Springer Environmental Science and Engineering

Gan-Lin Zhang
Dick Brus
Feng Liu
Xiao-Dong Song
Philippe Lagacherie *Editors*

Digital Soil Mapping Across Paradigms, Scales and Boundaries

Springer

Springer Environmental Science and Engineering

More information about this series at http://www.springer.com/series/10177

Gan-Lin Zhang · Dick Brus · Feng Liu
Xiao-Dong Song · Philippe Lagacherie
Editors

Digital Soil Mapping Across Paradigms, Scales and Boundaries

Springer

Editors
Gan-Lin Zhang
Institute of Soil Science
Chinese Academy of Sciences
Nanjing
China

Xiao-Dong Song
Institute of Soil Science
Chinese Academy of Sciences
Nanjing
China

Dick Brus
Soil Science
Alterra
Wageningen
The Netherlands

Philippe Lagacherie
National Institute for Agricultural Research
Paris
France

Feng Liu
Institute of Soil Science
Chinese Academy of Sciences
Nanjing
China

ISSN 2194-3214 ISSN 2194-3222 (electronic)
Springer Environmental Science and Engineering
ISBN 978-981-10-0414-8 ISBN 978-981-10-0415-5 (eBook)
DOI 10.1007/978-981-10-0415-5

Library of Congress Control Number: 2015960825

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer Science+Business Media Singapore Pte Ltd.

Foreword

Unprecedented demands are being placed on the world's soil resources (Koch et al. 2013). Responding to these challenging demands requires relevant, reliable, and applicable information. Indeed, soils have critical relevance to global issues such as food and water security and climate regulation and they are increasingly recognized as major contributors to a wide range of ecosystem services. Mankind depends upon soil for nearly everything. Our soil resource is being under threat, and we must improve our knowledge about the current state and trend of soil condition.

Traditional soil survey involves field reconnaissance and data collection to draw soil map unit boundaries (polygons) on maps. However, traditional soil survey programs are cost and time-consuming. Therefore, many parts of the world have no, or little, soil survey information. Also, as traditional soil survey mainly relies on expert knowledge, it cannot be easily reproduced and the uncertainties of the predictions are very difficult to estimate.

Digital soil mapping (DSM) has been proposed as a solution to increase cost-effectiveness of mapping soil classes and soil properties (McBratney et al. 2003), including an assessment of uncertainties. Basically, this method is based on the hypothesis that soil classes or properties can be predicted in a spatially explicit way, by using soil information and (1) spatially exhaustive proxies of soil formation factors and (2) spatially exhaustive sensors of some soil properties. Since the seminal paper from McBratney et al. (2003), enormous advances in DSM have been achieved, mainly thanks to the IUSS Working Group on Digital Soil Mapping. Indeed, DSM has substantially matured and we have reached major advances concerning suitable mapping and modelling procedure.

The DSM Working Group, currently led by Mogens Greve of Aarhus University, Denmark, holds biennial global workshops (Montpellier in 2004, Rio de Janeiro in 2006, Logan in 2008, Rome in 2010, Sydney in 2012); this book presents selected papers presented at the 6th Global Workshop on Digital Soil Mapping. It was held in November 2014 in Nanjing, China, skillfully organized by our Chinese colleagues. Prof. Zhang Ganlin and his colleagues were excellent hosts, and their hospitality was highly appreciated.

The participation of the workshop was successful, considering the contributions of 120 attendees originating from 15 countries from all continents, having 58 talks and 17 posters. The full papers published in this book are a selection from these presentations. They range from overviews of the DSM technology in general to specific applications in areas having more or less available soil information or areas where specific properties are investigated. In this book, recent findings are presented on the use of legacy data, soil sampling, covariates, soil spectroscopy, and 3D modelling in DSM. Particularly, sampling strategy and the uncertainty assessment of DSM products are major issues that are addressed and which should be accounted for in the future research. The coverages and scales of the applications described in this book range from the field, to landscape, national, continental, or world levels. Case studies in different parts of the world provide an excellent opportunity to evaluate DSM technique and test its utility.

These proceedings give a useful overview of the state of the art in DSM. I am convinced that it will be of broad interest for people involved in soil information delivery and utilization. It will be a valuable resource for many years to come for scientists, students, soil surveyors, and end users.

Dominique Arrouays
INRA-InfoSol Unit, France

References

Koch *et al.*, Soil security: solving the global soil crisis. *Global Policy*. 4, 434-441 (2013).
McBratney *et al*., On digital soil mapping. *Geoderma*. 117, 1-2, 3-52 (2003).

Preface

Digital Soil Mapping Across Paradigms, Scales and Boundaries contains papers presented at the 6th Global Workshop on Digital Soil Mapping, held November 11–14, 2014, at the Institute of Soil Science, Chinese Academy of Sciences of Nanjing, China. The organizing committee was chaired by Dr. Gan-Lin Zhang, professor of Institute of Soil Science, Chinese Academy of Sciences. Approximately 120 participants from 15 countries presented and discussed nearly 60 papers during the four-day session, demonstrating the global engagement in digital soil mapping.

Digital soil mapping is advancing on different fronts at different paces throughout the world, facilitating the development of digital soil information with increasing precision for many areas. To map the soils of the world to the every detail, we need is a glorious task of soil scientists, especially when it is done in a modern and fashionable way—mapping soils digitally. The goal of the sixth workshop is to review and discuss the state of the art in digital soil mapping and to explore the strategies for bridging research, methodologies, and environmental applications. The contents of predictive soil mapping, including the concepts, paradigms, models, and mathematical and computational tools, develop continuously and more and more researches and projects, in various sizes, resolutions, and geographic regions, are running in the world. There are also more and more scientists and users who are working in and shaping the frontiers of the field. It is certainly necessary once again to bring people together to exchange and share research results and to discuss the future of digital soil mapping, and we hope to recognize these distinct foci within the realm of digital soil mapping.

We have selected 29 papers from the workshop that focus on digital soil mapping research, environmental application, and operation. Part I is an introductory chapter which provides context for the whole book. The remaining papers are organized into the following parts: (II) Digital Soil Modelling; (III) Environmental

Application and Assessment; and (IV) Soil Sensors and Legacy Data. The CD-ROM accompanying this book contains the digital versions of all contributions with full colour. Whenever reference is made in the book to colour images, the reader is kindly requested to consult the CD-ROM.

Nanjing
November 2014

Gan-Lin Zhang
Dick Brus
Feng Liu
Xiao-Dong Song
Philippe Lagacherie

Contents

Part I
Digital Soil Modelling

Chapter 1
Digital Soil Mapping Across Paradigms, Scales, and Boundaries: A Review

Gan-Lin Zhang, Feng Liu, Xiao-Dong Song and Yu-Guo Zhao

Abstract Accurate spatial soil information is urgently needed for dealing with the global issues such as agricultural production, environmental pollution, food security, water security, and human health. This need has been motivating the development of digital soil mapping. We reviewed recent advances in digital soil mapping with respect to paradigms, scales, and boundaries, with the intent to improve our understanding on current status of soil mapping. Some important challenges thus research opportunities emerged recently were then outlined, such as 3D digital mapping of the soil properties beyond soil organic matter, soil mapping in areas with intensive human activities, and multi-source soil data integration for soil mapping.

1.1 Introduction

The series of the global workshops on digital soil mapping run under the umbrella of the International Union of Soil Sciences Working Group on digital soil mapping. The first global workshop on digital soil mapping was held in the year of 2004 at Montpellier, France. Its theme was "Digital Soil Mapping: An Introductory Perspective." A wide range of skills and tools that can be used for digital soil mapping were discussed in this workshop. The second workshop was held in the year of 2006 at Rio de Janeiro, Brazil. Its theme was "Digital Soil Mapping for Regions and Countries with Sparse Soil Data Infrastructures." The digital soil mapping techniques and applications that focused on areas with limited soil data were emphasized. The third workshop was held in the year of 2008 at Logan, America, with the theme of "Digital Soil Mapping: Bridging Research, Production, and Environmental Application." The soil mapping research, environmental application, and operation were discussed. The fourth workshop was held in the year of 2010 at Roma, Italy, with the theme of "From Digital Soil Mapping to Digital Soil

G.-L. Zhang (✉) · F. Liu · X.-D. Song · Y.-G. Zhao
State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China
e-mail: glzhang@issas.ac.cn

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_1

Assessment: Identifying key gaps from fields to continents." The issues of spatial scales were discussed. The theme of the fifth workshop was held in the year of 2012 at Sydney, Australia, with the theme of "Digital Soil Assessments and Beyond." Current and potential contributions of digital soil mapping to various assessments driven by stakeholders and global issues were emphasized. Then, it comes to the sixth global workshop on digital soil mapping. This event was organized by the Soil Science Society of China and the Institute of Soil Science, Chinese Academy of Sciences at Nanjing, China on November 9–11, 2014. Its theme was "Digital soil mapping across paradigms, scales, and boundaries." The advances in digital soil mapping paradigms, scales, and boundaries were emphasized in this workshop.

The state of the art of digital soil mapping has been reviewed several times mainly from different perspectives such as history, techniques, data production, and applications (McBratney et al. 2003; Lagacherie 2008; Grunwald 2010; Arrouays et al. 2014; Minasny and McBratney 2015). The objective of this paper is to present current status with respect to paradigms, scales, and boundaries and important issues on digital soil mapping that emerged more recently.

1.2 Soil Mapping Paradigms

A paradigm is a set of concepts or thought patterns, including theories, methods, and models. It provides solutions for a community of practitioners. In 1883, the Russian pedologist Dokuchaev put forward the famous theory on soil-forming factors, i.e., soil is formed over time as a consequence of climatic (CL), parent material (P), and biological processes (O), which he demonstrated that soils are products of soil-forming factors. Jenny (1941) further developed this into a soil-forming function, i.e., $S = f$ (clorpt…) by adding topographic relief as a factor. This equation suggests that, by looking for changes in these factors as the landscape is traversed, one can identify boundaries between different bodies of soils. The formulation has been used by a lot of soil investigators as a conceptual soil-forming model for understanding soil–landscape patterns within a region. Many studies have tried to quantitatively formalize the equation. Based on a review of various quantitative approaches to making digital soil maps, McBratney et al. (2003) proposed a quantitative framework suitable for digital mapping and modeling of soil classes and properties, i.e., the well-known SCORPAN model. It is an empirical model, and both factors and soil predictions are spatially and temporally explicit. To explicitly account for the role of anthropogenic factors in soil formation, Grunwald et al. (2011) and Thompson et al. (2012) proposed a new framework for soil mapping and modeling, i.e., the STEP-AWBH model. Water properties (e.g., surface runoff, infiltration rate) and human-induced forcings (e.g., contamination, greenhouse gas emissions) were added as new soil-forming factors. It is an enhanced quantitative framework for soil mapping and modeling. Its key features includes accounts for time-dependent variation of the factors and facilitates modeling of soil evolution and change.

1.3 Soil Mapping Scales

Soil varies over space and changes over time. At different spatial or temporal scales, soil can exhibit distinct processes and patterns. In order to meet the requirements of soil information for different levels of applications, digital soil mapping has been explored across various spatial or temporal scales. Temporal scales can span from hours to several decades and even one thousand years. Studies of digital soil mapping at specific temporal scale mainly focus on the changes of soil salinity, soil carbon, and soil thickness (Douaik et al. 2005; Follain et al. 2006; Lark et al. 2006; Sun et al. 2012; Ardekani 2013). Its purpose is to reveal the patterns of soil evolution. Spatial scales include global, continental, regional, catchment, landscape, and field. Digital soil mapping has been conducted at all these scales. The GlobalSoilMap.net Project launched in 2009 aims to produce a new digital soil map of the globe using digital soil mapping technologies. It will map most of the ice-free land surface of the world at a 90-m spatial resolution (Sanchez et al. 2009). The interpretation and functionality options will also be provided with the maps to support improved decisions for a range of global problems. However, limited attempts have been made at global scale especially for a high-resolution map. When there is no detailed map or soil samples are available in a region of interest, Mallavan et al. (2010) proposed a Homosoil method to extrapolate from other parts of the globe. In order to provide a consistent global soil data, Köchy et al. (2014) derived global distribution of soil organic carbon based on the Harmonized World Soil Database. Hengl et al. (2014) developed global 3D soil distribution data based on regression or regression-kriging methods. But due to some limitations, the prediction accuracies are relatively low. A few attempts have been made on the continental scale for all five continents (Henderson et al. 2001; Viscarra Rossel et al. 2011; Odgers et al. 2012; TÓth et al. 2013; Stevens et al. 2013; Dewitte et al. 2013; Láng et al. 2015). Scull and Okin (2007) discussed sampling challenges posed by continental-scale soil–landscape modeling and argued that the success of the sampling design in continental scale largely depend on the ability to anticipate the spatial variability of the variable being measured. Grunwald et al. (2011) incorporated anthropogenic forcings into a space-time modeling framework to provide a solution for soil mapping and modeling at continental scales. Some continental-scale mapping initiatives are also considered as national scale, because they cover the extent of a whole country, e.g., China, Australia, and the USA. A lot of studies have been made on the regional (Lacoste et al. 2011; Kerry et al. 2012; Wang et al. 2013; Heung et al. 2014; Guo et al. 2015), catchment (Zhu et al. 2001; Qin et al. 2011; Karunaratne et al. 2014; Wahren et al. 2015), landscape (Liu et al. 2013; Lacoste et al. 2014; Stockmann et al. 2015), and field scales (Ardekani 2013; Li et al. 2015; Bevington et al. 2016) due to the increasing requirements of soil information in agriculture and environmental management. Most digital soil mapping techniques have been developed for these scales.

1.4 Soil Mapping Boundaries

The soil-mapping boundaries can be the boundaries between different regions or nations and between soil science and other disciplines. First, if the soil-mapping area spans two or more countries, the soil data collected by different countries and different soil survey projects can be different in many aspects: data collection time (old vs. new soil survey data), data formats (profile points, polygon-based maps, and soil survey reports), sampling strategies (random, regular, or representative), sampling density (sparsely vs. densely distributed), sampling depth (topsoil vs. profile), laboratory analysis methods (e.g., laser diffraction techniques vs. pipette method for measuring soil texture), and soil classification systems (e.g., Canadian soil classification system vs. USDA soil taxonomy). Thus, soil data harmonization is necessary to get consistent soil data for digital soil mapping. Quite a few studies have explored the soil data harmonization techniques (Soon and Abboud 1991; Nemes et al. 2003; Pieri et al. 2006). The specifications of the GlobalSoilMap.net products (v2.3) have identified most of these problems and provide some regression equations for harmonizing multi-source soil data to a reference standard (Arrouays et al. 2014). Baruck et al. (2015) discussed the soil data harmonization issues for soil mapping across the eight Alps countries. The process of collecting soil data and mapping soils, as well as the soil classification systems used, significantly differs among the countries. The harmonization includes an upgrade of an existing international soil classification, e.g., the World Reference Base WRB (IUSS 2014). The harmonization is not only an international transborder problem. For example, in Italy within the Pedological Methods Program in the year 2000, criteria were established for making the soil map of Italy at a scale of 1:250,000. But in order to take into account local specificities, several regions developed their own soil survey manual. Second, the soil information products derived from digital soil mapping should not only meet the applications of soil science (e.g., agricultural production and management) itself but also those of other disciplines including hydrological, ecological, and climatic modeling and even pipeline network design. To what extent the products made by soil mappers can match the requirements of applications in other fields is still an issue to be addressed mainly due to the gaps between the disciplines.

1.5 Current Challenges

1.5.1 3D Digital Soil Mapping of Soil Properties

Most soil maps are continuous surface maps in two dimensions ignoring the fact that soil also varies with depth over a landscape. A few attempts have been made on 3D soil mapping (Liu et al. 2013; Minasny et al. 2013; Arrouays et al. 2014). Most considered it as multiple 2D soil-mapping operations at a set of predefined depth

intervals. These 2D mapping results are represented as depth averages (for concentrations) or sums (for stocks). These averages can be reconstructed into a full 3D soil property map. Although multiple 2D mapping is simple to implement, it is a pseudo 3D mapping approach and has two drawbacks (Liu et al. 2015). One is that soil variation pattern in the vertical dimension is neglected when performing separate horizontal soil predictions for each depth interval. The other is that depth function fitting is often applied twice in the mapping process. Any errors in the fitting are thus repeated and may be magnified. In addition, most 3D soil-mapping studies only focus on soil organic carbon, mainly because the profile distribution of this property is relatively simple and thus can be easily fitted by an exponential decay function. But, other soil properties such as soil texture and bulk density, to a big extent, have been ignored by the 3D soil-mapping studies mainly due to their complex distribution patterns with depth. Thus, it is necessary to study how to generate accurate 3D maps of these demanded soil properties in the next years.

1.5.2 Soil Mapping in Areas with Intensive Human Activities

Human activities are an important soil-forming factor, which exhibit both deterministic patterns (e.g., land-use patterns) and highly randomness (e.g., agricultural practices such as irrigation and fertilization). There are two types of areas with intensive human activities. One is the urban areas experiencing intensive urbanization, and the other is the cultivated areas experiencing intensive land uses. Urban soils present a diversity of specific processes and features, such as soil pollution and compaction, zoning, fertilization, sewage release, and combustion. These processes may result in high patchiness and short-distance heterogeneity. Very high short-distance soil variability within such areas and long distances between settlements limit the use of traditional spatial interpolation methods. Similarly, agriculture soils also have high spatial heterogeneity because of irrigation and fertilization and specific practices. Thus, the digital soil mapping in these two types of areas is challenging. In both, much attention is needed for anthropogenic soil-forming factors. Vasenev et al. (2013, 2014) explored the soil organic carbon mapping in a highly urbanized area. In addition to traditional factors, urban-specific factors, including size and history of the settlements and functional zoning, were used as auxiliary information for mapping soil organic carbon stocks.

1.5.3 Multi-Source Data Integration for Soil Mapping

Soil data used for digital soil mapping can be collected from multiple sources: legacy soil data from conventional soil survey and new soil sampling data from

recent soil survey projects. As mentioned above, data harmonization is needed before they are used for soil mapping. It includes the harmonization within a single soil data source and that between multiple sources. Both are challenging tasks. The soil-type cross-references from one soil classification system to another can only be performed at a coarse level (e.g., soil great group). It is usually difficult to convert clay, silt, or sand content from one soil texture classification standard to another. The conversion between different laboratory analysis methods is always empirical and dependent on the soil regions or types. In particular for the soil properties that are not steady over time (e.g., one or more decades), such as soil organic carbon and pH value, how to integrate legacy soil data with new soil sampling data for digital soil mapping remains a challenge. Sun et al. (2015) compared the changes of digital maps of soil organic matter generated from three sets of soil sampling data from three soil survey projects conducted at different periods. The proximal soil sensing and digital soil morphometrics are also important soil data sources which can provide a large amount of "soft" soil data for soil mapping. It is necessary to incorporate these data into existing soil sampling data for high-resolution digital soil mapping. But much work is still needed to be done. In addition to the data collected by specialists in soil science, Rossiter et al. (2015) argued that the citizen (non-specialists) can also assist digital soil mapping by providing soil samples or landscape knowledge. They proposed digital soil-mapping and citizen-science initiatives. The "citizen" can be farmers, land managers, civil engineers, gardeners, and participants in outdoor activities. They pointed out that a key issue for the citizen science is how to integrate observations from citizens and those from the professionals.

Acknowledgements This work was supported by the National Natural Science Foundation of China (41130530, 91325301, 41201207, 41401237, and 41571212).

References

Ardekani, M.R.M. 2013. Off- and on-ground GPR techniques for field-scale soil moisture mapping. Geoderma 200–201: 55–66.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., et al., 2014. GlobalSoilMap: toward a fine-resolution global grid of soil properties. In: Sparks, D.L. (Ed.), Advances in Agronomy 125. Academic Press, Burlington.

Baruck, J., Nestroy, O., Sartori, G., Baize, D., Traidl, R., et al. 2015. Soil classification and mapping in the Alps: The current state and future challenges. Geoderma, doi.org/10.1016/j.geoderma.2015.08.005

Bevington, J., Piragnolo, D., Teatini, P., Vellidis, G., Morari, F. 2016. On the spatial variability of soil hydraulic properties in a Holocene coastal farmland. Geoderma 262: 294–305.

Dewitte, O., Jones, A., Spaargaren, O., Breuning-Madsen, H., Brossard, M., Dampha, A., et al., 2013. Harmonization of the soil map of Africa at the continental scale. Geoderma 211–212: 138–153.

Douaik, A.,VanMeirvenne, M., Tóth, T. 2005. Soil salinity mapping using spatio-temporal kriging and Bayesian maximum entropy with interval soft data. Geoderma 128: 234–248.

Follain, S., Minasny, B., McBratney, A.B., Walter, C. 2006. Simulation of soil thickness evolution in a complex agricultural landscape at fine spatial and temporal scales. Geoderma 133: 71–86.

Grunwald, S. 2010. Current state of digital soil mapping and what is next. In: J.L. Boettinger et al. (eds.), Digital Soil Mapping, Progress in Soil Science.

Grunwald, S., Thompson, J.A., Boettinger, J.L. 2011. Digital soil mapping and modeling at continental scales: finding solution for global issues. Soil Sci. Soc. Am. J. 75: 1201–1213.

Guo, P.T., Li, M.F., Luo, W., Tang, Q.F., Liu, Z.W., Lin, Z.M. 2015. Digital mapping of soil organic matter ofr rubber plantation at regional scale: an application of random forest plus residuals kriging approach. Geoderma 237–238: 49–59.

Henderson, B., Bui, E., Moran, C., Simon, D., Carlile, P. ASRIS: continental-scale soil property predictions from point data. Technical Report 28/01, November 2001. CSIRO Land and Water, Canberra.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R. 2014. SoilGrids1 km-Global soil information based on automated mapping. PLOS ONE 9(8): e105992. doi:10.1371/journal.pone.0105992.

Heung, B., Bulmer, C.E., Schmidt, M.G. 2014. Predictive soil parent material mapping at a regional scale: a random forest approach. Geoderma 214–215: 141–154.

IUSS Working Group WRB, 2014.World reference base for soil resources 2014.World Soil Resources Report 106 (Rome).

Jenny H. 1941. Factors of Soil Formation, McGraw-Hill, New York.

Karunaratne, S.B., Bishop, T.F.A., Baldock, J.A., Odeh, I.O.A. 2014. Catchment scale mapping of measurable soil organic carbon fractions. Geoderma 219–220: 14–23.

Kerry, R., Goovaerts, P., Rawlins, B.G., Marchant, B.P. 2012. Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. Geoderma 170: 347–358.

Köchy, M., Hiederer, R., Freibauer, A. 2014. Global distribution of soil organic carbon, based on the Harmonized World Soil Database-Part 1: Masses and frequency distribution of SOC stocks for the tropics, and he world. Soil Discuss 1: 327–362.

Lacoste, M., Lemercier, B., Walter, C. 2011. Regional mapping of soil parent material by machine learning based on point data. Geomorphology 133: 90–99.

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C. 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213: 296–311.

Lagacherie, P., 2008. Digital Soil Mapping: A state of the art. In: Hartemink, A.E., McBratney, A.B., and Mendonça-Santos, M.L. (eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Láng, V., Fuchs, M., Szegi, T., Csorba, A., Micheli, E. 2015. Deriving World Reference Base Reference Soil Groups from the prospective Global Soil Product-a case study on major soil types of Africa.

Lark, R.M., Bellamy, P.H. & Rawlins, B.G. 2006. Spatio-temporal variability of some metal concentrations in the soil of eastern England, and implications for soil monitoring. Geoderma 133: 363–379.

Li, H.Y., Webster, R., Shi, Z. 2015. Mapping soil salinity in the Yangtze delta: REML and universal Kriging (E-BLUP) revisited. Geoderma 237–238: 71–77.

Liu, F., Rossiter, D.G., Song, X.D., Zhang, G.L., Yang, R.M., Zhao, Y.G., Li, D.C., Ju, B. 2015. A similarity-based method for three-dimensional prediction of soil organic matter concentration. Geoderma http://dx.doi.org/10.1016/j.geoderma.2015.05.013.

Liu, F., Zhang, G.-L., Sun, Y.-J., Zhao, Y.-G., Li, D.-C. 2013. Mapping the three-dimensional distribution of soil organic matter across a subtropical hilly landscape. Soil Sci. Soc. Am. J. 77: 1241–1253.

Mallavan, B.P., Minasny, B., McBratney, A.B. 2010. Homosoil: a methodology for quantitative extrapolation of soil information across the globe. In: J.L. Boettinger et al. (eds.), Digital Soil Mapping, Progress in Soil Science.

McBrantney, A.B., Mendonca Santos, M.L., Minasny, B. 2003. On digital soil mapping. Geoderma 117: 3–52.

Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Digital mapping of soil carbon. Advances in Agronomy 118: 1–47.

Minasny, B., McBratney, A.B. 2015. Digital soil mapping: A brief history and some lessons. doi:10.1016/j.geoderma.2015.07.017.

Nemes, A., Schaap, M.G. and Wösten, J.H.M. 2003. Functional evaluation of pedotransfer functions derived from different scales of data collection. Soil Sci. Soc. Am. J. 67:1093–1102.

Odgers, N.P., Libohova, Z., Thompson, J.A. 2012. Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale. Geoderma 189–190: 153–163.

Pieri, L., Bittelli, M., Pisa, P.R. 2006. Laser diffraction, transmission electron microscopy and image analysis to evaluate a bimodal Gaussian model for particle size distribution in soils. Geoderma 135: 118–132.

Qin, C.Z., Zhu, A.X., Qiu, W.L., Lu, Y.J., Li, B.L., Tao, P. 2011. Mapping soil organic matter in small low-relief catchments using fuzzy slope position information. Geoderma doi:10.1016/j.geoderma.2011.06.006.

Rossiter, D.G., Liu, J., Carlisle, S., Zhu, A.X. 2015. Can citizen science assist digital soil mapping. Geoderma 259–260: 71–80.

Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., de Mendonca-Santos, M.L. et al., 2009. Digital Soil Map of the World. Science 325(5941): 680–681.

Scull, P., Okin, G.S. 2007. Sampling challenges posed by continental-scale soil landscape modeling. Science of the Total Environment 372: 645–656.

Soon, Y.K. and Abboud, S. 1991. A comparison of some methods for soil organic carbon determination, Communications in Soil Science and Plant Analysis 22: 943–954.

Stevens, A., Nocita, M., Toth, G., Montanarella, L., van Wesemael, B. 2013. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PLoS ONE 8(6): e66409. doi:10.1371/journal.pone.0066409.

Stockmann, U., Malone, B.P., McBratney, A.B., Minasny, B. 2015. Landscape-scale exploratory radiometric mapping using proximal soil sensing. Geoderma 239–240: 115–129.

Sun, X.L., Zhao, Y.G., Wu, Y.J., Zhao, M.S., Wang, H.L. & Zhang, G.L. 2012. Spatio-temporal change of soil organic matter content in Jiangsu Province, China, based on digital soil maps. Soil Use and Management 28: 318–328.

Sun, X.L., Wu, Y.J., Lou, Y.L., Wang, H.L., Zhang, C., Zhao, Y.G., Zhang, G.L. 2015. Updating digital soil maps with new data: a case study of soil organic matter in Jiangsu, China. European Jouranl of Soil Science doi: 10.1111/ejss.12295.

Thompson, J.A., Roecker, S., Grunwald, S., Owens, P.R. 2012. Digital soil mapping: interactions with and applications for hydropedology. Hydropedology 1: 664–709.

TÓth, G., Gardi, C., Bodis, K., Lvits, E., Aksoy, E., Jones, A., Jeffrey, S., Petursdottir, T., Montanarella, L. 2013. Continental-scale assessment of provisioning of soil functions in Europe. Ecological Processes 2: 32. http://www.ecologicalprocesses.com/content/2/1/32.

Vasenev, V.I., Stoorvogel, J.J., Vasenev, I.I. 2013. Urban soil organic carbon and its spatial heterogeneity in comparison with natural and agricultural areas in the Moscow region. Catena 107: 96–102.

Vasenev, V.I., Stoorvogel, J.J., Vasenev, I.I., Valentini, R. 2014. How to map soil organic carbon stocks in highly urbanized regions. Geoderma 226–227: 103–115.

Viscarra Rossel, R.A. 2011. Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. Journal of Geophysical Research 116, F04023, doi:10.1029/2011JF001977.

Wahren, F.T., Julich, S., Nunes, J.P., Gonzalez-Pelayo, O., Hawtree, D., Feger, K.H., Keizer, J.J. 2015. Combing digital soil mapping and hydrological modeling in a data scarce watershed in north-central Portugal. Geoderma http://dx.doi.org/10.1016/j.geoderma.2015.08.023.

Wang, K., Zhang, C., Li, W. 2013. Predictive mapping of soil total nitrogen at a regional scale: a comparison between geographically weighted regression and cokriging. Applied Geography 42: 73–85.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D. 2001. Soil mapping using GIS, expert knowledge and fuzzy logic. Soil Sci. Soc. Am. J. 65:1463–1472.

Chapter 2
Spatial Prediction of Soil Antibiotics Based on High-Accuracy Surface Modeling

Wenjiao Shi, Tianxiang Yue, Xuewen Li and Zhengping Du

Abstract The spatial prediction of soil antibiotic is more difficult than other normal soil properties due to the diverse sources of soil antibiotics. Few studies have attempted to predict soil antibiotic residues in intensive vegetable cultivation areas. High-accuracy surface modeling (HASM) is regarded as an important new technique in the pedometrics and digital soil mapping fields. A total of 100 surface soil samples were collected from the north-central part of the Shandong Province of China. The antibiotic concentrations, including ciprofloxacin (CF), enrofloxacin (EF), norfloxacin (NF), and fluoroquinolones (FQs), were analyzed using high-performance liquid chromatography–tandem mass spectrometry. We employed splines to compare its performance with that of HASM method. The errors of HASM for NF, CF, EF, and FQ were less compared to splines. HASM has less mean absolute error (MAE) and root mean square error (RMSE) than splines. The RMSEs of splines for FQ, CF, EF, and NF were 3.02, 2.34, 3.46, and 2.64 times lager than those of HASM, respectively. Therefore, HASM can be considered as an alternative and accurate method for interpolating soil antibiotics. It can also make the map more consistent with the true spatial distributions.

Keywords HASM · Soil antibiotics · Interpolation accuracy · Geostatistics

W. Shi (✉)
Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, 11A, Datun Road, Anwai, Beijing 100101, China

T. Yue · Z. Du
State Key Laboratory of Resources and Environment Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, 11A, Datun Road, Anwai, Beijing 100101, China

X. Li
Department of Environment and Health, School of Public Health, Shandong University, Jinan, China

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_2

2.1 Introduction

A large number of antibiotics are released into soils via animal manures applied to the soil (Bound and Voulvoulis 2004), because antibiotics are widely used in treating disease, protecting animal health, and improving the feeding efficiency of animals (Xie et al. 2012; Sarmah et al. 2006; Aust et al. 2008; Li et al. 2013, 2014). Accumulation of antibiotics in the soil can damage the structure of bacterial communities, be absorbed by vegetables or crops thereby threatening the safety of agricultural products, or be leached to groundwater thereby affecting environmental health (Xie et al. 2012; Martínez-Carballo et al. 2007). The apparent immobilization of soil antibiotics in the surface soils to a greater depth was attributed to the presence of higher percentages of clay and organic matter in the surface soils with residues of antibiotics bound strongly to soil particles (Sarmah et al. 2006).

In order to avoid the risk of soil antibiotics on vegetable quality and human health, it is vital to research the spatial patterns of soil antibiotics (Sarmah et al. 2006). So, it is necessary to explore an effective surface modeling methods for soil antibiotics. Low accuracy of interpolation results of soil antibiotics will result in incorrect risk assessment. There are several interpolators in spatial prediction for soil properties, such as kriging (Stein and Corsten 1991; Stein et al. 1988; Webster and Oliver 2001; Goovaerts 1999), inverse distance weighting (IDW) (Panagopoulos et al. 2006; Weber and Englund (1992, 1994); Gotway et al. 1996), and splines (Webster and Oliver 2001). However, the spatial predictions of soil antibiotics are more difficult than other normal soil properties due to the diverse sources of soil antibiotics. Few studies have focused on the interpolation of antibiotic residues in soils of the intensive vegetable cultivation areas.

High-accuracy surface modeling (HASM) is a spatial interpolation technique based on the fundamental theorem of surfaces (Yue 2011), which has been successfully used in soil property interpolation (Shi et al. 2011, 2009, 2012). The HASM method combined with some ancillary information can improve the interpolation of soil properties (Shi et al. 2011). Due to the specific characters of the spatial distributions of soil antibiotics, we added the distribution characters of soil antibiotics in different vegetable areas in HASM method to predict the spatial distributions of soil antibiotics. The aims of this study are (i) to explore the distribution characters of soil antibiotics in different vegetable areas, (ii) to assess the feasibility of HASM combined with the statistical characters of soil antibiotics in different vegetable types in spatial interpolation of soil antibiotics, and (iii) to evaluate the performance of HASM in improving the soil property interpolation compared with classical methods such as splines.

2.2 Data and Methods

2.2.1 Data

The study area of this study is an important vegetable area in the north-central part of Shandong Province, China, covering 160 km^2. There are several typical vegetables grown in this area, such as cucumber, tomatoes, peppers, melons, eggplant, and some leaf vegetables. Several types of animal manure such as chicken manure and cow dung as organic fertilizer were applied in this area. The map of manure applications has been shown in the previous reference (Li et al. 2013). The whole study area was with manure application. The study area was covered by 100 sampling sites. The average distance between soil sampling locations is approximately 1 km. The sampling sites were designed to cover evenly the whole area and to include different vegetable types, different manure types, and different application years. Three types of fluoroquinolone (FQ) concentrations, including ciprofloxacin (CF), enrofloxacin (EF), and norfloxacin (NF), were analyzed using high-performance liquid chromatography–tandem mass spectrometry. The regents and sample analysis methods were in detail in the former studies (Xie et al. 2012; Li et al. 2013, 2014).

2.2.2 Methods

A full discussion on the theoretical aspects of HASM applied for the interpolation of soil properties was given by Shi et al. (2009, 2011, 2012). We only introduced the main steps here. HASM uses samplings of soil antibiotics to globally fit a surface through several iterative simulation steps. This surface is then used to interpolate soil antibiotic values at unknown points. The iterative simulation steps are summarized as follows (Yue et al. 2010): (1) interpolate within the domain of the sample data $(x_i, y_j, \bar{u}_{i,j})$, from which we can get interpolated values $\{\tilde{u}_{i,j}\}$ at point (x_i, y_j); (2) $u^0_{i,j} = \tilde{u}_{i,j}$ calculate the first fundamental coefficients $E^n_{i,j}$, $F^n_{i,j}$, and $G^n_{i,j}$ and the second fundamental coefficients $L^n_{i,j}$ and $N^n_{i,j}$ as well as coefficients in terms of $\left\{u^n_{i,j}\right\}$; (3) for $n \geq 0$, we can get $\left\{u^{n+1}_{i,j}\right\}$ by solving the HASM equations; and (4) the iterative process is repeated until simulation accuracy is satisfied.

Here, we calculated the mean values of antibiotic residues in the soils of different vegetable-type areas. Then, the mean values were grided as 100 m × 100 m over the whole study area. This layer was defined as $u^0_{i,j}$. So, the distribution characters of soil antibiotics in different vegetable-type areas were combined with HASM.

A total of 80 training points were randomly created as the interpolation data set, and the remaining 20 samples were used as the validation data set. We used the most common indices, the mean error (ME), the mean absolute error (MAE), and

the root mean square error (RMSE) as measures of interpolation quality (Shi et al. 2011, 2009, 2012).

As there are n sites belonging to the validation sample set, the ME, the MAE, and the RMSE are determined from the measured values $z(x_i)$ and the interpolated value $z^*(x_i)$. They are given by,

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}[z^*(x_i) - z(x_i)] \tag{2.1}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}[|z^*(x_i) - z(x_i)|] \tag{2.2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[z^*(x_i) - z(x_i)]^2} \tag{2.3}$$

The ME is a measure of the bias of the interpolation, which should be close to zero for unbiased methods, and the MAE and RMSE are measures of the accuracy of interpolation which should be as small as possible for accurate interpolation.

2.3 Results and Analysis

2.3.1 Distributions of Soil Antibiotics in Different Types of Vegetable Areas

In the three soil FQ residues, soil CF has the highest mean concentration with 104.4 μg/kg, and the followings were the concentrations of soil NF and EF with 55.7 and 15.8 μg/kg, respectively (Table 2.1). The standard deviations (SDs) were high for FQ (151.07 μg/kg), indicating that the FQs varied largely with different vegetables planted and different inputs (Li et al. 2014). More than 25 % samples of soil FQs were over 222 μg/kg, and the maximum (Max) was 682.69 μg/kg. The concentrations of CF in the surface soil were much higher than the other two types of FQs; the mean, median, and Max were 104.4, 72.4, and 651.6 μg/kg, respectively. Soil NF followed, and the mean, median, and Max were 55.7, 33.3, and 288.3 μg/kg, respectively. The concentrations of soil EF were the lowest, and the mean and median were only 15.8 and 7.0 μg/kg, respectively. Soil EF in 25 % samples was less than 0.6 μg/kg.

There is a large variability of soil FQs in different soil samples, so we analyzed the distributions of soil FQs in different types of vegetables (Table 2.1). Tomato–tomato and cucumber–cucumber were the two main planted types in the north part of study area, accounting for more than 50 % area. There were 21 (72.4 % of all the samples (29 samples) more than 200 μg/kg) samples more than 200 μg/kg for soil

Table 2.1 Descriptive statistics of antibiotics concentrations in surface soil (unit: μg/kg)

	Min	5 %	10 %	25 %	Median	75 %	90 %	95 %	Max	Mean	SD
FQ	9.1	20.5	31.2	63.8	132.2	222.3	413.8	509.9	682.7	175.2	151.1
NF	0.4	5.0	7.6	14.2	33.3	79.5	149.6	171.1	288.3	55.7	56.4
CF	2.4	8.7	13.0	29.3	72.4	141.0	221.2	302.1	651.6	104.4	117.5
EF	0.0	0.0	0.0	0.6	7.0	18.5	50.6	70.2	167.0	15.8	25.5

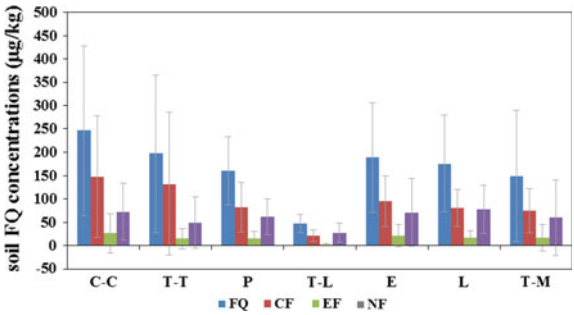

Fig. 2.1 Soil FQ concentrations in different vegetable planted areas. *C–C* cucumber–cucumber; *T–T* tomato–tomato; *P* pepper; *T–L* tomato–leaf; *E* eggplant; *L* leaf; *T–M* tomato–melon

FQs found in the two planted types. As shown in Table 2.1, higher FQ concentrations were mainly composed of higher CF residues, and a few higher NF samples located in the tomato–melon, eggplant, and leaf planted areas which were over 200 μg/kg.

The distributions of soil FQ residues in different planted vegetable types showed that FQ residues in the cucumber–cucumber-planted soil were the highest to 246.3 μg/kg and followed by those in tomato–tomato, eggplant, leaf, pepper, and tomato–melon planted areas with 196.6, 187.9, 176.1, 160.5, and 149.3 μg/kg (Fig. 2.1). The averages of soil CF in these two planted type area were also the highest, which were 147.5 and 132.4 μg/kg. The average of soil EF was fewer compared to the other two soil FQs, but the coefficients of variation were larger than 100 % only except for pepper and leaf planted areas.

2.3.2 *Comparisons of the Performance of HASM and Splines*

In order to compare the performance of HASM and splines, we computed ME, MAE, and RMSE (Table 2.2). HASM had higher accuracy in the two techniques, which performed better for the three indices than splines. The MEs of HASM for FQ, CF, EF, and NF were closer to zero than splines. Also, splines had larger

Table 2.2 The MEs, MAEs, and RMSEs of the validation set for HASM and splines (unit: μg/kg)

Soil antibiotics	Indices	HASM	Splines
FQ	ME	19.69	86.34
	MAE	46.66	168.13
	RMSE	68.73	207.23
CF	ME	-0.61	18.12
	MAE	44.03	109.88
	RMSE	58.97	137.72
EF	ME	6.54	21.62
	MAE	13.37	33.71
	RMSE	16.86	58.37
NF	ME	16.48	46.35
	MAE	33.69	75.98
	RMSE	39.16	103.28

MAEs and RMSEs for FQ, CF, EF, and NF. The MAE of HASM for FQ was 46.7 μg/kg and that of splines was 168.13 μg/kg. For RMSE, HASM also performed better than splines. The RMSEs of splines for FQ, CF, EF, and NF were 3.02, 2.34, 3.46, and 2.64 times greater than those of HASM, respectively.

2.3.3 The Interpolated Maps by HASM Methods

We mapped the distributions of FQ, CF, EF, and NF by HASM (Fig. 2.2). The map based on HASM has more details in the interpolation maps and also has consistent maximum and minimum with those of soil samples. Although the spatial patterns of the two techniques had similar spatial distribution tendency of higher FQ residues in the north part of the study area, HASM showed lower values of FQs consistent with the concentration of soil samples. However, splines has not showed this character in the interpolation maps. The EF concentrations in the soil of the study area were the lowest in the three types of the FQs, which had lower values less than 100 μg/kg except a small part of the north. Although a few higher values over than 200 μg/kg of NF were in the southeast of the study area, most of the NF concentrations in the soil were less than 100 μg/kg. In these three types of soil FQs, CF values were the highest. For example, there were about half areas of the study area more than 150 μg/kg, and 10 to 20 % areas of the study area were more than 200 μg/kg.

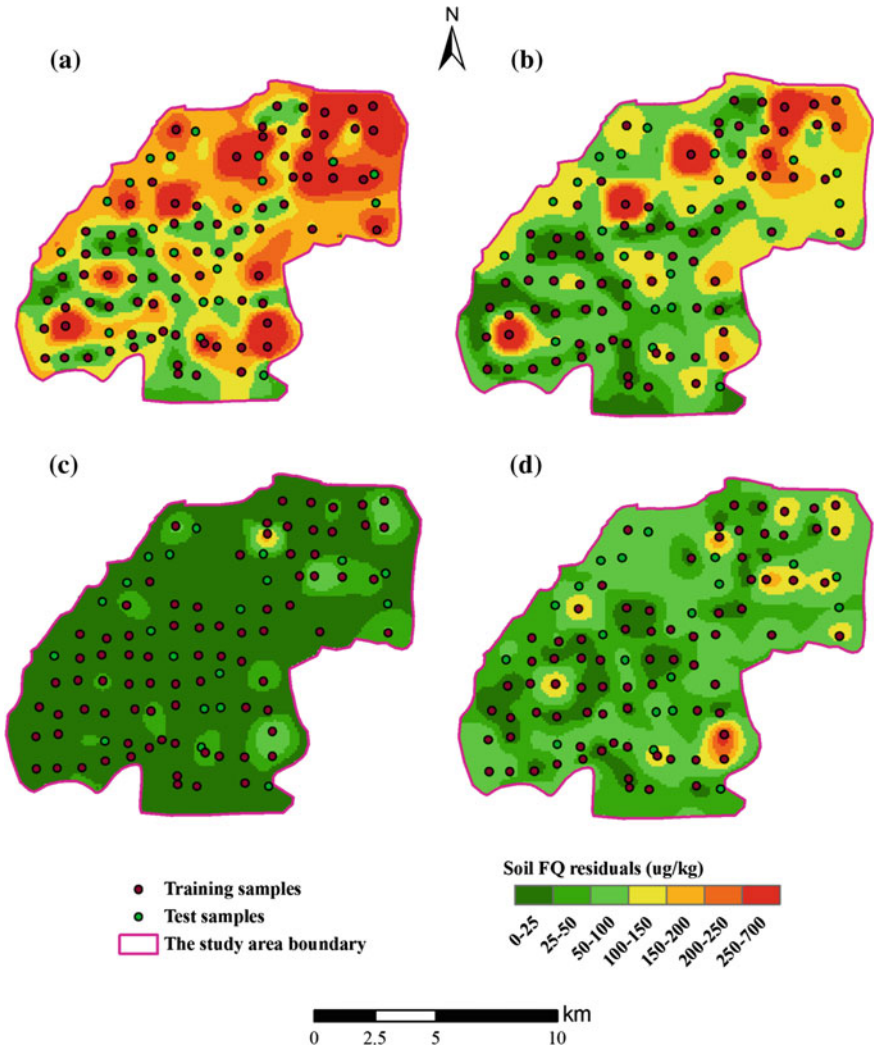

Fig. 2.2 The spatial patterns of soil FQ, CF, EF, and NF residues obtained by HASM. **a** FQ, **b** CF, **c** EF, **d** NF

2.4 Conclusions

China is the country which is the largest producer and consumer of antibiotics compared to the other countries in the world, and antibiotics can be easily added to soil particles with fertilizers of livestock and poultry feces. Fluoroquinolones (FQs) are the most frequently detected antibiotics in the feces of farm livestock and poultry in China. However, relatively few studies focus on the interpolation of soil

antibiotic residues in the intensive vegetable cultivation area. High-accuracy surface modeling (HASM) is regarded as an important new technique in the pedometrics and "digital soil" fields. The HASM method is a spatial interpolation technique based on the fundamental theorem of surfaces. In this study, a total of 100 surface soil samples were collected from the north-central part of the Shandong Province of China to test the performance of HASM with that of classical method. We found that the errors of HASM for NF, CF, EF, and FQ were less compared to splines. HASM has less MAE and RMSE than splines. The RMSEs of splines for FQ, CF, EF, and NF were 3.02, 2.34, 3.46, and 2.64 times larger than those of HASM, respectively. Therefore, HASM can be considered as an alternative and accurate method for interpolating soil antibiotics. HASM is not only as an alternative and accurate method for interpolating soil antibiotics, but also can make the map more consistent with the true spatial distributions.

Acknowledgements This study was supported by the State Key Laboratory of Resources and Environmental Information System, the National Natural Science Foundation of China (41371002 and 91325204). The authors are grateful to the reviewers for the constructive suggestions.

References

Aust MO, Godlinski F, Travis GR, Hao X, McAllister TA, Leinweber P, Thiele-Bruhn S (2008) Distribution of sulfamethazine, chlortetracycline and tylosin in manure and soil of Canadian feedlots after subtherapeutic use in cattle. Environmental Pollution 156(3): 1243–1251.

Bound J, Voulvoulis N (2004) Pharmaceuticals in the aquatic environment—A comparison of risk assessment strategies. Chemosphere 56(11): 1143–1155.

Goovaerts P (1999) Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89(1–2): 1–45.

Gotway CA, Ferguson RB, Hergert GW, Peterson TA (1996) Comparison of kriging and inverse-distance methods for mapping soil parameters. Soil Science Society of America Journal 60(4): 1237–1247.

Li X, Xie Y, Li C, Zhao H, Zhao H, Wang N, Wang J (2014) Investigation of residual fluoroquinolones in a soil–vegetable system in an intensive vegetable cultivation area in Northern China. Science of The Total Environment 468–469(0): 258–264.

Li X, Xie Y, Wang J, Christakos G, Si J, Zhao H, Ding Y, Li J (2013) Influence of planting patterns on fluoroquinolone residues in the soil of an intensive vegetable cultivation area in northern China. Science of The Total Environment 458–460(0): 63–69.

Martínez-Carballo E, González-Barreiro C, Scharf S, Gans O (2007) Environmental monitoring study of selected veterinary antibiotics in animal manure and soils in Austria. Environmental Pollution 148(2): 570–579.

Panagopoulos T, Jesus J, Antunes MDC, Beltrao J (2006) Analysis of spatial interpolation for optimising management of a salinized field cultivated with lettuce. European Journal of Agronomy 24(1): 1–10.

Sarmah AK, Meyer MT, Boxall A (2006) A global perspective on the use, sales, exposure pathways, occurrence, fate and effects of veterinary antibiotics (VAs) in the environment. Chemosphere 65(5): 725–759.

Shi W, Liu J, Du Z, Song Y, Chen C, Yue T (2009) Surface modelling of soil pH. Geoderma 150 (1–2): 113–119.

Shi W, Liu J, Du Z, Stein A, Yue T (2011) Surface modelling of soil properties based on land use information. Geoderma 162(3-4): 347–357.

Shi W, Liu J, Du Z, Yue T (2012) Development of a surface modeling method for mapping soil properties. Journal of Geographical Sciences 22(4): 752–760.

Stein A, Corsten LCA (1991) Universal Kriging and Cokriging as a Regression Procedure. Biometrics 47(2): 575–587.

Stein A, Hoogerwerf M, Bouma J (1988) Use of soil-map delineations to improve (Co-)kriging of point data on moisture deficits. Geoderma 43(2-3): 163–177.

Weber D, Englund E (1992) Evaluation and comparison of spatial interpolators. Mathematical Geology 24(4): 381–391.

Weber D, Englund E (1994) Evaluation and comparison of spatial interpolators II. Mathematical Geology 26(5): 589–603.

Webster R, Oliver MA (2001) Geostatistics for Environmental Scientists. West Sussex, England: John Wiley and Sons.

Xie Y, Li X, Wang J, Christakos G, Hu M, An L, Li F (2012) Spatial estimation of antibiotic residues in surface soils in a typical intensive vegetable cultivation area in China. Science of The Total Environment 430(0): 126–131.

Yue T, Song D, Du Z, Wang W (2010) High-accuracy surface modelling and its application to DEM generation. International Journal of Remote Sensing 31(8): 2205–2226.

Yue TX (2011) Surface Modelling: High Accuracy and High Speed Methods. New York: CRC Press.

Chapter 3
Incorporating Probability Density Functions of Environmental Covariates Related to Soil Class Predictions

Jenette M. Ashtekar, Phillip R. Owens, Zamir Libohova and Ankur Ashtekar

Abstract The distribution of continuous soil properties and their environmental covariates within soil classes are often times unknown or not evaluated. Understanding and defining the distribution of environmental covariates within soil classes is fundamental to the fuzzy logic inference mapping processes. Under knowledge-based applications of fuzzy logic mapping, the user typically utilizes predetermined distribution functions to define a representative relationship between soils and their covariates. If the predetermined distributions do not adequately describe the soil–covariate relationship, the misrepresentation can lead to inadequate prediction of soil properties while requiring a high level of user input. To move away from knowledge-based "guesses" of distributions, we present a new and innovative method of modeling the distribution of environmental covariates, specifically terrain attributes (TAs), within the landform-based soil classes. This eliminates the need for manually manipulated, user-defined curves and works to more accurately represent the distribution of TAs within soil classes. The fully automated method fits a variety of probability distribution functions (PDFs) to TA values within algorithm-derived landform classes. We compared the Pearson's correlation coefficient for goodness of fit to determine which PDF best models the distribution of TAs within soil classes. This fully automated method works to improve our understanding of how terrain attributes vary within soil classes, allowing for more accurate and reliable model predictions.

Keywords Digital soil mapping · Fuzzy logic · Probability density function · Landform classification

J.M. Ashtekar (✉) · P.R. Owens
Department of Agronomy, Purdue University, West Lafayette IN, USA
e-mail: jmashtekar@purdue.edu

Z. Libohova
National Soil Survey Center, United States Department of Agriculture, Natural Resources Conservation Service, Lincoln, NE 68508, USA

A. Ashtekar
AgSoil Analytics, Inc., West Lafayette, IN, USA

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_3

3.1 Introduction

Digital soil mapping (DSM) typically uses information ancillary to sampled soil data to classify and predict soil properties (McBratney et al. 2003; McKenzie and Austin 1993; McKenzie and Ryan 1999; Scull et al. 2003). This ancillary, covariate information often comes in the form of remotely sensed continuous rater data such as satellite imagery, digital elevation models (DEMs), and their derivatives. Digital elevation models and their derivatives are frequently used to represent important topographic features that drive soil formation and differentiation. These derivatives, called terrain attributes (TAs), have been used in DSM to aid in soil classification and property prediction (Behrens et al. 2010; Bruin and Stein 1998; Gessler et al. 2000; Moore et al. 1993).

Traditionally, soil mapping was performed by individual soil scientists who used their personally developed mental models to delineate taxonomic soil classes across the landscape. When parent material was held constant, surveyors would delineate based on landscape positions, following the toposequence, or catena models. Some forms of DSM seek to replicate the modeling process played out in the mind of the soil scientist by using computer technology to mimic and expand upon the surveyor's decision-making process.

Fuzzy logic modeling is a typically user-driven process which utilizes expert knowledge to estimate the distribution of TAs within user-defined soil classes. Under inference-based fuzzy logic mapping, membership curves, or functions, are used to quantify the membership of a given cell in each predicted soil class. Under the knowledge-based approach, curves are typically bell-shaped, s-shaped, and z-shaped, with full membership (curve peak) and half membership values defined by the user. The expert sets the curves based on their assumed concept of the TA distribution within theoretical soil classes, and the actual distribution is unknown (Ashtekar and Owens 2013; Zhu et al. 1996, Zhu et al. 2001, 1996). The success of this knowledge-driven method hinges on the user's ability to understand and define the relationships between soil classes, properties, and continuous TAs. When this knowledge is lacking, it becomes necessary to define membership curves using alternate means.

The goal of this study is to describe the actual distribution of TAs within soil classes represented through algorithm-driven landform classification. Because we know the spatial extent of each assumed soil class, we can fit probability density functions to the TAs within those classes, fitting curves to the data itself, not the Soil Scientist's mental concept. For fuzzy mapping purposes, we can rescale the PDF to become the fuzzy membership function.

The scope of this study was limited to exploring the automated fitting of different probability density functions (PDFs) to TA values within soil classes, as represented by algorithm-derived topographic landform classes and excludes the actually fuzzy mapping of soil classes and properties.

3.2 Methods

3.2.1 Site and Data Description

The study was conducted at the Southeastern Purdue Agricultural Center (SEPAC), located in southeastern Indiana, in the Midwestern region of the USA. The site encompasses an approximately 16 ha agricultural field currently managed under a corn soybean rotation. The field is characterized by loess over till.

The study site is interesting in that it encompasses areas of preglacial residual limestone interspersed with glacial till, overlain by a highly variable loess cap ranging from 5 cm on slopes to over 300 cm in low lying depressions. The area has a distinct, red paleosol layer formed in till underlying the loess.

A five-meter digital elevation model was obtained from the Indiana Spatial Data Portal (www.gis.iu.edu) from which a variety of terrain attributes were generating using SAGA GIS. The TAs selected for analysis were chosen for their relation to the soil classes. Certain values of a particular TA may be highly correlated to the occurrence of a particular soil type. For the purpose of field-level mapping, we focused on TAs relevant to water redistribution across the field and relative landscape position including slope, topographic wetness index (TWI), catchment area, multiresolution ridge top flatness (MRRTF), and multiresolution valley bottom flatness (MRVBF). We worked under the assumption that water redistribution is the main driver of soil differentiation and can be adequately represented by terrain attributes.

3.2.2 Landform Classification

Geomorphometry quantitatively characterizes the land surface topography, and geomorphometric algorithms can be used to segment and classify the landscape into landform units which may be relevant to soil-forming processes as well as overall soil function (Park and van de Giesen 2004; Park et al. 2001; Pennock 2003; Pennock et al. 1987; Pike et al. 2009). A variety of landform classification procedures are available, with most using computer manipulation of elevation, represented by raster DEMs, to segment the landscape (Iwahashi and Pike 2007; Jasiewicz and Stepinski 2012; MacMillan et al. 2000; Park et al. 2001; Pennock and Corre 2001; Pike et al. 2009).

For the purpose of automatically classifying landforms, two freely available landform classification algorithms were selected to act as surrogates for user-defined landform-based soil classes or previously defined class boundaries such as existing soil survey polygons. The Iwahashi and Pike and Geomorphons algorithms were selected for their applicability to any landscape, DEM grid size, and spatial extent (Iwahashi and Pike 2007; Jasiewicz and Stepinski 2012).

The automated landform classification approach outlined by Iwahashi and Pike (2007) is an unsupervised, empirical method which uses a nested means approach to classify topography automatically on the basis of three terrain attributes: slope gradient, surface convexity, and texture. The Geomorphons method, presented by Jasiewicz and Stepinski (2012), offers a unique approach to automated mapping of geomorphological units. While most other methods rely heavily on differential geometry for the generation of terrain attributes, from which landform classes are derived, Geomorphons identifies landforms from patterns in relative elevation, classifying landforms directly from the DEM itself.

The Iwahashi and Pike and Geomorphons classifications were run at a five-meter resolution for the study area. Iwahashi and Pike allows for the prediction of 8, 12, or 16 landform classes. Classification of 8 landforms was selected because it was assumed that 12 or more functionally district soil class are not present at the 16 ha SEPAC site.

3.2.3 Curve Fitting

Four probability density functions were generated for each TA within each landform including both the Iwahashi and Pike and Geomorphons classifications. The normal distribution (Eq. 3.1), log normal distribution (Eq. 3.2), exponential distribution (Eq. 3.3), and Weibull distribution (Eq. 3.4) were selected. To generate the functions, the TA values of each grid cell were first extracted by landform. From this dataset, 30 % of the pixels were randomly selected from each landform in an attempt to bring independence into the dataset. The resulting subset was used to fit each PDF in MATLAB. Parameter estimation for all PDFs was performed using maximum likelihood estimation (MLE).

$$y = f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{3.1}$$

$$y = f(x|\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}} \tag{3.2}$$

$$y = f(x|\mu) = \frac{1}{\mu} e^{\frac{-x}{\mu}} \tag{3.3}$$

$$f(x|a,b) = \frac{b}{a}\left(\frac{x}{a}\right)^{b-1} e^{-(x/a)^6} \tag{3.4}$$

The parameters of the normal, lognormal, and exponential probability density functions are mean, μ, and standard deviation, σ, given the data, x. The parameters of the Weibull distribution are, a, the scale parameter, and b, the shape parameter.

3.2.4 Goodness-of-Fit Assessment

To assess the goodness of fit of each distribution, chi-square (χ^2) and Kolmogrov–Smirnov (K-S) tests were performed. The chi-square test is used to determine whether the fitted distribution differs from the actual observed distribution and is calculated as follows:

$$\chi_c^2 = \sum_{i=1}^{m} \frac{n(f_s(x_i) - p(x_i))^2}{p(x_i)} \tag{3.5}$$

where n is the number of observations, m is the cell size, and $p(x)$ is the probability associated with each cell, $1/m$. The test was performed at a 95 % confidence level, where $\alpha = 0.05$. If the calculated test statistic, χ_c^2, is less than $\chi_{\alpha,\upsilon}$, the fitted distribution cannot be rejected (Massey 1951).

The Kolmogrov–Smirnov (K-S) test focuses on the deviation from the fitted cumulative density function (CDF).

$$\hat{F}(x_i) = \frac{i}{n} \tag{3.6}$$

where $x_{(i)}$ is the smallest ith value of the original annual maximum times series, x, and n is the total number of observations. The test statistic of interest is as follows:

$$d_2 = \max_{i=1}^{n} \left[\left| \hat{F}(x_i) - F(x_i) \right| \right] = \max_{i=1}^{n} \left[\left| \frac{i}{n} - F(x_i) \right| \right] \tag{3.7}$$

where $F(x(i))$ is the fitted CDF. If the test statistic, d_2, is less than the K-S limit, then the fitted distribution cannot be rejected. The K-S test was performed at a confidence level of 95 %, where $\alpha = 0.05$ (Norton 1945).

3.3 Results and Discussion

3.3.1 Landform Classification

Both the Geomorphons and Iwahashi and Pike algorithms were run for the SEPAC site using a five-meter DEM. The resulting classifications are shown in Figs. 3.1 and 3.2. Geomorphons predicted all 10 possible landforms, with class 1, flats, having only one cell classified. Because some classes encompassed too few grid cells for adequate curve fitting, these classes were combined logically with larger landform groups that were assumed to share common soil-forming conditions. The class numbers, corresponding landform names, total grid cells classified, and reclassification are shown in Table 3.1. Spurs and slopes were found to dominate the landscape. Flats, summits, and shoulders were grouped together.

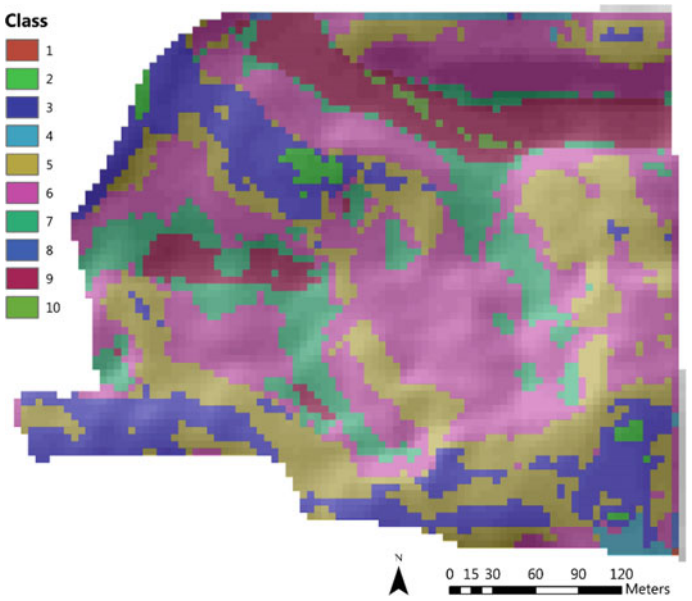

Fig. 3.1 Geomorphons classification. Algorithm derived classification of ten Geomorphons

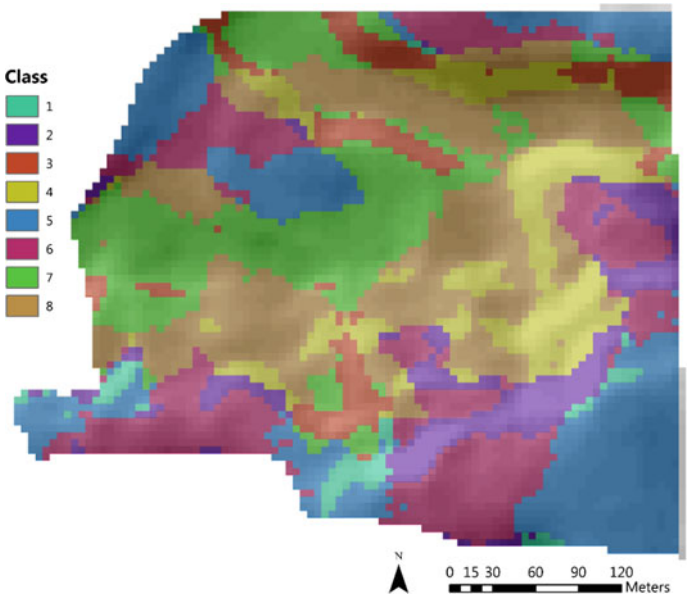

Fig. 3.2 Iwahashi and Pike classification. Algorithm derived classification of eight Iwahashi and Pike landforms

Table 3.1 Geomorphons classification

Class number	Landform name	Grid cells classified	Reclassification
1	Flat	1	1
2	Summit	56	1
3	Ridge	835	2
4	Shoulder	62	1
5	Spur	1373	3
6	Slope	2089	4
7	Hollow	684	5
8	Footslope	8	6
9	Valley	523	6
10	Depression	23	6

Landform name, number of cell classified, and reclassified class numbers

The shoulder landform typically occurs at the outer edges of the summit and acts as the transition from summit to side slope. Because these landforms are spatially contiguous, they were grouped with the assumption that the fuzzy nature of the modeling process would draw out the transitional characteristics of the shoulder position. Footslopes, valleys, and depressions were combined, given that they all represent environments in which water may accumulate. Though the Geomorphons algorithm is considered scale independent, some of the landforms, such as hollows, valleys, and ridges, would not typically be found at the field scale.

The Iwahashi and Pike algorithm was run to classify 8 landforms, shown in Fig. 3.2. Reclassification was not needed as all landforms encompassed a number of grid cells adequate for distribution fitting. The nested means approach of the Iwahashi and Pike algorithm ensures a more equal distribution of pixels among landform classes, when compared to Geomorphons, given that the data are not highly skewed. The class numbers, corresponding landform descriptions, and total grid cells classified are shown in Table 3.2. The gentle classes were found to dominate, which is consistent with the gently sloping farm field environment.

Table 3.2 Iwahashi and Pike classification

Class number	Landform description	Grid cells classified
1	Steeper/high convexity/fine texture	109
2	Steeper/high convexity/coarse texture	415
3	Steeper/low convexity/fine texture	348
4	Steeper/low convexity/coarse texture	578
5	Gentler/high convexity/fine texture	1039
6	Gentler/high convexity/coarse texture	966
7	Gentler/low convexity/fine texture	974
8	Gentler/low convexity/coarse texture	1225

IP class landform descriptions and number of grid cell classified per class

The Iwahashi and Pike and Geomorphons algorithms produced noticeably different landform patterns. Determining which algorithm works best to represent soil classes will require further study, including field sampling and fuzzy logic-based property mapping. Based on observations made in the field, the Iwahashi and Pike classification was able to identify a distinct soil-forming environment (class 5), clearly identifiable in the field and a broad, flat summit position having deep loess of approximately 125 cm, located in the southeast corner of the field.

3.3.2 Curve Fitting

Normal, lognormal, exponential, and Weibull PDFs were generated for each combination of landform class, TA, and distribution. The resulting goodness-of-fit assessments are shown in Tables 3.3 and 3.4.

For the Geomorphons landform classification, very few distributions passed the K-S and χ^2 goodness-of-fit tests. Of the 120 PDFs generated, 12 passed the K-S test, approximately 10.0 %, and four passed the χ^2, approximately 3.3 %. Of these, the Weibull distribution produced the largest number of passing PDFs, seven, and the exponential distribution the lowest, with two. Distribution of passes among classes showed no identifiable pattern, and no PDFs for class 2 were found to pass.

Table 3.3 Chi-squared and Kolmogrov–Smirnov test statistics for Geomorphons fitted distributions

Terrain attribute	Class	Weibull		Log normal		Normal		Exponential	
		X2	K-S	X2	K-S	X2	K-S	X2	K-S
Slope %	1	23.79	0.15	98.31	0.25	18.01	0.14	18.78	0.13
	2	27.54	0.06	371.44	0.17	40.42	0.08	132.29	0.13
	3	33.79	0.02*	148.15	0.08	80.64	0.06	610.9	0.21
	4	34.76	0.03	378.99	0.1	46.82	0.03	1189.48	0.23
	5	29.66	0.05*	103.86	0.12	64.47	0.05*	465.7	0.24
	6	25.43	0.05*	43.12	0.06	109.27	0.13	111.01	0.18
TWI	1	26.42	0.11*	46.38	0.16	33.59	0.13	65.97*	0.49
	2	357.63	0.16	218.16	0.14	298.52	0.16	2680.2	0.51
	3	754.15	0.19	391.97	0.14	666.78	0.18	5972.13	0.54
	4	927.39	0.15	347.2	0.09	714.86	0.13	8142.39	0.53
	5	90.4	0.07	29.29	0.03*	54.56	0.05	2091.41	0.48
	6	14.89	0.04*	15.44	0.05*	3.67*	0.03*	2549.27	0.5
Mod catch	1	30.19	0.23	25.07	0.16	62.77	0.38	44.65	0.31
	2	177.72	0.19	108.16	0.16	567.17	0.32	398.65	0.32
	3	235.24	0.26	585.49	0.21	486	0.38	237.47	0.41

(continued)

Table 3.3 (continued)

Terrain attribute	Class	Weibull		Log normal		Normal		Exponential	
		X2	K-S	X2	K-S	X2	K-S	X2	K-S
	4	389.51	0.18	437.6	0.13	1509.13	0.33	478.33	0.3
	5	50.32	0.1	23.81	0.05	298.37	0.27	182.49	0.21
	6	44.9	0.06	96.51	0.07	77.27	0.15	43.18	0.06*
MRRTF	1	83.15	0.2	161.95	0.28	84.68	0.2	444.47	0.38
	2	828.15	0.12	1340.23	0.13	572.09	0.19	789.21	0.13
	3	180.29	0.08	532.44	0.18	730.45	0.27	720.8	0.21
	4	192.57	0.07	292.82	0.13	1088.79	0.36	1446.98	0.39
	5	29.02	0.1	105.58	0.2	208.38	0.36	38.09	0.23
	6	20.42*	0.13	20.39*	0.09	169.53	0.37	14.69*	0.16
MRVBF	1	24.83	0.12*	7.72	0.08*	65.16	0.3	51.31	0.22
	2	157.75	0.08	169.73	0.13	522.01	0.2	156.18	0.08
	3	231.08	0.11	805.45	0.19	633.75	0.25	670.36	0.16
	4	271.17	0.08	851.33	0.12	1256.56	0.28	1124.03	0.32
	5	167.51	0.09	483.31	0.18	251.66	0.2	168.91	0.12
	6	302.24	0.17	418.18	0.22	334.74	0.12	536.48	0.25

Test statistics for each combination of soil class, terrain attribute (TA), and probability density function (PDF). TAs include slope %, topographic wetness index (TWI), modified catchment area (Mod. Catch), multiresolution valley bottom flatness (MRVBF), and multiresolution ridgetop flatness. K-S and χ^2 values denoted by * indicates that the data fit the distribution

Table 3.4 Chi-squared and Kolmogrov–Smirnov test statistics for Iwahashi and Pike fitted distributions

Terrain Attribute	Class	Weibull		Log Normal		Normal		Exponential	
		X2	K-S	X2	K-S	X2	K-S	X2	K-S
Slope %	1	22.35	0.13	7.24*	0.08*	9.39*	0.13*	198.95	0.44
	2	55.66	0.08	6.42*	0.03*	42.47	0.08	831.57	0.44
	3	24.74	0.06	8.37*	0.04*	11.66*	0.04*	963.74	0.46
	4	175.22	0.13	50.42	0.08	133	0.12	1521.7	0.47
	5	72.45	0.06	507.38	0.18	133	0.11	172.83	0.13
	6	54.73	0.05	273.74	0.11	26.85	0.04	794.02	0.25
	7	54.84	0.06	212.07	0.09	93.93	0.08	422.22	0.18
	8	97.14	0.07	403.44	0.12	58.16	0.05	906.2	0.23
TWI	1	33.45	0.16	15.24*	0.09*	10.27*	0.11*	94.85	0.59
	2	44.79	0.09	9.74*	0.04*	13.05	0.04*	3373.7	0.57
	3	222.91	0.16	116.51*	0.14	160.62*	0.16	1917.5	0.54
	4	123.45	0.13	31.25	0.06	57.48	0.08	3557.6	0.56
	5	238.58	0.11	127.14	0.1	176.6	0.12	3313.6	0.5

(continued)

Table 3.4 (continued)

Terrain Attribute	Class	Weibull		Log Normal		Normal		Exponential	
		X2	K-S	X2	K-S	X2	K-S	X2	K-S
	6	525.77	0.17	215.16	0.12	327.6	0.14	6429	0.55
	7	50.66	0.04*	120.25	0.07	62.89	0.05	2674.1	0.45
	8	264.7	0.09	93	0.06	185.1	0.08	3592.7	0.49
Mod catch	1	14.25	0.15	5.51*	0.12*	16.92	0.21	30.54	0.32
	2	23.42	0.08	5.07*	0.03*	27.69	0.12	131.06	0.29
	3	12.94	0.18	13.5	0.15	75.93	0.29	35.76	0.24
	4	79.23	0.15	38.65*	0.09	138.6	0.27	54.52	0.19
	5	207.95	0.15	142.56	0.09	535.7	0.35	589.78	0.23
	6	138.34	0.17	109.39	0.09	333.8	0.29	158.39	0.19
	7	48.06	0.04*	107.03	0.07	388.5	0.23	151.3	0.15
	8	92	0.11	40.87	0.07	657.7	0.29	696.59	0.28
MRRTF	1	5.08	0.11*	19.15*	0.21	22.63	0.24	12.44	0.27
	2	56.88	0.2	144.1	0.26	89.95	0.29	88.76	0.35
	3	9.93	0.15	66.3	0.27	78.46	0.28	29.09	0.27
	4	40.99	0.15	175.79	0.24	101.9	0.27	44.59	0.24
	5	1269.4	0.18	1923.8	0.17	786.3	0.21	1210.3	0.18
	6	133.83	0.07	92.44	0.05	619.3	0.24	317.89	0.13
	7	217.03	0.08	165.75	0.05	703.9	0.27	707.33	0.22
	8	96.82	0.06	44.84	0.03*	295.8	0.27	79.99	0.1
MRVBF	1	0.28*	0.08*	9.87*	0.21	26.68	0.28	7.21	0.21
	2	58.16	0.13	177.55	0.24	59.3	0.27	42.4	0.3
	3	5.05	0.07	49.16	0.15	98.08	0.28	54.32	0.28
	4	76.82	0.12	330.1	0.21	94.51	0.23	45.86	0.21
	5	171.65	0.1	167.61	0.07	414.9	0.23	170.77	0.09
	6	315.51	0.11	127.4	0.07	460	0.23	263.06	0.11
	7	243.36	0.11	326.38	0.12	416.5	0.21	253.29	0.11
	8	369.37	0.1	383.89	0.07	953.9	0.24	393.9	0.11

Values of test statistics for each combination of soil class, terrain attribute (TA), and probability density function (PDF). TAs include slope %, topographic wetness index (TWI), modified catchment area (Mod. Catch), multiresolution valley bottom flatness (MRVBF), and multiresolution ridgetop flatness. K-S and χ^2 values denoted by * indicates that the data fit the distribution

Iwahashi and Pike landforms performed better in goodness-of-fit assessment. Of the 160 PDFs generated, 16 passed the K-S test, approximately 11.3 %, and 16 passed the χ^2 test, approximately 10.0 %. The majority of the passing PDFs were from the normal and lognormal distributions. Only three Weibull and none of the exponential were found to fit the data.

3.4 Conclusions

The use of fitted probability density functions in fuzzy logic-based soil mapping has much potential. These functions have the ability to act as fuzzy membership curves, with no need of user definition of curve properties. By fitting a variety of curves and choosing the best one, it may be possible to improve fuzzy map prediction. Moreover, this method allows for curves to be generated automatically, lessening the need for expert supervision.

The high number of PDFs found to inadequately fit the data implies that the normal, lognormal, Weibull, and exponential distributions do not work well to model the distributions of terrain attributes within the algorithm-derived landforms. These distributions are generally unimodal and will not perform well for multimodal data. Further research is needed to explore multimodal PDFs for the purpose of better fitting membership functions.

References

Ashtekar, JM, Owens, PR (2013). Remembering Knowledge: An Expert Knowledge Based Approach to Digital Soil Mapping. Soil Horizons 54:1–6. doi:10.2136/sh13-01-0007

Behrens, T, Zhu, A-X, Schmidt, K, Scholten, T (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155:175–185. doi:10.1016/j.geoderma.2009.07.010

Bruin, S De, Stein, A (1998). Soil-landscape modelling using fuzzy c-means clustering of attribute data derived from a Digital Elevation Model (DEM). Geoderma 83:17–33. doi:10.1016/S0016-7061(97)00143-2

Gessler, PE, Chadwick, O A, Chamran, F, Althouse, L, Holmes, K (2000) Modeling Soil–Landscape and Ecosystem Properties Using Terrain Attributes. Soil Sci. Soc. Am. J. 64:2046–2056. doi:10.2136/sssaj2000.6462046x

Iwahashi, J, Pike, RJ (2007). Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. Geomorphology 86:409–440. doi:10.1016/j.geomorph.2006.09.012

Jasiewicz, J, Stepinski, T (2012). Geomorphons-a pattern recognition approach to classification and mapping of landforms. Geomorphology. 182:147-156. doi:10.1016/j.geomorph.2012.11.005

MacMillan, R, Pettapiece, WW, Nolan, SC, Goddard, TW (2000). A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets Syst. 113:81–109. doi:10.1016/S0165-0114(99)00014-7

Massey, FJ (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. J. Am. Stat. Assoc. 46:68–78. doi:10.1080/01621459.1951.10500769

McBratney, AB, Mendonça Santos, M, Minasny, B (2003). On digital soil mapping. Geoderma 117:3–52. doi:10.1016/S0016-7061(03)00223-4

McKenzie, N, Austin, M (1993). A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57:329-355. doi:10.1016/0016-7061(93)90049-Q

McKenzie, N., Ryan, PJ (1999). Spatial prediction of soil properties using environmental correlation. Geoderma 89:67–94. doi:10.1016/S0016-7061(98)00137-2

Moore, ID, Gessler, PE, Nielsen, G, Peterson, G (1993). Soil Attribute Prediction Using Terrain Analysis. Soil Sci. Soc. Am. J. 57:443-452. doi:10.2136/sssaj1993.03615995005700020026x

Norton, HW (1945). Calculation of Chi-Square for Complex Contingency Tables. J. Am. Stat. Assoc. 40:251–258. doi:10.1080/01621459.1945.10501855

Park, S, McSweeney, K, Lowery, B (2001). Identification of the spatial distribution of soils using a process-based terrain characterization. Geoderma 103:249–272. doi:10.1016/S0016-7061(01)00042-8

Park, S, van de Giesen, N (2004). Soil–landscape delineation to define spatial sampling domains for hillslope hydrology. J. Hydrol. 295:28–46. doi:10.1016/j.jhydrol.2004.02.022

Pennock, D (2003). Terrain attributes, landform segmentation, and soil redistribution. Soil Tillage Res. 69:15–26. doi:10.1016/S0167-1987(02)00125-3

Pennock, D, Corre, M (2001). Development and application of landform segmentation procedures. Soil Tillage Res. 58:151–162. doi:10.1016/S0167-1987(00)00165-3

Pennock, D, Zebarth, BJ, De Jong, E (1987). Landform classification and soil distribution in hummocky terrain, Saskatchewan, Canada. Geoderma 40:297–315. doi:10.1016/0016-7061(87)90040-1

Pike, RJ, Evans, IS, Hengl, T (2009). Geomorphometry: A Brief Guide 33:1–9. doi:10.1016/S0166-2481(08)00001-9.

Scull, P, Franklin, J, Chadwick, O, McArthur, D (2003). Predictive soil mapping: a review. Prog. Phys. Geogr. 27:171–197. doi:10.1191/0309133303pp366ra

Zhu, A-X, Band, LE, Dutton, B, Nimlos, TJ (1996). Automated soil inference under fuzzy logic. Ecol. Modell. 90:123–145. doi:10.1016/0304-3800(95)00161-1

Zhu, A-X, Hudson, B, Burt, J, Lubich, K, Simonson, D (2001). Soil Mapping Using GIS, Expert Knowledge, and Fuzzy Logic. Soil Sci. Soc. Am. J. 65:1463. doi:10.2136/sssaj2001.6551463x

Chapter 4
Mapping Horizontal and Vertical Spatial Variability of Soil Salinity in Reclaimed Areas

Yan Guo, Zhou Shi, Jingyi Huang, Laigang Wang, Yongzheng Cheng and Guoqing Zheng

Abstract In coastal China, there is an urgent need to increase the land area for agricultural production and urban development, where there is a rapid growing population. One solution is land reclamation from coastal tidelands, but soil salinization is problematic. As such, it is very important to characterize and map the within-field variability of soil salinity in space and time. Some proximal sensors such as the EM38 allow for the rapid and cost-effective in situ collection of high-resolution data. In this study, we used the EM38 to study spatiotemporal variability of soil salinity in a coastal paddy field. Geostatistical methods were used to determine the horizontal spatiotemporal variability of soil salinity over three consecutive years. The study found that the distribution of salinity was heterogeneous and the leaching of salts was more significant in the edges of the study field. By inverting the EM38 data using a Quasi-3D inversion algorithm, the vertical spatiotemporal variability of soil salinity was determined and the leaching of salts over time was easily identified. We concluded that the methodology of this study can be used as guidance for researchers interested in understanding soil salinity development as well as land managers aiming for effective soil salinity monitoring and management practices.

Keywords Soil salinity · EM38 · (Geo)statistical analysis · Quasi-3D inversion · Spatiotemporal variability

Y. Guo (✉) · L. Wang · Y. Cheng · G. Zheng
Institute of Agricultural Economics and Information, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China
e-mail: 10914063@zju.edu.cn

Z. Shi
Institute of Agricultural Remote Sensing and Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China

J. Huang
School of Biological, Earth and Environmental Science, The University of New South Wales, Kensington, NSW 2052, Australia

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_4

4.1 Introduction

Over the past decades, most of the tidelands in China have been reclaimed for agriculture and urban buffer zones (Huang et al. 2008). However, the highly saline coastal soil often causes adverse effects on agricultural productivity, particularly in the first 20 years of agricultural production. In order to better manage the reclaimed tidelands, it is important to determine the spatial and temporal variability of soil salinity in an accurate and efficient way.

In the last century, conventional visual observations with limited laboratory measurements have been used to map soil salinity variability. However, visual observations provide only qualitative information (Doolittle and Brevik 2014), and laboratory methods are often time-consuming, expensive, and labor-intensive (Corwin 2008). Generally, in order to characterize the soil spatial variability using geostatistical methods, approximately 100 sample points are required to estimate a spatial statistical model (Webster and Oliver 1992). For example, in an attempt to map soil salinity, Gallichand et al. (1992) collected 80 soil samples at two different depths on a regular grid and used 2D and 3D kriging to interpolate the conductivity of the saturated paste extract (i.e., ECe) in a field in Southern Alberta.

So, collection of easily obtainable auxiliary information cost-effectively is the first necessity. In this case, the proximal sensing electromagnetic induction (EMI) emerges as the need to detect soil salinity rapidly and effectively (Corwin 2008). EMI can acquire a large number of georeferenced and quantitative measurements that can be easily correlated with the spatial variability of salinity (Doolittle and Brevik 2014; Guo et al. 2015b). Under saline condition, soil apparent electrical conductivity (ECa) is mainly response to soil salinity (Lesch et al. 2005; Triantafilis et al. 2000). The most commonly used conductivity meter (EM38, Geonics Ltd., Ont, Canada) can measure ECa. The EM38 data have been used to map soil properties (e.g., ECe and soil moisture) using various calibration models (Corwin and Rhoades 1982, 1984; Cook and Walker 1992; Lesch et al. 1995a, b; Padhi and Misra 2011) at field (Lesch et al. 1995a, b), region (Akramkhanov et al. 2011), and catchment (Triantafilis et al. 2000) scales.

In addition to mapping the variability of soil salinity, many researchers have attempted to measure ECa at different soil depths with an inversion algorithm. A pioneering work in this field was undertaken by Hendrickx et al. (2002) and Li et al. (2013). In this literature, Tikhonov regularization was used to invert the EM38 data using measurements collected at different heights above the ground and in different directions. Though successful, the inversion was essentially a 1D inversion model and could not characterize the lateral variation of soil salinity. After several years, researchers developed 2D inversion algorithms to invert the EM38 data into 2D vertical slices (Vervoort and Annen 2006; Monteiro Santos et al. 2010; Mester et al. 2011; Viganotti et al. 2013) and 2D horizontal slices (Monteiro Santos et al. 2002). In recent years, the research with a combination of horizontal slices and vertical slices was employed to determine the 3D variability of soil conductivity

(Shiraz et al. 2013; Triantafilis et al. 2013; Guo et al. 2015a). With these inversion approaches, spatial variability of soil electrical conductivity and the correlated soil properties (e.g., salinity) can be presented in a 2D or 3D view.

Despite the successful application of EMI in soil salinity monitoring and assessments, few publications have reported on the spatiotemporal variability of soil salinity from a multidimensional view. The aim of this study is to map the spatiotemporal variability of soil salinity in a reclaimed costal paddy field using three years of EM38 data. Geostatistical analysis and a Quasi-3D inversion algorithm were combined to map the horizontal and vertical spatial variability of soil salinity in the study field.

4.2 Materials and Methods

4.2.1 Study Area

The study was conducted on a 4.25 ha paddy field in a coastal saline area located in the northern region of Shangyu City, Zhejiang Province, southeast of Hangzhou Bay, China. The climate is subtropical with an average annual temperature of 16.5 °C and an average annual precipitation of 1300 mm. Over the past 40 years, approximately 17,000 ha of coastal land has been reclaimed around Shangyu City in successive programs (Fig. 4.1). The soil is derived from recent marine and fluvial deposits. The study area was enclosed and reclaimed for rice cultivation in 1996. In this area, fields were separated by small embankments (bunds) which ensured flooded conditions within each field.

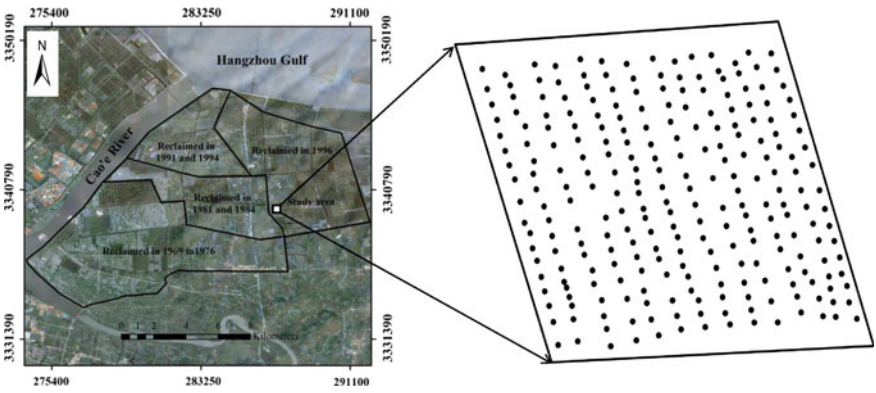

Fig. 4.1 Locations of the study area and ECa measurements

4.2.2 Data Collection and Processing

Measurements of ECa (mS/m) were taken with a Geonics EM38 (Geonics Ltd., Ont., Canada) in the horizontal mode of operation after the rice was harvested and the field was drained. Data files created with data logging system (DAS70-CX) can be used to position a survey according to locations recorded separately by a Global Positioning System (GPS), which can be combined with EM38 records through NMEA-0183 compatible data (i.e., GGA and GSA). The EM38 was placed on the ground, and georeferencing was provided by a Trimble GPS with differential correction within 2 m. ECa measurements were acquired on an approximate 20-m grid along the furrows in three consecutive years. There were 251, 256, and 339 ECa measurements collected in October 2009, November 2010, and November 2011, respectively. In order to calculate the coefficient of variation over time, EM38 measurements in 2010 and 2011 were harmonized onto a common grid consisting of the 251 ECa measurement sites in 2009 (see Fig. 4.1c) using the nearest neighbor algorithm available in ArcGIS 9.3.

It should be noted that EM38 measurements drift significantly when temperatures are over 40 °C and the drift is more obvious for small ECa readings (i.e., less than 100 mS/m) (Robinson et al. 2004). In this study, the temperature conditions were similar when the three surveys were taken (approximately 25 °C) and the study area was highly conductive, so we did not calibrate the ECa measurements to a standard temperature of 25 °C as suggested by Sheets and Hendrickx (1995). However, we still calibrated the equipment many times when conducting field measurements to reduce the error (Corwin and Lesch 2003).

4.2.3 Mapping Horizontal Spatiotemporal Variability of Salinity Using Geostatistical Approaches

Geostatistical methods are often used to define the variance structure, spatial distribution, and trend changes of soil properties. Generally, kriging is the most familiar univariate interpolation method, which uses the semivariogram to quantify the spatial variation of a regionalized variable, of which ordinary kriging (OK) is one of the most popular interpolation methods (Li and Heap 2011). It was defined as follows (Webster and Oliver 2007):

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \tag{4.1}$$

where $\gamma(h)$ is a semivariogram that measures the mean variability between two points x and $x + h$ as a function of their distance h; $Z(x_i)$ and $Z(x_{i+h})$ are the values of the variable Z at location x_i and $x_i + h$; and $N(h)$ is the number of pairs of sample points separated by the lag distance h.

The horizontal spatial variability of soil salinity (ECa) was interpolated by OK (Eq. 4.2) (Webster and Oliver 2007) with ArcGIS 9.3.

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{4.2}$$

where $Z^*(x_0)$ is the predicted ECa at location x_0; $Z(x_i)$ is the measured ECa at location x_i; λ_i is the weight assigned to the observation $Z(x_i)$; and n is the number of measurements.

As for the horizontal temporal variability, the coefficient of variation (CVt_i) over time at each measurement site was calculated to evaluate the stability of soil salinity (Eq. 4.3). The technique has been used by Shi et al. (2002) to assess the stability of soil properties in grasslands.

$$CVt_i = \frac{\sqrt{(n \times \sum_{t=1}^{n} ECa_{it}^2 - (\sum_{t=1}^{n} ECa_{it})^2)/n \times (n-1)}}{(\sum_{t=1}^{n} ECa_{it})/n} \tag{4.3}$$

where CVt_i is the coefficient of variation over three years at the ith ECa measurement site in the tth year and n is the number of ECa measurements.

4.2.4 *Mapping Vertical Spatiotemporal Variation of Salinity Using Quasi-3D Inversion*

To determine the distribution of true electrical conductivity (σ—mS/m) at different depths beneath the ECa measurements, an inversion software (EM4Soil) was used to convert ECa to σ. Herein, the Quasi-3D module (Q3Dm) of the software was employed following the procedure of Monteiro Santos et al. (2011) to invert the ECa data of the three consecutive years. Q3Dm is a 1D spatial constrained technique (1D SCI) and a forward modeling approach. It assumes that below each measured site, the 1D variation of the soil conductivity is constrained by the variation under neighboring sites. The modeling process is based on the cumulative function (Eqs. 4.1 and 4.2). The inversion algorithm is based on the Occam regularization method (Sasaki 1989; De Groot-Hedlin and Constable 1990).

First, gridding was applied onto the raw dataset using the gridding tool of the Q3Dm package. The gridding was based on the inverse distance-weighted method (EM4SOIL Manual, 2011). In this study, a weight value of 2.0 was selected and the grid consisted of 10 x-lines (west–east) and 8 y-lines (south–north) with grid spacing of 18 m. Then, the inversion of ECa data was performed using Algorithm 3 with a damping factor of 0.3, 10 iterations, a data error of 1.00, and a misfit target of 0.20. An initial two-layer laterally homogeneous model was predefined with initial

electrical conductivity of 10 mS/m for both layers, a depth of 0.6 m for the first upper layer, and a depth of 1.2 m for the bottom layer. ECa data of the three consecutive years were inverted separately.

4.3 Results and Discussion

4.3.1 Statistical Analysis of Multitemporal EM38 Data

Table 4.1 shows some basic summary statistics and quartile estimates for ECa in 2009, 2010, and 2011. The average values decrease substantially from 2009 (166.19 mS/m) to 2010 (134.02 mS/m) and 2011 (113.29 mS/m). Similarly, the quartile estimates of ECa show a decreasing trend from 2009 to 2011 with a lesser difference between 2010 and 2011 than 2009 and 2010. The Shapiro–Wilk (S-W) statistics are 0.925, 0.930, and 0.925 with *P*-values less than 0.01, which indicate significant deviation from normality. In such cases, Box–Cox transformation method was adopted to transform the data by monotonically increasing (or decreasing). In the next section, the datasets were also normalized by this method.

Figure 4.2 gives the curves of the cumulative distribution function (CDF) for the study area which illustrates visible temporal variations of soil salinity among the three years. For a given ECa value, CDF is largest in 2011 and smallest in 2009. In order to quantify the difference, we used the Tukey–Kramer multiple comparison

Table 4.1 Descriptive statistics of ECa (mS/m) in 2009, 2010, and 2011

Year	*n*	Mean	Stde	Min	25 %	Median	75 %	90 %	Max	S-W test
2009	251	166.19	3.50	51.3	145.4	179.3	195.6	209.18	226.7	0.925
2010	256	134.02	2.00	20.1	96.15	151.85	182.9	193.4	217.7	0.930
2011	339	113.29	3.01	10.5	73.2	140.2	157.9	168.5	181.8	0.925

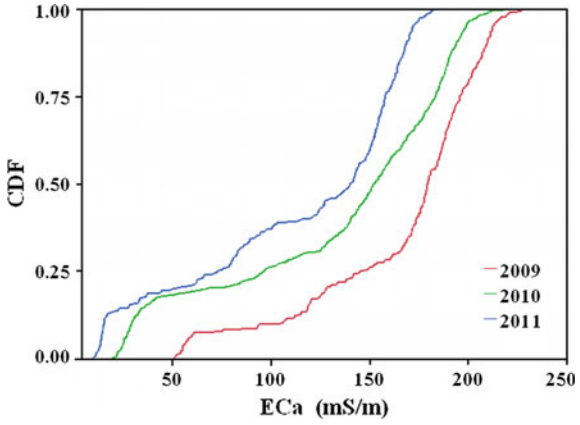

Fig. 4.2 Plot of cumulative distribution function (CDF) of ECa (mS/m) in 2009, 2010, and 2011

Table 4.2 Comparison of means of ECa (mS/m) for 2009, 2010, and 2011 using the Tukey–Kramer test

Year	Mean	2009	2010	2011
2009	166.19	−11.61	22.71	42.06
2010	134.02		−6.64	12.24
2011	113.29			−9.99

procedure. The Tukey–Kramer means comparisons are shown in Table 4.2. The values indicate the actual absolute differences in the means minus the least significant difference (i.e., abs-LSD). Here, there are two important things to note, first, that the mean with positive values indicates significant difference, and second, because the borders of the table are sorted by the mean, the most significant differences among the years appear in the upper right-hand corner. As shown in Table 4.2, the most significant change of ECa occurs between 2009 and 2011, followed by the period from 2009 to 2010, and then between 2010 and 2011.

4.3.2 *Horizontal Spatiotemporal Variability of EM38-Directed Soil Salinity with Geostatistical Approaches*

Analyses of spatial dependence were carried out on all the three datasets. The plot of experimental semivariances and the fitted semivariogram models for the ECa from 2009 to 2011 is shown in Fig. 4.3. And the parameters of these models are given in Table 4.3. The semivariances of the models indicate that the spatial behavior has good continuity in space and can be modeled quite well with exponential models which were selected by simulations using GS + 7.0 (Gamma Design Software, USA). However, different tendencies were found for models of the three years. The nugget value (C_0) decreases from 2009 to 2011, indicating that the variations of soil salinity over a short distance have become smaller and smaller.

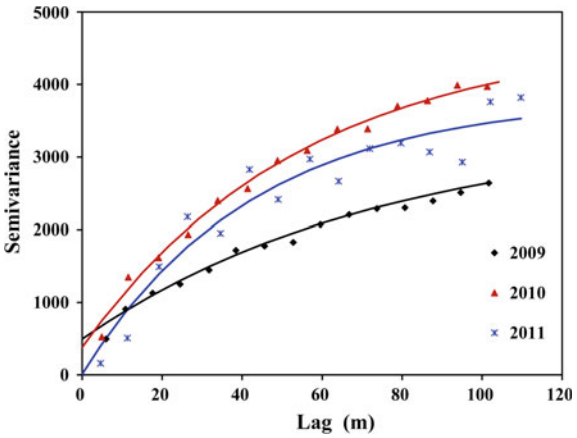

Fig. 4.3 Semivariance and fitted models (*solid lines*) for soil ECa (mS/m) in years 2009–2011

Table 4.3 Models and parameters of semivariogram for ordinary kriging of soil ECa (mS/m) in 2009, 2010, and 2011

Year	Semivariogram model	Nugget (C_0)	Sill ($C_0 + C$)	$C_0/(C_0 + C)$	Range (A)	r^2
2009	Exponential	495	2899	17.07	225.90	0.964
2010	Exponential	380	4302	8.83	165.00	0.912
2011	Exponential	10	3807	0.26	127.50	0.928

The ratios of C_0 to sill ($C + C_0$) decline sharply from 17.07 % (2009) to 0.26 % (2011). According to Shi et al. (2005), the ratio of C_0 to ($C + C_0$) reflects the spatial dependence of soil attributes (i.e., a ratio less than 0.25 indicates strong spatial dependence; a value between 0.25 and 0.75 denotes moderate spatial dependence; and a value greater than 0.75 indicates weak spatial dependence).

In this regard, we can conclude that the spatial autocorrelation of ECa was becoming stronger during the study period. This increase may be caused by the alternating irrigation and drainage practices necessary for rice cultivation. In addition, the relatively large nugget effect in the ECa data is most probably the consequence of an uneven distribution of soil salinity between ridge and furrow irrigation, perhaps associated with a small georeferencing error; Also the abrupt transitions in soil salinity, i.e., a short distance variability was not taken into account by the density of the sampling, and in this case, the nugget effect decreases, it can be assumed that the transitions, initially steep, soften between 2009 and 2011.

Maps of ECa in 2009, 2010, and 2011 generated by OK method are shown in Fig. 4.4a, b, and c, respectively. These smoothed contour maps show that ECa has decreased over the three years. For example, the maximum value of ECa is 181.8 mS/m in 2011 versus 226.7 mS/m in 2009 with the minimum value of ECa also decreasing from 2009 (51.3 mS/m) to 2011 (10.5 mS/m). As well as in a central block of the field (i.e., easting: 286,520–286,560 m; northing: 3,340,360–3,340,400 m), ECa was mostly larger than 200 mS/m in 2009, but the values decreased to 175–200 mS/m in 2010 and then dropped to 125–150 mS/m in 2011. The decreasing ECa value was most likely due to the irrigation and drainage practices for rice cultivation which leached the salts into a deep soil profile or the groundwater.

On the other hand, the spatial distribution of soil salinity also changed. In 2009, the largest ECa values (>200 mS/m) were found in the center of the field and values decreased with distance from the center. However, in 2010, the largest ECa values (>200 mS/m) were found in the right half of the field, and there was a distinctive difference in ECa between the left and right halves of the field. With regard to year 2011, any differences in ECa between the left and right field halves were not obvious and the field was mostly dominated by ECa values of 125–150 mS/m. The heterogeneous and changing salinity distribution of the study area may be caused by the presence of ditches in the study area because the study area is a paddy field and surrounded by ditches. The tillage with large-size tractor tends to result in uniform field surface topography irrigation and drainage for rice cultivation leach

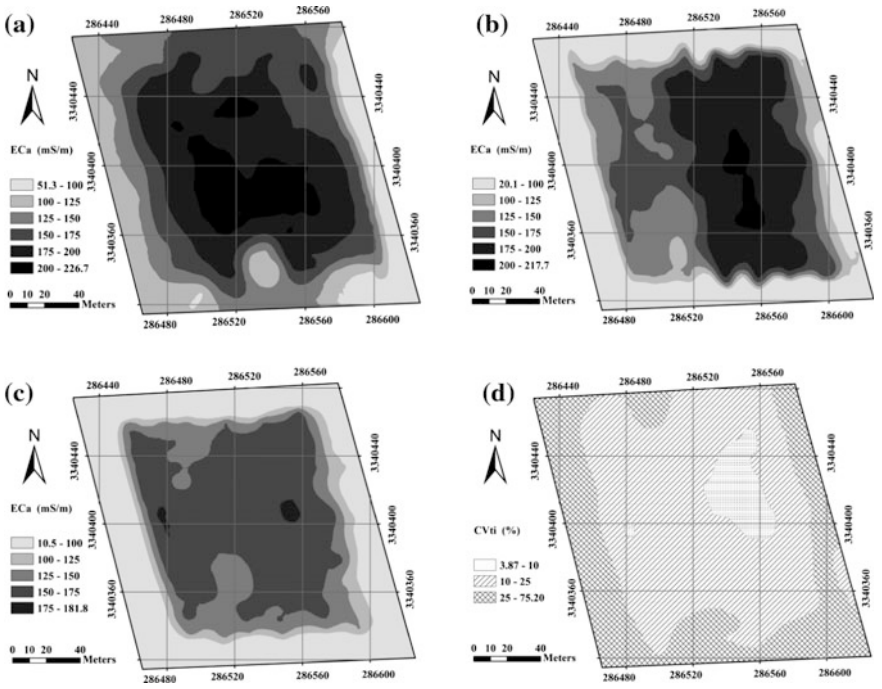

Fig. 4.4 Spatial variability of soil ECa (mS/m) in **a** 2009, **b** 2010, and **c** 2011. The plot of **d** coefficient of variation (CVti) over three years

the soil salt into deeper soil layers; and soil salt migrated with the water table to topsoil, ditches, and channels around the field. The low ECa level in the surrounding field may arise of the drainage ditches surrounding the field, and the same continuous agricultural practices (i.e., ridge building in the surroundings, irrigation, and drainage for rice) make the soil salinity content become lower in surroundings. Water table fluctuations are also mitigated by the presence of drainage ditches with an approximate depth of 0.8 m, and as such, the topsoil and subsoil ECa gradually decline. Therefore, the rate of leaching of salts will be a function of the distance to the ditches. This is consistent with the large coefficient of variation (CVt_i) values in the margins of the study area shown in Fig. 4.4d.

In order to assess the temporal stability of salinity, the calculated coefficient of variation (CVt_i) of each measurement over three years is shown in Fig. 4.4d. According to Shi et al. (2005), the variation should be considered stable with $CVt_i \leqslant 10\,\%$ and is moderately stable with $10\,\% < CVt_i < 25\,\%$ and unstable $CVt_i \geqslant 25\,\%$. Interestingly, it can be found that the area with a high salinity content (easting: 286,520–286,560 m; northing: 3,340,400–3,340,440 m) displays temporal stability, while the surrounding area shows temporal instability, especially the edges of the field with a lower salinity level. This is consistent with the reports by Shi

et al. (2005). The sharp change of salinity within the field edges may be due to the presence of irrigation ditches around the field (Fig. 4.1c) where large amounts of irrigation water allow salts to leach into deeper soils.

4.3.3 *Vertical Spatiotemporal Variability of Soil Salinity With Quasi-3D Inversion*

The Quasi-3D inversion of vertical spatial variability results is shown in Fig. 4.5. The vertical spatiotemporal variability of the soil salinity can be elaborated by the distribution of modeled σ. Seeing from the 2D cross section oriented west–east, the salinity decreased from topsoil to subsoil over the three years. For example, in 2009, it was around 200 mS/m at the depth of 0.4–1 m, while it decreased to 100–150 mS/m at the depth of 1.0–1.6 m. The consistent decrease of salinity from topsoil to subsoil is primarily determined by the annual rainfall amount (i.e., 1300 mm) and irrigation and leaching of paddy soils.

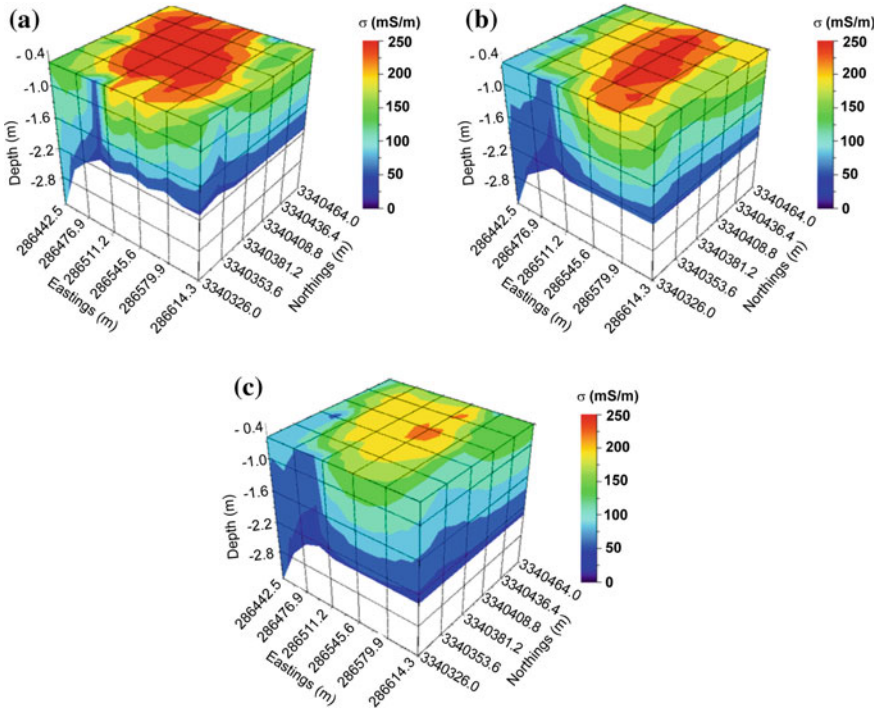

Fig. 4.5 3D models of soil electrical conductivity σ (mS/m) in **a** 2009, **b** 2010, and **c** 2011 across the study area

More specifically, the salts of the soil are found to migrate downward over the three years. For example, the area with easting from 286,511.2 to 286,545.6 m and depth from 0.5 to 1.6 m was primarily dominated by values between 100 and 150 mS/m in 2009. However, the conductivity of the area decreased to 75–150 mS/m in 2010. Furthermore, in the year of 2011, this conductivity was mostly 75–100 mS/m. This might be explained by the leaching of the salts from topsoil to subsoil, and it is also evident in 2D cross sections oriented south–north of the Quasi-3D models. The phenomenon is consistent with the vertical distribution of soil salinity of paddy field, where rice is cultivated. Additionally, the horizontal 2D cross sections at the top of the models of the three years are consistent with the kriging maps shown in Fig. 4.4. This implies that the two approaches for determining spatiotemporal variability of salinity are reliable and actually consistent with each other.

