

Legacy data integration into modern databases

E. Dobos¹, E. Micheli² and J. Kobza³

¹Associate Professor/ University of Miskolc, Hungary

² Professor/Szent István University, Hungary

³ Head of the Institute/Soil Science and Conservation Research Institute, Banska Bystrica, Slovakia

Driving forces

- EU/Globalization – need for harmonization
- INSPIRE
 - Environmental spatial data infrastructure
- *Soil Directive*
- Data needs for national use
- EU-wide database development
 - 1:1M scale soil database of Europe
 - 1:250K scale soil database of Europe
 - e-SOTER

The goals are ...

- to derive **common variable set** capable to answer certain questions on certain scales and **incorporating data from different sources** and characteristics
- to derive **methodology for database development** supporting the national and international data needs based on existing data sources / **harmonization** of the existing data

Input data sources vary in

- quality
- resolution/scale
- format
- time / age / temporal changes /
seasonal and longer periods

**NEEDS TO BE SCREENED (QA) AND
HARMONIZED**

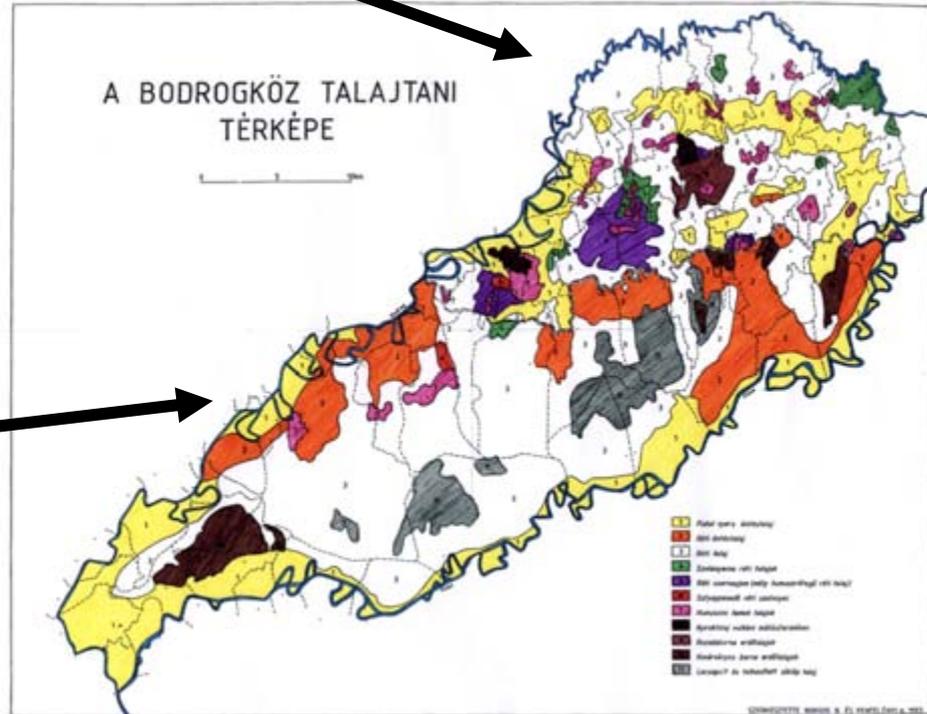
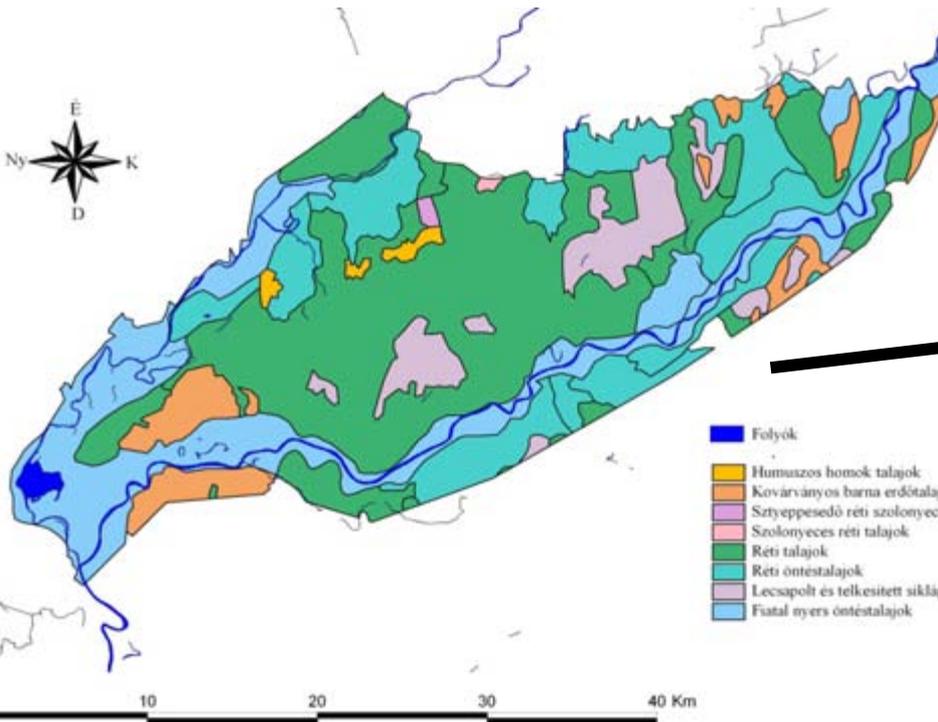
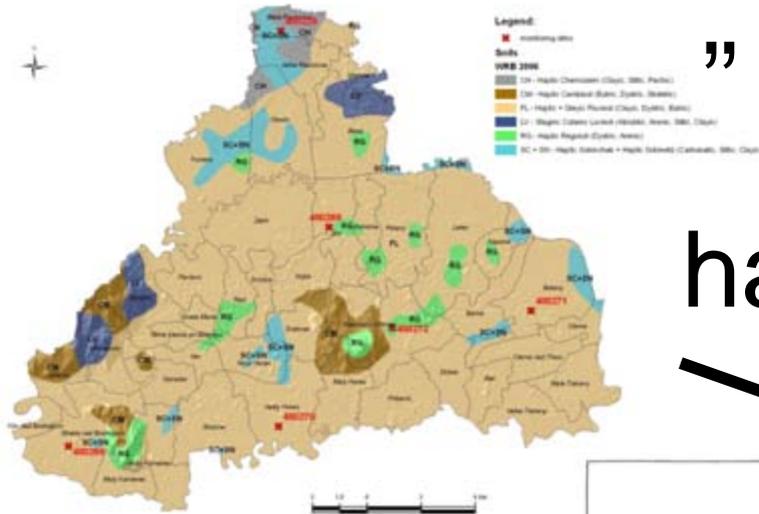
The problem of SEMANTIC data harmonization

