A novel approach for validating digital soil datasets with categorical data

Endre Dobos[1]*, Erika Micheli[2], Diana Bertoti[1], Vince Lang[2] and Karoly Kovacs[1]

[1] *Physical Geography, University of Miskolc, Hungary*

[2] *Soil Science and Agrochemistry, Szent Istvan University, Hungary*

ecodobos@uni-miskolc.hu

http://www.uni-miskolc.hu/~soil/

Visegrad Fund
www.visegradfund.org

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 method and validation dataset design developed for this purpose.

The validation dataset Validat.DSM

The e-SOTER project have a developed several soil data layers for Central Europe using existing profile data and several digital soil mapping techniques. Detailed description of the methodology can be found on the e-SOTER website (www.esoter.net). This is a novel product to replace the traditional soil maps, and is appropriate to answer specific soil related questions and provide inputs for several interdisciplinary models on a national and on the regional level. This process requires a scientific validation and a quantitative assessment and characterization of the dataset. Validat.DSM database is designed for this purpose.

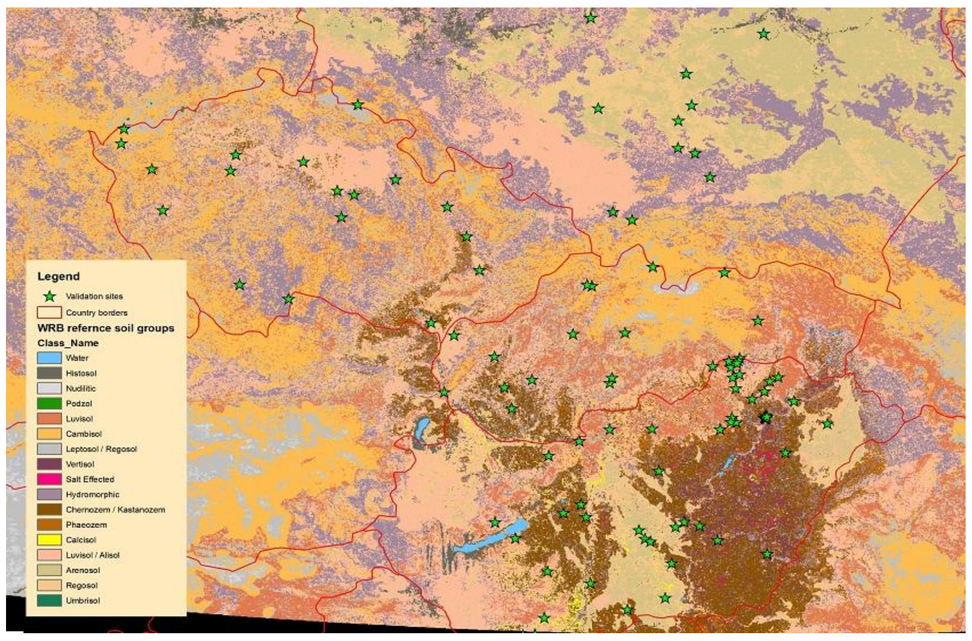


Figure 1. The position of the validation sites in Central Europe

Major steps to develop the validation dataset

1.Random point sampling to select the pixel for validation.

2.Automated procedure to optimaze the location.

3.Exclude areas where sampling is not possible (minimum 150 metres from all excluded areas, maximum 500 meters from the road system).

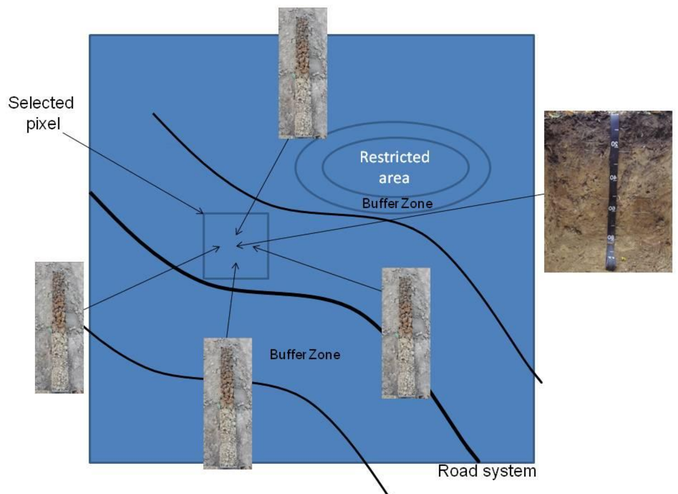
4.Selection of the pixel center where profile is excavated for detailed description.

