Atlas of Hungary: Natural Environment.* Ed.-in-chief: Kocsis, K., Budapest, Geographical Institute RCAES MTA, 82–93.

Pásztor, L. and Takács, K. 2014. Remote sensing in soil mapping. *Agrokémia és Talajtan* 63. (2): 353–370.

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. (2): 205–218.

Segal, D. 1982. *Theoretical Basis for Differentiation of Ferric-Iron Bearing Minerals, Using Landsat MSS Data.* Proceedings of Symposium for Remote Sensing of Environment. 2nd Thematic Conference on Remote Sensing for Exploratory Geology, Fort Worth, TX, USA. Ann Arbor, Environmental research Institute of Michigan, 949–951.

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

Soil Survey Staff 1999. *Soil taxonomy. A basic system of soil classification for making and interpreting soil surveys.* United States Department of Agriculture. Natural resources Conservation Service. Washington DC, U.S. Government Printing Office.

Szatmári, G. and Pásztor, L. 2016. Geostatisztika a talajtérképezésben (Geostatistics in soil mapping). *Agrokémia és Talajtan* 65. (1): 95–114. (In Hungarian)

Szatmári, G., Laborczy, A., Illés, G. and Pásztor, L. 2013. Large-scale mapping of soil organic matter content by regression kriging in Zala County. *Agrokémia és Talajtan* 62. (2): 219–234.

Tóth, G., Jones, A. and Montanarella, L. (eds.) 2013. *LUCAS Topsoil Survey. Methodology, data and results. JRC Technical Reports.* Luxembourg, Publications Office of the European Union, EUR26102. Scientific and Technical Research Series. Doi: 10.2788/97922

van Engelen, V.W.P. and Wen, T.T. 1995. *Global and National Soils and Terrain Digital Databases (SOTER), Procedures Manual*. Revised edition. Wageningen, FAO ISSS, ISRIC.

Várallyay, Gy. 2012. Talajtérképezés, talajtani adatbázisok (Soil mapping, databases in pedology). *Agrokémia és Talajtan* 61. Supplement. 249–268. (In Hungarian)

Várallyay, Gy., Hartyáni, M., Marth, P., Molnár, E., Podmaniczky, G., Szabados, I. and Kele, G. 1995. *Talajvédelmi Információs és Monitoring Rendszer. 1. kötet. Módszertan* (Information and Monitoring System for Soil Protection. Vol. 1. Methodology). Budapest, Ministry of Agriculture. (In Hungarian)

Worstell, B. 2000. *Development of soil terrain (SOTER) map units using digital elevation models (DEM) and ancillary digital data.* M.Sc. Thesis. West Lafayette, IN, USA. Purdue University Press.