Using machine learning to update soil survey

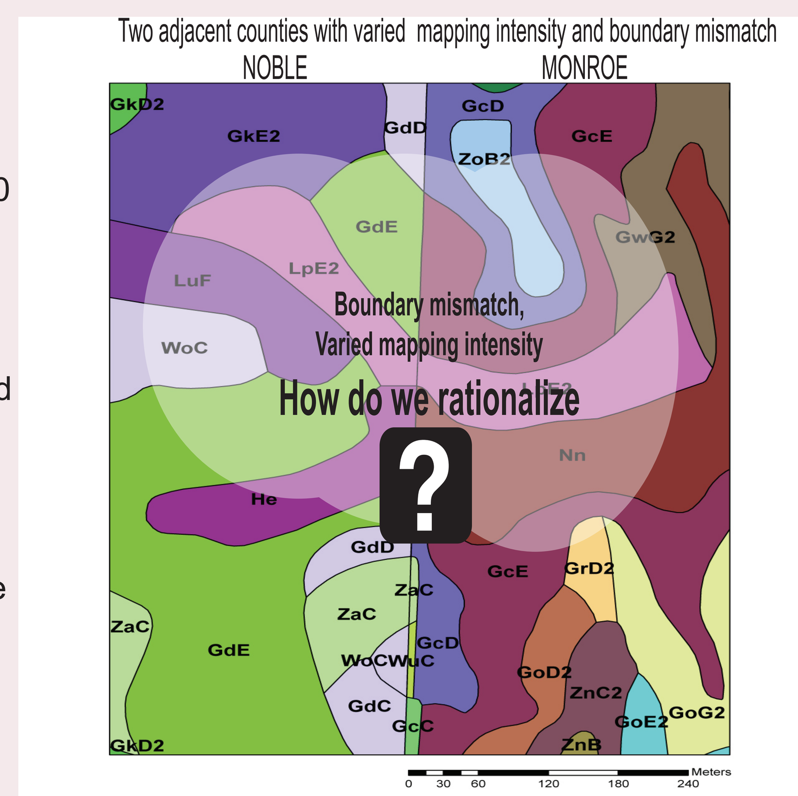
Sakthi K Subburayalu
School of Environment and Natural Resources

Conclusions

- Machine learning can be used to map component soil series within a map unit. Mapped components are clearly more closely related to the geography and the environment than impure mapunits.
- Soil maps derived by machine learning can serve as a guide to update the existent soil survey, and to address inaccuracies such as cross county mismatches, inconsistent mapping intensity and map errors.
- The prediction accuracy of component soil series was greatly affected by noise associated with the existent soil survey and the predictor variables, thus prepocessing of the dense geospatial data used for learning is critical.
- Random Forest algorithm promises to be a useful classifier for soil-landscape modeling involving complex environmental correlates.

Introduction

- The National Cooperative Soil Survey has served as a vital resource for land use planning and management for more than 100 years.
- The quality of soil information is critical for many uses and demands continual improvement.
- Discontinuities across county boundaries and varied level of mapping intensities pose a serious problem to seamless soil resource inventory on a Major Land Resource Area basis.
- Spatial and thematic inaccuracies reduce the reliability of soil information
- Machine learning when coupled with GIS offers potential for efficient analysis and effective updating of current soil survey information.



Machine learning framework
An overview of Materials and Methods

Data assembling
- DEM Generation and Terrain Analysis
- Bedrock Geology and Historic Vegetation from ODNR
- Precipitation and Evapotranspiration layer generation using NCDC's 30 year normal.
- SSURGO Digital Soils Data