5.4 augerings from 100 meters North, East, South and West from the profile.

6.This soil pit and the 4 augerings were described and all WRB diagnostic criteria, materials, horizons and features have been documented and the classification name was defined.

7.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.).

8.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 percents, like 20, 40, 60 80 and 100 percent (Table 2.). 100 percent was given for a certain diagnostic, when it could be found in all observations, while 40 percent 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 percent to a site.



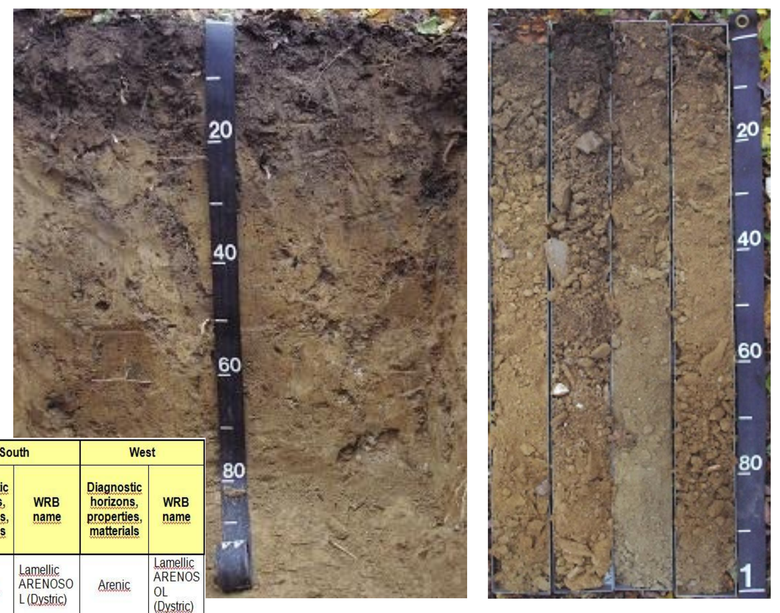


Figure 2. Standard photos of the profiles and the four augerings.
The soil trays from left to right are in clockwise order starting from North (N-E-S-W respectively).

Country ID	Profile ID	Site name	coord. x	coord. y	Colour	pH top	Texture	$CaCO_3$	Diagnostic horizons, properties, matterials	WRB name	North		East		South		West	
											Diagnostic horizons, properties, matterials	WRB name	Diagnostic horizons, properties, matterials	WRB name	Diagnostic horizons, properties, matterials	WRB name	Diagnostic horizons, properties, matterials	WRB name
HU	1	Apagy	567935	5311601	Ap: 0-20 cm: 10YR 5/4 Bt/C: 20-80 cm: 10YR 5/6 lamella: 10YR 4/4	4-5	sand	0	Arenic Dystric Lamellic	Lamellic ARENOSOL (Dystric)	Arenic Dystric Lamellic	Lamellic ARENOSOL (Dystric)	Arenic Dystric Lamellic	Lamellic ARENOSOL (Dystric)	Arenic Lamellic Dystric	Lamellic ARENOSOL (Dystric)	Arenic Lamellic Dystric	Lamellic ARENOSOL (Dystric)

Table 1. An example of the validation dataset

Country ID	Profile ID	coord. x	coord. y	Class Probability (%)												WRB Reference Soil Group Probability	
				Spodic Horizon	Argic Horizon	Cambic Horizon	Vertic Horizon (Vertisol)	Salic Horizon	Natric Horizon	Gleyic-Stagnic-Reducing condition	Mollic Horizon	Calcic Horizon	Calcic Horizon (Calcisol)	Dystric	Eutric	RSG	%
HU	1	567935	5311601											100		ARENOSOL	100
HU	2	528219	5281928								100	100			100	CHERNOZEM	100
HU	3	510872	5177845							100	100	100			100	CHERNOZEM	100
HU	4	446961	5206605								100	100			100	CHERNOZEM	100
HU	5	432353	5210713								100	100			100	CHERNOZEM	100
HU	6	394920	5193366										80		80	CALCISOL ARENOSOL	80 20
HU	7	420484	5168258							40	80	80	20		100	CHERNOZEM CALCISOL	80 20

Table 2. The interpreted validation dataset

ACKNOWLEDGEMENT
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 Visegrád 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).

