Geologia Croatica
Journal of the Croatian Geological Survey
and the Croatian Geological Society

2018 | 71/1 | 29–39 | 6 Figs. | 5 Tabs. |

www.geologia-croatica.hr

The role of geology in the spatial prediction of soil properties in the watershed of Lake Balaton, Hungary

Piroska Kassai and István Sisák

Department of Crop Production and Soil Science, University of Pannonia, Georgikon Faculty, Deák Ferenc u. 16., 8360 Keszthely, Hungary
(kassai.piroska@georgikon.hu; sisak@georgikon.hu)

doi: 10.4154/gc.2018.04

Article history:

Manuscript received January 05, 2018
Revised manuscript accepted February 19, 2018
Available online February 28, 2018

Abstract

There is no standard methodology which allows the incorporation of geological information into digital soil mapping (DSM) despite the great potential of geology as environmental covariate in DSM. To fill this gap, in this study, a geochemical parent material classification scheme was tested on the watershed area of Lake Balaton, for which soil maps at a finer scale have not yet been created. A parent material map was prepared on the basis of a 1:100 000 surface geology map in order to make the incorporation of soil modelling and mapping possible. Legacy data of 12400 soil sample points was used in order to examine the role of geology in the quantitative distribution of some soil properties and element content (liquid limit, soil organic carbon, $pH_{(KCL)}$, $CaCO_3$, Mg, Cu, Zn, Mn). Results confirm that the SiO_2 content of the parent material influences the properties of the derived soils.

In the second part of the study Random Forest models were developed for three major soil properties (liquid limit, soil organic carbon, pH) with the use of additional environmental covariates: elevation, slope, aspect, curvature, topographic position index (TPI), annual average temperature, annual average precipitation, remote sensing based normalized difference vegetation index (NDVI) and land cover information. The performance and accuracy of the models were evaluated on the basis of the coefficient of determination (R^2) and root mean square error (RMSE), calculated on a randomly selected validation dataset (20% of the database). The models performed with R^2 values of 0.72, 0.6 and 0.68 for liquid limit, soil organic carbon and pH respectively. The importance of variables was also examined in the RF models, and this demonstrated that while geology is among the best-performing predictors, in neither case is it the most important variable. Ninety metre resolution maps of the three major soil properties were compiled by making spatial predictions with the RF models developed here. For validation of the maps, an independent soil database was used, which showed that the prediction performed well on the cultivated area where the concordance correlation coefficients (CCC) were 0.73, 0.73 and 0.69 for liquid limit, pH and soil organic carbon respectively.

Keywords: parent material classification, soil properties, spatial prediction, digital soil mapping, Lake Balaton

1. INTRODUCTION

Over the last few decades advances in computing have created new opportunities for soil mapping, enabling the creation of more detailed soil maps at lower costs and in less time. The idea of digital soil mapping (DSM) is that existing soil information can be combined with auxiliary data using geoinformation and mathematical methods. By analyzing the relationship between soil and soil forming factors, outdated and less detailed soil maps can be improved. Furthermore, with the help of prediction models information may be obtained on areas where soil data are limited or soil surveys have never been conducted. In the "scorpan" framework (MCBRATNEY et al. 2003) seven important factors are defined and recommended for use in soil modelling and mapping:

$$S=f(s,c,o,r,p,a,n)$$

where: s=soil, c=climate, o=organism, r=relief, p=parent material, a=age and n=spatial position.

Auxiliary data are generally derived from thematic maps (relief, land cover, geology, and climate). These provide a wealth of data on environmental conditions, but in some cases it is difficult to extract useful information for DSM. This study focuses on geology, which is examined in detail less frequently as an auxiliary data layer, despite its importance in soil formation. JENNY (1941) defined parent material as the initial state of the soil system which is the substratum of soil formation, and the physical and chemical properties of the underlying rocks are important factors in the process. This geological background as a passive factor affects the formation of the soil profile (STEFANOVITS et al., 2005).

Generally, the most readily available source of parent material information can be found in surface geology maps. However, on the basis of comprehensive studies, geological maps are only used in less than a quarter of DSM projects (GRAY et al., 2016). This degree of neglect may generally be ascribed to the difficulty of interpretation of geological maps from a pedological point of view. Geology is also commonly omitted from the analyses on account of its lack of explanatory power in soil modelling (BEGUIN et al., 2017), though this may also be the result of the inappropriate use of geology as an environmental covariate; special expert knowledge is often required in order to convert the information contained in geological maps into meaningful soil covariates (BEHRENS et al., 2005).

The main problem is that the classification of the units on geology maps is generally based on the age and the genesis of rocks, while this information is irrelevant in DSM. In addition, usually too many units are distinguished on a geology map, making it difficult to explore the relationship between the geology and the derived soils. To alleviate this problem, geological units are usually reclassified based on rock type. In the FAO system there is a detailed classification scheme which allows the easy conversion of the units used in geological maps into parent material (PM) categories (FAO, 2006). However, this four-level classification places too great an emhasis on the genesis of rocks. Besides this, several important rock types are missing, and it is thus not well suited to the creation of PM categories.

Other PM classification schemes have also been developed which are more suitable for the purposes of DSM. SOTER is a harmonization and a mapping method combining landform classes and PM categories to define terrain units, which are then in turn represented by typical soils. Based on the results of the e-SOTER project a new PM classification has been developed which is the revision of the previous FAO scheme but it has a pedologically relevant structure (VAN ENGELEN & DIJKSHOORN, 2013). The basis of this five-level classification system is the geochemical character of rocks, and the classes are clearly distinguished, being reflected in the various derivative soils.

Undoubtedly the most important feature of any parent material is its geochemical character. The basic concept of this approach is that the SiO_2 content is one of the most important features in rock classification, and especially in igneous rocks, which are classified into ultrabasic, basic, intermediate and acid groups based on the percentage of SiO_2 in their geochemical composition. GRAY et al. (2016) developed a reclassification method for all the rock types, including sedimentary and metamorphic rocks, based on their SiO_2 content, and this method has proven to be very effective in modelling and mapping soil properties. They demonstrated that certain important soil properties (soil organic carbon, pH, cation exchange capacity, sum of bases, total P and clay %) stand in clear relationship to SiO_2 content. This relationship allows the use of geological information in a semi-quantitative form through the application of a 'silica index'which can be determined based on the median of the measured SiO_2 values in the new geological categories. The index works in the following way: in a parent material, high SiO_2 content generally means higher quartz content; this in turn has an inverse relationship with the clay and base cation content of the derivative soils. As soil texture is determined by the mineral fractions present in the soil (relative proportions of sand, silt and clay), it is obvious that the SiO_2 content of the parent material also plays a key role in soil physical properties.

Despite the effectiveness of GRAY's method, it does have several shortcomings. Any individual geochemical PM classification requires a great quantity of X-ray fluorescence (XRF) test data, in most cases not available locally, while the use of geochemical data measured elsewhere could distort the results. A typical case is that of loess, in which significant differences occur in SiO_2 content in different regions of the world (ÚJVÁRI et al. 2008). Regional averages of the SiO_2 content of these sedimentary rocks range from 60% to 80%; they cannot, therefore, be assigned to a single category, due to the high degree of diversity in their geochemical composition. Sedimentary rocks are generally problematic in this regard, and as a result, a unified geochemical PM classification scheme cannot be developed, so this approach cannot be standardized. Nevertheless, in cases where locally measured geochemical data are available, this method seems to be the most suitable for pedological purposes.

Despite the fact that DSM methods have already been widely used and recognition of the importance of geology is on the increase, only limited information is available concerning the proper use of geology as an auxiliary data source, and also about its role in the prediction of soil properties. One task is to resolve the contradictory nature of our understanding of geology as an environmental covariate: in some studies it is considered the strongest (GRAY et al. 2016) or one of the strongest (VAYSSE & LAGACHERIE, 2016, XIONG et al. 2014, WIESMEIER, 2011) predictors in DSM, while in most cases it cannot be used effectively.

