LEGACY SOIL DATA HARMONIZATION AND DATABASE DEVELOPMENT

E. Dobos[1], T. Bialkó[2], E. Micheli[3] and J. Kobza[4]

[1]Associate Professor/ University of Miskolc, Miskolc-Egyetemváros, 3515, Hungary,
Email:ecodobos@uni-miskolc.hu
[2] Head of Soil Protection Department/Plant Protection and Soil Conservation Authority of B.A.Z. county, Miskolc, Hungary. Email:BIALKO.TIBOR@borsod.ontsz.hu
[3] Professor/Szent István University, Gödöllő, Páter K u. 1. Hungary,
Email:micheli.erika@mkk.szie.hu
[4] Head of the Institute/Soil Science and Conservation Research Institute, Banska Bystrica, Slovakia, Email:kobza.vupop@bystrica.sk

Abstract
Many countries completed large scale (1:5,000 – 1: 25,000) soil surveys decades ago, and have since used their thematic and geographic information to derive thematic soil property layers of the same or smaller scale (1:100,000 and smaller). The new layers are often simply aggregates of the original soil polygons and inherit the same geographic relationships that were delineated in the original data source. In reality, this approach does not use all information of the input data. Instead of aggregating existing maps, the original, non-interpreted field survey point data can be gathered and used for deriving new property layers. The paper aims to summarize a soil database development project using legacy data for a transboundary area, representing two different systems of data collection, storage and management. Recent and archived soil profile data have been collected, including monitoring sites, soil nutrient status campaign data for different periods, and recorded soil profiles from previous soil mapping activities. These data sources have been transformed to have a common theoretical basis using commonly accepted pedotransfer rules and an integrated profile database has been formed. It was used to interpolate soil information and develop soil property maps and layers representing the WRB diagnostic properties and horizons. The creation of the property layers was based on statistical/geostatistical interpolations of the soil profile database using DEM derivatives, SPOT and Landsat satellite images as covariates to provide information for the natural setting of the area. The interpolated values for the numeric variables were estimated using regression kriging, while the classified variables were calculated using the maximum likelihood classification algorithm. It was concluded, that the development of WRB diagnostic criteria database is feasible using raw data of different origin and a set of harmonization and digital soil mapping tools.

Keywords: WRB; soil database development; data harmonization; remote sensing; digital terrain modeling

1. Introduction

Soil data of appropriate format and reliable accuracy are often the most limiting factor of soil related modeling and applications. Many countries have had several data collection campaigns serving different goals, like mapping or agricultural fertility testing. Besides a Canadian example to make use of legacy data in a digital database (see chapter 26) legacy data have also been used for several digital soil mapping applications to derive updated information (Rossiter, 2008, Baxter and Crawford, 2008, Mayr et al. 2008, Dobos et al. 2007, Bernoux et al. 2007, Mayr and Palmer, 2007,). The integration of several legacy data sources is a potential way to create a product with great value added without the need of strong field data collection. One of the key element of any digital soil mapping based database development procedure is the appropriate density of input calibration/training data (section 3 of chapter 29). However, the integration of interpreted maps is often difficult. Thus, a different approach is demonstrated here. A point database was created from each input dataset and an integrated, multi-origin point database was developed after the necessary harmonization and data filtering. Taxonomic harmonization was done using the WRB 2006 classification system (IUSS Working Group WRB, 2006). This database is used as calibration and training datasets for several digital soil mapping tools.

2. Materials and Methods

2.1. Study Area

The study area, called Bodrogköz, is located between the triangle of the Tisza, Bodrog and Latorica Rivers along the eastern section of the Hungarian-Slovakian border (Figure 1.). It represents a homogeneous landscape, a flood plain with some windblown sand dunes, typical for the Pannonian plain. The areas along the major rivers, the so-called natural levees, are a few meters higher than the area behind (backwater area). Its soil texture is much coarser than the backwater area, which is heavy clay overlaying the deeper sand strata. One to two meters of relative elevation difference results in different texture, chemical properties, and also soil and landuse types. The recent landscape-landuse-geomorphologic-parent material system of the area is very much interrelated and defines the soils in an almost deterministic manner. There is also a slight change along the NE-SW direction, which is the major axis driving the surface water flow as well. The NE edge is higher, while approaching to the SW edge the elevation tends to be lower, have more frequent flooding, hydromorphic impact, thus more leaching, lower pH, and higher humus content. This trend varies a little bit in the SW edge, where the two levees of the Tisza and the Bodrog meet and form a joint levee, with somewhat higher elevation, different water regime and coarser texture.

The area has been cultivated for over a thousand years, with a strong intensification starting in the 19^{th} century. The soils were developing under the strong impact of floods and high groundwater table. Flood protection and drainage systems have been constructed since the second half of the 19th century, which has changed the environmental system dramatically.

The landuse and the soil type are highly correlated. Low lying and high ground water areas have pasture and Gleysols on them, while the areas with lower ground water table have Vertisols, Arenosols and Luvisols. These soils are cultivated despite their high acidity and unfavorable textures.

The study area has temperate climate with an annual precipitation of 550 mm and a mean temperature of 10^{o} C. The altitude of the majority of the study area ranges from 90 to 120 meters. Only two small volcanic hills arise from the plain and reach 270 meters. The parent material is mainly alluvial clay and loamy fluvic material. The dominant landuse is farmland with some orchard, forest spots and wet pastures. The most common soil types are Vertisols, Arenosols, Gleysols, Fluvisols and Luvisols (Dobos and Kobza, 2008).

2.2. Digital and Field Data

2.2.1. Point data

The area has 1786 sampling sites, of which 1616 falls to the Hungarian and 164 onto the Slovakian side (Figure 1).

The highest number of points was imported from the Kreybig mapping campaign. These points were chosen as representative and complementary profiles for the 1:25,000 scale mapping, started in the late 1930s. 1161 points were processed, digitized and revalidated (Szabó et al, 2005). Five parameters, namely the 5 hour capillarity rise, pH(KCl), humus%, $CaCO_3$, and salt content were assigned to each point.

The 164 Slovakian sites were part of the 1:10,000 mapping campaign started in the 1980s and contained three parameters, namely the humus%, clay% and pH(KCl).

Official monitoring sites for both countries (18/6, Hu/Sk) were also used for interpolation and for data harmonization with the variables of $CaCO_3$, texture, humus%, and pH(KCl) (Várallyay et al. 1995).

