PROGRESS IN SOIL SCIENCE

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

Springer

Progress in Soil Science

Progress in Soil Science

Aims and Scope

Progress in Soil Science series aims to publish books that contain novel approaches in soil science in its broadest sense – books should focus on true progress in a particular area of the soil science discipline. The scope of the series is to publish books that enhance the understanding of the functioning and diversity of soils in all parts of the globe. The series includes multidisciplinary approaches to soil studies and welcomes contributions of all soil science subdisciplines such as: soil genesis, geography and classification, soil chemistry, soil physics, soil biology, soil mineralogy, soil fertility and plant nutrition, soil and water conservation, pedometrics, digital soil mapping, proximal soil sensing, soils and land use change, global soil change, natural resources and the environment.

For further volumes:
http://www.springer.com/series/8746

Janis L. Boettinger · David W. Howell ·
Amanda C. Moore · Alfred E. Hartemink ·
Suzann Kienast-Brown
Editors

Digital Soil Mapping

Bridging Research, Environmental Application, and Operation

Springer

Editors
Dr. Janis L. Boettinger
Utah State University
Dept. Plants, Soils, &
Climate
4820 Old Main Hill
Logan, UT 84322-4820
USA
janis.boettinger@usu.edu

Amanda C. Moore
U.S. Department of Agriculture
Natural Resources Conservation
Service
339 Busch's Frontage Road
Suite 301
Annapolis, MD 21409
USA
amanda.moore@md.usda.gov

Suzann Kienast-Brown
U.S. Department of Agriculture
Natural Resources Conservation
Service
Utah State University
Dept. Plants, Soils, &
Climate
4820 Old Main Hill
Logan, UT 84322-4820
USA
suzann.kienast@ut.usda.gov

David W. Howell
U.S. Department of Agriculture
Natural Resources Conservation
Service (Retired)
P.O. Box 709
Arcata, California 95518
USA
david@earthmapphoto.com

Prof. Dr. Alfred E. Hartemink
International Soil Reference
Information Centre (ISRIC)
Wageningen
Netherlands
Alfred.Hartemink@wur.nl

ISBN 978-90-481-8862-8 e-ISBN 978-90-481-8863-5
DOI 10.1007/978-90-481-8863-5
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2010925363

Cover image: Digital soil map illustrating the distribution of 24 soil classes in the Big Wash watershed in the Great Basin region of southwestern Utah, USA (adapted from Figure 15.2b, Chapter 15 from this book).

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

Digital Soil Mapping is by now a well-established branch of soil science, with regular meetings and a very active working group of the International Union of Soil Sciences (IUSS). Meetings of the group bring together scientists dealing with digital soil mapping in the broadest sense. These meetings allow for exchange of information among scientists on their research topics and are excellent opportunities for assessing the status of this relatively young area of research in soil science.

The chapters in this book were selected from papers presented at the 3rd Global Workshop on Digital Soil Mapping (DSM 2008) that was held in Logan (Utah, USA). The theme of the workshop was *Digital Soil Mapping: Bridging Research, Production, and Environmental Application*.

There is great interest in transferring the scientific achievements of the past years of digital soil mapping into operational data and information systems responding to the increasing demands for high quality soil data and information. The past collection of soil data was largely driven by a mono-functional view of soil as the basis for agricultural production. Under the leadership of the Food and Agriculture Organization of the United Nations (FAO) substantial progress has been made in collecting soils data and information in all continents, particularly in developing countries. Standardized systems for soil classification and soil profile description have facilitated the interoperability of information systems across national borders, paving the way for the creation of digital soil databases at global and continental scales based on advanced GIS technologies. Good examples of such systems are SOTER (SOil and TERrain Digital Database) coordinated by FAO and the European Soil Information System (EUSIS) of the European Union (EU).

The relatively recent recognition of the multi-functionality of soils, including important ecosystem services and socio-economic benefits, has emphasized the inadequacy of existing soil information systems worldwide. Traditional soil survey, based on soil profile descriptions, soil classification and extrapolation of data on a soil-landscape model developed by expert judgment, cannot respond to the new requirements coming from user communities other than agriculturalists.

New soil data and information are needed to address the emerging concerns about the functioning of soils systems in the delivery of services required by modern societies. Data on soil contamination, soil biota and their diversity, soil stability (landslides), soil hydraulic functions, soil carbon pools, soil erosion, salinization,

etc. are needed by policymakers dealing with the urgent priorities related to climate change, natural and man-made hazard prevention, food and feed health as well as food security, and bio-energy production.

The emergence of new legal frameworks for soil protection at national, regional, and global levels has made traditional soil survey techniques incapable of responding to stringent legal requirements. The delineation of areas with different soil properties needs to have a solid scientific and geostatistical basis so that it can be used by legislators. Priority areas for soil protection cannot solely be delineated on the basis of expert judgements, but must be based on quantitative data that can withstand legal challenges in court. For example, the EU Soil Thematic Strategy requires the delineation of priority areas for the various threats to soil functions. These delineations have legal and financial implications that affect landowners. Therefore, the definition of these areas requires the highest quality of soil data as well as solid scientific methods for producing soil data. Digital soil mapping will thus play a key role in implementing this legislation in the European Union. Similarly at the global scale, soils play a critical role in the implementation processes of Multilateral Environmental Agreements (MEAs). The United Nations Framework Convention on Climate Change (UNFCCC), the Convention on Biodiversity (CBD), and the United Nations Convention to Combat Desertification (UNCCD) increasingly recognise the crucial role of soils. Updated and accurate global soil data and information are urgently required for these emerging needs, such as information on soil organic carbon pools and their dynamics over time. Also, the specific initiative within the Group of Earth Observation (GEO) to establish a Global Soil Information System (GLOSIS) as part of the Global Earth Observation System of Systems (GEOSS) is a response to these new requirements from policymakers.

The digital soil mapping community has taken up the challenge to foster the development of a new generation of digital soil information at local, national and global scales. The establishment of the *GlobalSoilMap.net* consortium, pooling together the major players in digital soil mapping in the world, has initiated a process that will deliver a new digital soil map of the world at fine resolution. The first node getting active in the *GlobalSoilMap.net* project is the Africa Soil Information Service (AfSIS), coordinated by the Tropical Soil Biology and Fertility Institute of CIAT (CIAT-TSBF) and financed by the Bill and Melinda Gates Foundation. The experience of transferring digital soil mapping technologies into practice on such a scale is useful in making digital soil mapping operational at continental and global scales.

The chapters in this book provide a very useful and comprehensive overview of the status of digital soil mapping and are a further step in developing this branch of soil science. I strongly recommend their consultation and reading.

European Commission, Ispra, Italy Luca Montanarella

Preface

This book contains papers presented at the 3rd Global Workshop on Digital Soil Mapping held in Logan, Utah, USA, 30 September–3 October 2008. The workshop was organized under the auspices of the International Union of Soil Sciences Working Group on Digital Soil Mapping, and was hosted by Utah State University. The organizing committee was chaired by Dr. Janis Boettinger, professor of Pedology in Utah State University's Plants, Soils, and Climate Department. Financial and in-kind support for this workshop was provided by Utah State University and the US Department of Agriculture Natural Resources Conservation Service. Approximately 100 participants from 20 countries presented and discussed nearly 70 papers during the four-day session, demonstrating the global engagement in digital soil mapping.

The theme of this workshop was *Digital Soil Mapping: Bridging Research, Production, and Environmental Application*. Advances in digital soil mapping technology and methods occur at a rapid pace, facilitating the development of digital soil information with increasing precision for many areas around the world. In many cases we are fortunate to have a wealth of legacy soil data to work with. Legacy soil data can be used to improve digital soil mapping models and, in turn, digital soil mapping models can be used to help modernize and harmonize legacy soil data. Digital soil mapping has evolved to the point where it is has entered the operational realm, as a tool for improving accuracy, consistency, and efficiency of production soil mapping. At the same time, there is still a need for innovative soil information products to support environmental applications. Credible and innovative research is the basis for the development of digital soil mapping and soil assessment protocols. Development of practical soil mapping and environmental applications drives the need for continued progress in the field of digital soil mapping. With this workshop, we hoped to recognize these distinct foci within the realm of digital soil mapping.

We have selected 33 papers from the Logan 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) Research; (III) Environmental Application and Assessment; and (IV) Making Digital Soil Mapping Operational. Within the research section, papers are grouped by three key topics: (A) Environmental Covariates and Soil Sampling; (B) Soil Sensors and Remote Sensing; (C) and Soil Inference Systems. Mapping and modeling of organic carbon is the primary focus

of the section on environmental application and assessment and of major interest to the global soil science community as well as policy makers. Digital soil mapping in a production setting is presented with case studies from New Zealand, the European Union, Canada, the United States, and the *GlobalSoilMap.net* project.

This book complements and extends the ideas presented in *Digital Soil Mapping – An Introductory Perspective*, edited by Lagacherie, McBratney, and Voltz, (2007) and *Digital Soil Mapping with Limited Data*, edited by Hartemink, McBratney, and Medonça-Santos (2008). We hope that this book will inspire digital soil mapping researchers and practitioners at universities, agencies, and other organizations in their efforts to create and utilize soil information in a range of global issues like climate change, food production, energy, and water security. We are excited to see where global advancements in digital soil mapping research will take us in the project.

Logan, UT — J.L. Boettinger
Arcata, CA — D.W. Howell
Annapolis, MD — A.C. Moore
Wageningen, The Netherlands — A.E. Hartemink
Logan, UT — S. Kienast-Brown

November 2009

References

Hartemink, A.E., McBratney, A., and Medonça-Santos, M.L. (eds.), 2008. Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), 2007. Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science Vol. 31, Elsevier, Amsterdam.

Contents

Section B Soil Sensors and Remote Sensing

Section C Soil Inference Systems

Part III Environmental Application and Assessment

Part IV Making Digital Soil Mapping Operational

About the Editors

Dr. Janis L. Boettinger is professor of soil science (pedology) at Utah State University, USA. Her research, graduate student mentoring, and outreach programs focus on digital soil mapping, particularly on bridging research and operation of digital soil mapping in the USA National Cooperative Soil Survey program. Janis led the organization of the 3rd Global Workshop on Digital Soil Mapping in Logan, Utah, USA, in 2008.

Dr. Alfred E. Hartemink is a soil scientist at ISRIC – World Soil Information in Wageningen, the Netherlands. He has worked as soil surveyor and soil fertility specialist in Congo, Kenya, Tanzania, Indonesia, Papua New Guinea and Australia. He coordinates the GlobalSoilMap.net project that works on a new digital soil map of the world. Alfred is the Deputy Secretary General of the International Union of Soil Sciences and joint editor-in-chief of *Geoderma*.

David W. Howell worked as a field soil scientist and soil survey manager in the western USA prior to completion of graduate studies in 1999. David led implementation of geographic information systems (GIS) in soil survey applications and development of digital soil mapping in California for the US Department of Agriculture Natural Resources Conservation Service (USDA-NRCS). He participated in the Global Digital Soil Mapping Workshops in Montpellier, France; Rio de Janeiro, Brazil; and Logan, Utah. David now works in mapping, remote sensing, and travel photography in his own business called Earth Map Photo^{TM}.

Suzann Kienast-Brown is a soil scientist, GIS and Remote Sensing Specialist for the USDA-NRCS in Logan, Utah, USA. Suzann applies digital soil mapping in the USA Soil Survey Program and instructs a course on the application of remote sensing techniques in soil survey to USDA soil scientists. She received her Master of Science in Soil Science from Utah State University and is currently pursuing a doctoral degree focusing on digital soil mapping techniques for soil survey updates in the western USA.

Amanda C. Moore is a soil scientist with the USDA-NRCS. Her work has focused on the development and implementation of GIS methods for creating and using soils data and the integration of these methods into the USA soil survey program. Amanda is currently the State Soil Scientist for Maryland, Delaware, and the District of Columbia, USA.

Contributors

S.N. Bacon Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV 89512, USA, sbacon@dri.edu

J.S. Bailly INRA Laboratoire d'étude des Interactions Sol Agrosystème Hydrosystème (LISAH),UMR 1221 INRA-IRD-Supagro Montpellier, Montpellier, France, bailly@teledetection.fr

S.E. Baker Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV 89512, USA, sophie.baker@dri.edu

C. Ballabio Environmental and Land Sciences Department (DISAT) and Geology Department, University of Milano-Bicocca, 20126, Milano, Italy, cristiano.ballabio@unimib.it

F. Baret INRA, EMMAH, UMR 1114 INRA – University of Avignon, site Agroparc, 84914 Avignon, France, baret@avignon.inra.fr

J.R.F. Barringer Landcare Research, PO Box 40, Lincoln 7640, New Zealand, barringerj@landcareresearch.co.nz

S.D. Bassett Department of Geography, University of Nevada, Reno, NV 89577, USA, sbassett@unr.edu

P. Bellamy National Soil Resources Institute, Cranfield University, Bedfordshire, MK43 0AL, UK, p.bellamy@cranfield.ac.uk

R. Benton USDA Forest Service, Remote Sensing Applications Center, 2222 W. 2300 South, Salt Lake City, UT 84119, USA, robertbenton@fs.fed.us

R.L.L. Berbara UFRRJ – The Federal Rural University of Rio de Janeiro, BR 465, km 7, 23.890-000, Seropédica, RJ, Brazil, berbara@ufrrj.br

T. Bialkó Soil Protection Department, Plant Protection and Soil Conservation Authority of B.A.Z. County, Miskolc, Hungary, BIALKO.TIBOR@borsod.ontsz.hu

C. Blinn Department of Forestry, Virginia Tech, 216D Cheatham Hall, Blacksburg, VA, USA 24061, cblinn@vt.edu

P.K. Bøcher Department of Agroecology and Environment, Faculty of Agricultural Sciences, University of Aarhus, Aarhus, Denmark, Peder.Bocher@agrsci.dk

J.L. Boettinger Department of Plants, Soils, and Climate; Utah State University; 4820 Old Main Hill, Logan, UT 84322-4820, USA, janis.boettinger@usu.edu

B. Borchers Department of Mathematics, New Mexico Tech, 801 Leroy Place, Socorro, NM 87801, USA, borchers@nmt.edu

R. Bou Kheir Department of Agroecology and Environment, Faculty of Agricultural Sciences, University of Aarhus, Aarhus, Denmark, Rania.BouKheir@agrsci.dk

D. Brown Crop and Soil Sciences Department, Washington State University, Pullman, WA, USA, david_brown@wsu.edu

C.W. Brungard Department of Plant, Soils and Climate, Utah State University, 4820 Old Main Hill, Logan, UT 84322-4820, USA, c.w.b@aggiemail.usu.edu

T.F. Bullard Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV 89512, USA, Tom.Bullard@dri.edu

F. Carré Land Management & Natural Hazards Unit, DG JRC, 21020 Ispra (VA), Italy; INERIS, Scientific Division, Parc technologique Alata, BP 7, 60550 Verneuil en Halatte, France, Florence.CARRE@ineris.fr

S. Casalegno Land Management & Natural Hazards Unit, DG JRC, 21020 Ispra (VA), Italy; Predictive Models for Biomedicine and Environment Fondazione Bruno Kessler, Via Sommarive 18, I-38123 Povo (Trento), Italy, stefano@casalegno.net

D.G. Chandler Department of Civil Engineering, Kansas State University, Manhattan, KS 66506, USA, dcg@ksu.edu

I. Chiuchiarelli ARSSA, Agricultural Extension Service of Abruzzo Region, Piazza Torlonia 91, 67051 Avezzano AQ, Italy, igino.chiuchiarelli@tin.it

M.R. Coelho EMBRAPA Solos – Brazilian Agricultural Research Corporation, The National Centre of Soil Research, Rua Jardim Botânico, 1.024, 22.460-000, Rio de Janeiro, RJ, Brazil, mauricio@cnps.embrapa.br

R. Comolli Environmental and Land Sciences Department (DISAT), University of Milano-Bicocca, 20126, Milano, Italy, roberto.comolli@unimib.it

R. Corstanje National Soil Resources Institute, Cranfield University, Bedfordshire, MK43 0AL, UK, roncorstanje@cranfield.ac.uk

G. Coulouma INRA Laboratoire d'étude des Interactions Sol Agrosystème Hydrosystème (LISAH),UMR 1221 INRA-IRD-Supagro Montpellier, Montpellier, France

R.A. Coupé B.C. Ministry of Forests and Range, Williams Lake, B.C., Canada, Ray.Coupe@gov.bc.ca

G.K. Dalldorf Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV 89512, USA, Graham.Dalldorf@dri.edu

R.O. Dart EMBRAPA Solos – Brazilian Agricultural Research Corporation, The National Centre of Soil Research, Rua Jardim Botânico, 1.024, 22.460-000, Rio de Janeiro, RJ, Brazil, dart@cnps.embrapa.br

E. Dobos University of Miskolc, Miskolc (Egyetemváros), 3515, Hungary, ecodobos@uni-miskolc.hu

E.M. Engle Department of Earth and Environmental Science, New Mexico Tech, 801 Leroy Place, Socorro, NM 87801, USA, eengle@nmt.edu; emmengle@gmail.com

A. Farshad International Institute of Geo-information Sciences and Earth Observation (ITC), Enschede, The Netherlands, farshad@itc.nl

H. Fisk USDA Forest Service, Remote Sensing Applications Center, 2222 W. 2300 South, Salt Lake City, UT 84119, USA, hfisk@fs.fed.us

G.J. Forrester Landcare Research, PO Box 40, Lincoln 7640, New Zealand, forresterg@landcareresearch.co.nz

R. Francaviglia CRA, Research Centre for the Soil-Plant System, Via della Navicella 2-4, 00184 Rome, Italy, rosa.francaviglia@entecra.it

W. Fraser Soil Resource Group, Agriculture and Agri-Food Canada, Cereal Research Centre, 195 Dafoe Road, Winnipeg, Manitoba, Canada, R3T 2M9, xiaoyuan.geng@agr.gc.ca

B. Frazier Crop and Soil Sciences Department, Washington State University, Pullman, WA, USA, bfrazier@wsu.edu

X. Geng CanSIS, AESB, Agriculture and Agri-Food Canada, Neatby Building, 960 Carling Av., Ottawa, Ontario, Canada, K1A 0C6, xiaoyuan.geng@agr.gc.ca

C. Gomez INRA Laboratoire d'étude des Interactions Sol Agrosystème Hydrosystème (LISAH),UMR 1221 INRA-IRD-Supagro Montpellier, Montpellier, France, gomez@supagro.inra.fr

M.B. Greve Department of Agroecology and Environment, Faculty of Agricultural Sciences, University of Aarhus, Aarhus, Denmark, MetteB.Greve@agrsci.dk

M.H. Greve Department of Agroecology and Environment, Faculty of Agricultural Sciences, University of Aarhus, Aarhus, Denmark, Mogensh.greve@agrsci.dk

M.J. Grundy CSIRO Land and Water, Bruce E. Butler Laboratory, GPO Box 1666, Canberra ACT 2601, Australia, mike.grundy@csiro.au

S. Grunwald Soil and Water Science Department, University of Florida, 2169 McCarty Hall, PO Box 110290, Gainesville, FL 32611, USA, sabgru@ufl.edu

J.B.J. Harrison Department of Earth and Environmental Science, New Mexico Tech, 801 Leroy Place, Socorro, NM 87801, USA, bruce@nmt.edu

A.E. Hartemink ISRIC – World Soil Information, Wageningen, The Netherlands, alfred.hartemink@wur.nl

C. Haydu-Houdeshell USDA Natural Resources Conservation Service, 14393 Park Avenue, Suite 200, Victorville, CA 92392, USA, carrie-ann.houdeshell@ca.usda.gov

J. Hempel USDA Natural Resources Conservation Service, National Soil Survey Center, Federal Bldg. Rm 152, 100 Centennial Mall North Lincoln, NE, 68508, USA, jon.hempel@lin.usda.gov

J.M.H. Hendrickx Department of Earth and Environmental Science, New Mexico Tech, 801 Leroy Place, Socorro, NM 87801, USA, hendrick@nmt.edu

T. Hengl Computational Geo-Ecology (CGE), Faculty of Science Institute for Biodiversity and Ecosystem Dynamics, Universiteit van Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands, T.Hengl@uva.nl

A.E. Hewitt Landcare Research, PO Box 40, Lincoln 7640, New Zealand, hewitta@landcareresearch.co.nz

D.W. Howell USDA Natural Resources Conservation Service (Retired), P.O. Box 709, Arcata, CA 95518, USA; 3400 McMillan Court, Arcata, CA 95521 USA, david@earthmapphoto.com

N. Jeannée Géovariances, 49bis avenue Franklin Roosevelt, BP 91, 77212 Avon Cedex, France, jeannee@geovariances.com

Y. Jiao Soil Resource Group, Agriculture and Agri-Food Canada, Potato Research Centre, 850 Lincoln Rd., Fredericton, New Brunswick, Canada, E3B 4Z7, jiaoy@agr.gc.ca

S. Kienast-Brown USDA Natural Resources Conservation Service, Department of Plants, Soils, and Climate, Utah State University, 4820 Old Main Hill, Logan, UT 84322-4820, USA, suzann.kienast@ut.usda.gov

T. King Gray & Pape, Inc., 1318 Main Street, Cincinnati, OH 45202, USA, tking@graypape.com

J. Kobza Institute/Soil Science and Conservation Research Institute, Banska Bystrica, Slovakia, kobza.vupop@bystrica.sk

M.V. Konyushkova V.V. Dokuchaev Soil Science Institute, Pyzhevsky per. 7, Moscow 119017, Russia, mkon@inbox.ru

P. Lagacherie INRA Laboratoire d'étude des Interactions Sol Agrosystème Hydrosystème (LISAH),UMR 1221 INRA-IRD-Supagro Montpellier, Montpellier, France, lagache@supagro.inra.fr

R. Larsen Department of Agroecology and Environment, Faculty of Agricultural Sciences, University of Aarhus, Aarhus, Denmark, Rene.Larsen@agrsci.dk

O. Lemarchand Géovariances, 49bis avenue Franklin Roosevelt, BP 91, 77212 Avon Cedex, France, lemarchand@geovariances.com

R.F. Long USDA-Natural Resources Conservation Service, 481 Summer Street, Suite 202, St. Johnsbury, VT 05819, USA, robert.long@vt.usda.gov

J.F. Lumbreras EMBRAPA Solos – Brazilian Agricultural Research Corporation, The National Centre of Soil Research, Rua Jardim Botânico, 1.024, 22.460-000, Rio de Janeiro, RJ, Brazil, jflum@cnps.embrapa.br

S.R. MacCabe Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV 89512, USA, Shawn.MacCabe@dri.edu

R.A. MacMillan LandMapper Environmental Solutions Inc., Edmonton, AB, Canada; ISRIC - World Soil Information, Wageningen, The Netherlands, bobmacm@telusplanet.net; bob.macmillan@wur.nl

B.P. Mallavan Montpellier SupAgro, 2 place Pierre Viala, 34060 Montpellier Cedex 01, France, mallavanben@hotmail.fr

A. Marchetti CRA, Research Centre for the Soil-Plant System, Via della Navicella 2-4, 00184 Rome, Italy, alessandro.marchetti@entecra.it

T. Mayr National Soil Resources Institute, Cranfield University, Bedfordshire, MK43 0AL, UK, t.mayr@cranfield.ac.uk

A.B. McBratney Faculty of Agriculture, Food & Natural Resources, The University of Sydney, Sydney, NSW 2006, Australia, alex.mcbratney@sydney.edu.au

K. McCloy Department of Agroecology and Environment, Faculty of Agricultural Sciences, University of Aarhus, Aarhus, Denmark, Keith.McCloy@agrsci.dk

E.V. McDonald Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV 89512, USA, Eric.Mcdonald@dri.edu

J. McKay USDA Natural Resources Conservation Service, 481 Summer Street, Suite 202, St. Johnsbury, VT 05819, USA, jessica.mckay@vt.usda.gov

N.J. McKenzie CSIRO Land and Water, Bruce E. Butler Laboratory, GPO Box 1666, Canberra ACT 2601, Australia, Neil.McKenzie@csiro.au

J.P. McNamara Department of Geosciences, Boise State University, Boise, ID 83725, USA, JMCNAMAR@boisestate.edu

S.J. McNeill Landcare Research, PO Box 40, Lincoln 7640, New Zealand, mcneills@landcareresearch.co.nz

E. Meirik Crop and Soil Sciences Department, Washington State University, Pullman, WA, USA, ee.meirik@gmail.com

M.L. Mendonça-Santos EMBRAPA Solos – Brazilian Agricultural Research Corporation, The National Centre of Soil Research, Rua Jardim Botânico, 1.024, 22.460-000, Rio de Janeiro, RJ, Brazil, loumendonca@cnps.embrapa.br

E. Micheli Szent István University, Páter K u. 1., Gödöllő, Hungary, micheli.erika@mkk.szie.hu

B. Minasny Faculty of Agriculture, Food & Natural Resources, The University of Sydney, Sydney, NSW 2006, Australia, budiman.minasny@sydney.edu.au

T.B. Minor Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV 89512, USA, Tim.Minor@dri.edu

L. Montanarella European Commission, Land Management and Natural Hazards Unit, Institute for Environment and Sustainability, DG Joint Research Center, TP 280, Via Fermi 2749, I - 21027 Ispra (VA), Italy, luca.montanarella@jrc.ec.europa.eu; luca.montanarella@jrc.it

D.E. Moon CDT - Core Decision Technologies Inc., Richmond, B.C., Canada, CDT-Moon@Shaw.ca

R. Moonjun International Institute of Geo-information Sciences and Earth Observation (ITC), Enschede, The Netherlands, moonjun13562@itc.nl

A.C. Moore USDA Natural Resources Conservation Service, 339 Busch's Frontage Road, Suite 301, Annapolis, MD 21409, USA, amanda.moore@md.usda.gov

J.S. Noller Department of Crop & Soil Science, Oregon State University, ALS3017, Corvallis, OR 97331, USA, jay.noller@oregonstate.edu

R. Palmer National Soil Resources Institute, Cranfield University, Bedfordshire, MK43 0AL, UK, r.palmer@cranfield.ac.uk

G. Patterson Soil Resource Group, Agriculture and Agri-Food Canada, Atlantic Food and Horticulture Research Centre, 20 Tower Rd., Truro, Nova Scotia, Canada, B2N 5E3, pattersong@agr.gc.ca

N. Phillips Nona Phillips Forestry Consulting, Williams Lake, B.C., Canada, nophilli@shaw.ca

C. Piccini CRA, Research Centre for the Soil-Plant System, Via della Navicella 2-4, 00184 Rome, Italy, chiara.piccini@entecra.it

R.D. Ramsey Department of Wildland Resources, Utah State University, Logan, UT 84322-5230, USA, doug.ramsey@usu.edu

H.I. Reuter Land Management & Natural Hazards Unit, DG JRC, 21020 Ispra (VA), Italy; Gisxperts gbr, Eichenweg 42, D-06849 Dessau, Germany, hannes@gisxperts.de

M. Rivas-Casado National Soil Resources Institute, Cranfield University, Bedfordshire, MK43 0AL, UK, m.rivas-casado@cranfield.ac.uk

P. Roberts Crop and Soil Sciences Department, Washington State University, Pullman, WA, USA, probertswsu@gmail.com

L. Rodriguez Lado Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Duebendorf, Switzerland, luis.rodriguez-lado@eawag.ch

S.M. Roecker USDA Natural Resources Conservation Service, Victorville, CA 92392, USA, stephen.roecker@ca.usda.gov

R. Rupp Crop and Soil Sciences Department, Washington State University, Pullman, WA, USA, richard_rupp@wsu.edu

D.E. Sabol Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV 89512, USA, Don.Sabol@dri.edu

P. Sanchez Director, Tropical Agriculture and Rural Environment Director, Millennium Villages Project, The Earth Institute at Columbia University, Palisades, NY 10964-8000, USA, psanchez@ei.columbia.edu

H.G. Santos EMBRAPA Solos – Brazilian Agricultural Research Corporation, The National Centre of Soil Research, Rua Jardim Botânico, 1.024, 22.460-000, Rio de Janeiro, RJ, Brazil, humberto@cnps.embrapa.br

S. Santucci ARSSA, Agricultural Extension Service of Abruzzo Region, Piazza Torlonia 91, 67051 Avezzano AQ, Italy, sergio.santucci@tin.it

X. Shi Dartmouth College, 6017 Fairchild, Hanover, NH 03755, USA, xun.shi@dartmouth.edu

X.Z. Shi State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, No. 71 East Beijing Road, Nanjing, 210008 China, xzshi@issas.ac.cn

D.P. Shrestha International Institute of Geo-information Sciences and Earth Observation (ITC), Enschede, The Netherlands, shrestha@itc.nl

D. Smith USDA Natural Resources Conservation Service, 430 G Street #4164, Davis, CA 95616-4164, USA, dave.smith@ca.usda.gov

S. Smith Soil Resource Group, Agriculture and Agri-Food Canada, Pacific Agri-Food Research Centre, 6947 Highway 7, PO Box 1000, Agassiz, British Columbia, Canada, V0M 1A0, smithcas@agr.gc.ca

A.K. Stum USDA Natural Resources Conservation Service, 340 North 600 East, Richfield, UT 84701, USA, alex.stum@ut.usda.gov

D.G. Tarboton Civil and Environmental Engineering Department, Utah State University, Logan, UT 84322-4110, USA, david.tarboton@usu.edu

T.K. Tesfa Pacific Northwest National Laboratory, PO Box 999 Richland, WA 99352, USA, Teklu.Tesfa@pnl.gov

J.A. Thompson Division of Plant and Soil Sciences, West Virginia University, PO Box 6108, Morgantown, WV 26506-6108, USA, james.thompson@mail.wvu.edu

C. Unger Utah Geology Survey, 1594 W. North Temple, Salt Lake City, UT 84114, USA, coreyunger@utah.gov

C. Vaiphasa Chulalongkhorn University, Bangkok, Thailand, vaiphasa@alumni.itc.nl

B. VandenBygaart Soil Resource Group, Agriculture and Agri-Food Canada, Neatby Building, 960 Carling Av., Ottawa, Ontario, Canada, K1A 0C6, vandenbygaarta@agr.gc.ca

R.A. Viscarra Rossel CSIRO Land and Water, Bruce E. Butler Laboratory, GPO Box 1666, Canberra ACT 2601, Australia, Raphael.Viscarra-Rossel@csiro.au

A. Waddell Soil Resource Group, Agriculture and Agri-Food Canada, Cereal Research Centre, 195 Dafoe Road, Winnipeg, Manitoba, Canada, R3T 2M9, waddella@agr.gc.ca

M. Walsh Tropical Soil Biology and Fertility Institute (CIAT-TSBF), ICRAF Complex, UN Avenue, Gigiri, Nairobi, P.O. Box 30677-00100, Nairobi, Kenya, markusgwalsh@gmail.com

M.A. White Department of Watershed Science, Utah State University, Logan, UT 84322-5410, USA, mikew.usu@gmail.com

S. Williamson New Mexico State University – Jornada Experimental Range, P.O. Box 30003, MSC 3JER, NMSU, Las Cruces, NM 88003-8003, smwill@nmsu.edu

J. Zawadzka National Soil Resources Institute, Cranfield University, Bedfordshire, MK43 0AL, UK, j.zawadzka@cranfield.ac.uk

G.L. Zhang Institute of Soil Science, Chinese Academy of Sciences, 71 Beijingdonglu, Nanjing 210008, China, glzhang@issas.ac.cn

Y.C. Zhao State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, No. 71 East Beijing Road, Nanjing, 210008 China, yczhao@issas.ac.cn

Part I
Introduction

Chapter 1
Current State of Digital Soil Mapping and What Is Next

S. Grunwald

Abstract Digital soil mapping (DSM) involves research and operational applications to infer on patterns of soils across various spatial and temporal scales. DSM is not solely focused to map soils and their properties, but often environmental issues such as land degradation and global climate change, require assessing soils in context of ecosystem change and environmental stressors imparting control on soil properties. In this section an overview is provided of state-of-the art DSM applications and their constraints and potential is discussed. Future trends and challenges to map soils using digital approaches are outlined.

Keywords Environmental covariates · Soil sensors · Soil inference systems · Legacy soil data · Environmental assessment

1.1 Introduction

Digital soil mapping has evolved as a discipline linking field, laboratory, and proximal soil observations with quantitative methods to infer on spatial patterns of soils across various spatial and temporal scales. Studies use various approaches to predict soil properties or classes including univariate and multi-variate statistical, geostatistical and hybrid methods, and process-based models that relate soils to environmental covariates considering spatial and temporal dimensions. A comprehensive overview of digital soil mapping was provided by McBratney et al. (2003) and Grunwald (2006). Discussions of state-of-the-art digital soil mapping applications at different extents, geographic settings, and model resolutions (grains) were provided by Lagacherie et al. (2007) and Hartemink et al. (2008).

Research-focused digital soil mapping contrasts with agency-operated soil surveys. The dichotomy between research and agency-operated digital soil mapping is due to different sets of qualities. The former strives to find the best method/model to

S. Grunwald (✉)
Soil and Water Science Department, University of Florida, 2169 McCarty Hall,
PO Box 110290, Gainesville, FL 32611, USA
e-mail: sabgru@ufl.edu

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_1,

estimate soil characteristics exploiting digital, quantitative and emerging technologies with rigorous errors and uncertainty assessments. The latter aims to implement a standardized mapping protocol to characterize soils across a soil survey region. Soil taxonomic mapping of soil map units and development of soil information systems have played major roles in agency-operated soil surveys covering regional, national, and global scales. Whereas historically soil data needs were driven by food and fiber production (agriculture-centered period), more recent needs for soil data are more diverse with pronounced environmental-centered drivers requesting high-resolution, pixel-based soil products, which are associated with error assessment.

Traditional soil surveys explicitly incorporate pedological knowledge into the soil survey product, but have become costly and time-consuming when compared to emerging digital soil mapping approaches, such as diffuse reflectance spectroscopy (Lagacherie, 2008). This has evoked the thought to investigate in more detail how research and operational soil mapping can be fused. Grunwald (2009) presented a comprehensive analysis of recent digital soil mapping literature and pointed out that merging of quantitative, geographic, and pedological expertise is required to link production-oriented and research-oriented digital soil mapping. There is no universal soil equation or digital soil prediction model that fits all geographic regions and purposes, which complicates matters.

At the 3rd Global Workshop on Digital Soil Mapping organized by the International Union of Soil Sciences, Soil Science Society of America and Utah State University, Logan, UT, September 30–October 3, 2008, researchers, agency scientists, and practitioners met to share knowledge on digital soil mapping. This book compiles the outcomes from this Workshop in form of 34 chapters.

1.2 Research

1.2.1 Environmental Covariates and Soil Sampling

The section "Environmental Covariates and Soil Sampling" presents various chapters that focus on how environmental covariates are used to model soil properties. Factorial soil-landscape models form the conceptual framework for relating environmental covariates to soil properties as formalized in the *CLORPT* model (*CL*: Climate; *O*: Organism, vegetation; *R*: Relief; *P*: Parent material; and T: time) (Jenny, 1941) and the *SCORPAN* model (*S*: soil property or class; *C*: Climate; *A*: Age or time factor; and *N*: Space, spatial position) (McBratney et al., 2003) that are used to predict soil properties/classes (S_p). The *SCORPAN* model is made spatially and temporally explicit by predicting S_c (soil classes) or S_a (soil attributes) at a specific geographic location (x and y coordinates) and time. Grunwald (2006) extended the *SCORPAN* model by incorporating the vertical dimension (z), or depth of a specific soil property. Similarly, "environmental correlation" describes a method of relating environmental attributes (or environmental covariates) to S_p (McKenzie and Austin, 1993).

Although these conceptual models are accepted widely for digital soil mapping, the strength of relationships between environmental covariates and soil properties of interest differ by geographic region, observation/derivation method used to map environmental properties, spatial and temporal scales, and the specific soil property under investigation. In a comprehensive review, McBratney et al. (2003) found that the key environmental covariates for inferring S_a or S_c, were R (80% of studies) followed by S (35%), O and P (both 25%), N (20%), and C (5%). In contrast, in a review of 90 digital soil mapping journal articles Grunwald (2009) found that the contribution of S was 51%, C 6%, O 34%, R 24%, and P 6% to predict soil properties and classes.

To further investigate the behavior between environmental covariates and soil properties of interest various studies are presented in Section A. Chapter 2 discusses the use of environmental covariates in the Western USA derived from digital elevation models (DEMs) and remote sensing imagery (ASTER and Landsat) to infer on topography, climate, geomorphology, parent material, soil, and vegetation properties. These environmental properties are incorporated into soil prediction models to support soil mapping efforts, in particular, in the western USA region which still lacks initial soil mapping on private and public lands. In Chapter 3 a suite of topographic and land cover attributes to infer on soil depth using a Generalized Additive Model and Random Forest in a watershed in Boise, Idaho, USA is used. The importance of incorporating age (A factor) explicitly into digital soil models are emphasized in Chapter 4 fusing geological maps, age point data, and remote sensing data to infer on geochronology using a decision-tree analysis. The author indicates that incorporating the A factor explicitly into soil-prediction models as a co-variant is rare. A has been more often incorporated in implicit form carried in the age of parent material (P) and land form (R). In Chapter 5 different terrain attributes by varying grid and neighborhood sizes and investigate their effect on subsequent modeling of soil attributes are derived. Their study highlights that terrain attributes are specific to geographic land surfaces. Disparate neighborhood sizes correlate strongest with specific soil properties (soil carbon, rock fragment content, and clay content) suggesting that there is "no optimal" neighborhood size to model different soil properties. In Chapter 6 authors go after finding the optimal sample size for digital soil mapping in arid rangelands in Utah, USA. They employ conditioned Latin Hypercube sampling on five environmental covariates and identify an optimal sample size of 200–300 which is approximately 0.05–0.1% of the available potential sampling points in the 30,000 ha study area.

1.2.2 Soil Sensors and Remote Sensing

Sensing of both soils and environmental covariates is widely used in digital soil mapping studies. Lab-based or in-situ diffuse reflectance spectroscopy have been employed in the visible, near-infrared, and mid-infrared range to infer on a multitude of soil properties with varying success (Reeves, 2010). Other soil sensors map penetration resistance using cone penetrometers, apparent electrical conductivity, or

magnetic susceptibility (Grunwald and Lamsal, 2006). Grunwald (2009) found that out of 90 reviewed digital soil mapping studies 39% utilized soil or remote sensors, out of which 23.3% used soil sensors to complement analytical soil data which are more costly and labor-intensive to derive. In 16.7% of the studies, visible/near-infrared, mid-infrared, and/or Fourier-transform spectroscopy were used to infer different properties including soil organic carbon (SOC), texture, and others. Remote sensing applications that map soil properties, landscape or soil map boundaries, or environmental covariates, such as vegetation or climatic properties, can be readily incorporated into digital soil prediction models. A variety of satellite images are used in digital soil mapping projects including Landsat Enhanced Thematic Mapper (ETM), Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), Quickbird, IKONOS or others. These satellite images differ in their spatial resolution, spectral range and spectral resolution, which may affect the capabilities to infer on soil and environmental covariates. Remote sensing is confounded by the fact that the land surface is a mix of bare soil (with varying soil moisture content) and vegetation coverage which impact reflectance patterns sensed by aerial or satellite sensors. But aerial and satellite images provide dense grids of information across landscapes allowing to characterize SCORPAN factors.

Chapter 7 provides an overview of proximal soil sensors for digital soil mapping including electromagnetic induction, magnetic sensors, gravitometers, ground penetrating radar, magnetic resonance sounding, gamma-radiometrics, and diffuse reflectance spectroscopy. The use of hyperspectral imagery with 5 m spatial resolution to map clay content and calcium carbonate content in a Mediterranean region is presented in Chapter 8. In this chapter, special attention is given to derive soil data from a region that is partially covered by vegetation using hyperspectral images accounting for atmospheric effects. In Chapter 9 Quickbird imagery with 2.4 m spatial resolution are used to discriminate between different soil types including chernozem-like soils, light chestnut soils, and solonetzes (sodic soils). In Chapter 10 ASTER imagery with 15 m spatial resolution to infer on vegetation and correlate it to soil horizons are used in the North Cascades National Park in Washington State, USA. In Chapter 11 quantitative hydrologic parameters, such as root zone soil moisture obtained by land-surface energy models, are used for the identification of soil boundaries. They employ Landsat imagery with 30 m spatial resolution to infer on root zone moisture based on a multi-temporal analysis.

1.2.3 Soil Inference Systems

McBratney et al. (2002) provided an overview of soil inference systems, which take measurements we more-or-less know with a given level of (un)certainty, and infers data we do not know with minimal inaccuracy, by means of properly and logically conjoined pedotransfer functions (PTFs). In essence, the soil inference system has a source, an organizer, and a predictor. The inference system is a collection of logical rules selecting the PTFs with the minimum variance (McBratney et al., 2002).

In her recent digital soil mapping review study Grunwald (2009) found that the most popular soil inference methods (41.1%) were regressions followed by classification/discrimination methods (32.2%), and tree-based methods (e.g. Classification and Regression Trees, Random Forest) (13.3%). Other methods such as GIS-based modeling, neural networks, and fuzzy logic based models were less frequently used to predict soils. In her comprehensive review study, knowledge-based digital soil models that rely on expert knowledge were rare when compared to stochastic or deterministic methods to predict soils. Out of 90 reviewed journal articles 40.0% presented soil prediction results derived from only one method, whereas 60.0% used two or more quantitative methods to predict or model soil properties/classes. Grunwald (2009) found that 36.7% of 90 reviewed digital soil mapping journal articles used legacy data in their research.

In Chapter 12 Homosoil, a methodology for quantitative extrapolation of soil information across the globe is presented. Homosoil facilitates to map soils in places where soil information is difficult to obtain or does not exist. A major assumption of this conceptual approach is homology of soil-forming factors between a reference area and the region of interest. Gower's similarity index is used to quantify similarity in climate, physiography, and parent materials in a reference area and the rest of the world. In Chapter 13 Artificial Neural Networks and a Decision Tree model for predictive soil mapping based on the *SCORPAN* approach are employed in a poorly accessible 20 km^2 watershed in Thailand. In Chapter 14 a knowledge-based approach and a rule-based fuzzy inference engine, Soil Inference Engine (SIE), is used in two small watersheds in Vermont. In this study not only the predictive capability of the inference engine is evaluated to infer on soil series and drainage classes, but also the potential to transfer the prediction model to a watershed with similar landscape characteristics is assessed. In Chapter 15 Random Forest to predict soil classes using environmental covariates derived from Landsat ETM and a DEM in an arid region in Utah are employed. Chapter 16 explicitly incorporates legacy data into the soil predictive models (sand, silt, clay and organic carbon) implemented using Generalized Linear Modeling and Bayesian Belief Networks. Authors of this chapter emphasize the limitations of using legacy data that may not cover the existing feature space (i.e., the range of attribute values present in a given region) and may contain a mix of qualitative and quantitative data.

1.3 Environmental Application and Assessment

Historically, soil surveys have focused on soil descriptions and mapping of taxonomic soil data and standard soil properties. Recently, the emphasis has shifted from classification and inventory to understanding and quantifying spatially and temporally soil patterns to address environmental problems. This environmental-centered approach views soils as integral part of an ecosystem interacting with environmental factors generating complex patterns and processes that co-evolve through time. Environmental-centered digital soil mapping responds to critical societal needs including environmental quality assessment, soil degradation, soil quality,

and health as outlined in Hartemink (2006). Spatially-explicit soil carbon assessment over large landscapes has gained attention to help mitigate rising levels of greenhouse gases in the atmosphere.

In a review it was found that out of 90 investigated digital soil mapping studies, 40.0% focused on predictions of base soil properties such as texture, bulk density, and structure, 31.1% on soil carbon/global climate change, 24.4% on eutrophication/environmental quality assessment, 16.6% on hydrologic properties (such as soil moisture, saturated hydraulic conductivity, or soil water content), 8.9% on soil degradation (salinity, acidity, and erosion), and 15.6% on mapping of soil taxonomic/ecological classes (Grunwald, 2009). In particular, studies that focus on mapping of SOC and soil organic matter (SOM) are prominently represented in the recent digital soil mapping literature.

Chapter 17 addressed the problem of heavy metals in soils in a study site in the Italian Alps. They use multi-scale Support Vector Regression (SVR), a machine learning technique, to model distribution patterns of heavy metals. SVR is a non-parametric technique based on Structural Risk Minimization that aims to optimize model performance by minimizing both the error and the model complexity. Chapter 18 map carbon/nitrogen (C/N) ratio of forest soils aiming to evaluate soil functions and provide needed information to address climate change in Europe. Interestingly, in their study the classical Kriging approach performs better to model C/N ratios when compared to Neural Network modeling of C/N using environmental covariates. This may be explained by the scale of the "global soil mapping approach" extending over Europe. Chapter 19 compares various methods (Multiple Linear Regression, Universal Kriging, Regression Kriging, Artificial Neural Network–Kriging, Regression Tree, and Sequential Indicator Simulation) to model SOC in a province in China, with Regression Tree outperforming all other tested methods. In Chapter 20 SOM is estimated using Regression Kriging and various environmental covariates in central Italy. The topsoil SOC stocks are estimated using six different sets of SCORPAN factors implemented using Multi-linear Regression analysis and Regression Kriging in Rio de Jaineiro State (Chapter 21). Chapter 22 assesses the extent of organic soils in Denmark using Decision Tree Modeling and Indicator Kriging, which classified 58 and 52% correctly, respectively. In Chapter 23 wind erosion is assessed in the Danuve Basin using Regression Kriging and various environmental covariates bridging the gap between digital soil mapping and digital soil risk assessment.

1.4 Making Digital Soil Mapping Operational

Research-focused digital soil mapping contrasts with need-driven digital soil mapping and agency-operated soil surveys. Digital soil mapping studies are diverse with specialized, mathematical prototype models tested on limited geographic regions and/or datasets and simpler, operational digital soil mapping used for routine mapping over large soil regions.

Grunwald (2009) pointed out that numerous research-oriented digital soil mapping studies ranked high in terms of quantitative knowledge and expertise, but lacked pedological interpretations which may limit widespread adoption by practitioners. In her review study she documented various complex digital soil mapping methods, such as genetic programming, Simulation of Gaussian Fields, Markov Chain Random Fields, or mechanistic models, which require profound mathematical expertise. Minasny and McBratney (2007) suggested that for practical applications digital soil prediction methods, such as Regression Kriging, may be mathematically biased, however, they appear robust to predict soil properties in various soil regions. Both authors conclude that improvement in the prediction of soil properties does not rely on more sophisticated quantitative methods, but rather on gathering more useful and higher quality data.

There are multiple studies presented in this book that use Multivariate Regression, Regression Kriging, Tree-based models, or Neural Networks, which are methods that are versatile and easy to implement. In these studies much effort is invested in assembly of *SCORPAN* factors from various sources (legacy datasets, soil and remote sensors, derivatives from DEMs, and others). In many cases, data collection of environmental covariates is given more attention than the collection of soil samples. The presented studies further suggest that there is not one method emerging that performs best to estimate multiple soil properties/classes in different geographic regions. Factors that confound findings to estimate soil properties and classes include sampling design and sample density, quality and spatial resolution of soil and environmental covariates, scale (extent of study site, model grain), data aggregation, and integration methods. Critical is to evaluate the performance of soil prediction models using calibration and/or validation. Grunwald (2009) found that out of 90 investigated studies 21.1% used cross-validation, 46.7% used validation, and 35.6% did not use cross-validation, validation or any other performance test. Rigorous performance tests to evaluate soil predictions in various geographic soil-landscape settings are critical to minimize uncertainty in soil predictions at unsampled locations.

Chapter 24 presents the S-map designed to deliver a new digital soil map, database, inference system, and soil information system for New Zealand. The system builds on legacy data, older soil surveys, expert-knowledge, and digital soil mapping methods. Preliminary results contrast expert-clustering and data-driven clustering deriving soilscapes. Chapter 25 addresses the problem of legacy soil data harmonization and data base integration for a region covering the Hungarian-Slovakian border. They form an integrated database of profiles using pedotransfer rules and environmental covariates to employ Regression Kriging and Maximum Likelihood Classification to derive soil groupings, pH, and humus content. Chapter 26 provides an overview of how existing soil survey data and expert-knowledge is linked to implement digital soil mapping in Canada. Authors aim is to produce raster-based soil maps that utilize existing soil survey information managed by the Canadian Soil Information System (CanSIS). Chapter 27 demonstrates predictive ecosystem mapping for 8.2 million hectare of forestland in British Columbia, Canada. Their approach for operational modeling of ecological entities is based

on a combination of fuzzy membership functions and knowledge-based predictive rules. Chapter 28 presents an operational initiative facilitated by the Natural Resources Conservation Service building digital soil mapping capacity within the U.S. National Cooperative Soil Survey. In this pilot project ASTER satellite imagery and DEM data will be used to create soil predictive models in the Joshua Tree National Park, Mojave Desert. Complimentary, Chapter 29 presents a qualitative comparison of conventional soil survey product and one derived using environmental covariates and Random Forest to predict soil subgroups as part of the Mojave Desert initiative. In Chapter 30 the Optimum Index Factor (OIF) to multiple data types is applied to identify the optimum combination of bands from Landsat TM and DEM data. The OIF is used to determine which data layers, derived from elevation data and remote-sensing images, best represent the full range of biophysical characteristics in a study area in north-eastern Utah. The optimum data layers are combined into a multiband image used for classification and modeling, and ultimately to create a pre-map for the study area. Chapter 31 presents the TEUI-Geospatial Toolkit which is an operational GIS-based ecological inventory application used by the U.S. Department of Agriculture, Forest Service and other land management agencies. In Chapter 32 a GIS framework and rule-based system developed by experts is used to map shallow soil condition to model dust emissions in the arid southwest U.S. Chapter 33 describes the GlobalSoilMap project that aims to produce a new digital soil map of the world with a grid resolution of 90 m $\times$ 90 m. The global soil map will be freely available, web-accessible, and widely distributed. The first portion of the global soil map is focused on Sub-Saharan Africa. Chapter 34 provides methodologies for global soil mapping based on the current state of knowledge incorporating legacy data, extracting information from soil maps, combining soil maps and soil point data, SCORPAN, Kriging, extrapolating based on reference areas, and the Homosoil approach.

1.5 What Is Next in Digital Soil Mapping

The methodological digital soil mapping framework to map soils across the globe at fine grains ($\leq$90 m $\times$ 90 m) has been formalized (compare Chapters 33 and 34; and McBratney et al., 2003). Research-oriented digital soil mapping studies presented in this book and digital soil mapping literature (Grunwald, 2009; Hartemink et al., 2008; Lagacherie et al., 2007) provide evidence that soil taxonomic data and soil properties can be predicted successfully using sets of environmental covariates as shown in various soil-landscape settings. Availability of high-resolution and hyperspectral remote sensing data, high-quality DEMs as well as soil sensors, and multi-sensor systems have facilitated to improve soil prediction models. In research-oriented digital soil mapping error and uncertainty metrics accompany soil estimates to document the quality of digital soil maps, which have often not been provided in transparent format by operational soil survey programs. The trend to formalize pedological expertise in form of quantitative soil prediction models of various types is continuing in the research community.

Carré et al. (2007) suggested that the next step in digital soil mapping is digital soil assessment, which comprises three main processes: (1) soil attribute space inference, (2) evaluation of soil functions and the threats to soils, and (3) risk assessment and the development of strategies for soil protection. Digital Soil Risk Assessment consists of integrating political, social, economical parameters and general environmental threats to digital soil assessment outputs for building, modeling and testing some scenarios about environmental perspectives. This path responds to the pressing environmental issues that requires accurate and high-resolution, spatially-explicit soil data to conduct a holistic assessment of soil-environmental systems.

Future challenges will entail to adapt digital soil mapping to various soil-landscape settings accounting for spatial as well as temporal soil patterns. Digital soil mapping will need to encompass three-dimensional soil bodies extending along profiles and across landscapes. So far, digital soil mapping has focused on the topsoil but mapping of soil characteristics in the subsurface are critical to address nutrient enrichment and pollution problems, carbon sequestration in spodic horizons, and more. To bridge the gap between research and operational digital soil mapping programs will require fusing of expert-knowledge from soil surveyors and research scientists. Despite technological and methodological advancements in digital soil mapping it will be critical to collect reconnaissance soil data without relying too much on legacy soil data. Fusing of soil and environmental covariates and development of multi-sensor systems will be important to advance digital soil mapping. The methodology for digital soil mapping as outlined by Minasny and McBratney (Chapter 34) will help to homogenize future digital soil mapping products.

References

Carré, F., McBratney, A.B., Mayr, T., and Montanarella, L., 2007. Digital soil assessment: Beyond DSM. Geoderma 142:69–79.

Grunwald, S., 2006. What do we really know about the space-time continuum of soil-landscapes, pp. 3–36. In: Grunwald, S. (ed.), Environmental Soil-Landscape Modeling – Geographic Information Technologies and Pedometrics. CRC Press, New York, NY.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152:195–207.

Grunwald, S., and Lamsal, S., 2006. Emerging geographic information technologies and soil information systems, pp. 127–154. In: Grunwald, S. (ed.), Environmental Soil-Landscape Modeling – Geographic Information Technologies and Pedometrics. CRC Press, New York, NY.

Hartemink, A.E. (Ed.), 2006. The Future of Soil Science. International Union of Soil Sciences, Wageningen.

Hartemink, A.E., McBratney, A.B., and Mendonça-Santos, M.L. (eds.), 2008. Digital Soil Mapping with Limited Data. Springer, Berlin.

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

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.

Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), 2007. Digital Soil Mapping – An Introductory Perspective. Developments in Soil Science, Vol., 31. Elsevier, Amsterdam.

McBratney A.B., Mendonça-Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

McBratney, A.B., Minasny, B., Cattle, S.R., and Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109:41–73.

McKenzie, N.J., and Austin, M.P., 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57:329–355.

Minasny, B., and McBratney, A.B., 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. Geoderma 140:324–336.

Reeves, J.B., 2010. Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? Geoderma (in press).

Part II
Research

Section A
Environmental Covariates and Soil Sampling

Chapter 2
Environmental Covariates for Digital Soil Mapping in the Western USA

J.L. Boettinger

Abstract Unlike most of the USA, the western region still lacks initial soil mapping on many private and public lands. Spatially explicit soil data are needed to support a variety of land uses including energy development, mining, grazing, agriculture, forest products, recreation, wildlife habitat, species conservation and urban development. The wide diversity of lithology, geomorphology, climate, and vegetation and the vast, difficult-to-access terrain of western landscapes facilitated the development of digital soil mapping. Environmental covariates used in digital soil mapping in the western USA include derivatives of digital elevations models (DEM) and remotely sensed spectral data (RS). DEM-derived data are often used to represent climate and relief (geomorphology, microclimate). The low vegetation cover of arid, semiarid, and seasonally dry climates typical of the West facilitates the use of RS data and its derivatives. Landsat and ASTER data are the most common sources of RS data, and are used to represent organisms (vegetation type and density), parent material (lithology, mineralogy), and soil (mineralogy, wetness, other surface characteristics). Spatially explicit, digital environmental covariates have improved several aspects of soil survey in the western USA including pre-mapping, developing efficient and targeted field sampling plans, and implementation of spatial prediction models.

Keywords Remote-sensing · Spectral reflectance · Topographic attributes · Band combinations

2.1 Introduction

Large areas of the USA still need initial soil survey data on public and private lands, particularly in the West (Fig. 2.1). The USDA Natural Resources Conservation Service, which provides leadership to the National Cooperative Soil Survey in the USA,

J.L. Boettinger (✉)
Department of Plants, Soils, and Climate; Utah State University; 4820 Old Main Hill, Logan, UT 84322-4820, USA
e-mail: janis.boettinger@usu.edu

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_2,

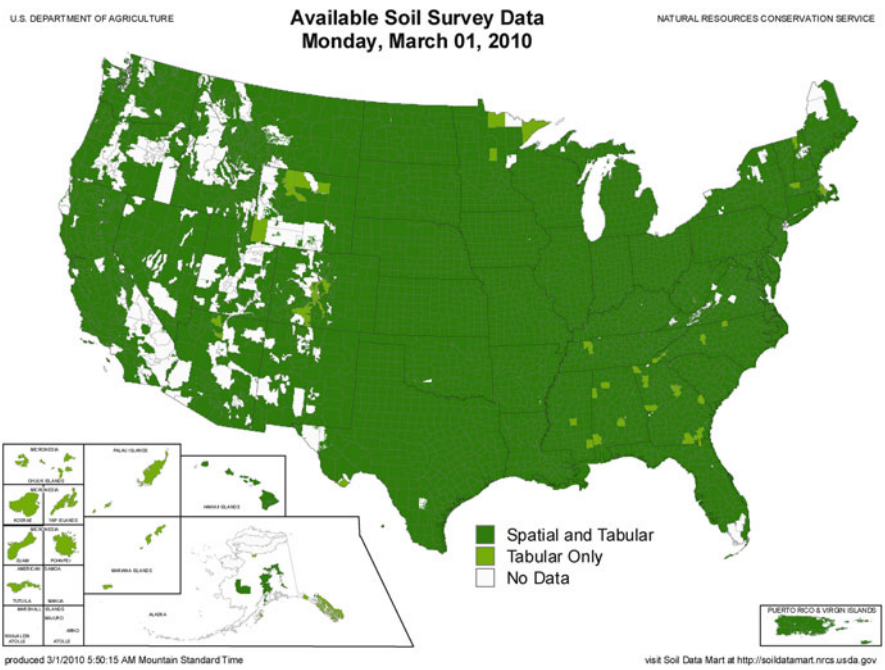


Fig. 2.1 Map showing the status of soil surveys in the USA and its territories. Areas shown in *white* still lack modern initial soil survey data

has a goal to complete the initial soil survey on Non-Federal, Native American, and high-priority Federal lands before the end of 2011. Other areas that have initial soil survey data now need updates of existing soil surveys to meet emerging demands for soil information, such as for urban planning, assessing environmental hazards and impacts of various land uses, predicting habitat of sensitive species, and the delivery of ecosystem services.

Federal (public) lands in the West managed by the US Department of Interior Bureau of Land Management and National Park Service and the US Department of Agriculture Forest Service that still lack initial soil survey are often vast tracts of land which are often very rugged with few roads (e.g., Fig. 2.2). While traditional uses of public lands were often of low intensity (i.e., grazing lands, recreation), recent increase in demand for domestic energy sources has placed intense development pressure on these lands, particularly those administered by the Bureau of Land Management. Traditional methods of soil survey commonly produce soil maps with broad map unit polygons representing soil associations and complexes with multiple soil and miscellaneous land type components. Therefore, even where soil survey data are available, the lack of specific spatially distributed soil information and associated estimates of uncertainty hinders land planning and management.

Until fairly recently, traditional soil survey was based largely on interpretation of aerial photography verified with field observations. Map unit concepts were

Fig. 2.2 Rugged, roadless landscape typical of public land managed by the Bureau of Land Management in southern Utah, USA

sometimes applied inconsistently across survey areas, with different soil scientists having different interpretations of soil-landscape relationships. Locations of pedon descriptions, transects, traverses, and observations were difficult to accurately geo-reference. Map unit concepts and rules of application were rarely documented and archived in places other than the minds of the soil scientists who moved on to map soils in other areas, making soil survey updates more difficult and time-consuming. Paper copies of pedon descriptions and other documentation were sometimes lost, damaged, or destroyed following completion of the soil survey.

The scientific foundation of soil mapping in the USA has traditionally been Jenny's (1941) conceptual model to explain that unique soils on the landscape are the products of unique sets of five soil-forming factors, including climate, organisms (mainly vegetation), relief, parent material, and time. Vegetation and relief were often interpreted from aerial photographs, in conjunction with topographic quadrangle maps published by the US Geological Survey (USGS). Parent material was interpreted from available surficial and bedrock geology maps published by USGS or state geological surveys. These data were usually of good quality, but it was cumbersome to spatially overlay these data layers and it was difficult to quantify relationships between soil-forming factors and the distribution of soils on the landscape.

Increased availability and user-friendliness of geographic information systems (GIS), software, and tools has improved the efficiency of initial and update soil survey. Most soil mapping currently involves the geo-referencing of data points and transects with global positioning systems (GPS), and spatial data can be viewed, overlaid, and queried using GIS software. Digital elevation models (DEM) are widely available; currently, most areas of the USA are represented by 30-m DEM, and some areas have 10-m DEM or higher resolution. Remotely sensed satellite imagery (RS) is obtained from remote or proximal sensors that record the electromagnetic reflection patterns of the Earth's surface (Lillesand and Kiefer, 2000) are also increasingly available. These DEM-derived and RS-derived data representing biophysical properties of the landscape and environment can be analyzed using various techniques.

Reviewing these advances, McBratney et al. (2003) and Scull et al. (2003) described the discipline of digital (predictive) soil mapping. McBratney et al. (2003) reframed Jenny's soil-forming factors into seven scorpan factors that describe the environment, where s = soil or soil properties), c = climate, o = organisms including vegetation and land cover, r = relief or topography, p = parent material, a = soil age (time), and n = spatial position. The goal is to represent all scorpan factors – environmental covariates – with spatially explicit digital data, assemble existing (legacy) and/or collect new georeferenced field and laboratory soil data, fit quantitative relationships, and predict the spatial distribution of soil classes and/or attributes. Digital soil mapping expects estimates of error to accompany the soil spatial predictions.

Public lands in the western USA are typically very diverse geologically, topographically, climatically, and ecologically, which facilitates the representation of the seven scorpan environmental covariates by a variety of DEM- and RS-derived data. Because there is little to no legacy soil data available for initial soil survey, the challenge of field sampling in vast and difficult-to-access areas remains. However, these environmental covariates can also be useful for stratifying study areas and developing objective field sampling plans that represent the environmental variability. The goal of this chapter is to demonstrate the potential usefulness of DEM- and RS-derived data to represent environmental covariates in the western USA.

2.2 Environmental Covariates Represented by Remotely Sensed Spectral Data: Organisms (o), Soil (s), and Parent material (p)

Publicly available and/or low-cost RS data are particularly useful for quantifying the spectral properties of organisms, soil, and/or parent material. This is especially true in arid to semiarid areas of the western USA where low vegetation cover exposes the surface of soils, sediments, and rocks, and where effective moisture can be represented by a range in vegetation density and type. While there are only seven bands from visible to short-wave infrared, spectral data from the Landsat Missions (Table 2.1; http://landsat.usgs.gov/) have commonly been used to represent organisms, soil, and parent material. With greater spatial and spectral resolution, spectral data from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) sensor (Table 2.1; http://asterweb.jpl.nasa.gov/) should become more common in future digital soil mapping studies.

The spectral properties of organisms, soil, and/or parent material can be exploited using simple or normalized difference band ratios (Jensen, 2005). Simple band ratios are ratios of the reflectance (represented as percent reflectance or a digital number) of one spectral band to another, with the equation taking the form of Band A/Band B. Normalized difference band ratios help reduce the effects of solar illumination intensity and angle, as well as topographic variations in slope and aspect. The equation takes the form (Band A–Band B)/(Band A+ Band B), which constrains the value of the ratio between -1 and $+1$.

Table 2.1 Spatial and spectral resolutions of Landsat and ASTER data used to represent environmental covariates in digital soil mapping projects in the western USA

Band	Spatial resolution	Spectral range	Common name
	m	μm	
Landsat			
1	30	0.450–0.515	Blue
2	30	0.525–0.605	Green
3	30	0.630–0.690	Red
4	30	0.775–0.900	NIR
5	30	1.550–1.750	SWIR (MIR)
6	60	10.40–12.50	TIR
7	30	2.090–2.350	SWIR (MIR)
Pan	15	0.520–0.900	Visible + NIR
ASTER			
1	15	0.520–0.600	Green
2	15	0.630–0.690	Red
3N	15	0.760–0.860	NIR (Nadir View)
3B	15	0.760–0.860	NIR (Backward Scan)
4	30	1.600–1.700	SWIR
5	30	2.145–2.185	SWIR
6	30	2.185–2.225	SWIR
7	30	2.235–2.285	SWIR
8	30	2.295–2.365	SWIR
9	30	2.630–2.430	SWIR
10	90	8.125–8.475	TIR
11	90	8.475–8.825	TIR
12	90	8.925–9.275	TIR
13	90	10.25–10.95	TIR
14	90	10.95–11.65	TIR

Abbreviations: NIR = near infrared; SWIR = short-wave infrared (formerly MIR = middle infrared, 1–3 μm; Lillesand and Kiefer, 2000); TIR = thermal infrared

2.2.1 Organisms (o)

The most obvious "organism" environmental covariate is vegetation, which is well represented by remotely sensed spectral data. Healthy green vegetation reflects near infrared (NIR; 0.7–1.1 μm) and absorbs visible electromagnetic radiation (0.4–0.7 μm), especially in the red region (e.g., Fig. 2.3). Dead or senescent vegetation reflect more in the visible region than healthy vegetation.

The Normalized Difference Vegetation Index (NDVI) is a normalized difference ratio model of the near infrared (NIR) and red bands of a multispectral image [NDVI = (NIR–RED)/(NIR + RED)]. A monochrome image of the NDVI illustrates areas that appear black which have no vegetation (e.g., water bodies, interstate highways), areas that appear white which have relatively high vegetation density (e.g., dense forest or meadow), and areas in gray shades representing intermediate vegetation density, with darker gray areas having lower vegetation cover and lighter gray areas having higher vegetation cover (e.g., Boettinger et al., 2008).

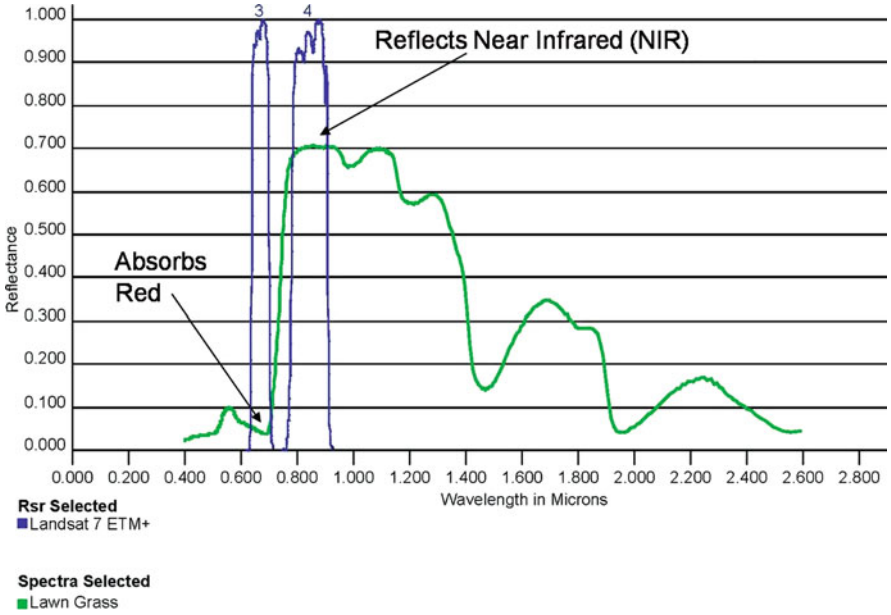


Fig. 2.3 Spectral reflectance of lawn grass (healthy green vegetation) viewed using the Landsat Missions' Spectral Characteristics Viewer (http://landsat.usgs.gov/tools_spectralViewer.php). The spectra illustrates the reflectance of near infrared radiation (NIR; Landsat 7 ETM+ band 4) and the absorbance of visible radiation, especially in the red region (Landsat 7 ETM+ band 3)

The NDVI is probably the most common representation of vegetation cover in digital soil mapping in the western USA (e.g., Sections 15.2.2 and 29.2.2). However, the soil-adjusted vegetation index (SAVI), which is similar to NDVI but includes a correction factor ("L") for soil, has also been used (e.g., Section 6.2.2). Tasseled cap and principal components analyses of Landsat data have also been useful covariates related to vegetation (see Chapter 3).

The "organism" environmental covariate in arid and semiarid regions may be represented by biological soil crusts. Biological soil crusts in the western USA are usually composed of cyanobacteria, lichen, and mosses. They can also include algae and other organisms, depending on parent material and climate (USGS Canyonlands Research Station, 2008). The cyanobacteria in biological soil crusts have distinct spectral properties. Most notably, the phycobilin pigment in cyanobacteria increases reflectance in the blue region (Karnieli et al., 1999). A normalized difference band ratio for cyanobacteria-dominated biological soil crust index (CI) was developed by Karnieli (1997) [CI = 1−(RED–BLUE)/(RED+BLUE)]. Our preliminary studies using Landsat spectral data in Canyonlands National Park in southern Utah indicate that the CI adds information about biological soil crust cover that appears potentially useful for digital soil mapping and assessment. The CI value is inversely related to NDVI, reflects soil properties in undisturbed sites, and indicates the degree of surface disturbance within a particular soil type. For lichen-dominated crusts, which are

more common on highly calcareous and gypsiferous soils in the western USA, Chen et al. (2005) developed biological soil crust index (BSCI), which exploits reflectance in the red, green and NIR bands.

The Advanced Spaceborne Thermal Emission and Reflectance Radiometer, or ASTER, on the Terra satellite platform has also been utilized to represent vegetation in the western USA. While there are only three spectral bands in the visible through NIR radiation, the spatial resolution is 15 m, which is finer than the 30-m resolution of Landsat spectral data (Table 2.1). Chapter 10 illustrates the use of ASTER to create a finer spatial resolution environmental covariate for vegetation in remote terrain of the Pacific Northwest.

2.2.2 Soil (s) and/or Parent Material (p)

The spectral properties of different mineral assemblages in parent materials (rocks, sediments) and soils can vary greatly. Figure 2.4 illustrates the spectra for calcite (calcium carbonate, $CaCO_3$), montmorillonite (an expandable 2:1 layer silicate clay), and hematite (iron oxide, Fe_2O_3). White- and light-colored minerals, such as calcite, have high reflectance in the visible portion of the electromagnetic spectrum. Some minerals, such as clay minerals, calcite, and gypsum are particularly responsive in the short-wave infrared (SWIR) range of the electromagnetic spectrum. Landsat band 5 and 7 and ASTER bands 4 through 9 have been quite useful for distinguishing variations in surface mineralogy of parent material and soil.

The Bureau of Land Management (BLM) developed three simple band ratios using Landsat data, known as the Soil Enhancement Ratios. Band 3 (Red) / Band 2

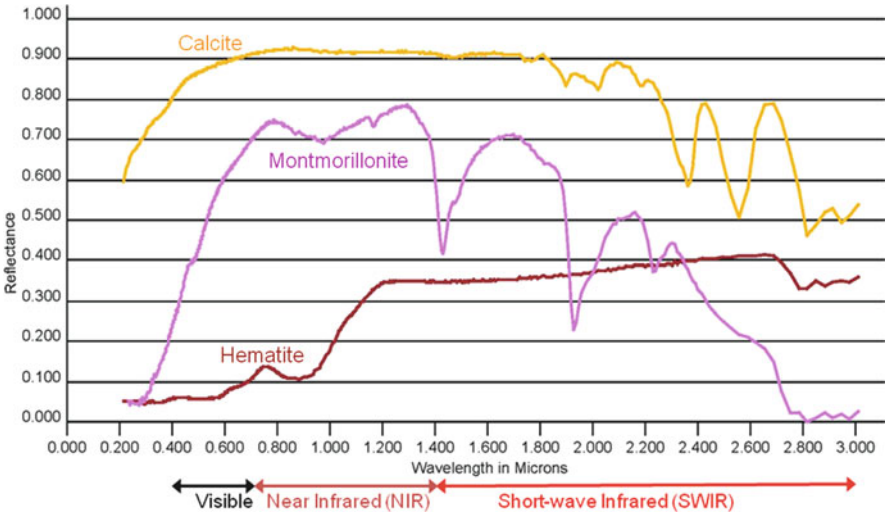


Fig. 2.4 Spectral characteristics of calcite ($CaCO_3$), montmorillonite (an expandable 2:1 layer silicate clay), and hematite (Fe_2O_3) viewed using the Landsat Missions' Spectral Characteristics Viewer (http://landsat.usgs.gov/tools_spectralViewer.php)

(Green) was thought to represent variations in carbonates, Band 3 (Red) / Band 7 (longer wavelength SWIR) was thought to represent iron, and Band 5 (shorter wavelength SWIR) / Band 7 (longer wavelength SWIR) was thought to represent hydroxyls of clays. While these exact physical relationships may not be valid for all study areas, these simple band ratios illustrate and help quantify mineralogical diversity of the soils and parent material on the landscape (e.g., Boettinger et al., 2008).

"Soils" can be represented by remotely sensed data. Chemical and/or physical properties of the soil surface or very near-surface may have diagnostic spectral properties. For example, in the arid climate of the San Rafael Swell in central Utah, there are soils with an accumulation of secondary gypsum within a few centimeters of the soil surface. Gypsum has diagnostic spectral response in the SWIR region, which is well represented by Landsat TM and ETM+ bands 5 and 7. A normalized difference ratio model of Landsat bands $(5 - 7)/(5 + 7)$ was used to successfully map the occurrence of gypsic soils in this area (Nield et al., 2007).

The "parent material" environmental covariate can be represented by remotely sensed spectral data. Chapter 15 illustrates the use of normalized difference band ratio of Landsat bands 5 and 2 to distinguish andesite from limestone in a Basin and Range landscape in southwestern Utah. In addition, normalized band ratios 4/5, 3/7, 5/1, and 4/7 distinguished unique soil-landscape-vegetation relationships thought to be useful in the random forests inference model. While rasterized vector coverages of bedrock and surficial geology may be useful covariates, spectral reflectance better quantifies the mineralogy and spatial position of the parent material. With the greater spectral resolution in both SWIR and TIR spectral regions, combinations and ratios of ASTER bands should be very useful for representing mineralogy of parent materials and soils in digital soil mapping. For example, calcitic, dolomitic, and quartzitic lithology, and rocks affected by contact metamorphism in southeastern California were differentiated using ASTER SWIR and TIR bands (Rowan and Mars, 2003). The ASTER mineral indexes (band combinations and ratios) developed by Geoscience Australia have been used to map mineralogy on a regional scale (Kalinowski and Oliver, 2004)

The spectral properties of materials of interest, especially minerals and rocks, can be explored using digital spectral libraries available on-line. These include the ASTER Spectral Library (http://speclib.jpl.nasa.gov/; Baldridge et al., 2009) and the USGS Spectroscopy Lab http://speclab.cr.usgs.gov/spectral-lib.html; Clark et al., 2007). For both the ASTER and USGS Spectral Libraries, one may view plots, which allows qualitative comparisons between spectra and band profiles in *ERDAS Imagine* or other image processing software. One may also view and download the spectral files with x, y wavelength and reflectance data, which can then be plotted against wavelength band intervals of sensors.

The use of proximal sensors and their spectral libraries (see Chapter 7) will likely be useful in scaling satellite and airborne imagery (e.g., HYMAP hyperspectral data; see Chapter 8) and aerial gamma radiometrics to field and laboratory measurements of soil properties as digital soil mapping progresses in the western USA.

2.3 Environmental Covariates Represented by Elevation Data: Relief (r) and Climate (c)

2.3.1 Relief (r)

The environmental covariate of "relief" can be represented by elevation, and can be determined from several remote sensing sources. These include photogrammetric, Shuttle Radar Topographic Mission (SRTM), Interferometric Synthetic Aperture Radar (IfSAR or InSAR), and Light Detection And Ranging (LiDAR). Most of the western USA has 30- and 10-m digital elevation models (DEM) available, and some areas have 5-m DEM.

Primary and secondary topographic derivatives from DEM data have been extremely useful for digital soil mapping in the high relief areas of the western USA. These include slope, curvature, and compound topographic index (also known as wetness index), etc. (e.g., Chapters 15 and 30). Chapter 3 briefly reviews commonly used topographic derivatives and introduces the use of 36 novel topographic attributes used to model soil depth in a western USA watershed, many of which represent landscape context including various representations of distance to ridge or stream.

2.3.2 Climate (c)

"Climate" is often difficult to quantify at the spatial scale needed in digital soil mapping in the western USA. Climate stations with long term climate records are widely spaced, and high relief and rugged terrain introduce orographic climate effect. While not very robust, elevation can help represent climate in remote areas subject to orographic climate effect (e.g., Chapter 15). Euclidean distance from the coast can also be useful to model climate for some locations (Howell and Kim, 2009).

Microclimate influenced by relief (slope gradient, aspect, etc.) can be quantified by estimating solar radiation, calculated from the georeferenced DEM (e.g., Chapter 30). Steep, north-facing slopes receive much less solar radiation than corresponding south-facing slopes in high relief, low-humidity environments typical throughout the western USA, resulting in lower evapotranspiration and higher effective moisture on north-facing slopes.

2.4 Soil Age (a)

Explicit and implicit estimates of soil age are very useful environmental covariates in digital soil mapping in landscapes of the western USA, but are often difficult to obtain. Chapter 4 reviews various approaches to add geochronology in digital soil mapping in southeastern Oregon. Improved prediction of soil on the landscape was obtained with addition of several implicit age covariates, including custom thematic maps of lithology and geologic age. In the absence of explicit soil age

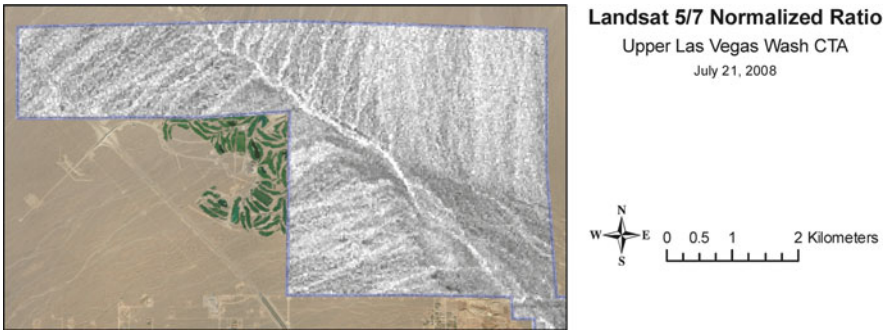


Fig. 2.5 The normalized difference ratio of Landsat bands 5 and 7 distinguishes recent channels from older surfaces on dissected alluvial fans emanating from the Spring Mountains in the southwest and the Sheep Mountains in the north to the basin floor near Las Vegas, southern Nevada. White areas indicate higher ratios and younger geomorphic surfaces, particularly evident as relatively active drainage channels

data, implicit topographic data representing landforms that are commonly related to soil relative age and remotely sensed spectral data that focuses on surface spectral characteristics represent surface or soil age. Chapter 15 illustrates the implicit use of landform surface age in predictive soil mapping by spatially modeling the shoreline of Pleistocene pluvial Lake Bonneville. Implicit soil age is represented by spectral reflectance focused on Fe-oxides in Chapter 4. Figure 2.5 illustrates the normalized difference ratio of Landsat bands 5 and 7, which distinguishes recent channels from older, stable surfaces on alluvial fans in southern Nevada.

2.5 Conclusions

In the western USA, the scorpan environmental covariates of organisms, parent material and soil, and possibly soil age, can be represented by remotely sensed spectral data, whereas properties related to relief or microclimate can be represented by topographic attributes derived from elevation. Explicit estimates of soil age would be very useful in digital soil mapping, but may not be available. While some covariates are routinely used, e.g., NDVI, novel covariates are being developed and tailored to specific study areas, e.g., new topographic attributes introduced in Chapter 3. The challenge of digital soil mapping of vast land areas in the western USA, particularly the remote and rugged public lands, is fertile ground for developing and utilizing a wide variety of environmental covariate data. Spatially explicit, digital environmental covariates have already improved several aspects of soil survey in the West including pre-mapping, developing efficient and targeted field sampling plans, and implementation of spatial prediction models (e.g., Chapters 28, 29, 30, 31, and 32).

Acknowledgments Sincere appreciation goes to my graduate students and colleagues who have advanced digital soil mapping in the western USA: Suzann Kienast-Brown, Colby Brungard, Alexander Stum, Shawn Nield, Amy Saunders, Nephi Cole, and Jedd Bodily. Special thanks to David Howell for his enthusiasm for digital soil mapping and helpful suggestions for improving this chapter.

References

Baldridge, A.M., Hook, S.J., Grove, C.I., and Rivera, G., 2009. The ASTER spectral library version 2.0. Remote Sensing of Environment 113:711–715.

Boettinger, J.L., Ramsey, R.D., Stum, A.K., Kienast-Brown, S., Nield, S.J., Saunders, A.M., Cole, N.J., and Bodily, J.M., 2008. Landsat spectral data for digital soil mapping, pp. 193–202. In: Hartemink, A.E., McBratney, A., and Mendonça-Santos, M.L. (eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Chen, J., Zhange, Y.M., Wang, L. Tamura, M., and Shimazki, H., 2005. A new index for mapping lichen dominated biological soil crusts in desert area. Remote Sensing of Environment 96:165–175.

Clark, R.N., Swayze, G.A., Wise, R., Livo, E., Hoefen, T., Kokaly, R., and Sutley, S.J., 2007. USGS digital spectral library splib06a: U.S. Geological Survey, Digital Data Series 231.

Howell, D.W., and Kim, Y.G., 2009. Soil temperature modeling for soil survey. Poster. Environmental Systems Research Institute User Conference, San Diego, CA, USA.

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

Jensen, J.R., 2005. Introductory Digital Image Processing. Prentice Hall, Upper Saddle River, NJ.

Kalinowski, A., and Oliver, S., 2004. ASTER Mineral Index Processing Manual. Remote Sensing Applications, Geoscience Australia. Available at http://www.ga.gov.au/image_cache/GA7833.pdf (Last verified 28 October 2009).

Karnieli, A., 1997. Development and implementation of spectral crust index over dune sands. International Journal of Remote Sensing 18:1207–1220.

Karnieli, A., Kidron, G.J., Glaesser, C., and Ben-Dor, E., 1999. Spectral characteristics of cyanobacteria soil crust in semiarid environments. Remote Sensing of Environment 69(1): 67–75.

Lillesand, T.M., and Kiefer, R.W., 2000. Remote Sensing and Image Interpretation. Wiley, New York, NY.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

Nield, S.J., Boettinger, J.L., and Ramsey, R.D., 2007. Digitally mapping gypsic and natric soil areas using Landsat ETM data. Soil Science Society of America Journal 71:245–252.

Rowan, L.C., and Mars, J.C., 2003. Lithologic mapping in the Mountain Pass, California area using Advanced Spaceborne Thermal, Emission and Reflection Radiometer (ASTER) data. Remote Sensing of Environment 84:350–366.

Scull, P., Franklin, J., Chadwick, O.A., and McArthur, D., 2003. Predictive soil mapping: a review. Progress in Physical Geography 27:171–197.

USGS Canyonlands Research Station, 2008. Biological soil crusts. http://www.soilcrust.org/index.htm (Last verified 13 October 2009).

Chapter 3
A Generalized Additive Soil Depth Model for a Mountainous Semi-Arid Watershed Based Upon Topographic and Land Cover Attributes

T.K. Tesfa, D.G. Tarboton, D.G. Chandler, and J.P. McNamara

Abstract Soil depth is an important input parameter in hydrological and ecological modeling. Presently, the soil depth data available in national soil databases (STATSGO, SSURGO) is provided as averages within generalized map units. Spatial uncertainty within these units limits their applicability for spatially distributed modeling. This work reports a statistical model for prediction of soil depth in a semi-arid mountainous watershed that is based upon topographic and other landscape attributes. Soil depth was surveyed by driving a rod into the ground until refusal at geo-referenced locations selected to represent the range of topographic and land cover variations in Dry Creek Experimental Watershed, Boise, Idaho, USA. The soil depth survey consisted of a model calibration set, measured at 819 locations over 8 sub-watersheds, and a model testing set, measured at 130 locations randomly distributed over the remainder of the watershed. Topographic attributes were derived from a Digital Elevation Model. Land cover attributes were derived from Landsat TM remote sensing images and high resolution aerial photographs. A Generalized Additive Model was developed to predict soil depth over the watershed from these attributes. This model explained about 50% of the soil depth spatial variation and is an important improvement towards solving the need in distributed modeling for distributed soil depth input data.

Keywords Generalized additive models · Explanatory variables · Land cover attributes · Soil depth · Topographic attributes

3.1 Introduction

Soil depth is one of the most important input parameters for hydrological and ecological models. Its spatial pattern, significantly affects soil moisture, runoff generation, and subsurface and groundwater flow (Freer et al., 2002; McNamara

T.K. Tesfa (✉)
Pacific Northwest National Laboratory, PO Box 999 Richland, WA 99352, USA
e-mail: Teklu.Tesfa@pnl.gov

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_3,

et al., 2005; Stieglitz et al., 2003). Consequently, its accurate representation is becoming increasingly important. It is highly variable spatially, and laborious, time-consuming and difficult to practically measure even for a modestly sized watershed (Dietrich et al., 1995). There is thus a need for models that can predict the spatial pattern of soil depth.

The national soil databases (SSURGO & STATSGO) have been the main sources of soil depth information used in hydrological and ecological modeling in the United States. In these soil databases, soils are spatially represented as discrete map units with sharp boundaries. A map unit may be comprised of more than one soil component but these components are not represented spatially within the map unit. As a result, soil attributes are spatially represented at map unit level as a mean or some other representative value of the components. Such a representation limits quantification of the variability of soil attributes within each class, and class boundaries generalize the spatial pattern of the soil properties, absorbing small scale variability into larger class units (Moore et al., 1993; Zhu, 1997). There is a need in spatially distributed modeling for fine scale models of soil depth that do not have these limitations. Past efforts to develop fine scale models include fuzzy logic, statistical and physically based approaches (Dietrich et al., 1995; Moore et al., 1993; Zhu, 1997).

In this chapter, we develop a statistical model for prediction of the spatial pattern of soil depth over complex terrain from topographic and land cover attributes in a mountainous semi arid watershed. Topographic and land cover attributes intended to have explanatory capability for soil depth were derived from a digital elevation model (DEM) and Landsat TM remote sensing images. A Generalized Additive Model (GAM) (Hastie and Tibshirani, 1990) was applied to predict soil depth based on these topographic and land cover attributes using soil depth data measured at 819 points at 8 sub-watersheds within Dry Creek Experimental Watershed (DCEW). This calibration data set was randomly divided into a training subset consisting of 75% of the data and a validation subset consisting of the remaining 25% that was used to estimate the prediction error for variable and model complexity selection (see Chapter 7). Soil depth data measured at an additional 130 random points within DCEW was used as an out of sample data set to test the model results. The Nash-Sutcliffe efficiency coefficient, which is widely used to assess the predictive accuracy of models, was used to evaluate the efficiency of the soil depth model.

3.2 Study Area

This study was carried out in the Dry Creek Experimental Watershed (DCEW), about 28 km^2 in area, located in the semi-arid southwestern region of Idaho, USA (Fig. 3.1). The area is composed of mountainous and foothills topography with elevations that range from 1,000 to 2,100 m (Williams et al., 2008). The landscape is typified by moderately steep slopes with average slope of about 25%, with steeper north facing slopes than south facing slopes, and is strongly dissected by streams.

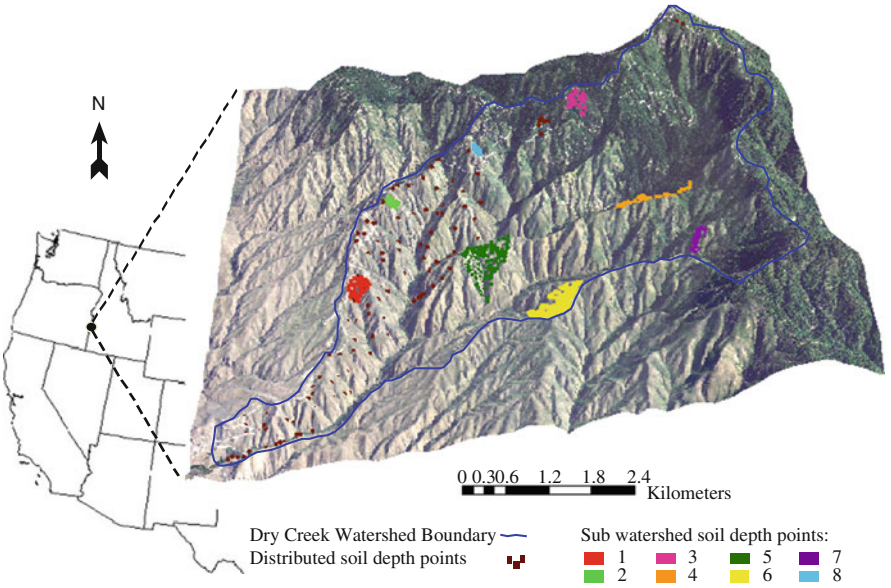


Fig. 3.1 Dry Creek Experimental Watershed (DCEW) near Boise, ID, in the Western USA. Points show locations where soil depth was sampled

The climate is a steppe summer dry climate at low elevation and moist continental climate with dry summers at high elevation (McNamara et al., 2005). Precipitation is highest in winter, as snow in the highlands and rain in the lowlands, and in spring in the form of rain. There are occasional summer thunderstorms. The average annual precipitation ranges from 37 cm at lower elevations to 57 cm at higher elevations (Williams, 2005). The average monthly temperatures are highest in July and lowest in January. Streamflow typically remains low in the early and mid winter and peaks in the early to mid spring due to snowmelt (McNamara et al., 2005).

Vegetation varies with elevation and landscape aspect (McNamara et al., 2005; Williams, 2005). Grass (south facing aspects) and sagebrush (north facing aspects) are dominant at lower elevations. Upper elevations are dominated by ponderosa pine (*Pinus ponderosa*) and Douglas-fir (*Pseudotsuga menziesii*) forest with patches of lodgepole pine (*Pinus contorta*) and aspen (*Populus tremuloides*). Middle elevations range from grass and shrublands to open forest of ponderosa pine and Douglas-fir.

Soils in this area are formed from weathering of the underlying Idaho Batholith, which is a granite intrusion ranging in age from 75 to 85 million years (Lewis et al., 1987; USDA, 1997). The dominant rock type is biotite granodiorite which consists of medium to coarse-grained rocks composed of plagioclase, quartz, potassium feldspar, and biotite (Johnson et al., 1988). The soils are classified into three general great groups according to US Soil Taxonomy: Argixerolls, Haploxerolls, and Haplocambids. These soils range from loam to sandy loam in texture and are generally well drained with high surface erosion potential (USDA, 1997). The Nat-

ural Resource Conservation's soil survey of the Boise Front (SSURGO soil database for survey area symbol ID903 obtained from Idaho NRCS office) provides a more detailed description of the soils underlying the watershed.

3.3 Methodology

3.3.1 Field and Digital Data

Eight sub-watersheds were selected to represent the elevation, slope, aspect and land cover variability present within the DCEW. Soil depth was surveyed at a total of 819 points within these sub-watersheds. Survey locations were chosen to represent the range of topographic and land cover variation in the sub-watersheds. At each survey location three depth replicates two to three meters apart were collected by driving a 220 cm long 1.27 cm diameter sharpened copper coated steel rod graduated at 5 cm interval into the ground using a fence post pounder until refusal. The survey was carried out in the early springs of 2005 and 2006, when the ground was relatively wet so that the rod penetrated more easily. The first author carried out this survey for 761 of the points in seven sub-watersheds, while soil depth data for 58 points in the eighth sub-watershed, had been previously collected using the same methods (Williams et al., 2008). The data from these 819 points are designated as the calibration dataset. A further 130 soil depth observations were collected using the same method at randomly distributed locations, at least 50 meters away from the selected sub-watersheds, over the remainder of the watershed. These are designated as the testing dataset (Fig. 3.1).

A wide range of topographic and land cover attributes were chosen as potential regression explanatory variables for the prediction of soil depth. Fifty five topographic variables (Table 3.1) were derived from the 1/3 arc second DEM obtained from the USGS seamless data server, which was projected to a 5 m resolution grid for the derivation of the topographic attributes. Of these, 36 were new topographic attributes that we derived following the approach described in Tarboton and Baker (2008). Ten land cover variables (Table 3.2) were derived from the Landsat TM imagery (path 41 row 30 obtained from the USGS) and an aerial photograph (obtained from NRCS Idaho State Office). Details on the derivation of these geospatial input variables are given in Tesfa et al. (2009).

3.3.2 Statistical Analysis

3.3.2.1 Normalization

Box Cox transformations (Equation (3.1)) were used to transform the measured soil depth (sd) and each explanatory variable so that their distribution was near normal.

Table 3.1 Topographic attributes derived from DEM; derivations and equations are in Tesfa et al. (2009)

Symbol	Description
elv**	Elevation above sea level
sca**	Specific catchment area from the D∞ method. This is contributing area divided by the grid cell size (from TauDEM[a] specific catchment area function)
plncurv**	Plan curvature is the curvature of the surface perpendicular to the direction of the maximum slope (From ArcGIS spatial analysis tools curvature function). A positive value indicates upwardly convex surface; a negative value indicates upwardly concave surface; and zero indicates flat surface
prfcurv	Profile curvature is the curvature of the surface in the direction of maximum slope (From ArcGIS spatial analyst tools curvature function) (Moore et al., 1993, 1991). A negative value indicates upwardly convex surface; a positive value indicates upwardly concave surface and zero indicates flat surface. See Table 29.1)
gncurv	The second derivative of the surface computed by fitting a fourth order polynomial equation to a 3×3 grid cell window (From ArcGIS spatial analyst tools curvature function) (Moore et al., 1993, 1991).
aspg	The direction that a topographic slope faces expressed in terms of degrees from the north (From ArcGIS spatial analyst tools aspect function).
slpg**	Magnitude of topographic slope computed using finite differences on a 3×3 grid cell window (From ArcGIS spatial analyst tools slope function).
ang**	The D∞ flow direction: This is the direction of the steepest outwards slope from the triangular facets centered on each grid cell and is reported as the angle in radians counter-clockwise from east (TauDEM Dinf Flow Directions function).
ad8	D8 Contributing Area: The number of grid cells draining through each grid cell using the single flow direction model (TauDEM D8 Contributing Area function)
sd8	The D8 slope: The steepest outwards slope from a grid cell to one of its eight neighbors reported as drop/distance, i.e. tan of the angle (TauDEM D8 Flow Directions function).
stdist	D8 Distance to Stream: Horizontal distance from each grid cell to a stream grid cell traced along D8 flow directions by moving until a stream grid cell as defined by the Stream Raster grid is encountered (TauDEM Flow Distance to Streams function).
Slpt	D∞ slope (Tarboton, 1997): The steepest outwards slope from the triangular facets centerd on each grid cell reported as drop/distance, i.e. tan of the slope angle (TauDEM Dinf Flow Directions function)
plen	D8 Longest Upslope Length: The length of the flow path from the furthest cell that drains to each cell along D8 flow directions. (TauDEM Grid Network Order and Flow Path Lengths function)
tlen	D8 Total Upslope Length: The total length of flow paths draining to each grid cell along D8 flow directions (TauDEM Grid Network Order and Flow Path Lengths function)
sd8a**	Slope averaged over a 100 m path traced downslope along D8 flow directions (from GRAIP[b], D8 slope with downslope averaging function)
p	The D8 flow direction grid representing the flow direction from each grid cell to one of its adjacent or diagonal neighbors, encoded as 1–8 counter-clockwise starting at east (TauDEM D8 Flow Directions function)
sar	Wetness index inverse: an index calculated as slope/specific catchment area (TauDEM wetness index inverse function)

Table 3.1 (continued)

Symbol	Description
sph8	D8 horizontal slope position
modcurv**	Curvature modeled based on field observed curvature using a regression equation on plan curvature, D8 horizontal slope position, wetness index inverse and general curvature, see Tesfa et al. (2009) for details
lhr*	Longest D∞ horizontal distance to ridge, see Tesfa et al. (2009) for details
shr*	Shortest D∞ horizontal distance to ridge, see Tesfa et al. (2009) for details
ahr*	Average D∞ horizontal distance to ridge, see Tesfa et al. (2009) for details
lhs*	Longest D∞ horizontal distance to stream, see Tesfa et al. (2009) for details
shs*	Shortest D∞ horizontal distance to stream, see Tesfa et al. (2009) for details
ahs*	Average D∞ horizontal distance to stream, see Tesfa et al. (2009) for details
lvr*	Longest D∞ vertical rise to ridge, see Tesfa et al. (2009) for details
svr*	Shortest D∞ vertical rise to ridge, see Tesfa et al. (2009) for details
avr**	Average vertical rise to ridge computed over multiple (D∞) paths from ridge to each point, see Tesfa et al. (2009) for details
lvs**	Longest vertical drop to stream computed over multiple (D∞) paths from point to stream, see Tesfa et al. (2009) for details
svs*	Shortest D∞ vertical drop to stream, see Tesfa et al. (2009) for details
avs*	Average D∞ vertical drop to stream, see Tesfa et al. (2009) for details
lsr*	Longest surface distance to ridge, see Tesfa et al. (2009) for details
ssr*	Shortest surface distance to ridge, see Tesfa et al. (2009) for details
asr*	Average surface distance to ridge, see Tesfa et al. (2009) for details
lss*	Longest surface distance to stream, see Tesfa et al. (2009) for details
sss*	Shortest surface distance to stream, see Tesfa et al. (2009) for details
ass*	Average surface distance to stream, see Tesfa et al. (2009) for details
lps*	Longest Pythagoras distance to stream, see Tesfa et al. (2009) for details
sps*	Shortest Pythagoras distance to stream, see Tesfa et al. (2009) for details
aps*	Average Pythagoras distance to stream, see Tesfa et al. (2009) for details
lpr*	Longest Pythagoras distance to ridge, see Tesfa et al. (2009) for details
spr*	Shortest Pythagoras distance to ridge, see Tesfa et al. (2009) for details
apr*	Average Pythagoras distance to ridge, see Tesfa et al. (2009) for details
lsph∞*	D∞ Longest horizontal slope position, see Tesfa et al. (2009) for details
ssph∞*	D∞ Shortest horizontal slope position, see Tesfa et al. (2009) for details
asph∞*	D∞ Average horizontal slope position, see Tesfa et al. (2009) for details
lspv**	Longest vertical slope position computed as longest vertical drop divided by the longest vertical drop plus longest vertical rise to ridge, see Tesfa et al. (2009) for details
sspv*	Shortest vertical slope position, see Tesfa et al. (2009) for details
aspv*	Average vertical slope position, see Tesfa et al. (2009) for details
lspp*	Longest Pythagoras slope position, see Tesfa et al. (2009) for details
sspp*	Shortest Pythagoras slope position, see Tesfa et al. (2009) for details
asp*	Average Pythagoras slope position, see Tesfa et al. (2009) for details
lspr*	Longest slope position ratio, see Tesfa et al. (2009) for details
sspr*	Shortest slope position ratio, see Tesfa et al. (2009) for details
aspr*	Average slope position ratio, see Tesfa et al. (2009) for details

* New topographic variables derived using enhanced terrain analysis

**Topographic variables selected for modeling soil depth

[a]TauDEM is the Terrain Analysis Using Digital Elevation Models software (http://www. engineering.usu.edu/dtarb/taudem)

[b]GRAIP is the Geomorphologic Road Analysis Inventory Package software (http://www. engineering.usu.edu/dtarb/graip)

Table 3.2 Landsat remote sensing image based data and their descriptions; equations are in Tesfa et al. (2009)

Symbol	Description
lc	Land cover map derived from Landsat TM image using supervised classification (this method is described in Section 10.2.4) in ERDAS IMAGINE. Land cover is represented as a numerical value encoded as follows: 1 Road, rock outcrop and bare, 2 Grass, 3 Mixed grass and shrub, 4 Shrub, riparian and deciduous forest, 5 Coniferous forest
*pc1***	First principal component from ERDAS IMAGINE principal component analysis of Landsat Thematic Mapper bands 1, 2, 3, 4, 5, and 7
pc2	Second principal component derived from principal component transformation of Landsat TM image in ERDAS IMAGINE (Jensen, 1996)
pc3	Third principal component derived from principal component transformation of Landsat TM image in ERDAS IMAGINE (Jensen, 1996)
tc1	First tasseled cap component derived from tasseled cap transformation of Landsat TM image in ERDAS IMAGINE (represents brightness)
tc2	Second tasseled cap component derived from tasseled cap transformation of Landsat TM image in ERDAS IMAGINE (represents greenness)
tc3	Third tasseled cap component derived from tasseled cap transformation of Landsat TM image in ERDAS IMAGINE (represents wetness)
ndvi	Normalized difference vegetation index calculated in ERDAS IMAGINE (Jensen, 1996) (see Table 29.1 and Section 20.2.3)
vi	Vegetation index calculated in ERDAS IMAGINE (Jensen, 1996)
cc	Canopy cover index calculated in ERDAS IMAGINE (Zhu and Band, 1994)

** Land cover variables selected for modeling soil depth

$$t(x) = \frac{(x^{\lambda} - 1)}{\lambda} \tag{3.1}$$

Here, $t(x)$ denotes the transform of variable x with transformation parameter λ. λ was selected to maximize the Shapiro-Wilks Normality Test W-statistic as implemented in R (R Development Core Team, 2007).