- *non-matching class definitions represents major limitations (Lack of direct, national to national translation methodology)*
- *heterogenous, undefined national use of the international soil terms*
- *creation of “melting-pot classes” capable of handling variable class limits via the aggregation and generalization of the input semantic data classes.*

The problem of GEOMETRIC data harmonization

- *the integration of - often - unrecorded methodologies instead of data sources*
- *no procedure exists to redraw the polygons according to the spatial extent of „the common” class definitions (except the manual ones), only “cosmetics” of the non-matching polygons along the borders are possible.*

„Traditional ” data harmonization



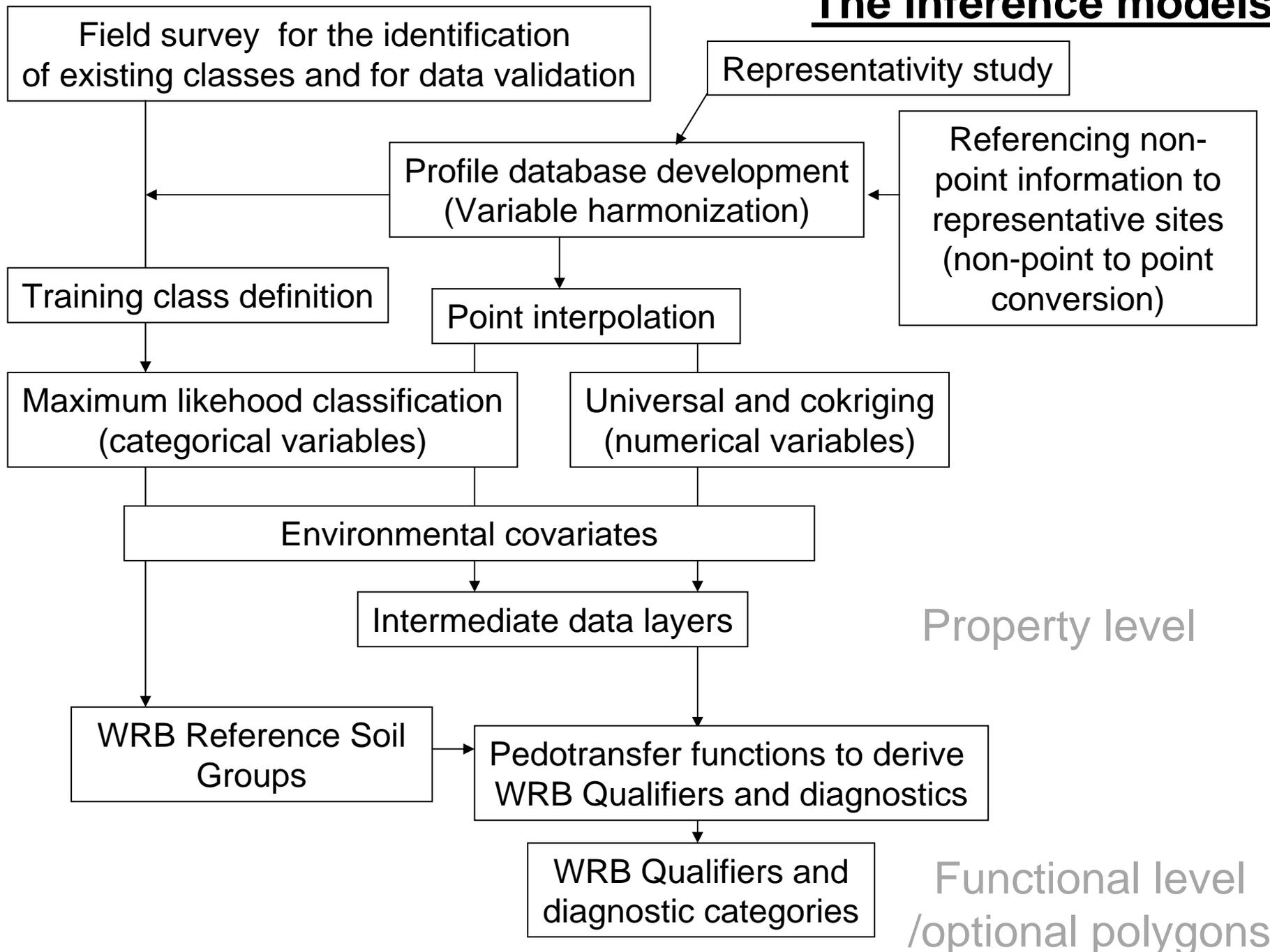
Geometric harmonization vs making use of the geometric data