Data from three soil nutrient survey campaigns (TVG) in Hungary between the late 1970's and 1987 were used as well. A total of 422 data points were generated having the following variables: the Arany-type cohesion measure (K_a), humus%, salt content, $CaCO_3$ and pH(KCl).

An additional 16 sites were also sampled as representative, calibration data – benchmark soil sites - and used in the harmonization process. The majority of these points were selected to revisit existing points of other datasets. These points were sampled and lab analyzed for the humus%, texture, pH, CaCO3 and salt content.

2.2.1.1. Point data derivation from averaged field data (TVG data processing)

In order to increase the data point density for areas where no reliable point data source was available non-point data sources were used as well, namely the TVG data. The TVG dataset is a non-point, field-based dataset, with 8 non-located composite samples taken along a recorded transect. Their average was assigned to a parcel, or a part of it, with the size ranging from 10 to 20 hectares. These data were first filtered for field homogeneity and only data representing homogeneous fields were processed and used in this project. Field homogeneity was tested in two ways. First, by looking at the site visually on orthophotos, SRTM terrain derivatives, and multitemporal/multispectral Landsat/SPOT/IKONOS images representing six different dates. Quantitative methods, like the spectral distance based region grow algorithm of the ERDAS Imagine was tested as well (ERDAS, 1999). However, due to the high diversity and variability of the input layers no successful method to define the thresholds has been developed yet. Thus, the thresholds were increased continuously up to the point when the expert judged and the measured results matched. If a match was not obtained within a certain range of threshold values, then the test failed. The second test was to check for deviation and outliers along the area selected in the first step. The measurements obtained in the transects falling within the selected areas were selected and the recorded measurements for each transects have been collected. The acceptable absolute deviation from the average was set by expert judgment for each variable, such as 0.5 for the pH. All measurements having greater deviation than the set value was considered outlier. If any outlier was identified then the area was dropped. If the tests were passed then the center or the most representative looking point of the area was selected and the average value was assigned to it. Data for 422 points were generated in this way having the following variables: the Arany-type cohesion measure (K_a), humus%, salt content, $CaCO_3$ and pH(KCl). This procedure unavoidably introduces some uncertainty to the process, therefore it was used only for areas having limited data.

2.2.1.2. Point data harmonization

The European Union, and also its member states, have several different ways of collecting and analyzing soil samples, and different ways of expressing the results. Therefore using these data sources is not straight forward. Much preprocessing is required to import all of these data into the same reference system. The preprocessing means both the spatial and the attribute data are transformed into a common system. This procedure is called, in our terminology, “harmonization”.

The first step of the harmonization procedure was the field work, when representative profiles were opened in the field, sampled, analyzed in lab, and classified according to WRB 2006. The sites were to represent the major reference/benchmark soils of the area. The site selection was based on existing soil maps, satellite and orthophoto images and on the major geomorphologic units. The joint field work was a crucial step for mentally harmonizing the group members from the two countries, to reach a common understanding of the soil variable interpretation and to develop a mental model of soil variability. Based on the expert/local knowledge learnt from the reference profiles, each input data type was translated to a common variable using existing transformation models or correlation functions developed within the project. The result of this section was a harmonized soil profile database, and a mental model of the soil resources.

Two major variables needed significant effort to harmonize, namely the taxonomic groups (WRB major reference groups, diagnostic horizons and criteria) and the texture. The taxonomic units were identified manually by the country representatives after field harmonization of the interpretation of the diagnostic properties. Numerous misclassified profiles were identified, screened and replaced by a commonly agreed unit. This work was crucial, and much less time-consuming than anticipated. Having the mental model and the field correlation efforts, it was quite easy and fast to screen the problematic profiles and modifiy/correct their classification units.

The property having the highest representation diversity was the texture. Clay % content, capillarity water rise in 5 hours, Arany-type cohesion measure (K_a) and interpreted texture classes were the input types of the different sources. Correlation rules developed by Buzás et al.

(1993) were employed to reach the common platform and convert all properties into the same variable. The less detailed variable, namely the classified texture unit was chosen to serve as the final variable, to which we could adjust/degrade the more detailed parameters. The correlation table is given in Table 1. The rest of the given parameters (humus, CaCO3 and salt %) were in the same units and were analyzed in the same way, so no further thematic harmonization was needed.

Table 1. The correlation table of the three input texture parameters.

	Clay%	**Ka**	**Capillarity water rise in 5 hours (mm)**
Coarse sand	**below 5**	**below 25**	**over 350**
Sand	**5-15**	**25-30**	**300-350**
Loamy sand	**15-20**	**30-37**	**250-300**
Loam	**20-30**	**37-42**	**150-250**
Clay loam	**30-40**	**42-50**	**75-150**
Clay	**40-45**	**50-60**	**40-75**
Heavy clay	**over 45**	**over 60**	**below 40**

Due to the temporal diversity of the input data sources some changes might have happened in the chemical properties over time and could resulted in a shift of the data values, which could significantly decrease the model performances. Therefore a set of t-tests for the humus content and pH were calculated to make sure that all input data sets represent the same population. The values of these two classes were close to normally distributed, skewness 0.6 and 0.5, while the Kurtosis was 3 and 3.3 for the pH and the humus, respectively. Because of the specific environmental setting of the study area – where parent material expresses the geomorphology and the terrain influence on the soils in the same time – a harmonized and simplified quaternary geology database was used to pre-stratify the area. The simplified parent material dataset contained four units: Holocene alluvial clay, Aeolian Dune Sand, Holocene reworked clay-loam alluvium and recent loamy, loamy-sand alluvium. The populations of the different point sources falling into the same parent material class polygons were tested for having the same means at 0.2 level of significance. This value was chosen as a lowest acceptable level. Both of the humus and pH tests were significant.

2.2.2. Other digital data sources

Two Landsat and two SPOT images were selected for the work, both representing different seasons and natural conditions. The SPOT images were taken in May and October of 2006, while the Landsat images were acquired in March, 1999 and in July 2006. The1999 image represents a flooded condition. These data sources were combined into a 22 band image, resampled to 120 meters and used as covariates for the interpolation and classification procedures. The pixel size degradation was carried out to decrease the impacts of artificial landscape patterns and increase the importance of the overall environmental condition.

High resolution digital data for validating the sites were also used. Digital orthophotos from the summers of 2002 (for the Slovakian side) and 2005 (for the Hungarian side) with 2 meter resolution were created to cover the entire study area. An IKONOS multispectral image with 4 meter resolution was also acquired for the entire area for the summer of 2007, when the field sampling campaign was running.