3.3.2.2 Model

We applied Generalized Additive Models (GAM) (Hastie and Tibshirani, 1990) to predict soil depth using the explanatory variables. GAM is a statistical approach that generalizes multiple regression by replacing linear combinations of the explanatory variables with combinations of nonparamtertic smoothing or fitting functions, estimated through a backfitting algorithm. The GAM model is:

$$E(sd|x_1, x_2, \ldots, x_p) = \alpha + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) \tag{3.2}$$

where, $x_1, x_2, \ldots, x_p$ are explanatory variables (predictors), sd is soil depth (response variable) and f_i are non-parametric smoothing splines that relate sd to the $x_1, x_2, \ldots, x_p$. The model assumes that the mean of sd is an additive combination of nonlinear functions of the explanatory variables $x_1, x_2, \ldots, x_p$. We used the GAM package as implemented in R (R Development Core Team, 2007).

3.3.2.3 Variable Selection and Model Complexity

Questions in developing a predictive regression model include which potential explanatory variables to use and what to do about interdependent explanatory variables. Many of the explanatory variables that we derived from the DEM (Table 3.1) were variants on similar quantities, so we were specifically concerned about the effect of explanatory variable correlation on model prediction error. A correlation matrix giving the cross correlation between all 65 explanatory variables was computed using all 819 data points in the calibration dataset. Random Forest (Breiman, 2001), a classification and regression package (this is described in Section 15.2.3) in R (R Development Core Team, 2007), was used to calculate a measure of explanatory variable importance (see Section 29.2.3.2) for the prediction of soil depth. Due to randomness in the Random Forest method the variable importance varies slightly each time it is run. We therefore ran Random Forest 50 times using all 819 data points in the calibration dataset with all 65 potential explanatory variables with soil depth as the response variable and averaged variable importance across these runs. Explanatory variables were then ordered based upon their importance measures.

The number of explanatory variables in a model is a measure of model complexity. We used the correlation matrix, together with the Random Forest importance values to develop sets of explanatory variables representing models of differing complexity by eliminating the variable of lesser importance from pairs of variables with correlation above a designated threshold (from 0.15 to 0.9 in increments of 0.05). Variables were filtered out working sequentially from high to low correlation until no pairs with correlation greater than the threshold remained. Lower thresholds result in fewer variables, so a range of models with differing complexity were developed. This approach reduced the correlation between variables selected for inclusion in a model. Models of differing complexity were also constructed using explanatory variables directly from the variable list ordered by importance. Figure 3.2 shows the explanatory variables with importance values greater than or equal to 0.009, ordered based on their average importance values from 50 RF runs with all 819 calibration data points and all 65 explanatory variables.

To evaluate appropriate model complexity, we randomly split the calibration sample of 819 data points into two parts, designated as the training and validation sets. The separate testing dataset of 130 points randomly distributed across the watershed was withheld from this process, so that it could be used for evaluation of the final model. GAM was applied, using the training data set of 614 data points to fit the models. Prediction error was computed for both the training and validation data set. The validation data set prediction error provides an out of sample estimate appropriate for trading off variance due to complexity with bias due to too few explanatory variables (see e.g. Hastie et al., 2001). The results from this analysis allowed us to select the explanatory variables and degree of model complexity.

3.3.2.4 Calibration and Testing

Once the explanatory variables and model with appropriate complexity had been selected, GAM was applied using the full calibration data set as input. It was used

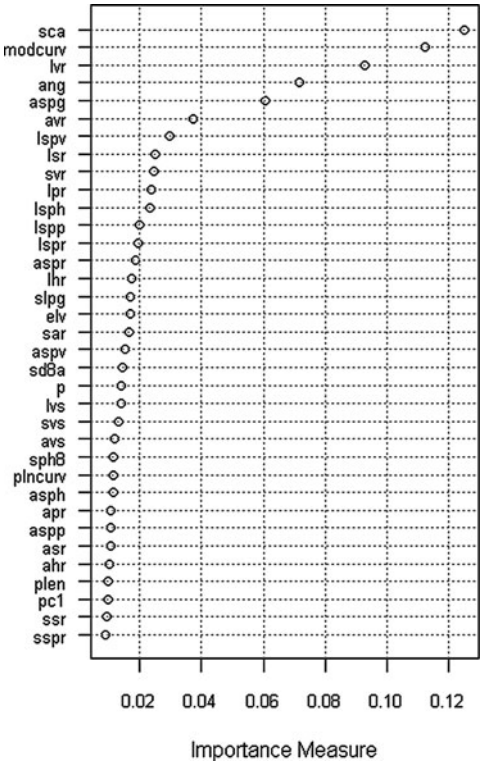


Fig. 3.2 Variable importance measure of the Box Cox transformed explanatory variables averaged from 50 RF model runs

to predict soil depth for the entire watershed. We then compared the testing dataset with the GAM soil depth values at testing locations using the Nash-Sutcliffe efficiency coefficient (NSE), which is a measure widely used to quantitatively assess the predictive accuracy of a model.

$$\mathrm{NSE} = 1 - \frac{\sum(\mathrm{SD_o} - \mathrm{SD_p})^2}{\sum(\mathrm{SD_o} - \mathrm{SD_m})^2} \tag{3.3}$$

where; $\mathrm{SD_o}$, $\mathrm{SD_p}$, and $\mathrm{SD_m}$ are observed (measured), predicted, and mean of observed (measured) soil depths respectively.

3.4 Results and Discussion

3.4.1 Variable Selection and Model Complexity

Figure 3.3 shows the variation of mean square prediction error for training and validation datasets versus model complexity in terms of the number of input variables. The continuous lines in this figure are from models developed using explanatory variables selected based on Random Forest importance directly. There is a new GAM model for each additional input variable. The symbols in this figure are from

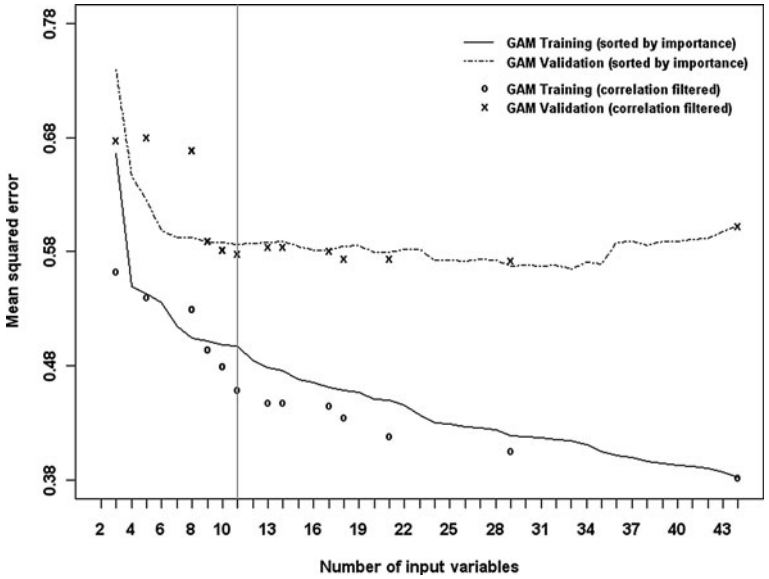


Fig. 3.3 Number of input variables (Model complexity) vs. mean squared error. Explanatory variables selected directly using importance (*continuous*) and filtered by correlation (*symbols*)

models developed using cross correlation as a filter to reduce inter-dependence among explanatory variables. There is a new GAM model with different number of input variables for each correlation threshold. Figure 3.3 reports training and validation errors separately.

For both the importance-selected and correlation-filtered models, the training error decreases progressively as additional input variables are added while the validation error decreases initially and then flattens out and starts to increase. The use of correlation-filtered explanatory variables resulted in lower error. The validation error starts to increase for complexity more than 11 correlation-filtered variables (Fig. 3.3). Although there are fluctuations on validation MSE that go slightly below the 11 variable complexity, for 18 and 21 input variables, in our judgment the point of diminishing returns has been reached at 11 input variables. Consequently we selected 11 correlation-filtered explanatory variables as representing the optimum GAM complexity for this dataset. Tables 3.1 and 3.2 list all the topographic and land cover explanatory variables derived for modeling soil depth. Variables derived using new DEM analysis methods are identified with single asterisk (*) and variables selected by this variable selection procedure are identified by double asterisks (**). Ten of the 11 selected explanatory variables are topographic variables, with three (avr, lspv, lvs), variables derived using the new DEM analysis methods.

3.4.2 Model Evaluation

Based on the selection of 11 correlation-filtered explanatory variables above, GAM was applied to the full calibration set of 819 data points. Figure 3.4 shows the scatter

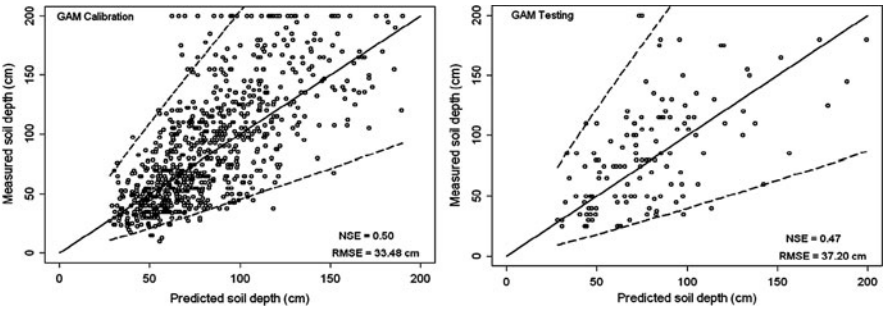


Fig. 3.4 Predicted soil depth vs. measured soil depth with plus and minus two standard error for calibration (*left*) and testing (*right*) data

plots of predicted versus measured soil depth for the calibration (left) and testing (right) data and their Nash-Sutcliffe Efficiency (NSE) and root mean squared errors (RMSE) after transforming back into space of soil depth. The testing data was not used at all in model development. In this figure, the diagonal (central) lines represent the 1:1 line (predicted = observed). The two diverging dash lines, above and below the 1:1 line, show the predicted soil depth plus and minus two standard errors representing 95% confidence intervals. These lines diverge as a result of the Box-Cox back transformation (Figs. 3.4 and 3.5).

Figure 3.5 shows the soil depth map created using GAM at 5 meter grid scale which improves the scale of soil depth representation as compared to the map unit based soil depth maps that can be created using conventional soil survey approach

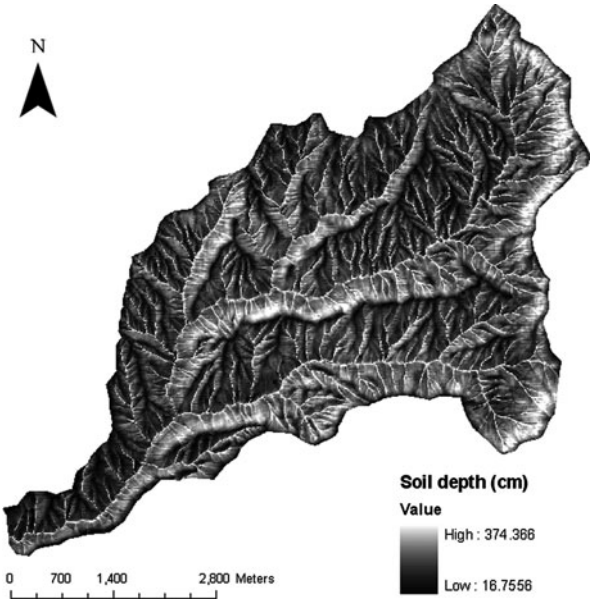


Fig. 3.5 Soil depth map predicted using GAM model

(see Sections 29.2.3.1 and 29.3.2). This models the ridges (convex areas) and south facing slopes as having shallower soils compared to the valleys (concave areas) and the north facing slopes respectively. This agrees with existing literature (e.g. Dietrich et al., 1995). As compared to soil depth maps created using conventional.

3.5 Conclusions

A statistical model has been developed that predicts soil depth using topographic and land cover attributes. The topographic attributes were found to be more important than the land cover attributes in predicting the soil depth. The model was able to explain about 50% of the measured soil depth variability in an out of sample test. New topographic variables derived from the DEM played an important role in this model. Considering the uncontrolled uncertainties due to the complex local variation of soil depth, DEM errors and GPS reading errors, this is considered an important improvement towards solving the need for distributed soil depth information in distributed hydrological and ecological modeling.

Acknowledgments The data collection part of this research was funded by the Utah Drought Management project, a project of USDA-CSREES Special Research Grant no. 2005-34552-15828 from the USDA Cooperative State Research, Education, and Extension Service. The modeling part was funded by the Inland Northwest Research Alliance (INRA). Field assistance was provided by Pam Aishlin and other students at Boise State University.

References

Breiman, L., 2001. Random forests. Machine Learning, 45:5–32.

Dietrich, W. E., Reiss, R., Hsu, M.-L., and Montgomery, D. R., 1995. A process-based model for colluvial soil depth and shallow landsliding using digital elevation data. Hydrological Processes, 9:383–400.

Freer, J., McDonnell, J. J., Beven, K. J., Peters, N. E., Burns, D. A., Hooper, R. P., Aulenbach, B., and Kendall, C., 2002. The role of bedrock topography on subsurface storm flow. Water Resources Research, 38(12):1269–1285.

Hastie, T., and Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.

Hastie, T., Tibshirani, R., and Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 533 pp., Springer, New York.

Jensen, J. R., 1996. Introductory to Digital Image Processing: A Remote Sensing Perspective, 2 ed., Prentice-Hall, Englewood Cliffs, NJ.

Johnson, K.M., Lewis, R.S., Bennet, E.H., and Kiilsgaard, T.H., 1988. Cretaceous and tertiary intrusive rocks of south-central Idaho, pp. 55–86. In: Link, P.K., and Hackett, W.R. (eds.), Guidebook to the Geology of Central and Southern Idaho: Idaho Geological Survey, Bulletin 27, University of Idaho, Moscow, Idaho.

Lewis, R.S., Kiilsgaard, T.H., Bennet, E.H., and Hall, W.H., 1987. Lithologic and chemical characteristics of the central and southeastern part of the southern lobe of the Idaho Batholith, pp. 171–196. In: Vallier, T.L., Brooks, H.C. (eds), Geology of the Blue Mountains region of Oregon, Idaho, and Washington – the Idaho Batholith and Its Border Zone: U.S. Geological Survey Professional Paper 1436.

McNamara, J.P., Chandler, D., Seyfried, M., and Achet, S., 2005. Soil moisture states, lateral flow, and streamflow generation in a semi-arid, snowmelt-driven catchment. Hydrological Processes 19(20):4023–4038.

Moore, I.D., Gessler, P.E., Nielsen, G.A., and Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Science Society of America Journal 57(2):443–452.

Moore, I.D., Grayson, R.B., and Ladson, A.R., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrological Processes 5(1):3–30.

R Development Core Team, 2007. R: A Language and Environment for Statistical Computing, edited, R Foundation for Statistical Computing, Vienna, Austria.

Stieglitz, M., Shaman, J., McNamara, J., Engel, V., Shanley, J., and Kling, G.W., 2003. An approach to understanding hydrologic connectivity on the hillslope and the implications for nutrient transport. Global Biogeochemical Cycles 17(4):1105–1120.

Tarboton, D.G., and Baker, M.E., 2008. Towards an algebra for terrain-based flow analysis, pp. 167–194. In: Mount, N.J., Harvey, G.L., Aplin, P., and Priestnall, G., (eds.), Representing, Modeling and Visualizing the Natural Environment: Innovations in GIS 13. CRC Press, Boca Raton, FL.

Tesfa, T.K., Tarboton, D.G., Chandler, D.G., and McNamara, J.P., 2009. Modeling soil depth from topographic and land cover attributes, Water Resources Research 45:W10438, doi: 10.1029/2008WR007474.

USDA, 1997. Soil survey of the Boise Front Project, Idaho: Interim and supplemental report, Boise, Idaho.

Williams, C.J., 2005, Characterization of the spatial and temporal controls on soil moisture and streamflow generation in a semi-arid headwater catchment, Masters Thesis, Boise State University.

Williams, C.J., McNamara, J.P., and Chandler, D.G., 2008. Controls on the temporal and spatial variability of soil moisture in a mountainous landscape: the signatures of snow and complex terrain. Hydrology and Earth System Sciences 5(4):1927–1966.

Zhu, A.X., 1997. A similarity model for representing soil spatial information. Geoderma 77(2–4):217–242.

Zhu, A.X., and Band, L.E., 1994. A knowledge-based approach to data integration for soil mapping. Canadian Journal of Remote Sensing 20:408–418.

Chapter 4
Applying Geochronology in Predictive Digital Mapping of Soils

Jay Stratton Noller

Abstract Explicitly adding time (geochronology) to the rubric of digital soil mapping enhances the results and accuracy of predictive maps for initial and update soil surveys. Experiments were performed on a data set used in the prediction of soils in the initial Malheur County soil survey, Oregon. Geochronological information was derived from (1) independently compiled Quaternary geological maps, (2) age point data, and (3) remotely sensed data. These data were incorporated in decision-tree analysis as area-class data in raster format. Experimental area consists largely of fluvial, lacustrine and/or volcanic materials and landforms. Expert soil survey maps are used as reference in making predictions with and without implicit or explicit age information. Addition of geochronological data produces predictive maps that are most closely aligned with expert maps. Improvements afforded by predictive soils mapping may be of greater magnitude in areas where the age stratification of the landscape is not obvious to the soil surveyor. Layers representing palaeogeographical changes in surficial processes and palaeoclimatological events may provide further enhancements.

Keywords Age · Dating · Geochronology · Decision-tree analysis · Volcanic soilscape

4.1 Introduction

Time is a widely recognized factor in conventional soil survey, typically documented as a descriptor of parent material (Soil Survey Staff, 1993), and is considered as the *t* factor of soil formation (Jenny, 1941). In digital soil mapping, which is seen as the next mode of global soil resource inventory and analysis (Lagacherie et al., 2007), this factor is expressed as *a* (for *age*) in the formulation of $\mathrm{Sc} = f(s, c, o, r, p, a, n)$ by McBratney et al. (2003).

J.S. Noller (✉)
Department of Crop & Soil Science, Oregon State University, ALS3017,
Corvallis, OR 97331, USA
e-mail: jay.noller@oregonstate.edu

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_4,

Age (a) is used by digital soil mapping in *implicit* or *explicit* form. Age has been used as a co-variant in digital soil mapping (e.g., Scull et al., 2005; see Section 13.2.4.5), but such explicit use is rare. The norm to date in digital soil mapping has been the implicit form carried in the age of parent material (p) and landform (r). In nearly all cases around the globe, parent material information is carried into digital soil mapping analysis as class data derived from geological and geomorphological (thematic) maps (e.g., Bui and Moran, 2001; see Section 13.2.4.4). These data are represented in GIS as polygon vector or raster representation, with age information conveyed as geological time-series names. From Precambrian to Holocene, the spectrum of time-series names generally connotes the age range in earth materials from ancient, hard and difficult-to-weather crystalline rocks to recent, soft and weatherable unlithified sediments.

Explicit forms of age information could be represented in GIS in the form of point observations or continuous rasters collected by a remote-sensing method. Point data would be derived from field-sampled soil profiles. Many of the known geochronological methods, particularly those yielding numerical age results (the most sought-after, high confidence type) (Table 4.1), could be used in this regard (Table 4.1). Spatially continuous observations of soil age, including proxies, from remote sensors would provide explicit measures of properties related to age. Currently, age is not directly determined from such sensor data without the use of a calibration data set, such as a chronosequence (e.g., Kahle et al., 1988). And so the best results for the continuous data set are a calibrated age result (Table 4.1) (Noller et al., 2000).

Age information for soil studies is not just a measure of time reported in years. It is important to consider the source of material that was dated, context (or stratigraphy) of the sample, method applied, type of result, and community confidence. These topics are considered in depth by Noller et al. (2000). A brief overview of key points follows.

The 35+ known and applied methods of Quaternary geochronology classify into six types based on (1) principles and laws, (2) biological, chemical and physical processes, (3) accumulated or complexes of results of processes, and (4) logical arguments (Noller et al., 2000) (Table 4.1). These six types of methods yield age estimates that are classified into four types of results (Table 4.1). At least eight of the dating methods would be amenable to remote detection (Table 4.1) and hence of considerable promise and use in digital soil mapping. Some key questions should be asked before considering using age in a digital soil mapping project. Is it appropriate, within the study context, to use age as environmental covariate? What type of temporal information is available and can it be used?

Digital soil mapping as a field of study developed and is practiced with considerable success in the absence of purposive use of age. In digital soil mapping study, is it appropriate, within the study context, to use age as an environmental covariate? If we look at the successes of digital soil mapping around the globe, one might be inclined to say it is not. Or is it that age has been used in implicit form, contributing to such successes without operator knowledge?

Table 4.1 Classification of Quaternary geochronologic methods and their application to digital soil mapping

Type of results

Numerical-age [a]

Calibrated-age [a]

Relative-age [a]

Correlated-age [a]

Type of method

Sidereal	Isotopic	Radiogenic	Chemical and biological	Geomorphic	Correlation
Dendrochronology[b]	Radiocarbon[c]	Fission track	Amino-acid racemization	***Soil-profile development***[c]	Stratigraphy[c]
Sclero-chronology and annual growth in other organisms (e.g. mollusks)	Cosmogenic isotopes[c] ^{36}Cl, ^{10}Be, ^{26}Al, ^{14}C, ^{3}He, and others[b]	Thermoluminescence	***Obsidian hydration and tephra hydration***[c]	***Rock and mineral weathering***[c]	Paleomagnetism
Varve chronology	K-Ar and ^{39}Ar-^{40}Ar[c]	Optically stimulated luminescence	Rock-varnish cation ratio	***Scarp morphology and other progressive landform modification***[c]	Tephrochronology
Historical records	Uranium-series	Infrared stimulated luminescence	***Lichenometry***[c]	***Rock-varnish development***[c]	Paleontology
	U-Pb, Th-Pb	Electron-spin resonance	***Soil chemistry***[c]	Rate of deposition	Archaeology
			^{10}Be accumulation in soils[c]	Rate of deformation	Astronomical correlation
				Geomorphic position	Tectites and microtectites
				Stone coatings ($CaCO_3$)	

[a] Triple-dashed line indicated the type of result most commonly produced by the methods below it; single-dashed line indicated the type of result less commonly produced by the methods below it

[b] ***Bold-italicized*** methods are particularly useful in digital soil mapping

[c] Methods used in this paper

Table after Noller et al. (2000)

4.2 Materials and Methods

4.2.1 Study Area

To examine the potential roles of age in digital soil mapping, this paper focuses on the initial soil survey of Malheur County, Oregon, with which my students and I have been engaged since 2005. The official soil survey began during 2006 following our development of digital soil mapping products to be used as pre-map layers by field survey crews. We sought to try several approaches to digital soil mapping with experiments focused on the Coyote Lake Basin and Jordan Valley area (Fig. 4.1). Experiments on age in digital soil mapping were run in the Jordan Valley Volcanic Field (JVVF), where c. 15 million years of lava flows and eruptive products have shaped a constructional landscape crossed by rivers draining the northern rim of the Great Basin of North America (Fig. 4.1).

The common landforms in the area consist of alluvial fans, fan remnants, pediments, playas, dune fields, and volcanic tablelands, cones and flows. Most of the soils in the study area are developed on volcanic parent materials (basalt and andesite, tuffaceous sedimentary rocks, welded tuffs), lacustrine deposits and fluvial sedimentary rocks. Elevation varies from 1,100 m to 2,100 m. Soil moisture regime is mainly aridic, with mean annual precipitation of 150 mm in low areas to 840 mm on high elevations. Soil temperature regime is mesic to frigid, with surface air temperatures ranging from monthly mean low temperature of −8°C (January) to monthly mean maximum of 33°C (August).

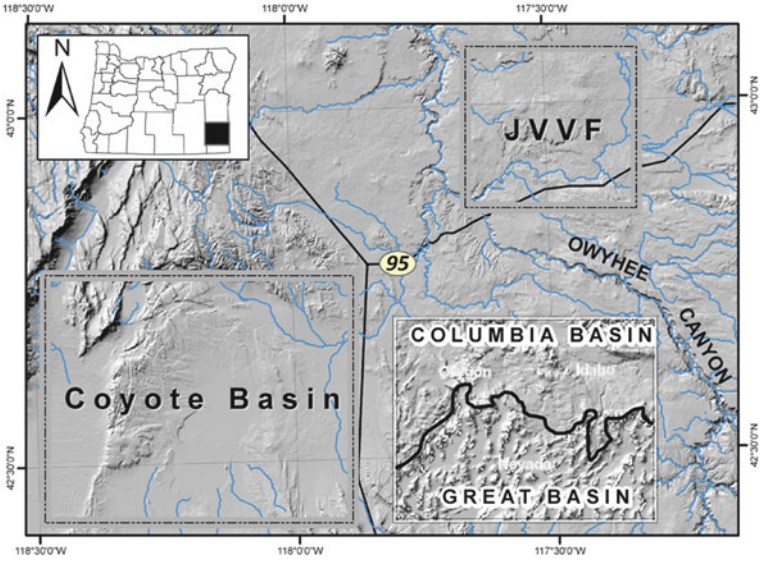


Fig. 4.1 Location of Malheur County, Oregon (*upper left inset*) study areas at Coyote Basin and Jordan Valley Volcanic Field (JVVF) (*outlined*), astride the northern rim of the Great Basin (*lower right inset*)

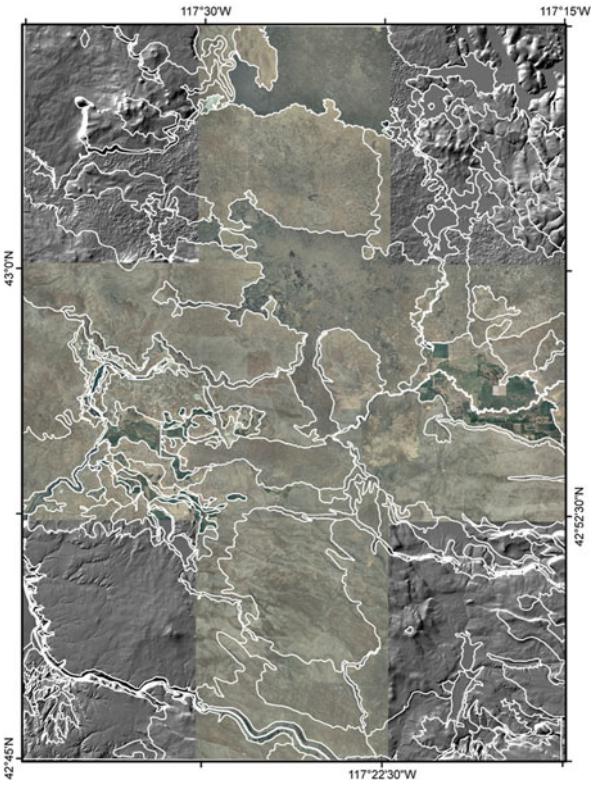


Fig. 4.2 The cross-pattern of the Three Mile Hill area quadrangles (shown in *color* NAIP imagery) is focus of age experiments. Surrounding quads are shown with hillshaded 5 m DTM. Soil Survey soil map units (unlabelled) are outlined in *white*

Aridisols is the most dominant soil order in the studied area, however there are some areas with weakly to slightly developed soils (Entisols and Inceptisols) and well-developed soils (Mollisols). Nearly all soils have andic properties, principally due to influx of ash from the many volcanoes located upwind in the Cascade Range.

Soil Survey in the study area is conducted using the map base of 1:24,000 scale topographic quadrangle maps (quads), and thus experiments were designed to conform to this scale and area of interest. This study focuses on five quads in the JVVF (Fig. 4.2), subject of other digital soil mapping experiments (Hash, 2008; Hash and Noller, 2009). Results of experiments were delivered to the NRCS for use in the Malheur County soil survey.

4.2.2 Geochronology

Estimates of age used in this study were collected from the literature, derived using established methods and handled and reported using geochronologic community protocols (Noller et al., 2000). Methods used in this study include ^{39}Ar-^{40}Ar geochronology of basalt rocks, radiocarbon (^{14}C) dating on organic matter in lake sediments, weathering of lava, and geomorphic expression of landforms. Argon and

radiocarbon geochronology data were collected from the literature (Bondre, 2006; Hart and Mertzman, 1983) and their geographic locations were inducted as a point-data layer in ArcGIS. Weathering of basalt lava flows yields a number of time-dependent characteristics, including changes in surface (glass) reflectance (Kahle et al., 1988), collapse and smoothing of flow microtopography (Farr, 1992), lichen growth (Stretch and Viles, 2002), and pedogenesis (Vaughan, 2008). Reflectance of exposed basalt surfaces is observed to be related to flow age (Noller, 2008) and is attributed to changes in Fe chemistry in the basalt surface layer. Such changes are known to occur elsewhere on basalt surfaces, with development of Fe oxides being an important weathering product (Jackson and Keller, 1970; Kahle et al., 1988; Wasklewicz, 1994). For this study, the reflectance of iron oxide is revealed through band 3/band 1 ratio of a Landsat ETM+ image (for this and other bands of similar application see Chapters 1 and 30).

Surficial geology of the study area was mapped at 1:24,000 scale (Noller, 2008), and the polygon vector data was attributed with an estimate of deposit/flow/landform age based on point geochronology and surface relative-dating techniques using field techniques, aerial photo interpretation, and IFSAR-derived 5 m DTM. Cross-cutting relations, vertical separation and degree of pattern ground features were the key criteria used to assess relative surface (and by extension *soil*) age. Lithology and age of surficial geologic map units used in this study are presented in Table 4.2.

Table 4.2 Geologic ages assigned to surficial geologic map units in the Jordan Valley area

Years B.P. (Thousands)	Geological period/epoch	Symbol
0	Active	A
0–4	Late Holocene	Hl
4–7.5	Middle Holocene	Hm
7.5–10	Early Holocene	He
10–15	Latest Pleistocene	Pll
10–125	Late Pleistocene	Pl
0–125	Late Pleistocene	Ql
125–700	Middle Pleistocene	Pm
700–1800	Early Pleistocene	Pe
0–1800	Quaternary	Q
1800–5000	Pliocene	P
700–5000	Plio-Pleistocene	Pp
1800+	Tertiary	T
1500–5000+	Tertiary-Quaternary	TQ

4.2.3 Experimental Setups

Experiments on age in digital soil mapping were run using decision-tree analysis (See5 – www.rulequest.com; Imagine 9.1 – gi.leica-geosystems.com) following the reference area approach (Lagacherie and Voltz, 2000; Lagacherie et al., 1995; Scull

et al., 2005) (see approach descriptions in Chapters 13, 14 and 34). Building on the decision-tree analysis work of Hash (2008) and Hash and Noller (2009) in the area, experiments were designed considering both implicit and explicit modes of age information exist for use in creating predictive digital soil mapping maps. Between these modes, a number of submodes were selected as experimental constructs: age implicit in remote sensing data; age implicit in lithological (thematic) map; age explicit in geochronological (thematic) map; and age explicit in geological (thematic) map. A control experiment was run without any dependent data layers that would directly carry geological information.

4.2.4 Age of the Soilscape

Establishing the age of soil is not an easy, straight-forward enterprise. Soils develop over a considerable amount of time and these biological, chemical and physical changes are not considered to be as "instantaneous" as the formation of its constituent minerals and/or organismal tissue. Most strategies to date soils focus on estimating the age of origin for its parent materials. The use of duration of pedogenesis, or rather maturity of soil profile, is a widely held concept in soil survey (Soil Survey Staff, 1993).

4.3 Results

Five experimental setups using decision-tree analysis in digital soil mapping yielded an array of results on the impact of the absence, presence and type of age information included with the environmental covariate data (Table 4.3). The base setup, experiment 1, was run with eight input layers (first eight listed in Table 4.3), without lithology and age information. Environmental data that might strongly carry this sort of information, e.g., Landsat imagery bands, were excluded from all runs. The experiments are numbered in order of the addition of increasingly more specific (explicit) age information. In experiments 2 through 5, only one layer of information is added to the base for a total of nine.

Absence of age information is taken here to mean geological map units classed according to lithology but not age. Landform map units are not differentiated on the basis of activity, preservation of original topography, and other age implicit or explicit forms. Assessment of accuracy of prediction without geology or soil age information in experiment 1 yielded the lowest values of the experiments (OA = 78%, Khat = 0.84) (Table 4.4). Traditional addition of a parent material or geology (lithology) layer in experiment 2 yielded improved results (OA = 82%, Khat = 0.88). In experiment 3, the addition of Fe-oxide reflectance (b1b3) layer yielded (OA = 79%, Khat = 0.84) values essentially unchanged from the experiment without geology and age. The lack of change between experiments 1 and 3 may be that while b1b3 is directly tracking the age of exposed lavas, it is also tracking the

Table 4.3 Environmental variables used to develop predictive models

Independent variable	Symbol	Cell size (m)	Source
Topography			
Slope		30	Derived from 10 m DEM
Aspect		30	Derived from 10 m DEM
Climate			
Mean annual temperature	MAT	800	PRISM climate data[a]
Mean annual precipitation	MAP	800	PRISM climate data
Mean Jan min. temp.	Tmin	800	PRISM climate data
Mean Jul max. temp.	Tmax	800	PRISM climate data
Vegetation			
Tasseled cap transformation			
Wetness index	wi	30	Landsat TM, acquired 7/5/1989
Normal. vegetation index	NDVI	30	Landsat TM, acquired 7/5/1989
Parent material, time			
Lithology	L	30	Original work this project
Geologic age	a	30	Original work this project
Geology (L + a)	La	30	Original work this project
Fe-oxides (Landsat b3b1)	L31	30	Landsat TM, acquired 7/5/1989

[a]PRISM data source is http://prism.oregonstate.edu

inverse of the vegetation cover (greenness), or soil cover which is greatest in areas of Tertiary bedrock and Holocene fluvial landscapes. In the final results, the b1b3 layer was ranked lower in producing the prediction than the vegetation layers.

The presence of explicit age information in experiment 4 significantly increased prediction (OA = 83%, Khat = 0.89) (Table 4.4). In the final go of experiment 5, the addition of geology and surface (soil) age yielded the most accurate of the experimental setups (OA = 86%, Khat = 0.94). Overall, incremental additions of age information yielded a corresponding incremental increase in the successfulness of prediction results.

Table 4.4 Results of age experiments in predictive soil map of Three Mile Hill Quadrangle, Oregon

	Accuracy assessment			
Experimental setup	Overall	Producer's	User's	Khat
1. No lithologic/age information	78.2	73.4	65.9	0.84
2. Lithology only	81.8	75.1	68.7	0.88
3. Implicit age information only[a]	78.7	73.4	66.7	0.84
4. Explicit age information only[b]	82.8	79.4	70.2	0.89
5. Lithology and (soil) age[c]	85.9	78.1	71.5	0.94

[a]Geomorphologic relative age (Fe-oxide reflectance) is linearly related to surface character; no other geology-related inputs

[b]Geochronologically established age classes of geological (thematic) map units; no other geology-related inputs

[c]Combined lithologic and age (geological thematic) map units

4.4 Discussion

Because age is in the mind of the soil surveyor as "t" of Jenny's (1941) soil-forming factors and in the digital soil mapping model as "a" of McBratney et al. (2003), it follows that inclusion of this factor will improve study results. However, due to a general lack of numerical methods for dating soils, age is commonly expressed as duration of exposure of the surface of the parent material to the atmosphere and pedogenic process (Birkeland, 1999). As discussed by pedologists, e.g., Jenny (1941), McBratney et al. (2003), and Birkeland (1999), time is involved in all environmental factors of soil formation and so it may remain difficult, if not impossible, to fully control for the age/time factor in digital soil mapping.

Dozens of sub-methods and strategies, under the umbrella of chemical, biological and geomorphic methods, are available for use to define both implicit and explicit data layers for digital soil mapping. Many of the rock and mineral weathering methods (Birkeland and Noller, 2000) could be applied to digital soil mapping study without much further development. For example, the population of exposed stones and boulders on the surface of a fluvial deposit is amenable to detection in imagery. Transfer of methods used by Quaternary geologists and physical geographers to map surficial geological and geomorphological map units could yield continuous raster (grid) datasets of age proxy data in the palette of environmental covariate data for digital soil mapping.

Age layers rank highly in their influence on decision-tree analysis results. This paper's demonstration of the importance of age in the milieu of environmental covariates corroborates the findings of Scull et al. (2005). In their soil decision-tree analysis study of the southwestern Great Basin area, age of surficial geologic map units ranks highly among the many independent data layers. Age was carried in the implicit form based on reflectance-age correlation.

Although the studies in which age appears in digital soil mapping are few, as suggested by McBratney et al. (2003) and a review of the literature for this paper, nearly all digital soil mapping studies incorporate data layers that carry implicit age information. It is reasonable to presume that some of these layers carry spatial information on (1) old vs young soils, surfaces and parent materials, (2) actively changing vs equilibrium landscapes, and (3) fast vs slow rates of surficial processes. In disaggregating their geoinformation for digital soil mapping study, Bui and Moran (2001) came very close to explicitly demonstrating this very point.

The addition of geological event(s) could improve prediction. Apart from the static data layers representing the typical digital soil mapping environmental covariates (Lagacherie et al., 2007), a sense of the dynamism in surficial processes can be expressed. Some soils developed under conditions quite different than today. By providing indications of the nature of soil parent materials and/or pedogenesis-altering conditions, paleoenvironmental maps interject information that is atypically carried in maps of geology, modern climate, and hydrology. Examples include glacial geology, paleofloods, and tephra distribution.

4.5 Conclusions

The addition of age information as inferred geological period/epoch classes and as numeric geochronological data yielded significant improvements in the accuracy of prediction in decision-tree analysis-produced digital soil maps. It may well be impossible to conduct digital soil mapping study without any sort of age information or dependence on time. However, it is recommended that the age factor be explicitly applied in all digital soil mapping studies.

Acknowledgments This study was supported by the USDA-NRCS Soils Survey National Geospatial Development Center, through CESU Agreement No. 68-3A75-4-101. The author is indebted to his graduate students, particularly Sarah Hash, who helped with the sublect analyses. Soil survey maps of the study area were provided by Alina Rice, Project Leader of the initial soil survey of Malheur County, Oregon, southern part. I appreciate the helpful review comments of Suzann Kienast-Brown. This is a contribution of the Oregon Agricultural Experiment Station.

References

Birkeland, P.W., 1999. Soils and Geomorphology. Oxford University Press, New York, NY.

Birkeland, P.W., and Noller, J.S., 2000. Rock and mineral weathering, pp. 293–312. In: Noller, J.S., Sowers, J.M., and Lettis, W.R. (eds.), Quaternary Geochronology: Methods and Applications: American Geophysical Union Reference Shelf Series, vol. 4, American Geophysical Union: Washington, DC.

Bondre, N.R., 2006. Field and geochemical investigation of basaltic magmatism in the western United States and western India. Ph. D. dissertation, Miami University, 263 pp.

Bui, E.N., and Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103:79–94.

Farr, T. G., 1992. Microtopographic evolution of lava flows at Cima Volcanic Field, Mojave Desert, California. Journal of Geophysical Research 97(B11):15,171–15,179.

Hart, W.L., and Mertzman, S.A., 1983. Late Cenozoic volcanic stratigraphy of the Jordan Valley area, southeastern Oregon. Oregon Geology 45:15–19.

Hash, S.J., 2008. Use of decision tree analysis for predictive soils mapping and implementation on the Malheur County, Oregon initial soil survey. Masters Thesis, Oregon State University, Corvallis.

Hash, S.J., and Noller, J.S., 2009. Incorporating predictive mapping to advance initial soil survey: an example from Malheur County, Oregon (Invited Paper). Soil Survey Horizons 50:111–115.

Jackson, T.A., and Keller, W.D., 1970. A comparative study of the role of lichens and "inorganic" processes in the chemical weathering of recent Hawaiian lava flows. American Journal of Science 269:446–466.

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

Kahle, A.B., Gillespie, A.R., Abbott, E.A., Abrams, M.J., Walker, R.E., Hoover, G., and Lockwood, J.P., 1988. Relative dating of Hawaiian lava flows using multispectral thermal infrared images: a new tool for geologic mapping of young volcanic terranes. Journal of Geophysical Research 93(B12):15239–15251.

Lagacherie, P., Legros, J.P., and Burrough, P., 1995. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. Geoderma 65:283–301.

Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), 2007. Digital Soil Mapping: An Introductory Perspective. Elsevier Science & Technology, Amsterdam.

Lagacherie, P., and Voltz, M., 2000. Predicting soil properties over a region using sample information from a mapped reference area and digital elevation data: a conditional probability approach. Geoderma 97:187–208.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

Noller, J.S., Sowers, J.M., and Lettis, W.R. (eds.), 2000. Quaternary Geochronology: Methods and Applications. American Geophysical Union Reference Shelf Series. American Geophysical Union, Washington, DC.

Noller, J.S., 2008. Preliminary Surficial geological maps of the Arock, Danner, Dry Creek Rim, Jordan Craters South, and Three Mile Hill 7.5–Minute quadrangles: A 1:24,000-scale geodatabase, Unpublished Report to the USDA-NRCS, National Geospatial Development Center, Morgantown, WV.

Scull, P., Franklin, J., and Chadwick, O.A., 2005. The application of decision tree analysis to soil type prediction in a desert landscape. Ecological Modelling 181:1–15.

Soil Survey Staff, 1993. Soil Survey Manual. U.S.D.A. Agriculture Handbook No. 18. U.S. Government Printing Office, Washington, DC.

Stretch, R.C., and Viles, H.A., 2002. The nature and rate of weathering by lichens on lava flows on Lanzarote. Geomorphology 47:87–94.

Vaughan, K.L., 2008. Pedogenesis at craters of the moon national monument and preserve, Idaho. Ph. D. dissertation, University of Idaho, Moscow. 156 p.

Wasklewicz, T.A., 1994. Importance of environment on the order of mineral weathering in olivine basalts, Hawaii. Earth Surface Processes and Landforms 19:715–734.

Chapter 5
Scale Effects on Terrain Attribute Calculation and Their Use as Environmental Covariates for Digital Soil Mapping

S.M. Roecker and J.A. Thompson

Abstract The digital representation of the Earth's surface by terrain attributes is largely dependent on the scale at which they are computed. Typically the effects of scale on terrain attributes have only been investigated as a function of digital elevation model (DEM) grid size, rather than the neighborhood size over which they are computed. With high-resolution DEM now becoming more readily available, a multi-scale terrain analysis approach may be a more viable option to filter out the large amount short-range variation present within them, as opposed to coarsening the resolution of a DEM, and thereby more accurately represent soil-landscape processes. To evaluate this hypothesis, two examples are provided. The first study was designed to evaluate the systematic effects of varying both grid and neighborhood size on terrain attributes computed from LiDAR. In a second study, the objective was to examine how the correlations between soil and terrain attributes vary with neighborhood size, so as to provide an empirical measure of what neighborhood size may be most appropriate. Results suggest that the overall representation of the land surface by terrain attributes is specific to the land surface, but also that the terrain attributes vary independently in response to spatial extent over which they are computed. Results also indicate that finer grid sizes are more sensitive to the scale of terrain attribute calculation than larger grid sizes. For the soil properties examined in this study, slope curvatures produced the highest coefficients of correlation when calculated at neighborhood sizes between 117 and 189 m.

Keywords Digital elevation model · Terrain analysis · Scale · Resolution · Neighborhood size

J.A. Thompson (✉)
Division of Plant and Soil Sciences, West Virginia University, PO Box 6108, Morgantown, WV 26506-6108, USA
e-mail: james.thompson@mail.wvu.edu

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_5,

5.1 Introduction

The utilization of digital elevation models (DEM) has proven to be invaluable to recent efforts in digital soil mapping (DSM) (see Chapter 2 for examples of DEM-derived attributes used to represent topography and climate in DSM applications). According to a recent survey of the literature by McBratney et al. (2003), DEM are by far the most heavily used environmental covariates. The popularity of this data source as a predictive variable in DSM stems from its simple data structure, widespread availability, and, most importantly, due to the strong influence of topography on landscape scale processes that influence soil variability. Multiple DEM-derived terrain attributes can be used to represent the Earth surface and serve as indices of pedogeomorphological significance, such as slope gradient, slope aspect, slope curvature, flow accumulation, and compound topographic parameters (see Table 3.1 and Table 29.1 for extensive lists of DEM-derived environmental covariates used for spatial modeling and predictive mapping). These digital representations of the Earth's surface so far have proven useful in explaining a substantial portion of soil variation with (geo)statistical models (see Section 17.2.2 and Table 30.3 for examples). In hopes of explaining more soil variation with terrain attributes, research has sought to improve their digital representation by evaluating the numerous methods for generating DEM (Chaplot et al., 2006), and the algorithms used to derive terrain attributes from them (Florinsky, 1998). Alternative approaches have examined the use of DEM simulation (Burrough et al., 2000) and wavelets (Gallant and Hutchinson 1997). A entirely different approach has sought to scale up the area (i.e., support) over which soil observations are made, so as to quantify a more representative area of the landscape from which their environmental correlation with terrain attributes are made (O'Connell et al., 2000). Perhaps the most significant source of error in DSM is attributable to the scale at which the Earth's surface is represented by terrain attributes.

The scale or spatial extent of terrain attributes is related to two factors, the grid size or horizontal resolution of the DEM used, and the window or neighborhood size over which they are calculated. While the effect of grid size has been studied extensively, the effect of neighborhood size has not. Perhaps the most important assertion that has been made by such studies is that no perfect DEM resolution exists (Claessens et al., 2005). Also, finer resolutions (5–10 m), in particular, do not produce more accurate predictions, but instead introduce excessive detail which is difficult to characterize with conventional automated methods (see Section 27.5.2). An alternative to coarsening the resolution of a DEM to match the scale of the soil-landscape processes being mimicked by terrain attributes is multi-scale terrain characterization. This alternative approach was first suggested by Wood (1996), whereby the window size used to calculate terrain attributes from a DEM is set by the user. The application of this approach has been demonstrated in a few instances, and it has been shown that soil-landscape relationships vary with neighborhood size (Park et al., 2001; Schmidt and Hewitt, 2004; Smith et al., 2006).

With high-resolution DEM now becoming more readily available, a multi-scale terrain analysis approach may be a more viable option to filter out the large amount short-range variation present within them, as apposed to coarsening the resolution of a DEM, and thereby more accurately represent soil-landscape processes. To evaluate

this hypothesis, two examples are provided. One case study was designed to evaluate the systematic effects of varying both grid and neighborhood size on terrain attributes computed from LiDAR. In a second case study, the objective was to examine how the correlations between soil and terrain attributes vary with neighborhood size, so as to provide an empirical measure of what neighborhood size may be most appropriate.

5.2 Materials and Methods

5.2.1 Case Study 1 – Systematic Effects of Varying Grid and Neighborhood Size on Terrain Attributes

5.2.1.1 Study Area

For this study, DEM datasets were chosen to represent two distinctly different landscapes within West Virginia. LiDAR elevation data with 1-m resolution were available for Gilmer County, in the Appalachian Plateau of central WV, and Jefferson County, in the Valley and Ridge of eastern WV (Fig. 5.1). The areal extent at each study area was a quarter-quarter (QQ) quadrangle (~600 ha).

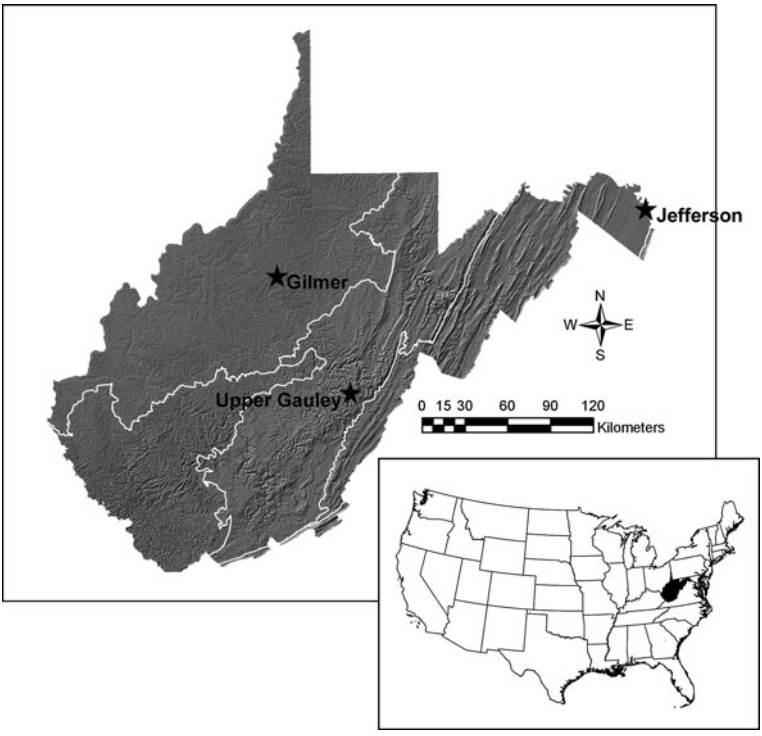


Fig. 5.1 Location of the three study areas across the Major Land Resource Areas within the state of West Virginia. *Inset*: Location of West Virginia in the continental US

5.2.1.2 DEM Resampling and Terrain Attribute Calculation

To allow for a direct comparison between the terrain attributes, specific DEM resampling and terrain attribute calculation procedures were used. This was necessary to ensure that the elevation values, grid sizes, and neighborhood sizes corresponded for each comparison. When resampling the DEM this was accomplished by using the nearest neighbor approach and selecting a set of nested grid and neighborhood sizes (Table 5.1). From each DEM, the terrain attributes slope gradient, northerness, profile curvature, tangential curvature, contour curvature, and mean curvature were calculated according to the formulas of Zevenbergen and Thorne (1987), with an adjustment to allow the neighborhood size to be specified by the user. As slope aspect is a circular measure, its values are not suitable for direct comparison. Therefore, slope aspect was transformed to northerness by, Northerness $= |180 - \text{Aspect}|$.

5.2.1.3 Comparison of Grid and Neighborhood Size Combinations

Both qualitative and quantitative comparisons were made to evaluate the systematic effects of varying both grid size and neighborhood size on terrain attributes. The qualitative comparisons were made by examining the scatter plots of the individual terrain attributes vs. neighborhood size and inspecting box plots of the terrain attributes at various neighborhood sizes. The quantitative comparisons were made by calculating the goodness of fit between the benchmark (3×3 window size) and expanded window sizes. The goodness of fit measures used were the mean difference (MD), root mean square difference (RMSD), and Pearson's correlation

Table 5.1 Experimental contrasts across grid size and neighborhood size

Grid size	1	3	9	27	81
Neighborhood size	Lag	Lag	Lag	Lag	Lag
3	1*				
5	2				
7	3				
9	4	1*			
15	7	2			
21	10	3			
27	13	4	1*		
45	22	7	2		
63	31	10	3		
81	40	13	4	1*	
135	67	22	7	2	
189	94	31	10	3	
243	121	40	13	4	1*
405	202	67	22	7	2
567	283	94	31	10	3
729	364	121	40	13	4

Lag is the radius of the neighborhood size, measured in the number of grid cells from the center cell. The lags that are labeled with an asterisk (*) serve as the bench mark, from which the comparisons were made

coefficient (r). Due to the high-resolution of the datasets involved, the number of corresponding points with the finest grid size totaled approximately 5,400,000. Therefore, for the sake of computational efficiency, only the cell centers from the 27-m grids were used to evaluate the finer grid sizes, which totaled ~7,400 points.

5.2.2 *Case Study 2 – Soil and Terrain Attributes Correlations Response to Neighborhood Size*

5.2.2.1 Study Area

For this study, the study area was the Upper Gauley watershed (UGW) on the Monongahela National Forest, which is located on the Appalachian Plateau of eastern West Virginia (Fig. 5.1). Within this watershed a soil dataset was available with which to examine the effect of neighborhood size on correlations between soil properties and terrain attributes. The elevation dataset for this site was derived from high-resolution aerial photography and had a grid size of 3-m, but was resampled to 9-m.

5.2.2.2 Soil Sampling and Analysis

Soils data were collected from 97 sites within the UGW. Sample locations were selected according to a stratified-random design; similar to that used by McKenzie and Ryan (1999). The stratifying variables used were geology (62% sandstone, 11% shale/sandstone, and 26 % shale), elevation (three quantile classes), and stream power index (five quantile classes). The intersection of these variables created 45 unique strata, within which two random sites were sampled. In order to avoid sampling extraneous features, exclusion rules were used to avoid roads (buffered to 20 m), streams (buffered to 10 m), developed areas, and patches of strata smaller than 4,000 m^2 (approximately 1 acre).

At each site a soil pit was excavated to 140 cm, described according to standard procedures (Schoeneberger et al., 2002), and georeferenced. From each horizon an approximately 300 g sample of soil was taken. These samples were analyzed for particle size, pH, extractable calcium and magnesium, and carbon. For the purpose of spatial analysis, the soil attributes subdivided into three depth increments (0–50, 50–100, and 100–150 cm) by taking a weighted average of the samples across the soil horizons.

5.2.2.3 Soil and Terrain Attribute Correlation

Procedures outlined by Florinsky and Kuryankova (2000) were used to examine the variation in environmental correlation between soil properties and terrain attributes when altering the neighborhood size used to calculate the terrain attributes from a high-resolution DEM. This procedure involves plotting the goodness of fit (correlation coefficient, r) between selected soil properties and individual terrain attributes

over a span of neighborhood sizes, and identifying the range in neighborhood size that occupies a smooth portion within the plots. Such an area of the correlation plot where the correlation values are relatively invariant as neighborhood size increases represents a range in scales where soil properties exhibit less spatial heterogeneity and more predictable variability, providing for more reproducible and interpretable results (Florinsky and Kuryankova, 2000). The terrain attributes that were considered for this analysis were slope gradient, northerness, profile curvature, tangential curvature, contour curvature and mean curvature.

5.3 Results and Discussion

5.3.1 Case Study 1 – Systematic Effects of Varying Grid and Neighborhood Size on Terrain Attributes

As the neighborhood size increases, the landscape features that are represented by the derived terrain attributes is altered, and short-range variation is filtered out in favor of broader hillslope trends. At small neighborhood sizes (e.g., $\leq$9 m), it is the microtopographic features that are represented by the terrain attributes (Fig. 5.2a). As such, there is noticeable short-range variability in terrain attribute

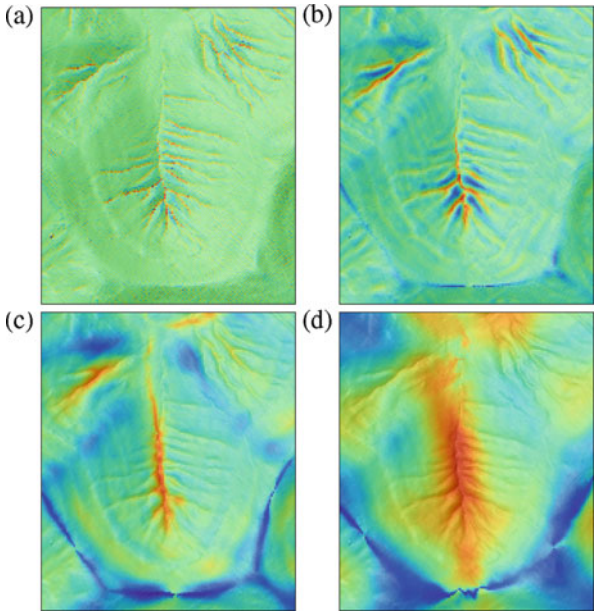


Fig. 5.2 Case Study 1. Tangential curvature derived from 3-m DEM calculated using neighborhood sizes of (**a**) 9 m, (**b**) 27 m, (**c**) 81 m, and (**d**) 243 m. The image represents a subcatchment from the QQ quadrangle from Gilmer County, WV

values (Fig. 5.2a) and wide ranges in the distribution of the terrain attribute values (Fig. 5.3). With increasing neighborhood size, the landscape features that are represented by the terrain attributes correspond more closely to recognizable landform elements (Fig. 5.2b and c), such as drainageways, footslopes, backslopes, and shoulders. At these intermediate neighborhood sizes (e.g., 15–81 m), the terrain attributes represent a smoother but more connected landscape (Fig. 5.2b and c) This smoothing of the landscape representation is also seen in the boxplots (Fig. 5.3), where the median value remains relatively stable, but the interquartile range decreases and the outliers become less extreme. When larger neighborhood sizes are used (e.g., $>$ 81 m), the landscape features depicted by the terrain attributes become oversimplified (Fig. 5.2d). When the neighborhood size becomes significantly large, the smoothing of the landscape increases and the terrain attributes may misrepresent landform elements because the neighborhood includes DEM data from outside the local landscape (e.g., from across a watershed divide). This distortion or oversimplification of the landscape represented by the terrain attributes appears in the boxplots of slope gradient by a loss in the stability in the median value and continued decrease in the maximum value above a neighborhood size of 81 m (Fig. 5.3a and e). More extreme variability is also seen in the contour curvature values beyond this same neighborhood size threshold of 81 m (Fig. 5.3c and d).

The effects of changing neighborhood size on derived terrain attributes is not the same for the two landscapes examined in this case study. The study area in Gilmer Co., WV, located on the Central Allegheny Plateau, exhibits much steeper slope gradients and more extreme slope curvature values compared to the lower-relief study area of the Northern Appalachian Ridge and Valley of the study area in Jefferson Co., WV. Accordingly, the effect of increasing neighborhood size is much less pronounced in the lower relief landscape of Jefferson Co. compared to the higher relief landscape of Gilmer Co. (Fig. 5.3a and e). However, there still appears to be a change in the distribution of terrain attribute values at larger neighborhood sizes, with a decrease in both median and maximum slope gradient values above a neighborhood size of 81 m.

While terrain attributes are affected by neighborhood size, the effect of grid size appears to be negligible. The same decreases in maximum values, decreases in interquartile range, and decreases in median value (above a neighborhood size of 81 m) are seen if slope gradient is calculated using a 1-m DEM (Fig. 5.3a and e), a 9-m DEM (Fig. 5.3b and f), or an 81-m DEM. These and other observations suggest that slope gradient is not sensitive to the effect of spatial extent up until 81-m. For these landscapes, it appears that a neighborhood size of 81-m corresponds with the threshold beyond which all the terrain attributes become increasingly less representative of the original landscape. Therefore a maximum grid size of 27-m, which produces a 3×3 moving window with an extent of 81 m, is recommended to adequately represent the land surface.

The goodness of fit measures (MD, RMSD, r) quantify the similarity of the terrain attributes derived from a given neighborhood size to the terrain attributes calculated using the nearest neighbors in the DEM. As such, higher MD and RMSD values and lower r values indicate greater differences between the two paired data

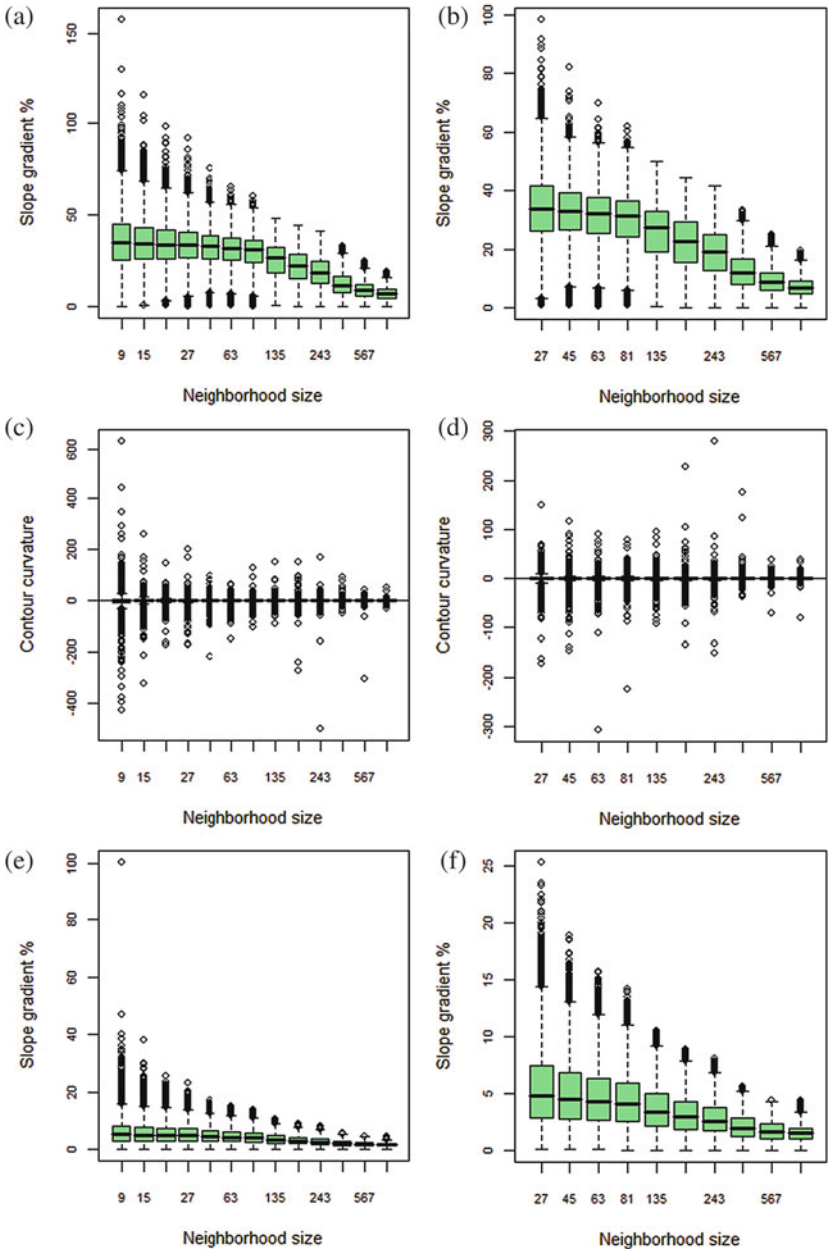


Fig. 5.3 Case Study 1. Boxplots of terrain attributes as calculated using differing neighborhood sizes and different grid sizes: (**a**) slope gradient, Gilmer Co., 3 m grid size; (**b**) slope gradient, Gilmer Co., 9 m grid size, (**c**) contour curvature, Gilmer Co., 3 m grid size; (**d**) contour curvature, Gilmer Co., 9 m grid size; (**e**) slope gradient, Jefferson Co., 3 m grid size; (**f**) slope gradient, Jefferson Co., 9 m grid size

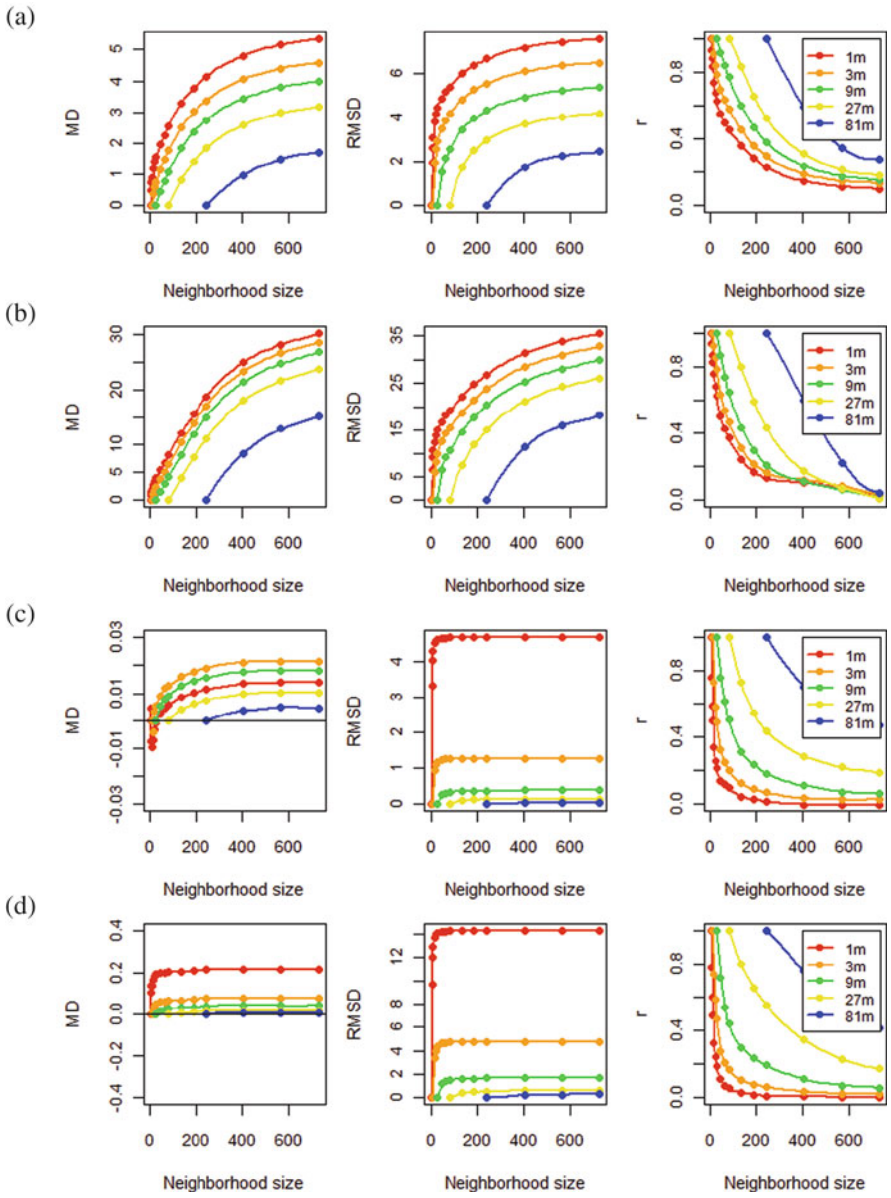


Fig. 5.4 Case Study 1. Goodness of fit plots of (**a**) slope gradient, Gilmer Co.; (**b**) slope gradient, Jefferson Co.; (**c**) tangential curvature, Gilmer Co.; and (**d**) tangential curvature, Jefferson Co

sets. Two trends are evident in the goodness of fit data. First, even small increases in neighborhood size lead to larger differences in the terrain attribute values, and therefore the representation of the landscape features. This is seen for all terrain attributes, for both study areas, and for all grid sizes (Fig. 5.4). While the differences increase rapidly at first, a neighborhood size is reached where the differences between the paired data sets approach (Fig. 5.4a and b) or reach (Fig. 5.4c and

d) a constant value. There inflection points are most pronounced with the slope curvature comparisons and with smaller grid sizes. For both study sites, and for all terrain attributers, this inflection point appears to occur at approximately the same neighborhood size: between 45 and 81 m. With increasing grid size this inflection point diminishes to the point where it disappears completely, so that with larger grid sizes it falls below the minimum window size of 3 × 3. A second trend is that

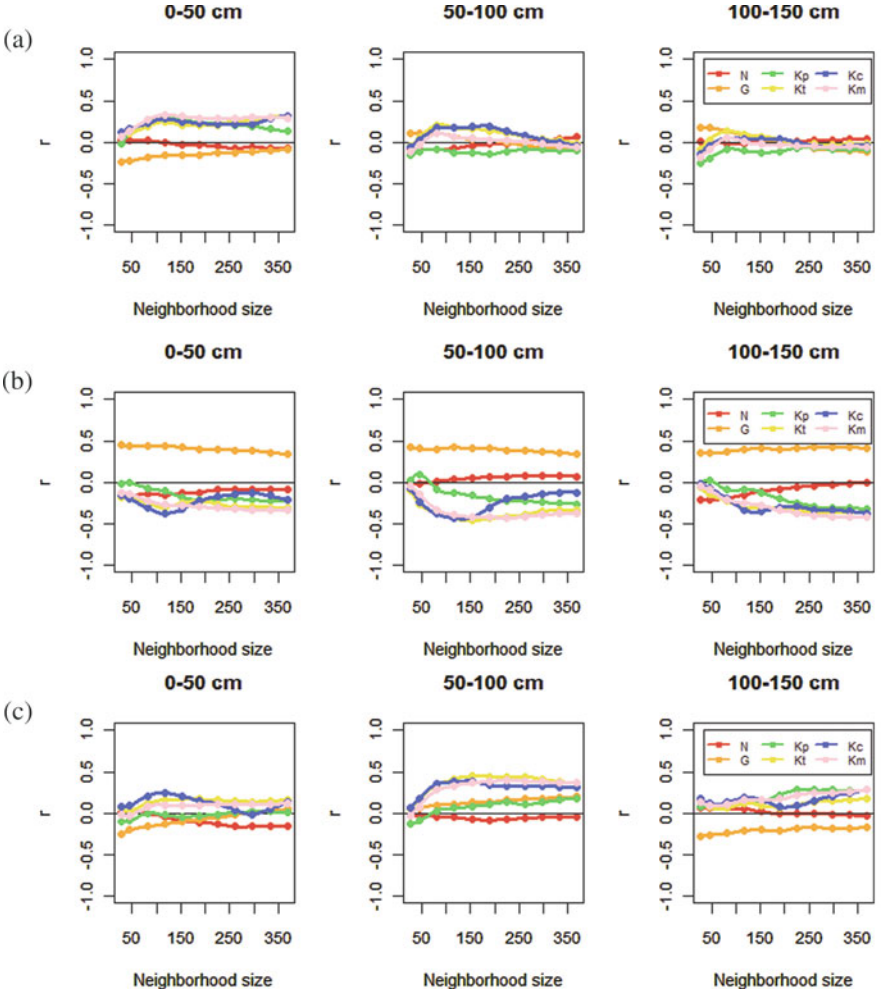


Fig. 5.5 Case Study 2. Correlation coefficient vs. neighborhood size for selected terrain attributes and (**a**) soil C, (**b**) rock fragment content, and (**c**) clay content (N = northness, G = slope gradient, Kp = profile curvature, Kt = tangential curvature, Kc = contour curvature, and Km = mean curvature)

the magnitude of the differences in terrain attribute values is much greater for the higher relief landscape of the Gilmer Co. study area (Fig. 5.4a and c) compared to the Jefferson Co. study area (Fig. 5.4b and d).

5.3.2 Case Study 2 – Soil and Terrain Attributes Correlations Response to Neighborhood Size

The results show that the correlation between the soil properties and surface curvatures are the most sensitive to the effects of neighborhood size (Fig. 5.5). Their correlation with the soil attributes ranged from approximately 0 to 0.5, with an optimal neighborhood size range of 117–189 m in most cases. That the correlations between the soil properties and surface curvatures were optimized is important because slope curvatures have the strongest correlation coefficients in most cases.

While significant correlations were also present between slope gradient and the soil properties, the effect of neighborhood size was negligible in most cases (Fig. 5.5). As for northerness, only one significant correlation was observed. The negligible effect of neighborhood size on the correlation between the soil properties and slope gradient may be explained by the results of the first case study. Increasing the neighborhood size used to derive terrain attributes had less of an effect on slope gradient than on slope curvature (Figs. 5.3 and 5.4). Consequently, correlation coefficients between soil properties and slope gradient do not appear to be sensitive to the effect of neighborhood size.

5.4 Conclusions

The spatial extent over which terrain attributes are derived has a considerable effect on their representation of the land surface and correlation with soil attributes. In the first study described here, it was shown that using a larger neighborhood size has a similar effect as using a larger grid size, without the unnecessary loss of detail caused by using a larger grid size. Still the amount of detail provided by the smallest grid sizes was excessive and computationally demanding. To help determine what neighborhood size might be most appropriate for DSM, the simple exploratory procedures used here proved to be informative. Ultimately, the relative size of the landforms within the study area should serve as a guide. Within the second study described here the correlation between the soil attributes and surface curvatures was optimized by increasing the neighborhood size. While no one neighborhood size showed the strongest correlation in all cases, a common range of optimal neighborhood sizes occurred over a range of 117–189 m. Slope gradient also showed significant correlations with some of the soil properties, but was not sensitive to changes in neighborhood size. These results suggest that surface curvatures are the most sensitive to altering the neighborhood size used to calculate terrain attributes and that curvature values poorly represent the land surface unless calculated over a neighborhood size commensurate with the size the landforms which they characterize.

References

Burrough, P.A., van Gaans, P.F.M., and MacMillan, R.A., 2000. High-resolution landform classification using fuzzy k-means. Fuzzy Sets and Systems 113:37–52.

Claessens, L., Heuvelink, G.B.M., Schoorl, J.M., and Veldkamp, A., 2005. DEM resolution effects on shallow landslide hazard and soil redistribution modelling. Earth Surface Processes and Landforms 30:461–477.

Chaplot, V., Darbouz, F., Bourennane, H., Leguedois, S., Silvera, N., and Phachomphon, K., 2006. Accuracy of interpolation techniques for the derivation of digital elevation models in relation to landform types and data density. Geomorphology 177:126–141.

Florinsky, I.V., 1998. Accuracy of local topographic variables derived from digital elevation models. International Journal of Geographic Information Science 12(1):47–61.

Florinsky, I.V., and Kuryankova, G.A., 2000. Determination of grid size for digital terrain modelling in landscape investigations – exemplified by soil moisture distribution at a micro-scale. International Journal of Geographical Information Science 14(8):815–832.

Gallant, J.C., and Hutchinson, M.F., 1997. Scale dependence in terrain analysis. Mathematics and Computers in Simulation 43:313–321.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

McKenzie, N.J., and Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89:67–94.

O'Connell, D.A., Ryan, P.J., McKenzie, N.J., and Ringrose-Voase, A.J., 2000. Quantitative site and soil descriptors to improve the utility of forest soil surveys. Forest Ecology and Management 138:107–122.

Park, S.J., McSweeney, K., and Lowery, B., 2001. Identification of the spatial distribution of soils using a process-based terrain characterization. Geoderma 103:249–272.

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

Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., and Broderson, W.D. (eds.), 2002. Field Book for Describing and Sampling Soils, Version, 2.0. Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE.

Smith, M.P., Zhu, A.X., Burt, J.E., and Stiles, C., 2006. The effects of DEM resolution and neighborhood size on digital soil survey. Geoderma 137:58–69.

Wood, J., 1996. The geomorphological characterization of Digital Elevation Models. PhD Thesis Department of Geography, University of Lancaster, UK.

Zevenbergen, L.W., and Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth Surface Processes and Landforms 12(1):47–56.

Chapter 6
Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA

C.W. Brungard and J.L. Boettinger

Abstract Conditioned Latin Hypercube Sampling (cLHS) is a type of stratified random sampling that accurately represents the variability of environmental covariates in feature space. As the smallest possible sample is important for efficient field work, what is the optimal sample size for digital soil mapping? An optimal sample size accurately represents the variability in the environmental covariates and provides enough samples for predictive models. This paper briefly reviews cLHS and investigates different sample sizes for representing five environmental covariates in a 30,000-ha complex landscape in the Great Basin of southwestern Utah. The cLHS code was run in Matlab^{TM} (Mathworks, 2008) and statistical analysis was performed using the R statistical language (R Development Core Team, 2009). Graphical analysis for continuous data and chi-square analysis of categorical data suggested optimal sample size for this study area is approximately 200 to 300 (0.05–0.1%).

Keywords Sampling · Latin hypercube · Digital soil mapping · Great Basin · Rangelands

6.1 Introduction

Current methods of sampling for soil survey in the USA depend on the subjective decisions of soil surveyors and involve few, if any, statistically identified sampling sites (Soil Survey Division Staff, 1993). Such methods introduce bias and impair attempts at statistical, classification, and/or interpolation methods for digital soil mapping. A statistically robust sampling method is needed to eliminate bias and improve predictive models.

Some soil scientists have used random (e.g. Howell et al., 2007) and stratified random (e.g., McKenzie and Ryan, 1999 and Section 29.2.2) methods as statistically

C.W. Brungard (✉)
Department of Plant, Soils and Climate, Utah State University, 4820 Old Main Hill, Logan, UT 84322-4820, USA
e-mail: c.w.b@aggiemail.usu.edu

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_6,

robust sampling strategies. While random and stratified random sampling may sample throughout the geographic space, the samples may not be distributed through the full range of the environmental covariates (the feature space). Additionally, random sampling methods often require large numbers of samples that are impractical to obtain with soil survey budgets and time constraints. Bui et al. (2007) found that extrapolation between sampling sites was grossly incorrect if the sampling was not representative of the landscape, and called for a method that ensures every combination of environmental covariates is covered. A sampling scheme that meets time and budget constraints, that is statistically sound, and that represents the covariate feature space is needed (see Section 16.4).

Minasny and McBratney (2006) proposed conditioned Latin Hypercube Sampling (cLHS) as an efficient method for sampling from the variability of the feature space of multiple environmental covariates. Conditioned Latin Hypercube Sampling is based on the concept of Latin Hypercube Sampling (LHS) where a sample is drawn from the covariates such that each variable is maximally stratified. Conditioned Latin Hypercube Sampling adds the condition that the sample chosen must actually occur on the landscape. Minasny and McBratney (2006) showed that cLHS closely represented the original distribution of the environmental covariates with relatively small sample sizes in a digital soil mapping project in the Hunter Valley of New South Wales, Australia. Minasny and McBratney (2007) compared cLHS to simple random sampling, stratified random sampling, sampling along the principal components, and spatial Latin Hypercube sampling, and demonstrated that cLHS most accurately reproduced the original distribution of the environmental covariates. However, the question remains: what is the optimal cLHS sample size?

Determining optimal cLHS sample size is important for accuracy and efficiency in production soil survey, especially for countries with limited soil survey budgets. An optimal sample size would minimize the costs of field data collection by minimizing sample number while providing both an accurate representation of the variability in the environmental covariates and enough samples for predictive models. This paper attempts to determine the optimal sample size for a 30,000-ha complex Great Basin landscape in southwestern Utah represented by five environmental covariates.

6.2 Materials and Methods

6.2.1 Study Area

Located in Beaver County, southwestern Utah, USA, the study area comprises approximately 30,000 ha (~70,000 acres) of mountains, hills, and associated alluvial fans typical of the arid Great Basin (Fig. 6.1). The study area includes parts of two small mountain ranges. The Star Range in the northeast of the study area is predominately tilted, faulted, and, in places, metamorphosed limestone, shale and sandstone, whereas the Shauntie Hills in the south are predominantly exposed

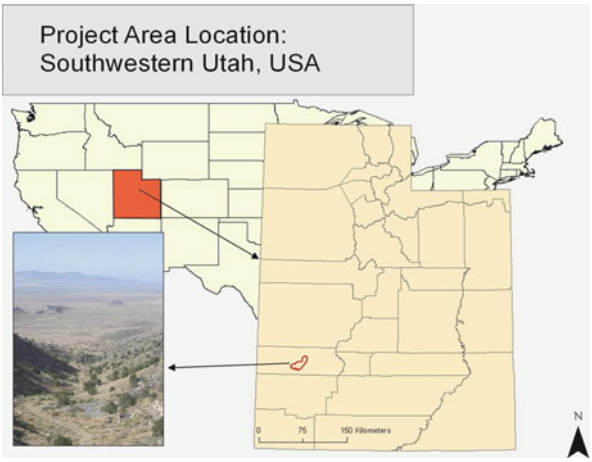


Fig. 6.1 Study area location in southwestern Utah. The photograph shows typical vegetation and relief

volcanic flows, covered in places with a relatively thin veneer of alluvium (Best et al., 1989). Elevation for the entire project area ranges from 1,500 to 2,100 m (4,900–6,900 ft).

Precipitation occurs as winter snow or high-intensity, convective thunderstorms in the late summer, yet water is scarce, averaging less than 23 cm annual precipitation. There are no perennial streams, but intermittent waterways periodically fill with flash flooding during particularly intense storms. A few extremely localized springs occur in the area. Consistent with the elevation and limited water this area supports desert shrub, shrub-grass, and sparse woodland vegetation (NRCS, 2008). Mining for precious metals mostly ended by the early 1900s, and land use has since been limited to sheep and cattle grazing.

Dominant soils in this area were classified as Aridisols and Entisols according to *Soil Taxonomy*. We estimated the soil temperature and moisture regime to be aridic and xeric, respectively.

6.2.2 Digital Data

Table 6.1 summarizes the digital data layers chosen to represent environmental covariates in the SCORPAN model (McBratney et al., 2003).

Slope, inverse wetness index, and aspect were calculated from a sink-filled, 10 m digital elevation model (DEM) from the United States Geological Survey (USGS) National Elevation Dataset (NED, 2006) using TauDEM (Tarboton, 2008). The Inverse Wetness Index, the inverse of the Compound Topographic Index (CTI), is used to avoid division by zero where slope is zero (Tarboton, 2008). Aspect was converted to transformed aspect (a measure of northness vs. southness) using ERDAS Imagine^{TM}(ERDAS, 2006) modeler (Stum, 2007).

Landsat 7 ETM+ images were obtained from the Intermountain Region Digital Image Archive Center (IRDIAC, 2008) for several dates. An image acquired July

Table 6.1 Digital data used to represent SCORPAN environmental covariates

Environmental covariate	Representative digital data	Source data
Soil (*s*)	None	None
Climate (*c*)	None	None
Organisms (*o*)	Soil adjusted vegetation index (SAVI)	Landsat 7 ETM+ (30 m)
	Land cover (vegetation type)	SWReGAP (30 m)
Relief (*r*)	Inverse wetness index (IWI)	10 m NED DEM
	Slope	10 m NED DEM
	Transformed aspect	10 m NED DEM
Parent Material (*p*)	Geology	USGS geological map (30 m)
Time/age (*a*)	None	None
Space (*n*)	None	None

31st 2000 was selected to show maximum green leaf area for calculating vegetation indices. The image was atmospherically corrected using the COST method (Chavez, 1996). Soil Adjusted Vegetation Index (SAVI) was calculated following the method of Huete using an L value of 0.5 (Huete, 1988).

Land cover type was obtained from the Southwest Regional Gap Analysis Program (USGS National Gap Analysis Program, 2004). The SWReGAP is a digital land cover map of the southwestern USA produced from Landsat imagery. Twelve common Basin and Range land cover types were identified.

Geology was obtained from a USGS 1:50,000 geology map. Five broad geologic types were digitized then converted to raster form. Because the Matlab code from Minasny (2007) allowed only one categorical variable, land cover and geology were combined into a single landcover-geology data layer.

A regularly spaced point grid was used to extract environmental covariate values every 30 m. The resulting text file was the input to the cLHS algorithm.

6.2.3 cLHS

The objective of cLHS for digital soil mapping is to find a set of values from several digital environmental covariates that satisfy the requirements of a Latin hypercube and that exist in the real world. The requirements of a Latin hypercube are that only one sample exists in each row and column, in *n* dimensions. Because particular combinations of values (samples) selected by Latin Hypercube Sampling (LHS) may not exist in the environmental covariate data a solution must be found. It is possible to keep repeating the LHS and searching until suitable combinations are found, but a better way is to select samples that form a Latin hypercube in the feature space. This becomes an optimization problem (Minasny and McBratney, 2006).

The cLHS is implemented as follows: The user decides upon a sample size, the algorithm divides the environmental covariate data into the same number of

equally probable strata, and a random sample of points from the input environmental covariate data is selected, combined, and tested against the demands of a Latin Hypercube. It is quite likely that this combination of random samples does not form a Latin Hypercube. A combination of values that is close to a Latin Hypercube and that occurs in the digital environmental covariate data is iteratively obtained by the objective function and annealing schedule contained in the cLHS code (Minasny and McBratney, 2006).

The cLHS algorithm developed by Minasny and McBratney (2006) was run using Matlab (Mathworks, 2008) software. Sample sizes of 50, 100, 200, 300, and 500 were produced. The run time for 50 samples and 125,000 iterations using an Intel Core 2 Duo 3.0 GHz with 256 MB RAM was approximately 2.1 h. Greater numbers of iterations require longer run times.

6.2.4 Optimal Sample Size

We generated box plots (Fig. 6.2) and density distributions (Fig. 6.3) to compare sample sizes against each continuous environmental covariate. Sample sizes were evaluated based on the similarity between sample size and covariate.

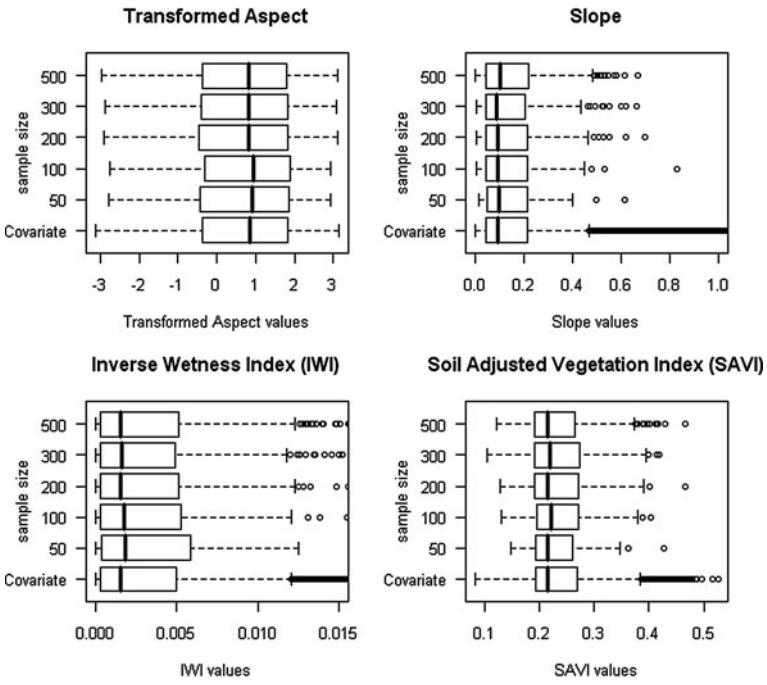


Fig. 6.2 Box-whisker plots comparing distribution of multiple sample sizes to distribution of full environmental covariate dataset

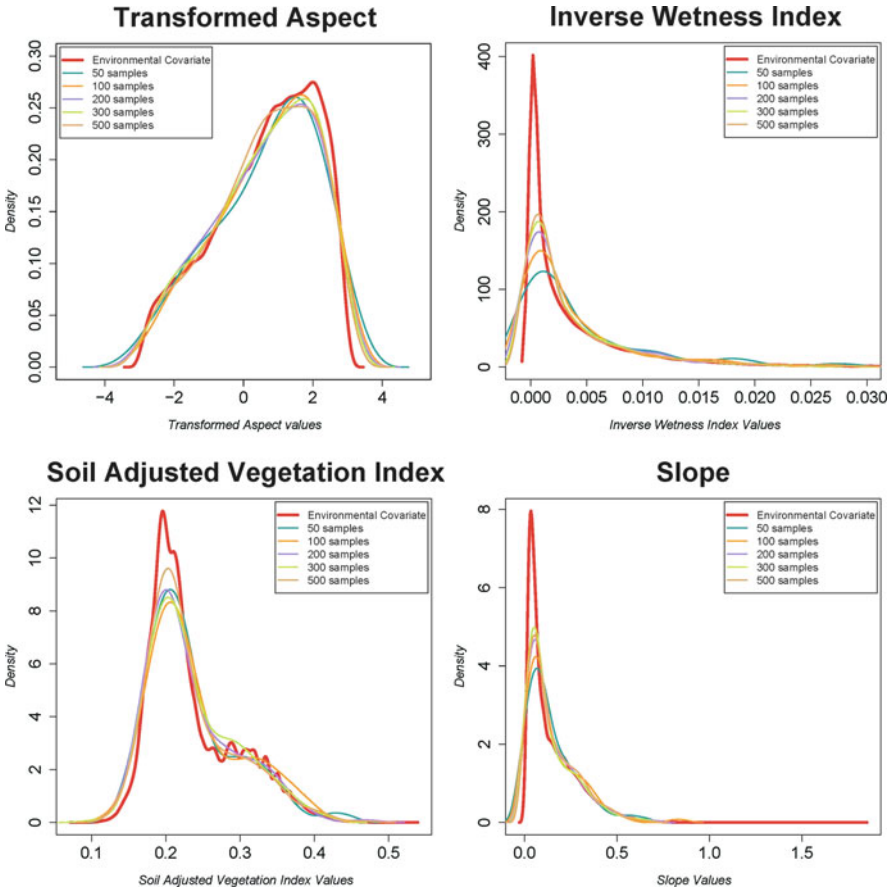


Fig. 6.3 Density plots comparing distribution of sample sizes for each environmental covariate

Table 6.2 Chi square test of the combined geology/land cover categorical variable distribution for multiple sample sizes

Combined geology and land cover categorical variable			
cLHS sample size	Chi-square values	*p*-values	Conclusion
50 samples	4.87	1.00	Accept null[a], sample is not statistically different from covariate
100 samples	12.68	0.999	Accept null[a], sample is not statistically different from covariate
200 samples	13.6	0.998	Accept null[a], sample is not statistically different from covariate
300 samples	4.13	1.00	Accept null[a], sample is not statistically different from covariate
500 samples	6.86	1.00	Accept null[a], sample is not statistically different from covariate

[a]Reject null at $a = 0.05$

We generated chi-square goodness of fit measures (Table 6.2) for the combined geology/land cover categorical covariate. If the sample is not statistically different from the covariate the sample is assumed to adequately represent the geology/land cover environmental covariate.

6.3 Results

Box plots are presented in Fig. 6.2. Box plots compare mean and inter-quartile range (IQR) between covariates and sample sizes. Whiskers on the box plots are 1.5*IQR. Box plots show that little difference exists between median and IQR for cLHS sample sizes and each covariate. Density curves for each covariate and each sample size are presented in Fig. 6.3. All sample sizes closely approximate the values of each environmental covariate.

Transformed aspect density curves are shown in Fig. 6.3. Sample sizes of 100, 200 and 300 closely follow the covariate distribution and approximate the "right leaning" shape of the environmental covariate. Five hundred samples least closely approximates the covariate distribution.

The large peak in the Soil Adjusted Vegetation Index (SAVI, Fig. 6.3) is most closely approximated by a sample size of 500. All sample sizes, except 300, capture the variation in SAVI values from 0.28 to 0.35.

Inverse Wetness Index (IWI) density curve (Fig. 6.3) comparison shows that for the peak at approximately 0.002 the similarity between covariate and sample increases with increasing sample size. The right tail of the covariate is closely approximated by all sample sizes.

All samples sizes closely approximate the slope density curve (Fig. 6.3) with 200, 300 and 500 more similar to the covariate than 50 or 100 samples.

A chi-square goodness of fit test for the combined geology/land cover categorical variable is presented in Table 6.2. All p-values are non-significant at $\alpha = 0.05$ indicating no significant difference between the sample and the environmental covariate.

6.4 Discussion

Based on the above discussion we propose 200–300 as the optimal sample size for this study area as these sample sizes provide the closest approximation of the distribution of all input environmental covariates. This is approximately 0.05–0.1% of the available potential sampling points (~290,000). While not grossly inaccurate, smaller sample sizes of 50 and 100 are not as representative of the whole set of environmental covariates. Optimal sample size also depends on the model used to predict soil distribution. Some predictive models, such as classification and regression trees and some geostatistical methods (see Section 22.4), are "data hungry" and require large amounts of input data. Sample sizes greater than 300 may not provide

significantly better representations of the landscape given the extra cost involved in data collection and may actually worsen the representation of the environmental covariate (see Fig. 6.3, Transformed Aspect, 500 samples).

6.5 Conclusions

Conditioned Latin Hypercube Sampling closely represents the original distribution of the environmental covariate feature space and as such is an appropriate method for selecting sampling sites for digital soil mapping, particularly in areas with limited soil survey budgets. Statistical comparison of multiple sample sizes allows the soil scientist to select an optimal sample size given time/cost, and interpolation method requirements. Approximately 0.05–0.1% of the available sampling area (200–300 samples from ~290,000 potential sampling points) is recommended as an adequate sample size for areas with similar variability in the environmental covariates. While some programming skill is required, we recommend that the soil scientist investigate multiple cLHS sample sizes to select the optimal sample size for mapping other areas.

References

Best, M.G., Lemmon, D.M., and Morris, H.T., 1989. Geologic Map of the Milford Quadrangle and East Half of the Frisco Quadrangle, Beaver County, Utah. USGS, Reston, VA, USA.

Bui, E.N., Simon, D., Schoknecht, N., and Payne, A., 2007. Adequate prior sampling is everything: lessons from the Ord River Basin, Australia, pp. 193–204. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, Vol. 31. Elsevier, Amsterdam.

Chavez, P.S. Jr., 1996. Image-based atmospheric corrections – revisited and revised. Photogrammetric Engineering and Remote Sensing 62(9):1025–1036.

ERDAS, 2006. ERDAS Imagine V9.1. ERDAS, Atlanta, GA, USA.

Howell, D., Kim, Y., Haydu-Houdeshell, C., Clemmer, P., Almaraz, R., and Ballmer, R., 2007. Fitting soil property spatial distribution models in the Mojave Desert for digital soil mapping, pp. 465–475. In: Lagacherie, P., McBratney, A.B., Voltz, M. (eds.), Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, Vol. 31. Elsevier, Amsterdam.

Huete, A.R., 1988. A soil adjusted vegetation index (SAVI). Remote Sensing of the Environment 25:295–309.

IRDIAC (Intermountain Region Digital Image Archive Center), 2008. IRDIAC image search. Retrieved September, 18, 2008 from http://earth.gis.usu.edu.

Mathworks, 2008. Matlab release 2008a. The Mathworks Inc., Natick, MA.

McBratney, A.B., Santos, M.L.M., and Minasny, B., 2003. On digital soil mapping. Geoderma 117(1–2):3–52.

McKenzie, N.J., and Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89(1–2):67–94.

Minasny, B., 2007. Software (cLHS code). Retrieved August, 2008 from http://www.usyd.edu.au/agric/acpa/people/budi/software.htm

Minasny, B., and McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32(9):1378–1388.

Minasny, B., and McBratney, A.B., 2007. Latin hypercube sampling as a tool for digital soil mapping**,** pp. 153–165. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, Vol. 31. Elsevier, Amsterdam.

McKenzie, N.J.N., and Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma, 89(1–2): 67–94.

NED, 2006. USGS National Elevation Dataset. Retrieved March 21, 2009 from http://ned.usgs.gov

NRCS, 2008. National Resource Conservation Service MLRA Explorer. Retrieved September 18, 2008 from http://www.cei.psu.edu/mlra

R Development Core Team, 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org (Last accessed 22 April 2010).

Soil Survey Division Staff, 1993. Soil Survey Manual. Soil Conservation Service. U.S. Department of Agriculture Handbook 18.

Stum, A.K., 2007. Aspect correction tool. Unpublished model. Soil Genesis Lab. Utah State University. Model available upon request.

Tarboton, D., 2008. Terrain Analysis Using Digital Elevation Models (TauDEM). Retrieved June, 8, 2009 from http://hydrology.neng.usu.edu/taudem.

USGS National Gap Analysis Program, 2004. Provisional Digital Land Cover Map for the Southwestern United States. Version 1.0. RS/GIS Laboratory, College of Natural Resources, Utah State University, Logan, UT.

Section B
Soil Sensors and Remote Sensing

Chapter 7
Using Proximal Soil Sensors for Digital Soil Mapping

R.A. Viscarra Rossel, N.J. McKenzie, and M.J. Grundy

Abstract There is a great need for soil data to be used in natural resource assessment and management, e.g. for environmental modelling and digital soil mapping (DSM) to better understand soil processes and reduce risks in decision making. Conventional soil survey cannot efficiently provide these data because the techniques are time consuming and expensive. Proximal soil sensing (PSS) can be used to acquire spatial and temporal data cheaply and with less effort. This paper reviews some the technologies that may be used for PSS and proposes a framework for their use with DSM.

Keywords Proximal soil sensing · Soil sensors · Diffuse reflectance spectroscopy · Soil spatial variability · Digital soil mapping

7.1 Introduction

The need for soil information is greater now than ever before. Concerns over food security and global climate change are transforming agriculture and the way in which we use and manage our soils. Nations are looking at ways to increase food production, increase the sequestration of carbon equivalents (or mitigate the loss of carbon) and identify new opportunities before existing lands become marginal. How we measure, model and map the soil is central to their response, but conventional soil survey cannot efficiently provide the amount of data that is required. The primary reason is well known; when spatial soil information is needed (particularly the functional attributes such as water balance, nutrient supply), conventional soil sampling and laboratory analyses are time-consuming and expensive (e.g. Viscarra Rossel and McBratney, 1998a).

R.A. Viscarra Rossel (✉)
CSIRO Land and Water, Bruce E. Butler Laboratory, GPO Box 1666,
Canberra ACT 2601, Australia
e-mail: Raphael.Viscarra-Rossel@csiro.au

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_7,

Digital soil mapping (DSM) may help to provide this information and at the required resolutions. However, further developments in, and adoption of, DSM depends on (i) the effective use of historical (or legacy) soil information and (ii) the development of rapid and inexpensive techniques to make new soil measurements. In this paper we deal with the latter, and identify technologies that might help. Remote sensing, for example, using multi- and hyper-spectral sensors can partly solve this problem (Gomez et al., 2008). The drawback of remote sensing for soil science (and soil mapping) is that the measurements and the models derived from them only account for the soil surface; in most cases, only the top few millimetres of the soil profile. Proximal soil sensing (PSS) (Viscarra Rossel and McBratney, 1998b) on the other hand can acquire soil sample information rapidly and cheaply and, unlike remote sensing, PSS can be used to measure both surface and subsurface soil properties. For DSM, both remote and proximal soil-sensing approaches are useful and potentially complementary. Examples of uses of remote sensing for DSM are shown in Chapters 8, 9 and 10; here we will discuss the use of proximal soil sensing in DSM.

