Digital Harmonisation of Adjacent Analogue Soil Survey areas - 4 Iowa Counties

Sun Wei[1,2], Alex McBratney [1*], Jon Hempel [3], Budiman Minasny [4], Brendan Malone [5], Tom D'Avello[6], Lee Burras [7], Jim Thompson[8]

[1*]The University of Sydney, Sydney, Australia, Email alex.mcbratney@sydney.edu.au
[2] China Agricultural University, Beijing, China. Email: sunwei1028@126.com
[3]USDA-NRCS-National Soil Survey Center, Lincoln, NE, Email jon.hempel@lin.usda.gov
[4]The University of Sydney, Sydney, Australia, Email budiman.minasny@sydney.edu.au
[5]The University of Sydney, Sydney, Australia, Email brendan.malone@sydney.edu.au
[6]USDA-NRCS-National Soil Survey Center, Morgantown, WV, Email tom.d'avello@wv.usda.gov
[7]Iowa State University, Ames, Iowa lburras@iastate.edu
[8]West Virginia University, Morgantown, West Virginia Email james.thompson@mail.wvu.edu

*corresponding author

Abstract

In conventional polygon-based mapping in the United States, it is often found that concepts across geopolitical boundaries (i. e. counties and states) don't match; this mismatch may include both lines and named map units. Of course, soils in the field do not change at these political boundaries. These soil-to-soil mismatches are the result of the past structuring of the soil survey program. For many years, soil surveys in the US were conducted on a soil survey area basis. Most times the soil surveys areas were based on county boundaries. Often adjacent counties were mapped years apart. Different personnel, different philosophies of soil survey science, new concepts of mapping and the availability of various technologies all play a part in why these differences occur. These differences maybe even more exaggerated at state lines as each state was administratively and technically responsible for the soil survey program within a given state.

We looked at various numeric approaches to harmonise concepts for adjacent map soil survey areas (4 adjoining counties in southern Iowa) that will produce raster component maps with a probabilities of series at each point. An example is given for four adjacent county soil survey areas in southern Iowa.

1. Introduction

In conventional polygon-based mapping, it is often found that concepts across geopolitical boundaries don't match; this mismatch may include both lines and named map units (Dobos et al., 2010). Of course, soils in the field do not change at these political boundaries. These soil-to-soil mismatches are the result of the past structuring of the soil survey program. For many years, soil surveys in the US were conducted on a soil survey area basis. Most times the soil surveys areas were based on county boundaries. Often adjacent counties were mapped years apart. These differences become even more exaggerated at state lines as the administrative and technical responsibility for the soil survey program was handled by each state individually.

These convention polygon based surveys delineate the landscape into map units. The map units are made up of components (e.g. soil series) that occur in predictable patterns in the landscape. Some map units represent areas dominated by a single soil series. However, most map units represent areas that are dominated by two or more soil series.

Digital soil mapping has been used to update conventional soil maps (Kempen et al., 2009), and here we looked at a digital approach to harmonise concepts for adjacent map soil survey areas (4 adjoining counties in southern Iowa) that will produce raster component maps with a probabilities of series at each point.

2. Methods

2.1 Study area and data

The study area is four adjacent counties in southern Iowa (Table 1). Each county was surveyed in different years by different surveyors. The map units are different for each map, and if we combine the 4 maps mismatches become apparent (Figure 1). It is also important to note that the concepts behind the mismatches cause inconsistencies beyond the actual border of the county. The north western quadrat is Monroe, (Oelman, D.B., 1984), the upper north eastern is Wapello, (Seaholm, J.E., 1981), south western is Appanoose (Lockridge, L.D., 1977), and south eastern is Davis (Lucassen, J.A., 1991). Different personnel, different philosophies of soil survey science, new concepts of mapping and the availability of various technologies all play a part in why these differences occur.

Table 1. Four adjacent counties in Iowa with different soil surveys.