Since a surface geology map is available for Hungary (GYALOG & SÍKHEGYI, 2005), and it is detailed enough for DSM purposes, the decision was made to examine the role of geology in the spatial distribution of soil properties. In this study the aim was to compile a parent material map on the basis of the surface geology map using GRAY's geochemical approach, then to use it in our DSM project. Models based on geology and other environmental covariates were developed for three selected soil properties - liquid limit, pH and soil organic carbon - which are important features of the soils and are also important factors in soil classification. Based on the developed models, predictions were made for our whole study area. One important purpose of this study is to evaluate the role of geology in soil modelling and mapping, but a further aim was to compile fine resolution maps of the three selected soil properties for the watershed area of Lake Balaton in order to serve the increasing need for more detailed soil data for environmental impact assesment in this area.

2. MATHERIALS AND METHODS

2.1. Study area

The watershed area of Lake Balaton (the largest freshwater lake in central Europe) is located between longitude 16°17'43''E and 18°11'26''E and latitude 47°6'47''N and 46°17'32''N (Fig. 1). Fine scale soil maps of the whole watershed have not yet been created, although detailed legacy soil information is available. For environmental reasons (especially for water quality protection) it is particularly important to create detailed soil maps, as a large part of the watershed lies on cultivated land. Detailed soil maps could support the more accurate assessment of diffuse phosphorus loads, and also aid in the targeting of erosion and runoff control measures.

From a geological point of view, the watershed can be divided into two parts. The Balaton Highland is located to the north of the lake, where older (Palaeozoic-Mesozoic), mainly carbonate rocks dominate, but basalt, sandstone, and methamorhic rocks are also present in notable quantities.

This elevated area together with the Keszthely Hills at the southwestern end of Lake Balaton represents the southern part of the Transdanubian Central Range. The other part of the watershed is located south and west of the lake (Somogy and Zala river subwatershed), lies at a lower altitude, and is mainly covered by younger (Neogene and Quaternary) alluvial and aeolian sediments.

From a pedological point of view the watershed also shows a great diversity. The following World Reference Base (WRB) units can be distinguished in the area: Rendzic Leptosols dominate the upland area, with Haplic Luvisols on the foothills, in the valleys and on the deeply weathered basalt plateaus. In the lower areas, the dominant soils are Haplic Luvisols, coloured with Eutric Cambisols in drier regions, Stagnic Luvisols in the western areas with high precipitation, Lamellic and Dystric Arenosols in

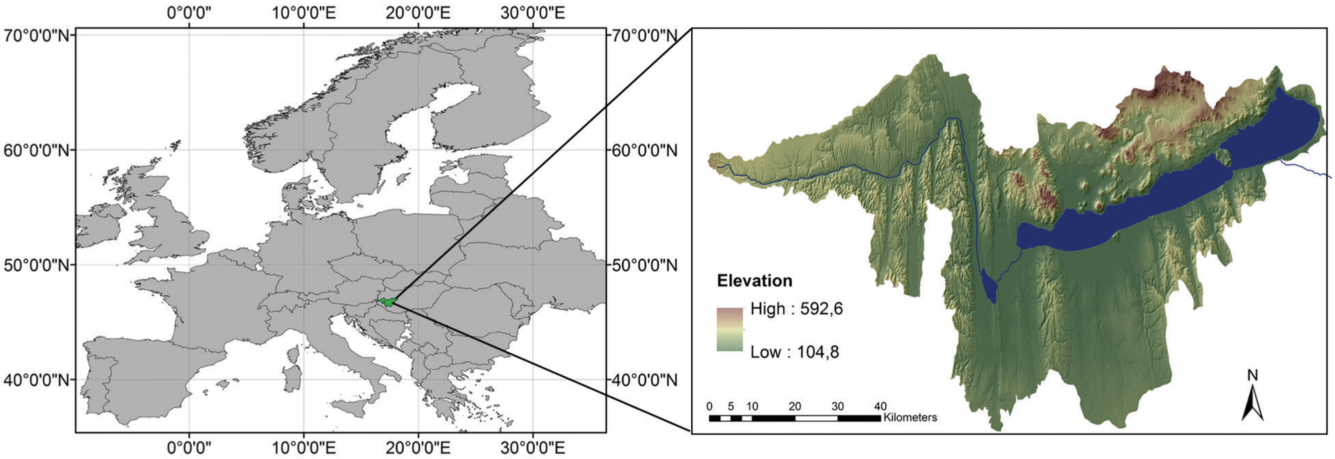

Figure 1. Location of the watershed of Lake Balaton in Europe (left) and the elevation of the area (right) based on the 90 m resolution digital elevation model (NASA, SRTM).

the southern part, Sapric Histosols near to the lake in former bays, and in general, Gleysols and Fluvisols in the valleys. Soil erosion by water is prevalent in certain regions with high relief.

The climate of this area is a typical European continental climate, with warm summers and fairly cold winters. The annual precipitation is about 400-600 mm.

2.2. The 1:100 000 scale surface geology map and the derived parent material map

The Hungarian Geology Institute published a 1:100 000 scale digital surface geology map, the result of more than ten years of harmonisation and updating, in 2005 (GYALOG&SÍKHEGYI, 2005). In lowland areas, 1:100 000 scale geological mapping had been completed in the course of the preceding 10-30 years. However, in hilly areas, finer scale legacy maps (e.g. 1: 10 000, 1:25 000 and 1:50 000) formed the basis of this new map. This is the reason why it appears far more detailed and accurate than soil maps with the same nominal scale (e.g. AGROTOPO, 1:100 000, 1994), especially in those hilly areas for which limited soil information is available.

The 1:100 000 scale surface geology map shows the presence of 143 geological units in the watershed of Lake Balaton: 3 Palaeozoic, 42 Mesozoic, 3 Palaeogene, 24 Neogene, and 71 Quaternary formations with more than 20 rock types. We classified these units into broad categories based on their approximate SiO_2 content. The basis of this new classification was the scheme developed by GRAY et al. (2016), modified to take into account the fact that different rock types dominate in the watershed than in the Australian study area.

The first step was the selection of some characteristic rocks and sediment types common in our study area. These were the following: aeolian sand, aeolian loess, fluvial sediments of varying grain size, sandstone, limestone, dolomite, marl, basalt, and sesquioxides. The average SiO_2 content of these rocks was determined, and the framework of a broad geochemical classification constructed. An attempt was made to place each geological formation into one or other of the new geological categories based on SiO_2 content. In several cases, legacy X-ray fluorescence (XRF) data were available (BUDAI&CSILLAG, 1999; ÚJVÁRI et al. 2008; BOHN, 1979; JUHÁSZ, 1962; GYALOG&HORVÁTH, 2004; EMBEY-ISZTIN et al. 1993), but data measured in our study area cannot be then simply assigned to all rock types. In such cases, research was conducted for legacy data of similar rocks from all over the world. To estimate the approximate SiO_2 content of the various rocks the free GEOROC database (SARBAS&NOHL, 2008) was employed, with additional legacy XRF data (PETTIJOHN 1963; CLARKE 1920; GU et al., 2013).

Our final classification structure contains nine groups, based on approximate SiO_2 content (Table 1).

Each original geological formation was assigned to one of these groups. In some cases this classification was difficult, especially where a combination of rock types or sediments were present. On the basis of the method employed by GRAY et al. (2016), in these cases an overall single class was determined; though this was actually an average of several classes. Sesquioxides and carbonate rocks were not excluded from our SiO_2 content based classification scheme, even though these classes have a low SiO_2 content.

A property other than SiO_2 content had to be taken into consideration in the last group, because organic material has a strong effect on derivative soils. For this reason, a separate group was created for those sediments with the highest organic matter content.

The parent material map compiled from these nine groups can be seen in Fig. 2.

2.3. Other environmental covariates

The basis of DSM is the empirical quantitative description of soil formation as a function of climate, organism, relief (topography), parent material, age and spatial position. In order to build predictive soil models, detailed information concerning these soil

Table 1. Broad geochemical classification of rocks and sediments.