2.2.2.1. Terrain information

The terrain was represented with the 90 meter resolution SRTM data. These data were preprocessed to remove the effect of forests, which was recognized as a major limitation factor. The removal required a forest coverage map. It was created using the SPOT images described above and field training samples. The training samples were taken based on high resolution orthophotos. Maximum likelihood classification algorithm was employed to classify the entire image. The classified image was resampled to the same resolution as the SRTM and then reclassified into two classes, forest and non-forest . This image was used to identify the forest plot edges and an estimated elevation difference was calculated based on the minimum and maximum values within a given size of search window. This edge contour with the estimated elevations was used for lowering the actual SRTM (with the canopy) data. The resulting image was used for the terrain characterization.

Except for the two hills, the area is almost totally flat. Thus the absolute elevation, and other commonly used parameters provided no useful information. Therefore two other topographic parameters were tested, the Topographic Position Index (TPI) (Weiss, 2002) and the Potential Drainage Density (PDD) index (Dobos, E and Daroussin J, 2007) to highlight the relative elevation, namely the low-lying and the elevated areas.

2.3. Inference Models

Figure 2 shows the flowchart of the inference system. The work had three major sections. The first step was the input data harmonization and the creation of the training/calibration point dataset. The second section was the creation of continuous property layers for the final and intermediate layers, like WRB Reference Soil Groups (RSG), texture, pH(KCl) and texture. (Alternative approaches for estimating soil properties based on various origin legacy data is described in sections 2 of chapters 16, 29 and 32.) After checking for potential trends, Universal Kriging and cokriging were used to interpolate the numerical data, namely the pH(KCl) and the humus content. Co-variables for the cokriging were selected by checking the cross-correlation of the variable to predict and the terrain parameters derived from the SRTM, and the best two were used. For the pH, 1611 observations were used and Universal kriging was selected as best performing model. The humus content was estimated with Universal cokriging using PDD as covariable with 657 observations.

Categorical variables, like the WRB Reference groups and the texture, which were only in classified format, were estimated by maximum likelihood classification using the 22 layers combined SPOT and Landsat image, with a degraded resolution of 120 meters. The spatial distribution patterns of both variables were clearly visible on the RGB composite images, thus good performance was expected. Regular accuracy measures, like RMS, standardized RMS and average standard error were calculated and error vs. measured plot was created to visualize the error trends. For the maximum likelihood classification the overall class performance (the correctly classified training pixels / the total number of training pixel), the Kappa statistics and the confusion matrix (user's and producer's accuracies) were calculated to characterize the accuracy (Congalton et al. 1991).

In the last section, the WRB diagnostic properties and horizons were estimated using the four intermediate data layers and pedotransfer functions. Pedotransfer functions are simple or more complex rules/relationships to estimate missing properties based on existing, correlated, and easy to collect/measure properties (McBratney et al. 2002). Table 2. summarizes the pedotransfer functions used for estimating the WRB qualifiers/diagnostics for the study area.

Table 2. The pedotransfer functions used for predicting the WRB qualifiers.

Predicted WRB Qualifiers	**Pedotransfer functions**
Vertic	**All areas where Vertisols exist**
Mollic	**Humus>1% and Eutric**
Arenic	**Having sandy texture**
Clayic	**Having clay texture**
Gleyic	**All areas where Vertisols, Fluvisols and Histosols occur**
Dystric	**pH(KCl)<5**
Eutric	**pH(KCl)>5**
Calcic	**CaCO3 % > 5**

3. Results and Discussion

3.1. Results from Model

3.1.1. The WRB reference soil groups

The WRB reference groups were estimated with maximum likelihood classification of the combined SPOT/Landsat images. Eight soil types appeared on the classified image with a very pronounced spatial distribution pattern (Figure 3). Fluvisols occur along the major rivers on the annually flooded areas. The backwater area behind the sandy levees, is covered by heavy textured Vertisols and Gleysols, with tiny islands of the remaining Histosols. The Northern part of the area is dominated by Luvisols, having a well developed B- horizon with strong clay skins. The small sand dunes have Arenosols and Cambisols on their lower sections. Histosols and Regosols occur as very small islands, representing small drained depressions and loamy plateaus.

This soil distribution pattern was evident from the satellite images. RGB composites of the images showed the extent of the major soil types for the experienced eyes. The visual interpretation of the classified image showed a very good match as well with our local knowledge and mental model. Quantitative tests are given below. However, the risk of having too strong "landuse pattern"-dominated classified soil image was a real possibility. This strong pattern was "softened" by resampling the image to 120 meter resolution and using PCA transformation. The first component of the PCA transformed image always emphasizes the landuse/landcover pattern, while the 2nd, 3rd, and 4th components are more related to secondary variability within-the-1st -component, within the land cover pattern. These secondary, hiding patterns are the ones we often need and are related to the soil characteristics. Using these tools limited the occurrence of the land cover pattern.

The transition zones between the Regosols, Arenosols and Fluvisols classes were often quite difficult to handle, the separation of these taxonomically similar soil types were not always easy to make, even in the field. The subtypes of the reference groups were very similar taxonomywise to the neighboring reference soil group, often representing the transitional types between the reference groups – like Fluvisols and Fluvic Cambisols. However, as classified units they occur far from each other in the classification system due to the hierarchy. This problem had a significant impact on the accuracy measures as well. However, this potential misclassification had more impact on the quantitative accuracy measures, than on the real usefulness of the map.

3.1.2. The property maps

The texture map shows settings similar to the WRB one. The active flood plains have loam and sandy loam texture. The inner part of the area is clay, with small islands of sand dunes occurring in the area. Organic materials and Histosols are very rare. The spatial patterns of the soil texture were easy to follow by simple visual interpretation of the composite satellite images as well. The RGB image of Figure 1. nicely shows the lighter colored levees of the recent and ancient rivers and the darker colored clayey (Gleysol-Vertisol) inland areas.

Similar spatial pattern can be identified in the humus content (Figure 4. and 5). The higher humus content occurs with the clayey soils, where the clay bounds it strongly and the longer water saturation retards the organic matter decomposition. An opposite trend can be identified in the pH map (Figure 6. and 7.). Low pH is linked to the same low lying, clayey areas, where leaching was very active up to the last century. Spherical models were used to fit the curve for both cases. Strong nugget showing significant local variation has been found (Figures 4. and 6.).