Much research is being conducted to develop proximal soil sensors and techniques that may be used for proximal soil sensing. This work includes investigation of the use of electromagnetic (EM) induction and electrical resistivity, magnetic susceptibility, ground penetrating radar (GPR), γ-radiometrics, soil colour, diffuse reflectance spectroscopy using visible–near infrared (vis–NIR) and mid infrared (mid-IR) energies, ion-selective field effect transistors (ISFET) and ion-selective electrodes (ISE) and mechanical-draft systems.

The aim of this paper is to discuss the potential for using proximal soil sensing in DSM.

7.2 Digital Soil Mapping and Proximal Soil Sensing

DSM uses soil information and related environmental covariates to fit quantitative relationships that describe soil variation through the creation of digital soil maps (Chapter 2, Lagacherie et al., 2007). Proximal soil sensors are devices that can provide digital, quantitative information for DSM. Although PSS is commonly associated with high-resolution soil mapping (resolutions < 20 m) for applications such as precision agriculture, it can also be used for DSM of larger areas, for example by spatially aggregating the sensor data. McBratney (1998) and Heuvelink and Pebesma (1999) provide relevant discussions on spatial aggregation of soil data. A further advantage of using high-resolution sensor data for DSM is that it can be used to characterize the short-range spatial variation of soil properties, which is often overlooked in DSM due to the limitations of conventional soil sampling and analysis.

The potential for using proximal soil sensing in DSM is shown in Fig. 7.1, using the proposed *scorpan*–soil spatial prediction functions framework described in the review by McBratney et al. (2003). The figure shows where PSS may be used.

To clarify our descriptions, we have defined two types of proximal soil sensors: those that may be used to measure covariates (PSS_C) for *scorpan*, and those that

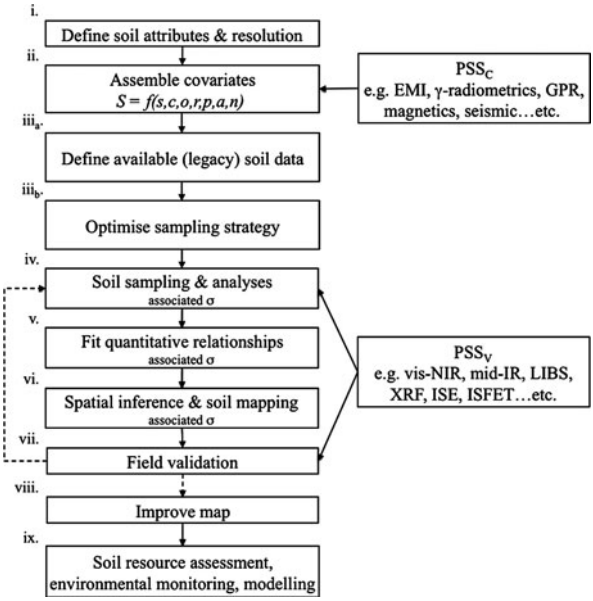


Fig. 7.1 A framework for DSM adapted from McBratney et al. (2003). It shows where proximal soil sensors can be used. We defined two types of proximal soil sensors: those that are used to measure covariates (PSS_C) and those that can be used for measuring soil attributes, or variates (PSS_V)

can be used to measure soil attributes, or variates (PSS_V) (e.g. soil chemical, physical, biological, mineralogical properties). The distinction is not absolute; we cite literature where techniques for PPS_C have been used to measure soil attributes (e.g. Viscarra Rossel et al., 2007). The important points are: (i) PSS can be used to measure covariates for *scorpan* as well as soil attributes, and (ii) PSS overcomes the difficulty and expense of soil measurements associated with conventional sampling and laboratory analysis.

Figure 7.1 shows the typical framework for DSM. The steps and methodologies are well documented by McBratney et al. (2003). In this framework, PSS could be incorporated at steps (ii), (iv) and (vii) (Fig. 7.1).

7.2.1 Acquiring Soil Data Using PSS_C

Environmental covariates to describe the *scorpan* factors (Chapter 2) may be acquired from various sources, such as, RS from satellite or airborne platforms, including imaging spectroscopy (e.g. Chapters 8) and airborne γ-radiometrics (Wilford, 2008). Proximal soil sensors such as those used in geophysical research can also be used. Most PSS_C use near-surface, non-invasive techniques that can be used for "on-the-go" surveying. Digital elevation models (DEM) and derived terrain attributes can also be used as environmental covariates because local variations in terrain affect the movement of sediments, water and solutes within the landscape. Table 7.1 lists some of the methods that might be used to covariates for *scorpan* using PSS_C.

Table 7.1 Proximal soil sensors that can be used to acquire soil covariates for the *scorpan* approach

Sensing technology	Depth (m)	Dependent property	Inferred properties	Technological development	Methodological development	Approximate cost $US
Electromagnetic induction (EMI)	<1–6 m	Resistivity, magnetic, permeability, permittivity	Conductivity, soil water and solutes, texture, temperature, salinity	Mature	Mature	>20 k
Electrical resistivity	<1–<30 m	Resistivity	Conductivity, soil water and solutes, texture, temperature, salinity	Mature	Mature	>20 k
Induced polarization	<1–<50 m	Resistivity, capacitance	Conductivity and polarization, soil water, hydraulic properties, lithology	Mature	Developmental	>50 k
Magnetics	<1 m to <10 km	Magnetization	Magnetic minerals Fe-oxides, structure, lithology	Mature	Mature/researchable	>40 k
Gravity	>1 m to <100 km	Density	Lithology, hydraulic properties	Mature	Researchable	>60 k
Seismic	<1 m to <500 km	Elastic moduli, acoustic impedance, density	Soil layering, soil structure, soil depth, soil density, lithology	Mature	Researchable	>50 k
Ground penetrating radar (GPR)	<1 m to <10 m	Resistivity, magnetic permeability, permittivity	Soil water, texture, soil depth	Mature	Researchable	>40 k
Magnetic resonance sounding	<1 m to <1.5 km	Proton density	Soil water, porosity, hydraulic properties	Emerging	Developmental	
γ-radiometrics	<1 m	Radioisotopes of Cs, K, U, Th	Total K, mineralogy, clay content, soil type	Mature	Researchable	>30 k

EM induction is a highly adaptable non-invasive technique that measures the apparent bulk electrical conductivity of the soil (EC_a). It has been used extensively in soil science since De Jong et al. (1979) first reported it. EM induction is particularly useful for mapping saline soils and for precision agriculture.

Electrical resistivity is used to determine the resistivity of the measured soil volume. Measurements of resistivity usually require four electrodes; two electrodes are used to apply the current (current electrodes) and two are used to measure the resulting potential difference (potential electrodes). The electrical resistivity of the soil is determined from these data and measurements of the electrical conductivity (EC_a) are possible because resistivity is the reciprocal of conductivity. Samouëlian et al. (2005) reviews the use of electrical resistivity surveys in soil science.

Induced polarisation (IP) measurements are essentially an extension of the four-electrode resistivity technique. IP operates by first applying an electric current between a current electrode pair and the resulting voltage induced in the soil is measured between a potential electrode pair. However, IP captures both the charge loss (conduction) and the charge storage (capacitance) characteristics of the soil. IP instruments have been used in hydrogeophysical applications, e.g. to look at hydraulic properties of soil in the vadose zone (Börner et al., 1996).

Magnetics sensors, or, magnetometers, measure variations in the strength of the earth's magnetic field and the data reflect the spatial distribution of magnetisation throughout the ground. Magnetisation of naturally occurring materials and rocks is determined by the quantity of magnetic minerals and by the strength and direction of the permanent magnetisation carried by those minerals (Hansen et al., 2005). Typically, magnetics has been used for the detection of geological bodies. However, there is increasing use of the technique for near-surface applications, for example, to better understand soil genesis and formation (Mathé and Lévêque, 2003); to detect anthropogenic pollution on topsoils; through their associations with Fe-oxides; and for rapid identification and mapping of soil heavy metal contamination (Jordanovaa et al., 2008).

Gravity data can be collected using gravimeters (or gravitometers) and provide information on the local gravitational field. There are two types: relative and absolute gravimeters. A relative gravimeter measures relative differences in the vertical component of the earth's gravitational field based on variations in the extension of an internal spring in the gravimeter. The technique has typically been used to determine the subsurface configuration of structural basins, aquifer thickness and geological composition. An absolute gravimeter measures the acceleration of free fall of a control mass. Absolute gravimetry can be used to measure mass water balances at regional or local scales (Nabighian et al., 2005).

Seismic reflection methods are sensitive to the speed of propagation of various kinds of elastic waves. The elastic properties and mass density of the medium in which the waves travel control the velocity of the waves and can be used to infer properties of the earth's subsurface. Reflection seismology is used in exploration for hydrocarbons, coal, ores, minerals, and geothermal energy. It is also used for basic

research into the nature and origin of rocks that make up the Earth's crust. It can be used in near-surface application for engineering, groundwater and environmental surveying. A method similar to reflection seismology, which uses electromagnetic waves is ground penetrating radar (GPR).

Ground penetrating radar uses the transmission and reflection of high frequency (1 MHz to 1 GHz) electromagnetic waves in the soil. The resolution of GPR images can be varied through the use of different antennae frequencies. Typically, higher frequencies increase the resolution at the expense of depth of penetration. Daniels et al. (1988) describes the fundamental principles of GPR. Knight (2001) provides an overview of GPR in environmental applications and Huisman et al. (2003) reviews its use for soil water determinations.

Magnetic resonance sounding uses a nuclear magnetic resonance (NMR) principle that is used in medical brain scanning (i.e. MRI or magnetic resonance imaging) to measure subsurface free water and hydraulic properties (Lubczynski and Roy, 2004). It is also known as surface nuclear magnetic resonance and can be used to measure water content and porosity to depths up to 1,500 m. Paetzold et al. (1985) used the technique to measure soil water content and concluded that the NMR signal is a linear function of volumetric water content and is not affected by clay mineralogy, soil organic matter, or texture. They concluded that the NMR signal is uniquely related to liquid water in soils and rocks.

The basis of passive γ-radiometrics is that γ-ray photons have discrete energies, which are characteristic of the naturally occurring radioactive isotopes from which they originate. By measuring the energies of these photons, it is possible to determine the source of the radiation. While many naturally occurring elements have radioactive isotopes, only potassium (^{40}K), caesium (^{137}Cs) and the decay series of uranium (^{238}U and ^{235}U and their daughters) and thorium (^{232}Th and its daughters) have long half-lives, are abundant in the environment, and produce γ-rays of sufficient energy and intensity to be measured by γ-radiometric sensors. The conventional approach to the acquisition and processing of γ-ray data is to monitor four broad spectral windows or regions of interest (ROI) corresponding to K, U, Th and the total count (e.g. Viscarra Rossel et al., 2007).

7.2.2 Acquiring Soil Attributes Using PSS$_V$

Proximal soil sensors can also be used as alternatives to conventional soil sampling and laboratory analyses, which as we have mentioned, are slower, can be difficult and more expensive than PSS. The rationale here is that although PSS may produce results that are not as accurate as conventional laboratory analysis, it facilitates the collection of larger amounts of quantitative spatial data using cheaper, simpler and less laborious techniques. Furthermore, the measurements may be made in situ, providing information of the soil at field conditions and in a timely manner. The interaction of PSS with GPS allows the signals to be processed and mapped digitally. Table 7.2 provides a summary of the techniques that are currently being developed to measure various soil properties, showing the approximate frequency

Table 7.2 Proximal soil sensors that can be used to measure soil attributes

EM Range	γ –rays			X-rays		UV – Vis – IR					Micro and Radio waves								
Frequency (Hz)	10^{22}			10^{18}		10^{16}	10^{15}	10^{13}	10^{12}		10^{10}	10^{8}	10^{7}			10^{6}	10^{2}			
Wavelength (m)	10^{-12}			10^{-10}		10^{-8}	10^{-6}	10^{-5}	10^{-4}		10^{-2}	10^{1}	10^{2}	10^{3}	10^{6}					
Technique	INS	Active γ	Passive γ	XRF	XRD	UV	Vis	NIR	mid-IR	LIBS	Micro	WSN	TDR	FDR	Capac	GPR	EMI	ER	ECh	Mech
Water content	P	P					c	P	P		P	P	P	P	P	P	C	C		
Nitrogen	P					P			P	P									P	
Phosphorus	P		c					c	c										P	
Potassium	P		P	P				c	c	P									P	
Other major nutrients	P			P				c	C	P									P	
Minor nutrients, elements	P			P		P	c	c	c	P									P	
Heavy metals	P			P				C	C	P										
Salinity and sodicity																	P	P	P	
CEC					c			C	C								c	c		
Soil pH								c	C										P	
Buffering capacity and lime requirement								c	C										C	
Carbon	P						C	P	P	P										
Clay content			C				c	C	C								C	C		
Silt content			c				c	c	c								c	c		
Sand content			C						P								c	c		
Clay minerals	C		c	C	P			P	P	C										
Fe-oxides	P		C	P		P	P	P	P	C										
Soil strength																				P
Bulk density		C																		C
Compaction		C														c	c	c		C
Rooting depth																c				

P indicates that the measurements are physically based while C indicates that the measurement is correlative. Lower case c indicates weak inference

This table serves only as a guide. It was derived from the literature and is not definitive as new methods are continually being developed

INS = inelastic neutron scattering; XRF = X-ray fluorescence; XRD = X-ray diffractometry; UV = ultra violet; Vis = visible; NIR = near infrared; mid-IR = mid infrared; LIBS = laser induced breakdown spectroscopy; Micro = microwave sensors; WSN = wireless sensor networks; TDR / FDR = time / frequency domain reflectometry; Capac = capacitance; GPR = ground penetrating radar; EMI = electromagnetic induction; ER = electrical resistivity; ECh = electrochemical; and Mech = mechanical

and wavelengths at which these sensors operate and whether the measurement is physically based (P) or correlative (C).

There are various techniques that can be used to measure soil attributes (Table 7.2). These range from those that are direct and physically based to those that are indirect and correlative (Table 7.2). For example, soil pH can be measured directly using ion-sensitive field effect transistors (ISFET) (Viscarra Rossel and Walter, 2004), or indirectly using vis–NIR spectroscopy. Indirect techniques rely on empirical calibrations and are therefore invariably less accurate than direct methods. Indirect methods also tend to work better under restricted local soil conditions and often fail when generalised. This conditional success is controlled by the type of soil: its mineralogy, particle-size distribution, presence of segregations (e.g. iron oxides and oxyhydroxides), soluble salts, water content, and the abundance and composition of organic matter. However, indirect techniques are also generally less expensive, technologically and methodologically better developed and more readily available to users.

7.2.3 Diffuse Reflectance Spectroscopy (DRS)

There is widespread interest for using DRS in DSM because it is rapid, requires minimal sample preparation, is non-destructive and requires no hazardous chemicals (Viscarra Rossel et al., 2006). However, the principal reason for this interest is that unlike other sensing techniques, several soil properties can be measured from a single scan (Table 7.2). This multi-parameter feature of diffuse reflectance spectroscopy implies that one spectrum holds information about various soil constituents and indeed, soil mid-IR and vis–NIR spectra are sensitive to both organic and inorganic soil composition. Nonetheless, for the most part, the techniques are correlative so that soil attributes cannot be measured directly from the spectra (Table 7.2). In these cases, to be useful quantitatively, spectra must be related to a set of known reference samples through a calibration model. The set of reference samples used in the models need to be representative of the range of soils in which the models are to be used. A discussion on the use of diffuse reflectance spectroscopy for DSM can be found in Viscarra Rossel and McBratney (2008).

An initiative to develop a global soil spectral library was started in April 2008 (Viscarra Rossel, 2009a). The aim was to help soil spectroscopy progress from being almost purely a research tool to become a more widely adopted and useful technique for soil analysis, PSS, soil monitoring and DSM. Currently, the spectral library contains 14695 vis–NIR (350–2,500 nm) spectra from 89 countries (Fig. 7.2), most with laboratory measurements of soil organic carbon and clay content. The library also contains laboratory measurements of other soil attributes (e.g. pH, CEC, carbonate content) as well as soil classification by FAO-WRB (FAO, 1998), land use and spatial co-ordinates. All spectra, however, do not have the full set of these data. The library is growing as more countries are joining the initiative.

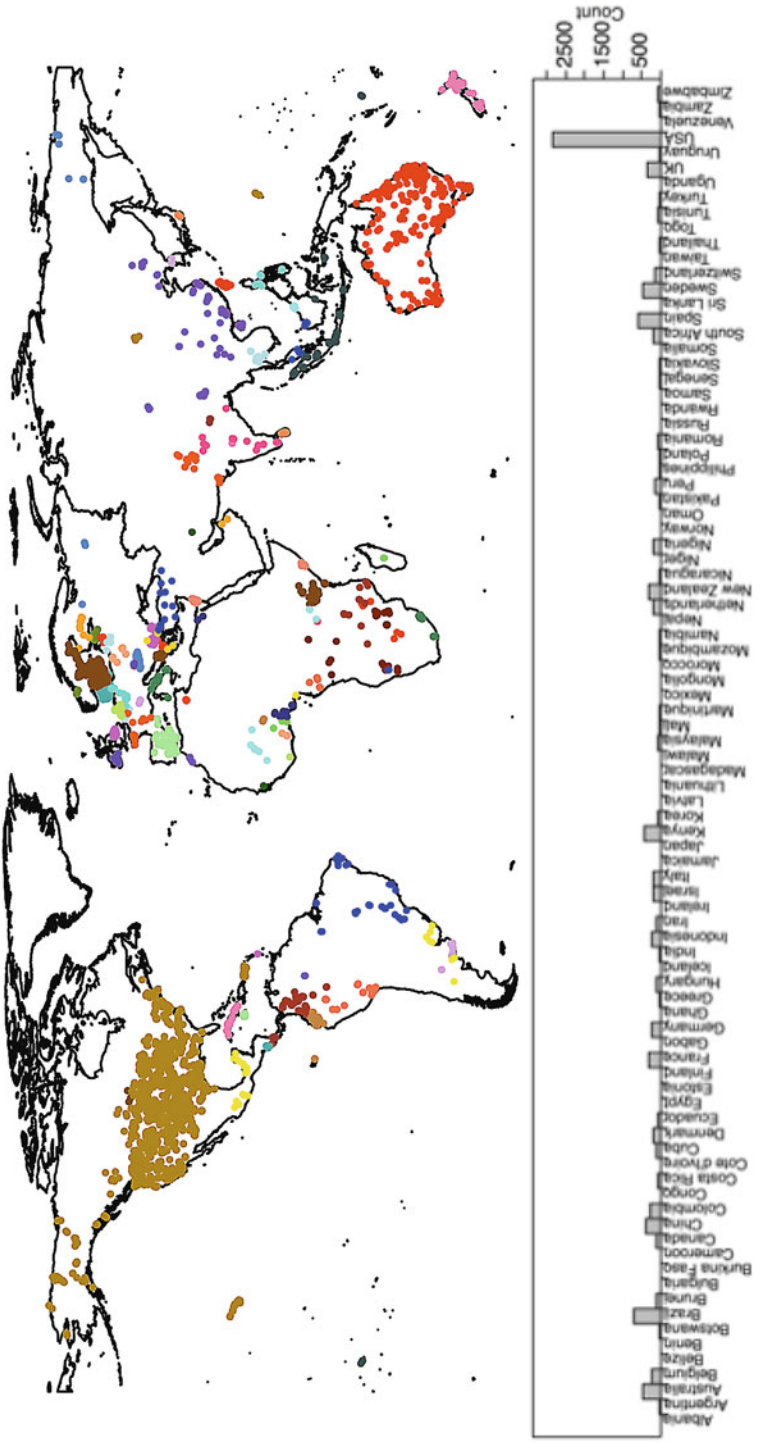


Fig. 7.2 Development of a global soil spectral library. The number of spectra from each of the 89 countries in the library. From Viscarra Rossel (2009a)

7.2.4 Multi-sensor Platforms

As every soil sensing technology has strengths and weaknesses and no single sensor can measure all soil properties (Table 7.2), the selection of a complementary set of sensors to measure the required suite of soil properties is important. Integrating multiple proximal soil sensors of both types (PSS_C and PSS_V) in a single multi-sensor platform can provide a number of operational benefits over single-sensor systems, such as:

- robust operational performance
- increased confidence as independent measurements are made on the same soil
- extended attribute coverage
- increased dimensionality of the measurement space (i.e., different sensors measuring various portions of the electromagnetic spectrum).

There are few reports of multi-sensor systems in the literature. For example, Christy et al. (2004) reported the use of a mobile sensor platform that simultaneously measures soil pH and EC. An NIR sensor has also recently been added to this multi-sensor platform (Christy, 2007). Taylor et al. (2006) reported the development of a multi-sensor platform consisting of two EMI instruments, an ER sensor, a γ-radiometer and a high resolution DGPS. Yurui et al. (2008) reported the development of a multi-sensor technique for measuring soil physical properties (soil water, mechanical strength and EC).

7.3 A Framework for Proximal Soil Sensing in Digital Soil Mapping

Figure 7.3 shows a framework for DSM using proximal soil sensors and a global soil spectral library.

Briefly, once the *scorpan* layers are assembled, the legacy data defined, one may select areas across the landscape where the sensing will take place. The multi-sensor platform is used to quantify soil variability in these areas using a suite of PSS_C. For example, PSS_C using a γ-radiometer and EM induction instruments may be used to characterise soil variability of areas in the landscape (Fig. 7.4). These data can be spatially aggregated and combined with the *scorpan* layers to help characterise soil spatial variability at different scales and to determine an optimal sampling strategy. The multi-sensor platform moves to each location and samples a soil core. A suite of PSS_V is used to measure the soil properties of the core at specific depth intervals (e.g. 1 m down the soil profile) and transfer functions are used to estimate the soil properties of interest. For example, using soil vis–NIR spectra to measure clay content and mineral composition (Fig. 7.5a) or mid-IR to measure soil organic carbon content (Fig. 7.5b). Spectroscopic estimates of clay content and organic carbon are made using soil spectral libraries (Fig. 7.3). The PPS_V allow collection of many more observations than would be possible using conventional

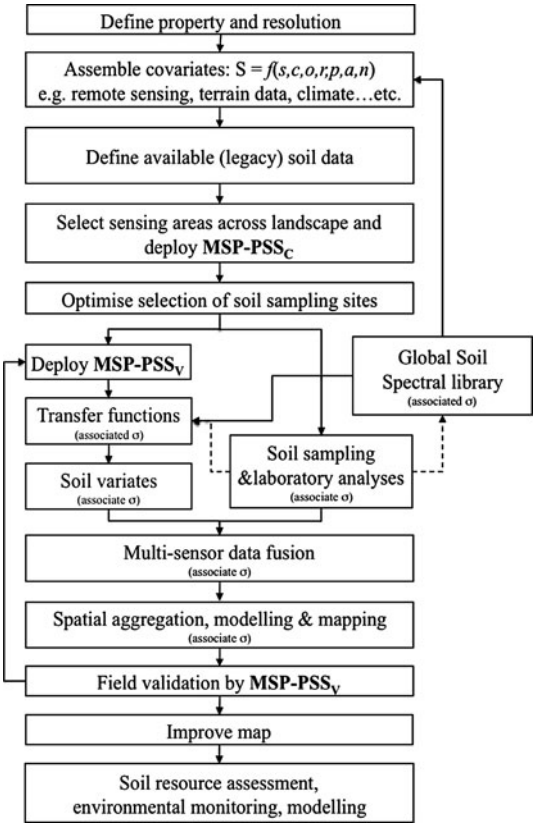


Fig. 7.3 A framework for digital soil mapping using proximal soil sensors (PSS). Here, PSS to measure soil covariates (PSS_C) and PSS to measure soil attributes (PSS_V) are mounted on a multi-sensor platform (MSP). A global spectral library can be used with the spectroscopic proximal sensors and hyperspectral remote sensors, for prediction of soil properties

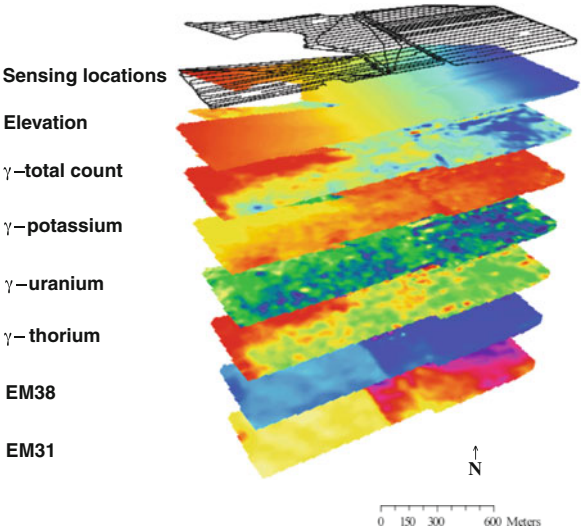


Fig. 7.4 Proximal soil sensors can be used to measure covariates (PSS_C) to quantify soil spatial variability in selected areas across a landscape. The figure shows the sensing locations, the DEM of the area, the total count, potassium, uranium and thorium regions of interest from a γ-radiometer, and the EC at different depths from an EM-38 and an EM-31. Data from Viscarra Rossel et al. (2007)

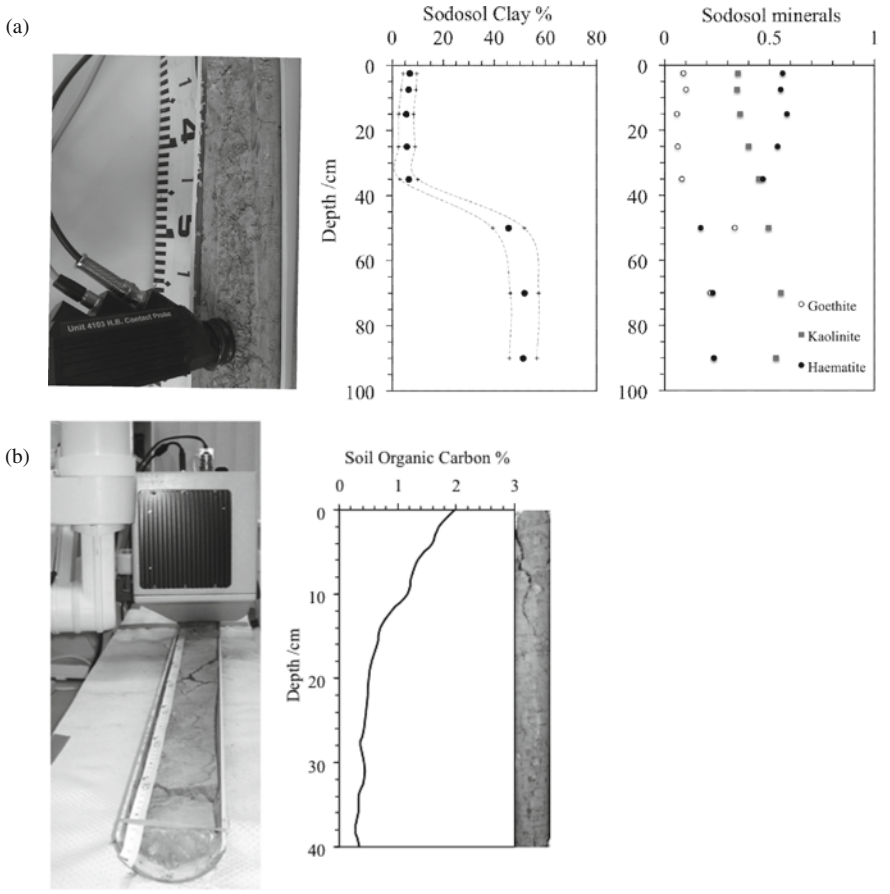


Fig. 7.5 Proximal soil sensing of (**a**) soil clay content and mineral composition using a vis–NIR spectrophotometer, adapted from Viscarra Rossel et al. (2009b) and (**b**) soil organic carbon using a mid-IR spectrophotometer

soil sampling and analyses. Nonetheless, at selected sites, soil samples are collected and taken back to the laboratory for analysis and verification. Of course, this additional data will also be used to improve the global spectral libraries and transfer functions in the current and future surveys. Importantly, a greater amount of spatial data is collected using the multi-sensor platform. The spectral libraries may also be used in the laboratory analysis of the sampled soil (e.g. soil organic carbon by laboratory-based mid-IR spectroscopy). Soil observations collected by the multi-sensor platform are then combined using multi-sensor data fusion. The data may then be spatially aggregated to the required support for modelling and mapping. The spatial information collected from the soil (i.e. both horizontally and vertically down the profile) can be used for the assessment of soil properties and processes.

7.4 Conclusions

Digital soil mapping encompasses a powerful suite of technologies that can be used for soil resource assessment and management. However, it is limited by expensive, time-consuming fieldwork and laboratory analysis. This paper suggests that:

- Further developments in, and adoption of, digital soil mapping depends on the use of proximal soil sensors to overcome the constraints imposed by conventional soil survey
- There are various techniques that can be used for proximal soil sensing, which can be used to measure covariates for *scorpan* and soil chemical, physical, biological, mineralogical attributes
- No single sensor can measure all soil properties, therefore integrating multiple proximal soil sensors in a single multi-sensor platform can provide operational benefits over single-sensor systems
- There is widespread interest in the use of diffuse reflectance spectroscopy because several soil properties can be measured from a single scan. Development of the global soil spectral library is important and will enhance the adoption of DSM.
- Proximal sensing can be integrated into a digital soil mapping framework.

References

Börner, F.D., Schopper, J.R., and Weller, A., 1996. Evaluation of transport and storage properties in the soil and groundwater zone from induced polarization measurements. Geophysical Prospecting 44:583–602.

Christy, C.D., 2007. Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. Computers and Electronics in Agriculture 61:10–19.

Christy, C.D., Collings, K., Drummond, P., and Lund, E., 2004. A mobile sensor platform for measurement of soil pH and buffering. Paper Number 041042, ASAE Annual Meeting. American Society of Agricultural and Biological Engineers, St. Joseph, Michigan.

Daniels, D.J., Gunton, D.J., and Scott, H.F., 1988. Introduction to subsurface radar. IEEE Proceedings – F Radar and Signal Processing 135:278–320.

De Jong, E., Ballantyne, A.K., Cameron, D.R., and Read, D.W.L., 1979. Measurement of apparent electrical conductivity of soils by an electromagnetic induction probe to aid salinity surveys. Soil Science Society of America Journal 43:810–812.

FAO, 1998. World Reference Base for Soil Resources. Food and Agriculture Organization of the United Nations, Rome.

Gomez, C., Viscarra Rossel, R.A., and McBratney, A.B., 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis–NIR spectroscopy: An Australian case study. Geoderma 146:403–411.

Hansen, R.O., Racic, L., Grauch, V.J.S., 2005. Magnetic methods in near-surface geophysics, pp. 151–175. In: Butler, D.K. (ed.), Near-Surface Geophysics: Investigations in Geophysics No. 13. Society of Exploration Geophysicists, Tulsa, OK.

Heuvelink, G.B.M., and Pebesma, E., 1999. Spatial aggregation and soil process modeling. Geoderma 89:47–65.

Huisman, J.A., Hubbard, S.S., Redman, J.D., and Annan, A.P., 2003. Measuring soil water content with ground penetrating radar: a review. Vadose Zone Journal 2:476–491.

Jordanovaa, N., Jordanovaa, D., and Tsachevab, T., 2008. Application of magnetometry for delineation of anthropogenic pollution in areas covered by various soil types. Geoderma 144:557–571.

Knight, R., 2001. Ground penetrating radar for environmental applications. Annual Review of Earth and Planetary Sciences 29:229–255.

Lagacherie, P., McBratney, A.B., and Voltz, M., 2007. Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, Vol. 31. Elsevier, Amsterdam.

Lubczynski, M., and Roy, J., 2004. Magnetic resonance sounding: new method for ground water assessment. Ground Water 42:291–303.

Mathé, V., and Lévêque, F., 2003. High resolution magnetic survey for soil monitoring: detection of drainage and soil tillage effects. Earth and Planetary Science Letters 212:241–251.

McBratney, A.B., 1998. Some considerations on methods for spatially aggregating and disaggregating soil information. Nutrient Cycling in Agroecosystems 50:51–62.

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

Nabighian, M.N., Ander, M.E., Grauch, V.J.S., Hansen, R.O., LaFehr, T.R., Li, Y., Pearson, W.C., Peirce, J.W., Phillips, J.D., and Ruder, M.E., 2005. The historical development of the gravity method in exploration. Geophysics 70:63ND–89ND.

Paetzold, R.F., Matzkanin, G.A., and De Los Santos, A., 1985. Surface soil-water content measurements using pulsed nuclear magnetic resonance techniques. Soil Science Society of America Journal, 49: 537–540.

Samouëlian, A., Cousin, I., Tabbagh, A., Bruand, A., and Richard, G., 2005. Electrical resistivity survey in soil science: a review. Soil and Tillage Research 83:173–193.

Taylor, J.A., McBratney, A.B., Viscarra Rossel, R.A., Minasny, B., Taylor, H., Whelan, B., and Short, M., 2006. Development of a multi-sensor platform for proximal soil sensing. 18th World Congress of Soil Science July 9–15, 2006 – Philadelphia, Pennsylvania, USA.

Viscarra Rossel, R.A., and McBratney, A.B., 1998a. Soil chemical analytical accuracy and costs: implications from precision agriculture. Australian Journal of Experimental Agriculture, Special Issue – Moving towards precision with soil and plant analysis. 38:765–775.

Viscarra Rossel, R.A., and McBratney, A.B., 1998b. Laboratory evaluation of a proximal sensing technique for simultaneous measurement of clay and water content. Geoderma 85:19–39.

Viscarra Rossel, R.A., and Walter, C., 2004. Rapid, quantitative and spatial field measurements of soil pH using an ion sensitive field effect transistor. Geoderma 119:9–20.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., and Skjemstad, J.O., 2006. Visible, near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131:59–75.

Viscarra Rossel, R.A., Taylor, H.J., and McBratney, A.B., 2007. Multivariate calibration of hyperspectral γ-ray energy spectra for proximal soil sensing. European Journal of Soil Science 58:343–353.

Viscarra Rossel, R.A., and McBratney, A.B., 2008. Diffuse reflectance spectroscopy as a tool for digital soil mapping, Chapter 13, pp. 165–172. In: Hartemink A.E., McBratney, A.B., and Mendonça-Santos, L. (eds.), Digital Soil Mapping with Limited Data. Developments in Soil Science Seris. Elsevier Science, Amsterdam.

Viscarra Rossel, R.A., 2009a. The soil spectroscopy group and the development of a global soil spectral library. NIR News 20:14–15.

Viscarra Rossel, R.A, Cattle, S.R., Ortega, A.S., and Fouad, Y., 2009b. In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. Geoderma 150:253–266.

Wilford, J., 2008. Remote sensing with gamma-ray spectrometry, pp.189–200. In: McKenzie, N.J., Grundy, M.J., Webster, R., and Ringrose-Voase, A.J. (eds.), Guidelines for Surveying Soil and Land Resources, Second edition. CSIRO Publishing, Collingwood, Victoria.

Yurui, S., Schulze-Lammers, P., Daokun, M., Ianhui, L., and Qingmeng, Z., 2008. Determining soil physical properties by multi-sensor technique. Sensors and Actuators A 147:352–357.

Chapter 8
The Use of Hyperspectral Imagery for Digital Soil Mapping in Mediterranean Areas

P. Lagacherie, C. Gomez, J.S. Bailly, F. Baret, and G. Coulouma

Abstract Hyperspectral imagery is considered as a promising source of data to overcome the lack of soil information that often hamper digital soil mapping. We have tested it in the vineyard plain of Languedoc (southern France) using an 5×5 m resolution HYMAP image and 52 calibration-validation points. Satisfactory predictions of clay content and calcium carbonate ($CaCO_3$) content were first obtained from HYMAP spectra over bare soils, partial least-squares regression performing better than continuum removal technique. These predictions were however less precise than using laboratory spectra. An examination of the possible factors that could explain this decrease showed that calibration uncertainties of the HYMAP sensor and of atmospheric effects were largely predominant. Secondly, since the HYMAP image was largely covered by vegetation with few pure bare soil pixels, an interpolation-aggregation procedure was proposed to obtain a 100×100 m digital soil map of the whole study area from a set of scattered bare soil fields with hyperspectral soil characterization. Interpolation was performed by a conditional simulation algorithm to estimate the within pixel soil pattern parameters. Validation results showed that satisfactory estimates of local means can be obtained whereas the variations of local variances were only partly represented.

In the near future, a new proof-of-concept zone will be implemented in Tunisia to confirm these encouraging results and to examine how hyperspectral imagery can be used in association with soil legacy data and digital terrain models to produce digital maps of soil properties in the Mediterranean areas.

Keywords Remote sensing · Reflectance spectroscopy · Partial least square regression · Uncertainty · Upscaling

P. Lagacherie (✉)
INRA Laboratoire d'étude des Interactions Sol Agrosystème Hydrosystème (LISAH), UMR 1221 INRA-IRD-Supagro Montpellier, Montpellier, France
e-mail: lagache@supagro.inra.fr

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_8,

8.1 Introduction

It has been largely recognized in the previous digital soil mapping workshops (Lagacherie, 2008) that poor soil datasets have been (and still remain) a factor that can severely limit digital soil mapping progress. It is thus important to include the use of soil sensors that can deliver precise soil property estimates over large areas in the digital soil mapping toolbox. Among the large set of possible soil sensors, visible and near infrared (vis-NIR) hyperspectral imagery is one of the most promising candidates since it is derived from reflectance spectroscopy, a laboratory technique that was proved as being a good alternative to the costly soil physical and chemical laboratory analysis for the estimation of a large range of soil properties (Viscarra Rossel et al., 2006; see also Section 7.2.3), and the few studies that exist in the literature show promising results (e.g., Ben-Dor et al., 2008; Gomez et al., 2008b). It is important to note that hyperspectral imagery can estimate the properties of the immediate soil surface only (e.g., the first millimetres). However, it provides a new soil covariate layer that may serve to estimate soil properties of deeper soil layers and to predict soil classes.

Hyperspectral imagery looks particularly promising in Mediterranean areas where bare soil surfaces and dry soil conditions are frequent, easing the interpretation of hyperspectral images. This is why our research team was early to work (since 2003) on hyperspectral imagery. The first exploratory results were presented in the first digital soil mapping workshop in Montpellier, France in 2004 (Madeira et al., 2007). Since then, further research has been undertaken on three questions that must be addressed for an effective use of hyperspectral imagery in digital soil mapping: (i) How to derive soil property estimates from Vis-NIR spectra? (Gomez et al., 2008a; Lagacherie et al., 2008), (ii) what are the main perturbing factors when passing from laboratory to remote sensing spectra? (Lagacherie et al., 2008), and (iii) how to obtain a digital soil map from a hyperspectral image with high fraction of pixels covered by the vegetation?

This paper summarizes the main findings of the research listed above.

8.2 The "Peyne Experiment"

8.2.1 Study Area

The study was carried out in la Peyne catchment, South of France (43° 29′ N and 3°22′ E) (Fig. 8.1), which is dominated by vineyard land cover. Marl, limestone and calcareous sandstones coming from Miocene marine and lacustrine sediments (Fig. 8.2) formed the parent material of several soil types observed in this area: Lithic Leptosols, Calcaric Regosols and Calcaric Cambisols. These sediments were partly covered by successive alluvial deposits ranging from the Pliocene to Holocene and differing in their initial nature and in the duration of weathering

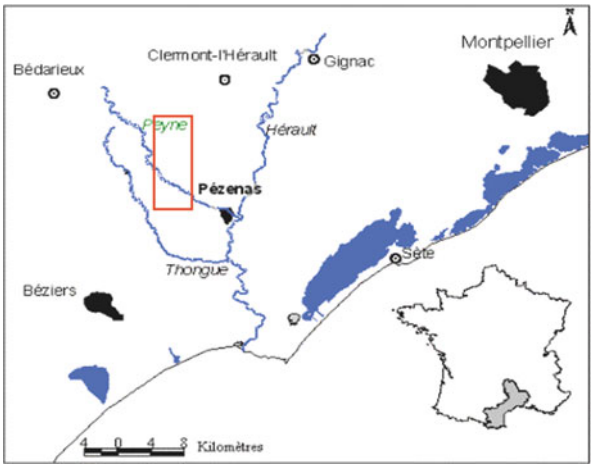


Fig. 8.1 Location of the study area

conditions (Fig. 8.2). They have produced an intricate soil pattern that includes a large range of soil types such as Calcaric, Chromic and Eutric Cambisols, Chromic and Eutric Luvisol and Eutric Fluvisols. Local transport of colluvial material along slopes added to the complexity of the soil pattern.

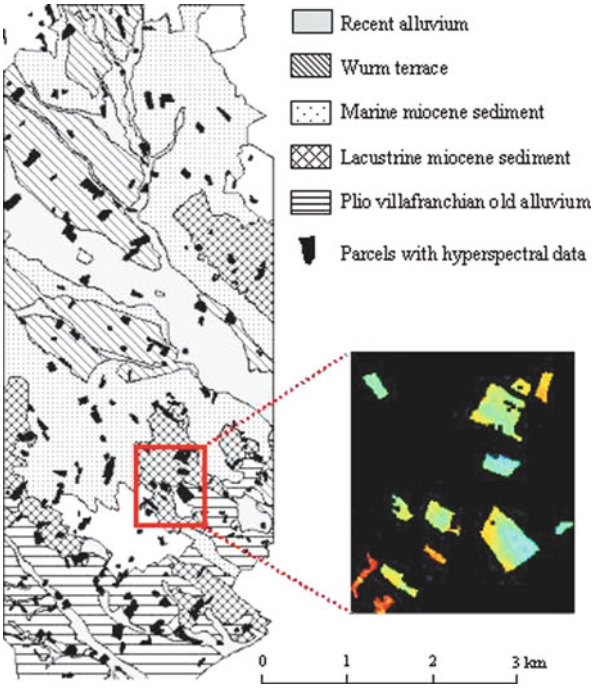


Fig. 8.2 Geology of the study area and the masked image showing only data on bare soil fields

8.2.2 The Hymap Image

The HYMAP airborne imaging spectrometer measured reflected radiance in 126 non-contiguous bands covering the 400–2,500 nm spectral domain with around 19 nm bandwidths and average sampling intervals of 17 nm in the 1,950–2,480 nm domain. The HYMAP image was acquired on July 13th 2003 with a spatial resolution of 5 × 5 m. This image was geometrically, atmospherically and topographically corrected (see details in Lagacherie et al., 2008).

Living (essentially vineyard) and dry vegetation were masked on the HYMAP image using respectively the NDVI and the cellulose absorption band (2,010 nm). The final result was a masked image covering 23.5 km^2 with data only on bare soil fields (Fig. 8.2).

8.2.3 Field Data

Clay and $CaCO_3$ contents were selected as examples of target soil properties since they are widely used by soil surveyors to describe soil types and are essential to quantify the soil erodibility. These two properties were determined by routine soil analysis for 52 sites collected over bare soil fields. The sampling was designed to capture the variability of the properties of interest within the study area (clay contents from 65 to 452 g/kg and $CaCO_3$ contents from 0 to 360 g/kg). Laboratory vis-NIR reflectance spectra were recorded for these 52 samples with an ASD pro FR Portable Spectro-radiometer, and field vis-NIR reflectance spectra were recorded for the 19 samples collected in 2005 with the same tool. Detailed measurement protocols are available in Lagacherie et al. (2008).

8.2.4 Methods

Two well-known techniques were used to infer the soil properties from laboratory and HYMAP spectra; the continuum removal analysis (CR) and the partial least square regression (PLSR).

CR is a means of normalizing reflectance spectra to allow comparison of individual absorption features from a common baseline (Clark and Roush, 1984). The CR technique presents the advantage of targeting specific absorption features that should be resistant across scales and observation conditions. After continuum removal, absorption band depth values are calculated from vis-NIR spectra to estimate mineral, rock, and soil properties. Specific absorption features of minerals, rocks, and soil properties, including clay and $CaCO_3$, have been widely studied under laboratory conditions. The absorption band depth values at 2,206 and 2,341 nm, calculated from laboratory spectra after continuum removal, can be used to estimate clay and $CaCO_3$ content, respectively.

PLSR is one of the most common multivariate statistical techniques for spectral calibration and prediction of soil properties (e.g., Viscarra Rossel et al., 2006). In the PLSR approach, the full spectrum is used to establish a linear regression model where the significant information contained in the vis-NIR spectra is concentrated in a few latent variables that are optimized to produce the best correlation with the desired property of interest.

8.3 Results

8.3.1 How to Derive Soil Property Estimates from Vis-NIR Spectra? (Gomez et al., 2008a; Lagacherie et al., 2008)

Table 8.1 shows the performance of the two tested methods for predicting clay and calcium carbonate content from laboratory and remote sensing (HYMAP) spectra. It shows that the two soil properties can be predicted with an acceptable precision although a decrease of precision is observed when passing from laboratory to HYMAP spectra. The PLSR technique performs better than the CR approach when the absorption peak selected in the CR approach corresponds to a chemical species that does not match perfectly the soil property of interest (e.g., OH^- for clay content), or when applied to the lower quality spectra provided by an airborne sensor like HYMAP. In these situations, PLSR is able to find surrogate spectral features that retain satisfactory estimations of the studied soil properties. However, these surrogate spectral features correspond to soil properties that have only area-specific correlations with the soil property of interest (e.g., soil colour with clay content), or to chemical species that cannot be related with any explainable soil features. This means that extrapolations to larger pedological contexts must be envisaged with care (Gomez et al., 2008a). Conversely, in the case of $CaCO_3$ estimations from laboratory spectra, CR overcame PLSR since it nearly equals the PLSR results while using a more parsimonious model. Although these results only concern two soil properties, this will certainly be extended in the near future to other soil properties that are known to be suitable for spectrometry analysis (granulometric fractions, carbon content, iron content, salinity, ...).

Table 8.1 Estimations of Clay and $CaCO_3$ contents of soil surface from laboratory and HYMAP spectra using the continuum removal method and the PLSR regression (after Gomez et al., 2008a)

		Laboratory spectra		HYMAP spectra	
		Clay	$CaCO_3$	Clay	$CaCO_3$
CR	R^2	0.73	0.92	0.58	0.47
	RMSE (g/kg)	44	52	82	132
PLSR	R^2	0.85	0.94	0.64	0.77
	RMSE (g/kg)	31	39	50	77

Abbreviations: CR: continuum removal method, PLSR : partial least squares regression method, RMSE: root mean square error, R^2: determination coefficient (obtained by cross-validation)

8.3.2 What Are the Main Perturbing Factors When Passing from Laboratory to Remote Sensing Spectra? (Lagacherie et al., 2008)

Nine intermediate stages from the laboratory up to HYMAP sensor measurements were considered for separately evaluating the factors of confusion that may decrease the estimation performances when going across scales and sensors (Fig. 8.3). These stages were either characterised by additional measurements or by simulations (see details in Lagacherie et al., 2008). The importance of each factor of confusion was evaluated by examining the correlations between the soil property predictions using the CR method with data from the two corresponding consecutive measurement stages.

Results show that the main uncertainty factor in scaling up laboratory to airborne measurements is the capacity of airborne reflectance measurements to be spectrally consistent (good inter-band relative radiometric accuracy), and well corrected from atmospheric effects, particularly regarding water vapour (first line in Fig. 8.3). A small effect on the degradation of the radiometric performances was observed as well. The influence of pebbles was found generally limited except for $CaCO_3$ estimates with calcareous pebbles and high $CaCO_3$ contents. More sites

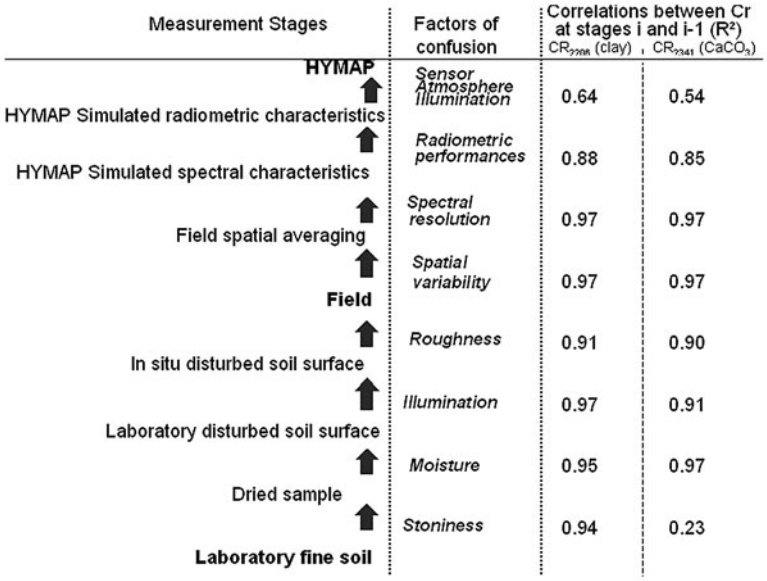


Fig. 8.3 The different possible factors of confusions across scale and sensors and their influence on soil property predictions. A low R^2 between consecutive stages means a strong influence of the corresponding confusion factor (after Lagacherie et al., 2008)

will be necessary to identify situations where pebbles have to be taken into consideration. The other factors considered (i.e., spectral and spatial resolutions, surface roughness, illuminations conditions and soil moisture) played a minor role.

8.3.3 How to Obtain a Digital Soil Map from a Hyperspectral Image Partly Covered by the Vegetation?

An interpolation-aggregation procedure has been developed to derive a 100 × 100 m resolution digital soil map of clay content of the entire study region from a hyperspectral image where pixels with vegetation cover were masked. In this situation, the source of data is the set of scattered bare soil fields with hyperspectral soil

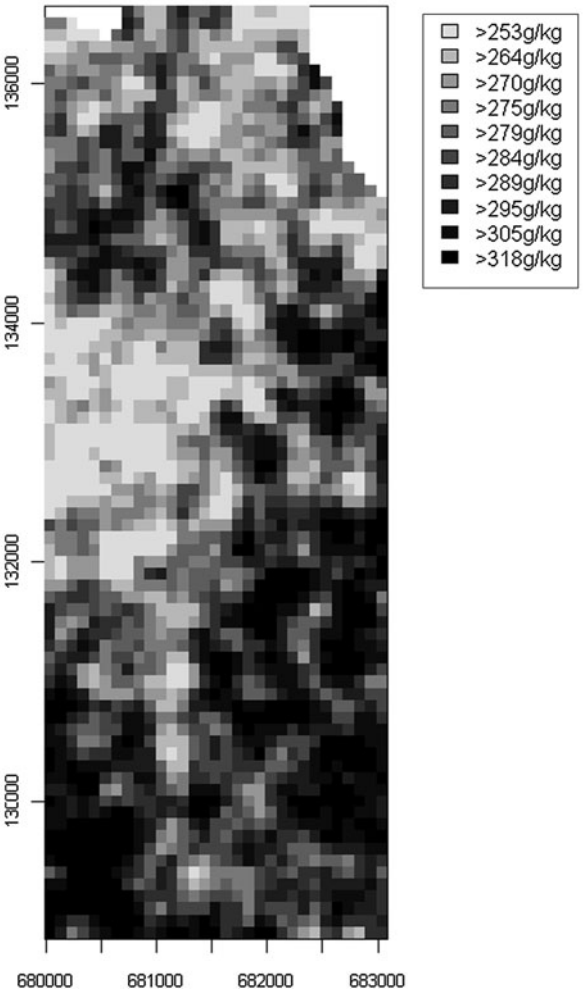


Fig. 8.4 Map of the local means of clay content at the soil surface

property estimations. The procedure includes two steps: (i) interpolation of the soil properties predicted from HYMAP spectra in the bare soil fields, and (ii) pooling of the interpolation outputs to obtain the soil pattern parameters (mean, variances and semi-variances at lags 0–50 m and 50–100 m) of the pixels of the targeted digital soil map. A conditional simulation approach was selected as the interpolation function since it can predict the local values of the property of interest while giving a realistic representation of its spatial structure.

The procedure was applied using the absorption band depth values at 2,206 nm ($CR_{2,206}$) that satisfactorily mapped clay content of the soil surface in bare soil fields as input data (Table 8.1). A double-log transformation of $CR_{2,206}$ was applied to obtain normal distributions. The conditional simulation was performed from the

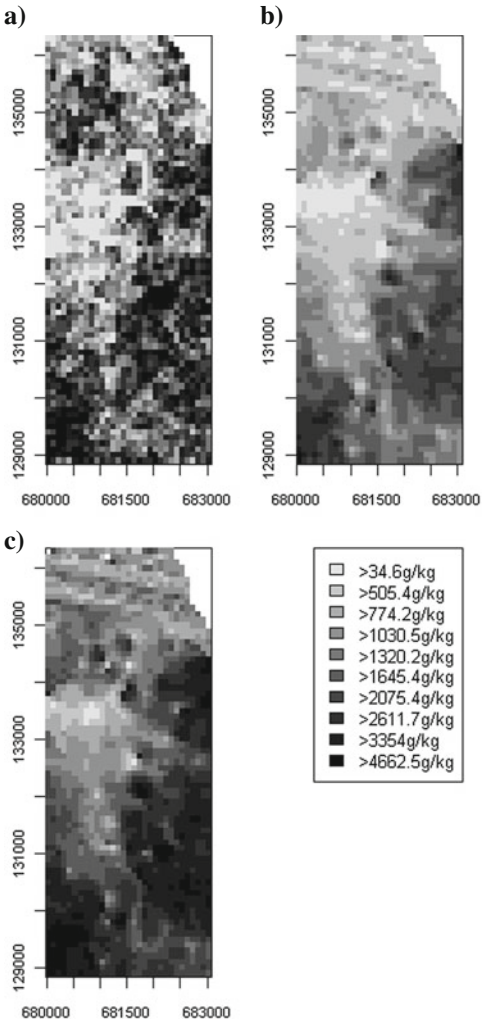


Fig. 8.5 Maps of the (**a**) local variances, (**b**) semi-variances at lag 0–50 m and (**c**) semi-variances at lag 50–100 m

variogram of the double-log transformed $CR_{2,206}$ computed from the set of bare soil pixels with $CR_{2,206}$ values. It was conditioned by bare soil pixels randomly sampled from this set. The Circulant Embedding (CE) simulation algorithm was selected for its ability of handling large datasets. Double log transformed outputs were finally back transformed into $CR_{2,206}$ estimates and the linear formula calibrated by Lagacherie (2008) was applied to obtain clay content values (clay% = 3,790* $(1 - CR_{2,206})$).

Figure 8.4 shows the map of local means of clay content at 100×100 m resolution variances that were obtained from the interpolation-aggregation procedure, and Fig. 8.5 shows the corresponding maps of local variances and semi-variances. The spatial distribution of these two soil pattern parameters seems to be in relation with the geological pattern of the region shown in Fig. 8.2.

As validation was not possible on pixels directly, a cross validation procedure was performed from the set of bare soil parcels. It showed that local means were estimated with an acceptable accuracy ($R^2 = 0.62$). The quality of estimations of the local variance and semi-variances was not so good ($R^2 = 0.30$, $R^2 = 0.29$, $R^2 = 0.22$ for local variance, local semi-variance at lag 0–50 m, local semi-variance at lag 50–100 m, respectively) since the procedure failed to reproduce the small number of erratic large local variances and semi-variances that was observed in the study region. However, the part of the variation of local variance in relation with geological variations seemed to be correctly represented (Fig. 8.5).

8.4 Conclusion

The following conclusions can be drawn from the first results of the Peyne experiment:

- Satisfactory predictions of bare soil surface properties can be obtained from hyperspectral imagery, although a decrease of precision is observed when passing from laboratory to airborne reflectance spectra.
- Calibration uncertainty of the airborne sensor and non-corrected atmospheric effects are by far the main factors that explain this decrease of precision.
- A digital soil map at medium resolution (100×100 m) can be derived from a hyperspectral image with part of the pixels covered by vegetation.

In the near future, a proof-of-concept zone located in the gouvernorat of Nabeul (northern Tunisia) will be implemented. It is expected to demonstrate that a digital soil map of the Mediterranean areas that fit the standards of the GlobalSoilMap.net project (see Chapter 33) can be produced using hyperspectral imagery associated with digital terrain model and legacy soil data. The added value of this digital soil map will be evaluated for the spatial assessment of soil vulnerability to erosion and the mapping wheat yield in water limited situations.

References

Ben-Dor, E., Taylor, R.G., Hill, J., Demattê, J.A.M., Whiting, M.L., Chabrillat, S., and Sommer, S., 2008. Imaging spectrometry for soil applications. Advances in Agronomy 97:321–392.

Clark, R.N., and Roush, T.L., 1984. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. Journal of Geophysical Research 89:6329–6340.

Gomez, C., Lagacherie, P., and Coulouma, G., 2008a. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. Geoderma 148(2):141–148.

Gomez, C., Viscarra Rossel, R.A., and McBratney, A.B., 2008b. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: an Australian case study. Geoderma 146:403–411.

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

Lagacherie, P., Baret, F., Feret, J.B., Madeira Netto, J.S., and Robbez-Masson J.M., 2008. Clay and Calcium carbonate contents estimated from continuum removal indices derived from laboratory, field and airborne hyper-spectral measurements. Remote Sensing of Environment 112:825–835

Madeira Netto, J.S., Robbez-Masson, J.-M., and Martins, E., 2007. Chapter 17. Visible-NIR hyperspectal imagery for discriminating soil types in the la Peyne watershed, France, pp. 219–234. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), Digital Soil Mapping, An Introductory Perspective. Developments in Soil Science, vol. 31. Elsevier, Amsterdam.

Viscarra Rossel, R.A., Walwoort, D.J.J., Mc Bratney, A.B., Janik, L.K., and Skjemstad, J.O., 2006. Visible, near infrared, mid-infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131:59–75.

Chapter 9
Automatic Interpretation of Quickbird Imagery for Digital Soil Mapping, North Caspian Region, Russia

M.V. Konyushkova

Abstract The digital analysis of a Quickbird image has been conducted to develop a procedure for automatic interpretation of soils within the north Caspian Depression. The soil cover in the study area has a spotted pattern and consists of contrasting soils (in terms of the humus content, salinity, pH, etc.): chernozem-like soils, light chestnut soils, and solonetzes (sodic soils). Multispectral data from the Quickbird satellite (September 13, 2006) of 2.4 m spatial resolution were used. Computer-based image analysis was conducted using the ILWIS Open GIS software (ITC, the Netherlands) and STATISTICA 6.0. This work represents the results of image interpretation for rangeland subjected to low grazing pressure. The ground truth data were collected in 2002–2004 and 2007. The original DN values (pixel brightness) of different soil types in the near-infrared (NIR), red, green, and blue bands were analyzed and NDVI values were calculated. Two methods to analyze DN values were used: descriptive statistics and discriminant analysis. The spectra showed that the brightness in the NIR band and the NDVI values are the most informative indices to discriminate soils. The indices were put into the image classification by threshold values obtained from descriptive statistics and by classification functions obtained from discriminant analysis. Both methods of image classification gave similar self-test and cross validation results, with an accuracy of classification of about 80%. Chernozem-like soils were best discriminated. Light chestnut soils and solonetzes were not delineated as well by automatic methods. The approach developed in this study can be used to map regions with contrasting soils changing within short distances, provided that the average soil area is 2–3 times more than a pixel area on the image (e.g., for solonetzic complexes of semiarid regions or for some cryogenic complexes in the tundra zone).

Keywords Solonetzic complex · Sodic soils · ILWIS software · Nonparametric Analysis

M.V. Konyushkova (✉)
V.V. Dokuchaev Soil Science Institute, Pyzhevsky per. 7, Moscow 119017, Russia
e-mail: mkon@inbox.ru

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_9,

9.1 Introduction

The soil cover in the northern Caspian Depression is mainly represented by soil combinations (solonetzic complexes, according to Russian terminology). The complexes consist of solonetzes, light chestnut soils, and chernozem-like soils alternating with one another at short distances of about 10–30 m. Landsat imagery, which is widely used for digital soil mapping (see Chapters 2, 27 and 32), is not of use in the regions with small soil polygons, as in our case where one pixel of the Landsat image often covers several contrasting soil types of the solonetzic complex.

Before the advent of the fine spatial resolution satellite imagery (Quickbird, SPOT, Ikonos, OrbView, etc.), several attempts were performed to map the soils of solonetzic complexes using aerial imagery (Simakova, 1959; Kornblum, 1985). Those approaches were based on the visual interpretation of aerial imagery. The drawbacks of the visual processing of aerial images for large areas are described in Section 31.1. Due to these drawbacks, no large-scale soil maps of the Caspian Depression have been developed.

The first attempt to develop a procedure for automated interpretation of aerial imagery for large-scale mapping of solonetzic complexes was made by Kozlovskii and Korolyuk (1980) and Kozlovskii et al. (1975). The authors demonstrated a theoretical possibility for automated image analysis. However, the absence of adequate technical facilities at that time made it impossible to put their approach into practice. Beyond the area of the Caspian Depression, automated interpretation of solonetzic spots was realized at the Hortobagy region (Hungary) with the use of field reflectometry and discriminant analysis (Toth and Pasztor, 1996).

The goal of this study was to develop a procedure for an automated interpretation of the high-resolution spaceborne (Quickbird) imagery for large-scale soil mapping in the northern part of the Caspian Depression. The methods of digital soil mapping made it possible to simplify the mapping of solonetzic complexes in comparison with conventional procedures.

9.2 Materials and Methods

9.2.1 Study Area

The study was conducted in the area of the Dzhanybek research station of the Russian Academy of Sciences. The station is located in the east of Volgograd oblast, close to the border with Kazakhstan, 49.35–49.43°N and 46.75–46.84°E (Fig. 9.1). This area belongs to the northwestern part of the Caspian Depression and to the geobotanical zone of desert grassland (desert steppes). The study area represents an extremely flat plain with absolute heights ranging from 27 m a.s.l. in the northeastern part to 25.5 m a.s.l. in the southwestern part. The soil cover is represented by solonetzic complexes (Fig. 9.2). Chernozem-like soils occupy shallow (10–30 cm deep) micro lows, solonetzes occupy flat-topped micro highs, and light chestnut soils occupy intermediate positions (micro slopes and micro saddles) (Fig. 9.3).

Fig. 9.1 Study area. Legend: ***1*** – Russia – Kazakhstan border; ***2*** – Boundary of Caspian Depression

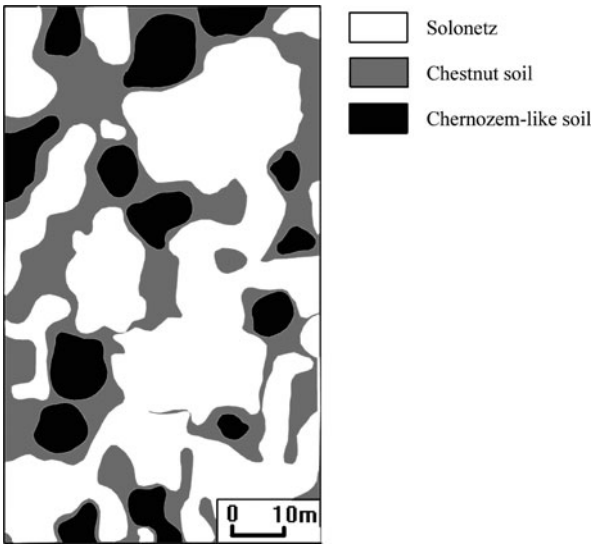


Fig. 9.2 Soil pattern of a solonetzic complex (Rode and Polskii, 1961)

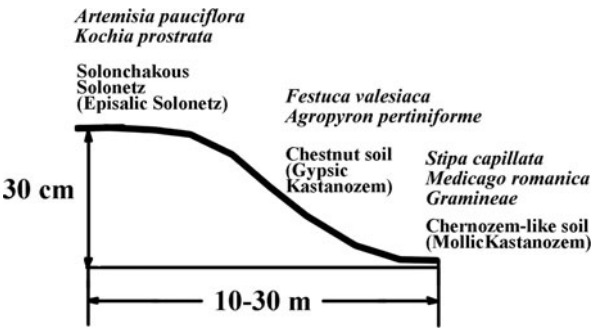


Fig. 9.3 Soil and vegetation change along the micro catena. Soil names according to the WRB (2006) are given in *parentheses*

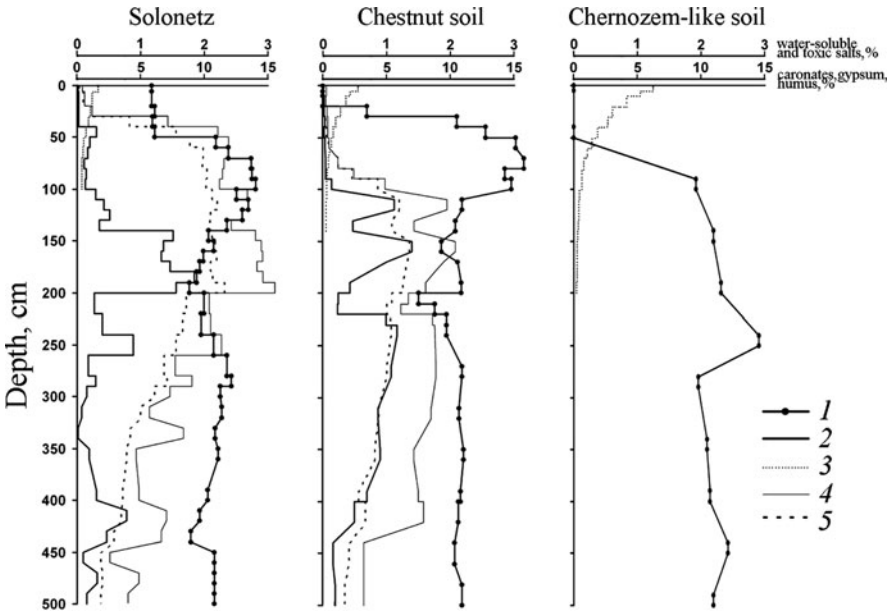


Fig. 9.4 Chemical properties of soils in solonetzic complex. Salts and gypsum are absent in chernozem-like soil. Legend: ***1*** – carbonates; ***2*** – gypsum; ***3*** – humus; ***4*** – water-soluble salts; ***5*** – toxic salts. The graphs are plotted according to the data of Rode and Polskii (1961)

These soils are very contrasting in terms of the humus, carbonates, gypsum, and salt contents (Fig. 9.4), which predetermines the high correlation between soils and vegetation (Fig. 9.3).

9.2.2 Digital and Field Data

High-resolution multispectral data from the Quickbird satellite (USA) acquired on September 13, 2006 were used (Table 9.1). The image had been subjected to standard corrections for radiometric and geometric sensor distortions and orthorectification. Image analysis was performed with the use of the ILWIS Academic 3.4 Open software (International Institute for Geo-Information Science and Earth Observa-

Table 9.1 Quickbird spectral ranges and spatial resolution

Sort of data	Band	Wavelength range, nm	Spatial resolution, m
Multispectral	Blue	450–520	2.44
	Green	520–600	2.44
	Red	630–690	2.44
	Near infrared	760–900	2.44

tion, ITC, the Netherlands). Statistical and discriminant analyses were performed using the STATISTICA 6.0 software.

The total area of investigation was 30 sq. km of rangelands subjected to low grazing pressure. Ground truth data were collected in 2002–2004 and 2007. Soil types and vegetation were determined and topographic leveling was performed along four 100 m-long transects crossing all the components of solonetzic soil complexes. Transects were geo-positioned using a Garmin Emap GPS receiver. Positioning of transects on the Quickbird image was performed in the field. Since the average area of a certain soil was several times more than the area of the Quickbird pixel (Fig. 9.2), it was possible to position ground truth data with sufficient accuracy. Besides, micro depressions with chernozem-like soils were good landmarks as they were clearly seen both on the image and in the field. The sample set included pixels both along transects and in their neighborhoods. The DN values of pixels associated with different soil types in the near-infrared (NIR), red, green, and blue bands were obtained by the "mapvalue" command in the ILWIS, and then NDVI values were calculated. The total amount of pixels was 200 per soil type (600 pixels for the entire soil combination). Out of that sample set, 60 pixels per soil type (180 total) were randomly selected for the control set. Thus, data for 420 pixels were put into analysis.

9.2.3 Inference Models

Two methods of digital analysis were used: model 1 – descriptive statistics (nonparametric) and model 2 – discriminant analysis (parametric). In the first method, minimum, maximum, lower, and upper deciles of brightness distribution in 4 spectral bands and of the NDVI distribution were calculated for 3 soil classes (solonetz, light chestnut and chernozem-like soils). The thresholds separating different soil types were then chosen. In the second method, classification functions were computed for each soil type.

9.3 Results and Discussion

9.3.1 Results from Models

9.3.1.1 Analysis of NDVI and DN Values Distribution

The ranges of DN values and NDVI of different soils are shown in Fig. 9.5. The values of different soils partially intersect due to the high variability (Fig. 9.5a). If extreme values (lower and upper deciles) of distribution are discarded, solonetz soils can be separated from the other two soils by the low DN values in the NIR band (less than 370 for solonetzes), and chernozem-like soils can be separated by the high NDVI values (more than 0.13) (Fig. 9.5b). Those thresholds have been put into the algorithm for automatic image classification. If NIR $>$ 370 and NDVI $>$

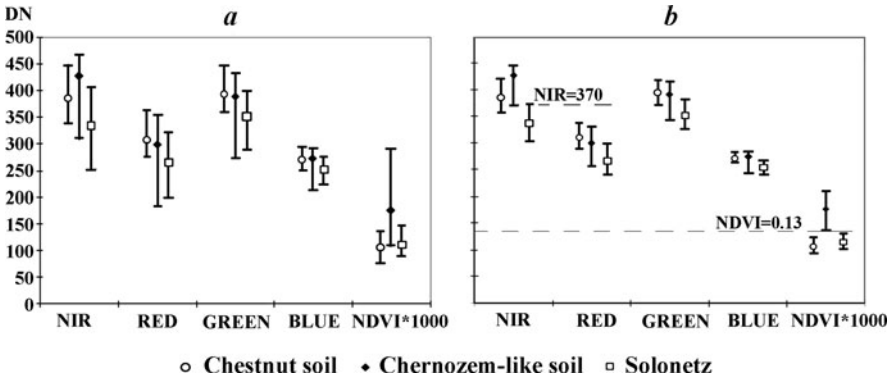


Fig. 9.5 Range plots of DN values in different spectral bands (***a*** – from minimum to maximum; ***b*** – excluding the lower and upper deciles of distribution; medians are depicted by *marks*). $N = 140$ for each soil class

0.13, then a pixel is classified as chernozem-like soil; if NIR <= 370 and NDVI <= 0.13, then a pixel is classified as solonetz; all remaining pixels are classified as light chestnut soils.

9.3.1.2 Discriminant Analysis

Two parameters have been put into the discriminant analysis: the brightness in the NIR band and the NDVI. In the STATISTICA program, the coefficients of classification functions have been calculated (Table 9.2).

The classification functions are calculated in the following way:

$$A = 0.548^{*}\text{NIR} + 439.203^{*}\text{NDVI} - 153.97;$$
$$B = 0.507^{*}\text{NIR} + 272.55^{*}\text{NDVI} - 114.186;$$
$$C = 0.4445^{*}\text{NIR} + 288.3561^{*}\text{NDVI} - 93.0152,$$

where NIR and NDVI are the brightness in the NIR band and the NDVI of the classified pixel, A is the classification score for chernozem, B is the classification score for light-chestnut soil, and C is the classification score for solonetz.

These functions allow us to compute classification scores for each pixel. Thus, each pixel has three classification scores (as many as there are soil types). A pixel belongs to the class, for which it has the highest classification score. Hence, if the

Table 9.2 Coefficients of classification functions. $N = 140$ for each soil class

Variable	CHERN	CHEST	SOLON
NIR	0.548	0.507	0.4445
NDVI	439.203	272.55	288.3561
Constant	−153.97	−114.168	−93.0152

Abbreviations: CHERN – chernozem-like soil; CHEST – light chestnut soil; SOLON – solonetz

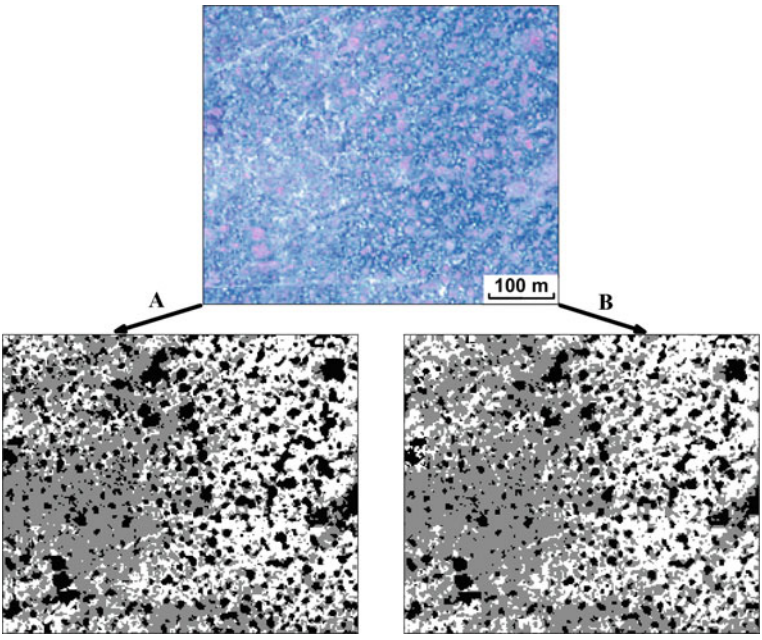


Fig. 9.6 The results of automatic interpretation of Quickbird imagery (at the *top*): ***A*** – classification by threshold values, ***B*** – discriminant classification. Soil designations are the same as in Fig. 9.2

maximum of *A*, *B*, and *C* is *A*, then a pixel is classified as chernozem-like soil; else, if the maximum of *B* and *C* is *B*, then a pixel is classified as light-chestnut soil; all remaining pixels are classified as solonetzes.

The results of the automatically interpreted Quickbird image are shown in Fig. 9.6. As seen from this figure, both classification methods give similar pictures.

9.3.2 Accuracy Assessment of the Models

The visual comparison of the Quickbird image and the digital maps illustrates that both methods of the classification provide quite satisfactory results.

Both models give similar self-test and cross validation results: the accuracy of classification is about 80% (Table 9.3). The accuracy of the discriminant classification is slightly higher (by 6%) than that of the threshold classification. Chernozem-like soils are best classified by both models (with an accuracy of about 90%). The final maps compiled by the two methods show unambiguously (with 100% accuracy) 87–93% of the actual area of chernozem-like soils. Despite the high accuracy of the automated classification of solonetzes (about 90%), they are not discriminated as well because they intermix with chestnut soils, and 30-40% of chestnut soils are misclassified as solonetzes.

Table 9.3 Self-test and cross validation results (model 1 – above the line; model 2 – under the line). $N = 60$ for each soil class

	Self-test					Cross validation			
	Soil class	CHERN	CHEST	SOLON	%	CHERN	CHEST	SOLON	%
Observed class	CHERN	116 117	24 19	0 4	*83* *84*	52 56	8 3	0 1	*87* *93*
	CHEST	4 0	97 115	39 25	*69* *82*	0 0	36 40	24 20	*60* *67*
	SOLON	2 0	29 29	109 111	*78* *79*	0 0	8 5	52 55	*87* *92*
	%	*95* *100*	*65* *70*	*74* *79*	***77*** ***82***	*100* *100*	*69* 83	*68* *72*	***78*** ***84***

Abbreviations are the same as in Table 9.2.

9.4 Conclusions

The use of high-resolution (2.4 m) multispectral Quickbird images has allowed us to compile a large-scale soil map of solonetzic complexes characterized by small areas of soils composing a complex. The maps developed with the use of automatic interpretation of Quickbird imagery display a detailed and accurate soil pattern of the study area.

The automatic interpretation of soils (solonetzes, light chestnut, and chernozem-like soils) has been performed using two parameters of Quickbird images, NIR band brightness and NDVI. Two methods used to classify soils, classification by threshold values and by discriminant functions, have given similar results. The accuracy of classification as determined by cross validation is 78% (threshold values) and 84% (discriminant functions).

It can be supposed that this approach can be used to create digital large-scale soil maps of regions with contrasting soils changing within short distances (e.g., solonetzic complexes of semiarid regions or some cryogenic complexes of tundra regions).

Acknowledgments This study was supported by the Russian Foundation for Basic Research (grant no. 07-04-00136).

References

Kornblum, E.A. (ed.), 1985. Procedure of the Large-Scale Soil Mapping of Solonetzic Complexes. Moscow. 95 p. [in Russian]

Kozlovskii, F.I., and Korolyuk, T.V., 1980. The use of opticostructural machine analysis for detailed soil-ameliorative mapping. Pochvovedenie 9:145–159. [in Russian]

Kozlovskii, F.I., Korolyuk, T.V., Panteleev, B.P., and Yanovskii, K.A., 1975. The method of machine analysis of aerial imagery for soil-ameliorative mapping, pp. 86–96. In: Egorov, V.V. (ed.) Soil-Amelioration Processes in the Regions of New Irrigation. Moscow. [in Russian]

Rode, A.A., and Polskii, M.N., 1961. Soils of the Dzhanybek research station, their morphological features, particle-size distribution, chemical composition, and physical properties, pp. 3–214. In: Rode, A.A. (ed.) Semidesert Soils of the Northwestern Caspian Region and Their Development. Akademia Nauk SSSR. Moscow. [in Russian]

Simakova, M.S., 1959. Procedure of soil mapping in the Caspian Depression with the use of aerial imagery, pp. 283–357. In: Tyurin, I.V., Liverovskii, Yu. A. (eds.) Soil Geographic Research and the Use of Aerial Imagery for Soil Mapping. RAS USSR, Moscow. [in Russian]

Toth, T., and Pasztor, L., 1996. Field reflectance measurements as means of distinguishing vegetation and different grades of salt concentration in the Hortobagy alkali grassland, pp. 23–36. In: Misopolinos, N., and Szabolcs, I. (eds.), Soil Salinization and Alkalization in Europe. ESSC Editions, Thessaloniki.

Chapter 10
ASTER-Based Vegetation Map to Improve Soil Modeling in Remote Areas

E. Meirik, B. Frazier, D. Brown, P. Roberts, and R. Rupp

Abstract Soil scientists are using spectral data from vegetation to help predict soils for the North Cascades National Park (NCNP) Complex in Washington State, USA. Vegetation is a proxy indicator for the soil forming factors organisms and climate. Two objectives of this research are: (1) can spectral remote sensing data be used to accurately map vegetation in the NCNP; and (2) are soil characteristics and properties significantly correlated with vegetation cover? We characterized soil profiles and vegetation cover at 70 sites within the 30,000 ha Thunder Creek Watershed (TCW) in the summer of 2007. These ground truth locations were used to manually interpret 1 m resolution National Agricultural Imagery Program (NAIP) color photography and classify 700 randomly selected validation locations as coniferous forest, shrub, meadow, heather, rock and snow. We applied Spectral Angle Mapper (ENVI version 4.3, ITT Visual Information Solutions; Boulder, CO) to July, 2007 ASTER data to classify vegetation cover for the TCW, with 3600 ground truth pixels and the 700 NAIP photograph pixels employed for calibration and validation respectively. This classification yielded an overall accuracy of 67% with coniferous forest having a producer's and user's accuracies of 84 and 91% respectively.

Standard Analysis of Variance (ANOVA) models were used to examine the correlation of genetic horizon thicknesses with vegetation cover. A-, O-, E- and Bs-horizon thickness were all highly significantly different ($N = 70$, $p < 0.01$) for these two classes. Moreover, the results confirmed conventional soil genesis theory with coniferous forest having thicker E, Bs and O-horizons but thinner A-horizons.

Keywords ASTER · Vegetation classification · Vegetation-soil relationships · Podzolization · Melanization

E. Meirik (✉)
Crop and Soil Sciences Department, Washington State University, Pullman, WA, USA
e-mail: ee.meirik@gmail.com

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_10,

10.1 Introduction

The Remote Area Soil Proxy (RASP) model was developed as a method of predictive soil mapping with a geographic information system (GIS) (Briggs, 2004; Rodgers, 2000; Ufnar, 2004). RASP is a heuristic method reliant on tacit knowledge gained through minimal strategically located fieldwork. The RASP model combines digital landform and land cover maps with attributes derived from a digital elevation model (DEM) to predict soil distribution in remote areas. Detail and accuracy of the final soil map are dependent on the digital inputs used to proxy the soil-formative factors. All past iterations of the RASP model have incorporated previously existing vegetation maps or unsupervised classifications of 30m resolution LANDSAT data. Methods such as illustrated in Chapter 31, Fig. 31.3 have been used.

This project focuses on quantifying soil-vegetation interactions thought to be critical to soil formation and creating a more useful and accurate digital vegetation map to proxy the soil properties. Improved resolution is sought from an orthorectified ASTER (Advanced Spaceborne Thermal Emission and Reflectance Radiometer) scene from 4 July 2007. The objectives of this paper are to analyze soil-vegetation interactions and answer two questions: (1) can spectral remote sensing data be used to accurately map vegetation in the North Cascades National Park (NCNP); and (2) are soil characteristics and properties significantly correlated with vegetation cover?

10.2 Materials and Methods

10.2.1 Study Area

The study area is Thunder Creek watershed (TCW) located in the North Cascades (NOCA) National Park complex, Washington (Fig. 10.1). The map is to serve as a proxy for soil properties resulting from vegetation.

Site stability and vegetation are the most significant predictors of soil development in TCW (Briggs et al., 2006). Various combinations of parent material and vegetation support Andisols, Spodosols, Inceptisols and Entisols of U.S. Soil Taxonomy (Soil Survey Staff, 2006). Stable landforms support coniferous forests, creating a cool and moist soil climate that facilitates podzolization and Spodosol formation. Andisols, Inceptisols and Entisols form on active landscapes under deciduous and herbaceous vegetation. Soil development corresponds to unique interactions of the five soil-forming factors (Jenny, 1941), which are observed in the field, analyzed in the lab and modeled in RASP.

10.2.2 Field Data

Field data were used to establish soil-vegetation relationships and guide vegetation mapping. Sampling procedures were designed to maximize time at each location,

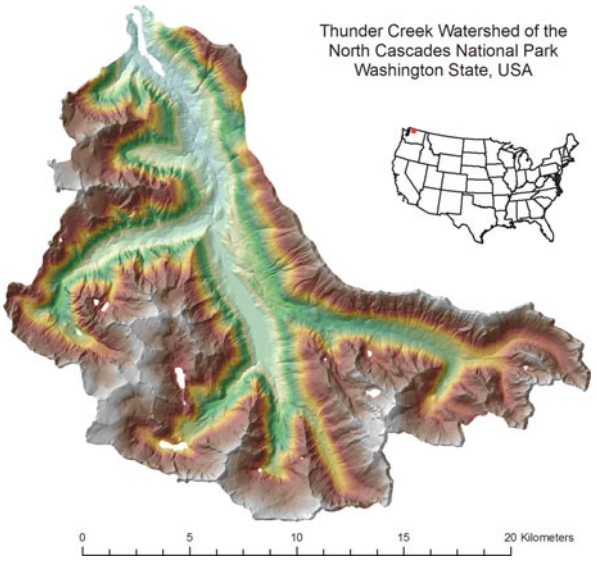


Fig. 10.1 Location of the Thunder Creek watershed in northern Washington, USA

having 10 sample sites in a cluster at each location, and having clusters designed to cross landform and cover variation at each location. Potential sample locations (clusters of sample sites) were pre-selected in areas of maximum variation based on prior mapping of landform and a preliminary classification of land cover. This procedure is efficient for inaccessible areas, but is somewhat different from other DSM projects as reviewed in Chapter 6.

The clusters of 10 sample sites were arranged in a star shape (Fig. 10.2). The central point was located at the theoretical center of an ASTER scene pixel. It is not possible to know the true location of a pixel center since there is assumed error in the orthorectified image (though not visible by alignment with the DEM) and in the GPS location ($\pm$ one half pixel). Cluster center points were located by GPS during daily time periods when satellites were most available. Each cluster consisted of three transects radiating from the central point, and each transect included three sample sites. The sample points were spaced one, three, and 10 pixels from the center point. One of the cardinal directions was assigned to each cluster prior to fieldwork. The first transect pointed in this cardinal direction and the other two were calculated using simple geometry and point projection.

Vegetation and land cover were described for a circular plot of 15 m diameter surrounding each soil pedon at each sample site. Plot size was chosen to represent one pixel of ASTER data. Pedon descriptions were written and soil samples collected and geomorphology and landform data were recorded for each plot as well. Elevation was determined from a topographic map and recorded by the Garmin GPS unit. Aspect was measured with a compass and slope with a clinometer.

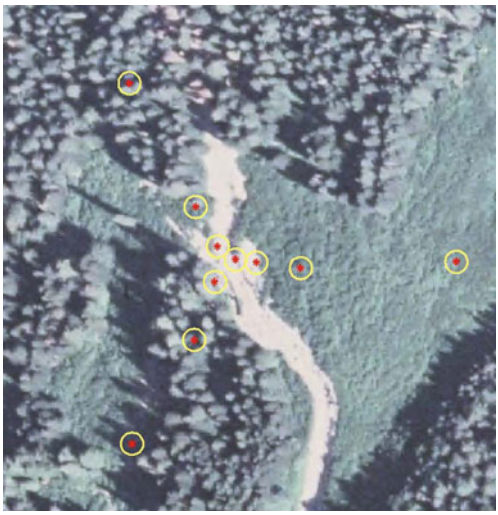


Fig. 10.2 GPS locations for cluster 26 with vegetation plot and soil pedon locations

10.2.3 Soil-Vegetation Interactions

Field data were used to compare genetic soil horizon thicknesses by vegetative cover. Interquartile range (IQR) box plots were constructed to compare horizon depth by vegetation class. Box plots were constructed for A-, O-, E- and Bs-horizons. The vegetation types mapped with ASTER data are compared: coniferous forest, heather, meadow and deciduous shrub. Vegetation classes were condensed for statistical analysis because of the low sample size for heather and meadow. Soil horizon depth was compared for deciduous and woody evergreen vegetation communities. Analysis of variance (ANOVA) models were applied to the log transformed data to quantify soil-vegetation relationships for the broad classes of deciduous and woody evergreens.

10.2.4 Image Classification Procedures

This project analyzes ASTER (Advanced Spaceborne Thermal Emission and Reflectance Radiometer) data to create land cover maps. ASTER data is provided at 15 m spatial resolution for visible and near infrared bands. Previous development of the RASP model used Landsat Thematic Mapper (TM) data with 30 m spatial resolution. Field investigators of NOCA have questioned the integrity of the Landsat TM-derived classification of vegetation. Visually the TM-based map appears pixilated with pure stands of pacific silver fir (*Abies amabilis*) adjacent to stands of western hemlock (*Tsuga heterophylla*) and Douglas fir (*Pseudotsuga menziesii*). While the legend lists species, in the natural setting it is more common to find mixed stands. The current research aims to independently assess that

accuracy of TM-based map and to present an alternative vegetation map using ASTER data.

Thematic maps were produced with image processing software by supervised classification of ASTER data. Field observations were used to drive the process. Regions of interest (ROIs) were digitized on the ASTER scene using georeferenced field observations. Each ROI is a combination of points and polygons for a known land cover type. Six land cover and vegetation types were distinguished and classified: coniferous forest, deciduous shrub, herbaceous meadow, heather, rock and talus, and snow and ice. ROIs were located more than three pixel widths into a cover type to avoid any influence from border pixels.

These land cover and vegetation classes were chosen because they communicate information about the soil and they are distinguishable on a 1 m resolution NAIP image. The georeferenced NAIP image and field observations were used to determine cover types at validation sample points. The NAIP image was not classified, but used to observe cover conditions at validation points. This supported our aim to create a reliable vegetation map with class distinctions for vegetation types that support different soil types.

The mean spectral signatures for each ROI are displayed in Fig. 10.3. The greatest separability is observed in the VNIR bands (1–3) and two SWIR bands (4 and 5), while spectra for the longer SWIR bands (6–9) are similar for these land cover classes. Consequently, bands 1–5 (VNIR and 2 SWIR) were used to classify the ASTER scene. The classifier that makes greatest use of spectral curve data is the Spectral Angle Mapper (SAM) (ENVI 4.3). We used the algorithm with a single threshold of 0.3 radians. Lakes were masked in the classification process with a polygon layer that was digitized from the ASTER scene.

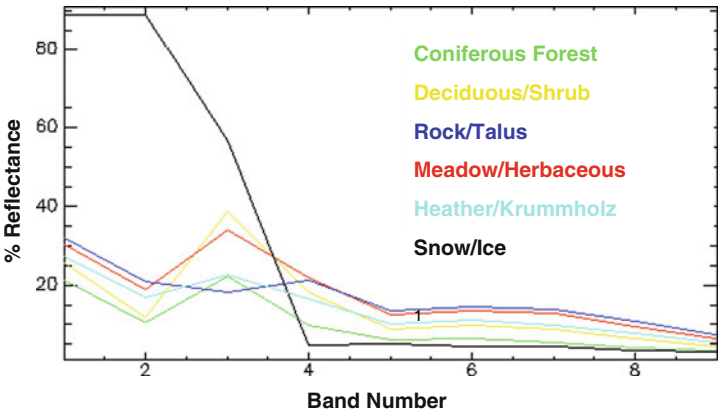


Fig. 10.3 Mean spectra for ROIs

10.3 Results and Discussion

10.3.1 ASTER Image Classification

Spectral Angle Mapper produced a thematic map of TCW with seven classes (Fig. 10.4). The six specified land cover types as well as an "unclassified" type are included. According to this classification coniferous forests cover the majority of TCW (34%), followed by snow/ice (28%) and heather (17%). Rock/talus covers 13% of TCW, followed by meadow (4%), shrub (3%), unclassified pixels (0.5%), and the lake mask (0.5%).

Classification accuracy was assessed by comparison to the NAIP validation dataset in the form of an error matrix (Table 10.1). Overall accuracy of this classification is 67%. Snow has the highest producer's accuracy (100%) followed by forest (84%), shrub (67%) and rock (67%). These classes have low errors of omission; they are being mapped with high accuracy. The heather class has a lower producer's accuracy (44%) and meadow is much lower (9%). These vegetation communities are being mapped as other classes. By looking at the error matrix, we see that meadows are commonly included in the heather class (37% of the time). Meadows are difficult to map using a single ASTER image because of seasonal variation and a short growing season. Meadow growth stage varies by elevation and aspect. A meadow on a north-facing slope may be covered in snow late into the summer, while one on a south-facing slope is green and flowering. This causes great variation in spectral signature for the meadow class, from dead and dry grasses to moist and

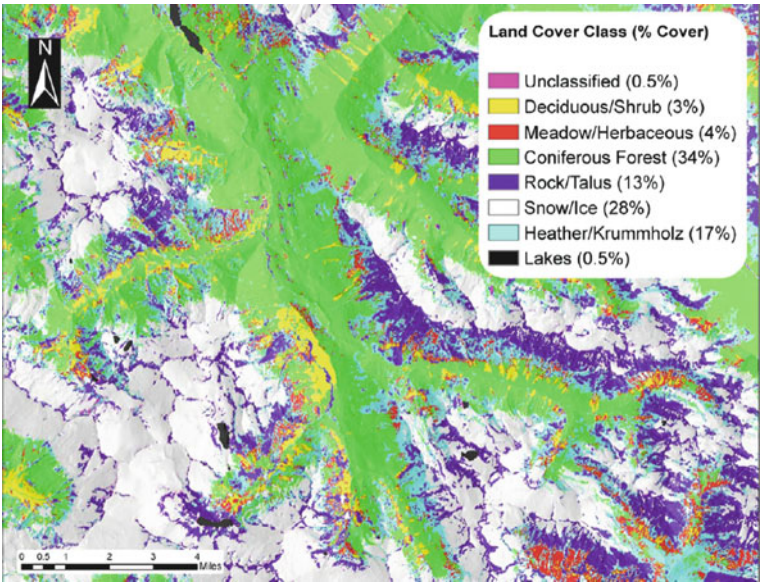


Fig. 10.4 Cover map of Thunder Creek watershed from ASTER data. Cover amounts listed as a percentage of the watershed

Table 10.1 Accuracy assessment for the SAM classification from ASTER data

		Actual NAIP								
		Snow	Forest	Shrub	Heather	Rock	Meadow	Total	UA	EC
Predicted-ASTER Classes	Snow	72	4	2	25	14	18	135	53	47
	Forest	0	242	10	5	1	9	267	91	9
	Shrub	0	1	59	0	0	4	64	92	8
	Heather	0	32	5	39	8	32	116	34	66
	Rock	0	1	0	17	50	16	84	60	40
	Meadow	0	9	11	3	0	8	31	26	74
	Unclass	0	0	1	0	2	0	3		
	Total	72	289	88	89	75	87	700		
	PA	100	84	67	44	67	9		67	
	EO	0	16	33	56	33	91			

green. Mapping accuracy of the meadow class could be increased by having more meadow classes, one to represent each of the phenological stages that meadows are found in on the image date. The disadvantage of having more than one meadow class is that it will increase the complexity of the over-all classification and require more validation points for a relatively minor landscape class.

User's accuracy and omission errors provide additional information for the classification. The snow class has a low user's accuracy (53%) and high commission errors (47%). The snow class encompasses 25 heather validation points. While all snow points were included in the snow category (PA 100%) we see that the snow class includes other vegetation types (UA 54%). This may be an artifact of varying snow levels between the ASTER and NAIP images. The shrub class had the highest UA (92%), followed closely by forest (91%) and meadow (60%). The rock class included many heather (17) and meadow points (16) and had relatively high error of commission (66%). The meadow class has a higher user's accuracy (34%) than producer's accuracy (9%). This means that the meadow class is adequate for mapping meadow; however the meadow areas are also being detected by other classes.

10.3.2 A-Horizon Thickness

A-horizon thickness was compared for vegetation classes (Fig. 10.5a). The median value is denoted by the thick line, while shaded boxes signify the IQR. The IQR is the middle 50% of the data, with the upper limit being half-way between the median and the largest value (0.75 quartile), and the lower limit being the midpoint of median and the lowest value (0.25 quartile). Box width is proportional to sample size (n), with the largest n being the widest. Hollow circles on the graph denote outliers in the dataset.

Forest and shrub categories had the greatest n, followed by rock, while meadow and heather have much smaller n; these differences are expressed by the width of

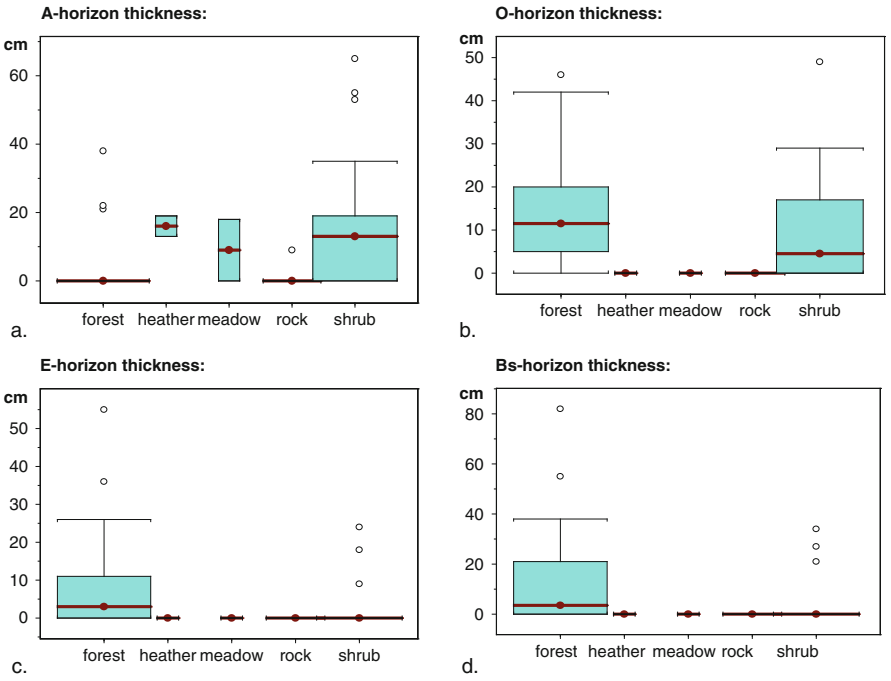


Fig. 10.5 Interquartile range box plots showing soil-vegetation interactions

each box. Shrub, meadow and heather classes have thick A-horizons, while forests generally had no A-horizon. Organic matter accumulation, or melanization, is occurring in shrub, meadow and heather but not in forest or rock. There are a few outliers in both the forest and shrub classes.

The ANOVA model indicated that A-horizon thickness is significantly different for deciduous and woody evergreen vegetation (p-value < 0.001). The mean difference for deciduous and woody evergreen A-thickness is 9.88 cm; on average deciduous vegetation had 9.88 cm thicker A-horizon than woody evergreen cover. Deciduous leaf litter leads to the accumulation of deep, dark, organic-rich A-horizons. Melanization is occurring beneath deciduous vegetation communities. Woody evergreen litter, on the other hand does not lead to A-horizon formation.

10.3.3 O-Horizon Thickness

Forests support the thickest O-horizons (0–47 cm) in TCW. Some shrub zones also have O-horizons (median 4 cm), while the heather, and meadow classes lack O-horizons (Fig. 10.5b). Organic matter is accumulating beneath coniferous forests and shrub, but not beneath meadow, and heather. This results from varying residence times and deposition rates for different litter types. Thick organic layers accumu-

late beneath forests because the trees are continually dropping needles and beneath shrub zones because of the annual fall leaf deposition. Meadows do not accumulate organic horizons on the soil surface because grasses break down more quickly than leaves and needles. Organic litter type varies for forest (coniferous needles) and shrub (deciduous leaves), causing varying sub-surface chemical reactions expressed in horizon sequence. Deciduous leaf litter is alkaline, promoting organic matter accumulation; while coniferous litter is acidic causing organic matter to complex with iron oxides and leach through the system. This process is known as podzolization and results in the eluvial (E)-illuvial (Bs) horizon sequence. The ANOVA model for total O-horizon thickness reveals that deciduous and woody evergreen vegetation are significantly different (p-value 0.007). On average, coniferous vegetation supports a 4.6 cm thicker O-horizon than woody evergreen cover.

10.3.4 E-and Bs-Horizon Thickness

The process of eluviation or leaching causes E-horizon formation. E-horizons are light colored because of an absence of mineral and organic coatings on grains and peds. Thick E-horizons develop beneath forested regions, but not under meadow or heather (Fig. 10.5c).

A few of the shrub plots had E-horizons, however these are outliers and not contained within the IQR. Field notes reveal that these pedons contained charcoal, and it is likely that the E-Bs horizon sequence is remnant of a forest which has burned. These outliers from the shrub class for E-horizon are also noted in the Bs-horizon plot (Fig. 10.5d). Eluviation is occurring beneath forests in TCW because of the acidic leaf litter. Organic acids readily complex with iron and aluminum oxides, and the complexes leach through the solum causing E-horizon formation. The complexes are deposited in the sub-soil in an area of illuviaiton promoting Bs-horizon formation. This eluvial-illuvial process is called podzolization and results in the E-Bs horizon sequence. Deciduous leaf litter is not associated with podzolization. Average E-horizon thickness is 5.05 cm thicker for woody evergreen communities than deciduous vegetation and the difference is significant (p-value < 0.001).