Geological classes	Approximate SiO_2 %	Rock types
1	80–90%	aeolian sand, quartzite
2	70–80%	alluvial sand, gravel, proluvial and deluvial sand, sandstone, loessy sand
3	65–70%	loess, sandy loess
4	60–65%	aleurite (alluvial, limnic, slope, deluvial)
5	50–60%	clay (alluvial, limnic, slope, deluvial)
6	45–50%	basalt
7	20–30%	bauxite, bauxitic clay, clayey bauxite,
8	1–10%	limestone, dolomite, marl
9	–	organic (peat)

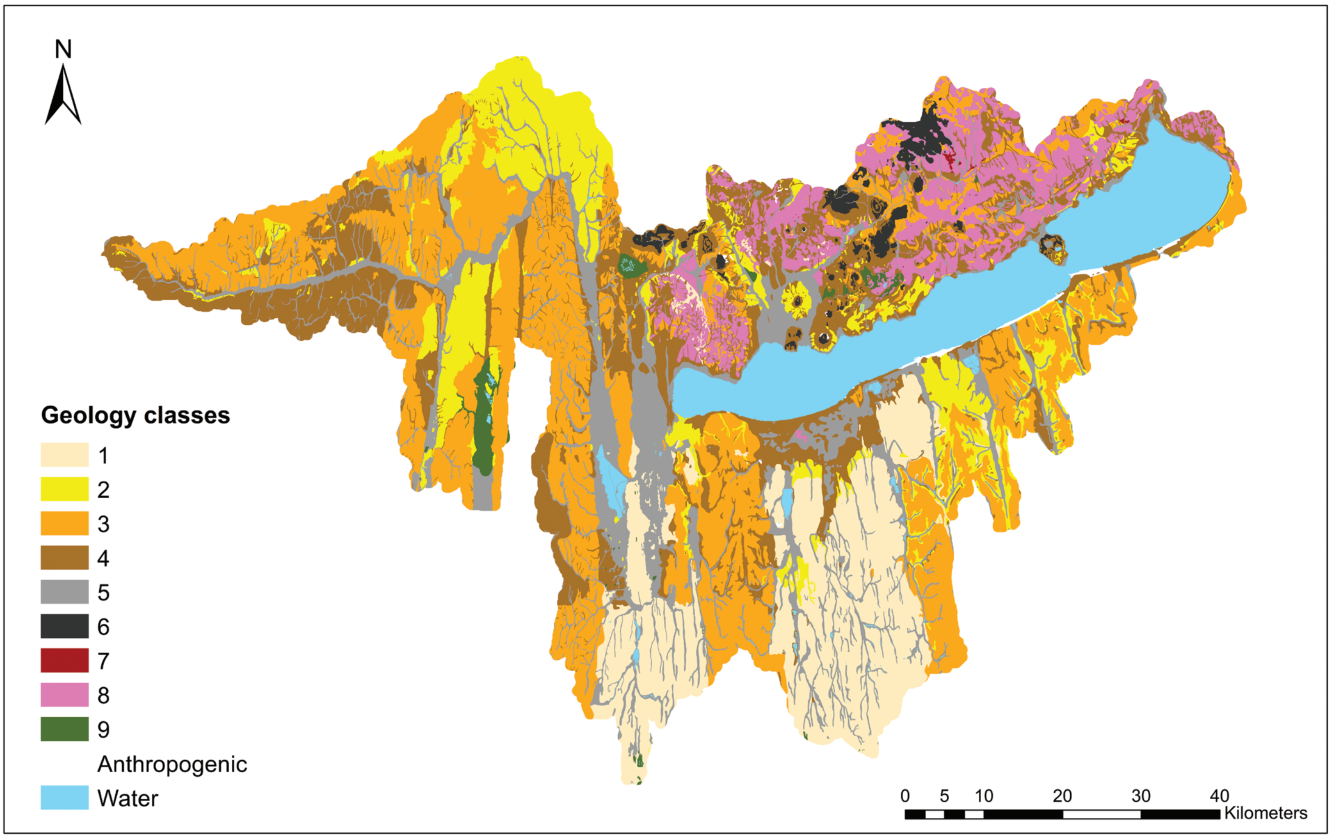

Figure 2. Parent material map of the watershed area of Lake Balaton. The categories (1-8) are in descending order based on the approximate SiO_2 content (see Table 1).

forming factors is required. Besides geology, a different set of climate, topography, vegetation and land cover covariates was used. These are summarized in Table 2. The geoinformatical processing of the environmental covariates was carried out in ArcGIS 10.0.0.

2.3.1. Topography:

a digital elevation model (SRTM DEM) at a resolution of 90 m was used in the analysis (JARVIS et al. 2008). Slope, aspect, and curvature were derived from the DEM, and the topographic position index (TPI) was also calculated for each cell following the method developed by WEISS (2001) and JENNESS (2016). A circle of 2000 m radius was used to determine the topographic position, which is an index for the quantitative description of geomorphology. The five topographic factors (elevation, slope, curvature, aspect, TPI) were used as quantitative variables in the description of relief.

2.3.2. Climate:

averaged annual precipitation and temperature data selected from the free WorldClim bioclimatic variables (FICK&HIJMANS, 2017) were used in the analysis. These rasters have a resolution of approximately 750 m.

2.3.3. Land cover:

CORINE land cover categories produced by the European Environment Agency (2013) based on satellite images from the land surface were used. In our study area, 22 of the original 44 CORINE categories appear. This land cover database provides concurrent information on both land use and vegetation.

2.3.4. Vegetation:

The Normalized Difference Vegetation Index (NDVI), derived from MODIS (NASA) satellite data at a resolution of 250 m (for the month of June 2016) was used in the analysis (MOD13Q1).

Table 2. Description of environmental covariates.

predictor variable	source	resolution	type	range
elevation	NASA SRTM DEM	90 m	q	104.8–592.6 m
slope	NASA SRTM DEM	90 m	q	0–61.8°
curvature	NASA SRTM DEM	90 m	q	–8.8–5.4
aspect	NASA SRTM DEM	90 m	q	0–360°
TPI	NASA SRTM DEM	90 m	q	–5.3–13.6
annual precipitation	WorldCLIM database	750 m	q	544–776 mm
annual temperature	WorldCLIM database	750 m	q	8.6–10.9 °C
land cover	European Environment Agency, CORINE	1:100 000	c	1–22
geology	Geological and Geophysical Insitute of Hungary	1:100 000	c	1–9
NDVI	NASA MODIS satellite 13Q1, 09-JUN-2016	250 m	q	1564–9296

q: quantitative, c: categorical

Table 3. Summary of the NPCPD soil data.

soil property	K_A	SOC	$pH_{(KCl)}$	$CaCO_3$	Mg	Zn	Cu	Mn
dimension	cm^3	%	–	%	ppm	ppm	ppm	ppm
range	24–81	0.30–3.19	3.52–7.76	0–20	10–440	0,4–11.8	0–13.2	2.1–132

2.4. Soil dataset

Soil data (12400 points) from the National Pedological and Crop Production Database (NPCPD) collected between 1985-1989 were used in the analysis. The nationwide soil property map series for agricultural areas in Hungary had also been previously created based mainly on this database (TÓTH et al. 2015). It contains liquid limit (K_A) according to ARANY (1943), humus, pH, $CaCO_3$, Mg, N, P, K, Cu, Zn, Mn values measured in the upper 25 centimetres of the soil. The detailed description of the database including the methodology of the soil tests were presented previously by KOCSIS et al. (2014). Here only the three mapped soil properties are discussed in detail. Organic carbon content was measured after wet combustion with kalium-bichromate sulfuric acid solution by the titration method and then humus content was calculated. Before the modelling and mapping processes were carried out, humus was converted back to soil organic carbon content (SOC) by dividing by 1.724 because SOC is more stable than humus and more comparative data are available about this parameter. The pH was measured in either distilled water or 1 M KCl solution in 1:2.5 soil to solution ratio supension after thorough wetting and two hours shaking. The 'K_A' stands for liquid limit according to ARANY (1943). It refers to the quantity of water (cm³) necessary for the creation of a fully water saturated "aggregate free" soil paste from a 100 g oven-dry soil sample to reach the "endpoint indicator", the "sticky point" (VÁRALLYAY, 2013). K_A was simplified from the traditional liquid limit determination (CASAGRANDE, 1932) for agricultural purposes. Liquid limit can be considered a good proxy for a particle size distribution test. It ranges from sandy (K_A=~25-30) to clayey soils (K_A>50). The major advantage of this measure is its simplicity and time-efficiency.

Some elements of the NPCPD database (plant available P, K, and N) were ignored in our analysis, as they may have been strongly influenced by temporarily and temporally variable effects such as those from fertilization. The soil information derived from the NPCPD database is summarized in Table 3.