- *Only manual harmonization procedures exist for the polygons, which require **tremendous amount of manual work**, practically the generation of entirely new database*
- *The majority of the available data is in polygon format. Despite of the difficulties of its interpretation, its geometric/locational information **is still of a high value**. There is no better on the market!*
- *The only way to develop a consistent polygon system is **to create a new one** using a common methodology. The question is how to define the polygon system: soil bodies vs. terrain/parent material units? Input semantic data has to be harmonized and transferred to the new polygons*
- *Another option is **the raster based approach** with thematic soil property layers. This requires DSM tools to employ, environmental covariates and harmonized training/calibration georeferenced soil data. This later data has to be extracted partly from polygon maps. Procedures can be developed!*

Suggested Methods

- Use an existing **common platform** (WRB)
- Use **raw profile data** with measured properties (if possible)
- **Derive point** information from non-point data sources like polygonal maps, for areas with less or no data
- Make use of the richness of historic, archived data via **integrating** the profile databases of different origin to increase the input data density
- Make use of **correlating soil data** and other environmental digital data sources explaining the spatial distribution of a soil property in question
- Make use of the **existing DSM tools** for extrapolating the soil information

The inference models



Pilot study...

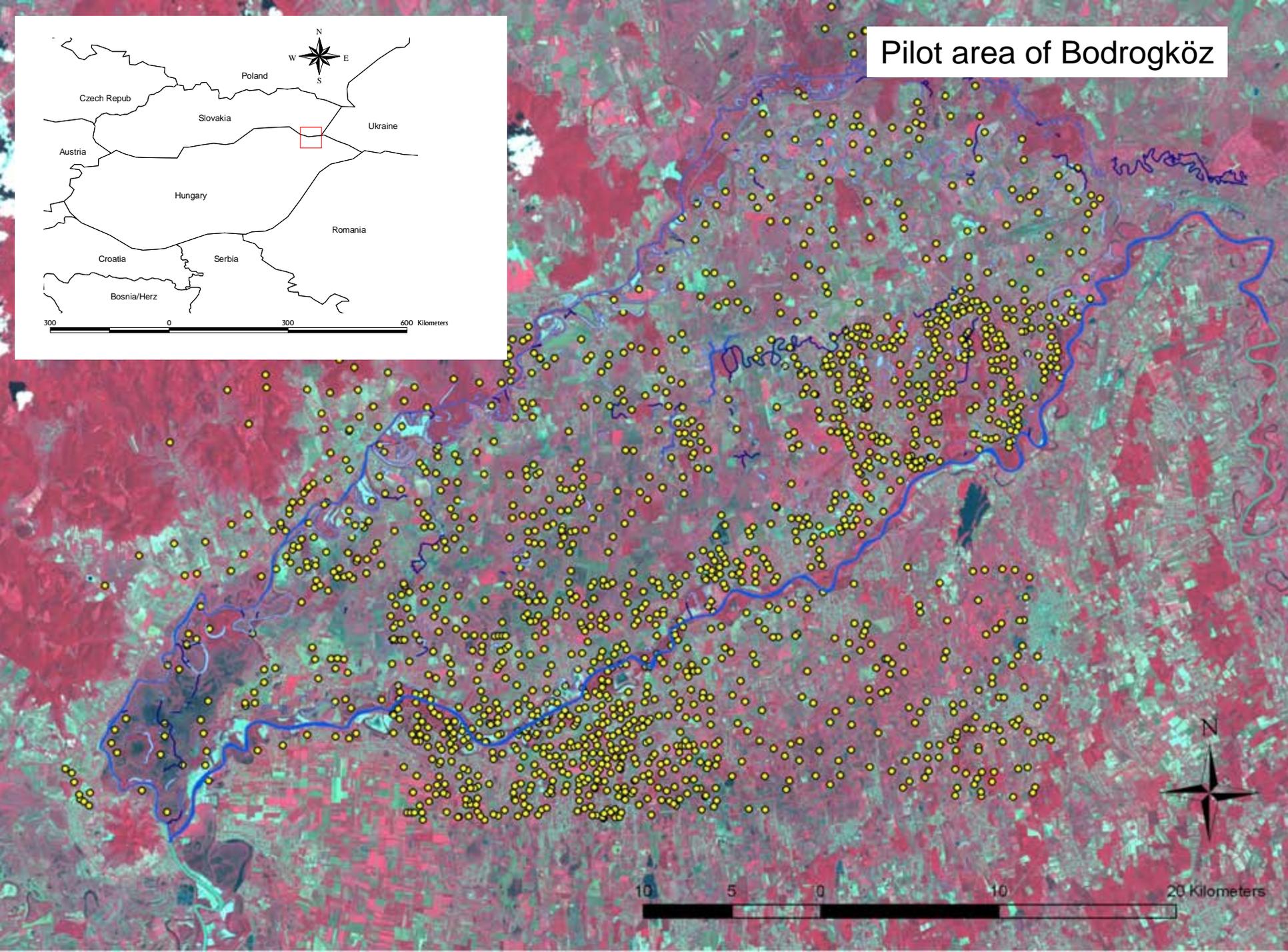
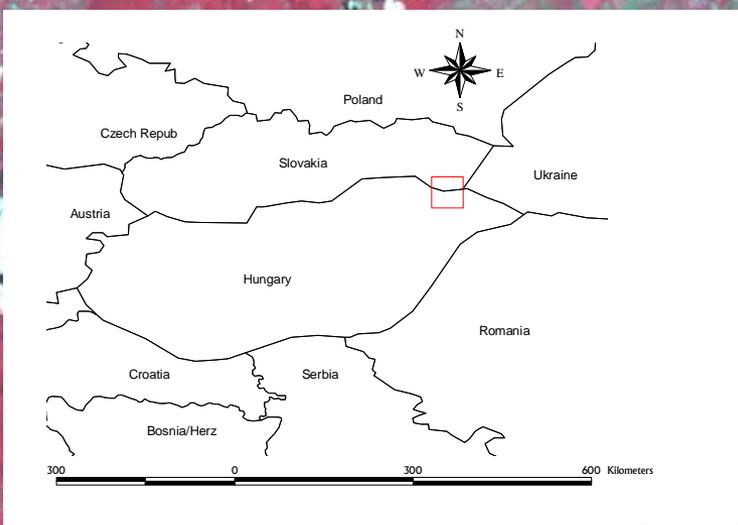
Cross-border pilot area between Hungary and Slovakia