A spatial trend in the E-W direction was identified for both the humus and the pH value distributions. These two trends show converse ways, the pH values are the lowest in the centre part and increases towards the ends, while the humus content change the opposite way. These are real trends, and were modeled with a second order de-trend algorithm. The phenomenon is easy to explain. The centre part is the most typical backwater area, far from the major rivers and partly separated from them by the natural levees. The flooding water flowing over this natural levee slow down, looses its heavy sediments and keeps only the small particles like clay. This clay is deposited in the backwater area. The trapped water cannot flow back, even after the flood is over, because the levee blocks its way back to the river. Therefore the water stays there longer and strongly leaches the soils – low pH –, while the high clay content and the long saturation decreases the decomposition of the organic matter and support the higher humus content.

The soils have seven major WRB diagnostic properties and horizons, which have common occurrence and strong importance in defining the soil use (Figure 8.). These diagnostics were created by manipulating the existing layers and combining their information according to the pedotransfer functions of Table 2.

The spatial patterns of the final maps do not match the Hungarian or Slovakian soil maps, which differ from each other and from the WRB classification anyway. However, the shape and extent of the soil regions coincide well with the geomorphologic and agro-environmental patterns of area, and match our mental model well. The WRB and texture maps correspond very well to each other, because they were derived from the same integrated satellite image. However, a very good genetic coincidence appears between the WRB/texture maps and the kriging based humus and pH data, which provides a visual support to the results as well.

3.2. Accuracy Assessment of Model

The Landsat and SPOT image based classification resulted in an overall classification performance of 77 % and a Kappa statistic of 0.7. The confusion matrix is given in Table 3. The User's accuracy ranged between 37 and 99 percent with an average of 64 %, while the Producer's accuracy was between 61 and 94 % with an average of 82 %.

The most severe misclassification occurred in the Histosols and Regosols classes. Both classes occur as small islands, often with a smaller extent than the pixel size used for its classification, which explains their low performance.

RMS, standardized RMS and the average standard error were calculated for the kriging based extrapolations. These values for the pH(KCl) are 0.76, 0.98 and 0.77 respectively, while the values for the humus estimation were 1.13, 1.03 and 1.1. The pH values range between 3.5 and 8. The humus values are between 0 to 8, but can go further up for extreme hydromorphic soils. Both estimations are smoothing the data, the estimation error increases towards the minimum and maximum values, while decreases to 0 around the average.

Table 3. The confusion matrix of the maximum likelihood classification of the WRB reference groups. The values in the matrix are percentages of training pixels from a given class classified into the resulting classes. The values of the "Total" line and column represents training pixel numbers.

Classified	Arenosol	Fluvisol	Histosol	Regosol	Luvisol	Vertisol	Cambisol	Total
Arenosols	**60.54**	3	1.09	2.21	2.38	6.19	5.85	2218
Fluvisols	0.48	**76.81**	0	0	0.42	0.4	0.6	7095
Histosols	4.16	5.55	**93.82**	0.44	3.22	5.91	1.38	1491
Regosols	4.68	5.63	0	**94.03**	2.87	1.15	0.34	1162
Luvisols	5.04	1.62	0.91	1.77	**84.81**	1.98	5.08	1659
Vertisols	17.69	3.93	3.64	0.22	1.4	**81.21**	8.18	5033
Cambisols	7.4	3.45	0.55	1.33	4.9	3.15	**78.57**	1653
Total	2499	9178	550	452	1429	5041	1162	20311

4. Conclusions

Archived legacy data have great value for database development. Huge amounts of data have been collected and recorded in many previous mapping and survey campaigns. These data are often interpreted into thematic polygon maps, and used for many applications. The integration of these types of data sources can improve the reliability and accuracy of our soil databases, and creates new generation data with added value. The best way to do so is to use the “raw” field survey observations as a profile database, or derive representative point data from averaged polygonal information. The integration and harmonization of these profile databases is the best and most consistent way of combining information of different origin and interpolate their information using digital soil mapping tools. It was also concluded that the diagnostic features, materials and horizons of WRB can be estimated from harmonized, variable origin, integrated data sources.

5. References

Baxter, S.J., Crawford, D.M. 2008. Incorporating legacy soil pH databases into digital soil maps. In: Hartemink, A.E., McBratney, A., Mendonca-Santos, M.L. (eds.) Digital soil mapping with limited data. Springer.

Bernoux, M., Arrouays, D., Cerri, C.E.P., Cerri, C.C. 2007. Regional organic carbon storage maps of the Western Brazilian Amazon based on prior soil maps and geostatistical interpolation. In: Lagacherie, P., McBratney, A.B., Voltz M. (eds.) Digital Soil Mapping. An introductory perspective. Developments in Soil Science. Vol 31.

Congalton, R. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote sensing of Environment. 37:35-46.

Dobos, E. and Daroussin, J. 2007. Calculation of potential drainage density index (PDD). In: Peckham, R. and Jordán, Gy. (eds.) Digital Terrain Modeling. Development and Applications in a Policy Support Environment. European Commission, Springer, pp.283-295. ISBN 978-3-540-36730-7.

Dobos, E., Micheli, E., Montanarella L. 2007. The population of a 500 meter resolution soil organic matter spatial information system for Hungary. In: Lagacherie, P., McBratney, A.B., Voltz M. (eds.) Digital Soil Mapping. An introductory perspective. Developments in Soil Science. Vol 31.

Dobos, E. and Kobza, J. 2008. A Bodrogköz talajai. (The soils of the Bodrogköz) In.: Dobos, E., and Terek, J., (eds.) Élet a folyók között. A Bodrogköz tájhasználati monográfiája. Miskolci Egyetem. ISBN:9789630642644

ERDAS Inc. 1999. ERDAS Field Guide. ERDAS Inc, Atlanta, Georgia, USA.

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

Mayr T., Palmer, B. 2007. Digital soil mapping: An England and Wales perspective. In: Lagacherie, P., McBratney, A.B., Voltz M. (eds.) Digital Soil Mapping. An introductory perspective. Developments in Soil Science. Vol 31.

Mayr, T., Palmer, R.C., Cooke, H.J. 2008. Digital soil mapping using legacy data in the Eden valley, UK. In: Hartemink, A.E., McBratney, A., Mendonca-Santos, M.L. (eds.) Digital soil mapping with limited data. Springer.

