Soil property mapping over large areas using sparse ad-hoc samples

A-Xing Zhu[A,B], Jing Liu[B,C], Chengzhi Qin[B], Shujie Zhang[B], Ya-ning Chen[D] and Xingwang Ma[E]

[A]Department of Geography, University of Wisconsin, Madison, Wisconsin, U.S.A.
[B]State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.
[C] School of Geography, Beijing Normal University, Beijing 100875, China.
[D]Key Laboratory of Oasis Ecology and Desert Environment, Xinjiang Institute of Ecology and Geography, Chinese Academy of Sciences, Urumqi, Xinjiang 830011, China.
[E]Institute of Soil and Fertilizer, Xinjiang Academy of Agricultural Sciences, Urumqi, Xinjiang 830000, China.

Abstract
This paper presents a new approach to predict soil properties and quantify uncertainty in the derived soil property maps over large areas using sparse and ad-hoc samples. According to the soil-landscape model, each soil sample contains corresponding relationships between soil and environment conditions. Under the assumption that the more similar the environment conditions between two locations the more similar the soil property values, each sample can be considered as a representative (individual representativeness) over areas of similar environmental conditions. The level of representativeness of an individual sample to an unsampled location can be approximated by the similarity in environmental conditions between the two locations. Based on this "individual representativeness" concept and with the use of Case-based Reasoning (CBR) idea, which solves new problems by referring to similar cases, soil property values at unsampled locations can be predicted based on their environmental similarity to the individual samples. Furthermore, the uncertainty associated with each prediction is related to the similarity and can thus be quantified. A case study located in Illy Region, Xinjiang, Northwest China, has demonstrated that the predicted map of soil organic matter of top layer is of good quality and the quantified uncertainty is positively correlated with prediction residual. This suggests that the approach can be an effective alternative for predicting soil property and reporting uncertainty in the resulting soil map over large areas with sparse and ad-hoc samples.

Key Words
Individual representativeness, soil-landscape model, case-based reasoning, digital soil mapping, uncertainty, SoLIM.

Introduction
Information on spatial variation of soil properties over large areas is a critical piece of input data for environmental modeling at the regional to continental scales (Abramopoulos *et al.* 1988; Bonan 1996; Dai and Zeng 1996; Chen and Dudhia 2001, Zhu and Mackay 2001). Yet, quality information on soil spatial variation over large areas is rather difficult to obtain due to the large number of field samples needed and the requirement of sound global representativeness imposed by the existing mapping techniques (Journel and Huijbregts 1978; Isaaks and Srivastava 1989; Cressie and Noel 1993; Goovaerts 1999; Mitas and Mitasova 1999; Schloeder *et al.* 2001; McBratney *et al.* 2003; Zhu *et al.* 2008). Due to the constraints of field conditions and project budget and the complexity of spatial variation of soil properties, field sampling can rarely meet these requirements (both the number of samples and the sound global representativeness). As a result, the collected samples are often sparse and ad-hoc (poor global representativeness) in nature. The soil property maps derived based on these samples using the existing mapping techniques are not only at low quality but also lack the information on the uncertainty introduced by samples' poor global representativeness. The lack of uncertainty information in the derived soil property maps also prevents proper uncertainty assessment of model outputs when the derived soil information is used as one of the inputs.

Methods
The approach is based on the concept of soil-landscape model (Jenney 1941; McBratney *et al.* 2000; McBrateney *et al.* 2003) which states that each sample contains certain corresponding relationship between soil and associated environmental conditions in parameter space. With the assumption that the more similar the environment conditions between two locations the more similar the soil property values, each sample can be considered as a representative over locations (not necessarily contiguous) with similar environmental conditions, that is, each sample owns "individual representativeness". The level of representativeness of an