There are a few outliers in the shrub category with Bs horizons. Bs horizons are not forming beneath heather or meadow. Illuviation of Fe and Al-oxides is occurring in forest soils but not beneath shrub zones. ANOVA results show that woody evergreen vegetation has an average of 9.79 cm thicker Bs-horizon than deciduous vegetation (p-value < 0.001). Woody evergreen vegetation is associated with the O-E-Bs horizon sequence, while deciduous vegetation supports O and A horizon development.

10.4 Conclusions

Supervised classification of ASTER data with SAM procedures is a good method of producing thematic maps in remote areas. We produced a classification with 67% overall accuracy and provided a reliable method of image pre-processing, training

and classification. Genetic horizon thickness was correlated with vegetation cover. A-, O-, E- and Bs-horizon thickness were all highly significantly different ($N = 70$, $p < 0.01$) for these two classes. The vegetation layer will be useful to predict soil properties in the next step of this work when these data are combined with landform and DEM derived covariates as discussed in Chapter 15.

References

Briggs, C.A.D., 2004. GIS-based mapping of soil distribution in Thunder Creek watershed, North Cascades National Park, WA. M.S. Thesis, Washington State University, Pullman.

Briggs, C.A.D., Busacca, A.J., and McDaniel, P.A., 2006. Pedogenic processes and soil-landscape relationships in North Cascades National Park, WA. Geoderma 137:192–204.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York, NY.

Rodgers, T.M. 2000. Modeling soils of the Sawtooth and Pasayten wilderness areas with a GIS. M.S. Thesis, Washington State University, Pullman.

Soil Survey Division Staff, 2006. Soil survey manual. Soil Conservation Service. U.S. Department of Agriculture, Handbook 18.

Ufnar, D.R., 2004. Mapping land type associations and soil distribution using GIS in the Cascade Range of Washington State. M.S. Thesis. Washington State University, Pullman.

Chapter 11
Digital Soil Boundary Detection Using Quantitative Hydrologic Remote Sensing

E.M. Engle, J.B.J. Harrison, J.M.H. Hendrickx, and B. Borchers

Abstract Creating accurate soil maps at small scales using traditional methods is a time consuming and expensive process. However, remote sensing techniques can provide spatially and spectrally contiguous data in a timely manner. For this study, 20 root zone soil moisture maps derived from Landsat images during the growing season were used for the detection of soil boundaries. A split moving window analysis along two demonstration transects in, respectively, a semi-arid desert and riparian area located in the Middle Rio Grande Valley of New Mexico showed that remotely sensed root zone soil moisture can reveal subsurface trends that can be used to identify soil boundaries which do not have a strong surface expression. Overall, the use of multiple remotely sensed root zone soil moisture images for soil boundary delineation shows great promise of becoming a valuable tool in the field of digital soil mapping.

Keywords Remote sensing · Boundary detection · Surface Energy Balance Algorithm for Land (SEBAL) · Split moving window · Soil mapping

11.1 Introduction

GIS and remote sensing are the basis of digital soil mapping (Lagacherie et al., 2007). For example, the Landsat Multispectral Scanner (MSS) and Thematic Mapper (TM) have been successfully used to map land cover, soils, terrains and man-made features such as dams and urban areas (Baban and Yusof, 2001). The use of the India Remote Sensing satellite Linear Imaging Self-scanning Sensor (IRS-1B LISS-II) can provide details about soil classes that are often not found on existing soil maps produced by more traditional means (Karale et al., 1991). While these are only two examples of the types of surveys and methods using remotely sensed data, they have one facet in common with most other studies: all use digital

E.M. Engle (✉)
Department of Earth and Environmental Science, New Mexico Tech, 801 Leroy Place, Socorro, NM 87801, USA
e-mail: eengle@nmt.edu; emmengle@gmail.com

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_11,

values from a single image that only provide information about the land surface, i.e. reflectances of the visible, near- and mid-infrared bands and long wave emission of the thermal infrared band. For example, the interesting studies in Chapters 8 (one HYMAP image), 9 (one Quickbird image), 10 (one ASTER image), 17 (one set of high resolution pseudo-RBG color composite images), 29 (one Landsat image), and 32 (one ASTER image) are based on one single image reflecting soil surface conditions on one specific day. In general, such data represent the top few centimeters of the soil surface at best or under full vegetative cover represent the characteristics of the vegetation that may or may not be related to soil type. Since in semi-arid New Mexico some soil boundaries have surface expressions while others do not (Gile, 1975a, b), it is expected that the use of digital values only will not be sufficient to detect all soil boundaries in the landscape.

Gile and his co-workers used the traditional approach for soil boundary detection that is based on qualitative evaluation of soil morphological characteristics with emphasis on texture. Because texture strongly effects soil moisture properties (Taylor and Ashcroft, 1972) it was expected that boundaries based on soil moisture conditions would show good agreement with those detected using soil morphological characteristics. Hendrickx and his colleagues analyzed several data sets along transects in southern New Mexico using the moving split window technique (Webster, 1973, 1978) and found good agreement with boundaries located qualitatively based on soil morphological characteristics and those located quantitatively based on soil water content measurements with depth (Hendrickx et al., 1990, 1986; Wierenga et al., 1987). An important observation of these studies was that using multiple days of soil moisture observations over longer periods yields more information that one single data set representative for 1 day only. Therefore, these studies firmly established that series of soil water content measurements with depth provide sufficient information for soil boundary detection in semi-arid New Mexico.

Unfortunately, taking soil water content measurements along transects on the km-scale requires much effort. Even when non-invasive electromagnetic induction is used for soil water content measurements (Kachanoski et al., 1988, 1990, 2002; Sheets and Hendrickx, 1995) the effort is too large to obtain data sets that can cover an entire watershed. Only by using operational remote sensing satellite imagery one can prepare regional root zone soil moisture maps at acceptable cost (Fleming et al., 2005; Scott et al., 2003). The objective of this study was to investigate whether remotely sensed root zone soil moisture maps can be used for soil boundary detection.

11.2 Study Area

Two field areas in central New Mexico, USA, were used in this study: the Sevilleta National Wildlife Refuge (NWR) and the Hilton Ranch (Fig. 11.1). These areas were chosen in part because the soils had been previously mapped by the Natural Resources Conservation Service (NRCS) and the landforms were mapped by New

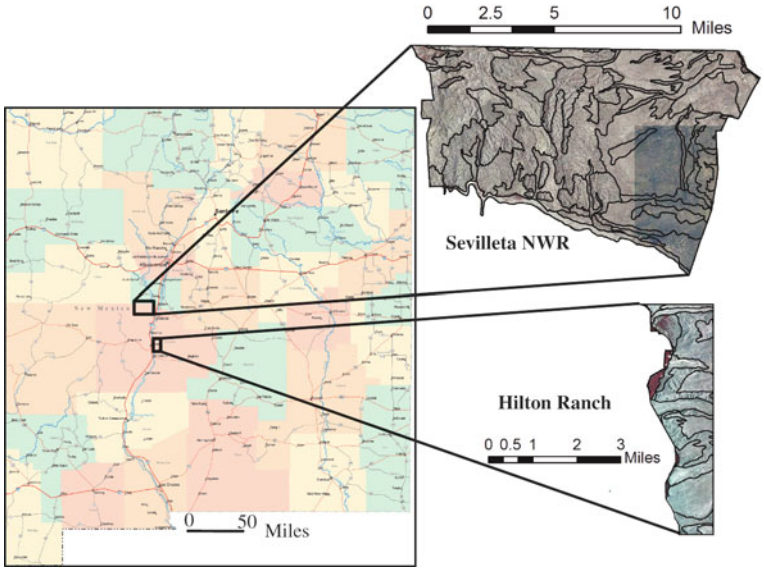


Fig. 11.1 Location of Sevilleta National Wildlife Refuge and Hilton Ranch in central New Mexico, USA

Mexico Tech Ph.D. geology student Alex Rinehart (unpublished data, 2009). These soil and landform maps were used to evaluate our novel method for detection of soil boundaries using remotely sensed satellite imagery.

The Sevilleta National Wildlife Refuge is located in central New Mexico and covers an area of approximately 1,000 km^2. This area contains four major ecosystems: the Chihuahuan desert, Great Plains grasslands, Colorado Plateau shrub-steppe and conifer woodlands. Landforms include alluvial fans, pediments and terraces of various ages and active channels. The NRCS map includes 17 soil associations and complexes (Johnson, 1984).

The Hilton Ranch is located on the east side of the Rio Grande opposite the town of Socorro, NM. The range of landforms is similar to those of the Sevilleta. However, due to its proximity to the Rio Grande more riparian vegetation is present along the floodplains. There are six soil complexes and associations in this area (Johnson, 1984).

11.3 Methods

Just as in our previous studies (Hendrickx et al., 1986; Wierenga et al., 1987) we used the split moving window technique (Webster, 1973, 1978) for soil boundary detection. However, instead of ground-measured soil water contents we employed a novel technique for determination of root zone soil moisture content from Landsat

Table 11.1 Date, path and row numbers of Landsat 5 and Landsat 7 images

Date	Path number	Row number	Satellite	Study area
4/7/2000	33	36	Landsat 7	Sevilleta, Hilton
5/6/2002	34	36	Landsat 7	Sevilleta, Hilton
5/9/2000	33	36	Landsat 7	Sevilleta, Hilton
5/12/2004	33	36	Landsat 5	Hilton
5/22/2005	34	36	Landsat 7	Sevilleta, Hilton
5/28/2004	33	36	Landsat 5	Hilton
5/31/2002	33	37	Landsat 7	Sevilleta, Hilton
6/4/2001	34	36	Landsat 7	Sevilleta, Hilton
6/13/2004	33	36	Landsat 5	Hilton
6/16/2002	33	36	Landsat 7	Sevilleta, Hilton
7/2/2005	33	36	Landsat 5	Hilton
7/6/2004	34	36	Landsat 5	Sevilleta, Hilton
7/28/2000	33	36	Landsat 7	Sevilleta, Hilton
7/31/2004	33	36	Landsat 5	Hilton
8/3/2005	33	36	Landsat 5	Sevilleta, Hilton
8/19/2002	33	36	Landsat 7	Sevilleta, Hilton
9/14/2000	33	36	Landsat 7	Sevilleta, Hilton
9/17/2004	33	36	Landsat 5	Hilton
9/30/2000	33	36	Landsat 7	Sevilleta, Hilton
10/14/1999	33	36	Landsat 7	Sevilleta, Hilton

images (Fleming et al., 2005; Scott et al., 2003). Twenty Landsat 5 TM and Landsat 7 ETM+ images captured during the growing season from April to October (Table 11.1) were used to map root zone soil moisture using the Surface Energy Balance Algorithms for Land (SEBAL). Fourteen of the images were used for the Sevilleta, due to lack of full coverage, and all 20 were used for the Hilton Ranch.

11.3.1 Surface Energy Balance Algorithm for Land (SEBAL)

Each image was processed through SEBAL by Hendrickx and postdoctoral research associate Sung-ho Hong. SEBAL is a remote sensing flux algorithm that solves the surface energy balance on an instantaneous time scale for every pixel of a satellite image (Allen et al., 2007a, b; Bastiaanssen, 2000; Bastiaanssen et al., 1998a, b, 2002). The method computes evapotranspiration and root zone soil moisture. It considers a user-defined wet and dry pixel to assume the sensible heat flux is zero and the latent heat flux is zero, respectively. The radiation balance can then be solved for each pixel in the entire image relative to those two points (Bastiaanssen et al., 1998a; Bastiaanssen, 2000). SEBAL is a physically based analytical method that evaluates the components of the energy balance and determines the ET rate as the residual

$$R_n - G - H = \lambda E \tag{11.1}$$

where R_n is the net incoming radiation flux density (W m^{-2}), G is the ground heat flux density (Wm^{-2}), H is the sensible heat flux density (W m^{-2}), and λE is the latent heat flux density (W m^{-2}), which is converted to the evapotranspiration (ET)

rate. The parameter λ is the latent heat of vaporization of water (J kg^{-1}) and E is the vapor flux density (kg m^{-2} s^{-1}). Evaporation E includes both bare soil evaporation and canopy transpiration. SEBAL uses an internal auto-calibration process that greatly eliminates the need for atmospheric corrections and it does not require actual measurements on the ground. The method computes the surface albedo, surface temperature and vegetation index from multispectral satellite data. The surface albedo is used to calculate net short wave radiation, and surface temperature for the calculation of net long wave radiation, soil heat flux and sensible heat flux. The vegetation index governs the soil heat flux by incorporating light interception by canopies, and is used to express the aerodynamic roughness of the landscape. The latent heat flux is computed as the residue of the surface energy balance. Air humidity measurements are not needed because evaporation is computed from the latent heat flux. SEBAL has been applied for water balance estimations (Pelgrum and Bastiaanssen, 1996), irrigation performance assessment studies (Roerink et al., 1997), and for weather prediction studies (van den Hurk et al., 1997).

Hendrickx and his research group have applied SEBAL in the United States, Panama, Morocco, and West Africa (Compaoré et al., 2008; Hendrickx and Hong, 2005; Hendrickx et al., 2005, 2006; Hong et al., 2009). Soil moisture conditions in the root zone can be determined from the evaporative fraction using the empirical relationship (Ahmad and Bastiaanssen, 2003):

$$S = \frac{\theta}{\theta_{\text{sat}}} = e^{\frac{\Lambda - 1}{0.42}} \tag{11.2}$$

where S is relative degree of saturation, θ is volumetric water content, and θ_s is volumetric water content at saturation. The validity of Eq. (11.2) has been tested in several studies (Ahmad and Bastiaanssen, 2003; Scott et al., 2003) including one in New Mexico (Fleming et al., 2005).

11.3.2 Split Moving Window Analysis

Sixteen transects were randomly selected in our field areas for analysis (Engle, 2009) but for this chapter we will only present data from transects 3 and 10. For the split moving window technique (Webster, 1973, 1978) a window size of 5 pixels was selected because it is sufficiently narrow to capture boundaries that occur over short distances but also adequate to minimize noise. A t-test was used to determine the statistical difference between the windows; boundaries are supposed to coincide with maximum t-values. In order to test the threshold at which a t-value is high enough to be considered a boundary, we used four "critical" t-values, 6, 8, 10, and 12.

In general, sharp boundaries such as landscape boundaries are distinct. However, gradational boundaries are harder to detect because they do not exhibit the sudden change in properties that generates a large t-value in the split moving window analysis. Transitional boundaries (boundaries that shift locations due to

Table 11.2 Classification scheme for detected boundaries

Percentage of days	Boundary strength	Range (m)	Boundary type
0 – 30%	Strong	0 – 100	Stable
30 – 60%	Intermediate	100 – 200	Intermediate/stable
60 – 100%	Weak	200 – 300	Intermediate/transitional
		300 – 400	Transitional

antecedent conditions) are also hard to detect because they may occur in slightly different locations on different days. All boundaries detected were classified based on two properties: the percentage of image days over which each boundary is present (the boundary strength) and the spatial range over which they occur (Table 11.2).

The split moving window technique was applied to four different sets of variables: (i) the first principal component of the digital values of the seven Landsat bands for each day (daily DV PCA); (ii) the first principal component of the digital values of the seven Landsat bands for all days (overall DV PCA); (iii) the root zone soil moisture values of each day (daily RZSM); and (iv) the first principal component of the root zone soil moisture values for all days (overall RZSM). The principal components were calculated using *ERDAS Imagine* and captured about 70% of the variability in the data.

11.4 Results and Discussion

Figures 11.2, 11.3 and 11.4 show the results of the split moving window technique along transects 3 and 10. Transect 3 (Fig. 11.2) crossed a number of landform and soil map boundaries including the ephemeral stream channel of the Rio Salado. In the daily data (Fig. 11.2), the northern boundary of the Rio Salado was clearly seen in all datasets while the southern boundary was not readily apparent. Some of the boundaries corresponded with landform and soil map boundaries. Other boundaries that did not correspond to landform or soil boundaries could have either been false detections or more importantly, boundaries that were not detected previously. The overall PCA data showed similar results (Fig. 11.3) but use of the daily data yielded more boundary information (Fig. 11.2). The northern boundary of the Rio Salado appeared in only one dataset (overall digital value PCA). The third and fourth boundaries seen at 2,280 and 2,500 m in the daily data appeared in the overall PCA data also. There were boundaries, mostly in the northern section of the transect, that did not correspond to preexisting boundaries. However, they were also identified in the daily data, which is further evidence that they are real boundaries that have not previously been mapped.

Transect 10 (Figs. 11.4 and 11.5) is an example of a transect crossing agricultural fields and the Rio Grande floodplain representing the most complex soil landscape in both study areas. As a result, many boundaries were detected in both the root zone soil moisture and the digital value PCA. Because the fields were often irrigated

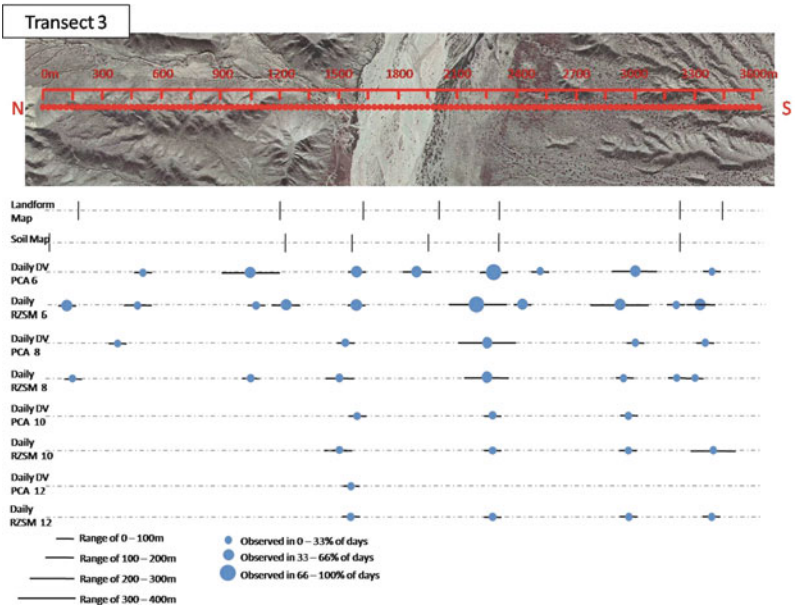


Fig. 11.2 Graphical representation of the boundaries generated using daily digital value (DV) PCA and daily root zone soil moisture (RZSM) at critical *t*-values of 6, 8, 10 and 12, respectively, compared to landform and soil boundaries mapped along transect 3. The dot size represents the percentage of days the boundary occurs and the line length represents the spatial range over which it occurs

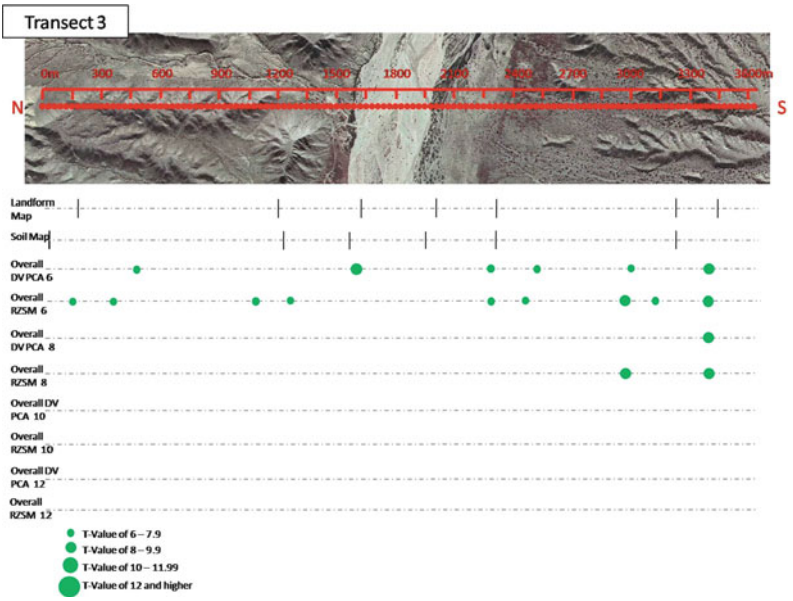


Fig. 11.3 Graphical representation of overall root zone soil moisture (RZSM) and overall digital value (DV) PCA data at critical *t*-values of 6, 8, 10 and 12 along transect 3. The size of the dot represents the *t*-value of each boundary

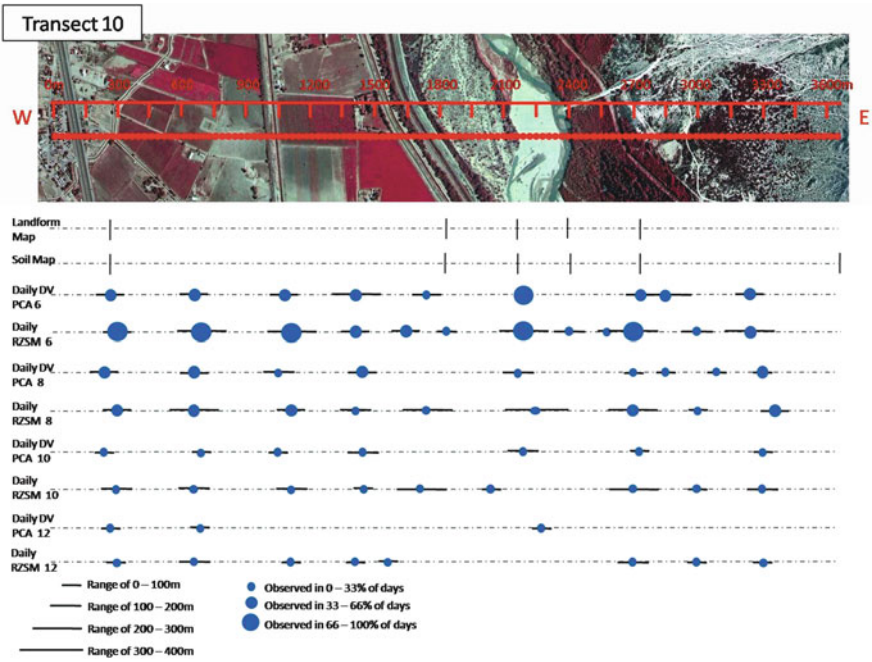


Fig. 11.4 Graphical representation of the boundaries generated using daily digital value (DV) PCA and daily root zone soil moisture (RZSM) at critical *t*-values of 6, 8, 10 and 12, respectively, compared to landform and soil boundaries mapped along transect 10. The *dot* size represents the percentage of days the boundary occurs and the *line* length represents the spatial range over which it occurs

separately, the moisture content in each field was different so that the boundaries detected were the edges of the field. Most of the boundaries detected had high *t*-values which attest to their strength. The start of the fields can be detected easily with this method due to the difference between the fields and the surrounding desert. The overall PCA data showed similar results (Fig. 11.5) but use of the daily data yielded more boundary information (Fig. 11.4).

In these two examples along a one-dimensional transect both the daily root zone soil moisture and daily digital value PCA (Figs. 11.2 and 11.4) were successful at detecting boundaries, while the overall datasets (Figs. 11.3 and 11.5) were not as efficient. There were cases where daily root zone soil moisture detected soil boundaries better than the daily digital value PCA. For example, the daily DV PCA 12 detected only one boundary compared to four boundaries detected by daily RZSM 12 in transect 3 (Fig. 11.2). Similarly, daily DV PCA 12 detected three boundaries compared to eight boundaries detected by daily RZSM 12 in transect 10 (Fig. 11.4). This suggests that the root zone soil moisture detected changes at depth that did not have a surficial expression detectable by Landsat digital values. Because most of these images were taken during the growing season, the root zone moisture conditions varied temporally and spatially across the study areas. By com-

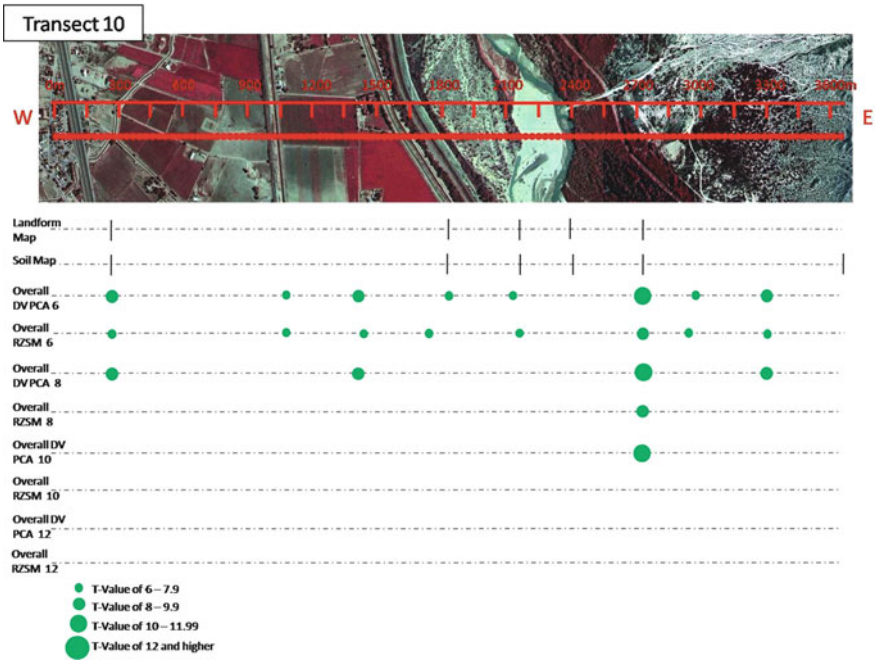


Fig. 11.5 Graphical representation of overall root zone soil moisture and (RZSM) overall digital value (DV) PCA data at critical *t*-values of 6, 8, 10 and 12 along transect 10. The size of the *dot* represents the *t*-value of each boundary

bining multiple images, we reduced this effect while still incorporating a sequence of varying levels of soil moisture. Thus, we are able to enhance the spatial trends while minimizing the temporal effects of localized wetting due to precipitation or irrigation.

Valuable information can be gained from the SEBAL-derived root zone soil moisture; but under certain environmental conditions, valuable information can also be taken from the daily digital value PCA. When all datasets were combined, the efficiency of the methodology at detecting confirmed boundaries decreased as the *t*-value increased (Figs. 11.3 and 11.5). This was expected because as the *t*-value increases the boundaries with lower differences across the windows will be filtered out leaving only the strongest boundaries.

The overall digital value PCA performed better than the overall root zone soil moisture data in transect 3 (Fig. 11.3), but not in transect 10 where the overall root zone soil moisture detected more boundaries. This suggests that root zone soil moisture images might be more useful in areas of high soil moisture such as close to the rivers and streams or agricultural areas. On the other hand, the daily digital value PCA can convey a great deal of information in areas where there is little change in the soil moisture because the digital values will be detecting surficial properties such as color when there are few other physical changes.

11.5 Conclusions

Analysis of multiple images collected over several years revealed consistent response patterns in all data sets. The boundaries of these response patterns as determined by a split moving window technique frequently coincided with soil map unit and/or landform boundaries. Root zone soil moisture identifies boundaries best under conditions when the moisture content is higher. The daily PCA data tends to identify landform boundaries and is more efficient when the soil moisture content is low indicating that in conditions where soil moisture variability is low the calculation using the SEBAL model may be unnecessary.

The advantage to this method is that it is not expert knowledge-based, unlike traditional soil mapping methods. At a low t-value over 70% of previously detected boundaries can be identified, however, the data suggests that there are previously undetected boundaries which may be also identified by this approach. Future work will focus on validating these boundaries and identifying the physical processes which produced changes in the satellite images. Overall, the use of remotely sensed root zone soil moisture for soil boundary delineation shows great promise of becoming a valuable tool in the field of digital soil mapping, especially when combined with the predictive soil maps described in Chapter 32.

Acknowledgments This research was sponsored by the Department of Defense and the National Aeronautics and Space Administration (NASA). Additional support was provided by the Sevilleta National Wildlife Refuge and the New Mexico Geological Society. We are grateful for the review and comments by Dr. Janis Boettinger which greatly improved this manuscript.

References

Ahmad, M.-u.-D., and Bastiaanssen, W.G.M., 2003. Retrieving soil moisture storage in the unsaturated zone using satellite imagery and bi-annual phreatic surface fluctuations. Irrigation and Drainage Systems 17:141–161, doi:10.1023/A:1025101217521.

Allen, R.G., Tasumi, M., Morse, A., Trezza, R., Kramber, W., and Lorite, I., 2007a. Satellite-based energy balance for mapping evapotranspiration with internalized calibration (METRIC) – applications. Journal of Irrigation and Drainage Engineering 133:395–406.

Allen, R.G., Tasumi, M., and Trezza, R., 2007b. Satellite-based energy balance for mapping evapotranspiration with internalized calibration (METRIC) – model. Journal of Irrigation and Drainage Engineering 133:380–394.

Baban, S.M.J., and Yusof, K.W., 2001. Mapping land use/cover distribution on a mountainous tropical island using remote sensing and GIS. International Journal of Remote Sensing 22:1909–1918.

Bastiaanssen, W.G.M., 2000. SEBAL-based sensible and latent heat fluxes in the irrigated Gediz Basin, Turkey. Journal of Hydrology 229:87–100.

Bastiaanssen, W.G.M., Ahmad, M.D., and Chemin, Y., 2002. Satellite surveillance of evaporative depletion across the Indus Basin. Water Resources Research 38:1273–1281.

Bastiaanssen, W.G.M., Menenti, M., Feddes, R.A., and Holtslag, A.A.M., 1998a. A remote sensing surface energy balance algorithm for land (SEBAL): 1. Formulation. Journal of Hydrology 212–213:198–212.

Bastiaanssen, W.G.M., Pelgrum H., Wang, J., Ma, Y., Moreno, J.F., Roerink, G.J., Roebeling, R.A., and Wal, T.v.d., 1998b. A remote sensing surface energy balance algorithm for land (SEBAL). Part 2: Validation. Journal of Hydrology 212–213:213–229.

Compaoré, H., Hendrickx, J.M.H., and Hong, S.-h., Friesen, J., van de Giesen, N.C., Rodgers, C., Szarzynski, J., and Vlek P.L.G., 2008. Evaporation mapping at two scales using optical imagery in the White Volta Basin, Upper East Ghana. Physics and Chemistry of the Earth, Parts A/B/C 33:127–140, doi:10.1016/j.pce.2007.04.021.

Engle, E.M., 2009. Digital soil boundary detection using quantitative hydrologic remote sensing, New Mexico Tech, Socorro NM.

Fleming, K., Hendrickx, J.M.H., and Hong, S.-H., 2005. Regional mapping of root zone soil moisture using optical satellite imagery. Proceedings of the International Society for Optical Engineering, SPIE 5811:159–170. Bellingham, WA 98227-0010, USA.

Gile, L.H., 1975a. Causes of soil boundaries in an arid region: I. Age and parent materials. Soil Science Society of America Journal 39:316–323.

Gile, L.H., 1975b. Causes of soil boundaries in an arid region: II. Dissection, moisture, and faunal activity. Soil Science Society of America Journal 39:324–330.

Hendrickx, J.M.H., Bastiaanssen, W.G.M., Noordman, E.J.M., Hong, S., and Calvo Gobbeti, L.E., 2005. Estimation of regional actual evapotranspiration in the Panama Canal watershed, pp. 315–324, In: Harmon, R.S. (ed.), The Rio Chagres: A Multidisciplinary Profile of a Tropical Watershed, Vol. 52. Springer, Dordrecht, The Netherlands.

Hendrickx, J.M.H., and Hong, S.-H., 2005. Mapping sensible and latent heat fluxes in arid areas using optical imagery. Proceedings of the International Society for Optical Engineering, SPIE 5811:138–146. Bellingham, WA 98227-0010, USA.

Hendrickx, J.M.H., Wierenga, P.J., and Nash, M.S., 1990. Variability of soil water tension and soil water content. Agricultural Water Management 18:135–148.

Hendrickx, J.M.H., Wierenga, P.J., Nash, M.S., and Nielsen, D.R., 1986. Boundary location from texture, soil moisture, and infiltration data. Soil Science Society of America Journal 50:1515–1520.

Hendrickx, J.M.H., Hong, S.-H., Friesen, J., Compaore, H., van de Giesen, N.C., Rodgers, C., and Vlek, P.L.G., 2006. Mapping energy balance fluxes and root zone soil moisture in the White Volta Basin using optical imagery. Proceedings of the International Society for Optical Engineering, SPIE 6239:238–249. Bellingham, WA 98227-0010, USA.

Hong, S.-H., Hendrickx, J.M.H., and Borchers, B., 2009. Up-scaling of SEBAL derived evapotranspiration maps from Landsat (30 m) to MODIS (250 m) scale. Journal of Hydrology 370:122–138; doi:10.1016/j.jhydrol.2009.03.002.

Johnson, W.R., 1984. Soil Survey of Socorro County Area, New Mexico. USDA Soil Conservation Service. US Government Printing Office, Washington, DC.

Kachanoski, R.G., De Jong, E., and Hendrickx J.M.H., 2002. Nonintrusive water content measurement in the field, pp. 497–501. In Dane, J., and Topp, C. (eds.), Methods of Soil Analysis. Part 1. Soil Science Society of America Madison, Madison, WI.

Kachanoski, R.G., De Jong, E., and Wesenbeeck, I.J.V., 1990. Field scale patterns of soil water storage from non-contacting measurements of bulk electrical conductivity. Canadian Journal of Soil Science 70:537–541.

Kachanoski, R.G., Gregorich, E.G., and Wesenbeeck I.J.V., 1988. Estimating spatial variations of soil water content using non-contacting electromagnetic inductive methods. Canadian Journal of Soil Science 68:715–722.

Karale, R.L., Venkataratnam, L. Sehgal, J.L., and Sinha, A.K., 1991. Soil mapping with IRS-1A data in areas of complex soils capes. Current Science, Special issue on Remote Sensing for Natural Resources Development 61:198–203.

Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), 2007. Digital Soil Mapping. An Introductory Perspective, Vol. 31, pp. 1–620. Elsevier, Amsterdam.

Pelgrum, H., and Bastiaanssen, W.B.M., 1996. An intercomparison of techniques to determine the area-averaged latent heat flux from individual in situ observations: a remote sensing

approach using the European Field Experiment in a Desertification-Threatened Area data. Water Resources Research 32:2775–2786.

Roerink, G.J., Bastiaanssen, W.G.M., Chambouleyron, J., and Menenti, M., 1997. Relating crop water consumption to irrigation water supply by remote sensing. Water Resources Management 11:445–465.

Scott, C.A., Bastiaanssen, W.G.M., and Ahmad, M.-U.-D., 2003. Mapping root zone soil moisture using remotely sensed optical imagery. Journal of Irrigation and Drain Enginneering 129:326-335.

Sheets, K.R., and Hendrickx, J.M.H., 1995. Non-invasive soil water content measurement using electromagnetic induction. Water Resources Research 31:2401-2409.

Taylor, S.A., and Ashcroft G.L., 1972. Physical edaphology. The physics of irrigated and nonirrigated soils. Freeman and Co., San Francisco, CA.

van den Hurk, B.J.J.M., Bastiaanssen, W.G.M., Pelgrum, H., and van Meijgaard, E., 1997. A new methodology for assimilation of initial soil moisture fields in weather prediction models using Meteosat and NOAA data. Journal of Applied Meteorology 36:1271–1283.

Webster, R., 1973. Automatic soil-boundary location from transect data. Mathematical Geology 5:27–37.

Webster, R., 1978. Optimally partitioning soil transects. Journal of Soil Science 29:388–402.

Wierenga, P.J., Hendrickx, J.M.H., Nash, M.H., Daugherty, L., and Ludwing, J., 1987. Variation of soil and vegetation with distance along a transect in the Chihuahuan desert. Journal of Arid Environments 13:53–63.

Section C
Soil Inference Systems

Chapter 12
Homosoil, a Methodology for Quantitative Extrapolation of Soil Information Across the Globe

B.P. Mallavan, B. Minasny, and A.B. McBratney

Abstract In many places in the world, soil information is difficult to obtain and can be non-existent. When no detailed map or soil observation is available in a region of interest, we have to extrapolate from other parts of the world. This chapter will discuss the Homosoil method, which assumes homology of soil-forming factors between a reference area and the region of interest. This includes: climate, physiography, and parent materials. The approach will involve seeking the smallest taxonomic distance of the *scorpan* factors between the region of interest and other reference areas (with soil data) in the world. Using the digital information of soil climate from the Climate Research Unit (CRU) (solar radiation, rainfall, temperature, and evapo-transpiration), topography from the HYDRO1k (elevation, slope, and compound topographic index), and lithology of the world on a $0.5° \times 0.5°$ grid, we calculated Gower's similarity index between an area of interest and the rest of the world. The rules calibrated in the reference area can be applied in the region of interest realising its limitations and extrapolation consequences.

Keywords Global soil mapping · Homoclime · Climate · Map extrapolation · Soil forming factors

12.1 Introduction

In many places around the world, legacy soil information is difficult to obtain and is practically nonexistent. With limited time and practically no resources to collect new soil samples in these areas, we have to develop a new methodology for estimating soil attributes based on our knowledge of soil forming factors and pedogenesis.

In Australia, at the system level (ASRIS, mapping scale 1:100,000) only about 60% of the intensive land-use zones, and less than 5% of rangelands, are covered by soil information (McKenzie et al., 2005). Countries in South America, South

A.B. McBratney (✉)
Faculty of Agriculture, Food & Natural Resources, The University of Sydney, Sydney, NSW 2006, Australia
e-mail: alex.mcbratney@sydney.edu.au

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_12, © Springer Science+Business Media B.V. 2010

East Asia, and Africa only have 10–30% of landscape coverage for maps finer than 1:100,000 (Hartemink, 2008). When no detailed maps or soil observations are available in a region of interest, we have to extrapolate from other areas or other parts of the world. If we are only dealing with global modelling at a coarse resolution, we can extrapolate soil observations that are available from other similar areas (that are geographically close) or by using spatial interpolation or spatial soil prediction functions (Schenk and Jackson, 2005). However if we wish to map the soil at a much finer resolution, such as the target resolution of GlobalSoilMap.net (90 m $\times$ 90 m), geographical extrapolation may not work (See Chapter 33). In this case, a new approach is required that allows extrapolation of spatial soil prediction functions, realising that similar areas could be in other parts of the region or other parts of the world.

Given severe prior soil information crisis, we introduce a new method for digital soil mapping called *Homosoil*, which aims to find areas in the world with similar soil-forming factors for the purpose of extrapolation of soil mapping rules.

The idea of extrapolating environmental variables has been explored previously, mainly to identify areas in the world with similar climates for crop production. Prescott (1938) coined the term "homoclime" referring to areas or regions in the world with similar climate. In Australia, studies of homoclimes have been carried out with reference to particular economic crops (Prescott, 1938, 1943). However these early studies simply compared some climatic variables that were considered to be critical and no quantitative similarity index was developed. Russell and Moore (1970) define homoclimates (similarity between weather stations) based on various similarity coefficients. They also noted that the detection of areas of similar climate could be of interest in pedology.

Jones et al. (2005) revisited the homoclime approach and defined the *homologue* approach to determine which crops could be grown at a specific site in the world. The principle is that crops will perform in a similar fashion in similar environments wherever they happen in space and time. The algorithm calculates the similarity between monthly climate variables and soil data at the place of interest and elsewhere in the tropical world.

Defining areas with similar soil-forming factors is not as straightforward as homoclime estimation. While climate defines the weathering regime of the soil, the parent materials and the age of the soil play a more important role in defining the type of soil that is formed. Topography and landuse also influence the distribution of soil. Nevertheless, results from spatial data mining from the Australian-wide mapping project ASRIS (Bui et al., 2006) showed that the state factors of soil formation form a hierarchy of interacting variables, with climate being the most important factor at a continental scale and different climatic variables dominating in different regions. It was shown that lithology is almost equally important to climate in defining broad-scale spatial patterns of soil properties, whereas shorter-range variability in soil properties appears to be driven more by terrain variables.

This paper will investigate methods to find areas in the world with similar soil-forming factors as expressed by environmental covariates. We will demonstrate

an algorithm for calculating the taxonomic distance of the environmental variables between the region of interest and other reference areas with good soil data coverage.

12.2 Conceptual Framework

The homosoil approach involves finding areas in the world with similar soil-forming factors. One can start by identifying areas where reliable and adequate soil maps or soil data are available, then use them to search for the homologous areas in environmental space. The methodology of this approach is represented in Fig. 12.1 and summarised as follows:

(1) Mapping areas in the world for soil donor and recipient sites. The first step is to delineate areas in the world where sufficient soil information is available. There are many areas in the world where soil information is not available at a resolution finer than 1:100,000. The level of soil information coverage in areas of the developing world is only available as estimates of percent coverage in the country based on questionnaires (Hartemink, 2008; Zinck, 1990). With the development of GlobalSoilMap.net consortium, we can now access global soil information and identify areas in the world where adequate soil information is available. These areas are

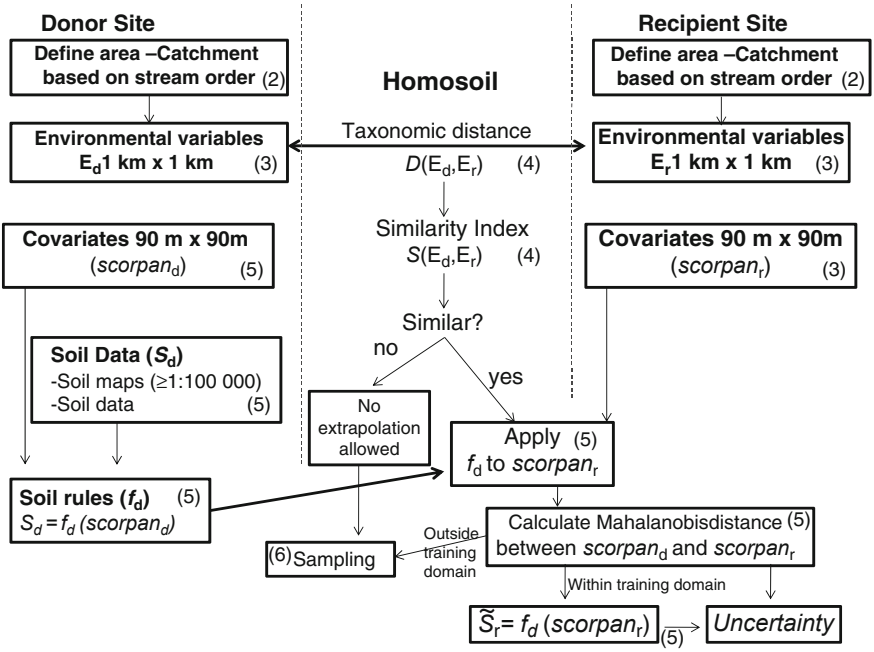


Fig. 12.1 A framework for developing Homosoil. *Numbers* in *brackets* represent the steps in the methods as outlined in the proposal

called "donor sites", where potentially, the soil mapping rules developed can be exported to other areas. Areas with inadequate soil information are called "recipient sites", where soil rules from the donor sites may be applied.

(2) Definition of donor and recipient sites. We need to define the appropriate scale for the donor and recipient sites, which can be represented as catchments. While calculation of the similarity index can done by a pixel-by-pixel comparison; for soil landscape analysis, it is more realistic to define the areas or sites as a catchment based on stream order. Mourier et al. (2008) found that the Strahler stream order is a crucial hierarchical descriptor of the distribution pattern of soil. Our initial investigation suggests that the 4th and 5th order catchment maybe appropriate for comparison of the soil-forming factors. This will need further investigation.

(3) Building a spatial database for environmental variables. A spatial database that serves as a proxy to represent soil-forming factors and will allow us to define areas of homosoil. This spatial database will contain information on environmental variables over large areas (e.g. a continent) with a resolution of 1 km $\times$ 1 km. The important environmental variables governing soil formation are: climate, topography, lithology, and age. At the global scale, only coarse climate and lithology maps are available at a resolution of $0.5^\circ \times 0.5^\circ$. The global climate data are available from the ERA-40 reanalysis and the Climate Research Unit (CRU) dataset (New et al., 1999) and the lithology data are available from a global digital map (Durr et al., 2005). However, for global soil map prediction, a finer resolution is required to capture the local environmental heterogeneity. Therefore we shall build a database of environmental variables (climate, lithology, topography, age) at a resolution of 1 km $\times$ 1 km.

We can derive global topographic information from elevation data provided by the SRTM. Elevation, slope, and compound topographic index (CTI) are important topographic variables. However regional-scale indices representing physiography, erosional and depositional areas may play more important roles in this broad-scale landscape analysis. Lithology must reflect the soil parent materials, and the age of the land surface is also an important factor in defining the type of soils. Age of the land can be defined in a number of ways, e.g. the time since the last glaciation (Adams, 1997).

(4) Calculating the taxonomic distances of environmental variables. The homology of soil-forming factors requires aggregation of the various components: climatic factors (homoclime), lithology (homolith), topography (homotop), and age (homochron).

Similarity Indices

The monthly climate data need to be converted to climatic indices (Booth et al., 1987) so that the taxonomic distance of the climate variables can be calculated. This also attempts to standardise climatic events across various parts of the world, such as the differences arising between the Northern and Southern hemispheres. For the climatic variables (e.g. rainfall, temperature, solar radiation and evapotranspiration) we can calculate various indicators, such as: annual mean, mean of the driest

period, mean of the wettest period, annual range, driest quarter mean, wettest quarter mean, coldest quarter mean, hottest quarter mean, seasonality, etc. Thus we obtain an abundance of climatic information. Using these climatic variables, we can calculate the climatic similarity (homoclime) between two areas either by Mahalanobis distance or by a Gower similarity measure (Booth et al., 1987; Gower, 1971). The Gower distance measures the similarity S_{ij} between sites i and j, each with p number of variables, standardised by the range of the variable k:

$$S_{ij} = \frac{1}{p} \sum_{k=1}^{p} \left(1 - \frac{|x_{ik} - x_{jk}|}{\text{range } k} \right)$$

where p denotes the number of climatic variables; $|x_i - x_j|$ represents the absolute difference of climate variables between site i and j. The similarity index has a value between 0 and 1. Initial investigations suggest that the Gower's similarity works better than the Mahalanobis distance, but further investigation is needed. All studies of homoclimes are based on comparison between two pixels on a grid, whereas for homosoil, the comparison will be based on two catchments.

Similarity Index for Soil-Forming Factors

Considering the scale and the resolution of this study and the available global and regional data, the climatic factor is probably the most important and reliable soil-forming factor (Bui et al., 2006). However other factors are still equally important, therefore we need to define our *Homosoil* based on *homoclimes*, areas with similar lithology (*homolith*), similar age (*homochron*), and similar topography (*homotop*). One approach could be to apply a hierarchical approach, i.e. by first calculating a similarity index based on the most important factor (e.g. climate or age) and then refining the selection of homologous areas based on the second most important factor (e.g. lithology), and so on. However this approach requires an assumption of the hierarchical nature of the soil-forming factors, which may not work in every area. Another approach includes assigning different weight to each of the factors. We will make a thorough comparison of the various approaches

(5) How to import soil mapping rules. The principle of the method for building soil mapping rules is based on the *scorpan* spatial soil prediction functions (McBratney et al., 2003). The function or mapping rules f is usually calibrated in an area where soil information is available. This function is then applied to the area of interest for interpolation (and limited extrapolation) of soil properties. In this project, we will extend the application of f into geographic areas other than the one used for calibration, in this sense f is used for extrapolation in other areas in the world. The requirement for this extrapolation is that both areas (the donor and recipient) should have the same *scorpan* factors, which cover both areas at a resolution of 90 m × 90 m. The *scorpan* factors define soil distribution at a local catchment scale. This is in contrast with the broader district scale environmental factors (1 km × 1 km) for defining soil-forming factors. The commonly used function

for f is the classification or regression-tree approach (Bui and Moran, 2001). In the Bui and Moran (2001) approach, the decision-trees have two levels: one is used to capture mapping rules from the training area, and another one is used to define the domain over which those rules can be extrapolated. We can also evaluate the appropriate domain of the mapping rules f using distance metrics of the *scorpan* factors (Tranter et al., 2009). The Mahalanobis distances will be used to determine the distance between the mean of the donor's *scorpan* factors and the recipient's *scorpan* factors. Distances exceeding a designated cutoff limit are deemed distinct from the calibration data and as such unsuitable for function application.

12.3 Examples

We illustrate the application of the homosoil approach using coarse resolution global climate and elevation data. This is the first attempt to demonstrate the homosoil concept.

12.3.1 Global Data

We compiled 3 data sources for the global climate, topography, and lithology data, all on a regular $0.5° \times 0.5°$ grid as matrix of 360×720 pixels. In total, there are 62 254 pixels with all data available. The data were kindly provided by Prof. Marc Bierkens and Dr. Rens van Beek from Utrecht University.

We used the long-term mean monthly temperature, rainfall, solar radiation and evapotranspiration data to represent the climate. We also use the DEM representing topography, and lithology, which gives broad information on the parent material. The climate data come from the ERA-40 reanalysis and CRU dataset. Maps represent the monthly average course over the year. The period corresponded to the datasets are 1961–1990 for the CRU (temperature, rainfall and solar radiation), and September 1957–August 2002 for the ERA-40 analysis (evapotranspiration). Temperature data is in Celsius degree, and we have the monthly value for the minimum temperature, mean temperature and maximum temperature. Rainfall value is in meter per day. Evapotranspiration is in meter per day and it is calculated with Penman formula. More details on the datasets are available on the website http://www.ipcc-data/obs/cru_climatologies.html.

The DEM is from the HYDRO1k dataset, which includes the mean elevation, slope, and compound topographic index (CTI). Details of the dataset are available at: http://eros.usgs.gov/#/Find_Data/Products_and_Data_Available/gtopo30/hydro (accessed 21 April 2010).

The lithology is from a global digital map (Durr et al., 2005) with 7 values which represent the different broad groups of parent materials. The lithology classes are: non- or semi-consolidated sediments, mixed consolidated sediments, silic-clastic sediments, acid volcanic rocks, basic volcanic rocks, complex of metamorphic and igneous rocks, and complex lithology.

12.3.2 Climatic Indices

The monthly climate data was first converted to climate indices (Booth et al., 1987) so that the taxonomic distance of the climate variables can be calculated. This also attempts to standardize the climatic events that can vary in various parts of the world such as the difference in the Northern and Southern hemisphere.

For each of the 4 variables climatic (rainfall, temperature, solar radiation and evapotranspiration), we calculated 13 indicators: annual mean, mean for the driest month, mean at the wettest month, annual range, driest quarter mean, wettest quarter mean, coldest quarter mean, hottest quarter mean, lowest ET quarter mean, highest ET quarter mean, darkest quarter mean, lightest quarter mean, and seasonality. Thus we obtain 52 climate variables.

12.3.3 Similarity Index

The climatic similarity between the indices values of two points of the grid is calculated by a Gower similarity measure (Booth et al., 1987; Gower, 1971). Considering the scale and the resolution of this study and the available global data ($0.5° \times 0.5°$), the climatic factor is probably the most important and reliable soil forming factor (Bui et al., 2006). Thus as our first attempt, we define our *Homosoil* based on three steps. First is to identify the *homoclime* around the world. The next step, within the *homoclime*, we find areas with similar lithology (*homolith*) and similar topography (*homotop*).This is summarized as a decision tree in Fig. 12.2. For the lithology

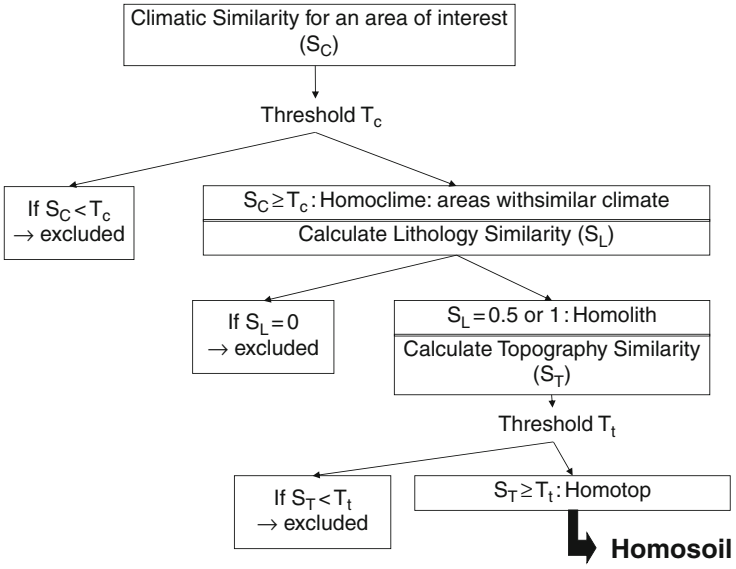


Fig. 12.2 A simple decision tree for homosoil

similarity index, a value of 1 is given when the same lithological class is encountered, 0.5 if the complex lithology class is encountered, and otherwise a value of 0 is given.

12.4 Results

12.4.1 Homoclime

To assess the method we calculate the similarity for a sample of 42 known locations, and see the coherence of the homoclime results. The locations are chosen following the climate classification of the ecozones (Schultz, 2005). Figure 12.3 shows the *homoclime* results for 4 different places in the world with variation in seasonal climate:

- A Mediterranean climate (Montpellier) where the rainfall and the temperature are seasonal
- Sydney where the temperature is seasonal but not the rainfall
- Harare where the climate is tropical with a seasonality of rainfall
- Manaus with an equatorial climate without seasonality.

The results showed that the Gower similarity index gives areas that are in accordance with the ecozones of the world (Schultz, 2005). It seems to be an appropriate method to the homoclime analysis.

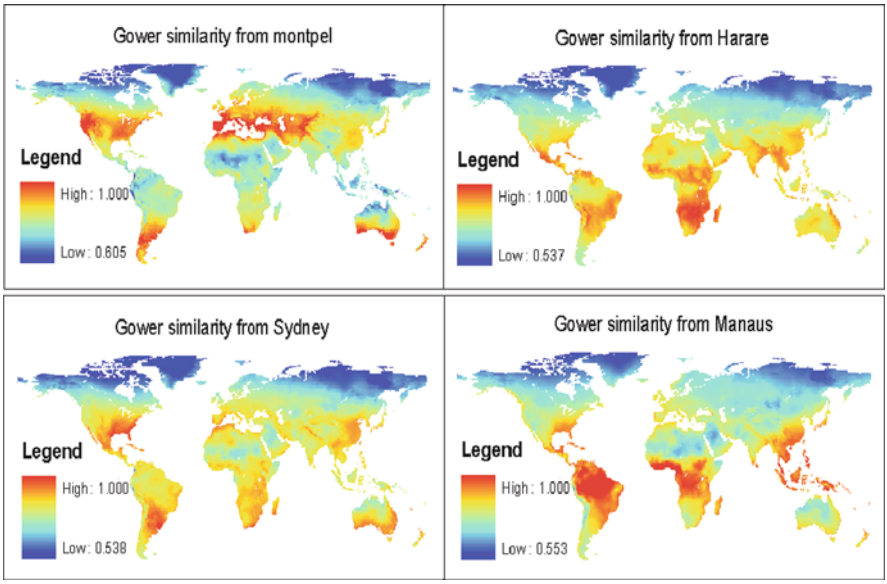


Fig. 12.3 Homoclime degree for 4 locations around the world

12.4.2 Homosoil

Figure 12.4 shows the similarity index for homoclime, homolith, and homotop for Harare. A combination for these factors gives the homosoil similarity index for the area. This is illustrated in Fig. 12.5 for defining homosoil for an area in the South of Somalia (1.5°N 42°E). The first step is to calculate the climate similarity index for Somalia; the second step is to select a threshold for the climate similarity index for homoclime. The selection of similarity based on a threshold of 1-std. dev. of the similarity index. The second step is within the homoclime, to select areas with lithology similarity $\geq$ 0.5. This is homolith. The final step is, within the area of homoclime and homolith, to select areas with similar topography. We select areas with the topography similarity index $>$ 1 – std. dev. The final selection represents area of homosoil.

Homosoil only identifies area with similar soil forming factors, given an area with no soil data, we find other homosoil areas. Within the homosoil areas, we identify are where soil data is available. The soil landscape rules or spatial prediction function from the known area then can be transferred to the area with no soil data. Figure 12.6 shows an illustration of the homosoil concept. Soil information in the Democratic Republic of Congo (2° N, 20.5° E) is not available, so we calculated homosoil areas in the world. It identifies an area in Peru (12.5° S, 69.5° W) as homosoil, and in this area soil information are available. Therefore the soil-landscape rules in Peru can be exported to areas in Congo.

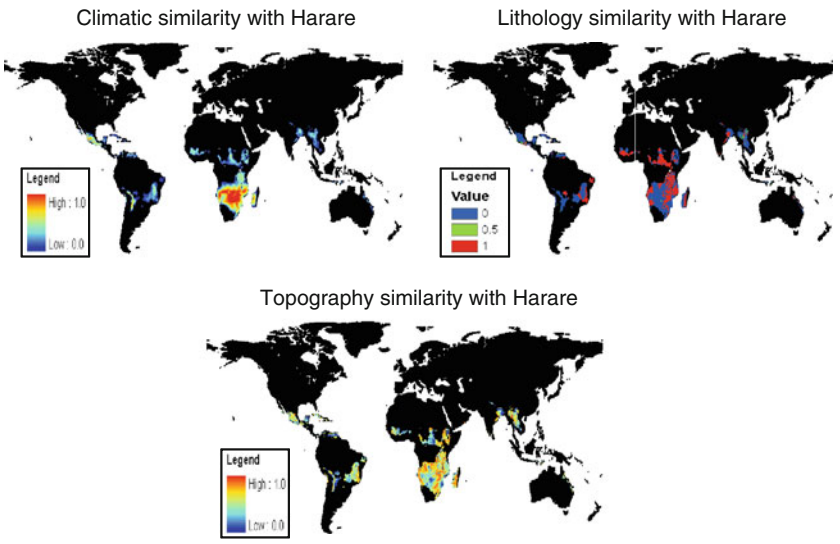
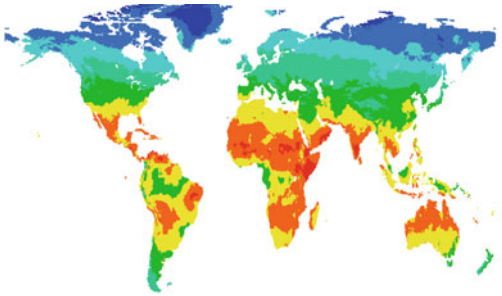


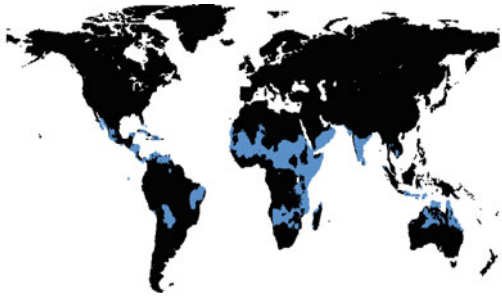
Fig. 12.4 Homoclime, homotop, and homolith for Harare

(a)



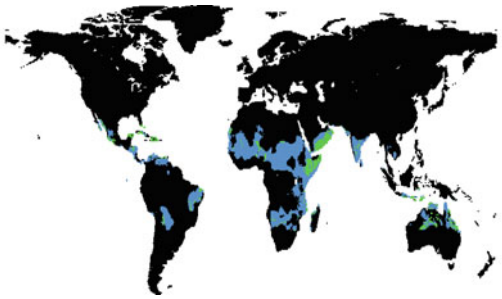
Climate similarity index for an area in the South of Somalia 1.5°N 42°E with values scaled between 0 and 1.

(b)



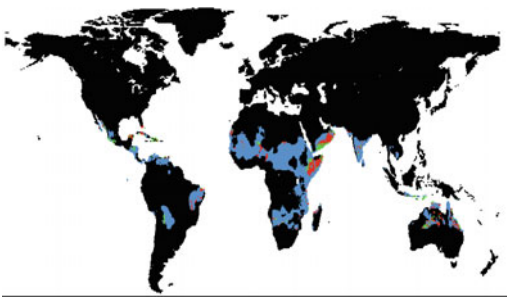
Homoclime. The first selection of similarity based on a threshold of 1-std dev of the similarity index.

(c)



Homolith
The second step is to select similar lithology ≥ 0.5.

(d)



Homotop, Homosoil.
The final step, within the area of homoclime and homolith, to select areas with similar topography. The areas in red represent the selection with value > 1- std dev of similarity index.

Fig. 12.5 Decision tree for selecting homosoil for an area in Southern Somalia

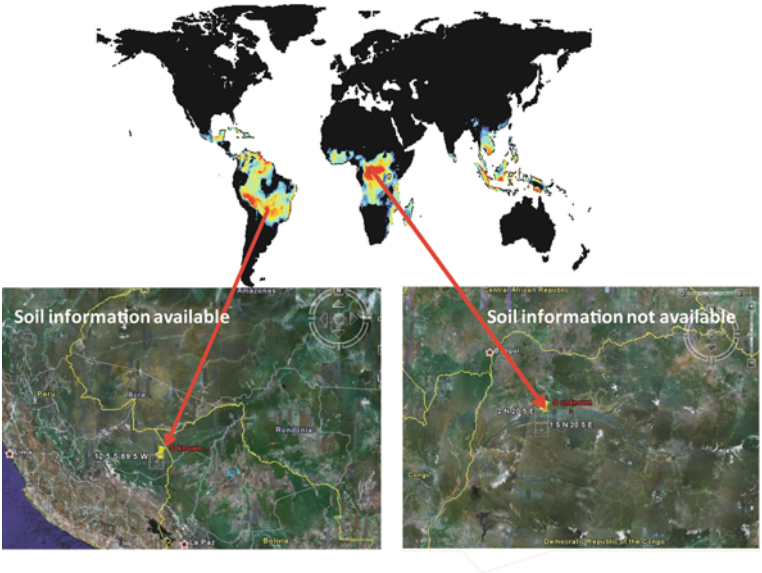


Fig. 12.6 An illustration of the homosoil concept. Soil information in the Democratic Republic of Congo (1.5°N, 20.5°E) is not available, meanwhile areas in Peru (12.5°S, 69.5°W) soil information is available. Therefore the soil-landscape rules in Peru can be exported to areas in Congo

12.5 Discussion and Conclusions

The homosoil approach allows us to export soil-landscape rules from one area to unmapped area. The first approach is based on the hierarchical nature of soil forming factors: climate, lithology and topography. In reality the interaction within these three factors and also biological factors maybe the same. We choose this approach as the first step considering the scale we are working, and at the moment, we can only validate the approach to well known global climate pattern. We have attempted to define homosoil areas with equal weighting for climate, lithology, and topography and the result is not promising, as the result showed patches of local homosoil near the location of interest.

Biological factors and local topography are important, but this factor hopefully will be taken care of when we use the local rules which are exported from other areas in the world. Of course, extreme landuse will influence soil properties, thus homosoil results with soil-landscape rules from an area with homosoil but completely different landuse may not work. An important factor of the age is not yet taken into account in this example (See Chapter 4 for the development of the age factor).

The decision-tree approach has other limitations, such as it is an aggregated result with no uncertainty estimation. The assumption is that it is not possible to have a similar area if the lithology group is different. This condition may restrictive con-

sidering the rough scale of the lithology and lithology groups. A better taxonomic distance for the lithological groups is required.

There are other important issues and questions that still need to be addressed:

- Which are the most important soil forming-factors? This of course partly depends on the scale and resolution of the global dataset, and their accuracy.
- How do we define the threshold that quantifies similarity for homoclime, homolith, and homotop? We need empirical studies to find the appropriate values.
- Uncertainty estimation, which will involve the uncertainty of climate similarity, lithology and topography. At the local scale, the uncertainty due to landuse and local topography may be more important.

References

Adams, J.M., 1997. Global land environments since the last interglacial. Oak Ridge National Laboratory, TN, USA, http://www.esd.ornl.gov/ern/qen/nerc.html

Booth, T.H., Nix, H.A., Hutchinson, M.F., and Busby, J.R., 1987. Grid matching: a new method for homoclime analysis. Agricultural and Forest Meteorology 39:241–255.

Bui, E.N., Henderson, B.L., and Viergever, K., 2006. Knowledge discovery from models of soil properties developed through data mining. Ecological Modelling 191:431–446.

Bui, E.N. Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103:79–94.

Durr, H.H., Meybeck, M., and Durr, S.H., 2005. Lithologic composition of the Earth's continental surfaces derived from a new digital map emphasizing riverine material transfer. Global Biogeochemical Cycles 19, GB4S10.

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–871.

Hartemink, A.E., 2008. Soil map density and a nation's wealth and income. In: Hartemink, A.E., McBratney, A.B., and Mendonca-Santos, M.L. (eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht.

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

Jones, P.G., Diaz, W., and Cock, J.H., 2005. Homologue, a computer system for identifying similar environments throughout the tropical world, version Beta a.0. Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

McKenzie, N.J., Jacquier, D.W., Maschmedt, D.J., Griffin, E.A., and Brough, D.M., 2005. ASRIS Australian Soil Resource Information System. Technical specifications. Version 1.5. October 2005. Australian Collaborative Evaluation Program.

Mourier, B., Walter, C., Merot, P., 2008. Soil distribution in valleys according to stream order. Catena 72:395–404.

New, M., Hulme, M., and Jones, P.D., 1999. Representing twentieth century space-time climate variability. Part 1: Development of a 1961–90 mean monthly terrestrial climatology. Journal of Climate 12:829–856.

Prescott, J.A., 1938. The climate of tropical Australia in relation to possible agricultural occupation. Transactions of the Royal Society of South Australia 62:229–240.

Prescott, J.A., 1943. The Australian homoclime of the zone of natural occurrence of *Parthenium argentatum*. Transactions of the Royal Society of South Australia 67:312–318.

Russell, J.S., and Moore, A.W., 1970. Detection of homoclimates by numerical analysis with reference to the brigalow region (eastern Australia). Agricultural Meteorology 7:455–479.
Schenk, H.H., and Jackson, R.B., 2005. Mapping the global distribution of deep roots in relation to climate and soil characteristics. Geoderma 126:129–140.
Schultz, J., 2005. The Ecozones of the World: the Ecological Divisions of the Geosphere. Springer-Verlag, Stuttgart, Germany.
Tranter, G., McBratney, A.B., and Minasny, B., 2009. Using distance metrics to determine the appropriate domain of pedotransfer function predictions. Geoderma 149:421–425.
Zinck, J.A., 1990. Soil Survey: Epistemology of a Vital Discipline. ITC, Enschede, the Netherlands.

Chapter 13
Artificial Neural Network and Decision Tree in Predictive Soil Mapping of Hoi Num Rin Sub-Watershed, Thailand

R. Moonjun, A. Farshad, D.P. Shrestha, and C. Vaiphasa

Abstract The demand for high-resolution soil mapping is growing increasingly, in particular for the purpose of land degradation studies. The objective of this study focuses on applying the methods for digital predictive soil mapping in inaccessible, land degradation-prone areas. Artificial Neural Network (ANN) and Decision Tree (DT) were employed within the GIS environment to comply with the complexity of the soil forming factors governing the soil formation. Following the principles of the geopedologic approach to soil survey, a digital predictive soil mapping was carried out in Hoi Num Rin sub-watershed, covering an area about 20 km^2. Both ANN and DT were applied to properly integrate the parameterized soil forming factors. To describe soil predictors to train the ANN and to formulate the decision trees, 4 organism types, 7 relief type units, 9 lithological units, 3 time series, 4 landscape units and 8 landform units were extracted from the map and databases. The results, the 10 soil class names were extrapolated to the unsampled areas to obtain the geopedologic map. In conclusion, the geopedologic approach to soil survey, which is based on understanding of landscape-soil relationship, is helpful to obtain spatial soil information in inaccessible areas, using ANN and/or DT are useful techniques in modeling the complex interactions among the soil forming factors. The difference, however, is that ANN, once it is well learnt, is faster, thus more recommendable in terms of time and cost saving.

Keywords Geopedologic approach · Predictive soil map · Digital soil map · Artificial neural network (ANN) and Decision Tree (DT)

R. Moonjun (✉)
International Institute of Geo-information Sciences and Earth Observation (ITC), Enschede, The Netherlands
e-mail: moonjun13562@itc.nl

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_13,

13.1 Introduction

There is an ever growing need for soil data in land degradation studies, particularly in deforested mountains. In Thailand, most of the available soil maps only cover valley floors. Conventional soil mapping in mountainous areas is said to be too expensive, mainly because of long fieldwork periods. Digital soil mapping, also known in the literature as predictive soil mapping, is considered as the solution. Digital (predictive) soil mapping is the computer-assisted production of digital maps of soil classes and/or soil properties (McBratney et al., 2003; Scull et al. 2003). Geographic information dystems (GIS), digital elevation models (DEM) and remote sensing are widely used in different ways in predictive soil mapping. So far, fuzzy logic (Burrough et al., 1992; McBratney and Odeh, 1997), statistical techniques (De Gruijter et al., 1997) (see also Chapter 19) and, to lesser extent, artificial neural networks (ANN) (Zhu, 2000) have been used.

In digital soil mapping, the equation $S = f(cl, o, r, p, t)$ is the theoretical backbone, where *cl*: climate, *o*: organisms, *p*: parent material, *r*: relief, and *t*: the time for soil development. The state factor theory was originally developed by Jenny (1941) to relate soil properties to the state factors abbreviated in *clorpt*. Digital terrain model (DTM) is an ordered array of numbers that may be translated to represent spatial distribution of some terrain attributes (forming the *clorpt* backbone), which can be derived from DEM (Bishop and Minasny, 2006) (see also Chapter 5).

McBratney et al. (2003) introduced the SCORPAN model, a modification of the Jenny's applied to digital soil mapping, to explain the relationships between soil and other spatially referenced environmental factors or covariates: $S_{(c,a)} = f(S, C, O, R, P, A, N,)$ (see also Sections 4.1 and 14.1). All environmental covariates are represented by Jenny's letters, except for soil age (time) abbreviated by A, and a new covariate "spatial position" represented by N. The lower case c and a with S symbolize "soil class" and "soil attribute", respectively.

The geopedologic approach to soil survey (Zinck, 1988/1989; Zinck and Valenzuela, 1990) opens the way to properly define the interactions between the state forming factors and the relationships between the soil and/or its selected properties with the environmental components (e.g., climate, geology, etc.). The approach includes image interpretation, sample area selection, and extrapolation of the soil information from the sample areas to the unsampled areas (Moonjun, 2007). Aerial photo interpretation (API) leads to the delineation of landforms (see Section 29.2.3.1). In the absence of API, DTM and Landsat remotely sensed data can be used.

The overall objective of this study was to develop a GIS-based method for soil mapping in poorly accessible areas, in an efficient and cost-effective manner. This study focuses on the geopedologic approach as the conceptual framework and on artificial neural network (ANN) and the decision tree (DT) for integration purposes. The intent of this chapter is to demonstrate some of the difficulties of how to predict the soil (class) content of pre-established (geopedologic-based) map units.

13.2 Material and Methods

13.2.1 Study Area

The study area, Hoi Num Rin, is located in the south of Mae Ka Chan district, in Chiang Rai province, Thailand, between 99°2′ and 99°7′ East longitude, and 19°25′ and 19°27′ North latitude, with an altitude varying between 700 and 1,720 m above sea level.

13.2.2 Materials and Methods for Deriving the Geoform Map

We acquired aerial photographs (1:25,000 and 1:4,000), a digital orthophoto (1:4,000; accuracy of 1 m), topographic map sheet (4748-II, of Royal Thai Survey Department; 1:50,000), and the geological map by the Mineral Resource Department (1:50,000; Fig. 13.1a). Various maps of the study area were prepared and produced by Land Development Department (LDD) (Soil Survey Division, 2004): digital contour map with an elevation interval of 2-m, drainage map, road map, pond map, and soil map with the map unit description (all maps are at scale 1: 25,000). Climate data of the study area (temperate, rainfall and humidity) were obtained

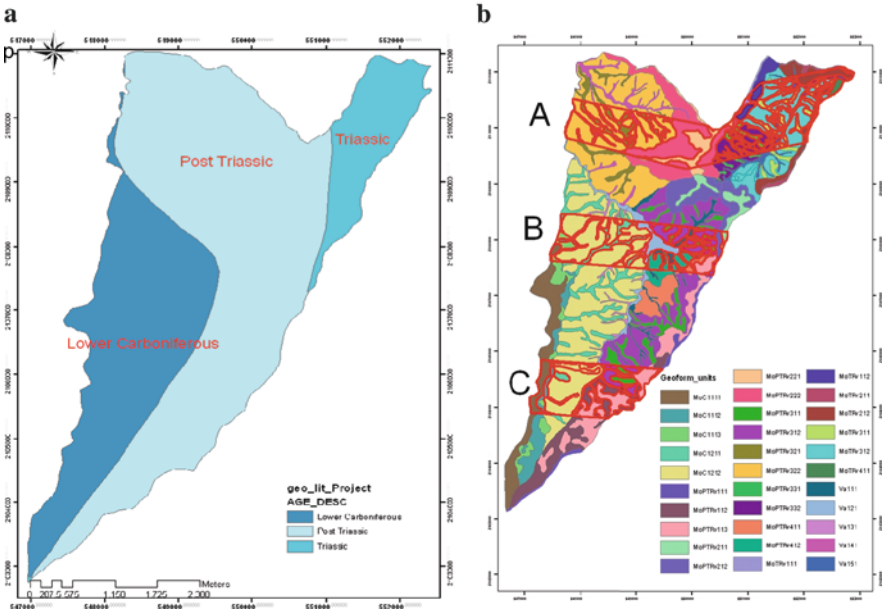


Fig. 13.1 Geological map (**a**) and geoform map with three sample areas (**b**)

from Hoi Num Rin Royal project station. Laboratory data (soil chemical properties for cation exchange capacity [CEC] and base-saturation) obtained from Office of Agriculture Research and Development Region 1.

All digital data were processed in ArcGIS and/or ENVI. A DEM was created from the contour line map, with contour interval of 2-m. The DEM was then used to generate some of the spatial data such as aspect, curvature, and hill-shade, which form backbone of some of the state factors (e.g., relief, parent material).

A visual image interpretation map at scale 1:25,000 (geoform map, see Fig. 13.1b) was prepared by applying the geopedologic approach (Zinck and Valenzuela, 1990). The flow diagram for creating the image interpretation map is shown in the pre-field work section of Fig. 13.2. Several iterations of maps were produced and/or improved. For instance, the image interpretation map (here referred to as the geoform map) was improved, adjusted also to the commonly used categorical legend, where landscape, relief-type, lithology and landform are used (Table 13.1).

Four major geoform units were defined in the mountainous areas, namely (1) the mountain in the Carboniferous lithologic formation (coded as MoC_1 on the geology map), characterized by ridges/hills stretching northeast to southwest; (2) the mountain in Triassic volcanic rock (coded as MoTRv); (3) the mountain in Post-Triassic volcanic rocks (coded as MoPTRv), and valley (coded as Va). Table 13.1 is a part of the legend of the geoform map in Fig. 13.1b where details on the types of relief, lithology, and landform are indicated.

13.2.3 Sample Areas (Training Map)

Sample areas were selected based on landscape, relief, lithology, and landform. Selection of the sample areas is an important step, as these are the areas where landform units are examined for their soil (class) content. Three sample areas were selected and labeled as A, B, and C (Fig. 13.1b). Area A in the northern part of the study area crossed through two dominant of lithologic formations indicated on the geology map as MoPTRv and MoTRv. The second area, B, in the middle part of the area occurs in the lithologic formations MoC_1 and MoPTRv. The third area, C, in the southern part crosses MoC_1 and MoPTRv lithologic formations. The sample areas were used to train the relationships between the soils and the environmental conditions. Several observations were made in these areas to get acquainted with the soil-landform relationships.

The three sample areas A, B, and C were studied in detail using 57 soil observations. There were 32 landform units represented in these areas (Fig. 13.1b, Table 13.1).

The sample area A included 6 soil units, classified at subgroup level as: Typic Hapludalfs, Typic Paleudalfs, Typic Paleudalfs/Typic Hapludalfs, Typic Hapludalf/Typic Paleudalfs, Typic Udifluvents, and Typic Udifluvents/Typic Argiudolls (Table 13.2).

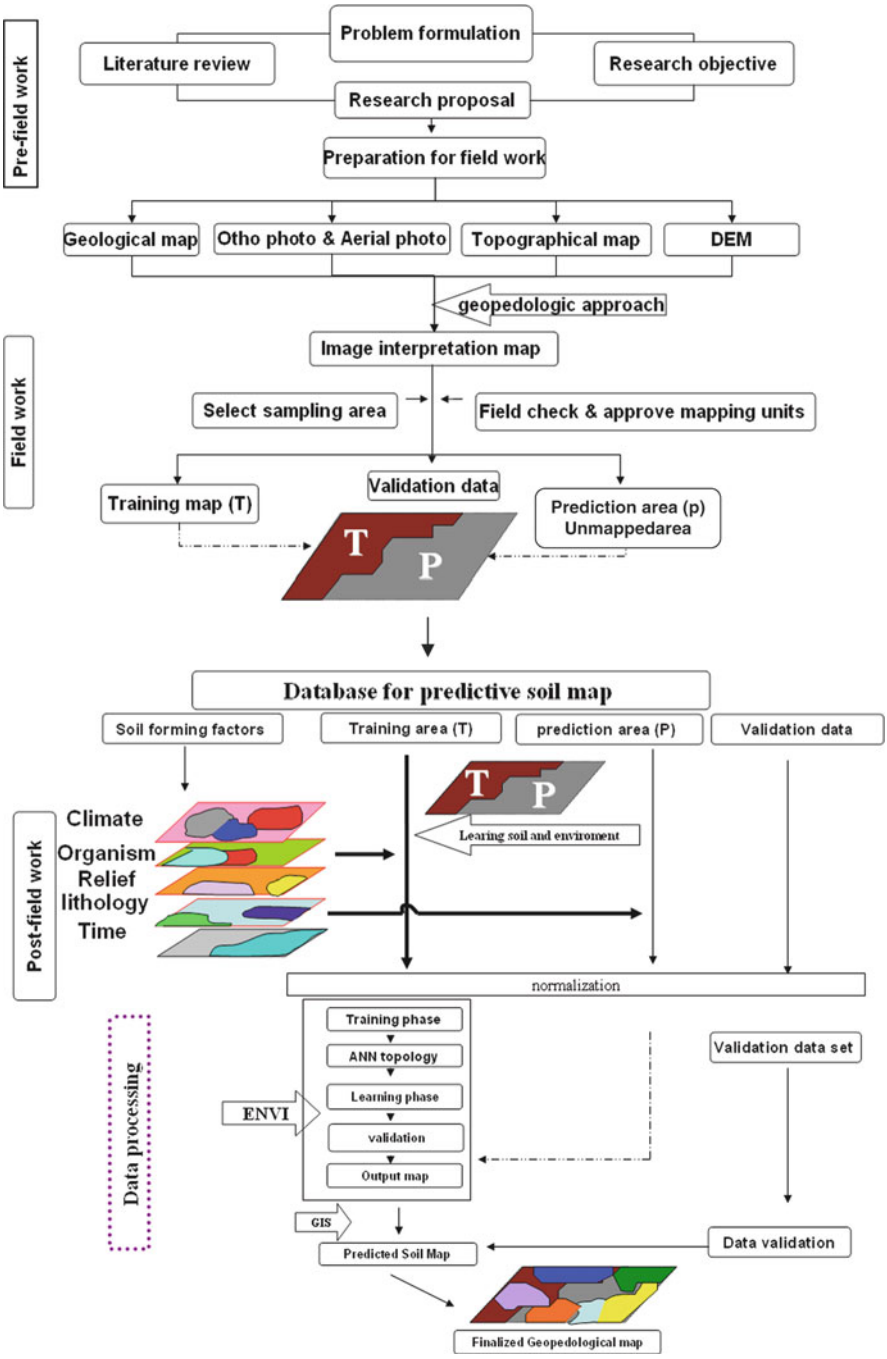


Fig. 13.2 Methodological framework

Table 13.1 The geoform units in Hoi Num Rin sub-watershed

Landscape	Relief	Lithology	Landform	Symbol
Mountain in carboniferous	Ridge	RS and CV from shale and phyllite	Summit and shoulder complex	MOC1111
			Steep/very steep of slope com	MOC1112
			Gentle/moderately of slope complex	MOC1113
	Spur	RS and CV from shale and phyllite	Summit and shoulder complex	MOC1211
			Middle and foot slope complex	MOC1212
Mountain in post Triassic volcanic rocks	High ridge	RS and CV from andesite	Summit and shoulder complex	MOPTRv111
			Steep/very steep of slope complex	MOPTRv112
			Gentle/moderately of slope complex	MOPTRv113
	Low ridge	RS and CV from andesite	Summit and shoulder complex	MOPTRv211
			Middle and foot slope complex	MOPTRv212
		RS and CV from shale/phyllite	Summit and shoulder complex	MOPTRv221
			Middle and foot slope complex	MOPTRv222
	Spur	RS and CV from andesite	Summit and shoulder complex	MOPTRv311
			Middle and foot slope complex	MOPTRv312
		RS and CV from shale/phyllite	Summit and shoulder complex	MOPTRv321
			Middle and foot slope complex	MOPTRv322
		RS and CV from granite	Summit and shoulder complex	MOPTRv331
			Middle and foot slope complex	MOPTRv332
	Glacies surface	CV from andesite	Erosional surface	MOPTRv411
			Accumulation surface	MOPTRv412
Mountain in Triassic volcanic rocks	High ridge	RS and CV from shale/phyllite	Summit and shoulder complex	MOTRg111
			Middle and foot slope complex	MOTRg112
	Low ridge	RS and CV from granite	Summit and shoulder complex	MOTRg211
			Middle and foot slope complex	MOTRg212
	Spur	RS and CV from granite	Summit and shoulder complex	MOTRg311
			Middle and foot slope complex	MOTRg312
	Hill	RS and CV from granite	Slope facet complex	MOTRg411
Valley	Vale	Alluviocolluvium from andesite	Bottom side complex	Va111
		Alluvium from andesite and shale	Bottom side complex	Va121
		Alluviocolluvium from shale	Bottom side complex	Va131
		Alluviocolluvium from shale and phyllite	Bottom side complex	Va141
		Alluviocolluvium from granite	Bottom side complex	Va151

RS = Residuum
CV = Colluvium

Table 13.2 Parameter maps

Class	Landscape	Relief-type	Lithology	Landform	Organism	Time
1	Mo in C_1	Glacis surface	Alluvio-colluvium from andesite	Accumulation (glacis) surface	Hill ever green forest/Pine forest	Old
2	Mo in PTRv	High ridge	Alluvio-colluvium from granite	Bottom/ side complex	Mixed deciduous forest-high moisture	Young
3	Mo in TRg	hill	Alluvio-colluvium from shale	Erosional (glacis) surface	Mixed deciduous forest/hill ever green forest	Mature
4	Valley	Low ridge	Alluvio-colluvium from shale and phyllite	Gentle/ moderately sloping of middle and foot slope facets complex	Hill ever green forest	
5		Ridge	Alluvium from andesite and shale	Middle and foot slope facets complex		
6		Spur	CV from Andesite	Slope facet complex		
7		Vale	RS and CV from Andesite	Steep/very steep part of middle and foot slope complex		
8			RS and CV from granite	Summit and shoulder (facets) complex		
9			RS and CV from Shale			
10			RS and CV from Shale/Phyllite			
11			RS from Shale			

RS = residuum, CV = colluvium

The sample area B included 7 soil units, classified at subgroup level as: Lithic Halpludults, Typic Hapludults, Typic Paleudults, Rhodic Paleudalfs, Typic Hapludalfs, Typic Paleudalfs and Typic Udifluents (Table 13.2).

The sample area C included 7 soil units, classified at subgroup level as: Lithic Hapludults, Typic Hapludults, Typic Paleudults, Rhodic Paleudalfs, Typic Hapludalfs, Typic Paleudalfs and Typic Udifluents (Table 13.2).

Soil map units in these sample areas were field-checked. There were 10 soil classes determined the soil maps of these sample areas served as the training map (see field work section of Fig. 13.2).

13.2.4 Data for Predictive Soil map

13.2.4.1 Climate

In this study, 15 years of climate data from Hoi Num Rin Royal Station and particle size classes (LDD method described in Treesuwan, 2002 and Kunaporn, 1991) were used to determine the soil moisture and temperature regime according to Soil Taxonomy (USDA, 1975). From the climate data, the study area has a udic moisture regime and hyperthermic soil temperature regime. Therefore, there is no variation in climate throughout the area.

13.2.4.2 Organism Map

Organisms are related to native vegetation, the primary forest of the study area. Vegetation is the best indicator of the environment wherein the soil has been formed. The data of the primary forest report by the LDD and the Hoi Num Rin Royal project were used, and the paper map was digitized. There were four types of forest vegetation (Table 13.2).

13.2.4.3 Relief Maps

Relief represents the Earth's surface and the position where the soils occur, the combination of topographic system and geological structure. Aerial photographs and the orthophoto were used to define the correlation between the different attributes and the slope-related features (gradient, form, curvature, etc.). Landscape can be interpreted by grouping soils with similar morphological and hydrological characteristics. Landform represents the position in a landscape, and comprises the lithological unit, and several other characteristics such as elevation, slope, orientation, stratification, rock exposure, relief unit, slope facets and soil type. There were four landscape units, seven relief-types, and eight landforms represented on the three relief maps (Table 13.2).

13.2.4.4 Lithology (Parent Material) Map

This is the most difficult layer to produce, as geological map is often available in small scale, wherein lithologic formations (quite complex in nature) are indicated.

However, a lithologic map includes detailed information, in terms of rock types, and sediment types (alluvium, colluvium, etc.). In this study, geology map, aerial photographs, orthophoto, slope, aspect and field observation were used to generate a lithological map, with a relatively high degree of uncertainty. The map contains 10 units (Table 13.2).

13.2.4.5 Time Map

The time or soil age (degree of soil development) map represents the age of the soil that is different from the geological time of the parent material. In this study, the development stage of soil (the line of soil evolution, e.g., from entic to cambic to argillic to oxic) was taken into account. We tried to find a correlation between the observed data (in the sample areas) and the environmental characteristics derived from the DEM. Three relative soil ages could be distinguished (Table 13.2): old (Ultisols, Oxisols), mature (Alfisols), and young (Inceptisols and Entisols).

13.2.5 Digital Predictive Soil Mapping

To extrapolate the knowledge, obtained from studying the soils and their associated environment at 69 soil profile sites in, in 32 landform units, to the unsampled areas, we used artificial neural network (ANN) and decision tree (DT) techniques.

13.2.5.1 Deriving a Soil Map from the ANN

Using the standard back-propagation algorithm in ENVI software, a number of three layer sets – an input layer, a hidden layer, and an output layer (Fig. 13.3) – was trained for the purpose of prediction of soils in unsampled areas. In the input layer the number of input nodes was fixed as the number of predictors, which are the six soil forming factor maps (Table 13.2). The number of hidden layers was 1, and the number of output nodes was fixed by the 10 soil classes.

The other network parameters were adjusted after the stage of learning to train the network and to select the parameters, which should give the best fit:

- Training threshold contribution is 0.9
- Training rate is 0.2; training momentum is 0.9
- Training RMS exit criteria (error) is 0.0001
- Number of training iterations is 1,000

The networks were trained until the error calculated over the testing data was judged to have reached a minimum. The best result of training process shows that the selected parameters and iteration of 1,000, 38.93 min time is required to train the network, with the training accuracy of 98% and the training error of 0.003.

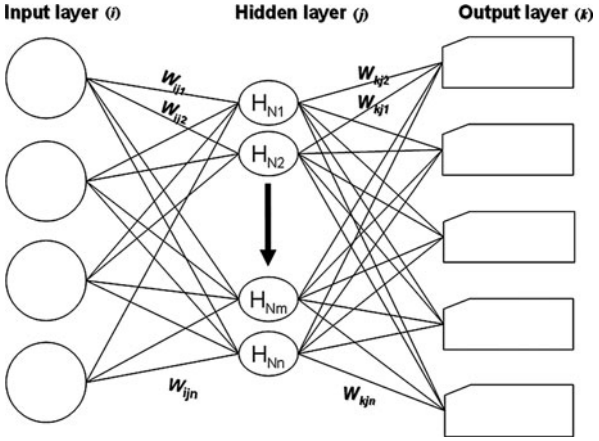


Fig. 13.3 Exemplified topology of feed-forward multi-layer neural network used

13.2.5.2 Deriving a Soil Map from the Decision Tree

To implement the decision process, the same parameters as used in the ANN (Fig. 13.4) were used. In this study, ArcGIS was used to establish a decision tree. There were 10 soil classes in three sample areas on the basis of which the tree was established. The unsampled areas were searched to find areas with the similar environments as the known sites in the sample areas. The decision tree (DT) algorithm was formulated based on the "if, then, else" logic. The ArcGIS raster tool was used for the calculation procedure and then the resulting layers (maps) were combined in the data management tool. The relationships between soils and their environments (based on soil forming factors) were learned in sample areas, thereafter predicted (extrapolated) to unsampled areas.

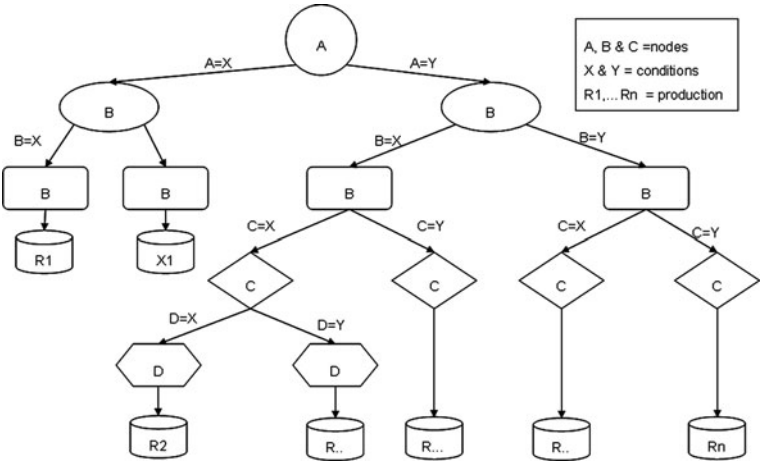


Fig. 13.4 The exemplifier of a simple decision tree structure

13.2.5.3 Map Validation

To validate the maps produced by ANN and DT, 57 auger observations of soil were made in the unsampled areas in the different landform units and landuse types. The data were taken to validate and result maps from the two methods were compared.

13.3 Results and Discussion

The resulting soil maps produced by ANN and DT are shown in Fig. 13.5a and b, respectively. Both ANN and DT could predict and transfer 10 soil classes (classified to subgroup level) from the sampled sites to unsampled areas.

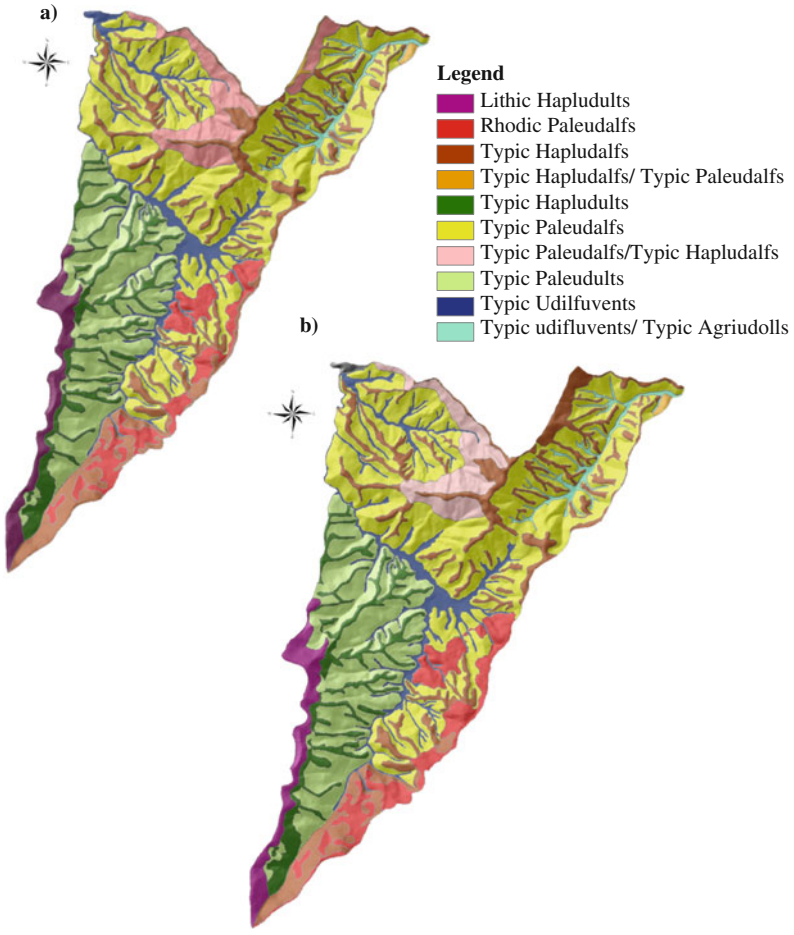


Fig. 13.5 The predictive soil map (**a**) derived from ANN and (**b**) derived from the decision tree technique

Table 13.3 Data validation

	Percent matching level					
Validation	Order	Order[a]	Suborder	Suborder[b]	Great group	Subgroup
Auger hole and ANN-based soilmap	87.50	95.00	94.00	100.00	75.00	90.00
Auger hole and DT-based Soilmap	72.30	89.23	80.00	98.46	67.69	89.23
Decision tree-based and ANN-based soil map	97.00	100.00	97.00	100.00	96.50	95.00
Overall	87.45	95.31	91.75	99.62	81.30	91.56

[a] and [b] after having combined Alfisols & Ultisols, and Inceptisols & Entisols

The results of comparison between the validation soil descriptions versus the ANN and DT based predictive soil maps are presented in Table 13.3. To compare soil classification maps, soil classification names were extracted to four levels (order, suborder, great group and subgroup) and compared by overlay map in each level in three ways: the auger hole descriptions versus the soil map from ANN method, the auger hole descriptions versus the soil map from DT method, and the soil map from ANN with the soil map made by DT method. The ANN-based map is quite satisfactory if no distinction is made between Alfisols and Ultisols (base saturation $\geq$ or $<$ 35%), and between Inceptisols and Entisols (relative degree of incipient development) (Table 13.3). Great group level is based on properties such as subsurface diagnostic horizons, and is thought to reflect the current processes and properties critical to plant growth, such as soil depth and degree of development (Soil Survey Staff, 1975). Subgroup is meant to differentiate several soil morphologic features, which cannot be defined at the great group level, such as redoximorphic features, and the depth of the control section, which represents soil depth related to the root zone (Soil Survey Staff, 2006). The high correlation between ANN and DT techniques likely occurs because similar predictors were used in both cases.

13.4 Conclusions

Soil maps in poorly accessible areas can be produced in a predictive way using the geopedologic approach to soil survey (an ITC approach), which is based on the understanding of soils/landscape relationships. Application of artificial neural network (ANN) and decision tree (DT) techniques considerably facilitated the extrapolation process. The ANN proved to be faster (important in terms of time-saving and cost effectiveness) than applying the decision tree. On the other hand, the ANN is more difficult in the training phase.

Acknowledgments The main author would like to express heartedly thanks to ARDA for the scholarship and to Land Development Department (LDD) for allowing me to study in The Netherlands.

References

Bishop, T.F.A., and Minasny, B., 2006. Digital soil-terrain modeling: the predictive potential and uncertainty, pp. 185–213. In: Grunwald S. (ed.), Environmental Soil-Landscape Modeling: Geographic Information Technologies and Pedometrics. Books in Soils, Plants, and the Environment. CRC/Taylor & Francis, Boca Raton, FL.

Burrough, P.A., MacMillan, R.A., and Van Deursen, W., 1992. Fuzzy classification methods for determining land suitability from soil-profile observation and topography. Soil Science 43(2):193–210.

De Gruijter, J.J., Walvoort, D.J.J., and van Gaans, P.F.M., 1997. Continuous soil maps – a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. Geoderma 77(2–4):169–195.

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

Kunaporn, S., 1991. The assessment of soil temperature regimes classes on hilly and mountainous areas in Northern Thailand (in Thai). Land Development Department, Bangkok.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

McBratney, A.B., and Odeh, I.O.A., 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. Geoderma 77:85–113.

Moonjun, R., 2007. Application of artificial neural network and decision tree in a GIS-based predictive soil mapping for landslide vulnerability study: a case study of Hoi Num Rin sub-watershed, Thailand. Enschede, ITC, The Netherlands, 104 pp. Available at : http://www.itc.nl/library/papers_2007/msc/aes/ruamporn.pdf (full thesis).

Treesuwan, N.M., 2002. Soil moisture regime interpretation in Thailand (based on climatological data and particle size class soil taxonomy 1996) (in Thai). Report No. 505. Soil Survey and Classification Division, Land Development Department, Bangkok.

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

Soil Survey Division, 2004. Soil survey, soil classification and land use planning report, Hoi Num Rin Royal project station, Chiang Rai. Office of Soil Survey and Land Use Planning, Land Development Department. Ministry of Agriculture & Cooperatives.

Soil Survey Staff, 1975. Soil taxonomy: a basic system of soil classification for making and interpreting soil surveys. Agriculture Handbook No. 436. Soil Conservation Service, United States Department of Agriculture, Washington DC, USA.

Soil Survey Staff, 2006. Keys to Soil Taxonomy, 10th ed. USDA-Natural Resources Conservation Service, Washington, DC.

Zhu, A.X., 2000. Mapping soil landscape as spatial continua: the neural network approach. Water Resources Research 36(3):663–677.

Zinck, J.A., 1988/1989. Physiography and soils (Parts 1 and 2). Unpublished Lecture notes, ITC (P.O. Box 6, 7500 AA), Enschede, The Netherlands. Available at http://www.itc.nl/library/papers_1989/tech/zinck_phy.pdf (Last verified 26 July 2009)

Zinck, J.A. and Valenzuela, C.R., 1990. Soil geographic database: structure and application examples. ITC Journal (No):270–294.

Chapter 14
Evaluation of the Transferability of a Knowledge-Based Soil-Landscape Model

J. McKay, S. Grunwald, X. Shi, and R.F. Long

Abstract Knowledge-based digital soil mapping has been used extensively to predict soil taxonomic and physico-chemical soil characteristics. Fuzzy logic knowledge-based models allow explicit integration of knowledge and expertise from soil mappers familiar with a region. Questions remain about the transferability of soil-landscape models developed in one region to other regions. Objectives of this study were to develop and evaluate a knowledge-based model to predict soil series and fuzzy drainage classes and assess its transferability potential between similar soil landscapes in Essex County, Vermont. Two study areas, study area W1, 3.5 km^2 in size and study area W2, 1.9 km^2 in size, were sampled at 128 and 42 sites, respectively. Both study areas are located in Essex County, Vermont. Rule-based fuzzy inference was used based on fuzzy membership functions characterizing soil-environment relationships to create a model derived from expert knowledge. The model was implemented using the Soil Inference Engine (SIE), which provides tools and a user-friendly interface for soil scientists to prepare environmental data, define soil-environment models, run soil inference, and compile final map products. Defuzzified raster predictions were compared to field mapped soil series and fuzzy drainage class properties to assess their accuracy.