4.4 Conclusions

Determining the spatiotemporal variability of the soil salinity requires accurate and effective mapping. Using ordinary kriging, horizontal distributions of ECa over three years show the heterogeneous variability of soil salinity as well as the leaching process of salts mainly due to precipitation and irrigation. During the rice cultivation, irrigation, and drainage, land tillage results in salts leaching into deep soil depths and surrounding ditches. Quasi-3D inversion of ECa provides detailed information of the vertical variation in soil salinity. These vertical variations of salinity due to irrigation ditches are consistent with observations of Rhoades et al. (1997). Spatiotemporal variability of soil salinity in paddy fields determined by the fast, cost-effective, and efficient EMI measurements provides valuable fine-grained information for scientific research on the salinity change associated with agriculture. It can be also used as a guide for field salinity management. For example, salinity level is about 150 mS/m, soybean and cowpea can normally grow, and we can take leaching and increasing the mulching film methods to reduce the accumulation of salt in the surface soil. If the salinity level was low (85 mS/m), cauliflower can be of normal growth. Reasonable agronomic measures, such as irrigation and fertilization, can be taken to adjust and control the salt content. Considering the actual situation and farming operation convenience, flatting land can be employed to adjust the influence of microtopography on salt migration; on the other hand, the rotation (rice and dryland crop), irrigation, and drainage salinity can be used to decrease the salt content (Guo et al. 2013). These methods also can be used in large-scale management, and stand by this point, different crops and agronomic measures can be adopted in different management zones. And the deeper mode of EM38 (i.e., EM38v) as well as other EMI instruments (e.g., DUALEM-421) can be incorporated to conduct Quasi-3D inversions for deeper soil profiles (Huang et al. 2014).

Acknowledgements This study is supported by National Natural Science Foundation of China (No. 41271234; No. 41101197), the Key National Projects of High-Resolution Earth Observing System (09-Y30B03-9001-13/15), the Science-Technology Foundation for Outstanding Young Scientists of Henan Academy of Agricultural Sciences (2016YQ21), and the Independent Innovative Project of Henan Academy of Agricultural Sciences.

References

Akramkhanov A, Martius C, Park SJ, Hendrickx JMH (2011) Environmental factors of spatial distribution of soil salinity on flat irrigated terrain. Geoderma 163: 55-62

Cook PG, Walker GR (1992) Depth profiles of electrical conductivity from linear combinations of electromagnetic induction measurements. Soil Sci. Soc. Am. J. 56: 1015-1022

Corwin DL (2008) Past, present, and future trends in soil electrical conductivity measurements using geophysical methods. In: Allred, B.J., Daniels, J.J., Ehsani, M.R. (Eds.), Handbook of Agricultural Geophysics. CRC Press, Taylor and Francis Group, Boca Raton, Florida, pp. 17-44

Corwin DL, Rhoades JD (1982) An improved technique for determining soil electrical conductivity — depth relations from above ground electromagnetic induction measurements. Soil Sci. Soc. Am. J. 46: 517-520

Corwin DL, Rhoades JD (1984) Measurements of inverted electrical conductivity profiles using electromagnetic induction. Soil Sci. Soc. Am. J. 48:288-291

Corwin DL, Lesch SM (2003) Application of soil electrical conductivity to precision agriculture: theory, principles, and guidelines. Agron. J. 95: 455-471

De Groot-Hedlin C, Constable SC (1990) Occam's inversion to generate smooth, two-dimensional models from magnetotelluric data. Geophysics 55: 1613-1624

Doolittle, JA, Brevik, EC (2014) The use of electromagnetic induction techniques in soils studies. Geoderma 223-225: 33-45

Gallichand J, Buckland GD, Marcotte D, Hendry MJ (1992) Spatial interpolation of soil salinity and sodicity for a saline soil in Southern Alberta. Can. J. Soil Sci. 72: 503-516

Guo Y, Huang JY, Shi Z, Li HY (2015a) Mapping spatial variability of soil salinity in a coastal paddy field based on electromagnetic sensors. PLoS One 10(5): e0127996

Guo Y, Shi Z, Huang JY, Zhou LQ, Zhou Y, Wang LG (2015b) Characterization of field scale soil variability using remotely and proximally sensed data and response surface method. Stoch Environ. Res. Risk Assess. DOI 10.1007/s00477-015-1135-0

Guo Y, Tian YF, Wu HH, Shi Z (2013) Zoning of soil management based on multi-sources data and fuzzy-k means. Acta Pedological Sinica 50(3): 441-447 (In Chinese)

Hendrickx JMH, Borchers B, Corwin DL, Lesch SM, Hilgendorf AC, Schlue J (2002) Inversion of soil conductivity profiles from electromagnetic induction measurements. Soil Sci. Soc. Am. J. 66: 673-685.

Huang MX, Shi Z, Gong JH (2008) Potential of multi-temporal ERS-2 SAR imagery for land use mapping in coastal zone of Shangyu city, China. J. Coastal Res. 24: 170-176

Huang JY, Davies GB, Bowd D, Monteiro Santos FA, Triantafilis J (2014) Spatial prediction of exchangeable sodium percentage at multiple depths using electromagnetic inversion modelling. Soil Use Manage 30: 241-250

Lesch S M, Corwin DL, Robinson DA (2005) Apparent soil electrical conductivity mapping as an agricultural management tool in arid zone soils. Comput. Electron. Agri. 46: 351-378

Lesch SM, Strauss DJ, Rhoades JD (1995a) Spatial prediction of soil salinity using electromagnetic induction techniques. 1. Statistical prediction models: a comparison of multiple linear regression and cokriging. Water Resour. Res. 31: 373-386

Lesch SM, Strauss DJ, Rhoades JD (1995b) Spatial prediction of soil salinity using electromagnetic induction techniques. 2. An efficient spatial sampling algorithm suitable for multiple linear regression model dentification and estimation. Water Resour. Res. 31: 387-398

Li J, Heap AD (2011) A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. Ecol. Inform. 6: 228-241

Li HY, Shi Z, Webster R, Triantafilis J (2013) Mapping the three-dimensional variation of soil salinity in a rice-paddy soil. Geoderma 195-196: 31-41

Mester A, van Der Kruk J, Zimmermann E, Vereecken H (2011) Quantitative two-layer conductivity inversion of multi-configuration electromagnetic induction measurements. Vadose Zone J. 10: 1319-1330

Monteiro Santos FA, Almeida EP, Castro R, Nolasco R, Mendes-Victor L (2002) A hydrogeological investigation using EM34 and SP surveys. Earth, Planets & Space 54: 655-662

Monteiro Santos FA, Triantafilis J, Taylor RS, Holladay S, Bruzgulis KE (2010) Inversion of conductivity profiles from EM using full solution and a 1-D laterally constrained algorithm. J. Environ. Eng. Geoph. 15: 163-174

Monteiro Santos FA, Triantafilis J, Bruzgulis K (2011) A spatially constrained 1D inversion algorithm for quasi-3D conductivity imaging: application to DUALEM-421 data collected in a riverine plain. Geophysics 76: B43-B53

Padhi J, Misra RK (2011) Sensitivity of EM38 in determining soil water distribution in an irrigated wheat field. Soil Till. Res. 117: 93-102

Rhoades JD, Lesch SM, LeMert RD, Alves WJ (1997) Assessing irrigation/drainage/salinity management using spatially referenced salinity measurements. Agric. Water Manag. 35: 147-165

Robinson DA, Lebron I, Lesch SM, Shouse P (2004) Minimizing drift in electrical conductivity measurements in high temperature environments using the EM-38. Soil Sci. Soc. Am. J. 68: 339-345

Sasaki Y (1989) Two-dimensional joint inversion of magnetotelluric and dipole-dipole resistivity data. Geophysics 54: 254-262

Sheets KR, Hendrickx JMH (1995) Noninvasive soil water content measurement using electromagnetic induction. Water Resour. Res. 31: 2401-2409

Shi Z, LiY, Wang RC, Makeschine F (2005) Assessment of temporal and spatial variability of soil salinity in a coastal saline field. Environ. Geol. 48: 171-178

Shi Z, Wang K, Bailey JS, Jordan C, Higgins AH (2002) Temporal changes in the spatial distribution of some soil properties on a temperate grassland site. Soil Use Manage. 18: 353-362

Shiraz FA, Ardejani FD, Moradzadeh A, Arab-Amiri AR (2013) Investigating the source of contaminated plumes downstream of the Alborz Sharghi coal washing plant using EM34 conductivity data, VLF-EM and DC-resistivity geophysical methods. Explor. Geophys. 44: 16-24

Triantafilis J, Laslett GM, McBratney AB (2000) Calibrating an electromagnetic induction instrument to measure salinity in soil under irrigated cotton. Soil Sci. Soc. Am. J. 64: 1009-1017

Triantafilis J, Ribeiro J, Page D, Monteiro Santos FA (2013) Inferring the location of preferential flow paths of a leachate plume by using a DUALEM-421 and a Quasi-Three-Dimensional inversion model. Vadose Zone J. 12: 117-125

Vervoort RW, Annen YL (2006) Palæochannels in Northern New South Wales: inversion of electromagnetic induction data to infer hydrologically relevant stratigraphy. Aust. J. Soil Res. 44: 35-45

Viganotti M, Jackson R, Krahn H, Dyer M (2013) Geometric and frequency EMI sounding of estuarine earthen flood defence embankments in Ireland using 1D inversion models. J. Appl. Geophys. 92: 110-120

Webster R, Oliver MA (1992) Sample adequately to estimate variograms of soil properties. J. Soil Sci. 43: 177-192

Webster R, Oliver MA (2007) Geostatistics for Environmental Scientists. John Wiley & Sons, England

Chapter 5
Mapping Soil Organic Matter in Low-Relief Areas Based on Time Series Land Surface Diurnal Temperature Difference

Ming-Song Zhao, Gan-Lin Zhang, Feng Liu, De-Cheng Li and Yu-Guo Zhao

Abstract Accurate estimates of the spatial variability of soil organic matter (SOM) are necessary to properly evaluate soil fertility and soil carbon sequestration potential. In plains and gently undulating terrains, soil spatial variability is not closely related to relief, and thus, digital soil mapping methods based on soil–landscape relationships often fail in these areas. It is necessary to find new environmental variables and methods to mapping soil attribute over the low-relief areas. Time series remotely sensed data, such as thermal imagery, provide possibilities for mapping SOM in such areas. In this study, Jiangsu Province was chosen as an example in eastern China and a total of 1519 soil samples (0 $\sim$ 20 cm layer) were collected from the Second National Soil Survey of Jiangsu Province. 8-day composited land surface diurnal temperature difference (DTD) was extracted from the time series of MODIS 8-day composited land surface temperature. 8-day averaged DTD was mean of 8-day composited DTD in the same periods between 2002 and 2011. Analysis showed that SOM content was significantly negative correlated with 8-day averaged DTD of different periods, of which higher correlation was in vegetation sparse periods. Averaged DTD of many periods and averaged DTD of specific periods were selected as two group of independent variable dataset. Linear regression, regression kriging (RK), and linear mixed model (LMM) fitted by residual maximum likelihood were used to model and map SOM spatial distribution. Ordinary kriging was used as a baseline comparison. The root-mean-squared error, mean error, and mean absolute error calculated from independent validation were used to assess prediction accuracy. Results showed that LMM are the best predictions, of which LMM using DTD of specific periods and DTD cell statistics

M.-S. Zhao
School of Surveying and Mapping, Anhui University of Science and Technology, NO. 168 Shungeng Road, Huainan 232001, Anhui Province, China

M.-S. Zhao · G.-L. Zhang (✉) · F. Liu · D.-C. Li · Y.-G. Zhao
State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, NO. 71 East Beijing Road, Nanjing 210008, Jiangsu Province, China
e-mail: glzhang@issas.ac.cn

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_5

as variables performed best. RK were somewhat worse than LMM. Linear regression performed worst. This suggests that time series remotely sensed data can provide useful auxiliary variable for mapping SOM in low-relief agricultural areas and LMM improved mapping SOM spatial distribution, which provided an effective approach for improving DSM in the low-relief areas.

Keywords Digital soil mapping · Land surface diurnal temperature difference · Linear mixed model · Low-relief areas · Soil organic matter, Jiangsu province

5.1 Introduction

Soil organic matter (SOM) is a crucial soil constituent related to soil physical, chemical, and biological processes, soil fertility and agricultural productivity. SOM is also a major component of the global carbon pool (Yadav and Malanson 2007). Current digital soil mapping (DSM) methods to map SOM are mostly based on quantitative soil–landscape relationship models using easily obtained regional environmental factors (McBratney et al. 2003; Qi et al. 2006), especially geomorphometry, vegetation, land cover, and parent material. However, models based on geomorphometry perform poorly in low-relief areas such as alluvial and coastal plains (Pei et al. 2010; Santos et al. 1997; Stoorvogel et al. 2009; Zhu et al. 2010). Moreover, in old agricultural areas such as eastern China long-term cultivation has weakened the relationship between soil properties and land cover (Ding et al. 1989; Zhu et al. 2010), and therefore, DSM methods based on soil–landscape relationships using geomorphometry and land cover as predictors are often ineffective in these areas.

Recently, some attempts have been made to map SOM in plains using DSM techniques and other predictors, such as using multi- and hyper-spectral remote sensing (RS) (Stevens et al. 2010) and the soil line Euclidean distance calculated from near-infrared remotely sensed data (Fox and Sabbagh 2002). Direct sensing of the soil has three disadvantages: (1) the soil surface is often obscured by vegetation; (2) the land surface may be obscured by clouds; and (3) only the few millimeters surface are sensed.

With the development of multi- and hyper-temporal RS, attempts have been made to use time series analysis to model spatial variability of soil properties. Chang et al. (2003) used the brightness temperature of multitemporal RS to identify soil texture in the southern Great Plains of North America based on an artificial neural network applied to multiple drying cycles. Zhu et al. (2010) developed a method called land surface dynamic feedbacks (LSDF) based on moderate-resolution imaging spectroradiometer (MODIS) imagery to differentiate the spatial variability of soil type after a major rainfall event in low-relief areas with partial vegetation cover in Heilongjiang and Xinjiang, China. Liu et al. (2012) mapped soil texture (sand, silt, and clay content) using LSDF derived from MODIS after a major rain event in

south-central Manitoba, Canada. Wang et al. (2012) predicted soil texture in Jiangyan using the changing pattern of land surface diurnal temperature difference (DTD) derived from MODIS land surface temperature (LST), based on fuzzy-c-means clustering method. These researches suggested that soil properties that affect water content can be related to LST, DTD, and their change pattern.

The theory behind these results is as follows. Water has a much higher thermal capacity than mineral or organic matter in soils, so that wetter soils have higher thermal capacity, given a constant composition (Verstraeten et al. 2006). Thus, intra-day changes of LST are reduced because of the increased thermal inertia; this is reflected in lower DTD. Wet soils also have slower decomposition of organic matter. Thus, the hypothesis is that in the long term, soils showing low DTD have high SOM content, and vice versa. Further, clay has a higher thermal inertia than sand; this fact implies a positive feedback to the moisture effect just noted: finer-textured soils retain more moisture and hold it more tightly due to their finer pore-size distribution. Further, clay provides both physical and chemical mechanisms protecting SOM from microbial breakdown, while soils high in sand generally have higher mineralization rates and thus lower SOM content (Hook and Burke 2000). The question remains to what degree these theoretical differences can be seen by RS.

Based on the direct, indirect, and interactive relationships between SOM, soil moisture, soil texture, and the change of surface soil temperature, our hypothesis is that time series DTD could reflect the spatial variability of SOM in the long term. The objectives of this study were (1) to examine this hypothesis and how much information of SOM can be explained by appropriately chosen DTD in low-relief agricultural areas and (2) to predict SOM content by regression kriging (RK) and linear mixed model (LMM).

5.2 Materials and Methods

5.2.1 Description of the Study Area

Jiangsu Province, located in eastern China (116° 18′–121° 57′ E, 30° 45′–35° 20′ N), was selected as the study area (Fig. 5.1). Jiangsu covers a total area of 10.26×10^4 km^2 with 69 % in plains, 14 % in low mountains and hills, and 17 % in lakes and rivers. The climate is characterized, by a typical transition from subtropical to temperate. The mean annual temperature ranges from 13 to 16 °C, mean annual precipitation ranges from 800 to 1200 mm, with an increasing trend from northwest to southeast of Jiangsu. The elevation is generally less than 40 m above sea level, except low hills in the northeast and southwest of Jiangsu. Double-cropping systems of rice–wheat (or rapeseeds) rotation in the middle and southern parts and wheat–maize (or soybean) rotation in the northern part dominate land use.

5.2.2 Soil Samples and Analysis

The soil dataset were obtained from the Second National Soil Survey of China conducted in the 1980s including typical soil profiles. The soil profile information includes sampling site, soil physical and chemical properties. This study focused on the topsoil SOM content (0–20 cm). A total of 1519 soil profiles were selected to represent all land use types and soil types (Fig. 5.1). 302 samples were randomly selected as the validation samples to evaluate the predictions, and 1217 remaining samples were training samples. SOM content was determined using wet combustion (Walkley–Black method).

5.2.3 Acquisition of DTD and Processing

8-day composited MODIS LST products (MOD11A2, 1 km resolution) were obtained from NASA LAADS Web (http://ladsweb.nascom.nasa.gov/data/search.html), including the day and night LST. 8-day composited DTD was derived from the composited day LST minus night LST. This research obtained 8-day composited LST of ten years (2002–2011), and corresponding DTD were calculated.

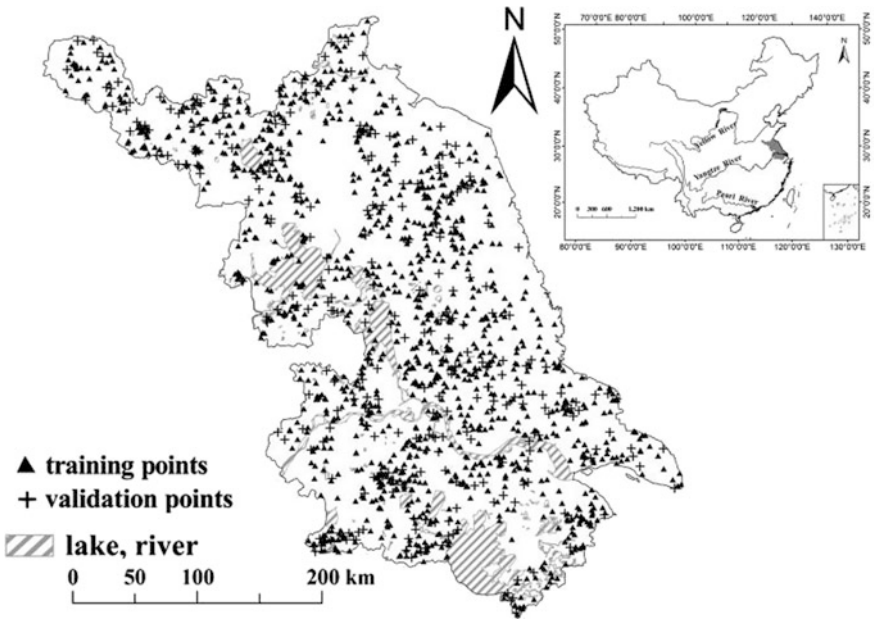

Fig. 5.1 Map of location, sampling sites of Jiangsu Province

Year	Period 1	Period 2	.....	Period 45	Period 46
2002	DTD2002001	DTD2002009	.....	DTD2002353	DTD2002361
2003	DTD2003001	DTD2003009	.....	DTD2003353	DTD2003361
⋮	⋮	⋮	⋮	⋮	⋮
2010	DTD2010001	DTD2010009	.....	DTD2010353	DTD2010361
2011	DTD2011001	DTD2011009	.....	DTD2011353	DTD2011361
	↓ Mean	↓ Mean		↓ Mean	↓ Mean
Mean DTD	DTD001	DTD009	.....	DTD353	DTD361

Fig. 5.2 Flow chart of DTD data processing

8-day averaged DTD was mean of 8-day composited DTD in the same periods between 2002 and 2011. For example, DTD001 was mean of DTD at period 1 from 2002 to 2011, the rest by analogy (Fig. 5.2). Finally, this research got a total of 33 average composited DTD at 33 periods because of poor data quality. These processed DTD can reflect long-term soil hydrothermal regimes and variation.

5.2.4 Linear Mixed Model

We use a multivariate linear model to fit the presumed deterministic component of the universal model. Here $\boldsymbol{Y}$ is the dependent variable, $\boldsymbol{X}$ is design matrix of independent variables, and $\boldsymbol{\beta}$ is the coefficient matrix. If the residual ε is independently and identically distributed with the same variance σ^2 as shown in Eq. (5.1), we can estimate the coefficients by ordinary least squares (OLS):

$$Y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I) \tag{5.1}$$

However, we cannot in general assume this structure for the residuals; rather, we assume they have a structure, as shown in Eq. (5.2). This results in the LMM as shown in Eq. (5.2):

$$Y = X\beta + \eta, \eta \sim N(0, V) \tag{5.2}$$

The residuals are considered themselves a random variable that represents both the spatial structure of the residuals from the fixed-effects model, and the unexplainable (short-range or measurement uncertainty) noise; the latter corresponds to the noise σ^2 of the linear model of Eq. (5.1). The new element here is V, a positive-definite variance–covariance matrix of the model residuals.

In the case of spatial correlation, we ensure positive definiteness by using an authorized covariance function C and assuming that the entries are completely determined by the distance between two points i and j:

$$V_{i,j} = C(x_i - x_j) \tag{5.3}$$

Further constraints on these equations and the solution are presented clearly by Lark and Cullis (2004) and several case studies (Lark et al. 2006; Lark 2012). These are called mixed models: Some effects (β) are fixed effect and others (η) are random effects but follow a known covariance structure. REML is used to maximize the likelihood of both sets of parameters (fixed β and random η) at the same time. The only prerequisite is to select the functional form of the covariance model. This is generally estimated by visual inspection of the residual variogram from an OLS model.

5.2.5 Data Processing and Analyzing

To make predictions, we (1) modeled the fixed effects using the **gls** function of the R (R Development Core and Team 2010) **nlme** "linear and nonlinear mixed effects models" package (Pinheiro and Bates 1996) using the REML option; (2) fitted a model of spatial correlation of the regression residuals considered as random effects, with the R **gstat** package (Pebesma 2004); (3) used SK on the GLS regression residuals, using the fitted variogram model; and (4) added the SK predictions and variances to the GLS predictions and variances at each prediction point.

Independent validation was used for model evaluation, using three indices: mean error (ME), mean absolute error (MAE), and root-mean-squared error (RMSE). ME measures the prediction bias, while RMSE and MAE both measure how close the prediction is to reality.

5.3 Results and Discussion

5.3.1 Exploratory Data Analysis

Table 5.1 shows the statistics of SOM content in Jiangsu. SOM content ranges from 1.3 to 52.4 g/kg, with the mean of 16.55 g/kg belonged to a medium level. The CV of SOM indicating the degree of variation and dispersion is 51.36 %, which belongs to moderate variability. The frequency distribution of SOM content follows a lognormal distribution.

There is clear global spatial structure of SOM content in the Jiangsu, showing a general increasing trend from north to south (Fig. 5.3). Some of this may be due to

Table 5.1 Summary statistics of SOM in Jiangsu Province (g/kg)

Range	Mean	Skewness	Kurtosis	CV[a] (%)
1.3–52.4	16.55 ± 8.49	1.12 (−0.48)[b]	1.75 (0.57)[b]	51.36

[a]*CV* Coefficient of variation
[b]Value in brackets was log transformed (n = 1.519)

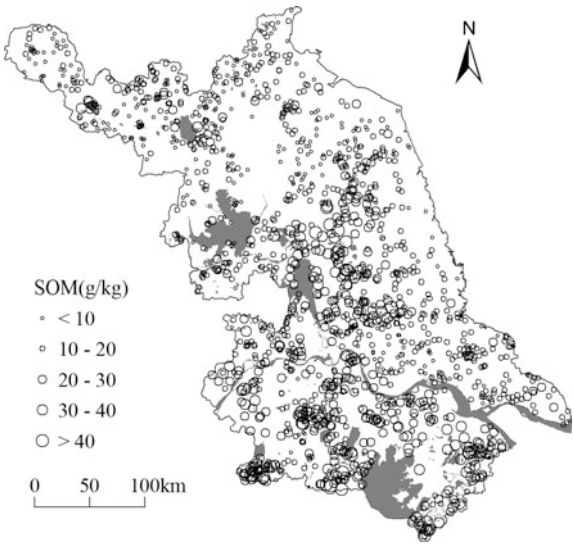

Fig. 5.3 Post-plot of SOM content in Jiangsu Province

soil environment and clay content. Different soil environment, such as soil moisture and temperature regimes, significantly alters organic matter accumulation and decomposition dynamics. High clay content in soil provided more physical protection (Hook and Burke 2000). In the central and southern part, it covers with dense river network and shallow underground water level; thus, soil environment is more humid in the long term. Additionally, clay and clay loam mainly distributed in these regions. These are contributed to SOM accumulation.

5.3.2 Relationship Between DTD and SOM

SOM content was negatively correlated with all composited DTD ($p = 0.01$) (Fig. 5.4). Scatter plots of SOM content versus composited DTD value show diffuse but highly significant negative linear correlations. Correlation was the best for DTD313 at autumn and became poorer as vegetation canopy coverage became dense, indicating the importance of selecting the proper DTD image. For paddy field, correlation between SOM and DTD is good, and correlation for dry land is poor, indicating that land use would affect the relation between DTD and SOM to a certain degree.

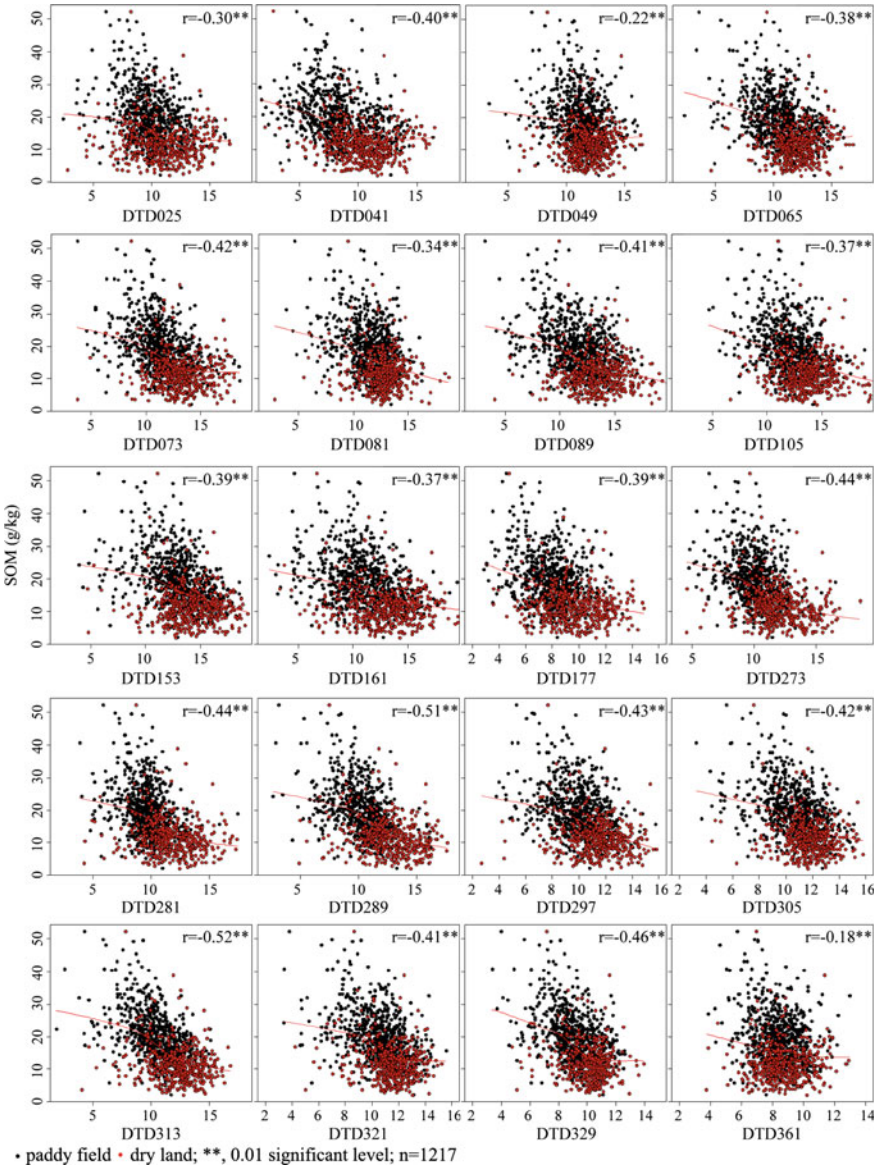

Fig. 5.4 Scatter plots of DTD and SOM content colored by land use, with empirical *smoothed lines*

Pearson correlation coefficients varied with time series (Fig. 5.5). High correlation was at autumn and winter, especially DTD289 and DTD313 had the highest correlation coefficient. At these periods, land covered with sparse vegetation. More diurnal variation of the LST would result in high DTD, which made the soil

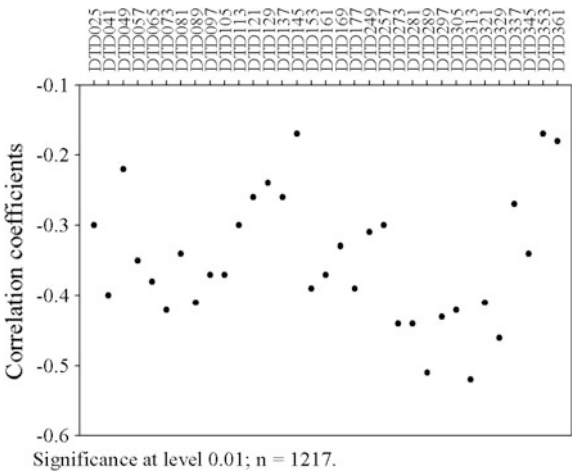

Fig. 5.5 Pearson correlation coefficients varied with time series

temperature rising and falling easily, meanwhile made SOM decomposition faster. Therefore, DTD and SOM content showed a negative correlation.

Most of the average composited DTD images show a decreasing trend from the north to the south (Fig. 5.6), which was opposite to the trend of SOM content (Fig. 5.3); this is the expected negative relation between DTD and SOM. Some may be due to geographical environment and soil texture. The central and south parts are plain and depression, with mean elevation less than 3 m. In these regions, it is covered with dense river network and lakes, with the drainage density about more than 2.01 km/km^2. The underground water level is shallow with the value of 0.6–0.8 m. Thus, soil environment is more humid in the long term, resulting in bigger soil thermal capacity and lower DTD. But in the north part, elevation ranged from 40 to 10 m with a decreasing trend from west to east gradually. It is covered with drainage density about 1.21 km/km^2 and deeper underground water level. Additionally, sandy soil and sandy loam mainly distribute in this region, and it has loose soil structure and much soil pore space. These lowers soil thermal capacity, leading to higher DTD.

5.3.3 Mapping of SOM Based on DTD and Linear Mixed Model

Linear regression analysis was used to identify the relationship between DTD and SOM content. This research adopted two solutions to selected DTD for modeling SOM. The first solution was that all DTD images were selected to model SOM by multiple linear stepwise regression. Only thirteen DTD images were selected as covariates for prediction at last (OLS_1 in Table 5.2), and OLS1 model could explain 39 % of SOM spatial variability. The second solution was that six DTD images at

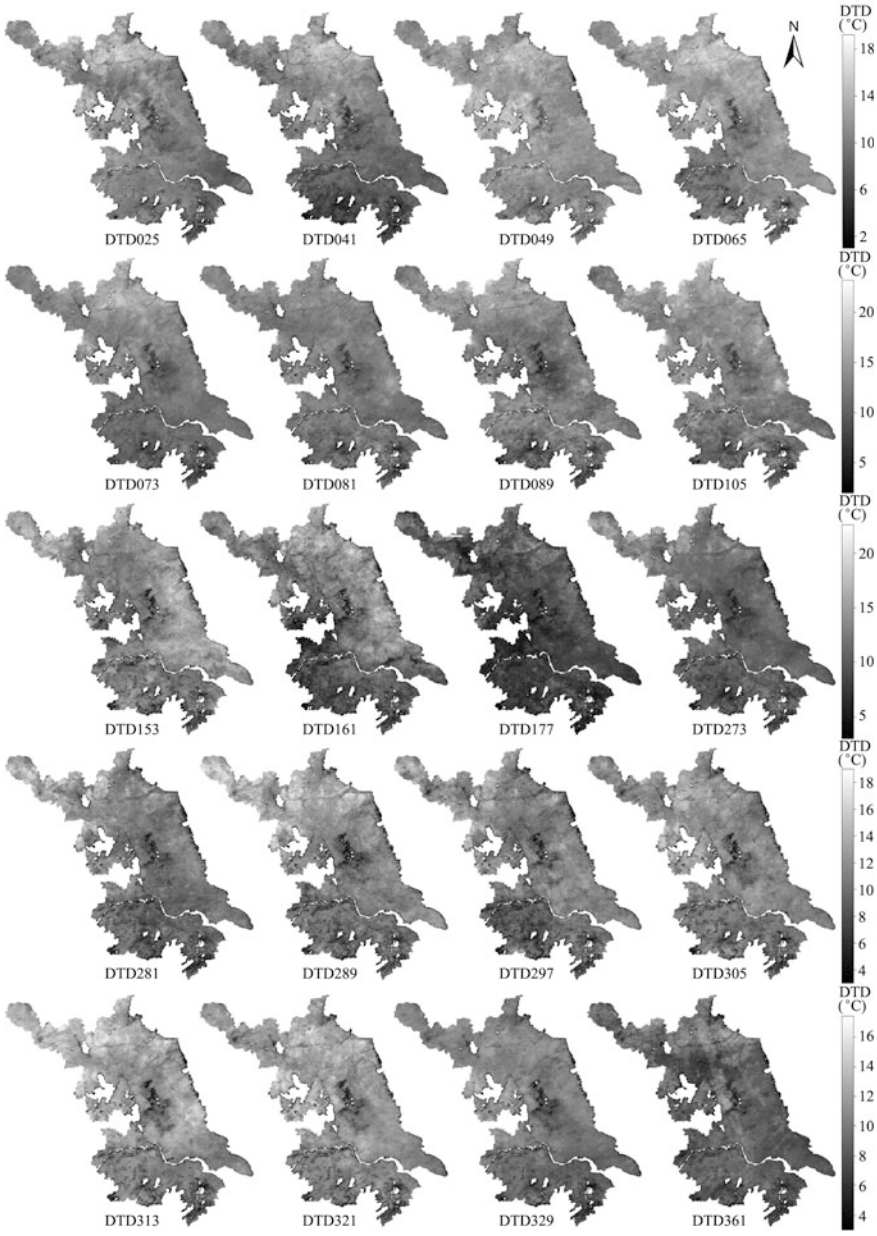

Fig. 5.6 Acquired 8-day composited and averaged DTD data

sparse vegetation canopy periods were selected for prediction by multiple linear regression (OLS_2 in Table 5.2). Although OLS_2 only had six variables, it could explain 34 % of SOM variability, indicating that selecting of proper DTD images

Table 5.2 Results of the models using OLS and GLS in Jiangsu

Variables	Coefficient	OLS_1	GLS_1	OLS_2	GLS_2
Intercept	β_0	31.50	25.80	39.48	29.38
DTD025	β_1	0.65	0.94	1.48	1.21
DTD145	β_2	0.57	−0.12	0.62	−0.09
DTD177	β_3	−0.54	−0.15	−0.58	−0.19
DTD289	β_4	−1.02	−0.27	−0.85	−0.13
DTD313	β_5	−1.40	−1.03	−1.53	−0.98
DTD329	β_6	−2.40	−1.40	−1.43	−1.13
DTD049	β_7	1.15	0.70	–	–
DTD361	β_8	0.86	0.04	–	–
DTD081	β_9	0.77	0.49	–	–
DTD089	β_{10}	−0.77	−0.32	–	–
DTD153	β_{11}	−0.40	−0.19	–	–
DTD281	β_{12}	0.67	0.32	–	–
DTD353	β_{13}	0.54	−0.10	–	–
R^2_{Adj}		0.39	–	0.34	–
St. Error		6.69	8.89	7.01	7.78
F-value		53.9	–	103.8	–
P-value		<0.001	<0.001	<0.001	<0.001
AIC		8090.89	7944.03	8193.93	7939.81

could fit SOM model well and avoid difficulty of collecting continuous time series image.

Two OLS and GLS models of SOM based on the two group covariates were fitted (Table 5.2). There are very large differences in coefficients between the OLS and GLS fit: (1) much lower intercept (β_0) for GLS fits; (2) much smaller regression slopes for GLS fits, indicating SOM does not change as much with DTD when spatial correlation among observations is accounted for. This shows that the OLS fit, which assumes independent residuals, is not appropriate in this area, with strong spatial autocorrelation of the residuals; this is confirmed by the substantially lower AIC values for the GLS models.

The GLS models were used for further analysis. Figure 5.7 shows the empirical variograms of the residuals from the LMMs, their models, and fitted parameters. Figure 5.7e shows the empirical variogram for raw SOM, along with their fitted models. Compared with variogram of raw SOM, nugget-to-sill ratio of GLS residual and effective range both became bigger, indicating spatial dependence lowers. Again, residual variograms of OLS and GLS were quite different: For GLS residual, the moderate nugget-to-sill ratio (68.06 and 67.66 %) and effective range more than 100 km show moderate spatial dependency; for OLS residual, a much higher nugget-to-sill ratio (86.35 and 84.62 %) and effective range lesser than 100 km show weak dependence.

The residuals from the LMMs are the random effects (η) in Eq. (5.3). The random effect depends on the covariance structure of the residuals with respect to points being

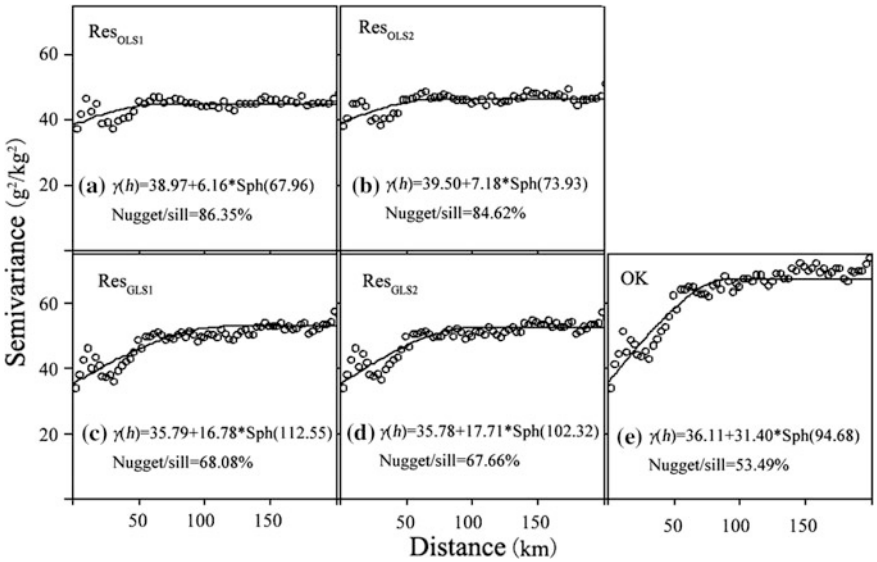

Fig. 5.7 Variograms and parameters for residuals and raw SOM (**a–d** residuals of OLS_1, OLS_2, GLS_1 and GLS_2; **e** raw SOM)

predicted, and among them. These variograms for different residuals were slightly different. This is consistent with the slight change from OLS to GLS models in this area. The nugget and sill of residuals from fixed-effect models became little lower when all DTD and selected DTD images were adding to the fixed-effects model GLS in sequence, and the spatial dependence, weakened substantially (Fig. 5.7c, d). The nugget-to-sill ratio increased when more fixed effects were explained, which was agreement with Chai et al. (2008). Comparing with the ordinary variogram of SOM, the spatial correlation distances for residuals were increased when the fixed effects were removed. The similar was reported by Chai et al. (2008).

The residuals from fixed effects of the LMMs and OLS model were interpolated by SK, using the fitted variogram models. These predictions were added to the fixed-effects predictions to get the final predictions (Fig. 5.8). Ordinary kriging was used to compare SOM mapping.

5.3.4 Comparison of Spatial Predictions and Validation

Table 5.3 shows the summary statistics of predicted SOM and Fig. 5.8 shows the spatial predictions. All predictions show a generally similar spatial distribution pattern: high SOM content in the central and south part with a decreasing trend from south to north, consistent with the post-plots (Fig. 5.3). However, the details differ. Predictions by OK are over-smoothed both spatially and in the attribute range.

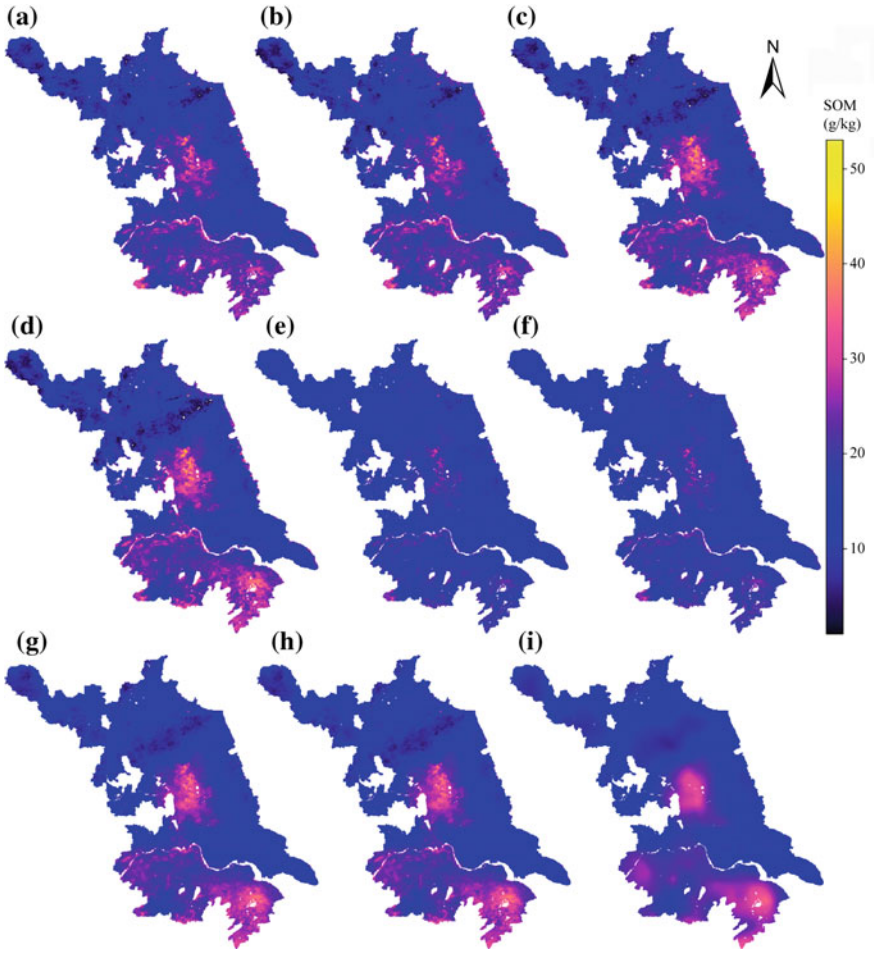

Fig. 5.8 Predictions of SOM content (**a**–**d** OLS_1, OLS_2, RK_1, RK_2; **e**–**h** GLS_1, GLS_2, LMM_1, LMM_2; **i** OK)

Predictions by RK and LMM have a wider range, similar quartiles and standard deviations, much closer to the original value range (Tables 5.1 and 5.3). Predictions by GLS are over a much narrower range than the known points. This is because when spatial structure of residual was not accounted for, the fixed-effect models correctly represented the limited predictive power of DTD as a deterministic predictor (Fig. 5.8), which is the best predictor if we have no local observations. Adding these, with their known spatial structure, in a universal model of soil variation, results in the most accurate map. Predictions by LMM were the best result.

Validation results are shown in Fig. 5.9. GLS models are worse, because it does not account for the observations' spatial structure. RK models are also worse than OK. LMMs are somewhat better than OK. In LMM, fixed effects by GLS_1 and

Table 5.3 Summary statistics of predicted SOM

SOM	Min	1st Qu[a]	Median	Mean	3rd Qu[b]	Max	SD[c]
SOM_{OLS1}	−1.28	12.38	16.08	16.41	20.30	48.11	5.42
SOM_{OLS2}	−1.85	12.53	16.36	16.45	20.27	50.60	5.55
SOM_{GLS1}	3.69	13.22	15.04	15.30	17.14	37.47	2.91
SOM_{GLS2}	5.41	14.04	15.86	16.09	17.97	38.67	2.92
SOM_{RK1}	−1.69	11.60	15.43	16.44	21.00	49.00	6.13
SOM_{RK2}	−2.16	11.64	15.46	16.46	21.09	48.98	6.16
SOM_{LMM1}	3.05	11.68	14.81	16.31	20.70	40.40	6.47
SOM_{LMM2}	2.98	11.72	14.86	16.34	20.76	40.29	6.61
SOM_{OK}	6.19	11.65	14.44	16.15	20.70	35.46	5.92

[a]The first quartile
[b]The third quartile
[c]Standard deviation

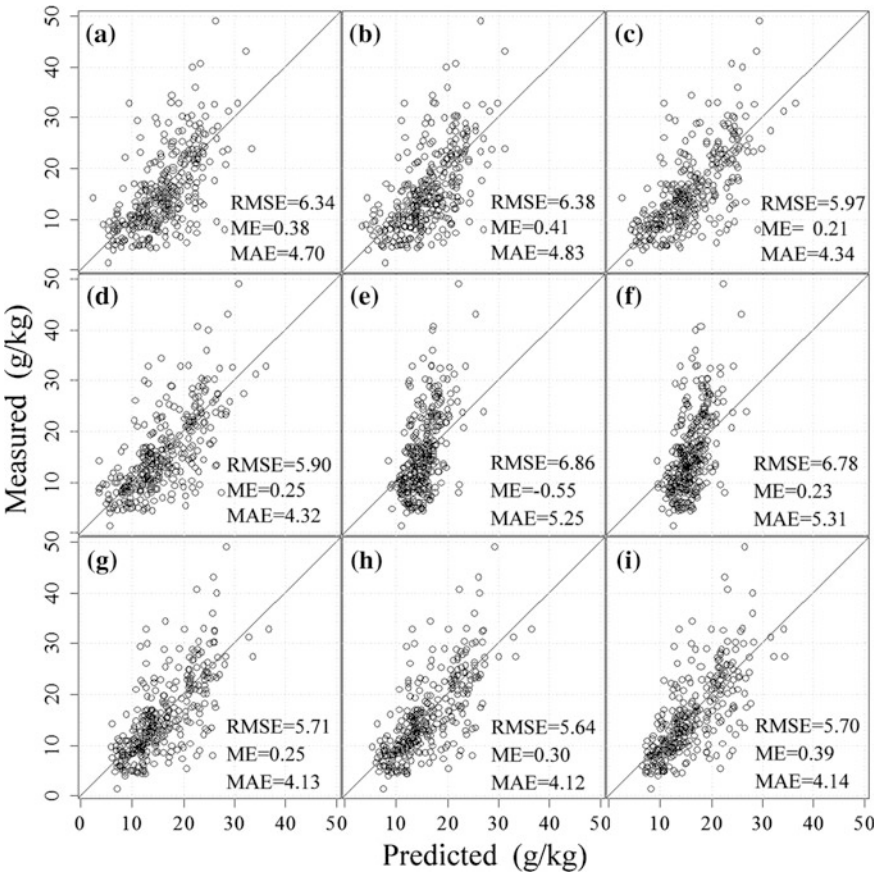

Fig. 5.9 Validation results and scatter plots (**a–d** OLS_1, OLS_2, RK_1, RK_2; **e–h** GLS_1, GLS_2, LMM_1, LMM_2; **i** OK)

GLS_2 were similar, and the final predictions by LMM were also similar, indicating that only selecting fewer DTD images of proper periods also got good prediction by LMM. Although RK and LMM have the similar predicting form, LMM performs better. This is because when fixed effect predicted by GLS, the spatial structure of random effect was considered by REML.

The method explained in this paper has inherent limits to its accuracy. The most obvious limit is the use of a MODIS pixel over a heterogeneous land surface, i.e., an average DTD. This accords with the results of Bartholomeus et al. (2011), who reported that the applicability of imaging spectroscopy in mapping soil organic carbon decreased rapidly when fields were partially covered with vegetation; Ben-Dor et al. (2009) report a similar result. Chinese agricultural practices with household as a unit, lacking unified management, lead to diversities of harvest time and land cover conditions. Although this study chose an appropriate observation period with sparse vegetation cover to avoid the impact as possibly, these factors will affect the relationship between DTD and SOM, which can bring uncertainty into SOM prediction. It needs to consider rectifying impacts of these factors on DTD for more accurate prediction and in large areas in further research.

5.4 Conclusions

This research examined the hypothesis that DTD extracted from MODIS LST could be used as an environmental covariable for mapping SOM in low-relief areas. This hypothesis was confirmed in Jiangsu, with the caution that proper DTD image must be selected, as the relationship between DTD and SOM weakened as the vegetation canopy became dense. Results showed that LMMs are the best predictions, of which LMM using DTD of specific periods as variables performed best. RK were somewhat worse than LMM. This suggests that time series remotely sensed data can provide useful auxiliary variable for mapping SOM in low-relief agricultural areas and LMM improved mapping SOM spatial distribution, which provided an effective approach for improving DSM in the low-relief areas.

Acknowledgments Project supported by the National Natural Science Foundation of China (41130530) (No. 41501226), the Foundation of State Key Laboratory of Soil and Sustainable Agriculture (Y412201431), the International Science and Technology Cooperation Project of China (2010DFB24140), and the "Strategic Priority Research Program" of the Chinese Academy of Sciences (XDA0505050303).

References

Bartholomeus H, Kooistra L, Stevens A, van Leeuwen M, van Wesemael B, Ben-Dor E, Tychon B (2011) Soil Organic Carbon mapping of partially vegetated agricultural fields with imaging spectroscopy. Int J Appl Earth Obs 13:81-88.

Ben-Dor E, Chabrillat S, Dematte JAM, Taylor GR, Hill J, Whiting ML, Sommer S (2009) Using Imaging Spectroscopy to study soil properties. Remote Sens Environ 113:S38-S55.

Chai XR, Shen CY, Yuan XY, Huang YF (2008) Spatial prediction of soil organic matter in the presence of different external trends with REML-EBLUP. Geoderma 148:159-166.
Chang DH, Kothari R, Islam S (2003) Classification of soil texture using remotely sensed brightness temperature over the southern great plains. IEEE T Geosci Remote 41:664-674.
Ding YX, Xu SR, Zhu KG (1989) Application of remote sensing techniques on 1:500,000 soil mapping in Nanjing, Jiangsu Province, China. Soils (in Chinese) 6:304-306.
Fox GA, Sabbagh GJ (2002) Estimation of soil organic matter from red and near-infrared remotely sensed data using a soil line Euclidean distance technique. Soil Sci Soc Am J 66:1922-1929.
Hook PB, Burke IC (2000) Biogeochemistry in a short grass landscape: Control by topography, soil texture, and microclimate. Ecology 81: 2686-2703.
Lark RM (2012) Towards soil geostatistics. Spatial Statistics 1:92-99.
Lark RM, Cullis BR (2004) Model-based analysis using REML for inference from systematically sampled data on soil. Eur J Soil Sci 55:799-813.
Lark RM, Cullis BR, Welham SJ (2006) On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur J Soil Sci 57:787-799.
Liu F, Geng XY, Zhu AX, Fraser W, Waddell A (2012) Soil texture mapping over low relief areas using land surface feedback dynamic patterns extracted from MODIS. Geoderma 171–172: 44-52.
McBratney AB, Santos MLM, Minasny B (2003) On digital soil mapping. Geoderma 117:3-52.
Pebesma EJ (2004) Multivariable geostatistics in S: the gstat package. Comput Geosci 30:683-691.
Pei T, Qin CZ, Zhu AX, Yang L, Luo M, Li BL, Zhou CH (2010) Mapping soil organic matter using the topographic wetness index: A comparative study based on different flow-direction algorithms and kriging methods. Ecol Indic 10:610-619.
Pinheiro JC, Bates DM (1996) Unconstrained parametrizations for variance-covariance matrices. Stat Comput 6:289-296.
Qi F, Zhu AX, Harrower M, Burt JE (2006) Fuzzy soil mapping based on prototype category theory. Geoderma 136:774-787.
R Development Core and Team (2010) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org; accessed 11/2/2012.
Santos MLM, Guenat C, Thevoz C, Bureau F, Vedy JC (1997) Impacts of embanking on the soil-vegetation relationships in a floodplain ecosystem of a pre-alpine river. Global Ecol Biogeogr Lett 6:339-348.
Stevens A, Udelhoven T, Denis A, Tychon B, Lioy R, Hoffmann L, van Wesemael B (2010) Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma 158:32-45.
Stoorvogel JJ, Kempen B, Heuvelink GBM, de Bruin S (2009) Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. Geoderma 149:161-170.
Verstraeten WW, Veroustraete F, van der Sande CJ, Grootaers I, Feyen J (2006) Soil moisture retrieval using thermal inertia, determined with visible and thermal spaceborne data, validated for European forests. Remote Sens Environ 101:299-314.
Wang DC, Zhang GL, Pan XZ, Zhao YG, Zhao MS, Wang GF (2012) Mapping soil texture of a plain area using fuzzy-c-means clustering method based on land surface diurnal temperature difference. Pedosphere 22:394-403.
Yadav V, Malanson G (2007) Progress in soil organic matter research: litter decomposition, modelling, monitoring and sequestration. Prog Phys Geog 31:131-154.
Zhu AX, Liu F, Li BL, Pei T, Qin CZ, Liu GH, Wang YJ, Chen YN, Ma XW, Qi F, Zhou CH (2010) Differentiation of soil conditions over low relief areas using feedback dynamic patterns. Soil Sci Soc Am J 74:861-869.

Chapter 6
Mapping Soil Thickness by Integrating Fuzzy C-Means with Decision Tree Approaches in a Complex Landscape Environment

Yuanyuan Lu, Ganlin Zhang, Yuguo Zhao, Decheng Li, Jinling Yang and Feng Liu

Abstract Predictive soil mapping depends on understanding the relationships between soil properties and environmental factors. However, in a complex soil landscapes, there is a shortage of suitable approaches to establish these relationships. The main objective is to predict soil thickness in an alpine watershed relating to soil environmental factors based on an unsupervised fuzzy clustering method (fuzzy c-means, FCM) and decision tree (DT) method. In this study, FCM method was used for stratifying the landscape, and then, a representative soil thickness was assigned to each class. For each class, a number of points were randomly chosen in proportion to representative areas, and then, the environmental factors at these point locations were extracted as a training data set (3626 points). For the training data set, DT method was used to obtain the critical threshold of soil–environment relationships. Finally, soil thickness map was created by applying the results of the DT across the region. An independently collected field sampling set (31 points) was used to evaluate the effectiveness of the proposed approach. For training set, 95.48 % of the total training data were correctly predicted. For validation set, the overall accuracy and Kappa coefficient could reach 74.2 % and 0.659, respectively. Evaluation accuracy of soil map was up to 74.2 %. In conclusion, it is suggested that soil–landscape modeling using FCM and DT methods can be efficiently used as a valuable research technique for spatial soil thickness prediction in a complex soil landscape where soil characteristics and properties are not available.

Keywords Soil thickness · Critical threshold of soil environment · Knowledge about soil–environment relationships · Predictive soil mapping

Y. Lu · G. Zhang (✉) · Y. Zhao · D. Li · J. Yang · F. Liu
State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China
e-mail: glzhang@issas.ac.cn

Y. Lu · G. Zhang
University of Chinese Academy of Sciences, Beijing 100049, China

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_6

6.1 Introduction

Soil thickness is one of the most important input parameters for hydroecological models especially in arid and semiarid regions (Buol et al. 2011; Boer et al. 1996). Meanwhile, soil thickness can directly reflect the degree of soil development and can also influence soil fertility condition and earth surface processes, such as vegetation growth, surface hydrology, and soil moisture (Zheng and Liu 2003; DeRose et al. 1991; Fuhlendorf and Smeins 1998; Meyer et al. 2007). However, soil thickness is characterized by high spatial variability, and to measure, it is laborious and time-consuming (Hudson 1992), especially in areas with complex landscape. Therefore, there is an urgent need for models to predict the spatial distribution pattern of soil thickness.

Currently, physical mechanisms, geostatistics, remote sensing inversion, and soil–landscape relationship reasoning are the four mostly used methods in predictive soil thickness mapping. The first three methods mentioned above require a long-term experiment with the in situ monitoring in small watersheds, or high demand for the quantity and distribution of sampling points, or the result can easily be affected by vegetation coverage. Therefore, these methods are difficult to be used for soil thickness mapping in areas with the condition of diverse environment and landscape ecology (Dietrich et al. 1995; Santos et al. 2000; Zhou 2012; Scull et al. 2003). In this study, relatively efficient method of soil–landscape relationships was used to obtain and establish the soil–environment relationship in a complex landscape environment. The basis of predictive soil thickness mapping based on soil–landscape relationship method is to understand and acquire the relationships between soil thickness and environmental factors, while existing models used to establish their relationships are limited (McBratney et al. 2000, 2003), especially in the complex landscape environment.

In the complex landscape environment, various environmental combinations are easy to influence the distribution of soil thickness. So it is difficult for soil scientists to clearly describe and clarify the relationships between soil thickness and environmental factors. Furthermore, it is also hard to collect sampling points according to the preconcerted layout scheme for limitations of the natural condition. In consequence, it is urgent for scientists to combine field investigation data with pedogenesis principles for obtainment of soil–environment relationships for predictive soil thickness mapping in the complex landscape environment.

In the present study, our major objective is to provide a new approach to obtain the critical threshold of soil environment and knowledge by integrating the methods of fuzzy c-means clustering (FCM) and decision tree (DT), in a complex landscape environment where traditional soil mapping methods are difficult to undertake.

6.2 Materials and Methods

6.2.1 Study Area

The study area (38° 12′ 19″N to 38° 16′ 12″N and 99° 50′ 09″E to 99° 53′ 52″E) is the Hulugou watershed, located in the upstream of Heihe River, a typical alpine basin, and covers a total area of 23.1 km^2 with rugged mountainous terrain on the southeastern part of the Qilian Mountain (Fig. 6.1a). Altitude of the region ranges from 2916 to 4600 m above the mean sea level, with a span about 1700 m, which is indicative of an extremely steep environmental gradient. The region belongs to the alpine continental climate zone and has a mean annual precipitation with an average range from 400 to 600 mm (Han et al. 2013). The representative soil types of the region are Cambisols, Primosols, Histosols, Isohumosols, and other types, according to the Chinese Soil Taxonomic Classification.

The representative characteristics of this study area are complicated and various landscapes, obvious differentiations of vertical gradient, and intensive spatial variability of soil properties. Meanwhile, the study area is located in the alpine landscape zone, which is a typical area for researches in the complex landscape environment. In conclusion, Hulugou watershed is chosen as the research area for the above reasons.

6.2.2 Data Sources

The basic geospatial data utilized in this research were all downloaded from the Cold and Arid Regions Science Data Center (http://westdc.westgis.ac.cn/). The data sets mainly included DEM and Landsat TM remote sensing image. A 30 m × 30 m

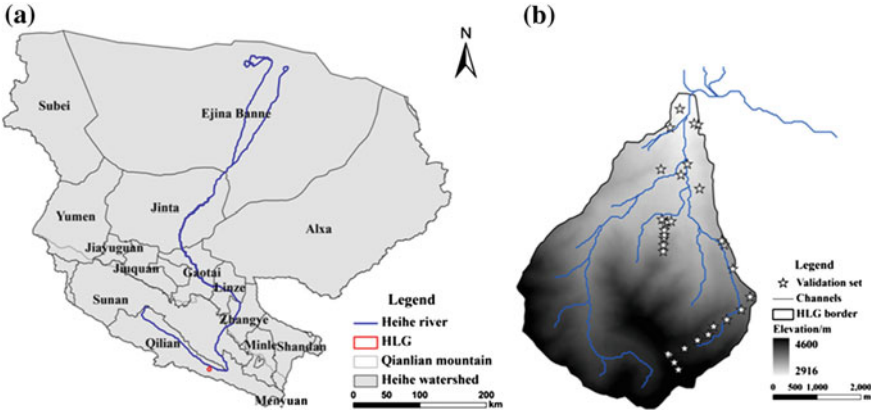

Fig. 6.1 The location and validation points of the study area. **a** Study area. **b** Validation points

DEM was obtained from the advanced spaceborne thermal emission and reflection radiometer global digital elevation model (ASTER GDEM). The primary topographic attributes such as elevation, slope, aspect, profile curvature, and plan curvature were derived from the ASTER DEM using the spatial analysis tools module in ArcGIS 9.3. The System for Automated Geoscientific Analyses (SAGA) software was used to calculate river basin topographic wetness index (TWI) (Ambroise et al. 1996). In addition, the distribution of vegetation also exhibited obvious differentiation under the influence of differentiation of elevation gradient. Hence, normalized difference vegetation index (NDVI) computed by infrared and near-infrared band of the Landsat TM images at 30-m resolution was used to indicate vegetation intensity (Rouse 1973).

In this study, 31 independent sampling profiles were collected as validation data set (Fig. 6.1b). Sampling points were designed to represent the range of topographic and vegetational variation in this watershed. The sampling design could cover various landscape units as much as possible. When the points were collected in the field,some descriptions were recorded in detail. The descriptions included landscape characteristics, vegetation distribution, rock content, and pedogenesis characteristics.

According to the division of soil thickness from the global workshop on digital soil mapping (GSM), soil thickness was divided into six grades in this study: 0–5 (level 1), 5–15 (level 2), 15–30 (level 3), 30–60 (level 4), 60–100 (level 5), and 100–200 cm (level 6), respectively.

6.2.3 Methodology

6.2.3.1 Construction of the Environmental Factors Database

In general, at the landscape scale, the main factors that play dominant roles in soil formation are topographical and hydrological conditions (Yang et al. 2007). Topography is an important element, which can influence the matter and energy exchange between soil and environment and can also influence other soil factors during the process of soil development (Huang 2000). Previous studies have shown that topographical attributes, such as elevation and slope, can basically represent the principal factors that influence the formation and development of soil thickness (Moore et al. 1993; Gessler et al. 1995; McIntosh et al. 2000; Park et al. 2001; McKenzie et al. 2000). Consequently, in this study, combining the previous researches with the characteristics of this study area, topographical attributes such as elevation, slope, aspect, profile curvature, plan curvature, and TWI derived from ASTER DEM were selected to construct the database of environmental factors. The distribution of derived environmental parameters in the Hulugou watershed was presented in Fig. 6.2.

In this study, statistical description and correlational analysis were executed to reduce the redundancy among multivariate. According to the correlation of

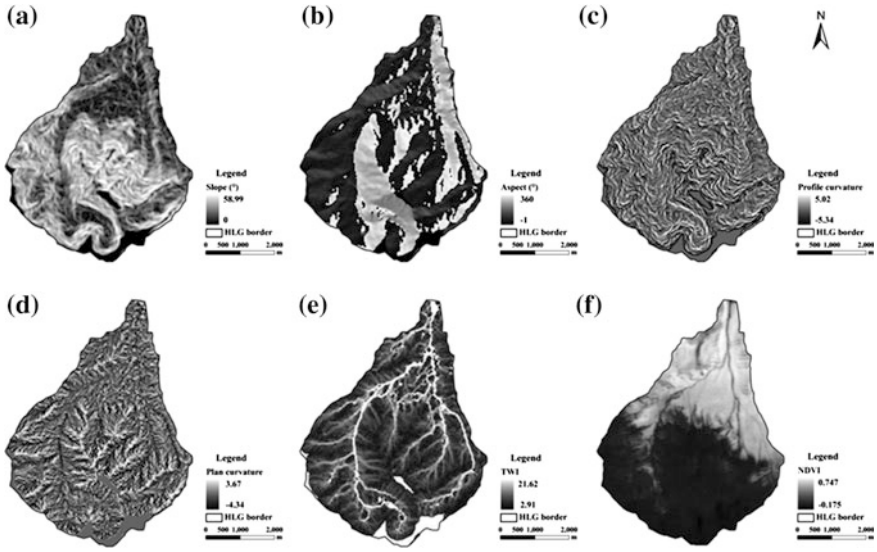

Fig. 6.2 Distribution of derived environmental parameters of Hulugou watershed. **a** Slope. **b** Aspect. **c** Profile curvature. **d** Plan curvature. **e** TWI. **f** NDVI

environmental attributes, the following three topographic variables (elevation, profile curvature, and TWI) were selected as input independent variables for fuzzy c-means cluster analysis.

6.2.3.2 Combinations Obtainment Using FCM Method

FCM is a widely used and effectively unsupervised fuzzy clustering method, which is based on the objective function, through optimization to solve fuzzy classification and clustering of the input data set (Sun et al. 2008). The principle of this method is to calculate the distance between each point and each prototype in multi-attributes-based space by using statistical methods, weighing the membership, ultimately obtaining minimum value of weighted error square and objective function (Zhu et al. 1996; Bezdek et al. 1984). The fuzzy objective functions are shown in Eqs. (6.1) and (6.2).

$$J_m(\boldsymbol{U}, \boldsymbol{v}) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m d_{ik}^2 \tag{6.1}$$

$$d_{ik}^2 = \|y_k - v_i\|_A^2 = (y_k - v_i)^{\tau} \boldsymbol{A} (y_k - v_i) \quad y \in \boldsymbol{Y} \tag{6.2}$$

Equations (6.1) and (6.2) contain a number of variables where U is the membership matrix of fuzzy clustering, v is the clustering center set, Y is the element data set for the environmental factors, n is the number of data in Y, c is the number of

cluster categories, m is weighted index (also called fuzzy degree index), d_{ik} is the weighted distance from point y_k to the cluster center point v_i, u_{ik} is the membership of the first k point belongs to the ith class, A is the distance weighting matrix, and J_m is the fuzzy classification error.

FCM was used to identify the unique combination that existed in the environmental data set. In general, the areas with high membership value were the typical combinations for various environmental conditions. The output of continuous classification was influenced by selection and determination of the number of cluster categories (c) and the weighted index (m). The partition coefficient (F) and normalized entropy (H) were used to determine the optimal value of cluster categories (Bezdek et al. 1984). Some studies have shown that the optimal weight index is in the interval between 1.5 and 2.5 (Yang et al. 2007; Odeh et al. 1992); therefore, a series of m values could be set to contrast corresponding results for choosing the optimal m value.

In this study, 3 environmental factors (elevation, profile curvature, and TWI) were calculated by FCM algorithm by applying 5 different weighed index (m = 1.5, 1.75, 2.0, 2.25, 2.5) settings. The optimal number of cluster categories was ensured by the relative change of F and H. In this study, C = 15 was chosen as the optimal number of cluster categories when m was equivalent to 1.75 by comprehensive analysis..

6.2.3.3 Extraction of the Training Data Set

The processes for extracting the training data set consist of the following steps. Firstly, the environmental factors which play dominant roles in determining soil thickness were defined, and then, the clustering center of various environmental factors combinations and membership of environmental conditions about clustering center were obtained by FCM analysis. Secondly, the general rules of soil sickness distribution were gained by analyzing distribution for different combinations of environmental factors, combined with field investigation data and pedogenesis principles. Next, the corresponding grade of soil thickness distribution under the condition of different combinations of environmental factors was determined. Then, a representative soil thickness to each class was assigned, and the corresponding relationship between soil thickness distribution and environmental factors combinations was obtained. And then, the typical areas of fuzzy membership threshold greater than a certain value (e.g., fuzzy membership > 0.5) were selected and a certain number of points which were proportional to area extent were randomly chosen. In addition, an approximated quantity of points was extracted with corresponding grade of soil thickness. Finally, the environmental factors at these typical points were extracted to build up the training data set.

In this study, elevation, slope, aspect, plan curvature, profile curvature, topographical wetness index, and NDVI were chosen as independent variables; the soil thickness grade of corresponding points was used as dependent variable.

6.2.3.4 Obtainment of Soil–Environment Relationships Using DT Method

DT algorithm is a machine-learning method used in data mining for constructing predictive model from data. The goal of DT is to create a model that predicts the value of a target variable based on several input independent variables with a tree structure. The tree is generated by partitioning the data recursively into a number of groups, each division being used to differentiate the response variable in the resulting nodes. Detailed information about the principles and characteristics of DT can be referred to papers by Loh (2011).

In this research, the C5.0 decision tree algorithm was used to predict soil thickness grade and then obtain the critical thresholds of environmental factors and the knowledge rule set of soil–environment relationships.

6.2.3.5 Evaluation of the Soil Thickness Map

Once the soil property values were calculated, they were compared with the observed soil property values to assess model performances. In order to test prediction accuracy, the validation set was used to estimate the accuracy of the model for prediction of soil thickness. Several indexes were used for quantitative assessment of models, such as the overall classification accuracy, Kappa index, producer accuracy, and user accuracy (Taghizadeh-Mehrjardi et al. 2014; Cohen 1960; Xu et al. 2011). In these indexes, the overall classification accuracy indicated the probability of consistency between classification results and the ground investigation data. Kappa index was used to estimate consistency of the categorized results. Producer accuracy was used to describe how successful the model is for prediction, and user accuracy can be used to show how well map predictions are represented in reality.

6.3 Results and Discussion

6.3.1 Extraction of the Training Data Set

The premise for acquiring the distribution rules of soil and environment was to obtain the optimal combinations of environmental factors. Fuzzy membership maps of representative combinations of environmental factors based on fuzzy c-mean clustering were shown in Fig. 6.3.

According to the analysis by running FCM algorithm, 15 environmental classes were identified to be the optimal number of classes in the study area based on the selection of partition coefficient and entropy of classification. Membership maps of the derived 15 environmental classes were generated, and locations with high

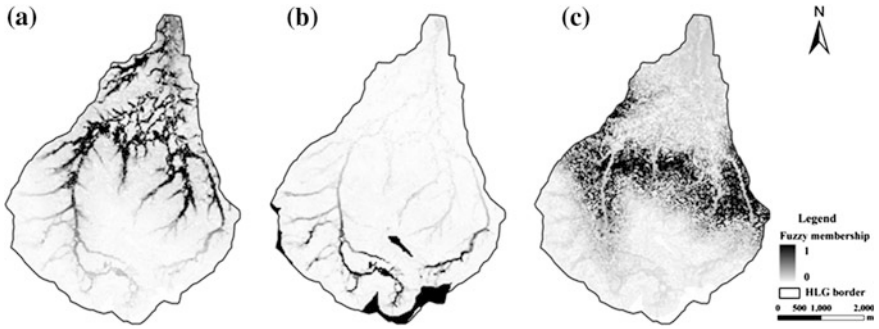

Fig. 6.3 Fuzzy membership maps of various combinations of environmental factors based on fuzzy c-mean clustering. **a** Class 1. **b** Class 7. **c** Class 8

membership values in these environmental classes were considered as locations of typical soil instances. Figure 6.3 was an example of analyzing and gaining the preliminary correspondence. As shown in Fig. 6.3a, class 1 is mainly distributed along the watercourse, surrounded by gully area, with a lower altitude than other clusters, terrain relief, relatively higher wetness index, good soil water condition, and relatively good soil development. Moreover, rock content in soil of class 1 is more sufficient than other categories due to short distance from the river. The overall soil thickness is approximately 30–60 cm, which belongs to the level 4. Class 7 mainly distributes in the high altitude alpine desert with concave slope area. Soil water condition is ordinary for class 7. From satellite images and field sampling record data, we can see that there is almost no soil development in areas covered by glaciers. So soil thickness of the corresponding category is 0, which belongs to the level 1. Meanwhile, class 8 mainly distributes in the bottom of alpine desert, which is in the transition zone between alpine desert and meadow belt. Although soil has initiatory development, the thickness is very thin in these areas. Thus, the soil thickness is mainly around 15–30 cm that belongs to the level 3. The remaining 12 clustering categories were assigned to corresponding soil thickness distribution grade using similar reasoning.

According to the principle that combinations of environmental categories can both cover the whole areas and have the smallest interaction areas, areas with clustering fuzzy membership higher than 0.5 were selected, and then typical points were extracted (according to the steps described in 6.2.3.3) to build the training data set in these areas. In this study, a total of 3626 typical points were extracted by using 15 cluster categories and 6 soil thickness grades (Fig. 6.4). It can be seen from Fig. 6.4 that the training points are relatively evenly distributed, with more points located in the transitional belt of alpine desert and meadow than in other zones. The reason for this is that in these places the variations of soil development and soil thickness are more dynamic. But nonetheless, the training data set can cover the different combinations of the soil environmental factors in this region.

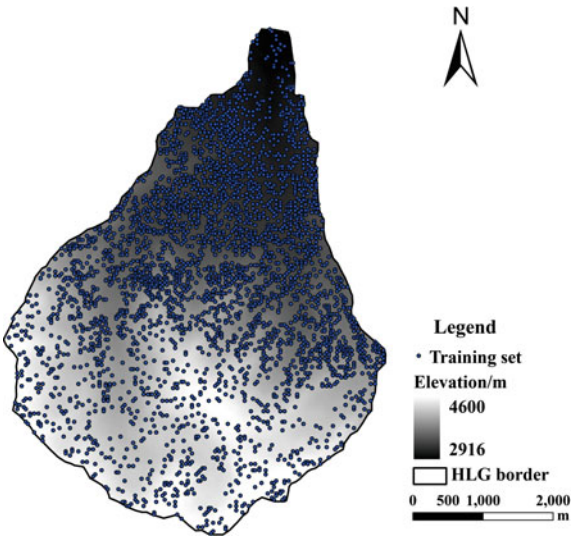

Fig. 6.4 Distribution of training points in Hulugou

6.3.2 *Knowledge Obtainment of Soil–Environment Relationships*

The formation and development of soil is the result of interactions of formative environmental factors, but there are some diversities at different spatial scales or by the influence of soil formative essentials. To build knowledge rule sets, the selection of principle for environmental factors was required. The principle for selecting environmental factors was that it can characterize the soil environmental conditions and can be easily obtained and utilized. In addition, the selected environmental factors can guide the subsequent field investigation. According to this principle, finally elevation, slope, aspect, plan curvature, profile curvature, and NDVI were chosen as input variables.

In the present study, the critical thresholds of soil environmental factors and the knowledge about soil–environment relationships were obtained by running training data set through the DT arithmetic. To obtain the minimum rate of error for classified results, debugging and building based on the C5.0 model were repeated many times for the training data set. Finally, the parameters were determined and the global pruning was used. The parameters were as follows:The pruning purity is 99 %, the number of leaf nodes is 21, and frequency for promotion and interaction is 10. The following five factors (elevation, slope, plan curvature, profile curvature, and NDVI) were chosen for constructing knowledge rule set about soil–environment relationships. However, the factor of aspect was not involved in the following steps for the reason that most of the areas in this study are located in the same direction of mountains; though there are differences for aspect, the differences are not large enough to influence the distribution of soil thickness grade. The confusion

Table 6.1 Confusion matrix of training set at various soil thickness grade

Prediction	Validation						
	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Total
Level 1	608	9	0	0	0	0	617
Level 2	0	606	1	0	0	0	607
Level 3	0	0	597	1	0	0	598
Level 4	0	0	17	511	38	37	603
Level 5	0	0	0	10	587	3	600
Level 6	0	0	0	39	9	553	601
Total	608	615	615	561	634	593	3626

matrix for training data set was shown in Table 6.1. From the table, we were able to calculate that 3462 points were correctly predicted, accounting for 95.48 % of the total training data set, and 164 points were wrongly predicted, with the percentage of 4.52 %.

About 25 rules about the knowledge of soil–environment relationships were summarized through the above analyses. Through the rules we can discover that the spatial variations of soil thickness were complicated and existed obvious cross-distribution on the space in this complex landscape environment. Overall, the main crucial factors influencing the distribution of soil thickness were elevation, followed by profile curvature and slope, while plan curvature and NDVI worked only on a smaller scale.

6.3.3 Validation of the Methods

6.3.3.1 Spatial Distribution of Soil Thickness

The rules obtained by combinations of FCM and DT method were applied to predict the distribution of soil thickness across the study area (Fig. 6.5). As shown in Fig. 6.5, almost all the areas of soil thickness in the range of 0–5 cm are distributed in high altitude areas, which is consistent with the actual field situation in this study area. This is because these areas are covered with glacier or of alpine desert landscape; there is basically no soil development. Within the range of 5–30 cm, soil thickness arises cross-distribution on the space. Because this region is located in the transition zone between alpine desert and meadow.There is a relatively preliminarily developed soil type.And the ecological environment is fragile. Besides the discrepancy of soil thickness is also influenced by micro-topography. The areas where soil thickness is more than 30 cm are located in low altitude and relatively flat district. Besides, the staggered distribution of soil thickness is influenced by the following factors, such as distance to the river, vegetation coverage, and uneven soil surface, and so on.

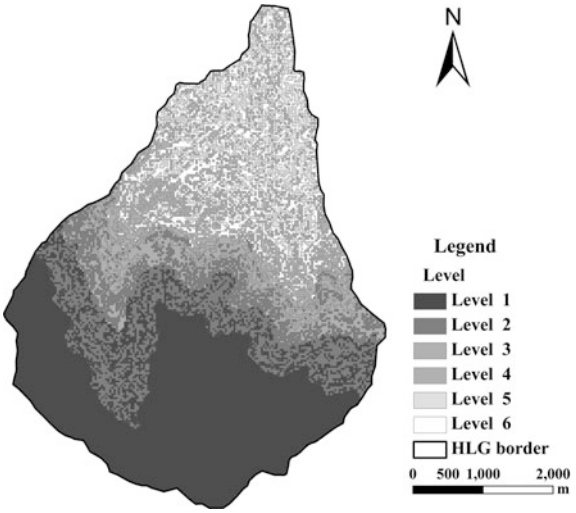

Fig. 6.5 Distribution of soil thickness of Hulugou

6.3.3.2 Evaluation of the Predictive Map

In this study, an independent set was used to examine the effectiveness of the approach for establishing the relationships between soil thickness and environmental factors. The result was presented in Table 6.2. The overall accuracy and Kappa coefficient can reach 74.2 % and 0.659, respectively. The accuracies of level 1 and level 6 were the best, followed by level 3 and level 4. It was difficult to evaluate accuracy for level 2 and level 5 due to limited number of validation points.

As shown in Table 6.2, the best accuracy was level 1, while the worst prediction was for level 2. Overall, 23 were correctly predicted out of 31 validation points. There were 6 points wrongly predicted in adjacent areas, which were related to the artificial partition of soil thickness hierarchy. In addition, another 2 points were wrongly predicted because of the locations of the points. For instance, the point, which was located in the transition zone between alpine desert and meadow, was attached to level 1, but it was wrongly predicted into level 3. Areas of soil thickness around the point that was covered with vegetation are 10–20 cm, but this was 0 for the location of the point.

Compared to results with other analogous researches (Henderson et al. 2005; McKenzie and Ryan 1999), the accuracy for this study was better than previous research. Additionally, distribution of predictive map was basically consistent with the actual distribution. For instance, Primosols was located in high altitude, while Histosols lied to the low-lying areas in intermediate altitude. Consequently, it is possible to predict the distribution of soil thickness in the complex alpine landscape environment by using the combination of FCM and DT methods.

Some restrictions or deficiencies existed when the methods in this study were used for the subsequent application in other areas.For example, the premise of the methods could be used for reference in other researches is familiar with the regional

Table 6.2 Confusion matrix of validation set at various soil thickness grade in study area

Predicted soil thickness grade	Validated soil thickness grade						Total	Producer accuracy (%)	User accuracy (%)
	1	2	3	4	5	6			
1	9	0	1	0	0	0	10	90.0	100
2	0	0	0	0	0	0	0	0	0
3	0	1	2	1	0	0	4	50.0	50.0
4	0	0	1	4	1	1	7	57.0	80.0
5	0	0	0	0	0	1	1	0	0
6	0	0	0	0	1	8	9	88.8	80.0
Total	9	1	4	5	2	10	31	74.2 %	

distribution of soil properties and possess with knowledge of pedogesis. Besides, at present, there is still no good method for choosing the number of optimal tree for classification (Scull et al. 2005). All the restrictions or deficiencies need to be further improved in subsequent research. Consequently, this study was carried out at catchment scale. The generality and transferability of the proposed method in this work to other areas still remain to be tested.

6.4 Conclusions

In this study, FCM and DT methods were integrated to construct the training data set, extract the knowledge set of soil–environment relationships, and predict the distribution map of soil thickness. This present study demonstrated the prediction of the spatial distribution of soil thickness in the complex landscape environment through the forementioned process. In application of this method, a variety of auxiliary variables derived from different sources have been used.

For the accuracy of validation set, the overall accuracy and Kappa coefficient could reach 74.2 % and 0.659, respectively, which was superior to other analogous research. The distribution of predicting soil thickness was consistent with the actual distribution. The accuracy was influenced by some uncontrolled uncertainties. The uncontrolled uncertainties were aroused by the complex local variations of soil thickness, the number of sampling points and DEM resolution. Without regard to these uncontrolled uncertainties, the accuracy in this study can be considered as an important improvement toward solving the need for distributed soil thickness information in the context of complex alpine landscape environment.

Acknowledgements Funding for this study was provided by National Natural Science Foundation of China (Project No. 41130530 and 91325301). The authors are grateful to Cold and Arid Regions Science Data Center at Lanzhou (CARD) for providing us with the basic geographic data. Furthermore, the authors would like to acknowledge the teachers and students in one team for their help and support in field investigation and soil samples collection and physical and chemical analysis.

References

Ambroise B, Beven K, Freer J (1996) Toward a generalization of the TOPMODEL concepts: Topographic indices of hydrological similarity. Water Resources Research, 32(7): 2135-2145.

Bezdek JC, Ehrlich R, Full W (1984) FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences 10(2): 191-203.

Boer M, Del Barrio G, Puigdefábres J (1996) Mapping soil depth classes in dry Mediterranean areas using terrain attributes derived from a digital elevation model. Geoderma 72(1): 99-118.

Buol SW, Southard RJ, Graham RC, McDaniel PA (2011) Soil genesis and classification. John Wiley & Sons.

Cohen JA (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20(1): 37-46.

DeRose RC, Trustrum NA, Blaschke PM (1991) Geomorphic change implied by regolith-slope relationships on steepland hillslopes, Taranaki, New Zealand. Catena 18(5): 489-514.

Dietrich WE, Reiss R, Hsu ML, Montgomery DR (1995) A process-based model for colluvial soil depth and shallow landsliding using digital elevation data. Hydrological processes 9(3–4): 383-400.

Fuhlendorf SD, Smeins FE (1998) The influence of soil depth on plant species response to grazing within a semi-arid savanna. Plant Ecology 138(1): 89-96.

Gessler PE, Moore ID, McKenzie NJ, Ryan PJ (1995) Soil-landscape modelling and spatial prediction of soil attributes. International Journal of Geographical Information Systems 9(4): 421-432.

Han CT, Chen RS, Liu JF, Yang Y, Liu ZW (2013) Hydrological characteristics in non-freezing period at the alpine desert zone of Hulugou watershed, Qilian mountains. Journal of Glaciology and Geocryology 35(6): 1536-1544. (in Chinese with English abstract)

Henderson BL, Bui EN, Moran CJ, Simon DAP (2005) Australia-wide predictions of soil properties using decision trees. Geoderma 124(3): 383-398.

Huang CY (2000) Agrology. Beijing: China Agriculture Press (in Chinese).

Hudson BD (1992) The soil survey as paradigm-based science. Soil Science Society of America Journal 56(3): 836-841.

Loh WY (2011) Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1(1): 14-23.

McBratney AB, Odeh IO, Bishop TF, Dunbar MS, Shatar TM (2000) An overview of pedometric techniques for use in soil survey. Geoderma 97(3): 293-327.

McBratney AB, Santos MM, Minasny B (2003) On digital soil mapping. Geoderma 117(1): 3-52.

McIntosh PD, Lynn IH, Johnstone PD (2000) Creating and testing a geometric soil-landscape model in dry steeplands using a very low sampling density. Soil Research 38(1): 101-112.

McKenzie NJ, Gessler PE, Ryan PJ, O'Connell DA (2000) The role of terrain analysis in soil mapping. In Terrain analysis: principles and applications. New York: Wiley.

McKenzie NJ, Ryan PJ (1999) Spatial prediction of soil properties using environmental correlation. Geoderma 89(1): 67-94.

Meyer MD, North MP, Gray AN, Zald HS (2007) Influence of soil thickness on stand characteristics in a Sierra Nevada mixed-conifer forest. Plant and Soil 294(1–2): 113-123.

Moore ID, Gessler PE, Nielsen GA, Peterson GA (1993) Soil attribute prediction using terrain analysis. Soil Science Society of America Journal 57(2): 443-452.

Odeh IOA, Chittleborough DJ, McBratney AB (1992) Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. Soil Science Society of America Journal 56(2): 505-516.

Park SJ, McSweeney K, Lowery B (2001) Identification of the spatial distribution of soils using a process-based terrain characterization. Geoderma 103(3): 249-272.

Rouse Jr JW (1973) Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. MD: NASA/GSFC Type III Final Report.

Santos MM, Guenat C, Bouzelboudjen M, Golay F (2000) Three-dimensional GIS cartography applied to the study of the spatial variation of soil horizons in a Swiss floodplain. Geoderma 97 (3): 351-366.

Scull P, Franklin J, Chadwick OA, McArthur D (2003) Predictive soil mapping: a review. Progress in Physical Geography 27(2): 171-197.

Scull P, Franklin J, Chadwick OA (2005). The application of classification tree analysis to soil type prediction in a desert landscape. Ecological Modelling 181(1), 1-15.

Sun XL, Zhao YG, Zhang GL, Li DC (2008) Optimization of clustering parameters in predictive mapping of soil organic matter. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE) 24(9): 31-37. (in Chinese with English abstract)

Taghizadeh-Mehrjardi R, Sarmadian F, Minasny B, Trianta filis J, Omid M (2014) Digital mapping of soil classes using decision tree and auxiliary data in the Ardakan region, Iran. Arid Land Research and Management 28(2): 147-168.

Xu WN, Wang PX, Han P, Yan TL, Zhang SY (2011) Application of Kappa coefficient to accuracy assessments of drought forecasting model: a case study of Guanzhong Plain. Journal of Natural Disasters 20(6): 81-86. (in Chinese with English abstract)

Yang L, Zhu AX, Li, BL, Qin, CZ, Pei T, Liu BY, Li RK, Cai QG (2007) Extraction of knowledge about soil-environment relationship for soil mapping using Fuzzy c-means (FCM) clustering. Acta Pedologica Sinica 44(5): 784-791. (in Chinese with English abstract)

Zheng ZP, Liu ZX (2003) Soil quality and its evaluation. Chinese Journal of Applied Ecology 14 (1): 131-134. (in Chinese with English abstract)

Zhou MX (2012) The Inversion of Lunar Regolith Layer Thickness By Using the Data Obtained from Microwave Radiometer On CE-1 Satellite. Nanjing: Nanjing University of Aeronautics and Astronautics, College of Electronic and Information Engineering. (in Chinese with English abstract)

Zhu AX, Band LE, Dutton B, Nimlos TJ (1996) Automated soil inference under fuzzy logic. Ecological Modelling, 90(2): 123-145.

Chapter 7
Multivariate Sampling Design Optimization for Digital Soil Mapping

Gábor Szatmári, Károly Barta and László Pásztor

Abstract In this study, we have extended the spatial simulated annealing (SSA) methodology to be able to simultaneously optimize a completely new sampling design for more than one pedological variable using regression kriging prediction-error variance (RKV) as optimization criterion. For this purpose, the following soil properties were chosen: soil organic matter content, rooting depth, calcium carbonate content, and plasticity index according to Arany. The number of new observations was set to 100. The methodology is illustrated with a legacy soil dataset and auxiliary information from a study site in Central Hungary. The combined structure of the regression models and the variogram of the dominant soil parameter were applied in the optimization process provided by SSA to calculate the quality measure (i.e., spatially averaged RKV). The resulted sampling design was evaluated by various statistical and point pattern analysis tools. The Kolmogorov–Smirnov test's results and the observed empty space function showed that the optimized sampling configuration represents properly both the feature and geographic space. Furthermore, the empty space function pointed out that there is an inhibition between the sampling points, which caused a "quasi"-regular point pattern. The extended SSA methodology is suitable to optimize the sampling design for more than one soil variable.

Keywords Model-based sampling · Spatial simulated annealing · Regression kriging prediction-error variance · Variography

G. Szatmári (✉) · L. Pásztor
Institute for Soil Science and Agricultural Chemistry, Centre for Agricultural Research, Budapest, Hungary
e-mail: szatmari@rissac.hu

K. Barta
Department of Physical Geography and Geoinformatics, University of Szeged, Szeged, Hungary

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_7

7.1 Introduction

Digital soil mapping (DSM) aims at spatial prediction of soil types and properties by combining soil observation at points with auxiliary information, such as contained in digital elevation models, remote sensing images, and climatological records (McBratney et al. 2003; Heuvelink et al. 2007). Hence, the direct observations of the soil are important for two main reasons: (1) they are used to characterize the relationship between soil property and auxiliary information and (2) they are used to improve the predictions based on the auxiliary information, by spatial interpolation of the differences between the observations and predictions (Heuvelink et al. 2007). Regression kriging (RK), also termed universal kriging or kriging with external drift (Hengl et al. 2007), illustrates pretty well that twofold application of the observations, since it combines a regression of the target pedological variable on covariates with kriging of the regression residuals. Nevertheless, RK assumes that sampling points properly cover (i.e., represent well) both geographic and feature space (Hengl 2007), where the latter is defined by the covariates.

Extensive work has been done on sampling strategy optimization for DSM over the past decades to satisfy the topical demands, which were suggested by soil surveyors, pedometricans, end users, etc. These demands can be, e.g., the expectation of the accuracy and/or uncertainty of the prediction(s), taking auxiliary information into account, optimization of the sampling design for more than one soil variable, taking previously collected samples into account, consideration of any kind of constraints such as the number of the new observations, inaccessible areas for sampling, budget, and/or accuracy constraints. One of the niggling techniques is spatial simulated annealing (SSA) that has been frequently applied in soil surveys to optimize the sampling design using RK prediction-error variance (RKV) as optimization criterion. SSA with RKV is sporadically able to satisfy the mentioned demands. The main drawback of SSA is that it can be used only for one target variable. However, in a soil survey the usual aim is to describe the spatial distribution of not just one but several pedological variables (Vašát et al. 2010). Vašát et al. (2010) and Szatmári (2014) extended the SSA methodology to be able to optimize the sampling design for more than one pedological variable; however, they used different approaches. Vašát et al. (2010) used the linear model of coregionalization to model the mutual spatial dependence of four target soil properties and applied that along the optimization procedure, while Szatmári (2014) optimized the sampling design for two soil variables through a combined regression structure and the variogram of the dominant soil parameter.

The objective of this paper is to extend the SSA methodology, following Szatmári (2014), to be able to optimize the sampling configuration for more than two soil properties using the spatially averaged RKV as optimization criterion. The methodology is tested and illustrated in a study site in Central Hungary using a legacy soil dataset and auxiliary information. Four basic soil properties (i.e., soil organic matter content, rooting depth, calcium carbonate content, and plasticity index according to Arany) were chosen for the implementation of the method.

7.2 Materials and Methods

7.2.1 Study Site and Legacy Soil Data

The study site is located in the central part of Hungary (Fig. 7.1). The area of interest, approximately 17 km^2, is mainly covered by Haplic Chernozems and Kastanozems with significant secondary carbonates. Calcisols and Regosols are found on the eroded steeper slopes, where the top-horizon is too thin for Mollic or it is completely missing. Colluvic material can be found at the bottom of the slopes, where Phaeozems or Regosols were formed.

The soil data were collected at the end of the 1980s in the framework of the National Land Evaluation Programme. The dataset consists of 117 topsoil (0–30 cm) observations for the area of interest. Four soil variables were chosen from the dataset to present, test, and evaluate the extended methodology, namely soil organic matter (SOM), rooting depth (RD), calcium carbonate content (CC), and plasticity index according to Arany (K_A). This last variable quantifies the amount of water in cubic centimeter added (by continuous mixing) to 100 g of air-dried soil sample to obtain the upper limit of plasticity (MSZ 08-0205:1978). The gained value is appropriate to infer the mechanical composition of the soil sample. The summary statistics of the four variables are presented in Table 7.1.

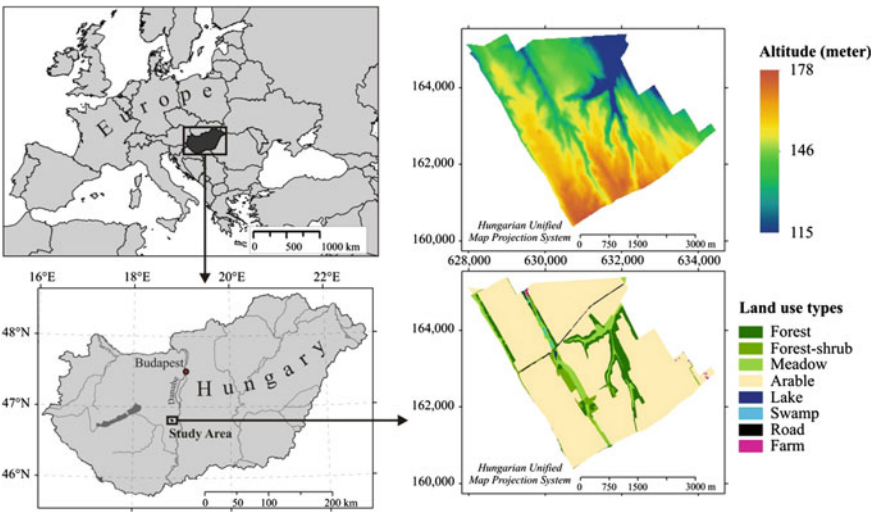

Fig. 7.1 The location of the study site in Central Hungary (*left*), the land use map and the digital elevation model of the study area (*right*)

Table 7.1 The summary statistics of the four soil variables computed from the legacy dataset

Soil variable	Mean	Minimum	Maximum	Std. deviation	Skewness
SOM (%)	2.90	1.51	4.44	0.56	−0.28
RD (cm)	73.54	5.00	150.00	27.46	0.09
CC (%)	8.22	0.5	29.00	5.41	0.54
K_A (cm^3)	41.74	26.00	58.00	4.18	0.30

SOM soil organic matter, *RD* rooting depth, *CC* calcium carbonate, K_A plasticity index according to Arany

7.2.2 Auxiliary Data

The exhaustive auxiliary information comes from a 20-m spatial resolution digital elevation model (DEM) and from the land use (LU) map of the study area (Fig. 7.1). The following morphometric parameters were derived from DEM: altitude, slope, slope length, aspect, profile and plan curvature, LS factor (Wischmeier and Smith 1978), topographic wetness index, vertical distance to channel network, and potential incoming solar radiations (direct and diffuse).

Products of the official aerial photography campaign of Hungary, taken in 2005, were used to derive the LU map. In contrast with the morphometric parameters, LU types are categorical variables. In consideration of that, each LU type was converted into indicator variable, respectively. A raster map was made for a given LU type with value domain showing 1 at the locations of the given LU and showing 0 for all other locations.

Principal component (PC) analysis was performed on the auxiliary data and the resulted PCs were used as covariates. It is a crucial step, because PCs are orthogonal and independent. Hence, those satisfy the requirements of the multiple linear regression analysis and decrease the multicollinearity effects (Hengl 2007).

7.2.3 Regression Kriging

In the last decade, RK has been more and more popular in DSM, as well as in SSA sampling optimization procedure using its prediction-error variance as optimization criterion. RK assumes that the deterministic component of the target soil variable is accounted for by the regression model, while the model residuals represent the spatially varying but dependent stochastic component. The estimation for Z variable at an unvisited location $\mathbf{s}_0$ is

$$\hat{Z}(\mathbf{s}_0) = \mathbf{q}_\mathbf{0}^\mathbf{T} \cdot \boldsymbol{\beta} + \boldsymbol{\lambda}_\mathbf{0}^\mathbf{T} \cdot (\mathbf{z} - \mathbf{q} \cdot \boldsymbol{\beta}), \tag{7.1}$$

where $\boldsymbol{\beta}$ is the vector of the regression coefficients, $\mathbf{q_0}$ is the vector of the covariates at the unvisited location, $\boldsymbol{\lambda_0}$ is the vector of the kriging weights, $\mathbf{z}$ is the vector of the

observations, and $\mathbf{q}$ is the matrix of covariates at the sampling locations. RKV at $\mathbf{s}_0$ is given by

$$\begin{aligned}\sigma^2(\mathbf{s}_0) = c(0) - \mathbf{c_0^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{c_0} \\ + (\mathbf{q_0} - \mathbf{q^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{c_0})^{\mathbf{T}} \cdot (\mathbf{q^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{q})^{-\mathbf{1}} \cdot (\mathbf{q_0} - \mathbf{q^T} \cdot \mathbf{C^{-1}} \cdot \mathbf{c_0}),\end{aligned} \tag{7.2}$$

where $c(0)$ is the variance of the residuals, $\mathbf{c_0}$ is the vector of covariances between the residuals at the observed and unvisited locations, and $\mathbf{C}$ is the variance–covariance matrix of the residuals. RKV is independent from the observed values (see Eq. 7.2), so it can be calculated before the actual sampling takes place, which can be considered as a beneficial property in point of costs and time. Furthermore, RKV incorporates both the prediction-error variance of the residuals and the estimation-error variance of the trend, which endeavor SSA algorithm to optimize the sampling configuration both in geographic and feature space (Heuvelink et al. 2007). However, it mainly depends on, how the prediction-error variance of the residuals and the estimation-error variance of the trend contribute to RKV.

7.2.4 Extended SSA Methodology and Its Settings

In brief, SSA is an iterative, combinatorial, model-based sampling optimization algorithm in which a sequence of combinations is generated by deriving a new combination from slightly and randomly changing the previous combination (van Groenigen et al. 1999). When a new combination is generated, the quality measure (i.e., spatially averaged RKV) is calculated and compered with the quality measure value of the previous combination (van Groenigen et al. 1999; Brus and Heuvelink 2007). The Metropolis criterion defines the probability that either accepts the new combination as a basis for the further computation, or rejects it, and the previous combination stays as a basis further (van Groenigen et al. 1999).

The SSA algorithm (using RKV as optimization criterion) requires for a given soil variable that the structure of the regression model (without the regression coefficients) and the variogram of the residuals are known (Heuvelink et al. 2007). The legacy soil dataset and the covariates were used to satisfy these requirements. Multiple linear regression analysis was performed to characterize the relationship between the target soil variables and the covariates. Stepwise method was applied to select the covariates into the regression models using significance level of 0.05. In the next step, the residuals were derived from the regression models and the corresponding variograms were calculated to model their spatial structures, respectively. According to Szatmári (2014), the combined structure of the resulted

regression models and the variogram of the residuals, which had the shortest spatial continuity, were applied in the optimization process provided by SSA to calculate the quality measure for a combination (i.e., sampling configuration). The reason of the variogram selection was to represent spatial continuity of the most variable residuals across the area of interest. It was called as "dominant parameter" by Füst and Geiger (2010). In case of multivariate sampling or monitoring network optimization, the dominant parameter has to control the optimal sampling distance between the sampling or monitoring locations in the geographic space; otherwise, the sampling configuration or the monitoring network will not be optimal for the most variable parameter (i.e., the sampling distance will be larger than the range of the dominant parameter).

The number of the new observations was set to 100, which can be considered as a sampling constraint for this study. Furthermore, this number of the new observations is commensurable with the sample size of the legacy soil dataset. The "initial temperature" for SSA was chosen in such a way that the average increase acceptance probability was 0.8 and the "cooling" was exponential. Furthermore, a stopping criterion was defined to rein up the simulation when the quality measure did not improve in many tries. The stopping criterion was set to 200. R software environment (R Development Core Team 2014) was used for the implementation.

7.2.5 Evaluation of the Optimized Sampling Design

The optimized sampling configuration was evaluated by various statistical and point pattern analysis tools. Kolmogorov–Smirnov test was used to examine for a given covariate, if the distribution from the optimized configuration is equal to the distribution from the complete area of interest. The null hypothesis of the statistical test was that the two distributions were drawn from the same distribution. The applied significance level was 0.01. The Kolmogorov–Smirnov test's results represent how the sampling points cover the feature space defined by the covariates.

The nearest-neighbor distribution function $G(r)$ and the empty space function $F(r)$ were calculated based on the optimized configuration to explore the type of interaction between the sampling points and to examine how they cover the geographic space. The $G(r)$ function measures the distribution of the distances from an arbitrary sampling point to its nearest sampling point, while the $F(r)$ function measures the distribution of all distances from an arbitrary point of the plane to its nearest sampling point (Bivand et al. 2008). In case of $F(r)$, the grid (with 20 m grid spacing) of the planned prediction locations was applied to measure the so-called empty space distances. The benefit of this practice is that it gives direct information on the kriging neighborhood (Szatmári et al. 2015).

7.3 Results and Discussion

7.3.1 Regression and Variogram Models

The main parameters of the resulted models are summarized in Table 7.2. The models explained 52, 39, 35, and 21 % of the total variation of RD, SOM, CC, and K_A, respectively.

The residuals were derived from the regression models, and the corresponding variograms were calculated to model their spatial structures, respectively. Table 7.3 summarizes the parameters of the fitted variogram models. The model type was spherical in all cases. The range values increase in the order CC < RD < SOM < K_A. Soil property CC is the most variable across the area of interest; hence, it is the dominant parameter in this case.

7.3.2 Sampling Optimization by the Extended SSA Methodology

A combined regression structure was used in the SSA algorithm, which was built upon the regression models of the four soil variables (Table 7.2). This combined structure was set in such a way that it contains all kinds of covariates from the four models, which occurs at least once in any of the models, according to Szatmári (2014).

Table 7.2 The main parameters of the fitted regression models

Soil property	R^2 (%)	No. of covariates	Sig.	List of covariates
SOM	39.01	4	$<10^{-7}$	PC1, PC2, PC14, PC18
RD	52.14	6	$<10^{-7}$	PC1, PC3, PC7, PC16, PC17, PC21
CC	34.93	8	$<10^{-7}$	PC2, PC3, PC7, PC11, PC13, PC14, PC16, PC19
K_A	20.87	3	7×10^{-6}	PC2, PC14, PC20

SOM soil organic matter, *RD*: rooting depth, *CC*: calcium carbonate, K_A plasticity index according to Arany

Table 7.3 The parameters of the fitted variogram models

Soil variable	Type	Nugget	Partial sill	Nugget/sill (%)	Range (m)
SOM	Spherical	0.042	0.150	21.87	1520
RD	Spherical	43.20	316.79	12.00	980
CC	Spherical	0.00	18.60	0.00	560
K_A	Spherical	4.90	8.62	36.24	2075

SOM soil organic matter, *RD* rooting depth, *CC* calcium carbonate, K_A plasticity index according to Arany

For example, PC21 occurs only in the RD model as covariate, in contrast with PC2, which occurs in the SOM, CC, and K_A models, respectively (see Table 7.2). Although, the PC21 occurs only once altogether, it has to be involved into the combined structure, as well as the PC2 covariate, because it is relevant in RD variable (even if it is irrelevant in SOM, CC, and K_A variables). According to this practice, the combined regression structure involved 13 covariates. The feature space is defined by these 13 covariates, where the sampling design should be optimized.