individual sample for an unsampled location can be approximated by the similarity in environmental conditions between the two locations. Based on this "individual representativeness" concept and with the use of Case-based Reasoning (CBR) idea, which solves new problems just by identifying existing similar cases while not requiring the presence of a global model for the entire problem domain (Aamodt and Plaza 1994; Watson and Abdullah 1994; Leake 1996; Watson 1998), soil property values at unsampled locations can be predicted based on the environmental similarities to the individual samples. Moreover, the uncertainty associated with each prediction is related to the similarity. For example, if a location is not similar or at a low degree of similarity to the current set of individual samples, the uncertainty associated with the predicted value for that location is high because none of the existing samples is a good representative of this location. Then, the uncertainty associated with the prediction at each location can be quantified by analyzing the nature of the similarity values to the individual samples (Zhu 1997).

The new approach consists of three major components: 1) The selection of environment variables (covariates) and characterization of associated environment conditions using these variables; 2) Calculation of similarity in environmental conditions; 3) Estimation of soil property value and quantify uncertainty based on the environmental similarity.

For environment characterization, the selected environment variables should be responsible for soil formation or co-varying with soil closely so that they can be used to indicate spatial variation of soil effectively. The approach uses a raster data model for spatial representation. For soil mapping over large areas the grid size is often large. The characterization of environmental conditions over large grid size depends on the variable. For variables (climate and geology) which do not vary rapidly over the area of a pixel, we use one value to represent the environmental conditions at each pixel. For variables (such as topographic variables and vegetation variables) that vary rapidly over a pixel area we use the probability density function estimated using the Kernal Density Estimation (KDE) method to characterize the environmental conditions at each pixel.

Similarity estimation was conducted at two levels: the individual environment variable level and the case (sample) level which integrates all similarities from the individual variable level. The methods for the first level depend on the data type and the characterization method of each variable. We adopted Gower distance for measuring similarity in climate variables, Boolean function for parent materials, and a consistent Measure (CM) for topographic variables and vegetation variables which are characterized using probability density functions (Zhu 1999). The methods for the second level depend on the perception of interaction of environment variables. With the knowledge that over large area climate conditions would control the general spatial distribution pattern of soil, parent material would then differentiate soils in the same climate zone, while specific topographic conditions would influence the local variation in the same parent material area, we adopted a hierarchy approach in this research to integrate the similarities from individual variables.

For uncertainty quantification and soil property prediction, similarities at each location to individual samples would form a similarity-vector characterizing the representativeness of sample cases at that location. By analyzing this similarity vector, uncertainty associated with the prediction related to samples' representativeness was quantified (Zhu 1997). Soil property value at an unsampled location was predicted using a similarity weighted average method which integrates similarities with sample attributes. The result from this approach contains two parts: a soil property map and the associated uncertainty map.

Results

A case study located in Illy Region, Xinjiang, Northwest China, has been conducted to examine the validity of this approach. The study area is about 50,000 km^2 in size. The variables used are: average annual precipitation, average annual temperature, average annual relative humidity, maximum and minimum monthly precipitation, maximum and minimum monthly temperature, and maximum and minimum monthly relative humidity, parent materials; elevation, slope gradient, profile curvature, surface area ratio and land position index. A cross validation method with 73 field observation points was used to evaluate the performance of the method. The RMSE between the predicted and the observed values is 0.32 which is much smaller than 3.16, the standard deviation of these 73 field points. The correlation coefficient between the values of uncertainty and the prediction residuals at these points is 0.537 which is significant at the 0.05 level.

Conclusions

This paper presented a new approach to predict soil property over large area based on "individual representativeness" of sparse ad-hoc samples. This approach does not require the global representativeness of the whole sample set and is able to quantify prediction uncertainty introduced by the poor global representativeness of the sparse and ad-hoc samples. The results suggest that this approach is an effective and accurate way to map soil properties over large areas and is capable of providing uncertainty associated with the derived property map. The uncertainty information is a valuable piece of information for evaluating the credibility of prediction at each location. We conclude that this approach can serve as an effective alternative for predicting soil property and reporting prediction uncertainty over large areas with sparse and ad-hoc samples.