In W1 the model was 73.7 and 88.8% accurate, respectively, in predicting soil series and fuzzy drainage classes using an independent validation set. In W2, similar results were achieved, with 71.4 and 89.9% accuracies, respectively. It was shown that the prediction model was transferable to a landscape with similar soil characteristics. For future soil prediction applications it is critical to identify constraints and thresholds that limit transferability of prediction models such as SIE to other soil-landscapes.

Keywords Digital soil mapping · Transferability · Soil inference engine · Vermont · Knowledge-based

J. McKay (✉)
USDA Natural Resources Conservation Service, 481 Summer Street, Suite 202,
St. Johnsbury, VT 05819, USA
e-mail: jessica.mckay@vt.usda.gov

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_14,

14.1 Introduction

For years, soil scientists have been working to build quantitative predictive models to a large extent based on the five factors of soil formation as described by Jenny (1941):

$$S = f(Cl, O, R, P, T) \tag{14.1}$$

where soil is a function of climate, organisms, relief, parent material, and time. McBratney et al. (2003) pointed out that soils can be predicted from their properties in combination with CLORPT factors, where soil properties can be derived from remote or proximal sensing or from expert knowledge.

Knowledge-based models are composed of three main elements: environmental data, a knowledge base, and an inference engine which combines the data and the knowledge base to infer logically valid conclusions about the soil (Skidmore et al., 1996). See Section 26.2 for a discussion on different types of expert knowledge.

The Soil Inference Engine (SIE) is an expert knowledge-based inference engine designed for creating soil maps under fuzzy logic. There are two main types of knowledge that SIE uses: rules, which are defined in parametrical space; and cases, which are defined in geographical space. Both rule-based reasoning (RBR) and case-based reasoning (CBR) can be used to perform inference. SIE also provides tools for result validation, terrain analysis, pre- and post-processing of raster data, and data format conversion (Shi et al., 2009).

Predictive models are often based on the catena concept (Milne, 1935), which indicates that soil profiles occurring on topographically associated landscapes will be repeated on similar landscapes.

One major question that remains in the field of soil landscape modeling is that of model transferability. In particular, knowledge-based empirical models are constrained by soil geographic space boundaries, whereas mechanistic pedogenic simulation models do not face this constraint. Boundary conditions that describe the attribute space used to develop specific soil prediction models are often not well defined in digital soil mapping studies (Grunwald, 2009). Local or site-specific predictions that are implemented within limited geographic extent are documented extensively in the literature (Baxter and Oliver, 2005; McKenzie and Ryan, 1999; Mitra et al., 1998) but their transferability to other regions is unclear.

Predictive capabilities are limited, especially over large areas, because the relationships between soil properties and landscapes are either nonlinear or unknown (Lagacherie and Voltz, 2000). Prediction of soil properties becomes even more difficult when factors other than topography begin to play more of a role, such as different parent materials or changes in climate (Thompson et al., 2006). However, Lagacherie et al. (2001), in studying the applicability of detailed soil surveys to larger areas, found some factors to be especially important, including the expert delineation of reference areas and the necessity of those reference areas of being highly representative of the prediction zone. Keeping these factors in mind, it is

reasonable to expect that expert soil scientists can achieve high model transferability. Section 24.6 contains a good discussion on the inherent bias in expert driven methods.

This study aims to take a soil prediction model developed for a relatively small study area in a complex landscape and test how well it transfers to another, similar study area a few kilometers away. The specific objectives are to populate a knowledge-based soil model to predict soil series (classified according to U.S. Soil Taxonomy (Soil Survey Staff, 2006)) and fuzzy drainage classes in one soil region in Vermont and validate model predictions in a nearby soil region.

14.2 Materials and Methods

14.2.1 Study Area

Two study areas were used to empirically evaluate the transferability of soil scientists' knowledge; specifically, the extent to which a soil-landscape model built by soil scientists for one area is applicable to a second, "similar" area. Both study areas W1 and W2 are in Essex County, Vermont and a comparison of their biophysical characteristics is given in Table 14.1.

Study areas W1 and W2 are assumed to be dominated by one catena of soils, and the model reflects this assumption. Three soil series that dominate these areas are Cabot, Colonel, and Dixfield (Table 14.2). Dixfield soils are found

Table 14.1 Study area comparison

	Study areas	
	W1	W2
U.S. geological survey (USGS) Quad	Averill lake	Bloomfield
Size (km^2)	3.5	1.9
Elevation (m)	Min: 468 Max: 833 Mean: 664 Std. Dev.: 51.9	Min: 375 Max: 618 Mean: 475 Std. Dev.: 49.67
Geology	Phyllite and schist (Gile mountain formation)	Phyllite and schist (Gile mountain formation)
Vegetation	Mixed northern-hardwood and spruce-fir forests	Mixed northern-hardwood and spruce-fir forests
Topography	Hills and narrow valleys	Hills and narrow valleys
Slope (%)	Min: 0.02 Max: 86.08 Mean: 15.42 Std. Dev.: 12.02	Min: 0.10 Max: 54.82 Mean: 12.93 Std. Dev.: 7.38
Mean annual temperature	6°C	6°C
Mean annual precipitation	97 cm	97 cm
Land use	Long term timber management	Long term timber management
Soils (general knowledge)	Deep, loamy basal till; some very poorly drained organic materials in depressions	Deep, loamy basal till; some very poorly drained organic materials in depressions

Table 14.2 Soil series modeled in W1 and W2

Series name	Drainage class	Taxonomic class according to U.S. soil taxonomy
Cabot	Poorly	Loamy, mixed, active, nonacid, frigid, shallow Typic Humaquepts
Colonel	Somewhat poorly	Loamy, isotic, frigid, shallow Aquic Haplorthods
Dixfield	Moderately well	Coarse-loamy, isotic, frigid Aquic Haplorthods

highest on the landscape, on the steepest, most convex slopes, and Cabot soils are found lowest on the landscape, on the flattest, most concave slopes. Colonel soils occur between Cabot and Dixfield both in terms of hillslope position and slope shape. Soils that occur to a lesser extent on the landscape were designated based on which of the three dominant series they most closely resembled morphologically.

14.2.2 Digital and Field Data

Two factors that were proven to provide a good basis for rules based on expert knowledge were slope and compound topographic wetness index. Other layers, such as vegetation, landform, and relative position were investigated and ultimately not used in this study. Investigation of potential data layers involved a visual assessment and comparison to known soil locations by expert soil scientists. Landform and relative position were too general, because the soils being investigated occur on multiple landforms and positions on the landscape. Vegetation showed limited correlation with soil series, and land cover was relatively homogenous, thus was not considered a promising predictor. On the other hand, wetness index is a powerful data layer that gives information not only on flow accumulation, but also indirectly on curvatures and, in addition, serves as a large scale relative position map. Both of the layers used were derived from a 5 m digital elevation model (DEM), aggregated from 1 m Light Detection and Ranging (LiDAR) data. The terrain attributes (slope and wetness index) were created using SIE. The neighborhood size for slope was 30 m, with a square neighborhood shape. The lag in calculating slope for wetness index (known as multi-path wetness in SIE) was 1 pixel.

Field sampling in W1 consisted of 128 points randomly chosen from a previously laid out grid design. Seventy percent (90 points) were used to aid model development and 30% (38 points) were used for model validation within W1.

In order to validate the model in W2, a sampling design similar to random catena sampling (McBratney et al., 2006) was used. Six randomly placed sampling sites along seven catenas were selected, for a total of 42 sampling points. At each of the selected sampling locations in both W1 and W2, a soil pit was exposed and taxonomic soil descriptions were derived to the soil series level. Drainage classes were identified at each site, as represented by depth to redoximorphic features.

14.2.3 Inference Models

For model development, the soil scientists provided global knowledge, which covers the entire mapping area (Shi et al., 2009), prepared data layers and rules, and ran SIE. The initial round of output maps was verified by the soil scientist and the rules were adjusted to fit known sample points. This verification and rule adjustment process was repeated multiple times. After the last round of output maps was verified, post-processing tools from SIE and other geographic information system (GIS) tools (specifically ArcGIS spatial analyst (ESRI, 2010)) were used to integrate the results and generate hardened (defuzzified) maps. The hardened maps are created by aggregating all three of the fuzzy membership maps for each study area using SIE to assign, at each pixel, the soil series with the highest fuzzy membership.

Once the model was fully developed for W1, it was run on the W2 study area and evaluated using an independent validation dataset consisting of 42 sample points.

The rules (Table 14.3) developed for the three soil series are straightforward and represent the understanding of the soils as they occur on the landscape in relation to one another. Figure 14.1 illustrates an example of the inference interface which shows the membership function.

Table 14.3 Rules for Cabot, Colonel, and Dixfield soils

	Full membership at		0.5 membership at		Curve shape		P function	
Series	Slope %	Wetness index	Slope %	Wetness index	Slope	Wetness	Slope	Wetness
Cabot	8	6.3	20	4.8	Z-shaped	S-shaped	Limiting factor	Limiting factor
Colonel	15	3.9	35	2.4, 5.4	Z-shaped	Bell-shaped	Limiting factor	Limiting factor
Dixfield	15	3.4	8	4.9	S-shaped	Z-shaped	Limiting factor	Limiting factor

14.2.4 Evaluation

The results were evaluated in two ways. First a one-to-one comparison of the hardened map and the soil series name at the calibration and validation sample points was done and represented in confusion matrices.

Second, the soil series delineations were further modified using fuzzy drainage class classifications as outlined in Table 14.4. A set of criteria (Table 14.4) was developed which allows illustration of typical and atypical conditions within each soil series based on drainage characteristics. For example, the Dixfield series falls into the “moderately well drained” drainage class, and has a range of characteristics defined that allows all soils that have redoximorphic features between 41 and 101 cm to be grouped in the same category. Some soils that are classified as Dixfield may

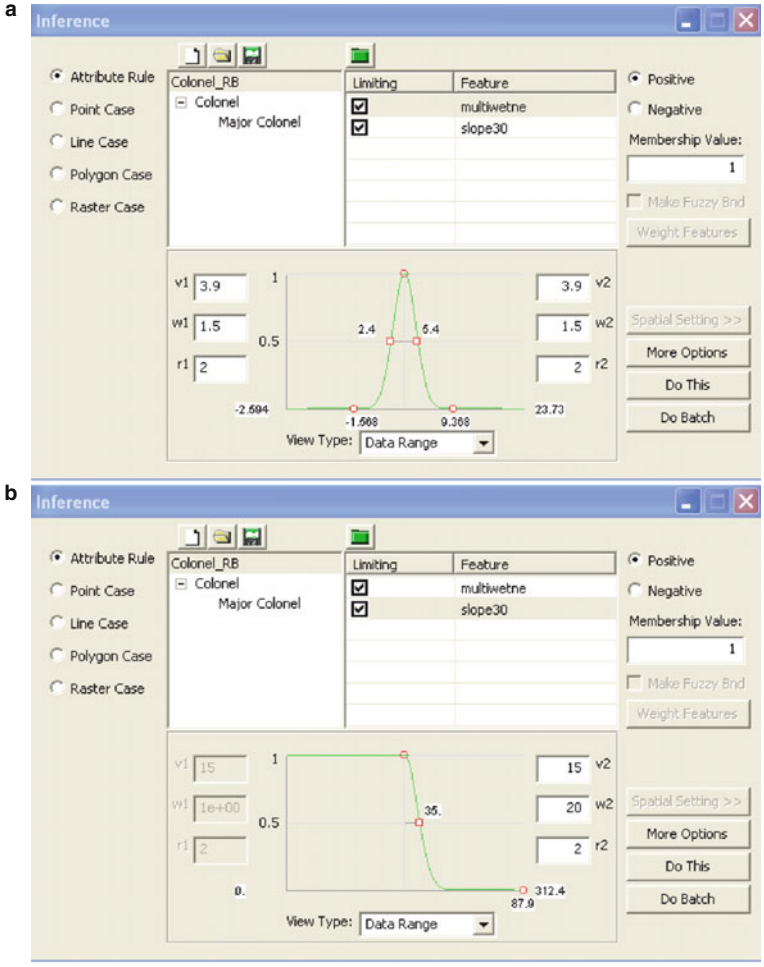


Fig. 14.1 Inference interface for Colonel (ArcSIE). (**a**) Bell-shaped curve for wetness index, (**b**) Z-shaped curve for slope

be considered more typical of Dixfield while some are still Dixfield but are on the dry fringe and others are on the wet fringe.

Every validation point was then assigned a fuzzy value (Table 14.5) based on a comparison of the SIE results and the evaluation of whether the field results were typical for the series' drainage class. Average accuracy numbers were determined based on these fuzzy membership designations.

Drainage class was the key characteristic, or interpretation, used to decide if the highest fuzzy membership was correct when compared to the mapped soil series at the sample points. In this particular soil survey area, parent materials were previously identified, so the difference in these catena members is essentially soil

Table 14.4 Evaluation criteria for fuzzy drainage class

Drainage class (soil series)	Typical characteristics	Wetter fringe characteristics	Drier fringe characteristics
Poorly drained (Cabot)	O horizon 0–15 cm, chroma 2 in profile	O horizon 15–20 cm	Chroma 3 within 76 cm of top of mineral soil; must be chroma 2 somewhere
Somewhat poorly drained (colonel)	Redox between 23 and 36 cm	Redox between 0[a] and 23 cm	Redox between 36 and 41 cm
Moderately well drained (Dixfield)	Redox between 56 and 86 cm	Redox between 41 and 56 cm	Redox between 86 and 102 cm

[a] Includes morphologically similar soils.

Table 14.5 Matrix of fuzzy membership designations comparing SIE results and fuzzy drainage classes

Field results									
SIE Output	Cabot (poorly drained)			Colonel (somewhat poorly drained)			Dixfield (moderately well drained)		
SIE output	Wet fringe	Typical	Dry fringe	Wet fringe	Typical	Dry fringe	Wet fringe	Typical	Dry fringe
Cabot	1	1	1	0.75	0.5	0.25	0	0	0
Colonel	0.25	0.5	0.75	1	1	1	0.75	0.5	0.25
Dixfield	0	0	0	0.25	0.5	0.75	1	1	1

wetness, which is represented by drainage class. There are no other significant interpretive differences that affect use and management of these soil series.

14.3 Results

14.3.1 Results from Model

The initial output maps from SIE show fuzzy results for each soil series (Figs. 14.2 and 14.3). Darker colors represent higher fuzzy memberships for that soil.

The final prediction maps (Figs. 14.4 and 14.5) for each study area are hardened maps of the SIE results, and also serve as a proxy for drainage class maps, because each soil type has a drainage class associated with it.

14.3.2 Accuracy Assessment of Model

The one-to-one comparison of the hardened map to the soil series found in the field yielded 42.6% accuracy for the calibration sites in W1.

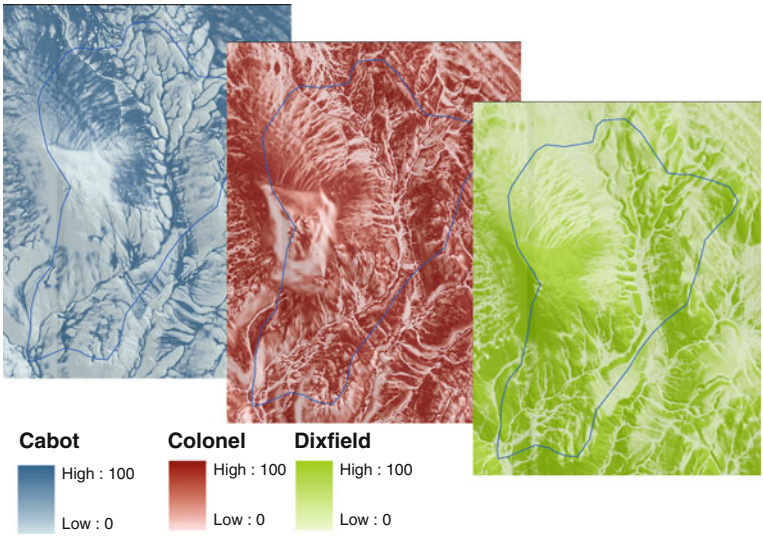


Fig. 14.2 Fuzzy prediction maps for study area W1 (McKay, 2008)

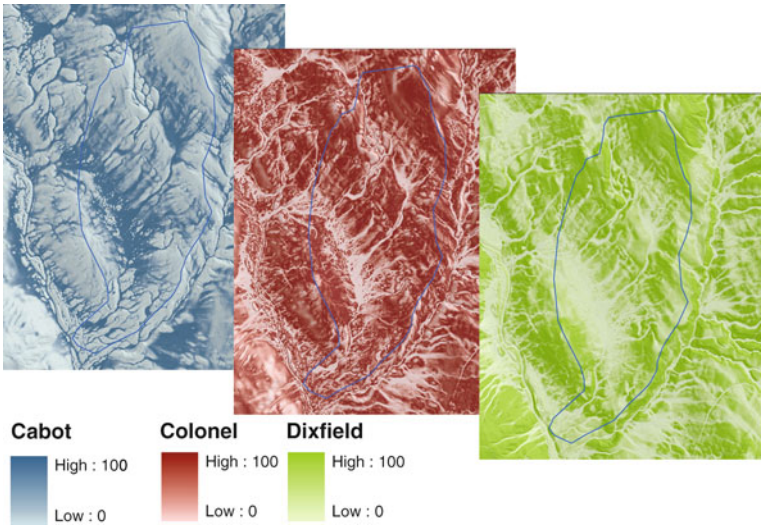


Fig. 14.3 Fuzzy prediction maps for study area W2 (McKay, 2008)

Fig. 14.4 Final prediction map of soil series for study area W1 (McKay, 2008)

The one-to-one comparison of the hardened map to the soil series found in the field yielded 73.7% accuracy overall in W1 (validation sites) and 71.4% accuracy overall in W2. The percent accuracy results by series name are represented in confusion tables (Tables 14.6, 14.7, and 14.8).

Nine iterations of statistics were done using different arrangements of points to represent calibration versus validation points within W1 (Tables 14.9 and 14.10) to accommodate for bias in splitting the dataset.

The fuzzy drainage class results show an overall average between classes of 88.8% accuracy in W1 and 89.9% accuracy in W2 (validation sets). The calibration points were 62.6% accurate overall when comparing fuzzy drainage class prediction results. While the calibration points still had lower accuracy numbers than the validation points, the drainage class results show higher accuracy (Tables 14.11, 14.12, and 14.13) than the one-to-one soil series comparison.

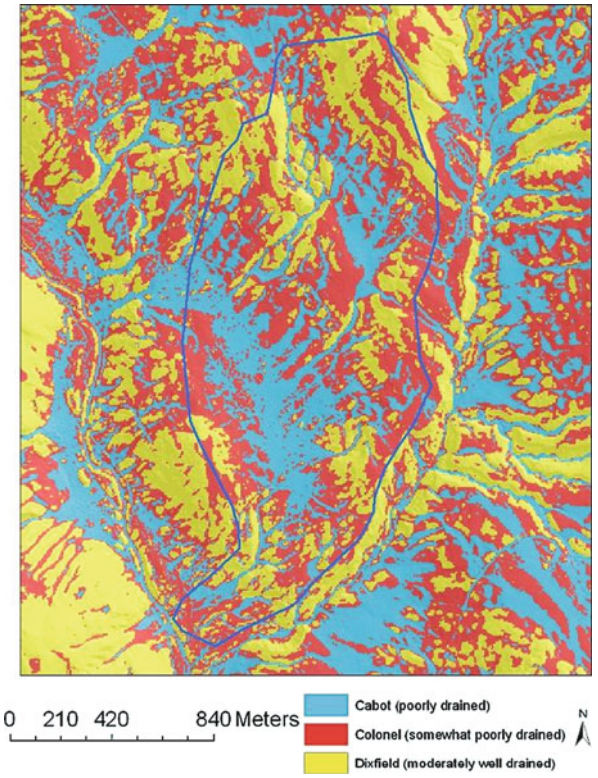


Fig. 14.5 Final prediction map of soil series for study area W2 (McKay, 2008)

Table 14.6 Calibration prediction results based on SIE compared to observed soil series using 90 model development sites in W1

		Observations		
Calibration sites (*n*:90) %		Cabot	Colonel	Dixfield
Predictions	Cabot	*42*	25	33
	Colonel	21	*47*	33
	Dixfield	9	52	*39*

Table 14.7 Validation prediction results based on SIE compared to observed soil series using 38 independent evaluation sites in W1

		Observations		
Validation sites (*n*:38) %		Cabot	Colonel	Dixfield
Predictions	Cabot	*73*	27	0
	Colonel	15	*77*	8
	Dixfield	0	30	*70*

Table 14.8 Validation prediction results based on SIE compared to observed soil series using 42 validation independent evaluation sites in W2

		Observations		
Validation sites (*n*:42) %		Cabot	Colonel	Dixfield
Predictions	Cabot	*69*	31	0
	Colonel	11	*63*	26
	Dixfield	0	10	*90*

Table 14.9 Calibration prediction results based on SIE compared to observed soil series using 90 model development sites in W1 using 9 calibration runs

		Observations		
Calibration sites (*n*:90) %		Cabot	Colonel	Dixfield
Predictions	Cabot	*42–56 (mean: 51)*	18–32 (mean: 26)	18–33 (mean: 23)
	Colonel	18–27 (mean: 22)	*40–58 (mean: 50)*	21–40 (mean: 28)
	Dixfield	0–10 (mean: 7)	36–55 (mean: 47)	*35–55 (mean: 47)*

Table 14.10 Validation prediction results based on SIE compared to observed soil series using 38 independent evaluation sites in W1 using 9 validation runs

		Observations		
Validation sites (*n*:38) %		Cabot	Colonel	Dixfield
Predictions	Cabot	*50–73 (mean: 60)*	9–45 (mean: 24)	0–27 (mean: 16)
	Colonel	6–31 (mean: 19)	*38–77 (mean: 52)*	8–40 (mean: 29)
	Dixfield	0–18 (mean: 5)	30–64 (mean: 43)	*36–70 (mean: 52)*

Table 14.11 Calibration prediction results based on SIE compared to observed drainage classes using 90 model development sites in W1

		Observations		
Calibration sites (*n*:90) %		Poorly drained	Somewhat poorly drained	Moderately well drained
Predictions	Poorly drained	*68*	32	0
	Somewhat poorly drained	17	*54*	29
	Moderately well drained	0	33	*66*

Table 14.12 Validation prediction results based on SIE compared to observed drainage classes using 38 independent evaluation sites in W1

		Observations		
Validation sites (n:38) %		Poorly drained	Somewhat poorly drained	Moderately well drained
Predictions	Poorly drained	*78*	21	0
	Somewhat poorly drained	7	*87*	6
	Moderately well drained	0	15	*85*

Table 14.13 Calibration prediction results based on SIE compared to observed drainage classes using 38 independent evaluation sites in W2

		Observations		
Validation sites (n:42) %		Poorly drained	Somewhat Poorly drained	Moderately well drained
Predictions	Poorly drained	*76*	24	0
	Somewhat poorly drained	6	*73*	21
	Moderately well drained	0	5	*95*

14.4 Discussion

The results from both the direct comparison between the hardened map and field results and the fuzzy drainage class comparison show that the model is highly transferable between the two study areas. The model should therefore transfer well to other areas that are similar to these study areas. As one or more environmental factors change, the transferability of the model can be expected to decline. The hierarchy of these environmental factors may depend on the scale of mapping (see also Chapter 12). The calibration points showed lower accuracy than the validation points (single run), which could be a result of the fact that the calibration set is larger than the validation set in W1; thus capturing more variability in the landscape. This effect was less pronounced when multiple iterations (runs) were used.

Assigning fuzzy drainage class memberships not only improves accuracy numbers; but it also points to the concept of a continuous field model, with soils changing gradually across the landscape rather than having discrete boundaries between one another.

Fuzzy results for each series can be used to depict the uncertainty associated with the hardened map.

There are constraints to this model. Of the five CLORPT factors (climate, organisms, relief, parent material, or time), relief most likely plays the biggest role in vari-

ability in this region, corresponding to the environmental factors that were used to create the model: slope and compound topographic wetness index. This is supported by the conceptual catena model that attributes variation in drainage and wetness to topographic changes.

14.5 Conclusion

In summary, a knowledge-based model such as SIE can be formed and transferred between similar areas effectively, as long as environmental factor constraints are recognized.

References

Baxter, S.J., and Oliver, M.A., 2005. The spatial prediction of soil mineral N and potentially available N using elevation. Geoderma 128:325–339.

ESRI, 2010. ArcGIS Spatial Analyst, Environmental Systems Research Institute – ESRI, Inc., Redlands, CA (http://www.esri.com/software/arcgis/extensions/spatialanalyst/index.html; Last accessed 22 April 2010).

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152(3–4):195–207.

Jenny, H., 1941. Factors of Soil Formation, A System of Quantitative Pedology. McGraw-Hill, New York, NY.

Lagacherie, P., Robbez-Masson, J.M., Nguen-The, N., and Barthès, J.P., 2001. Mapping of reference area representativity using a mathematical soilscape distance. Geoderma 101:105–118.

Lagacherie, P., and Voltz, M., 2000. Predicting soil properties over a region using sample information from a mapped reference area and digital elevation data: a conditional probability approach. Geoderma 97:187–208.

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

McBratney, A., Odgers, N., and Minasney, B., 2006. Random catena sampling for establishing soil-landscape rules for digital soil mapping. Proceedings of 18th World Congress of Soil Science (July 9–15, 2006). IUSS, Philadelphia (USA).

McKay, J., 2008. Using a knowledge-based system to test the transferability of a soil-landscape model in northeastern Vermont. M.S. thesis, University of Florida, Gainesville, FL.

McKenzie, N.J., and Ryan, J.P., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89:67–94.

Milne, G., 1935. Some suggested units of classification and mapping particularly for East African soils. Soil Research 4:183–198.

Mitra, B., Scott, H.D., Dixon, J.C., and McKimmey, J.M., 1998. Applications of fuzzy logic to the prediction of soil erosion in a large watershed. Geoderma 86:183–209.

Shi, X., Long, R., DeKett, R., and Philippe, J., 2009. Integrating different types of knowledge for digital soil mapping. Soil Science Society of America Journal 73:1682–1692.

Skidmore, A.K., Watford, F., Luckananurug, P., and Ryan, P.J., 1996. An operational GIS expert system for mapping forest soil. Photogrammetric Engineering and Remote Sensing 62:501–511.

Soil Survey Staff, 2006. Keys to Soil Taxonomy, 10th ed. USDA-Natural Resources Conservation Service, Washington, DC.

Thompson, J.A., Pena-Yewtukhiq, E.M., and Grove, J.H., 2006. Soil-landscape modeling across a physiographic region: topographic patterns and model transportability. Geoderma 133:57–70.

Chapter 15
Random Forests Applied as a Soil Spatial Predictive Model in Arid Utah

A.K. Stum, J.L. Boettinger, M.A. White, and R.D. Ramsey

Abstract We sought to predict soil classes by applying random forests (RF), a decision tree analysis, to predict 24 soil classes across an arid watershed of western Utah. Environmental covariates were derived from Landsat 7 Enhanced Thematic Mapper Plus (ETM+) and digital elevation models (DEM). Random forests are similar to classification and regression trees (CART). However, RF is doubly random. Many (e.g., 500) weak trees are grown (trained) independently because each tree is trained with a new randomly selected bootstrap sample, and a random subset of variables is used to split each node. To train and validate the RF trees, 561 soil descriptions were made in the field. An additional 111 points were added by case-based reasoning using photo interpretation. As RF makes classification decisions from the mode of many independently grown trees, model uncertainty can be derived. Furthermore, the probability that a pixel belongs to one or more classes in the legend can be determined. The overall out of the bag (OOB) error for discrete classes was 55.2%. The confusion matrix revealed that four soils that frequently co-occurred on landforms were frequently misclassified as each other. These soils were combined into six soil map units. To identify pixels that might belong to one of these newly created combinations of soil classes, minimum threshold probabilities were set. Employing probability by class can be an effective and objective method of determining membership in soil map unit associations and complexes mapped at the 1:24,000 scale.

Keywords Soil components · Soil map units · Digital soil mapping · Digital elevation model · Satellite imagery

15.1 Introduction

The incorporation of topographic and remotely sensed (RS) data into the study of soil systems has increased our ability to predict the spatial distribution of soils across the landscape (e.g., Chapter 2). Often, soils are traditionally represented as a the-

A.K. Stum (✉)
USDA Natural Resources Conservation Service, 340 North 600 East, Richfield, UT 84701, USA
e-mail: alex.stum@ut.usda.gov

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_15,

matic map made up of polygons representing individual map units. Each map unit represents the generalized distribution of soils in the landscape as an association, complex, or a consociation (Moran and Bui, 2002). Soils known to occur within a map unit are referred to as components. Each map unit delineation is a prediction of the soil based on the conceptual model of a soil scientist. Typically rangeland in the western United States is mapped at order three mapping (1:24,000). At this scale of mapping, most map units are associations and complexes (U.S. Department of Agriculture, 2007).

Spatial data analysis seeks to understand patterns or processes that occur in space, allowing predictions where no observations have been made (Bailey and Gatrell, 1995). Based on Jenny's soil-forming factors (1941), McBratney et al. (2003) proposed scorpan, an empirical formulation to quantitatively find relationships between spatially explicit data and soils. There are seven environmental covariates in the scorpan model. The five soil forming factors from Jenny (1941) are present: "c" climate; "o" organisms; "r" topography and landscape attributes; "p" parent material; and "a" time or age. Two scorpan environmental covariates are directed towards spatial predictive models: "s" soil or soil properties, and "n" space, spatial position, or relative position.

We sought to predict the distribution of individual soil components within a rangeland watershed in western Utah, USA. We incorporated topographic and remotely sensed data representing environmental covariates and applied random forests coupled with a geographic informations system (GIS) as a soil spatial predictive model.

15.2 Materials and Methods

15.2.1 Study Area

The study area is The Big Wash watershed in the Basin and Range physiographic province, northwest of Milford, Beaver County, Utah (Fig. 15.1). The area is predominantly fan remnants fed by mountains composed of Pre-Cambrian quartzite, Paleozoic dolomite, and Tertiary extrusive igneous rocks. The soil moisture regime is dominantly aridic bordering on xeric (xeric aridic) and the average precipitation ranges from 20 cm at the basin floor (1,490 m) to 41 cm atop the San Francisco Mountains (2,944 m). The soil temperature regime is mesic with an annual average temperature of 9.3°C at Milford (Soil Survey Staff, 2006).

15.2.2 Digital and Field Data

Field observations that trained the random forests were >90 m apart so no pixel was double-sampled (561 points). Each soil observation was classified to the family level of Soil Taxonomy and many were correlated to existing soil series and assigned to a class (see Table 15.2 and legend in Fig. 15.2) (Soil Survey Staff, 2006). Five

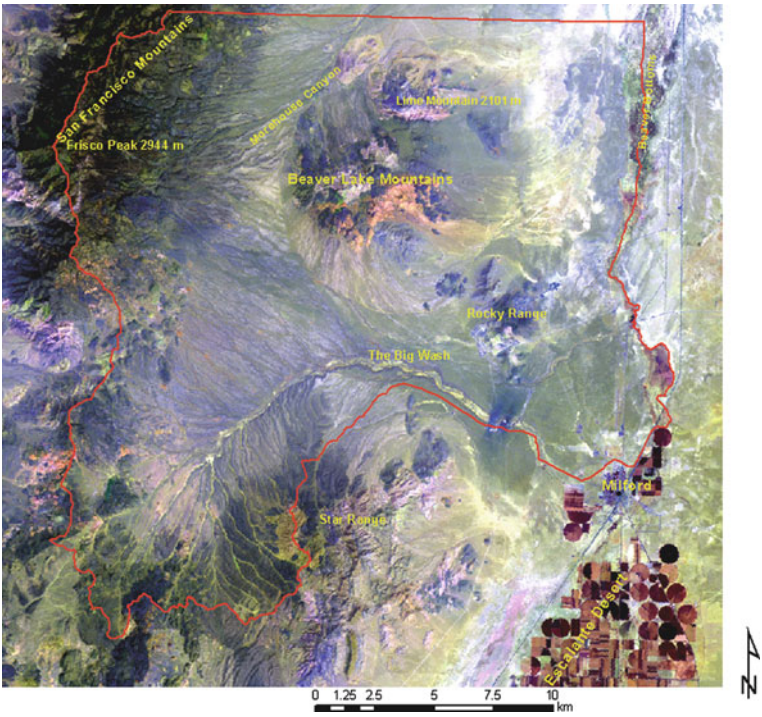


Fig. 15.1 The study area (*red*) encompasses ~47,000 ha in Beaver County, Utah. The Big Wash study area shown in Landsat 7 scene false color (RGB: 5, 7, 1)

classes in our legend were broader classes (associations) as there were insufficient observations to predict them individually. Several soil classes were under-sampled in the field and accurate prediction suffered. Three of these classes were easily indentified in aerial photography, classes 6, 15, and 18. A case-based reasoning approach (Shi et al., 2004) was applied to supplement these classes. This increased the training sample from 561 to 672 points.

The scorpan environmental covariates in the study area were represented by 21 spatially explicit digital data layers (Table 15.1). These covariates were derived from Landsat 7 Enhanced Thematic Mapper Plus (ETM+) and digital elevation models (DEM). Preparation of the Landsat image was done with ERDAS Imagine 9.1^{TM}(Leica, 2006). The digital elevation models were processed in ArcGIS 9.2^{TM}(ESRI, 2007). Digital data were projected into Universal Transverse Mercator (UTM) and North American Datum 1983 (UTM 12S North, datum NAD83).

The study area is covered by one Landsat 7 scene, path 038 and row 033, acquired July 31, 2000 (Fig. 15.1). The Landsat scene was standardized using the cosine theta (COST) method without tau (Chavez, 1996; Nield, 2007; RSGIS: script no. 3, 2003). The normalized difference ratio of bands 4/3 represented vegetation (NDVI), bands 5/2 distinguished most igneous formations (andesite) from sedimentary formations (limestone). In addition, normalized band ratios 4/5, 3/7, 5/1, and 4/7 exhibited

Table 15.1 Environmental covariates represented by digital data (raster)

Environmental covariates	Representative digital data	Source data
Vegetation	NDVI: Bands (4–3)/(4+3)	Landsat 7
Soil moisture regime	Aspect; Elevation; Bands 3 and 4	10-m DEM and Landsat 7
Relief	Slope	10-m DEM
	CTI	
	Filtered CTI (5x5)	
	Aspect ($-\pi$ to π)	
	Filtered slope (11 $\times$ 11) curvature	30-m DEM
Parent material and soil	Standardized Landsat scene Bands 1–7	Landsat 7
	Normalized difference ratios:	Landsat 7
	Bands (4–5)/(4 + 5)	
	Bands (3–7)/(3 + 7)	
	Bands (5–2)/(5 + 2)	
	Bands (5–1)/(5 + 1)	
	Bands (4–7)/(4 + 7)	
	Lake Bonneville Shoreline	10-m DEM

unique patterns wherein distinct landforms and vegetation communities could be identified and were thought to be useful in the model. The 10- and 30-m DEM were used to represent topography and some derivations from the DEM were filtered with a low pass filter (Gesch, 2007; Utah GIS Portal, 2009).

15.2.2.1 Customized Data Layers

Two customized data layers, the Lake Bonneville shoreline and the Xeric-Aridic soil moisture regime (SMR) raster layers, were created to stratify the study area into pedo-geomorphic regions (Soil Survey Staff, 2006).

The shoreline features of ancient Lake Bonneville range in elevation due to isostatic rebound. The elevation trend of the shoreline was estimated by simple linear regression to be 1.99×10^{-4} m rise in elevation per meter in distance northward. From this the approximate extent of ancient Lake Bonneville was determined. Landforms are younger (e.g., Chapter 4) and parent materials are generally finer textured below the Lake Bonneville shoreline.

Within the study area there is strong climosequence relative to aspect and elevation where relative soil moisture increases with elevation and northerly aspects. The break between xeric and aridic soil moisture regime is often characterized by the dominance of *Pinus edulus* (Single-leaf Pinyon) over *Juniperus osteosperma* (Utah Juniper). Spectrally, these two plant communities were distinguishable as two distinct clusters in a feature space plot of Landsat bands 3 and 4. Pixels within these two clusters were selected within the feature space viewer in Imagine to produce a supervised classification of pinyon and juniper plant communities (Leica, 2006). This supervised classification of pinyon and juniper was further refined by establishing an elevation threshold relative to the inverse tangent of aspect (transformed).

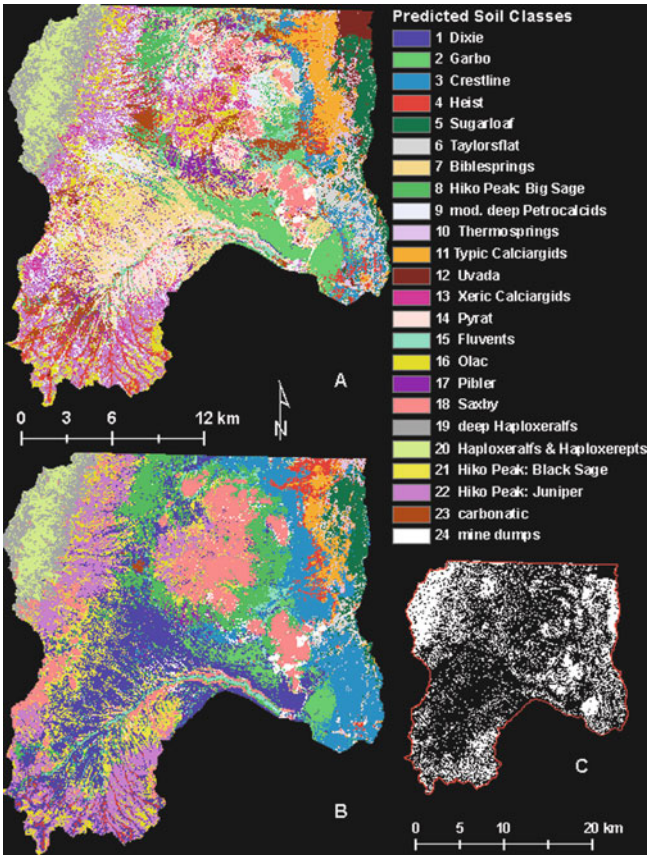


Fig. 15.2 (**a**) The results of grove 2A. (**b**) The results of grove 2B. (**c**) The white pixels indicate where models 2A and 2B were in agreement

15.2.3 Inference Models

Random forests (RF) is similar to the decision trees used in classification and regression trees (CARTs) (Breiman and Cutler, 2009). Random forests is as accurate as, or more accurate than, adaptive boosting, yet computationally lighter (Breiman, 2001; Gislason et al., 2006). Random forests is doubly random. Instead of growing one tree, many (hundreds to thousands) unpruned trees are grown. This ensemble of trees is a grove.

Each tree is trained from a random bootstrap sample, where a random subset of the sample dataset is used to grow (train) the tree and the remaining points are left to validate the tree. At each split a random subset of predictor variables is chosen (e.g., if there were 100 predictor variables, 10 could be selected at random). The strongest variable of this random subset is selected to split the data. Unlike boosting, each tree is grown independent of each other to the maximum depth. Like boosting, the modal

result of the entire grove determines the class membership. By making many weak trees, RF discerns other patterns in the data that are overlooked when few strong trees are grown.

Each sample point was attributed with the values of the independent variables using the spatial analyst toolbox→ extraction toolset→ sampling tool (nearest neighbor) in ArcGIS (ESRI, 2007), piercing through the stack of covariates. A raster mask (30-m resolution) of the study area was converted into a point shapefile. This was done so that each pixel could be attributed with the independent variables and imported into Random Forests where they were thrown down the trees for classification. These points were attributed with the sampling tool mentioned above.

Random Forests software by Salford Systems (2004) was used to grow the grove of trees to make the soil class predictions. Model outputs were validated with "out of the bag" (OOB) testing, which is the total number of misclassified cases divided by the total number of cases or for a particular class. Each tree was trained with an independent bootstrap sample, a random selection of sample points with replacement. Within an individual bootstrap sample some points are drawn one or more times while others may not be drawn. On average, one-third of cases are not selected for an individual bootstrap sample (Breiman, 2001). The points not drawn into a bootstrap sample (left out) are the "out of the bag" (OOB) samples. The OOB points are used to test the tree; OOB samples are thrown down the tree to predict the classification. Several parameters can be manipulated with Random Forests that can have significant impact on the model results, e.g., variable selection, class weighting, number of variables randomly selected at each node, bootstrap sample size, and number of trees grown for the grove. For all iterations we grew 500 trees and the bootstrap sample size was the same as the original sample. We ran several iterations comparing the OOB error with and without weighting and by changing the number of predictive variables selected at each node.

Each tree is grown with an independent bootstrap sample and a new set of OOB samples are thrown down that tree. After all trees are grown (500), each sample point is assigned a final classification based on the mode of predicted classes. The overall OOB error is based on the OOB classification of all the sample cases. For each point, the proportion of "votes" received for each class is reported, which represents the probability of membership in a given class. These ratios are used to estimate uncertainty, determine the extent of individual components, and create new soil map units (associations).

Since the probability that an individual point belongs to a soil class is known, the second and third most probable classifications can be determined. Minimum probability thresholds were established in determining the second and third most probable classifications. Three threshold values were established based on potential limitations to use and management; where the most limiting soils (e.g., Petrocalcids) had the lowest threshold (0.1). We applied these same probability thresholds to determine the possible inclusion of classes in addition to that of the final classification (Table 15.2).

The purpose in determining these alternate classifications is to determine a measure of proximity between pixels. For example if pixel i is predicted to be class α

Table 15.2 Summary of model output from grove 2B

Class	Number of cases	Count	Mean of first	OOB error (%)	Probability threshold	Component count
1	77	98,179	0.296	54.6	0.25	78,403
2	38	14,199	0.413	63.2	0.25	20,279
3	61	59,546	0.359	67.2	0.25	52,779
4	19	12,245	0.325	63.2	0.25	12,889
5	18	15,053	0.477	27.8	0.15	23,671
6	16	7,844	0.311	43.8	0.25	8,762
7[a]	13	156	0.186	100.0	Not considered	Not considered
8	65	57,216	0.326	43.1	0.15	119,136
9[a]	10	24	0.254	100.0	Not considered	Not considered
10[b]	5	2,604	0.289	80.0	0.15	6,626
11	10	11,837	0.362	50.0	0.25	12,203
12[b]	2	71	0.231	100.0	0.15	6,626
13	13	41	0.284	100.0	0.15	3,668
14	18	1,688	0.244	72.2	0.15	4,820
15	41	11,697	0.304	56.1	0.15	32,832
16	26	14,936	0.299	65.4	0.15	40,077
17	20	3,696	0.228	80.0	0.1	14,106
18	62	65,231	0.394	24.2	0.15	94,759
19	12	18,907	0.422	50.0	0.15	38,444
20	11	17,082	0.517	45.5	0.15	33,927
21	53	39,417	0.260	84.9	0.15	95,265
22	29	55,932	0.344	51.7	0.15	81,646
23	8	993	0.247	75.0	0.25	406
24	45	7,137	0.365	8.9	0.5	1,498

Number of cases is the observed occurrences of class. *Count* is the number of pixels in the study area that were predicted to belong to that class. *Mean of first* is the average probability of all pixels predicted to belong to the class. *OOB error* is the class OOB error. *Probability threshold* is the minimum probability. *Component count* is the number of pixels that had a probability greater than or equal to the specified threshold probability for the given class

[a]Classes 7 and 9 were not evaluated as individual components due to poor performance

[b]Classes 10 and 12 were combined as one class before determining the potential extent of individual components

with a probability of 0.60 and pixel y is predicted to be class β with a probability of 0.55 the two pixels are said to be different. But how different? If pixel i also has a 0.31 probability of belonging to class β and pixel y has a 0.37 probability of belonging to class α, pixels i and y are arguably similar.

Similar pixels can be indentified and aggregated (clumped) to represent map unit associations and complexes. This was done with the Knowledge Engineer in Imagine. A simple Boolean argument (rule) can be established where all pixels that have a high probability (≥ 0.25) of belonging to both classes α and β can be determined. Similarly, for each predicted soil class a raster layer of the probable predicted extent was created where all pixels greater than that class's threshold probability were identified.

15.3 Results and Discussion

Weighting had a much greater impact on the OOB error rate and the output of the classification than changing the number of independent variables selected at each node. We selected two groves based on their overall OOB error and OOB error by class. These models will be referred to as groves 2A and 2B. Classes were weighted inversely proportional to sample size in grove 2A and no weighting was applied in grove 2B. Three variables were randomly selected at each node in both 2A and 2B.

Grove 2A had a greater overall OOB error at 64.9% and grove 2B had a 55.2% OOB error. Weighting sacrificed the overall accuracy to improve the classification of minority classes. The outputs were also significantly different (Fig. 15.2a and b). Also, the OOB error rate by class was notably different (Table 15.2). Classes with smaller sample sizes had lower error rates than classes with larger sample sizes in 2A due to weighting; the opposite was true for grove 2B. This difference in class error was reflected in the predicted outputs of each grove. The predicted outputs of the two models were compared and only 37.6% of the pixels were classified the same across the study area (Fig. 15.2c). From a tacit perspective of the authors, grove 2A did not represent the soil distribution across the study area as well as grove 2B. For example, 9.1% of pixels were predicted to be class 7 in grove 2A but only 16 observations were made of this class, whereas only 0.12% of pixels were predicted to be class 1 which was observed 77 times. While weighting may be desirable for imbalanced data sets; it is undesirable for very imbalanced data sets (Chen et al., 2004). The rest of this paper will focus on grove 2B (Fig. 15.2).

Individual components of the landscape can still be easily identified in the model confidence images (Fig. 15.3). Some surfaces with unique signatures may not have been sampled sufficiently or were not represented at all in the sample. This can be

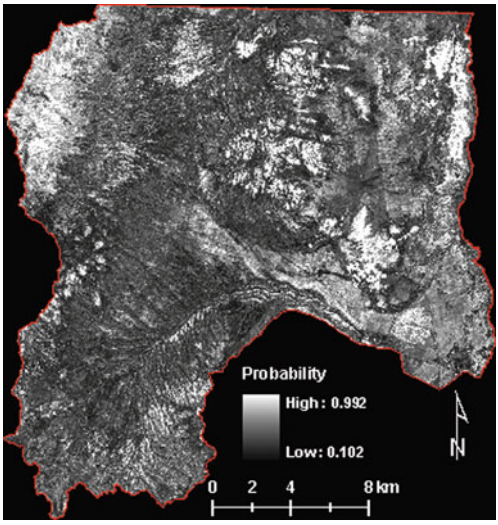


Fig. 15.3 Confidence image showing the probability that a pixel belongs to the class predicted by majority of all trees in the grove. Landform patterns can be discerned in the image, demonstrating that certain surfaces were strongly predicted while other surfaces may need further documentation

used as a guide for future sampling. Every pixel in the study area was predicted by at least one tree to belong to a class other than the final classification.

Many pixels had a low probability of belonging to any class. For example, 7.7% of the pixels had ≤ 0.20 probability of belonging to any class in the legend. An additional 17.3% of the pixels had 0.20–0.25 probability of belonging to any class in the legend. Model uncertainty may be the result of: (1) none of the predicted soil classes in the legend represented the soils in the pixel; (2) the pixel represents a transition soil between several other classes; (3) several soil classes may exist in an individual 30-m pixel; (4) there were insufficient predictive variables to distinguish spectrally and topographically similar soil classes, (5) or there was insufficient observation data to train the model.

Most pixels had a second most probable classification greater than the minimum threshold probability indicated in Table 15.2. Many pixels had a third component identified. Because relatively few pixels had a fourth class higher than the minimum threshold probability, the fourth most likely soil class was not determined.

While grove 2B (no weighting) had the lowest overall OOB error rate, it was still quite large (55.2%). In analyzing the error matrix there are several soil classes which co-occurred on landforms and were frequently misclassified as each other (Table 15.3). Four soil classes which co-occurred on landforms and were spatially extensive (predicted to cover $> 60\%$ of the area) were singled out to create six soil map units. These four soils, Dixie, Garbo, Crestline, and Hiko Peak made up six of the 18 consociations in our legend: classes 1, 2, 3, 8, 21, 22. Areas that were mapped in adjoining areas by the Natural Resources Conservation Service (NRCS) had these four soil series as major components in 19 different soil map units. We adapted six of these map units to create broader classes in our legend (the number of times the two soil components were misclassified as each other are in parentheses): Hiko Peak – Dixie (50); Crestline – Hiko Peak (33); Dixie – Garbo (22); Sevy – Crestline (18) (Sevy is very similar to Dixie); and all the phases of Hiko Peak with different dominant vegetation types were combined (20). Again, this was done by identifying

Table 15.3 Condensed error matrix with the six soil classes that were combined into broader thematic classes

Actual class	Total cases	Percent correct	Predicted as class						
			1	2	3	8	21	22	*All others*
1	77	45.5	**35**	6	2	11	5	6	12
2	38	36.8	16	**14**	1	4	1	1	1
3	61	32.8	16	2	**20**	15	3	2	3
8	65	56.9	10	0	8	**37**	1	0	9
21	53	15.1	14	0	5	12	**8**	3	11
22	29	48.3	4	0	0	0	4	**14**	7
All others	349	49.6	36	3	8	35	12	18	**173**
Total	672		131	25	44	114	34	44	324
Percent false positive			73.3	44.0	54.5	67.5	76.5	68.2	46.6

The other 18 classes are represented as "All Others". The numbers on the diagonal represent the times the observations were predicted correctly

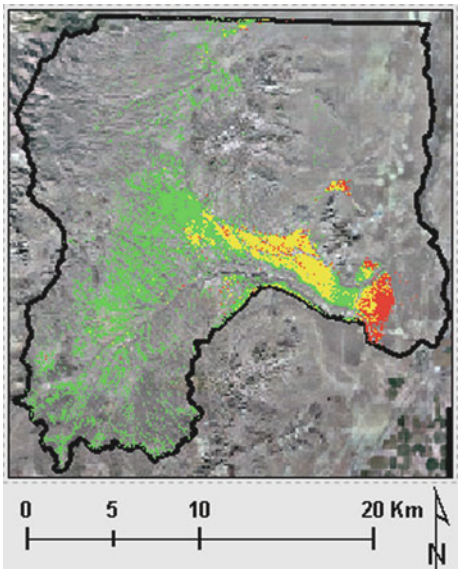


Fig. 15.4 Pixels that have a probability ≥ 0.25 of being Dixie (*green*); a probability ≥ 0.25 of being Garbo (*red*); and probabilities of ≥ 0.25 of being Dixie and ≥ 0.25 of being Garbo (*yellow*)

individual pixels that had a high probability of belonging to both components within the six map units. Pixels where two of these soil series overlapped were identified as the newly combined soil associations and complexes. The example of Dixie, Garbo, and Dixie-Garbo is illustrated in Fig. 15.4.

15.4 Conclusions

Random forests is a useful model to predict soil components as individual soil classes (see also Chapter 29). As the classification is the result of many weak trees, model uncertainty and alternate probable classifications of pixels can be determined. By defining alternate classes for individual pixels, the probable extent of minority classes expanded without sacrificing the prediction of larger classes. This provided an effective and objective method of determining membership in soil map unit associations and complexes mapped at the 1:24,000 scale.

References

Bailey, T.C., and Gatrell, A.C., 1995. Interactive Spatial Data Analysis. Prentice Hall, Harlow England.

Breiman, L., 2001. Random forests. Machine Learning 45:5–32.

Breiman, L., and Cutler, A., 2009. Random forests homepage. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (last verified 14 April 2010).

Chavez, P.S., Jr., 1996. Image-based atmospheric corrections – revisited and improved. Photogrammetric Engineering & Remote Sensing 62:1025–1036.

Chen, C., Liaw, A., and Breiman, L., 2004. Using random forests to learn unbalanced data. Technical Report 666, Statistics Department, University of California at Berkeley. [Online] Available: http://www.stat.berkeley.edu/tech-reports/666.pdf (last verified 14 April 2010).

ESRI GIS and Mapping Software, 2007. ArcGIS 9.2. ESRI, Redlands, CA.

Gislason, P.O., Benediktsson, J.A., and Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognition Letters 27:294–300.

Gesch, D.B., 2007. Chapter 4 – The national elevation dataset, pp. 99–119. In: Maune, D.F. (ed.), Digital Elevation Model Technologies and Applications: The DEM Users Manual, 2nd ed., American Society for Photogrammetry and Remote Sensing, Bethesda, MD.

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

Leica Geosystems, 2006. ERDAS Imagine V. 9.1. Leica Geosystems, Atlanta, GA

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

Moran, C.J., and Bui, E.N., 2002. Spatial data mining for enhancing soil map modeling. International Journal of Geographical Information Science 16(6):533–549.

Nield, S.J., Boettinger, J.L., and Ramsey, R.D., 2007. Digitally mapping gypsic and natric soil areas using Landsat ETM data. Soil Science Society of America Journal 71:245–252.

RSGIS, 2003. Remote Sensing and Geographic Information Systems Laboratory, Utah State University, http://earth.gis.usu.edu/imagestd/ (last verified 13 August 2009).

Salford Systems, 2004. Random Forests V. 1.0. Salford Systems, San Diego, CA.

Shi, X., Zhu, X-A., Burt, J.E., Qi, F., and Simonson, D., 2004. A case-based reasoning approach to fuzzy mapping. Soil Science Society of America Journal 68:885–894.

Soil Survey Staff, 2006. Keys to Soil Taxonomy, 10th ed., U.S. Department of Agriculture, Natural Resources Conservation Service, Washington, DC, pp. 26–28. [Online] Available: http://www.soils.usda.gov/technical/classification/tax_keys (last verified 14 April 2010).

U.S. Department of Agriculture, Natural Resources Conservation Service, 2007. National Soil Survey Handbook, title 430-VI sec. 627.03.d. [Online] Available: http://www.soils.usda.gov/technical/handbook/contents/part627.html#03 (last verified 14 April 2010).

Utah GIS Portal, 2009. National Elevation Dataset available for Utah, http://gis.utah.gov/elevation-terrain-data/national-elevation-dataset-ned (last verified 13 August 2009).

Chapter 16
Two Methods for Using Legacy Data in Digital Soil Mapping

T. Mayr, M. Rivas-Casado, P. Bellamy, R. Palmer, J. Zawadzka, and R. Corstanje

Abstract Legacy data are useful sources of information on the spatial variation of soil properties. There are, however, problems using legacy data, and in this paper we explore some of these problems. A common issue is often the uneven sample distribution over geographical and predictor space and the problems this generates for the subsequent modelling efforts. Furthermore legacy soil data often has a mixture qualitative and quantitative data. The current need is for quantitative data, whereas the available datasets are often qualitative; e.g. auger bores. In this paper we compare two methods: (i) a Generalized Linear modelling (GZLM) approach which uses scarce, measured soil property data and (ii) Bayesian Belief networks (BBN) which uses extensive but generic values of the soil property, linked to soil classes. We used digital soil mapping covariates such as small scale soil maps, geology, digital terrain model, climate data and landscape position in order to predict continuous surfaces for sand, silt, clay, bulk density and organic carbon. The objective is to present a qualitative comparison between the two methods, as a direct comparison was not possible due to the number and distribution of the legacy data. We found that the GZLM approach was significantly impacted by an uneven sampling of the predictor space. This study suggests that a more generalist approach such as BBN is better in the absence of few hard data but in the presence of many soft data.

Keywords Generalized linear models · Bayesian belief networks · Auger bores · Legacy data · Soil properties

16.1 Introduction

There is an increasing demand for spatial information on soil properties by environmental and agronomic scientists to better understand the effect of a changing environment, and also form an important input into the decision making by policy-makers. Funding to obtain new soil surveys is increasingly scarce and use must be

T. Mayr (✉)
National Soil Resources Institute, Cranfield University, Bedfordshire, MK43 0AL, UK
e-mail: t.mayr@cranfield.ac.uk

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_16,

made of existing data (legacy data). There are numerous problems that arise when using legacy data. For example, conclusions are obtained from data whose original sampling designs may not be appropriate for these post hoc analyses. Other practical issues that arise when legacy data are used are: the mixture of quantity, quality and types of data (ordinal, continuous and categorical); mapping problems such as non contiguous coverages, differences in taxonomic or other classification schemes and divergences in legends; scale and support issues (Chapter 5), where data are obtained and represented at different scales. These problems are known in digital soil mapping (DSM), see for example Bui and Moran (2001, 2003) who discuss and supply a number of possible remedies by which to obtain spatial information from existing data; also McBratney et al. (2003) who provide an overview of the possible methods to use legacy data.

In England and Wales, an important source of soil information is the 1:250,000 National Soil Map (NATMAP), in which the spatial distribution of 297 soil associations is represented. Each soil association is identified by a dominant soil subgroup and contains between one and ten individual soil series (soil types). The soil classification is based on predominant pedogenic characteristics, observed at a density of 2–3 per km^2. This information is catalogued as a set of descriptive features obtained by surveyors from auger bores. The information is spatially exhaustive with over 150,000 observations across the area and it is therefore an essential source of information on the spatial variability of soil and can be linked to specific soil properties via the soil series.

Much of the current demand is for quantitative rather than categorical information on the spatial variation of soil properties. This is partially model driven as soil information is used as inputs into environmental, climatic or agronomic models. This type of data is more expensive to obtain as laboratory analysis is required and therefore national data coverage tends to be patchy and scarce. In the National Soil Map, for instance, there are 6,000 observations that contain quantitative information like particle size distribution, bulk density and organic carbon.

There are disadvantages to using either form of data. Uncertainty is introduced when categorical data is converted to quantitative forms using generic soil properties associated to soil series. It is not certain, for instance, that a given value is the best estimate for that soil property in that series. On the hand, the sparse coverage of quantitative observations could introduce considerable spatial uncertainty.

In this paper, we report on an exploratory study in which we predict soil properties required for a related project in which they will be introduced to a biomass model (Yield-SAFE; van der Werf et al., 2007) from both forms of information. We obtain these predictions by modelling their relationship with a set of covariates for which we have exhaustive information, such as small scale soil maps, geology, a digital terrain model, climate data and landscape position. The basis of this approach is the "soil-landscape" paradigm, in which the spatial variation of soils can be expressed as a function of the underlying rocks, landforms, and vegetation patterns (Hudson, 1992).

The objective was therefore to explore and compare different methods for generating continuous primary soil property surfaces based on data typically that are

available at a national level. In order to predict the input variables required by the crop model, continuous surfaces for sand, silt, clay, bulk density, organic carbon and soil depth are required.

16.2 Material and Methods

Two approaches to digital soil mapping were taken: (i) using measured soil property data and (ii) using generic values of the soil property linked to soil classes. A 10 km × 10 km test area was selected which contained the highest number of measured soil property data available, and thus providing the best possible conditions for modelling the spatial distribution of primary soil property data using laboratory measurements. This was also one of the few areas for which a detailed soil map exists (1:25,000), and the map was used as a benchmark for comparing the output from the two approaches.

16.2.1 Study Area

The 10 km × 10 km test was identified. Situated in the county of Warwickshire, it contains the 1:25,000 Stratford-upon-Avon (SP25) map sheet which contains the largest number of representative profiles (39 profiles). This detailed soil map and associated data cover 100 km^2 in the Avon valley, a tributary of the river Severn in the SW Midlands of England. The whole area is low lying, generally between 40 and 100 m above Ordnance Datum and underlain by soft argillaceous sediments, mainly Jurassic and Triassic mudstones and clay shales. Locally some soils are formed in superficial deposits; alluvium and gravel river terraces of the Avon; and extensive spreads of glacial till. The majority of soils are slowly permeable and clay rich, developed either in solid rocks (Whimple and Worcester series in Triassic mudstones and Evesham and Denchworth in Jurassic Lias clays) or glacial till (Salop and Flint series). Permeable, well drained soils are restricted to the dissected patches of gravel river terraces (Wick series) or narrow outcrops of Lias Limestone (Haselor series) that form narrow ridges in the clay vales.

16.2.2 Digital and Field Data

16.2.2.1 Covariates

A wide range of variables were collated for the test area, including small scale soil maps, geology, digital terrain model, climate data and landscape position. The digital terrain model was processed using the LandMapR software (LandMapper Environmental Solutions Inc.) taking care to provide realistic critical upslope area for stream initiation based on existing river network coverage. In total, 54 covariates were assembled. Both the training data as well as the model implementation was undertaken by attaching all covariates to a point file centred on the 20 m grid of the digital terrain model (DTM).

16.2.2.2 Training Data

Quantitative data were obtained from representative profiles, which are sites where soil pits had been dug to characterise the soil (39 profiles). For these profiles, the descriptions are more detailed than the auger bores and laboratory analysis of soil samples on sand, silt and clay contents, pH, $CaCO_3$, organic carbon or loss on ignition, and in some cases (4 profiles) nitrogen, CEC, exchangeable cations (K, Na, Ca, Mg) and extractable Fe and Al. Undisturbed cores (222 cm^3) were collected in triplicate from each soil horizon in the soil pit and laboratory analysis of these cores, using sand suction baths and pressure membrane apparatus, provided data on bulk and packing density, total pore space, and retained (at 5 kPa suction) and available water (between 5 and 1,500 kPa) capacities. For the test area, 39 sites had analytical information on sand, silt, clay, organic carbon as well as soil depth information from the profile descriptions. Bulk density data were available for 15 sites.

Auger bore descriptions were recorded in the field by the soil surveyors. The information was recorded on RUFF cards and later transferred into the Land Information System (LandIS). The records hold information on the location and profile descriptions including the soil series identified by the surveyor. The standard terms used to describe soil profile characteristics are defined in the Soil Survey Field Handbook (Hodgson, 1997). These terms include the colour, particle-size class, stoniness, soil structure, consistence, porosity and roots in each horizon. For the test area, no auger bore descriptions were available electronically and the approach relied on auger bores outside the study area but within the same physiographic domains. In total, 14,936 observations were available.

16.2.3 Inference Approaches

Spatial estimates of sand, silt, clay, bulk density, organic carbon and soil depth were obtained using two inference approaches: one based on quantitative data using generalized linear modelling, the *direct* approach, and a second based on the qualitative soil series identification within the auger bore descriptions using Bayesian belief networks, the *indirect* approach. In each method, the predictors were the set of 54 covariates described earlier.

16.2.3.1 Generalized Linear Models, the *Direct* Approach

A generalized linear model was created to obtain predictions of soil properties at non measured locations. The advantage of using a GZLM is that it does not require normally distributed dependant variables. The model was developed using the covariates as independent variables and the soil properties at dependent variables. Model dimensionality was reduced to 26 as there was a degree of colinearity between some of the predictor variables. Those variables with more than two significant ($p < 0.05$) pair-wise correlations were not included in the analysis. The model used was:

$$y = \sum_{i=1}^{n} \beta_i x_i + \varepsilon,$$

where x_i is the ith covariate used to predict the soil property, y and ε is the error. The analysis was carried out using functions in the R software (R Development Core Team, 2005). The GZLM was fitted using an inverse Gamma link function and Gamma error distribution and this gave the best fit to the data. Predicted values were obtained for silt, clay and organic carbon on a 50 m grid for the test area. The sand predictions were made by subtracting the sum of the clay and silt predictions from 100% so that these variables were consistent. Sand was chosen to be estimated this way as it appeared to be the most difficult to predict using a GZLM.

16.2.3.2 Belief Networks, the *Indirect* Approach

We developed a set of Bayesian Belief Networks to predict soil series using all 54 predictor variables in NETICA^{TM} (NORSYS Software Corp). From these soil series, in combination with associated modal soil property values, we generated a set of continuous soil property surfaces for sand, silt, clay and organic carbon. In this approach, soil properties were calculated using weighted probability values, an approach proposed by Zhu et al. (2006), which couples the raster representation in the spatial domain with the probability representation in the attribute domain. For this study, the five highest probabilities were selected in order to remove contributions from entities with very small probability values.

NETICA uses Bayesian Belief Networks and influence diagrams, and provides a methodology for representing relationships between positions or variables, even if the relationships involve uncertainty, unpredictability or imprecision. The algorithms used are described in Spiegelhalter et al. (1993) and Neapolitan (1990). We used forward stepwise selection in this study to reduce the dimensionality of the covariates.

$$V_{ij} = \frac{\sum_{k=1}^{n} S_{ij}^{k} V^{k}}{\sum_{k=1}^{n} S_{ij}^{k}} \tag{16.1}$$

Where *Vij* is the predicted soil property value at location (i, j), $S_{i,j}^{k}$ 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 (Zhu et al., 2006).

16.3 Results

This digital soil mapping study relied entirely on legacy data. Due to the number and distribution of the legacy data, a direct comparison of both methods was not possible, so we present a feasibility study of the two approaches. With legacy data,

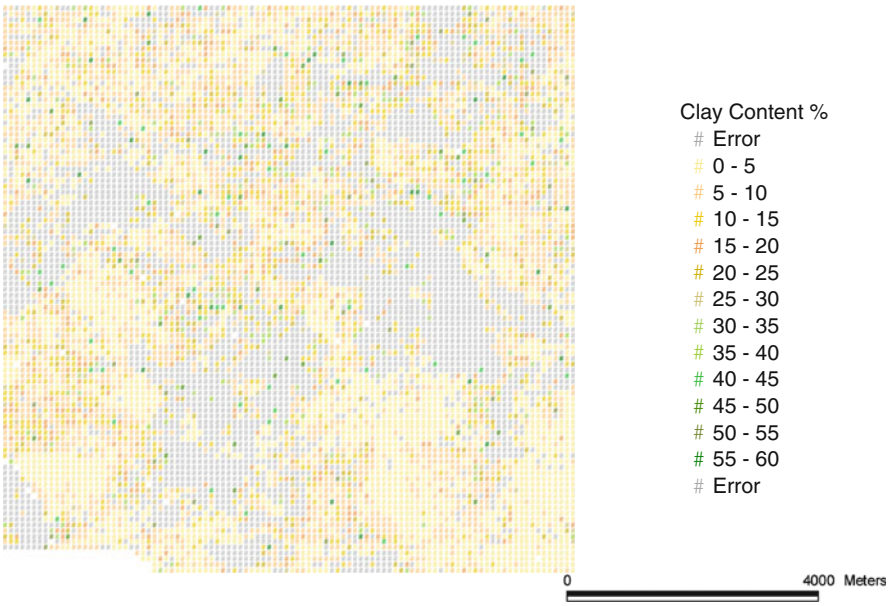


Fig. 16.1 Clay content (%) predictions using GZLM

the quality of the inference depends largely on the amount and spatial distribution of the raw data as it determines the feature space in which the model is valid. In the subsequent sections, we present and discuss the results for all predictions (sand, silt, clay and carbon) but supply figures of the predictions of clay and carbon only as examples of the predictive output (Figs. 16.1, 16.2, 16.3, 16.4, 16.5, and 16.6). The interest in the related study was for information over the soil profile; inputs are therefore weighted averages over the soil profiles.

16.3.1 Generalized Linear Models, the Direct *Approach*

The GZLM was fitted to the 39 observations using 24 variables for each of the soil properties silt, clay and carbon – of those variables, 10 made a significant ($p < 0.05$) contribution to the predictions for silt, 16 for clay and 3 for carbon. Sand could not be predicted using a GZLM as none of the variables made a significant contribution to the model. In Table 16.1 the goodness of fit of the model is presented as well as the prediction error for clay, silt and carbon. The GZLM modelling performs better for silt and clay than for soil organic carbon.

The results for the predictions of clay and carbon for a 50 m grid are presented in Figs. 16.1 and 16.2. Within the test area, there were values of the predictor variables that exceeded the range observed in the dataset of predictors used to formulate the model. In these cases we have reported this as no predictions (white) or where

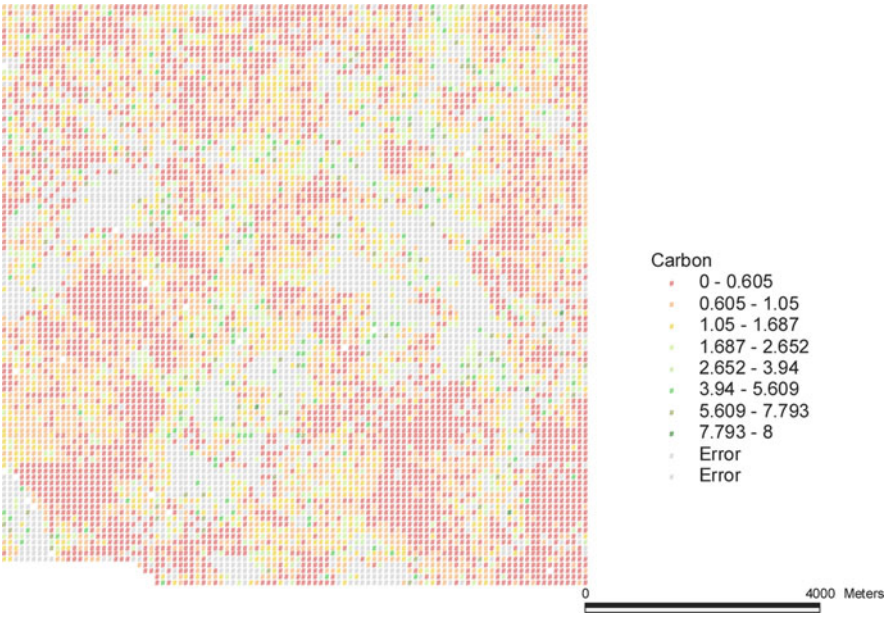


Fig. 16.2 Organic carbon content (%) predictions using GZLM

predicted values of clay were less than 0% or exceeded 60% (grey). For soil organic carbon, values identified as grey were those outside of the observed range in the test dataset (0–8%). These predictions were not possible as combinations of the variables were not present in the dataset on which the model was developed.

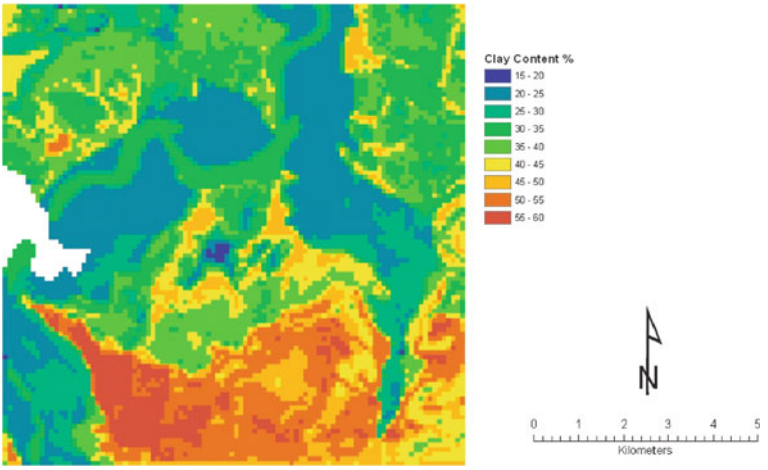


Fig. 16.3 Clay content (%) predictions using Bayesian belief networks

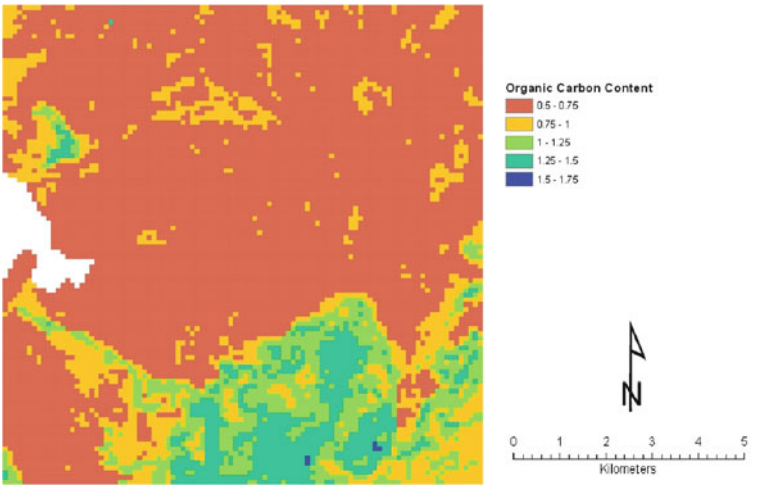


Fig. 16.4 Organic carbon content (%) predictions using Bayesian belief networks

16.3.2 Belief Networks, the Indirect *Approach*

The maximum membership model produced reasonably good performance measures compared to the two statistical models. The two statistical models (terrain-based and membership based) used all field points in the model development, which means that the error measures (or performance measures) are those of model development and not those of model validation. The maximum membership model only used one field sample per soil series (the sample with the maximum membership in that series) for model development. In addition, the performance measures for the maximum membership model are those of model validation (that is, only the field

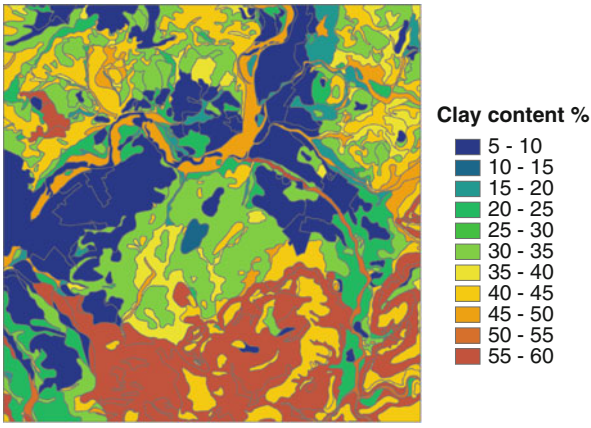


Fig. 16.5 Nominal clay content (%) for every map symbol from the detailed (1:25,000) soil maps

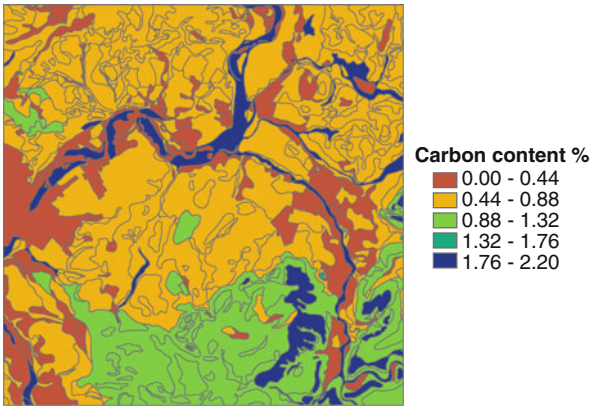


Fig. 16.6 Nominal organic carbon content (%) for every map symbol from the detailed (1:25,000) soil maps

Table 16.1 Summary of the goodness of fit of the GLZM and associated prediction error. In this case clay, silt and carbon were predicted. Note that the error of the predicted values was a factor of ten larger for the carbon model than for the silt and clay models

Variable	ME	MSE	RMSE
Silt	0.0006	0.0002	0.01
Clay	0.0009	0.0002	0.01
Soil organic carbon	0.0004	0.03	0.17

Mean error = predicted − observed
MSE = (predicted − observed)$^2/N$ = average ((predicted − observed)^2)
RMSE = sqroot (MSE)

samples not used in the model development, were used to compute the performance measures). In this sense the maximum membership model may have out performed both regression models (Zhu et al., 2006).

16.4 Discussion

A limitation of any model development is that the training dataset must contain and represent the full range of variation in the landscape (feature space) that exists in the area over which predictions are needed (see Section 6.1). This limitation was evident for the GZLM approach, which is a function of the sparse data on which this model was built. The problem was not presented with the Bayesian network as it was based on considerably more data. It illustrates a consistent problem with some of the more parametric techniques such as GZLM in that they are generally less capable of dealing with missing or sparse data and may therefore not be the most appropriate techniques for this type of exercise. We could have considered alternative approaches such as cokriging (Webster and Oliver, 2007) or more complicated mixed modelling approaches that consider the spatial autocorrelation structure present in the data. These approaches require, however, a larger number of data points (in excess of 100 according to the same paper) than was available in this area.

In Figs. 16.5 and 16.6 we present the detailed soil maps for the test area in which we have carbon and clay estimates for each polygon. These are our best estimates of the spatial variation in carbon and clay content. It shows that there is very little agreement between observed and predicted values obtained from either method. The GZLM approach, where predictions are available, provided a poor reproduction of the spatial patterns. The Bayesian Belief networks were more successful in reproducing some of the spatial features (Figs. 16.3 and 16.4). However, this method tended to underestimate larger clay content and over estimate carbon content. It is not surprising that the predictions from Bayesian Belief networks conform better to a landscape described by the detailed soil maps as this information is included in the Bayesian inference process in the form of a coarser set of soil associations. That is not the case for GZLM where the output does not adhere to soil boundaries.

Interpretation of validation statistics is equally fraught as there is no independent validation data. A comparison of Tables 16.1 and 16.2 suggests that the GZLM approach is much more precise than the Bayesian Belief networks. This finding is a function of the fact that they measure two different things: In the case of the GZLM, the validation is obtained at the point observations at which the model is formulated. The error reported in this table represents model error, or the uncertainty introduced by the model itself relative to the point observations. It is not a measure of the spatial accuracy of model performance. In the case of Bayesian Belief networks, these predict a soil class, to which a generic soil property value is associated. The error reported in Table 16.2 is therefore a composite error of model formulation, misclassification and the uncertainty introduced by using a generic value which may not be an accurate estimate of the mean value for that series.

There is also an associated divergence of support between predicted and observed and that introduces errors in the validation process. The independent, observed primary properties represent point observations whereas those predicted by the Bayesian model represent a mean value for a soil series. Validation is obtained by comparing these to the generic values associated to the soil series. The current validation approach is therefore misleading as it is unlikely that these generic values are robust independent estimates of the class mean. This divergence in support can be addressed using conditional simulations to obtain estimates at a coarser scale

Table 16.2 Summary of the goodness of fit of the Bayesian Belief Network and associated prediction error. In this case sand, silt, clay, bulk density and carbon were predicted

Variable	ME	MSE	RMSE
Sand	2.67	594.50	24.38
Silt	2.57	86.57	9.30
Clay	−2.89	185.50	13.61
Bulk density	0.074	0.038	0.20
Carbon	0.73	1.04	1.02

Mean error = predicted − observed
MSE = (predicted − observed)$^2/N$ = average ((predicted − observed)^2)
RMSE = sqroot (MSE)

(e.g. incorporating a block kriging step). The disadvantage of this approach, however, is that this assumes stationarity across the area, which may not be the case. We expect discontinuities between soil series and the variation in the primary soil properties will not be similar across all soil series. Stratification by landscape attributes coupled to sequential indicator simulation, may resolve units that can effectively be viewed as stationary and the primary property values can then be estimated within these. Although computationally complex, the approach could be considered in future endeavours.

It was beyond the scope of this study to compare a comprehensive suit of inference approaches. We used a Generalized Linear Model approach which does not consider any spatial correlation between observations. The model is not so widely used in digital soil mapping since it is not spatial and comparison with other techniques like regression-kriging would be of interest for future studies. For the indirect approach, comparison with alternatives such as extrapolation methods developed by Lagacherie et al. (2001) may be of interest.

There are also practical issues associated with legacy data that we encountered in this project. For instance, the geographic position of auger bores is generally recorded only to the nearest 100 m as no Global Positioning System was available at that time and surveyors generally paced their distances from reference points and were sampled to support a soil survey programme. A digital soil mapping programme requires a different sampling strategy which would be driven more by sampling the feature space than finding the extent or characterising individual soil polygons (see Section 6.3). Auger bores are also often sampled to represent the entire spectrum of a soil series rather than to obtain a central or mean description of that series. This can introduce considerable confusion when there are transition zones. We also encountered well known issues associated to using a DTM such as artefacts observed in the 2nd derivatives and divergencies in scale between that of the DTM and the scale at which the predictions were made. Finally, the absence of a particular type of glacial geology (drift) as another covariate meant that that we lacked an important driver of soil formation (Equation (14.1) in Chapter 14).

16.5 Conclusions

This paper compared two methods and illustrated some of the pitfalls associated when using legacy data for digital soil mapping. We found that parametric methods can be limiting when the input data is scarce and their sampling was not designed to cover the existing feature space. This study suggests that more generalist approaches such as Bayesian Belief networks might be more appropriate but due to the limitations of this study we were not able to do a rigorous comparison. This study does illustrate, however, that with legacy data, careful consideration has to be given to the merits of different inference approaches before selecting any given one.

Acknowledgments We wish acknowledge the European Joint Research Centre (JRC) for providing the funding for this project (Contract Number 382468 F1SC).

References

Bui, E.N., and Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103:79–94.

Bui, E.N., and Moran, C.J., 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.

Hodgson, J.M., 1997. Soil Survey Field Handbook, 3rd ed. Technical Monograph No. 5. Soil Survey and Land Research Centre, Silsoe.

Hudson, B., 1992. The soil survey as a paradigm-based science. Soil Science Society of America Journal 56:836–841.

Lagacherie, P., Robbez-Masson, J.M., Nguyen-The, N., and Barthes, J.P., 2001. Mapping of reference area representativity using a mathematical soilscape distance. Geoderma 101(3):105–118.

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

Neapolitan, R.E., 1990. Probability Reasoning in Expert Systems: Theory and Algorithms. John Wiley & Sons, New York.

van der Werf, W., Keesman, K., Burgess, P.J., Graves, A.R., Pilbeam, D, Incoll, L.D, Metselaar, K., Mayus, M., Stappers, R., van Keulen, H., Palma, J., and Dupraz, C., 2007. Yield-SAFE: a parameter-sparse process-based dynamic model for predicting resource capture, growth and production in agroforestry systems. Ecological Engineering 29:419–433.

R Development Core Team 2005. 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 (Last accessed 21 April 2010).

Spiegelhalter, D., Dawid, A.P., Lauritzen, S.L., and Cowell, R.G., 1993. Bayesian analysis in expert systems. Statistical Science 8(3):219–283.

Webster, R., and Oliver, M.A., 2007. Geostatistics for Environmental Scientists, 2nd edin. Wiley and Sons Ltd., New York, NY.

Zhu, A.X., Moore, A., and Burt, J.E., 2006. Prediction of soil Properties Using Fuzzy Membership. 2nd Global Workshop on Digital Soil Mapping, Rio de Janeiro, Brazil 4 to 7 July 2006.

Part III
Environmental Application and Assessment

Chapter 17
Mapping Heavy Metal Content in Soils with Multi-Kernel SVR and LiDAR Derived Data

Cristiano Ballabio and Roberto Comolli

Abstract Support vector regression (SVR) is a powerful machine learning technique in the framework of the statistical learning theory; while Kriging is a well-established prediction method traditionally used in the spatial statistics field. However, the two techniques share the same background of reproducing kernel Hilbert space (RKHS).

SVR has recently shown promising performance in different spatial mapping tasks. In the present work, the problem of spatial data mapping is addressed using a multi-scale SVR (MS-SVR) approach. This can be considered as a multi-resolution analysis of the observed process. The multi-scale SVR approach is particularly attractive for its capability to deal, at the same time, with the nonlinear regression of the dependent variable on auxiliary variables and with the spatial interpolation. This capability makes the MS-SVR an optimal choice for automatic mapping system.

In the present work MS-SVR was applied to soil heavy metal content mapping, in a study area site in the Italian Alps. The area complex landscape, modelled by both glacial and karsts phenomena, along with an heterogeneous nature of the parent material, makes the mapping of heavy metal content a difficult task to approach with linear regression or mixed geostatistical techniques.

The result obtained outlines the Multi-scale SVR as a powerful technique for general inference and automatic mapping, with the only constraint of the requirement of a multi-parameter optimization.

Keywords Support vector regression · Multi-scale modelling · LiDAR · Heavy metals

C. Ballabio (✉)
Environmental and Land Sciences Department (DISAT) and Geology Department,
University of Milano-Bicocca, 20126, Milano, Italy
e-mail: cristiano.ballabio@unimib.it

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_17,

17.1 Introduction

Geostatistics is widely used to predict soil properties. This is usually approached using Ordinary Kriging (OK), or Regression Kriging (RK) if suitable spatial covariates are available.

Although kriging estimation is known to be a Best Linear Unbiased Estimator (BLUE), the quality of the OK interpolation is usually strongly dependant on the analysis and the modeling of the spatial correlation variable; moreover, kriging is subject to several requirements: intrinsic stationarity and a number of observations large enough to estimate the variogram. If any of the assumptions of OK does not hold, the outcome of OK might be suboptimal (Cressie, 1993). RK tries to overcome some of these shortcomings by detrending the data with linear regression, but this leads to further assumptions about the relation between the studied property and the spatial covariates.

Despite these limitations, OK and RK are usually the best choice for spatial interpolation and both produce high quality results in term of accuracy and precision. However, when the dataset is sparse, and the relation with spatial covariates is not linear, RK is usually not feasible.

Dealing with noisy and sparse datasets is common in some fields of machine learning (i.e., image recognition), and several techniques were developed to gather information from this kind of data. One of the most promising approaches is the Support Vector Regression (SVR). SVR is a nonparametric technique, based on Structural Risk Minimization (SRM), which has shown promising performances in regression based digital soil mapping (Ballabio, 2009). SRM approaches the problem of model optimization, minimizing both the error and the model complexity. This makes it possible to achieve a good balance between bias and error.

In this work we propose an approach to soil properties mapping based on Multi-Scale SVR (MS-SVR), which can overcome most of the limitations of the SVR in mapping tasks, by using multiple basis expansion.

17.2 Materials and Methods

17.2.1 Study Area

The case considers an area of about 2 km^2 located in the Italian Central Alps, along the San Giacomo Valley, by the side of the artificial lake created by the Isola dam on the Liro Torrent (Fig. 17.1). The valley follows an almost N-S-striking tectonic lineament, parallel to the main geomorphologic feature, the Andossi high-plain. The Andossi area is a high-plain 2,000 m a.s.l., whose substrate is composed by a two different metamorphic units: a marble-dolomite unit, spanning the entire study area, and an altered schist formation intercalating the marble unit in the southern part of the area. In the same area a “cornieule” (vacuolar dolomitic rocks) alternates with the marble.

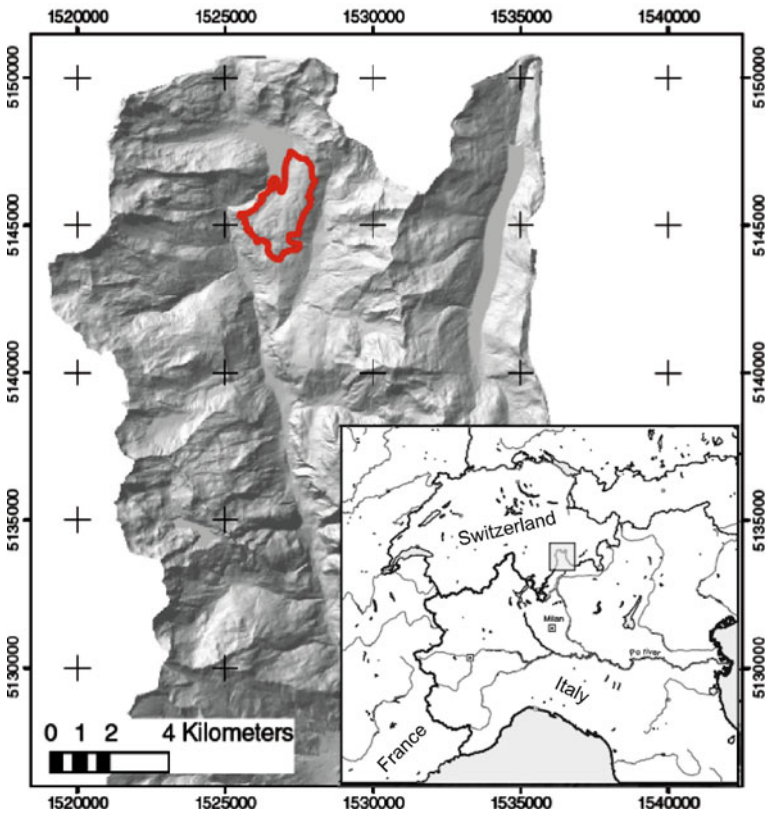


Fig. 17.1 Overlaid on the hill-shaded relief map of the upper Valchiavenna, the area delimited by the *dark grey outline* corresponds to the study area. The figure in the *bottom-right corner* of the larger picture depicts the position of the Valchiavenna (*light grey square*), across the Italian-Swiss border

The area has a complex geomorphology due to a polycyclic evolution. It is possible to identify two different kinds of geomorphologic processes that shaped the landscape of the area: a karst phase, acting during a tropical interglacial period (Messinian), and multiple glacial phases. The outcome of these processes is clearly evidenced by the presence of large complex dolines in the southern part of the area and by an extensive till cover in the northern-western part of the study area.

The area is a meadowland, used as a pasture during the summer season since historical times. Most of the area was deforested approximately 3,000 years ago, as a forest and shrub cover was documented before this period.

Cambisols and Podzols and a wide range of intergrades (WRB) occur in the area. Leptosols are most common on the steepest slopes. Cambisols and Podzols present in the area have a sandy to sandy-loam texture, a high content of organic carbon (OC) (up to $300\,\mathrm{g\,kg^{-1}}$ in the O horizons) and a low degree of stoniness. Podzol occurrence is likely to be related to the Holocene Climate Optimum and the presence of coniferous cover in the area.

17.2.2 Digital Data

A high resolution digital elevation model (DEM) of the area was obtained from a Light Detection and Ranging LiDAR survey. LiDAR is a technique able to capture high density points and high accuracy digital elevation data. LiDAR uses a pulse laser to determine distance between the sensor and the target. For terrain mapping purposes, an airborne LiDAR system is typically composed of the laser-sensor system, an inertial navigation system, and a differential GPS; a multiband optical sensor may also be present to gather high resolution ground images. LiDAR data allow the derivation of high accuracy and high resolution DEMs, with a minimum vertical Root Mean Square Error (RMSE) close to 15 cm in optimal conditions (Hodgson and Bresnahan, 2004).

Digital Elevation Models (DEM) have become very popular in the last two decades to describe the topographic surface and model relationships between different components of the landscape. DEMs can be produced in a number of different ways, most of them by direct field measurements of elevations at specific locations or by photogrammetry. LiDAR surveys are the current "state of the art" technique for DEM reconstruction.

LiDAR data were obtained from an airborne survey which was performed at the beginning of November 2004. The data were collected using an Optech 3100 ALTM with a nominal frequency of 10^5 laser pulses s^{-1} with an elevation RMSE of less than 45 cm. The on-field pixel resolution of the DEM is sub-metrical, but for practical purposes the obtained DEM was resampled to a 2-m pixel scale which corresponds to an appropriate scale for the terrain features present in the area. The effect of DEM scale in mapping is discussed in Chapter 5.

Along with LiDAR data, a set of high resolution images were collected using an Optech 4K02 ALTM camera. These images are a pseudo-RBG color composite, with a spectra shifted between near-infrared and green and a ground resolution of 25 cm. Despite the few bands available, the spatial resolution of these images makes it possible to use pattern-based techniques for segmentation and classification (Lucieer and Stein, 2005). The segmentation was performed using an IDL language script within the ENVI software environment. Land cover was classified into 8 classes using a supervised procedure similar to the one described in Section 10.2.4, the resulting map was included in the analysis along with DEM-derived maps. The resulting map of vegetation cover classes provides useful additional information for the mapping procedure (as also shown in Fig. 10.5).

The DEM-derived descriptors were:

- altitude,
- slope (Zevenbergen and Thorne, 1987) and its 1st and 2nd directional derivatives,
- Compound Terrain Index (CTI) (McKenzie and Ryan, 1999),
- Seasonal solar radiation (Dubayah and Rich, 1995),
- RUSLE "LS" factor (Van Remortel et al., 2001),
- Sediment Transport Index (STI) (De Roo, 1997).

These topographic descriptors were derived by using the routines implemented in the SAGA GIS software.

17.2.3 Field Data

Within the study area 140 soil samples were collected with a stratified random sampling scheme. Heavy metal content was measured by atomic absorption spectroscopy (AAS) after samples dissolution with aqua regia. The spatial distribution of each element was studied, but Cu proved to have the most spatially structured distribution and for this reason was chosen for this investigation. The Cu content in these soils is likely to be related exclusively to the nature of the parent material, as well as to the pedogenetic processes active in the area. The range of concentrations of Cu and several other elements (ppm) is shown in Table 17.1.

Table 17.1 Concentration range of some elements in the sampled soils

	Cu (ppm)	Cr (ppm)	Ni (ppm)	Zn (ppm)	Mn (ppm)
Min	3.28	2.07	N.d.	N.d.	30.69
Mean	36.27	16.35	28.26	35.40	324.67
Max	95.62	46.10	72.44	424.92	4, 667.67

N.d. = not detected.

17.2.4 Inference Models

Support Vector Regression (SVR) approximates a non-linear system with input $\mathbf{x} \in \mathrm{R}^p$ and output $\mathbf{y} \in \mathrm{R}$, as $f(x) = \langle \mathbf{w}, \mathbf{\Phi}(\mathbf{x}) \rangle + b$; where $\mathbf{\Phi}(\mathbf{x})$ is a mapping function from data space into feature space specified by a kernel function $k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{\Phi}(x), \mathbf{\Phi}(x') \rangle$.

The kernel function is based on the dot product of feature vectors, which in the linear forms can be expressed as $\langle \mathbf{x}, \mathbf{x}' \rangle = X^T X' = \sum_{i=1}^{n} x_i x_i'$. The kernel function is analogous to a covariance function, and the aim of the kernel projection is to transform non-linear data from a R^p space into linear data in a R^n space.

The SVR training produces a solution as $f(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n k(\mathbf{x}_n, \mathbf{x}) + b$ where the coefficients $\boldsymbol{\alpha}$ equal to 0 are the Support Vectors (SV) all of which define the training model.

Usually the Structural Risk Minimization (SRM) is achieved by and minimizing: $R = \frac{1}{2}||w||^2 + C \sum_{i=1}^{l} |y - (w, x_i)|_\varepsilon$ where $||w||^2$ is the Euclidean norm, and the second term in the equation bounds the complexity of the model (Drucker et al., 1997).

Under specific assumptions SVR and ordinary kriging are mathematically equivalent (Wensen and Yanguang, 2006).

The critical point in the SVR is the choice of the kernel function, which is user defined and whose choice influences the fitting of the model. The most common non-linear kernel function is the Gaussian Radial Basis (GRB): $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\sigma ||\mathbf{x}_i - \mathbf{x}_j||^2)$, which provides a good degree of generalization.

However, all single kernels suffer a shared problem: the model cannot globally fit the system if regions with different data trends are present.

This issue is especially felt when dealing with spatial data, since the typical scale dependence of the feature space derived from the covariates is usually different from the one deriving from the geographical space distribution.

To overcome this issue a Multi-Kernel SVR was proposed which utilizes a set of kernels sharing the same function, but spanning the feature space at different resolutions, thus the definition of "Multi-Scale SVR". This approach yields a more complex solution in which more than one kernel is used for the projection: $f(\mathbf{x}) = \sum_{m=1}^{M} \sum_{n=1}^{N} \alpha_n k(\mathbf{x}_n, \mathbf{x}) + b$.

The main drawback of the approach is further complication of the training (finding the $M \times N$ coefficients). This approach is easily implemented as each linear positive definite combination of kernels, is itself a regular kernel. On the other side, it is not known beforehand, which combination of kernels is optimal. To solve this problem is it possible to use a sequential training on the residuals; a single kernel is used to achieve a coarse approximation of the data, then other kernels are used to, sequentially, approximate the residuals of the fitted model.

This approach and can be implemented using a semiparametric Support Vector Machine (SVM) (Smola et al., 1998). The semiparametric SVM approach decomposes the function into two parts: $f(x) = \langle \mathbf{w}, \mathbf{\Phi}(\mathbf{x}) \rangle + \sum_{i=1}^{n} \beta_i \phi_i(x_i)$, where the second term expresses the previous knowledge about the model. In this way it is possible to update the model with the information previously collected by other kernels.

In this work the MS-SVR was performed using the Lib-SVM (Chang and Lin, 2001) and the R package kernlab (Karatzoglou et al., 2004), while we developed our own testing and kernel combination algorithms.

17.3 Results

17.3.1 Results from Model

The MS-SVR outperformed both single kernel SVR and regression-kriging (see also Chapter 22) for most of the mapped elements. Figure 17.2 compares the prediction of soil's Cu content obtained from different models with the actual data; in this case the MS-SVR appears to produce far less dispersed values than the other models.

This is not surprising since MS-SVR implements a non-parametric regression approach for both the covariates and the spatial component of the data, while the linear model limits the performance of RK and the single kernel of SVR is clearly suboptimal for this kind of environment.

While it is still possible to use non-parametric regression in the RK procedure, this can lead to unpredictable results, because the stationarity is usually not assured. On the other hand MS-SVR eschews this issue, not requiring any particular data distribution.

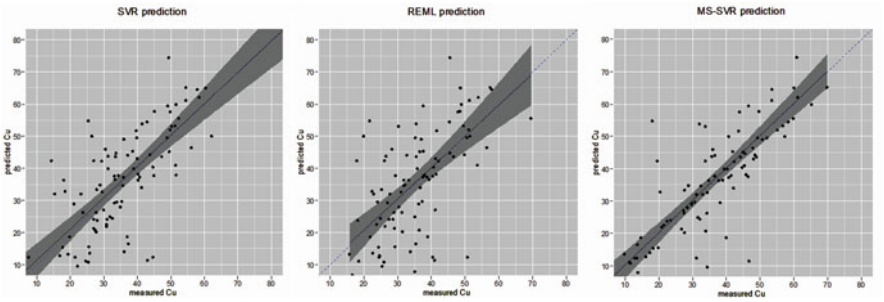


Fig. 17.2 Predicted/observed plot for single kernel SVR, RK and MS-SVR. The *dark grey* band represents the 0.95 confidence interval for the regression line (*solid line*)

There is a little difference between SVR and RK performance in the prediction of the Cu content is soils, but RK performs slightly better than single kernel SVR (Fig. 17.2).

Anyway, the result of single kernel SVR, shows an interesting property: regardless of the relative complexity of the prediction, the outcome is based only on a limited number of support vectors (between 8 and 28). Despite the simplicity of the model, the performance of single kernel SVR is very similar to that of the RK. This strongly suggests a relatively simple underlying physical model, although not a linear one.

As shown in Fig. 17.3, the residuals of MS-SVR are less dispersed than those of RK or SVR. In particular SVR tends to produce a more "depressed" shape of the distribution around zero; this feature is due to the choice of the loss function of the SVR model, which weights the data pairs falling outside the region defined by SV in model optimization.

