Hungarian Geographical Bulletin 62 (3) (2013) 313–320.

ValiDat.DSM, a new soil data validation dataset for Central Europe

Endre DOBOS[1], Erika MICHÉLI[2], Emil FULAJTÁR[3], Vít PENÍŽEK[4] and Marcin ŚWITONIAK[5]

Abstract

Digital soil maps are often derived using digital soil mapping tools, satellite imageries and digital terrain models as environmental covariates. Therefore several new datasets are raster based data representing soil classification categories, like WRB reference soil groups. Validating raster datasets with categorical data is not well researched and supported. No procedure and validation datasets exist that can take categorical diversity and similarity (taxonomic distance) into consideration. This approach would require an input validation dataset describing the categorical diversity of the spatial units to be validated. The aim of this study is to introduce a novel dataset developed for this purpose.

Keywords: ValiDat.DSM, soil validation, DSM, raster dataset, categorical data

Introduction

Digital soil mapping has become a very efficient tool in soil science, and several applications have been published (McBratney, A.B. *et al.* 2003; Lagacherie, P. *et al.* 2006). Many of these applications use environmental covariates like remotely sensed images and digital elevation models, which are raster based data sources with block support. Raster format is favoured by the many users as well. The majority of soil data users require data in raster format with values of certain properties, like pH, clay content or soil organic

[1] Institute of Geography, University of Miskolc. H-3515 Miskolc-Egyetemváros, Hungary. E-mail: ecodobos@uni-miskolc.hu
[2] Department of Soil Science and Agrochemistry, Szent István University. H-2100 Gödöllő, Páter Károly u. 1. Hungary, E-mail: micheli.erika@mkk.szie.hu
[3] Soil Science and Conservation Research Institute. Gagarinova 10, Bratislava, 827 13, Slovakia.
[4] Czech University of Life Sciences Prague. Kamýcká 129, Praha, 165 21, Czech Republic.
[5] Nicolaus Copernicus University. Gagarina 11, Torun, 87–100, Poland.

matter content. Qualitative data can be later classified and used as categorical data. The most typical categorical soil data is the soil type/classification category, like WRB (IUSS Working Group WRB, 2006) or the national classification systems.

Pixels represent a homogeneous spatial object having only one descriptive value or class allocated to it. However, the land surface area represented by a pixel has a more or less heterogeneous soil coverage. This heterogeneity is difficult to handle in a "one value environment". The quantitative variables often use the average value, while the categorical variables use the dominant class of the pixel area.

Both methods simplify the real heterogeneity of the area. Quantitative information can be further explained by descriptive statistics, like standard deviation, minimum, maximum, range etc. Explaining the diversity is more difficult for the categorical data. A potential way to characterize the pixel area is the fuzzy membership approach, when each potential class is represented by a corresponding layer representing the occurrence likelihood or spatial share of the given soil class within the pixel (A-Xing Zhua, *et al.* 2010; De Gruijter, J.J. and McBratney, A.B. 1988; McBratney, A.B. and Odeh, I.O.A. 1997; McBratney, A.B. *et al.* 1992, 2000). This is an appropriate way to keep the heterogeneity information, but user do not prefer this way of information presentation due to its data complexity. Fuzzy data sets are often simplified in the preprocessing steps by selecting the one with the highest share - namely the dominant class - and the rest of the information is lost.

The presentation and validation of the raster based, categorical soil data is not well developed. The e-SOTER project developed a novel approach to present categorical information on block support. The resulting dataset has several layers of occurrence probabilities of WRB diagnostic horizons/features/properties and an additional layer of the reference soil group (RSG) of the WRB system (IUSS Working Group WRB, 2007). However, no appropriate validation methodology and data exist so far.

This paper describes a novel approach for the development of a validation database, entitled as ValiDat.DSM and its potential use for validating digital soil mapping derived WRB reference soil groups and the occurrence probabilities of selected diagnostics. The sampling methodology combines an automated simple random sampling with slight adjustment for better accessibility and fit to the raster database and a systematic random sampling approach to populate the selected pixels with additional observations.