The variogram of the CC residuals had the shortest spatial continuity (i.e., range), which has to control the optimal sampling distance between the points in the geographic space; otherwise, the sampling configuration will not be optimal. The application of the CC residuals' variogram will infer that any of the planned prediction locations has at least one kriging neighbor. It also means that any of them has an influence from at least one sampling point (i.e., the geographic space is fully covered by the sampling locations), as it was presented by Füst and Geiger (2010) and Szatmári (2014). In addition, the optimized sampling design does not depend on the absolute value of the sill and the nugget; it merely depends on the nugget/sill ratio. If the nugget/sill ratio rises to 75 %, it has an influence on the sampling configuration (van Groenigen 2000). In the present study, there has been no any case, where the nugget/sill ratio would be close to 75 % (see Table 7.3). Hence, the variogram of the dominant parameter is appropriate to apply along the optimization procedure (Szatmári 2014).

The combined regression structure and the CC variogram were used to calculate (with Eq. 7.2) the RKV as optimization criterion. It must be admitted that the calculated RKV is an imaginary quality measure; however, it is appropriate to compare alternative sampling configurations and to optimize the sampling design for DSM (Szatmári 2014). The resulted sampling configuration by SSA is presented in Fig. 7.2.

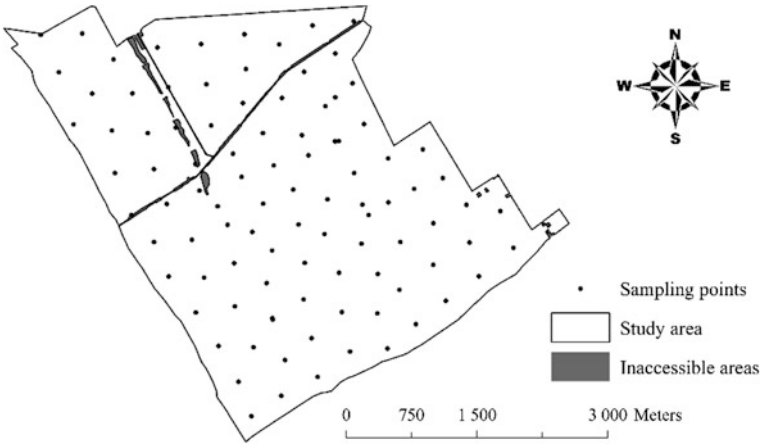

Fig. 7.2 The optimized sampling configuration with the inaccessible areas on the study site using the extended SSA methodology

7.3.3 Performance of the Sampling Configuration

The results of the Kolmogorov–Smirnov test are presented in Table 7.4. Only three cases gave that the two distributions are different (see Table 7.4). Hence, correlation test was performed on those three covariates to examine whether the value of the correlation coefficient is zero between the two distributions, following Hengl (2007). This null hypothesis was rejected in all cases. Moreover, the calculated Pearson correlation coefficients showed that there is a strong relationship between the two distributions in all cases. Based on the statistical tests' results, the feature space is properly covered by the optimized sampling design.

The observed $G(r)$ and $F(r)$ functions are presented in Fig. 7.3. Based on the observed functions, there is an inhibition (i.e., competition) between the sampling points, which caused a "quasi"-regular point pattern (see Fig. 7.2). It means that the

Table 7.4 The Kolmogorov–Smirnov test's results for the feature space

Covariate	ε_{emp}	Equal to? (YES/NO)
SPC1	0.1938	YES
SPC2	0.5696	YES
SPC3	0.0783	YES
SPC7	0.0038	NO
SPC11	0.0006	NO
SPC13	0.4540	YES
SPC14	0.3752	YES
SPC16	0.6172	YES
SPC17	0.0039	NO
SPC18	0.0104	YES
SPC19	0.0571	YES
SPC20	0.1314	YES
SPC21	0.4413	YES

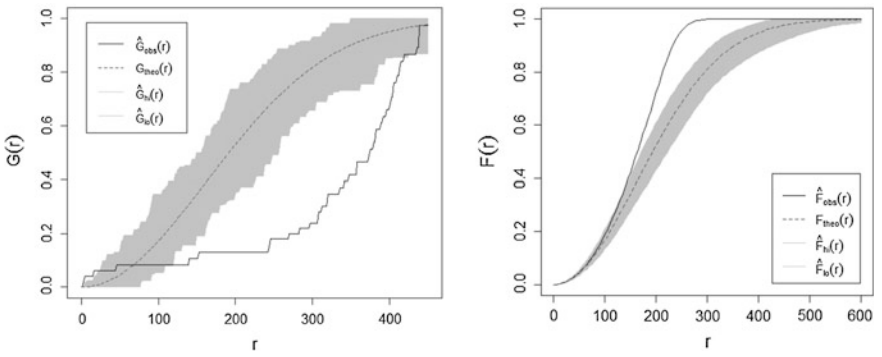

Fig. 7.3 The observed nearest-neighbor distribution function $G_{obs}(r)$ (*left graph*) and empty space function $F_{obs}(r)$ (*right graph*) for the optimized sampling configuration

Table 7.5 The summary statistics of the empty space distances

	Mean	Median	Minimum	Maximum	Std. deviation	Skewness
Empty space distances (m)	167.60	170.42	1.83	706.38	70.60	0.69

applied variogram model had the dominant influence along the optimization procedure rather than the regression structure, according to Heuvelink et al. (2007). However, it relates to our earlier expectation that the feature space is fairly homogeneous in point of topography and land use (Szatmári et al. 2015).

The calculated empty space distances are appropriate measures to characterize the relationship between the sampling points and the planned prediction locations (Szatmári et al. 2015). The summary statistics of these measures are presented in Table 7.5. The maximum value is larger than the range value of the CC residuals' variogram (see Table 7.3); it means that there is at least one prediction location (i.e., grid cell or pixel), which did not have any kriging neighbor. It is reasonable to examine, there are more such locations. Only 15 locations (i.e., grid cells or pixels), from the total of 42,037, were found which did not have any kriging neighbor. These 15 grid cells are located in the easternmost part of the study site, where many inaccessible areas are located for sampling too (see Fig. 7.2). On the other hand, the probability, to find the nearest sampling location to a given prediction location within the distance of 270.6 m, is equal to 0.95 (see Fig. 7.3 right graph). As a consequence, the geographic space is properly covered by the sampling configuration.

The optimized sampling design, provided by the extended SSA methodology, covered properly both the feature and geographic space, and thus, it can be applied to map the spatial distributions of the chosen four soil variables. While some components of our approach is less elaborated than that of provided by Vašát et al. (2010) (e.g., the calculated RKV is not suitable to use as absolute qualifier for the predictions), its significant advantage is that it is able to take auxiliary information into account along the optimization procedure.

7.4 Conclusions

The extended methodology of SSA, which was tested and evaluated with four soil variables in this paper, is able to simultaneously optimize the sampling design for more than one pedological variable using the RKV as optimization criterion. As it was presented, the extended methodology preserved the RKV beneficial properties and the optimized sampling design is appropriate for DSM purpose, because the sampling configuration covered properly both the feature and geographic space, which is essential in point of DSM.

Nevertheless, the calculated RKV, which was derived from the combined regression structure and the variogram of the dominant parameter, is an imaginary quality measure. Hence, it is not suitable to use as absolute qualifier for the prediction accuracies, as well as the joined uncertainties. To get around this problem, RKV can be recalculated based on the corresponding regression and variogram model for a given soil variable. The recalculated RKV will then satisfy the above-mentioned demands.

Acknowledgements Our work has been supported by the Hungarian National Scientific Research Foundation (OTKA, Grant No. K105167).

References

Bivand RS, Pebesma EJ, Gómez-Rubio V (2008) Applied Spatial Data Analysis with R. Springer, New York.

Brus DJ, Heuvelink GBM (2007) Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138:86-95.

Füst A, Geiger J (2010) Monitoringtervezés és –értékelés geostatisztikai módszerekkel I. Szakértői véleményen alapuló "igazoló" mintázás geostatisztikai támogatása [Setting up monitoring networks using geostatistics I. Geostatistical support for a judgmental sampling strategy – in Hungarian]. Földtani Közlöny 140:303-312.

Hengl T (2007) A Practical Guide to Geostatistical Mapping of Environmental Variables. ISBN 978-92-79-06904-8.

Hengl T, Heuvelink GBM, Rossiter DG (2007) About regression-kriging: from equations to case studies. Computers and Geosciences 33:1301-1315.

Heuvelink GBM, Brus DJ, de Gruijter JJ (2007) Optimization of sample configurations for digital mapping of soil properties with universal kriging. In: Lagacherie P, McBratney AB, Voltz M (Eds.) Developments in Soil Science, Vol. 31. Elsevier B.V., Amsterdam.

McBratney AB, Mendonça Santos ML, Minasny B (2003) On digital soil mapping. Geoderma 117:3-52.

MSZ 08-0205:1978 (Hungarian Standard for Determination of Physical and Hydrophysical Properties of Soils)

R Development Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. (http://www.R-project.org.)

Szatmári G (2014) Optimization of sampling configuration by spatial simulated annealing for mapping soil variables. In: Cvetković M, Novak Zelenika K, Geiger J (Eds.) 6th Croatian - Hungarian and 17th Hungarian geomathematical congress: "Geomathematics - from theory to practice". Croatian Geological Society, Zagreb.

Szatmári G, Barta K, Pásztor L (2015) An application of a spatial simulated annealing sampling optimization algorithm to support digital soil mapping. Hungarian Geographical Bulletin 64:35-48.

van Groenigen JW (2000) The influence of variogram parameters on optimal sampling schemes for mapping by kriging. Geoderma 97:223-236.

van Groenigen JW, Siderius W, Stein A (1999) Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87:239-259.

Vašát R, Heuvelink GBM, Borůvka L (2010) Sampling design optimization for multivariate soil mapping. Geoderma 155:147-153.

Wischmeier WH, Smith DD (1978) Predicting rainfall erosion losses: A guide to conservation planning. U.S. Government Printing Office, Washington DC.

Chapter 8
Applying Artificial Neural Networks Utilizing Geomorphons to Predict Soil Classes in a Brazilian Watershed

H.S.K. Pinheiro, P.R. Owens, C.S. Chagas, W. Carvalho Júnior and L.H.C. Anjos

Abstract The use of landscape terrain attributes associated with the field information in geographic information systems (GISs) helps to improve the methods applied in soil survey. Geomorphons is a novel technique to map surface elements from digital elevation model and visibility distance (search radius) of a central point in the landscape, which can adopt flexible scales. The main goal of this study was to evaluate the potential for incorporating Geomorphons, which is used to recognize landscape patterns and to improve the soil class predictions by artificial neural networks (ANNs). The procedures involved the acquisition of a cartographic database, creating digital models that represent landscape attributes relevant to paedogenesis on the research site (including Geomorphons of different search radius), sample collection and description of one hundred soil profiles in predefined locations, and finally the supervised classification by neural networks. The covariates used were as follows: elevation, slope, curvature, combined topographic index (CTI), euclidean distance, clay minerals, iron oxide, normalized difference vegetation index (NDVI), geology, and Geomorphons. All models for the terrain attributes have 30-m pixel resolution, and these variables correspond to neurons in

H.S.K. Pinheiro (✉) · P.R. Owens
Federal Rural University of Rio de Janeiro (Soil Department),
BR 465, Km 47, Seropédica, RJ CEP 23890-000, Brazil
e-mail: lenask@gmail.com

P.R. Owens
e-mail: prowens@purdue.edu

C.S. Chagas · W. Carvalho Júnior
Embrapa Soils, Jardim Botânico St. 1.024, Jardim Botânico,
Rio de Janeiro, RJ CEP 22460-000, Brazil
e-mail: cesar.chagas@embrapa.br

W. Carvalho Júnior
e-mail: waldir.carvalho@embrapa.br

L.H.C. Anjos
Soil Department- Agronomy, Federal Rural University of Rio de Janeiro,
BR 465, Km 47, Seropédica, RJ CEP 23890-000, Brazil
e-mail: lanjos@ufrrj.br

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_8

the input layer of the neural networks. The output layer of the supervised classification corresponded to the nine dominant soil classes in the study area. To define the appropriate scale of Geomorphons map, sixteen sets of neural networks contain each one of the terrain attributes plus a Geomorphons map calculated from different search radius. For comparative purposes, one of the sets included no Geomorphons. Selection of the appropriate Geomorphons search radius was based on the statistical indexes obtained from a confusion matrix. The results showed that the best classification used the Geomorphons map obtained by forty-five pixels of search radius, in combination with other variables. This classifier presented values to kappa index and global accuracy corresponding to 0.74 and 77.0, respectively.

Keywords Digital soil mapping · Artificial neural networks, ternary patterns · Geomorphometric attributes · GRASS

8.1 Introduction

Digital soil mapping (DSM) is a tool that can work to improve the products of soil surveys through geographic information systems (GISs) and knowledge of soil genesis, morphology, and classification. DSM can improve soil surveys by increasing the efficiency of cost and time, while improving overall map accuracy. These techniques can improve the classical procedures applied to soil surveys, incorporating pedometric concepts, allowing analysis from quantitative data and qualitative aspects of the physical environment. The flexibility of DSM products can provide easier interpretation and multifaceted presentations of soil–landscape information.

The morphometric parameters of landscape, derived from digital elevation models (DEMs) are particularly important to DSM for generating covariates that provide a consistent approach to representing landforms. In this sense, some studies about landforms have been developed based on recognizing and mapping terrain units and landscape patterns (Schmidt and Hewitt 2004; Iwahashi and Pike 2007; Ehsani and Quiel 2008).

In the last decade, there have been numerous studies using artificial neural networks (ANNs) for spatial correlations in soil classification and correlated soil properties (Tranter et al. 2007; Choi et al. 2010; Chen et al. 2011; Motaghian and Mohammad 2011; Carvalho Junior et al. 2011; Chagas et al. 2011). Some recent studies have focused on developing efficient techniques for auto-classification and mapping landforms elements (Iwahashi and Pike 2007; Ehsani and Quiel 2008). The Geomorphons approach provides the auto-classification of the ten most common recognized surface types, based on ternary patterns (Jasiewicz and Stepinski 2013). The values of ternary patterns were calculated based on a central pixel and the relative elevation of neighboring pixels.

The goal of this study was to evaluate the potential use of Geomorphons, in combination with other terrain covariables, as an input variable to predict soil classes through ANN.

8.2 Materials and Methods

8.2.1 Characterization of the Area

The study area corresponds to the Guapi-Macacu watershed in Rio de Janeiro, Brazil (Fig. 8.1). The watershed is an important natural unit and is appropriate for hydrological or environmental studies and analyses. The Brazilian National Policy on Water Resources (Law N° 9433/97) recognizes the watershed as a territorial unit to manage water resources and develop management plans for land use. Guanabara Bay, in the state of Rio de Janeiro, encompasses 12 watersheds of great importance with direct contribution to the bay (Cortes et al. 2010). The Guapi-Macacu watershed corresponds to approximately 31 % of the total land area of contribution and has a contribution area of 1250.78 km^2 and a perimeter of 199.2 km. Figure 8.1 shows the location of Guapi-Macacu watershed, in Rio de Janeiro State, Brazil.

The climate corresponds to the tropical rainy with dry winter (Aw) according to the classification proposed by Köppen (1948). The average temperature is 23 °C, and the average annual rainfall exceeds 1200 mm reaching 2600 mm in the highest quotes on the watershed (Ecologus-Agrar 2003; Dantas 2000).

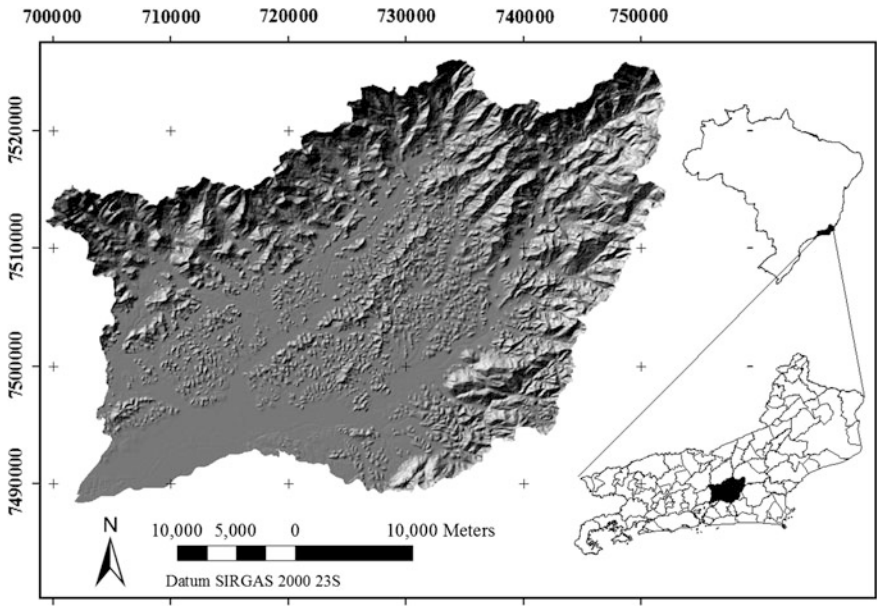

Fig. 8.1 Location of the Guapi-Macacu watershed, in Rio de Janeiro State, Brazil

8.2.2 Terrain Attributes (Input Variables)

The main geographic information system used was the ArcGIS Desktop v.10. Complementary analyses were performed in ERDAS Imagine v.9.1. (ERDAS Systems) and Geographic Resources Analysis Support System (GRASS). All layers were created at a 30-m spatial resolution and projected in Universal Transverse Mercator (UTM), horizontal datum SIRGAS 2000 Zone 23 S.

The DEM was created by interpolation of primary elevation data (contours, elevations points from the official database in Brazil) using the tool "TopotoRaster" at a 30-m resolution. The slope and curvature map were derived from the DEM, using the module "Spatial Analyst Tools: Surface." The euclidean distance of stream network was calculated by the tool "Distance." Compound topographic index (CTI) was generated in ArcINFO. Three indexes were generated using remotely sensed data from Landsat 5 TM (image from Sep 2011) in ERDAS, and they are as follows: normalized difference vegetation index—NDVI (Yang et al. 1997), clay minerals, and iron oxide, calculated as the ratio between the band 5 and band 7, and band 3 by band 4, respectively (Chagas et al. 2013).

The lithology of study supported the differentiation of soil types. This map was adapted from the geological survey charts of Rio de Janeiro state (CPRM 2001). The parent material map is important for soil mapping because it provides the boundary conditions for soil development.

Geomorphons algorithm was used to characterize terrain types based on the neighborhood of a central pixel, and consider not the relative elevation as well as the rate of change of their angles. Geomorphons was created using a flexible procedure, making possible the recognition of the same types of landforms at different scales. At the end of the process of auto-classification, the ten most commonly recognized surface types were identified (Fig. 8.2).

The Geomorphons maps were created using the Geomorphons add-on in GRASS (http://sil.uc.edu/). The procedure to calculate the landforms is flexible and produces different resulting maps according to the different distances of zenith and nadir angles, also called search radius, lookup distance or scale—L (Jasiewicz and Stepinski 2013). Fifteen maps of Geomorphons were generated with search radius

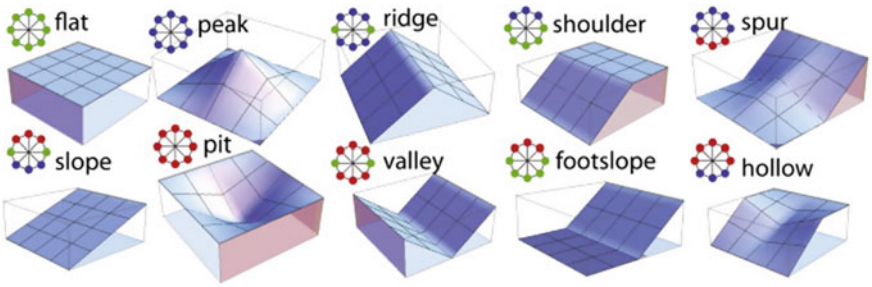

Fig. 8.2 The ten most common landforms (*Source* Jasiewicz and Stepinski 2013)

corresponding to 3, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100, 150, 300, and 500 pixels, respectively.

8.2.3 Soil Sampling and Profile Description

Conditioned Latin Hypercube Sampling (cLHS) was selected for its ability to capture the variation in soil property distribution, while maintaining some degree of randomness (Roudier et al. 2012; Minasny and McBratney 2006). The selection of sampling points involved elevation, slope, and curvature as conditions for the cLHS to select one hundred sample locations, within a 100 m buffer from the road (Carvalho Junior et al. 2014). The urban areas were excluded of the soil survey.

Based on the field survey, nine dominant soil orders were recognized and selected as output classes to be predicted by ANN. Soil taxonomic descriptions were adapted to World Reference Base for Soil Resources (WRB 2014).

To characterize the morphometric patterns, 500 pixels from each soil type were collected for each of the nine pedogenetic units, consistent with characteristics of the point of observations. A subset of 350 pixels was selected as training samples, and the remaining 150 pixels were used for validation. Zhu (2000) proposed a number of training samples around 30 times the number of output classes to account for the complexity of the relation between number of inputs variables and output classes. The author also suggested that the number of validation samples should be nearly half the number of training samples.

The selection of the training and validation samples was based on the variability of the terrain variables in each one of the nine soil orders. In this procedure lies an important step of the supervised classification once the success of the prediction depends directly of the coherence of the samples to represent the output classes, and the variables are used as an input (Pinheiro 2012).

8.2.4 Classification by Neural Networks

The definition of the network architecture was comprised of the following parameters: (1) number of layers, (2) number of neurons (perceptrons) in each layer, (3) type of connection between nodes, and (4) network topology. The input layer was represented by terrain covariables, and the output layer corresponded to the dominant soil orders.

For the first instance, the supervised classification by ANNs requires a training process where the network “learns” the conditions where each soil class occurs (Tso and Mather 2009). The learning algorithm was based on “backpropagation”

which allows the random distribution of the interneurons' weights between −1 and 1, varying learning rates and cycles. The parameters adopted in the training process were number of cycles (or iterations) corresponding to 2000 and learning rate equal to 0.2.

To select the appropriate Geomorphons search radius to be used as an input in predictive models, the training of sixteen set of ANNs were performed, each using all the terrain attributes in combination with each of the 16 differing Geomorphons maps. For comparison, one of the sets included no Geomorphons map (reference set). Table 8.1 presents the organization of the sets and respective variables used as an input in the predictive models.

To define the appropriated architecture of each set (described above), the training of seventeen networks with different numbers of nodes in the hidden layer (1–15, 20, and 30) was performed. The criterion used to add neurons in the hidden layer was based on the mean square error (MSE), which measures the difference between the estimated and the desired values for the training, according to Eq. 8.1.

$$\mathrm{MSE} = \frac{\sum (e - d)^2}{n} \tag{8.1}$$

where "*e*" represents the estimated value for each pixel; "*d*", the desired values; and "*n*," the number of learning cycles. Thus, the training should be stopped when the error have the lowest possible and no longer oscillates with new cycles (Chagas et al. 2011).

Table 8.1 Training sets using different sizes of Geomorphons search radius (L)

Set	Terrain variables in input layer
1	Elevation, slope, curvature, CTI, euclidean distance, geology, clay minerals, iron oxide, and NDVI (reference set)
2	All attributes from set 1 plus Geomorphons calculated with 3 cells of search radius
3	All attributes from set 1 plus Geomorphons calculated with 5 cells of search radius
4	All attributes from set 1 plus Geomorphons calculated with 10 cells of search radius
5	All attributes from set 1 plus Geomorphons calculated with 15 cells of search radius
6	All attributes from set 1 plus Geomorphons calculated with 20 cells of search radius
7	All attributes from set 1 plus Geomorphons calculated with 25 cells of search radius
8	All attributes from set 1 plus Geomorphons calculated with 30 cells of search radius
9	All attributes from set 1 plus Geomorphons calculated with 35 cells of search radius
10	All attributes from set 1 plus Geomorphons calculated with 40 cells of search radius
11	All attributes from set 1 plus Geomorphons calculated with 45 cells of search radius
12	All attributes from set 1 plus Geomorphons calculated with 50 cells of search radius
13	All attributes from set 1 plus Geomorphons calculated with 100 cells of search radius
14	All attributes from set 1 plus Geomorphons calculated with 150 cells of search radius
15	All attributes from set 1 plus Geomorphons calculated with 300 cells of search radius
16	All attributes from set 1 plus Geomorphons calculated with 500 cells of search radius

8.3 Results and Discussion

8.3.1 *Characterization of Soil Types and Occurrence Conditions*

The dominant soil orders in the study area were Ferralsols (28 %), Cambisols (18 %), Gleysols (15 %), Acrisols (24 %), Regosols, and Fluvisols (6 %). The output classes corresponding to the mapping units were as follows: (1) Haplic Acrisols (Clayic), (2) Haplic Acrisols (Chromic), (3) Haplic Cambisols, (4) Haplic Gleysols, (5) Endosalic Gleysols, (6) Haplic Ferralsols (Xanthic), (7) Haplic Ferralsols (Dystric), (8) Fluvisols, and (9) Regosols.

Figure 8.3 shows the variability of terrain attributes derived from the DEM to each map unit. To allow for the comparison of all the terrain variables at the same time, a rescale procedure was applied to restrict the variability of attributes from 0 to 1. More details about this procedure were described in Pinheiro (2012).

Haplic Acrisols Clayic occurs on gentle slopes and low elevations; in contrast, the Haplic Acrisols Chromic are common in higher elevations under wide slope conditions and were predominantly associated with alkaline rocks. The Haplic Cambisols dominated on concave forms, steep slopes, and high elevation, sometimes occurring in association with the Regosols and rock outcrops. The Gleysols occurred in low areas of recent sedimentation with low slopes and planar curvature, which were divided in two main units: Haplic Gleysols and Endosalic Gleysols. The difference in the landscape was mainly due to elevation and CTI, where the Endosalic Gleysols have higher CTI values and lower elevation values. Ferralsols occurs predominantly in convex landscapes and present reduced wetness index (CTI). The Fluvisols present high values of CTI, and its occurrence was observed surrounding the stream networks, the Macacu and Guapi-Acu rivers.

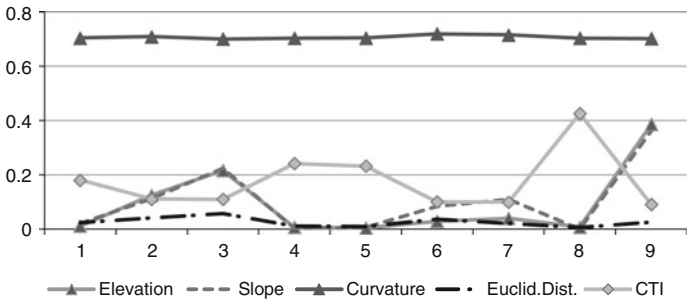

Fig. 8.3 Comparison of terrain attributes derived from the DEM over the nine map units. *1* Haplic Acrisols (Clayic); *2* Haplic Acrisols (Chromic); *3* Haplic Cambisols; *4* Haplic Gleysols; *5* Endosalic Gleysols; *6* Haplic Ferralsols (Xanthic); *7* Haplic Ferralsols (Dystric); *8* Fluvisols; *9* Regosols

8.3.2 *Inferred Classification Using Geomorphons as Discriminant Variable*

Sixteen sets were trained each with different network architectures and different number of neurons in the hidden layer (1–15, 20, and 30), while keeping the same number of neurons in the input layer (terrain variables) and output layer (soil classes). The analysis of supervised classification was based on statistical indexes such as MSE, kappa, and global accuracy, which determine the proportion of correct guesses (accuracy). Figure 8.4 shows the analysis of MSE to some of the training sets.

According to the graph, all sets present similar behavior with respect to MSE. Set 11 has the lowest MSE in the network with 30 nodes in the hidden layer (0.139). However, the decreasing error rates show reduced rates from 10 nodes in the hidden layer. Similar observations were reported by Chagas et al. (2011). Foody and Arora (1997) highlights that larger and complex networks can be more efficient to properly characterize a training set but are usually less efficient than simpler networks to generalize the output classes. Having said that, networks with largest number of neurons in the hidden layer do not necessarily imply a better performance of the neural network. At the end of the training step, a confusion matrix was created for each neural network, which determines the values to kappa index, overall and variance (Congalton and Green 1999). The comparison of neural networks was based on a significance matrix from these statistical indices. Table 8.2 shows the significance matrix between the neural networks of the set that used Geomorphons map with 45 cells of search radius as an input variable, and varying the number of neurons in the hidden layer (Set-11).

According to the data in Table 8.2, the network with seven neurons in the hidden layer was chosen to represent this set, because it represents better results for the kappa index (0.741) and variance (0.000166), and shows statistical difference when compared with other networks with the same input variables. For each set of input

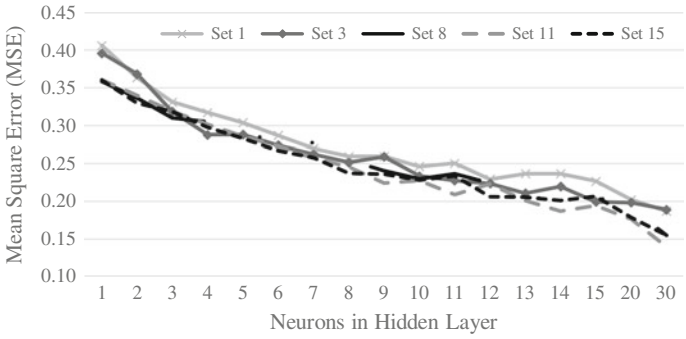

Fig. 8.4 Analysis of mean square error from different sets and numbers of neurons in the hidden layer

Table 8.2 Significance matrix from Set 11 with Geomorphons generated with the 45 pixels of search radius as an input variable

$N^{(1)}$	1	2	3	4	5	6	7	8
Kappa	0.688	0.665	0.695	0.685	0.662	0.669	0.741	0.654
$Var^{(2)}$	1.86	1.93	1.83	1.87	1.96	1.93	1.66	1.97
1	50.45							
2	1.18	47.87						
3	0.36	1.55	51.38					
4	0.16	1.03	0.52	50.09				
5	1.33	0.15	1.70	1.18	47.29			
6	0.98	0.20	1.34	0.82	0.36	48.16		
7	2.83*	4.01*	2.46*	2.98*	4.15*	3.80*	57.51	
8	1.74	0.56	2.10*	1.58	0.40	0.76	4.57*	46.60
9	0.00	1.18	0.36	0.16	1.33	0.97	2.82*	1.73
10	1.93*	0.76	2.30*	1.78	0.60	0.96	4.76*	0.20
11	2.63*	1.46	3.00*	2.48	1.30	1.66	5.46*	0.90
12	1.63	0.46	2.00*	1.48	0.30	0.66	4.45*	0.10
13	3.58*	2.40*	3.95*	3.42*	2.24*	2.60*	6.41*	1.84
14	0.10	1.08	0.47	0.05	1.23	0.87	2.93*	1.63
15	0.87	0.31	1.24	0.72	0.46	0.10	3.69*	0.86
20	1.84	0.66	2.20	1.68	0.50	0.86	4.66*	0.10
30	1.49	0.31	1.85	1.33	0.15	0.51	4.32*	0.25

$N^{(1)}$	9	10	11	12	13	14	15	20	30
Kappa	0.688	0.65	0.636	0.656	0.617	0.686	0.671	0.652	0.659
$Var^{(2)}$	1.88	2.00	2.04	1.99	2.08	1.87	1.93	1.99	1.94
9	50.18								
10	1.93	45.96							
11	2.63*	0.70	44.53						
12	1.63	0.30	1.00	46.50					
13	3.57*	1.63	0.94	1.93	42.78				
14	0.10	1.83	2.53*	1.53	3.47*	50.17			
15	0.87	1.06	1.76	0.76	2.70*	0.77	48.30		
20	1.83	0.10	0.80	0.20	1.74	1.73	0.96	46.22	
30	1.48	0.45	1.15	0.15	2.10*	1.38	0.61	0.35	47.31

$N^{(1)}$ = number of neurons in hidden layer
$Var^{(2)}$ = variance × 10^4
*Significance difference at 95 %

variables, the network with the best performance of the statistical indexes (kappa and variance), obtained from a confusion matrix, was selected. The contribution of each set was defined assuming that the reference set corresponds to Set 1, where the input data contain no Geomorphons map.

Table 8.3 Summary of the comparison between the best neural network of each set

Set	L	Neurons[a]	Kappa	Global accuracy	Variance[b]	Contribution
1	0	8	0.709	74.1	1.78	–
2	3	13	0.735	76.4	1.67	0.026
3	5	5	0.713	74.5	1.76	0.004
4	10	5	0.716	74.7	1.76	0.007
5	15	5	0.69	72.4	1.87	−0.019
6	20	4	0.703	73.6	1.8	−0.006
7	25	1	0.662	69.9	1.96	−0.047
8	30	2	0.68	71.6	1.88	−0.029
9	35	5	0.685	72	1.88	−0.024
10	40	4	0.686	72.1	1.86	−0.023
11	45	7	0.741	77	1.66	0.032
12	50	7	0.74	76.9	1.65	0.031
13	100	11	0.717	74.8	1.75	0.008
14	150	2	0.704	73.7	1.79	−0.005
15	300	5	0.716	74.7	1.76	0.007
16	500	5	0.719	75	1.75	0.01

L = size of search radius (cells) in Geomorphons map
[a]Neurons = number of neurons in hidden layer
[b]Variance = variance × 10^4

Table 8.3 shows the contribution of the best network of each set with different Geomorphons as an input variable compared with the reference set (without Geomorphons map).

The Set 11, a Geomorphons map created with 45 cells of search radius, had the highest positive value to contribution, when compared with the reference set and among others. The sets that used Geomorphons maps with search radius varying between 15 and 40 cells and the Set 14 with 150 cells of search radius had inferior performance when compared with the reference set.

A significance matrix with the best network of each set was generated from the kappa and variance values, to compare the performance between different neural networks with different Geomorphons search radius (Table 8.4).

The results showed that the best architecture was obtained from Set 11, which used the Geomorphons maps with a forty-five pixels of search radius as an input and seven neurons in the hidden layer. Statistical indices from the resulting classification showed superior performance, with values of kappa index, global accuracy, and variance corresponding to 0.741, 77.0, and 1.66 × 10^4, respectively. Although this set does not present a statistical difference between sets 5 and 12, a visual evaluation to analyze the coherence with other layers was performed, confirming that the Set 11 has a better performance than the others sets. Jasiewcz et al. (2014) used similar value to search radius (40 cells or 1200 m) to calculate the Geomorphons derived of DEM with 30 m pixels. Figure 8.5 shows the Geomorphons map calculated with 45 cells of search radius.

Table 8.4 Significance matrix between the best neural network of each set

Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	53.14															
2	1.40	56.88														
3	0.21	1.19	53.74													
4	0.37	1.03	0.16	53.97												
5	1.00	2.39*	1.21	1.37	50.46											
6	0.32	1.72	0.53	0.69	0.68	52.40										
7	2.43*	3.83*	2.64*	2.80*	1.43	2.11*	47.29									
8	1.52	2.92*	1.73	1.89	0.52	1.20	0.92	49.59								
9	1.26	2.65*	1.47	1.63	0.26	0.94	1.17	0.26	49.96							
10	1.21	2.61*	1.42	1.58	0.21	0.89	1.23	0.31	0.05	50.30						
11	1.73	0.33	1.51	1.35	2.71*	2.04*	4.15*	3.24*	2.98*	2.93*	57.51					
12	1.67	0.27	1.46	1.30	2.67*	1.99*	4.11*	3.19*	2.93*	2.88*	0.06	57.61				
13	0.43	0.97	0.21	0.05	1.42	0.74	2.86*	1.94	1.68	1.63	1.30	1.25	54.20			
14	0.27	1.67	0.48	0.64	0.73	0.05	2.17*	1.25	0.99	0.94	1.99*	1.94	0.69	52.62		
15	0.37	1.03	0.16	0.00	1.37	0.69	2.80*	1.89	1.63	1.58	1.35	1.30	0.05	0.64	53.97	
16	0.53	0.87	0.32	0.16	1.52	0.85	2.96*	2.05*	1.79	1.74	1.19	1.14	0.11	0.80	0.16	54.35

*Significance at 95 %

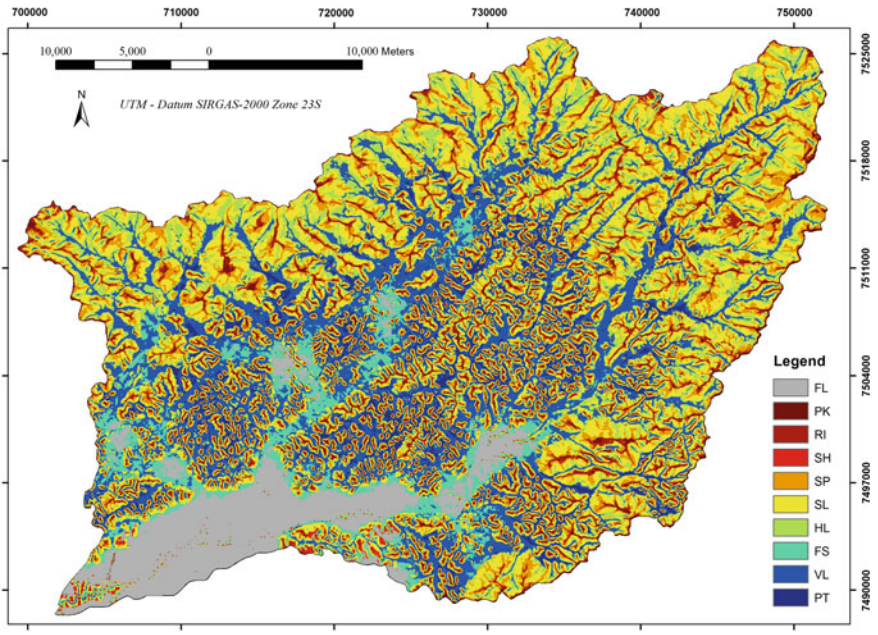

Fig. 8.5 Geomorphons map with 45 cells of search radius (Set 11). *FL* flat; *PK* peak; *RI* ridge; *SH* shoulder; *SP* spur; *SL* slope; *HL* hollow; *FS* footslope; *VL* valley; *PT* pit

The evolution of soils as a function of the water behavior in the landscape, which determines the favorable conditions for pedogenesis or morphogenesis, justifies the application of the Geomorphons approach to recognize the main landforms in study area. Landscape patterns need to be sensitive to natural processes and variations in the surface shape of landscapes, but also appropriated for the map scale and details. The Geomorphons map selected to represent the landforms (Fig. 8.5) shows coherence with the features, as observed in the field, relating the soils with the incipient degree of evolution (Regosols and Cambisols) with landforms as shoulder, peak, and ridge. The Acrisols and Ferrasols have a wide occurrence area and landform shapes, associated with the most of cases with slope landforms, gentle or steep. In contrast, the floodplains and drainage networks show direct relationship with flat and valley landforms, where Gleysols and Fluvisols occur.

8.4 Conclusion

The dominant soil orders in the Guapi-Macacu watershed were Ferralsols, Cambisols, Gleysols, Acrisols, Regosols, and Fluvisols. The interpretation of remote sensing images, DEM and thematic maps, combined with field observations and a literature review allowed for identifying consistent relationships between the

landforms and helped to understand the occurrence of different soil types in the study area.

The network selected to represent the soil distribution in the watershed is composed of ten discriminating variables, seven neurons in the hidden layer, and nine in the output layer, corresponding to identified soil classes. The parameters that justified the chosen network were the values of statistical indexes, such as global accuracy (77 %), kappa (0.741), and variance (0.000166). This set also showed a smaller value of MSE compared to all the different sets analyzed.

The Geomorphons map generated with a forty-five cell search radius was selected to represent the landforms as an input variable to predict soil classes in this watershed. The use of Geomorphons to represent landforms can improve the methods and data applied in soil surveys, providing greater information about the soil–landscape relationships.

Acknowledgements The study was supported by Purdue University—Department of Agronomy (USA), Federal Rural University of Rio de Janeiro, Soil Department—Agronomy, Embrapa Solos, and Coordination of Improvement of Higher Level Personnel—CAPES (Brazil).

References

Carvalho Junior W, Chagas CS, Fernandes Filho EI, Vieira CAO, Schaefer CEG, Bhering SB, Francelino MR (2011) Digital soilscape mapping of tropical hillslope areas by neural networks. Sci. Agric., Braz. 68(6): 691-696.

Carvalho Junior W, Chagas CS, Muselli A, Pinheiro HSK, Rendeiro NP, Bhering SB (2014). Conditioned Latin Hypercube method for soil sampling in the presence of environmental covariates for digital soil mapping. R. Bras. Ci. Solo 38:386-396.

Chagas CS, Carvalho Junior W, Bhering SB (2011) Integração de dados do quickbird e atributos do terreno no mapeamento digital de solos por redes neurais artificiais. R. Bras. Ci. Solo 35:693-704.

CHAGAS, C. S.; VIEIRA, C. A. O.; FERNANDES FILHO, E. I. (2013) Comparison between artificial neural networks and maximum likelihood classification in digital soil mapping. Rev. Bras. Ciênc. Solo 37 (2): 339-351. ISSN 0100-0683.

Chen T, Niu R, Li P, Zhang L, Du B (2011) Regional soil erosion risk mapping using RUSLE, GIS, and remote sensing: a case study in Miyun Watershed, North China. Environ Earth Sci doi:10.1007/s12665-010-0715-z.

Choi J, Oh H, Won J, Lee S (2010) Validation of an artificial neural network model for landslide susceptibility mapping. Environ Earth Sci. 60:473–483.

CONGALTON, R. G. and GREEN. K. (1999). Assessing the accuracy of remotely sensed data: principles and practices. New York: Lewis Publishers. 137p.

Cortes, M.B.V. (2010). Management of water for human consumption: microbiological and parasitological diagnosis of the Macacu, Caceribu and Guapi-Macacu rivers, State of Rio de Janeiro, Brazil. (Master Thesis). Universidade Federal Fluminense. Niterói, RJ.

CPRM - Companhia de Pesquisa de Recursos Minerais (2001). Serviço Geológico do Brasil. Mapas Geoambientais. Estado do Rio de Janeiro. Ministério de Minas e Energia, Brasília (DF). CD-ROM.

DANTAS. M.E. (2000) Estudo geoambiental do Estado do Rio de Janeiro. Geomorfologia do Estado do Rio de Janeiro. Ministério de Minas e Energia. Secretaria de Minas e metalurgia. CPRM – Serviço Geológico do Brasil. Brasília. 1 CD-ROM.

ECOLOGUS- AGRAR. (2003). Plano Diretor dos Recursos Hídricos da Região Hidrográfica da Baía de Guanabara. Rio de Janeiro, RJ. 3087p. CD-ROM.

Ehsani AH, Quiel F (2008) Geomorphometric feature analysis using morphometric parameterization and artificial neural networks. Geomorphology 99: 1–12.

Foody, G. M., Arora, M. K. (1997). An evaluation of some factors affecting the accuracy of classification by an artificial neural network. International Journal of Remote Sensing, 18: 799-810.

Iwahashi J, Pike RJ (2007) Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. Geomorphology 86: 409–440.

JASIEWICZ, J.; NETZEL, P.; STEPINSKI, T. F. (2014). Landscape similarity, retrieval, and machine mapping of physiographic units. Geomorphology 221: 104–112.

Jasiewicz J, Stepinski TF (2013) Geomorphons — a pattern recognition approach to classification and mapping of landforms. Geomorphology 182:147–156.

KÖPPEN, W. (1948). Climatologia: con un estudio de los climas de la tierra. Fondo de Cultura Econômica. México. 479p.

Schmidt J, Hewitt A (2004) Fuzzy land element classification from DTMs based on geometry and terrain position. Geoderma 121:243–256.

Tranter G, Minasny B, Mcbratney AB, Murphy B, Mckenzie NJ (2007) Grundy, M.; Brough, B. Building and testing conceptual and empirical models for predicting soil bulk density. British Society of Soil Science, Soil Use and Management. 23:437–443.

TSO, B., and MATHER, P. M. (2009). Classification Methods for Remotely Sensed Data (2nd ed.). Boca Raton, FL: CRC Press (356 pp.).

Minasny B, McBratney AB (2006) A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences. 32:1378-1388.

Motaghian HR, Mohammad IJ (2011) Spatial Estimation of Saturated Hydraulic Conductivity from Terrain Attributes Using Regression, Kriging, and Artificial Neural Networks. Pedosphere 21(2):170–177.

Pinheiro, H.S.K. (2012). Digital soil mapping by artificial neural network of the Guapi-Macacu watershed, RJ. (Master Thesis). Federal Rural University of Rio de Janeiro. Seropédica, RJ.

Roudier, P., Beaudette, D.E.; Hewitt, A.E. (2012). A conditioned Latin hypercube sampling algorithm incorportaing operational constraints. In: Digital Soil Assessments and Beyond. Proceedings of the 5th Global Workshop on Digital Soil Mapping, Sydney, Australia.

WRB. World Reference Base for Soil Resources (2014) FAO, Rome. 191p. (World Soil Resources Reports, No. 106).

Yang, W.; Yang, L.; Merchant, J.W. (1997). An assessment of AVHRR/NDVI-ecoclimatological relations in Nebraska. USA. International Journal of Remote Sensing, v.10. p.2161-2180. 1997.

ZHU, A.X. (2000). Mapping soil landscape as spatial continua: the neural network approach. Water Resources Research 36: 663-677.

Chapter 9
Comparison of Traditional and Geostatistical Methods to Estimate and Map the Carbon Content of Scottish Soils

Nikki Baggaley, Laura Poggio, Alessandro Gimona and Allan Lilly

Abstract The Scottish Government wish to preserve the carbon stocks already stored or sequestered in both organic and mineral soils and see land-use change as one of the key drivers affecting storage of soil organic carbon (SOC). A key component to develop any strategy to maintain the existing carbon stocks is the quantification of these stocks both in terms of the carbon content and its spatial distribution. To date, two different methods that use the same existing legacy data have been used to quantify carbon stocks in Scotland: a traditional approach and a hybrid generalised additive model (GAM)—geostatistical 3D model. Each of the methods revealed differences in the spatial patterns of SOC stocks. Understanding these differences will enable the development of more robust and accurate models that can be used to assess changes in stocks due to changing land use. Here, we compare these methods for the Scottish mainland, Western Isles, and Orkney. The traditional approach was based on calculating average organic carbon values from a subset (6000) of around 40,000 observations stored within the Scottish Soil Database. The total SOC stock was then determined by multiplying the areal extent of each soil series/land-use combination by the calculated profile stock. The uncertainty was also quantified based on standard error of the measured carbon contents and the uncertainty in the bulk density pedotransfer functions. A hybrid GAM-geostatistical 3D model combined the fitting of a GAM using a 3D smoother with related covariates and the kriging or Gaussian simulations of the residuals to spatially account for local details. The uncertainty was also calculated and was found to be large, indicating a wide range of credible values for each pixel. The deviation from the median ranges was between 5 and 75 % for the interpolated values depending on location.

N. Baggaley (✉) · L. Poggio · A. Gimona · A. Lilly
The James Hutton Institute, Aberdeen AB15 8QH, Scotland, UK

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_9

Keywords Carbon stocks · GAM-geostatistical 3D model · Soil legacy data · Soil mapping

9.1 Introduction

Improving soil and environmental management requires spatially explicit information about soil properties, processes, and variation across landscapes. Reliable estimates of regional soil organic carbon (SOC) stocks and their spatial variability and uncertainty are essential to better understand their vulnerability to direct and indirect climate and land-use change impacts (Mishra and Riley 2012). Knowledge of the spatial distribution of soil carbon is important for numerous reasons. It provides input values for simulation models and baseline values for the assessment of change as well as aiding in the understanding of the variables affecting carbon stocks and in the identification of areas where stocks are more vulnerable to changing environmental conditions or management. Numerous estimations of soil carbon stocks exist at different scales from global to local (see Minasny et al. (2013) for a recent review). Most of the studies considered in the review do not provide any measure of uncertainty for the results presented, and about half of them do not use validation. We compare two approaches to mapping soil carbon in which both consider uncertainty in the input data.

The first is a traditional approach using soil- and land-use polygons combined with representative soil profiles (referred to as SP approach in the text). This approach has been used in the USA to produce maps of carbon concentration for the global soil map (Odgers et al. 2012). Carbon concentrations and predicted bulk densities have been combined to predict carbon stocks and potential carbon storage in soils using 1:1,000,000 soil typological units for Europe (Stolbovoy and Montanarella 2008) and 1:250,000 scale map unit data in Scotland (Lilly and Baggaley 2013).

The second approach (Poggio and Gimona 2014) based on a 3DGAM (generalised additive model) coupled with 3D kriging (referred to as 3DGAM + GS in the text) was used and compared to other methods such as mass-preserving splines (Malone et al. 2009) and regression kriging (Hengl et al. 2004). The method takes into account the spatial neighbour information in both lateral and vertical dimensions, and at the same time, building relationships with the relevant covariates which were selected to describe the most important *scorpan* factors (McBratney et al. 2003).

Scotland provides an ideal landscape in which to test these 3D approaches to mapping carbon stocks as the soils includes mineral, organo-mineral soils, and deep peat soils (Fig. 9.1), with differing carbon contents. The land-use and vegetation

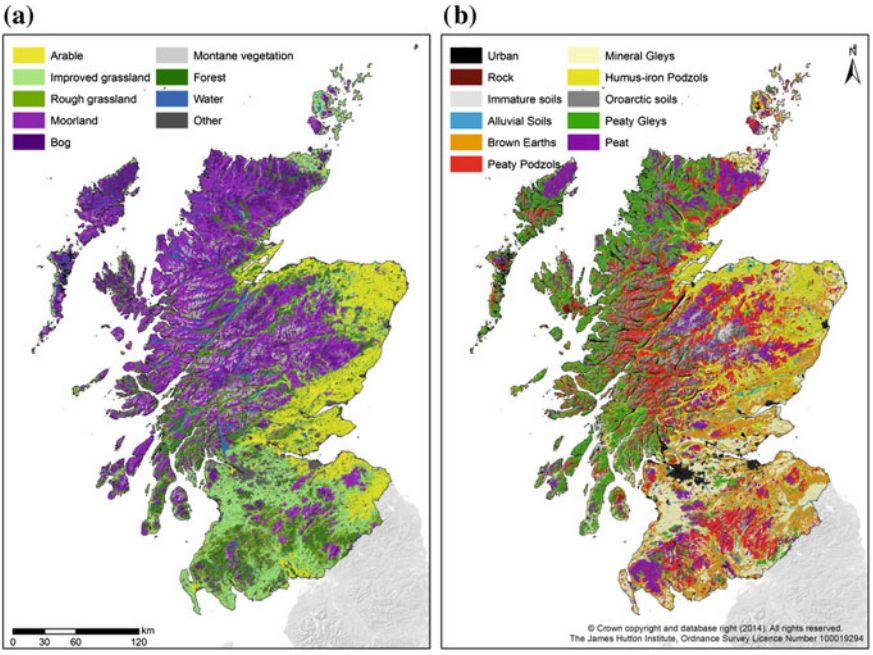

Fig. 9.1 Maps of the Scottish mainland, Western Isles, and Orkney **a** the dominant land-use types based on aerial photograph interpretation from the land cover of Scotland 1988 map (MLURI 1993). **b** The dominant major soil types based on the Scottish soil classification showing the distribution of organic, organo-mineral, and mineral soils Scotland (Soil Survey of Scotland Staff 1981)

types also vary and include arable, grassland, moorland, and forestry (Fig. 9.1) allowing a comparison between the mapped stocks under differing land uses defined from aerial photography and those derived from satellite sensors.

9.2 Methods

The SP approach was based on calculating average organic carbon values from a subset of around 40,000 observations stored within the Scottish Soil Database. The average SOC concentration was calculated for individual horizons deemed as typical of individual soil types (soil series taxonomic unit) and taking account of whether the soil was cultivated or not. The stock for each horizon (calculated from horizon thickness, average SOC concentration, and predicted bulk density) was summed to give an estimate for a modal soil profile typical of each individual soil

series to a depth of 1 m and spatially extrapolated using the 1:250,000 scale national soil map of Scotland (Soil Survey of Scotland Staff 1981) combined with land-use data (MLURI 1993). Uncertainty estimates around the mean carbon stock were calculated based on the 95 % prediction intervals on the pedotransfer functions used to calculate bulk density combined with the standard error on the measured carbon concentrations.

The 3DGAM + GS approach used a subset of 26,000 horizons (7800 profiles) from the same soil database, with 75 % selected for model development and 25 % randomly selected as a validation data set. Numerous covariates derived from globally available data, such as MODIS and SRTM, were considered. In this study, the continuous vertical and lateral distributions of carbon stocks in Scottish soils were modelled with a 3DGAM + GS approach. The approach used involves (1) GAM modelling of the trend with full 3D spatial correlation, i.e. exploiting the values of the neighbouring pixels in 3D space and (2) 3D kriging to interpolate the residuals. The values at each cell for each of the considered depth layers were predicted with a hybrid GAM-geostatistical 3D model, combining the fitting of a GAM to estimate the trend of the variable, using a 3D smoother with related covariates and Gaussian simulations of the model residuals as spatial component to account for local details. The total SOC stock for the profile was obtained summing the values at each depth. The uncertainty was calculated with a high number of simulations for both the trend predictions and the residuals interpolation.

In order to undertake a comparison, the stocks from the SP approach were aggregated to the same 1-km grid used in the 3DGAM + GS and grid squares where there was no soil present (e.g. urban areas, rock, and water) were masked out. The differences in the stocks for each grid were then mapped. To further explore the differences, we used the land-use and soil map polygons to calculate the total stocks based on broad soil and land-use classes.

9.3 Results

For the 3DGAM + GS, the uncertainty was large indicating a wide range of credible values for each pixel. The deviation from the median ranges was between 5 and 75 % for the interpolated values depending on location (Fig. 9.2b). The uncertainty limits calculated for a given map unit using the SP approach, based on bulk densities and measured carbon contents, were of a similar order of magnitude (Fig. 9.2a).

When the carbon stocks in the 1 km^2 were compared (Fig. 9.3), the greatest similarities were amongst the predominantly mineral soils of the agricultural areas in the east. There is greater variability between the two approaches in the north and west where organic and organo-mineral soils dominate and the terrain is more variable.

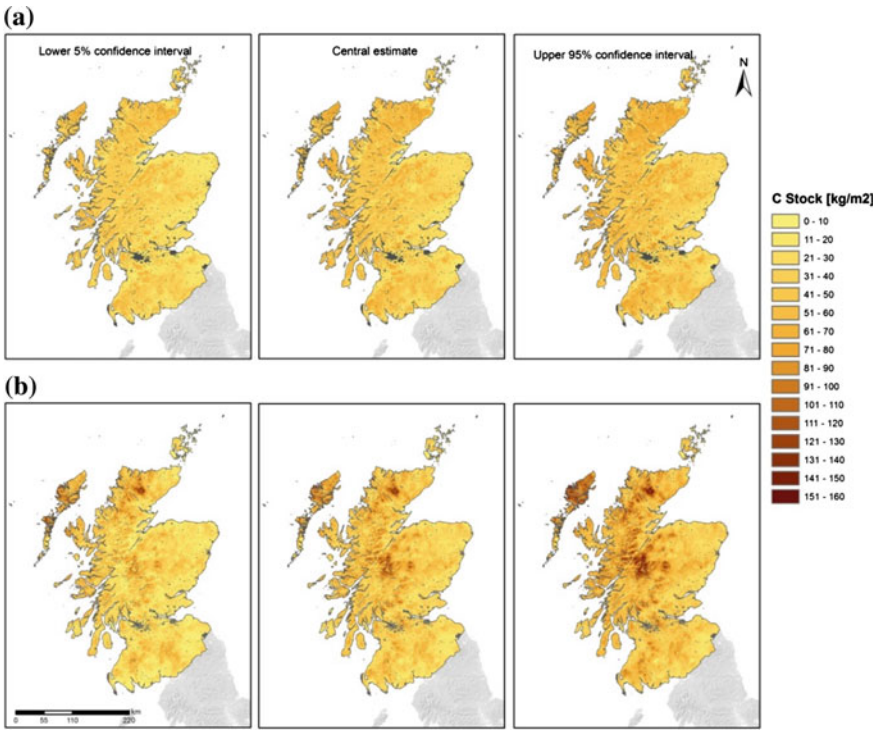

Fig. 9.2 **a** Carbon stocks using soil and land-use polygons and aggregated soil legacy data (SP approach: upper 3 maps) **b** carbon stocks using a hybrid GAM-geostatistical 3D model (3DGAM + GS: lower 3 maps)

When the stocks are compared by soil type (Fig. 9.4), the 3DGAM + GS predicts much greater stocks in the oroarctic (high altitude, cryoturbated soils) and immature soils (Rankers, Lithosols, and Regosols) compared with the SP approach, with almost double the stock of carbon in oroarctic soils being predicted by the 3DGAM + GS. The SP approach, however, predicts much greater stocks of carbon in areas of deep peat soils. The differences in peat soils are greatest in the far north where deep blanket peat bogs dominate the landscape (Fig. 9.1b).

When summarised by vegetation type, the two approaches show much more similarity in stocks (Fig. 9.5) and the greatest differences are seen under moorland with the 3DGAM + GS showing greater stocks than the SP approach. The SP approach, however, shows greater stocks under forest and bog.

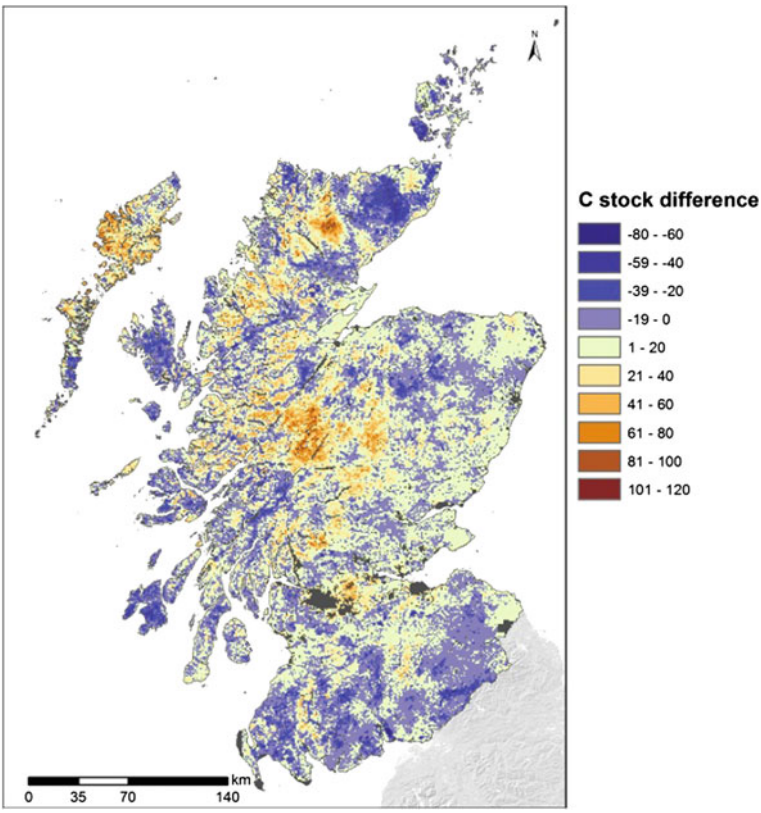

Fig. 9.3 Differences in the central estimates of carbon stock. Aggregated stocks from the SP approach subtracted from the results of the 3DGAM + GS. Grid cells where SP approach predicts greater stocks are shown in blue. Grid cells where 3DGAM + GS predicts grater stocks are shown in *orange*

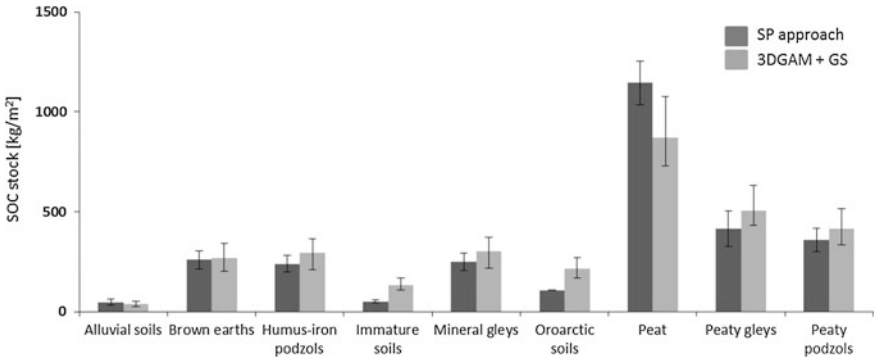

Fig. 9.4 Differences in the central estimates of carbon stock by main soil types

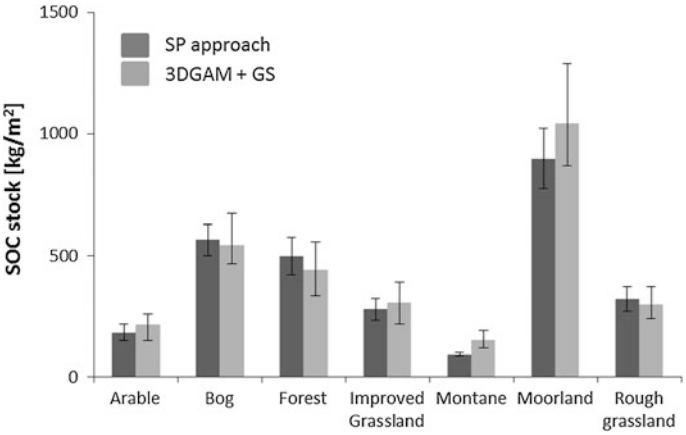

Fig. 9.5 Differences in the central estimates of carbon stock by broad habitats

9.4 Discussion

The spatial distributions of SOC stocks for both methods are quite different, especially for organo-mineral soils and soils under moorland vegetation. The differences between the central values, when aggregated by soil type, are greater than that when the data are aggregated by land use. This may due to the way individual vegetation communities were amalgamated into broad habitat types.

However, further work is needed to formally test the significance of the differences, between soils and land cover types. Further work will also compare the carbon stocks for each of the depths (0–5, 5–15, 15–30, 30–60, and 60–100 cm) defined for the global soil map. This will allow consideration of how the methods predict the varying carbon contents at depth and how the aggregation and smoothing processes in both approaches predict the stocks in a wide variety of soils with varying carbon contents in horizons throughout the soil profile. Further exploration of the difference in total soil depth and bulk densities through the soil profile will also be analysed.

The large differences in the two approaches to quantifying stocks in immature (often shallow soils) and soils of the high, exposed mountain tops may be due to the small-scale variations in topography or exposure in these areas where soils can change from having an organic-rich mineral topsoil to an organic topsoil within a distance of 5 m. In order to further compare the methods, an assessment using independent samples is needed.

The uncertainty limits in the SP approach are probably underestimated as we only consider the uncertainty in the carbon contents and in the predictions of bulk density and have currently ignored uncertainty in estimates of horizon thicknesses and stone contents. There is also >10 % of stocks for which the uncertainty could not be quantified as there was only one record for that individual soil in the database with

which to characterise the soil horizon or where bulk densities could be calculated with the pedotransfer functions. Further work will therefore seek to better quantify the uncertainty in this approach perhaps through identifying similar soils where there are more data or by amalgamating soils to a higher level in the classification system. Additionally, more work is required to identify why there appear to be greater differences between the two approaches for soils with organic surface layers, that is, Peats, Peaty gleys and Peaty podzols ((Soil Survey of Scotland Staff 1984), Histosols, Histic Gleysols, and Histic Podzols (IUSS Working Group WRB 2014)).

9.5 Preliminary Conclusions

- The stocks for the 3DGAM + GS are more smoothly distributed than those for the SP approach, as expected due to the method used for interpolation.
- The stocks in mineral soils under cultivation are the most similar, and the greatest differences in predicted carbon stocks are in peats and organo-mineral soils under moorland and montane vegetation. However, further work is needed in order to assess if the differences (i.e. between methods and between soil types or land uses) are statistically significant.
- The SP approach has lesser stock estimates than the 3DGAM + GS approaches in parts of the central highlands, south of the Great Glen (where there is a large proportion of immature and oroarctic soils).

Acknowledgements We acknowledge the funding from the Scottish Government's Rural and Environment Science and Analytical Services Division and the assistance of Biomathematics and Statistics Scotland. Maps contain Ordnance Survey data © crown copyright and database right (2014). All rights reserved. The James Hutton Institute Ordnance Survey License Number 100019294.

References

Hengl T, Heuvelink G, Stein A (2004) A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 122 (1–2), 75–93. doi: 10.1016/j.geoderma.2003.08.018

IUSS Working Group WRB (2014) World Reference Base for Soil Resources 2014. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106. FAO, Rome.

Lilly A, Baggaley NJ (2013). The potential for Scottish cultivated topsoils to lose or gain soil organic carbon Soil Use and Management 29:39-47. doi: 10.1111/sum.12009

Malone B, McBratney A, Minasny B, Laslett G (2009) Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154 (1–2), 138–152. doi:10.1016/j.geoderma.2009.10.007

McBratney A, Santos M, Minasny B (2003) On digital soil mapping. Geoderma 154 (1–2), 3–52. doi:10.1016/S0016-7061(03)00223-4

Minasny B, McBratney A, Malone B, Wheeler I (2013) Digital soil mapping of carbon. Adv. Agron. 118:1–47. doi:10.1016/B978-0-12-405942-9.00001-3

Mishra U, Riley W (2012) Alaskan soil carbon stocks: spatial variability and dependence on environmental factors. Biogeosciences 9:3637–3645. doi:10.5194/bg-9-3637-2012

MLURI (1993) The land cover of Scotland 1988. Macaulay Land Use Research Institute, Aberdeen. http://www.macaulay.ac.uk/explorescotland/lcs_mapformat.html

Odgers NP, Libohova Z, Thompson JA (2012) Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale Geoderma 189–190:153–163. doi: 10.1016/j.geoderma.2012.05.026

Poggio L, Gimona A (2014) National scale 3D modelling of soil organic carbon stocks with uncertainty propagation - An example from Scotland Geoderma 232–234: 284–299. doi: 10.1016/ j.geoderma. 2014.05.0040016-7061

Soil Survey of Scotland Staff (1981) Soil maps of Scotland at a scale of 1:250000. Macaulay Institute for Soil Research, Aberdeen.

Soil Survey of Scotland Staff (1984). Organization and Methods of the 1:250 000 Soil Survey of Scotland. Macaulay Institute for Soil Research, Aberdeen.

Stolbovoy V, Montanarella L (2008) Application of Soil Organic Carbon Status Indicators for policy-decision making in the EU. In: Threats to soil quality in Europe. (eds G. Toth, L. Montanarella & E.Rusco), pp. 87–99. Institute for Environment and Sustainability Land Management and Natural Hazards Unit, Joint Research Centre. European Commission. EUR 23438 EN, Luxembourg.

Part II
Environmental Application and Assessment

Chapter 10
Digital Soil Mapping for Hydrological Modelling

George M. van Zijl, Johan J. van Tol and Eddie S. Riddell

Abstract Digital soil mapping approaches can play a role in providing soil information in a format useful to hydrological modellers, thus filling a void in the current state of hydrology. In this paper, it is shown how an expert knowledge-based digital soil mapping approach was used to provide the soil-related input needed for a process-based hydrological model (ACRU) of the Stevenson Hamilton Research Supersite (SHRS) in the Kruger National Park, South Africa. First, a soil map was created for the entire 4001 ha study area. This soil map had a validation point accuracy of 73 %. Thereafter, the study area was divided into hillslopes. The hillslopes combined with the soil map were used to create a map showing the size and position of the hillslope-specific conceptual hydrological response models (CHRMs). The CHRM map was then used to configure ACRU and to model stream flow in a first-, second- and third-order catchment within the larger area. The stream flow modelling proved successful for the second- and third-order catchments, with Nash–Sutcliffe model efficiency coefficients (NS) of 0.79 and 0.73 for the two catchments, respectively. That the first-order catchment did not model well was explained by the level of detail of the soil mapping which was too coarse to model such a small catchment successfully. All configurations of ACRU modelled the third-order catchment very well (NS between 0.75 and 0.79), but failed to map single rain events consistently. This work showed that digital soil mapping can provide the soil information necessary to configure a process-based stream flow model successfully, provided that the scale of the mapping corresponds with the scale of the first-order controls of the process being modelled. It was indicated that the optimal time frame for this form of hydrological modelling is a hydrological season.

G.M. van Zijl (✉)
University of the Free State, PO Box 339, Bloemfontein 9301, South Africa
e-mail: george@dsafrica.co.za

J.J. van Tol
Department of Agronomy, University of Fort Hare, Alice 5700, South Africa

E.S. Riddell
Centre for Water Resources Research, University of KwaZulu-Natal, Scottsville 3209, South Africa

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_10

Keywords ACRU · Conceptual hydrological response model · Hydropedology · Kruger National Park · SoLIM

10.1 Introduction

Soil plays an integral role in hydrology, as it can transmit, store and react with water (Park et al. 2001). In the same way, water plays a primary role in the formation of soils. Soil genesis is a function of climate, organisms, relief, parent material and time (Jenny 1941), but it is largely the influence of the soil-forming factors on the hydrology which determines the influence of the soil-forming factors on soil formation. As a result of this interaction, soil carries the marks of the soil water regime under which it formed as morphological hydrological signatures, such as gleying, mottles, concretions and carbonate deposits. Valuable information regarding hydrological processes (Ticehurst et al. 2007; Van Tol et al. 2010) and hillslope hydrological behaviour (Lin et al. 2006) can be gained by careful interpretation of the hydrological signatures. Qualitative two-dimensional hillslope-based conceptual hydrological response models (CHRMs) can be created from the interpretation of such signatures. Integrating the two-dimensional CHRMs into three-dimensional catchments can assist in the making of predictions in ungauged basins (PUBs). Thus, hydropedological knowledge is increasingly sought after in the quest to make PUBs, because of the difficulty to observe and measure important hydrological processes (Sivapalan 2003).

However, soil information is often not optimally used in hydrological modelling. Part of the reason for this is that the quality of soil information at a spatial point is often in a form that cannot be successfully extrapolated to be representative of the hydrological response unit, due to heterogeneity and thus uncertainty. Thus, soil data usefulness to hydrological models on a basin scale is cumbersome and typically a considerable expense for modelling studies. Digital soil mapping can provide the answer to this dilemma. In this paper, an expert knowledge-based digital soil mapping approach was used to create a hillslope-based CHRM map for the Stevenson Hamilton Research Supersite (SHRS) within the Kruger National Park, South Africa. The CHRM map was used to configure ACRU, a process-based agrohydrological model (Schulze 1995). It uses a daily time step and multiple soil layers and can run in lumped or distributed mode. In lumped mode, average soil parameters are used across the catchment and no configuration of the soils of the landscape is necessary in the model. There are two distributed modes. The standard mode, ACRU2000, allows for two soil layers, an A and B horizon and a deep groundwater layer. In a revised version, ACRU-Int Lorentz et al. (2007) added an intermediate layer between the B horizon and the groundwater store. Soil inputs are as follows: thickness of soil horizon, water contents at the start of simulation, permanent wilting point, drained upper limit, saturation, plant available water, drainage rates and the soil erodibility factor.

The aims were to create an acceptable soil map of the area, to use this map to create a conceptual hydrological response unit (CHRU) map of the entire area and use this map to configure ACRU and then to assess the model outputs on different temporal and spatial scales. The hypothesis tested is that an accurate homogenized spatial representation of soil information will lead to more accurate model outputs.

10.2 Site Description

The Kruger National Park is a pristine savannah conservation area of some 2,000,000 ha in north-eastern South Africa (Fig. 10.1b), bordering Mozambique and Zimbabwe. In 2013, the Research Supersites were established (Smit et al. 2013), with the idea of attracting the various ecosystem-focused research programmes to the same area, in order to enable integrated research findings and allow data sharing, to inform conservation management. The SHRS (Fig. 10.1a) consists of 4001 ha and is located in the south-western part of the park, in the Renosterkoppies land type (Venter 1990), which is characterized by a highly dissected landscape with a high stream density cutting through the granite and gneiss of the Nelspruit suite. The mean annual precipitation is 560 mm/a (Smit et al. 2013), which falls predominantly

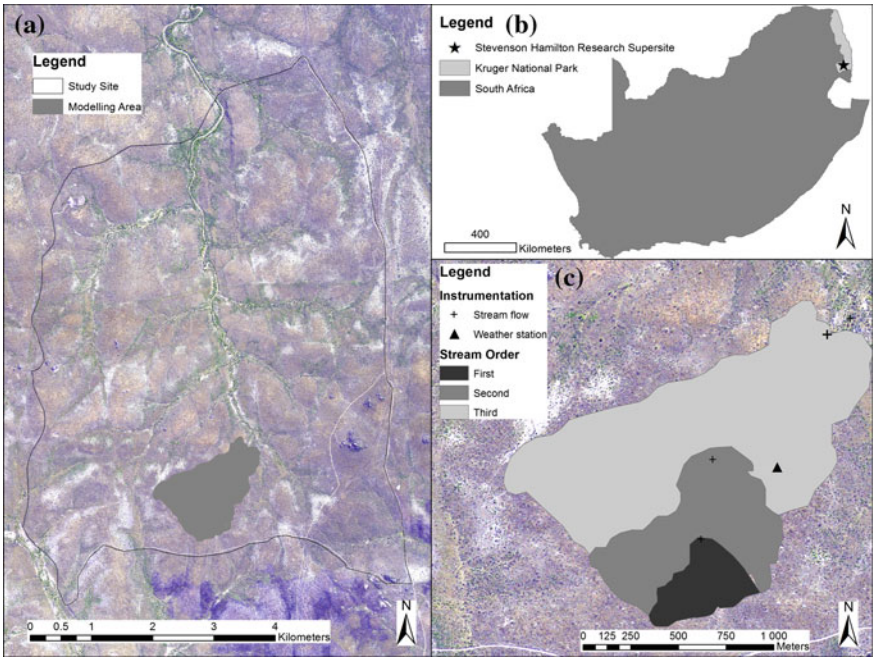

Fig. 10.1 The Stevenson Hamilton Research Supersite (**a**), the location of the Kruger National Park within South Africa (**b**) and the three catchments wherein stream flow was modelled (**c**)

in the summer months (September to March). Typical bushveld vegetation occurs, with a very good correlation between the woody vegetation, terrain position and soil type (Venter 1990). For the stream flow modelling, three small catchments where discharge data exist (ascertained through rated channels with pressure transducers) were used. This site included a first-, second- and third-order catchment (Fig. 10.1c).

10.3 Materials and Methods

Van Zijl and Le Roux (2014) created a functional hydrological soil map of the SHRS by applying an expert knowledge digital soil mapping approach, using SoLIM (Zhu 1997). Covariate data used included Spot 5 (Spot image 2013), Landsat 7 (USGS 2013), the Stellenbosch University DEM (SUDEM) (Van Niekerk 2012) (interpolated to both 10 m and 30 m resolutions) and remotely sensed evapotranspiration and biomass data for a series of dates (eLEAF 2013). One hundred and thirteen soil observations (Fig. 10.1c) were classified according to the South African soil classification system (Soil classification Working Group 1991). Stoniness, hand-estimated texture, mottles and structure were also noted per soil horizon. Of the 113 observations, the positions of 29 were predetermined using conditioned Latin hypercube sampling (Minasny and McBratney 2006), and another 25 were predetermined by smart sampling, using colour aerial photographs of the site. By visual inspection, the colour photographs were divided into five classes, each comprising of homogeneous colour units with a unique colour signature. Soil observations were then placed by hand within each of the colour units. Out of these potential observations, 25 sites were chosen which are based on accessibility, but also ensuring good spatial coverage of the site, as well as at least three observations within each colour unit. The remaining 59 were determined by the soil surveyors in the field. The latter group was used for the validation observations and the predetermined observations as training data. Soil observations were grouped into hydrological soil associations (Table 10.1) and mapped as such. This work is described in Van Zijl and Le Roux (2014). Validation was expressed as a point accuracy percentage, with a one pixel buffer around the soil mapping units (SMUs) allowed as in Van Zijl et al. (2012).

Following Van Tol et al. (2013), a hydrological response to each soil map unit was assigned, thus creating a hydrological soil map. Using the 30 m DEM, a hillslope map of the whole site was created using ArcGIS. By superimposing the hillslope map onto the hydrological soil, specific CHRMs could be devised for every hillslope, thus creating a CHRM map.

Van Tol et al. (2015) used this map to configure ACRU for three small catchments within the site, where hydrological measurements were available for the hydrological season 2012–2013. The three catchments are a first-order (10.8 ha), second-order (42.7 ha) and third-order (148.2 ha) catchment, respectively, which allowed for the spatial scale of modelling to be assessed. New to this paper, the model was also assessed on a temporal scale, using three specific large rain events,

Table 10.1 Descriptions of the soil map units

Soil association	Soil forms[a]	WRB reference groups[b]	Determining characteristics	CHRU[c]
Sodic site	Sterkspruit	Solonetz, planosols	Abrupt textural transition between the topsoil and subsoil. Redox morphology in C horizon	Responsive
Clayey interflow	Sepane, Bonheim	Luvisols, phaeozems	High clay percentage in B horizon. Redox morphology in C horizon	Interflow
Clayey recharge	Bonheim, Valsrivier, Swartland, Milkwood, Mayo	Phaeozems, luvisols, leptosols	High clay percentage in A and/or B horizon. No redox morphology in C horizon	Recharge
Sandy interflow	Tukulu, Pinedene, Westleigh, Avalon	Arenosols	Coarse textured A and/or E horizon. Redox morphology in C horizon	Interflow
Sandy recharge	Clovelly, Oakleaf, Mispah, Glenrosa	Arenosols, leptosols	Coarse textured A horizon. No redox morphology in C horizon	Recharge
Rock outcrops	Rock	Rock	Cracked rock outcrop	Recharge
Alluvial soils	Dundee, Oakleaf, Tukulu	Fluvisols, arenosols	Coarse textured soils from alluvial deposits	Recharge

WRB World Reference Base; *CHRU* conceptual hydrological response unit
[a]Soil Classification Working Group (1991)
[b]IUSS (2007)
[c]Van Tol et al. (2013)

representing early season (4 December), mid-season (25 December) and end-season (17 January), a full hydrological season (15 November–15 March) and a full year (15 April 2012–15 April 2013), which thus allows the examination of temporal scaling responses to antecedent catchment conditions. Model accuracy was assessed based on how well the modelled stream flow matched the observed.

10.4 Results and Discussion

The soil map (Fig. 10.2a) achieved an acceptable validation point accuracy of 73 % (Table 10.2). The majority of the incorrectly mapped observations were soils of the clayey recharge association, on either the sandy interflow or clayey interflow soil map units. The soil map was converted to a hillslope-based CHRM map (Fig. 10.2b), with eleven different CHRM-type hillslopes identified (Fig. 10.3) (Van Zijl and Le Roux 2014).

The statistical analysis of the stream flow modelling outputs for the third-order catchment for the hydrological season 2012–2013 (Table 10.3) shows that the soil

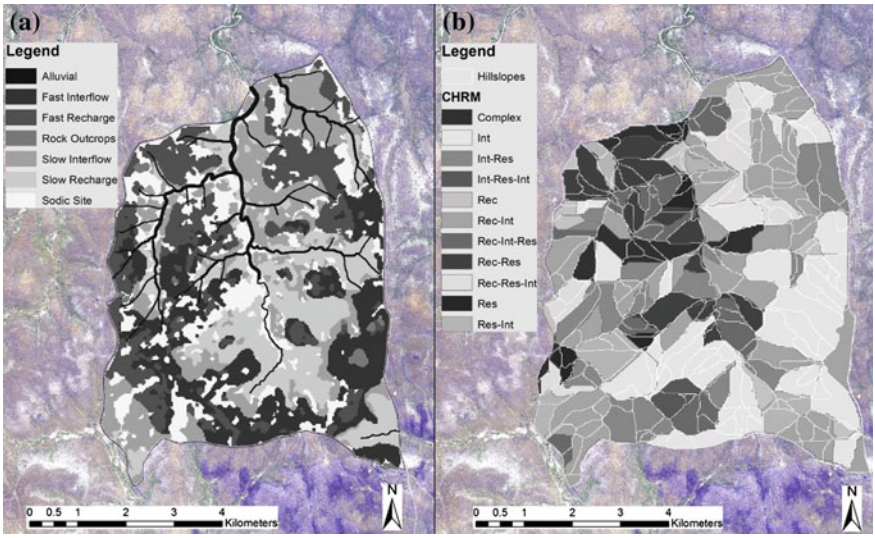

Fig. 10.2 The soil map of the area (**a**) and the conceptual response model (CHRM) map, also showing the hillslopes (**b**). *Int* interflow, *Res* responsive, *Rec* recharge

input does improve the stream flow modelling. For the second- and third-order catchments, both ACRU-Int and ACRU2000 achieved Nash–Sutcliffe model coefficients (NS) of above 0.72. This correlates well with the average NS of 0.71 which Royappen (2002) achieved in 13 catchments using ACRU and is slightly lower than the highest median model efficiency of Siebert and McDonnell (2013) of 0.8–0.85. Both ACRU-Int and ACRU2000 outperformed ACRU lumped in all three catchments, emphasizing the improvement that soil process information can yield in modelling hydrological processes. ACRU-Int also performed slightly better than ACRU2000, showing that increasing the soil detail by adding an extra soil horizon is also a valuable improvement.

The first-order catchment was, however, modelled less accurately than the second- and third-order catchments, with the outputs showing low R^2- and negative NS-values. These disappointing results are ascribed to the detail of mapping. When creating a soil map with sufficient detail for an area of 4001 ha, one will inevitably miss some detail which is important in a 10.8 ha area. This is what happened here, as the authors noted a small wetland occurring within the first-order catchment, not noted on the soil map. As the area on which the modelling was applied increased, so did the modelling accuracy. This is to be expected as the primary hydrological controls functioning in the larger area are closer to the level of detail at which the soils were mapped. As the area of modelling increases in size, the first-order hydrological controls will change (Bloschl and Sivapalan 1995). In this case, the first-order catchment level of detail of the soil map did not match the process scale of the model for the hydrological controls of that size catchment. Thus, the results of the modelling were not very good at this scale and represent a lower threshold of

Table 10.2 Confusion matrix of the soil map unit observations

		Map units								
		Sodic site	Clayey interflow	Clayey recharge	Sandy interflow	Sandy recharge	Alluvial	Total	Correct	%
Observations	Sodic	18	1	2	1	0	1	23	18	78
	Clayey interflow	0	3	0	0	0	0	3	3	100
	Clayey recharge	0	3	11	4	0	0	18	11	61
	Sandy interflow	0	0	1	5	0	0	6	5	83
	Sandy recharge	0	0	2	0	4	0	6	4	67
	Alluvial	1	0	0	0	0	2	3	2	67
	Total	19	7	16	10	4	3	59	43	73
	Correct	18	3	11	5	4	2	43		
	%	95	43	69	50	100	67	73		

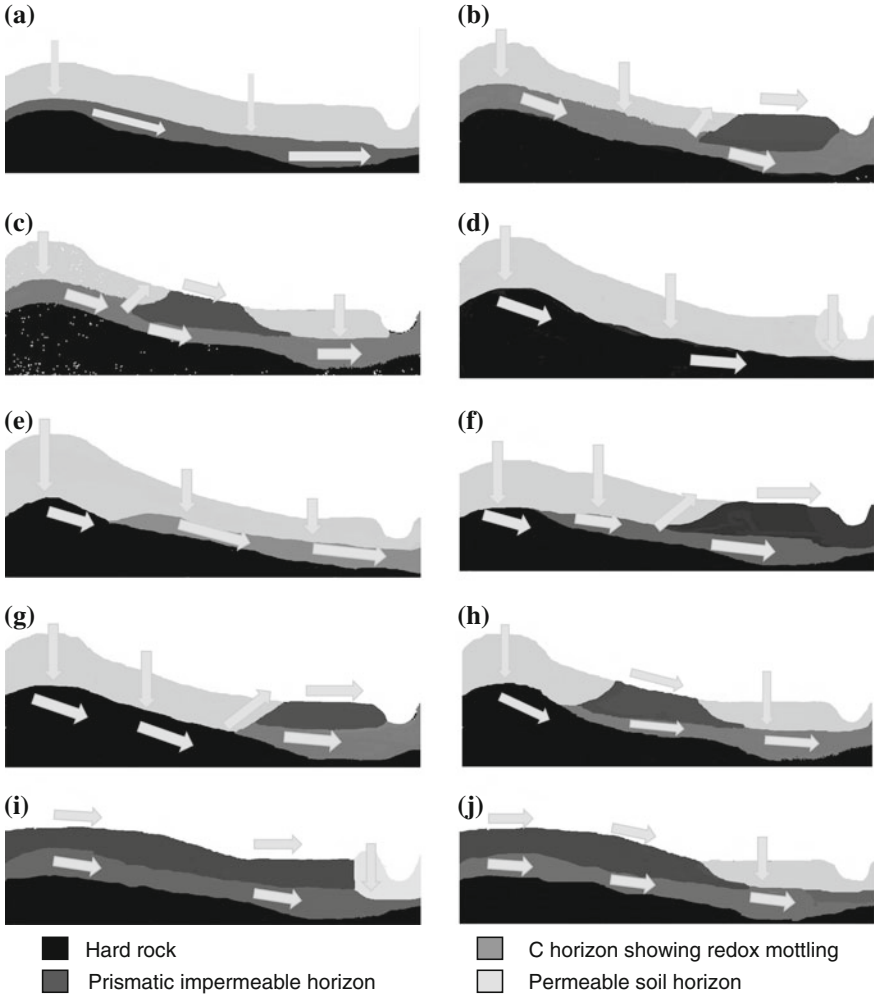

Fig. 10.3 The different hillslope CHRMs functioning within the study site. The percentage of area which each CHRM occupies is shown after the model name. **a** Interflow 25.4 %, **b** interflow–responsive 11.6 %, **c** interflow–responsive–interflow 1.2 %, **d** recharge 2.7 %, **e** recharge–interflow 25.3 %, **f** recharge–interflow–responsive 7.5 %, **g** recharge–responsive 9.0 %, **h** recharge–responsive–interflow 5.8 %, **i** responsive 2.0 %, **j** responsive–interflow 5.2 %

utility of digital soil mapping information. With increasing catchment size, at some point the level of detail of the soil map will correspond to the first-order hydrological controls functioning in the catchment. This is the optimal level of detail at which the mapping can be used as input for hydrological modelling. As the area modelled increases further, another factor such as climate will become the first-order hydrological control, and thus, the need for soil input into a hydrological model will diminish with increasing catchment size. The inconclusive results

Table 10.3 The statistical output for the stream flow values modelled against those observed, for the 2012–2013 hydrological season

Order	Mode	RMSE	R^2	NS
First	ACRU-Int	6.60	0.51	−0.71
	ACRU2000	6.22	0.57	−0.51
	Lumped	6.21	0.49	−7.62
Second	ACRU-Int	1.36	0.87	0.79
	ACRU2000	1.55	0.83	0.72
	Lumped	2.05	0.57	0.52
Third	ACRU-Int	2.63	0.91	0.73
	ACRU2000	2.67	0.90	0.72
	Lumped	2.89	0.82	0.67

From Van Tol et al. (2015)
NS Nash–Sutcliffe model efficiency coefficient

between the second and third orders suggest that the optimal spatial scale have not been reached.

The stream flow modelling output for the third-order catchment for the hydrological season 2012–2013 is shown graphically in Fig. 10.4. One can clearly see that the stream flow from rain events in the beginning of the season is overpredicted by all three modes of ACRU, while the stream flow from one very big rain event

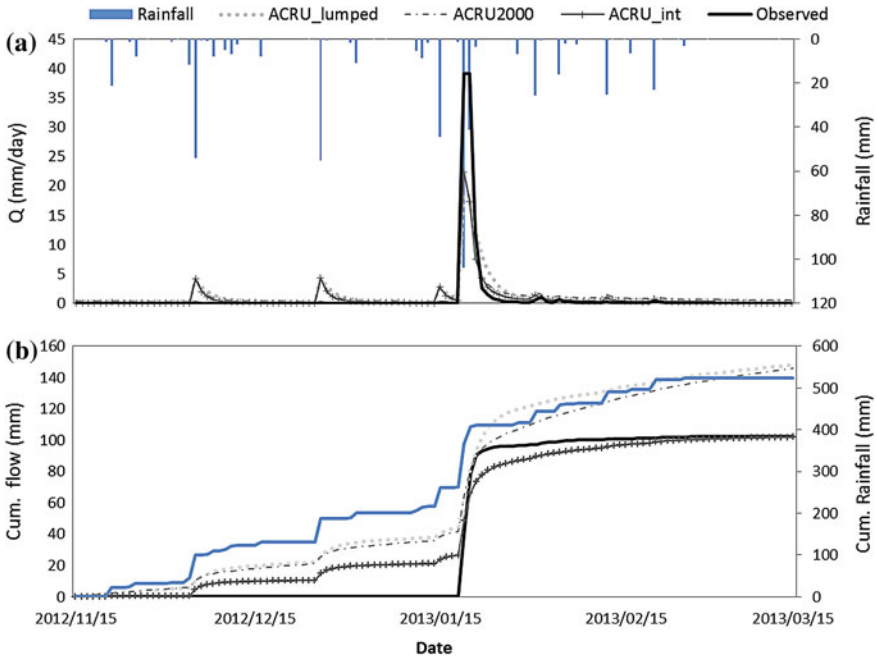

Fig. 10.4 The daily (**a**) and cumulative (**b**) stream flow model output for the third-order catchment for the hydrological season 2012–2013. From Van Tol et al. (2015)

later in the season was underpredicted. The cumulative flow of ACRU-Int closely resembled the measured cumulative flow, while ACRU2000 and ACRU lumped overestimated the flow. The same graphs for the second- and third-order flow are shown and discussed in Van Tol et al. (2015).

The modelling accuracy of specific rainfall events is shown in Table 10.4, in which mixed results were obtained. The NS peaked above 0.75 during each rain event. Specifically, the second-order catchment modelled with ACRU-Int achieved very good results. However, negative NS-values for at least one catchment for every rain event were also obtained, which means that the modelling was less accurate than an average value would have been. When looking at the R^2-values, it seems that ACRU lumped and ACRU2000 outperformed ACRU-Int frequently with very high values above 0.8 being recorded often. However, ACRU-Int also

Table 10.4 The statistical output for the stream flow values modelled against those observed, for three rain events during the hydrological season 2012–2013

Date	Order	Mode	RMSE	R^2	NS
2012/12/04–2012/12/09	First	ACRU-Int	4.41	0.32	−1.49
		ACRU2000	4.24	0.31	−1.30
		Lumped	4.19	0.59	−1.24
	Second	ACRU-Int	0.39	0.93	0.91
		ACRU2000	0.49	0.87	0.81
		Lumped	0.40	0.87	0.77
	Third	ACRU-Int	2.25	0.61	−2409.46
		ACRU2000	0.62	0.95	−2886.29
		Lumped	0.43	0.95	−3306.31
2012/12/25–2012/12/30	First	ACRU-Int	3.80	0.83	−1.83
		ACRU2000	3.67	0.83	−1.61
		Lumped	3.66	0.84	−1.59
	Second	ACRU-Int	0.89	0.65	0.80
		ACRU2000	0.45	0.86	0.63
		Lumped	0.38	0.85	0.28
	Third	ACRU-Int	2.33	0.60	−7578.78
		ACRU2000	0.61	0.94	−8666.15
		Lumped	0.45	0.94	−10301.94
2013/01/17–2013/01/25	First	ACRU-Int	6.79	0.96	−6.92
		ACRU2000	6.24	0.89	−5.68
		Lumped	5.74	0.72	−4.66
	Second	ACRU-Int	7.11	0.55	0.75
		ACRU2000	2.62	0.81	0.73
		Lumped	2.24	0.84	0.45
	Third	ACRU-Int	10.12	0.86	0.64
		ACRU2000	2.80	0.90	0.65
		Lumped	2.40	0.91	0.58

NS Nash–Sutcliffe model efficiency coefficient

achieved very high (0.93) R^2-values in some instances. Thus, the data are inconclusive as to whether or not modelling at a small timescale is accurately possible and whether or not a certain model configuration outperforms the others. When looking at the visual representation of the model outputs (Fig. 10.5), it is clear that

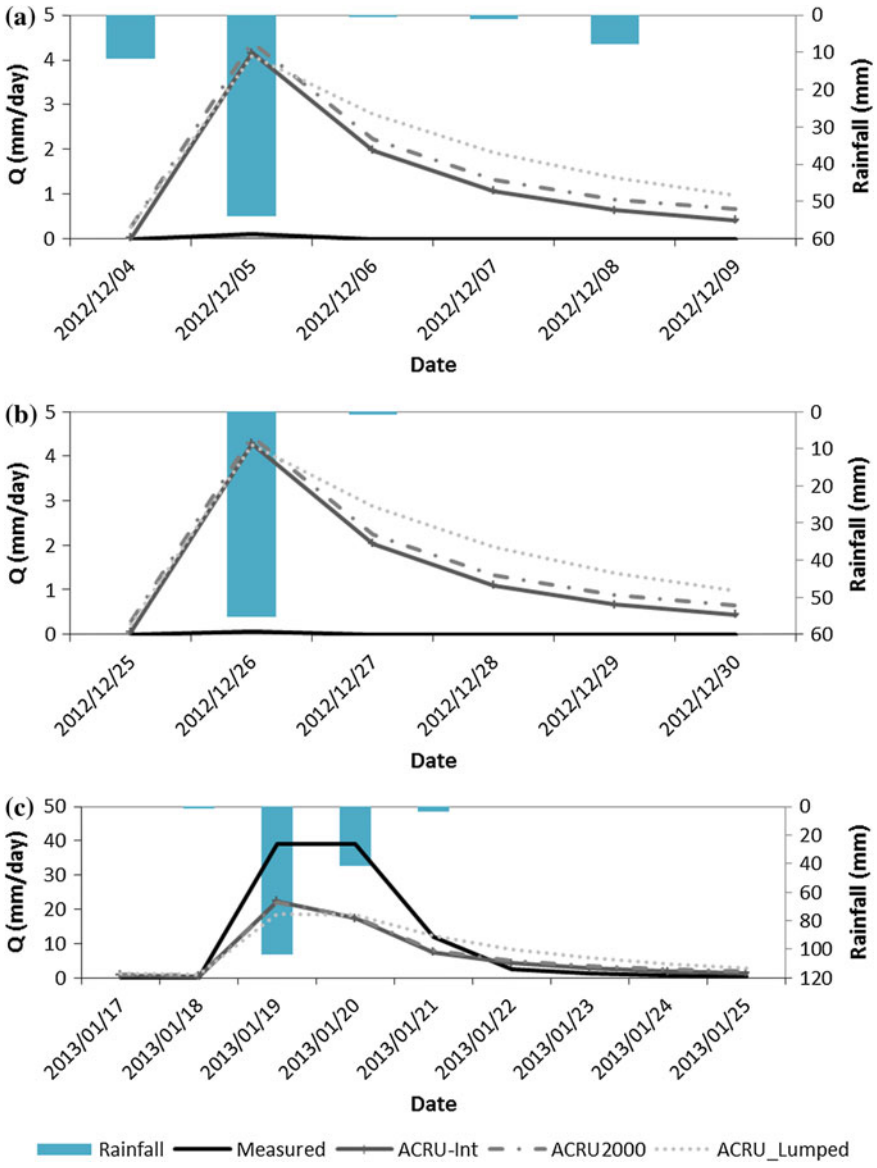

Fig. 10.5 Visual representation of the model outputs and measured stream flow values for three large rain events

in the beginning of the season, there is hardly any response to rainfall events and that the models incorrectly modelled such responses. However, for the rain event near the end of the season, all the models underestimated the stream flow. There can be two explanations for this. Either the water store of the soil is larger than anticipated, and thus the first two rain events served to fill up the soil water store, and as the soil could take no more water, most of the rainfall from the third rain event ended up as stream flow, or the sheer magnitude of the third rain event (145 mm in two days) would bring about greater stream flow than would be expected when setting up the model. Either way, the lack of consistently being able to model single rain events show that there is a gap in our understanding of hydropedological processes and our ability to model such events.

On a year scale (Table 10.5), the results are much more accurate than those on a single rain event scale. When moving from the first order to the third order, the model accuracy increased and ended with very good results for the third-order catchment. ACRU-Int slightly outperformed the other modes in the second and third orders, but performed worse in the first-order catchment. In the visual representation of the data (Fig. 10.6), one can clearly see that ACRU-Int predicted the cumulative discharge much closer to the observed values than the other two configurations. The overall accuracy of the models decreased from the hydrological season scale to the year scale, when measured with the R^2-value. This is due to the few rain events during the dry season that initiate a response in the models, but not observed to generate stream flow in reality. When measuring the model accuracy with the NS, it tells a different story, as NS-values increased for the full year for the first- and third-order catchment, but decreased for the second-order catchment. This is due to the high amount of time where no flow was recorded, influencing the average values to which the NS coefficient compares the model outputs.

Table 10.5 The statistical output for the stream flow values modelled against those observed, for the year 15 March 2012–15 March 2013

Order	Mode	RMSE	R^2	NS
First	ACRU-Int	3.26	0.42	−0.10
	ACRU2000	3.06	0.45	0.03
	Lumped	3.06	0.41	0.03
Second	ACRU-Int	2.10	0.40	0.47
	ACRU2000	2.13	0.36	0.46
	Lumped	2.30	0.23	0.37
Third	ACRU-Int	1.53	0.90	0.79
	ACRU2000	1.57	0.88	0.78
	Lumped	1.68	0.80	0.75

NS Nash–Sutcliffe model efficiency coefficient

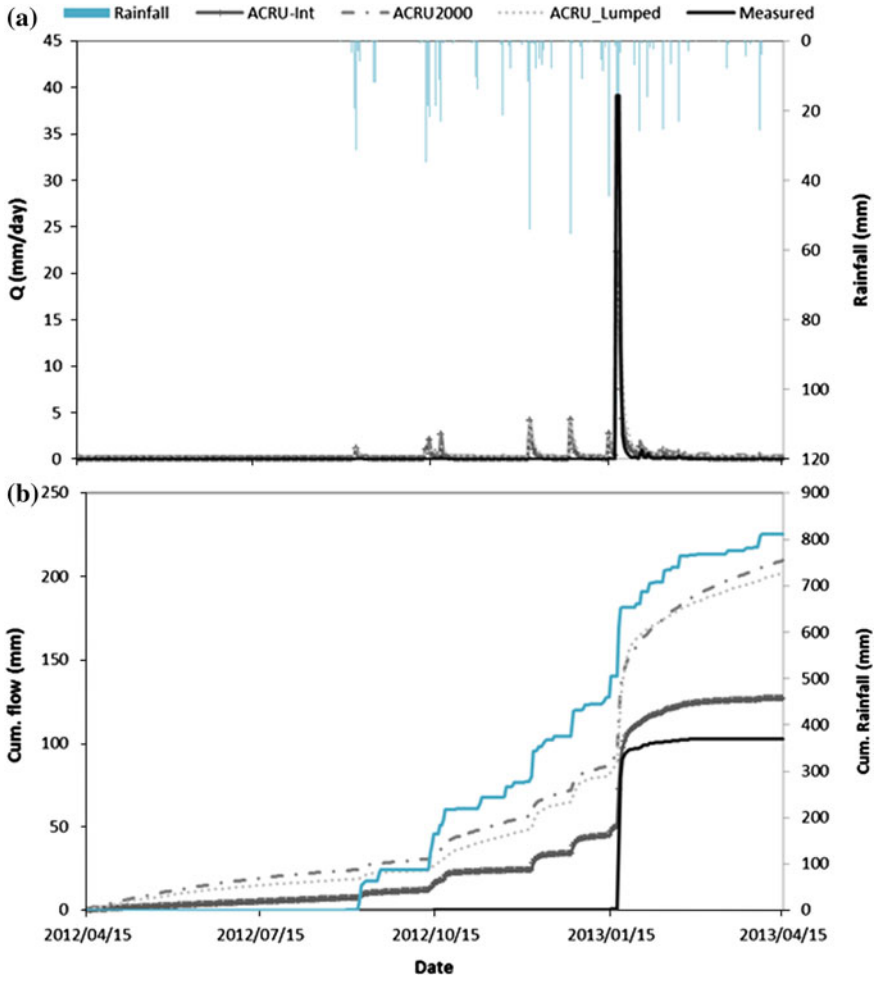

Fig. 10.6 Daily measured and modelled stream flows for the third-order catchment from 2012-04-15 until 2013-04-15 (**a**), as well as cumulative flow for the same period (**b**)

10.5 Conclusions

It was shown that a hydrological model (ACRU) could be configured using detailed soil information obtained by digital soil mapping. The more soil information was included into the model, the better the model performed. However, the model overestimated stream flow when low flow volumes were recorded and underestimated the flow with high flow volumes. This shows that there is still a lack of our understanding of hydrological processes within the soil or our ability to model those processes through hydro-pedotransfer functions. The optimal level size of area for

including soil mapping in hydrological modelling has not been confirmed but is larger than second-order catchments. In general, ACRU modelled the hydrological season well, but achieved less accurate results for a full year. Single rain events were modelled erratically. Future work should define the optimal size of area at which soil should be included into hydrological modelling and improve our understanding of soil-related hydrological processes and determine how we could model such processes.

Acknowledgements We would like to thank the South African Water Research Commission for funding this project, South African National Parks Board for hosting the research and Faith Jumbi, Ashton van Niekerk and Daniel Fudisi for collecting the hydrological data.

References

Blöschl G, Sivapalan M (1995) scale issues in hydrological modelling: A review. Hydrological Processes 9, 251-290.

eLEAF (2013) Data supplied by the Inkomati Catchment Management Agency on behalf of eLeaf (www.eleaf.com) and the WATPLAN EU project.

IUSS Working Group (2007) World reference base for soil resources 2006, first update 2007. World Soil Resources Reports No. 103. FAO, Rome.

Jenny H (1941) Factors of Soil Formation, a System of Quantitative Pedology. McGraw-Hill, New York.

Lin HS, Kogelman W, Walker C, Bruns MA (2006) Soil moisture patterns in a forested catchment: A hydropedological perspective. Geoderma 131: 345 – 368.

Lorentz SA, Bursey K, Idowu O, Pretorius C, Ngeleka K (2007) Definition and upscaling of key hydrological processes for application in models. WRC Report No. K5/1320. Water Research Commission, Pretoria.

Minasny B, McBratney AB (2006) A conditioned Latin hyper-cube method for sampling in the presence of ancillary information. Comput. Geosci. 32: 1378–1388.

Park SJ, McSweeney K, Lowery B (2001) Identification of the spatial distribution of soils using a process-based terrain characterization. Geoderma, 103: 249-272.

Royappen M (2002) Towards improved parameter estimation in streamflow predictions using the ACRU model. MSc dissertation, University of Natal.

Schulze RE (1995) Hydrology and agrohydrology: A text to accompany the ACRU 3.00 agrohydrological modelling system. WRC Report No 63/2/84. Water Research Commission, Pretoria.

Siebert J, McDonnell JJ (2013) Gauging the ungauged basin: The relative value of soft and hard data. Journal of Hydrologic Engineering. doi:10.1061/(ASCE)HE.1943-5584.0000861

Sivapalan M (2003) Prediction in ungauged basins: a grand challenge for theoretical hydrology. Hydrol. Process. 17: 3163 – 3170.

Smit IPJ, Riddell ES, Cullum C, Petersen R (2013) Kruger National Park research supersites: Establishing long-term research sites for cross-disciplinary, multiscaled learning. *Koedoe* **55** (1). doi.org/10.4102/koedoe. v55i1.1107.