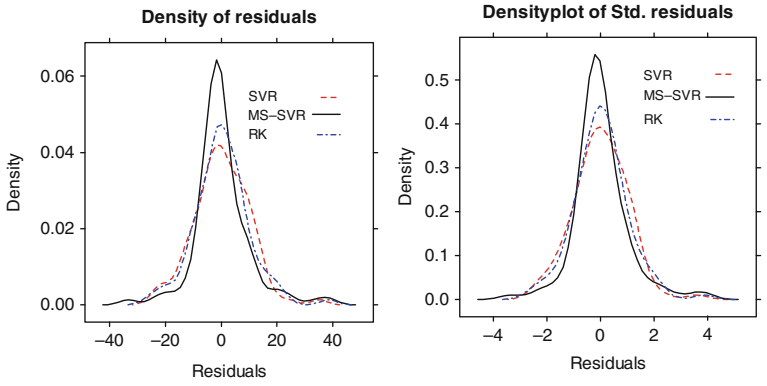


Fig. 17.3 Comparison of density of studentized residuals for SVR, RK and MS-SVR

Although MS-SVR residuals are less dispersed than those from RK, the model seems to perform slightly worse on outliers (Fig. 17.3), as few residuals with extreme values (± 40) are present; this is hardly a model issue, as these extreme values appear as outliers in each model. MS-SVR seems to favour a low residual mean, sacrificing some extreme value.

17.3.2 Model Validation

To assess model accuracy, a bootstrapped "leave one out" (LOO) procedure was used. The LOO procedure recursively trains the model on all, but one, observations. Then the performance of the model is tested by comparing the missing observation and its predicted value. The procedure is then repeated for all the available observations. Although usually "overoptimistic" for real model validation, due the high ratio between the number of observations used to train the model and the single observation used for testing it, LOO was chosen because of the small size of the dataset. Moreover, the aim of the evaluation was to compare different models, so the overoptimistic outcome of the LOO is likely to affect each model in equal measure.

A bootstrap procedure was used for the LOO error estimation to improve the reliability of error measurement, thus the procedure was repeated 100 times for each of the observations. This stabilizes the error, since the MS-SVR hyper-parameters optimization procedure can sometime lead to slightly different results.

As shown in Table 17.2, MS-SVR is usually a better predictor of Cu content. Although, in this instance, the values of R^2 and RMSE shown are somewhat deceiving; as MS-SVR performs worse than the other models on extreme values.

More information could be gathered from the distribution of studentized residuals: As shown in Fig. 17.3 and Table 17.2, the distribution of residuals from MS-SVR is less dispersed than that of the other models, but this feature comes at the expense of some skewness of the residuals. This is probably due to the re-tuning of the weights $\beta_{m-1} = \{\beta_o, \ldots, \beta_{m-1}\}$ of the parametric part.

Table 17.3 depicts the performance of the model when validated in a more realistic way. In this case the data set was randomly subdivided into training and testing sets with a ratio of two to three (3-fold cross validation). There is a general decrease in the performance of the models with 3-fold cross validations compared to the LOO procedure. However, in this situation the model is fitted on a suboptimal sample, thus limiting the prediction capabilities of any statistical model.

Table 17.2 Summary of statistics obtained from model comparison with the LOO procedure

	Prediction		Residuals			
	R^2	RMSE	Mean	Std	Kurtosis	Skewness
SVR	0.47	12.30	0.059	14.36	1.51	0.29
RK	0.45	14.50	0.027	14.63	1.69	0.36
MS-SVR	0.66	10.50	0.024	10.55	3.90	0.74

Table 17.3 Summary of statistics obtained from model comparison performing a cross-validation on 1/3 of the samples (3-fold cross validation)

	Prediction	
	R^2	RMSE
SVR	0.30	19.07
RK	0.29	18.08
MS-SVR	0.50	11.77

17.4 Discussion

Given its landscape complexity, the area was clearly undersampled. Therefore, the suboptimal performance of RK is probably due to the small number of observations available. On the other hand, SVR is known to perform very well on sparse datasets (Cherkassky and Mullier, 2007; Meyer et al., 2003; van der Walt and Barnard, 2006). So the superior performance of MS-SVR could be due to the particular structure of the analyzed data. Nevertheless, on simulated random fields, this performance difference is still present, albeit to a lesser degree.

The MS-SVR has been widely utilized for image segmentation, signal decomposition and gene analysis. Several algorithms were developed in the last year to simplify the optimization procedure (Bi et al., 2004; Sonnenburg et al., 2006).

The main drawback of the multi-scale approach resides in computational complexity. Each kernel increases the number of variables to optimize, as well as the number of hyper-parameters to be tuned. Anyway, the use of two or three kernels usually leads to very good results, thus limiting model complexity. For example, the MS-SVR outcome shown in Fig. 17.4 is obtained using a combination of two Gaussian kernels.

There are several optimization strategies available to try to lessen these shortcomings and the optimization of SVR is a very active field of research in the machine learning community.

Figure 17.4 shows the distribution of Cu content in the soil as predicted by the best performing MS-SVR model. The studentised residuals shown as a percent of the total value are plotted as bars over the map in correspondence of the sampling sites.

The prediction obtained from the model for the spatial distribution of the Cu content seems to reflect what is known about Cu chemistry in soils and the lithology of the area (Fig. 17.4).

In the northwest part of the study area, where an extensive cover of glacial till is present, the concentration of Cu is minimal; this is probably due to the relatively recent deposition of this material, as well as its chemical composition.

On the southern slopes, the presence of Cu seems to be related to the presence of the schist substrate, as well as the redeposition of fine material due to soil erosion. This is more evident in the karst area, which approximately corresponds to the southern half of the study area.

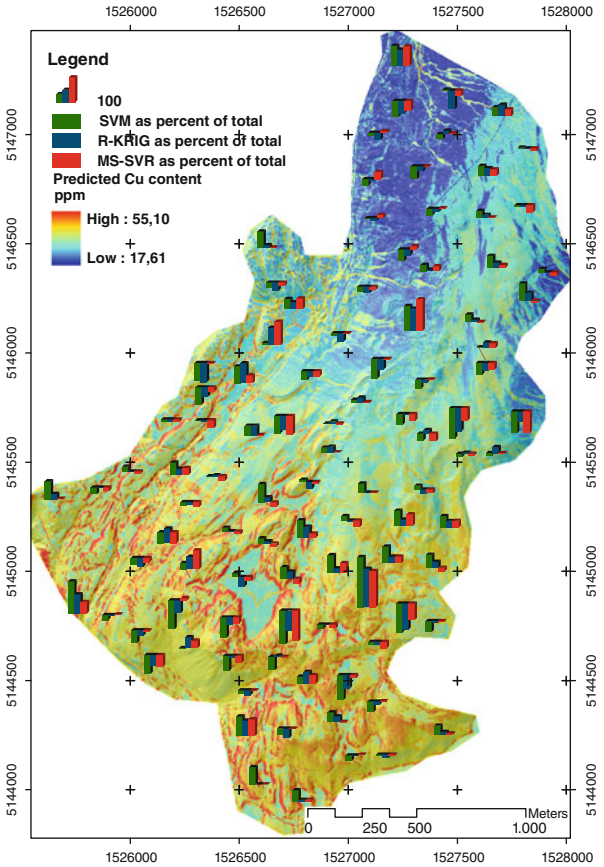


Fig. 17.4 Map of predicted Cu content in the topsoil. The studentized residuals for the three models are shown as *bar-plots*

17.5 Conclusions

Given its superior performance, MS-SVR is a promising approach to digital soil mapping.

Besides the quality of the results, the approach is particularly attractive for building automatic mapping systems. An automatic mapping system could be used for environmental modeling and surveillance, in case of hazardous events. For instance, in case of radioactive or chemical fallout, a fast mapping of the contamination in soils is needed for emergency management. In these occurrences, timing is relevant, and a system requiring few or no user input is especially useful, as it can be operated without the need of an expert, as long as data from soil samples are available.

Another possible use of automatic mapping systems is in the "real-time" production of maps for the end user querying geographic relational databases.

In both of these potential applications the use of a semiparametric SRV approach, both with single or multiple kernel, would make it possible to update existing well assessed models with new data.

Moreover all SVR related methods, could be joined with other kernel based methodologies, such as ranking, spectral clustering and kernel PCA algorithms. As one of the major limitations shown by all the models is the poor performance on extreme values. The implementation of kernel ranking and clustering can prove especially useful in the identification of spatial outliers and in the stratification of the data into more homogeneous groups, defined on the basis of the feature space properties. Further studies will be required to assess the full capabilities of these types of models in mapping tasks.

References

Ballabio, C., 2009. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma 151:338–350.

Bi, J., Zhang, T., and Bennett, K., 2004. Column-generation boosting methods for mixture of kernels. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'04), 521–526. Seattle, WA.

Chang, C.-C., and Lin, C.-J., 2001. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (last verified 25 April 2009).

Cherkassky, V., and Mullier, F., 2007. Learning from Data: Concepts, Theory and Methods, 2nd edition. Wiley, New York, Ch. Support Vector Regression, 439–445.

Cressie, N., 1993. Statistics for Spatial Data. Wiley Interscience, Hoboken, NJ.

De Roo, A. P. J., 1997. Modelling runoff and sediment transport in catchments using GIS. Hydrological Processes 12:905–922.

Drucker, H., Burges, C., Kaufman, L., Smola, A., and Vapnik, V., 1997. Support vector regression machines, pp. 155–161. Advances in Neural Information Processing Systems, vol. 9. NIPS, MIT Press, Cambridge, MA.

Dubayah, R., and Rich, P., 1995. Topographic solar radiation models for GIS. International Journal of Geographical Information Systems 9:405–419.

Hodgson, M. E., and Bresnahan P., 2004. Accuracy of airborne lidar-derived elevation: empirical assessment and error budget. International Archives of Photogrammetry and Remote Sensing, 70:331–339.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A., 2004. Kernlab – An S4 package for Kernel methods in R. Journal of Statistical Software 11:1–20.

Lucieer, A., and Stein, A., 2005. Texture-based landform segmentation of LiDAR imagery. International Journal of Applied Earth Observation and Geoinformation 6:261–270.

McKenzie, N., and Ryan, P., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89:67–94.

Meyer, D., Leisch, F., and Hornik, K., 2003. The support vector machines under test. Neurocomputing 55:169–186.

Smola, A., Friess, T., and Schölkopf, B., 1998. Semiparametric support vector and linear programming machines. Advances in Neural Information Processing Systems, vol. 11. MIT Press, Cambridge, MA.

Sonnenburg, S., Raetsch, G., Schaefer, C., and Scholkopf, B., 2006. Large scale multiple kernel learning. Journal of Machine Learning Research 7:1531–1565.

Van der Walt, C., and Barnard, E., 2006. Data characteristics that determine classifier performance, pp. 60–165. In: Proceedings of the 17th Annual Symposium of the Pattern Recognition Association of South Africa. Parys, South Africa.

Van Remortel, R., Hamilton, M., and Hickey, R., 2001. Estimating the LS factor for RUSLE through iterative slope length processing of digital elevation data within ArcInfo Grid. Cartography 3:27–35.

Wensen, A., and Yanguang, S., 2006. An equivalence between SILF-SVR and ordinary kriging. Neural Processing Letters 23:133–141.

Zevenbergen, L., and Thorne, C., 1987. Quantitative analysis of land surface topography. Earth Surface Processes and Landforms 12:47–56.

Chapter 18
Mapping the CN Ratio of the Forest Litters in Europe-Lessons for Global Digital Soil Mapping

F. Carré, N. Jeannée, S. Casalegno, O. Lemarchand, H.I. Reuter, and L. Montanarella

Abstract The Carbon/Nitrogen ratio (CN) of forest soils is one of the best predictors for evaluating the soil functions mainly involved in climate change issues.

The CN ratio of forest litters depends generally on tree species and forest management which are local factors, but also on broader environmental factors. Thus, the European forest litter CN ratio map is predicted using: (a) punctual CN ratio measurements collected systematically every 16 km in European forests and analyzed according to a common European laboratory method; (b) spatially continuous information on tree species abundance (derived from interpolation) and climate, landform and lithology at 1 km resolution.

The spatial modeling of the CN ratio is done according to complementary approaches: first, a classical kriging approach done on the CN ratio measurements; and second, a neural network approach using a set of nonlinear equations on the environmental predictors. Other multivariate geostatistical approaches were tested but not retained for final results due to lack of correlation between environmental factors.

Twenty percent of CN ratio measurements are kept for validation purpose. The two approaches are compared using coefficient of determinations and Root Mean Square Errors on the validation dataset. Surprisingly, the best approach is the classical kriging, meaning that the spatial structure and variability of CN ratio cannot be explained by the environmental factors, which show high local variation. This leads to a discussion of the quality of the data and to envisage possible risks for global digital soil maps.

Keywords CN ratio · Kriging · Neural network

F. Carré (✉)
INERIS, Scientific Division, Parc technologique Alata, BP 7, 60550 Verneuil en Halatte, France
e-mail: Florence.CARRE@ineris.fr

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_18,

18.1 Introduction

As an indicator of soil mineralization processes, the Carbon/Nitrogen (CN) ratio of the forest soils is one of the best predictors for evaluating soil functions such as biomass production and carbon storage capacity of forest soils. When integrated into risk assessment, these functions can serve for modeling scenarios of soil sustainability with climate change issues (gas fluxes emissions, biofuel production . . .). For instance, for a soil having a relatively high CN ratio, the mineralization process tends to be slower and the weak leaching of nitrogen results in a weak quantity of N gas flux emission.

The CN ratio is very dependent on the forest species (see Section 10.3.3 for more explanations on processes in the litter), management, and environmental factors (Burke et al., 1989) such as climate, relief, soil type and parent material. Furthermore, since forest management is done locally, the CN ratio variability has to be analyzed locally. The final objective is to model and map the local spatial distribution of the forest soil CN ratios for the entire area of Europe based on an European dataset. To this aim, different soil inference systems are tested (spatial and aspatial) based on the European CN ratio database and environmental data. In order to analyze the consistency of the data for such a large extent, a preliminary data analysis is performed. Then, three complementary soil inference systems are developed and compared. The results are then discussed relative to the dataset.

18.2 Material

18.2.1 The CN Data

There are 5,289 CN ratio measurements (see Fig. 18.1) available from the 1st Forest Focus campaign database (the official date is 1996 but the measurements are from 1987 up to 1995). The plot sampling and analysis are done according to the common manual which must be followed by the Member States (UNECE, 2003). The plots are analyzed systematically, usually every 16 km where forest exists. The plots are described according to 5 strata: the soil litter, the topsoil horizon (0–10 cm) and lower horizons (10–20 cm; 20–40 cm; 40–80 cm). Each strata measurement is a composite of at least 4 soil samples about 50 m from each other. For this study, we focus only on the litter dataset.

Out of the 5,289 measurements, 1,929 were lacking either the CN ratio information, or the coordinates. Moreover, due to possible errors of laboratory analysis, the 3,360 measurements were analyzed using classical statistics and an indicator of the spatial variability (variogram cloud analysis). Twenty outliers were detected and removed. At the end, there were only 3,340 remaining measurements of CN ratio for the litter.

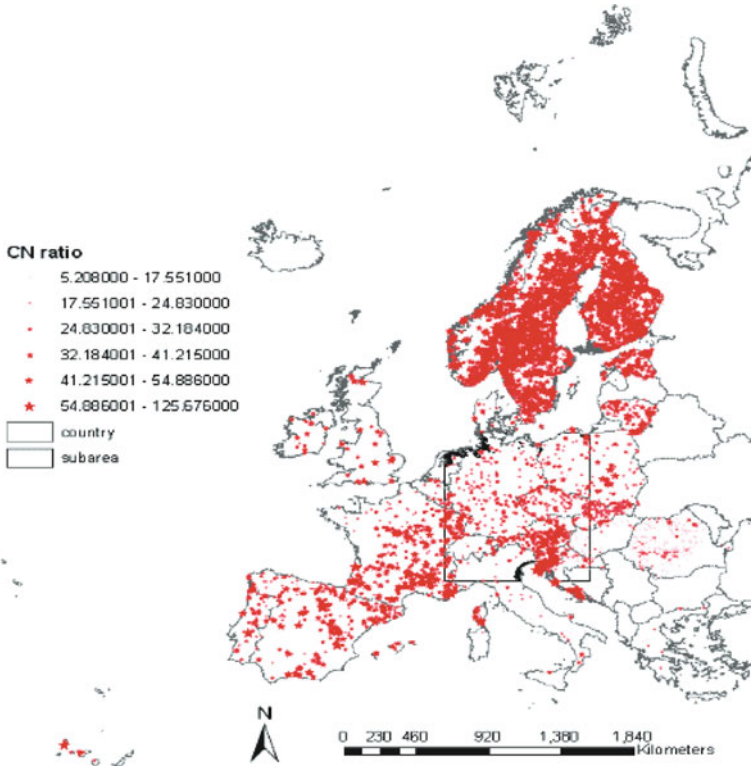


Fig. 18.1 The sample locations of CN ratio across Europe and the location of the study subarea

18.2.2 The Environmental Dataset

In order to improve the spatial prediction of the CN ratio for the forest soil litters, 39 soil covariates which can influence the CN ratio were considered. These are:

- The 23 main abundant tree species in Europe (representing 98% of the total number of trees) as of 1996. The tree species were punctual information for which only 58% had correspondence with soil sample locations. They were then transformed into raster grid cell by an Inverse Distance Weighted Interpolation, consistent with the European Forest species map estimation of 2004 (Casalegno, 2009). For the tree species prediction we also tried to use the FAPAR (Fraction of Absorbed Photosynthetically Active Radiation) of Europe for the period 1997–2000 (Gobron et al., 2006). Unfortunately, there was no correlation. We prefered a bottom-up approach compared to a top-down approach for predicting forest vegetation cover;
- Landform attributes derived from the SRTM (mean altitude, standard deviation altitude in a 1km raster grid cell, slope, aspect, curvature and moisture index-see Section 20.2.3 for more explanation on covariates);

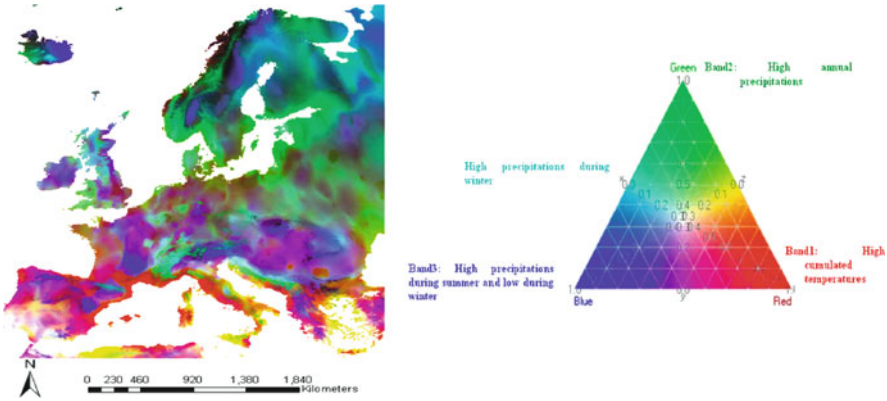


Fig. 18.2 (**a**) The bioclimatic map of Europe (1 km resolution); (**b**) the associated legend

- Principal components (see Fig. 18.2) on climatic data over the 1950–2000 period (annual mean temperature, mean diurnal range of temperature, temperature seasonality, isothermicity, annual rainfall) and on derived bioclimatic variables (ombrothermic index, drought stress index, thermicity index, continentality);
- Mean annual evapotranspiration, cumulated evapotranspiration and box moisture index, also derived from climatic data;
- Latitude and longitude coordinates.

All the raster grids were at the resolution of 1 km × 1 km and projected according to the European INSPIRE standards (ETRS89 Lambert Azimuthal Equal Area).

18.3 Methods for CN Ratio Modeling and Mapping

The modeling efficiency of the CN ratio using spatially continuous covariates was assessed on a subarea having transboundaries (Fig. 18.1 – see Central Europe), which contain 739 CN ratio measurements. 80% of the measurements were randomly selected for modeling (616 measurements) and 20% for validation (123). Different approaches were tested and developed.

(a) The first approach simply consisted of a classical ordinary kriging approach (Wackernagel, 1995) on the CN ratio measurements. The modeling was based on the following equation:

$$\mathrm{CN}(x_0) = \sum \lambda_i \, \mathrm{CN}(x_i)$$

Where λ_i are the kriging weights, $\mathrm{CN}(x_i)$ the measurements at location x_i and $\mathrm{CN}(x_0)$, the prediction at location x_0.

(b) The second approach was done using a neural network approach (see Section 13.2.5.1) on the CN ratio predictors. This two-stage classification

approach consisted of a set of nonlinear equations that predicted the CN ratio from the soil covariate predictors in a flexible way, using 8 node layers of linear regressions and S-shaped ($1/1 + e^{-x}$) functions. The modeling was based on the following equation:

$$\mathrm{CN} = \beta_0 + \sum \beta_i H_i$$

with $H_i = 1/(1 + e^{-xi})$ are the node layers (hidden layer)

$$x_i = \alpha_{i,0} + \sum \alpha_{i,j}{}^{*} p_j$$

$p_1, \ldots, p_n$ are the predictors and $\alpha_1, \ldots, \alpha_n$ are the regression coefficients.

Other multivariate geostatistical approaches were tested, such as cokriging and Min/Max Autocorrelation Factors, first using a reduction of the number of environmental predictors and then a kriging with external drift on the resulting MAF factors (Desbarats and Dimitrakopoulos, 2000). They were not retained for further analysis due to weak correlation between CN ratio and environmental factors and high small scale variability of the predictors.

The first approach (a) is spatial whereas the second one (b) is aspatial. For comparing the efficiency of the two approaches and to reach a conclusion on the need of using spatial or aspatial approaches, the percentage of good prediction (R^2 adjusted coefficient of determination) and RMSE were computed on the validation dataset.

Final mapping was limited to the areas covered by more than 10% of forests in Europe to get the final distribution of the CN ratios in the forest litters. Forest occupation data was derived from the Pan European Forest/Non-Forest map 2000 (Pekkarinen et al., 2009). All maps followed the standards of the predictors (INSPIRE resolution and projections).

18.4 Results

The three approaches are discussed and then compared.

18.4.1 The Classical Kriging Approach

The CN ratio variogram (Fig. 18.3) was obtained after a preliminary Gaussian transformation to reduce the impact of data variability. The variogram shows important small scale variability (50% of total variability), attributable to local forest management and potential measurement errors and/or inconsistencies in measurement procedures between adjacent countries (this point will be discussed later). Moreover, there was a clear spatial structure with a range approximately equal to 250 km (maximum distance of correlation).

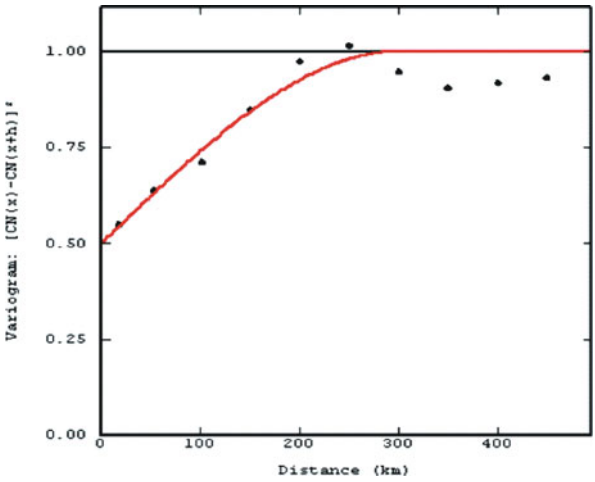


Fig. 18.3 The variogram of Gaussian transformed CN ratio measurements in the subarea. Experimental points in *black*, modeled variogram in *red*. Statistical variance is also displayed (*black line*)

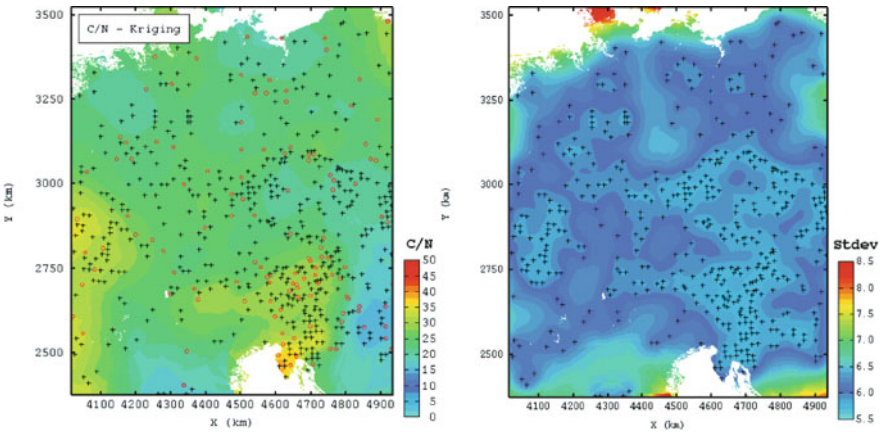


Fig. 18.4 (**a**) Map of the CN prediction derived from classical kriging; (**b**) map of the standard deviations of the kriged errors. *Red rings* are the validation points, *Black crosses* are the modeling points

The resulting maps of the kriging (CN ratio prediction in Fig. 18.4a) show a smoothed CN ratio prediction as expected, increased by the modeled nugget effect. The standard deviation error map allows for future possible sampling areas (Fig. 18.4b).

The adjusted coefficient of determination between predicted data and measurements was 0.60 and the Root Mean Square Error was 4.91 (Fig. 18.6a).

18.4.2 The Neural Network Approach

The neural network, contrary to the kriging, shows high local variability of the CN ratios (Fig. 18.5), respecting in some parts, boundaries of tree species and relief. The

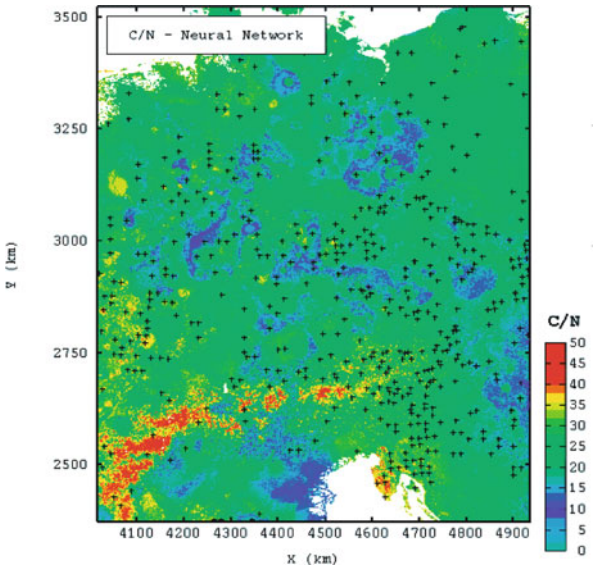


Fig. 18.5 Map of the CN ratio prediction derived from the neural network approach

adjusted coefficient of determination between predicted data and measured data was 0.40 and the Root Mean Square Error was 4.85 (Fig. 18.6b).

The comparison of the two maps (Figs. 18.4a and 18.5) reveals global trends in the CN ratios of Central Europe. The CN ratio is globally medium in this area but high in the northern part of Croatia and the eastern French border. However, the neural network map clearly shows shapes of environmental factors, particularly landform, but also some round spots which represent artifacts in the prediction of tree species. This map gives more importance to local variability than the kriged map. Both maps are defendable from a methodological point of view. For an assessment of the CN ratio, we would recommend the kriging method due to the reasons in the following discussion.

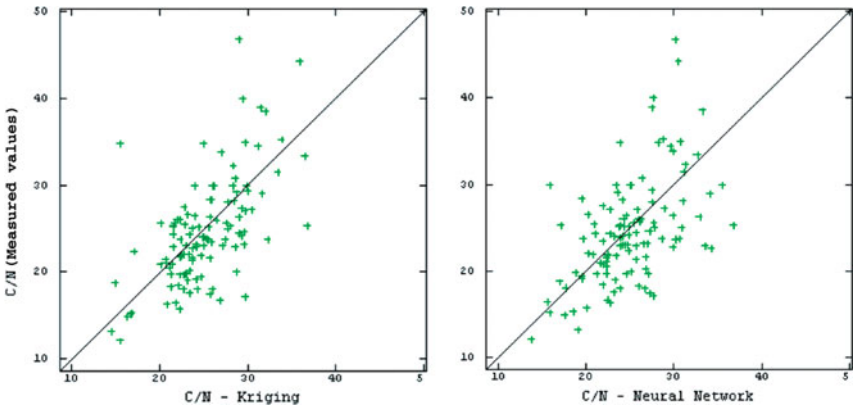


Fig. 18.6 (**a**) Scatter plot of validation against predicted values for the ordinary kriging; (**b**) scatter plot of validation against predicted values for the neural network

18.5 Discussion

The results of both methodologies are not satisfying and should be improved for risk assessments. Moreover, it is interesting to observe that taking into account spatial variability of the CN ratio without considering potential predictors sometimes gives better results than taking into account the predictors without considering spatial trends. This could be due to four main reasons:

(a) High local variability of the CN ratio which can be modeled at 1 km resolution. The four sampling repetitions are then not sufficient;
(b) The measurement errors due to sampling and laboratory analysis as shown by the variogram analysis, increased by an expected lack of consistency between sampling/analysis protocols on adjacent countries (see hereafter);
(c) The problem of Inverse Distance Weighted (IDW) interpolation of the tree species abundances. Indeed, tree abundances show a high spatial variability, and as a consequence, the IDW approach probably gives poor estimates of tree abundances outside of data locations.
(d) The models may not be efficient enough to predict the CN ratio.

These different issues are now studied. The prediction of tree species for the year 2004 shows that regression trees on environmental predictors should be preferably used (Casalegno et al., 2010). Concerning the study of local variability and other model tests, they should be studied once the quality of the data has been estimated.

For that, we launched inquiries asking each Member State the final methodology used for collecting and analyzing data. These inquiries demonstrated that each Member State followed its own methodology, the one they traditionally used concerning sampling strategy, even when common standards should be followed. Furthermore, the laboratory measurements are also very dependent on the calibration of the materials. It is then difficult to compare transboundary samples and also national samples, since sometimes different laboratories were used for the chemical analyses. Some work should be done on transboundary corrections of the measurements. To this aim, the new BIOSOIL project (IES, 2008) aims to collect data from existing and new samples in Member States and to analyze them at a central laboratory, in order to cross-check the previous measurements and to study soil and forest biodiversity in Europe. But, even using a central laboratory, possible errors due to transport and instability of soil samples could be introduced.

18.6 Conclusion

The spatial predictions of the CN ratio over European forest litters was done from CN ratio measurements across Europe and potential predictors related to climate, relief and tree species. Climate and relief are spatially continuous information whereas tree species have been interpolated by inverse distance weighting from

punctual information. Two predictions were completed: one using a classical kriging on the CN ratio measurements and another one using a neural network approach on the environmental predictors. Both methods are quite similar and show globally the same spatial trends (same order of magnitude of the predictions and higher values in the southern and western parts of Central Europe). Therefore, the use of spatial predictors is not necessary when kriging. This allows for discussing and reanalyzing the quality of the data. A first check on data quality demonstrates that even with a common base methodology and common criteria, errors are introduced in the measurements, due mainly to cultural heritage, but also due to technical issues (material calibration, transport and instability of the samples . . .). The issue Europe is facing is quite representative of potential problems we can face during the building of global digital soil maps. The solution provided here for a worldwide mapping of soil attributes with legacy data is then to deal with prior information, containing prior metadata (when existing). Then make a first spatial prediction which allows for identifying errors or gaps in the metadata and then make a final prediction. Rescuing the data appears to be the most difficult problem for the final mapping. In terms of economics, it should be necessary to check if the rescue cost is not more than cost of getting new data.

References

Burke, I.C., Yonker, C.M., Parton, W.J., Cole, C.V., Flach, K.D., and Schimel, S., 1989. Texture, climate, and cultivation effects on soil organic matter content in U.S. grassland soils. Soil Science Society of America Journal 53:800–805.

Casalegno, S., 2009. Forest species and forest types current distribution and habitat suitability under climate change. Retrieved from the European Commission Joint Research Centre website: http://forest.jrc.ec.europa.eu/climate-change. Data downloadable at: http://efdac.jrc.ec.europa.eu/climate (Last accessed 22 April 2010).

Casalegno, S., Amatulli, G., Camia, A., Nelson, A., and Pekkarinen, A., 2010. Vulnerability of Pinus cembra L. in the Alps and the Carpathian mountains under present and future climates. Forest Ecology and Management 259:70–761.

Desbarats, A. J., and Dimitrakopoulos, R., 2000. Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. Mathematical Geology 32:919–942.

IES, 2008. The Biosoil project. http://forest.jrc.ec.europa.eu/ForestFocus/biosoil.html (last verified 10 September 2008).

Gobron, N., Pinty, B., Melin, F., Taberner, M., Verstraete, M., Robustelli, M., and Widlowski, J.L., 2006. Evaluation of the MERISIENVISAT fAPAR product. Advanced Space Research 37, doi: 10.101 6lj.asr.2006.02.048.

Pekkarinen, A., Reithmaier, L., and Strobl, P., 2009. Pan-European forest/non-forest mapping with Landsat ETM+ and CORINE Land Cover 2000 data. ISPRS Journal of Photogrammetry and Remote Sensing 64:171–183.

UNECE, 2003. ICP Forests Manual on methods and criteria for harmonized sampling, assessment, monitoring and analysis of the effects of air pollution on forests, 2003. Part IIIa Sampling and Analysis of Soil and Part IIIb Soil Solution Collection and Analysis. United Nations Commission for Europe Convention on Long-Range Transboundary Air Pollution, International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests.

Wackernagel, H., 1995. Multivariate Geostatistics – An Introduction with Applications, Springer, Berlin.

Chapter 19
Spatial Prediction and Uncertainty Assessment of Soil Organic Carbon in Hebei Province, China

Yong-Cun Zhao and Xue-Zheng Shi

Abstract To quantify the spatial distribution of SOC in Hebei Province five models were compared: multiple linear regression (MLR), universal kriging (UK), regression-kriging (RK), artificial neural network combined with kriging (ANN-kriging), and regression tree (RT). The modelling was supported by 359 SOC density (total SOC by volume, SOCD) data points, as well as relief parameters derived from a 100 m × 100 m resolution DEM, and NDVI calculated from NOAA AVHRR data to map SOCD (to a depth of 1m) spatial distributions. Only 19.5% of the total SOCD variation can be explained by MLR method, the UK method resulted in a wider range of SOCD compared with MLR method. The UK method and RK method explain 53 and 65% of the total variation, respectively, and the local variation of lower SOCD in the southeast of the province was detected. The ANN-kriging and RT mapping both explained 67% of the total variation. Compared to ANN-kriging, the RT method has lower root mean square prediction error. The sequential indicator simulation (SIS) was applied for assessing topsoil SOCD (0–20 cm) uncertainty at unsampled locations. The conditional variance of 1,000 realizations generated by SIS was greater in mountainous areas where SOCD fluctuated the most, and the uncertainty was less on the plain area where SOCD was consistently low. The RT model is of best performance for mapping the spatial distribution of SOCD, and the SIS technique can quantitatively assess the local and spatial uncertainty of SOCD being greater than a given threshold.

Keywords Soil organic carbon (SOC) · Spatial prediction · Uncertainty assessment · Environmental correlation · Sequential indicator simulation

Y.C. Zhao (✉)
State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, No. 71 East Beijing Road, Nanjing, 210008 China
e-mail: yczhao@issas.ac.cn

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_19,

19.1 Introduction

Digital mapping of soil organic carbon (SOC) has many applications including site-specific crop management or in environmental modeling. Spatial prediction of SOC involves uncertainties that need to be made explicitly because they can be propagated into subsequent modelling. Existing methods for predicting the spatial distribution of soil properties includes linear regression, geostatistical methods, and advanced nonlinear regression methods such as regression trees, neural network (McBratney et al., 2003). Wherever secondary attributes are used for mapping the spatial distributions of SOC, they are most commonly sourced from DEM and remote sensing. Several studies have illustrated the potential for utilizing exhaustive secondary information such as terrain indices (Mueller and Pierce, 2003) or remote sensing (Chen et al., 2000) for making more precise maps of SOC.

Arrouays et al. (1998) generated SOC predictions in temperate forest soils of France using a simple linear regression model, and utilized the multiple linear regression model of parent material, elevation, and slope. Cheng et al. (2004) predicted the spatial distributions of SOC in Xingguo county of China. Thompson and Kolka (2005) developed soil-landscape models that quantify relationships between SOC and topographic variables derived from DEMs within a 1,500-ha watershed in eastern Kentucky, and the results showed that, despite low coefficients of correlation between measured SOC and individual terrain attributes, the developed robust linear regression models can explain up to 71% of SOC variability using three to five terrain attributes. With respect to geostatistical methods, hybrid geostatistical procedures that account for environmental correlation allow utilizing secondary information that is often available at finer spatial resolution than the sampled values of target variable. If the correlation between primary and secondary variables is significant, hybrid techniques generally result in more accurate local predictions than ordinary kriging or other univariate predictors (Simbahan et al., 2006). Terra et al. (2004) determined relationships between SOC, terrain attributes, soil electrical conductivity, soil texture and soil survey map units in a 9 ha coastal plain field, and compared the ordinary kriging, co-kriging, regression kriging and multiple regression method for mapping the spatial distributions of SOC. Ping and Dobermann (2006) evaluated the precision of ordinary kriging (OK) and regression kriging (RK) for mapping the spatial patterns of SOC in two irrigated maize fields in Nebraska; exhaustive ancillary variables used included relative elevation, slope, soil electrical conductivity, and remotely sensed soil surface reflectance. The results showed that the relative improvement of RK techniques over OK largely depended on the strength of the correlation between SOC and ancillary variables. Adaptive and non-parametric models used to predict the spatial patterns of soil properties may also be applied to predict the spatial distribution of SOC. These include artificial neural networks and regression trees. An important characteristic of the two techniques is their adaptive nature to learning for solving problems (Park et al., 2005). They are also an illustration of complex and non-linear relationships without rigorous assumptions regarding the distribution of samples (Bishop, 1995).

Hebei Province surrounds Beijing, the capital of China. Digital mapping of SOC is important for site-specific crop management and for environmental modeling and planning in the province. Therefore, the specific objectives of this study were: (1) to identify useful environmental variables for predicting spatial distributions of SOC in Hebei province; (2) to evaluate how the precision of SOC maps is affected by multiple linear regression, universal kriging, regression kriging, artificial neural network combined with kriging, and regression tree methods; and (3) to assess the uncertainty of mapping SOC in Hebei province.

19.2 Materials and Methods

19.2.1 Study Area

Hebei Province surrounds Beijing, the capital of China (Fig. 19.1). The province is characterized by high mountains and tablelands in the northwest and low plains in the southeast. The mountains and tablelands are mostly over 1,000 m in elevation with peaks exceeding 2,000 m, while the plains are only 3–5 m above sea level. The total land area of Hebei Province is 187,693 km^2. The tablelands occupy about 9% of the province, the mountains 50%, and the plains 41%. Hebei Province has a mean annual temperature ranging from 0 to 13°C and annual precipitation between 300 and 800 mm (temperate continental monsoon climate) (Ding, 1992). Significant differences in climate, geomorphology, geohydrological conditions, and parent materials throughout the province result in a great variety of vegetation and soil types and distribution patterns. Therefore the province is ecologically representative for much of northern China.

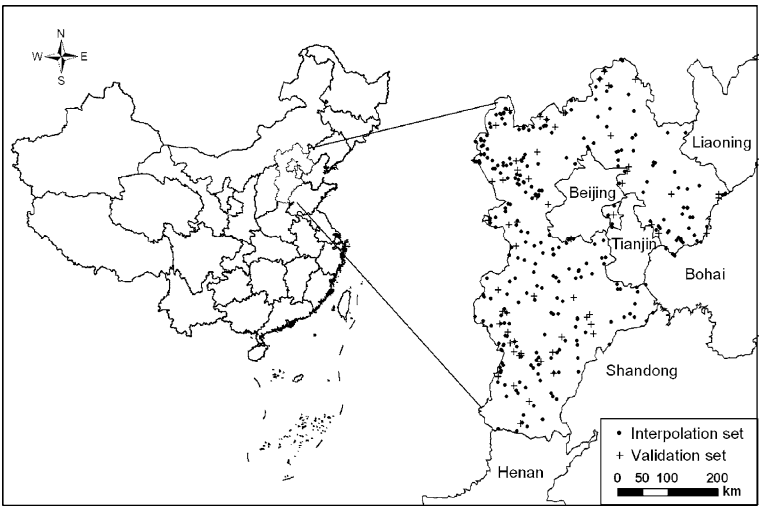


Fig. 19.1 Distribution pattern of SOC data in Hebei Province of China

19.2.2 Digital and Field Data

Digital data used in this study included relief parameters (elevation, slope, aspect, plan curvature, profile curvature, compound topographic index CTI) derived from a 100 m × 100 m resolution DEM, and NDVI calculated from NOAA-AVHRR images. The NDVI data was obtained from the Chinese Natural Resources Database using the equation NDVI = max $[(b2_i - b1_i)/(b2_i + b1_i)]$ $i = 1, \ldots, n$ (where b1 and b2 are the first and second band of AVHRR image, respectively, n is number of scenes for the province obtained in July, 1998).

Field soil data included 359 SOC density (total SOC by volume, SOCD) data points (see Zhao et al., 2006 for the SOCD calculation method), and each point has two SOCD values calculated to a maximum depth of 1 m and 20 cm, respectively. The data were from the second national soil survey of China, and the sampling density was about one profile per 500 km^2. Since 1 m and 20 cm are widely used depths in SOC storage estimates and mapping, the 359 SOCD data to 1 m depth was randomly divided into an interpolation (300 points) and validation set (59 points) for comparing different spatial prediction methods (Fig. 19.1). The 359 SOCD data of 20 cm depth were used to quantify uncertainties.

19.2.3 Inference Models

For mapping the spatial distribution of SOCD to 1 m depth, five inference models were used:

(i) Multiple linear regressions (MLR) with the classical least-squares regression model (see Section 21.2) SOCD data was taken as response variables, and relief parameters (i.e. elevation, slope, aspect, CTI and etc.) and NDVI were treated as independent variables for mapping the SOCD spatial patterns
(ii) Universal kriging (UK): the trend component of SOCD data was modeled as a function of sample coordinates, and residues of SOCD data were estimated by ordinary kriging.
(iii) Regression kriging (RK): the trend component of SOCD data was modeled using MLR, and then ordinary kriging estimates of the residues were added to the MLR predictions. A linear relationship was assumed between secondary variable and target variable.
(iv) Artificial neural network combined with kriging (ANN-kriging): the trend structure of SOCD data was modeled using a feed-forward back-propagation network (FFBP) model (this method is described in Section 18.3), and then the ordinary kriging estimates of the residues were added to the FFBP predictions. Here a non-linear relationship was assumed between the secondary variable and the target variable.
(v) Regression Tree (RT): The tree structure was generated by splitting the data into nodes in a binary fashion until the node was too homogenous or until there

were too few observations. Splitting was optimized by minimizing residual deviance. The advantage of the regression tree method over the linear model is the ability to deal with nonlinearity. In addition, the RT method requires no assumptions about the data and is able to deal with non-additive behaviors.

The prediction performance was evaluated using interpolation and validation sets. The interpolation set was used to derive the sum of squares of residuals (SSE). The true prediction accuracy was evaluated by comparing estimated values $\hat{z}(s_j)$ with actual observations at validation points $z(s_j)$ in order to assess the systematic error, calculated as mean prediction error (MPE):

$$\mathrm{MPE} = \frac{1}{l}\sum_{j=1}^{l}[\hat{z}(s_j) - z(s_j)] \tag{19.1}$$

The accuracy of prediction, calculated as root mean square prediction error (RMSPE):

$$\mathrm{RMSPE} = \sqrt{\frac{1}{l}\sum_{j=1}^{l}[\hat{z}(s_j) - z(s_j)]^2} \tag{19.2}$$

Where l is the number of validation points.

The sequential indicator simulation (SIS) was used for quantifying uncertainties of mapping topsoil SOCD (0–20 cm). Detailed descriptions about the SIS algorithm and procedures can be found in several geostatistical books (i.e. Deutsch and Journel, 1998; Goovaerts, 1997). One thousand times of SIS were carried out in this study and the search radii were set as the ranges of the semivariogram models. The sisim subroutine in GSLIB software package was used to perform the SIS.

19.3 Results and Discussion

19.3.1 Results from Model

19.3.1.1 Descriptive Statistics

The descriptive statistics (Table 19.1) showed considerable variation in the SOCD data. The randomly separated interpolation and validation sets are of similar statistical characteristics with the full data set, i.e. mean, median, standard deviation as well as coefficient of variation, indicating that the separation of interpolation and validation sets were representative and appropriate.

Table 19.1 Descriptive statistics for SOC density (SOCD) in Hebei Province of China

		Mean	Min	Median	Max	STD	CV
Field data		kg C m^{-2}					%
SOC density (0–1 m)	Full data set ($n = 359$)	8.03	0.20	6.73	39.28	6.37	79
	Interpolation set ($n = 300$)	8.02	0.20	6.81	38.98	6.10	76
	Validation set ($n = 59$)	8.08	0.25	6.79	39.28	6.26	77
SOC density (0–20 cm)	Full data set ($n = 359$)	2.75	0.13	1.94	17.15	2.44	89

Abbreviations: STD = standard deviation; CV = coefficient of variation

19.3.1.2 Spatial Predictions of SOCD (to 1 m Depth)

MLR, UK, and RK

Correlation analysis showed significant positive correlation between SOCD and elevation and slope with $r = 0.425$ and $r = 0.172$ ($p < 0.01$), respectively. Correlation between SOCD and NDVI was also significant with $r = 0.129$ ($p < 0.05$) whereas correlation between SOCD and CTI was negative with $r = -0.183$ ($p < 0.01$); correlations between SOCD and the other relief parameters were not significant. The multiple linear regression equation between SOCD and elevation, slope, CTI and NDVI was:

$$\mathrm{SOCD} = 4.96 + 0.0041\mathrm{Elevation} + 0.156\mathrm{Slope} - 0.0261\mathrm{CTI} + 1.66\mathrm{NDVI} \quad (19.3)$$

where $R^2 = 0.195$ ($p < 0.001$).

Figure 19.2a shows the SOCD distribution maps obtained by MLR method, which reflect the spatial relationships of SOCD, elevation, slope, CTI, and NDVI. The SOCD values are large in the northwestern mountainous areas where elevation is high and the land is sloping, while SOCD in the southeastern plain is low when the CTI is high. A strong smoothing effect can be observed in the MLR method.

There were significant positive correlations between sample ordinates Y, co-located log-transformed SOCD values, elevation, slope, CTI, and NDVI ($p < 0.05$). The sample abscissas X is also significantly correlated with elevation, CTI, and NDVI ($p < 0.05$) indicating that the universal kriging method can be used for mapping the SOCD distribution. Estimates of the UK method (Fig. 19.2b) showed similar SOCD distribution patters with MLR except that more detailed SOCD variations in the plains were detected using the UK method. Prediction using the RK method (Fig. 19.2c) showed detailed SOCD variations in the mountainous areas (northwestern parts) and plains of the province and the smoothing effect of RK method was much smaller compared to the MLR and UK method.

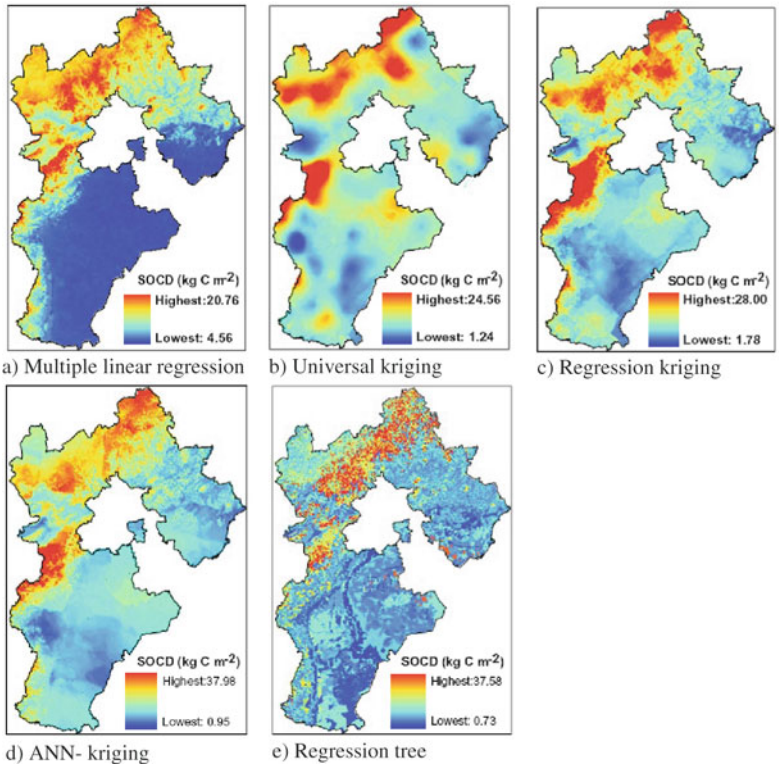


Fig. 19.2 Spatial distribution maps of SOC density (to a depth of 1 m) for Hebei Province

ANN-kriging and RT

Since the SOCD values in the province were closely related to the elevation, slope, CTI, and NDVI, a general function f (*Elevation, Slope, CTI, NDVI*)$\longrightarrow$SOCD was established. A multi-linear regression was applied to create this function, but this functional relationship probably is a nonlinear function because complex nonlinear interactions may exist between SOC and landscape as well as vegetation (Somaratne et al., 2005). The artificial neural network (ANN) and regression tree can be used to map nonlinear relationships.

The ANN structure used in this study is a multilayer perceptron (MLP), which is the most popular neural network structure in ecological modelling and soil science. In order to introduce nonlinearity during the ANN training, the hyperbolic tangent function is used. The MLP models with the following parameters were used in this study: four input neurons, elevation, slope, CTI, and NDVI, one hidden layer and output neuron describing SOCD trends. Backpropagation training with Levenberg-Marquardt followed by conjugate gradient algorithm was used in order to avoid local minima; the learning rate was 0.01. Variogram analysis of the obtained residuals for the trained neural networks with varying number of neurons in the hidden layer

showed optimal results (in the sense of modeling non-linear trends) using MLP with five neurons in a single hidden layer. Further increase of the number of hidden neurons led to extracting more detailed local peculiarities of the pattern reflected by the correlation range. There is no theory for determining the optimum numbers of hidden layers to approximate a given function. The decision was made in order to reduce the complexity of the models and prevent overfitting of the SOCD trends, the variogram of the SOCD residue of ANN-kriging method was fitted with a spherical model, with a nugget C_0 of 12, sill of 17.4, and range of 26,000 m. The RMSE of the fitted spherical model was 0.009. The variogram range of the residue of MLP prediction was much smaller than those of MLR and first order trend surface methods, indicating that the ANN method was better than the other two methods.

The spatial patterns of SOCD obtained by ANN-kriging (Fig. 19.2d) were more complex than when using RK method, and the range between the minimum and the maximum was wider than the RK method.

The tree structure was generated by partitioning the data recursively into a number of groups, whereby each division maximizes some measure of difference in the response variable in the resulting two groups (McBratney et al., 2003). The global distribution patterns of SOCD obtained by RT method (Fig. 19.2e) were similar to those of other four methods. Most local variations of SOCD can be detected by the RT method. The SOCD range with the RT method was much wider than the MLR and UK methods as the smoothing effect of kriging method narrowed the SOCD range.

19.3.1.3 Uncertainties for Mapping Topsoil SOCD (to 20 cm depth)

The E-type estimates (Fig. 19.3a) showed the global distribution trends of SOC density in topsoil of Hebei province, and the conditional variance (Fig. 19.3b) reflects the fluctuation of simulated SOC density values at any unsampled locations. The conditional variance of simulated values was zero at sampled locations because the SIS employed by this study is conditional simulation and the SOC density values of sampled locations are maintained. The conditional variance was larger in the high-valued parts of the study area where SOC density fluctuated the most so that the largest uncertainty was expected. The uncertainty decreased in the southeast of the study area where SOC density was low. Areas with larger uncertainty were mainly located in the mountainous parts, implying that topography has played a role in SOC density, and the topographic factors must be considered for more accurate predictions of SOC density of the province

19.3.2 Accuracy Assessment of Model

19.3.2.1 Validation for Spatial Prediction of SOCD (to 1 m depth)

The analysis of SOCD prediction accuracy (Table 19.2) showed that the sum of squares of SOCD residuals (SSE) using MLR method was large and only 19.5%

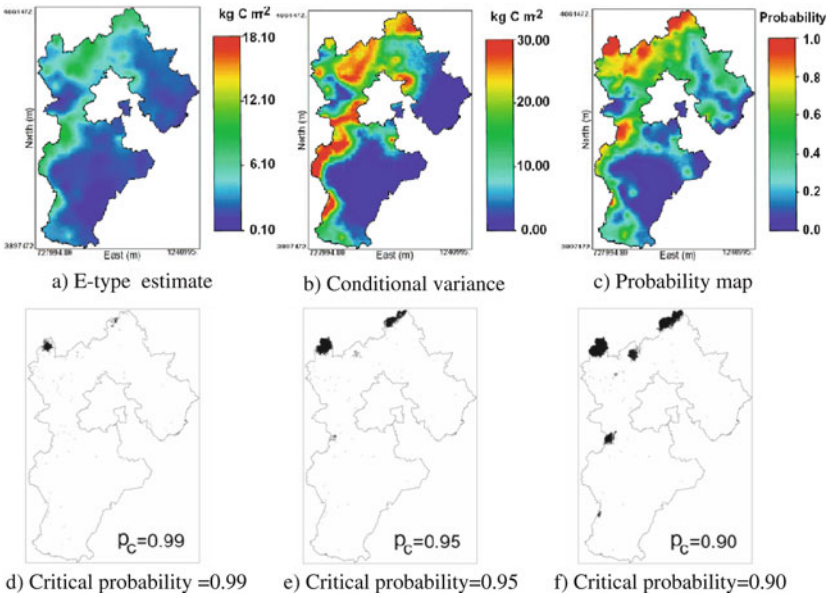


Fig. 19.3 The E-type estimates, conditional variance, probability map of SOC density (to a depth of 20 cm) being greater than 2.41 kg C m^{-2} (calculated from the 1,000 SOC density realizations generated by SIS), and the areas obtained by the rule Prob $[z(x') > 2.41] \geq p_c$ at three given critical probability values

of the total variation was explained. The UK method resulted in a lower SSE and a wider range of SOCD compared to the MLR method. The UK method and RK method explained 53 and 65% of the total variation of SOCD, respectively. After the ANN-kriging and RT methods were applied, the SSE decreased significantly and 67% of the total variation was explained by both ANN-kriging and RT method. Moreover, compared to the ANN-kriging method, the RT method had a lower root mean square prediction error.

Table 19.2 SOC density (0–1 m) prediction accuracy at validation points

	Interpolation set		Validation set	
Prediction method	TSS	SSE	MPE	RMSPE
MLR	11,127	8,832	2.13	9.80
UK	11,127	5,227	–1.57	8.80
RK	11,127	3,896	–1.35	8.06
ANN-kriging	11,127	3,683	–1.38	7.89
RT	11,127	3,679	–1.27	7.78

Abbreviations: TSS = total sum of squares; SSE = sum of squares of residuals; MPE = mean prediction error; RMSPE = root mean square prediction error; MLR = multiple linear regression; UK = universal kriging; RK = regression kriging; ANN-kriging = artificial neural network combined with kriging; RT = regression tree

19.3.2.2 Local Uncertainty and Spatial Uncertainty for Mapping Topsoil SOCD (to 20 cm depth)

Since the SOCD for the topsoil fluctuates in areas with high SOCD, a relatively higher SOCD threshold, $Z_t = 2.41\,\mathrm{kg\,C\,m^{-2}}$ (60% percentile), was selected to calculate the probability for SOCD values that are larger than a specified threshold value. This was also done to to assess the uncertainty of SOCD at any unsampled locations using the SIS method. Figure 19.3c shows that SOCD in most parts of the province is less than $2.41\,\mathrm{kg\,C\,m^{-2}}$. With a given critical probability p_c, e.g. 0.90, Fig. 19.3c can be used to obtain the areas where SOCD of all locations is greater than Z_t based on the rule $\mathrm{Prob}[z(x') > 2.41] \geq 0.90$ (black zones in Fig. 19.3f; where $z(x')$ is the unknown SOCD value at location x'). However, the probability map in Fig. 19.3c can not provide a measure on the reliability of the obtained areas because the ccdf obtained by SIS only provides a measure of local uncertainty, and a series of single-point ccdfs provides no measure of spatial uncertainty. The joint probability obtained from the realizations generated by the SIS method can be used to assess spatial uncertainty. If the given critical probability is 0.95, the joint probability will be only 0.12 for 3,240 simulated locations (Table 19.3) in the areas obtained by $\mathrm{Prob}[z(x') > 2.41] \geq 0.95$ (Fig. 19.3e). There is a high spatial uncertainty although the critical probability is high for a single location selected for delineating the areas. When the selected critical probability is 0.99, the areas obtained by $\mathrm{Prob}[z(x') > 2.41] \geq 0.99$ are presented in Fig. 19.3d. The probability is 800 out of 1,000 realizations where the SOCD of 975 simulated locations all being greater than Z_t and, therefore, the spatial uncertainty is much smaller.

Table 19.3 Assessment for spatial uncertainty of areas where SOC density (0–20 cm) is greater than $2.41\,\mathrm{kg\,C\,m^{-2}}$ based on joint probability

	Critical probability (p_c)		
	0.99	0.95	0.90
Number of cells	975	3,240	6,305
Joint probability	0.80	0.12	0.01

19.3.3 Comparisons of Prediction Methods in Other Parts of the World

Mathematical modeling has been widely used to the predict spatial distribution of SOC, and numerous ancillary variables, particularly parameters derived from DEMs and remotely sensed data, had been applied for improving the prediction accuracy of SOC distribution (Chai et al., 2008; Chen et al., 2008; Mueller and Pierce, 2003; Simbahan et al., 2006; Somaratne et al., 2005; Venteris and Slater, 2000; Wu et al., 2008). Generally, the MLR model with ancillary variables such as DEMs and satellite images had the worst performance compared to the ANN/geostatistics-related hybrid models. The spatial prediction of SOC in Coshocton

(Ohio) using MLR and terrain attributes derived from high resolution DEMs indicates that the model explained only 55% of the variation in A horizon SOC (Venteris and Slater, 2000). The MLR prediction in this study was even worse, with only 20 % of SOC variation explained. The geostatistics-related hybrid models that utilized secondary information usually produced SOC maps of higher quality than those that did not. For example the co-kriging with remotely sensed data (Wu et al., 2008) and RK model with terrain attributes (Chai et al., 2008) significantly improved the SOC prediction accuracy. The RK model used in this study also showed that its performance was obviously higher than that of MLR in predicting SOC distribution of Hebei Province, with 65% of SOC variation explained when the RK model was applied. As for ANN, results of SOC prediction in Sri Lanka (Somaratne et al., 2005) suggested that the predictive performance was also higher than that of MLR models. Our results for Hebei Province showed that the SOC variation explained up to 67% when the ANN model was combined with the ordinary kriging. This indicates that ANN-kriging hybrid model performed better in terms of consistently increasing prediction accuracy and flexiblity in modeling of the multivariate relationships between SOC and secondary information. The RT model used in this study explained 67% of SOC variation, and the interactions between SOC, terrain attributes, and vegetation were easy to interpret, but the interpretation of the results of neural network was more difficult.

19.4 Conclusions

Large variation exists in measured SOCD data of Hebei province. The SOCD is closely related to the elevation, slope, CTI, and NDVI. Validation results showed that the RT method is the best for predicting the spatial patterns of SOCD in our study area. Realizations generated by SIS can represent the possible spatial distribution patterns of SOC density without a smoothing effect. Once a threshold value of SOC density is given, SIS can quantitatively assess both local uncertainty and spatial uncertainty of SOC densities larger than the selected threshold.

The poor prediction performance for MLR probably resulted from the relatively low resolutions of DEM and NDVI, which may have blurred the relationships between SOC, topographic factors and vegetation (NDVI). This is the reason that the residues of SOC were further predicted by ordinary kriging to obtain better prediction results. The RT method had the lowest RMSPE. In fact, the RT method divided the study area into smaller regions where the topography and vegetation (NDVI) are relatively homogenous. The different linear functions were applied to each small homogenous regions, thus the complex landscape was decomposed first then the SOC was predicted. The ANN-kriging method utilized a nonlinear function to separate the SOC trend and then predict the SOC residues using ordinary kriging. The results of the RT method are easy to interpret whereas the results of neural network is more difficult.

Soil sampling and measurements for mapping the spatial patterns of SOC are costly and time-consuming, especially in large areas and over time. Spatially correlated secondary information, e.g. DEMs and satellite data provide indirect information on spatial distribution of many soil properties (see Section 20.2). Hence, the possibilities for future applications of the MLR, RK, RT, and ANN-kriging lie in their ability to integrate secondary information into spatial prediction, which may improve the mapping process, especially for areas with few measurements.

Acknowledgments Funding provided by the Natural Science Foundation of China (40601039), Knowledge Innovation Program of the Chinese Academy of Sciences (KZCX2-YW-409; KZCX2-YW-Q1-07).

References

Arrouays, D., Daroussin, J., Kicin, L., and Hassika, P., 1998. Improving topsoil carbon storage prediction using a digital elevation model in temperate forest soils of France. Soil Science 163:103–108.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford.

Chai, X., Shen, C., Yuan, X., and Huang, Y., 2008. Spatial prediction of soil organic matter in the presence of different external trends with REML-EBLUP. Geoderma 148:159–166.

Chen, F., Kissell, D.E., West, L.T., and Adkins, W., 2000. Field-scale mapping of surface soil organic carbon using remotely sensed imagery. Soil Science Society of America Journal 64:746–753.

Chen, F., West, L.T., Kissel, D.E., Clark R., and Adkins, W., 2008. Field-scale mapping of soil organic carbon with soil-landscape modeling, pp. 294–301. Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences Shanghai, P.R. China, June 25–27, 2008.

Cheng, X.F., Shi, X.Z., Yu, D.S., Pan, X.Z., Wang, H.J., and Sun, W.X., 2004. Using GIS spatial distribution to predict soil organic carbon in subtropical China. Pedosphere 14(4):425–431.

Deutsch, C.V., and Journel, A.G., 1998. GSLIB, Geostatistical Software Library and User's Guide. Oxford University Press, New York, NY.

Ding, D.Z., 1992. Soil Series of Hebei (In Chinese). Hebei Sci. and Technol. Press, Shijiazhuang, China.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York, NY.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

Mueller, T.G., and Pierce, F.J., 2003. Soil carbon maps: enhancing spatial estimates with simple terrain attributes at multiple scales. Soil Science Society of America Journal 67:258–267.

Park, S.J., Hwang C.S., and Vlek, P.L.G., 2005. Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. Agricultural Systems 85:59–81.

Ping, J.L., and Dobermann, A., 2006. Variation in the precision of soil organic carbon maps due to different laboratory and spatial prediction methods. Soil Science 171(5):374–387.

Simbahan, G.C., Dobermann, A., Goovaerts, P., Ping, J., and Haddix, M.L., 2006. Fine-resolution mapping of soil organic carbon based on multivariate secondary data. Geoderma 132:471–489.

Somaratne, S., Seneviratne, G., and Coomaraswamy, U., 2005. Prediction of soil organic carbon across different land-use patterns: a neural network approach. Soil Science Society of America Journal 69:1580–1589.

Terra, J.A., Shaw, J.N., Reeves, D.W., Raper, R., van Santen, E., and Mask, P.L., 2004. Soil carbon relationships with terrain attributes, electrical conductivity, and a soil survey in a coastal plain landscape. Soil Science 169:819–831.

Thompson, J.A., and Kolka, R.K., 2005. Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modeling. Soil Science Society of America Journal 69(4):1086–1093.

Venteris, E.R., and Slater, B. K., 2000. Spatial modeling of soil organic carbon by environmental correlation, Coshocton, Ohio. 4th International Conference on Integrating GIS and Environmental Modeling (GIS/EM4): Problems, Prospects and Research Needs. Banff, Alberta, Canada, September 2–8, 2000.

Wu, C., Wu, J., Luo, Y., Zhang, L., and DeGloria, S.D., 2008. Spatial prediction of soil organic matter content using cokriging with remotely sensed data. Soil Science Society of America Journal 73:1202–1208.

Zhao, Y.C., Shi, X.Z., Weindorf, D.C., Yu, D.S., Sun, W.X., and Wang, H.J., 2006. Map scale effects on soil organic carbon stock estimation in north China. Soil Science Society of America Journal 70:1377–1386.

Chapter 20
Estimating Soil Organic Matter Content by Regression Kriging

A. Marchetti, C. Piccini, R. Francaviglia, S. Santucci, and I. Chiuchiarelli

Abstract In Mediterranean countries soil organic matter (SOM) depletion is a key factor in land degradation. Here, climate (temperate winter/dry summer) and water scarcity give rise to faster mineralization rates and lower accumulation intensities, particularly in association with intensive and non-conservative agronomic practices.

The study area is located in central Italy, in the Soil Region 61.3 as defined by the European Soil Bureau, where soil erosion is the main cause of the low SOM content. In this area, about 250 georeferenced samples were collected from the surface horizon (plough layer) of agricultural soils. These samples have been analyzed for particle size distribution and soil organic carbon (SOC) content.

The use of regression kriging (RK) is proposed to predict SOC content and soil texture using the following attributes as predictors: (a) soil subsystems map (1:250,000) derived from pedological survey, (b) terrain parameters derived from DEM (elevation, slope, plan and profile curvature, TWI, incoming solar radiation), and (c) other indexes derived from Landsat TM imagery (e.g. NDVI, Grain Size Index, Clay Index).

Since the same level of SOM differently influences soil functions depending on soil texture, the values of SOM obtained from SOC were classified in four classes (very low, low, medium, high) based on the estimated USDA texture, and RK was applied. This map was compared with the soil subsystems map to evaluate the influence of the prevalent land use on SOM levels.

Keywords Soil organic matter · Land evaluation · Regression kriging · Italy

20.1 Introduction

In central and southern Italy the problem of soil organic matter (SOM) depletion is of particular concern due to the Mediterranean climatic conditions. In areas with Mediterranean climates, organic matter degradation is higher due to faster

A. Marchetti (✉)
CRA, Research Centre for the Soil-Plant System, Via della Navicella 2-4, 00184 Rome, Italy
e-mail: alessandro.marchetti@entecra.it

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_20,

mineralization rates and lower accumulation intensities resulting from often intensive and non-conservative agronomic practices. In these areas, deep tillage is often conducted to improve soil structure, permeability, and aeration, and to improve crop growth in clayey soils. These practices are especially common in the hills. This kind of soil management increases SOM degradation rates, due to higher aeration, and causes a dilution of SOM in the arable layer because of mixing with underlying horizons with lower SOM, thus leaving the soil susceptible to wind and water erosion (Bot and Benites, 2005).

When using SOM to evaluate "soil quality", we should consider that soil organic carbon (SOC, the carbon fraction of SOM) varies among environments and management systems, and generally increases: (a) with higher mean annual precipitation (Burke et al., 1989); (b) with lower mean annual temperature (Jenny, 1980); (c) with higher clay content (Nichols, 1984), (d) with an intermediate grazing intensity (Parton et al., 1987; Schnabel et al., 2001); (e) with higher crop residue inputs and cropping intensity (Franzluebbers et al., 1998), (f) with native vegetation compared with arable crops (Burke et al., 1989); and (g) with conservation tillage compared with conventional tillage (Rasmussen and Collins, 1991). Therefore, to prevent or at least to limit any negative effects that could arise from specific land uses, the knowledge of the current state of SOM is essential for site-specific crop management and for environmental modeling and planning.

In soil mapping, one of the primary difficulties is the lack of detailed soil surveys. As field data collection is often the most expensive part of a survey, usually only a limited number of locations is sampled, and point samples at local scale are more easily available. As a consequence, all soil information needs to be interpolated in space from a finite number of observations, inferring soil characteristics for the whole area of interest.

The OM content of the soil is strongly related to the land use, vegetation, climate and terrain features, which can be modelled with digital elevation model (DEM) and satellite data. The type and the amount of SOM are strongly related to the presence of water and the lateral redistribution of the surface material by erosion. Both of these phenomena are partially controlled by the terrain parameters (Dobos et al., 2006).

It is well established that kriging procedures based on auxiliary information usually result in maps with high accuracy, provided that the primary and secondary variables are well correlated (McBratney et al., 2003). The use of environmental covariates have improved several aspects of soil surveying in many parts of the world (see Chapter 2). Regression kriging (RK) is a spatial interpolation technique that combines a regression of the dependent variable on auxiliary variables with simple kriging of the regression residuals (Goovaerts, 1997; Hengl et al., 2007; Odeh et al., 1994). RK method generally produces realistic spatial representations, as the smoothing effect is much smaller than other interpolation methods (see Section 19.3). The aim of this work is to provide an evaluation of SOM from point-type data, estimating values in non-sampled locations by means of RK, and finally transferring them into GIS software to obtain a reliable framework and a valid tool for decision makers.

20.2 Materials and Methods

20.2.1 Study Area

The surveyed area (about 100 km^2), located in central Italy, comprises four municipalities in the Teramo province (Controguerra, Corropoli, Colonella and Martinsicuro). The most economically important zone falls in the Controguerra viticultural district (DOCG[1] area of Colline Teramane). The area is part of Soil Region 61.3 as defined by the European Soil Bureau (European Soil Bureau, 2001): Hills of Central and Southern Italy on Pliocene and Pleistocene marine deposits and Holocene alluvial sediments along the Adriatic Sea. In this framework, soil regions – areas with similar soil-forming conditions and as the largest units of soil description – are defined as typical associations of dominant soils occurring in areas which are limited by a special climate and/or a special association of parent material.

The main soils of the Soil Region 61.3 are: eroded soils with reorganization of carbonates (Eutric and Calcaric Regosols, Calcaric Cambisols, Haplic Calcisols); soils with clay accumulation (Haplic and Calcic Luvisols); soils with vertic properties (Vertic Cambisols and Calcic Vertisols); and alluvial soils (Calcaric, Eutric, and Gleyic Fluvisols). Climate is Mediterranean and Mediterranean suboceanic. Mean annual air temperature is between 12.5 and 16°C and mean annual precipitation is between 700 and 1,000 mm. The rainiest month is November and the driest months are July and August. The soil moisture regime is xeric, locally udic, and the soil temperature regime is thermic (Costantini et al., 2004).

Severe and continuous soil erosion is one of the major causes of the low organic matter content of many soils of this region, and causes the deterioration of an attractive traditional landscape, as well as of the agricultural value of the soils.

20.2.2 Field Data

The dataset used in this work is made up of 250 georeferenced samples collected by the Regional Agency for Agricultural Extension Services of Abruzzo Region (ARSSA) from the surface horizon of agricultural soils (plough layer, from the surface to about 50 cm in depth), in accessible agricultural lands. Sample collection was performed during spring and summer 2006.

The physical and chemical routine analyses included the determination of the SOC content according to the modified Walkley-Black method (Nelson and Sommer, 1982). SOM was evaluated from SOC (Jackson, 1965) by means of the formula:

$$\mathrm{SOM} = \mathrm{SOC} \times 1.724 \tag{20.1}$$

A descriptive statistical analysis was performed on collected data, determining also data distribution and the correlation matrix for the studied parameters.

[1] Denominazione di Origine Controllata e Garantita, attesting the origin and the quality of a wine

20.2.3 Spatial Analysis and Estimation

Auxiliary data for this area were derived from a DEM with a 40 m resolution provided by the Land Information Service of the Abruzzo Region, from Landsat 7 TM images (3 visibile bands and 4 infrared bands), and from the 1:250,000 Soil Subsystems Map of Abruzzo compiled by ARSSA (Chiuchiarelli et al., 2006).

From the DEM, the following morphometric parameters were derived (described in Section 22.2): elevation; slope gradient; curvature plan and profile (PLANC and PROFC); solar radiation; and Topographic Wetness Index (TWI). TWI is a parameter correlating topography and the water movement in slopes, used to display the spatial distribution of soil moisture and the shallow saturation degree: $\mathrm{TWI} = \ln(A_s/\tan\beta)$, where A_s is specific catchment area and β is slope (Beven and Kirkby, 1979).

From Landsat 7 TM images (July 2007, cloud cover 0%), the following indexes were derived: Grain Size Index (GSI); Clay Index (CI); and Normalized Difference Vegetation Index (NDVI). GSI is correlated with the fine sand content of the soil and is computed with the equation:

$$\mathrm{GSI} = (\mathrm{R} - \mathrm{B})/(\mathrm{R} + \mathrm{G} + \mathrm{B}) \tag{20.2}$$

where R is Red, B is Blue, and G is Green (Xiao et al., 2006). CI is correlated with the clay content of the soil and is computed with the equation:

$$\mathrm{CI} = \mathrm{MIR}/\mathrm{MIR2} \tag{20.3}$$

where MIR is Mid Infra Red (band 5) and MIR2 is Mid Infra Red (band 7) (Hengl, 2007). NDVI provides a quantitative and qualitative estimation of the vegetation and is computed with the equation:

$$\mathrm{NDVI} = (\mathrm{NIR} - \mathrm{R})/(\mathrm{NIR} + \mathrm{R}) \tag{20.4}$$

where NIR is Near Infra Red and R is Red (Colwell, 1974).

The last index, SST86, was derived from the 1:250,000 Soil Subsystems Map of Teramo Province compiled by ARSSA (Chiuchiarelli et al., 2006), reported in Fig. 20.1.

According to this map, eight different soil subsystems (Table 20.1) were defined. These soil subsystems have been grouped in a single index, obtained by converting the map into a raster.

The computational procedure we used to estimate SOC with RK is as follows:

1. set up and import predictor data layers (land-surface parameters and soil subsystems map);
2. match soil samples with land-surface parameters and setup the regression matrix; to avoid multicolinearity effects, instead of using directly the derived indexes as predictors, a Principal Component Analysis (PCA) was performed, since

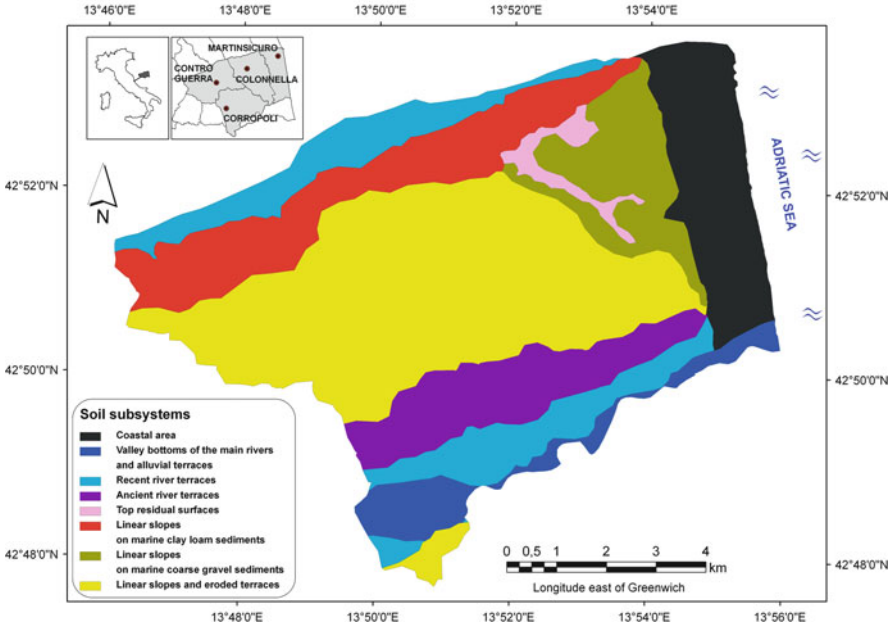


Fig. 20.1 Soil subsystems map of Teramo Province 1:250,000 (Chiuchiarelli et al., 2006)

principal components are orthogonal and independent. In the present study, 10 different components related to the 10 indexes were defined;

3. linear regression analysis and derivation of the regression residuals; regression coefficients were resolved by means of a maximun likelihood algorithm (Bailey et al., 2003);
4. analysis of residuals for spatial autocorrelation and fitting of the variogram model;
5. run the interpolation;
6. visualization and validation of the results using control points.

The derivation of indexes and the PCA was performed in SAGA 2.0.3 (SAGA User Group Association, 2008) and ILWIS 3.4 Open (52 North Initiative, 2007). For the regression analysis, the statistical software R 2.8.0 together with its main packages (i.e. *sp* and *rgdal* for spatial data preparation, and *gstat* for geostatistical modeling and prediction) was used (R Development Core Team, 2008).

The whole dataset was not used for the estimation, but it has been divided randomly in two numerically equivalent parts, the *training* part and the *test* part, so that a quota of the measured points could be used to validate the estimation results. For this operation, the Geostatistical Analyst extension of the software ArcGIS 9.2® was used. The goodness of the statistical estimation model is expressed by the root mean square prediction error (RMSE).

Table 20.1 Characteristics of the soil subsystems (Chiuchiarelli et al., 2006)

Map unit	Physiography and lithology	Morphometry	Land use and vegetation
A1a	Coastal area with incoherent substrates, sandy and sandy-gravelly.	Elevation: 0–20 m a.s.l. (99%). Slope: 85% from 0 to 5%.	Artificial surfaces: 59%. Agricultural areas: 28% (arable crops 16%, permanent crops 4%, heterogeneous areas 7%). Sea shores, dunes and sand plains: 3%
A2a	Valley bottoms of the main rivers, and alluvial terraces. Substrates of gravelly, sandy and clay loam alluvial sediments.	Elevation: 0–200 m a.s.l. (98%). Slope: 88% from 0 to 5%	Artificial surfaces: 17%. Agricultural areas: 64% (arable crops 38%, permanent crops 12%, heterogeneous areas 7%). Riparian formations: 13%
A2c	Recent river terraces higher than the present valley bottom. Substrates of gravelly-sandy and clay loam sediments, intercalated or underlying clay loam colluvial sediments.	Elevation: 0–200 m a.s.l. (91%). Slope: 80% from 0 to 5%.	Artificial surfaces: 28%. Agricultural areas: 64% (arable crops 40%, permanent crops 13%, heterogeneous areas 14%)
A2d	Ancient river terraces higher than the present valley bottom. Substrates of gravelly-sandy sediments.	Elevation: 20–300 m a.s.l. (82%). Slope: 83% from 0 to 13%.	Artificial surfaces: 10%. Agricultural areas: 83% (arable crops 36% and olive groves 22%)
A3b	Top residual surfaces, narrow and stretched, reduced to sub-level crests. Substrates of gravelly-sandy sediments.	Elevation: 100–400 m a.s.l. (95%). Slope: 67% from 0 to 13%.	Artificial surfaces: 16%. Agricultural areas: 85% (vineyards 25%, olive groves 20%, complex cultivation patterns 17%, arable crops 16%)
A4b	Linear slopes and secondarily unstable slopes with gully erosion. Substrates of marine clay loam sediments.	Elevation: 50–300 m a.s.l. (84%). Slope: 79% from 5 to 35%.	Agricultural areas: 79% (arable crops 29%, olive groves 22%, vineyards 11%)
A4d	Linear slopes and secondarily unstable slopes with gully erosion. Substrates of marine coarse gravel sediments.	Elevation: 100–300 m a.s.l. (85%). Slope: 76% from 5 to 35%.	Artificial surfaces: 6%. Agricultural areas: 78% (arable crops 40%, olive groves 12%, annual crops associated with permanent crops 12%)
A4e	Linear slopes and eroded surfaces of terraces with slightly undulating morphology.	Altitude: 50–300 m a.s.l. (97%). Slope: 98% from 0 to 35%.	Artificial surfaces: 5%. Agricultural areas: 91% (arable crops 47%, annual crops associated with permanent crops 19%)

Table 20.2 Soil organic matter evaluation (SILPA, 1999)

Texture			
		Loam	
		Sandy clay	Clay
	Sand	Sandy clay loam	Clay loam
Classes	Loamy sand	Silty loam	Silty clay
(USDA)	Sandy loam	Silt	Silty clay loam
Soil organic matter			
Level	%		
Very low	<0.8	<1.0	<1.2
Low	0.8–1.4	1.0–1.8	1.2–2.2
Medium	1.5–2.0	1.9– 2.5	2.3– 3.0
High	>2.0	> 2.5	> 3.0

Maps in GIS environment showing soil texture (USDA classification), SOC in kg m^{-2}, and four levels of SOM related to the three main classes of texture (SILPA, 1999) as specified in Table 20.2 were derived from the estimated values, by means of the software ArcGIS 9.2®.

20.3 Results and Discussion

20.3.1 Statistical Elaboration

In Table 20.3 basic statistics of the considered variables are reported. Frequency histograms were used to identify outliers for each parameter. These observations were removed, and the remaining data have frequency distributions that approach normal (Webster, 2001). New statistics are reported in Table 20.4.

A PCA was performed on the 10 indexes derived from auxiliary data (ELEVATION, SLOPE, PROFC, PLANC, TWI, SOLAR, GSI, CI, NDVI, SST86). In Table 20.5, the matrix of transformation coefficients, calculated from the covariance

Table 20.3 Basic statistics of sand, clay and SOC

	Count	Minimum	Maximum	Mean	Median	SD	Variance	Skewness	Kurtosis
Sand	250	0.00	66.80	20.52	19.30	11.32	128.1424	1.05	5.05
Clay	250	3.70	45.30	31.56	32.30	6.86	47.0596	−0.91	4.09
SOC	250	0.04	3.46	1.03	0.94	0.57	0.3249	1.77	7.48

Abbreviations: SD: standard deviation

Table 20.4 Basic statistics of sand, clay and SOC without outliers

	Count	Minimum	Maximum	Mean	Median	SD	Variance	Skewness	Kurtosis
Sand	241	0.0	46.2	19.23	19.2	9.2	85.32	0.287	3.207
Clay	241	15.0	45.3	32.07	32.5	6.1	37.25	0.546	3.003
SOC	238	0.04	2.07	0.94	0.91	0.39	0.62	0.22	1.77

Abbreviations: SD: standard deviation

Table 20.5 Principal components coefficients and significance for prediction

	Elevation	Slope	PLANC	PROFC	TWI	SOLAR	GSI	CI	NDVI	SST86	Sand	Clay	SOC
PC1	0.116	0.486	0.046	0.078	−0.674	0.114	0.157	−0.065	−0.002	0.496	*	***	
PC2	−0.054	0.309	0.050	0.071	−0.358	−0.055	0.099	−0.152	−0.201	−0.830			
PC3	0.131	−0.229	0.040	0.060	0.165	0.132	0.560	−0.604	−0.434	0.117	*		
PC4	0.568	0.034	−0.088	−0.083	0.088	0.567	0.212	0.047	0.489	−0.215	***	*	
PC5	−0.052	0.342	−0.479	−0.774	0.144	−0.052	0.035	−0.081	−0.142	0.030			
PC6	−0.027	−0.272	−0.857	0.367	−0.233	−0.027	0.014	0.006	0.016	−0.021			
PC7	0.020	0.644	−0.148	0.493	0.544	0.019	−0.099	−0.088	−0.058	0.054		*	
PC8	0.011	0.044	−0.010	0.048	0.085	0.013	0.535	0.762	−0.349	−0.010			**
PC9	0.379	−0.084	−0.007	−0.017	−0.065	0.371	−0.557	0.107	−0.621	0.010	*	*	*
PC10	−0.705	0.001	0.000	0.000	−0.001	0.709	−0.004	0.000	−0.003	0.000			
										R^2	0.83	0.96	0.88

Significance codes: *** = 0.001; ** = 0.01; * = 0.05

matrix, is reported. The PC10 component has been excluded a priori to avoid any rounding effect in PCA computation with ILWIS (Hengl, 2007). The coefficient matrix has shown that some components are mainly correlated with the DEM parameters (TWI and SLOPE), as they show the higher absolute values.

RK was applied to the sand and clay data to estimate soil texture, and to the SOC content of the training dataset. The correlation of the target variables with the principal components was very strong, as shown by R^2 values of the linear regression. The choice of predictors was performed by the step-wise regression, considering a significance level of 0.05. For sand estimation, PC4 component was chosen as predictor; for clay estimation, the sum of PC1 – PC4 – PC7 – PC9 components was used; all these components are linked to the soil type and position. For SOC estimation, the sum of PC8 and PC9 was used, essentially linked to grain size distribution of soil and to the elevation.

20.3.2 Estimated Maps

In Figs. 20.2 and 20.3, the estimated map of SOC and of USDA soil texture respectively are reported. USDA texture was estimated from sand and clay data using the training dataset. The boundaries of the soil subsystems are depicted on each map for an easier interpretation. It should be noted that the same SOC content and soil texture not necessarily correspond to the same soil type.

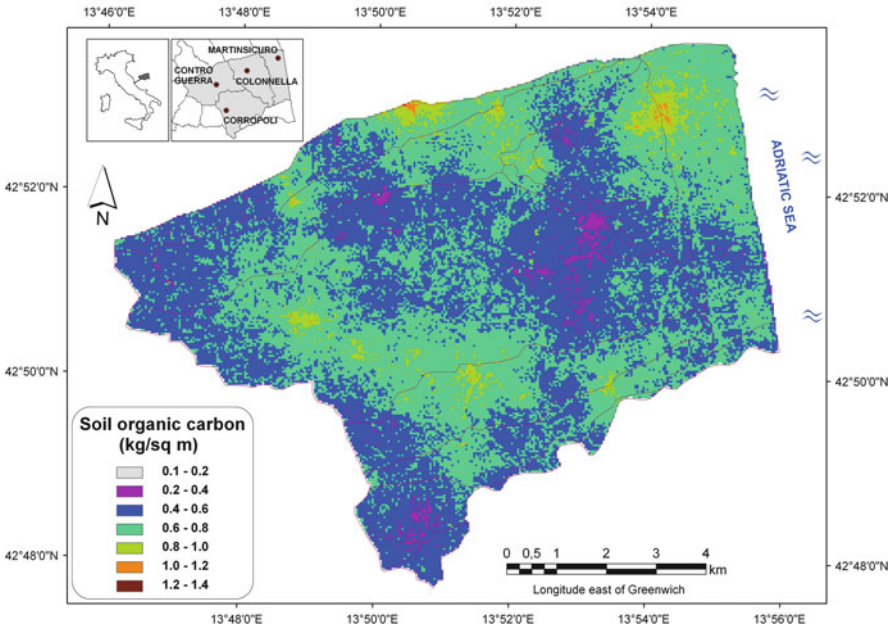


Fig. 20.2 Soil organic carbon content by RK

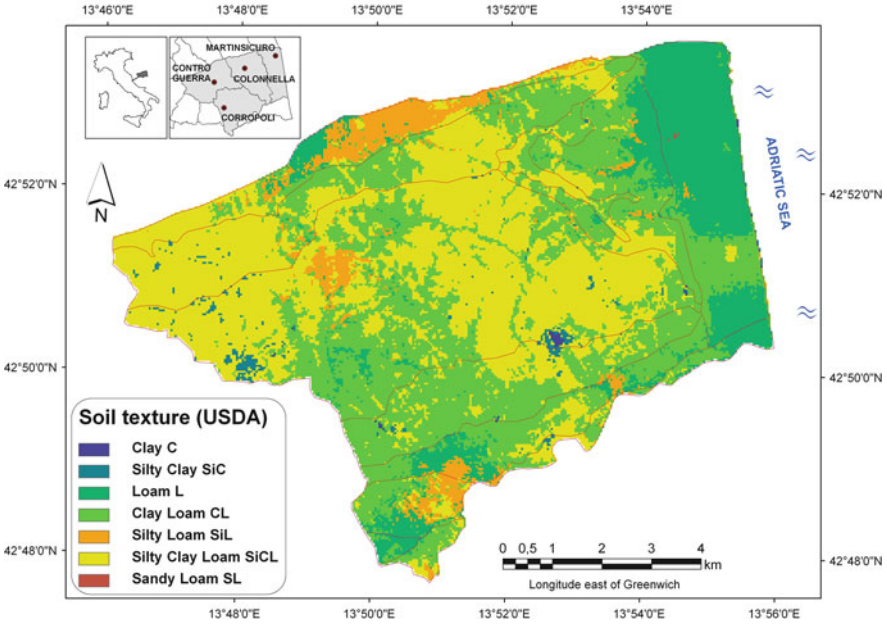


Fig. 20.3 Soil texture by RK (USDA)

Over 90% of the surveyed area has a SOC content between 0.4 and 0.8 kg m^{-2}, as indicated in Table 20.6. The prevailing texture classes in this area, as shown by the texture map and reported in Table 20.6, are clay loam and silty clay loam.

The precision of prediction can be assessed by comparing the values estimated from the training dataset with the correspondent values from the test dataset that were not used in the estimation process. The prediction errors for each parameter are reported in Table 20.7. For SOC, sand and clay, RMSE was 0.39, 8.96 and 5.44 respectively. These can be considered good results since they are lower than or close to the standard deviation of the data. Moreover, a prediction is good when RMNSE is close to one: this condition is fully verified for the three considered variables.

From an agronomic point of view, the climatic and especially the pedological context cannot be neglected in the evaluation of SOM, because the same level of

Table 20.6 Percentage distribution of classes in the surveyed area

Soil organic carbon (kg m^{-2})	% of the surface	Soil texture	% of the surface	SOM vs texture	% of the surface
0.1–0.2	0.005	Clay C	0.11	Very low	4.48
0.2–0.4	1.50	Silty clay SiC	0.98	Low	88.48
0.4–0.6	47.15	Loam L	14.95	Medium	6.95
0.6–0.8	47.45	Clay loam CL	38.22	High	0.09
0.8–1.0	3.76	Silty clay loam SiCL	40.13		
1.0–1.2	0.13	Silty loam SiL	5.60		
1.2–1.4	0.005	Sandy loam SL	0.01		

Table 20.7 Prediction errors

	Mean	SD	ME	RMSE	RMNSE
Sand (%)	21.32	11.40	−1.63	8.96	1.00
Clay (%)	31.39	7.23	0.39	5.44	1.01
SOC (kg m^{-2})	0.95	0.37	−0.01	0.39	1.10

Abbreviations: SD: standard deviation; ME: mean error; RMSE: root mean square error; RMNSE: root mean normalized square error

SOM can have different effects on soil functions in different soil types. Thus, from SOC values, SOM in g kg^{-1} was evaluated, and the values of SOM were classified in four classes (very low, low, medium, high) based on the estimated USDA texture, according to the scheme reported in Table 20.2. The obtained map is reported in Fig. 20.4.

The mean values of SOM calculated for each of the soil subsystems range from 1.43 to 1.70 g kg^{-1}. Subsequently, about 88 % of the area can be classified in the "low" SOM class in relation to soil texture (Table 20.6), regardless of the prevalent land use or the soil type.

The role of topography is very important (see Section 19.3): the low levels of SOM are strictly linked to the erosive processes active on hills (Fig. 20.5), and to the intensive agricultural practices adopted in these areas.

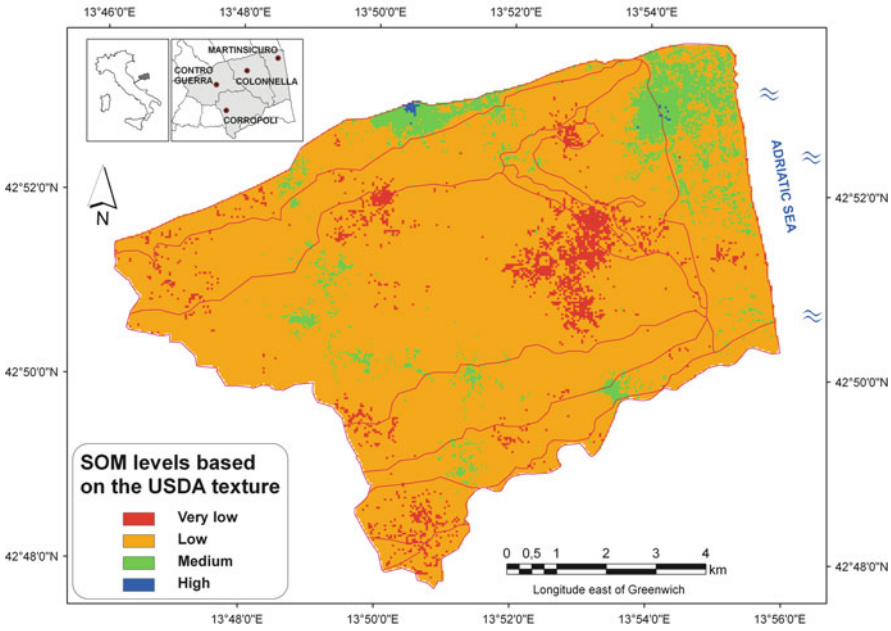


Fig. 20.4 SOM levels based on the USDA texture by RK

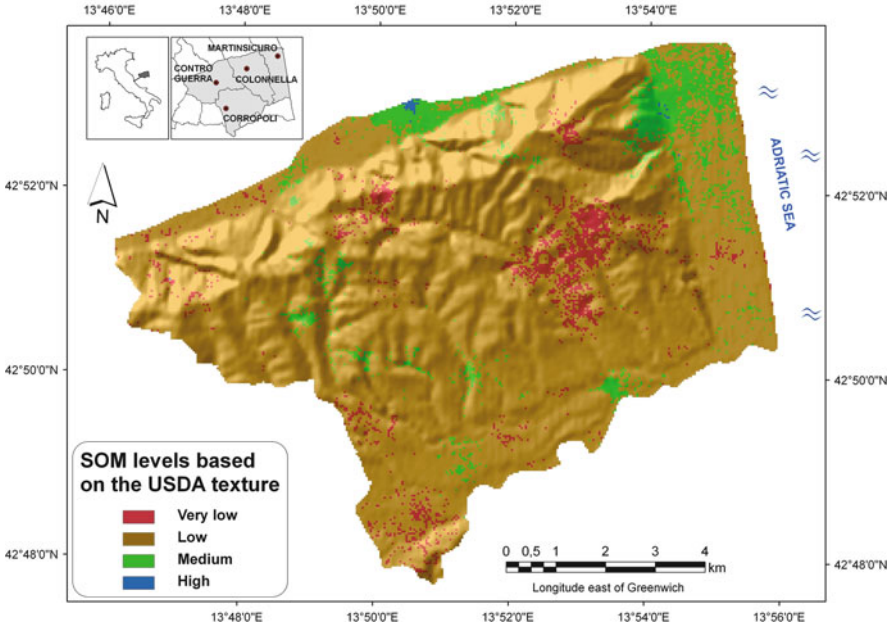


Fig. 20.5 SOM map draped over a hillshade of the area

20.4 Conclusions

In natural environments, SOM content is mainly controlled by climate, while in agricultural lands it is strongly affected by agronomic management. Thus, a spatial representation of SOM is essential to facilitate regional planning and to provide decision makers with a reference tool. In the Province of Teramo, the analysis of the available data, and the estimate in non-sampled locations by means of RK allowed us to create maps for SOC, soil texture, and SOM levels with a reasonable degree of accuracy.

RK produced a realistic spatial distribution of SOM, which showed a low content in relation with the texture in most of the surveyed area, regardless of the prevalent land use or the soil type. Here the adoption of measures to manage SOM content should be strongly recommended since the loss of SOM is often both a cause and a result of erosive processes on hills, and is usually coupled with the effects of intensive agricultural practices. Intervention in agronomical and environmental planning is required to provide soils with an appropriate OM content for the maintenance of their ecological and socio-economical functions and yield sustainability. Stimulating the adoption of good agronomical practices is also necessary to preserve soil resources, improve the SOM cycle, and significantly contribute to climate change mitigation through carbon sequestration via agricultural soils. Sound cropland management can play a positive role in reducing Green House Gas emissions, and carbon

dioxide emissions in particular, through a decrease of soil organic carbon losses, by increasing the organic matter input, or through a combination of both.

These results encourage to apply RK at a wider scale, and it is to be hoped that in the near future this method should be successfully integrated with traditional soil surveying procedures (see Section 28.1).

Acknowledgments This research was supported by the Italian Ministry of Agricultural, Food and Forestry Policies within the activities of the National Advisory Committee for Soil Quality.

The Authors thank the Service for Land Information and Telematics of Abruzzo Region for providing the DEM, and the staff of the Agrochemical Laboratory ARSSA of Avezzano for the valuable analytical support.

References

52 North Initiative, 2007. ILWIS 3.4 Open – Integrated Land and Water Information System. http://52north.org (last verified March 2009).

Bailey, M., Clements, T., Lee, J.T., and Thompson, S., 2003. Modelling soil series data to facilitate targeted habitat restoration: a polytomous logistic regression approach. Journal of Environmental Management 67(4):395–407.

Beven, K., and Kirkby, N., 1979. A physically based, variable contributing area model of basin hydrology. Hydrological Sciences Bulletin 24(1):43–69.

Bot, A., and Benites, J., 2005. The importance of soil organic matter. Key to drought-resistant soil and sustained food production. FAO Soil Bulletin 80, http://www.fao.org/docrep/009/a0100e/a0100e00.htm (last verified March 2009).

Burke, I.C., Yonker, C.M., Parton, W.J., Cole, C.V., Flach, K., and Schimel, D.S., 1989. Texture, climate, and cultivation effects on soil organic matter content in US grassland soils. Soil Science Society of America Journal 53:800–805.

Chiuchiarelli, I., Paolanti, M., Rivieccio, R., and Santucci, S., 2006. Suoli e Paesaggi d'Abruzzo – Carta dei Suoli della Regione Abruzzo. ARSSA Regione Abruzzo.

Colwell, B.J., 1974. Vegetation canopy reflectance. Remote Sensing of Environment 3:175–183.

Costantini, E.A.C., Urbano, F., and L'Abate, G., 2004. Soil regions of Italy. http://www.soilmaps.it (last verified March 2009).

Dobos, E., Carré, F., Hengl, T., Reuter, H.I., and Tóth, G., 2006. Digital soil mapping as a support to production of functional maps. EUR 22123 EN, Office for Official Publications of the European Communities, Luxemburg.

European Soil Bureau, 2001. Georeferenced soil database for Europe. Manual of procedures, Version 1.1. EUR 18092 EN, Office for Official Publications of the European Communities, Luxemburg, 178 pp.

Franzluebbers, A.J., Hons, F.M., and Zuberer, D.A., 1998. In situ and potential CO_2 evolution from a Fluventic Ustochrept in southcentral Texas as affected by tillage and cropping intensity. Soil Tillage Research 47:303–308.

Goovaerts P., 1997. Geostatistics for Natural Resource Evaluation. Oxford University Press, New York, NY.

Hengl, T., 2007. A practical guide to geostatistical mapping of environmental variables, http://bookshop.europa.eu/uri?target=EUB:NOTICE:LBNA22904:EN:HTML (last verified March 2009).

Hengl, T., Heuvelink, G.B.M., and Rossiter, D.G., 2007. About regression kriging: from equations to case studies. Computers & Geosciences 33(10):1301–1315.

Jackson, M.L., 1965. Soil Chemical Analysis. Prentice-Hall, Englewood Cliffs, NJ.