Methods

Overall validation procedure

An external validation dataset was developed for predicting the accuracy of categorical raster soil datasets. The *ValiDat.DSM* dataset has 114 validating sites from the four Visegrád Countries: 17 from the Czech Republic, 58 from Hungary, 23 from Poland and 16 from Slovakia *(Figure 1*). The sites/pixels for validation were randomly selected. All sites had 5 observations falling within a 450 by 450 meters pixel area. Having these 5 observations, proportions of the RSG within the pixel can be approximated with 20%, 40%, 60%, 80% and 100% coverage.

The coordinates of the sites are given in WGS_1984_UTM_Zone_34N projection system (Projection: Transverse Mercator, False Easting: 500,000, False Northing: 0, Central Meridian: 21, Scale Factor: 0.999600, Latitude of origin: 0, Linear Unit: Meter, Datum: D_WGS_1984).

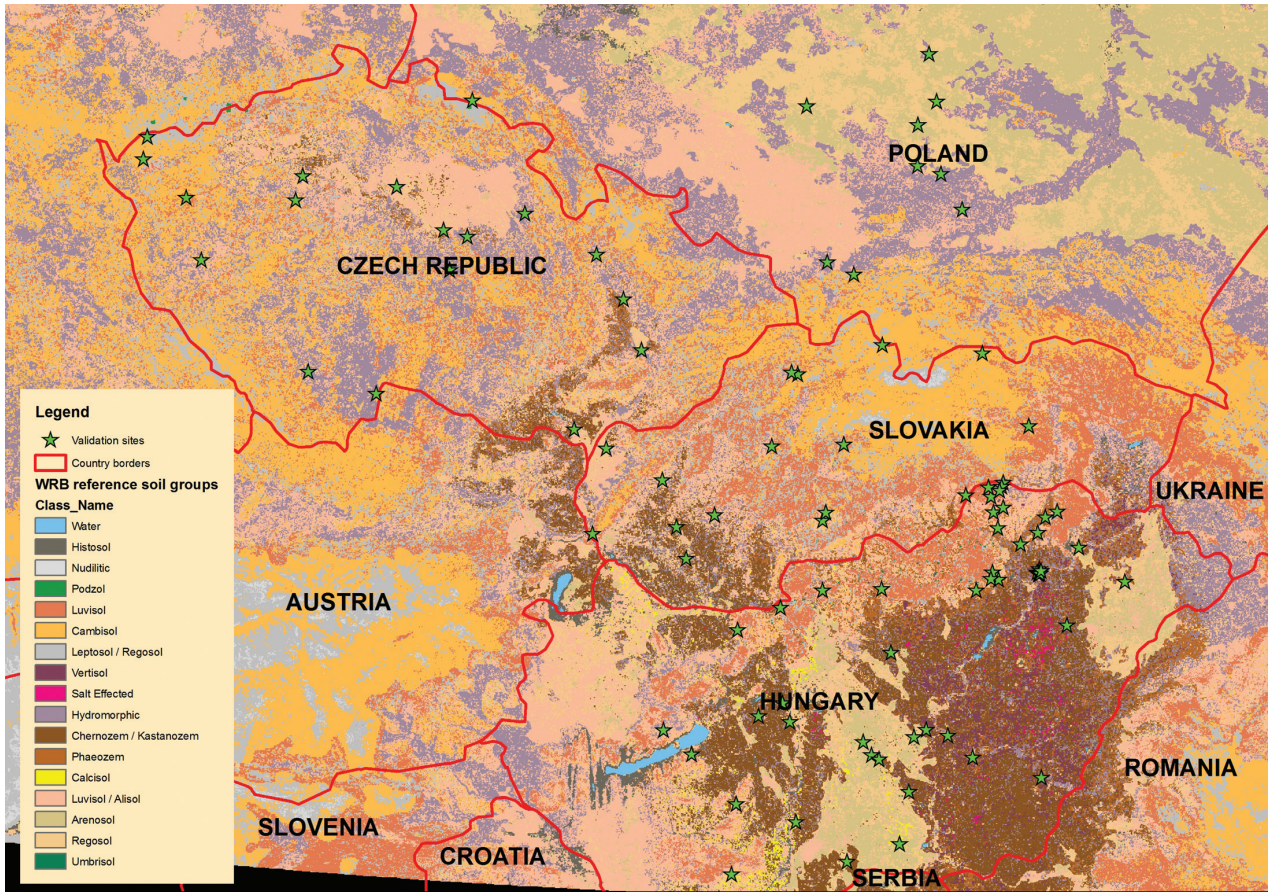

Fig. 1. The position of the validation sites

Field work and database design

Each sites had one profile opened in the centre of the selected pixel. This soil pit was described and all WRB diagnostic criteria, materials, horizons and features have been documented and the classification name was defined.