Soil Classification Working Group (1991) Soil Classification: A Taxonomic System for South Africa. Department of Agricultural Development, Pretoria, South Africa.

Spot Image (2013) SPOT satellite technical data. Available from http://www.spotimage.com/web/en/229-the-spot-satellites.php (Accessed 23 June 2013).

Ticehurst JL, Cresswell HP, Mckenzie NJ, Glover MR (2007) Interpreting soil and topographic properties to conceptualize hillslope hydrology. Geoderma 137: 279–292.

USGS (United States Geological Survey) (2013) Landsat images. URL: http://landsat.usgs.gov (Accessed 23 June 2013).

Van Niekerk A (2012) Developing a very high resolution DEM of South Africa. *Position IT* Nov-Dec: 55-60.http://www.eepublishers.co.za/images/upload/positionit_2012/visualisation_nov-dec12_developing-resolution.pdf.

Van Tol JJ, Le Roux PAL, Hensley M (2010) Soil indicators of hillslope hydrology in Bedford catchment. S. Afr. J. Plant Soil 27 (3): 242–251.

Van Tol JJ, Le Roux, PAL, Lorentz SA Hensley M (2013) Hydropedological classification of South African hillslopes. Vadose Zone J. 12 (4). DOI:10.2136/vzj2013.01.0007.

Van Tol JJ, van Zijl, GM, Riddell ES (2015) Application of hydropedological insights in hydrological modelling of the Stevenson Hamilton Research Supersite, Kruger National Park, South Africa. Water SA 41 (4): 525-533.

Van Zijl GM, Le Roux PAL, Smith HJC (2012) Rapid soil mapping under restrictive conditions in Tete, Mozambique. In: Minasny B, Malone BP, McBratney AB (eds.) Digital Soil Assessments and Beyond. CRC Press, Balkema. 335–339.

Van Zijl, GM, Le Roux PAL (2014) Creating a Conceptual Hydrological Soil Response Map for the Stevenson Hamilton Research Supersite, Kruger National Park. Water SA. 40: 331–336.

Venter FJ (1990) A classification of land management planning in the Kruger National Park. PhD thesis, Department of Geography, University of South Africa.

Zhu A-X (1997) A similarity model for representing soil spatial information. Geoderma 77: 217–242.

Chapter 11
Some Challenges on Quantifying Soil Property Predictions Uncertainty for the GlobalSoilMap Using Legacy Data

Zamir Libohova, Nathan P. Odgers, Jenette Ashtekar, Phillip R. Owens, James A. Thompson and Jon Hempel

Abstract The *GlobalSoilMap* project aims to create digital soil property maps in a raster format for six standard depths (0–5; 5–15; 15–30; 30–60; 60–100; 100–200 cm) and, for the first time, with estimates of uncertainty for predicted soil property maps. Data-driven methods and expert knowledge methods have been proposed, both of which present unique challenges. Initially, the majority of the predicted soil property maps will be derived from legacy soil data. The quantification of uncertainty, in particular, presents challenges due to the inherent nature of legacy data coming from different vintages (varying scales, formats, degree of completeness, differences in methods of observations, measurements, and classifications). We discuss the merits of each approach and potential practical and temporary solutions using two case studies from the USA, North America, and Llanos Orientales, Columbia, South America. Both case studies have limited data with insufficient point observations for a meaningful statistical approach for the estimation of prediction interval (PI) uncertainty. For the US case study, the available point measurements

Z. Libohova (✉) · J. Hempel
National Soil Survey Center, U.S. Department of Agriculture, Natural Resources Conservation Service, Washington, D.C, USA
e-mail: zamir.libohova@lin.usda.gov

N.P. Odgers
Department of Environmental Sciences, Faculty of Agriculture and Environment, The University of Sydney, Sydney, Australia
e-mail: nathan.odgers@sydney.edu.au

J. Ashtekar · P.R. Owens
Department of Agronomy, Purdue University, West Lafayette, USA
e-mail: goodman2@purdue.edu

P.R. Owens
e-mail: prowens@purdue.edu

J.A. Thompson
Division of Plant and Soil Sciences, West Virginia University, Morgantown, USA
e-mail: james.thompson@mail.wvu.edu

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_11

are not adequate for PI uncertainty quantification at soil map unit level and furthermore have been purposively collected to support the assignment of estimated mean, upper and lower property values to soil map units. We compared the estimated soil map unit upper and lower limits and 90 % CI from measured pedon for soil pH and found no significant differences between the two. The results suggest that the estimated upper and lower values from soil map units can be used for estimating the 90 % PI uncertainty at least initially until other independent measured point data become available. The available points in Llanos Orientales were collected for soil fertility evaluations and were independent of soil map unit polygons. However, they were surficial samples, clustered, and biased toward cultivated fields. As a result, only the 90 % CI was calculated and was found to be as wide as the range of the mean predicted soil property. These examples highlight few challenges in quantifying the 90 % PI and the need for more measured point data and flexible approaches when dealing with uncertainty quantification.

Keywords GlobalSoilMap · Digital soil mapping · Soil legacy data · Soil property maps · Uncertainty

11.1 Introduction

Quantifying the uncertainty associated with predicted soil property maps based on the GlobalSoilMap (GSM) specifications (GlobalSoilMap Science Committee 2013) presents numerous challenges (Odgers et al. 2012). These challenges relate mostly to the lack of sufficient point measured data that can be used for the calculation of the 90 % prediction interval (PI) around the mean. Thus, existing legacy data need to be used with Digital Soil Mapping (DSM) approaches (Lagacherie and McBratney 2007) that are tailored to the kind and quality of legacy data as discussed by Minasny and McBratney (2010). These approaches need to be flexible and combine data-driven approaches (Malone et al. 2011a, b) with expert knowledge methods (Lilburne et al. 2009).

The objective of this study is to highlight few challenges in quantifying soil property predictions uncertainty from two different scenarios that are commonly associated with the soil legacy data. The United States scenario deals with Soil Survey Geographic (SSURGO) database that provides estimated upper and lower limits to represent the typical range in the predicted soil property distributions of each soil series or soil map unit (USDA-NRCS 2013). The estimated upper and lower limits have been derived from a combination of laboratory-measured values and expert knowledge resulting in predicted values that are not necessarily independent of each other. In practical terms, this means that experts (i.e., soil mappers) derived these predictions by combining measured data with other field observations

to predict mean and ranges for soil polygon and soil type-related properties. The major hypothesis is that the estimated upper and lower limits are derived from the distribution of the measured point pedon data. If true, they can be used as 90 % PI of the SSURGO predicted soil property maps.

The Llanos Orientales scenario represents a case when soil map units do not have any estimated or measured values for the soil properties. However, there are laboratory-measured values that are most likely independent of the soil map units, meaning that were not used for attributing soil polygon maps with soil properties, but that are surficial samples collected for soil fertility testing. These samples are clustered mostly in agricultural fields and do not cover the entire area for which soil property maps are needed. Also the number of points is not sufficient for predictions at map unit level. The major objective for this case study was to use measured data from soil fertility campaign in combination with expert knowledge, fuzzy logic mapping (Zhu et al. 2001) combined with terrain analysis (Jasiewicz and Stepinski 2013), and homosoil approach (Mallavan et al. 2010) to generate predicted soil property maps and 90 % CI.

11.2 Materials and Methods

11.2.1 US Case Study

The analysis is based on soil pH values for soil series with mapping extend greater than 400,000 ha in order to assure an adequate sample size of sampled point data from USDA-NRCS National Cooperative Soil Survey Soil Characterization Database (NCSS-SCDB) (Libohova et al. 2013). The soil series were selected from the National Soil Information System (NASIS) database. We attempted to select soil series with more than 30–35 measured pedons data based on the central limit theorem assumption that the distribution of sampled measured values from a population tends to be close to that of the population for a sample size greater than 30–35 (Ott and Longnecker 2001). For the purposes of this comparison, a straightforward match at the soil horizon level of the soil pH between measured point pedon data and estimated values in SSURGO was difficult due to different naming of these horizons in NCSS-SCDB and NASIS. As a result, the comparisons were conducted only for the surface horizon in order to assure a satisfactory match between measured data from the NCSS-SCDB and estimated values from SSURGO/NASIS. Simple linear regression analysis was used to assess the overall significance of the relationship between the estimated upper and lower limits and the calculated 90 % CI from pedons to determine whether they were closely associated to support their use as 90 % PI.

11.2.2 Colombia Case Study

The Llanos Orientales occupies approximately 238,181 km^2 of northeastern Colombia. The area has a general soil map (1:250,000) and a semi-detailed soil map (1:25,000–1:50,000), which cover only 35 % of the study area. The existing soil, geology, and vegetation maps were used to understand soil–landscape relationships in order to be applied for the unmapped areas. A total of 64 sampled point data analyzed for soil organic carbon (SOC) were obtained from the Instituto Geografico Agustin Codazzi (IGAC) (http://www.igac.gov.co/igac) for 0–10 and 10–20 cm depths.

A pattern recognition landform classification (Jasiewicz and Stepinski 2013) based on the 90 m Shuttle Radar Topography Mission (SRTM) (Jarvis et at. 2008) digital elevation model was used to generate 10 geomorphons that were aggregated in five landscape positions (flat, ridges, slope, valley, and footslope). A zonal statistics for normalized height, topographic wetness index (TWI), and slope gradient were conducted for the aggregated landscape positions and the existing semi-detailed soil map units. Further, the landform classification offered more detail compared to the existing semi-detailed soil map; thus, only zonal statistics for aggregated landscape positions was used. However, the relationships between existing semi-detailed soil map and geology and vegetation were useful in predicting initial coarse scale soil map units for the unmapped areas (Ashtekar et al. 2013). The mean and standard deviation of each terrain attribute for the five aggregated landscape positions were used as rules to generate a fuzzy soil class map (Zhu et al. 2001). The values of SOC were assigned to each soil class (aggregated landscape positions) based on the mean from the measured sampled points that fell within each soil class. The derived mean values were multiplied with the fuzzy membership values of each soil class to make a continuous predicted SOC map. The 90 % CI for predicted SOC map was calculated by combining all measured sampled point data.

11.3 Results and Discussion

11.3.1 US Case Study

The mean measured pedon soil pH and estimated SSURGO mean were significantly correlated (Adjusted $R^2 = 0.4$; RMSE = 0.4; p-value < 0.01). More importantly, the lower boundary of 90 % CI from pedons and the SSURGO estimated lower limit was significantly correlated (Adjusted $R^2 = 0.86$; RMSE = 0.27; p-value < 0.001) as was the upper boundary of the 90 % CI from pedons and SSURGO estimated upper limit (Adjusted $R^2 = 0.67$; RMSE = 0.21; p-value < 0.001). As previously stated, these comparisons were not based on a straightforward relationship between the horizons for measured pedon data and the

estimated lower and upper limits. There were discrepancies between horizon nomenclatures due to various database transactions that have taken place over time (USDA-NRCS 2013). Most notably is the fact that the SSURGO database has aggregated various horizons in layers named H1–H3 based on their interpreted behavior but has also maintained the genetic horizon nomenclature for other soil series. This mixed naming convention has made it almost impossible to establish the straight links at matching genetic horizons between measured pedon data and SSURGO attribute data which is not uncommon when dealing with legacy data. In addition, the measured soil data were collected to support the soil mapping activities; thus, the sampling schemes were not necessarily designed for statistical purposes. As such the interpretation of statistical analysis needs to be done cautiously and within the context of legacy soil data. Indeed, Minasny and McBratney (2010) argue that the kind and quality of the legacy soil data determines the kind of DSM approaches which requires flexibility especially when legacy soil data are the only data available to support such mapping. However, the fact that the 90 % CI from measured pedon data and estimated upper and lower limits of predicted soil properties from SSURGO are highly correlated would justify the use of estimated SSURGO upper and lower limits as the 90 % PI at least initially until more measured point data becomes available for a truly statistical quantification of the 90 % PI according to GlobalSoilMap standards.

11.3.2 Colombia Case Study

The detailed soil map generated based on the geomorphons methodology (Jasiewicz and Stepinski 2013) showed much greater detail compared with the semi-detailed soil map (1:25,000–1:50,000) (Fig. 11.1). More importantly, the soil map generated from geomorphons extended to the entire Llanos Region, providing a better soil map resolution compared to the existing general soil map (1:250,000). The previously generated geology and geomorphology maps (Goosen 1971; Atehortúa et al. 2010) were overall in agreement with the new soil map but did not provide the level of detail needed for soil property maps at finer resolution. Using the homosoil approach (Mallavan et al. 2010), however, allowed for a detailed soil map based on the information provided from the geology, vegetation, and especially geomorphology maps in combination with the semi-detailed soil map.

The continuous SOC map generated from the fuzzy geomorphons soil map that was based on fuzzy logic approach (Zhu et al. 2001) highlights some differences associated with geomorphology and soil landscape position (Figs. 11.2a and 11.3a). The Plains overall had more SOC compared to the Dissected High Plains, mostly due to historical and management differences. The Dissected High Plains have experienced severe erosion since the last tectonic uplift that lead to the establishment of Meta River that divided the Llanos Region into two distinct geomorphic

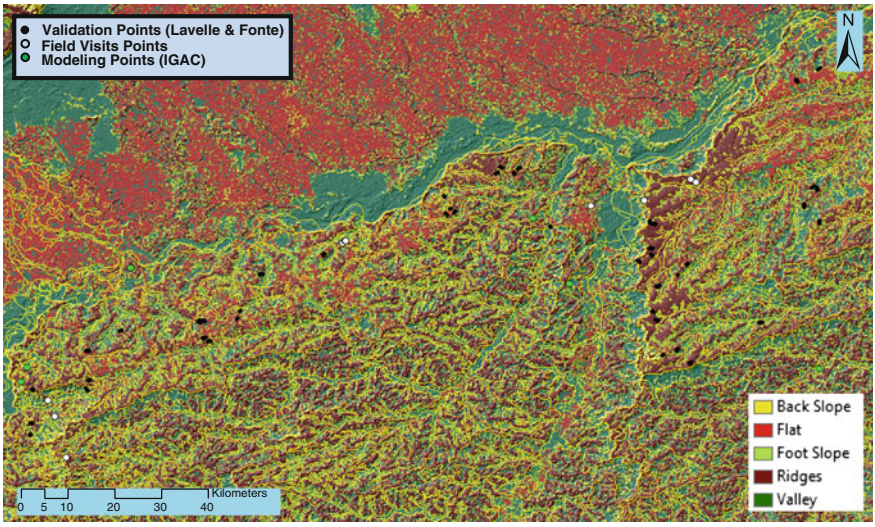

Fig. 11.1 Map of the major soil landscape positions based on geomorphons for a portion of Llanos Region. The delineations in *yellow* represent the existing semi-detailed soil map units. The *color-coded dots* represent observation points that were used to make property maps (modeling points) and validation [validation points—for more on validation, see Ashtekar et al. (2013)]

units (Goosen 1971). The Plains, on the other hand, have received continued fresh deposits from the uplift of Cordilleras Mountains northwest of the Llanos Region. Also the dominant vegetation of the Plains has been the savannah type which has favored the accumulation of the SOC compared to the Dissected High Plains where the opposite has occurred due to erosion caused by age of the landscape and anthropogenic factors (Goosen 1971).

The 90 % CI for the predicted SOC map were found to be as great as the variability in the predicted SOC itself (Figs. 11.2b and 11.3b). This is to be expected given the bias associated with the point observations that were used to generate the map. Most of the samples were taken near roads in agricultural fields displaying a degree of clustering (Ashtekar et al. 2013). As a result, the observation points failed to capture the large degree of variability present in the Llanos Region. Also, because of the insufficient sample points the 90 % CI instead of the 90 % PI was calculated, which does not meet the GlobalSoilMap standards for quantifying the uncertainty of predicted soil properties. The map is far from perfect; however, it provides an overall picture of the surface SOC distribution for the entire region. In addition, it highlights areas that may need more sampling thus serving as a platform for designing unbiased sampling schemes that would allow for a better representation of the variability of the entire region and quantification of uncertainty of predicted soil property maps.

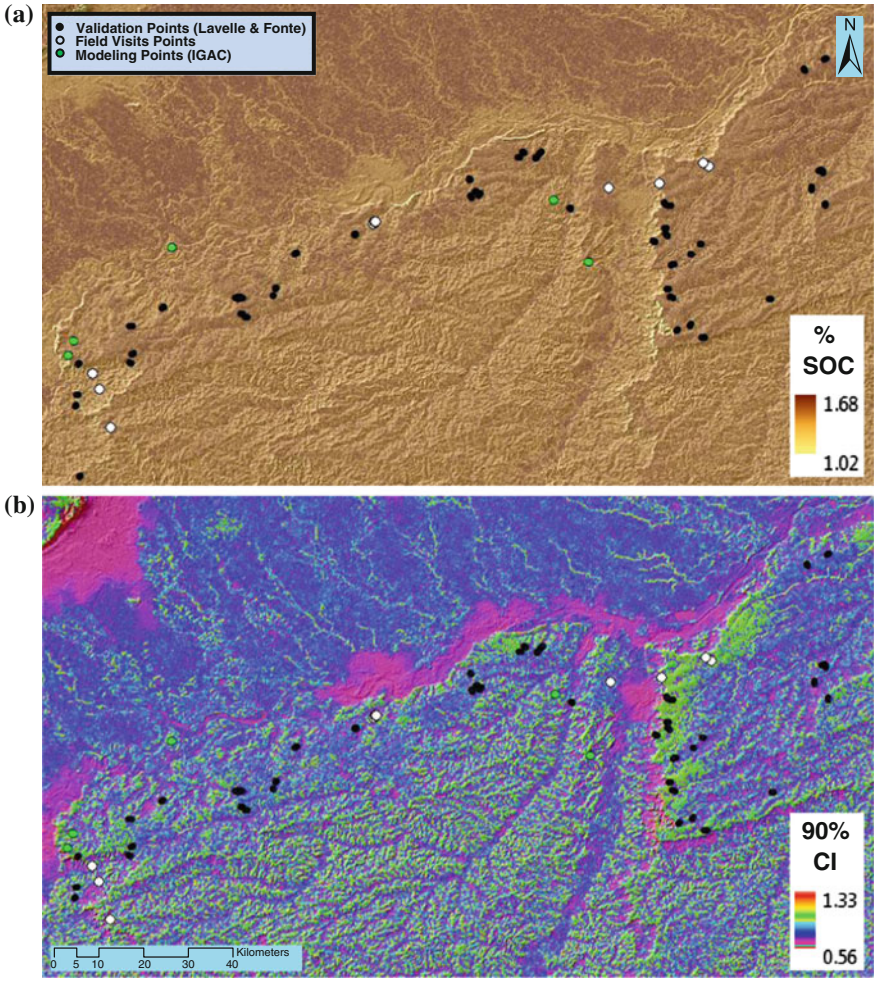

Fig. 11.2 **a** Continuous soil organic carbon (% SOC) map and **b** the 90 % confidence interval SOC map for a portion of Llanos Region for the 0–10 cm soil layer

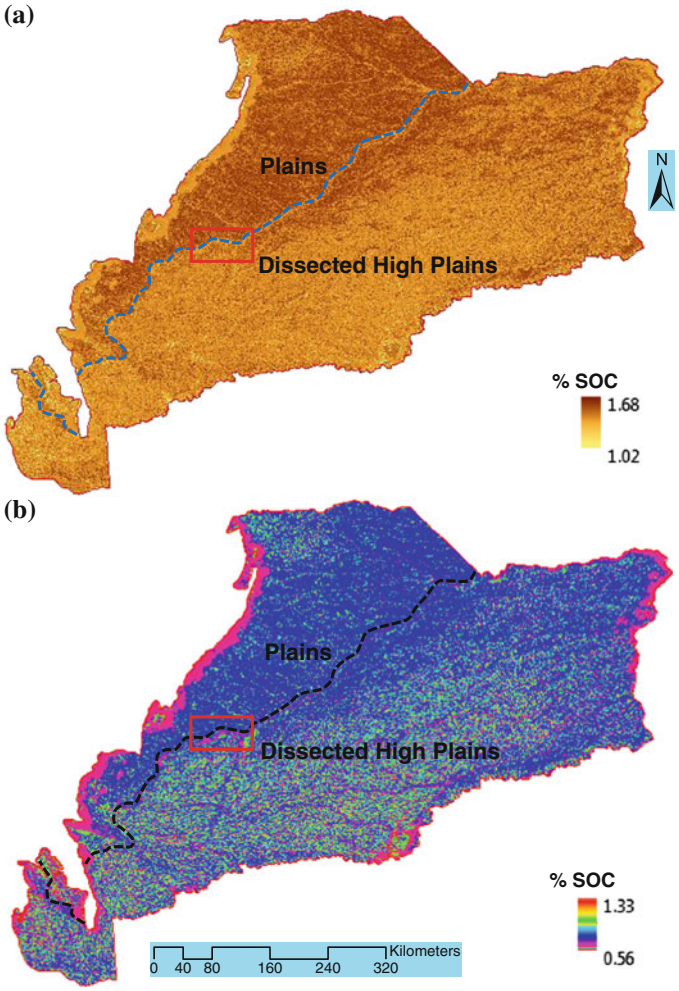

Fig. 11.3 SOC map 90 % CI for the 0–10 cm layer for the Llanos Region. The *dashed line* shows the current location of Meta River that was created as a result of tectonic uplift dividing the Llanos region into two distinct geomorphic units. The area in *red* represents the selected portion of the region where most of the point observations used to generate the SOC maps were located (Fig. 11.2a, b)

11.4 Conclusions

We have identified some challenges surrounding the development of maps of uncertainty predictions for *GlobalSoilMap* products based on legacy soil data collected in the USA and Llanos Orientales Region in Colombia. It is likely that some of the issues we identified here may be applicable to other jurisdictions also,

and that other jurisdictions will experience challenges unique to their own legacy soil data as well. Given the limitations associated with using legacy soil data, there may be a need to be pragmatic and make some assumptions as the data are used for tasks that were unforeseen at the time of their collection. As a result, uncertainty predictions about the reported soil property values will likely be high, but it will be quantified spatially exhaustively for the first time.

References

Ashtekar, J.M., P.R. Owens, R.A. Brown, H.E. Winzeler, M. Dorantes, Z. Libohova, M. Dasilva, A. Castro., 2013. Digital mapping of soil properties and associated uncertainties in the Llanos Orientales, South America. GlobalSoilMap Conference Proceeding, 2013. Orleans, France. Taylor & Francis group, London, UK.

Atehortúa, M., Sanabria, Y., Brito, J., and Rodrigues, S., 2010. LA GEOLOGÍA, GEOMORFOLOGÍA, PEDOLOGÍA Y USO DE LA TIERRA EN LAS MUNICIPALIDADES DE PUERTO LÓPEZ (COLOMBIA) Y UBERLÂNDIA (. Sci-ELO Brasil 22, 329–345.

GlobalSoilMap Science Committee. 2013. Specifications: Tiered GlobalSoilMap.net Products, Release 2.3.

Goosen, D., 1971. Physiography and soils of the Llanos Orientales, Colombia.

Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara, 2008, Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org).

Jasiewicz, J. and T.F Stepinski (2013) Geomorphons -a pattern recognition approach to classification and mapping of lanforms. Geomorphology 182, pp. 147-156.

Lagacherie, P., and A.B. McBratney (2007) Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping, 3-22. In Digital Soil Mapping: an Introductory perspective. http://www.sciencedirect.com/science/

Lilburne, L., Hewitt, A.E., & Ferriss, S. 2009. Progress with the design of a soil uncertainty database, and associated tools for simulating spatial realisations of soil properties. M. Caetano & M. Painho (eds), 7th International Symposi-um on Spatial Accuracy Assessment in Natural Resources and Environmental Science. Lisbon, Portugal, pp. 510–519.

Libohova, Z., S. Wills, and N.P. Odgers, 2013. Legacy Data Quality and Uncertainty Estimation for United States GlobalSoilMap Products. GlobalSoilMap Conference Proceeding, Orleans, France. Edited by Dominique Arrouays, Neil McKenzie, Jon Hempel, Ann C. Richer de Forges, and Alex McBratney. CRC Press 2014, pages 63–68. Print ISBN: 978-1-138-00119-0, eBook ISBN: 978-1-315-77558-6, DOI: 10.1201/b16500-18.

Malone, B.P., McBratney, A.B., Minasny, B., 2011a. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160(3-4), 614-626.

Malone, B.P., de Gruijter, J.J., McBratney, A.B. Minasny, B., & Brus, D.J. 2011b. Using additional criteria for measuring the quality of predictions and their uncertainties in a digital soil mapping framework. Soil Science Society of America Journal 75:1032–1043.

Mallavan, B.B., Minasny, B., and McBratney, A.B., 2010. Homosoil, a Methodology for Quantitative Extrapolation of Soil Information Across the Globe. In J.L. Boettinger et al. (eds.), Digital Soil Mapping in Soil Science 2.

Minasny, B. & McBratney, A.B. 2010. Methodologies for global soil mapping. In J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink & S. Kienast-Brown (eds), Digital soil mapping: bridging research, environmental application, and operation: 429–436. Springer Science+Business Media.

Odgers, N.P., Libohova, Z., Thompson, J.A., 2012. Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale. Geoderma 189–190, 153–163. doi: 10.1016/j.geoderma.2012.05.026.

Ott, R.L., Longnecker, M., 2001. An introduction to statistical methods and data analysis (5th edtition).

USDA-NRCS. 2013. National Soil Survey Handbook. Available at http://soils.usda.gov/technical/handbook/ (accessed 27 May 2013). United States Department of Agriculture-Natural Resources Conservation Service.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil Mapping Using GIS, Expert Knowledge, and Fuzzy logic. Soil Science Society of America Journal, 65(5), 1463-1472.

Chapter 12
Spatial Assessment of Soil Organic Carbon Using Bayesian Maximum Entropy and Partial Least Square Regression Model

Bei Zhang and Sabine Grunwald

Abstract There has been great interest in the estimation of soil carbon over the last decade to address critical environmental, agronomic, and sociopolitical issues. Soil proximal sensing has shown much potential for soil carbon assessment. Visible/near-infrared diffuse reflectance spectroscopy (VNIRS) has been introduced as a complementary data source in digital soil mapping due to its cost effectiveness. However, in many studies, the uncertainty in soil modeling using VNIRS has not been explicitly taken into account. Bayesian maximum entropy (BME) is a modern geostatistical method that incorporates auxiliary/soft data within a theoretical sound framework. Our objective was to employ VNIR data and BME to spatially estimate soil organic carbon (SOC). Another objective was to compare the performance to estimate SOC using BME to classical geostatistical methods. A total of 1012 soil samples from Florida, USA, were employed from a database that included pairs of SOC measurements derived by dry combustion and hyperspectral data with 1-nm resolution in the VNIR spectral range (350–2500 nm). Partial least square regression (PLSR) was used to model the relationship between VNIR data and SOC. For spatial estimations of SOC, we employed BME using "hard" (SOC measurements from the laboratory) and interval "soft" data (predictions of VNIR–PLSR model). For the purpose of comparison, ordinary kriging (OK) was used with only the hard data set (OK1) and the SOC estimates derived from the VNIRS–PLSR model (OK2) at point locations. Both BME and OK2 show distinctly different pathways of assimilating vague ("soft") data into the spatial modeling process. The three spatial estimation methods (BME, OK1, and OK2) were examined using the independent validation set by calculating bias, root mean square error (RMSE), residual prediction deviation (RPD), and ratio of performance to inter-quartile distance (RPIQ). The preliminary

B. Zhang
College of Resources and Environment, Huazhong Agricultural University, 430070 Wuhan, China

B. Zhang · S. Grunwald (✉)
Soil and Water Science Department, University of Florida, 2181 McCarty Hall, PO Box 110290, Gainesville, FL 32611, USA
e-mail: sabgru@ufl.edu

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_12

results show that BME performed generally as well as OK1, which may be due to the data splitting effects. However, both BME and OK1 were better than OK2. As BME can take advantage of data from the PLSR model, it offers the possibility to reduce the amount of laboratory-measured samples to map across a region. OK2 performed worse than OK1, which showed that using vague data into kriging leads to higher uncertainties. In this case, data from the VNIRS model may not help to improve the performance of predictions in kriging. These results underpin the potential of the BME approach in digital soil mapping.

Keywords Bayesian maximum entropy · Digital soil mapping · Soil carbon · Visible/near-infrared reflectance spectroscopy

12.1 Introduction

Soil carbon is considered as the largest pool of carbon in terrestrial ecosystems (Lal 2004) with multiple environmental cobenefits including fertility, productivity, and soil health that influence many agronomical, environmental, and political issues (Lacoste et al. 2014). Mapping the spatial distribution of soil carbon at a variety of spatial and temporal scales has been of great interest to address needs (Grunwald 2009; Minasny et al. 2013). A variety of methods have been used in soil carbon mapping, such as regression kriging (Vasques et al. 2010), geographically weighted regression (Zhang et al. 2011), and random forest (Wiesmeier et al. 2010). Visible/near-infrared reflectance spectroscopy (VNIRS) has been established as an alternative to more costly laboratory measurements to characterize soil properties. It is rapid and nondestructive and requires less sample preparation with less or no chemical reagents (McCarty et al. 2002; Viscarra Rossel et al. 2006; Brown 2007; Vasques et al. 2008). Modeling the quantitative relationships between soil attributes and spectral characteristics requires sophisticated statistical techniques (Viscarra Rossel et al. 2006). A variety of regression methods have been used for modeling soil VNIRS, such as principal component regression (PCR), partial least squares regression (PLSR), multiple linear regression (MLR), and artificial neural network (ANN) (Mouazen et al. 2010; Rossel and Behrens 2010). Among those techniques, PLSR is the most widely used (Brown et al. 2005; Vasques et al. 2008; Volkan Bilgili et al. 2010). The algorithm of PLSR is computationally faster than other methods; models are more interpretable and are relatively insensitive to over-fitting (Brodský et al. 2013). However, Brodský et al. (2013) found that PLSR modeling can cause uncertainty in the map of spatial prediction. More importantly, the uncertainty from spatial estimation by kriging can be substantial. Consequently, using VNIRS data directly in the kriging process may be not a good choice. The geostatistical methods that can incorporate auxiliary variables, such as regression kriging (RK) using VNIRS data to estimate soil properties (Ge et al. 2007), might be an alternative approach. However, if the relationship between auxiliary variable and target variable is not constant in all parts of the study area, the predictions might be even worse than just using plain kriging (Hengl

et al. 2007). It is critical to note that RK and PLSR fail to incorporate the prediction uncertainty explicitly into the modeling process. To explicitly incorporate soil spectral data into the modeling process to predict soil properties has been underexplored, but will be addressed in this study.

Bayesian maximum entropy (BME) proposed by Christakos (1990, 2000) is a modern geostatistical approach, which can integrate data with uncertainty into the modeling process, aiming to improve predictive capabilities compared with traditional estimation methods. In this framework, the term "hard data" refers to the most precise and accurate data with current instrumentation (e.g., soil analytical laboratory measurements), while "soft data" may represent varying levels of uncertain observations related to the target variables. The latter may be estimates of soil carbon derived from spectral data. Intervals of values or probability density functions are two ways to represent soft data. BME has been successfully applied in soil science (Bogaert and D'Or 2002; Douaik et al. 2004, 2005), environmental risk assessment (Lee 2005; Yu et al. 2009; Bogaert et al. 2009), environmental health (Puangthongthub et al. 2007; Money et al. 2009; Lee et al. 2009; Pang et al. 2010), and climate research (Lee et al. 2008).

Different kinds of soft data have been used in soil science, such as legacy soil map and raw measurement data (Bogaert and D'Or 2002; Douaik et al. 2004). There are no studies yet that have incorporated soil VNIRS data into the BME framework to improve predictions of soil properties.

The aims of this research were to: (i) investigate the performance of BME spatial estimation for SOC combined with VNIRS data, (ii) assess the performance of BME with soft data derived from VNIRS–PLSR models, and (iii) compare the accuracy of BME spatial estimation with traditional ordinary kriging as a reference.

12.2 Materials and Methods

12.2.1 Study Area

The study area is the State of Florida located in the southeastern Coastal Plain, USA, extending over six and one-half degrees of latitude (24.55–31.00 N, 80.03–87.63 W). The prevalent climate in Florida is humid subtropical, while the southern part has a tropical climate (add reference). The majority of soils in Florida are Spodosols (32 %), Entisols (22 %), Ultisols (19 %), Alfisols (13 %), and Histosols (11 %) (Natural Resources Conservation Service 2006). Land use and land cover are composed of wetlands (28 %), pinelands (18 %), and urban and barren lands (15 %), while agriculture, rangelands, and improved pasture occupy 9 %, 9 %, and 8 % of this state, respectively (Florida Fish and Wildlife Conservation Commission 2003). Florida is characterized by relatively flat topography with gentle slope varying from 0 to 5 % in most parts of the region. Only less than 1 % of the state has moderate slopes of 5–19 % (US Geological Survey 1999).

12.2.2 Data Preparation

We used soil data from a previous project "Rapid Assessment of Trajectory Modeling of Changes in Soil Carbon across a Southeastern Landscape" (courtesy of the soil database maintained by Dr. Grunwald's Pedometrics, Landscape Analysis, and GIS Laboratory). A detailed description of sampling design, laboratory analysis, and spectral scanning were provided by Xiong et al. (2014). Briefly, a total of 1012 soil samples were collected from March 2008 to August 2009 using a random-stratified sampling design based on land use—soil order strata. At each site, four 20 × 5.8 cm soil cores were collected within a 2-m-diameter area. These four soil samples were bulked in the field and then placed in a cooler until they could be transported to laboratory. SOC was analyzed in the laboratory using dry combustion (Shimadzu TOC-V/SSM-5000). For preliminary analysis, SOC values higher than the 75 % quantile of the whole data were removed. This pretreatment reduced the number of sample to 759.

Spectral data were derived from scanning of soil samples in the laboratory in the VNIR spectral range (350–2500 nm) at 1-nm intervals. Each sample was scanned four times. The average of these four scans was computed for every single sample. Three preprocessing methods were applied to the spectra data. First, the reflectance curves were smoothed across a moving window of 9 nm by using the Savitzky–Golay algorithm with a third-order polynomial. Then, to reduce the complexity of the data, the averages of reflectance values were taken across a 10-nm window. Last, the second-order derivate was applied with a 4 polynomials and 7-nm window. Those 3 steps reduced each of the spectral curves to 215 values.

12.2.3 Data Analyses

The BME approach was employed that provides a systematic and rigorous way to incorporate soft data in addition to hard data into the modeling process. According to Christakos (1990), BME balances two requirements: high prior information about the spatial variability and high posterior probability about the estimated map. The first requirement uses a variety of sources of prior information and involves the maximization of an entropy function. The second requirement leads to the maximization of a so-called Bayes' function.

To implement BME and kriging, all soil samples were randomly divided into four groups. The first one (model set) included laboratory-measured SOC and scanned spectral data and was used for establishing the VNIRS model using PLSR. By using the VNIRS–PLSR model, predictions of SOC were derived with spectral data from the second group (soft data set) to acquire prediction values and deviations. With the prediction outcomes of the VNIRS–PLSR model, each individual soft interval can be obtained. Specifically, for each sample in the soft data set, the upper limit of the interval equaled the SOC prediction value plus one deviation, and

the lower limit was set to the SOC predication value minus deviation. The last two groups were hard data and independent validation data, respectively. Although both of them were laboratory-measured SOC, the hard data set was only used for calibration and the other set was used to validate the performance of geostatistical estimations. Soft interval data and hard data were both integrated into the BME estimation process. The model set, soft data set, hard data set, and validation set included 190, 380, 114, and 75 samples, respectively.

For spatial estimations of SOC, we employed BME using "hard" (SOC measurements from the laboratory) and "soft" data (VNIR data). The BME analysis included three main stages (Christakos 1990; Douaik et al. 2005):

Prior stage: with the goal to maximize the information content using generalized knowledge which was implemented using the model set (i.e., pairs of laboratory-measured SOC and VNIR data) and PLSR to estimate the prior probability density function (PDF).
Meta-prior stage: By using the VNIRS–PLSR model, predictions of SOC were derived with spectral data from the second group (i.e., the soft data set) to acquire prediction values and deviations. With the prediction outcomes of the VNIRS–PLSR model, each individual soft interval can be obtained. Specifically, for each sample in the soft data set, the upper limit of the interval equaled the SOC prediction value plus deviation and the lower limit was set to the SOC predication value minus deviation.
Posterior stage: aiming to maximize the posterior PDF through updating of the prior PDF by taking into account the hard data set. The posterior and the prior PDFs are related through the conditional probability law based on Bayes' theorem.

For the purpose of comparison, ordinary kriging (OK) was used with only the hard data set (OK1). Then, OK was also employed using SOC estimates derived from the VNIRS–PLSR model (OK2). Both BME and OK2 show distinctly different pathways of assimilating vague data into the spatial modeling process. The three spatial estimation methods (BME, OK1, and OK2) were examined using the independent validation set by calculating bias, root mean square error (RMSE), residual prediction deviation (RPD), and ratio of performance to inter-quartile distance (RPIQ).

12.3 Results and Discussion

The descriptive statistics of SOC are reported in Table 12.1. Soil organic carbon (SOC) in Florida was highly variable with range values for all four sets with more than 17 g kg^{-1}. The means of all the four data sets were similar, except the validation set with a mean of 11.2 g kg^{-1} and median of 10.2 g kg^{-1}. This indicates a slight bias in the validation data set toward high SOC values which may have impacted the validation evaluation process. Moreover, the minimum value of the validation set was 3.9 g kg^{-1} that was substantially higher than the values in the soft

Table 12.1 Descriptive statistics of soil organic carbon in different data sets

	N	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis
		g kg^{-1}							
Whole set	759	10.3	4.4	9.4	1.3	21.3	20.0	0.6	−0.5
Model set	190	9.9	4.5	8.7	3.0	21.1	18.1	0.7	−0.6
Soft set	380	10.2	4.2	9.5	1.3	20.9	19.6	0.5	−0.5
Hard set	114	10.3	4.7	9.5	1.9	20.9	19.1	0.6	−0.6
Validation set	75	11.2	4.6	10.2	3.9	21.3	17.4	0.6	−0.7

N Number of observations in each set, *SD* Standard deviation

and hard sets. This may be another reason that impacts the result of validation especially when there are estimated values less than 3.9 g kg^{-1} derived from BME and kriging. As SOC in all the data set was positively skewed, the PLSR model was built using natural logarithm-transformed SOC values.

The PLSR model built from the model dataset performed fairly well with R^2 of 0.52 and RMSE of 0.32 g kg^{-1} (Fig. 12.1). Figure 12.2 shows the predicted SOC values using the VNIRS–PLSR model and the spectral data from the soft set. This shows a good model fit. In both models (Figs. 12.1 and 12.2), residuals in the high and low SOC range were found indicating the uncertainty arising from models.

The spatial covariance derived from the hard set and soft interval data was modeled by nesting two exponential models. The sills for these two models were 13.5 and 2, and the ranges were about 1000 m and 30,000 m, respectively. The ranges of these models showed that the spatial correlation of SOC in Florida was relatively large (Fig. 12.3).

Figures 12.4, 12.5, and 12.6 show the results of spatial estimation of BME, OK1, and OK2 in Florida. The basic patterns of these three maps are quite similar, with the high values of SOC mainly located in the southeast corner of Florida consisting of highly organic soils. The range in SOC was narrower for BME than kriging, with

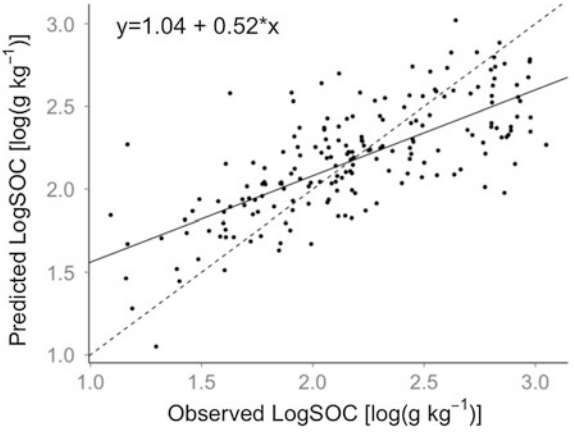

Fig. 12.1 Predicted and observed log-transformed soil organic carbon (log SOC) values derived from partial least square regression (PLSR) using the model set

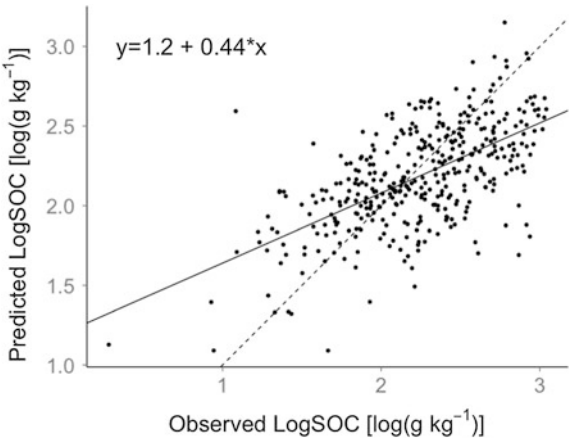

Fig. 12.2 Predicted and observed log-transformed soil organic carbon (log SOC) values derived from VNIRS–PLSR model using the soft data set

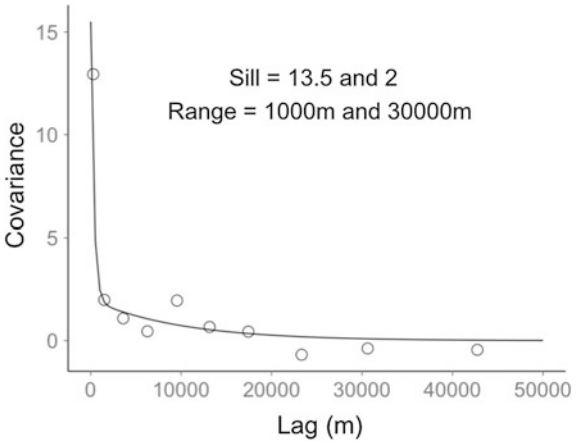

Fig. 12.3 The covariance structure of soil organic carbon (g kg^{-1}) and nested models

the maximum value only 15.4 g kg^{-1} that only covered about 75 % of the original range. This indicated that the method of BME was not sensitive to model high values. Another possible explanation may be the uncertainty associated with SOC of the PLSR model. As the goodness of fit for the model was just acceptable, the prediction SOC values may enlarge the uncertainty in BME. In contrast, the estimated SOC range by OK1 and OK2 corresponded well to the actual range in measured SOC within the State of Florida. The validation results indicate OK1 performance as well as BME (Table 12.2). The RMSE, RPD, and RPIQ of these two evaluations were almost the same. However, the SOC estimation map of BME was capable of showing more variations, whereas OK tended to smooth out SOC variation. The outcomes of OK2 were the worst, which indicate that assimilating vague data directly into kriging was not a good choice.

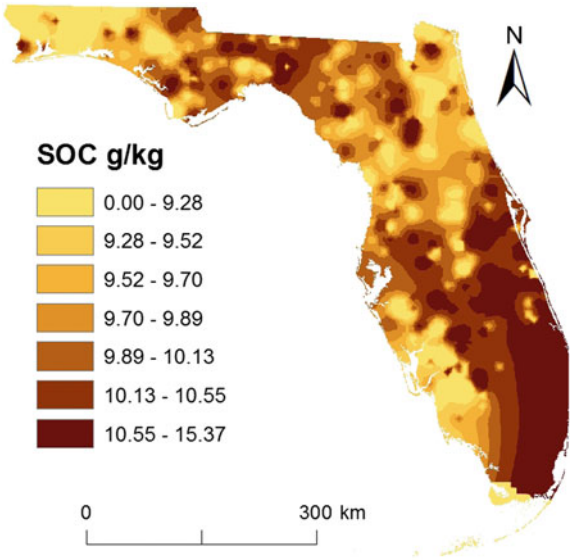

Fig. 12.4 Estimations of soil organic carbon (g kg^{-1}) using Bayesian maximum entropy

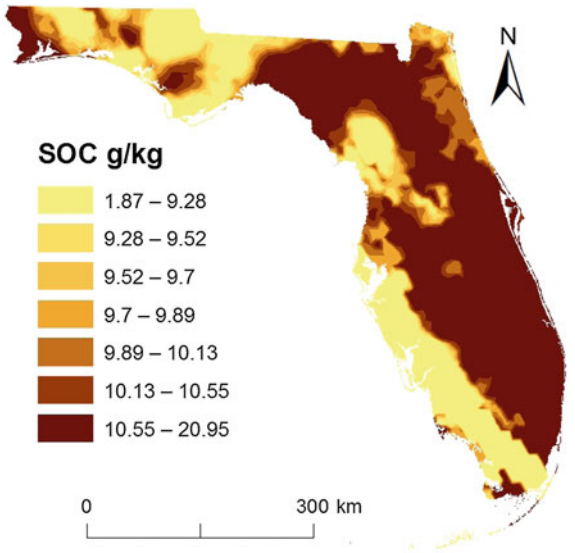

Fig. 12.5 Estimations of soil organic carbon in g kg^{-1} using ordinary kriging and the hard data set (OK1)

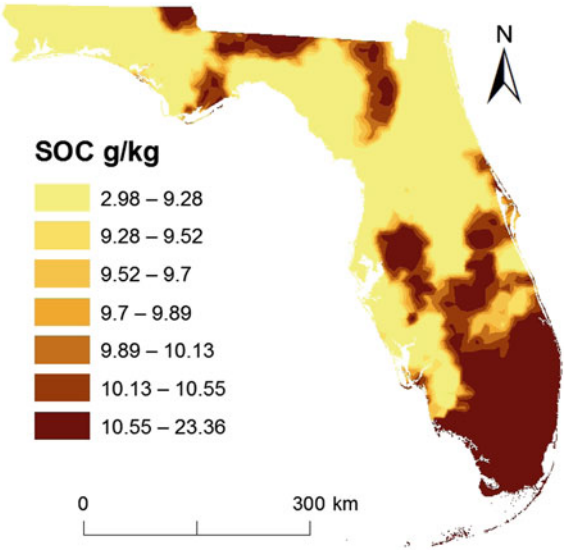

Fig. 12.6 Estimations of soil organic carbon in g kg^{-1} using ordinary kriging and the prediction from visible/near-infrared spectral (VNIRS)–partial least square regression (PLSR) model (OK2)

Table 12.2 Validation results for soil organic carbon (g kg^{-1}) derived from Bayesian maximum entropy, ordinary kriging with hard data (OK1), and ordinary kriging with predictions from partial least square regression model (OK2)

	RMSE[a]	Bias	RPD[a]	RPIQ[a]
BME	4.50	−1.27	1.02	0.63
OK1	4.45	−0.55	1.04	0.63
OK2	4.66	−1.86	0.99	0.61

[a]*RMSE* Root mean square error, *RPD* Residual prediction deviation, *RPIQ* Ratio of performance to inter-quartile distance

12.4 Conclusions

Theoretically, BME is expected to perform better than traditional univariate geostatistics (OK) because BME incorporates both—hard and soft data—into the modeling process. This was not found in the current preliminary study and needs further investigation probing into the causes. Preliminary findings suggest that the two methods, BME and kriging, performed almost the same using hard data. However, the spatial estimates of BME showed more details of SOC heterogeneity than OK1 and OK2. The VNIR spectral data used as soft data inputs in the BME modeling process possibly enhanced the capability to model SOC variability across Florida. The relatively small validation data set could not identify significant differences in performance between BME and OK in this study. Since the BME modeling process was influenced by many factors, such as preprocessing of VNIRS, the quality of the PLSR model, and parameter set during BME computing,

the accuracy of BME is expected to be improved by adjusting those factors. VNIR spectral data are easy to obtain and are poised to provide "vague" secondary data input to enhance the scarcity of hard (laboratory)-measured SOC data.

Acknowledgements We thank the following individuals for laboratory and field work to create the data sets used in this study: E. Azuaje, P. Chaikaew, A. Comerford, X. Dong, S. Moustafa, D. B. Myers, C.M. Ross, D. Sarkhot, L. Stanley, A. Stoppe, A. Quidez, and X. Xiong. We also thank co-PIs (Dr. N.B. Comerford and Dr. W.G. Harris) of the USDA-NIFA funded project "Rapid Assessment of Trajectory Modeling of Changes in Soil Carbon across a Southeastern Landscape" USDA-CSREES-NRI grant award 2007-35107-18368. And we also thank the China scholarship council and the Pedometrics, Landscape Analysis, and GIS Laboratory, Soil and Water Science Department, University of Florida, for hosting B. Zhang for a 2-year research visit.

References

Bogaert P, Christakos G, Jerrett M, Yu HL (2009) Spatiotemporal modelling of ozone distribution in the State of California. Atmospheric Environment 43:2471–2480. doi: 10.1016/j.atmosenv.2009.01.049

Bogaert P, D'Or D (2002) Estimating Soil Properties from Thematic Soil Maps. Soil Science Society of America Journal 66:1492–1500. doi: 10.2136/sssaj2002.1492

Brodský L, Vašát R, Klement A, Zádorová T, Jakšík O (2013) Uncertainty propagation in VNIR reflectance spectroscopy soil organic carbon mapping. Geoderma 199:54–63

Brown DJ (2007) Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma 140:444–453. doi: 10.1016/j.geoderma.2007.04.021

Brown DJ, Bricklemyer RS, Miller PR (2005) Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma 129:251–267. doi: 10.1016/j.geoderma.2005.01.001

Christakos G (1990) A Bayesian/maximum-entropy view to the spatial estimation problem. Math Geol 22:763–777. doi: 10.1007/BF00890661

Christakos G (2000) Modern Spatiotemporal Geostatistics. Oxford, New York

Douaik A, Van Meirvenne M, Tóth T (2005) Soil salinity mapping using spatio-temporal kriging and Bayesian maximum entropy with interval soft data. Geoderma 128:234–248. doi: 10.1016/j.geoderma.2005.04.006

Douaik A, Van Meirvenne M, Tóth T, Serre M (2004) Space-time mapping of soil salinity using probabilistic bayesian maximum entropy. Stoch Environ Res Risk Assess 18:219–227. doi: 10.1007/s00477-004-0177-5

Florida Fish and Wildlife Conservation Commission (FFWCC). Florida vegetation and land cover data derived from Landsat ETM+ imagery [Internet]. Tallahassee, FL; 2003. Available from: http://myfwc.com/research/gis/data-maps/terrestrial/fl-vegetation-land-cover/

Ge Y, Thomasson JA, Morgan CL, Searcy SW (2007) VNIR diffuse reflectance spectroscopy for agricultural soil property determination based on regression-kriging. Transactions of the Asabe 50:1081–1092.

Grunwald S (2009) Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152:195–207. doi: 10.1016/j.geoderma.2009.06.003

Hengl T, Heuvelink GBM, Rossiter DG (2007) About regression-kriging: From equations to case studies. Computers and Geosciences 33:1301–1315. doi: 10.1016/j.cageo.2007.05.001

Lacoste M, Minasny B, McBratney A, et al. (2014) High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213:296–311. doi: 10.1016/j.geoderma.2013.07.002

Lal R (2004) Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. Science 304:1623–1627. doi: 10.1126/science.1097396

Lee S-J, Balling R, Gober P (2008) Bayesian Maximum Entropy Mapping and the Soft Data Problem in Urban Climate Research. Annals of the Association of American Geographers 98:309–322. doi: 10.1080/00045600701851184

Lee S-J, Yeatts KB, Serre ML (2009) A Bayesian Maximum Entropy approach to address the change of support problem in the spatial analysis of childhood asthma prevalence across North Carolina. Spatial and Spatio-temporal Epidemiology 1:49–60. doi: 10.1016/j.sste.2009.07.005

Lee SJ (2005) Models of soft data in geostatistics and their application in environmental and health mapping. Dissertation, University of North Carolina at Chapel Hill

McCarty GW, Reeves JB, Reeves VB, et al. (2002) Mid-Infrared and Near-Infrared Diffuse Reflectance Spectroscopy for Soil Carbon Measurement. Soil Science Society of America Journal 66:640–646. doi: 10.2136/sssaj2002.6400

Minasny B, McBratney AB, Malone BP, Wheeler I (2013) Chapter One – Digital Mapping of Soil Carbon. BS:AGRON 118:1–47. doi: 10.1016/B978-0-12-405942-9.00001-3

Money ES, Carter GP, Serre ML (2009) Modern Space/Time Geostatistics Using River Distances: Data Integration of Turbidity and E. coliMeasurements to Assess Fecal Contamination Along the Raritan River in New Jersey. Environ Sci Technol 43:3736–3742. doi: 10.1021/es803236j

Mouazen AM, Kuang B, De Baerdemaeker J, Ramon H (2010) Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. Geoderma 158:23–31. doi: 10.1016/j.geoderma.2010.03.001

Natural Resources Conservation Service (NRCS), U.S. Department of Agriculture. U.S. General Soil Map (STATSGO2) [Internet]. Lincoln, NE; 2006. Available from: http://soils.usda.gov/survey/geography/statsgo/

Pang W, Christakos G, Wang J-F (2010) Comparative spatiotemporal analysis of fine particulate matter pollution. Environmetrics 21:305–317. doi: 10.1002/env.1007

Puangthongthub S, Wangwongwatana S, Kamens RM, Serre ML (2007) Modeling the space/time distribution of particulate matter in Thailand and optimizing its monitoring network. Atmospheric Environment 41:7788–7805. doi: 10.1016/j.atmosenv.2007.06.051

Rossel RAV, Behrens T (2010) Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158:46–54. doi: 10.1016/j.geoderma.2009.12.025

United States Geological Survey (USGS). National Elevation Dataset (NED) [Internet]. Reston, VA; 1999. Available from: http://ned.usgs.gov/

Vasques GM, Grunwald S, Comerford NB, Sickman JO (2010) Regional modelling of soil carbon at multiple depths within a subtropical watershed. Geoderma 156:326–336. doi: 10.1016/j.geoderma.2010.03.002

Vasques GM, Grunwald S, Sickman JO (2008) Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma 146:14–25. doi: 10.1016/j.geoderma.2008.04.007

Viscarra Rossel RA, Walvoort DJJ, Janik LJ, Skjemstad JO (2006) Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131:59–75. doi: 10.1016/j.geoderma.2005.03.007

Volkan Bilgili A, van Es HM, Akbas F, et al. (2010) Visible-near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. Journal of Arid Environments 74:229–238. doi: 10.1016/j.jaridenv.2009.08.011

Wiesmeier M, Barthold F, Blank B, Kögel-Knabner I (2010) Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340:7–24. doi: 10.1007/s11104-010-0425-z

Xiong X, Grunwald S, Myers DB, et al. (2014) Holistic environmental soil-landscape modeling of soil organic carbon. Environmental Modelling & Software 57:202–215. doi: 10.1016/j.envsoft.2014.03.004

Yu HL, Chen JC, Christakos G (2009) BME estimation of residential exposure to ambient PM10 and ozone at multiple time scales. Environ Health Perspect 117:537–544. doi: 10.1289/ehp.0800089

Zhang C, Tang Y, Xu X, Kiely G (2011) Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland. Applied Geochemistry 26:1239–1248. doi: 10.1016/j.apgeochem.2011.04.014

Chapter 13
Estimation of the Actual and Attainable Terrestrial Carbon Budget

P. Chaikaew, S. Grunwald and X. Xiong

Abstract Organic carbon is a key component of the terrestrial system that affects the physical, chemical, and biological processes. Changes in both the soil and the terrestrial carbon storage occur due to the interactions of natural ecological processes and anthropogenic activities. Research gaps to quantify soil and terrestrial carbon still exist. To discern between the actual and attainable carbon pools is critical to identify suitable adaptation and management alternatives to optimize natural carbon capital in the context of regional imposed changes, such as land use and climate change. Our objectives were to: (i) assess the spatially explicit relationships between observed soil organic carbon (SOC) and environmental factors and (ii) assess actual ($TerrC_{actual}$) and attainable ($TerrC_{attain}$) terrestrial carbon capital considering below-ground (soil) and above-ground (biomass) carbon. We collected 234 soil samples in the topsoil (0–20 cm) in 2008 and 2009 across the Suwannee River Basin in Florida, USA, based on a random design stratified by land cover/land use and soil suborders. For above-ground carbon assessment, we derived data from the LANDFIRE project which provided a high-resolution map of year-2000 baseline estimates of biomass carbon. A comprehensive set of 172 soil-environmental and human covariates was assembled from multiple data sources to predict and validate observed SOC stocks and $TerrC_{actual}$ using Random Forest (RF). The STEP-AWBH conceptual model (with S: Soil, T: Topography, E: Ecology, P: Parent material, A: Atmosphere, W: Water, B: Biota, and H: Human factors) provided the conceptual modeling framework to model $TerrC_{attain}$ that was implemented using RF and simulated annealing in combination. In the simulation, the STEP factors were kept constant, but the AWBH factors were varied by ±10, ±20, and ±30 %. The combined factors which amount to the highest modeled

P. Chaikaew · S. Grunwald (✉) · X. Xiong
Soil and Water Science Department, University of Florida, 2181 McCarty Hall, PO Box 110290, Gainesville, FL 32611, USA
e-mail: sabgru@ufl.edu

Present Address:
P. Chaikaew
Department of Environmental Science, Chulalongkorn University, 254 Payathai Road, Wang Mai, Pathumwan, Bangkok 10330, Thailand

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_13

terrestrial carbon stocks were postulated to equal the attainable terrestrial carbon stocks. Results suggest that the $TerrC_{attain}$ stocks showed slightly larger amounts than the $TerrC_{actual}$ stocks across the basin. The $TerrC_{actual}$ was 190 Tg C and the maximum $TerrC_{attain}$ was 195 Tg C. Biotic, soil, parent material, topographic, and water-related factors played important roles in determining SOC storage, while human factors were only weak predictors. Although mean annual precipitation and monthly mean temperature in summer months were significant to explain both SOC and terrestrial carbon stocks, they showed moderate/minor influence on carbon storage. The land use/land cover variables were the strongest factors predicting soil and terrestrial carbon stocks. These findings suggest that land use adaptions have much potential to achieve $TerrC_{attain}$, specifically conversions from cropland to land use systems with larger net primary productivity. Bare soils, which represent marginal soils, also have potential to elevate carbon storage through management adaptions that do not compete with other land uses.

Keywords Actual carbon stocks · Attainable carbon stocks · Terrestrial carbon · Soil organic carbon · Random forest · Simulated annealing · STEP-AWBH

13.1 Introduction

Carbon sequestration has become an important policy option to mitigate the increasing atmospheric greenhouse gases (GHG) that pose threats to a warming global climate. The terrestrial biosphere can sequester significant amounts of anthropogenic carbon dioxide (CO_2) by the natural carbon uptake process through plant biomass and soils. However, how ecosystem factors and carbon dynamics in terrestrial systems interact with each other that determine critically important ecosystem services is not well understood yet.

Numerous digital soil mapping studies have modeled soil organic carbon (SOC) across large regions (Bélanger and Pinno 2008; Wang et al. 2011; Wu et al. 2009; Xiong et al. 2014), and terrestrial carbon has been assessed at global and continental scale by Lal (2008) and Dickson et al. (2014). Yet, these carbon assessments focus on actual conditions without providing clues of attainable or potential carbon that could be sequestered in a landscape. To identify those site-specific adaptations that demonstrate most promise to attain the largest amount of carbon storable in soil and biomass is of interest because they can guide land management, adaptation, mitigation, decision making, and policy implementations to achieve a more carbon neutral global state.

The STEP-AWBH model was developed for pixel-specific assessment of soil properties from a suite of soil-environmental factors (Grunwald et al. 2011). This conceptual model embraces soil (S), topography (T), ecology (E), parent material

(P), atmosphere (A), water (W), biota (B), and human (H) factors together to account for the effects on soil genesis. The STEP-AWBH model is flexible enough to be implemented with various geostatistical techniques, ensemble regression, and data mining methods to predict soil and terrestrial properties. For example, Xiong et al. (2014) applied the STEP-AWBH modeling concept to model SOC stocks in Florida using various data mining techniques, such as regression trees, bagged trees, random forest (RF), and support vector machines, from a large set of predictors (210 STEP-AWBH variables).

This study adopts the STEP-AWBH model and applies it to assess the attainable capacity of a terrestrial ecosystem to store carbon. Similar to the SOC sequestration, the attainable terrestrial carbon ($TerrC_{attain}$) is set by factors that limit the inputs of carbon to the system (e.g., residue from vegetation, and fertilization that stimulates biomass production/net primary production), which can be modified by humans (e.g., implementation of best management practices, reduction of burning of fossil fuels for energy consumption, and wetland restoration) (Ingram and Fernandes 2001; Stockmann et al. 2013). The actual terrestrial carbon ($TerrC_{actual}$) is controlled by factors that modulate carbon storage (e.g., drainage, tillage, land use management, soil respiration, or photosynthesis) and depends on a combination of environmental landscape factors, past and current anthropogenic forcings, and socioeconomic drivers. We postulate that human-induced management strategies combined with environmental fluctuations of climate, land use, and hydrology contribute to $TerrC_{attain}$ that is constraint by site-specific soil-landscape conditions. Importantly, which of these coupled human-environmental combinations achieve the maximum attainable carbon storage in terrestrial systems is usually not known. The objectives of this study were to: (i) assess the spatially explicit relationships between measured SOC stocks and environmental factors and (ii) assess the environmental value of actual ($TerrC_{actual}$) and attainable ($TerrC_{attain}$) terrestrial carbon capital across the Florida portion of the Suwannee River Basin, (FL-SRB) consisting of below-ground and above-ground carbon.

13.2 Study Area

The FL-SRB is located in north-central Florida with an approximate area of 19,665 km^2 (Fig. 13.1). Dominant soils are sand-rich which inherently does not promote SOC accretion. On the other hand, soils formed in aquic conditions are carbon-rich which are prominent in the FL-SRB. The soil temperature regimes are mixed with 86 % of the area's soil classified as thermic and 14 % as hyperthermic (Natural Resources Conservation Service (NRCS), 2006). The ecological landscape conditions such as land use/land cover (LULC) and hydrology are complex.

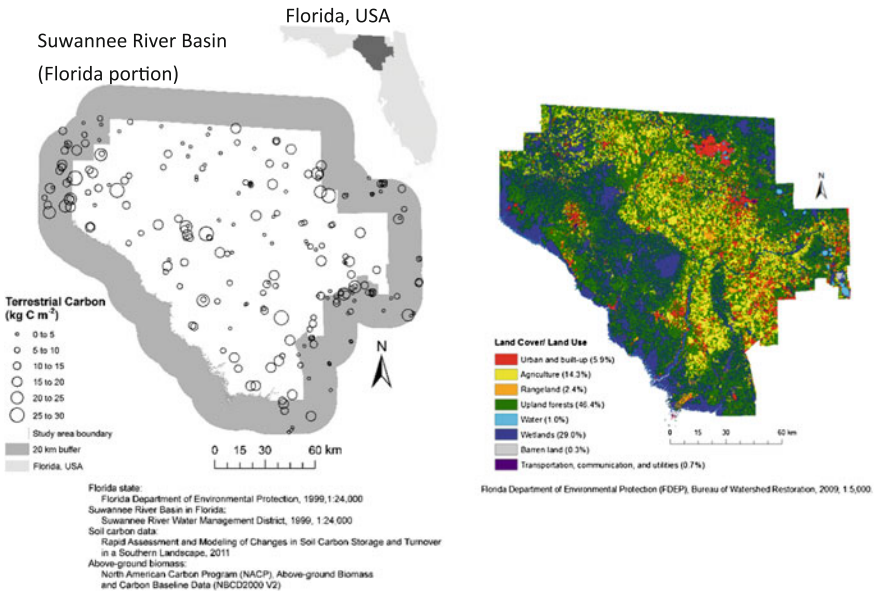

Fig. 13.1 Terrestrial carbon observations at 234 locations in the Suwannee River Basin (Florida portion) and land use/land cover classes

13.3 Materials and Methods

13.3.1 Above-Ground and Below-Ground Carbon Data

A total of 234 soil samples in the topsoil (0–20 cm) were collected between 2008 and 2009 across the FL-SRB and its buffer area (20 km around the FL-SRB) based on the random design stratified by LULC and soil suborders. Each soil sampling location was georeferenced with a differential global positioning system. Total carbon (TC) was measured by a combustion gas analyzer (Shimadzu TOC-V/SSM-5000) at 900°C, while inorganic carbon (IC) was analyzed by treating soil samples with 42.5 % phosphoric acid (H_3PO) and then combusting them at 200°C. The SOC concentration was calculated by subtracting the IC concentration from the obtained TC concentration (mg kg^{-1}). The SOC concentration for each site was converted to stock units (kg C m^{-2}) using measured bulk density and soil depth. For above-ground carbon assessment, we derived data from the LANDFIRE project which provided a high-resolution map of year-2000 baseline estimates of above-ground biomass carbon (National Biomass Carbon Data, NBCD 2000 Version 2) (Kellnorfer et al. 2013). The above-ground live and dry biomass in kg C m^{-2} was extracted at each of the soil sampling locations. The site-specific $TerrC_{actual}$ stocks were derived by the summation of the measured SOC stocks in

the top 20 cm of the soil and above-ground biomass from the NBCD database (kg C m^{-2}) based on 234 soil sampling locations.

13.3.2 Environmental and Anthropogenic Covariates

A comprehensive set of 172 environmental and human covariates representing the STEP-AWBH factors was compiled from multiple data sources using ArcGIS software. Predictors included 31 categorical and 141 continuous data types and are described in detail in Chaikaew (2014).

The S factor was described by 15 soil properties (e.g., soil taxonomic order and soil texture) and ten soil–water variables (e.g., surface soil moisture and drainage class); the T factor was represented by 9 topographic attributes (e.g., slope and compound topographic index); the E factor was described by 2 ecological variables (e.g., ecoregion and physiographic province); the P factor was represented by 2 parent material variables (e.g., environmental geology and surficial geology); the A factor was described by 3 atmospheric factors (e.g., precipitation, temperature, and solar radiation); the B factor was represented by 16 vegetation factors (e.g., LULC, canopy coverage, and cropland); and the H factor was described by 5 human covariates (e.g., population growth and household income).

13.3.3 Modeling the Relationships Between SOC and STEP-AWBH Factors

This study is embedded in the STEP-AWBH modeling concept which explicitly combines spatially and temporally explicit environmental and human variables that model the evolution of the soil ecosystem (Grunwald et al. 2011) (Eq. 13.1).

$$\mathrm{SA}(z, p_x, t_c) = f\left\{\sum_{j}^{n}\left[S_j(z, p_x, t_c), T_j(p_x, t_c), E_j(p_x, t_c), P_j(p_x, t_c)\right]\right\};$$

$$\int_{i=0}^{m}\left\{\sum_{j}^{n}\left[A_j(p_x, t_i), W_j(p_x,), B_j(p_x, t_i), H_j(p_x, t_i)\right]\right\} \tag{13.1}$$

where SA is the target soil (or terrestrial) realization, S represents the ancillary soil properties, T represents the topographic properties, E represents the ecological properties, P represents the parent material and geologic properties, A represents the atmospheric properties, W represents the water properties, B represents the biotic

properties, H represents the human-induced forcings, j is the number of predictors, $j = 1, 2,\ldots, n$, p_x is a pixel with size x (width = length = x) at a site specific on land, t_c is the current time, t_i is the time to t_c with time steps $i = 0, 1, 2, \ldots, m$, and z is the soil depth. The spatially explicit STEP factors capture the relative stable soil-forming factors within a human time frame, while the AWBH factors account for time-dependent variation (Thompson et al. 2012).

The RF regression method was used to identify the most powerful environmental predictive factors to model SOC. The model was randomly split into a calibration set (70 %, n = 164) and a validation set (30 %, n = 70). Model performance was assessed using the coefficient of determination (R^2), root-mean-square error (RMSE), and residual prediction deviation (RPD), and ratio of prediction error to interquartile range (RPIQ) was reported for error assessment of the RF model.

13.3.4 Assessing Terrestrial Carbon Stocks

The most powerful predictors (n = 43) in the first quantile of all STEP-AWBH variables of the SOC model were selected and used to predict the observed $TerrC_{actual}$ stocks using the RF model. To assess the $TerrC_{attain}$ stocks, we posit that the STEP factors are not expected to substantially change within a human time frame (e.g., past decades), whereas AWBH factors are likely to be variable in time and may increase or decrease. Thus, the latter were used to vary within a range of upper and lower bounds to simulate $TerrC_{attain}$.

The same 43 STEP-AWBH predictors were used to predict $TerrC_{attain}$ stocks at the 234 sites using a simulated annealing (SA) approach (Kirkpatrick et al. 1983). In the $TerrC_{attain}$ model, the STEP factors were kept constant. The AWBH factors were varied by ±10, ±20, and ±30 %, respectively, by keeping AWBH factors constant, except for one of them that was increased/decreased one-by-one with the respective percentage value, until all factor combinations were assessed within reasonable upper and lower bounds. The factor combination which amounted to the highest terrestrial carbon stocks simulated at the 234 sites for the FL-SRB was postulated to equal the attainable terrestrial carbon stocks that could be obtained based on dynamic AWBH variables and relatively stable STEP conditions.

To characterize the spatial distribution of actual terrestrial carbon stocks across the basin, the regression kriging (RK) technique (Odeh et al. 1995) was used. First, the residuals of $TerrC_{actual}$ at the 234 sites were kriged and then added back to the estimates of $TerrC_{actual}$ to create interpolated surfaces showing terrestrial carbon stocks.

13.4 Results

13.4.1 Variable Importance and Spatial Variation in Soil Organic Carbon

The variables that emerged in the first quantile of the RF-SOC model showing predictive power were as follow: biota > soil > topography. Minor, yet significant, predicting variables represented water, atmospheric properties, and parent material. Soil taxonomic variables, such as soil great group, suborder, and subgroup, were highly interrelated with SOC stocks as shown in the top ten explanatory variables. Areas with poorly to very poorly drained soils tended to accumulate SOC content, whereas areas with well-drained to excessively drained soils had a tendency to have net losses of SOC. These results suggest that topographic and soil/water-related variables play a dominant role to infer on SOC storage in the basin.

Slope was negatively correlated with SOC stocks. This indicates that the relatively flat downslope positions were correlated with high SOC, and vice versa, upslope positions showed the opposite behavior. Even though topographic terrain in Florida is relatively flat (0–5 % slopes), the steeper slopes are found in the northern region of Florida where the FL-SRB is located. Distance from streams or open water, available water capacity (0–25 cm), hydrologic group, ponding frequency class, and soil runoff potential were water variables (W) in the model that demonstrated strong connectivity with SOC stocks. The effect of wetness in soil is considered to be a major factor that controls vegetation growth and the decomposition process that are closely interlinked with SOC gain (Vasques et al. 2012).

The parent material was found to have high influence on SOC, while the ecology variables (E) were not strongly associated with SOC. The influence of parent material on SOC stocks occurs through different sources, such as soil weathering, mineralogy, water permeability, nutrient supply, mineral complexation, structure, that control the pH, and microbe's habitat affecting plant production and decomposition (Post et al. 2004).

Three climatic variables of long-term (2000–2008) monthly maximum temperature in summer (July, August, and September) were negatively correlated with SOC stocks. The opposite was found for annual average precipitation that was positively correlated with SOC stocks. Ample research has been conducted to study the interactions between climate and SOC which are still debated fiercely because interactions vary geographically around the globe (Bardgett 2011; Ontl and Schulte 2012; Poeplau et al. 2011; Xiong et al. 2014).

Human variables were not considered powerful predictors in the model as they ranked in the middle and bottom of all predictor variables. Fertilizer consumption (74th) had the most influence among the H factor, followed by best management practice (BMP) implementation (100th) of all variables, while population growth ranked near the bottom of all predictors. This indicates that human factors may have an indirect effect (e.g., through land use management and tillage operations) or little impact on SOC.

13.4.2 Model Performance

The prediction model of SOC stocks using 172 STEP-AWBH variables was able to account for 93 % of the variation in calibration mode and 44 % of the variation in validation mode across the basin. The RMSE value of 2.65 kg C m^{-2} in the validation set was higher than 1.33 kg C m^{-2} in the calibration set, and RPD values of 1.28 kg C m^{-2} in the validation set were lower than 2.58 kg C m^{-2} in the calibration set. Considering that the minimum of SOC stock was 1.1 kg C m^{-2} and the maximum of SOC stock in the basin was 28.8 kg C m^{-2}, the errors suggest that the model performed moderately well.

13.4.3 Estimates of the Terrestrial Carbon Stocks

The amount of carbon was almost twofold in the terrestrial ecosystem (mean of 9.2 kg C m^{-2}) in comparison with topsoil (mean of 5.2 kg C m^{-2}) at a site-specific basis. The results imply that about 44 % of carbon storage is found above ground and 56 % of carbon is stored in topsoils. This is a conservative estimate because additional carbon is present in subsoils. The relatively high terrestrial carbon stocks were observed in swamps and forests. The highest mean $TerrC_{actual}$ stock values were found in mixed wetland forests 16.6 kg C m^{-2}), followed by swamps (14.6 kg C m^{-2}) and pineland (11.0 kg C m^{-2}). The lowest $TerrC_{actual}$ stocks were present in row/field crops with a mean value of 1.2 Tg C (2.7 kg C m^{-2}).

The modeled $TerrC_{attain}$ estimates under different environmental forcings (i.e., AWBH forcings) clearly showed that factor combinations (i.e., climatic properties, water, and biota) concomitantly had strong effects on carbon storage. Assuming that predictors are changed by ±10 %, our model simulated a minimal increase in terrestrial carbon stocks with mean values of 9.3 kg C m^{-2} and median values of 8.8 kg C m^{-2}. The mean $TerrC_{attain}$ was 9.4 kg C m^{-2} and the median was 8.8 kg C m^{-2} when the environmental factors fluctuated by ±20. And the mean TerrCattain was 9.4 kg C m^{-2} and the median was 9.0 kg C m^{-2} when the environmental factors fluctuated by ±30.

The $TerrC_{actual}$ stocks in the model indicated that this terrestrial system was close to the saturation condition. The absolute increase from predicted $TerrC_{actual}$ to $TerrC_{attain}$ (±30 %) in terrestrial carbon stocks amounted to 4.3 Tg (mean) and 13.2 Tg (median) which are substantial amounts. The predicted $TerrC_{attain}$ (±30 %) for wetlands, pinelands, and hardwood forests were 13.7, 9.4, and 10.5 kg C m^{-2}, respectively, while the predicted $TerrC_{actual}$ for wetlands, pinelands, and hardwood forests were 11.9, 9.1, and 10.3 kg C m^{-2}, respectively. The $TerrC_{actual}$ and $TerrC_{attain}$ stocks in pineland and forests were quite similar in values with differences in $TerrC_{actual}$ and $TerrC_{attain}$ of 0.3 kg C m^{-2} (pineland) and 0.3 kg C m^{-2} (hardwood forests). The actual storage of carbon in the terrestrial ecosystem

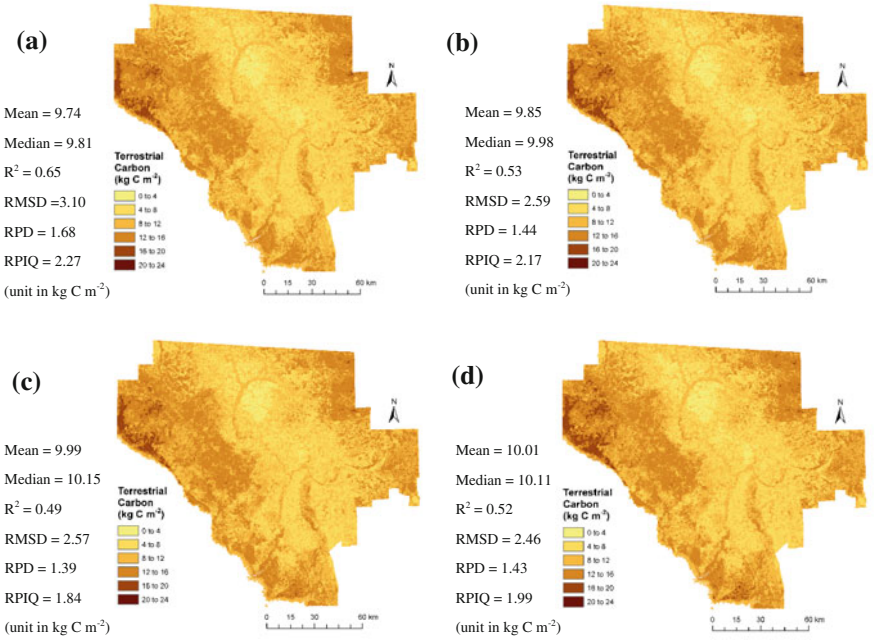

Fig. 13.2 Terrestrial carbon stock estimates derived from regression kriging (RK): **a** observed actual terrestrial carbon stocks; **b**, **c**, and **d** attainable terrestrial carbon stocks derived from simulated annealing considering environmental variables change by ±10, 20, and 30 %, respectively

amounted to 190 Tg C and the maximum storable attainable carbon was 195 Tg C. The spatial distribution of terrestrial carbon stocks is shown in Fig. 13.2.

13.5 Conclusions

We assessed the relationships between SOC stocks and STEP-AWBH factors in the FL-SRB and found that biotic, soil/water, and topographic factors played crucial roles in determining SOC storage, whereas human factors seemed to be fading from being strong predictors. Among the predictors, maximum temperature in summer time and mean annual precipitation also stood out as controlling factors for SOC storage. The whole basin stores tremendous amounts of carbon, multiple times larger than atmospheric carbon, with about 190 Tg of $TerrC_{actual}$ carbon stocks based on the RK model. There was no single environmental factor that imparted most control on SOC storage, but instead intricate combinations of STP-AWB variables.

Importantly, biotic factors played a larger role to associate with SOC stocks and terrestrial carbon compared to climatic factors. This inherently implies that global climate change will have less of an impact compared to land cover and land use change to achieve the attainable carbon level in this soil landscape. The model predicted that the $TerrC_{actual}$ stock values were close to the $TerrC_{attain}$ stock values in some instances (e.g., under wetlands) suggesting the presence of saturation in this area, while in other regions (e.g., under row/field crops) ample opportunities exist to enhance carbon sequestration.

The simulated $TerrC_{attain}$ values were "point specific (in kg C m^{-2})" assuming no fixed depth (i.e., soil depth/volume or vegetation height/volume) to store carbon. This vantage point liberates us from the constraint to consider a fixed soil volume (kg C m^{-2} and 0–20 cm depth) to accrete carbon up to a saturation limit. As is well known, hydric soils tend to accrete carbon through an increase in soil depth rather than enhancement of carbon density within existing peat layers. Similarly, attainable biomass carbon is not necessarily linearly linked to a specific height of the vegetation and may increase carbon not only through growth but also through changes in the vegetation density.

The major goal was to preserve the carbon stored in the terrestrial system of the FL-SRB and enhance them through optimized carbon management. Land use adaptations have much potential to reach $TerrC_{attain}$, specifically land use conversions from cropland to systems with larger net primary productivity (NPP). Bare soils, which represent marginal soils, also bear potential to elevate carbon storage if improved through management. Our study demonstrated a novel approach to assess terrestrial carbon through management, adaptation, and mitigation that is attainable in a subtropical basin consisting of a mosaic of different land uses embedded in a complex soil landscape. This approach is generalizable and transferable to any other landscape setting. Rather than offsetting CO_2 emissions, there are other cobenefits of increased levels of carbon sequestration to the ecosystem functions (e.g., improvements in crop productivity, soil security, food security, soil aggregation that enhances nutrient storage, and water holding capacity). The spatially explicit modeling of actual and attainable terrestrial carbon stocks allows identifying and targeting carbon poor areas to implement conservation and carbon management strategies. Hence, our approach has much value to outline pathways into a carbon-rich future.

Acknowledgements Soil data collection is supported by the "Rapid Assessment and Trajectory Modeling of Changes in Soil Carbon across a Southeastern Landscape" funded by USDA-NIFA, a core project of the North American Carbon Program. This research was conducted through a PhD fellowship funded by the Royal Thai Government.

References

Bardgett RD (2011) Plant-soil interactions in a changing world. Biol. Rep. 3. doi:10.3410/B3-16

Bélanger N, Pinno BD (2008) Carbon sequestration, vegetation dynamics and soil development in the boreal transition ecoregion of Saskatchewan during the Holocene. Catena 74: 65–72. doi:10.1016/j.catena.2008.03.005

Chaikaew P (2014) Assessment of climate regulation, carbon sequestration, and nutrient cycling ecosystem services impacted by multiple stressors. Ph.D. dissertation, University of Florida, Gainesville, FL, USA.

Dickson B, Blaney R, Miles L, Regan E, van Soesbergen A, Väänänen E, Blyth S, Harfoot M, Martin, CS, McOwen C, Newbold T, van Bochove J (2014) Towards a global map of natural capital: key ecosystem assets. UNEP, Nairobi, Kenya. Available at http://www.unep-wcmc.org/system/dataset_file_fields/files/000/000/232/original/NCR-LR_Mixed.pdf?1406906252

Grunwald S, Thompson JA, Boettinger JL (2011) Digital soil mapping and modeling at continental scales: finding solutions for global Issues. Soil Sci Soc Am J 75: 1201-1213. doi:10.2136/sssaj2011.0025

Ingram JSI, Fernandes ECM (2001) Managing carbon sequestration in soils: concepts and terminology. Agric. Ecosyst. Environ. 87: 111–117. doi:10.1016/S0167-8809(01)00145-1

Kellnorfer J, Walker W, Kirsch K., Fiske G., Bishop J, Lapoint L, Hoppus M, Westfall J (2013) NACP Abovegournd Biomass and Carbon Baseline Data, V.2 (NBCD 2000), U.S.A., 2000 dataset. Oak Ridge, TN

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220: 671–680. doi:10.1126/science.220.4598.671

Lal R (2008) Carbon sequestration. Philos Trans R Soc B Biol Sci 363: 815–830. doi:10.1098/rstb.2007.2185

Natural Resources Conservation Service (NRCS) (2006) Soil Survey Geographic (SSURGO) Database, Available at http://soildatamart.nrcs.usda.gov

Odeh IOA, McBratney AB, Chittleborough DJ (1995) Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67: 215–226. doi:10.1016/0016-7061(95)00007-B

Ontl TA, Schulte LA (2012) Soil carbon storage. Nat. Educ. Knowl. 3(10): 35.

Poeplau C, Don A, Vesterdal L, Leifeld J, Van Wesemael B, Schumacher J, Gensior A (2011) Temporal dynamics of soil organic carbon after land-use change in the temperate zone – carbon response functions as a model approach. Glob. Change Biol. 17: 2415–2427. doi:10.1111/j.1365-2486.2011.02408.x

Post WM, Izaurralde RC, Jastrow JD, McCarl BA, Amonette JE, Bailey VL, Jardine PM, West TO, Zhou J (2004) Enhancement of carbon sequestration in US Soils. BioScience 54: 895–908

Stockmann U, Adams MA, Crawford JW, Field DJ, Henakaarchchi N, Jenkins M, Minasny B, McBratney AB, Courcelles V de R de, Singh K, Wheeler I., Abbott L, Angers DA, Baldock J, Bird M, Brookes PC, Chenu C, Jastrow JD, Lal R, Lehmann J, O'Donnell AG, Parton WJ, Whitehead D, Zimmermann M (2013) The knowns, known unknowns and unknowns of sequestration of soil organic carbon. Agric. Ecosyst. Environ. 164: 80–99. doi:10.1016/j.agee.2012.10.001

Thompson JA, Roecker SM, Grunwald S, Owens PR (2012) Digital soil mapping: interactions with and applications for hydropedology, in: Lin, H.S. (Ed.), Hydropedology - Synergistic Integration of Pedology and Hydrology. Academic Press, Elsevier B.V., pp. 665–709

Wang Y, Fu B, Lü Y, Chen L (2011) Effects of vegetation restoration on soil organic carbon sequestration at multiple scales in semi-arid Loess Plateau, China. CATENA 85: 58–66

Wu QB, Wang XK, Ouyang ZY (2009) Soil organic carbon and its fractions across vegetation types: Effects of soil mineral surface area and microaggregates. Pedosphere 19: 258–264

Vasques GM, Grunwald S, Myers DB (2012) Associations between soil carbon and ecological landscape variables at escalating spatial scales in Florida, USA. Landscape Ecology J. 27: 355-367. doi:10.1007/s10980-011-9702-3

Xiong X, Grunwald S, Myers DB, Kim J, Harris WG, Comerford NB (2014) Holistic environmental soil-landscape modeling of soil organic carbon. Environ Model Softw 57: 202-215. doi:10.1016/j.envsoft.2014.03.004

Chapter 14
The Meta Soil Model—An Integrative Framework to Model Soil Carbon Across Various Ecosystems and Scales

S. Grunwald, P. Chaikaew, B. Cao, X. Xiong, G.M. Vasques, J. Kim, C.W. Ross, C.M. Clingensmith, Y. Xu and C. Gavilan

Abstract Over the past decades, a changing climate, land use shifts, socioeconomic development, and political decisions have had a tremendous impact on the spatial and temporal variation of soil carbon. How soil carbon interacts with such changing natural environmental and anthropogenic forcings within various ecosystem domains and spatial and temporal scales is still poorly understood. We discern different paradigms to model soil carbon and explore the meaning of such diversity in soil carbon paradigms situated within digital soil mapping (DSM) and beyond. The *Meta Soil Model* offers a container to hold multiple modeling paradigms that generate a variety of soil carbon realizations. The term soil realization acknowledges that there is not only one 'soil carbon map' or 'soil carbon model', but also several possible ones that approximate reality. The *Meta Soil Model* allows integrating, fusing, and synthesizing various soil carbon observations/maps/models through laboratory, field, or proximal/remote methods and ensembles other integration methods aiming to create more holistic representations of soil carbon. Besides explicit integration of soil carbon data/maps/models, the *Meta Soil Model* also facilitates side-by-side comparisons in a consistent and coherent framework. Here, we present a multiplicity of different DSM and modeling approaches and how they are integrated into a *Meta Soil Carbon Model*. Each approach is exemplified by

S. Grunwald (✉) · P. Chaikaew · B. Cao · X. Xiong · J. Kim · C.W. Ross
C.M. Clingensmith · Y. Xu · C. Gavilan
Soil and Water Science Department, University of Florida, 2181 McCarty Hall,
PO Box 110290, Gainesville, FL 32611, USA
e-mail: sabgru@ufl.edu

P. Chaikaew
Environmental Science Department, Chulalongkorn University, 254 Payathai Rd.,
Wang Mai, Pathumwan, Bangkok 10330, Thailand

G.M. Vasques
Embrapa Solos, Rua Jardim Botânico, 1024—Jardim Botânico, Rio de Janeiro-RJ
22460-000, Brazil

J. Kim
Department of Environmental Engineering, Chungnam National University,
432 College of Engineering III, 99 Daehak-Ro, Yuseong-Gu, Daejeon 305764, South Korea

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_14

a coherent model that entails the full suite of classical steps adopted in DSM to: (1) identify research questions and model approach, (2) develop a sampling design, (3) collect soil carbon data, (4) collect ancillary data in environmental and human domains, (5) analyze data (modeling), (6) create soil carbon predictions, estimates, or simulations and their uncertainties, and (7) test and validate soil carbon models. We present the integration pathways to build each of the exemplified *Meta Soil Carbon Models*. In conclusion, soil carbon can be viewed through various lenses—from above (through remote and/or proximal sensing), below (a soil pit or petri dish in the laboratory), or sideways (i.e., in new ways integrating multiple approaches). DSM and modeling is shifted into a new phase that is pluralistic in nature embracing a multiplicity of pathways focused to integrate data, methods, and knowledge and to understand about soils and ecosystems. In that sense, it is becoming more and more inter- and transdisciplinary, and through multiple comparisons, adaptations and validations, more robust, reliable and useful.

Keywords Soil organic carbon · Meta Soil Model · Digital soil mapping paradigms · Integration · Fusion · Soil models

14.1 Introduction

Over the past decades, a changing climate, land use shifts, population growth, and associated socioeconomic development and political decisions have had a tremendous impact on the spatial and temporal variation of soil carbon. This has spawned a profound number of soil carbon-related research from pedometrics, biogeochemistry, physical (e.g., sensing), and other vantage points. 'Soil carbon' publications have been exploding with currently over 1.76 million publications identified by Google Scholar using a generic search (205,000 using an exact phrase search) and 268,656 publications in the Web of Science. The topical focus of soil carbon studies has been diverse with a vast amount on soil carbon assessment (20.5 % in Web of Science and 29.6 % in Google Scholar), soil carbon modeling (17.6 % in Web of Science and 24.3 % in Google Scholar), and less so connoted explicitly as digital soil mapping (DSM) (1.2 % in Web of Science and 3.0 % in Google Scholar) (Table 14.1). A major amount of publications on soil carbon has been focused on soil carbon and management (33.4 % in Web of Science and 42.9 % in Google Scholar) and understanding soil carbon from a chemical perspective (25.4 % in Web of Science and 20.9 % in Google Scholar). Soil carbon has been stylized as a unifying theme playing a key role in inter- and transdisciplinary projects and programs, soil security (Bouma and McBratney 2013; McBratney et al. 2014), and food security (Lobell et al. 2008).

Considering this profound amount of knowledge on soil carbon, critical questions arise. How do we deduct and synthesize knowledge and understanding from these millions of publications and studies? Are there DSM methods/models that are universally applicable to gain insight into Spatio- and temporal soil carbon

Table 14.1 Summary of the number of publications (N) focused on soil carbon and associated research topics as identified by two prominent global scholarly literature tracking systems (status September 2014)

Keywords/topics	N in Web of Science[a]	N in Google Scholar[b,c]
Soil carbon/soil organic carbon/total soil carbon/soil organic matter	268,656/127,857/70,477/183,846	205,000[c]/106,000/3,510/590,000
Soil carbon and dynamics	30,673	76,200
Soil carbon and flux	23,395	31,100
Soil carbon and nitrogen	86,910	79,800
Soil carbon and phosphorus	25,489	29,300
Soil carbon and nutrients	43,600	55,600
Soil carbon and biogeochemistry/chemistry	2,094/68,113	26,700/42,800
Soil carbon and fractions/pools/aggregates	28,129/17,174/8,094	31,900/41,700/21,400
Soil carbon and modeling/model	47,239/47,239	49,800/111,000
Soil carbon and mapping/map/digital soil mapping	3,250/3,248/258	24,000/29,800/6,170
Soil carbon and assessment/quantify/budget	55,152/13,415/6,107	60,700/33,500/24,700
Soil carbon and spectroscopy/proximal sensing/remote sensing	8,265/33/1,391	14,500/613/13,000
Soil carbon and spatial	15,258	47,300
Soil carbon and change/sequestration/temporal/time	72,862/18,767/9,370/52,755	110,000/36,900/33,700/154,000
Soil carbon and spatio-temporal	2,813	18,200
Soil carbon and integration, integrate/synthesis/fusion/meta analysis	7,466/3,248/217/502	25,300/28,000/2,900/9,900
Soil carbon and nano/micro/pedon/field/site/regional/national/continental/global	424/8,622/180/64,227/44,742/7,809/5,864/1,509/21,216	1,400/19,800/1,100/78,400/76,400/53,000/58,000/14,700/79,800
Soil carbon and concentration/content/density/stock/storage		89,400/85,700/49,100/30,800/62,300
Soil carbon and management	89,611	88,000
Soil carbon and land use/land cover	24,688/5,532	76,700/66,300

(continued)

Table 14.1 (continued)

Keywords/topics	N in Web of Science[a]	N in Google Scholar[b,c]
Soil carbon and climate/climate change	38,867/25,806	85,800/47,900
Soil carbon and topography/terrain/relief	1,308/653/308	16,800/8,180/5,780
Soil carbon and risk/adaptation/mitigation/vulnerability/sustainability	7,129/2,737/2,299/476/2,895	23,700/23,900/21,800/15,200/29,800

[a]Web of Science by Thomas Reuter: The selection 'Topic' search was used to derive numbers without time restriction

[b]Google Scholar is a freely accessible Web search engine that indexes the full text of scholarly literature across an array of publishing formats and disciplines. The Google search was done selecting the keyword with option 'exact phrase' and searching anywhere in 'articles' without time restriction and excluded the options 'patents' and 'citations'

[c]A generic search for 'soil carbon' with the option 'with all of the words' yielded 1.76 million returns in publications, whereas the search for 'soil carbon' with the option 'with the exact phrase' yielded 205,000 returns in publications in Google Scholar

dynamics? Is the site-specific imprint that controls soil carbon dynamics too unique and geographically diverse hampering generalizations about soil carbon variation in space and time? The latter question arises specifically in context of wicked environmental problems at continental and global scale (Brown et al. 2010). In this paper, we hone into the depth of integrative soil carbon modeling across various ecosystems and scales. We present pathways for a more integrative soil carbon science that has the potential to raise more public awareness about soils, increase political action to sustain soil resources, and engage stakeholders and practitioners.

14.2 Soil Carbon Modeling Paradigms

Considering different paradigms, soil carbon can be perceived as (i) an empirical variable mapped at a specific time and geographic location—e.g., in state budget assessments or soil carbon maps, (ii) a response (dependent) variable to stressors or landscape factors (e.g., climate or management)—e.g., in soil factorial models, such as SCORPAN (McBratney et al. 2003) or STEP-AWBH (Grunwald et al. 2011), (iii) a relational variable interconnected with other soil ecosystem properties as part of an ecosystem—e.g., statistical and chemometric models, (iv) soil natural capital in ecosystem services context, (v) a state variable that describes the mathematical state of a dynamic ecosystem comprised of multiple interacting biological, physical, and chemical processes—e.g., mechanistic simulation or pedogenetic models, (vi) an active agent that interacts with the environment through positive and negative feedback processes cocreating their environment—e.g., in autopoietic and agent-based models, (vii) a metaphor or symbol (e.g., 'soil carbon $\sim$ black gold') evoking intuitive, emotional responses in people to protect, sustain, or degrade soils.

An example for model paradigm (i) was provided by Guo et al. (2006) who mapped soil organic carbon (SOC) and soil inorganic carbon (SIC) across the USA. They used the State Soil Geographical database (STATSGO), which represents soil carbon by map units (polygons), and provided carbon budgets for different regions in the USA. Approach (ii) is exemplified by the SCORPAN modeling approach that is used widely in DSM. For instance, Grimm et al. (2008) mapped SOC concentrations and stocks using environmental covariates and random forest (RF) in Panama. Grunwald (2009) provided a comprehensive review of studies that assessed soil carbon (and other properties) using various soil factorial models, spatial scales, and methods applied in diverse geographic regions. Another example of approach (ii) was provided by Vasques et al. (2010b) who mapped soil total carbon (TC), recalcitrant carbon, hydrolyzable carbon, hot water soluble (labile) carbon, and mineralizable soil carbon across a subtropical watershed in the southeastern USA. They applied lognormal block kriging and regression block kriging in their study. The conceptual STEP-AWBH modeling framework (paradigm ii) uses space-time soil environmental variables to predict a targeted soil property (e.g., soil carbon). It incorporates hydrologic and anthropogenic variables explicitly into the modeling

process. Xiong et al. (2014a, b) demonstrated the STEP-AWBH DSM approach using 210 potential space-time variables that exhaustively cover pedogenic and environmental factors. Their aim was to identify those variables with the strongest response to estimate SOC in Florida, USA. Several other prediction, estimation, and simulation methods have been used to infer on soil carbon, including geostatistical, fuzzy logic, neural network, Bayesian, data mining, and other methods. They all can be grouped under (ii) because they imply directionality where input data (predictor variables) are used to infer on a response (dependent) variable (e.g., SOC). This differs from approaches grouped under (iii) that focus on understanding relationships rather than making predictions of soil properties. These relationships identify correlations and reciprocity between properties acknowledging that soil properties are interdependent on other environmental properties and, vice versa, environmental properties form them. Importantly, no directionality or cause–effect relationships are implied in model (iii) types. Typical methods that represent (iii) are genetic algorithms, which are heuristic search methods to investigate input–output relationships. The model paradigm (iii) is prominent in soil spectroscopy where hyperspectral data (e.g., visible/near-infrared/mid-infrared spectral range) are related to laboratory-measured soil properties or pedo-transfer functions. Examples for approach (iii) investigating relationships between soil carbon and spectral data were provided by Gomez et al. (2008), Vasques et al. (2008), Minasny et al. (2009), and Ladoni et al. (2010). Another example (iii) where the emphasis was on understanding relationships between land use change, climatic factors, and SOC using general linear models was presented by Xiong et al. (2014a, b). The approach (iv) focuses on soil carbon capital assessed typically within the context of ecosystem services. For example, Egoh et al. (2008) mapped soil carbon storage as one service among others in a study in South Africa. SOC sequestration is considered a soil ecosystem service (or benefit to humans) as demonstrated in a comprehensive study using the Integrated Valuation of Ecosystem Services and Tradeoffs (InVest) tool in Oregon, USA (Nelson et al. 2009). Approach (v) can be exemplified by soil carbon simulation models, such as Century (Kelly et al. 1997) or Roth-C (Jones et al. 2005). These models adopt a system theoretical approach modeling ecosystem processes deterministically based on mechanistic, process-based understanding of biological, physical, chemical, pedogenic, hydrologic, and other interconnected systems. Manzoni and Porporato (2009) reviewed 250 biogeochemical models that simulate soil carbon and nitrogen mineralization across multiple spatial and temporal scales demonstrating the widespread use of models adopting strategy (v). Agent-based models, representing paradigm (vi), have appeared in other domains, such as agent-based land use modeling (Matthews et al. 2007), agent-based climate modeling (Moss et al. 2001), and human–environment interaction modeling (Schreinemachers and Berger 2011). Matthews (2006) applied an agent-based model (People and Landscape Model, PALM) in which soil carbon was one of the components considering household agents to simulate resource flows. The last one, paradigm (vii), uses a symbolic or metaphorical approach often framed as narratives to express the value and beliefs about soil carbon. These may be positively framed as concepts of safety or security (e.g., ‘soil carbon secures soils, functionality, food security, and survival of humanity’) or negatively framed as threat or risk (e.g., ’soil

carbon loss degrades food production and threatens livelihood of smallholder farm communities'). The perception of these narratives about soils is associated to beliefs and values that typically evoke people, stakeholders, action groups, and politicians to act, rest in denial, ignorance, indifference, or paralysis. The disparity among soil carbon paradigms (i–vii) is profound. To integrate these contrasting and antithetical perspectives is not about one approach being better and more profound than the other (e.g., stating that my soil carbon assessment is better or more right than yours). A truly integral approach honors all perspectives about soil carbon. Yet it takes a stance to find unity in the diversity of soil carbon paradigms. Esbjörn-Hargens and Zimmerman (2009) described such an integrative framework—Integral Methodological Pluralism (IMP)—based on three key principles: (i) nonexclusion, (ii) enactment, and (iii) enfoldment. These principles are at the core of the *Meta Soil Model* that was introduced by Grunwald (2013, 2014).

14.3 The Meta Soil Model—Integrative Modeling of Soil Carbon

Grunwald (2013) describes the *Meta Soil Model* that juxtaposes soil and environmental datasets and various methods grounded in different philosophical worldviews. It addresses the quintuplet questions—'Why,' 'For Whom,' 'What,' 'Who,' and 'How' we can describe and model soils and soil ecosystems. Therefore, it connects the purpose of soil maps and models with those who will use and value them (e.g., land managers, decision makers), identifies the soil attributes that are mapped and soil landscapes that are modeled, the individual experts or interdisciplinary teams that are performing the DSM, and methods that are applied. Thus, this framework is multidimensional (Fig. 14.1) and in a broad sense relates to the five dimensions of soil security: (i) capability, (ii) condition, (iii) capital, (iv) connectivity, and (v) codification (McBratney et al. 2014). Importantly, the *Meta Soil Model* offers a container to hold multiple modeling paradigms that generate a variety of soil realizations and integration pathways (Grunwald 2014). In its core, it is pluralistic and embraces synthesis and integration aiming to enhance our deep understanding of soils and its role in context of complex and wicked environmental problems at global scale. Synthesis is a key integrative concept, and it occurs when disparate data, concepts, or theories are combined in ways that yield *new* knowledge, values, insights, understanding, or explanations (Pickett et al. 2007; Carpenter et al. 2009; Peters 2010). The *Meta Soil Model's* intent is to synthesize, thereby generating *new* knowledge, values, and insight, and understand soil ecosystems connecting past, current, and future soils through integration pathways. Thus, it holds much premise to improve contemporary soil models. It provides an explicit quantitative construct complementing the loosely coupled dimensions of soil security. The term *'Meta'* ('after,' 'beyond') expresses the complexification of models where '*Meta Model*' (or surrogate model) refers to a 'model of a model'

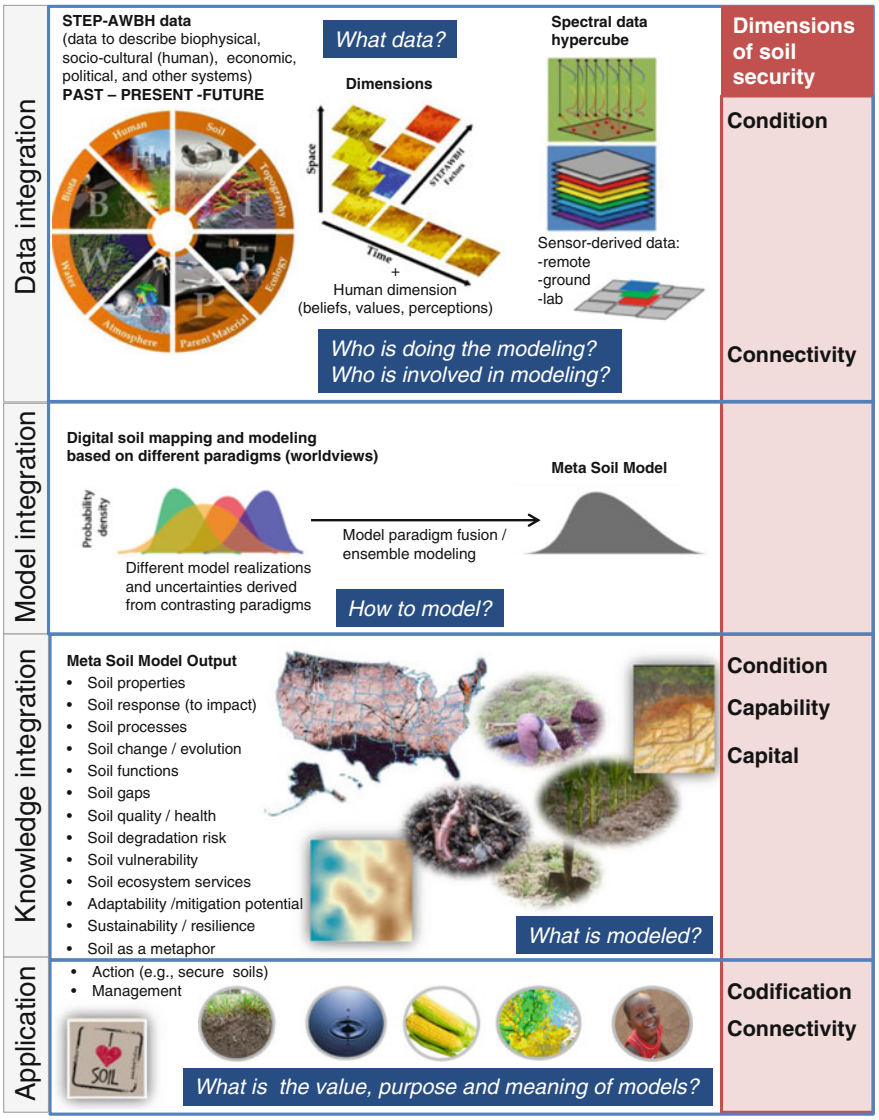

Fig. 14.1 The Meta Soil Model—an integrative model and how it relates to soil security

meaning that a more complex model is build that integrates submodels, data, algorithms, concepts, or other. For example, several decades ago, DSM pioneers applied ordinary kriging to model soil carbon, while recently more complex methods are applied, such as regression kriging, and ensemble soil carbon models. At the cutting edge of DSM at the current time are soil carbon models that even go beyond those combining data fusion and meta modeling (e.g., multitier soil carbon

frameworks derived from splining of soil carbon profile data, Bayesian estimation of soil carbon and uncertainty assessment, and coupling to human-induced impact assessment to quantify soil carbon change and evolution). These models strive for higher spatial and temporal resolutions and more explicit linkages between attribute (here: soil carbon), environmental factors, ecosystem processes, functions, services, and responses.