Jenny, H., 1980. The Soil Resource: Origin and Behavior. Ecological Studies, vol. 37. Springer, New York, NY, 377 pp.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

Nelson, D.W., and Sommer, L.E., 1982. Total carbon, organic carbon, and organic matter, pp. 539–579. In: Page, A.L. (ed.), Methods of Soil Analysis, 2nd ed. ASA Monograph 9(2) American Society of Agronomy, Madison, WI, USA.

Nichols, J.D., 1984. Relation of organic carbon to soil properties and climate in the southern Great Plains. Soil Science Society of America Journal 48:1382–1384.

Odeh I.O.A., McBratney A.B., and Chittleborough D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63:197–214.

Parton, W.J., Schimel, D.S., Cole, C.V., and Ojima, D.S., 1987. Analysis of factors controlling soil organic matter levels on grasslands. Soil Science Society of America Journal 51:1173–1179.

R Development Core Team, 2008. R: a language and environment for statistical computing. Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org (last verified March 2009).

Rasmussen, P.E., and Collins, H.P., 1991. Long-term impacts of tillage, fertilizer, and crop residue on soil organic matter in temperate semiarid regions. Advances in Agronomy 45:93–134.

SAGA User Group Association, 2008. SAGA 2.0.3 – System for Automated Geoscientific Analyses, http://www.saga-gis.org (last verified March 2009).

Schnabel, R.R., Franzluebbers, A.J., Stout, W.L., Sanderson, M.A., and Stuedemann, J.A., 2001. The effects of pasture management practices, pp. 291–322. In: Follett, R.F., Kimble, J.M., and Lal, R. (eds.), The Potential of US Grazing Lands to Sequester Carbon and Mitigate the Greenhouse Effect. Lewis Publishers, Boca Raton, FL, USA.

SILPA, Società Italiana dei Laboratori Pubblici di Agrochimica, 1999. Dall'analisi del terreno al consiglio di concimazione. ASSAM Regione Marche, Italy.

Webster, R., 2001. Statistics to support soil research and their presentation. European Journal of Soil Science 52:331–340.

Xiao, J., Shen, Y., Tateishi, R., and Bayaer, W., 2006. Development of topsoil grain size index for monitoring desertification in arid land using remote sensing. International Journal of Remote Sensing 12(27):2411–2422.

Chapter 21
Digital Soil Mapping of Topsoil Organic Carbon Content of Rio de Janeiro State, Brazil

M.L. Mendonça-Santos, R.O. Dart, H.G. Santos, M.R. Coelho, R.L.L. Berbara, and J.F. Lumbreras

Abstract A database with 431 soil profiles of Rio de Janeiro State was used in a research project entitled "Quantifying the magnitude, spatial distribution and organic carbon in soils of Rio de Janeiro State, using quantitative modeling, GIS and database technologies" (Projeto Carbono_RJ, sponsored by FAPERJ – Carlos Chagas Filho Foundation for Research Support in Rio de Janeiro State). These soil data were collected for other purposes and there were only limited soil bulk density data (103), which is needed for estimating soil organic carbon (SOC) stocks. Pedotransfer functions (PTFs) were estimated to be used in the modeling of organic soil carbon of topsoil (0–10 cm), using the *scorpan* model. The following environmental correlates were used as predictor variables: satellite data (Landsat ETM$^+$), lithology and soil maps, and a DEM and its derivatives. This dataset represents the best organized soil dataset in Brazil and is working as an educational trial for Digital Soil Mapping using a variety of methods for predicting soil classes and their properties. Multilinear analysis and regression-kriging were used to perform the modeling. Seven different models were built and compared through statistical methods. The main difference between the models was the set of predictor variables used to perform them. In general, all models performed well to predict the SOC stock. Nevertheless, model 6 was considered the best one since it presented the smallest AIC and RMSE as it used existing soil information (polygon soil maps) as a predictor variable, in addition to the variables used in the other models. The results obtained with this model were used to map topsoil carbon stock at a spatial resolution of 90 m.

Keywords Digital soil mapping · Soil · Soil organic carbon · Pedotransfer functions · *Scorpan* model · Regression-kriging

M.L. Mendonça-Santos (✉)
EMBRAPA Solos – Brazilian Agricultural Research Corporation, The National Centre of Soil Research, Rua Jardim Botânico, 1.024, 22.460-000, Rio de Janeiro, RJ, Brazil
e-mail: loumendonca@cnps.embrapa.br

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_21,

21.1 Introduction

Estimates of organic carbon stock in soils is important for a range of issues including climatic change, soil fertility and soil water storage capacity (see also Section 22.1). According to Batjes and Sombroek (1997), the soils of the world are one of the five main reservoirs of carbon, together with oceans, lithosphere, atmosphere and terrestrial biomass. Soils are essential for carbon sequestration representing approximately 75% of the carbon accumulation in the terrestrial ecosystem. The dynamics of carbon sequestration depend on many variables based on thermodynamic elements, characteristics of biomes and the responses to different land uses and management systems (Batjes, 1998). The soil works as source and reservoir (or sink) of carbon, depending on the relative rates of incorporation and decomposition of carbon by soil organisms. In order to estimate the net flow of carbon in terrestrial ecosystems, an understanding of the processes of soil formation and the spatial variability of organic carbon in the landscape is needed. Spatial variability data are important for estimating the stock of soil carbon and also for understanding the biophysical processes that can influence the flow of organic carbon in soils. The patterns and processes vary considerably in the landscape which limits extrapolation of point data as discussed in Chapter 26. Therefore, specific regionalized studies are important to assure a proper scale of study, as well as establish rules for the extrapolation of results and detailed understanding of soil carbon dynamics.

For prediction and mapping of the carbon stocks we used digital quantitative techniques named Digital Soil Mapping, defined by Lagacherie and McBratney (2007) as "the creation and population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge and from related environmental variables". The main use of the this approach is to replace the polygon-based soil maps of the past with digital maps of soil classes and properties and their associated uncertainties for areas previously mapped, or for new areas. These maps are stored and manipulated in a GIS environment, creating the possibility of vast arrays of data for analysis and interpretation.

Predictions of soil classes and properties in digital mapping are based on relationships among soils and the factors and processes of soil formation that enter in the equations as predictor variables. The logic of this reasoning is based on the equation of Jenny (1941) formulated from the recognition of the factors of soil formation, in a more quantitative formulation,

$$S = f(cl, o, r, p, t)$$

Where, S represents the soil, cl = climate, o = organisms, including human-induced activities, r = relief, p = parent material and t = time.

McBratney et al. (2003) generalized and formulated a new equation, with the objective of modeling the variables responsible for the processes of soil formation, through an empiric quantitative description of the relationships among other spatially geo-referenced factors (environmental covariates), used as spatial prediction

functions (see Chapter 2 and Section 16.1). This is the *scorpan* function with seven factors: s = soil and other properties of the soil in a given location; c = climate, climatic properties; o = organisms, vegetation or fauna or human activities; r = topography, attributes of the landscape; p = parent material, lithology; a = age, time factor; n = space, spatial location. Each factor is represented by a group of one or more continuous or categorical variables; for example, **r** for elevation, slope or other derived attribute of a DEM. The sources of data, the methods to estimate f from the equation, as well as the steps to execute the *scorpan* are presented and discussed in McBratney et al. (2003).

In this study, digital mapping techniques were used to predict the stocks of organic carbon of the topsoil (0–10 cm) in the Rio de Janeiro State. A multilinear analysis was used as the predictive model and several environmental variables as predictors. Seven different models were built and statistically compared. The choice of the best model was based on Akaike's Information Criterion (AIC) (Akaike, 1973), a quality index that represents a balance among the goodness of fit and the parsimony of the model. The best model is that with the smallest AIC. This model (M6) was then applied to the digital mapping of the soil carbon stock.

21.2 Materials and Methods

21.2.1 Study Area

The study area is the State of Rio de Janeiro located between the geographical coordinates 41° and 45° W and 20°30′ and 23°30′ S and is about 44,000 km^2 (Fig. 21.1). The area is characterized by eight large landscape types known as Serra da Bocaina, Coastal Plains, Mountainous Area, North-Northwest Fluminense, Paraíba do Sul River (Middle Valley), Serra Mantiqueira, Serra dos Órgãos, and Upper Itabapoana River (Plateau), described in Mendonça-Santos et al. (2008), where soil profiles have been studied in order to characterizes the soil organic carbon.

21.2.2 Digital and Field Data

The soil database that has been used to estimate soil classes in Rio de Janeiro State by Mendonça-Santos et al. (2008), was used in this study, to which we added 16 soil profiles from the RJ_Soil Carbon Project (Mendonça-Santos et al., 2005), summing 431 soil profiles for the whole state. The spatial distribution of soil profiles is shown in Fig. 21.1. Considering that these soil data were collected for other purposes, there were only scarce soil bulk density data, which is essential to estimate soil carbon stocks. Pedotransfer functions (PTFs) for the upland and lowland soils were used as auxiliary information for soil organic carbon estimation. The following covariates were used as predictor variables: Geocover^{TM} mosaic (bands 7, 4 and 2 in RGB), from NASA (2008); the NDVI (using band 2 instead of 3); Land

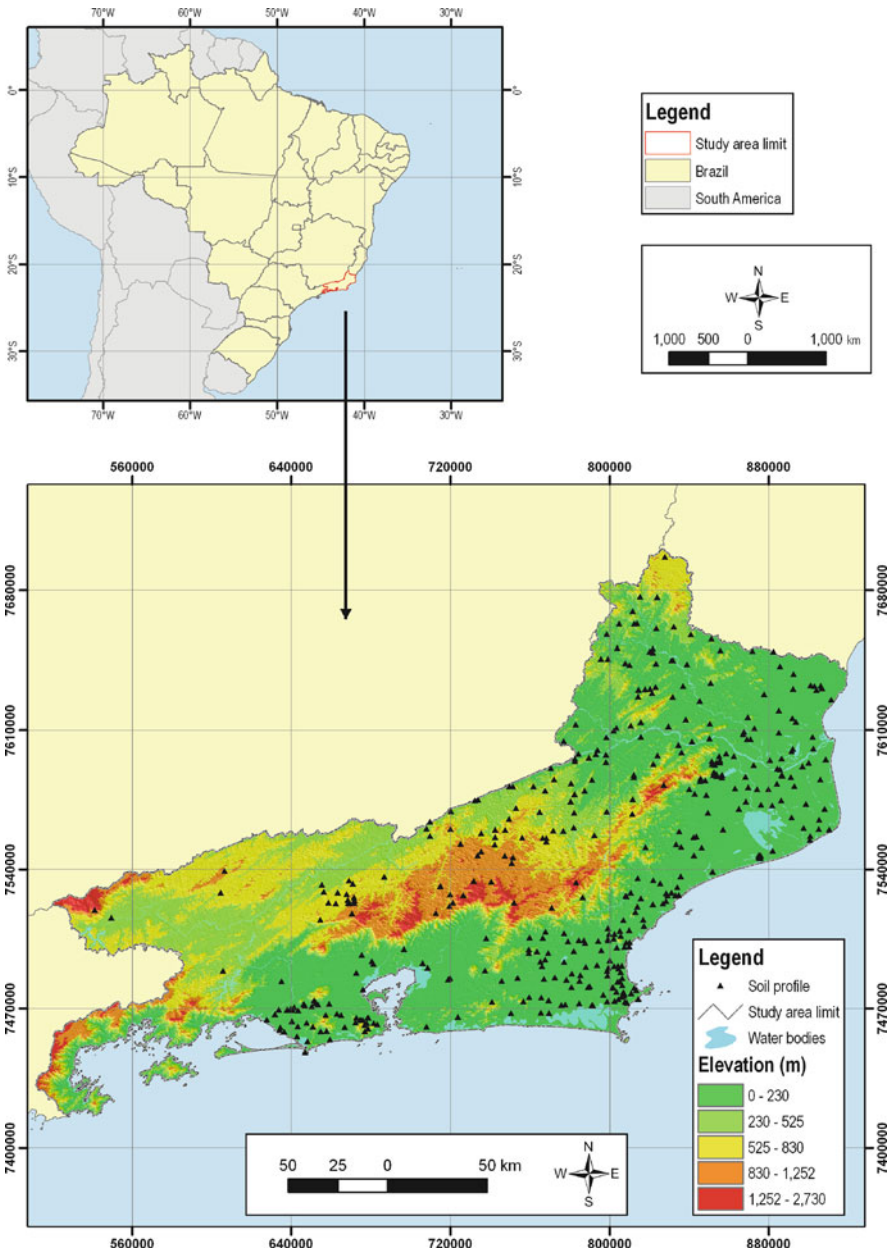


Fig. 21.1 The study area location and the soil profile distribution on the elevation map, extracted from the SRTM DEM (Jarvis et al., 2006) at 90 m pixel resolution

Use/Land Cover (LULC) map of Rio de Janeiro State, produced by Mendonça-Santos et al. (2003); Lithology class map (Rio de Janeiro, 2001) and the SRTM DEM 90 m, obtained from the CGIAR database at http://srtm.csi.cgiar.org (Jarvis et al., 2006) and modified by Mendonça-Santos et al. (2008) and its derivates extracted using the LandMapR software (MacMillan, 2003).

The soil dataset was complemented with the covariates of environmental factors for each soil data point. An ancillary dataset representing the whole study area was interpolated on a 90-m grid corresponding to the SRTM DEM, and populated with environmental and soil variables. Exploratory statistical analysis was performed on soil data. The modelling and prediction of soil carbon was performed using multilinear regression and regression-kriging. The output results were imported and mapped in a GIS environment.

21.2.3 Inference Models

Soil organic carbon stocks were calculated in mass per unit of volume, as follows:

$$StockC = C \times d \times p,$$

where: C is the carbon content (g/kg), d is soil bulk density (g/cm^3) and p is depth (cm).

Seven models of multilinear analysis were tested. The models were differentiated by the number of predictor variables used, the use or not of the stepwise procedure and the number of observations (profiles) used in the adjustment of each model. The performance of the models was statistically estimated using both RMSE (estimate of the standard deviation of the residual error) and the AIC (Akaike's Information Criterion), which is an index that considers the number of parameters used in the model. This index represents a commitment between the adjustment and the parsimony of the model. The model that presents the smallest AIC is the best. The AIC was calculated in agreement with Akaike (1973):

$$AIC = -2\, log\, like + 2m,$$

where *log like* is the logarithm of the prediction, and m it is the number of parameters used in the model.

For continuous variables, as in the case of soil carbon stock, AIC was calculated:

$$\mathrm{AIC} = N\, \ln\left(\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2\right) + 2\,m$$

Where N is the total number of soil profiles that were used in the model.

The model with the smaller AIC value was used for the final prediction of the soil organic carbon stock. The residues of the model were kriged and added to the predicted values (regression-kriging).

21.3 Results and Discussion

21.3.1 Results from the Models

Developed PTFs, together with the predictor variables, are shown in Table 21.1. Given the difference between carbon contents and soil texture of the soils in the lowland and the others soils (here denominated as mineral soils), it was necessary to build 2 PTFs. The PTFs were applied to estimate soil density and later on, to calculate soil organic carbon stocks.

The soil-landscape model *scorpan* was applied using the soil and landscape information. In Table 21.2 the different models are given with their details. Model M1 encompasses the extracted relief variables in LandMapR. In model M2 the same relief derivatives were used, but in this model, a stepwise procedure was undertaken in order to find which variables have larger correlation with the soil organic carbon stocks. In the models M3 and M4 the relief variables and the Geocover mosaic were used (Landsat 7 ETM+ with the bands 7-4-2, NDVI), with the difference that in model M4 a stepwise procedure was applied. The stepwise in model M4 did not allow the entrance of NDVI in the model. The variable NDVI excluded two profiles out of the model (in these two profiles the reflectance in the bands 4 and 2 was 0). In model M5, in addition to the terrain variables and the Geocover, the lithology map was also used. Model M6 encompasses all the variables of model M5 and a polygon soil map at a scale of 1:250,000. In the model M7 the variables of the model M5 were used, in addition to the LULC map.

21.3.2 Assessment of the Models Accuracy

The result of the carbon stock prediction, the performance of the indices AIC and RMSE (estimate of the standard deviation of the residual error) and the number of parameters (variables) used in each tested model is shown in Table 21.3.

Generally, all seven models presented a fair prediction of the carbon stocks considering that the differences among the indices of the seven models did not present a significant variation. The best result for modeling carbon stocks was model M6 because it had the smallest AIC and RMSE values (Table 21.3). Figure 21.2 shows the result produced by the best model (M6) for the stock of organic carbon (0–10 cm) of Rio de Janeiro State, including the estimates (predicted values), the residuals and the final result obtained by the sum of the kriged residues with the values predicted by multilinear regression (regression-kriging). The final map (Fig. 21.3), obtained by modeling and digital mapping, allows the modeled property

Table 21.1 Pedotransfer functions (PTFs) for soil bulk density, estimated from soil organic carbon content (%) and soil particles sizes (sand, silt and clay in %)

Layers	Predictors	Number of parameters	R^2	N	Average	Standard deviation	Max	Min	PTF_bulk density
0–10 Lowland	C content, fine sand, thick sand, silt, clay	5	0.947	131	1.10	0.20	1.09	1.12	0 + −0.00620188898666584∗ (“C cont_0–10”) + 0.00168408543309972∗ (“Thick sand 0–10”) + 0.00112348306614799∗ (“Fine sand 0–10”) + 0.00155733827838204∗ (“Silt 0–10”) + 0.00105136516518877∗ (“Clay 0–10”)
0–10 Upland soils minerals	C content, fine sand, thick sand, silt, clay	5	0.369	594	1.33	0.11	1.32	1.34	0 + −0.0104432003873496∗ (“C content_0–10”) + 0.00148435216740785∗ (“Thick sand 0–10”) + 0.00155020523466563∗ (“Fine sand 0–10”) + 0.00165886998238256∗ (“Silt 0–10”) + 0.00133985448863544∗ (“Clay 0–10”)

Table 21.2 Predictive models *scorpan* used to estimate the soil carbon stocks in topsoil (0–10 cm)

Models	Predictors variables – *SCORPAN model*	Stepwise	Number of soil profiles
M1	***R*** (ELEV, ASPECT, PLAN, PROF, QWETI, SLOPE)	____	429
M2	***R*** (ELEV, ASPECT, PLAN, PROF, QWETI, SLOPE)	ELEV, ASPECT, PLAN, QWETI, SLOPE	429
M3	***O*** (Landsat ETM^{+} -B7, B4, B2 e NDVI), ***R*** (ELEV, ASPECT, PLAN, PROF, QWETI, SLOPE)	____	427
M4	***O*** (Landsat ETM^{+}-B7, B4, B2 e NDVI), ***R*** (ELEV, ASPECT, PLAN, PROF, QWETI, SLOPE)	B7, B4, ELEV, ASPECT, PLAN, QWETI, SLOPE	429
M5	***O*** (Landsat ETM^{+}-B7, B4, B2 e NDVI), ***R*** (ELEV, ASPECT, PLAN, PROF, QWETI, SLOPE), ***P*** (Litology Map – vector format)	____	427
M6	***S*** (Soil Map – polygon), ***O*** (Landsat ETM^{+}-B7, B4, B2 e NDVI), ***R*** (ELEV, ASPECT, PLAN, PROF, QWETI, SLOPE), ***P*** (Litology Map – vector format)	____	427
M7	***O*** (Landsat ETM^{+}-B7, B4, B2, NDVI and LULC Map), ***R*** (ELEV, ASPECT, PLAN, PROF, QWETI, SLOPE), ***P*** (Litology Map – vector format)	____	427

to be spatially viewed in a continuous way in the grid determined by the availability of data and objectives of the work, facilitating observation of the variation and distribution of the stock of carbon in the landscape (see Fig. 20.5 for similar results).

Values of organic carbon stocks range from less than 3 to 70 Mg per hectare for the topsoil (0–10 cm). The variations had a strong correlation with the soil type and its position in the landscape (see also Chapter 22). The correlation with the landscape was clearer when analyzed in relation to the geoenvironments defined for Rio de Janeiro (Lumbreras et al. 2003) (Fig. 21.3). This map shows that unit II (Costal Plains) presents a higher stock of organic carbon in the lowland environments, such as mangroves (surroundings of Guanabara Bay, Sepetiba Bay, Guaratiba, among

Table 21.3 Comparison of the performance of the models

Models	RMSE	Number of parameters	AIC
M1	14.26907	6	2,286.603
M2	14.25226	5	2,284.604
M3	14.15396	10	2,279.976
M4	14.08333	7	2,267.386
M5	13.09959	19	2,215.548
M6	*11.95091*	*29*	*2,146.578*
M7	12.68869	28	2,196.807

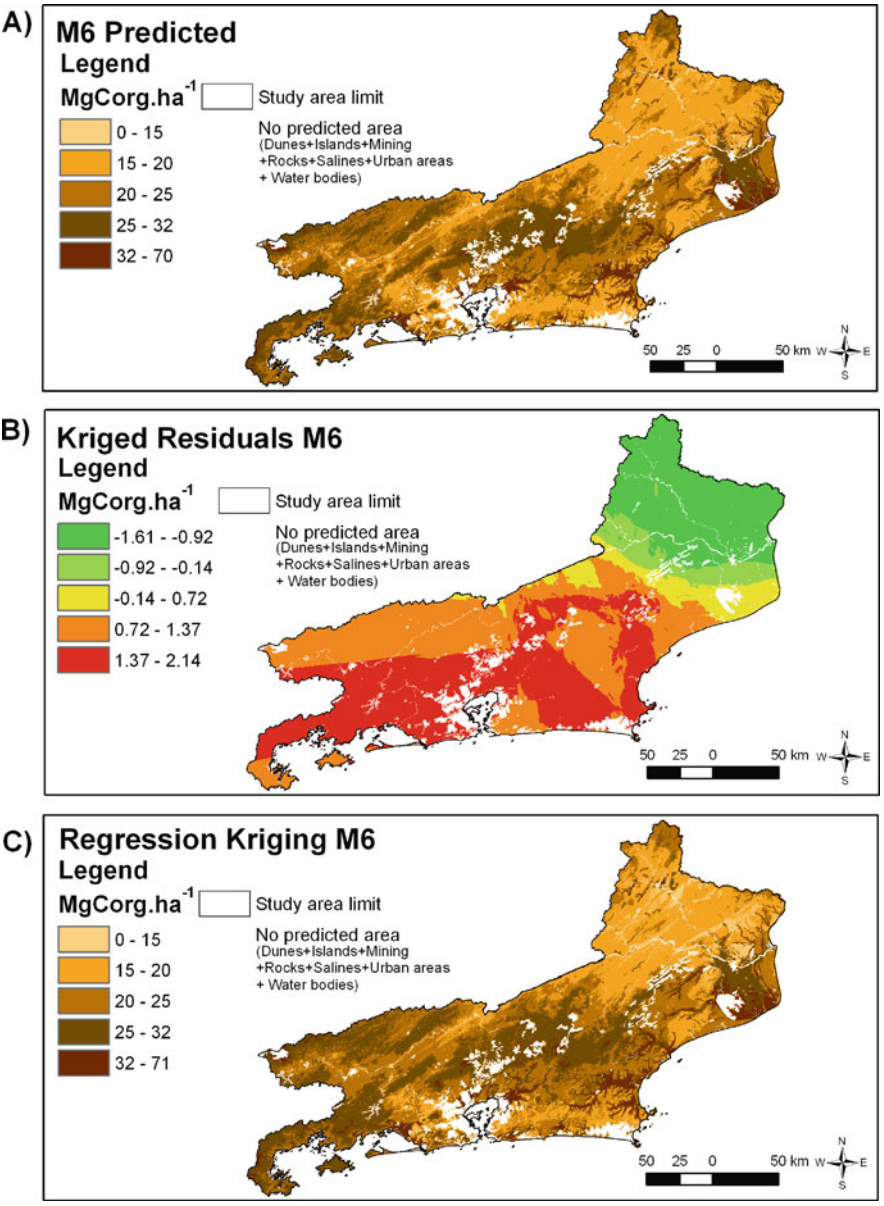


Fig. 21.2 Digital soil map of organic carbon (0–10 cm) of Rio de Janeiro State. (**a**) Result of the predictive *s.c.o.r.p.a.n.* soil-landscape modelling (multilinear regression); (**b**) kriging of the modelling residues; (**c**) final result obtained by the sum of the kriged residues with the values predicted by the multilinear regression (regression kriging)

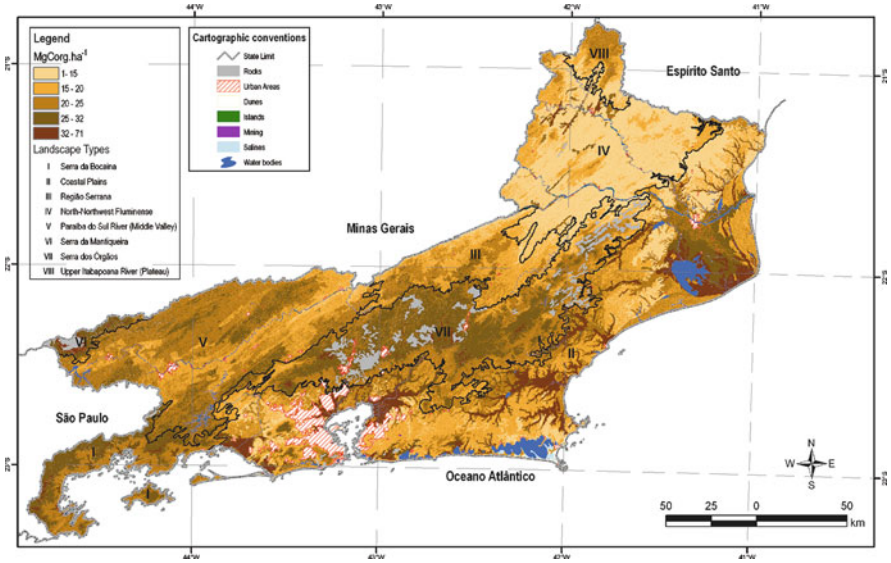


Fig. 21.3 The final map of the topsoil organic carbon of Rio de Janeiro State (0–10 cm) and landscape types

others), rivers, lakes (Lagoa Feia, Lagoa de Maricá among other) and areas close to the coast and its corresponding soils. On the other hand, the unit IV (North-Northwest Fluminense) presents the lowest stocks of organic carbon.

21.4 Conclusions

In this application of predicting the soil organic carbon stock, digital soil mapping is demonstrated, using the soil formation factors as predictor variables for the construction of different models. The work was designed to test the methodology of carbon stock prediction in the soil at the depth 0–10 cm. Seven predictive models were tested. The best result for carbon stock was obtained with application of the model 6 that had the lowest indexes AIC and RMSE (Table 21.3). This model used information from existing soil maps, satellite images, DEM and its derivates, and a lithology map.

The spatial distribution of soil organic carbon has clear relationships with the different geo-environments in the study area, e.g., the highest stocks of organic carbon occur in the lowlands areas. Establishing a baseline for soil organic carbon stock is very important for the definition of public policies maintaining agricultural systems and environmental protection, and thus considers the potential of soil as a carbon sink.

Acknowledgments We would like to thanks FAPERJ – Carlos Chagas Filho Foundation for Research Support of Rio de Janeiro State for financing the Thematic Project E-26/171.360/2001

entitled “Modeling of the magnitude and spatial distribution of the Organic Carbon in the State of Rio de Janeiro, using quantitative techniques, SIG and Data Base”. We also thankfully acknowledge Embrapa for supporting the Research Projects SEP 01.2002.202) and SEP 12.2002.001, entitled “Application of digital quantitative techniques to optimize soil mapping techniques for planning environmental management”, in which scope this work was developed.

References

Akaike, H., 1973. Information theory and an extension of maximum likelihood principle, pp., 267–281. In: Petrov, B.N., and Csaki, F. (eds.), Second International Symposium on Information Theory. Akademia, Kiado, Budapest.

Batjes, N.H., 1998. Mitigation of atmospheric CO_2 concentrations by increased carbon sequestration in the soil. Biology and Fertility of Soils 27:230–235.

Batjes, N.H., and Sombroek, W.G., 1997. Possibilities for carbon sequestration in tropical and subtropical soils. Global Change Biology 3:161–173.

Jarvis, A., Reuter, A., Nelson, E., and Guevara, E., 2006. Hole-filled SRTM for the globe Version 3. Available from the CGIAR-CSI SRTM 90m. Homepage: http://srtm.csi.cgiar.org (Last accessed 23 April 2010).

Jenny, H., 1941. Factors of Soil Formation. McGraw-Hill, New York, NY, 109 p.

Lagacherie, P., and McBratney, A.B., 2007. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping, pp. 3–22. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, vol. 31. Elsevier, Amsterdam, 600 p.

Lumbreras, J.F., Naime, U.J., Carvalho Filho, A., Wittern, K.P., Shinzato, E., Dantas, M.E., Palmieir, F., Fidalgo, E.C.C., Calderano, S.B., Medina, A.I.M., Pimentel, J., and Chagas, C.S., 2003. Zoneamento Agroecológico do Estado do Rio de Janeiro. Rio de Janeiro: (Embrapa Solos. Boletim de Pesquisa e Desenvolvimento n.33). 113 p.

MacMillan, R.A., 2003. LandMapR© Software Toolkit- C++ Version: User manual. LandMapper Environmental Solutions Inc., Edmonton, AB, 110 p.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

Mendonça-Santos, M.L., Jacques, P.D., Coelho, M.R., Pimentel, J., Santos, H.G. Dos, Almeida, P., Barbosa, E.L. de A., Costa, T.C. e C. Da, Ávila, R.M., Shinzato, E., Branco, P.C.M.P. de A., 2003. Mapeamento do Uso Atual e Cobertura Vegetal dos Solos do Estado do Rio de Janeiro. Rio de Janeiro: (Embrapa Solos. Boletim de Pesquisa e Desenvolvimento n. 22), 44 p.

Mendonça-Santos, M.L., Santos, H.G. Dos, Coelho, M.R., and Pares, J.G., 2005. Caracterização de paisagens e solos representativos do Estado do Rio de Janeiro para fins de estimativa de estoques de carbono no solo. Rio de Janeiro: (Embrapa Solos. Boletim de Pesquisa e Desenvolvimento n. 66). 80 p.

Mendonça-Santos, M.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, pp. 381–396. In: Hartemink, A.E., McBratney, A., and Mendonça-Santos, M.L. (eds.), Digital Soil Mapping with Limited Data. Springer, Amsterdam.

NASA, 2008. The Landsat program. Available in: http://landsat.gsfc.nasa.gov/about/ (Last verified 25 September 2008).

Rio de Janeiro, 2001. Geologia, geomorfologia, geoquímica, geofísica, recursos minerais, economia mineral, hidrogeologia, estudos de chuvas intensas, aptidão agrícola, uso e cobertura do solo, inventário de escorregamentos, diagnóstico geoambiental. Rio de Janeiro. CPRM: Embrapa Solos: DRM-RJ. 1 CD-ROM.

Chapter 22
Comparing Decision Tree Modeling and Indicator Kriging for Mapping the Extent of Organic Soils in Denmark

M.H. Greve, M.B. Greve, R. Bou Kheir, P.K. Bøcher, R. Larsen, and K. McCloy

Abstract Soil organic carbon (SOC) is of great importance in the global carbon cycle. In ratifying the Kyoto CO_2 Protocol, Denmark has chosen to fulfill article 3.4, which recognizes sources and sinks of biospheric carbon (including forest management, cropland management, and grassland management). As part of the four year 3.5 million $ contract with the Danish environmental authorities, University of Aarhus will assess the contemporary stock of organic carbon in the wet agricultural lands, serving as a baseline for future studies of soil carbon changes. The first step in the assessment process is to map the extent of the organic soils. This paper compares two prediction methods for this mapping. One method is decision tree modeling (DT) based on legacy soil information, DEM derivatives and RS indices, while the second method, indicator kriging (IK) is a geostatistical model of ground samples (point aggregation). On the decision tree modeling map 30% of the wetlands were classified as organic. We compared the 63% probability map from the IK to the decision tree modeling map. The similarity in the wetland areas between the DT map (pixel to pixel) and the 63% probability map was 63%. We compared both the 63% probability map and the DT map to an independent point dataset, and the DT classified 58% in the correct class, the 63% probability map classified 52% in the correct class.

Keywords Organic soils · Decision tree model · Indicator kriging · Legacy data · Digital soil mapping

M.H. Greve (✉)
Department of Agroecology and Environment, Faculty of Agricultural Sciences, University of Aarhus, Aarhus, Denmark
e-mail: Mogensh.greve@agrsci.dk

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_22,

22.1 Introduction

Soil organic carbon (SOC) is one of the most important carbon stocks globally and has large potential to affect global climate. Accordingly, many scientists from all over the world have initiated studies on SOC stock (e.g. Zhang and McGrath, 2004). SOC stock in surface soils worldwide has been estimated to be 2011 PgC (Bolin and Sukumar, 2000), twice the value in either living vegetation or atmospheric carbon. However, these estimates are highly uncertain largely because of data gaps for many regions of the world. SOC stock depends on local climate, landscape type, soil and other specific conditions; it is sensitive to human interference, and to changes in land use and soil management. Protecting or increasing the existing soil C pool by sequestering C from the atmosphere could become crucial in terms of future policies to mitigate the global greenhouse effect. As part of international efforts to stabilize atmospheric greenhouse gas concentrations, signatories to the Kyoto protocol are committed to establish national inventories of the C stock, and to estimate stock changes. In ratifying the Kyoto CO_2 Protocol, Denmark has chosen to fulfill article 3.4, which recognizes sources and sinks of biospheric carbon (including Forest Management, Cropland Management, and Grassland Management).

This requires reliable estimates of C stocks at one point in time for a baseline. Recently, C stock inventories have been established in France (Arrouays et al., 2001), in Europe (Batjes, 2002), in Denmark (Krogh et al., 2003), in North America (Lacelle, 1997), and in China (Zhang et al., 2004). The existing Danish inventory performed by Krogh et al. (2003) is based on historical soil samples mainly from the top soil dating from 25 to 35 years back (in Chapter 21 carbon stock is assessed using DSM techniques for a Brazilian case).

For this reason, this study was conducted to develop a method for delineation of highly organic soils based on existing soil samples and using different predictive geostatistical methods for arriving at the best approach for soil carbon mapping. Two different prediction methods were applied: Decision trees (DT) and indicator kriging (IK).

22.2 Materials and Methods

22.2.1 Study Area

The study area, covering 710,000 ha, is a 45 km wide strip crossing Jutland (within 56°9′–56°33′ N, and 08°6′–10°57′ E). See Fig. 22.1 for a map of the landscape types in the study area. It has a temperate climate with a winter mean temperature of 0°C, and a summer mean of 16° C (Danmarks Meteorologiske Institut, 1998). The average annual precipitation ranges from 800 mm in central Jutland (west) to 500 mm in the easternmost parts. In late autumn, winter and early spring, precipitation exceeds evapotranspiration and between 150 (east region) and 400 (west region) mm of water leaches through the soils (Aslyng, 1978).

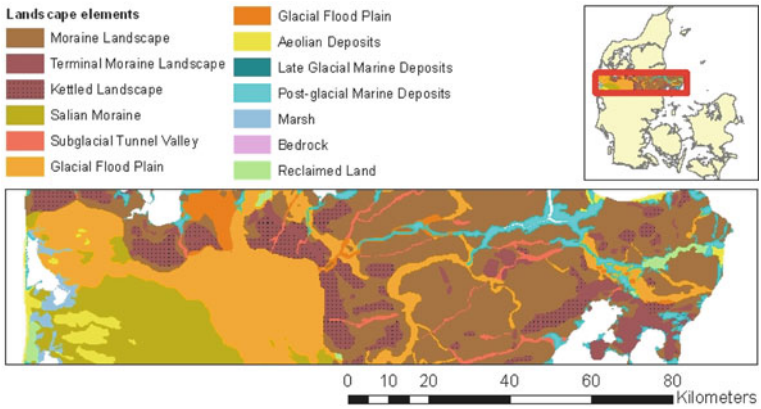


Fig. 22.1 Landscape types in the study area

About two-thirds of the area is intensively farmed with cereals as the main crop. The cultivated areas are fertilized and limed, and low-lying areas and heavy clay soils have been artificially drained. Ten percent is covered by forests, mostly spruce plantations. In the eastern part the area is dominated by loamy weichelian moraine, in the western part large sandy and coarse sandy glaciofluvial plains are dominating.

22.2.2 Delineation of Wetlands

To focus the investigation to wetlands, different sources of legacy polygon data were combined (Table 22.1) and the total wetland area of 15.5% was obtained for the study area. The data are very different in age and quality, but for this preliminary mapping we have allotted the same weight to all. The combined map of wetlands in the study area is shown in Fig. 22.2.

Table 22.1 Data used in this study for delineating the wetlands

Type of data	Description	Scale	Year of compilation	Reference
Wetland	Delineated from old topographic maps	1: 25,000	1910	Madsen et al. (1992)
Humus soil	The humus soil type in the Danish soil classification	1: 50,000	1975	Madsen et al. (1992)
Peat and gytje	Parent material at one meters depth	Ranging from 1: 25,000 to 1: 200,000	1880–2008	GEUS (2009)

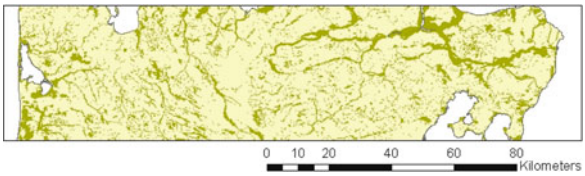


Fig. 22.2 Wetlands in the study area compiled by a combination of historical polygon data

22.2.3 Available Point Information

About 5,000 legacy point samples are available in the study area. In this study, the point dataset was divided into a training dataset (80%) used in the modeling stage, and a test dataset (20%) used in the validation phase. The point data derives from four different surveys (see Table 22.2). From two of the surveys, The Danish Soil Classification and The Danish Soil Profile Investigation, the Soil Organic Matter (SOM) is known from laboratory analysis. This continuous variable is then converted to a categorical variable using 10% SOM as a cut off value. In Denmark we have traditionally defined organic soils as having more than 10% SOM and mineral soil as having less than 10% SOM. The two other surveys give information on parent material in classes (e.g. peat, gytje, sand, silt and clay). The organic parent material classes are then reclassified to organic soils and the rest classified as mineral soils. The legacy point samples have very different age and quality, this aspect is not taken into account in this work. The resulting point dataset is containing 5,000 points of which 45% is classified as organic soils and 55% is classified as mineral soils.

Table 22.2 Legacy point sample input for both the DT and the IK

Type of data	Description	Year of compilation	Reference
The Danish soil classification	Top soil samples	1975	Madsen et al. (1992)
Ochre classification	Augerings with description of parent material in the wetland areas	1985	Madsen et al. (1985)
Well database	Well data from the upper one meter.	1950–2008	GEUS (2008)
The Danish soil profile investigation	A nationwide 7 km grid, with detailed information on the soil profile to a depth of 1.7 meter	1990	Ostergaard (1990)

22.2.4 Other Geospatial Data

22.2.4.1 Historical Maps

Landscape Type

The landscape map was compiled based on topographic maps at a scale of 1:100,000. The delineations between different landscapes were drawn on basis of the contour lines, former landscape maps, and geological surveys published at a

scale of 1:100,000. The landscape has been divided into nine different landforms (Fig. 22.1).

Parent Material

A nation-wide mapping of the geological origin of the sediments in the depth of 1 m was initiated in 1888 by the Danish Geological Survey, today known as GEUS. At present about 90% of the country is mapped and published at scales of both 1:100,000 and 1:20,000. The mapping divides the sediments in genetic classes like: till, eolian, outwash plains, and marine deposits. Furthermore, texture is indicated as clayey, sandy, or silty (e.g. clayey till, sandy melt water deposits). The modeling for our work was done using the update from February 2009 (GEUS, 2009).

Soil Type

The Danish Soil Classification was compiled from 1975 to 1978. The Danish Soil Classification is a textural classification of the topsoil (0–20 cm) into eight soil types, separated on the basis of the content of clay, fine sand, SOM and carbonate. The survey had national coverage and is based on approximately one sample per square kilometer.

22.2.4.2 Remote Sensing Derivatives

NDWI, NDVI and RS_Index3

Digital satellite data from the Landsat TM sensor, with a spatial resolution of 30 m acquired in April 1987, was used in this study for crop identification and wetland inventory, because the area covered with winter crops was much less in 1987 compared to more resent images, and therefore the percentage of bare soils was higher. The chosen images were free of clouds, and allowed us to interpret diverse types of soils in cultivated lands harvested at that time. Data selection was also influenced by the availability and relatively low cost of Landsat images. Data were registered and ortho-rectified (RMS error of about 0.7 pixels) using ground control points (GCPs) and digital road maps (Danish technical T0 maps) at a scale of 1:10,000. Two common band ratio indices were derived: (1) *the normalized difference vegetation index (NDVI)* which computes the normalized difference of brightness values from TM4 (near infrared band) and TM3 (red band) and is used for monitoring the amount of photosynthesizing present; and (2) the *normalized difference wetness index (NDWI)* created using near infrared and short infrared TM bands (TM4 and TM5) and used to compute soil moisture.

An index was developed also in this study for computing soil color, based on blue, green, red and NIR bands. The formula of this index can be written as follows: RS_Index3 = (3*TM4 + TM3 − TM2 − 3*TM1). Band TM1 is Blue, band TM2 is Green, band TM3 is Red and band TM4 is NIR.

22.2.4.3 Digital Elevation Derivatives

Generation of DEM

A digital elevation model (DEM) was generated by a private company (COWI) for the area based on data acquired using airborne LIDAR (Light Detection and Ranging) technology.

The initial DEM was produced as raster GRID with 2 m pixel resolution. To increase the efficiency in terms of storage and manipulation, and to acquire homogeneity and standardization with ancillary maps and satellite imageries applied in this study, the high-resolution DEM was coarsened to 24 m resolution (see Section 5.3 for a discussion on coarsening of DEM).

Derivation of Terrain Attributes from DEM

Terrain attributes derived from digital elevation models (DEMs) are commonly useful explanatory variables in predictive soil models (Gessler et al., 1995; Moore et al., 1993; Odeh et al., 1995; Skidmore et al., 1996). Eight primary terrain attributes were derived directly from a mosaic of DEMs with a resolution of 24 m using standard commands in ArcGIS 9.3 (attributes1-6) or TerraSTREAM (attributes 7 and 8): (1) elevation; (2) slope gradient (radians); (3) slope aspect; (4) profile curvature, (5) contour curvature, (6) mean curvature; (7) flow direction; and (8) contributing area (flow accumulation). Attributes 1–7 were calculated based on a window of eight pixels surrounding each pixel. This does not consider the characteristics of the upslope contributing area of each pixel, nor does it consider the relative position of each pixel within the toposequence. Therefore, the average value of these terrain attributes for the upslope contributing area (attribute 8) of each pixel was calculated.

Elevation (ranging between −4 and 123 m) is useful for classifying the local relief, and for locating points of maximum and minimum heights. It has a high correlation with SOC.

Slope gradient, S, characterizing the spatial rate of change of elevation in the direction of steepest descent, affects the velocity of both surface and subsurface flow, and hence organic carbon distribution. On steep slopes, dry soil conditions prevail due to more rapid removal of water causing an important decrease in SOC.

As for slope aspect, ψ (orientation of the line of steepest descent), it is useful for visualizing the direction of landscapes, and is frequently recorded in soil surveys. Over the study area, the aspect varies from maximum (359°) to minimum (0°). Aspect is divided into the eight major directions plus the non-oriented flat areas. Slopes exposed to the south and west are more subject to runoff for two reasons: (1) they are warmer with higher evaporation rates and lower moisture storage capacity, thus less forested than those exposed to the north and east, and (2) rainfall affects slope aspect depending on the direction of winds during rainfall, which commonly has a west and south–west trend in Denmark.

Plan, profile and mean curvature, K, measures the distribution of convex and concave areas; hence the propensity of water to converge or diverge as it flows

across the land. Concave slopes can concentrate more water and sediments indicating the potential accumulation of a large quantity of soil organic carbon. Convex slopes show an inverse effect, dispersing flow and limiting material accumulation, therefore a lesser quantity of soil tends to accumulate than on concave slopes. Flat areas (zero curvature) are without any effect on flow divergence or convergence.

The drainage network provides an important indication of water percolation rate. It is commonly accepted in some settings that the denser this network is, the lower the recharge rate and vice versa. Flow direction was calculated for all pixels by the direction of steepest descent from each cell. A stream network was derived by connecting all pixels that accumulate flow from 100 pixels or more. Flow accumulation grid and digitized outlets from the stream network were used to automatically subdivide the whole area into small watersheds. Each watershed was subdivided into two facets, separated by the streamline passing through the watershed.

In addition to the above attributes, a compound topographic wetness index (CTI) was also calculated for each pixel using the specific catchment area (As) and the slope degree (ß), according to the formula (CTI = ln[As/tanß]) (Moore et al., 1993).This index is a predictor of zones of soil saturation. Small values of CTI generally depict upper catenary positions and large values lower catenary positions with an overall range typically from 2 to 12 for zero-ordered planar areas.

22.2.5 Inference Models

The spatial data was aggregated using two different approaches: decision tree modeling and indicator kriging.

22.2.5.1 Decision Tree Modeling

The success of this type of modeling (i.e., decision-trees) relies on the strength of the relationship between soil properties and environmental variables available. Where these relationships are strong, the model will also be strong (see Chapters 20 and 21 for a comparable set of ancillary data used for prediction). Decision-tree models (DT) can handle both missing values, continuous and categorical predictors, are robust to predictor specification, and make very limited assumptions about the form of the regression model (Henderson et al., 2005; Scull et al., 2005). They are easier to interpret and discuss when a mix of continuous and discrete variables is used as predictors. However, the most significant advantage of tree-based models is the capacity to model non-additive and non-linear relationships in a relatively simple way. This is particularly useful for soil data where interactions between the response variable and environmental explanatory variables are often conditional on other explanatory variables.

Estimation of organic soil class can be achieved using classification-tree modeling through incorporating secondary spatial information into prediction (Mueller and Pierce, 2003). Terrain attributes (e.g., elevation, slope, aspect, curvature) may

Table 22.3 Input data for decision tree model

Variable	Type	Class	Reference
Point information (organic or mineral)	Target	Categorical	
Landscape type	Predictor	Categorical	Madsen et al. (1992)
Parent material	Predictor	Categorical	GEUS (2009)
Soil type	Predictor	Continuous	Madsen et al. (1992)
RS derivatives			
NDWI	Predictor	Continuous	Gao (1996)
NDVI	Predictor	Continuous	Rouse et al. (1973)
Index3	Predictor	Continuous	Pers. kom. Keith McCloy
DEM derivatives			
Elevation	Predictor	Continuous	
Profile curvature	Predictor	Continuous	Gallant and Wilson (1996)
Plane curvature	Predictor	Continuous	
Slope gradient	Predictor	Continuous	
Aspect	Predictor	Continuous	
Flow direction	Predictor	Continuous	
Flow accumulation	Predictor	Continuous	
Wetness index/CTI	Predictor	Continuous	Moore et al. (1993)

aid spatial estimation of soil carbon, because the relief has a great influence on soil formation (Bou Kheir et al., 2007, 2008; McKenzie and Ryan, 1999).

Besides digital elevation models, one of the most interesting sources of secondary information could be remote sensing (RS), if a relationship between soil properties and spectral data could be achieved. Remotely sensed data can be useful for improving existing coarse-scale soil survey information at a regional scale (Daniel et al., 2001). However, high carbon soils in Denmark cannot directly be differentiated from moist soils using satellite images, since both appear as dark soils with decreasing spectral reflectance as water or carbon content increases. However, some RS indices such as Normalized Difference Vegetation Index (NDVI) and Normalized Difference Wetness Index (NDWI) can be used to integrate vegetation status and biomass or to compute surface moisture, and will be tested for spatial prediction of soil carbon. Table 22.3 holds a comprehensive list of the predictor values used for building the DT.

The decision tree modeling was done using the software DTREG and the resulting data layer was compiled using Model Builder in ArcGIS Spatial Analyst.

22.2.5.2 Indicator Kriging

Indicator kriging (described by Oliver and Webster, 1991) provides estimates at unobserved location of the variable Z, based on the weighted average of adjacent

observed sites within a given area. The theory is derived from that of regionalized variables and can be briefly described by considering an intrinsic random function denoted by $Z(s_i)$, where s_i represents all sample locations, $i = 1, \ldots, n$. An estimate of the weighted average given by the ordinary kriging predictor at an unsampled site $z(s_0)$ is defined by:

$$z(s_0) = \sum_{i=1}^{n} \lambda_i z(s_i) \tag{22.1}$$

where: λ_i are the weights assigned to each of the observed sample sites.

These weights sum to unit so that the predictor provides an unbiased estimation:

$$\sum_{j=1}^{n} \lambda_i = 1 \tag{22.2}$$

The weights are calculated from the matrix equation:

$$\mathrm{c} = \mathrm{A}^{-1}\mathrm{b} \tag{22.3}$$

where A: a matrix of semivariances between the data points; b: a vector of estimated semivariances between the data points and the point at which the variable Z is to be predicted; and c stands for the resulting weights and the Lagrange Multipliers ψ (Triantafilis et al., 2001). In this study, a binary SOC value has been considered for all sample points as follows: if SOC is less than 10%, a score of 0 was attributed; if it is higher than 10%, a score of 1 was attributed.

The probability maps generated on the basis of the indicator kriging show the probability of a soil being organic, i.e. in a 63% probability map, all pixels classified as organic have a 63% or higher chance of being organic.

The analysis was done using the Geostatistical Analyst extension in ArcGIS.

22.3 Results and Discussion

22.3.1 Decision Tree Modeling

The decision tree predicts two classes: organic soil or mineral soil. The full tree has 129 terminal nodes. In pruning the tree using cross validation, the tree with 21 terminal nodes was optimal. The five branches leading to a classification in the organic soil group are shown in Fig. 22.3. The prediction variables used in the pruned tree and their relative importance are shown in Table 22.4.

Looking at the list of variables it is noteworthy that none of the remote sensing indices are on the list, even if we only looked at pixels with NDVI > 0, 1 indicating

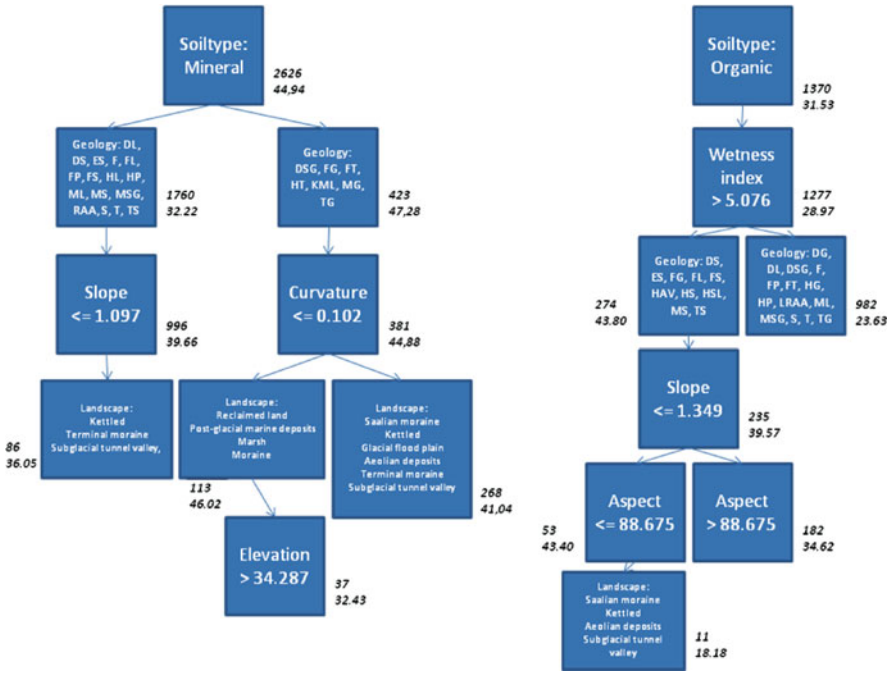


Fig. 22.3 The five organic branches in the decision tree. The numbers next to the nodes are: *top* – number of points, *bottom* – misclassification (percentage). Geology abbreviations refer to various parent materials

bare soils, the NDWI and Index3 performed better than NDVI. Probably because NDVI is a vegetation index and the other two RS indices are affected by soil.

Not surprisingly soil type turned out to be the most important variable since organic top soil is one of the classes on this map. The initial split is done using this variable.

The parent material is also important but more ambiguous and used in different levels of the analysis. This is probably because this map shows the parent material in the subsoil and a lot of the samples in the target variable are classified as organic due to high SOM in the topsoil only.

Of all the DEM derivatives, slope and wetness index had the highest priority and were of almost equal importance. Slope is used to split on two branches with slope less than 1.1 radians and 1.4 radians respectively for classification in the organic class. We had high expectation to the importance of the wetness index, but to our surprise this variable was of less importance than slope. When we looked closer into the data, we found that a lot of the large river valleys came out with very low wetness indices (with very high wetness indices at the foot slope) and a lot of the valleys were classified as relatively dry, despite of a very high groundwater table.

A test dataset comprising of 20% of the total points were used as an independent test of the DT map. Fifty eight percent of the points in the test dataset were classified correctly.

22.3.2 Indicator Kriging and the Probability Maps

The experimental variogram was calculated using a lag size of 2,000 m and 12 lags. A spherical model was fitted, and sill, nugget and range were estimated to be 0.21, 0.11 and 16,700 m.

A low sill to nugget ratio (0.52), indicating a weak spatial correlation, and the very long range leads to a SOC map with large scale (or REGIONAL) pattern and high smoothing effect. This result indicates that too few points are available for this technique. The map seems very unreliable by visual interpretation. Despite this first hand impression, a comparison to the independent test dataset shows that 52% of the test points were classified correctly by indicator kriging.

22.3.3 Comparing the IK Probability Map and the DT Distribution Map

It is not straightforward to compare the results from DT analysis with the results from IK. A map compiled on the basis of a DT analysis shows the *distribution* of organic and mineral soils. The probability maps generated on the basis of the IK show the *probability* of a soil being organic.

On the DT map, 30% of the wetland was classified as organic soils. We decided to compare the probability map with the same percentage of organic soils as the DT map, namely the 63% probability map.

The indicator map classifies the study area in large more or less uniform spots due to the very long range and high smoothing of the kriging. The DT map shows more fine scale variation due to variation in the input map with high statistical importance (Fig. 22.4).

The similarities between the two maps are relatively high, as 63% of the wetland area is classified similar by the two methods. Fifty three percent of the test points are classified similar.

Table 22.4 The relative importance of variables use in the pruned tree analysis

Variable	Importance
Soil type	100
Parent material/geology	39.8
Slope gradient	13.6
Wetness index/CTI	10.7
Landscape type	7.7
Elevation	2.2
Plan curvature	1.9
Aspect	1.7

In this case indicator kriging performs worse; this is probably due to too few sample points in the wetland areas for this method.

22.4 Conclusion

The modeling approaches (decision tree modeling and indictor kriging) adopted in this study were easily implemented with available GIS software, and are suitable for prediction of organic carbon at unobserved locations. They provided quick and simple methods for generating maps describing the extent of organic soils in Denmark.

Data derived from conventional soil surveys where the individual surveyors select sites for sampling are less applicable for statistical mapping since there is a risk of bias in the sample locations. The indicator kriging techniques require a fairly dense network of sampling sites and the result of the geostatistical analysis shows very long range (16,700 m) and a high nugget-to-sill ratio (0.52) indicating that too few points are available for this technique. To further pursue these techniques it will be important to increase the number of observations by including point information from other databases, or include the data from the coming survey finishing in 2010.

The DT analysis gave a good impression and a fairly good classification result. To further improve the results of the DT analysis we have to investigate the value of other DEM derivatives and the importance of the spatial resolution of the DEM, in

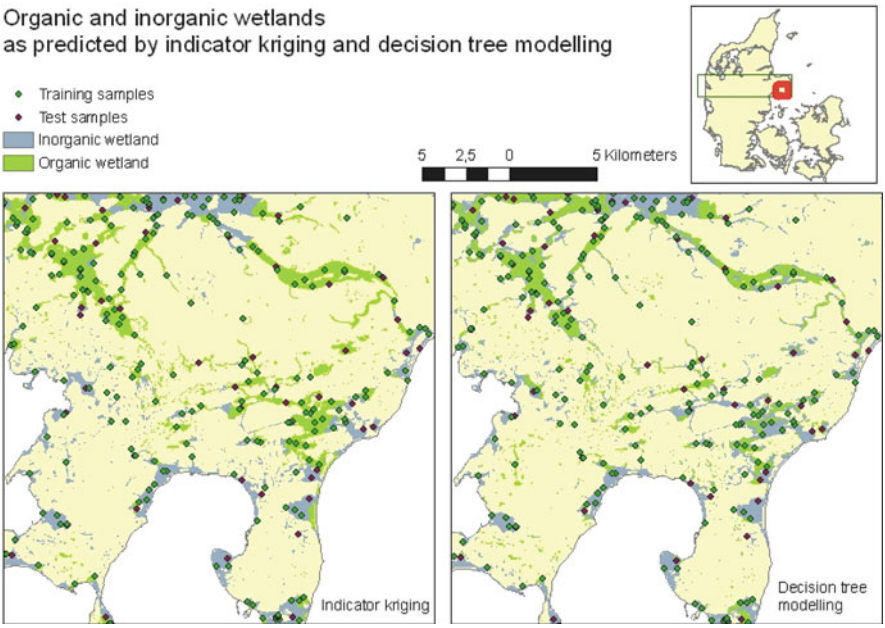


Fig. 22.4 Zoom-in on a part of the study area to visually interpret the differences between the indicator kriging map and the DT map

this study we used 24 m resolution, and we might get a better result if we increased the spatial resolution of the DEM or changed the size of the analysis window.

The geographic reliability of both the points and some of the predictors is very varied and in several cases unknown. This could be the cause of the mediocre classification results.

References

Arrouays, D., Deslais, W., and Badeau, V., 2001. The carbon content of topsoil and its geographical distribution in France. Soil Use Management 17:7–11.

Aslyng, H.C., 1978. Miljø og jordbrug. DSR Forlag, Copenhagen, Denmark.

Batjes, N.H., 2002. Carbon and nitrogen stocks in the soils of central and eastern Europe. Soil Use and Management 18:324–329.

Bolin, B., and Sukumar, R., 2000. Global perspective, pp. 23–51. In: Watson, R.T., Noble, I.R., Bolin, B., Ravindranath, N.H., Verardo, D.J., and Dokken, D.J. (eds.), Land-Use, Land-Use Change, and Forestry, A Special Report of the IPCC. Cambridge University Press, Cambridge, Massachusetts, USA.

Bou Kheir, R., Chorowicz, J., Abdallah, C., and Dhont, D., 2008. Soil and bedrock distribution estimated from gully form and frequency: a GIS-based decision-tree model for Lebanon. Geomorphology 93:482–492.

Bou Kheir, R., Wilson, J., and Deng, Y., 2007. Use of terrain variables for mapping gully erosion susceptibility in Lebanon. Earth Surface Processes and Landforms 32:1770–1782.

Daniel, K., Tripathi, N.K., Honda, K., and Apisit, E., 2001. Analysis of spectral reflectance and absorption patterns of soil organic carbon, 22nd Asian Conference on Remote Sensing, 5–9 November 2001, Singapore.

Danmarks Meteorologiske Institut. 1998. Danmarks Klima 1997. Danmarks Meteorologiske Institut, Copenhagen.

Gallant, J.C., and Wilson, J.P., 1996. TAPES-G: A grid-based terrain Survey Investigations Rep. 42, Version 3.0. USDA-NRCS, Linanalysis program for the environmental sciences. Computational Geosciences, NE. 22:713–722.

Gao, B., 1996. NDWI: A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment 58(3):257–266.

Gessler, P.E., Moore, I.D., McKenzie, N.J. and Ryan, P.J., 1995. Soil-landscape modelling and the spatial prediction of soil attributes. International Journal of Geographic Information Systems 94:421–432.

GEUS (Geological Survey of Denmark and Greenland) 2008. The Jupiter database, http://www.geus.dk/jupiter/index-dk.htm (Last accessed 23 April 2010).

GEUS, 2009. Danmarks digitale jordardskort 1:25.000. CD-ROM. De Nationale Geologiske Undersøgelser for Danmark og Grønland.

Henderson, B.L., Bui, E.N., Moran, C.J., and Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124:383–398.

Krogh, L., Noergaard, A., Hermanen, M., Humlekrog Greve, M., Balstroem, T., and Breuning-Madsen, H., 2003. Preliminary estimates of contemporary soil organic carbon stocks in Denmark, using multiple datasets and four scaling-up methods. Agricultural Ecosystem and Environment 96:19–28.

Lacelle, B., 1997. Canada's soil organic carbon database, pp. 93–102. In: Lal, R. (ed.), Soil Processes and the Carbon Cycle. CRC Press, Boca Raton, FL, USA.

Madsen, H.B., Jensen, N.H., Jakobsen, B.H., and Platou, S.W., 1985. A method for identification and mapping potentially acid sulfate soils in Jutland, Denmark. Catena 12(4):136–371.

Madsen, H.B., Nørr, A.H., and Holst, K.A., 1992. The Danish Soil Classification. Atlas of Danmark I3. Reitzel, Copenhagen.

McKenzie, N.J., and Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89:67–94.

Moore, I.D., Gessler, P.E., Nielsen, G.A., and Petersen, G.A., 1993. Terrain attributes: estimation methods and scale effects, pp.189–214. In: Jakeman, A.J., Beck, M.B., McAleer, M. (eds.), Modelling Change in Environmental Systems. Wiley, London.

Mueller, T.G., and Pierce, F.J., 2003. Soil carbon maps – Enhancing spatial estimates with simple terrain attributes at multiple scales. Soil Science Society of America Journal 67:258–267.

Odeh, I.O.A., McBratney, A.B., and Chittleborough, D.J., 1995. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67:215–226.

Oliver, M.A., and Webster, R., 1991. How geostatistics can help you? Soil Use and Management 7:206–217.

Ostergaard, H.S., 1990. Kvadratnettet for nitratundersøgelser i Danmark 1986-89 (in Danish). Danish Agricultural Advisory Centre, Skejby, Aarhus.

Rouse, J.W., Haas, R.H., Schell, J.A., and Deering, D.W., 1973. Monitoring vegetation systems in the great plains with ERTS, Third ERTS Symposium, NASA SP-351 I: 309–317.

Scull, P., Franklin, J., and Chadwick, O.A., 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecological Modelling 181:1–15.

Skidmore, A. K., Gauld, A., and Walker, P., 1996. Classification of kangaroo habitat distribution using three GIS models. International Journal of Geographical Information Systems 10: 441–454.

Triantafilis, J., Odeh, I.O.A., and McBratney, A.B., 2001. Five geostatistical models to predict soil salinity from electromagnetic induction data across irrigated cotton. Soil Science Society of America Journal 65:869–878.

Zhang, G., and McGrath, J., 2004. Geostatistical and GIS analyses on soil organic carbon concentrations in grassland of south-eastern Ireland from two different periods. Geoderma 119:261–275.

Zhang, S.R., Sun, B., Zhao, Q.G., Xiao, P.F., and Shu, J.Y., 2004. Temporal spatial variability of soil organic carbon stock in a rehabilitating ecosystem. Pedosphere 14(4):501–508.

Chapter 23
Modeling Wind Erosion Events – Bridging the Gap Between Digital Soil Mapping and Digital Soil Risk Assessment

H.I. Reuter, L. Rodriguez Lado, T. Hengl, and L. Montanarella

Abstract Wind erosion submits fine as well as coarse soil particles into the atmosphere, thereby affecting physical and chemical processes, affecting radiative forcing, chemical reactions and biological systems. This study was conducted to quantify wind erosion events by generating data for soil erodibility and wind erosivity for the Danube Basin. Estimates of surface soil texture were generated from ~8,000 soil profiles and 54 auxiliary datasets using the regression-kriging method. The quality of the regression equation was not satisfactory. Validation showed an RMSE of 8.6, 10.4 and 13.5 for clay, sand, and silt. Different texture scenarios were generated and the number of wind erosion events for the year 2006 was modeled using weather data from the European Centre for Medium range weather forecast. Vegetation cover fraction was approximated from Meteosat data. Magnitude and spatial extent of wind erosion estimations showed similar order compared to wind erosion estimations based on the European Soil Database.

Keywords Danube basin · DSFM · DSRA · Digital elevation model · Data harmonization

23.1 Introduction

Wind erosion is one of the soil threats outlined in the currently discussed Soil Thematic Strategy of the European Union (European Communities, 2006). Under dry conditions, the danger of wind erosion increases when plant coverage is sparse or non existent in spring and autumn (Lóki et al., 2005). Wind erosion occurs on all types of soil, however mainly on light, sandy soils (Chepil, 1960).

The transport capacity of the wind increases as a power function of the velocity, after exceeding a friction velocity threshold which depends on particle size

H.I. Reuter (✉)
Gisxperts gbr, Eichenweg 42, D-06849 Dessau, Germany
e-mail: hannes@gisxperts.de

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_23,

Fig. 23.1 Image from a natural event of a dust storm on 2005-07-24 in Iraq over a US military camp and from the MODIS satellite

(Bagnold, 1941). However, these transport equations are mostly developed for desert conditions and do not necessarily reflect the conditions occurring in agricultural soils (Fig. 23.1). The effects of the transport are driven by the movement of soil particles due to the wind force and can lead to crop damage, air pollution and decreased fertility of the top soil surface.

To estimate the extent of wind erosion in Europe, EEA-UNEP (EEA, 2000a) created a map of European wind erosion risk based on local empirical data and expert knowledge (see also Van Lynden, 1994). More advanced approaches exist to quantify erosion rates on agricultural soils. Some of them are physically based like the WEPS model (Hagen, 1996), WEAM (Shao et al., 1996) or the WEELS model (Böhner et al., 2003), whereas others like the RWEQ model (Fryrear et al., 1998) or the WFI approach (Beinhauer and Kruse, 1994) are based on empirical equations. One limitation is that these models usually require a number of input parameters, which are rarely available for large areas (e.g., the field management practices might be known for one single region, however they are usually not reported at the regional/federal level), or the model domain might be limited (e.g., WEPS only one field, WEELS only a maximum area of 25 km $\times$ 25 km). Global models used in the climate change community are at a relatively coarse scale and cannot be used for detailed assessments. For European-wide estimations (see also Chapter 18) of wind erosion events, the dominant soil surface texture could be obtained from the European Soil DatabaseV2.0, available at http://eusoils.jrc.it (EUR 19945EN) at 1:1 Million scale. However, wind erosion events are known to be localized and small scale soil mapping of the 1:1 Million might not be sufficient. In the course of the development of the Danube Soil Information System, soil profile information has been obtained from the different member states of the Danube catchments system in a harmonized and standardized way. Those soil profiles, in combination with a digital soil mapping/digital soil functional mapping/digital soil risk assessment approach, are assumed to represent a more detailed picture than a digital soil risk assessment based on the European Soil Database alone.

The objective is to investigate different texture scenarios coming from traditional soil mapping (ESDB) and digital soil mapping and to test the applicability of the workflow using digital soil mapping for digital soil assessment in the context of wind erosion modeling.

23.2 Materials and Methods

We used a regression kriging (RK) approach to estimate clay, silt and sand content (in %), and its related uncertainty, for the Danube basin by means of soil profile observations together with environmental information such as parent material, DEM, climate parameters, remote sensing among others. These texture estimations were then used to calculate the wind erosion susceptibility for the entire Danube river basin.

23.2.1 Study Area

The extent of the study area corresponds to the Danube basin (817,000 km^2) which contains Europe's second-longest river (2,860 km) after the Volga river. The Danube flows from the Black Forest area in Germany to the Black Sea. It drains the northern part of the Alps, the Pannonian and Rumanian lowlands, thereby providing drainage to large areas of southern middle Europe and southeast Europe. Climatic conditions are variable but mostly characterized as a moderate climate following the Köppen-Geiger climate classification (Peel et al., 2007). There is a wide range of soil types present, including Rendzinas (6%), Phaeozems (9%), Cambisols (30%), Luvisols (18%), Fluvisols (8%) and Chernozems (11%) with varying texture conditions. Several authors reported that as much as 650,000 km^2 in eastern and southeastern Europe have severe wind erosion problems (Gross and Bärring, 2003; Lóki et al., 2005).

23.2.2 Digital and Field Data

Soil profile data have been obtained from different member states of the Danube Basin with the aim of creating a Danube Basin Soil Information System (SIS). Although still not fully populated, the database represents a homogenized and standardized dataset of soil profiles containing sand, silt and clay content, OM, stones and a number of other parameters (see also Chapter 25). In total, 8870 points have been available in the Danube basin database for topsoil conditions. However, the samples not containing data for all parameters have been excluded (Fig. 23.2). For a more detailed description of the Danube Basin SIS please refer to Dobos et al. (2006).

In addition, we used a set of auxiliary variables, which are either directly influencing soil texture or might serve as a proxy for the factors in the scorpan model (McBratney et al., 2003). A short overview of the auxiliary variables that were used is given in Table 23.1.

All 54 auxiliary datasets have been organized in a multiband image from the different single input datasets (1.5GB). We rescaled the bands to the same data range to ensure similar weights in the PCA and converted them to 54 Principal Component

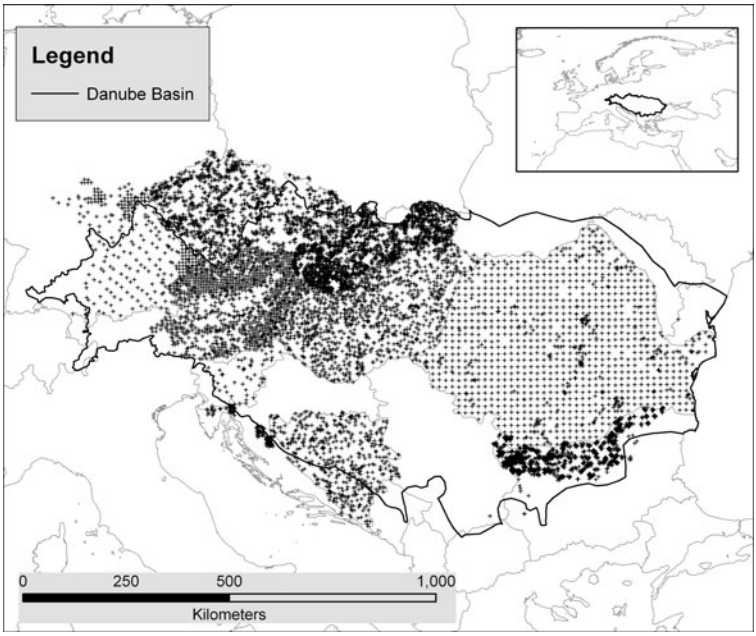


Fig. 23.2 Soil profile description available from the Danube Soil Information System. Note that (**a**) not all profiles contain a complete description and (**b**) several Danube-Basin member states did not deliver data into the system

raster maps using Principal Component Analysis in ENVI V4.3 (ITTVIS, 2008) in order to minimize possible collinearity between variables. For each of the soil sampling points we obtained the value of the 54 PCA components to be used in the inference model.

23.2.3 Inference Models

We used regression-kriging (RK, see Hengl, 2009) to estimate the sand, silt and clay content in the topsoil. Firstly, we performed a stepwise-linear regression (backward selection, significance 0.05) model for the measured texture components as dependent variables against the auxiliary environmental variables (Table 23.2). The derived regression equation was then applied using the 54, standardized, 1 km resolution auxiliary raster grids to obtain a continuous linear regression surface.

In a second step we interpolated the residuals of this regression model by ordinary kriging. The final map is an additive combination of both the stepwise regression and the ordinary kriging models. In case the texture components did not sum up to 100%, we calculated the sum of all estimated texture components and adjusted all three components accordingly.

The original dataset of observations was divided using a random sampling function into a model dataset, that includes the 80% of the samples and a validation

Table 23.1 Auxiliary variables used in the estimation of the soil texture parameters

Parameter	Description
Terrain	1 km SRTM30 dataset was used to derive a slope map, the Topographic Wetness Index (Moore et al., 1991) and total incoming solar insolation (Conrad, 2001) using SAGA 2.0. Land forms were calculated by a modification of the method proposed by Iwahashi and Pike (2007) that combines the topographic variables slope gradient, surface texture and local convexity to create 16 classes of landforms from steep to gentle landforms with fine/coarse texture and low/high convexity.
Geology	Main superficial geologic units in Europe (Pawlewicz et al., 2003) were reduced to 10 classes according to the genetic nature: (1) Granites, rhyolites and quartzites; (2) Paleozoic schists, phyllites, gneisses and andesites; (3) Shales and sandstones; (4) Mesozoic Ultramafic, basic phyllites, schists, limestones and evaporates; (5) Jurassic, Triassic and Cretaceous calcareous rocks; (6) Cenozoic serpentinites, gabros and sand deposits; (7) Tertiary basanites and andesites; (8) Neogene and Paleogene calcareous rocks; (9) Quaternary limestones and basaltic rocks; and (10) Other Ultramafic and undefined rocks. A cumulative earthquake's magnitude map was created by using 90,000 registered earthquakes during the period 1973–1994 (http://earthquake.usgs.gov/eqcenter/). We used the logarithmic measure of the "size" of an earthquake and then rasterized the point map to the 10 km grid by using the point density operation in ILWIS. This operation sums all earthquake magnitudes observed within a 10 km grid and gives a cumulative map of earthquake activity. The release of alkalinity by weathering of primary minerals in soils was calculated using the "Simple Mass Balance Method" as described in Rodriguez-Lado et al. (2007).
Satellite	Monthly averaged MODIS images of the Enhanced Vegetation Index EVI at 1 km resolution for the period 01/01/2004 to 31/12/2006 were obtained from the MODIS Terra imagery at the Earth Observing System Data Gateway. We performed a PCA on 19 complete mosaics and used the first five resulting principal components.
Urbanization	Lights at night for the year 2003 was obtained from the Defense Meteorological Satellite Program (http://www.ngdc.noaa.gov/dmsp/), which measures night-time light emanating from the earth's surface at 1 km resolution. This map is a proxy for urbanization and is now increasingly used for quantitative estimation of global socioeconomic parameters as well as for human population mapping (Doll et al., 2000; Sutton, 1997; Sutton et al., 1997). Similar, the map of distances to roads, airports and utility lines was calculated using the distance operation using the GIS layers from the GISCO database of the European Commission (http://eusoils.jrc.it/gisco_dbm/dbm/home.htm).
Land use	Corine Land Cover 2000 map of Europe was upscaled to a 1 km grid. For Switzerland, we used the Corine Land Cover from 1990 since no updated information was available. The CLC1990 classes for this country were adjusted to those described in the CLC2000 and both datasets were merged together and aggregated to 1 km resolution. The original 44 classes were simplified to 8 classes: (1) urban infrastructures; (2) agriculture; (3) forest; (4) natural vegetation; (5) beaches; (6) ice bodies, (7) wetlands and (8) water bodies. Class 5 (beaches) disappeared in the process of up scaling from 100 m to 1 km. Additionally data from the Global Land Cover classification (Fritz et al., 2003) has been used for areas where no Corine classification was available.
Climate	Mean annual temperature and accumulated precipitation maps were obtained from the very high resolution raster layers created by Hijmans et al. (2005) on a global scale at 1 km grid resolution. The annual potential evapotranspiration (PET) was calculated from monthly temperature data using the Thornthwaite method (1948). Runoff was calculated as the difference between annual accumulated precipitation and annual potential evapotranspiration.
Permafrost and lakes	We further masked out non-soil surfaces such as water bodies (rivers, lakes, sea etc.) and permafrost areas. A consistent European-wide water mask, indicating the percentage of water area inside a 1 km pixel, has been created based on the NASA SRTM V2 SWBDB dataset (Rabus et al., 2003), the CORINE land use classification, lakes contained in the GISCO data, base water reflection by the use of Image2000 dataset (de Jager et al., 2006), the GSHHS – Database (Wessel and Smith, 1996), and the Global Lakes and Wetlands Database (Lehner and Döll, 2004).

Table 23.2 Example of PCA coefficients for estimation of sand content

PCA	Estimate	Std. error	t value	$Pr(>\|t\|)$	PCA	Estimate	Std. error	t value	$Pr(>\|t\|)$
(Intercept)	−2.2E + 04	5.7E + 03	−3.9E + 00	1.03E − 04					
PCA1	4.4E − 01	1.1E − 01	3.9E + 00	1.14E − 04	PCA26	1.2E − 02	5.7E − 03	2.1E + 00	3.98E − 02
PCA2	−8.5E − 04	2.0E − 04	−4.3E + 00	1.81E − 05	PCA27	4.8E − 02	5.7E − 03	8.5E + 00	2.15E − 17
PCA3	−7.5E − 02	1.8E − 02	−4.2E + 00	3.20E − 05	PCA29	−8.1E − 02	8.1E − 03	−1.0E + 01	3.04E − 23
PCA4	9.3E − 01	2.3E − 01	4.1E + 00	3.96E − 05	PCA30	−3.3E − 02	2.3E − 02	−1.5E + 00	1.45E − 01
PCA5	−7.4E − 01	1.8E − 01	−4.1E + 00	4.07E − 05	PCA31	−8.4E − 02	9.7E − 03	−8.7E + 00	7.49E − 18
PCA6	1.5E − 01	3.6E − 02	4.0E + 00	6.19E − 05	PCA32	−4.7E − 02	1.7E − 02	−2.8E + 00	5.12E − 03
PCA7	1.8E − 02	4.4E − 03	4.1E + 00	4.38E − 05	PCA33	−1.4E − 01	1.2E − 02	−1.1E + 01	1.51E − 27
PCA8	−4.1E − 02	9.8E − 03	−4.2E + 00	3.29E − 05	PCA34	2.3E − 01	2.0E − 02	1.2E + 01	2.76E − 30
PCA9	−6.1E − 02	1.6E − 02	−3.8E + 00	1.34E − 04	PCA36	9.3E − 02	1.6E − 02	5.6E + 00	2.09E − 08
PCA10	2.7E − 02	6.4E − 03	4.2E + 00	2.93E − 05	PCA37	1.2E − 01	1.7E − 02	7.2E + 00	6.15E − 13
PCA11	−5.3E − 02	1.3E − 02	−4.2E + 00	2.34E − 05	PCA38	−3.1E − 02	1.1E − 02	−2.9E + 00	3.43E − 03
PCA12	2.1E − 02	5.1E − 03	4.2E + 00	3.35E − 05	PCA40	1.0E − 01	1.8E − 02	5.5E + 00	4.20E − 08
PCA13	3.7E − 02	8.8E − 03	4.2E + 00	2.47E − 05	PCA41	−1.9E − 01	2.0E − 02	−9.3E + 00	3.56E − 20
PCA14	8.0E − 02	1.9E − 02	4.3E + 00	2.05E − 05	PCA42	−7.2E − 02	2.3E − 02	−3.1E + 00	1.66E − 03
PCA15	−2.5E − 02	6.0E − 03	−4.2E + 00	2.52E − 05	PCA45	−2.4E − 01	3.5E − 02	−6.8E + 00	1.40E − 11
PCA16	−5.0E − 02	1.2E − 02	−4.2E + 00	3.34E − 05	PCA46	4.6E − 02	2.0E − 02	2.3E + 00	2.36E − 02
PCA17	3.7E − 02	9.2E − 03	4.0E + 00	5.92E − 05	PCA47	2.1E − 01	3.4E − 02	6.2E + 00	4.89E − 10
PCA18	−4.2E + 00	1.0E + 00	−4.2E + 00	2.86E − 05	PCA48	2.4E − 01	3.1E − 02	7.7E + 00	2.44E − 14
PCA19	−1.3E − 01	3.1E − 02	−4.2E + 00	2.46E − 05	PCA49	2.6E − 01	3.4E − 02	7.9E + 00	6.07E − 15
PCA20	−1.1E − 01	2.6E − 02	−4.3E + 00	2.01E − 05	PCA50	2.9E − 01	4.6E − 02	6.4E + 00	2.12E − 10
PCA21	3.8E − 02	9.1E − 03	4.2E + 00	3.21E − 05	PCA51	−6.9E − 02	3.4E − 02	−2.0E + 00	4.13E − 02
PCA23	−6.9E − 01	1.5E − 01	−4.5E + 00	6.88E − 06	PCA52	−5.1E − 02	2.0E − 02	−2.5E + 00	1.19E − 02
PCA24	−9.8E − 03	2.7E − 03	−3.6E + 00	2.96E − 04	PCA53	−1.4E − 01	3.9E − 02	−3.7E + 00	2.54E − 04

dataset, containing the remaining 20% of the samples. The multiple linear regression model was performed on the model points dataset.

23.2.4 Digital Soil Risk Assessment Model

Texture class percentages have been derived in four different ways: (1) using the Dominant Soil Surface Texture of the European Soil Database (ESDB); (2) using the estimated texture percentage based on the RK using the texture triangle, (3) as well a Best Case scenario, and (4) Worst Case (minimum clay, maximum sand content) scenario especially tailored for wind erosion. Depending on the various textured materials reported, different friction velocity thresholds were assigned: 5 m/s for coarse, 7 m/s for medium, 9 m/s for fine textured and 11 m/s for very fine textured soils. Organic soils were assigned a threshold value of 8 m/s. Finally, the wind erosion scenarios have been limited to soils which were under agricultural land use according to Corine 2000 land use classification (EEA, 2000b).

The Wind Force Integral (WFI) has been used as a parameter to compare erosivity of the meteorological conditions with respect to the underlying soil surface across different years (Beinhauer and Kruse, 1994). The WFI defines the potential transport capacity of the wind at the soil surface as a function of the wind force and the surface moisture providing that the following conditions are met (i) precipitation < 0.3 mm in that time step (e.g., no rain event), (ii) precipitation (m) $<$ evaporation (m), and (iii) the average wind speed (u) in m s^{-1} is above the soil friction velocity threshold (u_{thr}) in m s^{-1}. If all three conditions are met, the WFI is computed for the time steps (n) per day/month/or year:

$$\text{WFI} = \sum_{1}^{n} (u - u_{\text{thr}}) \times u^2$$

All climate data/WFI computations were performed using the Climate Data Operators (CDO) version 1.0.1 (Schulzweida, 2006). The authors define an erosive day (ED) as a day where the sum of all daily time steps (dn) of the WFI shows values greater than zero.

Weather data have been obtained from daily forecasts from the European Center for Medium Range Weather Forecast (ECMWF) for the years 2005–2007 in the meteorological GRIB format (1.8 GB). The required parameters have been extracted for the Danube Basin and joined to provide a single file with a 3 h time step (180 MB). Spatial resolution of this dataset has been remapped using bilinear interpolation to 0.1 degree resolution from the original 0.25 degree resolution. This remapping constitutes a limitation for the accuracy of the modeling. However no other dataset is available with a sufficient daily temporal and spatial resolution.

Finally, the surface vegetation cover was calculated using the daily surface vegetation cover fraction (FVC) derived from the Meteosat MSG/SEVIRI sensor. It allowed a quantitative classification of the structure of the vegetation canopy. Areas of no-data in the FVC dataset have been filled by time series interpolation. The

original spatial resolution of the dataset (5 km × 5 km) has been resampled to a 0.1 degree resolution to match other datasets. As a rule of thumb based on expert knowledge, we assumed that, for vegetation cover percentages above 20%, wind erosion would be negligible. To account for subscale variability of the FVC we chose a value of 30% as a threshold for wind erosion events.

23.3 Results and Discussion

23.3.1 Results from Digital Soil Mapping Model

The results from the developed stepwise regression model based on the PCA of the 54 auxiliary variables are quite poor with only a multiple R-squared of ∼0.2 for sand, silt and clay. After the stepwise regression we retained for the prediction 41 PCA bands for silt and clay and 46 bands for sand. An analogous analysis for some soil chemical parameters shows an explanation between 40 and 50% (Rodriguez et al., 2008). Most of the variability was not captured by the deterministic model, implying that (i) the auxiliary variables selected were not adequate to estimate soil textural parameters at the basin scale, (ii) soil profile data were non-representative due to poor sampling quality; and/or (iii) the soil profile data related to different scales/periods. The adjusted R-squared values obtained in this work were 0.21, 0.17 and 0.17 for sand, silt and clay, respectively. Still, the final digital soil mapping model is a combination of the regression model and the interpolation by ordinary kriging of their respective residuals. A map of the predicted clay contents is shown in Fig. 23.3.

23.3.2 Accuracy Assessment of Digital Soil Mapping Model

The accuracy of the results was evaluated by comparing the texture values in the validation dataset (2,000 points) against the RK model predictions. The RMSE values for clay, sand, and silt were 8.6, 10.4 and 13.5, respectively. These values seem on the first hand quite high. This uncertainty can be partially explained since a point scale dataset is being compared with averaged texture values for a 1,000 m block (the support size of our RK models). To obtain an estimation of the accuracy of the input data we compared each soil profile with each neighbor in a maximum radius of 2,000 m, obtaining a RMSE of 9.9, 11.9 and 15.6 for clay, silt and sand respectively. Thus, the RMSE values from our model validation were quite similar to the internal variability found in the Danube Basin SIS.

23.3.3 Results for Digital Soil Risk Assessment Model

Estimated days for wind erosion events in 2006 ranged from 1 to 2 for the best case scenario (BC), up to 3 days for the average scenario (AV) and up to 6 days

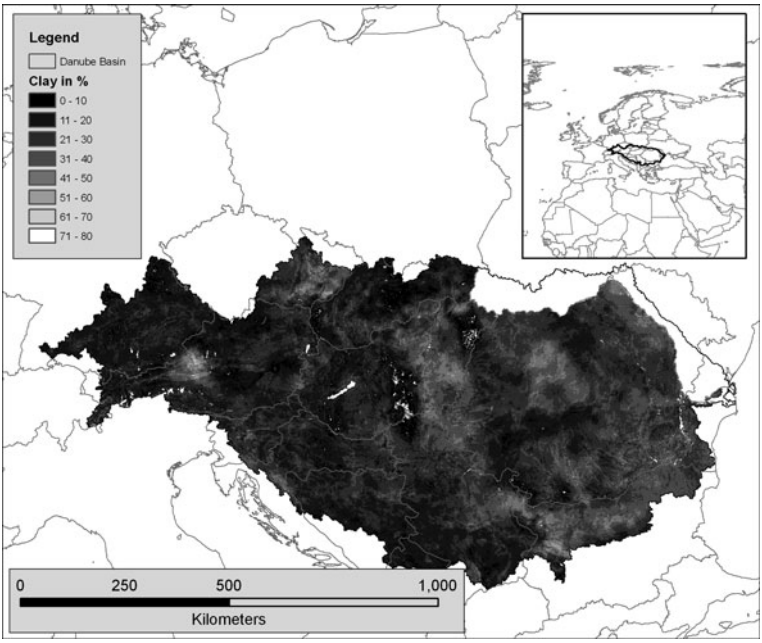


Fig. 23.3 Predicted clay content using regression-kriging. Note that for parts of the Danube Basin in Moldavia and Ukraine no prediction could be performed due to missing input data

for the worst case scenario (WC) which covered a total area of $\sim 14 + \text{E3}\,\text{km}^2$. In the AV scenario (Fig. 23.4), only half of the area would be affected (i.e., $8.5 + \text{E3}\,\text{km}^2$), whereas the BC scenario affected an area of $\sim 2.7 + \text{E3}\,\text{km}^2$. The spatial dimension and magnitude are close to the estimations performed with the ESDB dataset, where $\sim 4 + \text{E3}\,\text{km}^2$ have been influenced with one and two erosion days. The extent of areas affected for Bulgaria, Czech Republic, Hungary and Slovakia is $\sim 2.5 + \text{E5}\,\text{km}^2$ (Funk and Reuter, 2006). This extent would represent the maximum area assuming bare earth dry conditions without surface cover and wind speeds above the threshold velocity. As the modeled area extent is an order of magnitude smaller, the results seem plausible as "near real-time" modeling has been performed.

A validation of the estimated days against real measurements can not be performed as no measurements of wind erosion events were available. A possible approach for validation might be to use background or rural PM10 air quality station data from the different EEA member states. However, especially in Eastern Europe the density of the monitoring stations is rather limited for the year 2006.

Another limitation arises from the applied static threshold for friction velocity based on texture classes. The authors assumed that a threshold friction velocity for a sandy soil is the same across the Danube Basin, which may not be the case due to additional influencing soil components (biological, chemical). For example, organic matter content, which is known to stabilize the soil surface (e.g., crusting), is one parameter which could have been integrated, but was not considered here. Furthermore, besides the more stable soil properties, dynamic properties are more

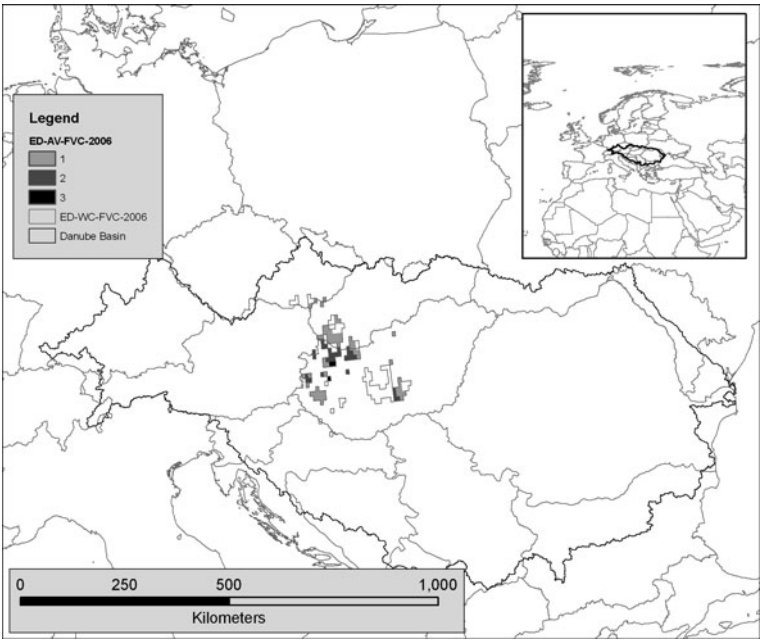


Fig. 23.4 Number of wind erosion days (ED) in 2006 in the Danube Basin using a digital soil mapping generated soil texture map (AV), 3-hourly weather forecast from the ECMWF and daily surface vegetation cover fraction (FVC) using the wind force integral. Note the extent of the Worst Case Soil Texture Scenario (ED-WC-FVC-2006)

difficult to obtain and use for modeling purposes. The approach certainly lacks the incorporation of factors like soil moisture content. Still, the use of evaporation in the WFI improves the results at least for the top soil surface of sandy soils. The authors are aware that any changes in land use and changes in land use cover percentage are neglected between the time of creation of the land use classification and the simulation. Still, these changes should be negligible on the rather coarse simulation scale. Finally, temporal changes in development of soil crusting (Goosens, 2004) and changes in time of the threshold friction velocity have been taken into account, which might be influenced by management activities (ploughing, seeding).

23.4 Conclusions

We performed a digital soil mapping – digital soil functional mapping – digital soil risk assessment workflow for the Danube Basin based on $\sim$ 8,000 single soil profiles and 54 auxiliary datasets for the soil threat wind erosion on a daily basis in the year 2006. To perform such application, multiple data sources had to be fused (climate, FVC, soil profiles, geology, DEM) into the modeling process allowing for "near

real time" wind erosion events prediction. We expect that in the future many more of these multi-data fusion applications will be used.

The digital soil mapping model was not satisfactory with the chosen auxiliary parameters that can only explain a small percentage of the total variability. We can already conclude that for the Danube Basin, estimations for soil physical parameters need a different set of auxiliary parameters than a European soil chemical estimation. Further investigations need to be performed to determine parameters or methods which allow for increased prediction accuracy for the digital soil mapping model. RK and WC scenario showed consistently slightly higher extent and magnitude compared to the ESDB scenario, whereas the BC scenario shows similar values. This is the big advantage of the digital soil mapping/digital soil risk assessment approach - as uncertainties can be specified – to outline differences in results which can be expected for policy relevant decision making. If new soil information becomes available in the context of the GlobalSoilMap.net (see Chapter 33) approach, the method as outlined in here, can be applied elsewhere to support decision making.

The digital soil functional mapping/digital soil risk assessment allowed for an approximation of the wind erosion events in 2006. However it became clear that certain changes in temporal variability of crusting or changes in friction velocity due to management activities require a substantial effort to (i) understand processes at the Danube Basin scale and (ii) to generate models for future scenarios. Such algorithm-coded models are required to perform policy relevant support to decision making at European-wide scales.

References

Bagnold, R.A., 1941. The Physics of Blown Sand and Desert Dunes. Methuen, London, 265 pp.

Beinhauer, R., and Kruse, B., 1994. Soil erosivity by wind in moderate climates. Ecological Modelling 75/76:279–287.

Böhner, J.B., Schäfer, W., Conrad, O., Gross, J., and Ringeler, A., 2003. The WEELS model: methods, results and limitations. Catena 52:289–308.

Conrad, O., 2001. Tools for (grid based) digital terrain analysis V1.0 Program module for SAGA GIS 2.0, Göttingen

Chepil, W.S., 1960. Conversion of relative field erodibility to annual soil loss by wind. Soil Science Society of America Proceedings 24:143–145.

de Jager, A., Rimavičiūtė, E., and Haastrup, P., 2006. A water reference for Europe, pp. 259–286. In: Haastrup, P., and Würtz, J. (eds.), Environmental Data Exchange Network for Inland Water. Elsevier, Amsterdam.

Dobos, E., Carré, F., Hengl, T., Reuter, H.I., and Tóth, G., 2006. Digital soil mapping as a support to production of functional maps. EUR 22123 EN, 68 pp. Office for Official Publications of the European Communities, Luxemburg.

Doll, C.H., Muller, J.-P., and Elvidge, C.D., 2000. Night-time imagery as a tool for global mapping of socioeconomic parameters and greenhouse gas emissions. AMBIO: A Journal of the Human Environment 29:157–162.

European Communities, 2006. Proposal for a directive of the European Parliament and of the Council Establishing a Framework for the Protection of Soil and Amending Directive 2004/35/EC.COM(2006) 232.2004/35/EC.COM(2006) 232. European Commission, Brussels, Belgium.

EEA, 2000a. Down to Earth: soil degradation and sustainable development in Europe. Environmental issues Series, Number 16, European Environmental Agency, 32

EEA, 2000b. Corine land cover 2000 (CLC2000) seamless vector database, http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=950 (Last accessed 29.07.2009)

EUR 19945 EN, 2004. The European soil database V2.0 European Commission, Directorate General Joint Research Centre.

Fryrear, D.W., Saleh, A., Bilbro, J.D., Schomber, H.M., Stout, J.E., and Zobeck, T.M., 1998. Revised Wind Erosion Equation (RWEQ). Wind erosion and Water conservation Research Unit. USDA-ARS, Lubbock, Texas, WEWC Technical Bulletin Nr. 1

Funk, R., and Reuter, H.I., 2006. Wind erosion in Europe, pp. 563–582. In: Boardman, J., and Poesen, J. (eds.), Soil Erosion in Europe. Wiley, Chichester.

Goosens, D., 2004. Effects of soil crusting on the emission and transport of wind-eroded sediment: field measurements on loamy sandy soil. Geomorphology 58:145–160.

Gross, S., and Bärring, L., 2003. Wind erosion in Europe, where and when, pp. 13–28. In: Warren, A. (ed.), Wind Erosion on Agricultural Land in Europe. European Commission, Directorate-General for Research, Office for Official Publications of the European Communities, Luxembourg.

Hagen, L.J., 1996. WEPS, USDA Wind Erosion Prediction System to meet user needs. Journal of Soil and Water Conservation 46:106–111.

Hengl, T., 2009. A Practical Guide to Geostatistical Mapping, 2nd ed. University of Amsterdam, www.lulu.com, 291 p. ISBN 978-90-9024981-0.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., and Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25:1965–1978.

Iwahashi, J., and Pike, R.J., 2007. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. Geomorphology 86(3–4):409–440.

ITTVIS, 2008. Visual Information Solutions – ENVI Release 4.3, www.ittvis.com (Last accessed 5 May 2010)

Lehner, B., and Döll, P., 2004. Development and validation of a global database of lakes, reservoirs and wetlands. Journal of Hydrology 296:1–22

Lóki, J. Rajkai, K., Czyz, E.A., Dexter, A.R., Diaz-Pereira, E., Dumitriu, E., Enache, R., Fleige, H., Horn, R., de la Rosa, D., and Simota, C., 2005. SIDASS project: Part 4. Wind erodibility of cultivated soils in north-east Hungary. Soil and Tillage Research 82(1):39–46.

McBratney, A.B., Santos, M.L.M., and Minasny, B., 2003. On digital soil mapping. Geoderma, 117(1–2):3–52.

Pawlewicz, M.J., Steinshouer, D.W., and Gautier, D.L., 2003. Map Showing Geology, Oil and Gas Fields, and Geologic Provinces of Europe including Turkey. U.S. Geological Survey Open-File Report OFR-97-470-I URL: http://pubs.usgs.gov/of/1997/ofr-97-470/OF97-470I/ (Last accessed 23 April 2010).

Peel, M.C., Finlayson, B.L., and McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrology and Earth System Sciences 11(5):1633–1644.

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. Photogrammetric Engineering and Remote Sensing 57:241–262.

Rodriguez Lado, L., Hengl, T., and Reuter, H.I., 2008. Heavy metals in European soils: a geostatistical analysis of the FOREGS Geochemical database. Geoderma 148:189–199.

Shao, Y., Raupach, M.R., and Leys, J.F., 1996. A model for predicting Aeolian sand drift and dust entrainment on scale from paddock to region. Australian Journal of Soil Research 34:309–342.

Sutton, P., 1997. Modeling population density with night-time satellite imagery and GIS. Computers, Environment and Urban Systems 21:227–244.

Schulzweida, U., 2006. Climate Data Operators- Release 1.0.1, http://www.mpimet.mpg.de/ (Last accessed 5 May 2010)

Thornthwaite, C.E., 1948. An approach towards a rational classification of climate. Geographical Review 38:55–94.
Van Lynden, G.W.J., 1994. European Soil Resources - Current status of soil degradation in Europe: Causes, impacts and need for action, Nature and Environment No. 71, Council of Europe Press, Strasbourg, 93 p.
Wessel, P., and Smith, W.H.F., 1996. A global self-consistent, hierarchical, high-resolution shoreline database. Journal of Geophysical Research 101:8741–8743.

Part IV
Making Digital Soil Mapping Operational

Chapter 24
Soilscapes Basis for Digital Soil Mapping in New Zealand

A.E. Hewitt, J.R.F. Barringer, G.J. Forrester, and S.J. McNeill

Abstract S-map is designed to deliver a new digital soil map, database, inference system, and soil information system for New Zealand. The strategy integrates legacy data and new data with digital soil mapping techniques. Legacy data include a database of analyzed pedons; older soil surveys; and experienced pedologists with knowledge of soil variability and soil–landscape relationships. Given the available spatial prediction layers, digital soil mapping is most suited to hilly and mountainous land. Definition of soilscapes is a key step for effective survey planning – soilscapes will be used to match mapping techniques to land and soil type, analyse legacy data gaps to plan further sampling, and identify potential areas suitable for mining predictive relationships from legacy soil surveys. A two-stage approach is proposed. In stage one first-approximation soilscapes are defined using legacy data, and used to plan and complete a digital soil map. In stage two the new soil data are used to derive improved second-approximation soilscapes. Two trial methods were used to define and map soilscapes: data-driven clustering, and expert clustering from legacy data. The expert clustering result was evaluated by multivariate regression. The factors used to guide clustering and classification included parent rock, elevation, precipitation, slope, and the presence of thick (>2 m) loess. The data-driven-clustering output expressed major soilscape transitions, but in detail produced many inappropriate clusters. An expert-driven-clustering output was less problematic but multivariate regression suggested that the adopted variables did not adequately describe the distinctions made by the expert for different soilscapes.