McBratney, A.B.,Minasny, B., Cattle, S.R., Vervoort, R.W. 2002. From pedotransfer functions to soil inference systems. Geoderma. 109, 41-73.

Rossiter. D.G. 2008. Digital soil mapping as a component of data renewal for areas with sparse soil data infrastructure. In: Hartemink, A.E., McBratney, A., Mendonca-Santos, M.L. (eds.) Digital soil mapping with limited data. Springer.

Szabó, J., Pásztor, L. & Bakacsi, Zs. 2005. Egy országos átnézetes talajinformációs rendszer kiépítésének igénye, lehetősége és lépései. Agrokémia és Talajtan 54:1-2.

Várallyay, Gy., M. Hartyáni,. P. Marth, E. Molnár, G. Podmaniczky, I. Szabados & Kele. G. 1995. Talajvédelmi Információs és Monitoring Rendszer. 1 kötet. Módszertan. (Soil Monitoring Sytem for Soil Protection. The Manual of Procedures) Földművelésügyi Minisztérium, Budapest

Weiss, A. 2001. Topographic Position and Landforms Analysis. Poster presentation, ESRI User Conference, San Diego, CA.

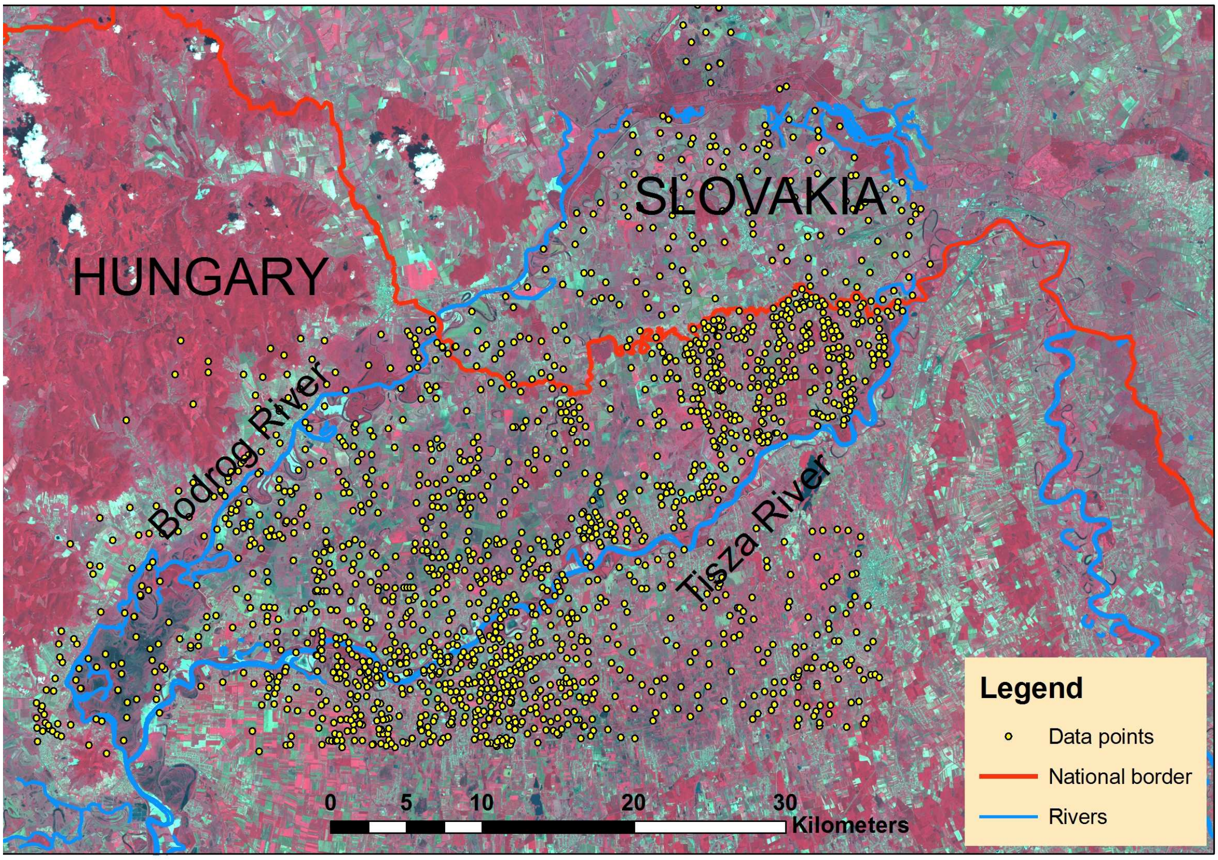

Figure 1. The location of the study area and the sampling sites (in yellow) over a Landsat RGB composite image (Bands 4,3,2).

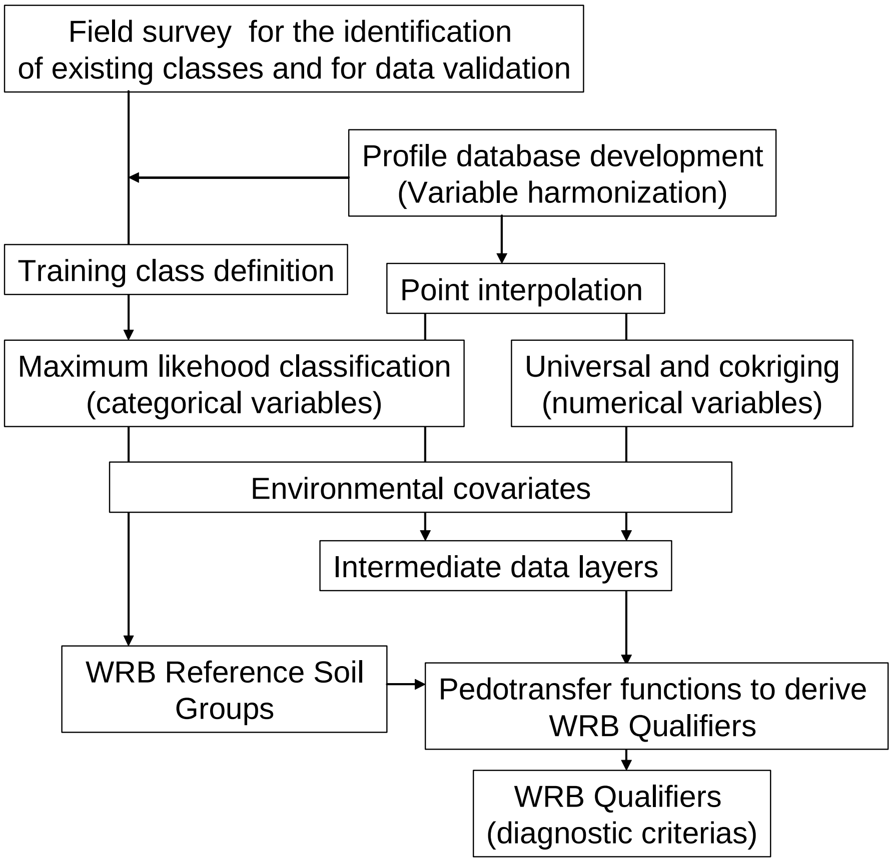

Figure 2. The flowchart of the inference system used within this project.