Preprocessing
- Removal of noisy data such as missing data values and outliers

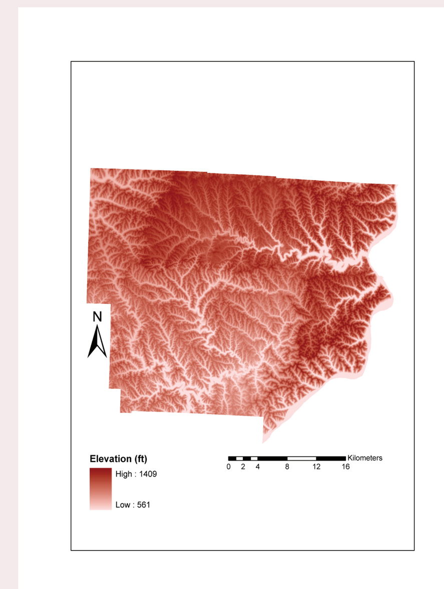
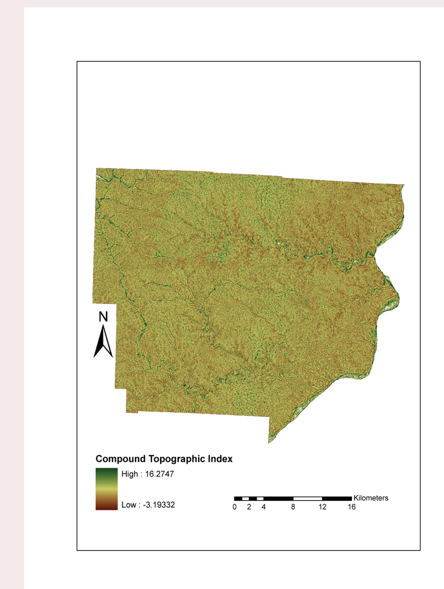
Learning
- Data mining using J48 and Random forest classifiers

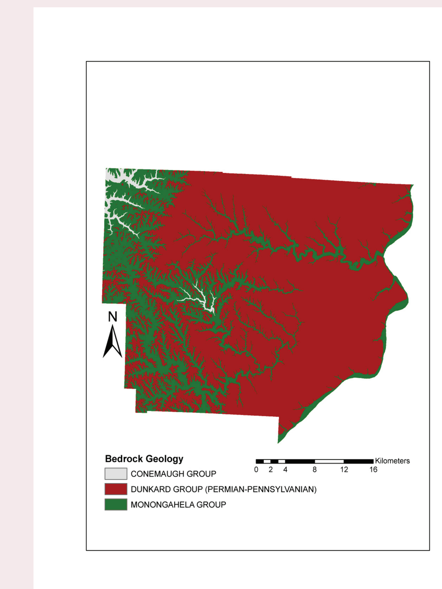
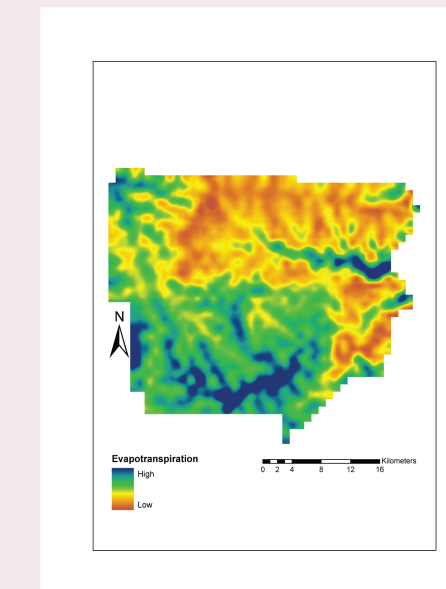
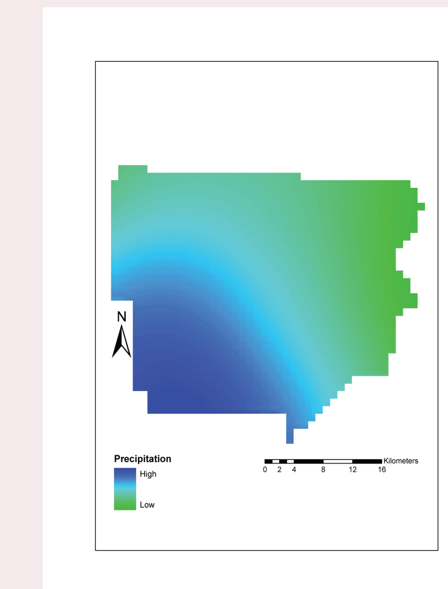
Validating
- Map Predicition