Chemical properties were identified using only field tools, like HCl 10% solution for $CaCO_3$ content, pH indicators, alpha-alpha-dipiridil test for free iron detection. The interpretation and the translation of the results into quantitative information were done using expert knowledge and the soil description guidelines of the FAO (FAO, 2006). Field work for the core set (65 profiles) was done by an international expert group representing all four countries.

Four additional augerings were deepened 100 m North, East, South and West from the pit. The material taken out from the hole has been put into a 1 m long tray keeping the original depth. By this way a disturbed profile has been created and was taken back to the pit, where all four were put next to each other in a clockwise order starting from North. Documenting photos were taken from the trays and the pit as well (*Photo 1a, b*). All four disturbed profiles have been described in the same way as the pit.

In some cases, where the disturbed material did not let us recognizing the diagnostic the features important for the classification (like the lamellas or clay coatings), the existence or lacking of them was assumed based on the pit description. At the end a table was compiled with five observations and all diagnostic properties, features, horizons and material have been listed for each of the observations (*Table 1*).

Based on the five observations per site, a table with the RSG classes and diagnostics were listed with an appropriate proportion rounded up to 20%, like 20%, 40%, 60% 80% and 100% (*Table 2*).

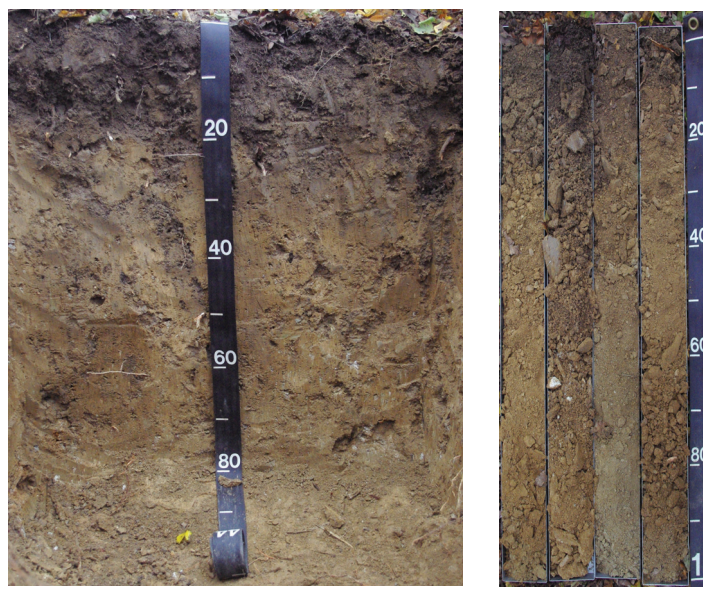

Photo 1a, b. Standard photos of the profiles and the four augerings. The soil trays from left to right are in clockwise order starting from North (N–E–S–W respectively)

*Table 1. An example of the validation dataset**

Colour	pH top	Texture	$CaCO_3$	Diagnostic horizons, properties, matterials	WRB name	North and East		South and West	
						Diagnostic horizons, properties, matterials	WRB name	Diagnostic horizons, properties, matterials	WRB name
Ap: 0–20 cm: 10YR 5/4	4–5	Sand	0	Arenic	Lamellic	Arenic	Lamellic	Arenic	Lamellic
Bt/C: 20–80 cm 10YR 5/6	–	–	–	Dystric	ARENOSOL	Dystric	ARENOSOL	Lamellic	ARENOSOL
Lamella: 10YR 4/4	–	–	–	Lamellic	(Dystric	Lamellic	(Dystric)	Dystric	(Dystric)

* Country ID: HU; Profile ID: 1; Site name: Apagy; Coordinatas. x = 567935; y = 5311601

Table 2. The interpreted validation dataset for seven profiles in Hungary

Profile ID	Co-ordinates		Class probability* %						WRB Reference Soil Group probability	
	x	y	Gleyic-Stagnic-Reducing condition	Mollic Horizon	Calcic Horizon	Calcic Horizon (Calcisol)	Dystric	Eutric	RSG	%
1	567935	5311601	–	–	–	–	100	–	Arenosol	100
2	528219	5281928	–	100	100	–	–	100	Chernozem	100
3	510872	5177845	100	100	100	–	–	100	Chernozem	100
4	446961	5206605	–	100	100	–	–	100	Chernozem	100
5	432353	5210713	–	100	100	–	–	100	Chernozem	100
6	394920	5193366	–	–	–	80	–	80	Calcisol	80–20
7	420484	5168258	40	80	80	20	–	100	Arenosol Chernozem Calcisol	80–20

* Values for Spodic, Argic, Cambic, Vertic (Vertisol), Salic and Natric Horizons were 0%.