Since the NPCPD represents mainly cultivated land, additional data were required, and especially those concerning forested areas. Fifty forest points were selected from the land evaluation database maintained by the land registery offices in Hungary. They record the exact geographical positions and soil profile descriptions of reference sites, which are then used to judge the relative primary production value of all plots of productive land.

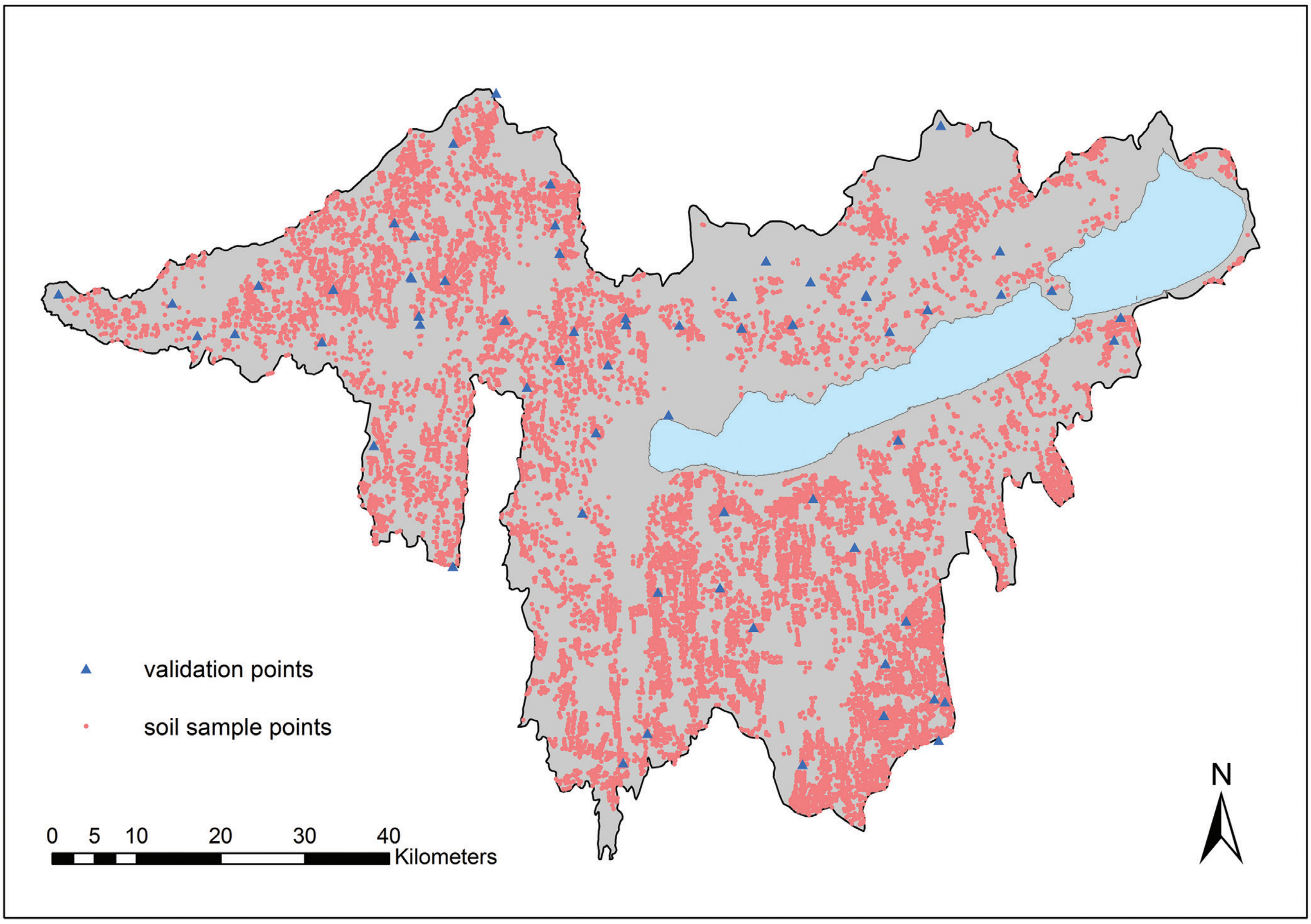

Figure 3. Spatial distribution of the soil sample points and the validation points.

Table 4. Correlation matrix of the numerical predictor variable.

	aspect	curvature	precipitation	elevation	slope	temperature	TPI	NDVI
aspect	1	0	0.05	0.07	0.09	–0.03	0.02	–0.04
curvature		1	0.04	0.08	004	–0.03	0.16	0.01
precipitation			1	0.29	0.05	–0.54	0.19	–0.05
elevation				1	0.3	–0.67	0.38	–0.07
slope					1	–0.23	0.09	0.06
temperature						1	–0.12	0.08
TPI							1	–0.07
NDVI								1

The final soil property maps were validated based on the topsoil data from the Soil Conservation Information and Monitoring System (SIMS), which is independent of the NPCPD, similar methods were, however, used in measurement of the soil properties (VÁRALLYAY et al., 2010). SIMS consists of 1,236 soil profile descriptions nationwide with highly detailed physical and chemical test data. Sixty eight points fall within the watershed, and from these the humus, pH and liquid limit data were used (Fig. 3).

2.5 Spatial prediction of soil properties using the Random Forest algorithm

Firstly a correlation matrix was calculated in order to identify the highly correlated variables and avoid the distortion effect of redundant features in our analysis.

For the modelling of major soil properties (liquid limit, pH, soil organic carbon) the Random Forest algorithm was used. The use of this method, developed by BREIMAN (2001), is very widespread in environment modelling because it has lots of benefits. It can also be used for classification and regression tasks, and there are no restrictions on the types of variables. The Random Forest model consists of a number of decision trees from which a predefined number of randomly selected independent variables (different subsamples of the data) are used. The final result is derived from the mean predictions made on the basis of individual trees. In this study the soil properties are dependent variables and the environment covariates are independent (predictor) variables. For each soil property, an individual model was constructed using the predictors. In the analysis, 80% of observations were randomly selected and used for RF modeling and the other 20% for model validation.

For each model, the optimum number of randomly selected independent variables was determined. The Recursive Feature Elimination (RFE) method was applied in feature selection in order to identify the most relevant predictors; in the final models only these variables were used. RFE algorithm ranks the variables by its importance and automatically selects a subset of the most predictive features (KUHN, 2008).

The coefficient of determination (R^2), defined by the percentage of variation explained by the model, was employed to evaluate prediction performance, and RMSE (root mean square error) to measure the accuracy of the models.

For the analyses the ‘Random Forest’ (LIAW&WIENER, 2002) and the ‘caret’ (KUHN, 2008) package were used in the R environment (R CORE TEAM, 2017).

2.6. Evaluation of the role of geology in the spatial distribution of soil properties

The effect of geology was analyzed independently of other environment covariates, and its importance in the soil property models was also evaluated.

Since accurate geochemical data are not available for all geological formations, it was decided not to use the ‘silica index’ (GRAY et al., 2016), but instead use broad categories based on the approximate SiO_2 content of the various rocks. For this reason, analyses of variance (ANOVA) - in which the nine geological classes are factors and accurate values for SiO_2 content are not required - were carried out in order to examine the effect of geology on soil properties.

The other approach to the evaluation of the role of geology was the measurement of variable importance in the soil property models in which a comparison with other environment covariates was also possible.

3. RESULTS

3.1. Correlation analysis of the predictor varaiables

The correlation matrix (Table 4.) of the numerical predictor variables showed that in the dataset used here the variables are not highly correlated (maximum absolute correlation is 0.67, but values are generally lower than 0.4).

3.2. Model performance and accuracy

The results show that in the RF models the environmental predictor variables explain 69.65%, 57.78% and 65.31% of the variation of the liquid limit, SOC, and pH, respectively if all independent variables are used. In order to construct models with a higher degree of accuracy and better performance, the RFE method was employed. After removing non-relevant variables based on the results of RFE, the performance of these models increased to 72.4%, 60.3% and 68.4%. Aspect, curvature and NDVI were ignored in all three models, as was slope in the case of liquid limit and pH. The final models were constructed from 500 regression trees using 4 randomly selected variables from the relevant predictors.

The observed RMSE values of the validation samples were 5.29 for the liquid limit, 0.27 for SOC and 0.54 for pH in the final models.