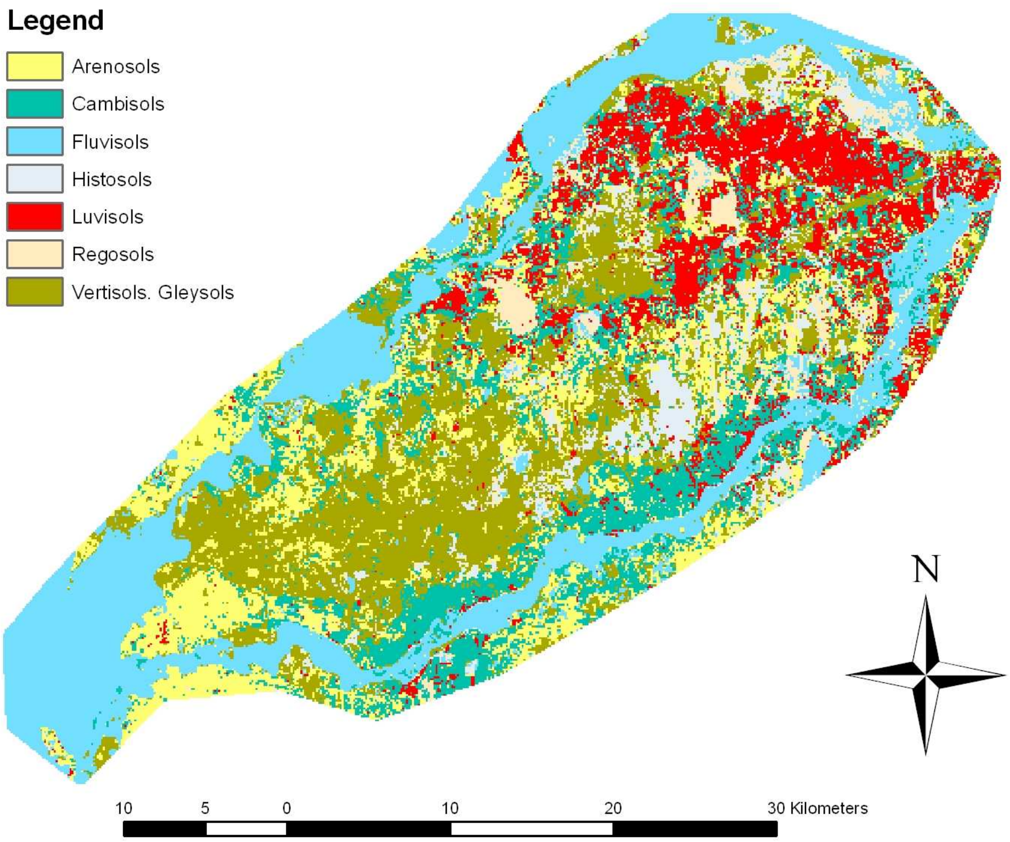

Figure 3. The WRB Reference Soil Groupings map

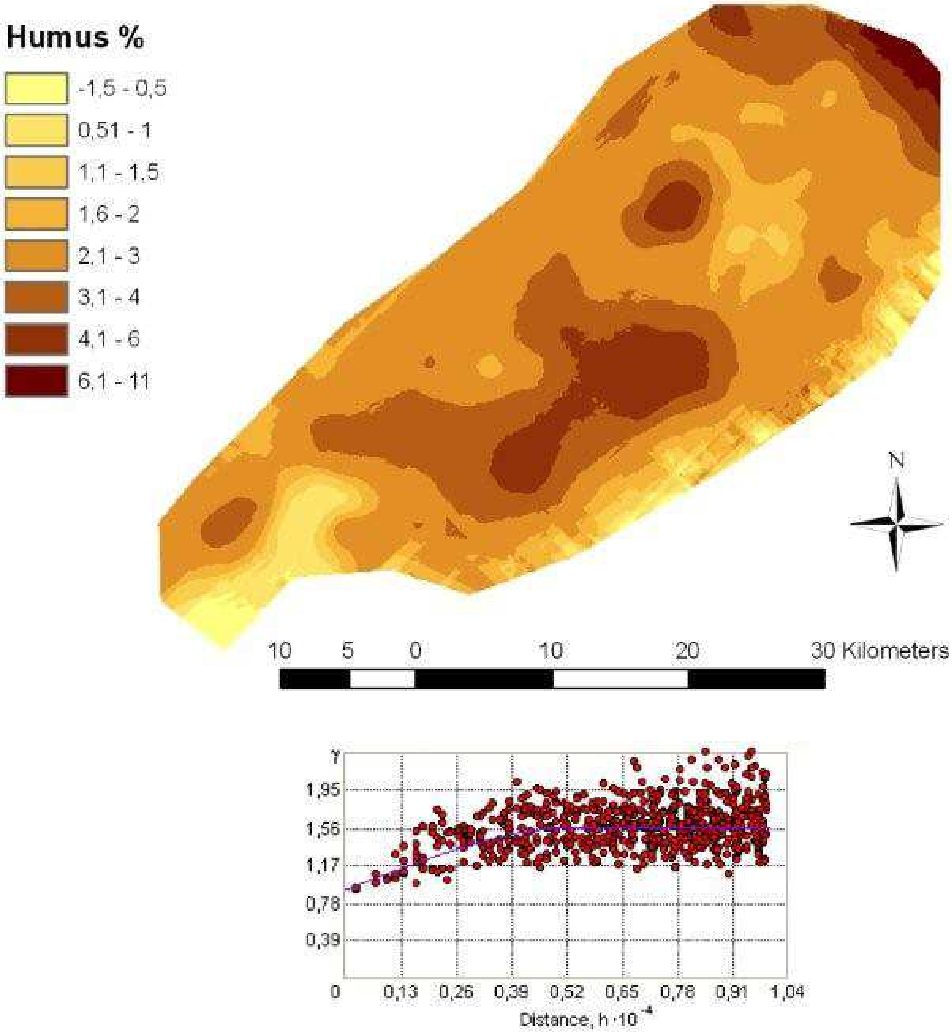

Figure 4. The humus content layer and the calculated semivariogram, where ɤ(h) is the semivariance function of the humus content in the function of the lag distance (h).