100% was given for a certain diagnostic, when it could be found in all observations, while 40% was given when 2 out of the five showed the certain feature. The RSG column lists all RSG observed in the site having the proportion list as well, where the proportions are rounded in the same way as for the diagnostics and sums up to 100% to a site.

Site selection methodology

The sites have been selected randomly. These sites had to be moved to the closest pixel centres and checked for accessibility, potential disturbance or other restricting factors. The whole site optimization procedure was programmed in ArcGIS, no personal bias could have a significant impact on the site selection.

Required input data:

– randomly generated sampling points by ArcGIS random point generator;

– the raster dataset to be validated (in this case the e-SOTER Central European window);

– vector-based GIS databases of the settlements, road and railroad networks, water bodies, nature conservation and other protected areas of the country.

The site location optimization process

A 5 pixel circle shape neighbourhood around the selected point was selected as potential sampling pixels. Because neither the pits, nor the auger sites should be within settlements or on roads, railroads, or any similar locations or even close to them, a 50 m limit was set as minimum distance from the lines or polygons symbolizing them in the vector databases. This limit was increased to 150 m because the auger sites are 100 m far from the profile pit, so to keep the minimum 50 m distance in the case of the auger sites, the pit should be at least 150 m far from the excluded areas. Every points falling within a distance of 150 m from roads, railroads, or settlements were deleted from the possible sampling points, just like the points that were closer to the water bodies, or protected areas then this limit.

At the end an accessibility test was performed on the data. A 500 m maximum allowed distance was set up from the closest road to make sure that the field sampling group does not have to spend too much time on approaching the points and transport the gears there. These two steps of filtering result a set of potentially selectable pixels. The closest to the original randomly selected point was selected as validation pixel.

Results and discussion

Validation of categorical information, like WRB RSG, is a complex problem. Congalton, R.G. (1991) and Brus, D.J. *et al.* (2011) reviewed the most common tools and approaches. Taxonomic adjacency or genetic relationship within a certain set of soil forming factors makes a significant difference in the level of misclassification (Phillips, J.D. 2013). Misclassifying a pixel to a related RSG or to a "nonsense" RSG does not mean the same level of uncertainty. Minasny, B. and McBratney, A.B. (2007) have published an approach to quantify the differences between the soil classes by estimating the taxonomic distances for the WRB RSG classes. This approach is very promising to solve the problem of taxonomic adjacency and quantify

the taxonomic differences. However, the variables and their weights used to calculate the taxonomic distances are needed to be further refined for a more realistic picture. Besides of the lack of an advanced procedure for validation, the most limiting factor is the lack of appropriate, unbiased datasets describing the within-pixel variability, that can be used as ground truth for the validation. The aim of the ValiDat.DSM is to support new initiatives to develop a more appropriate and "standardisable" way of categorical soil data validation.

The ValiDat.DSM dataset has three major forms of information. *Table 1.* shows the field recording sheet. It describes the profile physical and chemical properties needed for the WRB classification procedure and all diagnostics that was identified in the profile and the official WRB classification category. The second half of the table records all diagnostics for the four augerings done 100 m North, East South and West from the profile and also the WRB classification names. This table can be used to understand the site when data used for scientific purposes. *Table 2.* is derived from table 1 by interpreting the soil variability expressed in a selected set of diagnostics important in Central Europe and by the WRB RSG (Dobos, E. *et al.* 2010, 2011, 2013). The tabulated information is complemented with soil profile photos and photos on the landscape and the four auger sites in one picture. This latter one is a magnificent tool for soil diversity representation.