3.3. The effect of geology on soil properties and its importance in RF models

One of the most important goals of the analysis was to evaluate the role of geology in the spatial distribuiton of soil properties. Firstly, the effect of geology was analyzed independently of other environment covariates using the ANOVA method. Significant differences were found in geological classes in the case of all the soil properties of the NPCPD. Tukey post hoc tests show which groups may genuinely be distinguished from each other (Table 5.) Although not all classes differ significnatly, the means show an interesting trend, namely, that almost all soil properties in the NPCPD show an inverse relationship with the approximate SiO_2

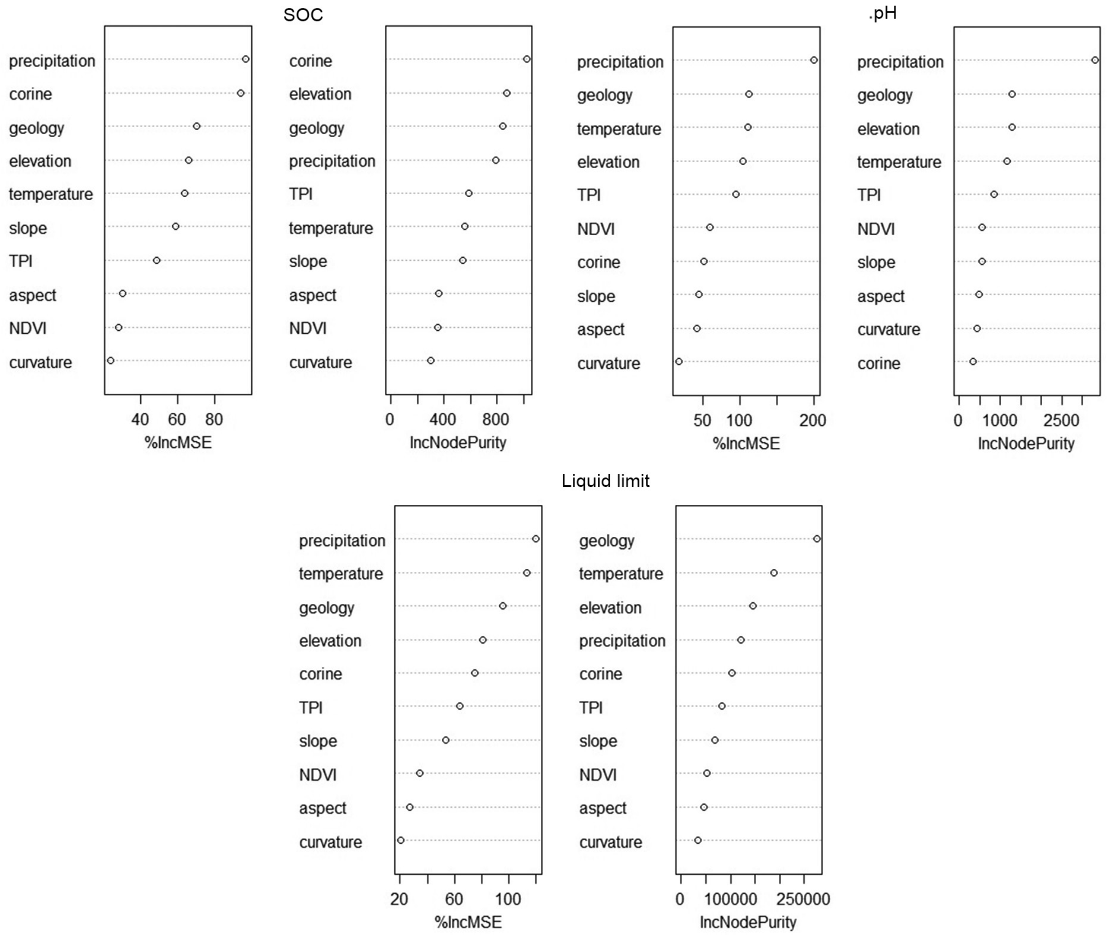

Figure 4. Variable importance plots of the RF models based on the precentage increase of MSE and node purity.

content. Consequently, this PM classification provides relevant information for pedological purposes, and it may be expected to be well suited for use in the modelling and mapping of soil properties.

The actual importance of geology in the major soil property models was evaluated based on the results of the RFE and variable importance plots of the RF models. The RFE method showed that geology is a relevant predictor in the case of all three major soil properties. Variable importance plots ranked the independent variables based on the increase of MSE and node purity in the RF models after removing the variables (PRASAD et al., 2006). This method enables the examination of the importance of geology in comparison to the other variables. The results of these may be seen in Fig. 4. The best performing predictors can be seen on the top of the plots. Judging by the importance of the variables, geology is among the top predictors, but unlike its position in GRAY's results, it is not the most important variable in any case.

3.4. Major soil property maps compiled using the RF models

In order to compile soil property maps rasters were created at a resolution of 90 m, and predictions for each cell were made using the developed RF models. The 'raster' R package (HIJMANS, 2016) was used in the creation of the final maps.

The results of these are three soil property maps for the watershed of Lake Balaton (Fig. 5), which are at present the most detailed soil maps of the area.

Table 5. The results of the Tukey post hoc tests.

Geology classes	Liquid limit (cm^3)	SOC (%)	pH (KCl)	$CaCO_3$ (%)	Mg (ppm)	Zn (ppm)	Cu (ppm)	Mn (ppm)
1	28.31^{a}	0.75^{a}	5.14^{a}	0.20^{a}	91.95^{a}	1.49^{a}	1.95^{a}	$103.36^{b,c}$
2	34.89^{b}	$0.89^{a,b}$	5.81^{b}	$0.73^{a,b}$	134.20^{b}	$1.96^{a,b}$	$2.71^{a,b}$	$89.12^{a,b,c}$
3	$38.60^{b,c}$	$0.92^{a,b}$	5.81^{b}	$0.53^{a,b}$	176.44^{c}	1.72^{a}	$2.77^{a,b}$	$97.84^{a,b,c}$
4	$42.12^{c,d}$	1.08^{b}	$6.15^{b,c}$	$1.24^{b,c,d}$	197.18^{c}	$2.07^{a,b}$	$2.87^{b,c}$	$77.88^{a,b}$
5	$44.03^{d,e}$	1.35^{c}	$6.16^{b,c}$	$1.66^{b,c,d}$	199.95^{c}	$2.43^{b,c}$	$2.86^{b,c}$	73.38^{a}
6	$44.40^{d,e}$	$1,10^{b}$	$6.17^{b,c}$	$2.13^{d,e}$	245.33^{d}	$2.51^{b,c}$	$3.27^{b,c}$	$98.28^{a,b,c}$
7	$45.80^{d,e}$	$1,42^{c}$	$6.55^{c,d}$	$1.92^{c,d,e}$	269.51^{d}	2.64^{c}	3.71^{c}	108.06^{c}
8	47.81^{e}	1.41^{c}	6.69^{d}	3.01^{e}	260.78^{d}	$2.53^{b,c}$	3.71^{c}	$82.09^{a,b}$
9 (organic)	46.71^{e}	1.38^{c}	$6.48^{c,d}$	$1.69^{c,d}$	185.71^{c}	$2.09^{b,c}$	$3.27^{b,c}$	$80.76^{a,b}$

a,b,c,d,e: the means with the same letter are not significantly different (p-value < 0.05).