We argue that the *Meta Soil Model* is poised to extend contemporary soil model applications because it generates various realizations of soil properties (Fig. 14.1), and thus, allows derivation of more complex soil ecosystem assessments, such as soil risk, vulnerability of soils, adaptability, and sustainability. The term soil realization acknowledges that there is not only one 'soil carbon map' or 'soil carbon model', but several possible ones that approximate reality. The *Meta Soil Model* facilitates fusion and synthesis of various soil carbon observations/maps/models through laboratory, field, or proximal/remote data processing methods and ensembles Bayesian or other integration methods aiming to create more holistic representations of soil carbon. Grunwald et al. (2014a) presented the underlying data infrastructure to populate a *Meta Soil Carbon Model* exemplified for the USA. This framework is transferable to any region and scale and to develop other type of *Meta Soil Models* (e.g., *Meta Soil Security Model*). Grunwald et al. (2014b) provided an explicit description of 20 different integration pathways that compose the *Meta Soil Model*. It is applicable to define a *Meta Soil Carbon Model*.

14.4 Moving Toward a Meta Soil Carbon Model

We applied the *Meta Soil Carbon Model* scheme in various ecosystems and across temporal and spatial scales in the southeastern USA including: (i) SOC space-time mapping (Ross et al. 2013); (ii) predictive SOC modeling from environmental covariates (Vasques et al. 2010b, 2012a; Xiong et al. 2014a, b); (iii) modeling of SOC–environmental covariate relationships (Vasques et al. 2012b; Xiong et al. 2014a, b), soil carbon–soil relationships (Ahn et al. 2009; Bliss et al. 2013), and soil carbon–spectral (visible/near-infrared and mid-infrared spectral data) relationships (Vasques et al. 2008, 2009, 2010a; Knox et al. 2015); (iv) soil carbon ecosystem services modeling soil carbon as natural carbon capital (Chaikaew 2014); (v) process-based modeling of soil carbon dynamics using Century (Kwon and Grunwald 2014) and DayCent to model SOC vulnerability and adaptability to climate change (ongoing); and (vi) agent-based modeling of soil carbon dynamics (future). Next, we present two examples of integration pathways embedded within the *Meta Soil Carbon Model*.

14.4.1 Meta Soil Carbon Modeling *Based on Data Mining and Ensemble Modeling of SOC Stocks in Florida* (Fig. 14.2)

Data mining is the practice of examining large datasets in order to reveal new patterns and generate new information and knowledge. Typically, it involves combinations of methods (i to vi) such as machine learning, statistical methods, and artificial intelligence. Xiong et al. (2014a, b) used 210 different space-time environmental covariates to model SOC stocks in Florida, USA. Figure 14.2 identifies the integration pathways (as outlined by Grunwald et al. 2014b) adopted in this study to build a *Meta Soil Carbon Model* for Florida. They employed a two-tier approach consisting of 'all-relevant variable strategic selection' that revealed the ecosystem processes to explain the variation in SOC stocks and 'minimum-optimal variable selection' to identify parsimonious models with similar accuracy when compared to exhaustive, highly parameterized SOC models. In this study, ensemble models were employed to derive SOC predictions. The major factors explaining SOC variation in Florida were vegetation and soil water gradient. Topography and climate showed moderate effects on SOC variation (Xiong et al. 2014a, b). These results confirmed findings by Vasques et al. (2012b) who identified soil hydrologic factors and biotic properties as most critical to explain the spatial variation in SOC in Florida These results differ from other geographic regions where land use, vegetation, and topography impart most control on SOC variability. The *Meta Soil Carbon Model* is used in the Florida Forever Conservation Program.

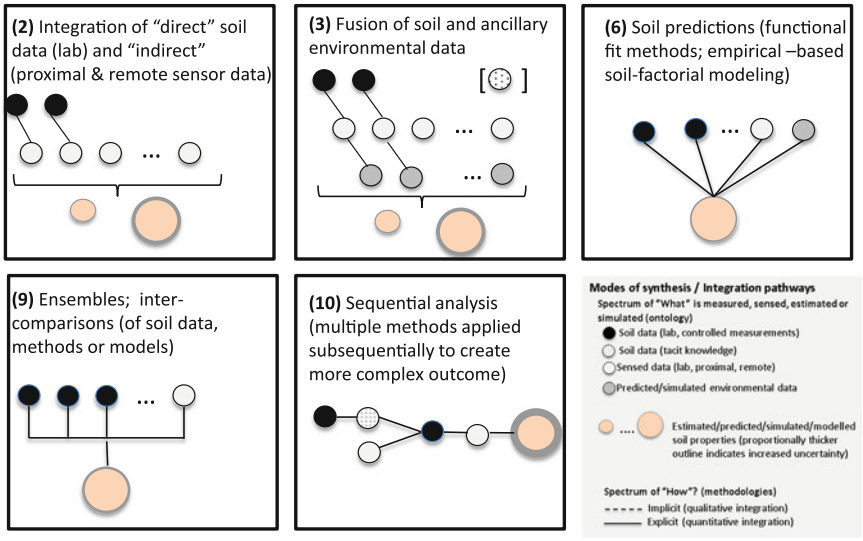

Fig. 14.2 Integration pathways as codified by Grunwald et al. (2014b) to build a Meta Soil Carbon Model for Florida based on Xiong et al. (2014a, b). In this Meta Soil Carbon Model, integration pathways (*2*), (*3*), (*6*), (*9*), and (*10*) were used (see Grunwald et al. 2015 for detailed description of integration pathways)

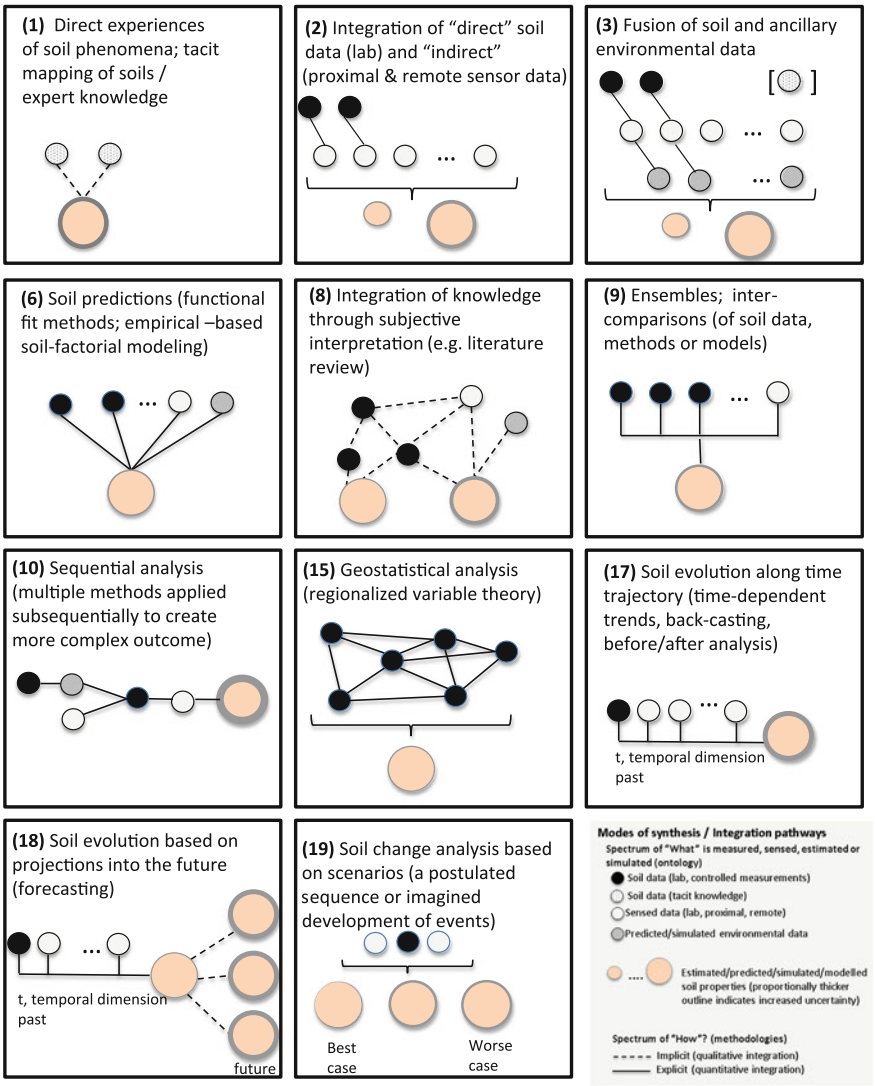

Fig. 14.3 Integration pathways as codified by Grunwald et al. (2014b) to build a Meta Soil Carbon Model for the Suwannee river basin (Florida portion) based on Chaikaew (2014). In this Meta Soil Carbon Model, integration pathways (*1*), (*2*), (*3*), (*6*), (*8*), (*9*), (*10*), (*15*), (*17*), (*18*), and (*19*) were used (see Grunwald et al. 2015 for detailed description of integration pathways)

14.4.2 Meta Soil Carbon Modeling *Focused on Ecosystem Services in a Basin in North-Central Florida* (Fig. 14.3)

Historical and current SOC stocks were assessed using ordinary and block kriging, and SOC sequestration and change analysis incorporated uncertainty assessment explicitly in the modeling process. Actual SOC stocks were predicted from a large set of soil-environmental covariates (n: 172) and RF and regression kriging (Chaikaew 2014). Random forest is an ensemble regression method that allows modeling high-order, complex relationships between soil-environmental covariates and SOC stocks. Actual and attainable terrestrial carbon was assessed using a conjoint RF-simulated annealing approach that allows quantifying the adaptability to imposed external environmental change through three scenarios (Chaikaew et al. 2014). The quantitative SOC stock assessment was integrated with findings from a survey that incorporated a choice experiment to assess the perceptions, values, and beliefs of residents in the basin using a Bayesian belief network (BBN) to model three ecosystem services, of which SOC sequestration was one service (Chaikaew 2014). The BBN modeling entailed four scenarios projecting future states of the ecosystem (BU, Business as Usual; GP, Go toward Projection; GE, Gain Economic value; GA, Go with Environmental Awareness) (Chaikaew 2014).

14.5 Final Remarks

Soil carbon can be viewed through various lenses—from above (through remote and/or proximal sensing), below (a soil pit or petri dish in the laboratory), or sideways (i.e., in new ways integrating multiple approaches). DSM and modeling is shifted into a new phase that is pluralistic in nature embracing a multiplicity of pathways focused to integrate data, methods, and knowledge and to understand about soils and ecosystems. In that sense, it is becoming more and more inter- and transdisciplinary, and through multiple comparisons, adaptations and validations, more robust, reliable and useful.

We are in the process to extend the *Meta Soil Carbon Model* approach to select regions in Southeast Asia and Latin America focused on fusion of laboratory, ground, and remotely sensed spectral datasets. We envision a *global Meta Soil Carbon Model* to emerge that combines spectral and DSM and modeling techniques.

Acknowledgements This study was supported by United States Department of Agriculture (USDA)—Cooperative State Research, Education and Extension Service (CSREES)—National Research Initiative (NRI) grant award 2007-35107-18368 'Rapid Assessment and Trajectory Modeling of Changes in Soil Carbon across a Southeastern Landscape' (National Institute of Food and Agriculture (NIFA)—Agriculture and Food Research Initiative (AFRI)). This project is a core project of the North American Carbon Program. In addition, funding for this research was provided from various projects including grant award 68-3A75-4-73 Mod 2 'Linking Experimental

and Soil Spectral Sensing for Prediction of Soil Carbon Pools and Carbon Sequestration at Landscape Scales' (Cooperative Ecosystem Studies Unit, Natural Resources Conservation Service (NRCS), US Department of Agriculture (USDA)). And grant award 68-7482-11-532 'US Soil Carbon Assessment' funded by NRCS-USDA.

References

Ahn M-Y, Zimmerman AR, Comerford NB, Sickman JO, Grunwald S (2009) Carbon mineralization and labile organic carbon pools in the sandy soils of a North Florida watershed. Ecosystems 12:672–685. doi: 10.1007/s10021-009-9250-8

Bliss CM, Comerford NB, Graetz DA, Grunwald S, Stoppe AM (2013) Land use influence on carbon, nitrogen, and phosphorus in size fractions of sandy surface soils. Soil Sci 178:654–661. doi: 10.1097/SS.0000000000000032

Bouma J, McBratney A (2013) Framing soils as an actor when dealing with wicked environmental problems. Geoderma 200–201:130–139. doi: 10.1016/j.geoderma.2013.02.011

Brown VA, Harris JA, Russel, J (eds) (2010) Tackling wicked problems: through the transdisciplinary imagination. Earthscan Publ., London, UK

Carpenter SR, Armbrust EV, Arzberger PW, Stuart Chapin III, F, Elser JJ, Hackett EJ, Ives AR, Kareiva PM, Leibold MA, Lundberg P (2009) Accelerate synthesis in ecology and environmental sciences. BioScience 59:699–701

Chaikaew P (2014). Assessment of climate regulation, carbon sequestration, and nutrient cycling ecosystem services impacted by multiple stressors. Ph.D. dissertation, University of Florida, Gainesville, Florida, USA.

Chaikaew P, Grunwald S, Xiong X (2014). Estimation of the actual and attainable terrestrial carbon budget. Global Workshop on Digital Soil Mapping, Nanjing, China, Nov. 11-14, 2014.

Egoh B, Reyers B, Rouget M, Richardson DM, Le Maitre DC, van Jaarsveld AS (2008) Mapping ecosystem services for planning and management. Agric Ecosyst Environ 127:135–140. doi: 10.1016/j.agee.2008.03.013

Esbjörn-Hargens S, Zimmerman, ME (2009) Integral ecology: Uniting multiple perspectives on the natural world. Integral Books Publ., Boston, MA

Gomez C, Viscarra Rossel RA, McBratney AB (2008) Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma 146:403–411. doi: 10.1016/j.geoderma.2008.06.011

Grimm R, Behrens T, Märker M, Elsenbeer H (2008) Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma 146:102–113. doi: 10.1016/j.geoderma.2008.05.008

Grunwald S (2009) Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152:195–207. doi: 10.1016/j.geoderma.2009.06.003

Grunwald S (2014) Part I - Conceptualization of a Meta Soil Model. In: Arrouays D, McKenzie N, Hempel J, Richer de Forges AC, McBratney AB (eds) Global Soil Map - Basis of the Global Spatial Soil Inference System. Taylor & Francis Publ, New York, NY, pp 233–238

Grunwald S (2013) Part I – Conceptualization of a Meta Soil Model. Global Soil Map Conference, Orleans, France, Oct. 7-11. 2013.

Grunwald S., G.M. Vasques and R.G. Rivero. 2015. Fusion of soil and remote sensing data to model soil properties. In: Sparks, D.L. (Ed.), Advances in Agronomy, Vol. 131, pp. 1–109.

Grunwald S, Cao B, Xiong X, Ross CW, Patarasuk R, Hempel J, West LT, Andrews SS, Wills S, Loecke TD (2014a) Part II - Integration of data to work towards a Meta Soil Carbon Model in the U.S. In: Arrouays D, McKenzie N, Hempel J, Richer de Forges AC, McBratney AB (eds) Global Soil Map - Basis of the Global Spatial Soil Inference System. Taylor & Francis Publ, New York, NY, pp 239–244

Grunwald S, Thompson JA, Boettinger JL (2011) Digital soil mapping and modeling at continental scales: Finding solutions for global issues. Soil Sci Soc Am J 75:1201–1213. doi: 10.2136/sssaj2011.0025

Grunwald S, Vasques GM, Rivero RG (2014b) Fusion of soil and remote sensing data to model soil properties. Adv. Agron. J. (in press)

Guo Y, Gong P, Amundson R, Yu Q (2006) Analysis of factors controlling soil carbon in the conterminous United States. Soil Sci Soc Am J 70:601. doi: 10.2136/sssaj2005.0163

Jones C, McConnell C, Coleman K, Cox P, Falloon P, Jenkinson D, Powlson D (2005) Global climate change and soil carbon stocks; predictions from two contrasting models for the turnover of organic carbon in soil. Glob Change Biol 11:154–166. doi: 10.1111/j.1365-2486.2004.00885.x

Kelly RH, Parton WJ, Crocker GJ, Graced PR, Klír J, Körschens M, Poulton PR, Richter DD (1997) Simulating trends in soil organic carbon in long-term experiments using the century model. Geoderma 81:75–90. doi: 10.1016/S0016-7061(97)00082-7

Kwon HJ, Grunwald S (2014) Inverse modeling to link carbon pools in CENTURY with measured soil properties. Soil Sci (in review)

Knox NM, Grunwald S, McDowell ML, Bruland GL, Myers DB, Harris WG (2015) Modelling soil carbon fractions with VNIR and MIR spectroscopy. Geoderma 239-240: 229–239

Ladoni M, Bahrami HA, Alavipanah SK, Norouzi AA (2010) Estimating soil organic carbon from soil reflectance: a review. Precis Agric 11:82–99. doi: 10.1007/s11119-009-9123-3

Lobell DB, Burke MB, Tebaldi C, Mastrandrea MD, Falcon WP, Naylor RL (2008) Prioritizing climate change adaptation needs for food security in 2030. Science 319:607–610. doi: 10.1126/science.1152339

Manzoni S, Porporato A (2009) Soil carbon and nitrogen mineralization: Theory and models across scales. Soil Biol Biochem 41:1355–1379. doi: 10.1016/j.soilbio.2009.02.031

Matthews R (2006) The People and Landscape Model (PALM): Towards full integration of human decision-making and biophysical simulation models. Ecol Model 194:329–343. doi: 10.1016/j.ecolmodel.2005.10.032

Matthews RB, Gilbert NG, Roach A, Polhill JG, Gotts NM (2007) Agent-based land-use models: a review of applications. Landsc Ecol 22:1447–1459. doi: 10.1007/s10980-007-9135-1

McBratney AB, Mendonça Santos M., Minasny B (2003) On digital soil mapping. Geoderma 117:3–52. doi: 10.1016/S0016-7061(03)00223-4

McBratney A, Field DJ, Koch A (2014) The dimensions of soil security. Geoderma 213:203–213. doi: 10.1016/j.geoderma.2013.08.013

Minasny B, Tranter G, McBratney AB, Brough DM, Murphy BW (2009) Regional transferability of mid-infrared diffuse reflectance spectroscopic prediction for soil chemical properties. Geoderma 153:155–162. doi: 10.1016/j.geoderma.2009.07.021

Moss S, Pahl-Wostl C, Downing T (2001) Agent-based integrated assessment modelling: the example of climate change. Integr Assess 2:17–30. doi: 10.1023/A:1011527523183

Nelson E, Mendoza G, Regetz J, Polasky S, Tallis H, Cameron Dr, Chan KM, Daily GC, Goldstein J, Kareiva PM, Lonsdorf E, Naidoo R, Ricketts TH, Shaw Mr (2009) Modeling multiple ecosystem services, biodiversity conservation, commodity production, and tradeoffs at landscape scales. Front Ecol Environ 7:4–11. doi: 10.1890/080023

Peters DPC (2010) Accessible ecology: synthesis of the long, deep, and broad. Trends Ecol Evol 25:592–601

Pickett STA, Kolasa J, Jones CG (2007) Ecological understanding: the nature of theory and the theory of nature. Academic Press Publ., Waltham, MA

Ross CW, Grunwald S, Myers DB (2013) Spatiotemporal modeling of soil organic carbon stocks across a subtropical region. Sci Total Environ 461–462:149–157. doi: 10.1016/j.scitotenv.2013.04.070

Schreinemachers P, Berger T (2011) An agent-based simulation model of human–environment interactions in agricultural systems. Environ Model Softw 26:845–859. doi: 10.1016/j.envsoft.2011.02.004

Vasques GM, Grunwald S, Harris WG (2010a) Spectroscopic models of soil organic carbon in Florida, USA. J Environ Qual 39:923. doi: 10.2134/jeq2009.0314

Vasques GM, Grunwald S, Myers DB (2012a) Associations between soil carbon and ecological landscape variables at escalating spatial scales in Florida, USA. Landsc Ecol 27:355–367. doi: 10.1007/s10980-011-9702-3

Vasques GM, Grunwald S, Myers DB (2012b) Influence of the spatial extent and resolution of input data on soil carbon models in Florida, USA. J Geophys Res Biogeosciences 117:n/a–n/a. doi: 10.1029/2012JG001982

Vasques GM, Grunwald S, Sickman JO (2008) Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma 146:14–25. doi: 10.1016/j.geoderma.2008.04.007

Vasques GM, Grunwald S, Sickman JO (2009) Modeling of soil organic carbon fractions using visible–near-infrared spectroscopy. Soil Sci Soc Am J 73:176. doi: 10.2136/sssaj2008.0015

Vasques GM, Grunwald S, Sickman JO, Comerford NB (2010b) Upscaling of dynamic soil organic carbon pools in a north-central Florida watershed. Soil Sci Soc Am J 74:870–879. doi: 10.2136/sssaj2009.0242

Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B. (2014a) Holistic environmental soil-landscape modeling of soil organic carbon. Environ Model Softw 57:202–215.

Xiong X, Grunwald S, Myers DB, Ross CW, Harris WG, Comerford NB (2014b) Interaction effects of climate and land use/land cover change on soil organic carbon sequestration. Sci Total Environ 493:974–982. doi: 10.1016/j.scitotenv.2014.06.088

Chapter 15
Example of Bayesian Uncertainty for Digital Soil Mapping

Laura Poggio, Alessandro Gimona, Luigi Spezia and Mark J. Brewer

Abstract Any model for digital soil mapping suffers from different types of errors, including interpolation errors, so it is important to quantify the uncertainty associated with the maps produced. The most common approach is some form of regression kriging (RK) or variation involving geostatistical simulation. Another way of assessing the spatial uncertainty lies in the Bayesian approach where the uncertainty in the results is described by the posterior density. The aim of this paper is to present an example of a Bayesian approach for uncertainty estimation when mapping the topsoil organic matter content in the Grampian region of Scotland (UK, about 12,100 km^2). The chosen approach uses (Bayesian) latent Gaussian models fitted using integrated nested Laplace approximation (INLA) and the stochastic partial differential equation (SPDE) models approach for coping with spatial correlation (INLA_SPDE). For practical comparison purposes, the results of INLA_SPDE were compared with the results of an extension of the scorpan kriging approach, i.e., (1) combining generalized additive models (GAM) with Gaussian simulations and (2) traditional RK. The results were assessed using in-sample and out-of-sample measures and compared for distribution similarity, spatial structure reproduction, computational load, and uncertainty ranges. We conclude that the Bayesian framework using INLA offers a viable alternative to existing methods and an improvement over traditional RK.

Keywords Bayesian · Soil organic matter · MODIS · Uncertainty

L. Poggio (✉) · A. Gimona
The James Hutton Institute, Craigiebuckler, AB15 8QH Aberdeen, Scotland, UK
e-mail: laura.poggio@hutton.ac.uk

L. Spezia · M.J. Brewer
Biomathematics and Statistics Scotland, Craigiebuckler, AB15 8QH Aberdeen, Scotland, UK

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_15

15.1 Introduction

Any model for digital soil mapping suffers from different types of errors, including interpolation errors, and it is therefore important to quantify the uncertainty associated with the maps produced. Recent developments in digital soil mapping include methodologies to evaluate and map the spatial uncertainty. The most common approach is some form of regression kriging (RK) or variation involving geostatistical simulation. This represents a frequentist approach with uncertainty calculated from a (large) number of realizations. Another way of assessing the spatial uncertainty lies in the Bayesian approach where the uncertainty in the results is described by the posterior density. Markov chain Monte Carlo (MCMC) algorithms are normally used for Bayesian computation when dealing with complex stochastic systems. MCMC is flexible and able to deal with virtually any type of data and model, but involves computationally and time-intensive simulations, e.g., Gaussian univariate Bayesian spatial regression models (Banerjee et al. 2008). The integrated nested Laplace approximation (INLA; Rue et al. 2009) method has been recently developed as a computationally efficient alternative to MCMC. INLA is designed for latent Gaussian models, a very wide and flexible class of models, including (generalized) linear mixed spatial models. INLA can be combined with the SPDE approach proposed by Lindgren et al. (2011) in order to implement efficient models for spatial point data.

The aim of this paper is to present an example of a Bayesian approach for uncertainty estimation when mapping the topsoil organic matter content in the Grampian region of Scotland (Fig. 15.1a). The chosen approach uses (Bayesian) latent Gaussian models fitted using INLA and the SPDE approach for coping with spatial correlation (INLA_SPDE).

For practical comparison purposes, the results of INLA_SPDE were compared with the results of an extension of the scorpan kriging approach—a geostatistical model combining generalized additive models (GAM; Wood 2006) with Gaussian simulations (GAM+GS; Poggio and Gimona 2014).

15.2 Data

15.2.1 Test Area

The Grampian region of Scotland (UK, about 12,100 km^2) covers the whole of NE Scotland (Fig. 15.1a) and has a variety of landscapes and soils. It includes large river catchments and the Cairngorm mountains, with some of the highest peaks in Scotland.

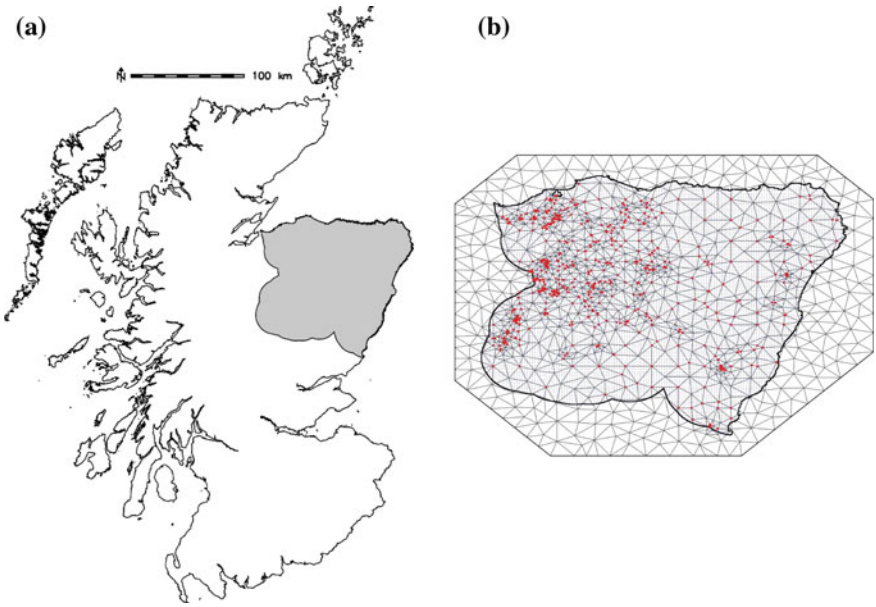

Fig. 15.1 Test area location and selected mesh for the test area. **a** Grampian region and **b** INLA_SPDE mesh

15.2.2 Response Variable

A total of 1183 profiles derived from the National Soil Inventory of Scotland (NSIS, Lilly et al. 2010) were available in the selected test area. The samples were collected on a regular 10-km grid. Only the topsoil values were used. In order to provide validation of the models, the data available were randomly split into training and validation sets with a ratio of 75:25. In this study, the soil property considered was proportion of soil organic matter and the data were transformed using the logit transformation. The data were divided into three more homogeneous groups derived from the database: organic, intermediate, and mineral (Fig. 15.2). The groups were considered separate (i.e., having different means) according to Tukey's honest significant differences (Miller 1981).

15.2.3 Covariates

The freely and globally available covariates included were selected to describe the most important scorpan factors, namely topography; vegetation; climate; time; and geographic position.

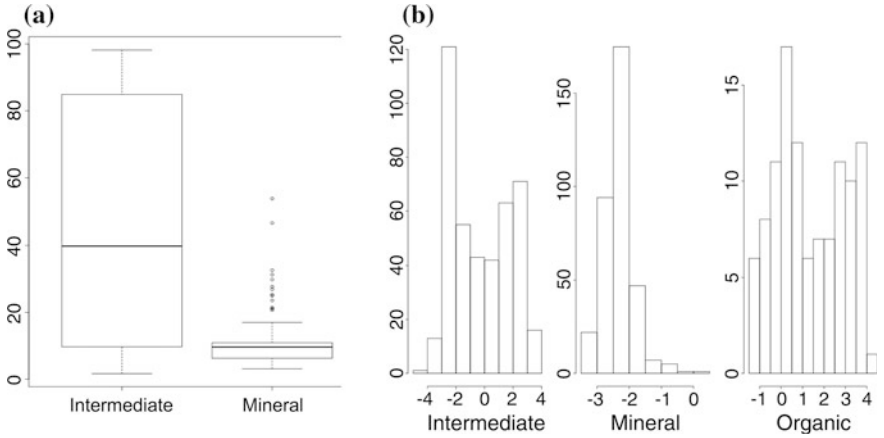

Fig. 15.2 Soil organic groups in the input data. **a** Boxplots with the proportion of soil organic matter on the *y*-axis and **b** histograms with the frequency of logit proportions on the *y*-axis and the logit proportions of soil organic matter are on the *x*-axis

Morphological features

The digital elevation model (DEM) used as a covariate in the fitted models was Shuttle Radar Topography Mission (SRTM). SRTM has a spatial resolution of 90 m with global coverage, and it was further processed to fill in no-data voids (Jarvis et al. 2006; Rodriguez et al. 2006). The measures used were as follows:

- altitude (meters);
- slope: steepest slope angle, calculated using the D8 method (O'Callaghan and Mark 1984).

In order to match the resolution of the other covariates, the medians in each grid cell of 1 × 1 km were used.

MODIS

The Terra Moderate Resolution Imaging Spectroradiometer (MODIS) 8 and 16 day composite products were used to derive a set of indices selected for their capability to differentiate spectral responses from different bare soils, vegetation cover and mixed situations. The 12-year (2000–2012) time series of data was acquired from the NASA FTP Web site (ftp://e4ftl01u.ecs.nasa.gov/MOLT/). The single images were processed to fill the cloud gaps (Poggio et al. 2012), and finally, the medians

over the 12 considered years were used as covariates. The indices calculated were as follows:

1. Enhanced Vegetation Index (EVI; Huete et al. 2002)
2. Normalized difference water index (NDWI; Gao 1996) calculated with NIR (near infrared) and 2130 short-wave infrared (SWIR) band (Gu et al. 2008):

$$\text{NDWI} = \frac{\text{NIR} - \text{SWIR}}{\text{NIR} + \text{SWIR}} \tag{15.1}$$

15.3 Methods

This study used Bayesian latent Gaussian models, fitted using INLA and with the SPDE model approach for coping with spatial correlation (INLA_SPDE). Below, Eq. (15.2) describes the model used for INLA_SPDE in this study in R-like syntax (R Core Team 2013).

The approach involved the following: (1) SPDE (Lindgren et al. 2011) to model the spatial structure of the data and (2) INLA (Rue et al. 2009) to model the observed data with the support of the relevant covariates. Our analysis adapted the tutorial described in Blangiardo et al. (2013) for spatial data. The model is specified like so

$$\begin{aligned}\text{SOM} \sim \text{Intercept} + \text{Random effects} + f(\text{covariates}, \text{model} = \text{RW2}) \\ + f(\text{spatial effect}, \text{model} = \text{spde})\end{aligned} \tag{15.2}$$

The spatial structure is modeled using a mesh, i.e., the corresponding finite element representation of a continuously indexed spatial random field with piecewise linear basis functions over a triangulated mesh (Lindgren et al. 2011). The mesh is created performing a constrained refined Delaunay triangulation for a set of spatial locations. The triangle vertices are placed at the observation locations, and then, further vertices are added to satisfy triangulation quality constraints. The mesh can be adjusted with various parameters: (1) offset: defines how much the considered domain should be extended within and outside the borders of the test area; (2) maximum edge; (3) minimum angle: set the triangle structure; and (4) cutoff: set the minimum allowed distance between points. Depending on the values chosen for the mesh arguments, the total number of vertices changes with a trade-off between the accuracy of the spatial representation and the computational costs. The net result is that the mesh generated forms a network for a Markov spatial correlation model, which is more flexible than (say) a grid structure as the mesh adapts correctly to different densities of points in different regions of space. In this example, the prior were non-informative.

The results of INLA_SPDE were compared with the results of an extension of the scorpan kriging approach, a geostatistical model combining generalized additive

Table 15.1 Differences between INLA_SPDE and GAM+GS

	INLA_SPDE	GAM+GS
Approach	Bayesian	Frequentist
Predictor	Conditional autoregressive with latent model RW2 (Random Walk)	GAM
Spatial term structure	Mesh	Variogram of the residuals
Spatial model	SPDE	Gaussian simulations
Random effects	Three groups of decreasingly organic soils (organic, intermediate, and mineral)	Not used

models (GAM; Wood 2006) with Gaussian simulations (GAM+GS; Poggio and Gimona 2014).

The GAM+GS approach involves the following: (1) modeling the trend with full spatial correlation and (2) Gaussian simulations to interpolate the residuals. The values at each cell were defined using a hybrid GAM–geostatistical model, combining the fitting of a GAM to estimate the trend of the variable, using a smoother with related covariates, and Gaussian simulations of the model residuals as a spatial component to account for local detail.

Table 15.1 presents a summary comparison between the chosen INLA_SPDE approach and the comparison GAM+GS method.

Finally, traditional RK with linear models (RK_LM; Hengl et al. 2004) and regression trees (RK_RT; e.g., Kheir et al. 2010) were also performed, to enable further comparison for the median values only.

15.3.1 Validation

The results of the approaches are assessed using in-sample and out-of-sample measures and compared for distribution similarity, spatial structure reproduction, computational load, and uncertainty ranges. In particular:

1. Root mean square error (RMSE),
2. R^2 derived from a linear model between observed and modeled data R^2LM, and
3. Standardized squared prediction errors Θ

$\Theta_{(x)}$ was taken as an index suggested by Lark (2000) to check that the used model is a valid representation of the spatial variation of the property:

$$\Theta_{(x)} = \frac{\left(\hat{z}_{(x)} - z_{(x)}\right)^2}{\sigma^2_{(x)}} \tag{15.3}$$

$z_{(x)}$ is the measured value and $\hat{z}_{(x)}$ is the estimated value and $\sigma^2_{(x)}$ is its estimated variance. A mean of $\Theta_{(x)}$ close to 1 and a median close to 0.455 indicate a good fit and ensure an unbiased variance for the kriging (Lark 2000), respectively.

15.3.2 Software Used

The analyses were performed using open source software:

1. GRASS GIS (GRASS Development Team 2014) for data management, preparation, and visualization;
2. the R software (R Core Team 2013). The following packages were used: (a) `raster` for data management, preparation, and visualization (Hijmans and van Etten 2013); (b) `mgcv` for GAM (Wood 2006); (c) `geoR` (Ribeiro and Diggle 2001) for fitting the variograms of the residuals; (d) `gstat` (Pebesma 2004) for kriging; (e) `rgdal` for data management (Keitt et al. 2009); (f) `R-INLA` for INLA_SPDE approach (Rue et al. 2014). (g) `randomForest` (Liaw and Wiener 2002) for regression trees; and (h) `GSIF` (Hengl 2014) for RK.

15.4 Results

The results obtained with INLA_SPDE were compared with out-of-sample assessment and GAM+GS as an example of a more traditional scorpan kriging approach. Table 15.2 presents the main summary parameters. The two approaches give similar results. However, they both showed a considerable improvement compared to RK. The relationships between the variable of interest and the covariates were not linear. Therefore, as expected, methods using splines and nonlinear approaches perform better.

INLA_SPDE reproduces more closely the values of the soil property, while GAM+GS reproduces more closely the spatial structure and the spatial variability (Θ median closer to 0.445 and Fig. 15.3b). The quantile–quantile plots in Fig. 15.3a confirmed that both approaches provide similar results with INLA_SPDE slightly closer to the 1-to-1 line. The maps obtained are presented in Fig. 15.3. The spatial

Table 15.2 Assessment of the results: INLA_SPDE and GAM+GS with RK comparison (using both LM and RF approaches)

	INLA_SPDE	GAM+GS	RK_LM	RK_RF
RMSE	25.41	26.63	39.79	39.99
R^2LM	0.39	0.38	0.09	0.11
Θ (mean)	1.02	0.96	2.86	2.69
Θ (median)	0.27	0.36	1.84	1.69

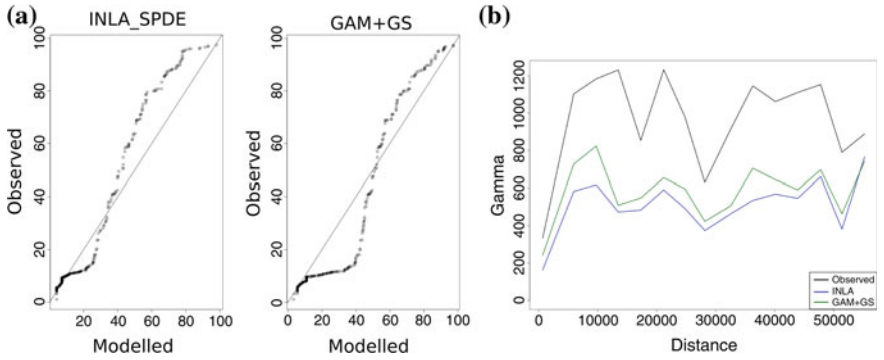

Fig. 15.3 Assessment of the results: INLA_SPDE and GAM+GS. **a** Qq-plots and **b** variograms

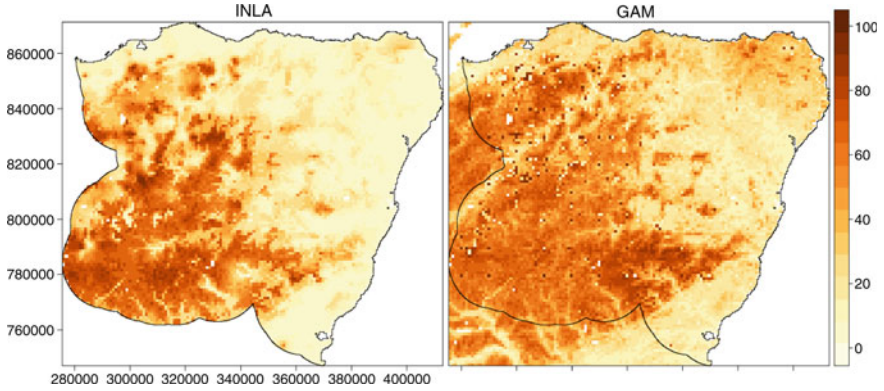

Fig. 15.4 Maps: INLA_SPDE and GAM+GS

patterns are rather different, with GAM+GS predicting larger areas with higher organic matter content (Fig. 15.4).

Comparing the results from the distribution (Fig. 15.5), the two approaches have very similar results for the mean and the median. However, they tend to diverge when considering percentiles. This is also confirmed by the higher values of GAM +GS for the percentage of the validation set that is outside the considered confidence intervals at different percentiles (Table 15.3). GAM+GS was run with only 500 simulations. A higher number of simulations could result in more validation values within the confidence intervals.

INLA_SPDE approach is rather sensitive to the chosen mesh. Figure 15.6 shows the ratio of the different results obtained with different meshes, while Table 15.4 summarizes the validation parameters, i.e., RMSE and R^2. The results are most sensitive to choice of the cutoff parameter, which determines the minimum distance between points; in the current example, choosing a value of 1000 would seem to be too large, as it does not allow the capture of smaller-scale autocorrelation. Note that

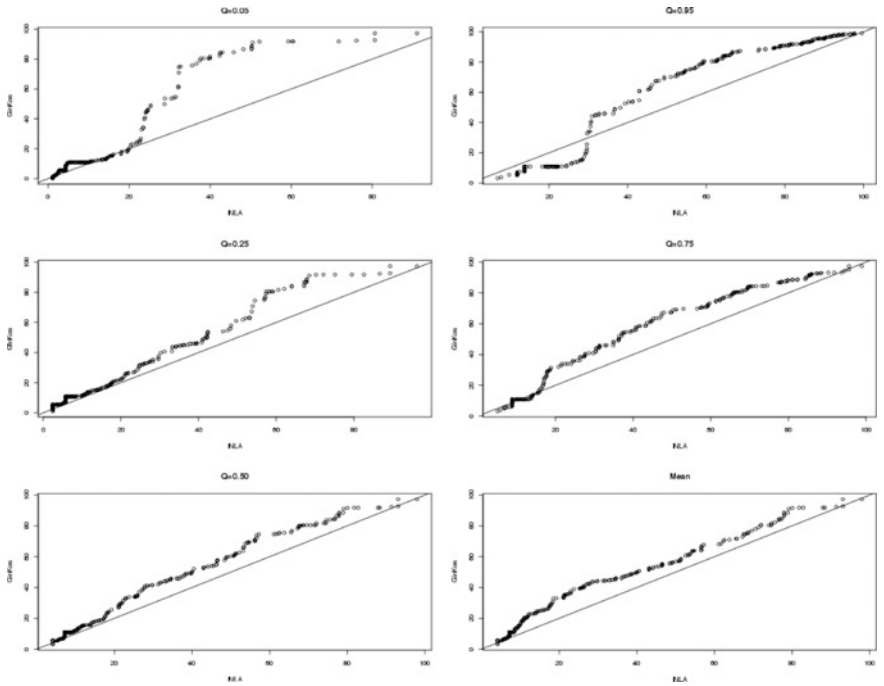

Fig. 15.5 Comparison of percentiles between INLA_SPDE and GAM

Table 15.3 Proportion of the validation set values outside CI credible intervals (for INLA_SPDE) and confidence intervals (for GAM+GS) at different percentiles: CI90 = between 5 and 95; CI80 = between 10 and 90; CI50 = between 25 and 75

	INLA_SPDE	GAM+GS
CI90	0.21	0.38
CI80	0.33	0.33
CI50	0.56	0.65

the process of choosing the right mesh is analogous to the process of choosing the right variogram.

GAM+GS featured better performances when a covariate with the membership of one of the three groups (i.e., organic, intermediate, and mineral) was used. The explained deviance increased to 47.3 % from 41.7 %, and the AIC decreased to 3214 from 3220. However, in order to obtain a full spatial prediction, it is necessary to have a map indicating which pixels belong to which group. This map is not needed when using INLA_SPDE, but it would improve model performance. The map to derive the three groups could be derived from a traditional soil map (e.g., reclassification of soil groups) providing an important synergy with legacy soil data.

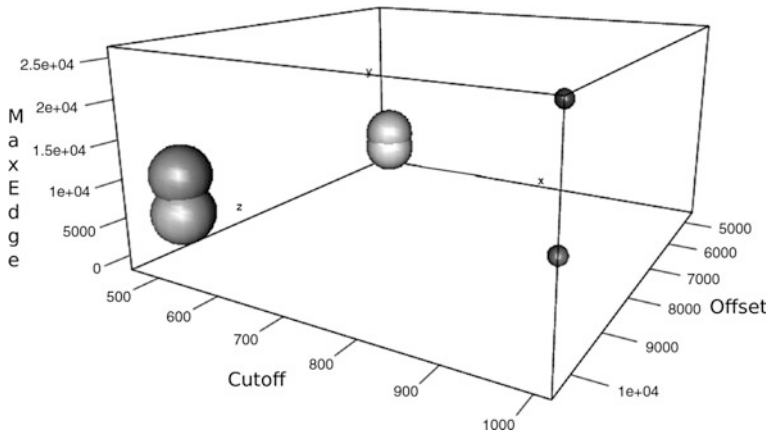

Fig. 15.6 Validation: mesh parameters. The size of the spheres is proportional to the goodness-of-fit measure, i.e., *larger sphere* indicates a better validation result

Table 15.4 Validation: mesh parameters for INLA_SPDE

	Cutoff	Offset	Maxedge	R^2	RMSE
2	500	10,000	10,000	0.41	25.07
3	1000	10,000	10,000	0.03	38.71
4	1000	10,000	25,000	0.02	39.38
5	500	10,000	5000	0.43	25.75
6	500	5000	5000	0.43	25.87
7	500	5000	2500	0.44	25.84

The computational load is lower in INLA_SPDE: 90 s versus 4 min (for 500 simulations of GAM+GS) for about 12,000 pixels. GAM+GS could be further optimized, but it is likely that more simulations are needed to fully characterize the distribution. Finally, as INLA_SPDE is a full Bayesian approach, prior information and soft data can be integrated into the analysis.

15.5 Discussion, Conclusions, and Future Work

Often, the interest of a statistical analysis is estimating the effect of a set of relevant covariates on the observed data, while accounting for the spatial correlation implied in the model. There are several advantages to the Bayesian approach, mainly (1) the specification of prior distributions allows the formal inclusion of information that can be obtained through legacy data or from expert opinion and (2) it is relatively easy to specify a hierarchical structure on the data and/or parameters, which in turn makes prediction for new observations and missing data imputation relatively straightforward.

Table 15.5 Summary of the main differences in results and implementation between INLA_SPDE and GAM+GS

	INLA_SPDE	GAM + GS
Value reproduction	+	–
Spatial variability reproduction	–	+
Spatial modeling (e.g., anisotropy)	–	+
Uncertainty	+	+
Distribution	+	–
Possibility to include priors	+	–
Computation	+	–
Simplicity of implementation	–	+

+ indicates an advantage for the considered approach. This table is built with the information from the considered test case and some expert opinion

In this paper, we presented an example of the use of the INLA_SPDE approach for DSM and we compared it with a more traditional approach derived from the family of scorpan kriging, such as GAM+GS. Table 15.5 summarizes the main differences in results and implementation of the two approaches. Further work is needed to apply INLA_SPDE to 3D modeling to take into account the vertical as well as the lateral variability of soil properties. INLA_SPDE is also suitable for integration of information derived from legacy soil data and maps to be used as prior information.

INLA_SPDE proved to be an interesting framework comparable with an approach such as GAM+GS and a considerable improvement compared to traditional RK. The main advantages are the possibility to include soft data, the computational load, and the possibility to use a subset of covariates for prediction, i.e., to build the model with covariates only available at points' locations. However, its implementation and the spatial modeling are still more complex than a scorpan kriging-like approach and some aspects of spatial modeling (e.g., anisotropy) are not yet fully included.

Acknowledgements This work was funded by the Scottish Government's Rural and Environment Science and Analytical Services division. Many thanks to The James Hutton Institute teams which sampled and analyzed the soils in past years and set up the database. MODIS data are distributed by the Land Processes Distributed Active Archive Centre (LP DAAC), located at the US Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center (lpdaac.usgs.gov). Thanks are due to Prof. David Elston (BioSS) for comments on an early version of this manuscript and to Jane Morrice for proofreading.

References

Banerjee S, Gelfand A, Finley A, Sang H (2008) Gaussian predictive process models for large spatial datasets. J R Stat Soc Series B 70:825–848

Blangiardo M, Cameletti M, Baio G, Rue H (2013) Spatial and spatio-temporal models with r-inla. Spatial and Spatio-temporal Epidemiol 7:39–55

Gao B (1996) NDWI - a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens Environ 58:257– 66

GRASS Development Team (2014) Geographic Resources Analysis Support System (GRASS GIS) Software, version 6.5.0svn. http://www.grass.osgeo.org

Gu Y, Hunt E, Wardlow B, Basara JB, Brown JF, Verdin JP (2008) Evaluation of MODIS NDVI and NDWI for vegetation drought monitoring using Oklahoma Mesonet soil moisture data. Geophys Res Lett 35(22):L22401

Hengl T (2014) GSIF: Global Soil Information Facilities. R package version 0.4-1. http://CRAN.R-project.org/package=GSIF

Hengl T, Heuvelink G, Stein A (2004) A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 122 (1-2):75–93

Hijmans RJ, van Etten J (2013) raster: raster: Geographic data analysis and modeling. R package version 2.1-25. http://CRAN.R-project.org/package=raster

Huete A, Didan K, Miura T, Rodriguez EP, Gao X, Ferreira LG (2002) Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens Environ 83(1-2):195–213.

Jarvis A, Reuter H, Nelson A, Guevara E (2006) Hole-filled seamless SRTM data V3. Technical report International Centre for Tropical Agriculture (CIAT).

Keitt T, Bivand R, Pebesma E, Rowlingson B (2009) rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 0.6-21. http://CRAN.R-project.org/package=rgdal

Kheir RB, Greve MH, Bocher PK, Greve MB, Larsen R, McCloy K (2010) Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. J Environ Manage 91(5):1150–1160

Lark M (2000) A comparison of some robust estimators of the variogram for use in soil survey. Eur J Soil Sci 51:137–157

Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2 (3):18–22. http://CRAN.R-project.org/doc/Rnews/

Lilly A, Bell J, Hudson G, Nolan A, Towers W (2010) National Soil Inventory of Scotland 1 (NSIS1): site location, sampling and profile description. (1978-1998). Technical report Macaulay Institute.

Lindgren F, Rue H, Lindstrom J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach (with discussion). J R Stat Soc Series B 73 (4):423–498

Miller RG (1981) Simultaneous Statistical Inference. Springer.

O'Callaghan JF, Mark DM (1984) The extraction of drainage networks from digital elevation data. Comput Vision Graph 28:323–344

Pebesma E (2004) Multivariable geostatistics in S: the gstat package. Comput Geosci UK 30:683–691

Poggio L, Gimona A (2014) National scale 3D modelling of soil organic carbon stocks with uncertainty propagation - An example from Scotland. Geoderma 232-234:284–299

Poggio L, Gimona A, Brown I (2012) Spatio-temporal MODIS EVI gap filling under cloud cover: an example in Scotland. ISPRS J Photogramm Remote Sens 72:56–72

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org/

Ribeiro P, Diggle P (2001) geoR: a package for geostatistical analysis. R-NEWS 1 (2), 14–18. http://CRAN.R-project.org/doc/Rnews/

Rodriguez E, Morris C, Belz J, Chapin E, Martin J, Daffer W, Hensley S (2006) An assessment of the SRTM topographic products. Technical report. JPL D-31639, NASA-Jet Propulsion Laboratory.

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). J R Stat Soc Series B. 71 (2):319–392

Rue H, Martino S, Lindgren F, Simpson D, Riebler A, Krainski ET (2014) INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximation. R package version 0.0-1410248378

Wood S (2006) Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC Press

Chapter 16
An Unsupervised Fuzzy Clustering Approach for the Digital Mapping of Soil Organic Carbon in a Montaneous Region of China

Lei Zhu, Jiandong Sheng, Hongtao Jia and Hongqi Wu

Abstract Spatial distribution of soil attributes is the basic information required for land surface process simulating and ecological modeling. Purposive sampling method based on typical points which employed environmental factors has been widely used in digital soil mapping (DSM) to acquire soil spatial properties at different scales. Clustering analysis of soil environmental covariates was performed to explore for sampling points representative of different grades of soil spatial distribution and to formulate a sampling designing method based on representativeness grade. This method was used to predict soil organic matter (SOM) content in the surface layer of grassland soil within a 4 km^2 area of the Bayanbulak District, Xinjiang Uyghur Autonomous Region. Six terrain factors, including elevation, slope, aspect, planform curvature, profile curvature, and topographic wetness index, were clustered by fuzzy c-means method. Fuzzy membership distribution of 9 groups of environmental factors was derived to position 18 soil samples in the area with membership larger than 0.9. Then, SOM map was predicted with fuzzy membership model. Finally, 35 individual soil samples (16 regular sampling points, 9 cross-sectional sampling points, and 10 sampling points according to altitude) were collected as the verify point. The results showed that purposive sampling combined with FCM is a low cost and efficiency mapping method with satisfactory prediction precision and model stability and could be possibly applied to small-scale region with the similar landscape conditions.

Keywords Purposive sampling · Fuzzy clustering · Soil organic matter · Bayanbulak

L. Zhu (✉) · J. Sheng · H. Jia · H. Wu
Xinjiang Key Laboratory of Soil and Plant Ecological Processes, College of Grassland and Environmental Sciences, Xinjiang Agricultural University, Urumqi 830052, Xinjiang, China
e-mail: sjd_2004@126.com

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_16

16.1 Introduction

Spatial distribution of soil properties provides essential information for agricultural and environmental management applications (Zhu et al. 2010). Conventional soil survey maps are not only time-consuming, but also unable to meet the requirement of many environmental modeling and land management applications. In recent years, with the development of remote sensing, geographic information systems, artificial intelligence and fuzzy reasoning technology, and digital soil mapping (DSM) have made great progress (Zhu et al. 2001; McBratney et al. 2003; Zhang et al. 2004, 2012; Yang et al. 2007; Sun et al. 2013). Based on fuzzy logic theory, Zhu et al. have established fuzzy inference model, soil land inference model (SoLIM), which has been promoted as a standard technique by US Department of Agriculture (USDA) for soil survey (Zhu et al. 2005). The model which has been applied in different area, such as Wisconsin Raffelson River basin in USA and Heshan Farm of Heilongjiang Province in China, mainly focused on watershed scale and acquired higher soil map accuracy than traditional (Zhu et al. 2008; Yang et al. 2009). However, more case studies are needed to prove the model applicability in small scale.

Bayanbulak alpine grassland is China's second largest prairie grasslands. Research on the distribution characteristics of soil organic carbon in this region and its relationship with environmental factors can provide scientific basis for rehabilitation of degraded rangelands and play the potential of grassland ecosystem better.

In this paper, taking a small area of typical grassland in Bayanbulak, Xinjiang, as an example, fuzzy c-means clustering (FCM) method is used to establish the relationship between terrain factors and surface soil organic matter (SOM) in the study area. Then, we design typical sampling point on the basis of the fuzzy membership degree distribution. Through the indoor test analysis, finally, simulate surface SOM, and validate result accuracy through independent sample.

16.2 Study Area

The study area is located in Bayanbulak Town, Hejing County, Xinjiang Uygur Autonomous Region, China. It has an area of about 4 km^2, with a length of 2.01 km in the east–west direction and a width of 2.00 km in the north–south direction. With the north side higher than the south, the elevation is between 2603 and 2995 m. In this area, the average temperature is 4.8 °C. The absolute winter minimum could reach −48 °C in January, while the absolute summer maximum could climb to 30.5 °C in July. The annual average wind speed is 2.7 m/s, the annual precipitation is 276.2 mm, the annual evaporation range from 1022.9 to 1247.5 mm, the sunshine duration is about 2466–2616 h, the heat energy is 562.8 kJ/cm^2/year, the day of snow lying is normally 150–180 days throughout the year, and there is no absolute

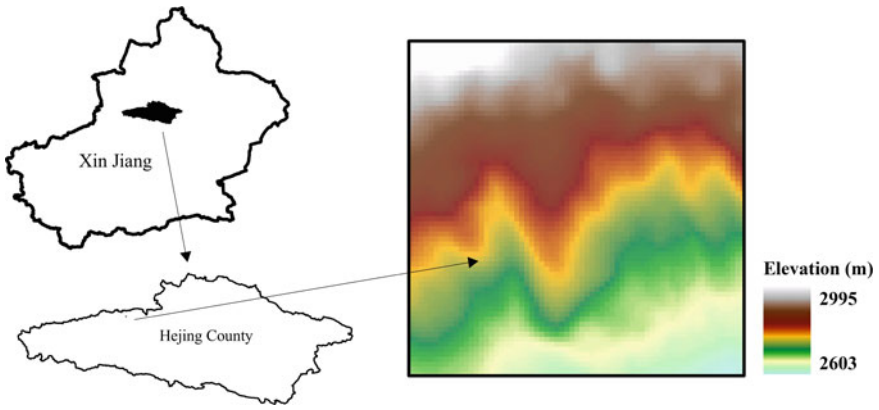

Fig. 16.1 Location map of study area

frost-free period in this area. Totally, it is the typical alpine climate. The soil in the study area is the subalpine steppe soil, which is intensively covered by alpine meadow, mainly including *Stipa purpurea*, *Kobresia capillifolia,* and *Polygonum viviparum*. (Fig. 16.1).

16.3 Methods

16.3.1 Establish Environmental Factors Database

Under landscape scale, the main factors of soil changing are the topographic and hydrological conditions (Yang et al. 2007). Especially in a small study area, topographic features can basically represent the main influencing factors of soil formation and development (McSweeney et al. 1994). Because the study area is just nearly 4 km^2, climate condition can be regarded as uniform. In addition, parameterization of the regional geological age is difficulty and semiquantitative analysis and sample-based methods are difficult to generate raster data with sufficient accuracy, so time factor is not considered in this study. The environmental conditions for the study area were characterized at 30-m resolution and the following environmental variables were used: elevation, slope gradient, slope aspect, profile curvature, planform curvature, and topographic wetness index (Fig. 16.2).

16.3.2 Cluster Combination Based on Fuzzy Logic

Using FCM method (FCM), we get a set of combinations of environmental factors. Fuzzy membership degree (FMD), which is derived from different combination of

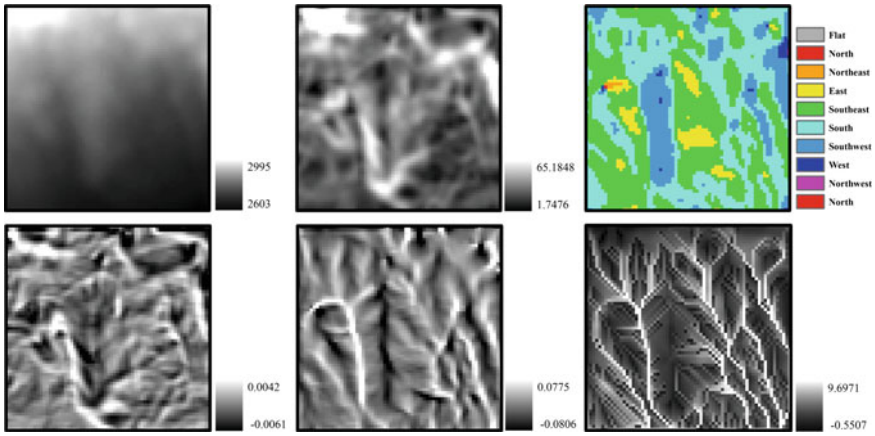

Fig. 16.2 Environmental factors of study area

environmental factors, shows the spatial change of soil properties. And according to FMD, the central area in a combination and the transition zone between different combinations could both be distinguished. Finally, we design one or two typical points with higher membership degree in the center of each environmental factor combination (Yang et al. 2010).

Through partition coefficient and normalized entropy with the change of cluster number, FCM can determine the optimal clustering number (Bezdek et al. 1984; Yang et al. 2007; Zhu et al. 2008). In general, with the increase of the cluster number, if decreasing quantity of partition coefficient from $(c - 1)$ to c is smaller than from c to $(c + 1)$, the clustering result can be considered relatively stable, and the corresponding cluster number is the optimum. In this study, the optimal cluster number is 9 and fuzzy weighted is 2. Membership degree of nine environmental factor combinations were calculated by SoLIM. Figure 16.3 shows class1 and class 5 as examples.

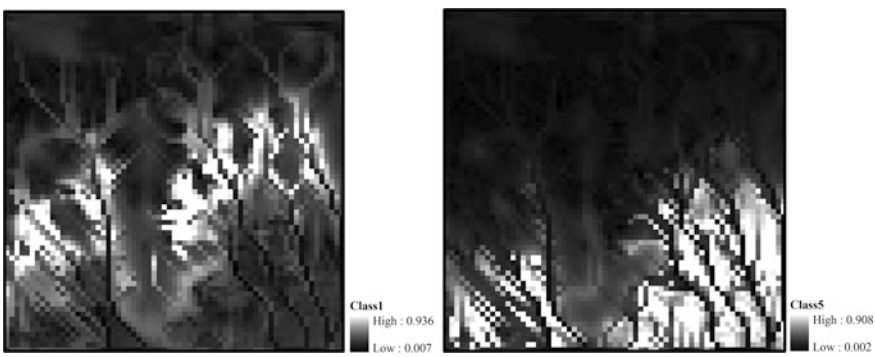

Fig. 16.3 Membership degree of environmental factors combination (class1 and class 5)

16.3.3 Obtain Surface SOM of Typical Points for Each Combination

Based on clustering of six environmental factors in the study area, distribution of fuzzy membership degree class of nine environmental factor combinations was captured (Fig. 16.3). Then, 14 typical points were selected in the center of each environmental factor combination (Yang et al. 2010). Finally, surface soil sample (0–20 cm) was collected for each typical points by field sampling, and SOM of each sample points was obtained through laboratory analysis.

16.3.4 Simulate Surface SOM

We used a fuzzy membership-weighted average model in which the soil property value at a location is the weighted average of the typical soil property values of the prescribed soil types with the weights being the fuzzy membership values (similarity values) (Eq. 16.1) (Zhu et al. 1997).

$$V_{ij} = \frac{\sum_{k=1}^{n} S_{ij}^{k} V^{k}}{\sum_{k=1}^{n} S_{ij}^{k}} \tag{16.1}$$

where V_{ij} is the predicted soil property value at location i, j, S_{ij}^{k} is the fuzzy membership value in soil type k for the soil at the given location, and v^{k} is the typical soil property value for soil type k. This model is based on the assumption that the higher the membership of a local soil in a given soil series, the closer the property values at that location will be to the typical property values of the series (Zhu et al. 2010).

16.4 Results

16.4.1 Simulation of Surface SOM

According to the membership function of the soil and the environment factors, surface (0–20 cm) SOM in the study area was simulated combined with environmental factor databases and experimental data (Fig. 16.4). The results showed that surface SOM in the study area varies continuously and changes with the terrain.

16.4.2 Accuracy Measure

Predicted SOM distribution results were compared with the measured data in order to assess model performances. A total of 35 independent samples were collected in

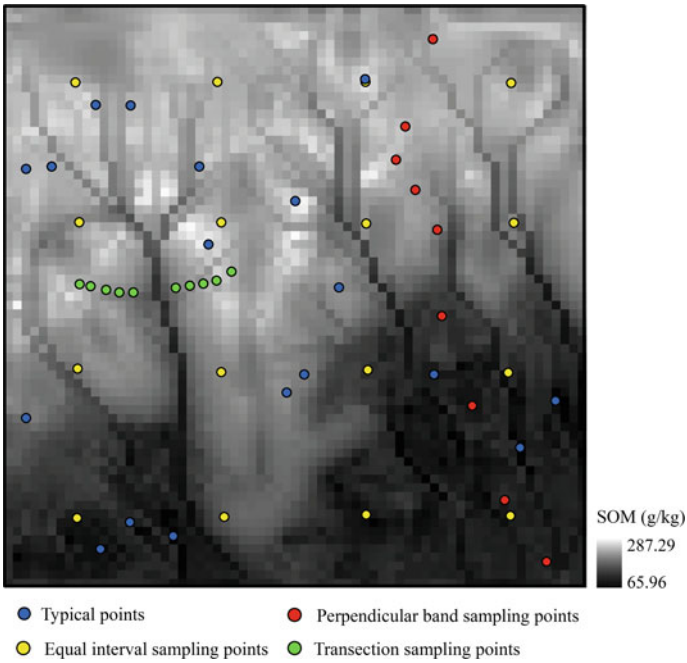

Fig. 16.4 Simulated distribution of surface SOM and distribution of verification points

the study area including 16 equal interval sampling points (500 m) for testing the whole simulation (ZR), 9 transection sampling points designed to cross the path of the hills and valleys and tested whether simulation results can better reflect the characteristics of the spatial gradient of soil properties (ZT), and 10 perpendicular band sampling points (ZG), which was used to test the change of soil properties along the elevation gradient. Mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2) were set up for evaluation of mapping results by using above independently validation points (See Table 16.1 and Fig. 16.5). The results showed that the inference results matched measured value well on the whole. Besides, simulation accuracy was relatively high for locations with a short distance from the cluster center or areas with more significant changes in the terrain.

Based on the environmental similarity, formula (16.2) is calculated to assess the speculated uncertainty caused by the sample representativeness:

$$\text{Uncertainty}_{ij} = 1 - \max\left(S_{ij}^1, S_{ij}^2, \ldots, S_{ij}^n\right) \tag{16.2}$$

In oder to predict the uncertainty, we put the environmental similarity of the predictive points and the center points into formula 16.2. Then we got the spatial distribution of predicted uncertainty, as shown in Fig. 16.6.

Table 16.1 Evaluation results of the surface SOM of study area

	MAE	RMSE
Total	35.26	45.61
ZR	67.36	45.65
ZT	21.85	27.71
ZG	44.00	58.18

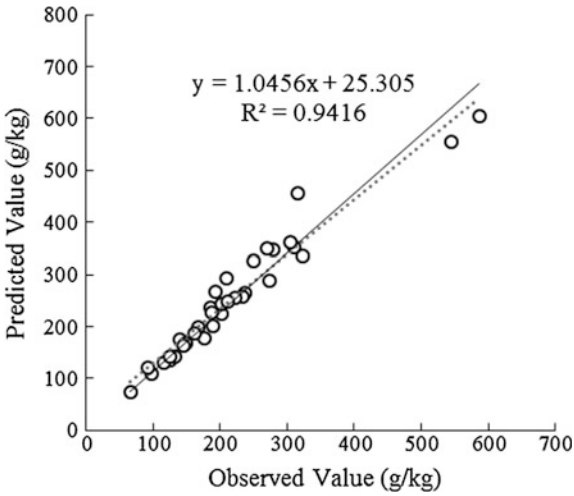

Fig. 16.5 Scatter diagram of validation points

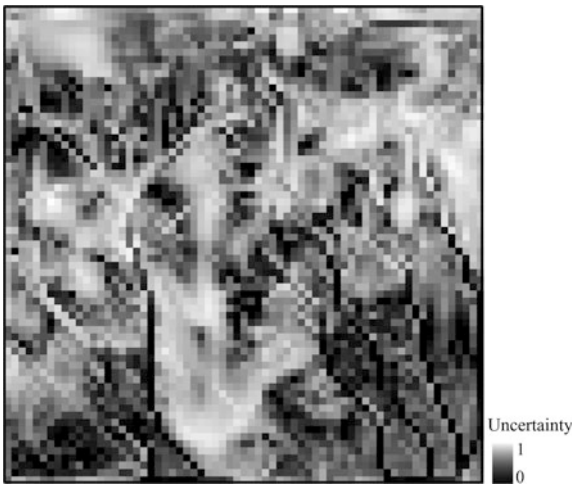

Fig. 16.6 Uncertainty distribution map

16.5 Conlusions

In this study, taking a small area of typical grassland in Bayanbulak, Xinjiang, as an example, FCM method is used to establish the relationship between terrain factors and surface SOM in the study area. Then, we design typical sampling point on the basis of the fuzzy membership degree distribution. Through the laboratory analysis, finally, simulate surface SOM, and validate result accuracy through 35 independent sample points. The results showed that FCM method could rationally and effectively classify the combination of terrain factors, and it is a low cost and efficiency mapping method with satisfactory prediction precision and model stability and could be possibly applied to areas with the similar landscape conditions. Besides, prediction precision was relatively higher for locations were a short distance from the cluster center or areas with more significant changes in the terrain.

Purposive sampling design method based on typical points is efficient for spatial simulation of regional soil properties. The method is also effective for small scale of the study area. In future study, vegetation information (such as the surface of the biomass) should be added to the environmental factor clustering model, which will help improve the quality of model.

Acknowledgements This study is funded by Postdoctoral Foundation of Xinjiang Agricultural University. The authors are grateful to researchers who contribute to the development and maintenance of the SoLIM software.

References

Bezdek JC, Ehrlich R, Full W (1984) FCM: The fuzzy c-means clustering algorithm. Computers and Geosciences 10 (2/3): 191–203.

Mcbratney AB, Mendonca Santos ML, Minasny B (2003) On digital soil mapping. Geoderma 117 (1-2): 3–52.

McSweeney K, Slater BK, Hammer RD, Bell JC, Gessler PE, Petersen GW (1994) Towards a new framework for modeling the soil-landscape continuum. In: Amundson R. (ed.) Factors of Soil Formation: A Fiftieth Anniversary Publication. Madison, Wisconsin: Soil Science Society of America. 127–154.

Sun XL, Zhao YG, Liu F, Wang DC, Liang CP (2013) Digital Soil Mapping and Advance in Research. Chinese Journal of Soil Science 44(3): 752–759.

Yang L, Zhu AX, Li BL, Qin CZ, Pei T, Liu BY, Li RK, Cai QG (2007) Extraction of knowledge about soil-environment relationship for soil mapping using Fuzzy c-Means(FCM) clustering. ACTA PEDOLOGICA SINICA 44(5): 784–791.

Yang L, Zhu AX, Qin CZ, Li BL, Pei T, Liu BY (2009) Soil property mapping using fuzzy membership—a case study of a study area in Heshan Farm of Heilongjiang Province. ACTA PEDOLOGICA SINICA 46(1): 9–15.

Yang L, Zhu AX, Qin CZ, Li BL, Pei T, Qiu WL, Xu ZG (2010) A Purposive Sampling Design Method Based on Typical Points and Its Application in Soil Mapping. PROGRESS IN GEOGRAPHY 29(3): 279–286.

Zhang H, Zhang GL, Gong ZT (2004) The Progress of Quantitative Soil- Landscape Modeling - A Review. Chinese Journal of Soil Science 35(3): 339–346.

Zhang SJ, Zhu AX, Liu J, Yang L (2012) Sample-based Digital Soil Mapping Methods and Related Sampling Schemes. Soils 44(6): 917–923.

Zhu AX (2008) Detail Digital Soil Survey: Models and Methods (in Chinese) Science Press Beijing China. ISBN: 978-7-03-021521-5, 227 p.

Zhu AX, Band LE, Vertessy R, Dutton B (1997) Derivation of soil properties using a soil land inference model (SoLIM). Soil Science Society of America Journal 61: 523–533.

Zhu AX, Hudson B, Burt J, Lubich K, Simonson K (2001) Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of American Journal 65: 1463–1472.

Zhu AX, Li BL, Yang L, Pei T, Qin CZ, Zhang GL, Cai QG, Zhou CH (2005) Predictive soil mapping based on a GIS, expert knowledge, and fuzzy logic framework and its application prospects in China. ACTA PEDOLOGICA SINICA 42(5): 844–851.

Zhu, AX, Qi F, Moore A, Burt JE (2010) Prediction of soil properties using fuzzy membership values. Geodema 158: 199–206.

Chapter 17
Application of Digital Soil Mapping Techniques to Refine Soil Map of Baringo District, Rift Valley Province, Kenya

Rita Juma, Tamás Pőcze, Gábor Kunics and István Sisák

Abstract Detailed and precise description of soil information is important for both developed and developing countries. Africa is highlighted as the most soil data-challenged land surface in the world and it is the area most in need of improved soil information. Our objective was to compile a detailed soilscape class map for the Baringo area in Kenya by using auxiliary variables (digital elevation model, satellite images, and climate maps). In the first step, we extracted landscape–soil relationships based on soil classes from KENSOTER database. We applied soil spatial prediction based on nine standardized predictor variables: x and y coordinates of the sampling points, two principal components from the seven bands of satellite images explaining 83 % of the total variance, three principal components from the 42 variables of climate database explaining 96 % of the total variance, and slope and elevation from digital elevation model. In the first phase (rule extraction), explanatory and target maps were sampled at 999 random points. In the next phase (prediction), 14 major combined soil classes were predicted based on randomly placed 10,000 points. Distances between point values and centroids of the soil classes were calculated, and the closest were scored with 1 and the others with 0. The scores were kriged to obtain continuous probability estimates. Final map was derived based upon the highest probabilities. Our approach had the clear advantage that real-world variability was represented by stacked layers of smooth probability estimates for the soil classes instead of blurred outputs where neighboring pixels can be differently allocated. Our method is suitable to update old and less detailed soil maps or predict new ones for similar environments in the presence of fine resolution auxiliary information. Validity of the prediction should be appropriately tested.

Keywords Limited soil data · Landscape–soil class rule extraction · KENSOTER · Stacked probability layers

R. Juma · T. Pőcze · G. Kunics · I. Sisák (✉)
Georgikon Faculty, Department of Plant Production and Soil Science,
University of Pannonia, 16 Deák Ferenc Street, Keszthely 8360, Hungary
e-mail: talajtan@georgikon.hu

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_17

17.1 Introduction

Detailed and precise soil information is very important for both developed and developing countries. Unfortunately, in many countries this information is not available if so then the existing soil databases are incomplete, not exhaustive or precise enough, and the direct assessment of these resources is therefore constrained by limited spatial data—particularly soil data (Dent and Bai 2008) especially in many tropical countries where limited infrastructure is emphasized (Gonzalez et al. 2008; Hartemink et al. 2008).

The African continent is simultaneously highlighted as the most soil data-challenged land surface in the world and as the area most in need of improved soil information (Eswaran et al. 1997; Palm et al. 2007; Rossiter 2008). Time and cost involved in soil survey are probably one of the reasons for the scarcity of more detailed and updated soil maps since traditional soil survey is known to be very expensive, laborious, and time-consuming (Nachtergaele and Van Ranst 2003). Besides, the number of soil scientists and financial resources is also scarce in these particular regions.

This emphasizes a clear need for quantitative soil information and need for efficient alternative methods to utilize historical and limited soil databases to produce detailed and precise quantitative soil maps of fine resolution at lower costs, in less time and with higher accuracy. One response to this demand is digital soil mapping, where soil maps are produced digitally based on environmental variables (McBratney et al. 2003). In the recent digital era, digital soil mapping plays a more and more important role in this context (Lagacherie and McBratney 2006; Rossiter 2004); as Information Technology continues to improve rapidly, in particular GIS, remote sensing, expert systems, as well as prediction models and digital soil mapping techniques (Cook et al. 2008). Applying digital soil mapping based on existing data emerges as a potential alternative to help to address the increasing demands (Bacic 2008) and the limitations.

In Kenya, majority of soil maps were prepared by conventional methods with soil information commonly available in reconnaissance scale, and based on broadly based classifications that are of general, rather than specific application. Furthermore, not adequate data are available, and if available, a good percentage lacks detail in both spatial location and soil attribute information. According to Nachtergaele and Van Ranst (2003) and Zinck (1995) soil surveys coverage in 44 low- and middle-income countries, Kenya has 100 small-scale (1:500,000 to ±100,000); 25 medium-scale (1:100,000 to ±50,000) only; and no large-scale (1:50,000 to ±10,000) maps. Besides, their concepts and goals vary and time of compilation stretches way back to colonial era (1960s).

In spite of the recent and speedy developments in digital soil mapping methods, high-resolution soil data are seldom digitized and evaluated and if so, they are lacking systematic quality assessment. Our study exhibits how to update a rough soil map to a detailed digital fine scale soil class map by exploiting existing auxiliary data and digital soil mapping techniques. The result of this study may be

suitable to update old and less detailed soil maps or predict new ones for similar environments in the presence of fine resolution auxiliary information. Such description of up-to-date soil status is needed by various areas of speciality; for instance, researchers and policymakers need accurate and consistent soil information to support policy development for environmental integrity, economic development, and food and water security. Besides, this study may also contribute to projects such as GlobalSoilMap.net (Sanchez et al. 2009) which is currently working on African soils.

17.2 Materials and Methods

17.2.1 Study Area

The Study area is Baringo district, one of the districts in Rift Valley Province of Kenya. The district has a population of 264,978 (1999 census). Geographically, the area is situated from 00° 13′S to 1° 40′N and 35° 36′E to 36° 30′E. The district covers an area of 10,162 km^2, of which about 108 km^2 is covered by water surface, as shown in Fig. 17.1.

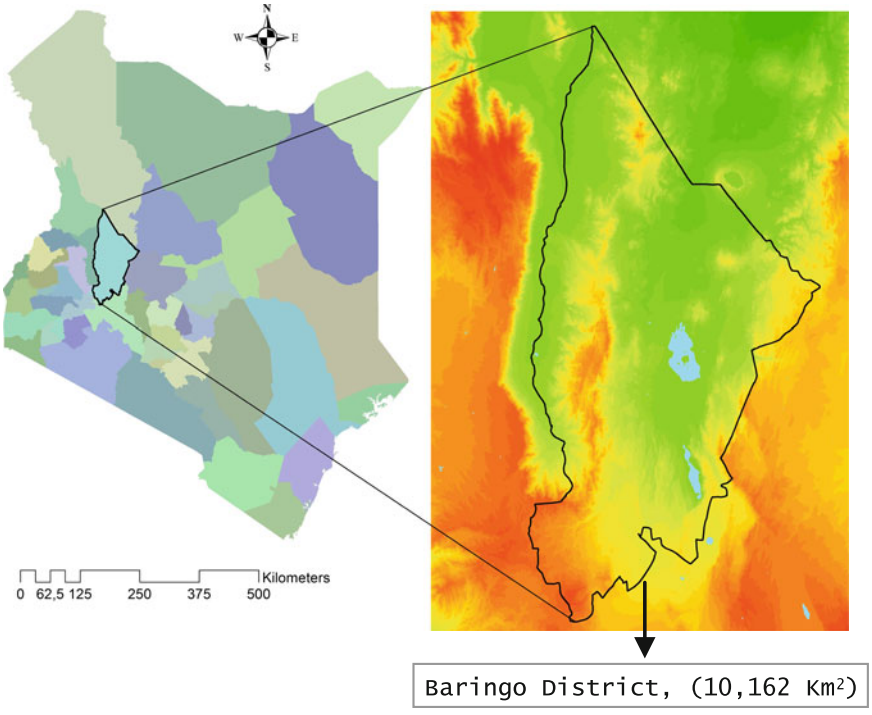

Fig. 17.1 SRTM elevation map of Baringo district in Kenya

Baringo district falls into the Kenyan Highland Zone, described by Morgan (1973) as having two peaks of rainfall patterns: long and short rain. The long rains start from the end of March to the beginning of July, and the short rains from the end of September to November. Average annual rainfall ranges from 600 mm in the lowlands to 1000–1500 mm while the annual mean minimum and maximum temperatures range from 16 to 18 and 25 to 30 °C respectively, with period between January and March as the hottest. The major topographic features in the district are river valleys and plains, the Tugen Hills, floor of the Rift Valley, and the northern plateau. The altitude varies between 752 m in the lowlands and 2600 m in the Tugen Hills.

The hills occurring in a north–south bearing mainly consist of volcanic rocks with steep slopes dissected by gullies. On the eastern and western parts of the hills are the escarpments and rivers flowing down these hills past through very deep gorges.

Vegetation change is remarkable along the topographic gradients, including temperate forests in the highlands to desert shrubs, such as drier acacia-species, on the valley floors. The highlands of the south and southwest of the catchment area and the summits of Tugen Hills are partly occupied by evergreen forest, farms, and pastures. The top of the eastern rift escarpment is covered by evergreen bushland with semi-deciduous wooded grassland at the foot of the hills. The soil consists of clay and clay loams with various depths.

17.2.2 Data Collection

The input soil data were collected from KENSOTER database (Batjes and Gicheru 2004) and consisted of 2570 soil profiles taken during the 1990s distributed throughout the country. The KENSOTER dataset is compiled by the Kenya Soil Survey (KSS) and ISRIC in accordance with the SOTER methodology developed for national and local agricultural planning purposes (Van Engelen and Wen 1995).

Digital elevation model (DEM) was obtained by the Shuttle Radar Topography Mission (Rabus et al. 2003) and downloaded from the free data service site with 90 × 90 m ground resolution. DEM is a numerical representation of topography, usually made up of equal-sized cells, each with a value of elevation. Its simple data structure and widespread availability have made it a popular tool for land characterization and soil distribution analysis (Blöschl and Sivapalan 1995; Chaplot et al. 2000; McBratney et al. 2003). Using the TOPOgrid function with ArcInfo Workstation GIS available in ArcGIS 9.3 package (ESRI, USA), a digital elevation model with 30 arc sec, corresponding to a pixel size of approximately 90 m, was generated, based on 1:50,000 topographic map obtained from USGS global topographic Data (GTOPO30)—SRTM (Shuttle Radar Topography Mission) (Rabus et al. 2003). First, it was prepared from the digitized contours and spot height using option in 3D Analyst and later converted to raster to give DEM. The DEM was used, directly or as a component, to compute slope gradient in percent.

Satellite images were generated from Landsat 5 (Jan 2010).

Surface Climate and rainfall distribution dataset were acquired from Almanac Characterization Tool (ACT) database (Mitchell and Jones 2005) with a roughly 5 × 5 km ground resolution, along with monthly rainfall from the CRU TS dataset (Mitchell and Jones 2005). The climate variables used were monthly mean values for minimum temperature, maximum temperature, precipitation, solar radiation, evaporation, etc. These climate surfaces can also be used to generate secondary information, e.g., bioclimatic parameters such as mean temperature of warmest period, precipitation of driest quarter., which are useful in determining the climatic envelope for such processes as soil formation among others.

We then converted all the data layers into a GIS database with WGS 1984 projection and with Projected Coordinate Systems of WGS 1984 PDC Mercator.prj.

17.2.3 Data Evaluation: Extracting Relationships from Existing Databases

We used the first and second most frequent soil types of KENSOTER database, combined them into a complex soil category whenever it was necessary and the output was used as target class-variable in our study.

A total of 999 random sample points were selected within the entire study area using Arcview Data management tools function. Surface spot and intersect tools were used to get the values for each and every sample points from all data layers (KENSOTER complex soil classes, 7 bands from satellite images, elevation and slope from SRTM DEM, and 42 indicators from ACT climate database). These attached values were then analyzed to develop relationships and predict the refined map based on the derived rules. The data were sampled in ArcGIS 9.3 environment (ESRI, USA) then exported and analyzed statistically using SPSS version 13 and MS Excel.

Principal component analysis is often preferred as a method for data reduction. Some describe it as a method of fitting a linear subspace to multivariate data by minimizing the chi distances (Jolliffe 2002). PCA is mostly used as a tool in exploratory data analysis and for making predictive models. The main application of principal component techniques in this study was to reduce the number of predictor variables to few and relatively easy to manage data. In this context, we reduced number of variables in Landsat and ACT climate data layers to two and three factors, respectively.

Following the steps above, we acquired nine explanatory variables (*x and y coordinates of the 999 random sampling points, elevation and slope, two factors from satellite bands, and three factors from climate data*) to predict complex soil classes in KENSOTER. These variables were then standardized to determine the centroid for each complex soil class in the defined space.

17.2.4 Data Evaluation: Prediction

For prediction phase, we sampled the predictive maps of variables in the study area at 10,000 random points (approximately one point per square kilometer). We calculated distances between point values and the centroids of the KENSOTER soil classes deduced from the previous step. Each complex soil class had one probability variable. The closest centroid to the point scored a value of 1 and the others 0 in the respective probability variables. On this way, we got 10,000 probability estimates for the soil classes with values 1 and 0.

The scores were then kriged to obtain continuous probability estimates for all the complex soil classes, and we subsequently combined the individual probability maps into a complex soil prediction map based on the highest probability values. Eventually, the final map was then evaluated with majority block statistics and it was converted into a vectorial map for evaluation and interpretation purposes.

17.3 Results

14 combined soil classes had more than 10 points among the initial 999 random points (Table 17.1) which we used as a proxy for defining major (>1 % area) and minor (<1 % area) soil categories. We considered that less than 10 points are insufficient to establish a reliable average for the centroids; thus, we performed our analysis for the major soil categories only. Four combined classes exist. There are Calcaric Regosols alone and in combination with Chromic Luvisols, there are Lithic Leptosols alone and Leptosols in combination with Haplic Calcisols, and Calcic Solonetz are combined either with Calcaric Fluvisols or with Calcisols. There is also a major non-soil category: relatively unaltered lava flows. We expected that soil categories alone or in combination with other categories should fall relatively close to each other in the established nine-dimensional space. However, this was only true between Calcaric Regosols and their complexation with Chromic Luvisols (distance < 1) and not for other combinations. However, centroids for lava flows and Calcaric Fluvisols were very close to each other (distance < 1) and that was the case for Haplic Andosols and Chromic Cambisols, too. These similarities had consequences for the final prediction as it will be discussed later. Distances between 1 and 2 existed for 16 of the possible 91 pairs of combinations.

In the next step, we performed prediction by using standardized explanatory variables at 10,000 random points, allocating probabilities to points and kriging allocated values. The result for one combined soil class (Leptosols + Haplic Calcisols) is shown in Fig. 17.2. The core relationships between soil classes and explanatory variables did not present themselves in the whole area from where it was drawn, and in turn, it was present also in other areas where originally other soil categories were indicated in the KENSOTER database. Transitional areas were represented with lower probabilities.

Table 17.1 Representation of soil categories in the initial sampling set

Combined soil classes	Number of points	Soil units (FAO-UNESCO 1974)
RGc + LVx	208	Calcaric Regosols + Chromic Luvisols
LPq	186	Lithic Leptosols
LP + CLh	128	Leptosols + Haplic Calcisols
RGc	123	Calcaric Regosols
SNk + FLc	88	Calcic Solonetz + Calcaric Fluvisols
Lav	42	(Lava flows)
CMe	36	Eutric Cambisols
NTu	35	Humic Nitisols
FLc	31	Calcaric Fluvisols
NT	24	Nitisols
CMx	18	Chromic Cambisols
ANh	16	Haplic Andosols
CMu	13	Humic Cambisols
SNk + CL	12	Calcic Solonetz + Calcisols
Major soil classes	**960**	
RGe	9	Eutric Regosols
FLe	7	Eutric Fluvisols
GLe	7	Eutric Gleysols
ANm	5	Mollic Andosols
CMc	5	Calcaric Cambisols
LXh	2	Haplic Lixisols
Minor soil classes	**35**	
not defined	4	(Lake surface)

Table 17.2 shows the changes between KENSOTER map and the final predicted map. The proximity of certain classes with respect to the explanatory variables resulted strong changes in area percentages. This is the major reason for the increased ratio of lava surfaces. Shallow and weakly developed soils (Leptosols, Regosols) usually lost their shares due to the large percentage of moderately steep slopes at lower elevations where the method predicted presence of Cambisols and other soil classes with more distinct profiles but still in initial phase of their development. The final predicted map is shown in Fig. 17.3.

17.4 Discussion

Digital soil mapping (DSM) provides a framework to formalize the use of existing information. The conceptual framework of DSM (McBratney et al. 2003) is based on the original model of Jenny (1941). Numerous DSM studies adopted the SCORPAN framework as the underlying conceptual model to predict soil properties and classes:

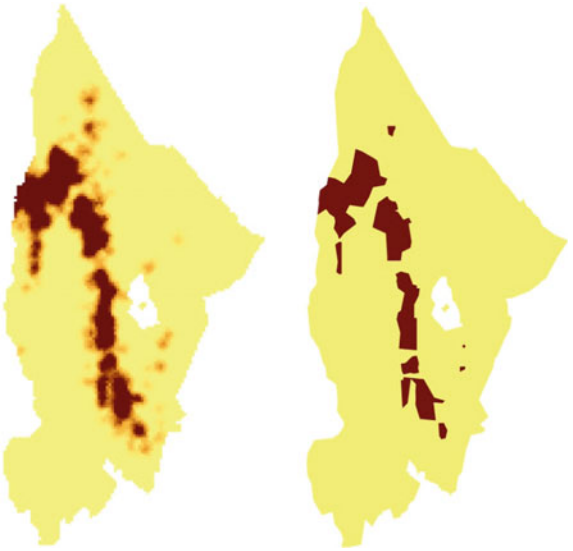

Fig. 17.2 Probability map for the Leptosols + Haplic Calcisols combined soil class and its final allocated area

Table 17.2 Changes in the predicted soil map compared to the KENSOTER database

	KENSOTER		Predicted map	
Soil classes	Number of polygons	% of total area	Number of polygons	% of total area
Lithic Leptosols	3	19.38	15	10.75
Calcaric Regosols + Chromic Luvisols	2	17.44	4	10.43
Calcaric Regosols	10	13.25	14	10.96
Leptosols + Haplic Calcisols	3	11.7	10	9.25
Calcic Solonetz + Calcaric Fluvisols	8	8.7	11	11.55
Calcaric Fluvisols	6	4.39	18	7.18
Lava flows	4	4.31	22	10.92
Humic Nitisols	3	4.11	5	5.34
6 minor soils	15	3.84	0	0
Nitisols	1	3.65	14	4.06
Eutric Cambisols	7	3.03	10	10.43
Chromic Cambisols	1	2.53	18	5.16
Haplic Andosols	3	1.33	8	1.59
Calcic Solonetz + Calcisols	2	1.28	2	0.7
Humic Cambisols	1	1.07	4	1.69
Total	71		155	

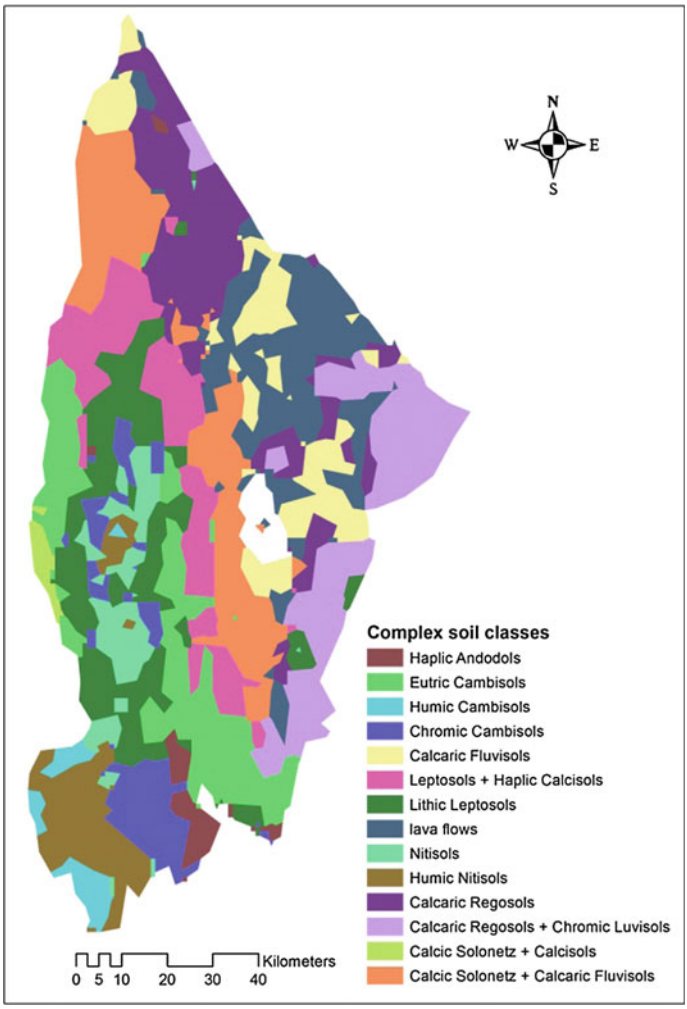

Fig. 17.3 Predicted soil class map for Baringo district, Kenya

$$\mathrm{Sc} = f(\mathrm{s,c,o,r,p,a,n}) \quad \text{or} \quad \mathrm{Sa} = f(\mathrm{s,c,o,r,p,a,n})$$

where Sc is a set of soil classes and Sa is soil attributes (properties) and the seven factors for soil spatial prediction are given as: s: refers to soil information at the same location either from a prior map or from remote sensing or expert knowledge, c: climate; o: organism (vegetation, fauna, or human activities); r: relief, the local topography (such as elevation, slope gradient, topographic wetness index); p: the parent materials; a: means age, and n: the geographical position. These factors are referred to as environmental covariates. McBratney et al. (2003) recognized that it is a rather rare case when all the soil-forming factors are represented in a study and

this is the case for our work, too. Time is difficult to include in the analysis but parent material (lithology) was also missing from our evaluated datasets. We can assume that conversion of soil categories into lava surfaces could have been better explained by inclusion of such data. Digital elevation models allow to calculate large number of terrain parameters which are more or less correlated (Behrens et al. 2010). We retained in our study only elevation and slope as predictor variables because an additional investigation indicated irrelevance of further variables. Climate was exhaustively described by 42 variables but only three factors of principal component analysis represented 96 % of their total variance. Surface cover and land use were captured by seven Landsat bands and 83 % of the total variance was represented by two PCA factors.

A comprehensive review of DSM is provided in McBratney et al. (2003) and an overview of pedometric techniques is used in DSM by McBratney et al. (2000) and Grunwald (2006). Many methods have been developed to extract or determine relationships between soil properties and terrain variables (McBratney et al. 2000, 2003). These methods can be grouped into four major types based on data sources: (1) methods for obtaining knowledge on relationship from local scientists; (2) methods for establishing relationships from field samples; (3) methods for discovering relationships from existing soil maps; and (4) methods for extracting relationships from typical pedons (typical classes).

The third approach (using existing soil maps to obtain landscape–soil relationship rules) has got special attention recently (Bui et al. 1999, 2002; Qi et al. 2006; Mayr et al. 2008) and disaggregation of existing soil maps was the objective of several studies. Häring et al. (2012) disaggregated spatially complex soil map units with the decision-tree method. In Hungary, Pásztor et al. (2014) and Sisák and Benő (2014) used classification trees to refine soil maps with the help of more detailed ancillary data.

Digital soil mapping approaches which utilize soil information from existing (usually small or medium scale) soil maps and field observations perform much better than pure theoretical constructions (Mendonça-Santos et al. 2008). Soil maps are physical representations of the mental models of the mappers on how soil-forming factors interact (Bui 2004). They provide us a path through the almost infinite number of theoretically possible combinations to the most probable outcome. In countries where small- or medium-scale soil maps exist, their statistical analysis may help to define homogenous soil regions or soilscapes and representative areas for detailed soil surveys (Behrens et al. 2009; Schmidt et al. 2010). By using these concepts, our study was a soilscape disaggregation exercise since we dismissed minor soil classes but refined probable areas for major classes.

In the study of Häring et al. (2012), original soil polygons were not overwritten only subdivision was allowed to the contrary of the papers of Bui and Moran (2001) and Yang et al. (2011) with which our work has some methodological similarities. However, the authors of both papers used unsupervised classification (*k*-means clustering and fuzzy *c*-means clustering) to establish environmental variable centroids for subclasses the number of which was fairly well known from pedological descriptions. In our study, only one centroid was calculated for each soil class so

classification was not our objective. Both other studies calculated distances between pixel values and cluster centroids while we calculated distances between values in random sampling points and centroids. Then, we obtained pixel-wise allocation rules with kriging. The two approaches have rather different results. Pixel-based calculations may produce extremely unhomogenous surfaces in the first step while our method produces relatively smooth probability estimates for each soil classes and the stacked layers of different classes with different probabilities represent the real-word variability. Grinand (2008) observed that soil class prediction accuracy can only be approximated correctly if test samples are collected at a certain distance from the training samples when predicting unvisited areas. Nauman and Thompson (2014) have found very similar results. The prediction accuracy was rather low for point profiles but it increased considerably when 60 m surrounding was considered. This may lead us to the conclusion that it is almost impossible to exactly predict soil class at a given location from small-scale soil maps at least in a variable terrain. However, we can allocate probabilities of occurence as our results suggest.

Our digital map has the same advantage as those reported by Bui and Moran (2001) and Yang et al. (2011) that existing boundaries in the map (seam along aligned map sheets or borderline of large polygons) can be overwritten and corrected by the derived rules.

Further refinement of the methodology is possible. We predicted a map only for soils with primary probabilities but it is still possible to produce similar maps for soils with secondary and tertiary probabilities and the resulting maps can be combined. Further possibility for modification when instead of 1 and 0 for the closest and the other centroids, we use membership measures with value of 1 for the closest centroid and continuous lower values proportional to the distances for the others (Yang et al. 2011). However, 1 and 0 scoring has the clear advantage of the nonambiguities.

17.5 Conclusion

With help of auxiliary variables, we were able to predict a refined soilscape map of Baringo district, Kenya, compared to the original KENSOTER database. We sampled the surface at random points and kriged the result instead of pixel-based calculations. Our approach had the clear advantage that real-world variability was represented by stacked layers of smooth probability estimates for the soil classes instead of blurred outputs where neighboring pixels can be differently allocated. We derived a soilscape map from classes of primary probabilities but calculation with secondary and tertiary probabilities is also possible and the resulting maps can be combined. 1 and 0 scoring of the most probable and the remaining soil classes seemed to have the clear advantage of nonambiguity but alternative calculation of membership grade is also possible.

Acknowledgements This Master Thesis project was realized within the framework of 2010/2011 Hungarian Government and FAO Masters Scholarships program. Special thanks to the entire Georgikon Faculty, FAO, and Hungarian Government for the support and scholarship to the first author.

References

Bacic, I. L. Z. (2008). *Demand driven land evaluation*. In: Hartemink, E.A., McBratney, A., Mendonca-Santos, M.L. (eds.) Digital soil mapping with limited data. Springer, Singapore. p. 151-162.

Batjes, N.H. and Gicheru, P. (2004). *Soil data derived from SOTER for studies of carbon stocks and change in Kenya (GEF-SOC project; Version 1.0).* Technical report 2004/01. ISRIC - World Soil Information, Wageningen.

Behrens, T., Zhu, A. X., Schmidt, K. and Scholten, T. (2010). *Multi-scale digital terrain analysis and feature selection for digital soil mapping*. Geoderma 155(3-4): 175-185.

Behrens, T., Schneider, O., Lösel, G., Scholten, T., Hennings, V., Felix-Henningsen, P. and Hartwich, R. (2009). *Analysis on pedodiversity and spatial subset representativity – the German soil map 1:1 000 000.* Journal of Plant Nutrion and Soil Science 172: 91–100.

Blöschl, G. and Sivapalan M. (1995): *Scale issues in hydrological modeling - a review*. Hydrological Processes 9(3-4): 251-290.

Bui, E. (2004). *Soil survey as a knowledge system*. Geoderma 120(1-2): 17–26.

Bui, E. N., and Moran, C. J. (2001). *Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data*. Geoderma 103(1-2): 79–94.

Bui, E.N., Loughhead, A. and Corner, R. (1999). *Extracting soil –landscape rules from previous soil surveys*. Australian Journal of Soil Research 37: 495-508.

Bui, E.N., Henderson, B., Moran, C.J. and Johnston, R. (2002). *Spatial data mining for enhanced soil map modelling*. International Journal of Geographical Information Science 16(6): 533-549.

Chaplot, V., Walter, C. and Curmi, P. (2000). *Improving soil hydromorphy prediction according to DEM resolution and available pedological data.* Geoderma 97: 405-422.

Cook, S.E., Jarvis, A. and Gonzalez, J. P. (2008). *A new global demand for digital soil information.* In: A.E. Hartemink, A. McBratney and M.L. Mendonça-Santos (Eds.) Digital Soil Mapping with Limited Data, Springer, Singapore. p. 31-42.

Dent, D.L. and Bai, Z. G., (2008). *Assessment of land degradation Using NASA GIMMS: A Case Study in Kenya.* In: Hartemink, E.A., McBratney, A., Mendonca-Santos, M.L. (Eds.) Digital Soil Mapping with Limited Data, Springer, Singapore. p. 247–258.

Eswaran, H., Almaraz, R., VandenBerg, E. and Reich, P. (1997). *An assessment of the soil resources of Africa in relation to productivity*. Geoderma 77(1): 1-18.

FAO-UNESCO, (1974). *Soil Map of the World (1 : 5 OOO OOO)*. Vol. 1: Legend. UNESCO, Paris.

Gonzalez, J.P., Jarvis, A., Cook, S.E., Oberthur, T., Rincon-Romero, M., Bagnell, J.A. and Dias, M. B. (2008). *Digital soil mapping of soil properties in Honduras using readily available biophysical datasets and gaussian processes*. In: Hartemink, E.A., McBratney, A., Mendonca-Santos, M.L. (eds.) Digital soil mapping with limited data. Springer, Singapore p. 367-380.

Grinand, C., Arrouays D., Laroche B. and Martin M.P. (2008). *Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context*. Geoderma 143(1-2): 180–190.

Grunwald, S. (Ed), (2006). *Environmental Soil-Landscape Modeling — Geographic Information Technologies and Pedometrics*. CRC Press, New York.

Häring, T., Dietz E., Osenstetter S., Koschitzki T. and Schröder B. (2012). *Spatial disaggregation of complex soil map units: a decision-tree based approach in Bavarian forest soils*. Geoderma 185: 37-47.

Hartemink, A.E., McBratney, A.B. and Minasny, B. (2008). *Trends in soil science: looking beyond the number of students*. Journal of Soil and Water Conservation 63: 76–83.

Jenny, H., (1941). *Factors of Soil Formation. A System of Quantitative Pedology.* McGraw-Hill Book Company, New York, NY, USA. 281 pp. ISBN: 0486681289

Jolliffe, I. T. (2002). *Principal Component Analysis*. 2nd Ed. Springer, New York.

Lagacherie, P., and McBratney, A. B. (2006). *Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping*. Developments in soil science 31: 3-22.

Mayr, T.R., Palmer R.C. and Cooke H.J. (2008). *Digital Soil Mapping using legacy data in the Eden Valley, UK*. In: Hartemink, E.A., McBratney, A., Mendonca-Santos, M.L. (eds.) Digital soil mapping with limited data. Springer, Singapore. p. 291-301.

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S.and Shatar, T.M., (2000). *An overview of pedometric techniques for use in soil survey*. Geoderma 97: 293–327.

McBratney, A.B., Santos, M.L.M., and Minasny, B., (2003). *On digital soil mapping*. Geoderma 117: 3–52.

Mendonça-Santos, M.D.L., Santos, H.G., Dart, R.O. and Pares, J.G. (2008). *Digital mapping of soil classes in Rio de Janeiro State, Brazil: data, modelling and prediction*. In: Hartemink, E. A., McBratney, A., Mendonca-Santos, M.L. (eds.) Digital soil mapping with limited data. Springer, Singapore. p. 381-396.

Mitchell, T.D. and Jones (2005). *An improved method of constructing a database of monthly climate observations and associated high-resolution grids*. International Journal of Climatology 25: 693-712. DOI:10.1002/joc.1181

Morgan, W.T.W. (1973). *East Africa*. Longman, London.

Nachtergaele, F.O. and Van Ranst, E. (2003). *Qualitative and quantitative aspects of soil databases in tropical countries*. In: G. Stoops (Editor), Evolution of tropical soil science: Past and future. Koninklijke Academie voor Overzeese Wetenschappen, Brussel, p. 107-126.

Nauman, T. W., & Thompson, J. A. (2014). *Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees*. Geoderma 213: 385–399.

Palm, C., Sanchez, P., Ahamed, S. and Awiti, A. (2007). *Soils: A contemporary perspective*. Annual Review of Environment and Resources 32: 99-129.

Pásztor, L., Dobos, E., Szatmári, G., Laborczi, A., Takács, K., Bakacsi, Z., & Szabó, J. (2014). *Application of legacy soil data in digital soil mapping for the elaboration of novel, countrywide maps of soil conditions.* Agrokémia és Talajtan 63(1): 79-88.

Qi Feng, Zhu A-Xing, M., Harrower, J. And Burt (2006). *Fuzzy soil mapping based on prototype category theory*. Geoderma 136(3-4): 774-787.

Rabus, B., Eineder M., Roth A. and Bamler R. (2003). *The Shuttle Radar Topography Mission — a new class of digital elevation models acquired by spaceborne radar*. ISPRS Journal of Photogrammetric and Remote Sensing 57: 241-262.

Rossiter, D. G. (2004). *Digital soil resource inventories: status and prospects*. Soil use and management 20(3): 296-301.

Rossiter, D.G., (2008). *Digital Soil Mapping as a Component of Data Renewal for Areas with Sparse Soil Data Infrastructures*. In: Hartemink, E.A., McBratney, A., Mendonca-Santos, M.L. (eds.) Digital soil mapping with limited data. Springer, Singapore. p. 69-80.

Sanchez, P.A., Ahamed S., Carre F., Hartemink A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.D., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T.G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., and Zhang, G.L. (2009). *Digital soil map of the world*. Science 325:680–681.

Schmidt, K., Behrens, T., Friedrich, K. and Scholten, T. (2010). *A method to generate soilscapes from soil maps*. Journal of Plant Nutrition and Soil Science 173(2): 163–172.

Sisak, I., & Benő, A. (2014). *Probability-based harmonization of digital maps to produce conceptual soil maps*. Agrokémia és Talajtan 63(1): 89–98.

Yang, L., Jiao Y., Fahmy S., Zhu A. X., Hann S., Burt J. E., and Qi, F. (2011). *Updating Conventional Soil Maps through Digital Soil Mapping*. Soil Science Society of America Journal 75(3): 1044–1053.