- Characteristics of the area
 - Physiographically homogeneous area (Bodrogköz)
 - Alluvial plain with Arenosols, Regosols, Fluvisols, Gleysols and Vertisols
 - Size: 50 by 50 km

Soil data sources

- Soil monitoring sites of Hungary and Slovakia
- Slovakian soil survey sites
- Soil profiles of previous soil mapping campaigns (Kreybig, RISSAC)
- Nutrient survey campaign data from the 80's (agricultural plot based data)
- Our own field observation data

Pilot area of Bodrogköz



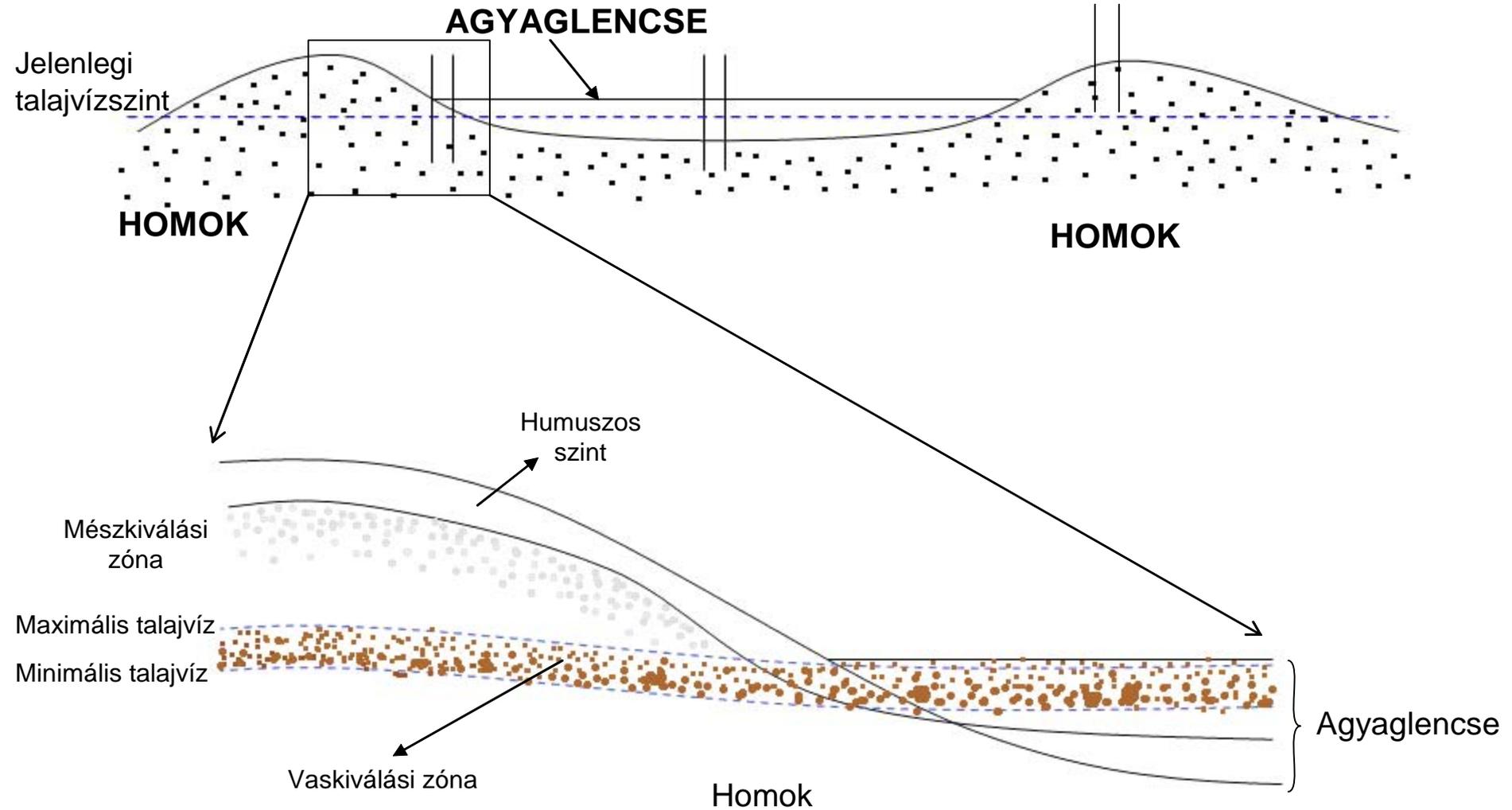
Other digital data sources used as environmental covariates