The Validation procedure

1.Development of a similarity table (Table 3.)

Property based calculation of the taxonomic similarity (MINASNY, B., & MCBRATNEY, A. B. 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes, Geoderma 142. 285-293. New York. ISBN 0-486-68128-9.)

Expert knowledge based similarity weights

2. Calculation of the taxonomy based classification accuracy

$$TAPi = \frac{\sum_{j=1}^{n} hj * pj}{100}$$

TAP*i* is the taxonomy based classification accuracy for the site i
n is the number of Reference Soil Groups observed at the site
j is the reference Soil Group identifier
h is the similarity factor between the estimated and observed classes
pj is the spatial share of the íRSG within the validation siteis

3. Calculation of the weighted overall accuracy

$$SÁP = \frac{\sum_{i=1}^{n} TAPi}{n}$$

SÁP weighted overall accuracy
TAPi TAPi is the taxonomy based classification accuracy for the site i
n is the number of validation sites

Similarity factor (h)		
WRB reference soil groups	**0,9**	**0.7**
Histosols	-	Gleysol
Leptosols	-	Umbrisols
Vertisols	Solonetz	Gleysols, Stagnosols, Chernozem/Kastanozem, Phaeozem
Salt-affected soils	Vertisol	Gleysols, Fluvisol
Gleysols	-	Fluvisol, Chernozem/Kastanozem, Phaeozem, Vertisol, Salt-affected soils
Chernozems/ Kastanozem	Phaeozem	Vertisols, Solonetz, Gleysols, Calcisols
Phaeozems	Chernozem/ Kastanozem	Vertisols, Solonetz, Gleysols
Calcisols	Arenosol	Chernozem/Kastanozem, Regosol
Alisols	Luvisol	Cambisol, Stagnosol
Luvisols	Alisol	Cambisol, Stagnosol
Arenosols	Calcisol, Regosol	Podzol
Cambisols	-	Regosol, Luvisol, Podzol
Regosols	Arenosol	Cambisol, Fluvisol, Solonczak
Umbrisol	-	Leptosol
Podzol	-	Cambisol/ Arenosol
Fluvisol	-	Regosol/Gleysol
Stagnosols	Gleysols	Luvisols, Alisols, Vertisols

Table.3. WRB RSG class similarity table

ValiDat.DSM data		e-SOTER estimated data	
Estimated Reference Soil Group	**(p)**	**Reference Soil Group**	**Estimates taxonomic accuracy** (TAP)
ARENOSOL	100	Regosol	0,9
CHERNOZEM	100	Vertisol	0,7
CHERNOZEM	100	Vertisol	0,7
CHERNOZEM	100	Chernozem/ Kastanozem	1
CHERNOZEM	100	Chernozem/ Kastanozem	1
CHERNOZEM	100	Chernozem/ Kastanozem	1
CALCISOL ARENOSOL	80 20	Arenosol	0,92
CHERNOZEM CALCISOL	80 20	Chernozem/ Kastanozem	0,94
CALCISOL	100	Arenosol	0,9
CHERNOZEM	100	Chernozem/ Kastanozem	1
REGOSOL	100	Hydromorphic	0
KASTANOZEM CALCISOL PHAEOZEM	40 40 20	Chernozem/ Kastanozem	0,86

Table 4. Estimated taxonomic accuracies for the validation pixel

Conclusion

Validation of pixel based categorical data, especially the soil classification categories are difficult for several reasons. Pixels are block supported spatial elements with often one value assigned to, aver-age values of quantitative properties or dominant classes. In reality, pixels with relatively large size, large 500 meter resolutions, almost never represent a pure area, there is a significant level of heterogeneity behind. Traditional datasets handled this problem with using soil associations. Is this heterogeneity within 450 meter pixel really that much to worry about? The results proved that almost 80 percent of the pixels are quite homogeneous as far as the RSG is considered only. That is proved by the accuracy indicators as well. However, it may happen that are-as with much more variable soil forming environment show different results.

Soil taxomonic adjacency is a crucial factor for characterizing the accuracy of a dataset. The level of similarities has to be quantified in order to integrate this information into the accuracy assessment algorithms. Taxonomic distance is the most promising approach to express these similarities, but the variables and their probability of occurrence within the RSG classes have to be refined. It was also concluded, that local knowledge is still needed to revise the similarity matrix and adjust it to the local taxonomic and environmental settings.