Van Engelen, V. and Wen, T.T. (1995). *Global and National Soils and Terrain Digital Databases (SOTER)*. Procedures Manual, FAO, Rome, 1995. pp. 125.

Zinck, J. A. (Ed.). (1995). *Soil survey: perspectives and strategies for the 21st century* (Vol. 80). Food & Agriculture Org..

Chapter 18
Predictive Mapping of Soil Organic Matter at a Regional Scale Using Local Topographic Variables: A Comparison of Different Polynomial Models

Xiao-Dong Song, Gan-Lin Zhang and Feng Liu

Abstract Borrowing the idea of software engineering, this paper aimed to evaluate the mapping accuracy of soil organic matter (SOM) content from the "black box" perspective by combing regression kriging (RK) with local terrain attributes calculated by different polynomial models. When calculating local terrain attributes, we applied two neighborhood shapes (square and circular) and six frequently used algorithms (Evans-Young, Horn, Zevenbergen–Thorne, Shary, Shi, and Florinsky). Overall, 35 combinations of first- and second-order derivatives were produced as secondary information for RK. For comparison, the ordinary kriging (OK), ordinary cokriging (COK), and universal kriging (UK) were also utilized to map the SOM spatial distribution. The results of the study showed that the RK application outperforms OK, COK, and UK in improving the prediction quality of SOM content in a region where the soil properties were strongly influenced by the toposequence and the altitude was with a wide range. The most accurate mapping result was obtained by the combination of the Evans-Young algorithm and Zevenbergen–Thorne algorithm for the calculation of first- and second-order derivatives, respectively. The mapping results from the higher-order approach (Zevenbergen–Thorne and Florinsky) yielded less prediction errors and the circular-neighborhood method could enhance some algorithms for the calculation of local terrain attributes.

Keywords Digital soil mapping · Soil organic matter · Regression kriging · Local terrain attributes · Environmental correlation

X.-D. Song · G.-L. Zhang (✉) · F. Liu
State Key Laboratory of Soil and Sustainable Agriculture,
Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China
e-mail: glzhang@issas.ac.cn

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_18

18.1 Introduction

In the past thirty years, significant advances have been made in information technology, especially in Geographic Information System (GIS), remote and proximal sensors, and digital elevation models (DEMs), which have significantly boosted the vitality of soil science (McBratney et al. 2003). Taking DEM as an example, a number of fundamental topographic attributes have been proposed to quantitatively identify landform classes and features within geomorphology (Wilson 2012), and thus, diverse algorithms are presented focusing on specific goals and scenarios. Much attention, therefore, has been devoted to predict soil properties by using the terrain attributes. A large number of studies have shown that prediction methods incorporating these pieces of secondary information outperform generic geostatistical models (e.g., ordinary kriging) (Bishop and McBratney 2001).

Some of topographic attributes are distinguished from non-local or regional parameters, and hence are referred to as local terrain attributes, which are derived directly from DEMs without additional inputs and usually calculated by moving a three-by-three window (Behrens et al. 2010; Florinsky 1998; Shary et al. 2002; Wilson 2012), such as slope, curvature, roughness, and elevation percentile. After a traversal across DEM, a new grid with the same dimension will be produced, whose cells are each filled with a calculated value of land surface parameter. For morphometric variables, the terms local and non-local are usually used regardless of the study scale or DEM resolution and associated with the mathematical sense of a particular variable (Florinsky 2011).

Among local terrain parameters, slope and aspect, twelve kinds of curvatures (Shary 1995) are also called first- and second-order derivatives, respectively, as they are defined by the formulae depending on the first- and second-order partial derivatives of altitudes. Multifarious mathematically modeling methods have been developed to calculate these derivatives from a gridded DEM focusing on various landscapes (Evans 1980; Horn 1981; Minár et al. 2013; Shary 1995; Shary et al. 2002; Shi et al. 2007; Zevenbergen and Thorne 1987). As the accuracy of the variables is unavoidably influenced by the DEM data and calculation algorithms, numerous studies have been published to estimate the accuracy of these algorithms (Schmidt et al. 2003; Warren et al. 2004), analyze the relationships between errors of derived parameters with DEM data characteristics (Chang and Tsai 1991; Gao 1997), and compare computed slope gradients with actual field measurements (Bolstad and Stowe 1994; Warren et al. 2004). Nevertheless, none of those studies is within the context of soil mapping and their results are hardly applicable to knowledge-based digital soil mapping (Shi et al. 2012). The selection approaches of terrain attributes also have not received the attention they deserve in soil science literature (Behrens et al. 2010).

The purpose of this research was to evaluate the mapping performance of soil organic matter (SOM) that results from RK technique combined with local terrain attributes based on different polynomial models. Nine terrain attributes were calculated from grid DEMs: elevation, topographic wetness index (TWI), slope,

aspect, plan curvature, profile curvature, tangent curvature, maximal curvature, and minimal curvature. The local terrain attributes were derived from six quadratic and Lagrange polynomials and two types of neighborhood shapes. Among the six algorithms, the Evans-Young algorithm (Evans 1980; Young 1978), the Horn algorithm (Horn 1981), and the Shary algorithm (Shary 1995) are based on a quadratic polynomial, and the Zevenbergen–Thorne algorithm (Zevenbergen and Thorne 1987), the Shi algorithm (Shi et al. 2007), and the Florinsky algorithm (Florinsky 2009) are based on a Lagrange polynomial. At the beginning of interpolation, Pearson correlation and partial correlation analyses were performed to scan the relations between SOM and all variables. We then compared the results of ordinary kriging (OK), ordinary cokriging (COK), universal kriging (UK), and regression kriging (RK). Furthermore, we discussed the combination of local terrain variables for RK which achieved acceptable quality for predicting the spatial variation of SOM contents and potentially other soil properties.

18.2 Materials and Methods

18.2.1 Data

The study area, the upper and middle reaches of the Heihe River Basin, is located along the northeast margin of the Qinghai-Tibetan Plateau in China at the intersection of the Tibetan Plateau, the Inner Mongolia-Xinjiang Plateau, and the Loess Plateau (Fig. 18.1). With the geographical boundary of about 97°20′–101°51′E and 37°41′–39°59′N, this area stretches for 340 km from the northwest to the southeast at a width between 115 and 180 km. Soil sampling was conducted in July to August 2012, including regular sampling and purposive sampling (Zhu et al. 2008) based on the concept of soil–environment relationships. A total of 223 topsoil (0–20 cm) samples recorded in above collections were compiled in a digital database. These data points were randomly split into calibration (80 %; n = 178) and validation (20 %; n = 45) datasets using the subset function of Geostatistical Analyst in ArcGIS (ESRI 2010).

18.2.2 Calculation of Local Terrain Variables and Other Related Terrain Variables

Terrain variables used in this study for the estimation of SOM were slope gradient, slope aspect, profile curvature, maximal curvature, minimal curvature (Table 18.1), elevation, and TWI. SRTM DEM data were employed and geo-referenced from three-arc second resolution to 90 m × 90 m resolution. The principal differences among most algorithms for the computing of local terrain variables are the number

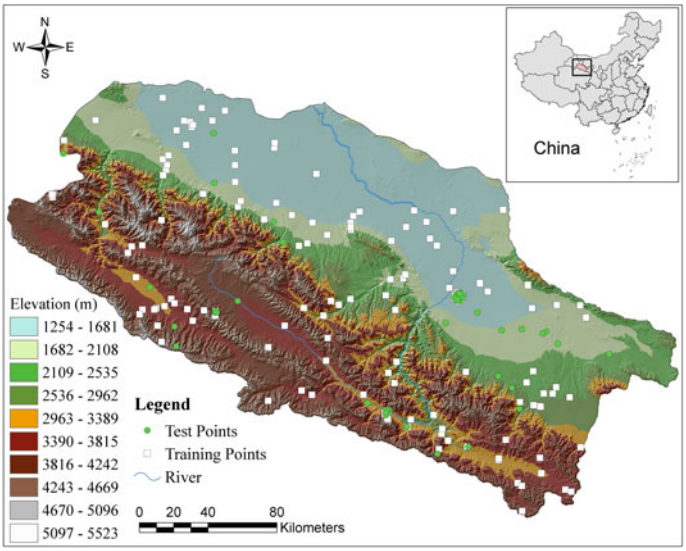

Fig. 18.1 Location of the study area and distribution of soil sampling sites

Table 18.1 Descriptive statistics of measured soil organic matter stock (SOM), log-transformed SOM (LnSOM), of the study area

Variables	SOM g kg^{-1}	LnSOM g kg^{-1}
Mean	33.61	2.99
Median	19.44	2.97
Minimum	1.21	0.19
Maximum	269	5.96
Standard deviation	41.49	1.02
Coefficient of variation (%)	140.10	34.11
Coefficient of skewness	3.91	0.19
Coefficient of kurtosis	20.06	0.28

SOM the soil organic matter stock; *LnSOM* log-transformed soil organic matter stock

of grid cell used and the weight given to each of those cell values. In general, most algorithms utilize some elevation values in a three-by-three window centered on the elevation cell in question, so that one can find all the unknown coefficients for a polynomial. However, a three-order polynomial should be fitted over all points in a 5 × 5 neighborhood for approximation of all the coefficients (Florinsky 2009; Minár et al. 2013).

The approximations for regular grid DEMs used were bivariate second-, third-, and partial fourth-order polynomials. In this paper, the first- and second-order terrain attributes were selectively calculated using the circular and square neighborhood, which resulted in a total of 39 layers (14 first-order derivatives and 25 s-order derivatives). The first-order derivatives, slope and aspect, were computed

by seven algorithms: the Horn, Zevenbergen–Thorne, and Florinsky algorithms using the square neighborhood, the Shi and Evans-Young algorithms using both square and circular neighborhood. Five kinds of curvatures (Table 18.1) were achieved by five algorithms: the Zevenbergen–Throne, Shary, and Florinsky algorithms with square neighborhood, and the Evans-Young algorithm with both square and circular neighborhood. Hence, 35 combinations of first- and second-order derivatives were grouped. All combinations were incorporated into the multiple linear regression of RK, so as to test which group would yield the best performance. The formula of aforementioned variables could be found in literatures (Florinsky 2011; Horn 1981; Shary 1995; Shary et al. 2002; Shi et al. 2007; Zevenbergen and Thorne 1987).

For convenience, in the rest of this paper, a specific terrain variable and all attributes with the same order are abbreviated to "*Variable* _ *Method* _ *Neighborhood*" and "*Method*" + "*n*" + "*Neighborhood,*" respectively, where *n* is the order of local topographic attributes. For example, Slp_EY_C denotes the slope gradient using circular neighborhood and the Evans-Young algorithm; FY_2_Q is the second derivatives calculated by the Florinsky algorithm with square neighborhood. Most of the layers were generated by the Terrain Analysis function of ArcSIE®, and other algorithms were implemented in C++ using GDAL library.

18.2.3 Methods

Four geo-statistical methods were involved in this study, including ordinary kriging (OK), cokriging (COK), universal kriging (UK), and regression kriging (RK). As a most general and widely used method of kriging, OK was employed to characterize the spatial variation of SOM and map overlays. If an interpolation is merely based on sample dataset, OK is commonly applied. OK uses the spatial correlation structure of the dataset to calculate weights for linear prediction from known points. Therefore, this method may require dense sample data for an interpolation with reasonable accuracy. In addition to OK, UK, COK, and RK are hybrid interpolation methods in which the variation of soil properties is quantified by deterministic and stochastic (empirical) models and can incorporate one or more ancillary variables in the estimation.

Cross-validation procedure was conducted to evaluate the accuracy of different models through three statistical measurements of the prediction error. The accuracy of estimates was assessed by the mean absolute error (MAE), the root mean squared errors (RMSE), and mean relative error ratio of performance to deviation (RPD). These indices were derived according to Eqs. (18.1), (18.2), and (18.3), respectively:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left[\left|Z^*(x_i) - Z(x_i)\right|\right] \tag{18.1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[Z^*(x_i) - Z(x_i)\right]^2} \tag{18.2}$$

$$\mathrm{RPD} = \frac{\mathrm{STD}}{\mathrm{RMSE}} \tag{18.3}$$

where $Z(x_i)$ is the observed value of Z at locations x_i, $Z^*(x_i)$ the predicted value at the same location, n the number of samples, and STD the standard deviations of the variable. MAE and RMSE were used to estimate the accuracy of the predictions which should be as low as possible for accurate interpolation. The RPD was employed so as to interpret the prediction ability of each model.

18.3 Results

18.3.1 Exploratory Data Analysis

The summary statistics for SOM and log-transformed SOM (LnSOM) are presented in Table 18.1. The observed SOM content in surface soils varied from 1.21 to 386.00 g kg^{-1}, with a mean value of 34.61 g kg^{-1}. The coefficient of variation (CV) was 140.10 g kg^{-1}, indicating that SOM for all samples had a very large variability. The value of skewness was 3.91 g kg^{-1}, suggesting that samples had a positively skewed distribution (Fig. 18.2a). The Kolmogorov–Smirnov (K–S) test (p-value = 0.000 < 0.05) rejected the null hypothesis of normality for samples. The SOM stock data were transformed by natural logarithm to create an

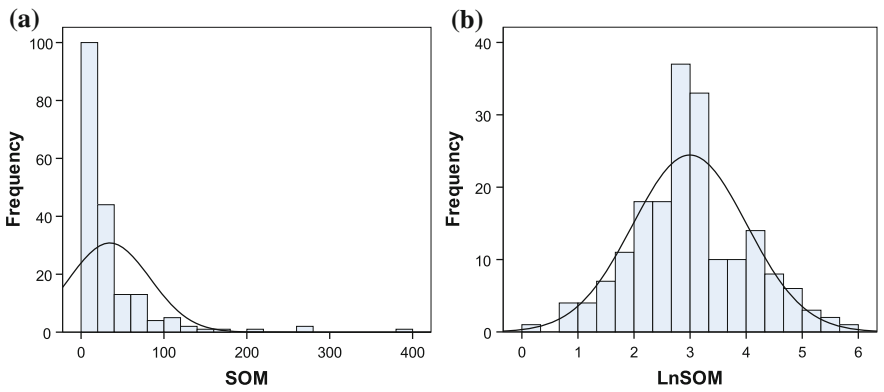

Fig. 18.2 Histogram of raw (**a**) and processed (**b**) datasets of SOM

approximately normal distribution, with mean (2.99 g kg^{-1}) and median (2.97 g kg^{-1}). Coefficients of skewness and kurtosis of lognormal SOM stock dropped from original values to 0.19 and 0.28 g kg^{-1}, respectively. Finally, the prediction values of SOM were back-transformed to original units.

Pearson's correlation analysis was carried out to explore the relationship between LnSOM and the terrain attributes based on the Evans-Young algorithm using square neighborhood (Table 18.2). These correlations were significant at the 0.01 level, suggesting that topography has important impacts on the distribution of SOM. The step-wise regression therefore was executed, aiming to derive the best subset of predictor variables and reduce the number of predictors (Table 18.2).

18.3.2 Prediction Accuracy of Different Kriging Methods

The aforementioned 45 validation datasets were used to assess the performance of different kriging methods with elevation, TWI, and local terrain attributes (Table 18.3). The prediction accuracy of SOM in this study was improved using RK with various combinations of local topographic attributes. The smallest and the largest prediction errors were produced by RK(EY1S_ZT2S) and RK (FY1S_EY2S), respectively. Compared with the worst method, the MAE and RMSE produced by RK(EY1S_ZT2S) method decreased by 6.46 g kg^{-1} and 20.23 g kg^{-1}, respectively, and the MRE increased by 0.84. The results of validation indicated that the combination of EY1S_ZT2S for the deriving of the local terrain attributes could remarkably improve the prediction accuracy of SOM prediction in this study area. RK(HN1S_ZT2S) also achieved a considerable accuracy, while the Horn and Zevenbergen–Throne algorithms might be the most widely used to calculate the first- and second-order derivatives due to the integration of mainstream GIS software. In the case of MAE, no values were close to zero, suggesting that there was a biased prediction. The RMSE values were slightly smaller than the standard deviations of the soil sample values (41.49 for SOM), and most of the RPD values were larger than 1.4. The inclusion of more auxiliary information in the RK regression models significantly improved the prediction performance.

Another important finding was that the performances of RK method whose second-order terrain attributes (SI1S_FY2S, ZT1S_FY2S, and SI1C_FY2S) were calculated by the third-order polynomial (Florinsky 2009) method outperformed most of the RK combinations and other kriging methods. Simultaneously, all the RPD values of RK combinations with FY2S, EY2C, and ZT2S were greater than 1.4, whereas the combinations with SA2S and EY2S were smaller than 1.4. Among RK results, RK with EY2S achieved the poorest performance, whereas all the RK with EY2C produced acceptable errors (RPD > 1.4). For all RK combinations, the circular neighborhood did not perform consistently better than the square neighborhood. This confirmed the previous conclusion (Shi et al. 2007) that the circular-neighborhood method may be more advantageous when used together with a specified neighborhood size, especially on a high-resolution DEM.

Table 18.2 The Pearson correlations between the logit-transformed soil organic matter content (LnSOM) and terrain variables calculated by the Evans-Young method

	Elev	TWI	Asp_EY_S	Slp_EY_S	MaC_EY_S	MiC_EY_S	PlC_EY_S	PrC_EY_S	TaC_EY_S
LnSOM	0.625**	−0.166*	0.266**	0.445**	0.205**	−0.347**	−0.093	0.139	−0.187*
Elev	1	−0.274**	0.318**	0.563**	0.451**	−0.263**	−0.281**	−0.038	−0.306**
TWI		1	−0.205**	−0.530**	−0.390**	0.017	0.412**	0.022	0.328**
Asp_EY_S			1	0.307**	0.183*	−0.175*	−0.036	−0.043	−0.114
Slp_EY_S				1	0.439**	−0.297**	−0.421**	0.123	−0.633**
MaC_EY_S					1	0.028	−0.651**	−0.532**	−0.548**
MiC_EY_S						1	−0.368**	−0.536**	−0.065
PlC_EY_S							1	0.091	0.802**
PrC_EY_S								1	−0.050
TaC_EY_S									1

Elev elevation; *TWI* topographic wetness index; *Asp* slope aspect; *Slp* slope gradient; *MaC* maximal curvature; *MiC* minimal curvature; *PlC* plan curvature; *PrC* profile curvature; *TaC* tangent curvature; *EY* Evans-Young algorithm; *S* square neighborhood

*Correlation is significant at the 0.05 level (2-tailed)

** Correlation is significant at the 0.01 level (2-tailed)

Table 18.3 Assessment of the various methods for predicting soil organic matter

Methods	MAE	RMSE	RPD	Methods	MAE	RMSE	RPD
RK(EY1S_ZT2S)	15.50	23.08	1.80	RK(EY1S_EY2C)	18.03	28.68	1.45
RK(SI1S_FY2S)	16.25	24.85	1.67	RK(SI1S_EY2C)	16.98	28.73	1.44
RK(ZT1S_FY2S)	16.97	26.36	1.57	RK(SI1C_SA2S)	19.53	37.41	1.11
RK(SI1C_FY2S)	16.98	26.39	1.57	RK(ZT1S_SA2S)	19.55	37.45	1.11
RK(HN1S_ZT2S)	17.28	27.40	1.51	RK(SI1S_SA2S)	20.10	38.29	1.08
RK(EY1C_ZT2S)	17.26	27.47	1.51	RK(EY1C_SA2S)	20.53	38.41	1.08
RK(FY1S_ZT2S)	17.31	27.73	1.50	COK	19.15	38.85	1.07
RK(FY1S_FY2S)	17.43	27.78	1.49	RK(HN1S_SA2S)	20.77	38.99	1.06
RK(EY1C_EY2C)	17.00	27.81	1.49	RK(ZT1S_EY2S)	21.08	39.32	1.06
RK(HN1S_EY2C)	17.16	27.86	1.49	RK(SI1C_EY2S)	21.09	39.35	1.05
RK(EY1C_FY2S)	17.50	27.92	1.49	RK(FY1S_SA2S)	20.80	39.47	1.05
RK(EY1S_FY2S)	18.40	28.06	1.48	OK	21.29	39.52	1.05
RK(HN1S_FY2S)	17.56	28.07	1.48	RK(EY1S_SA2S)	21.23	39.73	1.04
RK(FY1S_EY2C)	17.19	28.28	1.47	UK	20.71	40.29	1.03
RK(SI1S_ZT2S)	17.19	28.35	1.46	RK(EY1C_EY2S)	21.32	40.84	1.02
RK(ZT1S_ZT2S)	16.95	28.44	1.46	RK(SI1S_EY2S)	21.49	41.47	1.00
RK(ZT1S_EY2C)	16.69	28.48	1.46	RK(HN1S_EY2S)	21.68	41.76	0.99
RK(SI1C_ZT2S)	16.96	28.49	1.46	RK(EY1S_EY2S)	22.47	42.60	0.97
RK(SI1C_EY2C)	16.72	28.55	1.45	RK(FY1S_EY2S)	21.96	43.31	0.96

RK regression kriging; *OK* ordinary kriging; *COK* cokriging; *UK* universal kriging; *MAE* mean absolute error; *RMSE* root mean squared error; *RPD* ratio of performance to deviation. *EY* Evans-Young algorithm; *ZT* Zevenbergen–Throne algorithm; *HN*: Horn algorithm; *SA* Shary algorithm; *SI* Shi algorithm; *FY* Florinsky algorithm; *S* square neighborhood; *C* circular neighborhood

It is clearly seen that RK produced the SOM maps with more marked fluctuation than those of OK, COK, and UK (Figs. 18.3 and 18.4), especially when the maps were draped over the DEM they were based on. The obvious differences between the SOM maps generated by RK and other three kriging methods were the predicted SOM values in south part of study area (Qilian Mountain). The maps produced by RK showed more details of SOM content in spatial variation, which convincingly indicated the significant influences of toposequence as only the terrain attributes were used within the multiple linear regression.

One of the overall aims of this study was to compare the accuracies of SOM maps derived from RK with various combinations of local terrain attributes. Different from quantitative surface analysis (Jones 1998; Zhou and Liu 2004), it is an application-specific scenario for the mapping of SOM in regional area where the topography undulates greatly. Different combinations of first- and second-order derivatives provide diverse SOM maps due to their describing abilities of the general geomorphometry of land surface. In common with one of the objectives of geomorphometry, to a certain extent, digital soil mapping aims to quantitatively describe and model the variation of soil properties in terrestrial ecosystem. This

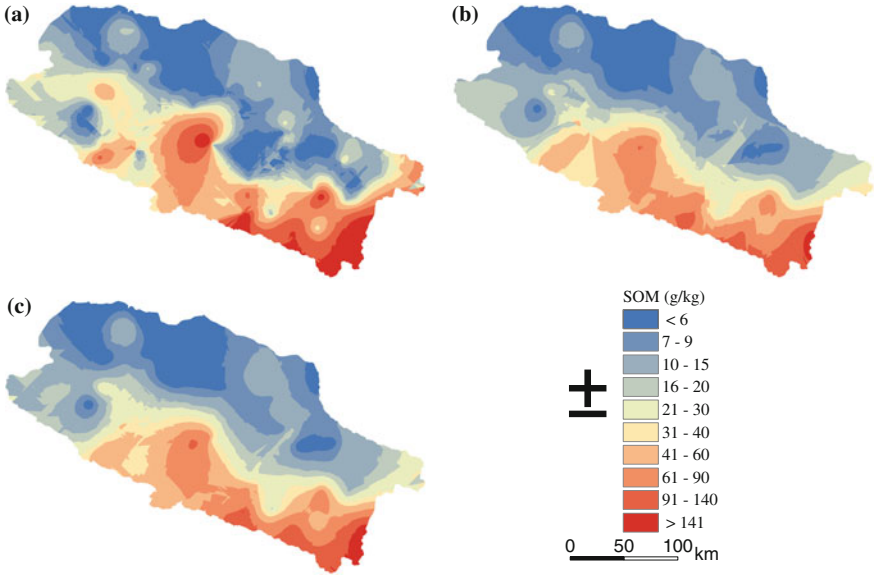

Fig. 18.3 The spatial predictions of soil organic matter content (g kg^{-1}) by universal kriging (**a**), ordinary kriging (**b**), and ordinary cokriging (**c**)

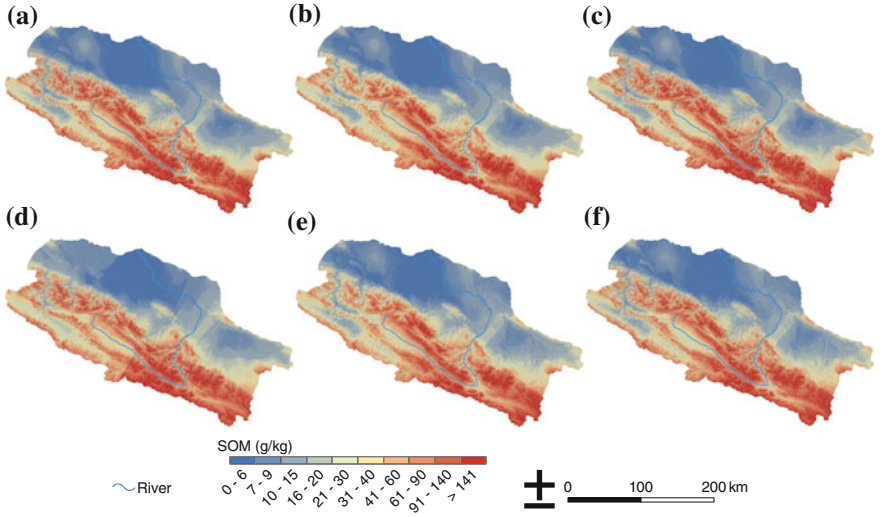

Fig. 18.4 Predicted soil organic matter (SOM) maps using regression kriging. *Note* The prediction values are draped over the DEM they are based on

quantitative description could be seemed as a scientific approach to evaluating the land surface modeling, which is reflected directly by the correlations between topographic variables and soil properties. Generally, it is confirmed especially when

the soil patterns are not affected by the agriculture and other anthropogenic activities.

It is helpful to arrive at a conclusion that we could achieve an optimal combination of first- and second-order derivatives based on disparate algorithms rather than the same algorithm. Other contrastive studies of polynomial models also found that modeling results from higher-order approaches show higher sensitivity to local variations (Florinsky 2009; Schmidt et al. 2003), such as the Zevenbergen–Throne algorithm and the Florinsky algorithm. This was coincided with the results of cross-validation listed above. There were 8 and 14 RK methods whose second-order derivatives were calculated by the Zevenbergen–Throne and Florinsky algorithms in the top 10 and 20 combinations. The main advantage of the Florinsky algorithm is the local denoising by approximating the polynomial to elevation values of the 5 × 5 window which could enhance the calculation of partial derivatives. Likewise, the modified Zevenbergen–Thorne algorithm with circular-neighborhood method is more sensitive to noise in the DEM, whereas the square-neighborhood method is less sensitive (Shi et al. 2007).

18.4 Conclusions

The contrast results of the current study could be deemed as the benchmark of different algorithms of local topographic variables. Nevertheless, it does not mean that the best method for the calculation of local parameters in this study will outperform others with different spatial resolutions and neighborhood sizes, especially when the DEM datasets are generated variously due to the vital accuracy of DEM. Comparing with traditional application, we can conclude that the performance of predictive methods that can incorporate auxiliary variables might be improved by using the same local terrain variable calculated by different methods. However, although the "black box" approach of digital soil mapping is working in hindsight, a more accuracy soil map of large poorly accessible area or difficult terrain might be achieved, which takes up a little time and energy rather than high sampling costs. In conclusion, our findings are important to select the algorithms of local morphometric variables for the RK technique or other prediction methods especially for the high-relief sites. Our study also provides a promising approach to choose the ancillary variables for mapping the spatial variation of other soil properties.

Acknowledgments The study was supported financially by the National Natural Science Foundation of China (grant No. 41130530, No. 91325301, No. 41401237) and partly by the Jiangsu Province Science Foundation for Youths (No. BK20141053). The authors are grateful to Prof. Ian S. Evans for sharing his published materials and to Dr. Florinsky for the interpretation of his method.

References

Behrens T, Zhu AX, Schmidt K, Scholten T (2010) Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155: 175-185.

Bishop TFA, McBratney AB (2001) A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103: 149-160.

Bolstad PV, Stowe T (1994) An evaluation of DEM accuracy: elevation, slope and aspect. Photogrammetric engineering and remote sensing 60: 1327-1332.

Chang KT, Tsai BW (1991) The effect of DEM resolution on slope and aspect. Cartography and Geographic Information Systems 18: 69-77.

ESRI. (2010). ArcView and ArcInfo. Version 10.0. California, Redlands: ESRI.

Evans IS (1980) An integrated system of terrain analysis and slope mapping. Zeitschrift für Geomorphologie, N.F., Supplementband 36: 274-295.

Florinsky IV (1998) Accuracy of local topographic variables derived from digital elevation models. International Journal of Geographical Information Science 12, 47-62. Florinsky, I.V., Eilers, R.G., Manning, G.R., Fuller, L.G., 2002. Prediction of soil properties by digital terrain modelling. Environmental Modelling & Software 17: 295-311.

Florinsky IV (2009) Computation of the third-order partial derivatives from a digital elevation model. International Journal of Geographical Information Science 23: 213-231.

Florinsky IV (2011) Digital terrain analysis in soil science and geology. Elsevier Academic Press, Amsterdam, p. 7-16.

Gao J (1997) Resolution and accuracy of terrain representation by grid DEMs at a micro-scale. International Journal of Geographical Information Science 11: 199-212.

Horn BKP (1981) Hill shading and the reflectance map. Proceeding of the IEEE 69: 14-47.

Jones KH (1998) A comparison of algorithms used to compute hill slope as a property of the DEM. Computers & Geosciences 24: 315-323.

McBratney AB, Mendonça Santos ML, Minasny B (2003) On digital soil mapping. Geoderma 117: 3-52.

Minár J, Jenčo M, Evans IS, Minár J, Kadlec M, Krcho J, Pacina J, Burian L, Benová A (2013) Third-order geomorphometric variables (derivatives): definition, computation and utilization of changes of curvatures. International Journal of Geographical Information Science 27: 1381-1402.

Schmidt J, Evans IS, Brinkmann J (2003) Comparison of polynomial models for land surface curvature calculation. International Journal of Geographical Information Science 17: 797-814.

Shary PA (1995) Land surface in gravity points classification by complete system of curvatures. Mathematical Geology 27: 373-390.

Shary PA, Sharaya LS, Mitusov AV (2002) Fundamental quantitative methods of land surface analysis. Geoderma 107: 1-32.

Shi X, Zhu AX, Burt J, Choi W, Wang RX, Pei T, Li BL, Qin CZ (2007) An experiment using a circular neighborhood to calculate slope gradient from a DEM. Photogrammetric Engineering and Remote Sensing 73: 143-157.

Shi X, Girod L, Long R, DeKett R, Philippe J, Burke T (2012) A comparison of LiDAR-based DEMs and USGS-sourced DEMs in terrain analysis for knowledge-based digital soil mapping. Geoderma 170: 217-226.

Warren SD, Hohmann MG, Auerswald K, Mitasova H (2004) An evaluation of methods to determine slope using digital elevation data. Catena 58: 215-233.

Wilson JP (2012) Digital terrain modeling. Geomorphology 137: 107-121.

Young M (1978) Terrain analysis: program documentation. Report 5 on Grant DA-ERO-591-73-G0040, 'Statistical characterization of altitude matrices by computer'. Department of Geography, University of Durham, England. 27 pp.

Zevenbergen LW, Thorne CR (1987) Quantitative analysis of land surface topography. Earth Surface Processes and Landforms 12: 47-56.

Zhou Q, Liu, X (2004) Analysis of errors of derived slope and aspect related to DEM data properties. Computers & Geosciences 30: 369-378.

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

Chapter 19
Estimating Soil Carbon Sequestration Potential in Fine Particles of Top Soils in Hebei Province, China

Xianghui Cao, Huaiyu Long, Qiuliang Lei and Shuxia Wu

Abstract Accurate evaluation of carbon sequestration potential (CSP) plays an important role in mitigating the buildup of atmospheric carbon dioxide. This study evaluated topsoil CSP of Hebei, using data collected during the recent soil inventory in 2010–2011. The results showed that shajiang black soils, irrigation silting soils, and coastal solonchaks were found the highest C content, and the values of them are 109.46 ± 14.70, 108.96 ± 30.24, and 146.91 ± 19.43 t C/ha, respectively. However, in terms of total potential of sequestration, although average potential of brown earths, cinnamon soils, and fluvo-aquic soils is not the highest, total potential of them is higher, and the values of them are 161.11, 475.12, and 409.76 Tg, respectively. From the perspective of the spatial pattern of CSP, the soils of 80–120 t C/ha that included bog soils, shajiang black soils, solonchaks, and irrigation silting soils possess the largest area (60.39 % of total soil area) and distributed mainly in the middle part of Hebei. The results will make it clear to understand the status quo of CSP, and the different types of soils play different roles in sources and sinks of CO_2.

Keywords Topsoil · Carbon sequestration · SOC · Carbon saturation

X. Cao · H. Long · Q. Lei (✉) · S. Wu
Institute of Agricultural Resources and Regional Planning,
Chinese Academy of Agricultural Sciences, Beijing 100081, China
e-mail: leiqiuliang@caas.cn

X. Cao
e-mail: 820646658@qq.com

H. Long
e-mail: hylong@caas.ac.cn

S. Wu
e-mail: 330889755@qq.com

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_19

19.1 Introduction

The world pays more attention to the climate change in recent years. The atmospheric CO_2 concentration has increased by 0.31 times from 1750 to 1999 and is currently increasing at the rate of 1.5 ppmv/year (McCarthy and Intergovernmental Panel on Climate Change. Working Group II 2001). However, terrestrial soils play an important role in the atmospheric carbon dioxide budget, which includes 1500 Pg of organic carbon, or 2.5–3 times as much organic carbon as the global atmosphere or terrestrial vegetation (Batjes 1996; Follett 2010). Consequently, these estimates will help establish better soil management practices, which could improve soil quality and mitigate the effects of global warming (Lal 2004a).

Currently, national or regional scale soil organic carbon (SOC) density and carbon sequestration research, especially in agricultural soils, have attracted significant attention (Vleeshouwers and Verhagen 2002; Marland et al. 2003). Numerous studies have been conducted to estimate agricultural soil sequestration potentials and explore management options to enhance carbon sequestration at national and regional levels (West and Post 2002). For example, no-tillage can contribute to the reduction of soil carbon significantly by reducing the loss of soil aggregates and the exposing of young and unstable organic matter to microbial decomposition in order to enhance CSP (Paustian et al. 2000). Recently, the DNDC model has been used to evaluate the significance of RMPs contributing to increase soil carbon sequestration and to explore effective carbon sequestration options by using a regional mode (Li et al. 2004; Tang et al. 2006; Zhang et al. 2006). Some research also have combined site-level process-based model with GIS, which extrapolated point measurements to regional scales (Falloon et al. 2000; Zimmerman et al. 2004). Nevertheless, soil carbon sequestration is a complicated process that is affected by many factors, such as organic carbon inputs from crop residue, climatic and soil property, the original carbon content, and soil type. At present, few studies focus on the CSP of different soil types.

So, in this study, quantitative estimation of CSP of different types of top soils is analyzed in Hebei Province of China according to the current investigation data in Hebei Province. We aimed to understand the status quo of CSP to provide the basis for choosing the measures of sustainable soil management and evaluate the capacity of soil carbon sequestration to make it clear that different types of soils play different roles in sources and sinks of CO_2.

19.2 Materials and Methods

19.2.1 Study Area

This research was conducted in Hebei Province (36°–43°N, 113°–120°E) in China. Hebei Province covers about 190,000 km^2 and the area of the cultivated soils accounted for 39.91 % of the total area. The altitude of north is higher than that of

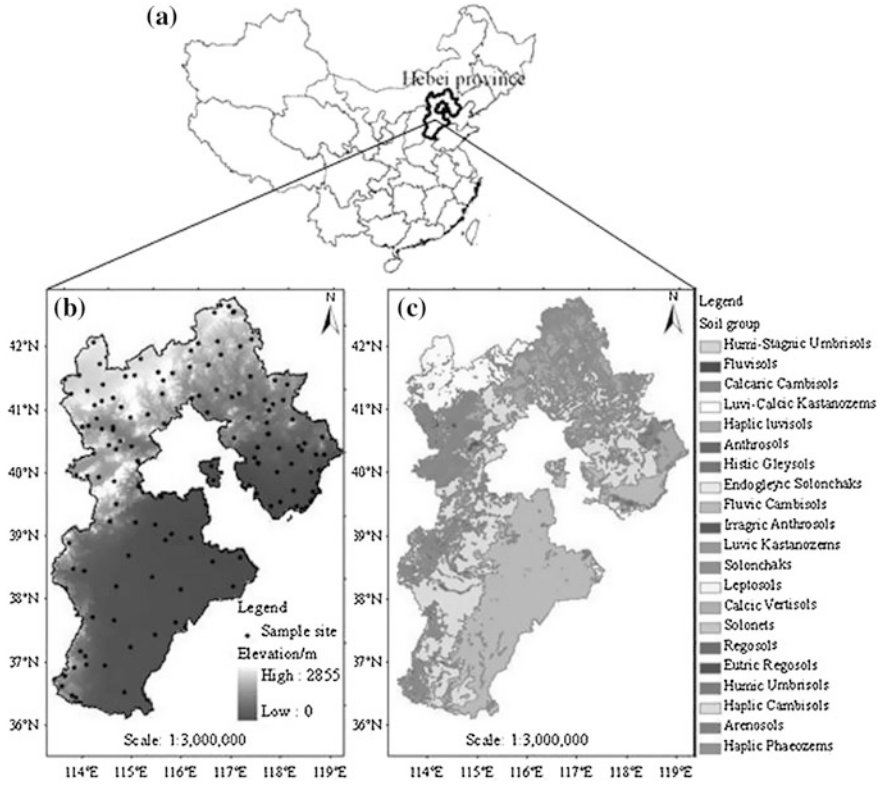

Fig. 19.1 The location of Hebei Province in China (**a**), elevation map of Hebei Province and distribution of soil sampling sites (**b**), and soil map of Hebei Province (**c**)

south, and the climate is temperate and warm temperate continental monsoon climate. The annual mean temperature and precipitation are 9.67 °C and 536 mm, respectively. There are 21 types of dominant soils, the names of which are shown in Fig. 19.1 and the covering area of each soil is in Table 19.1.

19.2.2 Data Source

The data were based on the current soil survey conducted in 2010–2011 in Hebei Province. There were 166 profiles according to the method of traditional sampling design. The sampling points of the overall 21 soil types identified by the soil survey were divided into uncultivated and cultivated soils according to the soil survey. Locations of the sampling points were shown in Fig. 19.1. The basic data of soils investigated mainly include SOM content, bulk density (some samples), mechanical composition, and the area of region.

Table 19.1 Saturation level of SOC and carbon sequestration potential of different soil groups in Hebei of China

Soil groups	Area (k ha)	Content of clay and silt (%)	Saturation level of SOC (t C/ha)	Average potential of carbon sequestration (t C/ha)	Total potential of sequestration (T g)	Percentage of potential of each soil type (%)
Castanozems	1277.35	45.29 ± 9.36	82.36 ± 12.03	30.27 ± 11.55	38.67	2.72
Brown earths	2308.51	65.33 ± 14.86	114.82 ± 24.47	69.79 ± 36.99	161.11	11.34
Meadow soils	69.98	35.11 ± 3.12	69.01 ± 4.56	21.21 ± 2.14	1.48	0.10
Bog soils	71.04	81.99 ± 5.78	141.79 ± 1.37	99.62 ± 19.3	7.08	0.49
Lithosols	386.34	61.43 ± 16.71	101.09 ± 19.23	41.49 ± 21.92	16.03	1.13
Black soils	1.56	54.63 ± 6.02	86.83 ± 5.44	14.03 ± 7.25	0.02	0.01
Skeletal soils	1451.57	57.04 ± 15.02	105.07 ± 26.81	64.08 ± 37.81	93.02	6.55
Aeolian soils	182.84	17.12 ± 7.03	46.76 ± 10.95	27.85 ± 8.03	5.09	0.36
Shajiang black soils	65.44	85.73 ± 7.24	149.09 ± 7.39	109.46 ± 14.70	7.16	0.50
Solonetzs	81.69	59.72 ± 6.54	102.32 ± 7.23	46.86 ± 3.45	3.83	0.27
Solonchaks	952.29	68.02 ± 10.8	126.73 ± 16.55	97.61 ± 21.58	92.95	6.54
Irrigation silting soils	84.45	75.27 ± 19.49	138.33 ± 30.82	108.96 ± 30.24	9.20	0.65
Neo-alluvial soils	74.46	35.34 ± 18.65	72.71 ± 26.56	41.65 ± 23.66	3.10	0.22
Castano-cinnamon soils	734.97	54.79 ± 17.92	106.67 ± 29.77	79.51 ± 32.9	58.43	4.11
Cinnamon soils	5080.39	67.79 ± 17.53	125.05 ± 28.41	93.52 ± 35.04	475.12	33.43
Fluvo-aquic soils	4251.11	66.93 ± 19.38	124.95 ± 32.38	96.39 ± 39.31	409.76	28.83
Red primitive soils	0.86	63.41 ± 5.89	119.68 ± 12.61	90.59 ± 10.54	0.08	0.01
Paddy soils	48.97	75.22 ± 11.57	135.08 ± 18.71	98.82 ± 19.83	4.83	0.34

(continued)

Table 19.1 (continued)

Soil groups	Area (k ha)	Content of clay and silt (%)	Saturation level of SOC (t C/ha)	Average potential of carbon sequestration (t C/ha)	Total potential of sequestration (T g)	Percentage of potential of each soil type (%)
Coastal solonchaks	203.30	89.04 ± 4.98	166.35 ± 16.98	146.91 ± 19.43	29.86	2.10
Gray forest soils	105.51	56.87 ± 14.31	94.23 ± 16.06	32.31 ± 17.14	3.41	0.24
Mountain meadow soils	43.19	49.92 ± 5.78	80.36 ± 6.98	20.13 ± 1.47	0.87	0.06
Total	17475.90				1421.10	100

19.2.3 Calculation Methods

19.2.3.1 Calculation of SOC Content and SOC Density in Topsoil (0–30 cm)

This paper mainly studied the soils of thickness of soil horizon (0–30 cm). However, the thickness of some top soils exceeds 30 cm and that of some top soils is less than 30 cm. For the thickness of soil horizon exceeding 30 cm, SOC of 0–30 cm is original values. However, SOC of 0–30 cm can be calculated by the thickness weight method. The soil organic matter content of the samplings was converted to SOC by multiplying a constant (0.580).

$$\mathrm{SOC}_{0-30\,\mathrm{cm}} = \frac{0.58\left[O_i H_i + O_j(30 - H_i)\right]}{30} \tag{19.1}$$

where $\mathrm{SOC}_{0-30\ \mathrm{cm}}$ means the SOC content of 0–30 cm; H_i is the thickness of soil horizon that is less than 30 cm; O_i means the SOM content of H_i; and O_j means the SOM content of the thickness of soil horizon ranging from Hi to 30 cm.

The topsoil SOC density (Dsoc) was estimated using Eq. (19.2).

$$D_{\mathrm{oc}} = \mathrm{SOC}_{0-30\,\mathrm{cm}} \times \gamma \times H \times \left(1 - {}^{\delta_{2\,\mathrm{mm}}}/_{100}\right) \times 10^{-1} \tag{19.2}$$

where Doc (t/ha) is the total amount, γ (g/cm^3) is the bulk density, H is the soil depth (cm), and $\delta_{2\ \mathrm{mm}}$(%) is the 2-mm coarse fraction of the soil. Partial data of bulk density were missing, and bulk density was estimated by regression analysis between the available bulk density and SOC content for a given layer.

19.2.3.2 Calculation of Carbon Sequestration Potential

According to clay and silt content of different soil types, maximum amount of SOC associated with the particles (<20 μm) can be calculated by (Hassink 1997):

$$C_{\mathrm{sat}} = 4.09 + 0.37 \times \%(\mathrm{clay} + \mathrm{silt}) \tag{19.3}$$

where C_{sat} means the saturated carbon content of clay and silt (g/kg).

Based on total organic carbon content and the proportion of stable carbon of clay and silt (<20 μm) (x), the saturated carbon content of clay and silt can be calculated. The percentage content (x) of saturated carbon content in total organic carbon content ranges from 85 to 89 %. Generally, the percentage content (x) is 85 ± 2.5 %, and the carbon sequestration of soil void can be calculated by:

$$S_{\text{def}} = C_{\text{sat}} - x\text{SOC}_{0-30\,\text{cm}} \tag{19.4}$$

where S_{def} means carbon sequestration amount of soil void (g/kg) and $\text{SOC}_{0-30\ \text{cm}}$ means the current content of SOC (g/kg).

The potential capacity of carbon sequestration can be estimated by:

$$S_c = S_{\text{def}} \times \text{BD} \times (1 - \text{RF}) \times d \times 10^{-1} \tag{19.5}$$

19.2.3.3 Calculation of Soil Bulk Density

Because bulk densities did not fully follow the measurement of SOM in the current soil survey, the missing bulk densities could be estimated by establishing correlation between the available bulk densities and SOC content. The regression between soil bulk density and SOC content does rest with soil types. Of the data available, 30 topsoil samples had the value of both SOC content and bulk density (γ). The regression between bulk density and SOC is established as follows:

$$\gamma = 1.5915 \times e^{-0.012 \times \text{SOC}} \left(R^2 = 0.7816\right) \tag{19.6}$$

The regression Eq. (19.6) was used to calculate the missing bulk density value.

19.3 Results and Discussion

19.3.1 Current SOC Density in Different Soil Types

The estimated SOC density of individual soil types ranged from 22.25 ± 7.71 to 85.65 ± 9.75 t C/ha widely (Fig. 19.2). The difference of organic carbon density of different soil types is significant. SOC density of black soils is the highest (85.65 ± 9.75 t C/ha), and the second highest is the gray forest soils (72.84 ± 18.96 t C/ha). SOC density of aeolian soils is the lowest (22.25 ± 7.71 t C/ha). SOC density of the other soil types is between 31.95 ± 0.39 and 70.85 ± 8.64. The differences of SOC density may be due to different soil environment and soil characteristics. Generally, SOC density of luvisols, semi-luvisols, calcium soil, and semi-hydromorphic soil is higher. However, SOC density of primitive soil, saline-alkali soil, and anthrosol is lower. For instance, black soils and gray forest soils that belong to semi-luvisols have higher carbon density. The main reason is that lower annual average temperature (<1 °C) slows down the decomposition rate of organic matter so that organic matter converts into lots of humus. Aeolian soils that form mainly in the extreme arid area have lower carbon density. Low soil moisture and vegetation coverage rate result in low decomposition rate of plant residue and high mineralization rate so that the accumulation rate of organic matter is very low in soil (Fig. 19.2).

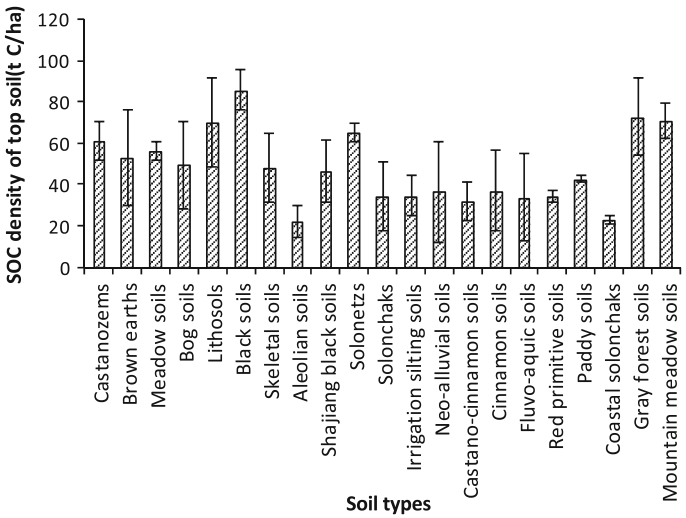

Fig. 19.2 Soil organic carbon density of different soil types

19.3.2 Carbon Sequestration Potential and Saturation Level of SOC in Different Soils

In total, 166 soil profiles from the 2010 to 2011 survey were identified as having sufficient data to derive the clay and silt content for both the 0–30 cm. According to the formulas (19.1) and (19.3), saturation level of SOC can be calculated. The soils of different types have different saturation level of SOC. As shown in Table 19.1, on average, the saturation level for the different soil types varied widely, ranging from 46.76 ± 10.95 to 166.35 ± 16.98 t C/ha. Saturation level of coastal solonchaks is the highest and that of aeolian soils is the lowest. The positive correlation between the clay–silt content and the saturation level were shown in Table 19.1. Several studies showed that the saturated SOC is mainly related with clay and silt content. A number of published studies indicate that the ability of associating with clay and silt particles is one of the principal factors responsible for physical protection of organic matter in soils (Theng 1979). It has been determined that the capacity of silt and clay of protecting organic matter is finite and the capacity can be determined. The saturated SOC is related to the percentage of the clay and silt fraction in the total size fraction (Hassink 1997).

There was wide variation in average potential of carbon sequestration for different types of soils, ranging from 14.03 ± 7.25 to 146.91 ± 19.43 t C/ha. Potential of shajiang black soils, irrigation silting soils, and coastal solonchaks is higher, and the values of them are 109.46 ± 14.70, 108.96 ± 30.24, and 146.91 ± 19.43 t C/ha, respectively. The higher CSP of these soils could be mainly due to their occurrence in low lands and their high clay that contribute to accumulate SOC (Pan et al. 2004; Li et al. 1992). However, in terms of total potential of sequestration, although

average potential of brown earths, cinnamon soils, and fluvo-aquic soils is not the highest, total potential of them is higher, and the values of them are 161.11, 475.12 and 409.76 Tg, respectively. Total potential of the three types of soils accounts for 11.34, 33.43, and 28.83 %, respectively. Brown earths (2308.51 kha), cinnamon soils (5080.39 kha), and fluvo-aquic soils (4251.11 kha) possess larger area of Hebei, which can explain this phenomenon. In terms of total potential, cinnamon soils, brown earths, and fluvo-aquic soils will play an essential role in carbon sequestration to mitigate global warming in the future.

19.3.3 Spatial Distribution of Carbon Sequestration Potential and Saturation Level of SOC

Statistics based on 166 soil samplings for the saturation level of SOC (t C/ha) map of Hebei (Fig. 19.3a) show that saturation level of SOC in different polygons varied dramatically, with the lowest saturation level (46.76 t C/ha) and the highest level (166.35 t C/ha). The total soil area is 17475.93 kha in Hebei Province, China, with the largest area of 100–150 t C/ha (88.79 %) and with the most small area of 0–50 t C/ha (1.06 %). Figure 19.3 shows that the distribution of saturation level is

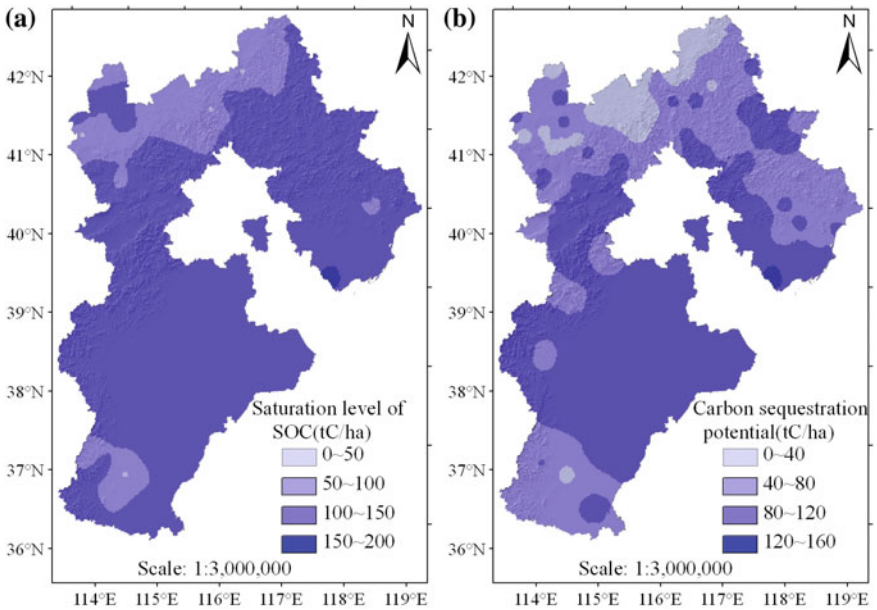

Fig. 19.3 Spatial pattern of saturation level of SOC (**a**) and potential of carbon sequestration (**b**) among different soil types in Hebei Province of China

uneven in Hebei. The soils of 100–150 t C/ha dominated saturation level, distributing mainly in the south and northeast of Hebei. The soil types of this region mainly include brown earths, bog soils, skeletal soils, shajiang black soils, red primitive soils, and cinnamon soils. However, the distribution scope of the other three grades is narrow. The soils of 50–100 t C/ha including castanozems, meadow soils, and black soils mainly distributed in the north of Hebei. The soils of 0–50 and 150–200 t C/ha present the condition of scattered distribution. Although the soil area of 0–50 t C/ha (aeolian soils) is close to that of 150–200 t C/ha (coastal solonchaks), saturation level of the soils ranging from 150 to 200 t C/ha is more than three times than that of the soils ranging from 0 to 50 t C/ha. The main factors affecting this distribution pattern of SOC saturation level are climate, degree of vegetation coverage, terrain, degree of the land use, and human activities.

Figure 19.3b depicts the spatial patterns of CSP for different soil groups in Hebei Province, China. The highest C sequestration potential (120–160 t C/ha) that accounted for about 1.16 % of total soil area occurred mostly in northeastern Hebei; and the lowest C sequestration potential (0–40 t C/ha) that accounted for about 9.62 % of total soil area can be found in northern Hebei. However, the soils of 80–120 t C/ha that included Bog soils, shajiang black soils, solonchaks, and irrigation silting soils possess the largest area (60.39 % of total soil area) and distributed mainly in the middle part of Hebei. Such regional distribution is closely associated with the pattern of climate, cropping systems, and soil properties (Li et al. 2004; Li et al. 2005; Sun et al. 2010). The middle parts of Hebei are dominated by high-clay soils, whereas other regions of Hebei are dominated by relatively low temperature and soils with low silt and clay content.

Overall, the middle regions accounting for 60.39 % of the total Hebei have higher CSP and should be considered as future carbon sequestration items (Fig. 19.3b). From spatial distribution, it was possible to identify the distribution regions of carbon sequestration.

19.3.4 Carbon Sequestration Potential of Different Land Uses

Different land-use types present different CSP (Fig. 19.4). CSP in the grassland, forestland, and arable land ranged from 14.03 ± 7.25 to 85.69 ± 21.54 t C/ha, 19.87 ± 6.45 to 108.64 ± 32.01 t C/ha, and 25.94 ± 7.78 to 112.96 ± 25.58 t C/ha, respectively. It is easy to see that average CSP of three land-use types was different significantly. And average CSPs of them are 42.37(grassland), 61.92(forestland), and 83.89(arable land) t C/ha, respectively. It is clear that average CSP of arable land is the highest. Generally, SOC content of agricultural soils is lower than that of natural soils, which is mainly due to lower carbon input (as a result of annual harvest and removal of crop residue, etc.), higher organic carbon decomposition (owing to frequent tillage), increasing soil erosion (Lal 2004a), and other factors.

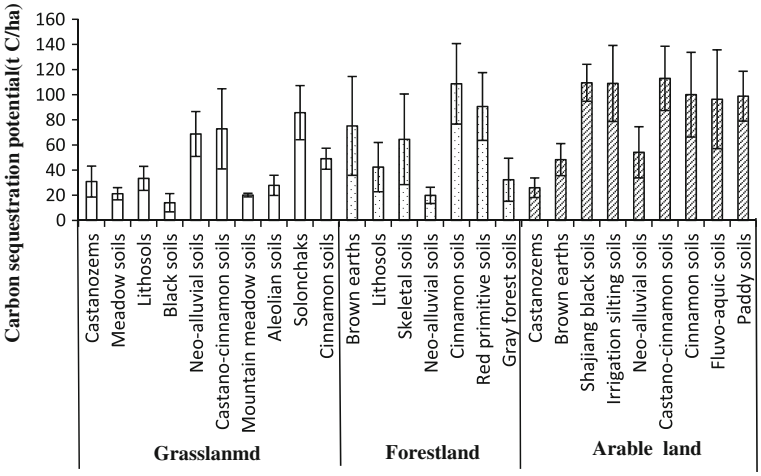

Fig. 19.4 Carbon sequestration potential of different land-use types

Several evidences also showed that amounts of crop residue carbon returning back to soil contribute to CSP ultimately (Srinivasarao et al. 2012, 2014). Probably, it could be considered as the adoption of RMPs (Xu et al. 2011). This indicates that arable land plays an essential role in the CSP.

19.3.5 Analyses and Suggestions for Management

Soil is considered as one of the most important sources and sinks of greenhouse gases that cause global warming and climate change (Janssens et al. 2003). Soils account for approximately 20 % of total carbon dioxide emissions due to soil and root respiration, including 12 and 60 % of methane and anthropogenic nitrous oxide emissions, respectively (IPCC 2007). Soil has potential for mitigating global warming by sequestrating carbon (Pathak et al. 2011). According to our results, different soils possess various carbon sequestration capacities. For example, Endogleyic solonchaks have a distinct carbon sequestration advantage over other soils (Brevik and Homburg 2004). Thus, these soils have a very high potential (126.48–146.91 t C/ha) for accumulating large amounts of carbon, at high rates and over long periods because they continuously accumulate organic-rich sediments.

The average CSP in soils of different ecosystems has been estimated in Hebei, emphasizing agricultural soils. According to the previous experience, compared with grassland and forestland, the study found that arable land has higher capacity of carbon sequestration (Smith et al. 1998). So it is necessary to strengthen the management of agricultural soil. Currently, crop straws are removed from fields after harvest, and crop roots remain in the soil. No-tillage contributes to enhance the carbon sequestration. It is well documented that greater tillage intensities result in

greater SOM decomposition rates due to the effects of tillage on macro-aggregate breakage, soil aeration, and crop residue burial (Huang et al. 2010; Mishra et al. 2010). Therefore, no-tillage with minimum SOM decomposition interference is preferable for increasing CSP. Of course, although grassland and forestland possess only 16.3 and 16.7 % of soil area of Hebei, the importance of them in the mitigating the CO_2 concentration of atmosphere cannot be ignored. It is necessary to protect forest and grass ecosystem to mitigate the climate change combining with controlled agriculture ecosystem.

19.4 Conclusion

It is well known that several factors such as soil type, production, and management influence CSP; and it is important to identify CSP in different soil groups. This paper conducts the study in SOC density, carbon saturation level, and CSP of 21 types of soils in Hebei.

The research found that whether the CSP of soils is high or low is not determined by single factor of SOC density or carbon saturation level. However, CSP is the comprehensive result of SOC density and carbon saturation level. Shajiang black soils, irrigation silting soils, and coastal solonchaks are higher, and the values of them are 109.46 ± 14.70, 108.96 ± 30.24, and 146.91 ± 19.43 t C/ha, respectively. However, in terms of total potential of sequestration, although average potential of brown earths, cinnamon soils, and fluvo-aquic soils is not the highest, total potential of them is higher, and the values of them are 161.11, 475.12, and 409.76 Tg, respectively. This indicates that they will play a important role in mitigating climate change in the future. Of course, we know of not only the CSP of soils but also the spatial pattern of CSP of soils. The soils ranging from 80 to 120 C/ha that included bog soils, shajiang black soils, solonchaks, and irrigation silting soils possess the largest area (60.39 % of total soil area) and distributed mainly in the middle part of Hebei. This suggests that soils of middle part should be paid more attention. From the aspect of land-use types, agricultural soils have higher capacity of carbon sequestration. In summary, these provide information for sustainable soil management of Hebei and sources and sinks of CO_2.

Acknowledgements We acknowledge the "Ministry of Science and Technology (MoST), China" for funding this study. The results and opinions presented in this paper are those of the authors.

References

Batjes N H (1996) Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science* **47**:151-163. doi: 10.1111/j.1365-2389.1996.tb01386.x.

Brevik E C, Homburg J A (2004) A 5000 year record of carbon sequestration from a coastal lagoon and wetland complex, Southern California, USA. *Catena* **57**:221-232. doi: 10.1016/j.catena.2003.12.001.

Falloon P, Smith P, Coleman K, Marshall S (2000) How important is inert organic matter for predictive soil carbon modelling using the Rothamsted carbon model? *Soil Biology & Biochemistry* **32**:433-436. doi: 10.1016/S0038-0717(99)00172-8.

Follett R F (2010) Symposium: Soil Carbon Sequestration and Greenhouse Gas Mitigation. *Soil Science Society of America Journal* **74**:345-346. doi: 10.2136/sssaj2009.cseqghgsymp.intro.

Hassink J (1997) The capacity of soils to preserve organic C and N by their association with clay and silt particles. *Plant and Soil* **191**:77-87. doi: 10.1023/A:1004213929699.

Huang S, Sun Y N, Rui W Y, Liu W R, Zhang W J (2010) Long-Term Effect of No-Tillage on Soil Organic Carbon Fractions in a Continuous Maize Cropping System of Northeast China. *Pedosphere* **20**:285-292.

IPCC (2007) *Climate change: the scientific basis*. Cambridge Univ. Press, Cambridge, U.K.

Janssens I A, Freibauer A, Ciais P, Smith P, Nabuurs G J, Folberth G, Schlamadinger B, Hutjes R W A, Ceulemans R, Schulze E D, Valentini R, Dolman A J (2003) Europe's terrestrial biosphere absorbs 7 to 12 % of European anthropogenic CO_2 emissions. *Science* **300**:1538-1542. doi: 10.1126/science.1083592.

Lal R (2004a) Soil carbon sequestration impacts on global climate change and food security. *Science* **304**:1623-1627. doi: 10.1126/science.1097396.

Li C S, Mosier A, Wassmann R, Cai Z C, Zheng X H, Huang Y, Tsuruta H, Boonjawat J, Lantin R (2004) Modeling greenhouse gas emissions from rice-based production systems: Sensitivity and upscaling. *Global Biogeochemical Cycles* **18**. doi:10.1029/2003gb002045.

Li C, Frolking, S., Frolking, T.A (1992) A model of nitrous oxide evolution from soil driven by rainfall events: I, model structure and sensitivity. *Geophys. Res.* **97**:9759–9776.

Li C S, Frolking S, Butterbach-Bahl K (2005) Carbon sequestration in arable soils is likely to increase nitrous oxide emissions, offsetting reductions in climate radiative forcing. *Climatic Change* **72**:321-338. doi: 10.1007/s10584-005-6791-5.

Marland G, West T O, Schlamadinger B, Canella L (2003) Managing soil organic carbon in agriculture: the net effect on greenhouse gas emissions. *Tellus Series B-Chemical and Physical Meteorology* **55**:613-621. doi: 10.1034/j.1600-0889.2003.00054.x.

McCarthy J J (2001) *Climate change: impacts, adaptation, and vulnerability : contribution of Working Group II to the third assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK; New York. x 1032 p.

Mishra U, Ussiri D A N, Lal R (2010) Tillage effects on soil organic carbon storage and dynamics in Corn Belt of Ohio USA. *Soil & Tillage Research* **107**:88-96. doi: 10.1016/j.still.2010.02.005.

Pan G X, Li L Q, Wu L S, Zhang X H (2004) Storage and sequestration potential of topsoil organic carbon in China's paddy soils. *Global Change Biology* **10**:79-92. doi: 10.1111/j.1365-2486.2003.00717.x.

Pathak H, Byjesh K, Chakrabarti B, Aggarwal P K (2011) Potential and cost of carbon sequestration in Indian agriculture: Estimates from long-term field experiments. Field Crops Research 120:102-111. doi: 10.1016/j.fcr.2010.09.006.

Paustian K, Six J, Elliott E T, Hunt H W (2000) Management options for reducing CO2 emissions from agricultural soils. *Biogeochemistry* **48**:147-163. doi: 10.1023/A:1006271331703.

Smith P, Powlson D S, Glendining M J, Smith J U (1998) Preliminary estimates of the potential for carbon mitigation in European soils through no-till farming. *Global Change Biology* **4**:679-685. doi: 10.1046/j.1365-2486.1998.00185.x.

Srinivasarao C, Deshpande A N, Venkateswarlu B, Lal R, Singh A K, Kundu S, Vittal K P R, Mishra P K, Prasad J V N S, Mandal U K, Sharma K L (2012) Grain yield and carbon sequestration potential of post monsoon sorghum cultivation in Vertisols in the semi arid tropics of central India. *Geoderma* **175**:90-97. doi: 10.1016/j.geoderma.2012.01.023.

Srinivasarao C, Lal R, Kundu S, Babu M B, Venkateswarlu B, Singh A K (2014) Soil carbon sequestration in rainfed production systems in the semiarid tropics of India. *Sci Total Environ* **487**:587-603. doi: 10.1016/j.scitotenv.2013.10.006.

Sun W J, Huang Y, Zhang W, Yu Y Q (2010) Carbon sequestration and its potential in agricultural soils of China. *Global Biogeochemical Cycles* **24**. doi: Artn Gb3001,Doi 10.1029/2009gb003484.

Tang H J, Qiu J J, Van Ranst E, Li C S (2006). Estimations of soil organic carbon storage in cropland of China based on DNDC model. *Geoderma* **134**:200-206. doi: 10.1016/j.geoderma.2005.10.005.

Theng (1979) *Formation and properties of clay-polymer complexes*. Elsevier, Amsterdam.

Vasques GM, Grunwald S, Sickman JO (2008) Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. *Geoderma* 146(1-2):14-25.

Vleeshouwers L M, Verhagen A (2002) Carbon emission and sequestration by agricultural land use: a model study for Europe. *Global Change Biology* **8**:519-530. doi: 10.1046/j.1365-2486.2002.00485.x.

West T O, Post W M (2002) Soil organic carbon sequestration rates by tillage and crop rotation: A global data analysis. *Soil Science Society of America Journal* **66**:1930-1946.

Xu S, Shi X, Zhao Y, Yu D, Li C, Wang S, Tan M, Sun W (2011) Carbon sequestration potential of recommended management practices for paddy soils of China, 1980-2050. *Geoderma* **166**:206-213. doi: 10.1016/j.geoderma.2011.08.002.

Zhang F, Li C, Wang Z, Wu H (2006) Modeling impacts of management alternatives on soil carbon storage of farmland in Northwest China. *Biogeosciences* **3**:451-466.

Zimmerman P R, Price M H, Updegraff K L, Capehart W J (2004) C-lock: An online system to maximize the value of agricultural carbon sequestration for producers and purchasers. *Abstracts of Papers of the American Chemical Society* **227**:U1094-U1094.

Part III
Soil Sensors and Legacy Data

Chapter 20
Digital Soil Morphometrics via a Low-Cost Radiometer for Estimating Soil Organic Carbon and Texture

Alexandre ten Caten, Ricardo Simão Diniz Dalmolin, André Carnieletto Dotto, Jean Michel Moura-Bueno, Evandro Loch Boeing, Jose Lucas Safanelli, Walquiria Chaves Silva and Bruno Fellipe Bottega Boesing

Abstract There is scientific evidence toward the incorporation, in a near feature, of diffuse reflectance spectroscopy (DRS) as an everyday laboratory tool for soil attribute determination. Nevertheless, research still has to be conducted toward the capabilities of limited ranges of the spectra (i.e., 325–1075 nm), as well as the use of more affordable spectrometers. This study aimed at evaluating the capacity of a 15,000 USD spectrometer for estimating soil organic carbon (SOC) and texture. Soil

A. ten Caten (✉)
Agriculture, Biodiversity and Forest Department, Federal University of Santa Catarina, Campus Curitibanos, Ulysses Gaboard highway, km3, Curitibanos, SC 89520-000, Brazil
e-mail: alexandre.ten.caten@ufsc.br

R.S.D. Dalmolin
Soil Department, Federal University of Santa Maria, 1000 Roraima Avenue, Santa Maria, RS 97105-900, Brazil
e-mail: dalmolin@ufsm.br

A.C. Dotto · J.M. Moura-Bueno
Federal University of Santa Maria, 1000 Roraima Avenue, Santa Maria, RS 97105-900, Brazil
e-mail: andrecdot@gmail.com

J.M. Moura-Bueno
e-mail: bueno.jean1@gmail.com

E.L. Boeing · J.L. Safanelli · W.C. Silva · B.F.B. Boesing
Federal University of Santa Catarina, Campus Curitibanos, Ulysses Gaboardi highway, km3, 89520-000 Curitibanos, SC, Brazil
e-mail: dro.loch@gmail.com

J.L. Safanelli
e-mail: zecojls@gmail.com

W.C. Silva
e-mail: walquiria.chs@gmail.com

B.F.B. Boesing
e-mail: brunofellipebb@gmail.com

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_20

samples were collected in 10 Ferralsol profiles of basaltic parental material in Serra Geral Formation in southern of Brazil. Spectral signatures were collected in 45 air-dried soil samples previously sieved through 2-mm mesh and 45 soil samples grounded in an agate mortar. Sample preparation through pestle grounding showed a slight gain in modeling accuracy. The best results of partial least squares regression (PLSR) were achieved for SOC with an error of prediction of 2.44 g kg^{-1}, R^2 of 0.88, and RPD of 2.85. These results are an indication of the applicability of a low-cost spectrometer for soil attribute determination through DRS. This approach could lead to a wider adoption of the technique, especially in laboratories were there are budget limitations and are in need of this important soil attribute determination.

Keywords Diffuse reflectance spectroscopy · Soil reflectance · Proximal soil sensing · Digital soil mapping · Near-infrared measurement

20.1 Introduction

In recent years, researchers have been evaluating a wide range of possibilities to increase soil scientists' capacity in collecting data to attend an increasing demand for soil information for environmental modeling. Particle size and soil organic carbon (SOC) are two attributes of fundamental importance when defining soil use and management. With the increasing demand for food and energy, knowledge of the physical and chemical soil characteristics imposes a greater ability to sample this natural resource which renews at such slow pace.

In this context, digital soil mapping (DSM) has proven to be an efficient approach for building soil class and properties datasets (McBratney et al. 2003). To help in this task, proximal soil sensing (PSS) has facilitated the collection of a larger amount of soil spatial data using faster and less laborious techniques (Viscarra Rossel et al. 2009). Having reached a mature level of acceptance and application by soil science community, both approaches toward soil mapping, DSM and PSS, have recently given birth to a new discipline in soil science: Digital soil morphometrics (DSMh) (Hartemink and Minasny 2014).

DSMh takes advantage of an enormous evolution of PSS equipments. Attempts have already been made to measure properties and attributes of soil profiles in situ (Viscarra Rossel et al. 2009; Waiser et al. 2007) and ex situ in laboratory-controlled conditions (Vasques et al. 2014). It is claimed that soil attributes such as horizons, texture, color, structure, moisture, redoximorphic features, consistence, carbonates, rock fragments, and pores can be determined by PSS in the DSMh Pedology approach (Hartemink and Minasny 2014).

Diffuse reflectance spectroscopy (DRS) operating in visible and near-infrared (VNIR) spectral region has gained attention as a PPS technique that could deliver soil data with required speed and accuracy even for the development of on-the-go sensors. The literature has shown the evidences of full-spectrum data compression techniques' capabilities, such as partial least squares regression (PLSR) toward a deep understanding of soil-spectroscopy relationships (Viscarra Rossel et al. 2010).

In the spectral region of 325–1075 nm, very important spectrum features, related to the presence of iron oxides such as goethite and hematite, are located (Stenberg and Viscarra Rossel 2010). Humic acids as part of SOC fraction can also be responsible for broad absorptions in visible portion of spectra. Using a spectrometer in the spectral region of 325–1075 nm, Melendez-Pastor et al. (2008) found a high correlation ($r > 0.75$) for soil spectral signature in visible region (380–700 nm) and its attributes such as silt, sand, electrical conductivity, carbonates, and organic matter, showing DRS as a reliable technic for collecting soil data more quickly and with little environment impact. Viscarra Rossel et al. (2006) have demonstrated the potential of DRS operating in visible region (400–700 nm) for organic carbon prediction, with a RMSE of 0.18 dag kg^{-1} and R^2 of 0.60. Authors have highlighted that the cost of mid-infrared equipment (25,000–2500 nm) could not be justified for carbon prediction, since predictions using only the visible region are comparably accurate and not as expensive as the former.

One of the drawbacks for a wider application of PPS is the equipment cost. VNIR (400–2500 nm) spectrometers might cost as much as 60,000 USD. This underlines the importance of research into the use of more affordable PSS equipment. Besides, there has to be broader collaboration toward development of PSS technique in a worldwide range of soils. The objective of this study was to evaluate the capacity of a 15,000 USD spectrometer, with limited spectral range acquisition (325–1075 nm), for estimating SOC and texture.

20.2 Materials and Methods

A total of 10 Ferralsol profiles were morphological described in the southwest part of Marombas River watershed located near the center of Santa Catarina State, south of Brazil (Fig. 20.1). Parental material in the region consists mainly of basaltic rocks of Serra Geral Formation. The climate is subtropical with mild summer and mean annual temperatures of 16 °C. Köppen climate classification system for the area is Cfb. Annual precipitation is about 1.600 mm. Altitude of watershed varies from 900 to 1300 m above sea level. Natural vegetation belongs to the Mixed Ombrophylous Forest (or Araucaria Forest). The total area of the watershed is approximately 950 km^2, and predominant land cover consists of 22 % of agriculture (garlic, onion, soy beans, and maize), 37 % of cultivated forest (*Pinus taeda*), 33 % of natural forest (with *Araucaria angustifolia*), and 8 % of grassland and pasture. Prevalent soil types are Ferralsol, Nitisol, Cambisol, Leptosols, and Regosols.

All 10 profiles were sampled and classified following the Brazilian System of Soil Classification (SiBCS). Every profile was sampled in its pedogenetic soil horizons following SiBCS, and in total, 45 soil samples were collected from top until to 2 m deep. In each profile, morphological features as soil color were recorded, using a Munsell® color book, for the purpose of soil classification. Chemical and physical attributes were determined in 45 soil samples after air-dried, grounded, and sieved through 2 mm mesh according to Embrapa (1997). Half of

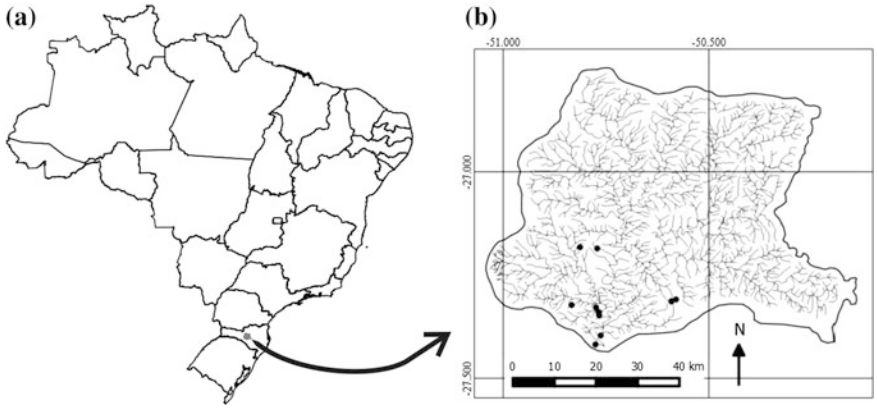

Fig. 20.1 **a** Location of study area in the southern state of Santa Catarina in Brazil. **b** Location of the 10 sampled Ferralsol profiles in Marombas River watershed

every soil sample was used for grounding in an agate mortar, for at least 10 min, in order to test micro-aggregation influence.

Diffuse reflectance spectroscopy of sieved and grounded samples was collected with a spectrometer FieldSpec HandHeld II (ASD Inc.) with a spectrum range acquisition of 325–1075 nm and spectral resolution of <3 nm at 700 nm. Soil scanning was conducted inside a black painted box (dimensions L/750 × H/400 × W/400 mm), in order to allow the illumination to be controlled (Fig. 20.2). Inside the box, soil samples were put in a Petri dish. Spectrometer was installed on the top of the box with a conical field of view of 10° in a distance of 400 mm from samples. With this

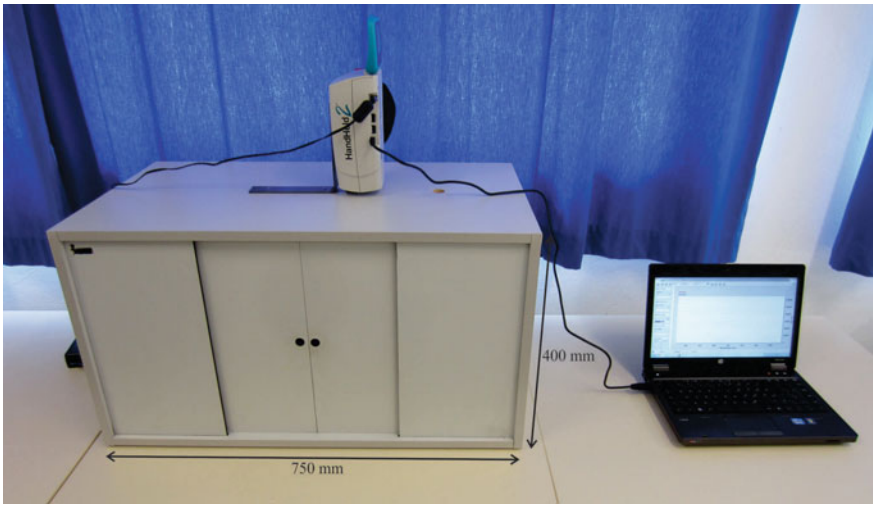

Fig. 20.2 Spectrometer mounted on the top of the sampling box for sample illumination control

configuration, the spectrometer sampling area in the Petri dish was 40.7 cm^2. A light source of 70 W quartz–tungsten–halogen lamp with integrated reflector was placed inside the box. Light source was placed 400 mm away from the soil sample and inclined 30° from lamp nadir. Four composite scans (each one is an average of 100 internal scans) were obtained for each sample from the four quadrants of Petri dish by rotating it 90°. Final spectrum was calculated by averaging all four composite scans.

Before statistical analyses, spectra noisy ends were removed (325–400 and 980–1075 nm) (Fig. 20.3a). The remaining reflectance spectra were smoothed by a Savitzky–Golay second-order polynomial across a moving window of nine bands (Torrent and Barrón 2002). Furthermore, an average across a 5 nm moving window was applied to all spectra. Savitzky–Golay first-order derivatives were calculated on resulting soil reflectance spectra using a first-order polynomial across a nine-band moving window (Fig. 20.3b). Equally, Savitzky–Golay second-order derivativeswere calculated on soil reflectance spectra using a second-order polynomial across a nine-band moving window (Fig. 20.3c).

First- and second-order derivatives have the capability of eliminating baseline fluctuation and background noise, and at the same time, they enhance the reflectance features through peak values. Both derivatives procedures followed the best results achieved by Vasques et al. (2008) in preprocessing spectral curves tests. Derivatives were calculated in sieved (<2 mm) and agate mortar grounded

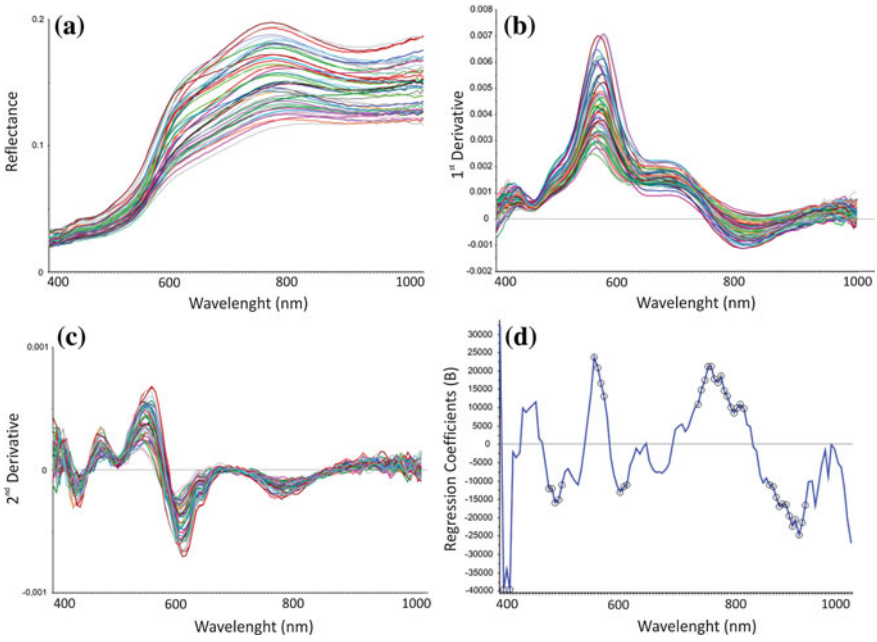

Fig. 20.3 **a** Reflectance data of the 45 soil samples; **b** 1st derivative of reflectance; **c** 2nd derivative of reflectance; **d** regression coefficients in PLSR for clay prediction. All graphs were produced by data collected in sieved (<2 mm) soil samples

reflectance samples. Dependent variable dataset was formed by SOC and texture analyses. Reflectance first- and second-order derivatives for sieved and grounded samples formed the independent variable dataset.

The multivariate technique PLSR was used for data modeling. PLSR takes the advantage of spectral information and relates it to soil attributes, which can be seen through peaks of higher regression coefficients (positive or negative) in some wavelength when spectral data are applied to clay prediction, specially at 420, 550, 730, and 880 nm (small circles on Fig. 20.3d). PLSR is a similar approach to principal components analysis (PCA) reducing the dimensionality of a large number of potential correlated variables, thus avoiding the problems of multicollinearity and minimizing the lost of information of original variables. Furthermore, PLSR has the advantage that it also takes into consideration the dependent variables, in this case soil attributes information, when calculating the principal components. PLSR were computed using the orthogonalized PLSR algorithm and evaluated through a cross-validation test. Tabulated statistics were R^2, root-mean-squared error for cross-validation (RMSECV) and ratio of standard deviation of soil attributes to RMSECV (RPD). PLSR was used for data modeling in the Unscrambler X 10.3 software (CAMO Technologies Inc., Woodbridge, NJ).

20.3 Results and Discussion

Soil samples had an average of 50.3, 253.0, and 696.6 g kg^{-1} of sand, silt, and clay, respectively (Table 20.1). The soil texture content of samples characterizes those soils as being of clay soil class. The calcic plagioclase and pyroxene basalt weathered completely and formed clay minerals through hydrolysis process of the parental material contributing to the soil texture.

Attribute SOC reached a maximum of 23.78 g kg^{-1} (Table 20.1) due to constant supply of new organic material in vegetated areas. The altitude of the region causes annual average temperature to be around 16 ºC, thus collaborating to maintain a higher SOC content in top layers. Clay soil texture also plays a role in protecting organic carbon from declining through covering of organic molecules. It was found that SOC decreases with depth, a fact which follows from continuing deposition of organic matter on superficial horizons and low solubility of humic and fulvic acids to migrate in depth into the soil profile.

Table 20.1 Descriptive statistics of 45 soil samples used in the study

Attribute	Minimum	Mean	Maximum	Standard deviation
Sand (g kg^{-1})	15.10	50.35	154.95	35.62
Silt (g kg^{-1})	165.43	253.02	355.00	46.85
Clay (g kg^{-1})	556.39	696.65	777.09	55.66
SOC (g kg^{-1})	1.74	12.56	23.78	7.04

SOC soil organic carbon

Table 20.2 Summary of PLSR results for the four soil attribute datasets

Soil attribute	RMSECV	R^2	RPD
1st derivative <2 mm sieved samples			
Sand (g kg^{-1})	21.66	0.63	1.62
Silt (g kg^{-1})	35.80	0.44	1.29
Clay (g kg^{-1})	36.27	0.58	1.52
SOC (g kg^{-1})	2.44	0.88	2.85
1st derivative pestle grounded samples			
Sand (g kg^{-1})	20.19	0.69	1.74
Silt (g kg^{-1})	28.69	0.64	1.60
Clay (g kg^{-1})	36.60	0.58	1.51
SOC (g kg^{-1})	2.77	0.84	2.51
2nd derivative <2 mm sieved samples			
Sand (g kg^{-1})	22.26	0.60	1.58
Silt (g kg^{-1})	30.05	0.60	1.54
Clay (g kg^{-1})	35.40	0.60	1.55
SOC (g kg^{-1})	2.62	0.86	2.66
2nd derivative pestle grounded samples			
Sand (g kg^{-1})	21.93	0.62	1.61
Silt (g kg^{-1})	27.27	0.66	1.70
Clay (g kg^{-1})	32.04	0.67	1.70
SOC (g kg^{-1})	2.95	0.82	2.36

SOC soil organic carbon

Generally, preparation of soil samples through agate mortar grounding represented a slight improvement in PLSR prediction capability for soil texture (Table 20.2). Regarding derivatives, soil attributes showed different results among the three studied particle sizes. Clay prediction best performed when soil was grounded and second-order derivative applied (RMSECV, R^2, and RPD of 32.04 g kg^{-1}, 0.67, and 1.70, respectively). This was also the case for silt which was best predicted through grounded soil samples and second-order derivative (RMSECV, R^2, and RPD of 27.27 g kg^{-1}, 0.66, and 1.70, respectively). On the other hand, sand prediction achieved best results with first-order derivatives and grounded samples (RMSECV, R^2, and RPD of 20.19 g kg^{-1}, 0.69, and 1.74, respectively). Stenberg and Viscarra Rossel (2010) state that soil coarser structure increases scattering and reduces reflection, which could ultimately lead to poorer model predictions. Our results showed that soil grounding produced a finer structure and improved prediction of soil texture in weathered soils such as Ferralsols.

Further preparation of samples through grounding showed no improvement for SOC prediction. Best results were achieved for this attribute through 2-mm sieved samples and first-order derivatives, reaching values of RMSECV, R^2, and RPD of 2.44 g kg^{-1}, 0.88, and 2.85, respectively. Taking into account that soil grounding in agate mortar requires an extra effort in soil preparation, DRS can be collected in sieved (<2 mm) Ferralsols soil samples without loss of prediction power of SOC.

Results for SOC are superior to the ones found by Nocita et al. (2013). These authors, using spectral data from 350–2500 nm, determined SOC with a RMSE of 4.72 g kg^{-1} and a R^2 0.78. This shows that, even using a limited part of spectra, promising prediction results for SOC could be delivered with a low-cost spectrometer. Waiser et al. (2007) reported an R^2 of 0.81 and an error of prediction for fine clay (<2 mm) of 34 g kg^{-1} when using first-order derivative of visible near-infrared reflectance spectra (350–2500 nm). RPD values, from 1.29 to 1.74, for soil texture show that using a more affordable equipment deserves further investigation (Table 20.2).

RPD values for SOC varied from 2.36 to 2.85, with mean 2.59, showing the potential of a low-cost spectrometer for this important biological soil attribute determination. According to Vasques et al. (2008), these high RPD values indicate that models are robust enough to predict SOC when applied to soils from the same geographical area and within the same characteristics Thus, modeling SOC with a limited spectra spectrometer could also take profit of national and/or global soil spectral libraries, nevertheless using only part of the vis-NIR region of the spectra.

This research on the potential of a low-cost spectrometer for soil attribute determination through DRS is still being conducted. Sampling intensity will be increased 10 fold in the near future. Besides Ferralsols, the soil classes Nitisol, Cambisol, Leptosols, and Regosols will also be sampled. Following the recommendations of Vasques et al. (2010), an approach through generating PLSR models separately, for every soil class, might improve the prediction capabilities. This could be spatially important for a limited range spectrometer like the one used in this study. Tests will also be carried out toward evaluating the potential of the 400–980-nm spectral region for in situ soil attribute determination, specially organic carbon content.

20.4 Conclusions

We have demonstrated the capabilities of a cheaper spectrometer operating in spectral range of 325–1075 nm in predicting soil attributes sand, silt, clay, and SOC. Through PLSR, clay content was predicted to an accuracy of 32.04 g kg^{-1}, R^2 of 0.67, and RPD value of 1.70. Results for the prediction of SOC content reached and accuracy of 2.44 g kg^{-1}, R^2 of 0.88, and RPD of 2.85. The literature shows that full range VNIR spectrometers (400–2500 nm) have reached higher absolute accuracy values than those demonstrated here. However, our results have to be considered if there are budget constraints and lower cost equipment is the only option available.

Further research has to be conducted toward the adequacy of cheaper spectrometers for soil attribute prediction, as well as possible preprocessing transformation options which could improve the prediction capability.

Acknowledgements This study was supported by the Foundation for Research Support of Santa Catarina State (FAPESC) n°2012000094 with funds for equipment acquisition and soil analysis. The National Council of Technological and Scientific Development (CNPq) financed the research grant for co-author (1) number 442718/2014-4 and for co-author (2) as well as undergraduate fellowships of co-authors (4). First author and co-authors (3) acknowledge support of the Coordinating Office for the Advancement of Higher Education (CAPES) for graduation fellowships and support for paper presentation at 6th Global Workshop on Digital Soil Mapping. We are also in debt with the anonymous reviewer for his (hers) useful comments which improved the quality of the manuscript significantly.

References

Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA (1997) Centro Nacional de Pesquisa de Solos. Manual de Métodos de Análise de Solo. 2. ed., Rio de Janeiro

Hartemink AE, Minasny B (2014) Towards digital soil morphometrics. Geoderma 230–231:305-317. doi: 10.1016/j.geoderma.2014.03.008

McBratney AB, Mendonça Santos ML, Minasny B (2003) On digital soil mapping. 117(1-2): 3-52 doi: 10.1016/S0016-7061(03)00223-4

Melendez-Pastor I, Navarro-Pedreño J, Gómez I, Koch M (2008) Identifying optimal spectral bands to assess soil properties with VNIR radiometry in semi-arid soils. Geoderma 147(3-4): 126-132 doi: 10.1016/j.geoderma.2008.08.004

Nocita M, Stevens A, Noon C, Wesemael B van (2013) Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy. Geoderma 199:37-42 doi: 10.1016/j.geoderma.2012.07.020

Stenberg B, Viscarra Rossel RA (2010) Diffuse reflectance spectroscopy for high-resolution soil sensing In: Viscarra Rossel RA, McBratney AB, Minasny B (Eds.) Proximal Soil Sensing pp 29-47. Springer, Amsterdam

Torrent J., Barrón, V. (2002) Diffuse reflectance spectroscopy of iron oxides. Encyclopedia of Surface and Colloid Science. New York: Marcel Dekker, Inc. 1438-1446.

Vasques GM, Demattê JAM, Viscarra Rossel RA, Ramírez-López L, Terra, FS (2014) Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths. Geoderma 223-225:73-78 doi: 10.1016/j.geoderma.2014.01.019

Vasques GM, Grunwald S, Harris WG (2010) Spectroscopic models of soil organic carbon in Florida, USA. J Environ Qual 39(2): 923-934 doi: 10.2134/jeq2009.0314

Vasques GM, Grunwald S, Sickman JO (2008) Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma 146(1-2): 14-25 doi: 10.1016/j.geoderma.2008.04.007

Viscarra Rossel RA, Cattle SR, Ortega A, Fouad Y (2009) In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. Geoderma 150(3-4): 253-266 doi: 10.1016/j.geoderma.2009.01.025

Viscarra Rossel RA, McBratney AB, Minasny B (2010) Proximal Soil Sensing. Springer, Amsterdam

Viscarra Rossel RA, Walvoort DJJ, McBratney AB, Janik LJ, Skjemstad JO (2006) Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties Geoderma 131(1-2): 59-75 doi: 10.1016/j.geoderma.2005.03.007

Waiser TH, Morgan CLS, Brown DJ, Hallmark CT (2007) In Situ Characterization of Soil Clay Content with Visible Near-Infrared Diffuse Reflectance Spectroscopy. Soil Sci Soc Am J 71(2): 389-396 doi: 10.2136/sssaj2006.0211

Chapter 21
Transferability and Scaling of VNIR Prediction Models for Soil Total Carbon in Florida

Congrong Yu, Sabine Grunwald and Xiong Xiong

Abstract The assessment of soil total carbon (TC) across large land areas is critical to derive global and regional soil carbon budgets and better understand the interactions between carbon and other biogeochemical cycles. But the cost and time involved in measurements of TC with standard laboratory methods are impractical. Research has suggested that visible/near-infrared (VNIR) diffuse reflectance spectroscopy can provide robust and accurate estimations for TC. The applicability, transfer, and scalability of VNIR-derived soil models are still poorly understood. The objectives of this study in Florida, USA, were to (i) compare two methods to predict soil TC using five fields (local scale) and a pooled (regional scale) VNIR spectral dataset, (ii) assess the model's transferability among fields, and (iii) evaluate the up- and downscaling behavior of TC prediction models. A total of 560 TC-spectral sets were modeled by partial least squares regression (PLSR) and support vector machine (SVM). The transferability and up- and downscaling of models were limited by the following factors: (i) the spectral data domain, (ii) soil attribute domain, (iii) methods that describe the internal model structure of VNIR-TC relationships, and (iv) environmental domain space of attributes that control soil carbon dynamics. All soil logTC models showed excellent performance based on both methods (PLSR and SVM) with $R^2 > 0.86$, bias < 0.01 %, root-mean-square prediction error (RMSE) = 0.09 %, residual predication deviation (RPD) > 2.70 %, and ratio of prediction error to inter-quartile range (RPIQ) > 4.54. PLSR performed substantially better than SVM to scale and transfer models. Upscaled soil TC models performed somewhat better in terms of model fit (R^2), RPD, and RPIQ, whereas downscaled models showed less bias and smaller RMSE based on PLSR. But no universal trend was found indicating which of the four investigated factors (i–iv) had the most impact that constraints transferability and scalability. The findings from this study have implications for the development of

C. Yu · S. Grunwald (✉) · X. Xiong
Department of Soil and Water Science, University of Florida, Gainesville, USA
e-mail: sabgru@ufl.edu

C. Yu
State Key Lab of Hydrology-Water Resources and Hydraulic Engineering,
College of Hydrology and Water Resource, Hohai University, Nanjing, Jiangsu, China

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_21

'universal' spectral-based soil models aiming to predict soil properties for a diverse set of different soils formed in different environmental conditions covering a wide range of geographic settings, at its extreme the whole globe. Those 'universal' spectral libraries are based on the premise that soil predictions (e.g., soil TC) can be made anyplace because they are built using soil spectral datasets that characterize exhaustively the attribute feature space. This assertion is limited by the fact that a large number of interacting factors of soils, spectra, and environmental properties are needed to represent the exhaustive sample population which has not materialized yet. Given the many factors that can impinge on empirically derived soil spectral prediction models, as demonstrated by this study, more focus on the applicability and scaling of them is needed.

Keywords Soil organic carbon · Visible/near-infrared spectroscopy · Transferability · Scalability · Modeling

21.1 Introduction

Research has suggested that visible/near-infrared (VNIR) diffuse reflectance spectroscopy can provide robust and accurate estimations for TC and carbon fractions (Viscarra Rossel et al. 2006; Vasques et al. 2009, 2010; Nocita et al. 2011; Sarkhot et al. 2011; McDowell et al. 2012a, b). Spectral soil carbon models are poised to contribute to spatially explicit regional and global carbon assessment. However, knowledge gaps still exist in terms of the prediction quality across different soils and landscapes, transferability, and scalability of such models. Scaling and transfer concepts and their implications for modeling were presented by Blöschl and Sivapalan (1995), Wu et al. (2006), and Grunwald et al. (2011).

Given the multitude of potential factors that may impact the application of VNIR soil carbon models to make predictions for unknown samples, the underlying motivation for this research was to design an experimental study to investigate the transfer and up- and downscaling behavior of soil TC-VNIR models. The specific objectives were to (i) compare the performance of two methods to predict soil TC using five fields (local) and a pooled (regional) VNIR spectral dataset, (ii) assess the model's transferability among five representative field sites in Florida, (iii) evaluate the upscaling and downscaling behavior of TC prediction models, and (iv) examine the constraining factors in model transferability and scaling.

21.2 Data and Methods

Five fields (each of size ~ 0.25 km^2) were selected that represent prominent soil–land-use types in Florida, USA (Xiong 2013). Table 21.1 provides a description of the main landscape characteristics of each field. A total of 112 samples (0–20 cm depth)

Table 21.1 Characteristics of the five fields

	Variables	Study areas				
		Field 1	Field 2	Field 3	Field 4	Field 5
	Sampling location	Ordway-Swisher Biological Station	San Felasco Hammock Preserve State Park	Econfina Creek Water Management area	Santa Fe River Ranch	Myakka River State Park
	Longitude	81° 59′9″W	82° 27′31″W	85° 33′51″W	82° 29′40″W	82° 17′16″W
	Latitude	29° 41′23″N	29° 43′59″N	30° 26′42″N	29° 55′45″N	27° 11′22″N
Topography	Elevation (m)[a]	42.8	43.5	23.9	28.8	8.7
	Slope (%)[a]	1.2	1.2	2.9	2.4	0.2
Climate	Max temperature (°C)[b]	27.5	27.1	26.3	27.2	29.2
	Min temperature (°C)[b]	14.0	13.8	12.9	13.6	16.3
	Precipitation (mm)[b]	1325	1345	1634	1360	1464
Parent material	Surficial geology[c]	Cypresshead	Coosawhatchie	Citronelle	Coosawhatchie	Shelly sediments of plio-pleistocene
Organism	Land use	Xeric upland forest	Mesic upland forest	Pineland	Improved pasture	Rangeland
	NPP (kg C m^{-2})[d]	7.91	13.60	9.07	7.50	8.13
	NDVI[d]	3.81	7.90	3.81	9.50	4.31
	Dry biomass (kg m^{-2})[e]	2.76	12.50	5.53	–	6.68
Soil	Soil suborder[f]	Psamments	Aquults-psamments-udepts-udults	Psamments	Udults	Aquods
	AWC (cm cm^{-1})[f]	1.2	2.1	1.5	2.2	1.7
	Clay content (%)[f]	1.2	5.2	3.7	4.6	1.9
	Sand content (%)[f]	98.6	93.2	93.1	90.8	96.8

Variable descriptions, abbreviations, and sources

[a]National Elevation Dataset (NED), United States Geological Survey (USGS) (1999)

[b]Long-term maximum and minimum annual average temperature, long-term annual average precipitation between 1971 and 2000 from Parameter-elevation Regressions on Independent Slopes Model (PRISM) climate group

[c]USGS 1998; Florida Fish and Wildlife Conservation Commission (2003)

[d]Net primary productivity (NPP), normalized difference vegetation index (NDVI) from Moderate-Resolution Imaging Spectroradiometer (MODIS) for North American Carbon Project 2005

[e]National Biomass and Carbon Dataset (NBCD) 2000

[f]Soil suborder, available water holding capacity at 0–25 cm (AWC), clay content and sand content at 0–20 cm from Soil Survey Geographic Database (SSURGO), Natural Resources Conservation Service (NRCS) (2009)

in each field were collected (whole dataset comprising five fields $n = 560$) using the unbalanced nested spatial sampling design described by Lark (2011). In each field, at first, nine main centers gridded at 200-m intervals were chosen to constitute the highest level of the hierarchy. Secondly, at each main center, one additional sampling point (subnode) was collected 67 m away in a random direction. In similar pattern, the 2nd, 3rd, and 4th hierarchical sampling points were fixed at locations 22, 7, and 2 m away from their parent nodes, respectively. The approximately threefold hierarchy has been proven to be effective in capturing soil variation and avoiding overlaps among different branches (Webster and Oliver 2007).

Soil TC content was measured by dry combustion in the laboratory using a Shimadzu TOC-5050 analyzer (Table 21.2). The samples were 2 mm sieved and then oven dried at 40–45 °C for 12 h before scanning with a QualitySpec Pro Spectroradiometer (Analytical Spectral Devices Inc., Boulder, CO) in the VNIR spectral range (350–2500 nm) with a 1-nm-interval spectral resolution. For each sample, four replicate scans were taken at each of the four quadrants of a petri dish by rotating the sample at angles of 90°. The spectrometer was recalibrated to remove the baseline at every 10 samples with white spectralon. An average reflectance spectral curve was obtained for each sample that was used for modeling. Two preprocessing transformations were applied to the soil reflectance curves: First, the reflectance curves were smoothed across a moving window of nine nm using the Savitzky–Golay algorithm with a third-order polynomial to reduce the random noise (Savitzky and Golay 1964). Second, the first-degree Savitzky–Golay derivative with a search window of seven measurements and second-order polynomial was applied to the smoothed curves.

Two different multivariate regression techniques were applied to develop spectral models to predict logTC: partial least squares regression (PLSR) (Martens and Næs 1989) and support vector machine (SVM) (Vapnik 2000). First, leave-one-out

Table 21.2 Descriptive statistics of measured soil total carbon (original values: TC, logarithm-transformed values: logTC)

		TC (%)						logTC (log %)			
Datasets	*n*	Min.	Median	Mean	Max.	CV	Skew.	Mean	SD	CV	Skew.
Whole	560	0.31	1.04	1.18	3.55	0.55	0.99	0.01	0.24	30.56	0.06
Field 1	112	0.32	0.56	0.59	1.12	0.28	1.02	−0.24	0.11	−0.47	0.37
Field 2	112	0.70	1.63	1.77	3.35	0.36	0.78	0.22	0.15	0.68	0.15
Field 3	112	0.31	0.62	0.68	2.32	0.42	3.28	−0.20	0.14	−0.70	1.28
Field 4	112	0.56	1.05	1.10	2.84	0.30	2.25	0.030	0.11	4.48	0.74
Field 5	112	1.02	1.69	1.76	3.55	0.26	0.80	0.23	0.11	0.47	0.09
CAL	392	0.33	1.02	1.17	3.55	0.55	1.06	0.01	0.23	31.77	0.11
VAL	168	0.31	1.07	1.19	3.21	0.55	0.86	0.01	0.24	28.53	−0.02
SUB-W	112	0.32	1.04	1.20	2.85	0.53	0.69	0.02	0.24	15.29	−0.04

CAL = the dataset used to calibrate the models; VAL = the dataset used to validate the models; SUB-W = the 112 observations randomly chosen from the five fields (Fig. 21.1); *n* = number of observations; SD = standard deviation; CV = coefficient of variation; skew. = skewness

(LOO) cross-validation was employed to evaluate the model performance of the CAL datasets (70 % or $n = 392$ of the whole dataset). Second, independent validation was used to assess the model performance using the VAL datasets (30 % or $n = 168$ of the whole dataset). The coefficient of determination (R^2) was used as the goodness-of-fit statistic. The root-mean-square error (RMSE), residual prediction deviation (RPD) (Williams 1987), ratio of performance to inter-quartile distance (RPIQ) (Bellon-Maurel et al. 2010), and bias (Davies and Fearn 2006) were provided as complementary error statistics to evaluate the performances of different prediction models.