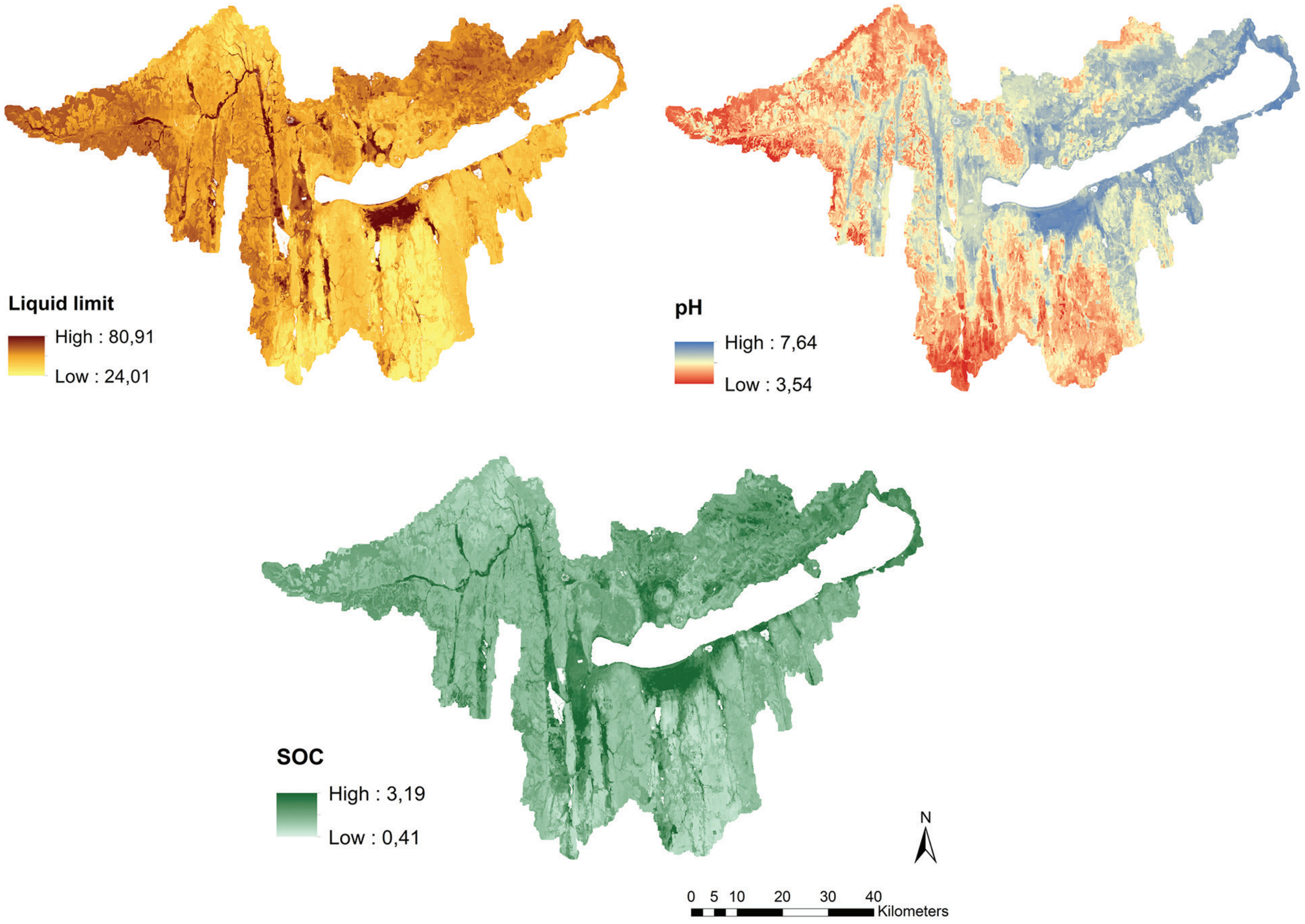

Figure 5. Spatial prediction of major soil properties.

3.5. Validation of the maps

For the validation of the maps, data of 64 soil sample points from the Soil Conservation Information and Monitoring System (SIMS) were used.

Lin's concordance was employed to compare predicted and observed values. The concordance correlation coefficients (CCC) were 0.55, 0.59 and 0.50 for liquid limit, pH and SOC, respectively. But if points were selected only from cultivated areas, the CCC values increased to 0.73, 0.73 and 0.69 (Fig. 6.). The main reason seems to be that the dataset forming the base of our soil property models contains data primarily from cultivated land, and thus it is in such locations the prediction models perform better.

4. DISCUSSION

In the present study, digital maps of three major soil properties, liquid limit, pH and soil organic carbon (SOC) were created for the watershed area of Lake Balaton based on legacy soil information using the Random Forest algorithm. While different sets of environmental covariates were employed in the prediction, particular attention was paid to the role of geology. A parent material map was prepared through the application of a geochemical approach, with the reclassification of the original map units from the 1:100 000 surface geology map. It was confirmed that SiO_2 content based PM categories are capable of providing relevant information for soil prediction. Nonetheless, a great deal of XRF data are required in order to arrive at the correct classification, and it is precisely this which is the main drawback to the method proposed by GRAY et al. (2016). In the case of the classification constructed here, this type of inaccuracy may also distort the final results. This, in turn, may well be one of the reasons why in the derived RF models geology is not the strongest predictor, in contrast to the conclusions arrived at by GRAY et al. (2016). The other possible explanation for this is the scale dependency of the importance of geology, i.e. the study area in the present paper is not large enough, and is less diverse from a geological point of view. Another factor is that the spatial distribution of the soil sample points can weaken the result, as the soil database used covers mainly cultivated land where the geology is less diverse. Although the database was broadened with additional points on forest land, this was not sufficient to eliminate the imbalance in distribution of the sample points; it is possible that this also reduced the importance of geology in the models developed.

Scale dependency and the imbalance in the spatial distribution of the sampling points may well have led to a result in which geology is not the most important predictor, but this assumption cannot be verified due to a lack of comparative data. Although a similar geochemical PM classification scheme has yet to be used in other studies, it is remarkable that in general, geology is not the most important predictor in DSM projects. Unfortunately, geology is not often examined in detail as a predictor variable in DSM, and thus only a few data are available with which it would be possible to compare the results arrived at here. Usually, different variables dominate in each model, depending on the target variables. This is a realistic consequence, since soil formation is the function of several environmental factors.

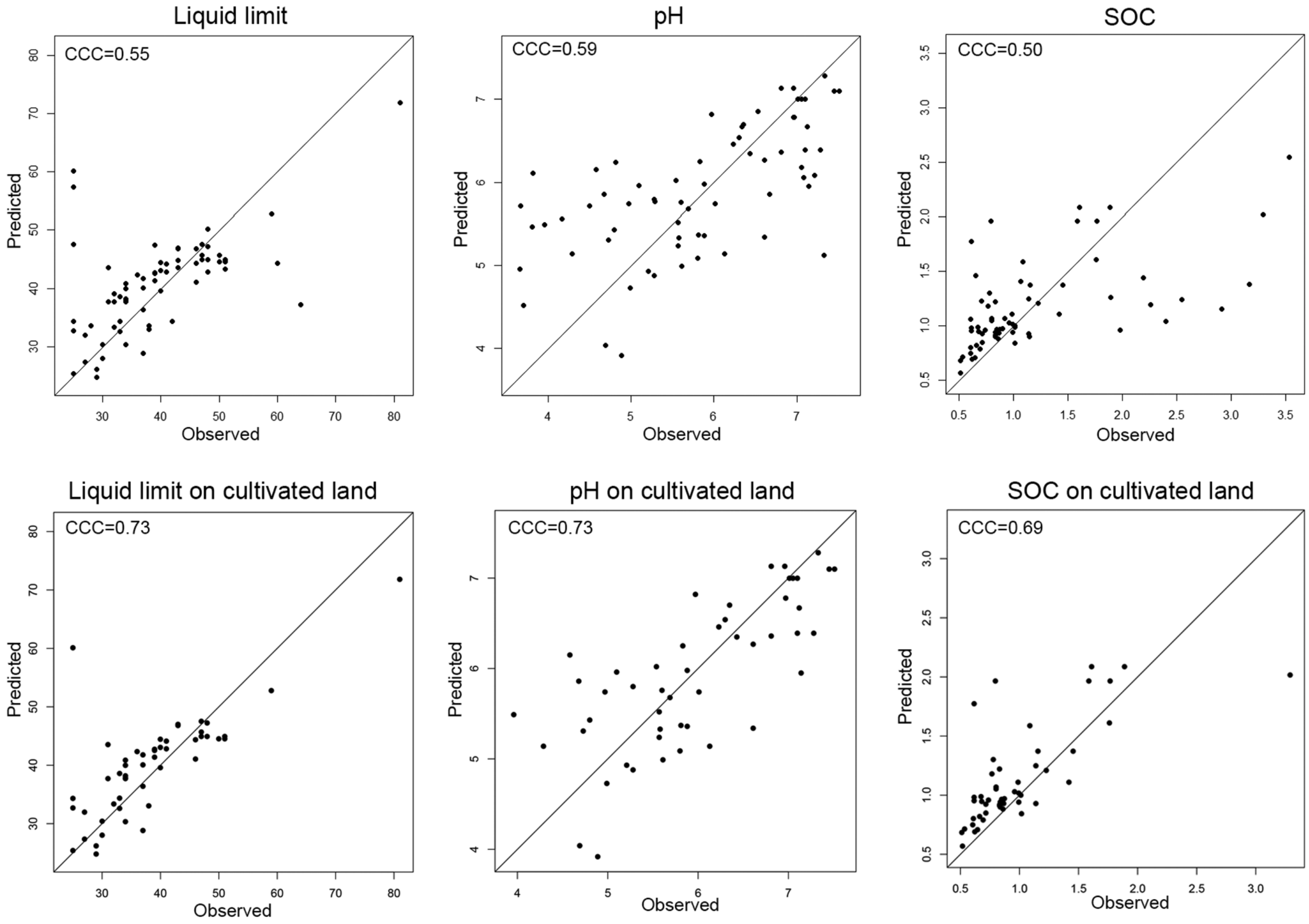

Figure 6. Concordance correlation plots for liquid limit (cm^3), pH and SOC (%).