Reviewing

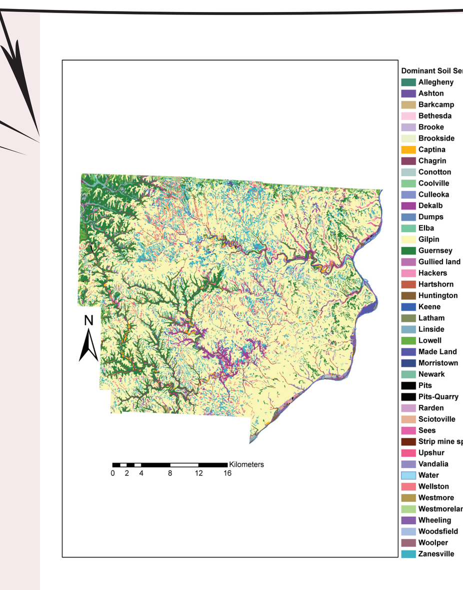
A Case study on Monroe County - Results and Discussion

Data Assembly

Soil = f (terrain, vegetation, geology, Climate) based on SCORPAN approach. (McBratney et al. 2003)

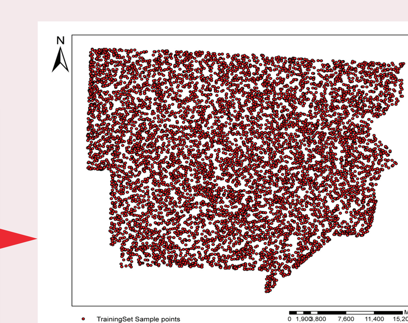


- A 10m DEM was generated using USGS contour lines and NHDC stream line data
- 24 different terrain attributes including slope, aspect, relative landscape position and other secondary derivatives were calculated from the DEM
- Historic vegetation layer was obtained from ODNR

- Bedrock geology map was obtained from ODNR
- Evapotranspiration and Precipitation map was generated using NCDC 30 year climate normal
- The base soil layer is the SSURGO data from NRCS

Preprocessing

- A spatially constrained random sampling of the soil map was done proportional to the area of the different soil series, by excluding samples near map unit boundaries where the uncertainity of a soil class could be high.
- Missing data values and outliers were removed using Hampel outlier indentifier (Davies et al., 1993)

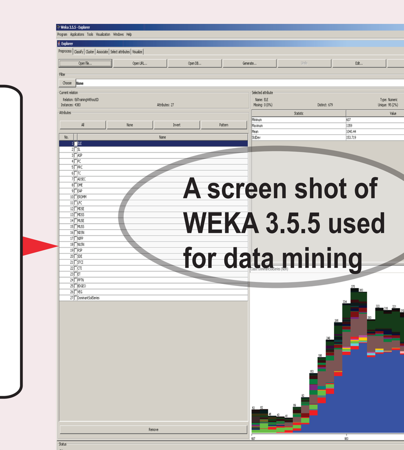


Preprocessed training set

Learning

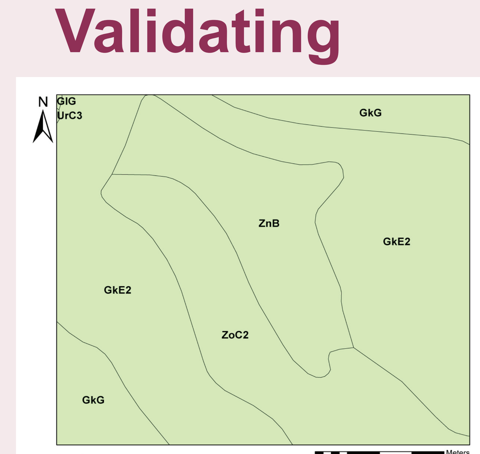
Two machine learning algorithms were applied to train the dataset and the built model was then used to generate a prediction map of Monroe County.

- Random forest (RF) algorithm (Brieman, 2001) which combines multiple trees built on bootstrap samples of training data and arrives at a aggregated tree model
- J48, a WEKA version of Quinlan's C 4.5 decision tree algorithm.