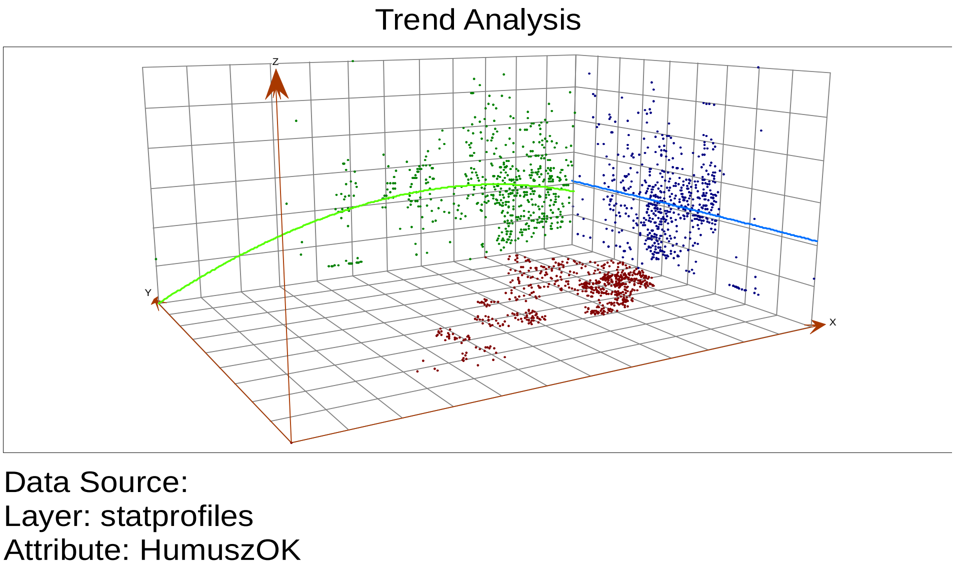

Figure 5. The trend analysis diagram for the humus content. The colors indicate the different dimensions, red is the horizontal plain, blue is the North-South direction, while the green color is the East-West one. The red points show the horizontal distribution/location of the points, while the blue and green ones refer to the pH values along the NS and the EW directions respectively.

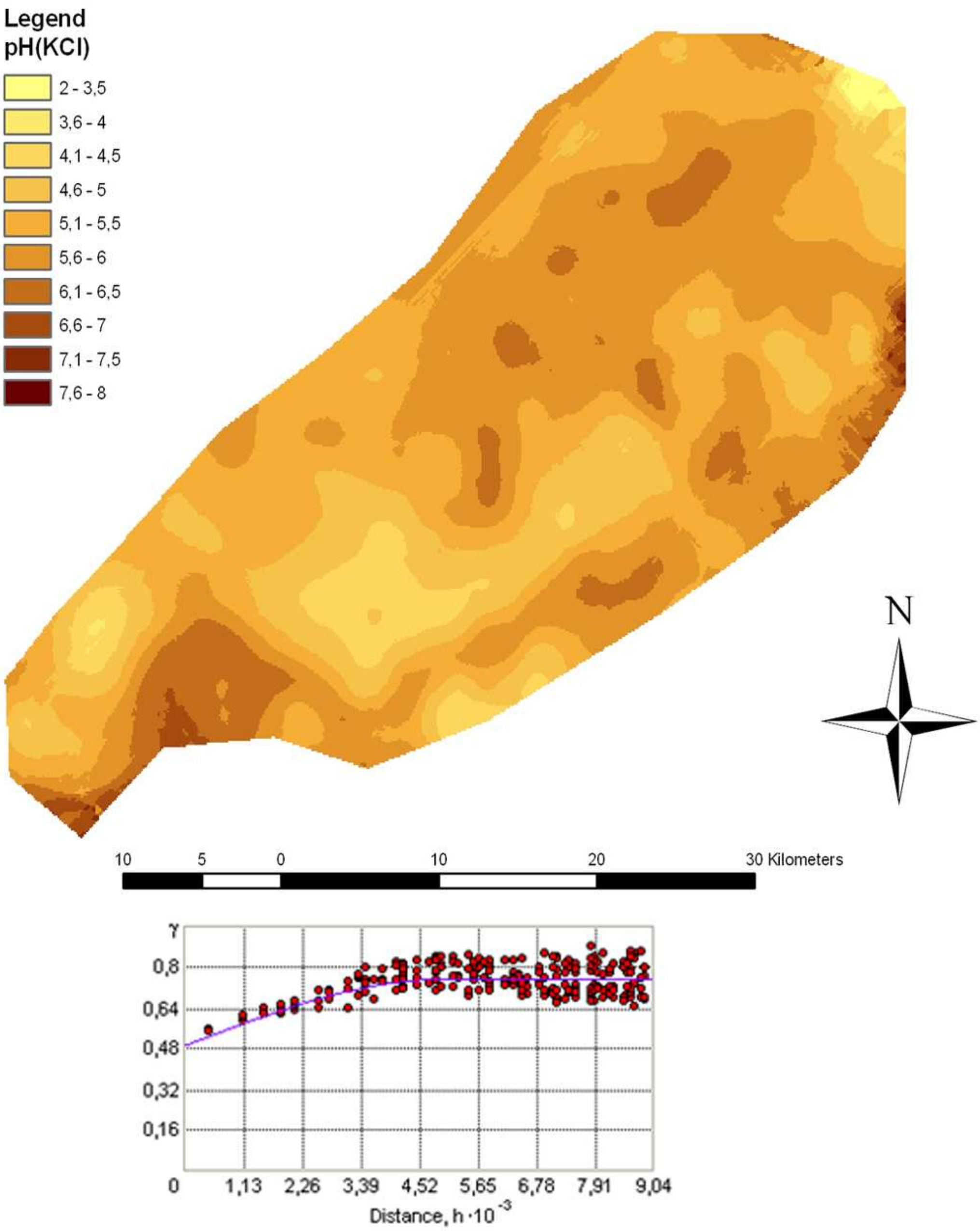

Figure 6. The pH(KCl) layer and the semivariogram, where ɣ(h) is the semivariance function of the pH in the function of the lag distance (h).