According to VAYSSE & LAGACHERIE (2014), the prediction of SOC is driven primarily by climate co-variables. ARAUJO et al. (2017) also demonstrated in a recent study that the effect of climate on SOC can be stronger than the parent material driven textural and mineralogical control of SOC retention because the soil organic matter decomposition rate is strongly dependent on climatic conditions (e.g. in a relatively cool environment the stabilization of soil organic matter is more efficient). Besides climatic variables, remote sensing vegetation indices (NDVI, EVI) and land use information may also have greater predictive value in SOC models (DHARUMARAJAN et al., 2017). Only SILVA et al. (2016) found that parent material plays a decisive role in determining soil characteristics and that geology has a stronger effect on SOC than climate under humid-semi-arid conditions. Although WIESMEIER at al. (2011) also reported a high value for the predictive role of geology in modelling soil organic matter, it should be mentioned that they do not use any climatic variables, and that even then, land use was a stronger predictor than geology in their model. The results of the present study confirm that both geology and climate, and especially annual precipitation, have a strong effect on SOC, but land use and vegetation data (land cover) are also important factors.

In the case of pH, we arrived at similar conclusions. Some studies have reported that geology has the strongest influence on the pH of the derived soils (e.g. VAYSSE & LAGACHERIE 2014), but climatic factors are also important, and elevation is often a key factor (DHARUMARAJAN et al., 2017), similar to the pH model in this paper.

Liquid limit comparative data are not available for the evaluation of the role of geology in soil because this physical property is in general used only in Hungary. Since liquid limit is related to soil texture, it is obvious that this property also stands in close relation to climate and geology, as it is essentially these two factors which determine weathering processes. Similar results were reported by RODRÍGUEZ-LADO & LADO (2017) about the scaling properties of particle size distributions where parent material and climatic indices are the most influential variables. In our liquid limit model, climate is a stronger factor than geology, as was also the case in the pH and SOC models.

Summarizing the results, the importance of geology in soil prediction has been definitely established, and SiO_2 content based PM classification may be considered suitable for DSM purposes, despite the fact that geology turned out not to be the most important variable. Nevetheless, it is also important to note that, although our study area was relatively small in geographical extent, climatic variables (especially annual precipitation) still have the strongest influence on soil properties. The explanation for this may be the young age of the sediments covering the greater part of the watershed, as these are less weathered and landscape formation is still in its initial phase, with the presence of intensive weathering processes.

The relatively small effect in the studied area of certain topographical variables (aspect, curvature, slope) on soil properties is also interesting, although other authors have reported similar results (CAMERA et al. 2017, DHARUMARAJAN et al. 2017, WIESMEIER et al. 2011).

The most important results of the present study are the three final soil property maps, which are the most detailed soil maps for the watershed area of Lake Balaton currently in existence and is to be hoped that they can be used in future assessments.

Finally, an important outcome of the validation of the maps is that the spatial distribution of the existing soil sample points can and does result in great differences in the degree of uncertainty of the predicted maps in different regions. This is one of the main disadvantages of the use of legacy data, and it indicates the need for new soil surveys in the area. The validation results showed that our predictions were more accurate on cultivated land, from which most of the soil data originated. This is in line with previous work (TÓTH et al., 2015), in which this database was used for nationwide soil modelling and mapping and the predictions proved to be reliable only on cultivated land. However, based on these results the areas from which additional soil samples should be taken can easily be delineated and this, in turn, would improve any future maps.

5. CONCLUSION

The method of digital soil mapping developed in the last few decades is one of the most important innovations in soil science. However, the effective use of DSM requires lots of information about the role of soil forming factors including parent material.

In this study a geochemical approach developed by GRAY et al (2016) was applied in order to convert geology to a usable soil covariate. A legacy database (NPCPD) with 12400 soil test data and the 1:100 000 scaled surface geology map of Hungary facilitated the performance of a comprehensive analysis of the relationship between geology and soils and evaluation the role of geology in soil modelling and mapping. Well performing predictive models could be built using geology and further environmental covariates (climate, topography, land cover) for three major soil properties (liquid limit, soil organic carbon, pH) in our study area. Based on the results, it has been confirmed that the geochemical classification of the rocks is useful for DSM purposes, but accurate geochemical data would be essential to improve this method. Nevertheless it should be mentioned that the potential use of more accurate geochemical data does not necessarily result in determining that geology acts as the most influental variable in soil prediction because additional factors are also involved including scale dependency, other soil forming factors, spatial distribution of the soil data. In our study area it seems that climate (especially annual precipitation) has a greater impact on soil properties than geology.

By using the developed models, predictions were made which resulted in the most detailed soil maps available at present for the watershed of Lake Balaton. However, map validation showed that the uncertainty is much higher in the areas (forests and hilly areas) where limited soil data is available but a new soil survey in these areas could improve both the models and the predictions.

ACKNOWLEDGEMENT

We acknowledge the financial support of Széchenyi 2020 under the EFOP-3.6.1-16-2016-00015.

REFERENCES

AGROTOPO (1994): AGROTOPO database. Spatial Soil Information System. RISSAC HAS. Budapest. https://maps.rissac.hu:3344/webappbuilder/apps/2/

ARANY, S. (1943) A talajok rögszerkezete [*The aggregate structure of soils-in* Hungarian].– Mezőgazdasági Kutatások, XVI., 220–249.

ARAUJO, M. A., ZINN, Y. L. & LAL, R. (2017): Soil parent material, texture and oxide contents have little effect on soil organic carbon retention in tropical highlands.– Geoderma, 300, 1–10. doi: 10.1016/j.geoderma.2017.04.006

BARANYAI F., FEKETE, A. & KOVÁCS, I. (1987): A magyarországi talajtápanyag-vizsgálatok eredményei [*Results of the soil nutrient assesment in Hungary* – in Hungarian].– Mezőgazdasági Kiadó, Budapest, 189 p.

BEGUIN, J., FUGLSTAD, G.A., MANSUY, N. & PARÉ, D. (2017): Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches.– Geoderma, 306, 195–205. doi: 10.1016/j.geoderma.2017.06.016

BEHRENS, T., FÖRSTER, H., SCHOLTEN, T., STEINRÜCKEN, U., SPIES, E.D. & GOLDSCHMITT, M. (2005): Digital soil mapping using artificial neural networks. – Journal of Plant Nutrition and Soil Science, 168, 1–13. doi: 10.1002/jpln.200421414

BUDAI, T. & CSILLAG, G. (eds.): (1999): A Balaton-felvidék földtana [*Geology of the Balaton Highland* – in Hungarian].– Magyar Állami Földtani Intézet, Budapest, 257 p.

BOHN, P. (1979): A Keszthelyi-hegység regionális földtana [*Geology of the Keszthely Hills* – in Hungarian].– Geologica Hungarica. Series Geologica 19, 197 p.

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

CASAGRANDE, A. 1932. Research on the Atterberg limits of soils.– Public Roads, 13/8, 121–136.