1. Digital orthophoto (mosaiced for the entire area)
2 meter pixel size (2005 HU, 2002 SK)
2. Landsat (30 meter pixel size)
July. 05. 2006.
March. 28. 1999. (flooded)
3. SPOT (20 meter pixel size)
May. 23. 2006.
October.13. 2006.
4. IKONOS (4 meter pixel size)
July. 23. 2007.
5. SRTM (90 meter resolution) and its terrain derivatives

Data harmonization is impossible without field calibration!!!



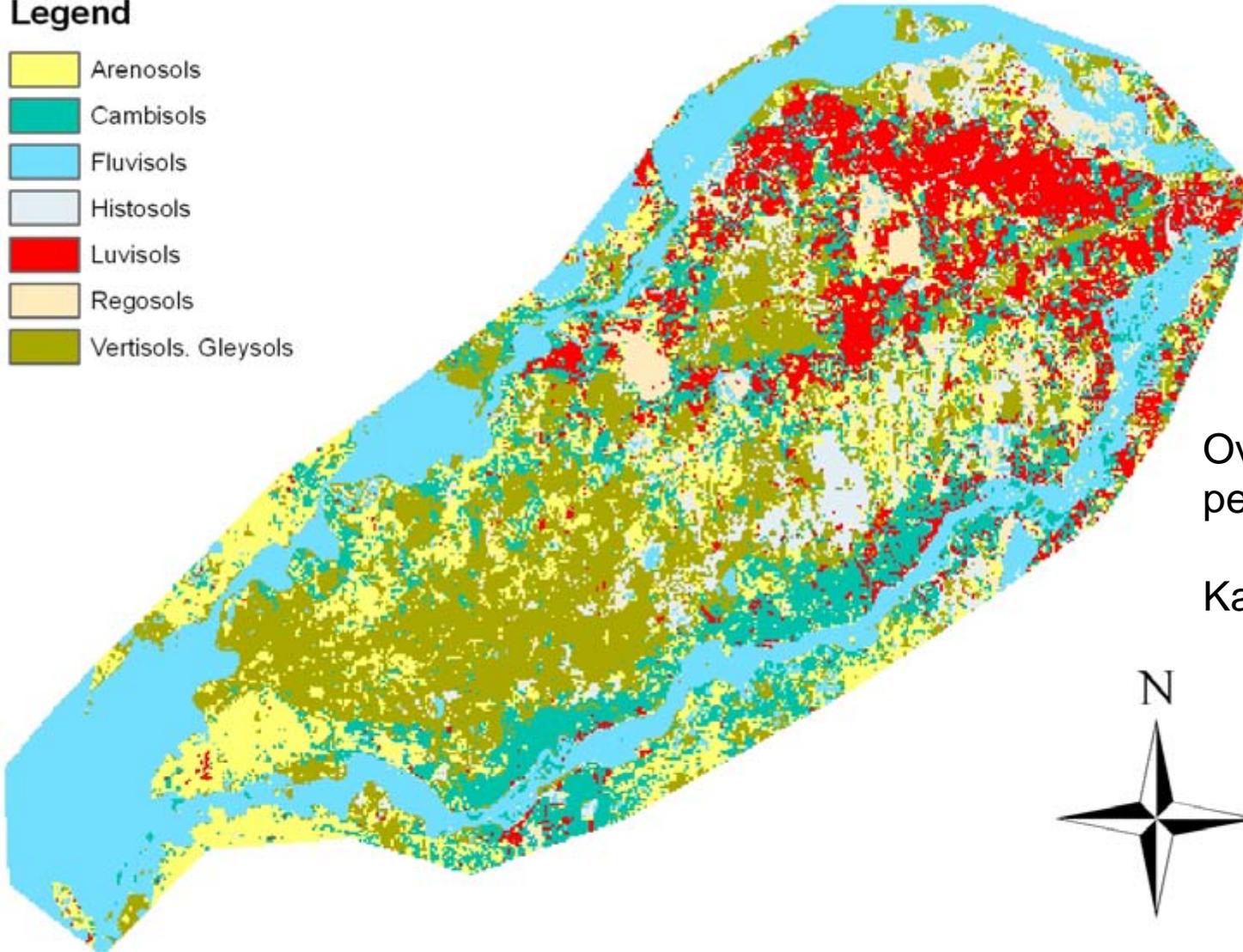
The „menthal model”



Property level...