The transferability and scalability analyses were conducted using PLSR and SVM models. In this study, the definitions of 'model transfer,' 'scale transformation,' and 'up-/downscaling' as provided by Blöschl and Sivapalan (1995) and Wu et al. (2006) were adopted. Hence, 'transferability' denotes the transfer (or application) of a VNIR-based soil TC prediction model (Models 1 to 5) developed at one field site (Fields 1, 2, 3, 4, and 5, respectively) to another field site (Fields 1, 2, 3, 4, and 5, respectively) (Fig. 21.1b). The model performance at calibration sites was assessed using LOO cross-validation, and transferability was assessed using R^2, RMSE, RPD, and RPIQ. In this paper, 'scalability' denotes a change in the extent (size) of the geographic area represented by models, 'upscaling' refers to an escalation of the area (i.e., from smaller to larger extent), and 'downscaling' refers to the contraction of the area (i.e., from larger to smaller extent) Wu et al. 2006) (Fig. 21.1c, d). In the scaling analysis, a pooled subset-whole (SUB-W) dataset was created ($n = 112$) randomly selected from the whole dataset ($n = 560$) (Fig. 21.1a). The observation size of the SUB-W was equal to that of each field ($n = 112$),

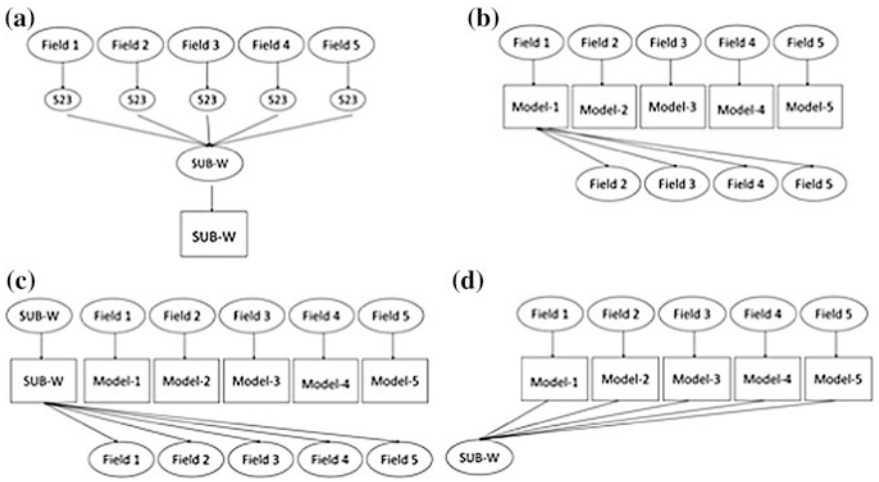

Fig. 21.1 The principle scheme of the transferability and scaling analysis: **a** the sample source of SUB-W dataset; **b** transferability at field scale; **c** downscaling analysis; **d** upscaling analysis. *Note* S23 in Fig. 21.1a represented the 23 samples randomly chosen from each of the five fields to calibrate the regional model

eliminating any bias or negative effects caused by the different sample sizes on the comparative analysis. To assess the downscaling behavior, the regional SUB-W models were applied to each of the five fields (Fig. 21.1c). And vice versa, to assess the upscaling performance, the TC models using PLSR and SVM developed for each of the five fields were applied to the regional SUB-W dataset (Fig. 21.1d). The same error statistics as outlined above were used to evaluate scaling behavior of TC models.

The Gower similarity coefficient (Gower 1971), as outlined in Mallavan et al. (2010), was employed to measure the similarity in soil-forming factors among fields according to Eq. (21.1). Important variables that were included in the similarity analysis are shown in Table 21.1.

$$S_{ij} = \frac{1}{p}\sum_{k=1}^{p}\left(1 - \frac{\left|x_{ik} - x_{jk}\right|}{\text{range } k}\right) \tag{21.1}$$

where S_{ij} is the Gower similarity coefficient between two sites i and j; k represents the soil-forming factors (i.e., environmental covariates); p is the number of variables; range k is the value range of variable k in the whole study area.

21.3 Results and Discussion

21.3.1 Prediction Performance of Spectral Prediction Models

The TC predictions derived from PLSR and SVM across all five fields (Table 21.3) and at the five field sites (Table 21.4) showed moderate performance. The R^2 was ≥0.86, and the RPIQ ≥4.54 for the whole dataset (Table 21.3). Brown et al. (2005) found that VNIR models developed using boosted regression trees (BRTs) outperformed PLSR to predict soil organic carbon (SOC) and soil TC, while McDowell

Table 21.3 Summary statistics for the spectral models of logTC produced by partial least squares regression (PLSR) and support vector machine (SVM) derived from calibration (CAL) using 70 % of all the samples (n = 392) and validation using 30 % of the samples (n = 168)

	LOO cross-validation using CAL		Validation using VAL				
	R^2	RMSE (log %)	R^2	Bias (log %)	RMSE (log %)	RPD	RPIQ
PLSR	0.88	0.08	0.86	0.004	0.09	2.70	4.54
SVM	0.87	0.09	0.88	0.01	0.09	2.78	4.67

LOO cross-validation = leave-one-out cross-validation; R^2 = coefficient of determination; RMSE = root-mean-squared deviations; RPD = residual prediction deviation; RPIQ = ratio of prediction error to inter-quartile range

Table 21.4 Summary statistics of leave-one-out cross-validation for partial least squares regression (PLSR) and support vector machine (SVM) models of logTC (log %) developed in SUB-W and the five field datasets

Models	PLSR		SVM	
	R^2	RMSE (log %)	R^2	RMSE (log %)
Model SUB-W	0.82	0.10	0.84	0.10
Model 1	0.69	0.06	0.55	0.08
Model 2	0.62	0.10	0.59	0.11
Model 3	0.46	0.10	0.33	0.11
Model 4	0.56	0.07	0.59	0.08
Model 5	0.61	0.07	0.52	0.08

R^2 = coefficient of determination; RMSE = root-mean-squared deviations; SUB-W = the 112 observations randomly chosen from the five fields (Fig. 21.1)

et al. (2012b) found no significant difference among PLSR and random forest (RF) ensemble regression trees to predict soil TC on Hawaiian soils. Minasny and McBratney (2008) and Minasny et al. (2009) in Australia found excellent predictions for SOC and TC using regression rules (Cubist approach). In contrast, Vasques et al. (2010) identified SOC predictions made by ensemble regression trees as more accurate than those derived from PLSR in an investigation in Florida. This suggests that depending on the geographic soil region, one method may outperform several others to make SOC or TC predictions from VNIR spectra.

21.3.2 Factors that Impact the Transferability and Scalability of Prediction Models

Overall, PLSR models performed better to transfer and scale than SVM models (Tables 21.5, 21.6, 21.7, and 21.8). This implies that linear relationships between VNIR spectra and soil TC (quantified by PLSR) were more pronounced than nonlinear, complex relationships (quantified by SVM). Reasons that constrain the transferability and scaling of soil prediction models may be explained by differences in the (i) spectral data domain space, (ii) soil attribute domain space, (iii) methods that determine the internal model structure of VNIR-TC relationships, and (iv) environmental domain space of attributes that control soil carbon dynamics (i.e., soil-forming factors).

21.3.2.1 Spectral Data Domain Space

The transferability and scaling of models may be also dependent on the spectral data domain. The VNIR models to predict TC selected variables in the spectral regions of the absorption features of C–H, N–H, and O–H groups, similar to the VNIR models presented by Vasques et al. (2008, 2009, 2010). These spectral

Table 21.5 The transferability of partial least squares regression (PLSR) models developed in one of the five study fields to predict the soil logTC (log %) of the other four fields

Models	Validation datasets (n = 112)	R^2	Bias (log %)	RMSE (log %)	RPD	RPIQ
Model 1	Field 2	0.53	−0.15	0.19	0.83	1.27
	Field 3	0.51	−0.16	0.21	0.66	0.68
	Field 4	0.17	−0.34	0.36	0.32	0.33
	Field 5	0.11	0.01	0.17	0.64	0.95
Model 2	Field 1	0.15	0.01	0.11	0.99	1.36
	Field 3	0.39	−0.10	0.18	0.74	0.77
	Field 4	0.15	−0.01	0.17	0.67	0.70
	Field 5	0.17	−0.23	0.27	0.40	0.59
Model 3	Field 1	0.12	0.28	0.31	0.37	0.51
	Field 2	0.09	−0.46	0.58	0.28	0.42
	Field 4	0.02	−0.21	0.34	0.34	0.35
	Field 5	0.02	0.01	0.20	0.54	0.80
Model 4	Field 1	0.34	0.34	0.35	0.32	0.44
	Field 2	0.29	0.05	0.15	1.09	1.67
	Field 3	0.32	0.19	0.23	0.59	0.61
	Field 5	0.34	0.19	0.21	0.51	0.75
Model 5	Field 1	0.24	0.37	0.39	0.29	0.41
	Field 2	0.28	0.05	0.16	0.98	1.49
	Field 3	0.25	−0.23	0.28	0.48	0.50
	Field 4	0.22	0.07	0.14	0.82	0.85

R^2 = coefficient of determination; RMSE = root-mean-squared deviations; RPD = residual prediction deviation; RPIQ = ratio of prediction error to inter-quartile range

signatures are produced by the overtones and combinations of absorption molecular vibrations (e.g., C–H, O–H, H_2O, and CO_{3^-}) in mid-infrared regions (Brown et al. 2005). The features associated with TC can be masked or distorted by Fe-oxides and secondary clays which are commonly found in soils (Clark 1999). This alludes to a critical issue of VNIR-modeling that other soil properties, such as texture, nutrient content, and minerals, may mask or interfere with the prediction of a given property of interest (e.g., soil TC) and thus impact the transferability of models. In this study, the soil texture differed only slightly among the five fields with sand content ranging between 90.8 and 98.6 % and clay content between 1.2 and 5.2 %. Hence, the effect of soil texture imposed on TC-spectral signatures was likely minor. However, the soil suborders differed among sites with Entisols (Psamments), Ultisols (Aquults, Udults), Inceptisols (Udepts), and Spodosols (Aquods) (Table 21.1), suggesting that the mineralogy, sesquioxides, and other chemical and physical soil properties differed substantially among sites.

Table 21.6 The transferability of support vector machine (SVM) models predicting soil logTC (log %) developed in one of the five study fields to predict the soil logTC (log %) of the other four fields

Model	Test datasets (n = 112)	R^2	Bias (log %)	RMSE (log %)	RPD	RPIQ
Model 1	Field 2	0.13	−0.47	0.49	0.32	0.49
	Field 3	0.12	−0.06	0.15	0.93	0.96
	Field 4	<0.01	−0.28	0.30	0.38	0.39
	Field 5	0.31	−0.48	0.49	0.22	0.33
Model 2	Field 1	0.06	0.39	0.40	0.28	0.39
	Field 3	0.17	0.37	0.39	0.35	0.36
	Field 4	<0.01	0.14	0.18	0.62	0.65
	Field 5	0.21	−0.06	0.12	0.88	1.30
Model 3	Field 1	0.33	0.09	0.14	0.81	1.12
	Field 2	0.01	−0.38	0.41	0.39	0.60
	Field 4	0.02	−0.19	0.22	0.52	0.54
	Field 5	0.27	−0.39	0.41	0.27	0.40
Model 4	Field 1	<0.01	0.28	0.30	0.38	0.52
	Field 2	<0.01	−0.18	0.24	0.67	1.02
	Field 3	0.06	0.23	0.27	0.51	0.53
	Field 5	0.18	−0.20	0.22	0.49	0.72
Model 5	Field 1	0.04	0.46	0.47	0.24	0.33
	Field 2	<0.01	0.00	0.16	1.00	1.53
	Field 3	0.05	0.41	0.43	0.32	0.33
	Field 4	<0.00	0.19	0.22	0.51	0.53

R^2 = coefficient of determination; RMSE = root-mean-squared deviations; RPD = residual prediction deviation; RPIQ = ratio of prediction error to inter-quartile range

21.3.2.2 Soil Attribute Domain Space

The soil attribute space, i.e., the upper and lower bounds and dispersion of soil TC used to build spectral-based prediction models (Table 21.2), may explain some of the transferability and scalability behavior of models. Typically, the soil attribute domain space expands as the geographic size of the modeled region increases (Grunwald et al. 2011). The range of soil TC values of CAL, VAL, and the SUB-W sets matched reasonably well the min. of 0.31 % and max. of 3.55 % of the Whole dataset. However, the differences in soil TC among field sites were profound. Ideally, the boundary conditions of attributes used for model development of a transfer function (or calibration spectral model) matches the boundary conditions of a transfer set. Brown et al. (2005) demonstrated the limitations of spectral-based model transfer to predict soil carbon in fields in Montana, USA, where the SOC values differed widely among field sites (min. of 1.93 g kg^{-1} to max. of 15.82 g kg^{-1}).

In this study, PLSR Models 1, 3, and 4 that resembled the TC range of SUB-W most closely with TC min. of 0.32 % and TC max. of 2.85 % did not show

Table 21.7 The downscaling performance of the partial least squares regression (PLSR) and support vector machine (SVM) models predicting soil logTC (log %) developed at regional scale (SUB-W) predicting samples at field scales

Model	Validation datasets (n = 112)	R^2	Bias (log %)	RMSE (log %)	RPD	RPIQ
PLSR models						
Model SUB-W (n = 112)	Field 1	0.42	<0.01	0.11	1.07	1.47
	Field 2	0.47	−0.02	0.13	1.27	1.93
	Field 3	0.32	0.07	0.16	0.87	0.90
	Field 4	0.51	−0.04	0.10	1.10	1.14
	Field 5	0.20	−0.03	0.16	0.68	1.00
SVM models						
Model SUB-W (n = 112)	Field 1	0.35	0.08	0.12	0.92	1.26
	Field 2	0.55	−0.08	0.13	1.19	1.81
	Field 3	0.26	0.08	0.14	0.95	0.99
	Field 4	0.65	0.01	0.07	1.63	1.69
	Field 5	0.51	−0.03	0.08	1.32	1.94

R^2 = coefficient of determination; RMSE = root-mean-squared deviations; RPD = residual prediction deviation; RPIQ = ratio of prediction error to inter-quartile range; SUB-W = the 112 observations randomly chosen from the five fields (Fig. 21.1)

persistent responses in terms of transferability (Table 21.5). For example, Model 3 (developed in Pineland and Psamments) failed to transfer well to Field 4, whereas the opposite was found for the transfer behavior of Model 4 (developed in Improved Pasture and Udults) to Field 3. These findings were confounded in down- and upscaling mode (Tables 21.7 and 21.8).

Besides the upper and lower bounds of attributes that matter for successful model transfer and scaling, it is also the internal variability (variance) of soil attributes that potentially impacts behavior. McBratney (1998) and Grunwald et al. (2011) asserted that an increase in the variance of soil attributes can impact the model building process, transferability, and scalability of soil properties. In this study, the coefficient of variation (CV) ranged from 0.26 % (Field 5) to 0.42 % (Field 3) which was lower than in the pooled sets (0.53 % in SUB-W and 0.55 % in whole, respectively). Effects of variability in TC on transfer and scalability of TC models are evident (compare CVs in Table 21.2 and results in Tables 21.5, 21.6, 21.7, and 21.8).

21.3.2.3 Model Structure

Regression methods use different strategies to relate predictors (here: spectral data) and a response variable (here: soil TC). The underlying strategies for predictor selection are different for PLSR and SVM impacting transfer and scale responses.

Table 21.8 The upscaling performance of the partial least squares regression (PLSR) and support vector machine (SVM) models predicting soil logTC (log %) developed at field scale predicting samples at regional scale (SUB-W)

Models	Validation dataset	R^2	Bias (log %)	RMSE (log %)	RPD	RPIQ
PLSR models						
Model 1	SUB-W (n = 112)	0.53	−0.12	0.22	1.09	1.97
Model 2		0.64	−0.03	0.15	1.58	2.87
Model 3		<0.01	−0.17	0.50	0.48	0.86
Model 4		0.57	0.18	0.23	1.02	1.84
Model 5		0.36	0.06	0.23	1.02	1.86
SVM models						
Model 1	SUB-W (n = 112)	0.08	−0.27	0.35	0.67	1.22
Model 2		0.19	0.15	0.27	0.90	1.63
Model 3		0.04	−0.18	0.29	0.81	1.48
Model 4		0.11	0.03	0.23	1.05	1.91
Model 5		0.15	0.21	0.31	0.77	1.40

R^2 = coefficient of determination; RMSE = root-mean-squared deviations; RPD = residual prediction deviation; RPIQ = ratio of prediction error to inter-quartile range; SUB-W = the 112 observations randomly chosen from the five fields (Fig. 21.1)

If the internal model structure that describes the relationship between spectral predictors and soil TC is not stable when it is scaled, it suggests scale-variant behavior. The PLSR and SVM models predicting logTC showed significant differences in the selection of spectral predictors in Models 1 to 5 and the SUB-W Model (results not shown). Thissen et al. (2004) has found major differences in the selection of spectral predictors that are inherent to the modeling process of PLSR and SVM, specifically in cases where the physico-chemical composition of the soil samples differed. In this study, PLSR was more robust than SVM to transfer models among sites. The PLSR models mainly focused on three regions to identify spectral predictors: ~350, ~1860, and ~2200 nm, which represented the reflection region of organic matter (Galvao and Vitorello 1998); O–H, water, C–H, C–N, C–O, N–H (Vasques et al. 2008); and calcium carbonate (2206 and 2341 nm) (Lagacherie et al. 2008), and various C–O (Brown et al. 2005). On the other hand, the top 50 important spectral wavelengths of the SVM models were found around ~670, ~1400, ~1800, and ~2200 nm. Although SVM is advantageous to model complex, high-dimensional spectral datasets because it can model nonlinear structures it performed poorly to transfer and upscale models (Tables 21.6 and 21.8). This can be explained by the high susceptibility of SVM to overfitting (Hernández et al. 2009). The substantially larger amount of spectral values selected as important in the SVM model compared to the PLSR model suggests overfitting.

21.3.2.4 Environmental Domain Space of Attributes

Soil carbon gains/losses have been linked to various environmental factors such as climate (Hook and Burke 2000), land use/land cover (John et al. 2005), soil moisture/hydrology (Vasques et al. 2012a, b), and topography (Yimer et al. 2006). Mallavan et al. (2010) asserted that the more similar regions (fields) are in terms of soil–environmental properties the more likely it is to successfully transfer a soil prediction model. The soil–environmental factors of fields differed widely in terms of topography, climate, parent material, organism/biota, and soils (Table 21.1). The homology among soil–environmental conditions explained a substantial amount of the ability to transfer TC models to other field sites and scales in this study (Fig. 21.2 and Table 21.9). Minasny et al. (2009) found that the transfer of mid-infrared spectral SOC prediction models among three different regions in Australia did not perform well due to differences in parent material and climate in which soils have formed in Queensland, New South Wales, and Victoria. Although the R^2 of transferred models were still moderate, all models showed significant bias. Studies that test not only for similarity in soil TC among sites, but also consider the similarity in environmental factors that form soil carbon are still rare in the soil science literature.

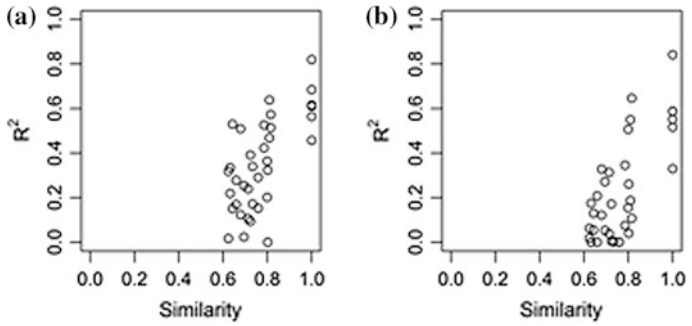

Fig. 21.2 The coefficient of determination (R^2) of each model transferred to other fields and scale versus the Gower similarity coefficient between the model development field/scale and the model application field/scale: **a** partial least squares regression (PLSR); **b** support vector machine (SVM)

Table 21.9 Gower similarity coefficients of soil–environmental factors among fields and across scales (SUB-W)

	Field 1	Field 2	Field 3	Field 4	Field 5	SUB-W
Field 1	1.00	0.64	0.68	0.73	0.71	0.78
Field 2	–	1.00	0.72	0.76	0.66	0.81
Field 3	–	–	1.00	0.62	0.69	0.80
Field 4	–	–	–	1.00	0.63	0.81
Field 5	–	–	–	–	1.00	0.80
SUB-W	–	–	–	–	–	1.00

SUB-W = the 112 observations randomly chosen from the five fields (Fig. 21.1)

21.4 Conclusions

This study showed that, although the spectral models to predict soil TC with different methods (PLSR and SVM) were successful in calibration and validation modes at five different fields nested within a large sand-dominated region in the USA, the transferability and up- and downscaling of models were limited by several factors. All of them interacted with each other impacting the transferability of models among field sites, upscaling, and downscaling behavior of spectral soil prediction models. These findings have implications for the development of 'universal' spectral-based soil models aiming to predict soil properties for a diverse set of different soils formed in different environmental conditions covering a wide range of geographic settings, at its extreme the whole globe. Those 'universal' spectral libraries are based on the premise that soil predictions (e.g., soil TC) can be made anyplace because they are built using soil spectral datasets that characterize exhaustively the attribute feature space. This assertion is limited by the fact that a large number of interacting factors of soils, spectra, and environmental properties are needed to represent the exhaustive sample population which has not materialized yet. Furthermore, the stationarity in mean and variance in local (field) calibrations of spectral soil prediction models are usually easier to meet though can have severe effects on scale-variant behavior of models at escalating spatial scales. The confounding trends in TC up- and downscaling behavior found in this study suggest that scale matters indicating the need for further soil scaling studies. Given the many factors that can impinge on empirically derived soil spectral prediction models, as demonstrated by this study, more focus on the applicability and scaling of them is needed. This study juxtaposed local and regional predictions, transferability, and scalability of soil TC models derived from VNIR spectra within a subtropical region in the southeastern USA. The constraints of soil spectral models identified in this research may also be found in other regions and spectral libraries that intent to have universal applicability.

Acknowledgements We thank the technical staff members of the Environmental Pedology Laboratory, Soil and Water Science Department for assistance with the soil carbon analysis.

References

Bellon-Maurel V, Fernandez-Ahumada E, Palagos B, Roger JM, McBratney AB (2010) Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC Trend. Anal. Chem. 29(9): 1073-1081.

Blöschl G., Sivapalan M (1995) Scale issues in hydrological modelling: A review. Hydrol. Process. 9(3-4): 251-290.

Brown DJ, Bricklemyer RS, Miller PR (2005) Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma 129 (3-4): 251-267.

Clark RN (1999) Spectroscopy of rocks and minerals, and principles of spectroscopy. p. 3-58. In Rencz, A.N. (ed.), Manual of remote sensing, Vol. 3, remote sensing for the earth science. John Wiley and Sons, New York, USA.

Davies AMC, Fearn T (2006). Back to basics: calibration statistics. Spectrosc. Eur. 18: 31-32.

Florida Fish and Wildlife Conservation Commission, FFWCC (2003) Florida vegetation and land cover data derived from Landsat ETM + imagery. Available at: http://myfwc.com/research/gis/data-maps/terrestrial/fl-vegetation-land-cover/.

Galvao LS, Vitorello I (1998) Role of organic matter in obliterating the effects of iron on spectral reflectance and colour of Brazilian tropical soils. Int. J. Remote Sens. 19(10): 1969-1979.

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27(4): 857-871.

Grunwald S, Thompson JA, Boettinger JL (2011) Digital soil mapping and modeling at continental scales: finding solutions for global issues. Soil Sci. Soc. Am. J. 75(4): 1201-1213.

Hernández N, Kiralj R, Ferreira MMC, Talavera I (2009) Critical comparative analysis, validation and interpretation of Support Vector Machine and Partial Least Square Regression models in a QSAR study on HIV-1 protease inhibitors. Chemometr. Intell. Lab. Syst. 98(1): 65-77.

Hook PB, Burke IC (2000) Biogeochemistry in a shortgrass landscape: control by topography, soil texture, and microclimate. Ecology 81(10): 2686-2703.

John B, Yamashita T, Ludwig B, Flessa H (2005) Storage of organic carbon in aggregate and density fractions of silty soils under different types of land use. Geoderma 128(1-2): 63-79.

Lagacherie P, Baret F, Feret JB, Madeira Netto J, Robbez-Masson JM (2008) Estimation of soil clay and calcium carbonate using laboratory, field and airborne hyperspectral measurements. Remote Sens. Environ. 112(3): 825-835.

Lark RM (2011) Spatially nested sampling schemes for spatial variance components: Scope for their optimization. Comput. Geosci. 37(10): 1633-1641.

Mallavan BP, Minasny B, McBratney AB (2010) Homosoil, a methodology for quantitative extrapolation of soil information across the globe. p. 137–150. In Boettinger, D.J.L., Howell, D.W., Moore, A.C., Hartemink, P.D.A.E., Kienast-Brown, S. (eds.), Digital soil mapping. Progress in soil science. Springer, The Netherlands.

Martens H, Næs T (1989) Multivariate calibration. 1st ed. John Wiley & Sons.

McBratney AB (1998) Some considerations on methods for spatially aggregating and disaggregating soil information. Nutrient Cycling in Agroecosystems 50(1): 51-62.

McDowell ML, Bruland GL, Deenik JL, Grunwald S (2012a) Effects of subsetting by carbon content, soil order, and spectral classification on prediction of soil total carbon with diffuse reflectance spectroscopy. Appl. Environ. Soil Sci. 2012. Available at http://www.hindawi.com/journals/aess/2012/294121/abs/.

McDowell ML, Bruland GL, Deenik JL, Grunwald S, Knox NM (2012b) Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. Geoderma 189-190(0): 312-320.

Minasny B, McBratney AB (2008) Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemometr. Intell. Lab. Syst. 94(1): 72-79.

Minasny B, Tranter G, McBratney AB, Brough DM, Murphy BW (2009) Regional transferability of mid-infrared diffuse reflectance spectroscopic prediction for soil chemical properties. Geoderma 153(1-2): 155-162.

Natural Resources Conservation Service (NRCS), U.S. Department of Agriculture, USDA (2009). Soil survey geographic database (SSURGO). Available at: http://soils.usda.gov/survey/geography/ssurgo/.

Nocita M, Kooistra L, Bachmann M, Müller A, Powell M, Weel S (2011) Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa. Geoderma 167-168(0): 295-302.

Sarkhot DV, Grunwald S., Ge Y, Morgan CLM (2011) Comparison and detection of total and available soil carbon fractions using visible/near infrared diffuse reflectance spectroscopy. Geoderma 164(1-2): 22-32.

Savitzky A, Golay MJE (1964) Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36(8): 1627-1639.

Thissen U, Pepers M, Üstün B, Melssen WJ, Buydens LMC (2004) Comparing Support Vector Machines to Partial Least Square Regression for spectral regression applications. Chemometr. Intell. Lab. Syst. 73(2): 169-179.

United States Geological Survey, USGS (1999) National Elevation Dataset (NED). Available at: http://ned.usgs.gov/.

Vapnik V (2000) The nature of statistical learning theory. Second. Springer, New York, USA.

Vasques GM, Grunwald S, Harris WG (2010) Spectroscopic models of soil organic carbon in Florida, USA. J. Environ. Qual. 39(3): 923-934.

Vasques GM, Grunwald S, Myers DB (2012a) Influence of the spatial extent and resolution of input data on soil carbon models in Florida, USA. J. Geohys. Res.: Biogeosci. 117(G4): 1-12.

Vasques GM, Grunwald S, Myers DB (2012b) Associations between soil carbon and ecological landscape variables at escalating spatial scales in Florida, USA. Landscape Ecol. 27(3): 355-367.

Vasques GM, Grunwald S, Sickman JO (2008) Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma 146(1-2): 14-25.

Vasques GM, Grunwald S, Sickman JO (2009) Modeling of soil organic carbon fractions using visible-near-infrared spectroscopy. Soil Sci. Soc. Am. J. 73(1): 176-184.

Viscarra Rossel RA, Walvoort DJJ, McBratney AB, Janik LJ, Skjemstad JO (2006) Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131(1-2): 59-75.

Webster R, Oliver MA (2007) Geostatistics for environmental scientists. John Wiley & Sons, Chichester, England.

Williams P (1987) Variables affecting near-infrared reflectance spectroscopic analysis. p. 143–167. In Near-infrared technology in the agricultural and food industries. American Association of Cereal Chemists, St. Paul, Minnesota.

Wu J, Jones KB, Li H, Loucks OL (Eds) (2006) Scaling and uncertainty analysis in ecology: methods and applications. Springer, Dordrecht, The Netherlands.

Xiong X (2013) Geo-spatial Modeling of soil organic carbon and its uncertainty. Ph.D. dissertation, University of Florida, Gainesville, Florida, USA.

Yimer F, Ledin S. Abdelkadir A (2006) Soil organic carbon and total nitrogen stocks as affected by topographic aspect and vegetation in the Bale Mountains, Ethiopia. Geoderma 135: 335-344.

Chapter 22
Digital Soil Resource Inventories: Status and Prospects in 2015

David G. Rossiter

Abstract Eleven years ago, the author published a paper (Soil Use and Management 20(3): 296–301) titled "Digital soil resource inventories: status and prospects," which concluded that, at the time, the quantity and quality of digital soil survey information at global, national, regional, and local scales was increasing dramatically, however, with several problems such as (1) lack of metadata, (2) limited interpretations for professionals who are not soil specialists, (3) geodesic incompatibility with other digital data, (4) frequent reorganization of Web sites, and most seriously (5) much digital data were proprietary and only available for sale or under license. The current paper updates the situation to mid-2015, with an inventory of publically available soil geographic databases, their coverage, the type of information, and intended purposes. These are summarized in a portal maintained by the author (http://www.css.cornell.edu/faculty/dgr2/research/sgdb/sgdb.html). With regard to the deficiencies identified eleven years ago, metadata provision is much improved; more products come with interpretations; geodetic incompatibility has largely been overcome by metadata and conversion programs; Web sites still change frequently and are often confusing; and much data are still proprietary or not generally accessible. Over the next several years, several disruptive technologies are predicted to radically change how online soil survey information is collected, compiled, and disseminated. The question of open access to primary data is not resolved.

Keywords Soil geographic databases · Spatial data infrastructure

D.G. Rossiter (✉)
Section of Soil and Crop Sciences, Cornell University, Ithaca, NY, USA
e-mail: dgr2@cornell.edu

D.G. Rossiter
ISRIC-World Soil Information, Wageningen, The Netherlands

D.G. Rossiter
Chinese Academy of Sciences Soil Science Institute, Nanjing, China

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_22

22.1 Introduction

Eleven years ago, this author published a paper (Rossiter 2004) titled "Digital soil resource inventories: status and prospects," surveying the state of digitally available primary soil information (point observations, polygons, and grids) as well as secondary information, i.e., interpreted for end users. The present paper has the same objective. The intervening ten years have been a decade of dramatic progress in information technology, large disciplinary data initiatives such as OneGeology[1] and WorldClim,[2] and interdisciplinary spatial data infrastructures such as Infrastructure for Spatial Information in the European Community (INSPIRE).[3] This paper reviews to what extent the soil mapping and soil data provision community have participated in this progress. The view is from an interested user, searching the Internet for publically available primary soil survey information. If I have missed something, it is likely too difficult for others to find.

22.2 Forms of OnLine Soils Information

These can be categorized as (1) freely downloadable GIS-ready coverages, with adequate metadata to allow users to produce their own products such as customized maps, model inputs, and interpretations; (2) same but available only on off-line digital media, typically DVD-R; (3) commercially available (under license or for purchase) in both formats; (4) viewable and printable online but not available as a digital product; (5) non-georeferenced scanned maps as images, sometimes with their original accompanying documentation (e.g., soil survey reports), in both formats. Of these, the most useful is the first form. A variant of (1) is data provided dynamically as a Web Feature Service (WFS). A variant of (4) is data provided dynamically as a Web Map Service (WMS), where the GIS data remain with a map server but can be integrated into the user's GIS.

Another categorization is by the originating institution. Comprehensive general-purpose soil resource inventories (SRI), also called soil surveys, usually with interpretations, have traditionally been produced by national soil survey organizations. Other government institutions, for example, forestry or irrigation departments, have sometimes made special-purpose surveys. Development projects and consultants have made surveys of limited areas, often as interpreted rather than primary products, e.g., suitability for irrigation projects. These sources have been

[1] http://www.onegeology.org/.

[2] http://www.worldclim.org/.

[3] http://inspire.ec.europa.eu/.

combined into synoptic products by institutions with international mandate, notably the FAO,[4] the European Soil Bureau,[5] and ISRIC—World Soil Information.[6]

Yet another categorization is by type of information: (1) soil types in some classification system; (2) "static," or at least slowly changing, soil properties; (3) dynamic soil properties, notably soil moisture and temperature; and (4) interpretations directly usable by modelers and land managers.

22.3 Users of OnLine Primary Soils Information

Potential users include (1) soil mappers within the producing organization; (2) land use specialists within the producing organization, using the primary information to make interpretations; (3) government departments responsible for land use planning, public lands management, and taxation; (4) soil mappers in other organizations, using these maps as a basis for more detailed or generalized products; (5) land use specialists in other organizations, e.g., development consultants; (6) land managers and their consultants; (7) environmental modelers of, e.g., surface energy balance or watershed hydrology, (8) outdoor recreation enthusiasts such as hunters and hikers.

Some of these users prefer, or can only understand, interpreted information. Primary soil survey data are widely used in environmental modeling, e.g., pollution risk assessment (Sekhon et al. 2014), soil hydrology (Toth et al. 2012), gas flux (Yao et al. 2013), and watershed hydrology (Yu et al. 2014), just to mention a few recent examples. Modelers typically need primary, rather than interpreted, information, because they build their own interpretive models.

22.4 Status of Primary Soil Survey Information

Here, I review the current status of online static or slowly changing soil survey information over world, regional, national, and local extents.

22.4.1 Area Coverages (Polygons and Grids)

22.4.1.1 World

The most detailed compiled and edited product is the harmonized world soil database (HWSD)[7] (IIASA et al. 2012), supported by the FAO and compiled by

[4] http://www.fao.org/soils-portal/en/.

[5] http://eusoils.jrc.ec.europa.eu.

[6] http://www.isric.org/.

[7] http://www.iiasa.ac.at/Research/LUC/luc07/External-World-soil-database/HTML/index.html.

IIASA. This is a gridded product (21,600 × 43,200) with a consistent 30 arc-second (approximately 1 km^2 at the equator) resolution. Although 1 km^2 corresponds to the minimum legible delineation (MLD) of a 1:200k map, considering a 5 × 5 grid cell window as the MLD, the resulting map scale is 1:1M.

Data sources for the HWSD include SOTER, European Soil Database, Soil Map of China, the WISE profile database, and the digitized 1:5M scale FAO–UNESCO soil map of the world. This latter was produced in stages from 1971 to 1981 and thus is seriously outdated. The resulting raster database consists of 21,600 rows and 43,200 columns, which are linked to harmonized soil property data. The use of a standardized structure allows for the linkage of the attribute data with the raster map to display or query the composition in terms of soil units and the characterization of selected soil parameters (organic Carbon, pH, water storage capacity, soil depth, cation exchange capacity of the soil and the clay fraction, total exchangeable nutrients, lime and gypsum contents, sodium exchange percentage, salinity, textural class, and granulometry). Although the product is consistently formatted, there are extreme differences in the level of categorical and cartographic detail, depending on the source. Surprisingly, some well-studied areas (USA, Canada) are only represented by the 1:5M source and not by the much more detailed national soil survey databases. ISRIC is currently updating the HWSD with improved basis polygons, a single classification system (FAO Revised Legend 1988), estimates of uncertainty, and seven soil depth slices (layers) of representative synthetic profiles, following the SOTER specifications. Soil parameter estimates are recomputed for each component soil unit using an elaborate taxotransfer scheme that evolved from earlier work with FAO, IIASA, and ISRIC and contributions by ISRIC to HWSD via the SOTER program and WISE project. The above procedure considers 20 soil properties, five textural classes (SOTER criteria), seven depth layers up to 2 m depth, and broad climate as an important covariate in the taxotransfer scheme.

A global product produced by digital soil mapping methods is SoilGrids1 km from ISRIC—World Soil Information[8] (Hengl et al. 2014). This is a collection of consistent soil property and class maps of the world at 1-km resolution, produced using documented statistical models, from primary data (points and polygons) provided by soil survey organizations and environmental covariates which cover the whole world, including long-term NDVI time series and WorldClim layers. The soil polygon covariate is the HWSD, so that areas with poor HWSD resolution (e.g., USA) have much less spatial precision than those with the best HWSD resolution (e.g., China). Newer editions of SoilGrids may replace the HWSD with either an updated HWSD or may omit it altogether; although it covers the whole world, it is not a consistent coverage. The authors have chosen 3D regression with splines for continuous soil properties and multinomial logistic regression for soil classes. Both of these provide uncertainty: confidence limits from the kriging prediction variance for continuous properties and probability of membership for soil classes. An advantage of this

[8] http://www.isric.org/content/soilgrids.

product is that it is easily updatable: provide improved soil polygons, points, or environmental covariates and rerun the models. The data are available for download[9] and via an API[10] for incorporation into user-written applications. It can also be viewed via a SoilInfo tablet and smart phone application,[11] "providing free access to soil data anytime anywhere…for everyone." The mapping method can be used at finer resolutions (see AfSoilsGrid250m, below), depending only on the availability of covariates at these resolutions and sufficient calibration points.

The GlobalSoilMap.net consortium[12] (Arrouays et al. 2014) has since late 2007 been working toward a gridded soil map of the world at a nominal 100-m resolution. Specifications (Science Committee 2013) include a consistent geometry and tiling method, depth increments, properties, and uncertainty description. Each regional node is free to use any method to populate the grid according to the specifications. The first publically available product is from Australia (see below).

The Global Soil Partnership (GSP)[13] is a FAO-coordinated consortium "to improve governance of the limited soil resources of the planet… in accordance with the sovereign right of each State over its natural resources." One of its five "Pillars of action" is the fourth: to "enhance the quantity and quality of soil data and information: data collection (generation), analysis, validation, reporting, monitoring and integration with other disciplines." As part of this, Omuto et al. (2012) produced a report on the status of global and regional soil information, and a working group wrote an action plan (late 2014), which has been transformed (mid-2015) to an implementation plan, and it is hoped (subject to financing) to a global soil information system.

22.4.1.2 Regional

A product with a long history is SOTER,[14] a collaborative activity of ISRIC, FAO, and UNEP, endorsed by the International Union of Soil Sciences (Oldeman and van Engelen 1993) and used for a wide variety of regional assessments (e.g., Batjes et al. 2007). This is a well-structured soil geographic database: polygons at scales 1:5M to 1:250k, depending on quality of source data with a linked relational database. This is hierarchical from terrain, through terrain component, to soil components, to profiles, and to representative horizons. Each product is internally harmonized across country boundaries, using a consistent mapping concept based on terrain units, and a consistent soil classification. Regions available are Central and Eastern Europe, southern Africa, central Africa, Latin America, and the Caribbean. The concept of soil units within terrain units is not always in accordance

[9] http://soilgrids.org/.
[10] http://rest.soilgrids.org/.
[11] http://soilinfo.isric.org/.
[12] http://globalsoilmap.net.
[13] http://www.fao.org/globalsoilpartnership/en/.
[14] http://www.isric.org/projects/soil-and-terrain-database-soter-programme.

with the soilscape (e.g., in volcanic areas), leading to some difficulties in delimiting and characterizing units. Soil components are not necessarily mappable at the target scale, in which case their proportions are estimated.

Dewitte et al. (2013) report on the Soil Atlas of Africa,[15] which was produced as an update to the HWSD. Nominal scale is 1:3M, corresponding to a minimum legible area (MLA) of 225 km^2. It is available as PDF, as e-book, and for download as GIS coverage[16] on request.

ISRIC has used a similar methodology to the global SoilGrids1 km to produce AfSoilsGrid250 m,[17] a finer-resolution product for the non-desert areas of Africa.

Europe is served by the European Soil Bureau (ESB), which has set up a European Soil Data Centre (ESDAC) to fulfill its responsibility for responding to the European Commission for policy support (Panagos et al. 2012). ESDAC includes the European Soil Portal[18] with access to the European Soil Database (ESDB), a 1:1M harmonized coverage. Single-property 1 × 1 km and 10 × 10 km grids have been extracted from this. Several "soil threats" gridded maps are available, including heavy metals in topsoils, soil salinization, susceptibility to compaction, organic C, and erosion estimates. The ESB operates under a complicated legal framework (EU-wide and national) and strives to make the primary data as open as legally possible; for restricted products at least, the metadata is supplied, so that a potential user can judge the fitness for use.

22.4.1.3 National

A few soil survey organizations provide free download of their polygons (map units) with associated attribute tables, e.g., Canada (CanSIS[19]), the USA, Australia (ASRIS[20]), and New Zealand (S-Map[21]). Some provide gridded data, e.g., Australia. Point observations (profiles) are only available for the USA; this very detailed database (with the extensive laboratory tests required by USDA Soil Taxonomy) also includes some non-USA observations.

SOTER is available at scales of 1:1 M (Argentina, Burundi, Cuba, Kenya, RSA, Rwanda, Senegal and the Gambia, Tunisia) and 1:2M (DRC).

The USA has two almost complete polygon coverages: SSURGO 2.2 (semi-detailed, source scale 1:12,000–1:25,000) and STATSGO2 (1:250,000, generalized from SSSURGO). These are provided to the public by the Web Soil

[15] http://eusoils.jrc.ec.europa.eu/library/Maps/Africa_Atlas/.

[16] http://eusoils.jrc.ec.europa.eu/library/maps/africa_atlas/data.html.

[17] http://www.isric.org/data/AfSoilGrids250m.

[18] http://eusoils.jrc.ec.europa.eu/.

[19] http://sis.agr.gc.ca/cansis/.

[20] http://www.asris.csiro.au/.

[21] http://smap.landcareresearch.co.nz/home.

Survey interface,[22] which allows the user to specify an area of interest. Two other interfaces to the same data source are provided by the California Soil Resource Lab[23]: SoilWeb, which uses Google maps, and SoilWebEarth, which uses Google Earth to allow a 3D view of the soilscape. The USA has gridded the SSURGO product at 30-m resolution (gSSURGO) and is experimenting with a disaggregation (dSSURGO) to this resolution using environmental covariates as a training set (Chaney et al. 2015); however, this last-named is not yet publically available.

Many European countries have digital databases of polygons and/or points, but these are not immediately available online. Some have provided data viewers or static maps online, for example, the Dutch.[24] Depending on national data policies, they may be provided by commercial contract, use agreement, cooperative project, or publically available. Some products are generalizations of more detailed products that are kept for internal use. For example, the Base de Données Géographique des Sols de France,[25] available as ArcInfo coverages on CD-ROM for the cost of reproduction and postage, is a 1:1M generalization of several detailed products (Connaissance Pédologique de la France, Secteurs de Référence) at 1:50,000 or 1:100,000. These are only available under agreement to regional partners and cooperation projects.

Australia has produced the first national map to GlobalSoilMap specifications: the Soil and Landscape Grid of Australia.[26] This is managed as part of ASRIS. In addition to the soil properties, it also provides many landscape attributes, e.g., the Prescott Index measure of water balance and solar radiation.

22.4.1.4 Local

The national products listed in the previous section can be queried for any locality. There are a few purely local digital products. For example, SOTER is available at scale 1:250k for the upper Tana River basin, Kenya.

22.4.1.5 Standards

Each product has its own standards, which are in general well documented. The three international standards are for the HWSD, GlobalSoilMap.net, and SOTER (Pourabdollah et al. 2012).

[22] http://websoilsurvey.nrcs.usda.gov/.

[23] http://casoilresource.lawr.ucdavis.edu/soilweb-apps/.

[24] http://maps.bodemdata.nl.

[25] http://gissol.fr/programme/bdgsf/bdgsf.php.

[26] http://www.clw.csiro.au/aclep/soilandlandscapegrid/.

22.4.2 Points

Georeferenced point observations (typically of soil profiles), generally with accompanying laboratory data, are especially valuable as uninterpreted primary information on soils at known locations. The largest freely available sets are as follows:

1. US National Soil Survey Center Soil Characterization Data[27]: (1) analytical data for more than 20,000 USA and 1100 other pedons and (2) standard morphological pedon descriptions for about 15,000 of these.
2. Soil Profile Databases for Europe[28] (SPADE); actual or inferred profiles for each soil typological unit in the 1:1M SGDBE.
3. Land Use/Cover Area frame Statistical Survey (LUCAS)[29]; selected properties of approximately 20,000 topsoil samples from 25 European countries (coarse fragments, particle size distribution, pH, organic carbon, carbonates, P, total N, extractable K, CEC, multispectral properties).
4. WISE[30]: The result of various ISRIC projects, this contains about 11,000 non-harmonized profiles with attributes, of which 1125 have been harmonized.
5. Africa Soil Profiles database[31] from the Africa Soil Information Service (AfSIS) and ISRIC; about 15,000 profiles.
6. World Soil Profiles[32] from ISRIC, about 32,000, allows users to submit their own profiles and create their own data entry templates.

22.4.3 Scans

Maps from the ISRIC collection have been scanned as non-georeferenced images by the European Soil Bureau and published online and as DVD as the European Digital Archive on Soil Maps of the World (EuDASM)[33] (Panagos et al. 2011). The accompanying reports have not been included; many of these can be found in the Wageningen University library via a search interface[34] and are currently being scanned by ISRIC. A similar project but with focus on British soil survey activities (in the UK and former colonies) is the World Soil Survey Archive and Catalogue (WOSSAC) hosted by Cranfield University (England).[35]

---

[27] http://ncsslabdatamart.sc.egov.usda.gov/.

[28] http://eusoils.jrc.ec.europa.eu/projects/spade/.

[29] http://eusoils.jrc.ec.europa.eu/projects/Lucas/.

[30] http://www.isric.org/data/isric-wise-international-soil-profile-dataset.

[31] http://www.isric.org/content/africa-soil-profiles-database.

[32] http://worldsoilprofiles.org.

[33] http://eusoils.jrc.ec.europa.eu/esdb_archive/EuDASM/EUDASM.htm.

[34] http://www.isric.org/content/search-library-and-map-collection.

[35] https://www.wossac.com/.

22.5 Status of Dynamic Soil Information

The International Soil Moisture Network[36] collects volunteered datasets; one of the largest is from the former Soviet Union, digitized by the Global Soil Moisture Databank of the Rutgers University. The Soil Climate Analysis Network (SCAN) from the (USA) National Water and Climate Center[37] does give downloadable time series of soil moisture, temperature, and snowpack for scattered stations across the USA and some for Puerto Rico and the US Virgin Islands. Texas A&M University has produced a harmonized and quality-controlled historical soil moisture database for the USA and some Canadian provinces.[38]

22.6 Progress Over the Past Decade

Comparing the current situation to that in 2004:

1. There is much more digital geoinformation and more is publically available. This is despite the low level of new soil survey activity.
2. Online access and user interfaces to find and obtain geoinformation are much improved.
3. Metadata provision is much improved. Although most of the above-listed databases do not use formal metadata standards, almost all have sufficient information for proper use. Some products include uncertainty estimates in metadata and some as interpolated layers.
4. Interpretations for professionals who are not soil specialists are excellent in some databases, notably the USA and Australia. Some organizations, such as the European Soil Bureau, use their databases to make separate interpretive products that are directly useable.
5. Geodesic incompatibility with other digital data are far less of a problem. Many data providers have standardized on WGS84 geographic coordinates. The Europe-wide databases have standardized on ERTS89 coordinate system. But with easy conversions provided by the GDAL program[39] bundled with R[40] and most GIS, and the collection of coordinate reference systems in the EPSG database,[41] combined with much better metadata, this problem becomes minor.
6. Frequent reorganization of Web sites is still a big problem. The new organization may be better but data that could be found previously is now relocated.

[36] http://ismn.geo.tuwien.ac.at/.
[37] http://www.wcc.nrcs.usda.gov/scan/.
[38] http://soilmoisture.tamu.edu/.
[39] http://www.gdal.org/.
[40] http://www.r-project.org/.
[41] http://www.epsg-registry.org/.

A related problem is the increasing number of datasets per site (in itself a good thing); this often makes finding a particular dataset more difficult.

7. Much digital data are still proprietary and only available for sale or under license. Some is considered public but not made available to the general public in digital form, only as a view in a Web mapping application or as a Web Mapping Service (WMS) layer for use in GIS.
8. There is a new generation of "geoportals" which provide an entry point to find digital geodata from multiple themes, including soils, for example, INSPIRE from the European Commission.[42] These give exposure to soil data to users who might not find them otherwise, and to users who are looking for different coverages of the same area for integrated modeling.

22.7 Prospects

What will the next ten years bring us? Clear trends in the GIS world are as follows: (1) massive increase in data storage and processing power, allowing models to be run on large grids with many layers; (2) many new high-resolution sensors from satellite systems, providing almost unmanageable data streams, many of these useful as soil mapping covariates; (3) new ultra-resolution airborne and field sensors, including low-cost drone-borne LIDAR and spectrometers; (4) large, cheap networks of point sensors with automatic recording, e.g., soil moisture; and (5) increasingly powerful Web services over a faster Internet backbone, reaching most clients via very high-capacity links. Sensors will increase not only in number and coverage, but also in temporal resolution. The data volume will be too large for manual processing; this will require automated methods of quality control and summary, as is being developed for streaming sensors in environmental monitoring networks (Campbell et al. 2013).

In the digital soil mapping world, the massive increase in spatiotemporal covariates will require new models. The temporal aspect is particularly interesting: There is no reason for a soil map to be static. For example, why do we use generalized soil moisture and temperature regimes, when we are able to give a detailed model of the soil moisture and temperature status, over depths as well as across the landscape, at temporal resolutions matching the sensors?

In the soil survey world, increasing emphasis will be on soil functions rather than properties. For example, (1) soil health and resilience, including soil biodiversity; (2) soil-related human health risks and benefits; (3) soil functioning within the hydrosphere and at the earth–atmosphere interface. These may require new concepts and models, but surely will require spatially detailed properties that drive such models. Some may be directly mapable.

[42] http://inspire-geoportal.ec.europa.eu/.

There will be a strong push for harmonized global coverages, especially useful for global modeling, at increasingly finer spatial resolutions. Examples are GlobalSoilMap, Pillar 4 of the Global Soil Partnership,[43] and the Harmonized world soil database. It is unclear that the current observation density can support reliable products at the desired resolutions.

It is unclear how the near future will develop in terms of data access and data sharing. Many countries still have restrictive laws and do not recognize that primary data can have a large multiplier effect on the economy and general welfare of the citizenry, through unanticipated uses. Some institutions struggling with funding still see primary data as a revenue source, rather than as an advertisement for their specialist knowledge in aiding interpretations and modeling.

References

Arrouays D, Grundy MG, Hartemink AE, et al (2014) GlobalSoilMap. Advances in Agronomy. Elsevier, pp 93–134

Batjes NH, Al-Adamat R, Bhattacharyya T, et al (2007) Preparation of consistent soil data sets for modelling purposes: Secondary SOTER data for four case study areas. Agriculture, Ecosystems & Environment 122:26–34.

Campbell JL, Rustad LE, Porter JH, et al (2013) Quantity is nothing without quality: automated QA/QC for streaming environmental sensor data. BioScience 63:574–585. doi: 10.1525/bio.2013.63.7.10

Chaney, NW, Hempel, JW, Odgers, N, McBratney, AB, & Wood, EF (2015). dSSURGO: Development and validation of a 30 meter digital soil class product over the 8-million square kilometer contiguous United States. Geophysical Research Abstracts 17:EGU2015–11042.

Dewitte O, Jones A, Spaargaren O, et al (2013) Harmonisation of the soil map of Africa at the continental scale. Geoderma 211-212:138–153. doi: 10.1016/j.geoderma.2013.07.007

Hengl T, de Jesus JM, MacMillan RA, et al (2014) SoilGrids1 km — Global Soil Information Based on Automated Mapping. PLoS ONE 9:e105992. doi: 10.1371/journal.pone.0105992

IIASA; FAO; ISRIC; ISS-CAS; JRC (2012) Harmonized World Soil Database (version 1.2). FAO and IIASA, Rome, Italy and Laxenburg, Austria

Oldeman LR, van Engelen VWP (1993) A world soils and terrain digital database (SOTER) — An improved assessment of land resources. Geoderma 60:309–325. doi: 10.1016/0016-7061(93)90033-H

Omuto C, Nachtergaele F, Vargas Rojas, Ronald (2012) State of the art report on global and regional soil information: where are we? Where to go? ix, 69. FAO, Rome.

Panagos P, Jones A, Bosco C, Kumar PSS (2011) European digital archive on soil maps (EuDASM): preserving important soil data for public free access. International Journal of Digital Earth 4:434–443. doi: 10.1080/17538947.2011.596580

Panagos P, Van Liedekerke M, Jones A, Montanarella L (2012) European Soil Data Centre: Response to European policy support and public data requirements. Land Use Policy 29:329–338. doi: 10.1016/j.landusepol.2011.07.003

Pourabdollah A, Leibovici DG, Simms DM, et al (2012) Towards a standard for soil and terrain data exchange: SoTerML. Computers and Geosciences 45:270–283. doi: 10.1016/j.cageo.2011.11.026

[43] http://www.fao.org/globalsoilpartnership/the-5-pillars-of-action/4-information-and-data/en/.

Rossiter DG (2004) Digital soil resource inventories: status and prospects. Soil Use and Management 20:296–301.

Science Committee (2013) Specifications: Tiered GlobalSoilMap.net products; Release2.3. GlobalSoilMap.net

Sekhon BS, Bhumbla DK, Sencindiver J, McDonald LM (2014) Using soil survey data for series-level environmental phosphorus risk assessment. Environmental Earth Sciences 72:2345–2356. doi: 10.1007/s12665-014-3144-6

Toth B, Mako A, Guadagnini A, Toth G (2012) Water retention of salt-affected soils: Quantitative estimation using soil survey information. Arid Land Research and Management 26:103–121. doi: 10.1080/15324982.2012.657025

Yao H, Campbell CD, Chapman SJ, et al (2013) Multi-factorial drivers of ammonia oxidizer communities: evidence from a national soil survey. Environmental Microbiology 15:2545–2556. doi: 10.1111/1462-2920.12141

Yu X, Duffy C, Baldwin DC, Lin H (2014) The role of macropores and multi-resolution soil survey datasets for distributed surface-subsurface flow modeling. Journal of Hydrology 516:97–106. doi: 10.1016/j.jhydrol.2014.02.055

Chapter 23
Evaluating the Relative Importance of Legacy Soil Sampling and Spatial Models in Digital Soil Mapping Performances: A Case Study in Languedoc-Roussillon (Southern France)

Philippe Lagacherie and Kévin Vaysse

Abstract A growing set of real-life applications of digital soil mapping (DSM) is now available across the planet. These DSM applications need to be thoroughly analyzed for identifying the corrective actions that will provide the best increase in performances. In Languedoc-Roussillon, the analysis of performances of three DSM models applied for mapping 29 soil properties showed that DSM performances were mainly driven by the ability of the spatial sampling to capture the variability of soil properties, itself driven by the sampling density and the intrinsic scale of the soil property variations. In this region, increasing the sampling density of soil measurements appeared therefore as the priority instead of looking for a more efficient DSM model. We recommend the extension of our approach for analyzing further DSM results.

Keywords Soil map · Random forest · Kriging · Uncertainty · Variogram · Spatial structure

23.1 Introduction

In recent years, there have been a lot of real-life experiments of digital soil mapping (DSM) that have covered a great diversity of pedological contexts across the planet. These experiments have revealed a great variability of DSM performances, with,

P. Lagacherie (✉) · K. Vaysse
INRA, LISAH, Montpellier, France
e-mail: philippe.lagacherie@supagro.inra.fr

K. Vaysse
SIG-LR, Montpellier, France

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_23

however, a majority of weak soil predictions. A meta-analysis over this set of experiments should be highly valuable in view of identifying the corrective actions that could best improve the current DSM performances.

However, it can be difficult to identify which factor is the most limiting with regard to DSM performances in a given pedological context, especially among limitations caused by inadequate spatial sampling with regard to the local patterns and those caused by the use of irrelevant spatial models for representing these variations. This issue was examined in Languedoc-Roussillon (southern France) where it can be observed a great diversity of soil property patterns, which mimics a large range of pedological contexts for DSM applications.

23.2 Materials and Methods

A DSM experiment was conducted over the Languedoc-Roussillon region (27,236 km^2). It used as input soil data a small-scale (1:250,000) soil map and a set of 2024 measured legacy soil profiles (one profile per 13.5 km^2). Spatial sets of soil covariates covering the entire region and available over the whole French territory were built for accounting of variations in relief (SRTM digital elevation model and derivatives), climate (WorldClim database layers), lithology (1:50,000 scale geological map of France), and land use (derived from Landsat 7). Twenty-nine soil properties selected among those specified in the GlobalSoilMap project (Arrouays et al. 2014) were considered for predictions [soil depth and clay, silt, sand, organic carbon, pH, CEC, coarse fragment at (0–5 cm), (5–15 cm), (15–30 cm), and (30–60 cm)]. They exhibited a large range of spatial structures characterized by the proportion of spatially structured variance that can be computed from their experimental variogram by the complement to 1 of the nugget-to-sill ratio.

Three DSM models were tested: area-weighted means from the small-scale soil map, random forest calibrated from the measured soil profiles, and random forest plus kriged residuals. Independent validations of DSM were performed over 105 sites of the French soil-monitoring network located at the nodes of a 16 km × 16 km grid intersecting the study area. More details on DSM models are provided in Vaysse and Lagacherie 2015.

23.3 Results

The validations showed great variations of performances (from R^2 = 0.00 to R^2 = 0.79). Variations of performances between soil properties were found greater than between DSM models. Concerning the latter, random forest and random forest with kriged residuals were equally efficient and outperformed area-weighted means. Concerning the former, the performances of soil properties predictions were clearly correlated with the proportion of spatially structured

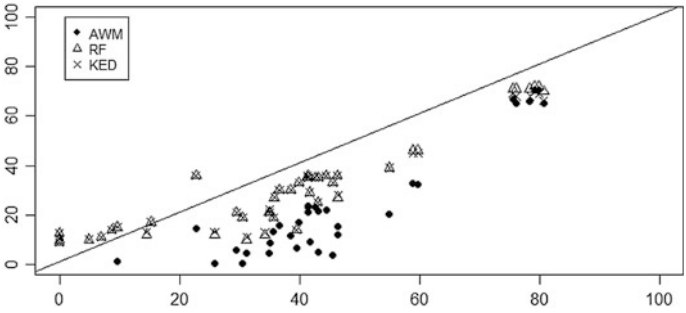

Fig. 23.1 Relation between the proportions of spatially structured variances of soil properties (1—nugget/sill, *x*-axis) and the proportions of variance captured by digital soil mapping (R^2_v, y-axis). Each dot is a soil property predicted by the one of the three following models: area-weighted mean (AWM), random forest (RF), and kriging with external drift (KED) modeled by RF

variance derived from the experimental variograms obtained from the set of legacy-measured profiles (Fig. 23.1). For soil properties (e.g., coarse fragment contents) that exhibited the smallest proportion of spatially structured variance (i.e., with large proportions of short scale soil variations), all DSM models performed poorly. The converse was observed for pH that exhibited the greatest proportion of spatially structured variance. In intermediate situations (e.g., organic carbon), performances were also intermediate, with increasing contrasts between the DSM models.

23.4 Conclusions

These results suggest that, in situations of sparse soil sampling like in Languedoc-Roussillon, DSM performances were mainly driven by the ability of the spatial sampling to capture the variability of soil properties, itself driven by the sampling density and the intrinsic scale of the soil property variations. The choice of the spatial model for mapping the soil property seemed to less impact these DSM performances. Increasing the sampling density of soil measurements appeared therefore as the priority instead of looking for a more efficient DSM model.

From this experiment, we recommend (i) to use the proportion of spatially structured variance as an indicator that can predict a potential level of DSM performances according to the available (legacy) spatial sampling) and (ii) to complement the classical evaluations of DSM model performances by a new ratio (R^2_{spat}) that measure the proportion of spatially structured variance explained by the tested model.

Acknowledgements This research was granted by the French National Institute of Agronomical Research (INRA) and the French Research and Technology Agency (ANRT). The authors are also indebted to BRGM (French Geological Survey), Jean-François Desprats for providing geological

maps at the 1:50,000 scale and the French Scientific Group of Interest on soils, "GIS Sol," and the US INFOSOL (INRA Orléans) for providing soil measurement data from a RMQS survey.

References

Arrouays, D., Grundy, M. G., Hartemink, A. E., Jonathan, W., Heuvelink, G. B. M., Hong, S. Y., … James, A. (2014). GlobalSoilMap : Toward a Fine-Resolution Global Grid of Soil Properties. *Advances in Agronomy*, *125*, 93–134.

Vaysse, K., & Lagacherie, P. (2015). Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). *Geoderma Regional*, *4*, 20–30.

Chapter 24
Improved Soil Mapping in British Columbia, Canada, with Legacy Soil Data and Random Forest

C. Bulmer, M.G. Schmidt, B. Heung, C. Scarpone, J. Zhang, D. Filatow, M. Finvers, S. Berch and S. Smith

Abstract The need for improved soil inventory information in the province of British Columbia (BC), Canada, was addressed using a random forest (RF) classifier that was informed using legacy soil data. RF models were prepared for 110 ecodistrict subdivisions of BC, and predictions were subsequently assembled into a final soil parent material map mosaic covering the entire province. The ecodistricts are part of a framework for ecosystem classification in BC and in Canada, and delineate areas with relatively homogeneous biophysical and climatic conditions. Training areas for predicting soil parent materials were identified using single-component polygons from legacy terrain, soil, and ecosystem maps. For parent material mapping, we intersected training points amalgamated from all legacy surveys with a suite of 18 topographic covariates derived from a 100-m digital elevation model (DEM). For each ecodistrict, two versions of the resulting training dataset were submitted to the RF classifier. A 'balanced' dataset contained equal numbers of training data points for all parent material classes representing all legacy data derived from single-component polygons. A 'constrained' dataset was also derived where conditions were imposed on selected topographic attributes of the training points to reflect known geomorphic processes and to ensure consistent

C. Bulmer (✉)
BC Ministry of Forests Lands and Natural Resource Operations, 3401 Reservoir Road, Vernon, BC, CanadaV1B2C7
e-mail: chuck.bulmer@gov.bc.ca

M.G. Schmidt · B. Heung · C. Scarpone · J. Zhang
Department of Geography, Simon Fraser University, Burnaby, BC, Canada

M. Finvers · S. Berch
BC Ministry of Environment, Victoria, BC, Canada

S. Smith
Agriculture and Agri-Food Canada, Summerland, BC, Canada

D. Filatow
BC Ministry of Environment, Kelowna, BC, Canada

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_24

mapping criteria were applied across multiple legacy soil survey projects. RF predictions of soil parent material resulted in 100-m gridded class maps for BC that incorporate expert knowledge extracted from legacy soil inventories.

Keywords Random forest · Soil parent materials · Soil development · Legacy soil data

List of Abbreviations

BC	British Columbia
BEC	biogeoclimatic ecosystem classification
DEM	digital elevation model
MDA	mean decrease accuracy
MDG	mean decrease in gini
OOB	out-of-bag error
RF	random forest

24.1 Introduction

There is a need for improved soil survey information in the Canadian province of British Columbia (BC). Existing soil databases were derived from soil surveys that were carried out at various levels of detail over a period of more than 75 years, but cover less than 50 % of the 945,000 km^2 area within the provincial boundary. Terrain inventories are also available to provide information on soil parent materials, but large gaps remain in both spatial coverage and detail at a time when the demands for information on BC's natural resources are increasing.

Predictive mapping techniques have shown great potential for extending legacy soil information to new areas (Bui and Morgan 2003) and for increasing the spatial resolution of existing inventories by generating gridded attribute and class maps and databases (e.g., Sarmento et al. 2012). In BC, digital mapping techniques have been used successfully to predict forest ecosystems in the Cariboo region (MacMillan et al. 2007), parent materials in the lower Fraser Valley (Heung et al. 2014) and soil classes in the Okanagan Valley (Smith et al. 2012). Among the many choices available for modeling soil landscape relationships and predicting soil properties, random forest (RF) classification has proven to be a particularly useful approach (Heung et al. 2014; Stum et al. 2010; Häring et al. 2012; Subberayalu and Slater 2013). Random forest implements an ensemble of decision trees, with randomized bootstrap sampling and selection of potential predictors for node splitting. The aggregation of the resulting classifications has successfully predicted outcomes in a wide variety of applications.

The objective of this study was to develop 100-m gridded soil parent material and soil development class maps at the order level for the entire province of BC by capturing knowledge contained in legacy terrain (surficial geology) and soil

polygon maps. This marks the first-time predictive mapping has been used in BC at this scale and extent. A global gridded soil class and attribute map product (Hengl et al. 2014) covers the province but has not yet been regionally validated.

The purpose of this paper is to (1) describe how RF predictions were used to develop new soil parent material maps for BC, (2) describe the covariates that were used for the maps and how a collection of individual predictions for landscape subdivisions was mosaicked into a final map product for parent material, and (3) discuss some of the challenges of using legacy maps as knowledge sources for predictive mapping.

24.2 Materials and Methods

24.2.1 Physiography, Soils, and Ecology of British Columbia

British Columbia occupies the western coast of North America between the 49th and 60th north parallels, a distance of approximately 1200 km, and extends inland across the Western Cordillera for more than 750 km. BC's landscape consists of an arrangement of mountain systems and plateaus that have resulted from tectonic forces operating at the leading edge of the North American continental plate over very long periods of geologic time (Church and Ryder 2010). In the past 2 million years, glaciation and contemporary geomorphic processes have created predictable associations of local landform and surficial material that have been superimposed on the original character of the larger mountain systems, valleys, and plateaus.

Soil development in BC reflects the operation of pedogenic processes mediated by climate, topography, and vegetation at localized scale. During the ca. 10,000 years BP since glacier ice receded from most of BC, soil processes have transformed the upper portion of the (surficial) parent material to a depth of approximately 1 m. BC's rugged topography is an important factor driving the tremendous diversity of soil parent materials, climate, and vegetation, and the soil properties observed today reflect that diversity (Valentine et al. 1978).

Canada's National Ecological Framework (Ecological Stratification Working Group 1996) provides an approach for subdividing landscapes into nested ecozones, ecoregions, ecodistricts, and soil landscapes. Because the ecodistricts of the nested ecological framework represent contiguous areas with similar physiography and ecology, their use as rule sheds defining the limits for applying predictions from a single RF model was considered appropriate for this study. This nested ecological framework is complemented by BC's biogeoclimatic ecosystem classification (BEC) system (Meidinger and Pojar 1991) where climate defines broad ecological zones expressed through changes in plant species composition and physiognomy, and local variation is described in relation to changes in vegetation, soil, and topography. Therefore, the mapping units identified within BEC system provides a useful predictor of soil development.

24.2.2 Legacy Soil Surveys and Terrain Mapping

Although some of the first soil surveys in BC were carried out on forested lands (Anderson and Smith 2011), a major focus for many subsequent soil surveys was to identify and characterize new areas for settlement and provide improved agricultural opportunities (e.g., Kelley and Spilsbury 1939; Sprout and Kelly 1964; Luttmerding 1980). As a result, detailed and comprehensive soil information is currently available for most of the major agricultural valleys and lowlands in BC, while a patchwork of soil and terrain coverage exists outside those areas (Fig. 24.1). In addition, the entire province is covered by the generalized Soil Landscapes of Canada mapping at 1:1,000,000 scale (Schut et al. 2011). Government led soil survey activity in BC slowed dramatically after 1990, but the legacy soil data and terrain mapping carried out according to the methods outlined in Province of BC (2010) embodies a wealth of expert knowledge of landscape patterns of soil parent material and development that can inform predictive maps.

24.2.3 Digital Data

Geographic information for training data and all model inputs described in this paper were derived from public sources. Terrain and soil databases were obtained from Data BC www.env.gov.bc.ca/tei/access_tei.html in shape file format. Terrain

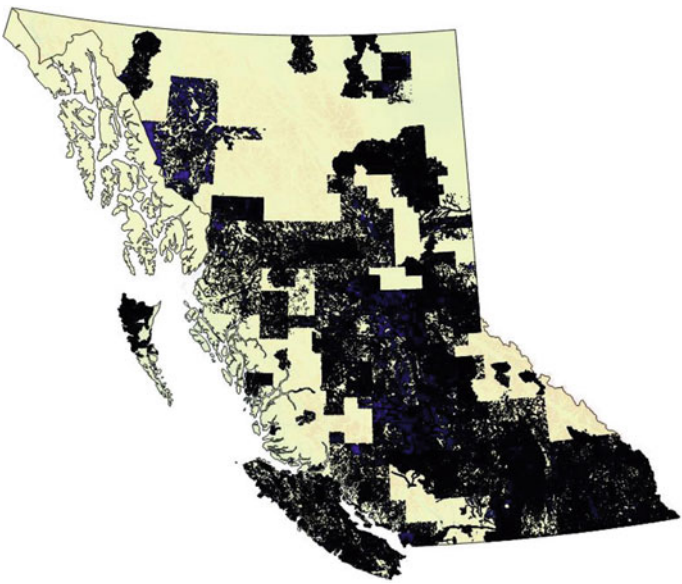

Fig. 24.1 Availability of detailed training data for soil parent materials in British Columbia

and soil polygons can have from one to three components. Polygons were selected where a single component occupied more than 90 % of the polygon area, or where two or more components with the same soil/terrain attributes comprised more than 90 % of the area. A 100-m-resolution digital elevation model (DEM) was obtained from http://www.hectaresbc.org and preprocessed using the procedure described by Heung et al. (2014). A 1-km DEM was also prepared by resampling the preprocessed 100-m DEM. Two datasets were used for validating the results. The BC Soil Information System (Sondheim and Suttie 1983) is an inventory of field measurements incorporating soil site characteristics and was commissioned to track soil conditions in BC. The BEC point dataset contains a collection of individual point samples of ecosystem conditions obtained by BC's BEC inventory program.

Soil Parent Materials

A total of 35 terrain derivatives were prepared from the 100- and 1-km DEMs to describe 13 different landform classes. Initial modeling results guided the selection of 18 of the terrain derivatives for use in production of the final predicted map (Table 24.1). Selection was based on the average (combined) mean decrease in accuracy (MDA) and mean decrease in Gini (MDG). Reducing the number of

Table 24.1 DEM derivatives used for RF prediction of soil parent material

Terrain attribute	Min	Max	Mean	Std. Dev.	Description
C_N_B_L	0	2388	939	434	Channel network base level (masl)
ELEV	0	3628	1118	529	Elevation (masl)
HD_2_CH	0	53480	1976	2441	Horizontal distance to channel (m)
HTNRM_K	0	0.99	0.44	0.29	Normalized height 1-km grid
MB_IND	−0.5	1.09	0.004	0.10	Mass balance index (index)
MDSLP_K	0	0.88	0.26	0.15	Midslope position 1-km grid (masl)
MRRTFKM	1.6	8.6	3.0	2.9	Multiresolution index of ridge top flatness 1 km
MRVBFKM	0	8.9	3.2	3.16	Multiresolution index of valley bottom flatness 1 km
MRVBFHA	0	8.26	0.94	1.68	Multiresolution index of valley bottom flatness
OPENNEG	0.92	1.63	1.45	0.11	Topographic openness—negative (index)
OPENPOS	0.98	1.65	1.44	0.10	Topographic openness—positive (index)
RHSP_KM	0.02	0.98	0.51	0.27	Relative hydrologic slope position 1-km grid
SL_HT_K	0	1229	241	127	Slope height 1-km grid (m)
SLOPEHT	0	1956	127	154	Slope height (m)
SLOPEUS	0	1.23	0.25	0.22	Slope from unsmoothed DEM (m/m)
V_D_C_N	0	3469	178	233	Vertical distance above channel network (m)
VALLY_D	0	1998	171	186	Valley depth (m)
VY_DP_K	0	1580	259	170	Valley depth from 1-km DEM (m)

predictors resulted in a small (<1 %) reduction in out-of-bag (OOB) error estimates compared to models with 35 predictors. Input predictors for parent material were restricted to topographic information only because parent material is thought to be highly correlated with landform due to BC's recent glacial past.

24.2.4 Modeling Approach

Our modeling approach for this study was described in Heung et al. (2014). Training datasets were prepared by random selection of points within the single-component polygons from legacy soil and terrain surveys and intersecting them with a range of topographic climatic and other attributes. Individual RF models and predictions were prepared for 100 ecodistrict subdivisions of BC where training data were available. For ten remaining ecodistricts with no training data, a model from an adjacent ecodistrict was used in the prediction. To incorporate the influences of training data from neighboring ecodistricts, a 5-km buffer was applied to each ecodistrict and the training points from neighboring ecodistricts that fell within the buffered areas were included for the prediction of the centroid ecodistrict.

Two versions of the parent material training dataset were input to the classifier. The 'balanced' dataset contained equal numbers of training data points for all parent material classes representing all legacy data derived from single-component polygons. The 'constrained' dataset also contained equal numbers of points per class, but with a restricted range of values allowed for selected topographic attributes to reflect known geomorphic processes associated with specific materials and to ensure consistent mapping criteria were applied across multiple legacy survey projects (Table 24.2).

Table 24.2 Constraints applied to topography derivatives for the constrained RF model of parent material

Parent material	Required condition	Rationale
Colluvium	SLOPEUS > 0.25	Movement of material on steep slopes is primarily driven by gravity
Fluvial	SLOPEUS < 0.10	Fluvial materials are deposited in flat areas
Fluvial	MRVBFHA > 0.5	Fluvial materials are deposited in flat low-lying areas
Fluvial	V_D_C_N < 0.0005*VY_DP_K	Fluvial materials are associated with the modern drainage network
Glaciofluvial	V_D_C_N > 0.0005*VY_DP_K	Glaciofluvial materials occur higher in the landscape than fluvial
Till	SLOPEUS < 0.35	Till is found on gentle and moderate slopes

24.2.5 Final Map Preparation

The individual maps for soil parent material were assembled into a final mosaic after clipping each one to the original ecodistrict boundary. The resulting map can be thought as an assemblage of buffered patches, allowing for new information or improvements to be incorporated into a single patch (rule shed), without compromising areas of the map outside of the rule shed boundary. Also, predictions for further subdivision within ecodistricts could be accomplished by considering the more detailed soil landscape level as rule sheds, allowing for improved modeling at finer resolution. There are more than 2500 soil landscape polygons mapped in BC (Schut et al. 2011).

24.3 Results and Discussion

Constrained models had lower OOB error rates compared to balanced models. Variability in OOB error rates between ecodistricts (Fig. 24.2) illustrates how consistently the individual models within the ecodistrict rule shed relate the topographic derivatives to the parent material class and could reflect (1) differences in availability, quality, and/or distribution of the training data; and/or (2) differences in the underlying relationships between the available topographic attributes and parent material distribution due to unique physiographic conditions within ecodistricts.

Overall, the most important variable for improving node purity was elevation (Fig. 24.3), highlighting the tremendous topographic diversity in this part of western North America. The derived indices for multiresolution ridge top and valley bottom flatness were important determinants of model accuracy. The variable importance results also illustrate the need to incorporate small-scale landscape context in modeling parent materials in BC because topographic derivatives

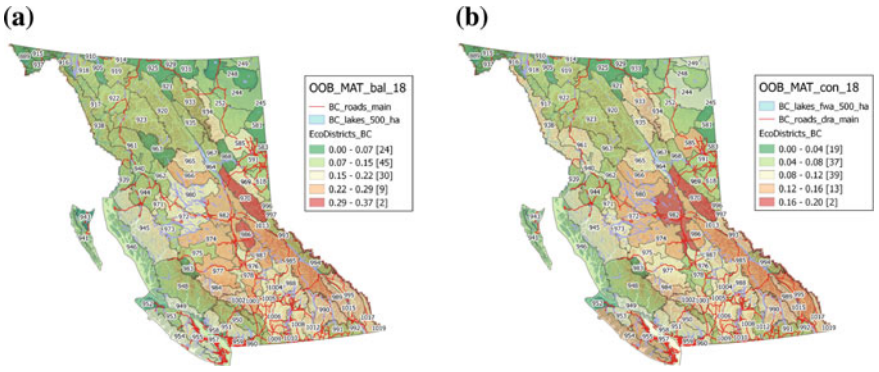

Fig. 24.2 Out-of-bag error rates by ecodistrict for **a** balanced and **b** constrained models for soil parent material

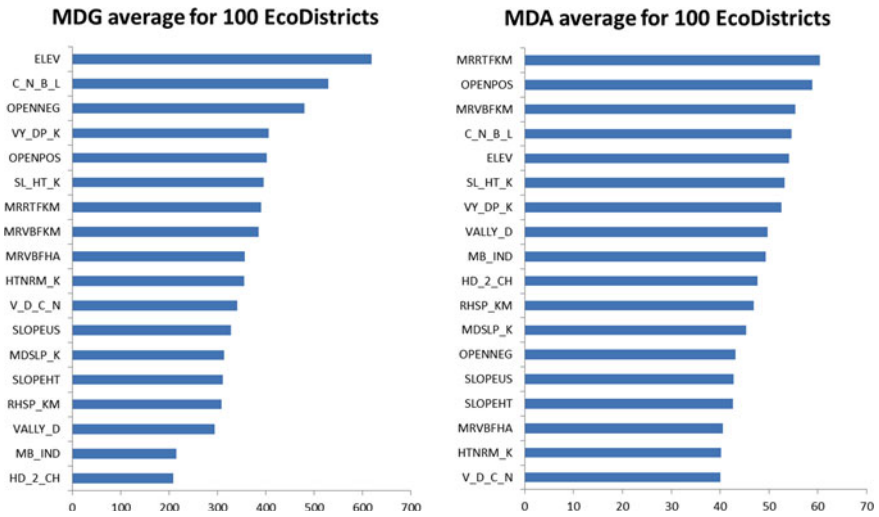

Fig. 24.3 Variable importance for parent material classification with balanced RF: average for 100 ecodistricts

calculated from the 1-km DEM tended to have higher importance than those same derivatives calculated from a 100-m grid. Many of the 100-m terrain derivatives we calculated did not appear in the reduced list of 18 predictors for final map production.

More than half of the ecodistricts shared very similar suites of important predictors, but a small number had RF models where only one or two of the five most important predictors were among the most important provincially (Fig. 24.4). Ecodistricts with similar variable importance metrics could potentially be combined into larger rule sheds, but for some ecodistricts, unique physiographic characteristics

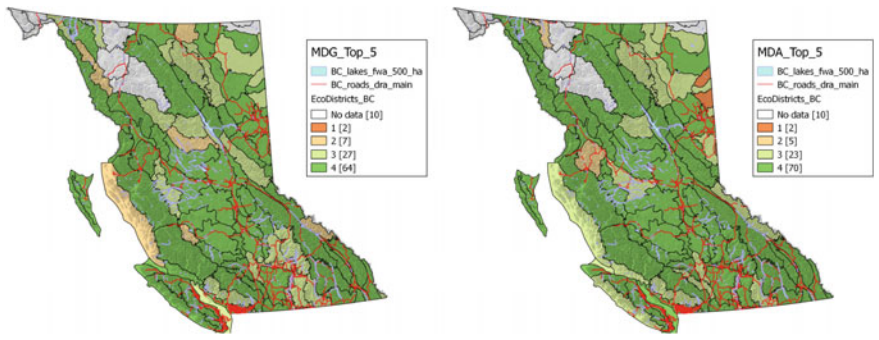

Fig. 24.4 Number of the top 5 provincial topographic derivatives that ranked in the top 5 for individual ecodistricts. Ecodistrict models relying on similar topographic attributes for node purity and accuracy have higher values

or other factors contribute to very different models, and possibly very different relationships between landscape characteristics and parent material distribution.

One of the major challenges in using legacy maps for training models is the inconsistent or changing concepts and definitions used in creating mapping units. For example, as a result of an evolving understanding of landscape processes in BC, different concepts of colluvium were sometimes employed by mappers on adjacent map sheets with similar topography (Fig. 24.5). The balanced RF model (Fig. 24.6a) predicted a more continuous distribution of colluvium in this same area. The constrained model limited training points for colluvium to those with steeper slopes (Fig. 24.6b) and also predicted a continuous, but more restricted distribution of colluvium. These results illustrate the capability of RF classification to produce consistent results even where the underlying training data were derived from mapping projects that were carried out at different times and by different surveyors.

Another challenge in using legacy data was observed in the prediction of classes with a very specific ecological niche. Organic parent materials, legend class 'UNDO' shown in green in Fig. 24.7, are closely associated with the presence of wetlands, but appeared to be over predicted by approximately 20 % on a provincial basis for both balanced and constrained RF models (Fig. 24.7a, b). An attempt to

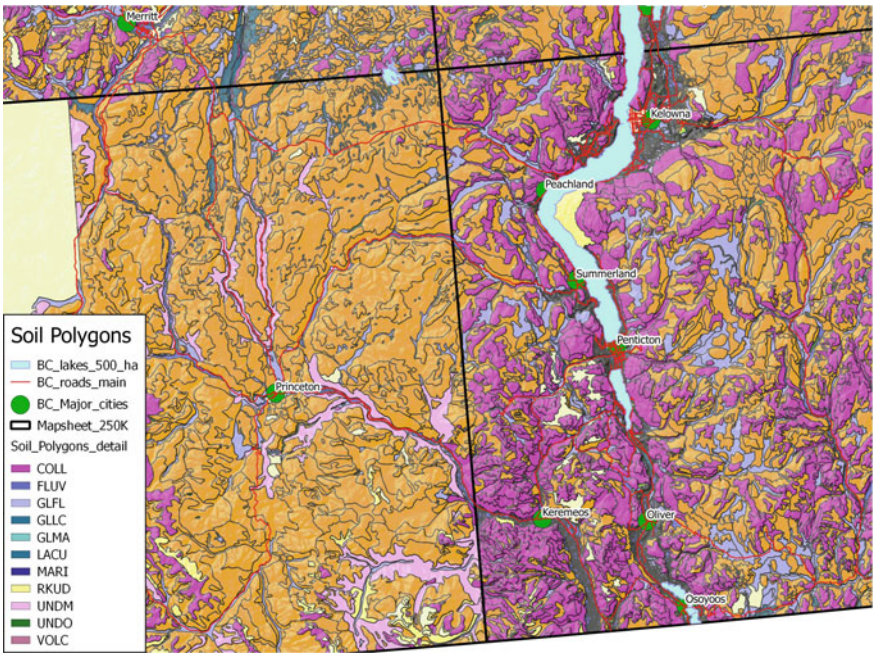

Fig. 24.5 Legacy map polygons for a portion of southern BC, showing different mapping concepts for colluvium (legend class 'COLL'). *Black lines* outline borders of adjacent map sheets. Parental material abbreviations: COLL = Colluvium, FLUV = Fluvial, GLFL = Glaciofluvial, GLMA = Glaciomarine, LACU = Lacustrine, MARI = Marine, RKUD = Bedrock, UNDM = Undifferentiated mineral, TILL = Glacial till, UNDO = Undifferentiated organic, VOLC = Volcanic

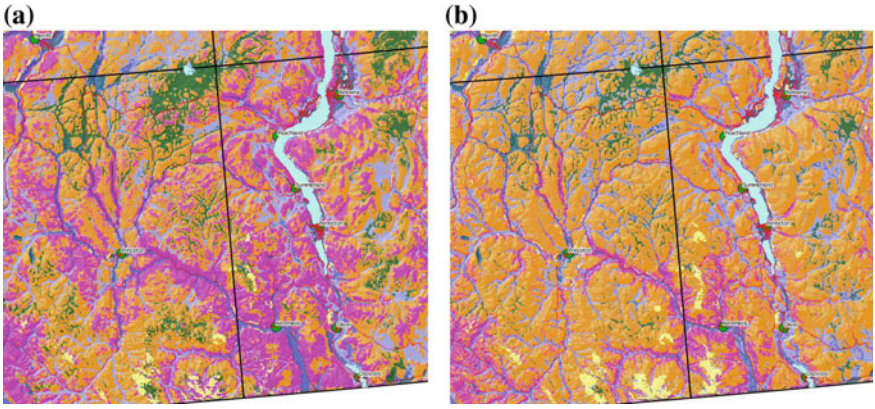

Fig. 24.6 Parent material predictions for colluvium (see legend in Fig. 24.5) employed in adjacent mapping projects, **a** the balanced prediction, and **b** the constrained prediction

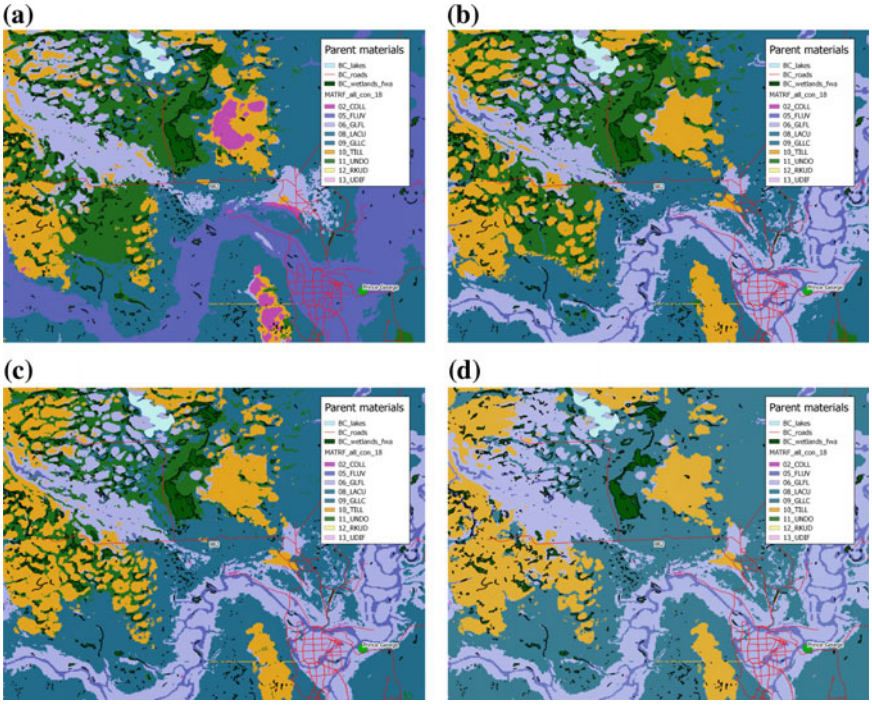

Fig. 24.7 Parent material predictions for a portion of ecodistrict 981, showing **a** balanced RF with organic parent materials (legend class '11_UNDO'), **b** constrained RF, **c** constrained RF with geographically restricted training points, and **d** organic parent materials forced on to known wetland locations

Table 24.3 Validation results for RF prediction of selected soil parent materials using balanced (Bal) and constrained (Con) models, expressed as percent correct

		BCSIS dataset			BEC points dataset		
Material	Symbol	Points	Bal %	Con %	Points	Bal %	Con %
Colluvium	02_COLL	2299	44.2	21.3	1767	53.0	25.2
Fluvial	05_FLUV	7310	56.2	56.7	1234	47.5	48.2
Lacustrine	06_LACU	834	31.3	31.6	262	3.1	4.2
Till	10_TILL	6264	34.6	45.3	1839	36.4	61.9
Organic	11_UNDO	1403	20.3	19.3	733	27.6	19.2
Marine	15_MARI	7241	35.8	34.8	–	–	–
Total points (average)		25351	41.1	41.5	5835	41.2	40.0

Fluvial materials were combined with glaciofluvial, and lacustrine was combined with glaciolacustrine

refine the predictions by providing RF with a more geographically focused training dataset consisting solely of points included within wetland polygons also resulted in significant over-prediction (Fig. 24.7c). We concluded that the best solution was to abandon predictive methods and to simply assign the organic materials class to all grid cells underlying existing wetland boundaries that were previously determined by landscape classification using aerial photographs (Fig. 24.7d).

We used two independent datasets to validate the results for parent material (Table 24.4). No ground sampling of the validation points was carried out. Results were generally similar for balanced and constrained models. The highest agreement occurred for the till material class, when validated against the BEC points. The percent correct in this study was consistently lower than values observed by Heung et al. (2014). The map produced by Heung et al. (2014) was in an area with consistent coverage of detailed mapping information, training data, and validation points. Our results reflect in part the inconsistent mapping conventions described earlier, but could also reflect problems with the validation data. In particular, many of the BCSIS points were collected decades ago before GPS was widely available and are known to have poor spatial accuracy. We believe that these results provide a lower bound on the reliability of the parent material predictions provided by RF.

24.4 Conclusion

The results of this project illustrate that digital soil mapping techniques provide practical approaches for improving soil information and its use for resource management in BC. Our approach facilitates improvement of the overall map mosaic as new data becomes available, or as alternative modeling approaches are tested and employed.

Although validation with two independent datasets produced some mixed results, we believe that the RF model provides valuable information for predicting

soil parent material and development, and with further refinement, it will prove to be an essential part of our overall goal to improve soil information in BC.

Legacy data contains a wealth of information, but considerable judgment is required to utilize it in predicting soil properties. For organic materials, simply mapping known locations of wetland organic deposits proved more efficient and accurate than predicting them with the covariates we used.

Our results should be considered a first step toward building improved digital datasets for soils in BC.

Acknowledgements This study is supported by British Columbia's Forest Science Program, and research programs within BC Ministry of Forests Lands and Natural Resource Operations, BC Ministry of Environment, and Agriculture and Agri-Food Canada. Simon Fraser University provided ongoing support for graduate students working on this project.

References

Anderson DW, and Smith CAS (2011) A history of soil classification and soil survey in Canada: Personal perspectives. Can. J. Soil Sci.:91:675-694.

Bui EN, and Moran CJ (2003) A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia. Geoderma 111:21–44.

Church M, and Ryder J (2010) Physiography of British Columbia. *In* RG Pike et al 2010. A compendium of forest hydrology. Land Management Handbook 66. BC Min. Forests. Available on the internet at www.for.gov.bc.ca/hfd/pubs/Docs/Lmh/Lmh66.htm (viewed 2014.09).

Ecological Stratification Working Group (1996). A National Ecological Framework for Canada. Agriculture and Agri-Food Canada, Research Branch and Environment Canada, State of Environment Directorate. Ottawa 125p.

Häring T, Dietz E, Osenstetter S, Koschitski T, and Schroeder B (2012) Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. Geoderma 185-186:37-47.

Heung B, Bulmer CE, and Schmidt MG (2014) Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma 214-215:141-154.

Hengl T, de Jesus JM, MacMillan RA, Batjes NH, Heuvelink GBM, et al. (2014) SoilGrids1 km — Global soil information based on automated mapping. PLoS ONE 9(8): e105992. doi:10.1371/journal.pone.0105992.

Kelley CC, and Spilsbury RH (1939) Soil survey of the lower Fraser Valley. Technical Bulletin 20. Canada Department of Agriculture, Ottawa. Available on the internet at www.sis.agr.gc.ca/cansis/publications/surveys/bc/bc1/index.html (viewed 2014.09).

Luttmerding HA, (1980) Soils of the Langley-Vancouver map area. Report No. 15. British Columbia Soil Survey. Victoria. Available on the internet at www.sis.agr.gc.ca/cansis/publications/surveys/bc/bc15/index.html (viewed 2014.09).

MacMillan RA, Moon DE, and Coupé RA (2007) Automated predictive ecological mapping in a Forest Region of B.C., Canada, 2001–2005. Geoderma 140:353-373.

Meidinger DV and Pojar J (1991) Ecosystems of British Columbia. BC Ministry of Forests, Special Report No. 6. Victoria. Available on the internet at www.for.gov.bc.ca/hfd/pubs/Docs/Srs/Srs06.htm (viewed 2014.09).

Province of BC (2010) Field Manual for Describing Ecosystems in the Field. Land management handbook No. 25 - 2nd edition. BC Ministry of Forests and Range, Victoria. Available on the internet at www.for.gov.bc.ca/hfd/pubs/docs/Lmh/Lmh25-2.htm (viewed 2014.09).

Sarmento EC, Giasson E, Weber E, Flores CA, and Hasenack H (2012) Prediction of soil orders with high spatial resolution: response of different classifiers to sampling density. Pesquisa Agropecuária Brasileira 47(9): 1395-1403. Available on the internet at www.scielo.br/scielo.php?pid=S0100-204X2012000900025&script=sci_arttext (viewed 2014.09).

Schut P, Smith S, Fraser W, Geng X, and Kroetsch D (2011) Soil Landscapes of Canada: Building a National Framework for Environmental Information. Geomatica 65(3):293-309.

Smith CAS, Daneshfar B, Frank G, Flager E, and Bulmer C (2012) Use of weights of evidence statistics to define inference rules to disaggregate soil survey maps. p 215-220 *In* Digital Soil Assessments and Beyond – Minasny, Malone & McBratney (eds) CRC Press, Leiden, The Netherlands. 466 pp.

Sondheim M, and Suttie K (1983) User Manual for the British Columbia Soil Information System, 1. BC Ministry of Forests Publication R28-82053, Victoria, BC.

Sprout PN, and Kelley CC (1964) Soil survey of the Kettle River Valley. Report No. 9. British Columbia Soil Survey, Victoria. Available on the internet at www.sis.agr.gc.ca/cansis/publications/surveys/bc/bc9/index.html (viewed 2014.09).

Stum AK, Boettinger JL, White MA, and Ramsey RD (2010) Random forests applied as a soil predictive model in Utah. *In* Digital Soil Mapping: Bridging Research, Environmental Application, and Operations. Progress in Soil Science 2, Springer.

Subberayalu SK, and Slater BK (2013) Soil series mapping by knowledge discovery from an Ohio County soil map. Soil Sci. Soc. Am. J. 77:1254-1268.

Valentine KWG, Sprout PN, Baker TE, and Lavkulich LM (1978) The Soil Landscapes of British Columbia. BC Ministry of Environment, Victoria. Available on the internet at www.env.gov.bc.ca/soils/landscape/index.html (viewed 2014.09).

Chapter 25
Disaggregation of Legacy Soil Maps to Produce a Digital Soil Attribute Map for the Okanagan Basin, British Columbia, Canada

Scott Smith, Denise Neilsen, Grace Frank, Eve Flager, Bahram Daneshfar, Glenn Lelyk, Elizabeth Kenney, Chuck Bulmer and Deepa Filatow

Abstract The Okanagan Basin is undergoing extensive hydrologic modeling in an effort to better understand regional water supply and demand issues. To assist in providing spatially explicit soil data to the modeling effort in this 8000-km^2 mountainous watershed in southern British Columbia, a digital soil map based on a 25-m DEM was compiled using a variety of methods. Legacy soil polygon maps exist for the basin at various scales. Because we lacked a comprehensive set of point (pedon) data required for geostatistical predictions, our objective was to disaggregate the legacy maps so as to assign to each grid cell soil class likelihood values and then soil attributes derived from the Canadian Soil Information System formatted to follow the GlobalSoilMap.net specifications. To do this, we used virtual point data generated by sampling homogeneous (single component) soil polygons and a range of covariates. On the upland where a good knowledge of the ecological distribution of soil series existed, we used an expert system of fuzzy logic inference using the ArcSIE software add-on. Predictors for modeling included high-resolution forest ecological zone maps, surficial geology maps, legacy soil

S. Smith (✉) · D. Neilsen · G. Frank · E. Flager
Agriculture and Agri-Food Canada, Science and Technology Branch, Summerland BC V0H 1Z0, Canada

B. Daneshfar
Agriculture and Agri-Food Canada, Science and Technology Branch, Ottawa, ON, Canada

G. Lelyk
Agriculture and Agri-Food Canada, Science and Technology Branch, Winnipeg, MB, Canada

E. Kenney
Consulting Pedologist, Agassiz, BC, Canada

C. Bulmer
British Columbia Ministry of Forests, Lands and Resource Operations, Vernon, BC, Canada

D. Filatow
British Columbia Ministry of Environment, Kelowna, BC, Canada

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_25

maps, and several derived terrain attributes. A limiting factor function was used to integrate the fuzzy membership values of all covariates to produce a single value for each soil series for each cell. For the intensively mapped valley floor, it was not possible to accurately disaggregate soil series from complex map polygons, many of which were defined based on subtle changes in subsurface soil texture for which we did not have meaningful predictors. For this region of the basin, we used simple polygon averaging to generate soil attributes for underlying grid cells. In a final step, we aggregated the 25-m data to a 100-m grid to provide data at a scale more suitable for some of the hydrologic modeling requirements.

Keywords Disaggregation · Legacy polygon soil maps · Fuzzy logic · Soil series · Attributes

25.1 Introduction

Ongoing modeling efforts in the Okanagan Basin of south central British Columbia require improved soil data. The basin forms the headwaters for the 8200-km^2 watershed of the Canadian portion of the Okanagan River system which flows southward through Washington state where it empties into the Columbia River (Fig. 25.1). The basin has a large range in elevation with cooler temperatures and higher precipitation found at higher elevations, and warmer, drier conditions found at low elevations. The climatic gradient results in the formation of soil types from grassland (Chernozems) to strongly acidic forest soils (Podzols). These soils are characterized by having a wide range of soil physical and chemical properties.

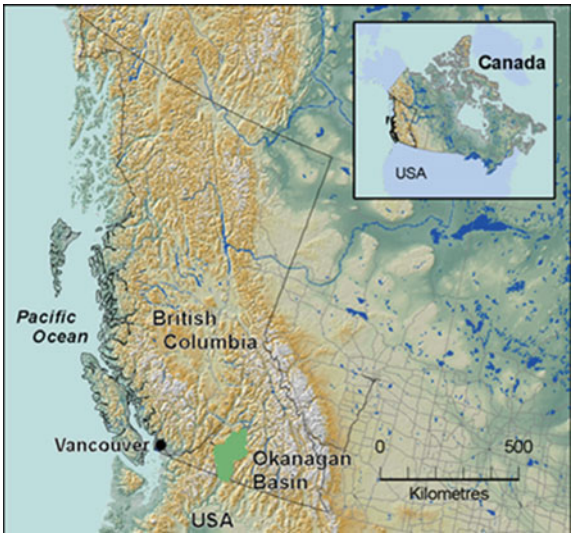

Fig. 25.1 Map of the location of the study area in mountainous south central British Columba. The area of the Okanagan Basin is highlighted in *green*

There are legacy soil maps at scales from 1:20,000 to 1:125,000 available for the basin. However, in many instances, traditional soil map data have proven to be difficult to use in environmental modeling applications when most inputs are in gridded format (Smerdon et al. 2010). In response to this need for improved data suitable for a variety of modeling efforts, we generated soil information in raster rather than vector format and organized by specific depth intervals rather than by soil horizon. Given the lack of reliable soil point data in the study area, disaggregation of the legacy soil polygon maps was our best option to provide raster-based soil class and attribute mapping suitable for modeling input. The objectives of this project were to (1) use methods tested in earlier research in a subwatershed (Smith et al. 2012) to produce a basin-wide raster soil class map using a 25-m digital elevation model (DEM) and selected environmental covariates, (2) assign attribute values for each soil class from data stored in the Canadian Soil Information System, and (3) reformat the horizon-based attribute values to a standardized depth interval for all soils following as closely as possible the GlobalSoilMap.net specifications (Science Committee 2011).

25.2 Methods

25.2.1 Harmonizing Legacy Soil Maps

When disaggregating multiple legacy soil maps that vary in scale and vintage, it is first necessary to spatially harmonize these into a seamless map coverage through polygon edge matching, re-projection, and data correlation. In our study area, we started with five original soil survey maps (Table 25.1). The spatial harmonization work and the construction of the soil attribute databases for all soil names (series) used on the maps were major tasks completed by Kenney and Frank (2010). Based on information contained in the original soil survey reports and from records in provincial pedon databases, a representative soil profile with horizon attributes was generated to characterize each soil series in the study area. While some of these pedons were sampled from within the study area, they lacked locational information necessary for their use as part of a geostatistical prediction approach. This compiled profile information was used to produce records within the British Columbia Soil

Table 25.1 List of legacy soil maps used as the basis for disaggregation

Soil survey name	Map scale	Projection	Vintage
Vernon	1:50,000	Geographic NAD 83	1986
Penticton	1:50,000	Geographic NAD 83	1986
Okanagan/Similkameen	1:20,000	UTM Zone 11	1986
North Okanagan	1:31,680	UTM Zone 11	1960
Tulameen	1:126,720	Geographic NAD 27	1974

Name Table and Soil Layer Table, standard tables following the specifications of the Canadian Soil Information System (Schut et al. 2011). The seamless polygon map and attribute records within the associated tables provided the starting point for our digital soil mapping efforts.

Even though all of the original maps had been spatially aggregated to produce a seamless coverage, there still remained the significant task of further correlating all soil series names used in the legacy mapping within the study area. This correlation took on two forms; determining the permitted geographic extent of use of names (how far from its original map origin could we predict a soil to occur) and examining the full set of names looking for redundancies (often two different soil series originating from two different soil surveys had almost identical pedological definitions). We found that within our study area, we had fairly discrete sets of soil series that covered distinct geographic regions of the basin, and another set of soils with rather wide use in legacy mapping. For each of these distinct regions, we selected a set of soil series to use in our predictions that characterized, to the best extent possible, the range of pedological niches that existed on the landscape (Table 25.2). We defined four distinct regions as individual 'rulesheds'. The boundaries of these rulesheds were defined based on the type and extent of the legacy map sheets with the exception of ruleshed 1 which is a subwatershed in which we conducted earlier research and retained the results for this basin-wide project. Soil series were generally not exclusive to a single ruleshed. A few series occurred in all rulesheds.

The reconnaissance mapping (Penticton, Vernon, Tulameen sheets) for the upland portions of the basin had polygon areas averaging 500 ha with up to three soil series described, often associated with different parent materials or distinct soil moisture regimes. In these upland areas, which were designated as rulesheds 1–3 (Fig. 25.2), polygon disaggregation was considered feasible given the covariates available. It is important to note that we did not disaggregate polygons individually, but rather we made spatial prediction of individual soils from a pool of soil series selected for a ruleshed.

For the area of detailed mapping (Okanagan-Similkameen, North Okanagan) covering the valley floor (ruleshed 4) where intensive irrigated agriculture is

Table 25.2 Design of rulesheds and handling of soil series defined on legacy soil survey maps

Physiographic setting	Uplands			Valley floor
Ruleshed	1	2	3	4
Name	Trout creek	Basin south	Basin north	Detailed
Total number of soil series used	24	39	33	118
Unique soil series	4	7	12	112
Shared soil series	20	32	21	6

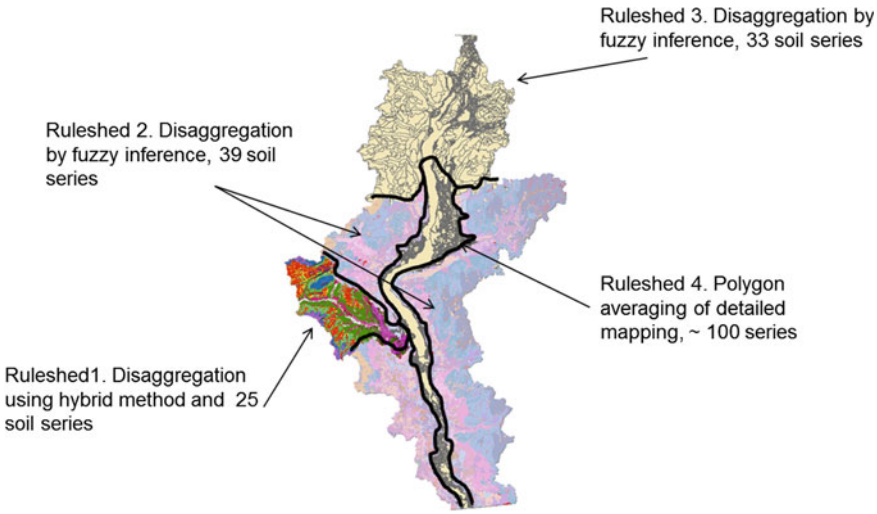

Fig. 25.2 Maps showing rulesheds used to constrain geographic extent and population of soil names predicted in different regions of the basin. Ruleshed boundaries were largely defined based on scale and extent of legacy map sheets

practiced, polygons averaged 17 ha in size with 60 % of these represented by a single soil series. In polygons where more than one series was listed, these often represented subtle textural variations and lithological discontinuities on a single parent material. We did not have covariates available to discriminate these differences, thereby making meaningful disaggregation in this ruleshed unfeasible. There were over 100 soil series defined in the detailed mapping on the valley floor. The bulk of these series did not occur in the upland areas.

25.2.2 Extracting Knowledge from Legacy Soil Maps

Several disaggregation methods were tested in the Trout Creek subwatershed (ruleshed 1) between 2010 and 2012. The project team gained experience with several statistical techniques including random forest, weights of evidence, logistic regression, and fuzzy inference (Smith et al. 2012). In this work, we developed a technique that we subsequently extended for use in rulesheds 2 and 3 whereby we established virtual sampling points within the soil polygons from the seamless map coverage to develop relations between individual soil series and sets of environmental covariates (Table 25.3). Polygons composed of 100 % of a single soil series were selected; where these were not available, polygons were sampled if they consisted of at least 80 % of that soil series. We spatially refined the sampling by using defining criteria from the soil survey reports provided for each series—specifically the ecological zone the series belonged to and the geologic parent

Table 25.3 Listing of covariates used in prediction methods

Covariate type	Description	Reference/source
Digital elevation models	Canadian digital elevation data 25 m	http://www.geobase.ca/geobase/en/data/cded/description.html
	Hectares BC 100 m	
Terrain derivatives (from 25-m DEM)	Aspect	
	Elevation	
	Topographic position index	Jenness (2006)
	Slope	
	SAGA wetness index	
	LandMapR facet classes	MacMillan (2003)
Remotely sensed	30-m land cover derived from LandSat and RadarSat imagery	Agriculture and Agri-Food Canada (2009)
Environmental maps	1:2,000 surficial geology	Filatow and Finvers (2009)
	1:50,000 BEC subzones	BC Ministry of Forests, land and resources (2011)
	1:100,000 generalized lithology mapping	

material. High-resolution vector maps of surficial geology and biogeoclimatic zones were available for the entire basin and used as covariates both in the spatial prediction and in refining the polygon sampling. Within a selected polygon, we only sampled points that also aligned to our defining criteria. For each soil series, up to 200 training points were generated from the refined polygons. If we could not find polygons suitable for sampling for a particular soil series, then we were simply not able to predict its occurrence even if it occupied a distinctive ecological niche. Fortunately, this occurred in only a few cases.