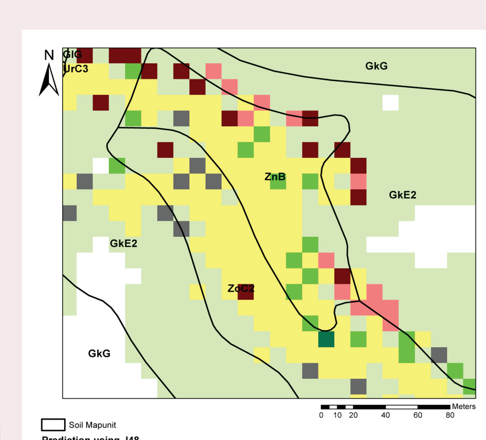


A screen shot of WEKA 3.5.5 used for data mining

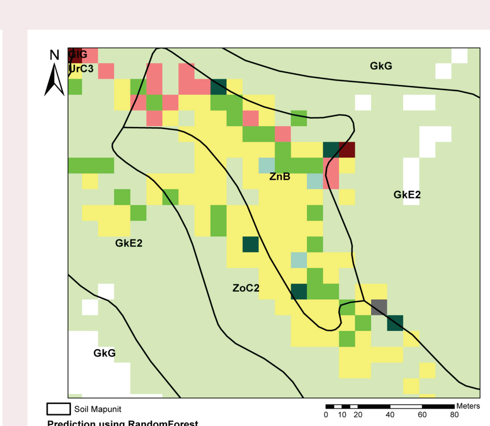
Validating



SSURGO Map
Impure map units show conso-ciations and associations but dont reveal position of components



Prediction using J48
Maps components within mapunits but with some spatial incoherency



Prediction using RF
Components predicted with greater spatial coherency

Reviewing

- The prediciton accuracy for J48 was slightly lower (about 50%) than that of RF (about 60%).However it should be noted that the error rate is a measure based on assuming that the samples belonged to the dominant component series only. It is likely that the model is often predicting minor component soil series correctly, in which case the actual accuracy would be greater than the measured one.
- The predicted map shows that it is possible to locate the position of component soil series within a mapunit. For example, the mapunit ZnB on the map to the left is Zanesville, 2 to 6 percent slopes, moderately eroded.A minor component soil series of this map unit is Keene. The predicted map shows that the most likely position of Keene is in the Northwest corner of this polygon.
- The predicted map can also be used to check the spatial accuracy of the map unit boundaries. For example, the map on left shows that it is possible that the boundary of ZnB is slightly shifted in XY direction as evident from the dominant soil series Zanesville of ZnB, supported by the presence of Wellston which is a component of the adjacent mapunit GkE2

References and further readings

- Moran, C. J., and Bui E.N., 2002. Spatial data mining for enhanced soil map modelling. *Int. J. Geographical Information Science*,16(6) 533-549.
- McBratney,A.B., Mendonca Santos,M.L., and Minasny, B., 2003. On digital soil mapping. *Geoderma*, 117, 3–52
- Davies,L., and Gather, U., 1993. The identification of multiple outliers. *Joual of the American Statistical Association*, 88(423), 782-792
- Brieman, L., 2001. Random Forests. *Machine learning*, 45, 5-32

Acknowledgement

The author thanks his faculty advisor Dr.Brian Slater for his valuable support and guidance throughout this study

Abstract

Soil survey in the recent past seems to be taking a paradigm shift with the advent of various geospatial and pedometric techniques. The impetus for this includes both the usability and limitations associated with the traditional soil survey products. This reseacrh explores the use of machine learning and GIS tools for updating an existing soil surveyof Monroe county in southeastern Ohio. A soil landscape modeling framework was adopted to predict soil series based on a number of high-resolution geospatial environmental correlates. Base data layers included the existent soil survey, climate attribute surfaces (precipitation and evapotranspiration), historic vegetation, terrain attributes derived from digital elevation model and bedrock geology. The old soil survey was randomly sampled to generate pre-classified training set containing target soil series and thier environmental correlates. Two machine learning algorithms (J48 classifier and Random forest classifier) were used to build classification models. The built models were then applied to the entire county to generate digital soil maps. The models predicted the correct dominant soil series about 60 percent of the time. When compared with the existent soil survey map, the digital soil map was able to predict even the components with in a soil mapunit. Machine learning can efficiently be used to get new insights into the traditional soil maps and can be used as a guide for further field investigations.