County	Year published	No. map-units	No. soil series	Area(km^2)	No. pixels
Appanoose	1977	152	70	1355	12 775 998
Monroe	1984	183	73	1127	11 158 728
Wapello	1981	172	71	1130	11 111 687
Davis	1991	194	76	1321	12 899 906
Total			189		

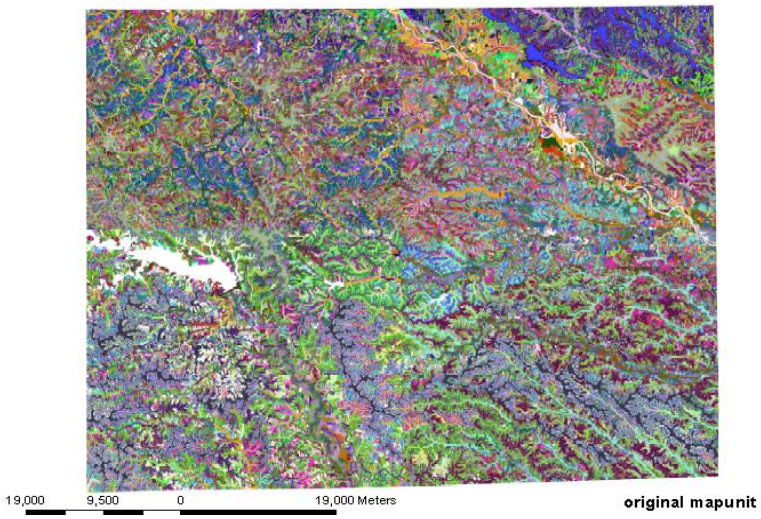


Fig. 1 A merged soil map from four adjacent counties in Iowa. (Mismatches are evident along county boundaries.)

2.2 General framework

In this paper, we will use an iterative numerical approach to harmonise the four maps. We will harmonise the maps by predicting the soil series at each pixel. A schematic diagram showing the generic framework is given in Fig.1. To achieve this task, we programmed an algorithm in C++, linking the sampling to the See5 data-mining software. No high-level statistical program can handle such a big dataset.

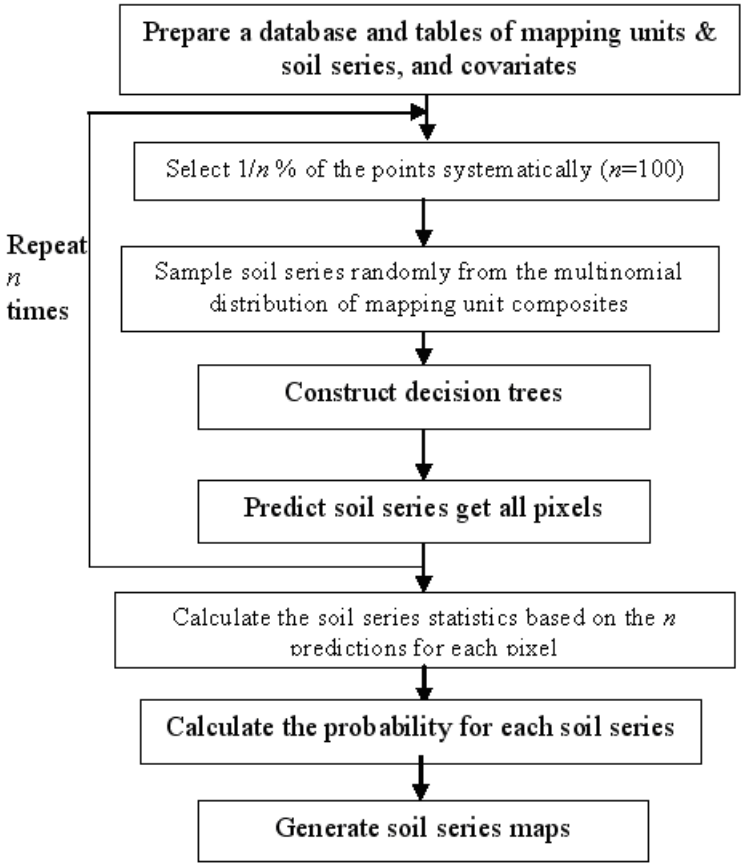


Fig.2 Flow diagram: framework for soil map harmonisation.

The steps for our digital harmonization techniques are as described as follows.

1. Prepare a database and table of mapping units & soil series.

The covariates used in this study are based on a US Geological Survey National Elevation Dataset (NED) 10 Meter Digital Elevation Model (DEM) for Appanoose, Davis, Monroe and Wapello counties. The datasets have a spatial resolution of 10 m. The county SSURGO databases (Soil Survey Staff, Appanoose County, Iowa), (Soil Survey Staff, Davis County, Iowa), (Soil Survey Staff, Monroe County, Iowa) and (Soil Survey Staff, Wapello County, Iowa) were converted to a raster coinciding with the DEM. The following covariates were calculated from DEM using the SAGA program: elevation (DEM), slope in degrees (SLOPE), length-slope factor (SLOPELEN), topographic wetness index (WETN), altitude above channel network (ACON), cos[aspect+90^o] (ASPECT). We also created a table of the mapunits and their soil series composites for each county. Each county is done separately because the same map unit in different counties consists of different soil components.