WRB Reference Soil Groups of the Bodrogköz

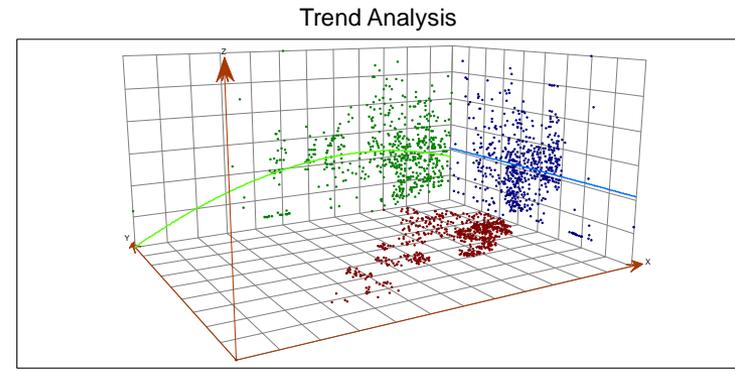
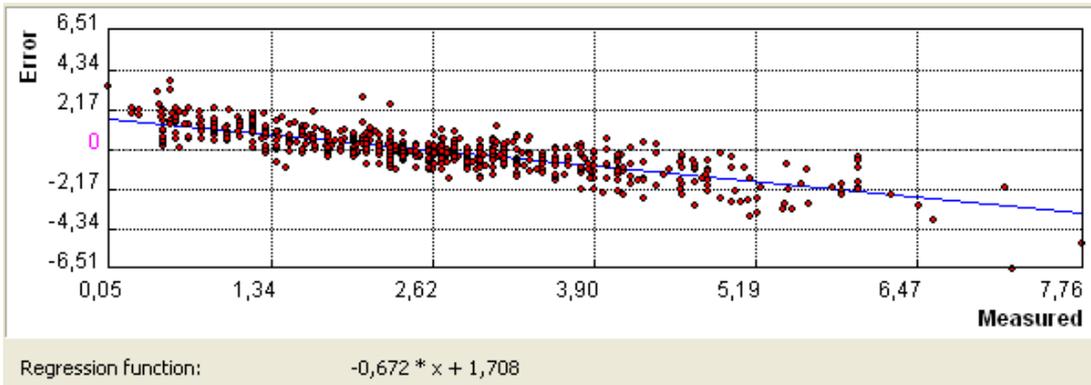
Legend



Overall classification
performance: 77%

Kappa: 0,7

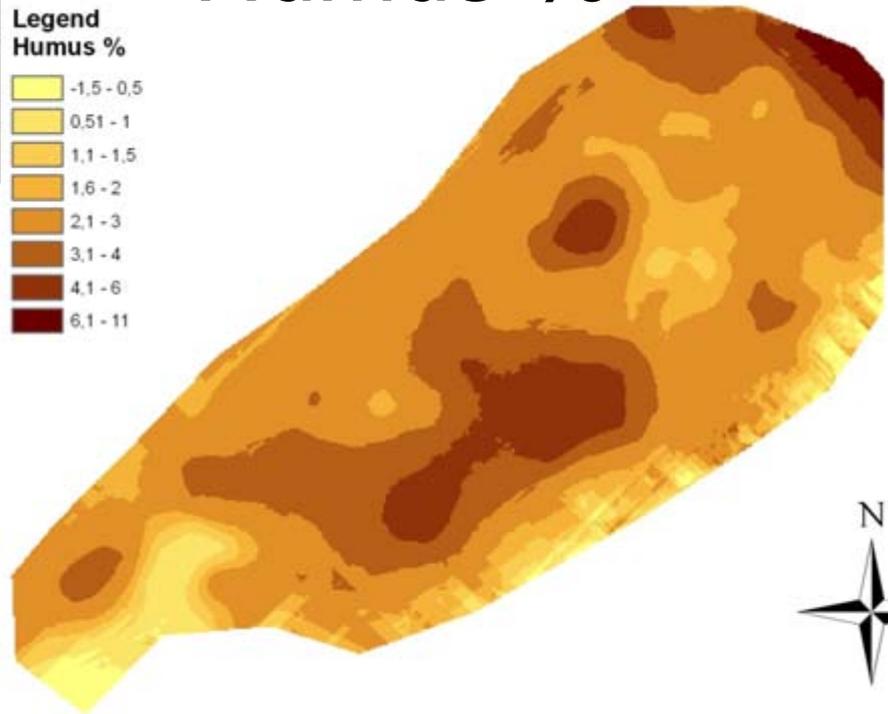
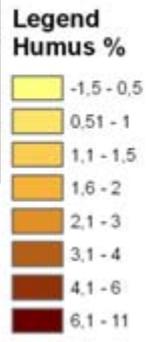




Data Source:
 Layer: statprofiles
 Attribute: HumuszOK

Geostatistical procedure	Universal cokriging with PDD
Number of observations	657
RMS	1,131
Average standard error	1,088
Standardised RMS	1,035

Humus %



Semivariogram | Covariance

Semivariogram/Covariance Surface

Show search direction

Angle direction: 0,0

Angle tolerance: 45,0

Bandwidth (lags): 3,0

Semivariogram/Covariances: Var1 & Var1

Modeling

Model: 1 Model: 2 Model: 3

Circular Spherical Tetraspherical Pentaspherical Exponential Gaussian Rational Quadratic Hole Effect K-Bessel J-Bessel Stable

Major range: 5500

Anisotropy

Minor range:

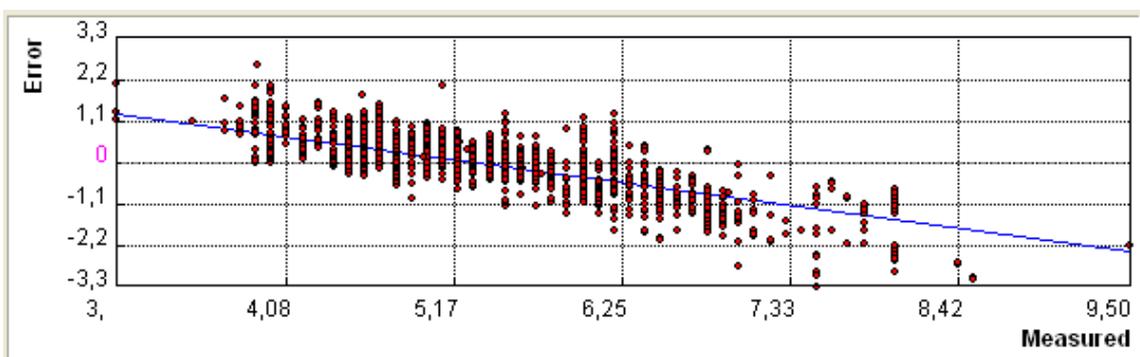
Direction:

Parameter: Partial sill: 0,65501

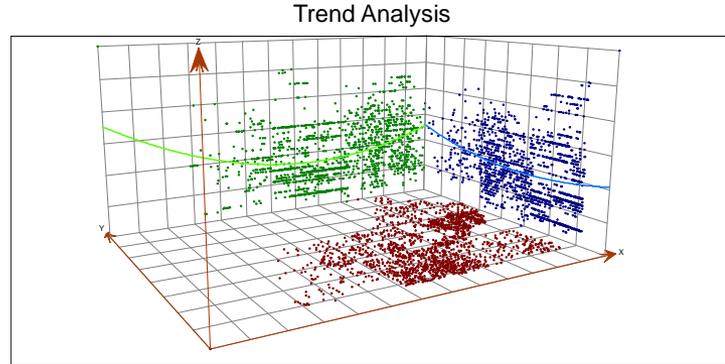
Nugget: 0,90451 Shifts

Lag size: 500 X: None Y: None

Number of lags: 20



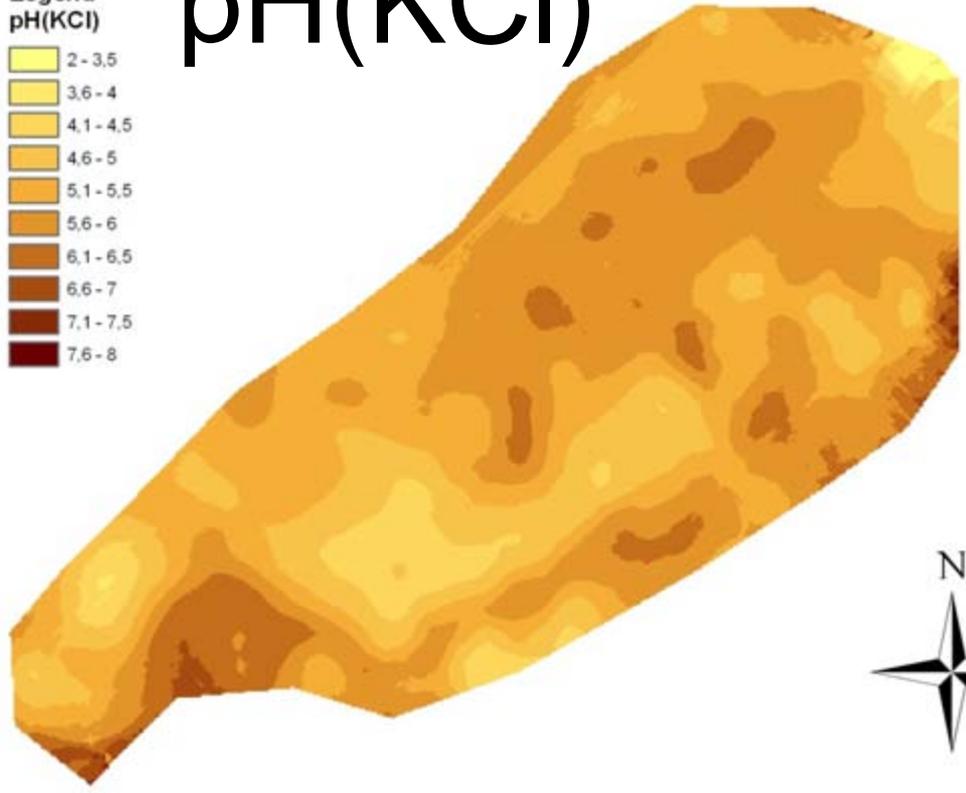
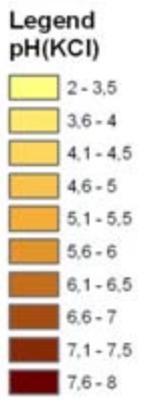
Regression function: $-0,562 * x + 2,980$



Data Source:
Layer: statprofiles
Attribute: pHKCI OK

Geostatistical procedure	Universal kriging
Number of observations	1611
RMS	0,7589
Average standard error	0,7767
Standardised RMS	0,983

pH(KCI)



Semivariogram | Covariance

Semivariogram/Covariance Surface

Show search direction

Angle direction: 0,0

Angle tolerance: 45,0

Bandwidth (lags): 3,0

Semivariogram/Covariances: Var1 & Var1

Model: 1 | Model: 2 | Model: 3

Circular
 Spherical
 Tetraspherical
 Pentaspherical
 Exponential
 Gaussian
 Rational Quadratic
 Hole Effect
 K-Bessel
 J-Bessel
 Stable

Major range: 5033,45

Anisotropy

Minor range: []

Direction: []