CAMERA, C., ZOMENI, Z, NOLLER, J.S., ZISSIMOS, A.M., CHRISTOFOROU, I.C. & BRUGGEMAN, A. (2017): A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization.– Geoderma, 285, 35–49. doi: 10.1016/j.geoderma.2016.09.019

CLARKE, F.W. (1920) The Data of Geochemistry.– Bulletin of the U.S. Geological Survey. Washington, 832 p.

DHARUMARAJAN, S., HEGDE, R. & SINGH, S.K. (2017): Spatial prediction of major soil properties using Random Forest techniques - A case study in semi-arid tropics of South India.– Geoderma Regional, 10, 154–162. doi: 10.1016/j.geodrs.2017.07.005

EMBEY-ISZTIN, A., DOBOSI, G., JAMES, D., DOWNES, H., POULTIDIS, C. & SCHARBERT, H.G. (1993): A compilation of new major, trace element and isotope geochemical analyses of the young alkali basalts from the Pannonian Basin.– Fragmenta Mineralogica and Palaeontologica, 16, 5–26.

EUROPEAN ENVIRONMENT AGENCY (2013): CORINE Land Cover (CLC) 2006, Version 17, Kopenhagen K, Denmark. CRC/TR32 Database (TR32DB).

FAO (2006): Guidelines for Soil Descriptions. 4th edition. Food and Agriculture Organization of the United Nations. Rome, 77 p.

FICK, S.E. & HIJMANS, R.J. (2017): Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas.– International Journal of Climatology. doi: 10.1002/joc.5086

GYALOG, L. & SÍKHEGYI, F. (eds.)(2005): Magyarország földtani térképe, M=1:100 000. [*The Geological Map of Hungary, 1:100 000* – in Hungarian].– Magyar Állami Földtani Intézet kiadványa. Budapest.

GYALOG, L. & HORVÁTH, I. (2004): A Velencei-hegység és a Balatonfő földtana [*Geology of the Velence Hills and the Balatonfő* – in Hungarian].– Magyar Állami Földtani Intézet, Budapest, 316 p.

GRAY, J.M., BISHOP, T.F.A. & WILFORD, J.R. (2016): Lithology and soil relationships for soil modelling and mapping.– Catena, 147, 429–440. doi: 10.1016/j.catena.2016.07.045

GU, J., HUANG, Z., FAN, H., JIN, Z., YAN, Z. & ZHANG, J. (2013): Mineralogy, geochemistry, and genesis of lateritic bauxite deposits in the Wuchuan–Zheng'an–Daozhen area, Northern Guizhou Province, China.– Journal of Geochemical Exploration, 130, 44–59. doi: 10.1016/j.gexplo.2013.03.003

HIJMANS, R.J. (2016): raster: Geographic Data Analysis and Modeling. R package version 2.5-8. https://CRAN.R-project.org/package=raster

JARVIS, A., REUTER, H.I., NELSON, A. & GUEVARA, E. (2008): Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org).

JENNESS, J. (2006): Topographic Position Index extension for ArcView 3.x,v.1.3a. Jenness Enterprises. [online: www.jennessent.com/arcview/tpi.htm]

JENNY, H. (1941): Factors of soil formation.– McGray-Hill Book Company, New York, 281 p.

JUHÁSZ, Á. (1962): A balatonfelvidéki permi homokkőösszlet kvarcporfiranyagának eredete [*The origin of the quartz porphyry material in the Permian sandstones of Balaton Highland* – in Hungarian].– Földtani Közlöny, 92/2, 160–173.

KOCSIS, M., TÓTH, G., BERÉNYI ÜVEGES, J. & MAKÓ, A. (2014): Az Agrokémiai Irányítási és Információs Rendszer (AIIR) adatbázis talajtani adatainak bemutatása és térbeli reprezentativitás-vizsgálata [*Presentation of soil data from the National Pedological and Crop Production Database and investigations on spatial representativeness* – in Hungarian].– Agrokémia és Talajtan, 63, 223–248.

KUHN, M. (2008): Building predictive models in R using the caret package.– Journal of Statistical Software, 28/5, 1–26. doi: 10.18637/jss.v028.i05

LIAW, A. & WIENER, M. (2002). Classification and Regression by random Forest.– R News 2/3, 18–22.

MCBRATNEY, A.B., MENDONCA SANTOS, M.L. & MINASNY, B. (2003): On digital soil mapping.– Geoderma, 117, 3–52. doi:10.1016/S0016-7061(03)00223-4

PETTIJOHN, F.J. (1963): Chemical composition of sandstone – excluding carbonate and volcanic sands. In: FLEISCHER, M. (ed.): Data on Geochemistry. Geological Survey Professional Paper.– United States Department of Interior. US Government Printing Office, Washington.

PRASAD, A.M., IVERSON, L.R. & LIAW, A., (2006): Newer classification and regression tree techniques: bagging and random forests for ecological prediction.– Ecosystems, 9, 181–199. doi:10.1007/s10021-005-0054-1.

R CORE TEAM (2017): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RODRÍGUEZ-LADO, L. & LADO, M. (2017): Relation between soil forming factors and scaling properties of particle size distributions derived from multifractal analysis in topsoils from Galicia (NW Spain).– Geoderma, 287, 147–156. doi: 10.1016/j.geoderma.2016.08.005

SARBAS, B. & NOHL, U. (2008): The GEOROC database as part of a growing geoinformatics network.– In: BRADY, S.R., SINHA, A.K. & GUNDERSEN, L.C. (eds.): Geoinformatics 2008-Data to Knowledge, Proceedings: U.S. Geological Survey Scientific Investigations Report 2008-5172 (2008), 42–43.

SILVA, Y.J.A.B., NASCIMENTO, C.W.A., BIONDI, C.M., STRAATEN, P., SOUZA JR., V.S. & FERREIRA, T.O. (2016): Weathering rates and carbon storage along a climosequence of soils developed from contrasting granites in northeast Brazil.– Geoderma, 284, 1–12. doi: 10.1016/j.geoderma.2016.08.009

STEFANOVITS, P.- FILEP, GY.- FÜLEKY, GY. (2005): Talajtan.– Mezőgazda Kiadó, Budapest, 470 p.

TÓTH, G., HENGL, T., HERMANN, T., MAKÓ, A., KOCSIS, M., TÓTH, B. & BERÉNYI ÜVEGES, J. (2015): Magyarország mezőgazdasági területeinek talajtulajdonság-térképei [*Soil property maps of the agricultural area in Hungary* – in Hungarian]. Institute for Environment and Sustainability, JRC, Ispra, 63 p., doi: 10.2788/318926

ÚJVÁRI, G., VARGA, A. & BALOGH-BRUNSTAD, ZS. (2008): Origin, weathering, and geochemical composition of loess in southwestern Hungary.– Quaternary Research, 69, 421–437. doi:10.1016/j.yqres.2008.02.001

VAN ENGELEN, V. & DIJKSHOORN, J. (eds.) (2013): Global and National Soils and Terrain Digital Databases (SOTER), Procedures Manual, version 2.0. ISRIC Report 2013/04. Wageningen, the Netherlands: ISRIC — World Soil Information, 192 p.

VAYSSE, K. & LAGACHERIE, P. (2015): Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France).– Geoderma Regional, 4, 20–30. doi: 10.1016/j.geodrs.2014.11.003

VÁRALLYAY, GY., SZABÓNÉ KELE, G., BERÉNYI ÜVEGES, J., MARTH, P., KARKALIK, A. & THURY, I. (2010): Soil conditions in Hungary based on the data from the Soil Conservation Information and Monitoring System (SIMS).– Ministry of Agricultural and Rural Development. Budapest, 91 p.

VÁRALLYAY, GY. (2013): Soil Scientific Basis of Agricultural Water Management. - University of Debrecen, Service Sciences Methodology Centre, Debrecen, 219 p.

WEISS, A. (2001): Topographic Position and Landforms Analysis. Poster presentation, ESRI User Conference, San Diego, CA.

WIESMEIER, M., BARTHOLD, F., BLANK, B. & KÖGEL-KNABNER, I., (2011): Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem.– Plant Soil, 340, 7–24. doi:10.1007/s11104-010-0425-z

XIONG, X., GRUNWALD, S., MYERS, D.B., KIM, J., HARRIS W.G. & COMERFORD, N.B. (2014): Holistic environmental soil-landscape modeling of soil organic carbon.– Environmental Modelling & Software, 57, 202–215. doi:10.1016/j.envsoft.2014.03.004