2. Sample the points systematically

Because of the high number of data points to process (a total of 48 million pixels), we selected 1% of the raster along with the corresponding covariates to be processed by the decision-tree program. We sampled 1% from the map systematically, and it is repeated 100 times to make sure every point in the map was covered and used in decision trees and prediction. Thus there are 100 training datasets. Because each mapping unit is composed of one or several soil series, we sampled the soil series using random sampling from the mapping unit multinomial distribution. This is done by generating a uniform random number between 0 and 1, and the soil series being selected based on the empirical distribution.

3. Generating decision trees and predicting the maps

We used See5 (RuleQuest Research, 2009), a sophisticated data-mining tool for discovering patterns that delineate categories, assembling them into classifiers, and using them to make predictions. See5 was implemented to execute decision trees and predictions with our data.

4. Statistics for each pixel

We generated 100 maps, the maps were predicted from different sampling realizations and different decision trees. Here we summarised the results for each pixel as showed in Fig.3, we calculated the multinomial distribution of soil series, allowing us to estimate the dominant soil series and its probability. We can also calculate the probability of occurrence for each soil series.

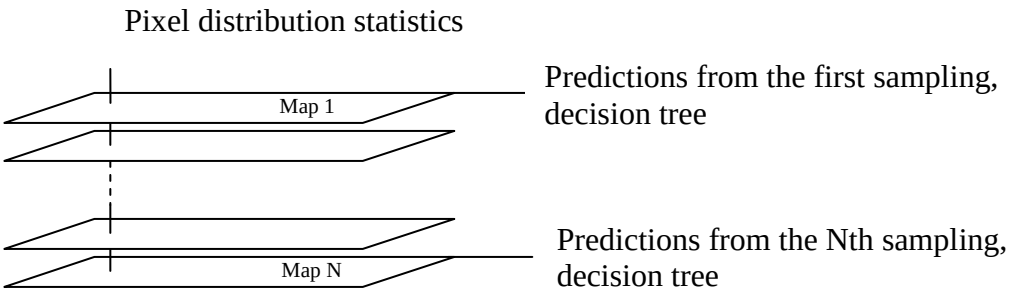


Fig.3 The diagram of the pixel statistics.

3. Results and Discussion

3.1 Soil series prediction

Fig. 4 shows the most probable soil series in the study area along with its probability. The accuracy of the See5 decision tree in training of the soil series is on average 68%. This is similar to the results of map purity of conventional soil maps obtained in many studies. Here we are able to generalise the soil-landscape rules for the four counties. The decision tree was also unable to predict 8 soil series in the area (Ainsworth, Chelsea, Dickinson, Fayette, Klum, Lamont, Nordness, Rushville). These soil series have low probability and probably do not have strong topographic signatures that can be picked up by the decision tree.

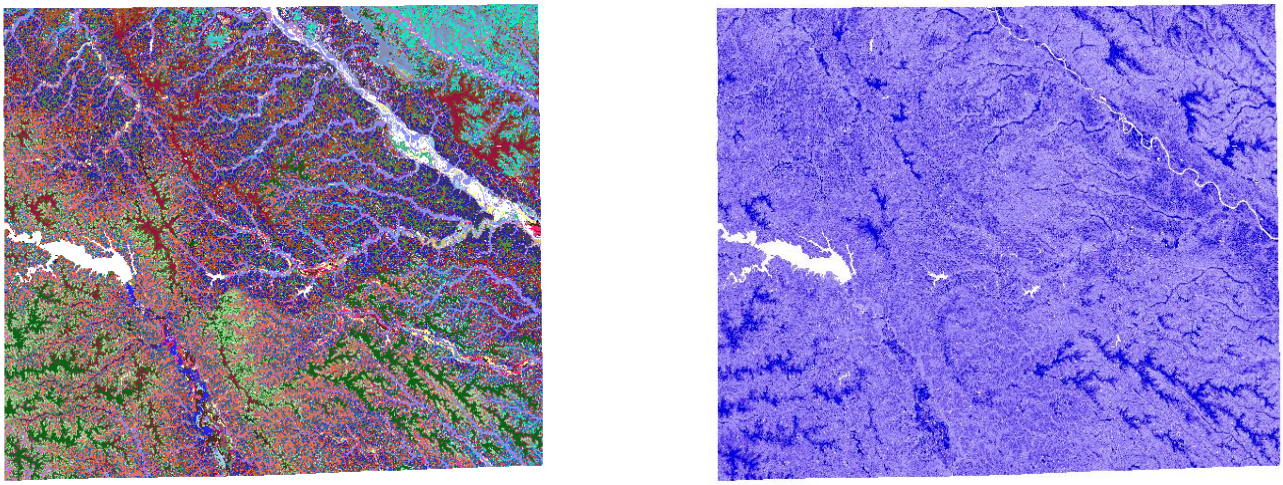
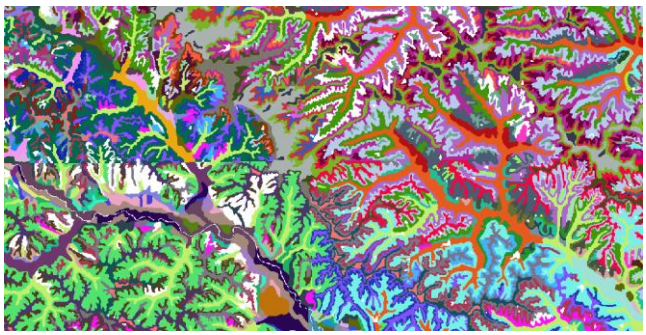