Acknowledgements

This study is supported by National Natural Science Foundation of China (No. 40601078), National Basic Research Program of China (No. 2007CB407207), and Key Projects in the National Science & Technology Pillar Program during the Eleventh Five-year Plan Period (2007BAC15B01-1).

References

Aamodt A, Plaza E (1994) Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications* **7**, 39-59.

Abramopoulos F, Rosenzweig C, Choudhury B (1988) Improved Ground Hydrology Calculations for Global Climate Models (GCMs): Soil Water Movement and Evapotranspiration. *Journal of Climate* **1**, 921-941.

Bonan GB (1996) A land surface model (LSM version 1.0) for ecological, hydrological, and atmospheric studies: technical description and user's guide. NCAR Tech. Note NCAR/TN-417+STR, 150 p.

Chen F, Dudhia J (2001) Coupling an Advanced Land Surface–Hydrology Model with the Penn State–NCAR MM5 Modeling System. Part I: Model Implementation and Sensitivity. *American Meteorological Society* **129**, 569-585.

Cressie N, Noel AC (1993) Statistics for Spatial Data. John Wiley & Sons, Inc, New York, 900 p.

Dai Y, Zeng QC (1996) A land surface model (IAP94) for climate studies. Part I: Formulation and validation in off-line experiments. *Advances in Atmosphere Science* **14**, 433-460.

Goovaerts P (1999) Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma* **89**, 1-45.

Isaaks EH, Srivastava RM (1989) Applied Geostatistics. Oxford University Press, New York, 561p.

Jenny H (1941) Factors of Soil Formation: A System of Quantitative Pedology. Dover Publ., New York, 281p.

Journel AG, Huijbregts CJ (1978) Mining Geostatistics. Academic Press, London, US, 600 p

Leake DB (1996) Case-based reasoning. *The knowledge engineering review* **9**, 61-64.

McBratney AB, Odeh IOA, Bishop T, Dunbar M, Shatar T (2000) An overview of pedometric techniques of use in soil survey. *Geoderma* **97**, 293-327.

McBratney AB, Mendonça Santos M, Minasny B 2003. On digital soil mapping. *Geoderma* **117**, 3-52.

Mitas L, Mitasova H (1999) Spatial Interpolation. In 'Geographical Information Systems: Principles, Techniques, Management and Applications'. (Eds P Longley, MF Goodchild, DJ Maguire, DW Rhind) pp. 481-492. (GeoInformation International: New York).

Schloeder CA, Zimmerman NE, Jacobs MJ (2001). Comparison of methods for interpolating soil properties using limited data. *Soil Sci. Soc. Am. J.* **65**, 470-479.

Watson ID, Abdullah S (1994) Developing Case-Based Reasoning Systems: A Case Study in Diagnosing Building Defects. In, Proc. IEEE Colloquium on Case-Based Reasoning: Prospects for Applications, Digest No: 1994/057, 1/1-1/3.

Watson I.D (1998) Applying Case-Based Reasoning: techniques for enterprise systems. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA.

Zhu AX (1997) Measuring uncertainty in class assignment for natural resource maps using a similarity model. *Photogrammetric Engineering & Remote Sensing* **63**, 1195-1202.

Zhu AX (1999) A personal construct-based knowledge acquisition process for natural resource mapping using GIS. *International Journal of Geographic Information Science* **13**, 119-141.

Zhu AX, Mackay DS (2001) Effects of spatial detail of soil information on watershed modeling. *Journal of Hydrology*, **248**, 54-77.

Zhu AX, Yang L, Li BL, Qin CZ, English E, Burt JE, Zhou CH (2008). Purposive sampling for digital soil mapping for areas with limited data. In 'Digital Soil Mapping with Limited Data'. (Eds AE Hartemink, AB McBratney, ML Mendonca Santos) pp.233-245. (Springer-Verlag: New York)