25.2.3 Predictive Methods

The following covariate map layers were used to enable spatial predictions: A filtered 25-m DEM generated from provincial 1:20,000 topographic contour mapping available from the Geobase Canada Web site, several terrain derivatives, a 1:20,000 vector surficial geology layer outlining individual soil parent materials, a 1:20,000 scale map of ecological subzones, and a 30-m raster land cover layer. The harmonized soil maps and 11 covariate layers were compiled into ArcGIS™ v10.1 for spatial analyses.

For ruleshed 1, we retained the original predictions based on a hybrid method using outputs from weights of evidence analyses to inform the setting of inference rule curves in ArcSIE as described in Smith et al. (2012). This method yielded good

prediction accuracy for soil classes and outperformed all other methods we tried in the ruleshed.

The boundary between ruleshed 2 and 3 was based on the boundary of two legacy map sheets (Penticton and Vernon) which used somewhat different soil legends; hence, different sets of soil series were predicted in the two rulesheds. For ruleshed 2, we selected some 39 soil series for prediction; for ruleshed 3, we selected 33 series. Most series occurred in both rulesheds, but there were several soils that were unique to only one ruleshed (Table 25.2). In both instances, we used fuzzy inference modeling using the spatial inference engine (SIE) extension within ARCGIS v.10.1.

ArcSIE©, a program and user interface that uses fuzzy logic to assign a membership value for individual soil series to each DEM grid cell (Shi 2010), was used to create the digital soil map in rulesheds 1, 2, and 3. The fuzzy logic model used, also called a similarity model by Zhu (1997), generates a membership value by integrating optimality values of environmental features at a location. The optimality values are defined by a Gaussian-style function curve that represents a rule created by the user to describe how an environmental feature (covariate) relates to a soil series. The Gaussian-style curve is defined by the user in the ArcSIE interface by setting the v, w, and r values of the curve. The v values define the limits of most optimal values for an environmental covariate, and the w and r values define the shape of the curve. For details of the curve function, readers are referred to Shi (2010). In this study, we automated this process to the extent possible and used an iterative approach in making predictions. We modified curves through expert knowledge to refine predictions and for some soil series reduced the number of covariates used in the modeling which improved results. Zhu et al. (2010) discuss in detail the use of fuzzy logic approaches in digital soil mapping to capture expert knowledge to produce both a class map and to predict spatial variation of soil attribute through membership (likelihood) calculations.

The membership values for each environmental covariate for each soil series were integrated using a limiting factor function within the software, which is the equivalent of the fuzzy AND operator. This operator chooses the lowest optimality (membership) value of all environmental features as the overall optimality value for the location. When the inference is run, a membership grid is produced for each of the soil series included in the rule base for the ruleshed. The ArcSIE Harden Map tool was used to generate the 'hardened' soil series grid, where each cell of the grid contains a value that represents the soil series with the highest membership value for that location. The membership value is not a probability of whether a certain soil class (series) occurs at a location or not. It is an index which measures the similarity between the properties of a given soil series and the environmental properties at a given location. In mathematical terms, the index measures the level to which the pixel can be considered a member of the set representing the assigned soil.

We undertook no predictive mapping in ruleshed 4, and the area of detailed mapping on the valley floor where disaggregation was not feasible.

25.2.4 Linking Soil Attributes to Grid Cells

Predictions result in likelihood (membership) values of every soil series in the ruleshed. Attributes may be assigned based on the most likely soil series or weighted by likelihood value for all or some subset (top 3 or 5 series) of series as was explored by Lelyk et al. (2014). As described earlier, each series is represented by a set of horizon attributes from the British Columbia Soil Layer Table, a subset of the Canadian Soil Information System. Horizon attribute data were fit to specific depth intervals following the concepts of Malone et al. (2009) modified using a horizon weighted averaging method as described by Lelyk et al. (2014). Examples of this transformation for attributes for two soils in the study area are given in Fig. 25.3.

In ruleshed 4, we used simple polygon averaging to generate attribute values for the 25-m grid cells. In instances where only a single soil series was listed in a polygon, that series and its attributes were simply assigned to all grid cells spatially underlying the polygon. Where multiple components were listed, the polygon

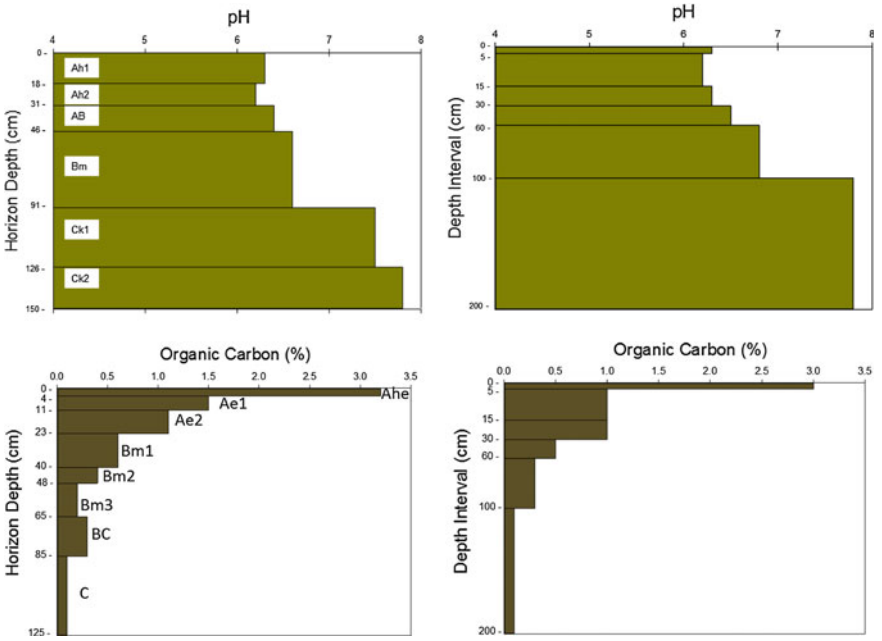

Fig. 25.3 Attribute transformation from horizon-based values to specified depth interval-based values. The upper panels illustrate the pH transformation for the Alleyne soil series (a mid-elevation forested soil) and the lower panels illustrate the organic carbon transformation for the Armstrong soil series (a lower-elevation grassland soil). Both soils occur in several rulesheds in the basin

averaging approach assumes that we are unable to predict the internal short range variation in any soil property within the polygon and that the best estimate of the most likely value for a soil property in that map unit is a weighted mean of the values for all soils in the polygon. The weighting factor is derived from the estimated proportion or extent of each soil component in the polygon. Details of the polygon averaging method we used are given in Hempel et al. (2012). To create a soil class map at 25 m resolution for ruleshed 4, we simply assigned the dominant soil name to all grids underlying the polygon. As stated earlier, two-thirds of the detailed polygons contained only one soil series.

25.3 Results

25.3.1 Soil Class Maps

Several outputs resulted from our methods. The rulesheds were merged to produce a single predicted soil class map for every 25-m grid cell in the 8000-km^2 watershed of the Okanagan Basin. The most likely soil series in each cell was used to create the map and the grid database also contains a confusion index value (Burrough et al. 1997) based on the ratio of likelihood values for the top two most likely classes which provides a simple measure of uncertainty for the prediction class.

25.3.2 Soil Attribute Maps

The attributes included were those defined in the GlobalSoilMap specifications (Science Committee 2011) for primary soil attributes although we report these in units used in the Canadian Soil Information System. These attributes include soil pH (in $CaCl_2$), soil organic carbon (%), sand, silt and clay (%), coarse fragment content (%), electrical conductivity (dS/m), and available water holding capacity (% vol). The spatial distribution patterns of attributes become clearly evident when the data are mapped to the 25-m grid cells. For example, soil pH correlates closely to precipitation. Areas of the basin with highest precipitation, such as the subalpine forest zone, have the lowest soil pH values as shown in Fig. 25.4. On a more local scale, the grids provide good spatial resolution of the extent of highly alkaline soils associated with glaciolacustrine deposits found along the shores of Okanagan Lake (see inset map, Fig. 25.4). A major limitation to the accuracy of these attribute maps is the fact that we assign a single attribute value to a soil class based on a national soil database value and this is represented everywhere that soil class is predicted.

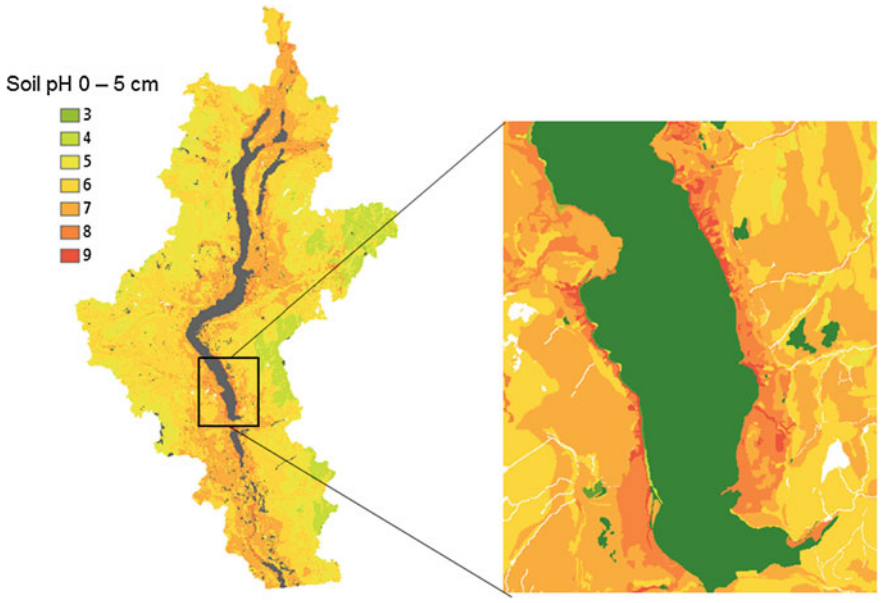

Fig. 25.4 A digital soil pH map for the Okanagan Basin. Regional and local patterns of soil pH are evident as depicted on a 25-m grid. A total of eight attribute maps were produced. Large geographic feature in valley bottom is Okanagan Lake

25.3.3 Scaling-up

Some modeling efforts are better served by data provided at lower resolution. Several users have requested data at courser resolution. To satisfy this need, 25-m grid data were scaled up to 100 m resolution. This step, referred to as conflation by Malone et al. (2013), is a relatively simple process involving the averaging of the fine-scaled values (25-m grid) to generate an overall mean of the target variable across a coarse-scale grid (100 m) over the same map extent. With respect to soil class, we list likelihood values for the most common classes occurring in the 100-m cell based on the roll-up of the sixteen underlying 25-m class values (Fig. 25.5). The end result is a product more easily managed by some environmental models such as the integrated catchment model such as MIKE SHE (Refsgaard et al. 2010) and other semi-distributed hydrology models used for regional water supply and demand assessments.

In a similar approach, attribute values for each soil depth interval are generated for each 100-m cell by weighted averaging values from the underlying 25-m cells. In this way, a mean and a range of values are reported for each cell giving the user some sense of the data variability and uncertainty for each cell.

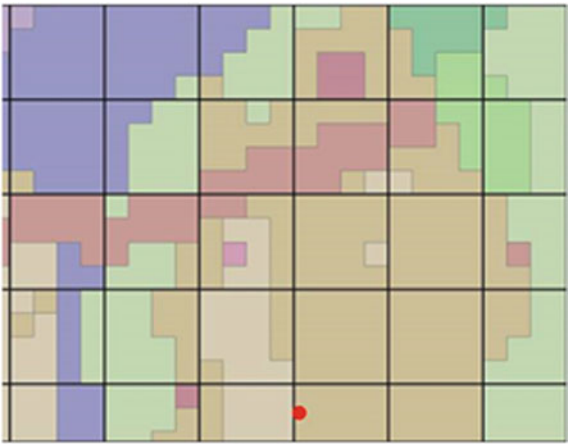

Fig. 25.5 A 100-m grid superimposed onto the 25-m grid. Most of the 100-m cells are dominated by a single soil class but not all. Scaling-up provides a method to better quantify the range of classes and attribute values in each cell

25.4 Conclusion

Producing digital soil maps entirely from legacy polygon maps presents significant challenges. The methods used in this project drew heavily on expert knowledge of local pedological conditions and would be difficult to apply if that knowledge did not exist. Because of the manual intervention in the creation of many (but not all) covariate rule curves used in the inference modeling, the disaggregation can be both labor-intensive and difficult to optimize. These are drawbacks to using this type of approach as opposed to geostatistical processes that can benefit by simply adding more point data to the process. Our methods did, however, effectively transform the legacy class and attribute data from vector to raster format.

We retained the soil class map to facilitate some existing land suitability algorithms used by Agriculture and Agri-Food Canada that draw upon horizon data contained in the Soil Layer Table. Running these crop suitability modules against gridded data rather than map polygons greatly enhances the spatial resolution of the model output. Attribute mapping at both the 25- and 100-m resolutions greatly facilitates environmental modeling both and resolutions. Ultimately the value of the Okanagan Basin digital soil map products will be measured by their performance in effectively delivering soil information to the modeling activities in the region. Gridded data, suited for some applications, will not replace entirely the use of polygon maps which have historically been used in a range of land planning and zoning applications.

A final step remains to field validate both our class predictions and attributes as given on this digital map. Validation will highlight where we might need to modify our predictive methods and allow us to inform users of the uncertainties associated with the values presented.

Acknowledgements This study was supported by funding from Agriculture and Agri-Food Canada, Science and Technology Branch Project Numbers 13–1166 and 10–1079 and in-kind contributions from the BC Ministry of Environment and Ministry of Forests, Land and Resource Operations. The authors are grateful to Bob MacMillan for his training efforts, encouragement, and technical advice.

References

Agriculture and Agri-Food Canada, 2009, Land Cover for agri-cultural regions of Canada, circa 2000 Available http://open.canada.ca/data/en/dataset/16d2f828-96bb-468d-9b7d-1307c81e17b8

British Columbia Ministry of Forests, Lands and Natural Resource Operations 2011. Biogeoclimatic zone and subzone maps. Forest Analysis and Inventory Branch, Victoria, BC Available at http://www.for.gov.bc.ca/hre/becweb/resources/maps

Burrough, P.A., van Gaans,P.F.M. and Hootsmans R. 1997. Continuous classification in soil survey: spatial correlation, confusion and boundaries. Geoderma 77: 115-135.

Filatow, D and M. Finvers 2009. British Columbia terrain, soil and ecosystem mapping databases now available online. Streamline Watershed Management Bulletin 12: 13-17.

Hempel, J. W., Z. Libohova, N. P. Odgers, J.A. Thompson, C.A.S. Smith and G.L. Lelyk. 2012. Versioning of GlobalSoilMap.net raster property maps for the North America Node. Pages 429-434 **in** B. Minasny, B.P. Malone and A.B. McBratney (eds). Digital Soil Assessments and Beyond. CRC Press, Leiden, The Netherlands. 466pp.

Jenness, J. 2006. Topographic Position Index extension for ArcView3.x, v. 1.3a. Jenness Enterprises. Available at: http://www.jennessent.com/arcview/tpi.htm

Kenney E. and G. Frank 2010. Creating a seamless soil dataset for the Okanagan Basin, British Columbia. in Proceedings of the Western Regional Cooperative Soil Survey Conference, Las Vegas, NV. USDA-Natural Resources Conservation Service. Available http://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/partnership/ncss/?cid=nrcs142p2_053514

Lelyk, G.W., R.A. MacMillan, S. Smith and B. Daneshfar 2014. Spatial disaggregation of soil map polygons to estimate continuous soil property values at a resolution of 90 m for a pilot area in Manitoba, Canada. p 201-207 **in** Arrouays et al. (eds) GlobalSoilMap: Basis of the global spatial soil information system. CRC Press/Balkema. Leiden, The Netherlands. 477pp.

MacMillan, T. A. 2003. LandMapR© Software Toolkit- C++ Version: Users Manual. LandMappR Environmental Solu-tions Inc., Edmonton, AB. 110pp.

Malone, B.P., A.B. McBratney, B. Minasny and G.M Laslett 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154, 138–152.

Malone, B.P., McBratney, A. B. and Minasny, B. 2013. Spatial scaling for digital soil mapping. Soil Sci. Soc. Am. J. 77: 890-902.

Refsgaard, J.C., Storm, B. and Clausen, T. 2010. Système Hydrologique Europeén (SHE): review and perspectives after 30 years development in distributed physically-based hydrological modelling. Hydrology Research 41: 355–377 doi:10.2166/nh.2010.009

Schut, P., Smith, S., Fraser, W., Geng, X. and Kroetsch, D. 2011. Soil Landscapes of Canada: Building a national framework for environmental information. Geomatica 65(3):293-309.

Science Committee 2011. GlobalSoilMap.net Product specification Version 1 Release 2.1. URL http://www.globalsoilmap.net/specifications.

Shi, X. 2010. ArcSIE user's guide. Lebanon, NH. 119pp. Available at http://www.arcsie.com/download.htm

Smerdon, B. D., Allen, D. M. and Neilsen, D. 2010. Evaluating the use of a gridded climate surface for modelling groundwater recharge in a semi-arid region (Okanagan Basin, Canada). Hydrological Processes 24(21):3087-3100.

Smith, C. A. S., Daneshfar, B., Frank, G., Flager, E. and Bulmer, C. 2012. Use of weights of evidence statistics to define inference rules to disaggregate soil survey maps. Pages 429-434 **in**

B. Minasny, B.P. Malone and A.B. McBratney (eds). Digital Soil Assessments and Beyond. CRC Press, Leiden, The Netherlands. 466pp.

Zhu A-Xing 1997. A similarity model for representing soil spatial information. Geoderma 77: 217-242.

Zhu A-Xing, Lin Yang, Baolin Li, Chengzhi Qin, Tao Pei and Baoyuan Liu. 2010. Construction of membership functions for predictive soil mapping under fuzzy logic. Geoderma 155: 164–174.

Chapter 26
Comparison of Different Strategies for Predicting Soil Organic Matter of a Local Site from a Regional Vis–NIR Soil Spectral Library

Rong Zeng, Yu-Guo Zhao, Deng-Wei Wu, Chang-Long Wei and Gan-Lin Zhang

Abstract Soil spectral libraries were established all over the world to help build the base for predicting soil properties by proximal soil sensing. Previous studies indicated that it was important to select optimum subsets when predicting soil properties of a local site from a large spectral library. Thus, how to determine optimum subsets from the spectral library becomes crucial. This study compared different strategies for predicting soil organic matter of a local site from a regional Vis–NIR soil spectral library. Different calibration subsets and two calibration models [local and global partial least squares regression (PLSR)] were assessed for prediction of the target set: (1) different calibration subsets were compared (Pro_cali, samples in the province; Hb_cali, samples in Huaibei area, geographically close, and with similar parent material compared to the target set; Local_cali, samples located in the same county of the target set); (2) the spiking effects were investigated by selecting different numbers of local samples from Local_cali using Kennard–Stone algorithm to be spiked with different calibration sets (Pro_cali and Hb_cali); (3) local PLSR and global PLSR calibrations were compared for prediction accuracy. Model performances were assessed in terms of coefficient determination between observed and predicted values (R^2), root-mean-squared error for prediction (RMSEP), and the ratio of percentage deviation (RPD). In general, this study concluded that (1) prediction performances of different calibration subsets indicated that Hb_cali can be a good alternative to replace Local_cali for prediction, when local samples are not available; (2) the spiking effects depended on the number of spectra spiked, also it did not always lead to higher prediction

R. Zeng · Y.-G. Zhao · D.-W. Wu · C.-L. Wei · G.-L. Zhang (✉)
State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, People's Republic of China
e-mail: glzhang@issas.ac.cn

R. Zeng · D.-W. Wu · C.-L. Wei · G.-L. Zhang
University of the Chinese Academy of Sciences, Beijing 100049, People's Republic of China

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_26

accuracy; and (3) global PLSR and local PLSR exhibited similar prediction accuracy in this case study, more research were needed to compare the performances of these two models.

Keywords Vis–NIR · Regional spectral library · SOM · Calibration subsets · Spiking

26.1 Introduction

The development of visible and near-infrared (Vis–NIR) spectroscopy has provided an alternative to predict soil properties, because it is cost-effective, time-saving, and nondestructive compared to traditional laboratory analysis (Brown et al. 2006; Rossel et al. 2006). In order to improve prediction accuracy, Vis–NIR soil spectral libraries have been built at scales of local, regional, country, continental, and global (Shepherd and Walsh 2002; Brown 2007; Rossel et al. 2008; Rossel et al. 2009; Shi et al. 2014). When we need to predict soil properties of a specific local site from a large spectral library, we need to find the subsample (which could be the whole library) which will give the best prediction (Araujo et al. 2014; Gogé et al. 2014). The accuracy achieved by simply calibrating all the spectra data in a large library to predict for a local site is generally not good (Brown 2007; Wetterlind and Stenberg 2010). Because the library contains information from a wide variety of soils not similar to those of interest in the local area, using it directly introduces noise with respect to the local samples and thus reduces prediction accuracy.

There are several methods to localize spectral libraries, among which subset selection and spiking were reported in recent research to be quite successful (Guerrero et al. 2010; Kuang and Mouazen 2013). Subset selection is used to select subsamples from a large spectral library according to different rules, such as distance proximity and spectral similarity. Local calibration is subset selected based on distance proximity from the target site, while spiking becomes a compromise alternative by adding a certain number of local samples into calibration sets because it is usually not practical to acquire a large number of local samples. Previous studies have suggested that the use of subset selections and spiking can help extract useful information from a large library to help explain the variance of a target property for a specific site (Guerrero et al. 2010; Kuang and Mouazen 2013). Brown (2007) achieved improved prediction accuracy for soil organic carbon (SOC) and clay estimation of samples in a 2nd-order Ugandan watershed by spiking a global Vis–NIR soil spectral library with very few local samples. Guerrero el al. (2014) investigated the effects of selection and extra-weighting on the spiking subset, even the addition of only 8 local samples can lead to improved accuracy of SOC predictions.

In addition, local partial least squares regression (PLSR) has also been suggested to be a useful method for predicting soil properties of a local site from a large

spectral library (Naes et al. 2002; Fearn and Davies 2003). Because global PLSR calibration generates one model for the target set while the local PLSR model generates one model for each sample in the target set. Previous study indicated that local PLSR model can solve the nonlinearity of a large spectral library and generally yielded higher prediction accuracy compared to global PLSR model (Sankey et al. 2008; Gogé et al. 2012; Nocita et al. 2014).

The aim of this study was to compare different strategies for predicting SOM of a local site from a regional Vis–NIR soil spectral library. The main objectives are as followed: (1) Compare the prediction performances of different calibration subsets; (2) Investigate the effect of spiking after adding local samples into calibration subsets; and (3) Compare the model performances between the method of global PLSR and local PLSR.

26.2 Materials and Methods

26.2.1 Regional Spectral Library

The regional spectral library was built from 1580 soil samples collected in Anhui Province, China (Fig. 26.1), supported by previous soil investigation projects. Sample points almost cover the whole area of Anhui Province, with three counties

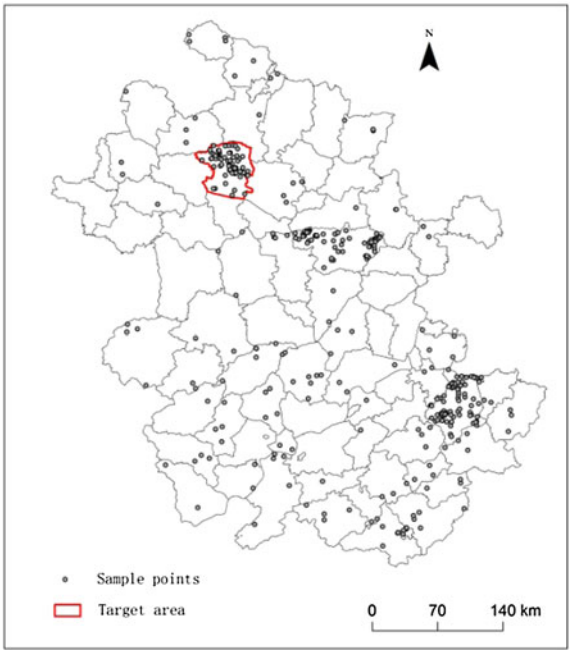

Fig. 26.1 Sample distribution map

densely (Mengcheng, Dingyuan, and Xuanzhou) sampled. Mengcheng County was selected as the prediction target area (highlighted in red in Fig. 26.1).

Soil samples were air-dried, ground, and passed through a 100-mesh sieve and oven-dried for 24 h 350–2500-nm spectra were measured using Cary 5000 under controlled laboratory conditions.

26.2.1.1 Target Set

Samples located in Mengcheng County (202 samples) was chosen as the target dataset in this study. This dataset was divided into local calibration (Local_cali, 152 samples) and local prediction (Local_pre, 50 samples) set using Kennard–Stone algorithm.

26.2.1.2 Calibration Set

In order to evaluate the impacts of different calibration subsets on the model prediction ability, three calibration sets were built: (1) local calibration set: the aforementioned Local_cali with 152 samples, which are mostly geographically closer to the target set; (2) Huaibei calibration set (Hb_cali, samples in Mengcheng County were not included): samples located in the north of Huaihe River, which are geographically close and have similar parent material compared to the target set; and (3) province calibration set (Pro_cali): all samples except those in Mengcheng County.

To investigate the effects of spiking on model performances, different numbers of spectra were selected from Local_cali using Kennard–Stone algorithm to be spiked with two other calibration subsets (Hb_cali and Pro_cali).

26.2.2 Spectral Preprocessing

Several spectral transformations were explored on the whole dataset using cross-validation: absorbance, first derivative, second derivative. Absorbance yielded the highest prediction accuracy and was used for model calibration and prediction in this study.

26.2.3 Local PLSR Model and Global PLSR Model

PLSR models were used in this study. Local PLSR calibration and global PLSR calibration were compared. The main difference between local PLSR and global PLSR is that global PLSR calibrates one model for all samples in Local_pre, while

local PLSR calibrates one model for each sample in Local_pre. The models were compared using following steps.

(1) Global model: 300 (different numbers of similar spectra were compared by cross-validated PLSR models, the highest accuracy was achieved when the number was 300) most similar spectra compared to Local_pre were selected from Pro_cali using spectral angle mapper (SAM) to make up the calibration set. Based on this dataset, a calibration model was constructed and then used for prediction of Local_pre.
(2) Local model: For each sample in Local_pre, 300 most similar spectra were selected from Pro_cali using SAM to make up the calibration set, which was subsequently used for prediction of the aforementioned sample.
(3) Spiking effects: The spiking effects were evaluated by selecting different numbers of spectra from Local_cali using Kennard–Stone algorithm to be spiked with the calibration subsets described in (1) and (2).

Global PLSR was performed in Unscrambler 9.3 (Camo Software AS), while local PLSR was performed using MATLAB R2012a (Mathworks, Massachusetts, USA). The flowchart of this study is illustrated in Fig. 26.2.

PLSR model accuracy was assessed in terms of coefficient determination (R^2) between observed and predicted values, root-mean-squared error for prediction

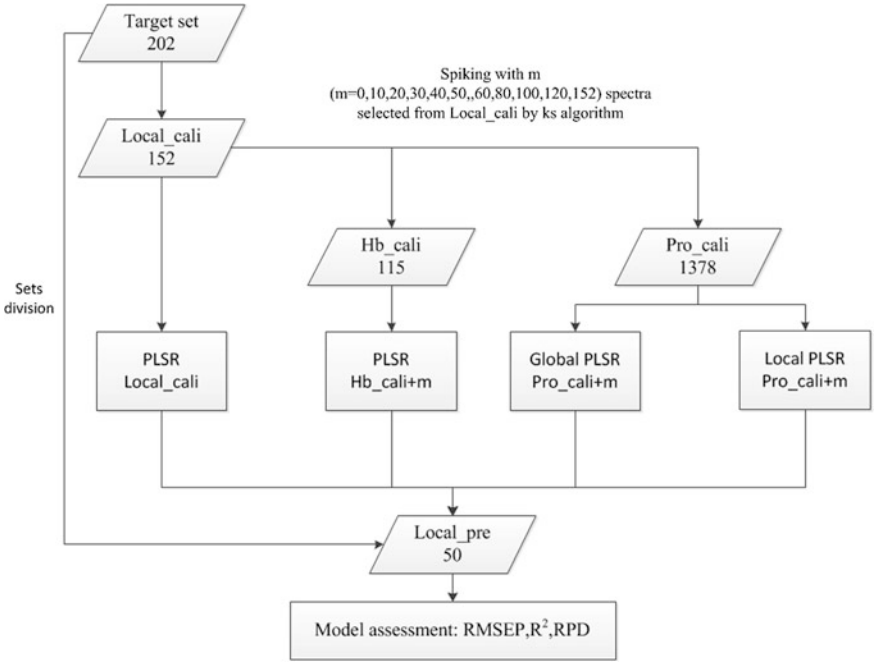

Fig. 26.2 Flowchart

(RMSEP), and the ratio of percentage deviation (RPD), calculated using the following equations:

$$R^2 = \frac{[\mathrm{cov}(\hat{y}_i, y_i)]^2}{\mathrm{var}(\hat{y}_i).\mathrm{var}(y_i)} \tag{26.1}$$

$$\mathrm{RMSEP} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{26.2}$$

where n is the number of sample in the target set, y_i is the observed value of sample i, and $\widehat{y}_i$ is the predicted value of sample i.

$$\mathrm{RPD} = \mathrm{SD}/\mathrm{RMSEP} \tag{26.3}$$

where SD is the standard deviation of observed values.

26.3 Results and Discussion

26.3.1 Statistical Summary of SOM Content for Different Datasets

SOM contents in Pro_cali dataset vary dramatically from 0.03 g kg^{-1} to 60.14 g kg^{-1}, while the contents in other there datasets are within the range of 0.25–30.60 g kg^{-1} (Table 26.1). Similar SOM contents between Hb_cali and Local_cali can be explained by the proximity in distance and similar parent material. Local_pre is more or less within the range of the three calibration datasets only except the maximum value (29.03 g kg^{-1}) is a bit larger than the maximum value (26.36 g kg^{-1}) of Hb_cali dataset.

Table 26.1 Statistical summary of SOM content for different datasets

Datasets	*N*	Min g kg^{-1}	Max g kg^{-1}	Mean g kg^{-1}	SD g kg^{-1}	CV	Skew
Pro_cali	1378	0.03	60.14	12.02	9.93	0.83	1.49
Hb_cali	115	0.25	26.36	8.75	6.13	0.70	1.01
Local_cali	152	0.27	30.60	11.58	7.32	0.63	0.40
Local_pre	50	0.73	29.03	11.64	7.26	0.62	0.39

N number of samples; *Min* minimum; *Max* maximum; *SD* standard deviation; *CV* coefficient of variation

26.3.2 Prediction Accuracy at Different Model Scales

Prediction accuracy were compared among three calibration datasets in terms of RMSEP, R_p^2, and RPD (Table 26.2). Local_cali achieved the highest prediction accuracy, which is in accordance with most previous studies (Brown 2007; Wetterlind and Stenberg 2010). "Local" calibration datasets indicate the similarity or homogeneity between samples for calibration and prediction in many aspects, such as soil type, parent material, land use, and SOM contents. The similarity features of local samples can help build models better explain the variance of the target set.

Pro_cali yielded the lowest accuracy as expected because of its big diversity in sample distribution, but the accuracy was acceptable (R_p^2 of 0.77 and RPD of 1.30). The prediction accuracy of Hb_cali was very similar to that of Local_cali, which provides a very good alternative for predicting SOM contents of a local site from a regional, national, or global library, when local samples are not present. Because in real scenario, local samples are often difficult to acquire considering the limitation of projects' budgets, Hb_cali was constructed based on expert knowledge and legacy data of soil parent materials and land use.

26.3.3 Spiking Effects

26.3.3.1 Pro_Cali Spiked with Local Samples

As previously mentioned, local samples are usually difficult to obtain in large number for independent model calibration. However, they can still be useful by spiking with other available datasets. Different numbers of local samples (m = 10–152) were added into Pro_cali to investigate the spiking effects (Table 26.3). The prediction accuracy generally increased with m in spite of some fluctuations. But the accuracy improvement was slightly compared to the models built upon Pro_cali alone. Furthermore, regardless of the number of local samples added, the accuracy achieved is always lower than that of Local_cali and Hb_cali. The results of this study were different from some previous studies; other studies demonstrated

Table 26.2 Prediction accuracy at different model scales

Calibration datasets	N	RMSEP g kg^{-1}	R_p^2	RPD
Local_cali	152	3.05	0.83	2.38
Hb_cali	115	3.60	0.82	2.02
Pro_cali	1378	5.54	0.77	1.30

N number of samples; *RMSEP* root-mean-squared error of prediction; R_p^2 coefficient determination of prediction; *RPD* ration of percentage deviation

Table 26.3 Prediction accuracy of models based on Pro_cali spiked with *m* local sample from Local_cali

m	R_p^2	RMSEP g kg^{-1}	RPD
0	0.77	5.54	1.3
10	0.77	4.84	1.5
20	0.77	4.18	1.74
30	0.74	3.81	1.91
40	0.77	4.1	1.77
50	0.76	4.52	1.61
60	0.77	4.72	1.54
80	0.77	3.7	1.96
100	0.74	3.84	1.89
120	0.79	3.95	1.84
152	0.77	3.90	1.86

m number of samples chosen from Local_cali; *RMSEP* root-mean-squared error of prediction; R_p^2 coefficient determination of prediction; *RPD* ration of percentage deviation

improved prediction accuracy after adding local samples, even with a small number (Brown 2007; Guerrero et al. 2014). The reason may be that the number of local samples added (maximum is 152) is small compared to the size of Pro_cali (1378). The spiking of local samples only give a small weight on the calibration dataset (Pro_cali), leading to the slight change of model performances.

26.3.3.2 Hb_Cali Spiked with Local Samples

The spiking effects for Hb_cali were quite different from that of Pro_cali. Model accuracy decreased gradually when the number of local samples (*m*) increased (Table 26.4). This is opposite to what we expected. It seems that spiking does not necessarily always lead to better prediction accuracy. It also depends on the distribution and relationship between target set and spiked set.

26.3.4 Comparison of Global PLSR and Local PLSR

The performances of global PLSR and local PLSR were compared in terms of RMSEP and R^2 in Fig. 26.3. For RMSEP of global PLSR, there were some fluctuations as m changed, while for local PLSR, RMSEP increased firstly as local spectra were added, then decreased and stabilized as m increased from 50 to 152. As for R^2, global PLSR performed slightly better than local PLSR regardless of the number of added spectra. In general, there was slight difference between these two models.

Table 26.4 Prediction accuracy of models based on Hb_cali spiked with *m* local sample from Local_cali

m	R_p^2	RMSEP g kg^{-1}	RPD
0	0.82	3.60	2.02
10	0.77	5.67	1.28
20	0.74	5.8	1.25
30	0.72	5.45	1.33
40	0.72	4.79	1.52
50	0.66	4.44	1.64
60	0.67	4.6	1.58
80	0.64	4.97	1.46
100	0.64	4.64	1.56
120	0.69	4.49	1.62
152	0.66	4.98	1.46

m number of samples chosen from Local_cali; *RMSEP* root-mean-squared error of prediction; R_p^2 coefficient determination of prediction; *RPD* ration of percentage deviation

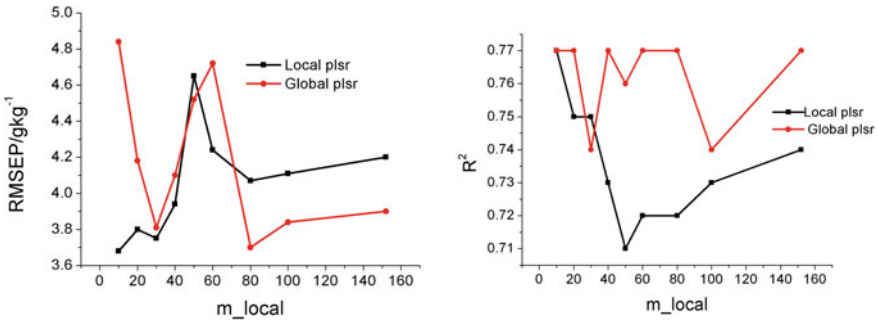

Fig. 26.3 Local PLSR versus global PLSR

26.4 Conclusions

In general, this study showed that (1) Pro_cali-based models achieved the lowest but reasonable prediction accuracy, while Hb_cali and Local_cali achieved similar prediction performances. Prediction performances of different calibration subsets indicated that Hb_cali can be a good alternative to replace Local_cali for prediction, when local samples are not available; (2) the spiking effects depended on the number of spectra spiked, also it did not always lead to higher prediction accuracy; and (3) global PLSR and local PLSR exhibited similar prediction accuracy in this case study, more research were needed to compare the performances of these two models.

Acknowledgements The study was supported by the National Science Foundation of China (41130530, 91325301). The authors are grateful to Dr. David Rossiter for his comments and suggestions.

References

Araujo SR, Wetterlind J, Demattе JAM, Stenberg B (2014) Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. European Journal of Soil Science **65**(5): 718–729.

Brown DJ (2007) Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma **140**(4): 444–453.

Brown DJ, Shepherd KD, Walsh MG, Mays MD, Reinsch TG (2006) Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma **132**(3–4): 273–290.

Fearn T, Davies A (2003) Locally-biased regression. Journal of Near Infrared Spectroscopy **11**(6): 467–478.

Gogé F, Gomez C, Jolivet C, Joffre R (2014) Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? Geoderma **213**(0): 1–9.

Gogé F, Joffre R, Jolivet C, Ross I, Ranjard L (2012) Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. Chemometrics and Intelligent Laboratory Systems **110**(1): 168–176.

Guerrero C, Stenberg B, Wetterlind J, Rossel RAV, Maestre FT, Mouazen AM, Zornoza R, Ruiz-Sinoga JD, Kuang B (2014) Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the spiking subset. European Journal of Soil Science **65**(2): 248–263.

Guerrero C, Zornoza R, Gómez I, Mataix-Beneyto J (2010) Spiking of NIR regional models using samples from target sites: Effect of model size on prediction accuracy. Geoderma **158**(1–2): 66–77.

Kuang, B, Mouazen AM (2013) Effect of spiking strategy and ratio on calibration of on-line visible and near infrared soil sensor for measurement in European farms. Soil and Tillage Research **128**(0): 125–136.

Naes, T, Isaksson T, Fearn T, Davies T (2002) A user friendly guide to multivariate calibration and classification. NIR publications.

Nocita M, Stevens A, Toth G, Panagos P, van Wesemael B, Montanarella L (2014) Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biology & Biochemistry **68**: 337–347.

Rossel RAV, Jeon YS, Odeh IOA, McBratney AB (2008) Using a legacy soil sample to develop a mid-IR spectral library. Soil Research **46**(1): 1–16.

Sankey JB, Brown DJ, Bernard ML, Lawrence RL (2008) Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. Geoderma **148**(2): 149–158.

Shepherd KD, Walsh MG (2002) Development of Reflectance Spectral Libraries for Characterization of Soil Properties. Soil Science Society of American Journal **66**(3): 988–998.

Shi Z, Wang Q, Peng J, Ji W, Liu H, Li X, Rossel RAV (2014) Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. Science China Earth Sciences **57**(7): 1671–1680.

Rossel RAV (2009) The Soil Spectroscopy Group and the development of a global soil spectral library. NIR news **20**(4): 14–15.

Rossel RAV, Walvoort DJJ, McBratney AB, Janik LJ, Skjemstad JO (2006) Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma **131**(1–2): 59–75.

Wetterlind J, Stenberg B (2010) Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. European Journal of Soil Science **61**(6): 823–843.

Chapter 27
Variations for the Implementation of SCORPAN's "S"

László Pásztor, Annamária Laborczi, Katalin Takács, Gábor Szatmári, Zsófia Bakacsi and József Szabó

Abstract Development of DSM can be notably attributed to frequent limitations in the availability of proper soil information; consequently, it has been typically used in cases featured by limited soil data. Since SCORPAN equation includes other or previously measured properties of soil, the usage of legacy soil data supports the applicability of DSM and improves the accuracy of DSM products as well. Nevertheless, the occurrent abundance of available soil information poses new demands on and at the same time opens new possibilities in the application of DSM methods. A great amount of soil information has been collected in Hungary in the frame of subsequent surveys and assessments. The majority of these legacy soil data were integrated in various spatial soil information systems. Our paper presents three approaches for the application of Hungary's most extended legacy soil data source in goal-oriented digital soil mapping.

Keywords Disaggregation · Homosoil · Legacy data · Spatial soil information system · Soil-related map

27.1 Introduction

Heaps of evidence furnish proof that significant amount of soil-related information has been demanded worldwide (Bullock 1999; Mermut and Eswaran 2000; Tóth et al. 2008; Sanchez et al. 2009; Baumgardner 2011; Pásztor et al. 2014). Soil maps were typically used for a long time to satisfy the needs. Presently, both the degree and the nature of the current demands have changed. Traditionally, primary soil properties and the agricultural functions of soils were focused on, and areal soil maps provided the base information. More recently rather secondary soil properties, various processes, functions and services, furthermore systems related to soils play

L. Pásztor (✉) · A. Laborczi · K. Takács · G. Szatmári · Z. Bakacsi · J. Szabó
Institute for Soil Science and Agricultural Chemistry,
Centre for Agricultural Research, Budapest, Hungary
e-mail: pasztor@rissac.hu

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_27

more important role (Omuto et al. 2013), reflecting that information related to other soil functions also becomes important (Blum 2005). However, this renewed information requirement might be heavily fulfilled with recent data collections, as compared to traditional soil mappings (Montanarella 2010). High costs of recent data collection together with the spreading of geographical information technology made spatial soil information systems and digital soil mapping the primary source of spatial soil data taking over the role of traditional soil maps. Notwithstanding, legacy soil data provide huge pool of appropriate information, which can be exploited by proper DSM methodologies.

Development of DSM can be notably attributed to frequent limitations in the availability of proper soil information (Hartemink et al. 2008). The SCORPAN equation (McBratney et al. 2003) includes other or previously measured properties of soil, the usage of legacy soil data supports the applicability of DSM and improves the accuracy of DSM products (Lagacherie 2008). Nevertheless, the availability of spatial soil information poses new demands on and opens possibilities in the application of DSM methods.

In Hungary, presently soil information demands are serviced with the available datasets either in their actual form or after certain specific and often enforced, thematic, and spatial inference (see, e.g., Dobos et al. 2010; Pásztor et al. 2013a; Sisák and Benő 2014; Szabó et al. 2007; Szatmári et al. 2013; Waltner et al. 2014). Considerable imperfection may occur in the accuracy and reliability of the map products, since there might be significant discrepancies between the available data and the expected information. The DOSoReMI.hu (Digital, Optimized, Soil Related Maps and Information in Hungary; Pásztor et al. 2015) project was started intentionally for the renewal of the national soil spatial infrastructure in Hungary.

27.2 Materials and Methods

27.2.1 Digital Kreybig Soil Information System, the Abundant Pool of Kreybig Legacy Soil Data

Digital Kreybig Soil Information System (DKSIS; Pásztor et al. 2010) integrates the full dataset collected in the frame of Hungary's most detailed nationwide soil survey led by Kreybig (1937). DKSIS consists of two types of geometrical datasets. Soil mapping units (SMU) were defined and delimited based on robustly categorized chemical and physical soil properties of the rooting zone. The mapping did not regionalize basic characteristics, and soil properties (such as pH, SOM, $CaCO_3$ content) have been available at profile level. Traditionally, the supporting SMU has been featured by the properties of its representative soil profile. However, the legacy dataset of soil profiles is much more extended, which can be efficiently used for the compilation of soil property maps by appropriate DSM methods. Nevertheless, SMUs themselves can also support the spatial inference, according to

the results presented in the next subsection. DKSIS covers the whole area of Hungary. Detailed profile descriptions are available for about 22,000 plots, which is spatially transferred for further, approximately 250,000 locations. The structure of DKSIS is presented in details by Pásztor et al. (2010).

27.2.2 Implication of Various DKSIS Components into DSM

Three basically different approaches for the application of DKSIS legacy soil data source are put forward in the followings, which are summarized in Fig. 27.1. The presented methods provide support for different challenges.

Unmapped soil properties of DKSIS, which are available only for profiles, can be spatially inferred by various DSM techniques applying spatially exhaustive, auxiliary environmental variables. They can be widely interpreted, that is spatial

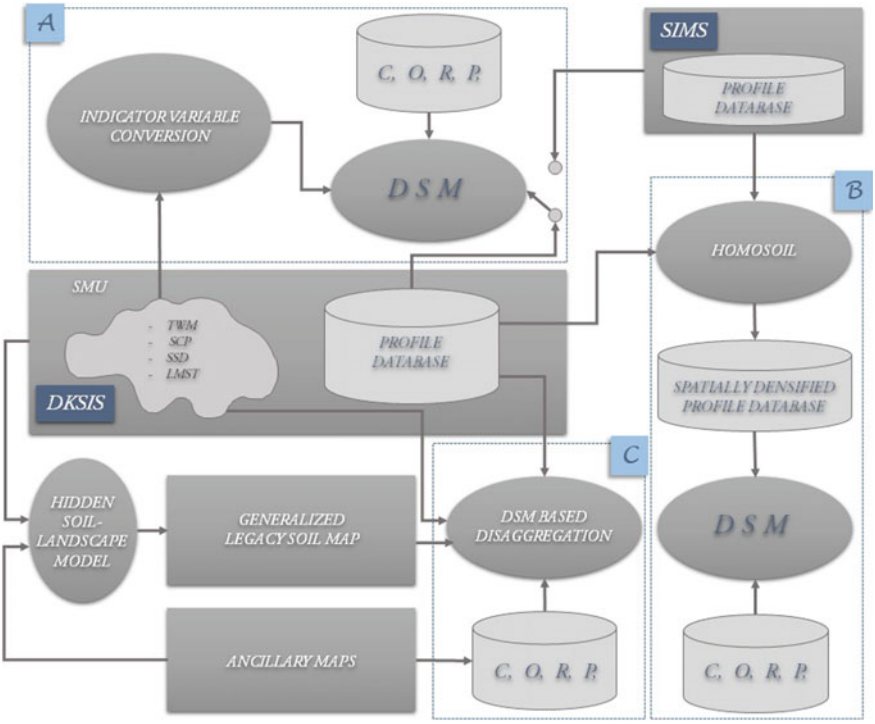

Fig. 27.1 Framework of the three presented approaches for the application of DKSIS legacy soil data source in digital soil mapping. Attributes of DKSIS SMUs: combined texture and water management categories (TWM), overall soil chemical properties (SCP), shallow soil depth (SSD), and landscape management soil type (LMST)

information on independent soil features could also be involved. In Hungary, spatially the most detailed representation of the soil cover with nationwide coverage is provided by the soil mapping units of DKSIS. As a consequence, elaboration of novel countrywide soil property maps may rely on this spatial pattern. DKSIS SMUs were introduced into regression-kriging (RK), which is widely used for the spatial inference of quantitative soil properties (e.g., Hengl et al. 2004; Dobos et al. 2010; Illés et al. 2011; Szatmári and Barta 2013) in the form of indicator variable for the compilation of soil property maps.

Numerous formerly elaborated thematic soil maps are not available in Hungary in the recently required scale. The original maps were compiled (i) in analogue environment and (ii) applying hardly identifiable soil–landscape models and unrecorded rules, so their reproducibility is problematic. Their theme, however, represents a widely used, embedded information source, which is expected to be produced in larger scales.

Various possibilities were studied for the solution of the problem. Decision trees proved to be adequate data mining technique to increase the spatial resolution of categorical soil maps disaggregating their SMUs. Classification and regression trees have numerous advantages. They can be applied for the understanding of soil–landscape models involved in existing soil maps as well as for the post-formalization of the rules applied during the survey and map compilation (Moran and Bui 2002; Scull et al. 2005; Bou Kheir et al. 2010; Giasson et al. 2011; Greve et al. 2012). The relationships identified and expressed in decision rules make the compilation of spatially refined, disaggregated maps possible using detailed, spatially exhaustive, ancillary co-variables. Among them, a special role is played by larger scale, spatial soil information.

The agro-ecological units in the AGROTOPO (1994) database, compiled as a result of a substantial scientific synthesizing work (Várallyay et al. 1985), were elaborated dominantly on the basis of mapping units originating from Kreybig soil maps, applying appropriate spatial and thematic generalization. Consequently, the Kreybig pattern contains significant and potentially utilizable information on the heterogeneity of these agro-ecological units, as do the elevation models characterizing the relief features. The availability of AGROTOPO and DKSIS spatial soil information systems and appropriate digital elevation models for the whole country has huge potential, which can be exploited in an integrated manner for the disaggregation of the thematic soil layers stored exclusively by AGROTOPO.

The third approach is presented with more details in the next session.

27.2.3 Extending the Spatial Validity of Sparse Soil Profile Data Based on Homosoil Concept

The Homosoil method, introduced recently in DSM literature by Mallavan et al. (2010), is proposed to be used when it is difficult to obtain soil information or these are

nonexistent. According to their suggestion, the assumed homology of soil-forming factors between a reference area and the region of interest can judge the extrapolation of soil-related information even from distant parts of the globe. The concept of soil homology, however, was already used during the Kreybig soil survey, based on a more justifiable manner, considering local and personally identified similarities. The resemblance in soil profiles was used for their coding within the distinct parts of a map sheet. If a soil profile with similar geographical position and very similar properties had already been described, the code of that soil profile was associated with the given soil profile and no new sampling was carried out. The utilization of this special feature of the Kreybig legacy soil information in digital soil mapping was first suggested by Pásztor et al. (2006).

The 1234 observation locations of the Hungarian Soil Information and Monitoring System (SIMS; Várallyay 2002) are characterized by detailed and up-to-date quantitative parameters, like particle-size distribution data. The sampling pattern was, however, not designed for mapping purpose. As a consequence, the sampling density does not allow the compilation of soil maps with finer scale than 1:1,500,000 (roughly 1500 m grid resolution; Hengl 2006) on the sole interpolation of SIMS data. To achieve maps with better spatial resolution, the regionalization of SIMS should be supported by spatially more detailed, auxiliary information. There are three possibilities: (1) application of widespread DSM procedures using exclusively auxiliary environmental co-variables; (2) usage of crisp legacy soil maps (as it was presented in the previous subsection); and finally (3) extension of the spatial validity of the sparsely available, "top-ranked" soil data, based on the supplementary soil information. The soil profile dataset of DKSIS proved to be adequate for this purpose.

The basic idea of our approach is the identification of homologous sampled locations within a reasonable region, where specific soil properties of a SIMS site may be predicted to be valid. The profiles of DKSIS in the geographical neighborhood of a given SIMS location were tested according to a simple homology rule. The sites within the region fulfilling the rule were identified as homosoil sites. The concept of this spatial inference method is outlined in Fig. 27.2.

For the regionalization of (hydro-)physical soil properties, the thematic rule was based on the homology of measured physical features. The DKSIS soil profile database does not contain direct information on particle-size distribution. Physical soil parameters given by horizons, however, can be used for a raw, two-layer (topsoil–subsoil) FAO texture classification (Bakacsi et al. 2012). SIMS profiles were similarly categorized. DKSIS profiles with identical two-layer classes within the same physiographical unit (Dövényi 2010) in the vicinity of the similar SIM profile were identified as SIMS homosoil sites. Physiographical units of Hungary stratify the surface of the country into 230 spatial entities delineating relatively homogeneous areas concerning terrain and main physiographical features of land. Vicinity was taken into account by area of influence provided by Thiessen polygons. As a result, "families" of DKSIS profiles form in the neighborhood of SIMS profiles delineating a wider area where its properties are considered valid (Fig. 27.3). In the present case finally, roughly 14,000 locations could be used for

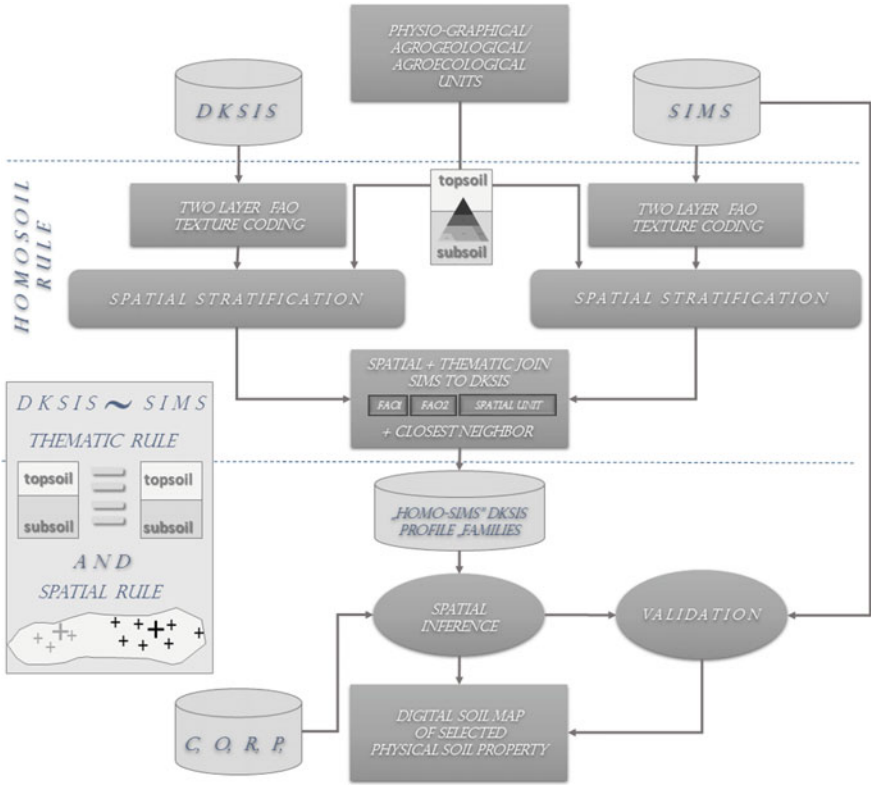

Fig. 27.2 Flowchart of the introduced homosoil method

specific spatial inferences of quantitative soil physical parameters, as opposed to the original 1234 SIMS sites. Basic physical soil properties, such as sand and clay contents, have been mapped this way with a grid resolution of 0.5 min.

27.3 Results

The results of the application of DKSIS SMUs in regression-kriging for the compilation of soil property maps are discussed in detail by Pásztor et al. (2014); here, only some relevant statements are put forward.

- Categorical data of DKSIS SMUs can be effectively applied as indicator variables.
- Usage of larger scale, spatial soil data in the course of RK-based compilation of SOM maps significantly increases the accuracy as compared to the case, when only pure environmental co-variables are applied.

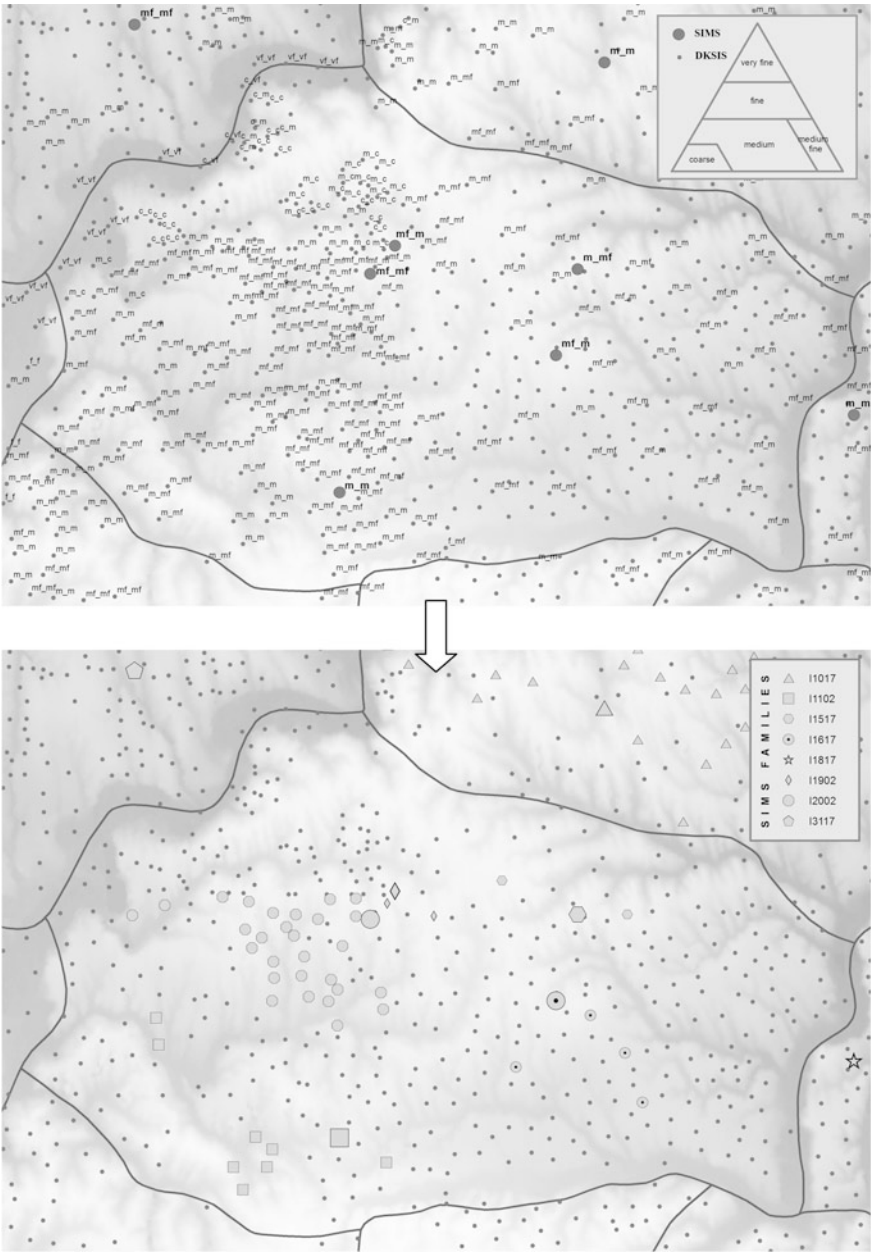

Fig. 27.3 Formation of DKSIS profile families in the neighborhood of SIMS profiles based on the applied homosoil rule. *Above* SIMS and DKSIS profiles are labeled with a two-letter code according to their topsoil–subsoil FAO texture classification (*c* coarse, *m* medium, *mf* medium fine, *f* fine, *vf* very fine) if the classification was feasible. *Down* DKSIS profile families (*small shapes*) form in the neighborhood of their parent, homologous SIMS locations (*bigger*, *identical shapes*); non-related DKSIS profiles (*gray dots*) and physiographical units (*dark polygons*) are also displayed

– Maps created using models, which also include soil co-variables, are less smoothed and show more realistic spatial structure.
– Usage of different soil co-variables (even if they originate from the same SSIS) in a model with identical ancillary variables can result in notable variances in the final map.
– Application of multiple soil layer does not increase inevitably the mapping accuracy.

The disaggregation of categorical soil maps with the aid of auxiliary spatial soil information was carried out in cases with different thematic and spatial extent. The results have been recently presented in detail by Pásztor et al. (2013b). The disaggregated version of the nationwide, soil productivity map was compiled with the aid of decision trees using the DEM100 derivatives and the SMUs of DKSIS as environmental auxiliary co-variables. The refined map has been successfully used for the delineation of Areas with Excellent Productivity in the framework of the National Regional Development Plan (http://www.terport.hu/webfm_send/4211).

The other great challenge has been the characterization of the soil cover in terms of genetic soil types at a scale of 1:50,000–1:25,000, which is required by various users for different purposes. For the fulfilling of these demands, the genetic soil-type layer of AGROTOPO was disaggregated for pilot areas based on the DKSIS SMUs and further environmental auxiliary variables using decision trees and random forests (Pásztor et al. 2015).

Extension of the spatial validity of sparse soil profile data based on the homosoil concept was validated as follows. Original SIMS points were excluded from spatial inference, which was carried out based on solely DKSIS points with transferred properties. The measured values of SIMS were then used for the validation of the specific inference results. As an example, Fig. 27.4 presents the case when sand

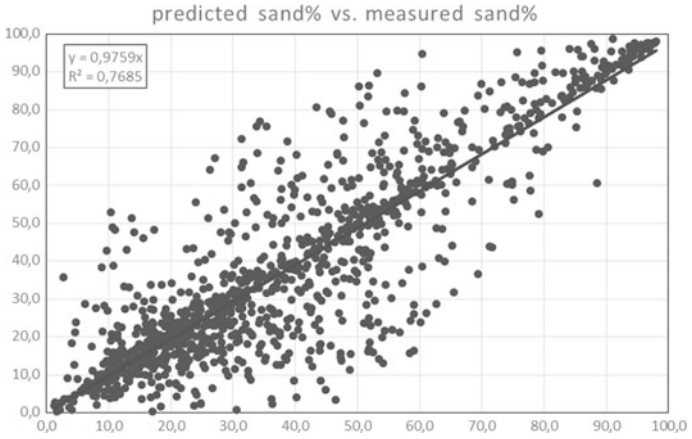

Fig. 27.4 Validation of the homosoil method. The scatter plot displays sand% of the topsoil predicted using ordinary kriging at the SIMS locations versus measured in SIMS profiles

percentage of the topsoil was interpolated using ordinary kriging for the whole area of Hungary. The calculated root-mean-squared error (RMSE) of the resulted map according to the validation with SIMS profiles is 12.5.

27.4 Discussion

The three presented variations for the implementation of SCORPAN's "S" in various DSM methods are basically differing approaches for the application of legacy soil data in the course of spatial inference. They were used and are proposed to be further applied for solving different challenges.

Application of DKSIS SMUs as indicator variables in regression-kriging was proposed to involve the expert knowledge incorporating in the delineation of soil mapping units as well as the inferred spatial stratification of soil cover into the elaboration of digital soil property maps based on the quantitative data, which is available for DKSIS soil profiles.

Disaggregation of categorical soil maps with the aid of auxiliary spatial soil information was proposed to recreate formerly elaborated thematic soil maps with higher spatial resolution. The earlier maps were compiled in analogue environment and applying subsequently hardly identifiable soil–landscape models and unrecorded rules. If their theme is expected to be produced in larger scales, they are proposed to be disaggregated using the legacy soil data, which was relied on in the original map compilation process.

Extension of the spatial validity of sparse soil profile data based on homosoil concept is proposed in the case when there are sparsely available, "top-ranked" soil data, which are originating from non-mapping purpose sampling and their spatial validity should be identified and potentially extended. The mapped soil property available for sparse profiles could be transferred to more densely sampled sites using some simple thematic and spatial rules.

27.5 Conclusions

The spatial pattern provided by DKSIS SMUs proved to be an informative co-variable in the form of indicator variables for spatial inference even in geostatistically dominated DSM methods like regression-kriging.

Further application of disaggregating methods is planned for solving similar problems, while the lessons and experience gained will also be exploited. It is also hoped to achieve progress by expanding the pool of environmental co-variables applied and by testing the performance of further classification methods.

The result of the extension of spatial validity of sparse soil profile data based on the homosoil concept is promising for the applicability of the concept; however, further refinements are already considered beyond that evidently, mapping of

various soil properties may require the application of different homology rules and can be executed using varied DSM components.

The drawn conclusions are relied on in our countrywide mapping activities in the frame of the DOSoReMI.hu project.

Acknowledgements Our work has been supported by the Hungarian National Scientific Research Foundation (OTKA, Grant No. K105167). Authors thank J. Matus for her indispensable contribution.

References

AGROTOPO (1994) AGROTOPO database of RISSAC. RISSAC HAS, Budapest, http://maps.rissac.hu/agrotopo_en

Bakacsi Zs, Kuti L, Pásztor L, Vatai J, Szabó J, Müller T (2012) Method for the compilation of a stratified and harmonized soil physical database using legacy and up-to-date data sources. Agrokémia és Talajtan 59:39–46

Baumgardner M F (2011) Soil databases. In: Handbook of Soil Sciences: Resource Management and Environmental Impacts (Eds.: Huang P M, Li Y & Sumner M E): 21–35. CRC Press, Boca Raton

Blum W E H (2005) Functions of soil for society and the environment. Reviews in Environmental Science and Biotechnology 4:75–79

Bou Kheir R, Bøcher P K, Greve M B, Greve M H (2010) The application of GIS based decision-tree models for generating the spatial distribution of hydromorphic organic landscapes in relation to digital terrain data. Hydrology and Earth System Sciences 14:847–857

Bullock P (1999) Soil resources of Europe – An overview. In: Soil Resources of Europe (Eds.: Bullock P, Jones R J A, Montanarella L) European Soil Bureau Research Report 6:15–25. Office for Official Publications of the European Communities. Luxembourg

Dobos E, Bialkó T, Michéli E, Kobza J (2010) Legacy soil data harmonization and database development. In: Digital Soil Mapping Bridging Research Environmental Application, and Operation. (Eds.: Boettinger J L Howell DW, Moore A C, Hartemink A E, Kienast-Brown S) 309–323. Springer, Heidelberg.

Dövényi Z, (ed.) (2010) Magyarország kistájainak katasztere. (In Hungarian) Budapest, MTA FKI

Giasson E, Sarmentoi E C, Weber E, Flores C A, Hasenack H (2011) Decision trees for digital soil mapping on subtropical basaltic steeplands. Scientia Agricola (Piracicaba, Braz.) 68(2): 167–174

Greve M H, Kheir R B, Greve M B, Bøcher P K (2012) Quantifying the ability of environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: The case study of Denmark. Ecological Indicators 18:1–10

Hartemink A E, McBratney A B, Mendonça-Santos M. de L (Eds.) (2008) Digital Soil Mapping with Limited Data. Springer, The Netherlands

Hengl T, Heuvelink G, Stein A (2004) A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 122(1–2):75–93

Hengl T (2006) Finding the right pixel size. Computers & Geosciences 32:1283-1298

Illés G, Kovács G, Heil B (2011) Comparing and evaluating digital soil mapping methods in a Hungarian forest reserve. Can J Soil Sci 91(4):615–626

Kreybig L (1937) The survey, analytical and mapping method of the Hungarian Royal Institute of Geology (in Hungarian and German). M Kir Földtani Intézet Évkönyve 31:147–244.

Lagacherie P (2008) Digital soil mapping: A state of art. In: Digital Soil Mapping with Limited Data. (Eds.: Hartemink A E, McBratney A B, Mendonça-Santos M de L): 3–14. Springer, The Netherlands

Mallavan B P, Minasny B, McBratney A B (2010) Homosoil, a methodology for quantitative extrapolation of soil information across the Globe. In: Digital Soil Mapping Bridging Research Environmental Application, and Operation. (Eds.: Boettinger J L Howell DW, Moore A C, Hartemink A E, Kienast-Brown S): 137–150. Springer, Heidelberg

McBratney A B, Mendonça-Santos M L, Minasny B (2003) On digital soil mapping. Geoderma 117:3–52

Mermut A R, Eswaran H (2000) Some major developments in soil science since the mid-1960s. Geoderma 100:403–426

Montanarella L (2010) Need for interpreted soil information for policy making. In: 19th World Congress of Soil Science, Soil Solutions for a Changing World, 1–6 August 2010, Brisbane, Australia.

Moran C J, Bui E N (2002) Spatial data mining for enhanced soil map modelling. International Journal of Geographic Information Science 16: 533–549

Omuto C, Nachtergaele F, Rojas R V (2013) State of the Art Report on Global and Regional Soil Information: Where are we? Where to go? Global Soil Partnership Technical Report. FAO, Rome

Pásztor L, Szabó J, Bakacsi Zs (2010) Digital processing and upgrading of legacy data collected during the 1:25 000 scale Kreybig soil survey. Acta Geodaetica et Geophysica Hungarica 45:127–136

Pásztor L, Szabó J, Bakacsi Zs, László P, Dombos M (2006) Large-scale soil maps improved by digital soil mapping and GIS-based soil status assessment. Agrokémia és Talajtan 55:79–88

Pásztor L, Szabó J, Bakacsi Zs, Laborczi A (2013a) Elaboration and applications of spatial soil information systems and digital soil mapping at the Research Institute for Soil Science and Agricultural Chemistry of the Hungarian Academy of Sciences. Geocarto International 28(1):13–27

Pásztor L, Bakacsi Zs, Laborczi A, Szabó J (2013b) Downscaling of categorical soil maps with the aid of auxiliary spatial soil information and data mining methods. (In Hungarian) Agrokémia és Talajtan 62:205–218

Pásztor L, Szabó J, Bakacsi Zs, Laborczi A, Dobos E, Illés G, Szatmári G (2014) Elaboration of novel, countrywide maps for the satisfaction of recent demands on spatial, soil related information in Hungary. In: Global Soil Map: Basis of the Global Spatial Soil Information System (Eds.: Arrouays D, McKenzie N, Hempel J, Richer de Forges A C, McBratney A): 207–212. Taylor & Francis Group, London

Pásztor L, Laborczi A, Takács K, Szatmári G, Dobos E, Illés G, Bakacsi Zs, Szabó J (2015) Compilation of novel and renewed, goal oriented, digital soil maps using geostatistical and data mining tools. Hungarian Geographical Bulletin 64(1):49-64.

Sanchez P A, et al. (2009) Digital soil map of the world. Science 325:680–681.

Scull P, Franklin J, Chadwick O A (2005) The application of classification tree analysis to soil type prediction in a desert landscape. Ecological Modeling 181:1–15

Sisák I, Benő A (2014) Probability-based harmonization of digital maps to produce conceptual soil maps. Agrokémia és Talajtan 63(1): 89-98.

Szabó J, Pásztor L, Bakacsi Zs, László P, Laborczi A. Application of the Kreybig Digital Soil Information System to solve land use problems at regional level. (In Hungarian) Agrokémia és Talajtan 56(1):5-20 (2007)

Szatmári G, Barta K (2013) Digital mapping of the organic matter content of chernozem soils on an area endangered by erosion in the Mezőföld region (In Hungarian) Agrokémia és Talajtan 62:47–60.

Szatmári G, Laborczi A, Illés G, Pásztor L (2013) Large-scale mapping of soil organic matter content by regression kriging in Zala County. (In Hungarian) Agrokémia és Talajtan 62:219–234

Tóth G, Montanarella L, Stolbovoy V, Máté F, Bódis K, Jones A, Panagos P, Van Liedekerke M (2008) Soils of the European Union. EUR 23439 EN. Office for Official Publications of the European Communities, Luxembourg

Várallyay Gy (2002) Soil survey and soil monitoring in Hungary. In: European Soil Bureau Research Report 9:139–149. ESB, Ispra

Várallyay Gy, Szűcs L, Zilahy P, Rajkai K, Murányi A (1985) Soil factors determining the agro-ecological potential of Hungary. Agrokémia és Talajtan. 34:90–94

Waltner I, Michéli E, Fuchs M, Láng V, Pásztor L, Bakacsi Zs, Laborczi A, Szabó J (2014) Digital mapping of selected WRB units based on vast and diverse legacy data. In: Global Soil Map: Basis of the Global Spatial Soil Information System (Eds.: Arrouays D, McKenzie N, Hempel J, Richer de Forges A C, McBratney A): 313–318. Taylor & Francis Group, London

Chapter 28
Monitoring Ecological Environment in Nansi Lake Area Using Remote Sensing

Ling-xia Li, Feng-mei Zhang, Chao Wang and Dong-wei Wang

Abstract Nansi Lake is the biggest lake in Huaibei area of China, which has relatively complete structure. It is an inland freshwater macrophytic lake, and large numbers of ecological protection works have been done there. In order to understand this region's ecological status, we monitored the ecosystem types of Weishan County, Nansi Lake nature reserve, and its circumjacent area in 2000, 2005, and 2010 using satellite remote sensing technology. The results showed that the ecosystem protection work of this area has achieved the desired effect. The wetland ecosystem structure has changed, and the area of lake has increased. We also found that the artificial surface area has increased. It means more ecosystem protection works still need to be done by the local government in the future.

Keywords Remote sensing · Ecological status · HJ-1 satellite · Nansi Lake

28.1 Introduction

Ecological environment holds important position in the environmental protection and natural resources development. However, due to the pressure of survival and economy and also the poor ecological protection concept, the ecological environ-

L. Li (✉) · F. Zhang
Haihe River Water Conservancy Commission, Tianjin, China
e-mail: 1814365449@qq.com

F. Zhang
e-mail: 120734876@qq.com

C. Wang
Management Bureau of Weishan Irrigation District, Liaocheng, China
e-mail: lizixuan_1981@126.com

D. Wang
China Aerospace Science and Technology Consulting Company Limited, Beijing, China
e-mail: greenhuman2000@163.com

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_28

ment has been severely damaged. The traditional way of evaluating a region's ecological change is the field investigation, but it requires a lot of human power and time and the precision is restricted. Remote sensing technology is a kind of large-scale monitoring method. It can carry out a multiple time period regional ecological type monitoring. Liu et al. (2007) used the index NDVI for the lower reaches of Tarim River region, which was calculated based on CBERS-1 satellite data in 2000, 2002, and 2004, to distinguish different ecological types. And the transformation matrix of ecological types was used to determine the transformation probability of different ecological types. Li et al. (2008) did the land-type interpretation using the Landsat TM data of the Yellow River source in 1990, 2000, and 2004 through the method of supervised classification. It correctly reflected the region's ecological changes. Zhang et al. (2013) analyzed the ecological types of the Pubugou Hydropower Station in 2003, 2007, and 2011 through man–machine combination methods and also demonstrated the region's ecological environment changes. These researches show that remote sensing technique is a rapid and effective method for multiple time period regional ecological type monitoring. In particular, all the methods are involved in the monitoring of wetland environment.

In order to know the ecological changes of Nansi Lake region from 2000 to 2010 and evaluate the effect of ecological protection, we monitored the ecological types of this region in 2000, 2005, and 2010 using remote sensing technology.

28.2 Remote Sensing Monitoring of Nansi Lake Area

28.2.1 Study Area

Nansi Lake locates in the south of Jining, Shandong province, and it is a famous shallow barrier lake. Meanwhile, it is the largest freshwater lake in Shandong area. Nansi Lake is connected by the following four lakes: Nanyang Lake, Zhaoyang Lake, Dushan Lake, and Weishan Lake. It is a multifunctional lake that integrated with flood control, water logging control, water supply, aquaculture industry, shipping, and touring. It is 126 km long from north to south, 5–25 km wide from west to east, and 311 km perimeter. The largest water storage area is 1266 km^2. The average water depth is 1.46 m. The highest water level in history is 36.48 m. The maximum storage capacity is 5.3 billion cubic meters. The total basin area is 31,700 km^2 and the basin across 32 counties (cities), 4 provinces. The water system of Nansi Lake region is very complex, which caused a lot of trouble for the ecological protection work. Figure 28.1 shows the river system of this area. The numbers of rivers that in and out Nansi Lake are 53 and 3, respectively.

In order to strengthen the ecosystem protection of Nansi Lake region, governments built a county-level nature reserve in 1982. And then, city-level and province-level nature reserves are established in 1996 and 2003 successfully. Meanwhile, the local government carries out also publicity and education work actively. The protection has been strengthened through ecological conservation projects.

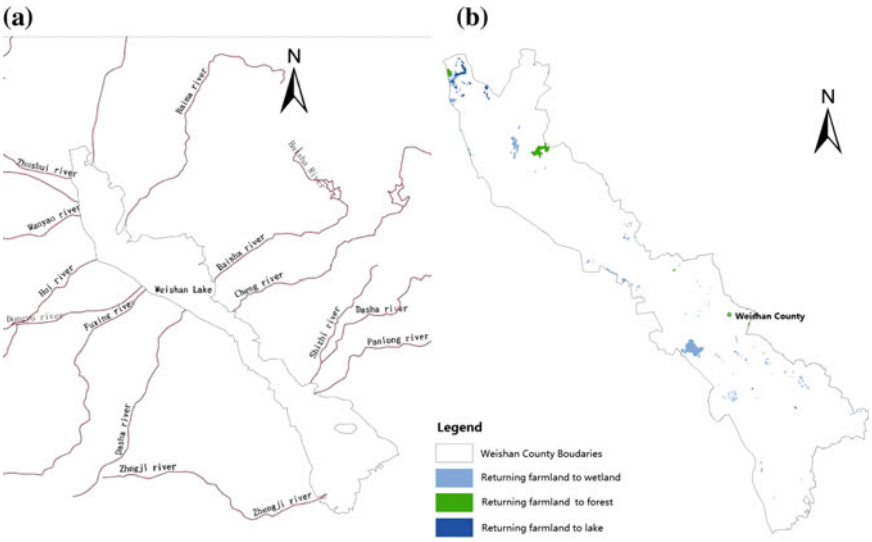

Fig. 28.1 **a** The river system of Nansi Lake Region. **b** The spatial distribution of returning farmlands to wetlands, lakes, and forest from 2000 to 2010 in Weishan County (the wetland in this figure includes the entire wetland types except lakes)

28.2.2 Data and Methods

In this paper, we used remote sensing method to monitor the ecological change of Nansi Lake region. The HJ-1A satellite images acquired in 2000, 2005, and 2010 were used. These images that we downloaded from China Centre of Resource Satellite and Application (CCRSA) are shown in Fig. 28.2. HJ-1A satellite equipped with a CCD sensor of four bands. We chose B2 (0.43–0.52 μm), B3 (0.52–0.60 μm), and B4 (0.76–0.9 μm) bands to obtain the best band combination (Wang et al. 2011). The resolution of the CCD sensor is 30 m, which is sufficient for ecological monitoring.

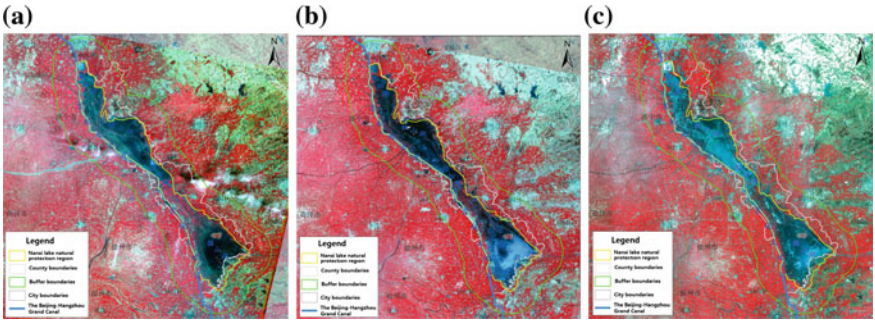

Fig. 28.2 The remote sensing images of Nansi Lake region in **a** 2000, **b** 2005, and **c** 2010

In order to achieve our goals of remote sensing monitoring, we chose Mahalanobis distance classification method to classify the images. Before performing Mahalanobis distance classification, we conducted a field investigation to choose the most suitable training samples. The area corresponding to training samples was measured by investigators with handheld GPS, and then, we generated vector files including all kinds of ecological types. After imported the vector files into the ENVI, we generated the training samples which Mahalanobis distance classification method needed by matching the vector files with HJ-1A images. The classification system was chose from Ouyang's paper (2015) for the system based on medium-resolution remote sensing data. This system included 9 first classes, 21 second classes, and 46 third classes. It was mainly based on the similarity of ecosystem characteristics and also considering the climate and topography factors. Because of this, the classification system was adaptive to the ecological classification in China area. In this study, we first distinguished between forest, grassland, farmland, and wetland through the first-level classification. Then, the wetland was further classified into marsh, lake, reservoirs/swag, river, and canal/water channel.

28.3 Results

The situations of land utilization in natural protection area of Nansi Lake and the buffer area about 10 km around are analyzed based on statistical analysis results. The final results are shown in the following.

28.3.1 The Ecological Status of Nansi Lake Nature Reserve

The total area of Nansi Lake Nature Reserve is 1275.47 km^2. The mainly ecosystem type is wetland ecosystem which accounts for more than 80 %. Most of the wetland types are lakes, reservoirs, and marshes. According to statistical results in 2010, the above wetland types are accounted for 52.21, 31.98, and 14.86 %, respectively. During the years from 2000 to 2010, the proportion of lakes is continuously increasing and the proportion of marshes is continuously depleting. The details are shown in Table 28.1 and Fig. 28.3.

Table 28.1 The statistic table of wetland ecosystem area in Nansi Lake Nature Reserve in 2000, 2005, and 2010

Ecosystem type	Area (km^2)		
	2000	2005	2010
Marsh	206.35	169.69	157.06
Lake	521.50	547.67	551.93
Reservoirs/swag	314.21	337.52	338.11
River	8.33	8.58	8.75
Canal/water channel	0.77	0.96	1.38
Total area	1051.16	1064.42	1057.22

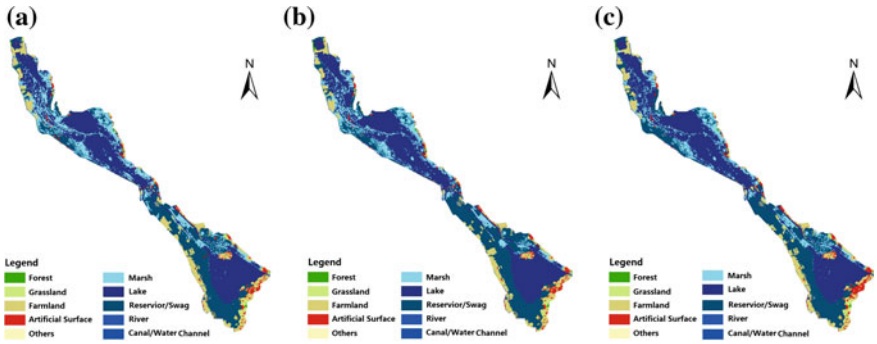

Fig. 28.3 The distribution of ecosystem of Nansi Lake region in **a** 2000, **b** 2005, and **c** 2010

28.3.2 The Ecological Status of Weishan County

The total area of Weishan County is 1766.06 km^2. Its main ecosystem type is wetland ecosystem, and the whole Nansi Lake is in it. The wetland, farmland, artificial surface, and forest account for 62.97, 25.75, 7.79, and 2.57 %, respectively in 2010. The grassland and other ecosystem types are less than 1 % in total. The details are shown in Table 28.2 and Fig. 28.4.

During the years from 2000 to 2010, the farmland area significantly decreased and the area of wetland and forest obviously increased. A total reduction of farmland area is 38.42 km^2. It is mainly converted into artificial surface, wetland, and forest. The conversion areas are 29.45, 14.44, and 4.15 km^2, respectively. Meanwhile, the area of wetland that transformed into farmland is 9.37 km^2.

28.3.3 The Ecological Status in Nansi Lake and Circumjacent Areas

The total area of Nansi Lake and buffer areas is 4311.60 km^2. The main ecosystem types are wetland and farmland, which account for 28 and 48 % of the total area,

Table 28.2 The statistic table of ecosystem area in Weishan County in 2000, 2005 and 2010

Ecosystem type	Area (km^2)		
	2000	2005	2010
Forest	41.30	45.38	45.30
Grassland	14.62	13.24	14.26
Wetland	1105.33	1119.77	1112.16
Farmland	493.26	459.18	454.84
Artificial surface	109.30	126.27	137.53
Others	2.25	2.23	1.96
Total area	1766.06	1766.06	1766.06

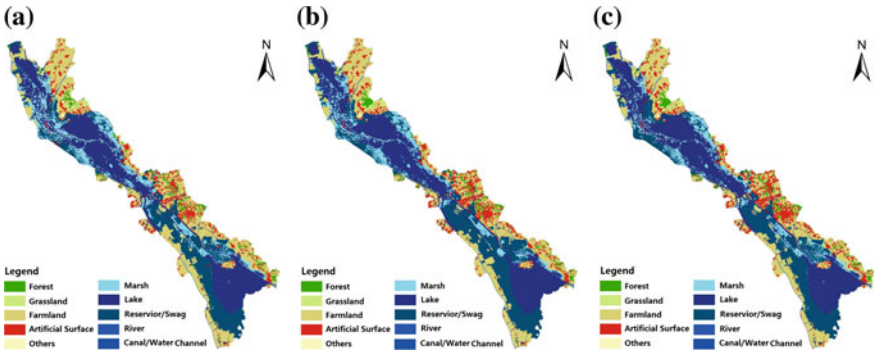

Fig. 28.4 The distribution of ecosystem of Weishan County in **a** 2000, **b** 2005, and **c** 2010

Table 28.3 The statistic table of ecosystem area in Nansi Lake and circumjacent areas in 2000, 2005 and 2010

Ecosystem type	Area (km^2)		
	2000	2005	2010
Forest	199.76	204.01	206.97
Grassland	109.94	102.77	107.59
Wetland	1207.48	1225.62	1222.28
Farmland	2215.30	2141.27	2086.44
Artificial surface	575.10	633.93	684.60
Others	4.03	3.99	3.72
Total area	4311.60	4311.60	4311.60

respectively. Table 28.3 describes the area of different ecosystem types in the year of 2000, 2005, and 2010. During the years 2000–2010, the wetland area presents a minor increase and farmland area decreased. The area of artificial surface increases significantly, and the proportion is 18.9 %. The details are shown in Fig. 28.5.

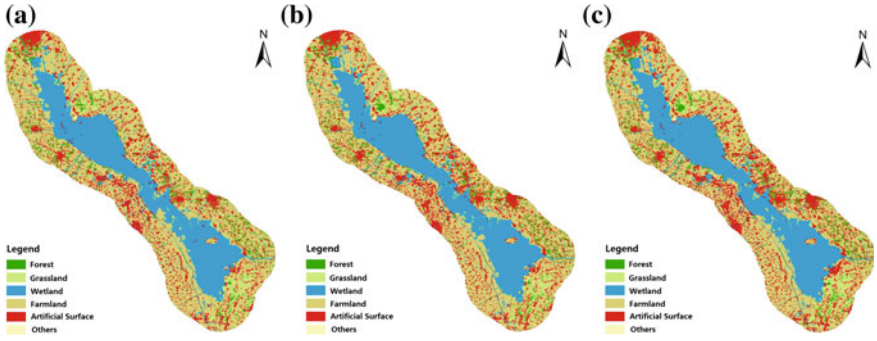

Fig. 28.5 The distribution of ecosystem of Nansi Lake and the surrounding area in **a** 2000, **b** 2005, and **c** 2010

28.4 Discussion

The monitoring results showed that wetland, farmland, artificial surface, and forest are four main ecological types of Weishan County, which account for 99.08 % area of the county in 2010. The most important ecological type is wetland, which accounts for 62.97 %. By comparing the monitoring results in 2000, 2005, and 2010, we found that wetland area of Nansi Lake Nature Reserve had increased slightly, but the subtypes of wetland had changed. For example, the area of marshes has decreased from 206.35 km^2 in 2000 to 157.06 km^2 in 2010, and the area of lakes from 521.50 km^2 to 551.93 km^2. We also derived the conclusion from our study that the ecosystem protection work of Weishan County has achieved the expected results in the 10 years during 2000–2010. The areas of wetlands and forests have increased 38.42 km^2. Although the local government had done large amounts of work in ecosystem protection, we found artificial surface area has increased by 18.9 % from 2000 to 2010, which demonstrates that human activity is still the main threat to local ecosystem. The rapidly increased artificial surface also shows that much more work is still needed by the local government in the future.

References

Li FX, Fu Y, Li LX, Xiao JS. (2008). Remote sensing monitor and driving factors of ecological environment change in the source region of the Yellow river. Ecology and Environment, 17(6): 2297- 2303.

Liu H, Chen YN, Yang XM. (2007). Monitor of ecological response along lower reaches of Tarim River based on remote sensing. Arid Land Geography, 30(2): 203-208.

Ouyang ZY, Zhang Y, Wu BF, Li XS, Xu WH, Xiao Y, Zhang H. (2015) An ecosystem classification system based on remote sensor information in China. Acta Ecologica Sinica, 35: 219–225.

Wang JW, Liu G, Ma HT, Xu HJ. (2011) Optional Bands Combination of HJ-1A/B Satellite to Macroscopic Monitoring. China Science and Technology Information, 16: 21–23.

Zhang SS, Qin R, Jiang D, Li J, Luo YY (2013) Monitoring and analyzing eco-environmental impacts of Pubugou Hydropower station project with HJ-1 satellite data. Journal of Gansu Science, 25(1):68-72

Chapter 29
Extraction and Integration of Different Soil Nutrient Grading Systems for Soil Nutrient Mapping

Shuxia Wu, Weili Zhang, Aiguo Xu and Qiuliang Lei

Abstract These works present a model to integrate and harmonize different nutrient grading indexes originating from various existing soil polygon maps. The soil nutrient grading indexes might be different from one national soil survey to another and even be different for the different counties in the same national soil survey in China. Soil nutrients mapping in large regions, such as national or provincial regions, had to be done after the integration of those grading indexes. The hardcopy of the soil nutrient maps for most of the counties of China was collected and vectorized. These maps were mostly scaled at 1:50,000–1:500,000 and could be used as the input data for the integration of the grading index. Next, a model named Soil Nutrient grading system Integration Model (SNI-Model) was established using ARCGIS10.0 and was written in C#. The SNI-Model did not change or revise the properties of the spatial soil nutrient maps and saved the original grading index for every soil polygons. Also, it was designed in a two color systems for more easy reading. SNI-Model consisted of five modules and could be easily used to intelligent integrate the soil grading indexes for soil nutrient mapping. The SNI-Model is a general model and may also be applicable to the environment, ecology, and other research areas to resolve the similar problems.

Keywords Soil nutrient · Model · Data integration · Mapping · Grading system

29.1 Introduction

From the 2th national soil survey of China, a series of soil nutrient maps (1:50–100 K) on county level were completed, including soil organic matter, total nitrogen, available nitrogen, available phosphorus, available potassium content, pH value, and several soil microelements. These maps could be very useful to study the

S. Wu (✉) · W. Zhang · A. Xu · Q. Lei
Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Science, Beijing 100081, China
e-mail: wushuxia@caas.cn

G.-L. Zhang et al. (eds.), *Digital Soil Mapping Across Paradigms, Scales and Boundaries*, Springer Environmental Science and Engineering,
DOI 10.1007/978-981-10-0415-5_29

temporal change of soil quality characteristics (Xi et al. 1994; Zhao 1996; Sun et al. 2003; Nyssen et al. 2008; Yuan et al. 2013). The nutrient contents differed a lot among counties because the climate and topography varied in different regions of China. That is, the maps from the second soil survey executed in different counties might have used different grading systems (National Soil Survey Office 1992). Also, there were differences among each grade's minimum and maximum value and ranges. For example, soil total nitrogen content in some counties used five classification systems and some used a six-grading system (Table 29.1). In some counties, multi-level compound systems were used, there could be about 9 classification grades or more. Concluding, (1) the minimum and maximum value and range between the different grading systems could be different, (2) one system may contain another, or (3) systems could be intersecting each other (Table 29.2).

Therefore, the soil nutrient mapping in large regions, such as national or provincial regions, had to be done after the integration of those grading indexes. When doing the integrating, feature attributes should not be changed and similar color code should be used for those similar grading indexes, to ensure full consistency mapping in large regions. As the existing GIS mapping software package failed to provide the required functionality to do this analysis, the purpose of this study is to build a computer model which should be intelligent, process-oriented, and easy for human–computer interaction to extract and integrate the different grading indexes in soil nutrient mapping.

Table 29.1 Examples for the grading system of soil total nitrogen content (%)

Grade	National standard	Grade	Jiangsu, Wujin	Grade	Jiangsu, Yuhuatai
I	>0.2	I	>0.2	I	>0.15
II	0.15–0.2	II	0.15–0.2	II	0.15–0.125
III	0.1–0.15	III	0.125–0.15	III	0.125–0.1
IV	0.075–0.1	IV	0.1–0.125	IV	0.1–0.075
V	0.05–0.075	V	<0.1	V	<0.075
VI	<0.05				

Table 29.2 Examples for compound grading systems of soil total nitrogen (%)

Grade	National standard	Grade	Shanxi, Datong	Grade	Henan, Mengjin
I	>0.2	I	>0.2	I	>0.2
II	0.15–0.2	II	0.15–0.20	II	0.15–0.20
III	0.1–0.15	III	0.1–0.15	III1	0.125–0.15
				III2	0.1–0.125
IV	0.075–0.1	IV1	0.085–0.1	IV1	0.085–0.1
		IV2	0.075–0.085	IV2	0.075–0.085
V	0.05–0.075	V1	0.06–0.075	V1	0.06–0.075
		V2	0.05–0.06	V2	0.05–0.06
VI	<0.05	VI	<0.05	VI	<0.05

29.2 Model Construction

29.2.1 Model Design Principles

The Soil Nutrient grading system Integration Model (SNI-Model) was constructed to extract every grading index from the different nutrient maps, compare and integrate the different grading systems to one legend, and allocate a color code to each grading index in a certain large area for mapping. The integrated model was designed by following the three main principles:

1. The new mapping after integration should keep all of the properties of the spatial nutrient features of the original maps, and these properties could not be modified or lost;
2. Legend expression should be normalized and should not cause confusion for different nutrient classification;
3. Two different levels of boundaries were considered when allocating color codes: One was considering the whole region such as integrating the different county maps to a national map or one province map, and the other one was designed as national standard subdivision maps to ensure the perfection of color under different mapping purposes.

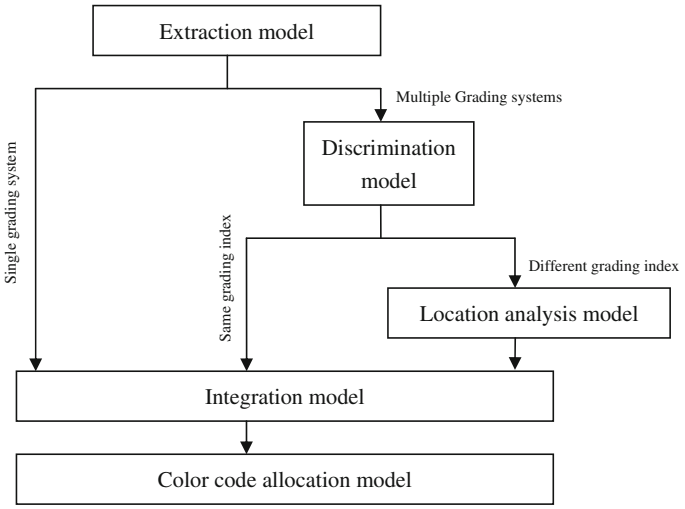

Fig. 29.1 Sub-models and their relationships within the SNI-Model

29.2.2 Construction of the SNI-Model

The SNI-Model consists of five sub-models to achieve the expression of spatial mapping. The names of sub-models and their relationships were shown in Fig. 29.1, namely (a) extraction model, (b) discrimination model, (c) location analysis model, (d) integration model, and (e) color code allocation model.

29.2.3 Function of Sub-models

The "extraction model" was designed to extract every grading system and each grading index to a database from all of the original county nutrient maps, by setting the control table and using a human–computer interaction (HCI) technique (Clouard et al. 2011). The legend of each county's nutrient map was stored as a grading system in the database. Next, for each grading system, the grading index was extracted and the maximum, minimum, mean, and range were calculated. Each of the grading indexes was assigned a sequence number. A field named GrName was added to the attribute database, which was used to express the grading index. If there was only one grading system, then the model run into the integration model; otherwise, it would run into the "discrimination model."

The "discrimination model" was applied to compare each two grading indexes. A table was created to store the comparison results, and the GrName which was created in the "extraction model" was used as the standard to judge whether the two grading indexes were the same or not. If the two GrNames were the same, then the value was stored to the database table. For different GrNames, the values were sorted by the minimum and mean of the grade index in ascending order and stored to the database table. The number of the sorted results could be used as the location of the grading index in the next model.

The "location analysis model" was used to express the location of each grading index in the legend of whole map or subdivision maps by setting the tolerance value through the HCI function on the basis of the results of pairwise comparison in "discrimination model." In this model, the relationship between the grading indexes from all the maps was revised and re-assigned to a new location ID to express the location of each grading index and stored to a location result database. The "integration model" was created to summarize all the results of the "extraction," "discrimination," and "location analysis model" and stored the results to a new result database. Here, each grading index was recoded with a position number Nr_P, and a standardized expression of each different grading index was generated in this table for mapping. In this model, the number of decimal was set for the standardized expression field.

The "color code allocation model" could merge the legends with same grading systems to one and arrange the different grading systems in parallel from small to large according to the minimum and mean of the grading index. Two color code assignment methods were used for perfect visual comfort purpose: One was

designed for the whole region including all the sample counties, and the other one was for single subdivision map when the whole region was divided to various subdivisions. For perfect visual comfort purpose, different equations were used to assign the color codes (Eqs. 29.1a, b). The color IDs were saved in a parameter database, the colorID was used by the model for the grading indexes, and certain color codes were allocated. The final database contained all of the attributes of the original nutrient maps and was used as the new attribute table, and the field of color code was used for mapping.

$$\text{CorNr w} = \text{ROUND}\left[1 + \left(\sum \text{CorID} \Big/ \sum \text{GRw}\right) \times (\text{Nr\_Pw} - 1)\right] \quad (29.1a)$$

$$\text{CorNr s} = \text{ROUND}\left[1 + \left(\sum \text{CorID} \Big/ \sum \text{GRs}\right) \times (\text{Nr\_Ps} - 1)\right] \quad (29.1b)$$

CorNr w or CorNr s	The color code which the grading index was assigned for perfect vision view under the whole region mapping or single subdivision mapping;
∑CorID	The total number of the color IDs in the color database;
∑GRw or ∑GRs	The total number of the grading indexes in the whole region or the single subdivision mapping;
Nr_Pw or Nr_Ps	Position number of the grading indexes in the whole region or the single subdivision mapping;
ROUND	Make the value to integer by rounding.

29.2.4 Model Development

The SNI-Model used (1) C# as the programming language in the NET Framework 4 Extended development environment, (2) functions of mapping software (ArcGIS), (3) database software (Access), and (4) the interface control package (DotNet Bar) completed the development of the system and the functional modules, according to the model design.

29.3 Example of SNI-Model Application

Soil organic matter content, an important nutrient in soil survey, is an important index of soil fertility and is useful to investigate the soil formation, distribution, and classification. In this example, 17 organic matter maps (1:500 K) on province lever were used as the input database for SNI-Model to extract, integrate the grading systems, and allocate color ID to each grading index, on basis of national standard

Table 29.3 Extraction, integration, and expression of organic matter content in the studied 17 provinces

Nr_Pw	Grade index (GRw)	CorNrw
1	<0.6	1
2	0.6–0.8	2
3	<1.0	3
4	0.6–1.0	5
5	0.8–1.0	6
6	1.0–1.2	7
7	1.0–1.5	8
8	1.0–2.0	9
9	1.2–1.5	10
10	1.5–2.0	12
11	>2.0	13
12	2.0–2.5	14
13	2.0–3.0	15
14	2.5–3.0	16
15	3.0–3.5	17
16	3.0–4.0	19
17	3.5–4.0	20
18	>4.0	21
19	4.0–4.5	22
20	4.0–7.0	23
21	4.5–5.0	24
22	>5.0	26
23	7.0–10.0	27
24	>10.0	28

subdivision maps (1:1000 K). The 17 provinces were Tianjin, Hebei, Shanxi, Inner Mongolia, Heilongjiang, Shanghai, Zhejiang, Fujian, Jiangxi, Shandong, Henan, Hubei, Guangdong, Guangxi, Sichuan, Guizhou, and Shaanxi. Forty national standard subdivisions (1:1000 K) were covered by these 17 provinces. A province covered 1–15 subdivisions (1:1000 K) and a subdivision covered 1–6 provinces. The color table used in SNI-Model consisted of color system, color code number, and RGB value of each color code. There were 28 color codes in the color system which was selected to express the grading index in this example in SNI-Model.

SNI-Model was run in stepwise, and then every grading system and grading index of all the 17 province maps were extracted and integrated in two levels: national level and subdivision level. And two series color codes were generated for each subdivision. In Table 29.3, the extraction and integration results from all the studied 17 provinces were showed. There were 24 grading index in total, and each one was allocated a place number and color code number. In Tables 29.4 and 29.5, the extraction and integration results for subdivision G48 and J49 were listed, respectively, including the grading index extracted and the two levels of color code for different visual purposes.

Table 29.4 Extraction, integration, and expression of organic matter content in G48 and Sichuan, Guangxi, and Guizhou provinces

Nr_Ps	Grade index (GRs)	CorNrs	CorNrw
1	<0.6	1	1
2	0.6–1.0	6	5
3	1.0–2.0	10	9
4	2.0–3.0	15	15
5	3.0–4.0	20	19
6	>4	24	21

Table 29.5 Extraction, integration, and expression of organic matter content in J49 and 5 provinces

Nr_Ps	Grade index (GRs)	Inner Mongolia	Shaanxi	Hebei, Shanxi, Henan	CorNrs	CorNrw
1	<0.6	<0.6	<0.6	<0.6	1	1
2	0.6–0.8		0.6–0.8		3	2
3	0.6–1.0	0.6–1.0		0.6–1.0	5	5
4	0.8–1.0		0.8–1.0		7	6
5	1.0–1.2		1.0–1.2		9	7
6	1.0–2.0	1.0–2.0		1.0–2.0	11	9
7	1.2–1.5		1.2–1.5		13	10
8	1.5–2.0		1.5–2.0		15	12
9	2.0–3.0	2.0–3.0	2.0–3.0	2.0–3.0	17	15
10	3.0–4.0	3.0–4.0	3.0–4.0	3.0–4.0	19	19
11	>4.0		>4.0	>4.0	21	21
12	4.0–7.0	4.0–7.0			23	23
13	7.0–10.0	7.0–10.0			25	27
14	>10.0	>10.0			27	28

29.4 Discussion

In this study, the SNI-Model was established, and the different soil nutrient grading systems were extracted, integrated, and allocated color codes through HCI on the basis of different sub-models, following that attributes of original maps should not be changed and similar color code should be used for those similar grading indexes. The characteristics of the integrating technique could be the following: (a) HCI was used when key issues had to be judged, for example, on whether the two grading system indexes were the same or not, and to realize the massive data integration of different grading indexes in a short time; (b) every grading system or grading index of the soil nutrient maps could be extracted and integrated, and each grading index will be assigned a color code from the model; (c) two color code assignment methods were used for perfect visual comfort purpose; and (d) SNI-Model is a general model and

could be also applicable to the environment, ecology, and other research areas and to resolve the similar problems could also be available to different scale's mapping.

References

Zhao QG (1996) Modern soil science and sustainable development of agriculture. ACTA PEDOLOGICA SINICA 33(1):2-12

Xi Chengfan, Zhang Siyan (1994) Brief introduction on achievements in national soil survey project since 1979. ACTA PEDOLOGICA SINICA 31(3):330-335

Nyssen J, Temesgen H, Lemenih M, Zenebe A, Haregeweyn N (2008) Haile M. Spatial and temporal variation of soil organic carbon stocks in a lake retreat area of the Ethiopian Rift Valley. Geoderma 146(1-2):261–268

Sun B, Zhou SL, Zhao QG (2003) Evaluation of spatial and temporal changes of soil quality based on geostatistical analysis in the hill region of subtropical China. Geoderma. (1 - 2): 85 – 99

Yuan ZX, Wang Y (2013) Study on space-time change of arable layer soil nutrients in liangzhou district based on GIS. Gansu Agricultural Science and Technology, (4):28-30

National Soil Survey Office (1992) "Chinese soil survey technology." Agriculture Press, Beijing

Régis Clouard, Arnaud Renouf, Marinette Revenu (2011) Human-computer interaction for the generation of image processing applicationsOriginal Research Article, International Journal of Human-Computer Studies, 69(4): 201-219