Fig. 4 Predicted most probable soil series, and its probability.

(a)



(b)

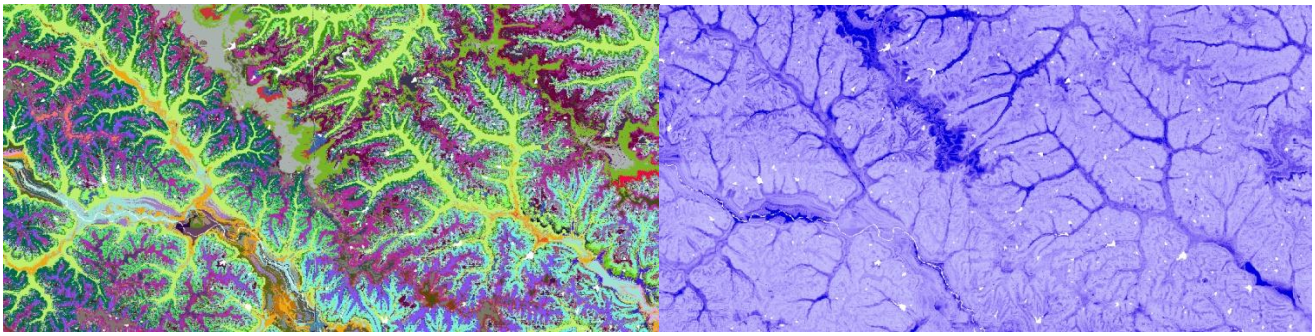


Figure 5. A closer look at the junction point in the middle of the 4 combined maps, (a) the original map units, and (b) the most likely soil series map and its associated probability. The length of the image is approximately 14 km.

3.3 Individual soil series

The advantage with this approach is that we are also able to create the probability of occurrence for each of the soil series. Figure 6 shows the example for the classes that occur most often in the area. As expected, each soil series occupies different positions in the landscape.