This information can be used as field/ground truth data for validating soil categorical information with estimated proportions or occurrence probabilities. Having information on the spatial share of the soil classes within the pixels, advanced techniques can be used to assess the real reliability of the datasets. The validation can be done considering the taxonomic adjacencies/distances (Phillips, J.D. 2013; Minasny, B. and McBratney, A.B. 2007) between the WRB RSG classes and defining similarity factors to express their relationship in the quantification of the level of misclassification/uncertainty. This dataset can be used for research purposes as well for soil variability studies within different soil forming environments important for soil mapping and for the definition of the minimum set of sampling sites for mapping and validation.

Conclusions

The ValiDat.DSM dataset has been initiated for Central Europe with the contribution of four countries, Czech Republic, Hungary, Poland and Slovakia. The dataset is freely available after registration in the project site (http://www.uni-miskolc.hu/~soil/index.html). Data is presented there in several ways, excel sheet format and the documenting sets of photos and in several kind of GIS environment for visualization helping the users understanding the spatial relationships.

The dataset is a good tool for validating DSM derived soil datasets and for scientific researches on soil variability within different soil forming conditions. At the end each validation sites – pixel area – have 5 observations. Therefore the overall purity – defined as the proportion of the mapped area covered by a certain soil class – can be predicted.

Acknowledgements: Our work has been supported by FP7 project "Regional pilot platform as EU contribution to a Global Soil Observing System" Grant agreement no.: 211578, financed by the European Commission"; by the Hungarian National Scientific Research Foundation (OTKA, Grant No. K105167); by the „Validation of the Central European Soil database" Strategic Grant of the Visegrad Fund, No. 31210072, and by the BONUS-HU Grant No. OMFB-01251/2009 and by the "Excellent Research Faculty" Grant of the Hungarian Ministry of Human Resources" (registration no.: 17586-4/2013/TUDPOL).

REFERENCES

A-Xing, Z., Lin, Y., Baolin, L., Chengzhi, Q., Tao, P. and Baoyuan, L. 2010. Construction of membership functions for predictive soil mapping under fuzzy logic. *Geoderma* 155. (3–4): 164–174.

Brus, D.J., Kempen, B. and Heuvelink, G.B.M. 2011. Sampling for validation of digital soil mapping. *European Journal of Soil Science* 62. 394–407.

Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment* 37. 35–46.

De Gruijter, J.J. and McBratney, A.B. 1988. A modified fuzzy k-means method for predictive classification. In *Classification and Related Methods of Data Analysis.* Ed.: Bock, H.H. Amsterdam, Elsevier, 97–104.

Dobos, E., Seres, A., Vadnai, P. 2010. *Az e-SOTER digitális talajtérképezés módszertana* (Methodology for digitalized soil mapping "e-SOTER"). V. Magyar Földrajzi Konferencia. Pécs. 4–6. Nov. University of Pécs. Manuscript.

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

Dobos, E., Vadnai, P., Michéli, E., Láng, V., Fuchs, M. and Seres A. 2011. *Új generációs nemzetközi talajtérképek készítése, az e-SOTER módszertan* (Making international soil maps of new generation, methodology "e-SOTER"). Térinformatikai Konferencia. Debrecen. 19–20 May, 2011. University of Debrecen. Manuscript.

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

IUSS Working Group WRB, 2006. *World reference base for soil resources 2006*. 2nd edition. World Soil Resources Reports No. 103. Rome, FAO,

IUSS Working Group WRB, 2007. *World reference base for soil resources 2006, update 2007*. 2nd edition. World Soil Resources Reports No. 103. Rome, FAO.

Lagacherie, P., McBratney, A.B., and Voltz, M. eds. 2006. *Digital soil mapping: an introductory perspective*. Amsterdam, Elsevier. 600 p.

McBratney, A.B. and Odeh. I.O.A. 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma* 77. (1–2): 85–113

McBratney, A.B., de Gruijter, J.J. and Brus. D.J. 1992. Spatial prediction and mapping of continuous soil classes. *Geoderma* 54. (1–2): 39–64.

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

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S. and Shatar. T.M. 2000. An overview of pedometric techniques for use in soil survey. *Geoderma* 97. (3–4): 293–327.

Minasny, B. and McBratney, A.B. 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. *Geoderma* 142. (3–4): 28–293.

Phillips, J.D. 2013. Evaluating taxonomic adjacency as a source of soil map uncertainty. *European Journal of Soil Science* 64. 391–400.