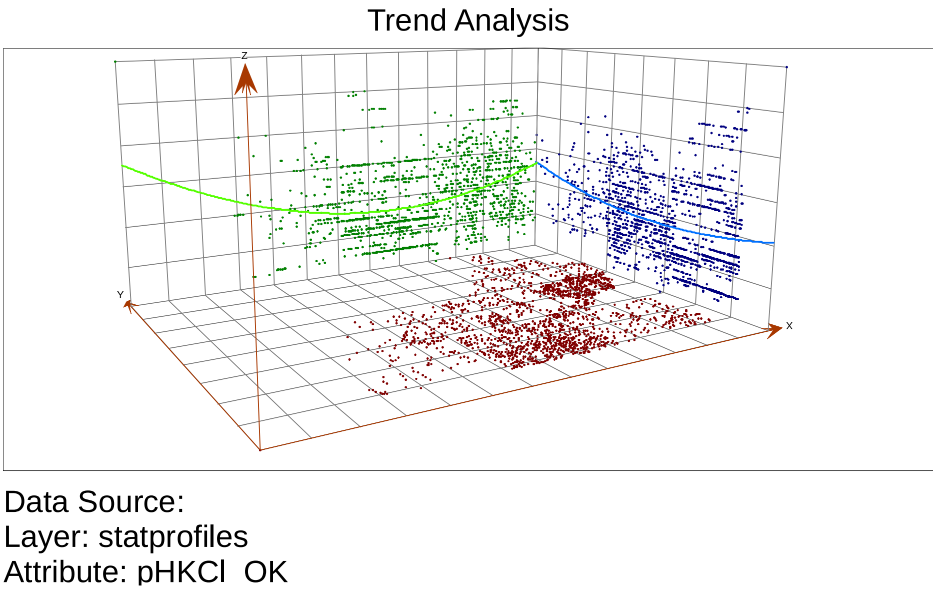

Figure 7. The trend analysis diagram for the pH. The colors indicate the different dimensions, red is the horizontal plain, blue is the North-South direction, while the green color is the East-West one. The red points show the horizontal distribution/location of the points, while the blue and green ones refer to the pH values along the NS and the EW directions respectively.

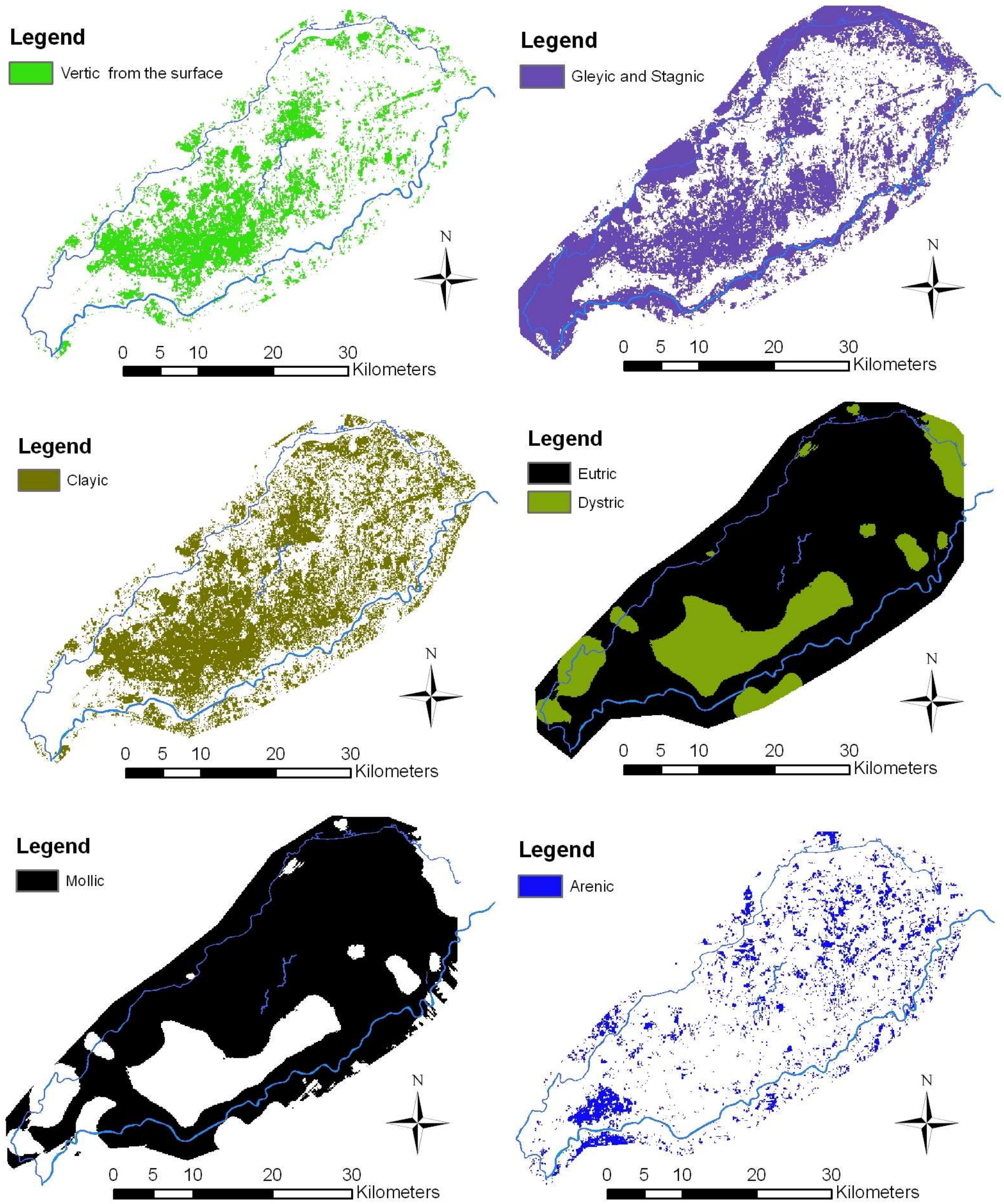

Figure 8. The WRB diagnostic properties and horizons.