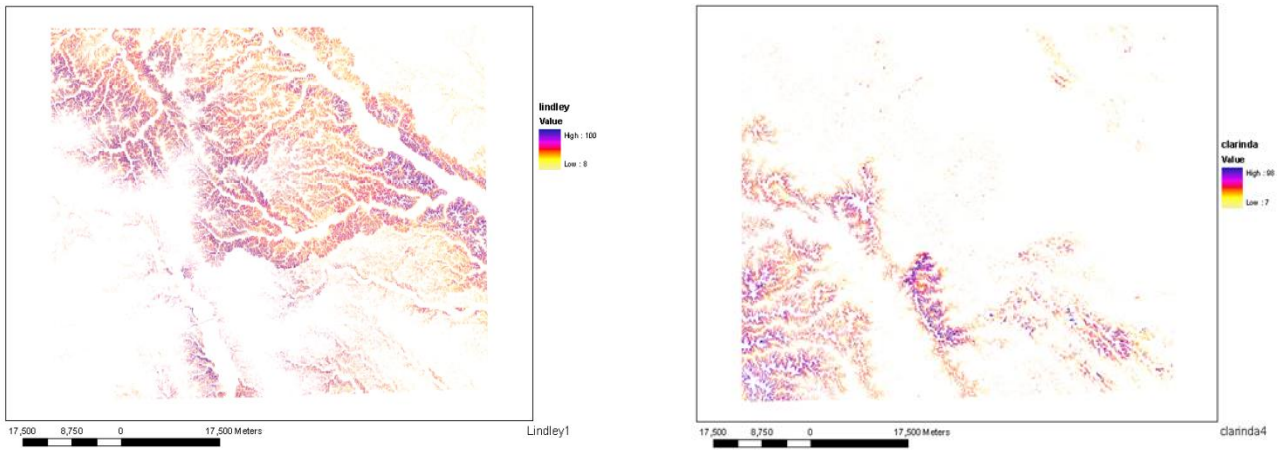


Fig. 6 Predicted probability of soil series occurrence.

3.4 Quality of prediction

At the time of writing this paper, we don't have enough validation samples to be able to show the quality of prediction. However, we recognise the limitations of this study and suggest ways of improving its accuracy and quality:

- Improvement using taxonomic distance.

Currently, we treat the soil series only as 'labels' and their prediction only considers the minimisation of the misclassification error. However soil series have taxonomic relationships between

each other, and in some instances the errors in prediction of certain series that are closely related are less serious than the others. Our previous study (Minasny and McBratney, 2007) showed that we can incorporate the taxonomic distance between soil series into a supervised classification routine.

- Improvement using expert knowledge

Since each of the maps was conducted by different surveyors, the compositions of the map units are also different in each county. Here, we used random sampling for the multinomial distribution of soil series composition to define a priori series at each pixel. However, in the real world the soil series would probably occupy certain parts of the map unit. Expert knowledge can be used to better define the a priori most probable soil series at each pixel, and this should improve the posterior probabilities.

4. Conclusions

A digital harmonisation algorithm is proposed and demonstrated in four adjacent counties in Iowa. Here we propose to map the individual soil series instead of mapping units. This has the advantage of creating soil landscape rules that can be applied in all of the areas. We can also create probability maps for the soil series. This is useful if we wish to make soil property maps from the soil series. The soil property at a particular pixel can be calculated as the weighted average of the soil series using the central concepts, where the posterior probabilities can be used as weights.

References

Dobos, E., Bialkó, T., Micheli, E., Kobza, J. 2010. Legacy Soil Data Harmonization and Database Development. J.L. Boettinger et al. (eds.), Digital Soil Mapping, Progress in Soil Science 2, Springer.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151, 311-326.

Lockridge, L.D., 1977. Soil Survey of Appanoose County, Iowa. U.S. Deparment of Agriculture, Soil Conservation Service in cooperation with Iowa Agriculture and Home Economics Experiment Station; Cooperative Extension Service, Iowa State University and Division of Soil Conservation, Department of Agriculture and Land Stewardship.

Lucassen, J.A., 1991. Soil Survey of Davis County, Iowa. U.S. Deparment of Agriculture, Soil Conservation Service in cooperation with Iowa Agriculture and Home Economics Experiment Station; Cooperative Extension Service, Iowa State University and the Division of Soil Conservation, Iowa Department of Agriculture and Land Stewardship.

Minasny, B, McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma 142, 285-293.

Oelman, D.B., 1984. Soil Survey of Monroe County, Iowa. U.S. Department of Agriculture, Soil Conservation Service in cooperation with Iowa Agriculture and Home Economics Experiment Station; the Cooperative Extension Service, Iowa State University and the Department of Soil Conservation, State of Iowa.

Rulequest Research, 2009. See5. Version 2.07. RuleQuest Research. St Ives, Australia.

Seaholm, J.E., 1981. Soil Survey of Wapello County, Iowa. U.S. Department of Agriculture, Soil Conservation Service in cooperation with Iowa Agriculture and Home Economics Experiment Station, Cooperative Extension Service, Iowa State University and Department of Soil Conservation, State of Iowa.

Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Soil Survey Geographic (SSURGO) Database for Appanoose County, Iowa. Available online at http://soildatamart.nrcs.usda.gov accessed 1/2010.

Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Soil Survey Geographic (SSURGO) Database for Davis County, Iowa. Available online at http://soildatamart.nrcs.usda.gov accessed1/2010.
Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Soil Survey Geographic (SSURGO) Database for Monroe County, Iowa. Available online at http://soildatamart.nrcs.usda.gov accessed 1/2010.
Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Soil Survey Geographic (SSURGO) Database for Wapello County, Iowa. Available online at http://soildatamart.nrcs.usda.gov accessed 1/2010.