Parameter: [] Partial sill: 0,26379

Nugget: 0,4878

Shifts

Lag size: 753,03

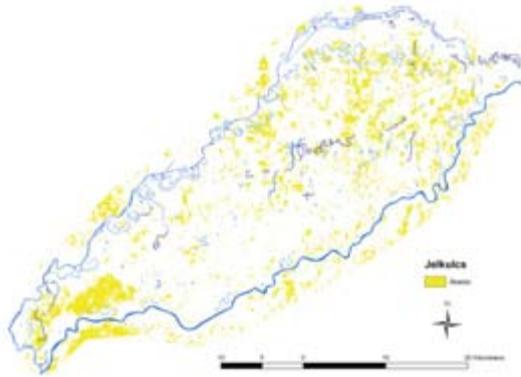
X: [] Y: []

Number of lags: 12

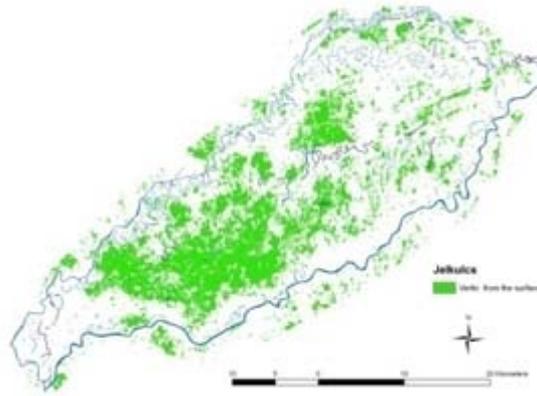
Functional level

The qualifiers/diagnostic categories and the pedotransfer functions used to derive them

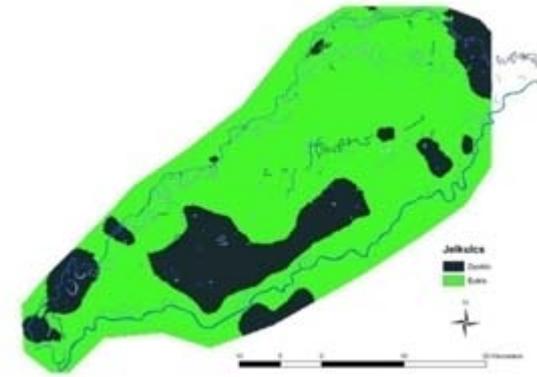
Qualifiers/Diagnostics	Pedotransfer functions
Vertic	Vertisols
Mollic	Humus>1% and Eutric
Arenic	Sand texture
Clayic	Clay texture
Gleyic	Vertisols, Fluvisols and Histosols
Dystric	pH(KCl)<5
Eutric	pH(KCl)>5
Calcic	CaCO ₃ % > 5



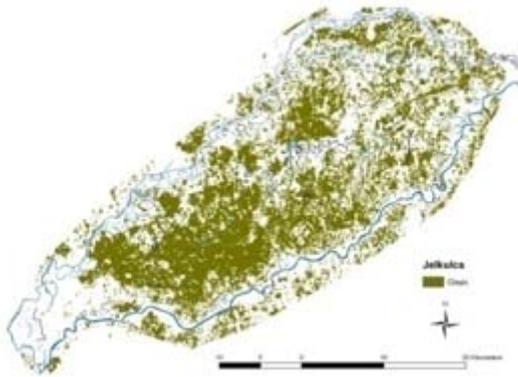
Arenic



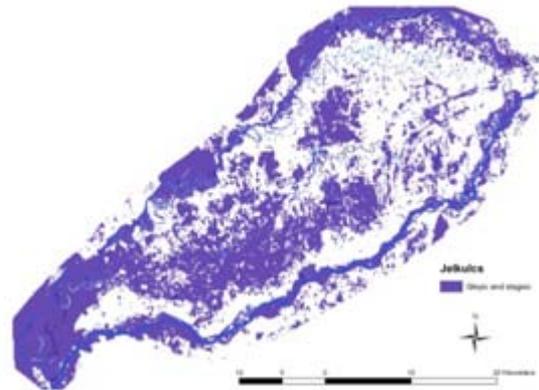
Vertic



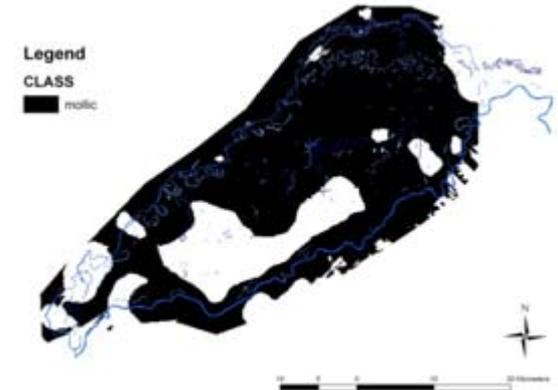
Dystric-Eutric



Clayic



Gleyic-Stagnic



Mollic

WRB qualifiers and diagnostic categories supporting environmental legislation

Conclusions

- Different data sources have different quality measures. Screening and „data-tuning” is crucial for the harmonization procedure. Importance of field harmonization and data validation!!!
- Thematic resolution of soil databases can be increased with increasing the input data density via integration of different sources
- Non-point, but spatially referenced information can be transformed to point data for data densification
- Importance of metadatabase
- WRB diagnostics and qualifiers provide appropriate basis for common variable platform



**Thank You for
your attention!**