Keywords Legacy data · Soilscapes · Digital soil mapping

A.E. Hewitt (✉)
Landcare Research, PO Box 40, Lincoln 7640, New Zealand
e-mail: hewitta@landcareresearch.co.nz

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_24,

24.1 Introduction

New Zealand is under increasing pressure from intensifying land use. The profitable dairy industry has grown rapidly over time, to the extent that immature pine plantations have been clear-felled for new dairy operations. For some time New Zealand land use patterns have been highly dynamic as enterprises have responded to market signals in a relatively free regulatory environment. Now however, issues of water quality, water scarcity, green house gas emissions, and soil quality have emerged and land managers are facing pressures of environmental compliance enforced by export customers and domestic law. Research funding has been released to provide better planning and management mitigation models.

How then do we respond? We have a small number of senior pedologists, a rich set of soil survey reports, and a good-quality national soils database. Digital soil mapping techniques hold the promise of being able to deliver new soil information cost-effectively. Clearly the way forward is to take the best of our legacy data and integrate these with the best of our digital soil mapping techniques. This has to be an important topic for this conference and this first session is welcomed.

This paper outlines the nature of the New Zealand soil data legacy and relates our efforts to apply legacy data to the recognition of soilscapes as an aid in the planning of digital soil mapping.

The origin and usage of the term soilscape is reviewed by Jamagne and King (2002). In this paper we follow the soilscape definition of Lagacherie et al. (2001) where a soilscape is "a landscape unit including a limited number of soil classes that are geographically distributed according to an identifiable pattern". In the absence of prior mapping of soils to identify those soil patterns, our aim is to estimate the distribution of soilscapes by mapping areas of similar arrays of related land attributes and environmental factors of soil formation. The concept of soilscapes differs from "soil associations" by explicit inclusion of characteristic landscape attributes.

24.2 Legacy Data

24.2.1 Legacy Datasets in New Zealand

The New Zealand soil survey has a vigorous history (Tonkin, 2007) but it faced drastic funding cuts from 1988. Full-time staff dedicated solely to soil survey and related land resource mapping reduced from 49 to only 0.8 in 2007. The funding drought has now partly eased and we may now build a new digital map using legacy data from the earlier soil survey era.

The legacy data are:

Soil maps and reports National soil map coverage at 1:253,440 scale (inch: 4 miles) was achieved by mapping of soil associations (called soil sets) in 1954 for the North Island and 1968 for the South Island (Soil Survey Staff, 1954, 1968). More detailed surveys followed at scales from 1:125,000 to 1:10,000, of counties,

prominent arable areas, and parts of catchments. This legacy comprises a patchy distribution of soil maps of varying scale, varying quality, and in parts poor soil correlation. Section 16.1 further describes the characteristics of legacy data. Other studies that have used legacy data are referred to in Section 24.1 above.

An independent land inventory and land capability survey, the New Zealand Land Resource Inventory (NZLRI) (NWASCO 1979), was conducted rapidly to achieve national coverage in 1979. It mapped five "themes": soil, slope, erosion status, vegetation, and rock type. This compiled all soil information published at the time, to a scale of 1:63,360. The soil theme of the NZLRI remains the only digital national soil coverage and receives extensive use. A relational join of features from NZLRI, and the National Soils Database (NSD), with input of expert judgement, allocated soil chemical and physical attribute classes to polygons of the NZLRI. Sixteen attribute fundamental soil layer (FSL) maps were generated and are accessible through the geospatial data integration portal: http://gisportal.landcareresearch.co.nz/webforms/catalogue1.aspx.

Pedon databases. The National Soils Database has soil chemical analyses from sampled pedons at 2,500 sites. Only 13% have soil physical analyses (bulk density, water release, etc.). A further 800 sites have topsoil soil-quality-indicator analyses (Sparling and Schipper, 2002). Samples are archived and available for further analysis. In recent years, the soil carbon database has added 410 sites analyzed for carbon and bulk density but potentially available for other analyses from archived samples.

Pedologists. Apart from the NZLRI and fundamental soil layer compilations, the soil survey legacy data remain underutilized because they are scattered among many reports, in which the data are inconsistent, qualitative, and with a nomenclature that can be interpreted only by specialists. A key to extracting the richness of these data is to use senior pedologists who understand the soils, landscapes, and the nature of the data. Experienced pedologists are part of the legacy.

24.3 S-map

24.3.1 S-map Goal

S-map is a national soils database, soil map and soil information inference system for New Zealand (Lilburne et al., 2004). The goal is one complete national digital soil map with easily accessible, relevant data and inferred key information to support sustainable development and management. The map scale will be at 1:50,000, or more detailed.

S-map is designed to incorporate legacy data (Hewitt et al., 2006) by taking what is best from the past and incorporating both new data and pedologist knowledge. Legacy data are transformed into data that are interpreted by an inference system of pedotransfer functions and used to generate soil information outputs including Web-accessible, dynamic, soil fact sheets (http://smap.landcareresearch.co.nz). Pedologists sift through old soil reports extracting soil profile data in the form of newly defined soil functional horizons, estimate simple probability distributions of key

attributes within classes, and estimate proportions of classes within new mapped polygons. Uncertainty ratings are attached to these estimates to express the relative contribution of data and judgment. Uncertainty estimates are vital for the continuing progress of the new database because when national coverage is achieved, parts with lower certainty may then be prioritized and become the focus for further database investment.

24.3.2 S-map Strategy

Because effective spatial predictors are scarce in low-relief land, opportunities for digital soil mapping are limited. Remote sensing solutions have yet to be adequately investigated, but meanwhile soil survey is proceeding with conventional methods. In land of moderate to high relief the national digital elevation model is the primary source of spatial prediction layers together with climate surfaces, land cover and land use. In the high-relief land, rule-based models are being used to relate soil associations to land elements modelled from the 25-m-resolution DEM (Barringer et al., 2008; Schmidt and Hewitt, 2004; Schmidt et al., 2005). Work proceeds on the derivation of higher resolution DEMs using data from the PRISM sensor of the ALOS satellite.

Definition and delineation of soilscapes is a key step for effective planning and conduct of an S-map strategy to map the soils of New Zealand.

24.4 Soilscape Rationale

Soil mapping is more efficient if planned and focused on areas defined by soils and landscapes rather than land administrative boundaries. We then need to stratify New Zealand into soil and landscape areas that will most effectively serve operational soil mapping. These strata, which we call soilscapes, should ideally delineate similar soil–landscape predictive relationships, sampling strategies, survey methods, and likely land use versatility. See Section 28.2 for a discussion of an analogous concept (landscape stratification).

The primary digital soil survey planning applications of soilscapes include (1) matching mapping techniques to land and soil type, (2) analyzing the density of legacy observations to locate gaps for further sampling effort, and (3) indicating potential areas that may be mapped by extrapolation of soil–landscape relationships mined from quality legacy soil surveys. The value of an investment in data mining may be judged by the size of the potential area of extrapolation.

An important secondary application of soilscapes is for planning and implementing sustainable development and land management. Many applications of spatial soil information require greater spatial and categorical generalization than provided by soil maps of 1:50,000 or more detailed scales. Examples of such applications of soilscapes include: recognition of areas of similar hydraulic behavior in catchment studies (Elsenbeer, 2001), definition of study areas for pedotransfer function research (Scheinost et al., 1997), the presentation of national soil resources

(National Soil Resources Institute, 2001), and its inclusion in the anthropological concept of cultural soilscapes (Wells, 2006). Soilscapes have even been used as a defining topic for artistic expression (Noller, 2009).

24.5 Methods

To achieve these goals a two-stage approach is proposed. In stage one first-approximation soilscapes are defined using legacy data, and used to plan and complete digital soil mapping coverage. In stage two the new soil data are used to derive improved second-approximation soilscapes. We are at the first-approximation-soilscapes stage.

We chose a collection of variables to classify and map soilscapes based on the assumption that similar soilscapes should display similar suites of soil types, and be subject to similar factors that drive soil pattern. One of the aims of this study, then, was to determine whether this collection of classification variables adequately describes the similarity within soilscapes, and distinctions between soilscapes, and if the collection of variables adequately codified the expert's knowledge of soilscapes.

Significantly different suites of soils were recognised by indicator soil orders or soil groups of the New Zealand Soil Classification (Hewitt, 1998). For example Semiarid Soils, Pallic Soils, Brown Soils, and Podzols were used to indicate the climate sequence that dominates the soil pattern of the South Island. Each of these indicator classes was associated with a suite of cohorts that typically included Gley Soils, Recent Soils, and Raw Soils. The soil data inputs were soil sets as compiled by the NZLRI from the 1:253,440-scale soil maps of the North Island and South Island (Soil Survey Staff, 1954, 1968).

Significant factors that control soil patterns in New Zealand are rock type, tectonic uplift rate, and rainfall (Basher and Tonkin, 1985). To infer these with available mapped data we used classes of rock composition (classes aggregated from rock types by silica, nickel, and calcium carbonate contents), elevation, precipitation, slope, and the presence of thick loess.

In this paper, we describe three methods for exploring the current soils database and the factors that we assume will discriminate soilscapes. First, we trialed a data-driven clustering analysis in an attempt to aggregate previously classified soil associations with essentially similar soilscape characteristics. Second, we used an expert's knowledge to partition the soil types into a modest number of similar groups based on landscape interpretation. Finally, we used the expert's aggregate partition of soil types in a regression analysis to find explanatory distinctions for model structure within the potential soilscapes.

24.5.1 Data-Driven Clustering

Although the term can have several different meanings, here we take the term clustering to be concerned with discovering groupings among the different cases available to us (Venables and Ripley, 2002). The clustering procedure used a

hierarchical data-driven method with hierarchical cluster routines from the *cluster* package (Maechler et al., 2005) in the R statistical computing environment (R Development Core Team, 2009). Hierarchical clustering was used to agglomerate 1:50,000-scale polygons of the NZLRI, using as inputs the rock and indicator soils, as well as unprocessed ordinal data for precipitation, altitude, and slope. The North Island and South Island were analyzed separately.

A practical limitation of this approach was the very large number of NZLRI polygons, since a data-driven approach such as this requires synthesis of a distance matrix of size N^2 for a dataset of length N. To avoid this difficulty, the polygons were reclassified into groups based on the soil set classes and parent rock composition. For each of these classes, mean values were estimated for precipitation, altitude, slope and rock composition, nickel content, and loess depth, yielding 437 classes for the South Island and 543 for the North Island. The cluster analysis in R was used to determine similarity of these classes based on their average properties. Maps were prepared for assessment by truncating the cluster trees at different levels in the hierarchy to control the level of disaggregation in the landscape.

24.5.2 Expert Clustering

Soil map units were aggregated from the1:253,440-scale legacy soil sets. A total of 546 soil sets for the South Island (Soil Survey Staff, 1968) were aggregated into 55 soilscapes. Aggregation was based on the same factors used in the data-driven clustering method, with the addition of landform dissection, and precipitous slopes.

24.5.3 Multivariate Regression Exploration of the Expert Clusters

Two aggregate soilscapes produced by the expert clustering were analyzed in a regression or classification procedure (as appropriate) that attempted to predict the categorical soilscape within the expert's classification of soilscape. This process used six explanatory variables, including parent rock type, precipitation, altitude, and slope, as well as a variety of regression models (e.g. linear model, tree regression). The models were tested for single potential soilscapes; that is, a model was tested for its ability to predict a soilscape given the other explanatory variables. Then, models were tested between groups; that is, tested for their ability to predict several soilscapes. The model to test for single potential soilscapes was generally simpler, since testing between groups requires the modeller to decide how to apportion the error between different soilscapes. The quality measures for this step included the false classification rates for the different soilscapes.

24.5.4 Distribution of Legacy Data by Soilscape

The distribution of quality soil surveys and pedons of the National Soils Database were compared with soilscape areas. Soilscapes represented in some part by a

quality soil survey were identified as potential areas for extrapolation of the legacy data by data mining.

24.5.5 Judging Success of Outputs

We judged the success of soilscapes qualitatively by degree of fit of soilscape classes and their map delineations to known transitions in indicator soil classes, soil patterns, and soil–landscape relationships. The level of stratification was judged by assessing relationships with well-understood conceptual soil–landscape models. If a soilscape spanned areas where two or more distinct sets of soil–landscape relationships exist then it was judged to be understratified. Alternatively, two or more potential soilscapes that subdivided areas of similar soil–landscape relationships were judged to be overstratified.

24.6 Results

24.6.1 Data-Driven Clustering

Data-driven clustering expressed well the major transitions across the North and South islands, but there were many map units that did not adequately stratify expected soil–landscape relationships. These resulted from the relative priority assigned by the cluster technique to the input factors. For example, high-altitude landscapes were clustered by elevation at high level in the hierarchy irrespective of rock types that control soil types and patterns. Clearly, the hierarchy requires judicious trimming.

Parts of the map were understratified. For example, on the Canterbury Plains (Figs. 24.1 and 24.2) areas of Pleistocene gravel outwash sheets were not separated from the contrasting highly variable Holocene fluvial landscapes. Generally the data-driven clustering performed poorly in flat or gently undulating land because spatial predictors adequate for discrimination of soilscapes were either lacking or not employed for that land.

24.6.2 Expert-Driven Aggregation of Legacy Data

The expert-driven map output was less problematic than the data-driven clustering output. This was to be expected because of the inherent bias in the expert-driven method. Whereas the data-driven clustering method understratified soilscapes on the Canterbury Plains (Fig. 24.2b), the expert-driven method overstratified them by delineating soil texture and depth patterns within the Holocene fluvial landscapes (Fig. 24.2a). It mapped out soil classes (that would be predicted by a soil–landscape model) rather than merely estimating the area where that model should be applied.

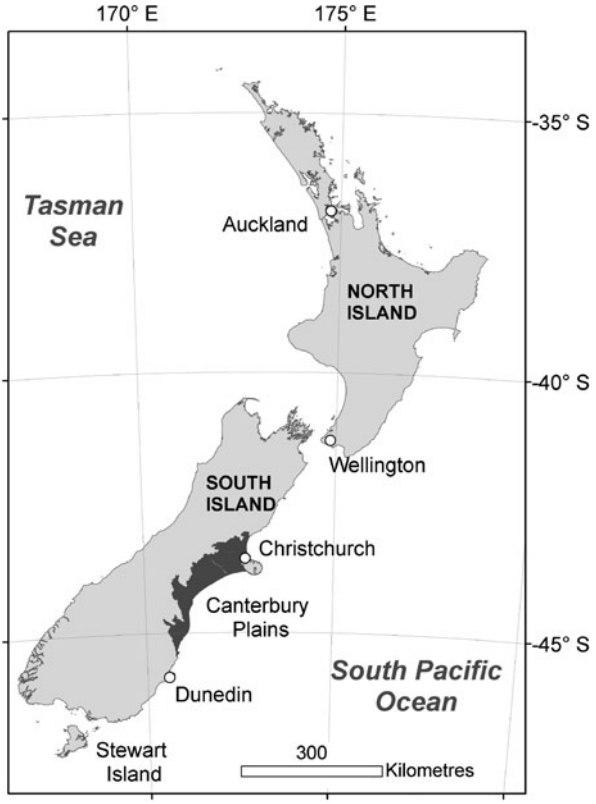


Fig. 24.1 Map of New Zealand. The Canterbury Plains feature in more detail in Fig. 24.2

24.6.3 Multivariate Regression Exploration of the Expert Clusters

Since the multivariate regression approach uses a combination of machine and expert analysis, its assessment contains elements of the results of Sections 24.6.1 and 24.6.2 above. The explanatory factors were highly collinear, so they were transformed using principal components analysis and regression was based on all available principal components. In general, a single soilscape could be predicted with good accuracy using the regression approach within a single soilscape (typically better than 75% classification accuracy on a per-polygon basis). This result suggests that the set of adopted explanatory variables adequately captured the variability of a single class. However, adding additional soils to the regression or classification procedure resulted in considerable misclassification. This suggests that the adopted variables do not adequately describe the distinctions made by the expert for different soilscapes.

There are several possible reasons for this result. First, there may be too few explanatory variables available to describe the distinctions between soilscapes. Second, the measures used to codify the presumed logic of the expert in the distinction of each soilscape may have been faulty. For example, using mean altitude as an

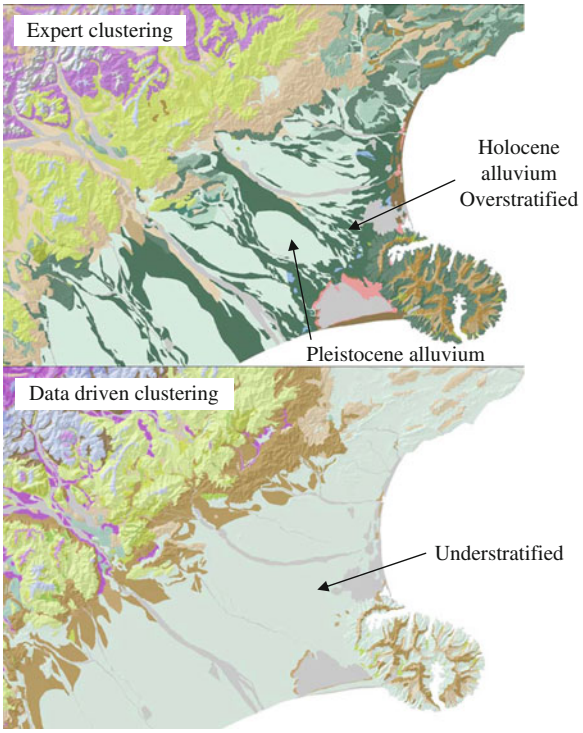


Fig. 24.2 Maps of provisional soilscapes for a portion of the Southern Alps, Canterbury Plains, and Banks Peninsula, South Island. (**a**) The expert-driven clustering over stratified Holocene alluvium areas of the plain where one clearly recognised soil landscape relationship was split into two soilscapes. (**b**) The data-driven clustering method understratified the Canterbury Plains because more than one distinct set of soil-landscape relationships exist

input variable may fail to adequately distinguish soilscapes that were distinguished primarily by change in altitude.

24.6.4 Distribution of Legacy Data by Soilscape

Of the 55 soilscapes provisionally defined for the South Island, 11 occupied areas covered in part by quality soil surveys with potential for data mining. These were areas of strongly rolling, hill or steep land where the available DEM may be readily used. A further 9 soilscapes on low-relief land had potential for data mining if remote sensing techniques can be applied.

24.7 Conclusions

Comparison of outputs from the two methods will be useful in producing a satisfactory first-approximation soilscape coverage for New Zealand. The data-driven

clustering output both confirmed and challenged the groupings of the expert-driven clustering technique. The regression analysis refined this by examining the integrity of the expert clusters. Reappraisal of the clusters will prompt a revision of the expert-driven aggregation incorporating these insights.

Although inclusion of new input factors, prior processing of factors, weighting, and improved pruning of the cluster hierarchy should result in a better product, it will still carry elements of subjectivity in the choice of input attributes and their weighting. High accuracy for the first-approximation soilscapes cannot be expected because of comparative lack of knowledge at the beginning of this process. It is accepted that in their application in planning digital soil mapping it is sufficient that the soilscapes are only indicative.

This study has proposed that the primary purpose of a first-approximation map of soilscapes lies in its application as a digital soil mapping planning tool. The soilscapes will provide a template for application of the digital soil mapping planning methods described in Chapter 34.

We conclude that where legacy data and experienced pedologists with knowledge of different patterns of soil variability across the landscape are available, then it is effective to use a method for mapping soilscapes that is either totally expert driven or data driven with strong expert supervision. It is important that the living legacy of senior pedologists is captured while that legacy still lives. Involvement in soilscape mapping and evaluation of results is one way this can be achieved.

The purpose of second-approximation soilscapes lies in providing a generalization of patterns of soil variability and behavior for wide-ranging production and environmental applications of soil information. For these consequent applications, the soilscapes used to plan soil digital mapping may be updated and modified on the basis of new evidence afforded by the newly generated digital data.

Acknowledgments Funding for this work was provided by Foundation for Research Science and Technology, contract CO9X0306. We are grateful to Alfred Hartemink and Ian Lynn for their comments on the manuscript.

References

Barringer, J.R.F., Hewitt, A.E., Lynn, I.H., and Schmidt, J., 2008. National mapping of landform elements in support of S-Map, a New Zealand soils database, pp. 443–458. In: Zhou, Q., Lees, B., and Tang, G. (eds.), Advances in Digital Terrain Analysis. Springer, Dordrecht.

Basher, L.R., and Tonkin, P.J., 1985. Soil formation, soil erosion and revegetation in the central South Island hill and mountain lands, pp. 154–169 . In: Campbell, I.B. (ed.), Proceedings of the Soil Dynamics and Land Use seminar, Blenheim, New Zealand, May 1985.

Elsenbeer, H., 2001. Hydrologic flow paths in tropical rainforest soilscapes – a review. Hydrological Processes 15:1751–1759.

Hewitt, A.E., 1998. New Zealand Soil Classification. Landcare Research Science Series No.1, 2nd edition, Manaaki Whenua Press, Lincoln, New Zealand.

Hewitt, A.E., Lilburne, L., Lynn, I.H., and Webb, T.H., 2006. Capturing heritage soil survey data for pedometric analysis and modeling: the S-map approach. Abstract to Proceedings of the World Congress of Soil Science, Philadelphia, July 2006.

Jamagne, M., and King, D., 2002. The current French approach to a soilscapes typology, pp. 157–178. In: Eswaran, H., and Ahrens, R. (eds.), Soil Classification: A Global Desk Reference. CRC Press, Boca Raton, FL.

Lagacherie, P., Robbez-Masson, J.M., Nguyen-The, N., and Barthès, J.P., 2001. Mapping of reference area representativity using a mathematical soilscape distance. Geoderma 101:105–118.

Lilburne, L., Hewitt, A.E., Webb, T.H., and Carrick, S., 2004. S-map: a new soil database for New Zealand. In: SuperSoil 2004: 3rd Australian New Zealand Soils Conference, University of Sydney, Australia, www.regional.org.au/au/asssi/supersoil2004/ (Last accessed 21 April 2010).

Maechler, M., Rousseeuw, P., Struyf, A., and Hubert, M., 2005. Cluster analysis basics and extensions. Unpublished package for R software. http://cran.r-project.org/web/packages/cluster/index.html (Last accessed 21 April 2010).

National Soil Resources Institute, 2001. NATMAP Soilscapes: Easy Soils Map. LandIS, Cranfield University, Cranfield, UK, http://www.landis.org.uk/data/nmsoilscapes.cfm (Last accessed 21 April 2010).

NWASCO, 1979. Our land resources. Bulletin accompanying NZLRI worksheets. Wellington, National Water and Soil Conservation Organisation. 70p.

Noller, R., 2009. The art and science of soilscape painting. http://www.soilscapes.blogspot.com (Last accessed 21 April 2010).

R Development Core Team, 2009. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (Last accessed 19 April 2010).

Scheinost, A.C., Sinowski, W., and Auerswald, K., 1997. Regionalization of soil water retention curves in a highly variable soilscape, I. Developing a new pedotransfer function. Geoderma 78:129–143.

Schmidt, J., and Hewitt, A.E., 2004. Fuzzy land element classification from DTMs based on geometry and terrain position. Geoderma 121:234–256.

Schmidt, J., Tonkin, P.J., and Hewitt, A.E., 2005. Quantitative soil-landscape models for the Haldon and Hurunui soil sets, New Zealand. Australian Journal of Soil Research 43:127–137.

Soil Survey Staff, 1954. General survey of the soils of the North Island, New Zealand. Soil Bureau Bulletin 5. DSIR, Wellington. 286 p., including maps.

Soil Survey Staff, 1968. General survey of the soils of the South Island, New Zealand. Soil Bureau Bulletin 27. DSIR, Wellington. 404 p., including maps.

Sparling G.P., and Schipper, L.A., 2002. Soil quality at a national scale in New Zealand. Journal of Environmental Quality 31:1848–1857.

Tonkin, P.J., 2007. A history of soil survey and selected aspecs of soil conservation in New Zealand. New Zealand Soil News 55:59–70 (part one), 102–115 (part two).

Venables, W.N., and Ripley, B.D., 2002. Modern applied statistics with S. Springer, New York, 495 p.

Wells, E.C., 2006. Cultural soilscapes, pp. 125–132. In: Frossard, E., Blum, W.E.H., and Warkentin, B.P. (eds.), Function of Soils for Human Societies and the Environment. Geological Society, London.

Chapter 25
Legacy Soil Data Harmonization and Database Development

E. Dobos, T. Bialkó, E. Micheli, and J. Kobza

Abstract Many countries completed large scale (1:5,000 – 1:25,000) soil surveys decades ago, and have since used their thematic and geographic information to derive thematic soil property layers of the same or smaller scale (1:100,000 and smaller). The new layers are often simply aggregates of the original soil polygons and inherit the same geographic relationships that were delineated in the original data source. In reality, this approach does not use all information of the input data.

Instead of aggregating existing maps, the original, non-interpreted field survey point data can be gathered and used for deriving new property layers. The paper aims to summarize a soil database development project using legacy data for a transboundary area, representing two different systems of data collection, storage and management. Recent and archived soil profile data have been collected, including monitoring sites, soil nutrient status campaign data for different periods, and recorded soil profiles from previous soil mapping activities. These data sources have been transformed to have a common theoretical basis using commonly accepted pedotransfer rules and an integrated profile database has been formed. It was used to interpolate soil information and develop soil property maps and layers representing the WRB diagnostic properties and horizons. The creation of the property layers was based on statistical/geostatistical interpolations of the soil profile database using DEM derivatives, SPOT and Landsat satellite images as covariates to provide information for the natural setting of the area. The interpolated values for the numeric variables were estimated using regression kriging, while the classified variables were calculated using the maximum likelihood classification algorithm. It was concluded, that the development of WRB diagnostic criteria database is feasible using raw data of different origin and a set of harmonization and digital soil mapping tools.

Keywords WRB · Soil database development · Data harmonization · Remote sensing · Digital terrain modeling

E. Dobos (✉)
University of Miskolc, Miskolc (Egyetemváros), 3515, Hungary
e-mail: ecodobos@uni-miskolc.hu

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_25,

25.1 Introduction

Soil data of appropriate format and reliable accuracy are often the most limiting factor of soil-related modeling and applications. Many countries have had several data collection campaigns serving different goals, like mapping or agricultural fertility testing. Besides a Canadian example using legacy data in a digital database (see also Chapter 26), legacy data have also been used for several digital soil mapping applications to derive updated information (Baxter and Crawford, 2008; Bernoux et al., 2007; Dobos et al., 2007; Mayr and Palmer, 2007; Mayr et al., 2008; Rossiter, 2008). The integration of several legacy data sources is a potential way to create a product with great value added without the need of strong field data collection. One of the key elements of database development is the appropriate density of input calibration/training data (see Section 29.3). However, the integration of interpreted maps is often difficult. Thus, a different approach is demonstrated here. A point database was created from each input dataset and an integrated, multi-origin point database was developed after the necessary harmonization and data filtering. Taxonomic harmonization was done using the WRB 2006 classification system (IUSS Working Group WRB, 2006). This database is used as calibration and training datasets for several digital soil mapping tools.

25.2 Materials and Methods

25.2.1 Study Area

The study area, called Bodrogköz, is located between the triangle of the Tisza, Bodrog and Latorica Rivers along the eastern section of the Hungarian-Slovakian border (Fig. 25.1). It represents a homogeneous landscape, a flood plain with some windblown sand dunes, typical for the Pannonian plain. The areas along the major rivers, the so-called natural levees, are a few meters higher than the area behind (backwater area). Its soil texture is much coarser than the backwater area, which is heavy clay overlaying the deeper sand strata. One to two meters of relative elevation difference results in different texture, chemical properties, and also soil and landuse types. The recent landscape-landuse-geomorphologic-parent material system of the area is very much interrelated and defines the soils in an almost deterministic manner. There is also a slight change along the NE-SW direction, which is the major axis driving the surface water flow as well. The NE edge is higher, while approaching the SW edge the elevation tends to be lower, have more frequent flooding, hydromorphic impact, thus more leaching, lower pH, and higher humus content. This trend varies a little bit in the SW edge, where the two levees of the Tisza and the Bodrog meet and form a joint levee, with somewhat higher elevation, different water regime and coarser texture.

The area has been cultivated for over a thousand years, with a strong intensification starting in the nineteenth century. The soils were developing under the strong

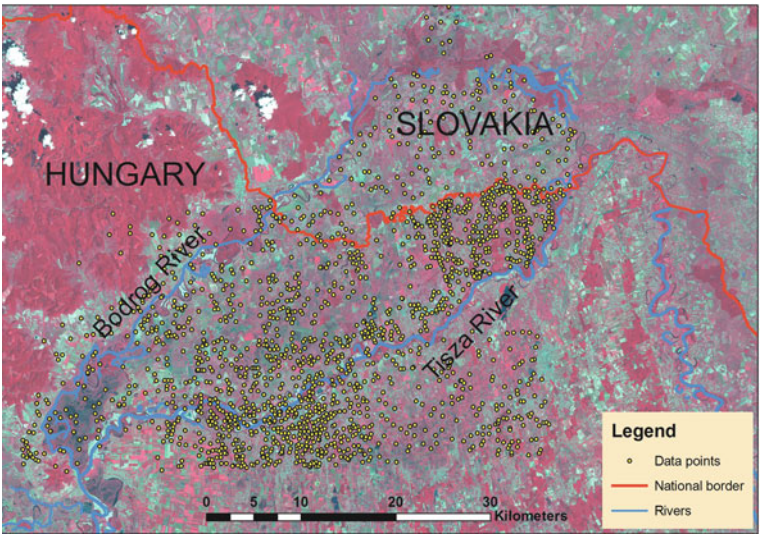


Fig. 25.1 The location of the study area and the sampling sites (in *yellow*) over a Landsat RGB composite image (Bands 4,3,2)

impact of floods and high groundwater table. Flood protection and drainage systems have been constructed since the second half of the nineteenth century, which has changed the environmental system dramatically.

The landuse and the soil type are highly correlated. Low lying and high ground water areas have pasture and Gleysols on them, while the areas with lower ground water table have Vertisols, Arenosols and Luvisols. These soils are cultivated despite their high acidity and unfavorable textures.

The study area has temperate climate with an annual precipitation of 550 mm and a mean temperature of 10°C. The altitude of the majority of the study area ranges from 90 to 120 m. Only two small volcanic hills arise from the plain and reach 270 m. The parent material is mainly alluvial clay and loamy fluvic material. The dominant landuse is farmland with some orchard, forest spots and wet pastures. The most common soil types are Vertisols, Arenosols, Gleysols, Fluvisols and Luvisols (Dobos and Kobza, 2008).

25.2.2 Digital and Field Data

25.2.2.1 Point Data

The area has 1,786 sampling sites, of which 1,616 fall to the Hungarian and 164 onto the Slovakian side (Fig. 25.1).

The highest number of points was imported from the Kreybig mapping campaign. These points were chosen as representative and complementary profiles for

the 1:25,000 scale mapping, started in the late 1930s. A total of 1,161 points were processed, digitized and revalidated (Szabó et al., 2005). Five parameters, namely the 5 h capillary rise, pH(KCl), humus%, $CaCO_3$, and salt content were assigned to each point.

The 164 Slovakian sites were part of the 1:10,000 mapping campaign started in the 1980s and contained three parameters, namely the humus%, clay% and pH(KCl).

Official monitoring sites for both countries (18/6, Hu/Sk) were also used for interpolation and for data harmonization with the variables of $CaCO_3$, texture, humus%, and pH(KCl) (Várallyay et al., 1995).

Data from three soil nutrient survey campaigns (TVG) in Hungary between the late 1970s and 1987 were used as well. A total of 422 data points were generated having the following variables: the Arany-type cohesion measure (K_a), humus%, salt content, $CaCO_3$ and pH(KCl).

An additional 16 sites were also sampled as representative, calibration data – benchmark soil sites – and used in the harmonization process. The majority of these points were selected to revisit existing points of other datasets. These points were sampled and lab analyzed for the humus%, texture, pH, $CaCO_3$ and salt content.

Point Data Derivation from Averaged Field Data (TVG Data Processing)

In order to increase the data point density for areas where no reliable point data source was available non-point data sources were used as well, namely the TVG data. The TVG dataset is a non-point, field-based dataset, with 8 non-located composite samples taken along a recorded transect. Their average was assigned to a parcel, or a part of it, with the size ranging from 10 to 20 ha. These data were first filtered for field homogeneity and only data representing homogeneous fields were processed and used in this project. Field homogeneity was tested in two ways. First, by looking at the site visually on orthophotos, SRTM terrain derivatives, and multitemporal/multispectral Landsat/SPOT/IKONOS images representing six different dates. Quantitative methods, like the spectral distance based region grow algorithm of the ERDAS Imagine was tested as well (ERDAS, 1999). However, due to the high diversity and variability of the input layers no successful method to define the thresholds has been developed yet. Thus, the thresholds were increased continuously to the point when the expert and measured results matched. If a match was not obtained within a certain range of threshold values, then the test failed. The second test was to check for deviation and outliers along the area selected in the first step. The measurements obtained along transects within the selected areas were collected and recorded. The acceptable absolute deviation from the average was set by expert judgment for each variable, such as 0.5 for the pH. All measurements having greater deviation than the set value were considered outliers. If any outlier was identified then the area was dropped. If the tests were passed then the center or the most representative point of the area was selected and the average value was assigned to it. Data for 422 points were generated in this way having the following variables: the Arany-type cohesion measure (K_a), humus%, salt content, $CaCO_3$ and pH(KCl).

This procedure unavoidably introduces some uncertainty; therefore it was used only for areas having limited data.

Point Data Harmonization

The European Union and its member states have several different ways of collecting and analyzing soil samples, and different ways of expressing the results. Therefore using these data sources is not straight forward. Much preprocessing is required to import all of these data into the same reference system. The preprocessing means both the spatial and the attribute data are transformed into a common system. This procedure is called, in our terminology, "harmonization".

The first step of the harmonization procedure was the field work, when representative profiles were opened in the field, sampled, analyzed in lab, and classified according to WRB 2006. The sites were selected to represent the major reference/benchmark soils of the area. The site selection was based on existing soil maps, satellite and orthophoto images and on the major geomorphologic units. The joint field work was a crucial step for mentally harmonizing the group members from the two countries, to reach a common understanding of the soil variable interpretation and to develop a mental model of soil variability. Based on the expert/local knowledge learnt from the reference profiles, each input data type was translated to a common variable using existing transformation models or correlation functions developed within the project. The result of this section was a harmonized soil profile database, and a mental model of the soil resources.

Two major variables needed significant effort to harmonize, namely the taxonomic groups (WRB major reference groups, diagnostic horizons and criteria) and the texture. The taxonomic units were identified manually by the country representatives after field harmonization of the interpretation of the diagnostic properties. Numerous misclassified profiles were identified, screened and replaced by a commonly agreed unit. This work was crucial, and much less time-consuming than anticipated. Having the mental model and the field correlation efforts, it was quite easy and fast to screen the problematic profiles and modify/correct their classification units.

The property having the highest representation diversity was the texture. Clay % content, capillary water rise in 5 h, Arany-type cohesion measure (K_a) and interpreted texture classes were the input types of the different sources. Correlation rules developed by Buzás (1993) were employed to reach the common platform and convert all properties into the same variable. The less detailed variable, namely the classified texture unit was chosen to serve as the final variable, to which we could adjust/degrade the more detailed parameters. The correlation table is given in Table 25.1. The rest of the given parameters (humus, $CaCO_3$ and salt %) were in the same units and were analyzed in the same way, so no further thematic harmonization was needed.

Due to the temporal diversity of the input data sources some changes might have happened in the chemical properties over time and could result in a shift of the data values, which could significantly decrease the model performance. Therefore a set of *t*-tests for the humus content and pH were calculated to make sure that all

Table 25.1 The correlation table of the three input texture parameters

	Clay%	Ka	Capillary water rise in 5 h (mm)
Coarse sand	Below 5	Below 25	Over 350
Sand	5–15	25–30	300–350
Loamy sand	15–20	30–37	250–300
Loam	20–30	37–42	150–250
Clay loam	30–40	42–50	75–150
Clay	40–45	50–60	40–75
Heavy clay	Over 45	Over 60	Below 40

input data sets represent the same population. The values of these two classes were close to normally distributed, skewness 0.6 and 0.5, while the Kurtosis was 3 and 3.3 for the pH and the humus, respectively. Because of the specific environmental setting of the study area – where parent material expresses the geomorphology and the terrain influence on the soils in the same time – a harmonized and simplified quaternary geology database was used to pre-stratify the area. The simplified parent material dataset contained four units: Holocene alluvial clay, Aeolian Dune Sand, Holocene reworked clay-loam alluvium and recent loamy, loamy-sand alluvium. The populations of the different point sources falling into the same parent material class polygons were tested for having the same means at a level of significance equal to 0.2. This value was chosen as a lowest acceptable level. Both of the humus and pH tests were significant.

25.2.2.2 Other Digital Data Sources

Two Landsat and two SPOT images were selected for the work, both representing different seasons and natural conditions. The SPOT images were taken in May and October of 2006, while the Landsat images were acquired in March, 1999 and in July 2006. The1999 image represents a flooded condition. These data sources were combined into a 22 band image, resampled to 120 m and used as covariates for the interpolation and classification procedures. The pixel size degradation was carried out to decrease the impacts of artificial landscape patterns and increase the importance of the overall environmental condition.

High resolution digital data for validating the sites were also used. Digital orthophotos from the summers of 2002 (for the Slovakian side) and 2005 (for the Hungarian side) with 2 m resolution were created to cover the entire study area. An IKONOS multispectral image with 4 m resolution was also acquired for the entire area for the summer of 2007, when the field sampling campaign was running.

Terrain Information

The terrain was represented with the 90 m resolution SRTM data. These data were preprocessed to remove the effect of forests, which was recognized as a major limitation factor. The removal required a forest coverage map. It was created using the SPOT images described above and field training samples. The training samples

were taken based on high resolution orthophotos. Maximum likelihood classification algorithm was employed to classify the entire image. The classified image was resampled to the same resolution as the SRTM and then reclassified into two classes, forest and non-forest. This image was used to identify the forest plot edges and an estimated elevation difference was calculated based on the minimum and maximum values within a given size of search window. This edge contour with the estimated elevations was used for lowering the actual SRTM (with the canopy) data. The resulting image was used for the terrain characterization.

Except for the two hills, the area is almost totally flat. Thus the absolute elevation and other commonly used parameters provided no useful information. Therefore two other topographic parameters were tested to highlight the relative elevation, namely the low-lying and the elevated areas; the Topographic Position Index (TPI) (Weiss, 2001) and the Potential Drainage Density (PDD) index (Dobos and Daroussin, 2007).

25.2.3 Inference Models

Figure 25.2 shows the flowchart of the inference system. The work had three major sections. The first step was the input data harmonization and the creation of the training/calibration point dataset. The second section was the creation of continuous property layers for the final and intermediate layers, like WRB Reference Soil Groups (RSG), texture, pH(KCl) and texture. (Alternative approaches for estimating soil properties based on legacy data of various origin are described in Sections 16.2, 29.2, and 32.2.) After checking for potential trends, Universal Kriging and cokriging

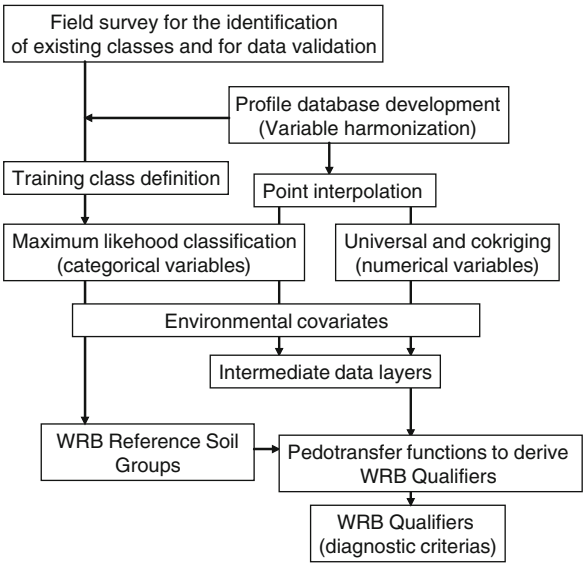


Fig. 25.2 The flowchart of the inference system used within this project

Table 25.2 The pedotransfer functions used for predicting the WRB qualifiers

Predicted WRB qualifiers	Pedotransfer functions
Vertic	All areas where Vertisols exist
Mollic	Humus>1% and Eutric
Arenic	Having sandy texture
Clayic	Having clay texture
Gleyic	All areas where Vertisols, Fluvisols and Histosols occur
Dystric	pH(KCl)<5
Eutric	pH(KCl)>5
Calcic	$CaCO_3$ % > 5

were used to interpolate the numerical data, namely the pH(KCl) and the humus content. Co-variables for the cokriging were selected by checking the cross-correlation of the variable to predict and the terrain parameters derived from the SRTM, and the best two were used. For the pH, 1,611 observations were used and Universal kriging was selected as best performing model. The humus content was estimated with Universal cokriging using PDD as covariable with 657 observations.

Categorical variables, like the WRB Reference groups and the texture, which were only in classified format, were estimated by maximum likelihood classification using the 22 layers combined SPOT and Landsat image, with a degraded resolution of 120 m. The spatial distribution patterns of both variables were clearly visible on the RGB composite images, thus good performance was expected. Regular accuracy measures, like RMS, standardized RMS and average standard error were calculated and error vs. measured plot was created to visualize the error trends. For the maximum likelihood classification the overall class performance (the correctly classified training pixels / the total number of training pixel), the Kappa statistics and the confusion matrix (user's and producer's accuracies) were calculated to characterize the accuracy (Congalton, 1991).

In the last section, the WRB diagnostic properties and horizons were estimated using the four intermediate data layers and pedotransfer functions. Pedotransfer functions are simple or more complex rules/relationships to estimate missing properties based on existing, correlated, and easy to collect/measure properties (McBratney et al., 2002). Table 25.2 summarizes the pedotransfer functions used for estimating the WRB qualifiers/diagnostics for the study area.

25.3 Results and Discussion

25.3.1 Results from Model

25.3.1.1 The WRB Reference Soil Groups

The WRB reference groups were estimated with maximum likelihood classification of the combined SPOT/Landsat images. Eight soil types appeared on the classified image with a very pronounced spatial distribution pattern (Fig. 25.3). Fluvisols

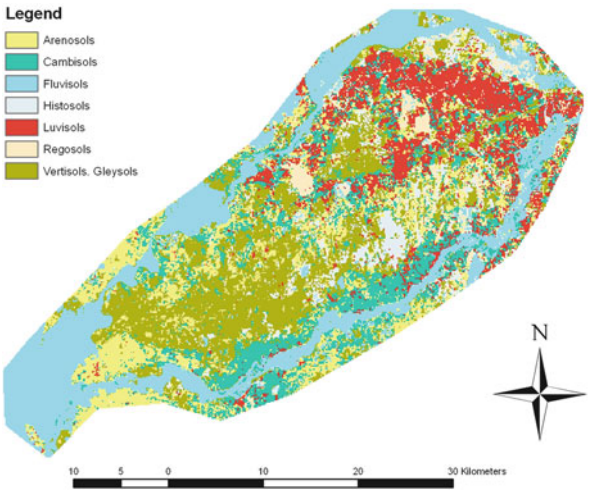


Fig. 25.3 The WRB Reference soil groupings map

occur along the major rivers on the annually flooded areas. The backwater area behind the sandy levees is covered by heavy textured Vertisols and Gleysols, with tiny islands of the remaining Histosols. The Northern part of the area is dominated by Luvisols, having a well developed B- horizon with strong clay skins. The small sand dunes have Arenosols and Cambisols on their lower sections. Histosols and Regosols occur as very small islands, representing small drained depressions and loamy plateaus.

This soil distribution pattern was evident from the satellite images. RGB composites of the images showed the extent of the major soil types for the experienced eyes. The visual interpretation of the classified image showed a very good match as well with our local knowledge and mental model. Quantitative tests are given below. However, the risk of having too strong "landuse pattern"-dominated classified soil image was a real possibility. This strong pattern was "softened" by resampling the image to 120 m resolution and using PCA transformation. The first component of the PCA transformed image always emphasizes the landuse/landcover pattern, while the 2nd, 3rd, and 4th components are more related to secondary variability within-the-1st -component, within the land cover pattern. These secondary, hidden patterns are the ones we often need and are related to the soil characteristics. Using these tools limited the occurrence of the land cover pattern.

The transition zones between the Regosols, Arenosols and Fluvisols classes were often quite difficult to handle, the separation of these taxonomically similar soil types were not always easy to make, even in the field. The subtypes of the reference groups were very similar in taxonomy to the neighboring reference soil group, often representing the transitional types between the reference groups – like Fluvisols and Fluvic Cambisols. However, as classified units they occur far from each other in the classification system due to the hierarchy. This problem had a significant impact on

the accuracy measures as well. However, this potential misclassification had more impact on the quantitative accuracy measures, than on the real usefulness of the map.

25.3.1.2 The Property Maps

The texture map shows settings similar to the WRB one. The active flood plains have loam and sandy loam texture. The inner part of the area is clay, with small islands of sand dunes occurring in the area. Organic materials and Histosols are very rare. The spatial patterns of the soil texture were easy to follow by simple visual interpretation of the composite satellite images as well. The GB image of Fig. 25.1 nicely shows the lighter colored levees of the recent and ancient rivers and the darker colored clayey (Gleysol-Vertisol) inland areas.

Similar spatial pattern can be identified in the humus content (Figs. 25.4 and 25.5). The higher humus content occurs with the clayey soils, where the clay bounds it strongly and the longer water saturation retards the organic matter decomposition. An opposite trend can be identified in the pH map (Figs. 25.6 and 25.7). Low pH is linked to the same low lying, clayey areas, where leaching was very active up to the last century. Spherical models were used to fit the curve for both cases. Strong nugget showing significant local variation has been found (Figs. 25.4 and 25.6).

A spatial trend in the E-W direction was identified for both the humus and the pH value distributions. These two trends show converse ways, the pH values are the lowest in the centre part and increases towards the ends, while the humus content changes the opposite way. These are real trends, and were modeled with a second

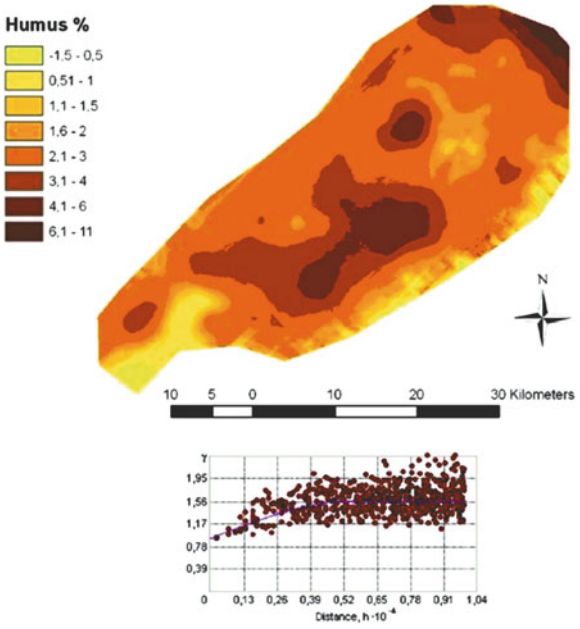


Fig. 25.4 The humus content layer and the calculated semivariogram, where $\delta(h)$ is the semivariance function of the humus content in the function of the lag distance (h)

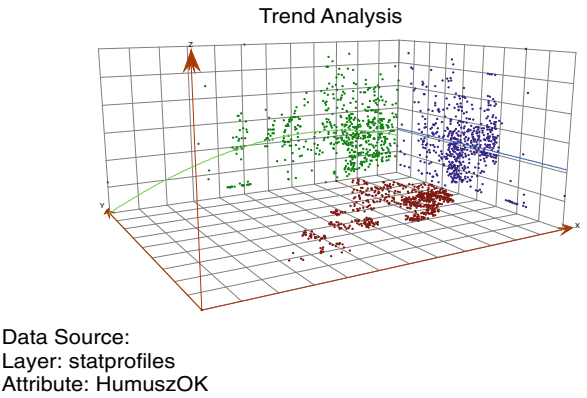


Fig. 25.5 The trend analysis diagram for the humus content. The colors indicate the different dimensions, *red* is the horizontal plain, *blue* is the North-South direction, while the *green* color is the East-West one. The *red* points show the horizontal distribution/location of the points, while the *blue* and *green* ones refer to the pH values along the NS and the EW directions respectively

order de-trend algorithm. The phenomenon is easy to explain. The centre part is the most typical backwater area, far from the major rivers and partly separated from them by the natural levees. The flood water flowing over this natural levee slows down, loses its heavy sediments and keeps only the small particles like clay. This clay is deposited in the backwater area. The trapped water cannot flow back, even after the flood is over, because the levee blocks its way back to the river. Therefore

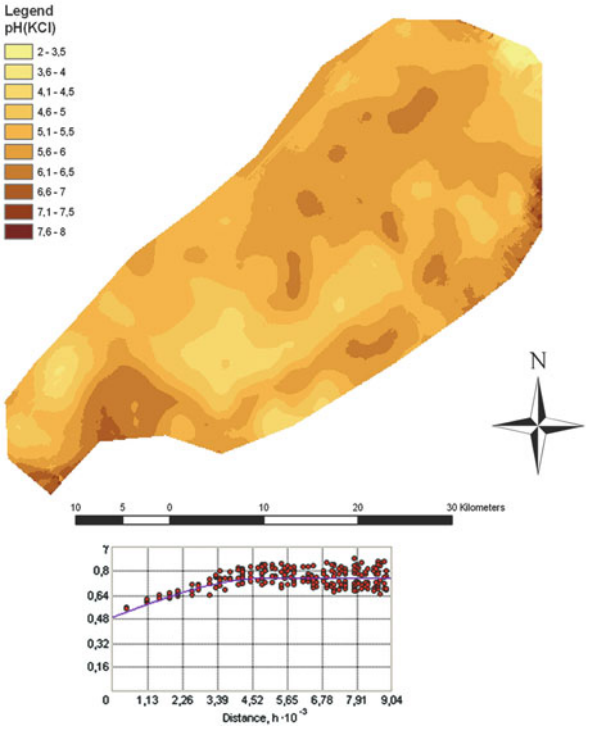


Fig. 25.6 The pH(KCl) layer and the semivariogram, where $\delta(h)$ is the semivariance function of the pH in the function of the lag distance (h)

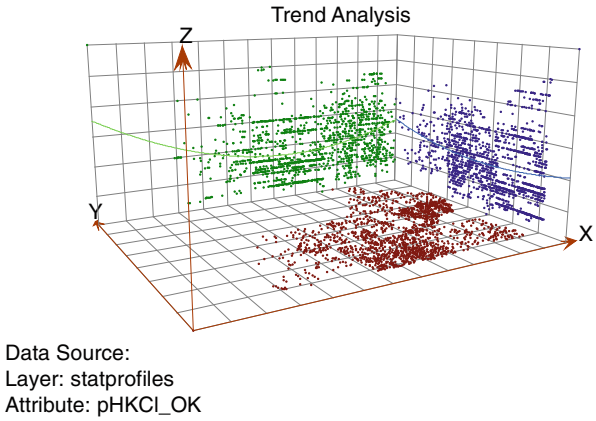


Fig. 25.7 The trend analysis diagram for the pH. The colors indicate the different dimensions, *red* is the horizontal plain, *blue* is the North-South direction, while the *green* color is the East-West one. The *red* points show the horizontal distribution/location of the points, while the *blue* and *green* ones refer to the pH values along the NS and the EW directions respectively

the water stays there longer and strongly leaches the soils lowering pH, while the high clay content and the long saturation decreases the decomposition of the organic matter and support the higher humus content.

The soils have seven major WRB diagnostic properties and horizons, which have common occurrence and strong importance in defining the soil use (Fig. 25.8). These diagnostics were created by manipulating the existing layers and combining their information according to the pedotransfer functions of Table 25.2.

The spatial patterns of the final maps do not match the Hungarian or Slovakian soil maps, which differ from each other and from the WRB classification. However, the shape and extent of the soil regions coincide well with the geomorphologic and agro-environmental patterns of area, and match our mental model well. The WRB and texture maps correspond very well to each other, because they were derived from the same integrated satellite image. However, a very good genetic coincidence appears between the WRB/texture maps and the kriging based humus and pH data, which provides a visual support to the results as well.

25.3.2 Accuracy Assessment of Model

The Landsat and SPOT image based classification resulted in an overall classification performance of 77% and a Kappa statistic of 0.7. The confusion matrix is given in Table 25.3. The User's accuracy ranged between 37 and 99% with an average of 64%, while the Producer's accuracy was between 61 and 94% with an average of 82%.

The most severe misclassification occurred in the Histosols and Regosols classes. Both classes occur as small islands, often with a smaller extent than the pixel size used for its classification, which explains their low performance.

RMS, standardized RMS and the average standard error were calculated for the kriging based extrapolations. These values for the pH(KCl) are 0.76, 0.98 and 0.77 respectively, while the values for the humus estimation were 1.13, 1.03 and 1.1. The

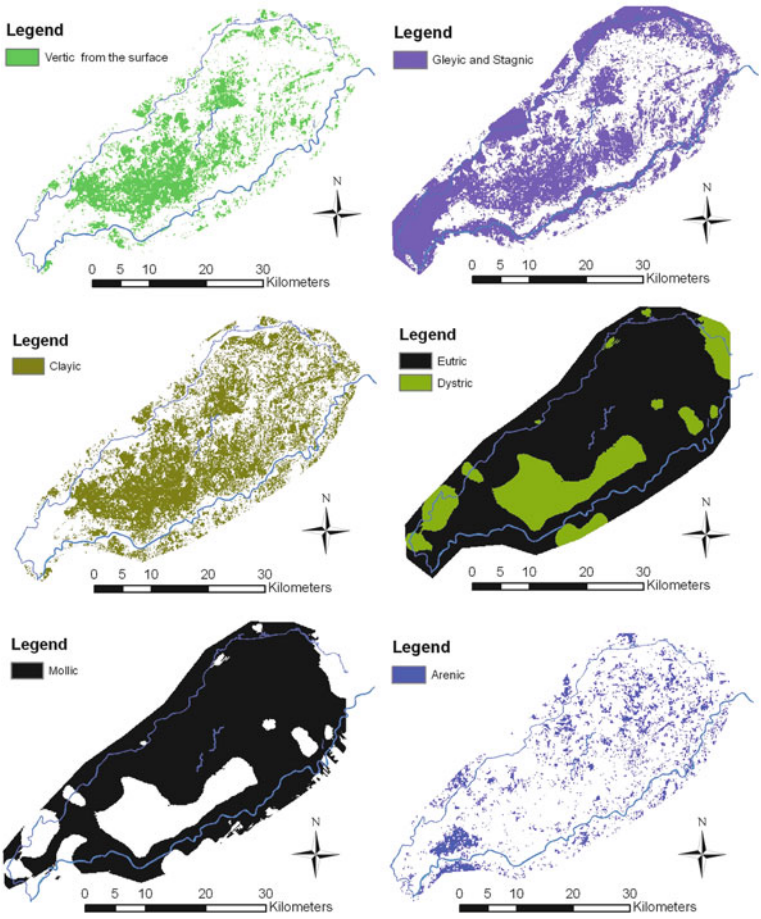


Fig. 25.8 The WRB diagnostic properties and horizons

Table 25.3 The confusion matrix of the maximum likelihood classification of the WRB reference groups

Classified	Arenosol	Fluvisol	Histosol	Regosol	Luvisol	Vertisol	Cambisol	Total
Arenosols	**60.54**	3	1.09	2.21	2.38	6.19	5.85	2, 218
Fluvisols	0.48	**76.81**	0	0	0.42	0.4	0.6	7, 095
Histosols	4.16	5.55	**93.82**	0.44	3.22	5.91	1.38	1, 491
Regosols	4.68	5.63	0	**94.03**	2.87	1.15	0.34	1, 162
Luvisols	5.04	1.62	0.91	1.77	**84.81**	1.98	5.08	1, 659
Vertisols	17.69	3.93	3.64	0.22	1.4	**81.21**	8.18	5, 033
Cambisols	7.4	3.45	0.55	1.33	4.9	3.15	**78.57**	1, 653
Total	2499	9178	550	452	1429	5041	1162	20, 311

The values in the matrix are percentages of training pixels from a given class classified into the resulting classes. The values of the "Total" line and column represents training pixel numbers

pH values range between 3.5 and 8. The humus values are between 0 and 8, but can go further up for extreme hydromorphic soils. Both estimations are smoothing the data, the estimation error increases towards the minimum and maximum values, while decreases to 0 around the average.

25.4 Conclusions

Archived legacy data have great value for database development. Huge amounts of data have been collected and recorded in many previous mapping and survey campaigns. These data are often interpreted into thematic polygon maps, and used for many applications. The integration of these types of data sources can improve the reliability and accuracy of our soil databases, and creates new generation data with added value. The best way to do so is to use the "raw" field survey observations as a profile database, or derive representative point data from averaged polygonal information. The integration and harmonization of these profile databases is the best and most consistent way of combining and interpolating information of different origin using digital soil mapping tools. It was also concluded that the diagnostic features, materials and horizons of WRB can be estimated from harmonized, variable origin, integrated data sources.

References

Baxter, S.J., and Crawford, D.M., 2008. Incorporating legacy soil pH databases into digital soil maps. In: Hartemink, A.E., McBratney, A., and Mendonca-Santos, M.L. (eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Bernoux, M., Arrouays, D., Cerri, C.E.P., and Cerri, C.C., 2007. Regional organic carbon storage maps of the Western Brazilian Amazon based on prior soil maps and geostatistical interpolation. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), Digital Soil Mapping. An Introductory Perspective. Developments in Soil Science, vol. 31. Elsevier, Amsterdam.

Buzás, I. (ed.), 1993. Handbook of Soil and Agrochemical Methods, 1. Physical, Water Management and Mineralogical Analysis of Soils. INDA 4231 Kiadó, Budapest, Hungary, 155 pp. (in Hungarian).

Congalton, R., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37:35–46.

Dobos, E., and Daroussin, J., 2007. Calculation of potential drainage density index (PDD) , pp. 283–295. In: Peckham, R., and Jordán, Gy. (eds.), Digital Terrain Modeling. Development and Applications in a Policy Support Environment. European Commission, ISBN 978-3-540-36730-7. Springer, Dordrecht.

Dobos, E., Micheli, E., and Montanarella, L., 2007. The population of a 500 meter resolution soil organic matter spatial information system for Hungary. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), Digital Soil Mapping. An Introductory Perspective. Developments in Soil Science, vol. 31. Elsevier, Amsterdam.

Dobos, E., and Kobza, J., 2008. A Bodrogköz talajai (The soils of the Bodrogköz). In: Dobos, E., and Terek, J., (eds.), Élet a folyók között. A Bodrogköz tájhasználati monográfiája. Miskolci Egyetem, ISBN:9789630642644.

ERDAS Inc., 1999. ERDAS Field Guide. ERDAS Inc, Atlanta, GA, USA.

IUSS Working Group WRB, 2006. World reference base for soil resources 2006. World Soil Resources Reports No. 103. FAO. Rome.

Mayr T., and Palmer, B., 2007. Digital soil mapping: an England and Wales perspective. In: Lagacherie, P., McBratney, A.B., and Voltz M. (eds.), Digital Soil Mapping. An Introductory Perspective. Developments in Soil Science, vol. 31. Elsevier, Amsterdam.

Mayr, T., Palmer, R.C., and Cooke, H.J., 2008. Digital soil mapping using legacy data in the Eden valley, UK. In: Hartemink, A.E., McBratney, A., and Mendonca-Santos, M.L. (eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht.

McBratney, A.B., Minasny, B., Cattle, S.R., and Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109:41–73.

Rossiter. D.G., 2008. Digital soil mapping as a component of data renewal for areas with sparse soil data infrastructure. In: Hartemink, A.E., McBratney, A., and Mendonca-Santos, M.L. (eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Szabó, J., Pásztor, L., and Bakacsi, Zs., 2005. Egy országos átnézetes talajinformációs rendszer kiépítésének igénye, lehetõsége és lépései. Agrokémia és Talajtan 54:1–2.

Várallyay, Gy., Hartyáni, M., Marth, P., Molnár, E., Podmaniczky, G., Szabados, I., and Kele, G., 1995. Talajvédelmi Információs és Monitoring Rendszer. 1 kötet. Módszertan. (Soil Monitoring System for Soil Protection. The Manual of Procedures) Földmûvelésügyi Minisztérium, Budapest.

Weiss, A., 2001. Topographic Position and Landforms Analysis. Poster presentation, ESRI User Conference, San Diego, CA.

Chapter 26
Toward Digital Soil Mapping in Canada: Existing Soil Survey Data and Related Expert Knowledge

Xiaoyuan Geng, Walter Fraser, Bert VandenBygaart, Scott Smith, Arnie Waddell, You Jiao, and Gary Patterson

Abstract Digital soil mapping involves the creation of new raster-based soil attribute datasets from existing soil and environmental data, coupled with other spatial knowledge of soil distribution. The *GlobalSoilMap.net* project is intended to provide a digital soil map of the world on a 90 m raster base derived from existing soil data sources in each country. This paper provides an overview of the existing soil information holdings in the Canadian Soil Information System (CanSIS) in terms of their scale, coverage, and potential suitability for digital soil mapping applications. A description of a possible approach to the capture and transformation of legacy soil survey knowledge for digital soil mapping purposes is also provided. Most historical soil inventory maps and reports in Canada have been produced for the southernmost 20% of Canada's land area by a variety of federal and provincial agencies at scales ranging from 1:20,000 to 1:250,000. Many of these datasets are available in digital format as part of the National Soil Data Base (NSDB) within CanSIS. The NSDB detailed soil map coverages provide the most precise spatial data source for building raster based digital soil mapping products, but coverage is incomplete. The Soil Landscapes of Canada (SLC) map series provides complete coverage for all of Canada, at a scale of 1:1 million. SLC maps are less spatially precise, but provide the national coverage needed for applications like the *GlobalSoilMap.net* project. Pedon datasets provide spatial information at specific points, but the sampling density is very low, and not well spatially distributed. Understanding the status and relevance of the NSDB data holdings, and how they can be effectively combined with expert knowledge, digital terrain models, and other data sources are important for organizing our approach to future digital soil mapping work in Canada.

Keywords Canadian soil · Legacy soil data · Digital soil mapping · Soil survey · Expert system

X. Geng (✉)
CanSIS, AESB, Agriculture and Agri-Food Canada, Neatby Building, 960 Carling Av., Ottawa, ON, Canada K1A 0C6
e-mail: xiaoyuan.geng@agr.gc.ca

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_26,

26.1 Introduction

Soil forms on a continuum of topography and surficial material and as such, soil spatial variability should be inherent in models describing soil genesis and development. In practice, soil survey is often characterized by treating soils as geographic bodies. In the last 80 years, soil survey in Canada has focused on the identification of bodies of related soils that can be recognized as natural units and on their prediction and delineation on maps (Coen, 1987). Although the resulting maps and reports are valuable information sources for users such as policy makers, land managers or farmers, the true nature of soil spatial distribution is not well represented. This lack of spatial specificity has meant that these map polygons are often not suitable for system modeling or other needs for continuum soil property data (Behrens and Scholten, 2006; Carre et al., 2007).

Currently there exists only limited effort in field survey in Canada for the purposes of soil mapping, even though it is greatly acknowledged that such data are necessary and useful for organizations and individuals working at local through national scales. Recently there have been advances in techniques for the extrapolation of existing soil map information along with the development of inference models that can predict the spatial distribution of soil properties and/or classes at varying scales that may not require expensive and timely field soil survey (Grinand et al., 2008; Henderson et al., 2005). This field of work is broadly being termed digital soil mapping.

In this paper we aim to review the data holding of the Canadian Soil Information System (CanSIS), to briefly assess the adequacy and usability of the legacy soil survey data for *GlobalSoilMap.net* (see also Chapter 33) and other digital soil mapping applications and to explore the methods to extract accumulated expert knowledge that is embedded within the existing soil survey data. We expect that this work will provide a much-needed start to national scale research and development on digital soil mapping and will build on and apply the earlier foundational work on digital soil mapping in Canada (MacMillan et al., 2004).

26.2 Materials and Methods

26.2.1 Legacy Soil Survey and Expert Knowledge

An inventory of current holdings of the Canadian Soil Information System was conducted. The extent, scale and data structure of existing soil maps were categorized and summarized into a set of map figures (Figs. 26.1 and 26.2) as part of the initial overall assessment of suitability of soil survey maps and pedon data for digital soil mapping.

Walter et al. (2007) identified five broad domains of expert knowledge that are possibly useful for digital soil mapping: (1) relative distribution of soil entities within the landscape; (2) identification of soil development factors such as

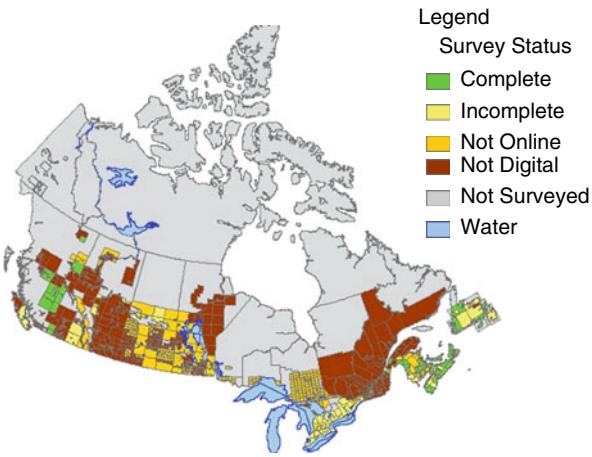


Fig. 26.1 Coverage of detailed/semi-detailed soil surveys in Canada. The Canadian Soil Information System (CanSIS) holds over 3,000 individual soil maps at scales ranging from 1:20,000 to 1:250,000

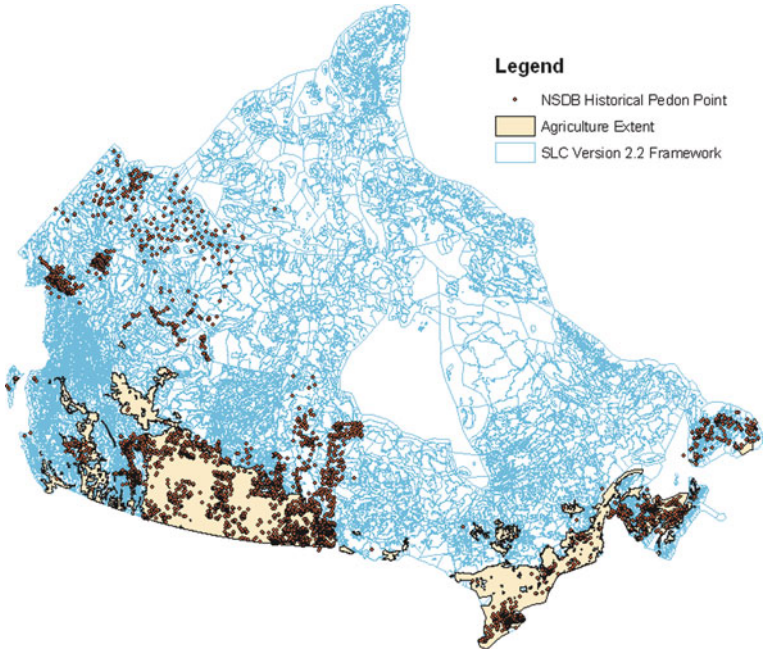


Fig. 26.2 Coverage of the national-scale Soil Landscapes of Canada (SLC) mapping. SLC version 2 covers all of Canada as shown by the polygon outlines. Within the agricultural extent of the country mapping has been upgraded and published as SLC version 3. Locations of pedon samples stored within CanSIS are heavily biased toward agricultural regions. SLC version 1, not shown, was a non-digital hard copy map series with national coverage

topography, parent material etc.; (3) correlation between soil properties; (4) spatial structures of soil properties; and (5) temporal dynamics of soil processes. The Canadian System of Soil Classification incorporates soil bio-geographic genesis and physiographic concepts. In the National Soil Database (NSDB) within CanSIS, some of the domain knowledge has been retained and embedded across the NSDB relational entities. Typically within a given soil polygon, major landscape components are identified and linked to a soil name and representative pedon/layer tables. However, currently there is no explicit spatial reference for soil polygon components within the NSDB data structure. Nevertheless, it may be a feasible and cost-effective measure to recover the lost geospatial linkage of soil classes and properties from the legacy soil survey for digital soil mapping (see Section 25.1). We tested state-of-the-art GIS capabilities to organize, visualize, assess and extract soil-landscape relations. The extracted soil knowledge can be stored within the well-structured Extensible Markup Language (XML) format to be enriched and validated with additional related environmental covariates. The stored structured knowledge can be further organized into semantic form as interoperable inputs for various digital soil mapping methods as reviewed by McBratney et al. (2003).

26.2.2 Evaluation of Effective Scale

Canadian soil resource data has been produced at various scales from many vintages for different purposes. According to Forbes (1982), the various presentation scales of resource data can be partially evaluated using the concept of "effective map scale". Effective map scale is defined as a precision measure that indicates if the intricacy of the polygons depicted on the soil map corresponds to the claimed presentation scale (Hengl and Husnjak, 2006). Effective map scale number (**ESN**) can be calculated as:

$$\mathbf{ESN} = \mathbf{NSN}^{*}(\mathbf{IMR}/\mathbf{2}) \tag{26.1}$$

Where

NSN is the nominal scale number of the presentation scale
IMR is the index of maximum reduction, that is, the factor by which the scale of the map could be reduced before the average size delineation (**ASD**) is equal to the minimum legible delineation (**MLD**) (Hengl and Husnjak, 2006) . In Canada, **MLD** was defined as 0.5 cm^2.

The ASD (converted to cm^2 at a given presentation scale) and IMR are derived as (Hengl and Husnjak, 2006):

$$\mathbf{ASD} = \sum_{j=1}^{m} Aj/m \tag{26.2}$$

Where

Aj is the area of the jth polygon and m is the total number of polygons

$$\mathbf{IMR} = \sqrt{\mathrm{ASD/MLD}} \tag{26.3}$$

We selected a number of maps at detailed and semi-detailed scale and the national Soil Landscapes of Canada scale to derive the ESN in order to assess the effective scale of these key Canadian soil mapping products.

26.3 Results and Discussion

26.3.1 Mapping Procedures

Soil survey in Canada has followed methods not unlike those in other countries. Major soil types have traditionally been identified, described and delineated on a map based on direct field observations, supplemented by indirect inferences based on aerial photo interpretation. The steps usually followed by a soil survey can be summarized into 4 main stages:

(1) The soils observed within a survey area are arranged into a limited number of soil names (series or associations) on the basis of properties that are relevant to the survey objectives;
(2) Each major named soil is described and sampled as part of the field pedon investigation and submitted for laboratory analyses, the results of which are used to populate soil attribute tables;
(3) Soil map polygons are delineated and labeled to describe the portions of the landscape that are associated with each soil name;
(4) Named soils are identified with a unique code (e.g. Province Code + Soil_Code + Modifier + Land_Use) which is used to reference relational soil information within the NSDB.

26.3.2 Status of Canadian Soil Survey and the National Soil Database Holding

The Canadian Soil Information System (CanSIS) publishes various scales of soil and soil derived (interpretive) maps via the CanSIS website (http://sis.agr.gc.ca/cansis). Detailed soil maps exist for approximately 20% of Canada's total land area, consisting mainly of the agricultural areas in each Canadian province (Fig. 26.1).

Over the last decade, Agriculture and Agri-Food Canada, together with provincial partners, have attempted to create provincial seamless digital maps at standard scales to replace the many individual detailed map sheets produced at a range of scales and covering local, county or rural municipal jurisdictions. For these new seamless coverages, not only are map polygons correlated and standardized but also

are the soil naming conventions and the attribute data structures. These maps are some of our most highly suited for digital soil mapping applications.

Soil Landscapes of Canada (SLC) maps are generalized maps published at 1:1,000,000 scale. Several versions of the SLC maps have been published over the last 25 years reflecting changes in data availability, management and technology (Fig. 26.2). These versions as well as various detailed map products, reports and technical manuals are distributed freely via the CanSIS website.

CanSIS also holds about 6,500 pedons records from across Canada (average of 1 pedon/1,400 km^2). However, most of the sampled pedons are located in the southern regions of Canada where agriculture land predominates (Fig. 26.2). Many additional pedon records are held outside of CanSIS by other agencies and organizations but are not readily accessible. The overall distribution of pedon sample locations is inconsistent and the usability of the pedon records has yet to be fully assessed in terms of location accuracy and attribute completeness.

26.3.3 Canadian Soil Resource Data Structure and Linkages

The data model used to manage Canadian soil data has evolved to meet soil resource data business needs and changing technology capacity. The current data model of the NSDB within the Canadian Soil Information System (CanSIS) is illustrated in Fig. 26.3. Pedon data including both point feature class and tabular

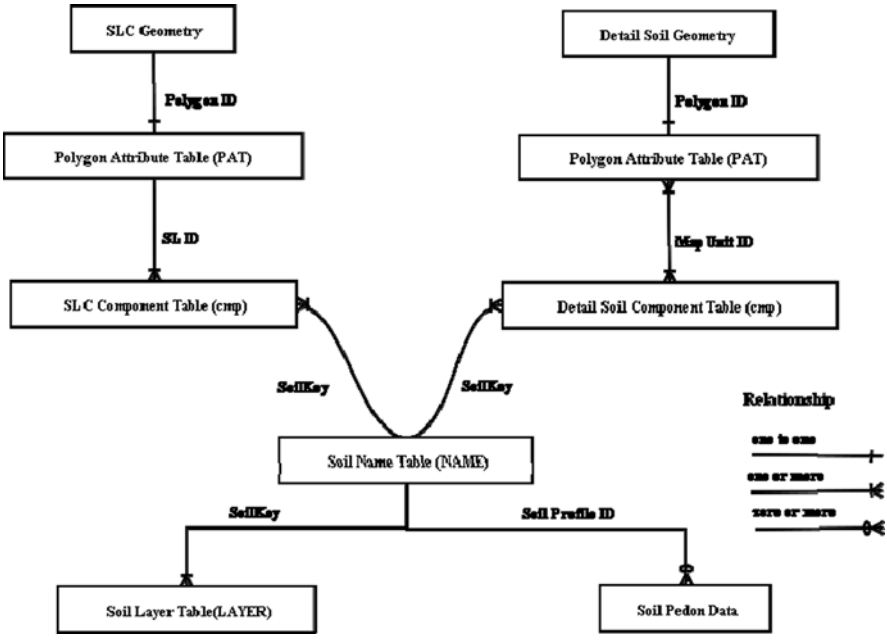


Fig. 26.3 Data model for the holdings of the National Soil Database contained within CanSIS

field and laboratory information are the foundation of the soil layer table and higher level data derivations. The layer table is constructed for each soil name with attributes calculated by averaging values contained in all pedons representing that name. In cases where no analytical pedon data exist for a soil name, layer table attributes are estimated or derived by pedo-transfer functions. The layer tables are used to support both detailed and SLC mapping and accessed by any algorithm or scientific model run against the NSDB. Further details of the Canadian soil resource data structure can be found through the CanSIS web site (http://sis.agr.gc.ca/cansis/nsdb/detailed/data_model.html).

26.3.4 Canadian Soil Survey and Effective Scales

Canada is a very large country with a relatively small population. As a result of the dominant societal needs in Canada, 1:20,000 detailed and 1:50,000 semi-detailed surveys were most often conducted in those regions of the country with agricultural or in some provinces, forestry capability. As previously described this scale of mapping covers less than a quarter of the country. Although so-called exploratory or reconnaissance mapping was also conducted at scales of between 1:125,000 and 1:250,000, the second major mapping scale is the 1:1 million scale Soil Landscapes of Canada series. These SLC maps are compiled by province and then merged to form national coverages. We examined these two dominant map types for their effective scale of spatial presentation (Table 26.1).

Many years ago pedologists defined a minimum delineation area or minimum legible delineation (MLD) for CanSIS maps as $0.5\,cm^2$ on the published map scale. We used this standard in the calculation of the effective scale number (ESN). With one exception, detailed and semi-detailed map polygons have average size delineations (ASD) between 2 and $8\,cm^2$ which translates into ESN values slightly smaller than published presentation scales. One of our selected maps, published at a 1:50,000 scale, had a very large average polygon size and an ESN of 1:168,800. With respect to the SLC mapping, significant variation occurs with respect to the density of information. Although the SLC maps are published at a consistent scale for all regions of Canada, the effective map scale based on polygon size for most provinces is much smaller than the published presentation scale. The only exception is for the province of New Brunswick where the published scale and effective scale are nearly identical. In provinces with northern regions devoid of extensive resource development and hence detailed soil information (i.e., Manitoba, Ontario, Quebec and Newfoundland and Labrador), the effective scale number is smaller than 1:2M. For the northern territories (Northwest Territories, Yukon and Nunavut) the effective SLC mapping scale is closer to 1:3M. We have not fully conducted position accuracy assessment on the legacy soil survey datasets. However, it is clear that we will have to treat each dataset individually given the variation in map data intensity observed at all scales of mapping.

Table 26.1 Effective scale number (ESN) of selected Canadian soil maps published at scales from 1:20,000 to 1:1,000,000

		MLD (0.5 cm^2)	Mapped area		ASD			
Soil map_ID	Presentation scale	ha	ha	# of polygons	ha	cm^2	IMR	ESN
Detailed and semi-detailed maps								
BCD070	1:20K	1	18,580	1,288	14	4	3	1:26,800
MBD065	1:20K	1	5,028	167	30	8	4	1:38,800
NBD066	1:20K	1	8,593	595	14	4	3	1:26,800
NSD008	1:50K	1	217,525	6,998	31	8	4	1;39,400
NSD005	1:50K	6	286,396	1,565	183	7	4	1:95,700
NFD008	1:50K	6	64,411	113	570	23	7	1:168,800
NBD005	1:50K	6	347,587	5,768	60	2	2	1:54,800
NFD002	1:50K	6	70,411	368	191	8	4	1:97,700
Soil landscapes of Canada maps								
Alberta	1:1M	2,500	66,362,706	1,044	63,565	6	4	1:1,782,800
British Columbia	1:1M	2,500	101,000,000	2,651	38,182	4	3	1:1,381,700
Saskatchewan	1:1M	2,500	70,326,956	1,575	44,652	4	3	1:1,494,200
Manitoba	1:1M	2,500	70,505,963	454	155,299	16	6	1:2,786,600
Ontario	1:1M	2,500	120,000,000	866	138,369	14	5	1:2,630,300
Quebec	1:1M	2,500	172,000,000	1,389	123,474	12	5	1:2,484,700
New Brunswick	1:1M	2,500	9,271,646	471	19,685	2	2	1:992,100
Nova Scotia	1:1M	2,500	10,016,471	419	23,905	2	2	1:1,093,300
Prince Edward Island	1:1M	2,500	1,842,357	33	55,829	6	3	1:1,670,800
Newfoundland	1:1M	2,500	61,634,667	677	91,040	9	4	1:2,133,500
Northwest Territories	1:1M	2,500	192,000,000	999	191,845	19	6	1:3,097,100
Nunavut	1:1M	2,500	408,000,000	1,804	226,111	23	7	1:3,362,400
Yukon	1:1M	2,500	51,846,687	346	149,845	15	5	1:2,737,200

26.3.5 Data Mining of Legacy Soil Survey

Soil survey information contains various levels and aspects of knowledge about soil distribution, environmental co-variables and soil properties. However the original geographic location of identified soil types or classes within a map polygon is not known. One only knows that there are one or more components or major soil types in a soil survey polygon. Extracting and representing the embedded knowledge for raster-based spatial inference or interpolation is the objective of legacy soil survey data mining.

Given the ever increasing geospatial thematic information for Canada, we are working toward the very first steps of digital soil mapping by gathering the expert knowledge embedded in the legacy soil survey using in-house tools (Fig. 26.4). By integrating soil survey data with other currently available environmental thematic data like digital elevation models and land cover, a component soil can be allocated to a specific landscape (geographic) location (MacMillan et al., 2000).

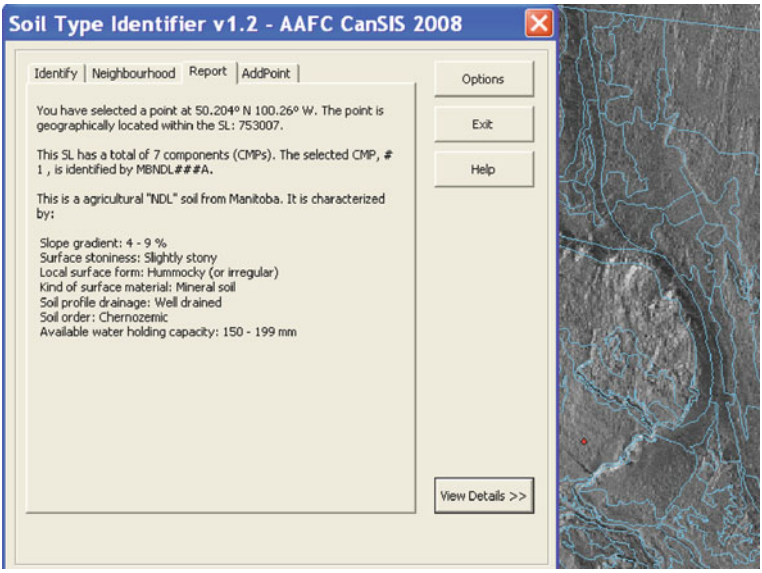


Fig. 26.4 Graphic user interface of the GIS-based tool for capturing expert knowledge about soil-environmental relationships inherent within legacy soil survey data

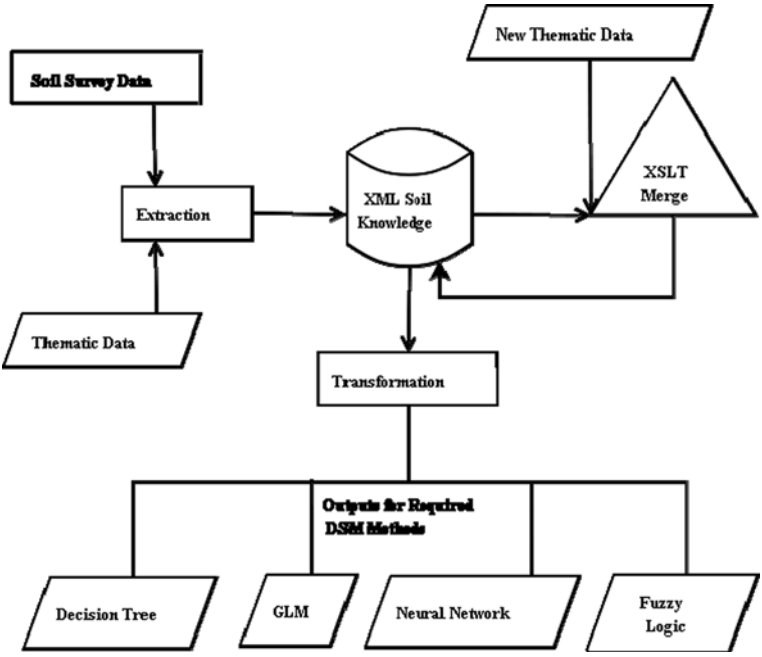


Fig. 26.5 Process flow for captured expert knowledge within legacy soil survey data. The Soil Type Identifier tool in Fig. 26.4 is shown as the extraction process which relates soil survey data to existing thematic data and moves this into a soil knowledge database. These relations can be enhanced by merging new thematic data as they become available. The XML database allows easy transformation of data into formats suitable for use in a variety of potential statistical operations

To enable this, a data mining tool kit has been developed. The Soil Type Identifier tool compiles available spatially-specific (raster) environmental covariates underlying a given soil map polygon. An experienced pedologist then samples specific points on the polygon for which covariate data are brought forward and using expert knowledge, identifies the most likely soil component (name) that applies to that specific set of conditions. Using this knowledge extraction process, the tool builds a database of soil name – environmental covariate relationships (Fig. 26.5). The captured geographic locations of a soil are then used as a linkage for both quantitative and qualitative regression between soil classes/properties and environmental covariates. This knowledge database can be further enriched by adding additional soil pedon data, updating existing records, and merging new thematic data such as new earth observation imagery to sampled points where the covariate relationships can be broadened. This should greatly facilitate the application of various inference system approaches (Henderson et al., 2005; McBratney et al., 2003; Zhu et al., 2001; see also Chapter 34) such as decision trees or neural networks (Fig. 26.5). Testing of the tool kit and use of the captured expert knowledge stored in XML form for digital soil mapping is currently being conducted.

26.4 Conclusions

Existing Canadian soil resource information has many challenges for its incorporation into new digital soil mapping approaches. Detailed and semi-detailed soil map coverages are incomplete and are in a variety of formats and scales. The creation of seamless detailed coverages is underway but only complete for the agricultural portions of some provinces. The national Soil Landscapes of Canada map (version 2) is complete but the number of soil components and associated attribute data in each map polygon are limited making it difficult to use this product to populate gridded datasets like that proposed for *GlobalSoilMap.net*. For most northern regions of the country, the effective map scale of the SLC is much smaller than the published presentation scale further adding to the challenge. The latest version of the SLC (version 3) is more data-rich and scale effective but its coverage is limited. The soil pedon data holdings within CanSIS are of relatively low density and spatially inconsistent, again presenting challenges to using pedon data as a substantive source of information for digital soil mapping initiatives. Nonetheless, legacy soil data mining tools are being developed to systematically formulate the quantitative soil-environment relations inherent within the current CanSIS digital map files. We expect these should be useful for compilation and use in future digital soil mapping efforts in Canada.

Acknowledgments This research and development work is administrated by the Agriculture and Agri-Food Canada, and funded by the Government Related Interest Program (GRIP) of the Canadian Space Agency and the National Land and Water Information Service (NLWIS). We thank Patrick Rollin, PTPR Unlimited Inc., and David Howlett for their programming and data retrieval work.

References

Behrens, T., and Scholten, T., 2006. Digital soil mapping in Germany – a review. Journal of Plant Nutrition and Soil Science 169(3):434–443.

Carre, F., McBratney, A.B., and Minasny, B., 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma 141(1–2):1–14.

Coen, J.M. (ed.), 1987. Soil Survey Handbook. Vol. 1 Technical Bulletin 1987-9E, ISBN 0-662-15374-X, Agriculture Canada, Ottawa, ON.

Forbes, T.R., Rossiter, D., and Wambeke, A.V., 1982. Guidelines for Evaluating the Adequacy of Soil Resource Inventories. SMSS Tech. Monogr. #4. Cornell University Department of Agronomy, Ithaca, NY.

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 143:180–190.

Henderson, B.L., Bui, E.N., Moran, C.J., and Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124:383–398.

Hengl, T., and Husnjak, S., 2006. Evaluating adequacy and usability of soil maps in Croatia. Soil Science Society of America Journal 70:920–929.

MacMillan, R.A., Pettapiece, W.W., Nolan, S.C., and Goddard, T.W., 2000. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets and Systems 113(1):81–109.

MacMillan R.A., Jones R.K., and McNabb, D.H., 2004. Defining a hierarchy of spatial entities for environmental analysis and modeling using digital elevation models (DEMs). Computers, Environment and Urban Systems 28(3):175–200.

McBratney, A.B., Santos, M.L.M., and Minasny, B., 2003. On digital soil mapping. Geoderma 117(1–2):3–52.

Walter, C., Lagacherie, P., and Follain, S., 2007. Integrating pedological knowledge into digital soil mapping, pp. 281–300. In: Lagacherie P., McBratney, A.B., Voltz, M. (eds.), Digital Soil Mapping an Introductory Perspective. Elsevier, Amsterdam.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., and Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of America Journal 65(5):1463–1472.

Chapter 27
Predictive Ecosystem Mapping (PEM) for 8.2 Million ha of Forestland, British Columbia, Canada

R.A. MacMillan, D.E. Moon, R.A. Coupé, and N. Phillips

Abstract Operational predictive ecosystem mapping (PEM) at a scale of 1:20,000 is described for an area of 8.2 million ha in the former Cariboo Forest Region of British Columbia (B.C.), Canada. Mapping was conducted over 5 years by a small team consisting of a knowledge engineer, a local ecological expert, a project technical monitor, a project manager and a number of short-term contractors. The total cost for all project activities was $2.8 million Canadian dollars or 34 cents per ha. The rate of progress was 2 million ha per year for the 2 person modeling team. The predictive map was assessed for accuracy in terms of its ability to provide reliable estimates of the proportions of ecological site types within small areas. Accuracy assessments were made using 345 km of independently classified ecological observations collected along 230 randomly selected, closed linear field traverses of 1.5 km total length. The final PEM maps achieved an average accuracy of 69% across the entire map area. We summarize and generalize our experiences by recasting them in the form of ten principles that we feel are applicable to all efforts to make predictive mapping operational. We hope that these principles will stimulate discussion among practitioners of digital soil mapping and may help others to consider how best to achieve their own success in operational digital soil mapping.

Keywords Operational predictive mapping · Basic principles · Expert knowledge · Area-class maps · Accuracy assessment

27.1 Introduction

This paper presents a post-completion description and assessment of operational predictive ecosystem mapping (PEM) at a scale of 1:20,000 for an area of 8.2 million ha in the former Cariboo Forest Region of British Columbia (B.C.), Canada.

R.A. MacMillan (✉)
LandMapper Environmental Solutions Inc., Edmonton, AB, Canada;
ISRIC – World Soil Information, Wageningen, The Netherlands
e-mail: bobmacm@telusplanet.net; bob.macmillan@wur.nl

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_27,

As with most similar exercises reported in the literature (see Zhu et al., 2001), the rationale for adopting predictive methods for the present project was to try to lower costs and speed up rates of production while maintaining or exceeding currently achievable levels of map accuracy.

Under licenses granted by the provincial government, forest industry companies manage the forested land base within three timber supply areas (TSAs) that cover the entire extent of the former Cariboo Forest Region. Part of their management responsibility involves collecting and maintaining information about the forest and land resources of the areas under their management. Several legislated and mandated procedures and decisions require the use of 1:20,000 scale ecosystem maps at the level of Site Series (Pojar et al., 1987) that have been demonstrated to have obtained a minimum level of classification accuracy of 65% according to a provincially approved accuracy assessment protocol (Meidinger, 2003 or Moon et al., 2005). The forest industry companies that initiated and financed this project required a methodology for operational production of predictive ecosystem maps that could achieve or exceed the minimum required level of accuracy and that would be as efficient and cost effective as possible.

As implemented for this project, PEM represents an exercise in predicting the spatial distribution of discrete ecological classes by applying knowledge-based fuzzy classification rules to selected predictor input layers. The intent is to produce area-class output maps that capture the spatial distribution of previously defined and described ecological-landform class spatial entities. These ecological-landform class entities are direct equivalents to conceptual soil-landform models as described by Hudson (1991) and Bockheim et al. (2005).

27.2 Materials and Methods

27.2.1 Study Area

The study area consists of the entire extent (8.2 million ha) of the former Cariboo Forest Region in south central BC, Canada (Fig. 27.1). This area is approximately 416 km east-west by 304 km north-south.

The physiography of the area (Fig. 27.1) ranges from spectacular glacier-covered alpine and sub-alpine mountain ranges with annual precipitation in excess of 2,000 mm to dry grassland valleys with annual precipitation of less than 400 mm. The majority of the area consists of the broad, level to gently rolling, Fraser Plateau with elevations of 900–1,500 m situated between mountain ranges that rise to elevations of over 3,500 m in the Coast Mountains on the west and to 2.500–3,000 m in the Columbia Mountains on the east.

27.2.2 Digital Input Data for Supporting Predictive Mapping

The operational PEM mainly used a 25 m DEM and a limited number of terrain derivatives computed from the DEM in the prediction process (Table 27.1). Other

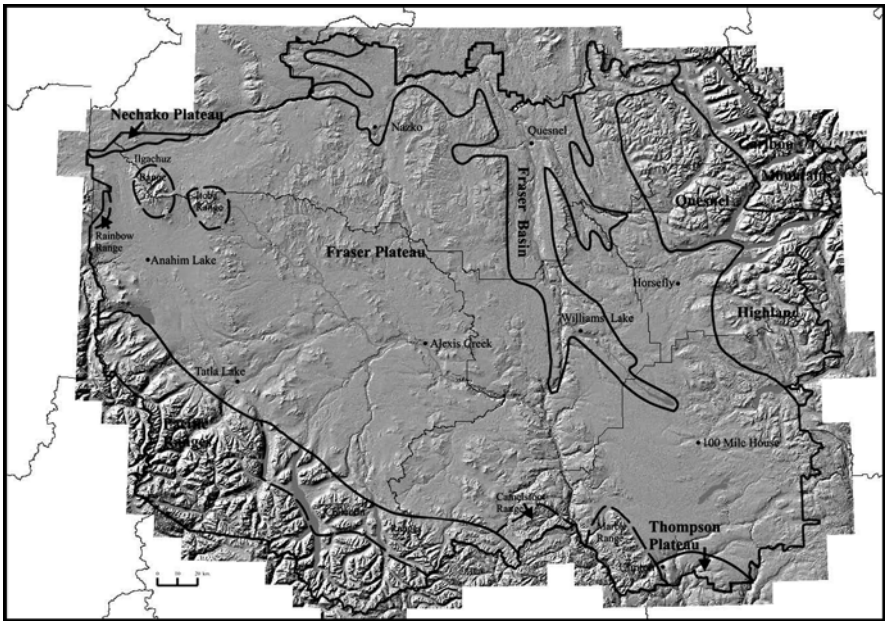


Fig. 27.1 Location, extent and physiographic subdivisions of the Cariboo PEM project area

layers of input data used in the PEM included a LandSat 7 ETM, false color composite image, a manually prepared map of Biogeoclimatic Ecosystem Classification (BGC) unit boundaries at the level of Subzones and Variants and a manually interpreted map of parent material depth, texture and exceptions classes.

27.2.3 Field Data for Supporting Assessments of Map Accuracy

No field data were collected or analyzed to assist in the preparation or finalization of rules used to classify ecological entities. By design, all predictive rules were based on trying to capture and utilize existing expert knowledge as presented in published field guides, ecological keys, landscape profile diagrams and textual descriptions of previously defined ecological entities.

A great deal of field data was collected to support a post-mapping assessment of the relative accuracy and reliability of the completed PEM maps. These field data were collected along closed triangular linear transects (Fig. 27.2) using a stratified random design selected to ensure a sufficient number of observations were obtained for each project area. Accuracy assessment procedures followed a double-blind protocol in which the ecological experts collecting the field observations had no knowledge of what had been mapped for any given area and the PEM modeler had no prior knowledge of what locations were to be sampled for assessing accuracy.

The field sampling entity was a closed, three-sided, line-intercept traverse of 500 m on each side (see Fig. 27.2). A random number between 0 and 360°

Table 27.1 List of digital input layers used in the prediction of ecological class entities

Input No.	Source DBF File	Input Name	Description of Digital Input Layer	Concept(s) that the Input Layer was used to Represent	Reference for Method of Calculation	Example Attribute
1	formfile	LNQAREA	Log of Upslope Area (m2)	Used to estimate relative landform position	Quinn et al., 1991	Crest
2	formfile	QWETI	Quinn Wetness Index (dimensionless)	Used to estimate relative moisture regime	Quinn et al., 1991	Dry_WI
3	formfile	SLOPE	Slope gradient in percent (%)	Used as a direct measure of slope steepness	Eyton, 1991	Steep
4	formfile	NEW_ASP	Values for aspect rotated 45° counter clockwise to put the 0° aspect value at 325°	Rotated aspect was used to infer degree of exposure as for warm (SW) or cool (NE) aspects		NE_Aspect
5	relzfile	PCTZ2ST	Percent change in elevation (Z) of a cell relative to the closest stream and ridge cells to which it is connected by paths of simulated surface flow.	This variable was used as the main measure of local relative landform position or relative relief. Think of it as expressing relative distance upslope in percent.	MacMillan et al., 2007	Crest2Mid
6	relzfile	PCTZ2PIT	Percent change in elevation (Z) of a cell relative to the closest pit and peak cells to which it is connected by paths of simulated surface flow.	This variable was used as a secondary measure of a more general or regional relative landform position.	MacMillan et al., 2007	Near_Base
7	geofile	DEPTH	Code for depth to bedrock estimated by contract air photo interpreters	This input variable was used to identify areas that were shallow to bedrock (< 50 cm).	TFIC, 2007	Deep
8	geofile	TEXTURE	Code for texture class as estimated by contract air photo interpreters	This input variable was used to identify areas of coarse, fine or organic textures	TFIC, 2007	Organic
9	geofile	SEEPAGE	Manualy interpreted areas of elevated likelihood of seepage.	Seepage areas were used as a flag to alert KB rules to treat cells as wetter than normal.	TFIC, 2007	Hi_Seep
10	relzfile	Z2St	Absolute vertical change in elevation (Z) from a cell to the nearest stream channel cell to which it is connected by a path of simulated surface flow.	Used as a measure of absoute vertical change in elevation relative to a local base level at the nearest cell recognized as a channel cell.	MacMillan et al., 2007	Hi_Ridge

Table 27.1 (continued)

Input No.	Source DBF File	Input Name	Description of Digital Input Layer	Concept(s) that the Input Layer was used to Represent	Reference for Method of Calculation	Example Attribute
11	relzfile	Z2Pit	Absolute vertical change in elevation (Z) from a cell to the nearest stream pit cell to which it is connected by a path of simulated surface flow.	Used as a measure of absoute vertical change in elevation relative to a local base level at the nearest cell recognized as a pit cell.	MacMillan et al., 2000	Hi_Bench
12	geofile	Z2wet	Vertical distance (m) of a cell from the nearest cell classified as a wetland or water body to which it is connected by a path of simulated surface flow.	This variable was used as a measure of vertical distance above a manually mapped wetland or water body. It was used to create elevation sensitive buffers around wetlands.	MacMillan et al., 2007	WetZ_LT05
13	geofile	L2Wet	Horizontal distance (m) of a cell from the nearest cell classified as a wetland or water body to which it is connected by a path of simulated surface flow.	This variable was used as a measure of horizontal distance back from a manually mapped wetland or water body. It was used to create buffers around wetlands.	MacMillan et al., 2007	WetL_LT200
14	geofile	N2Wet	Buffered horizontal distance of a cell from the nearest cell that was manually mapped as a pasture or meadow in only 2 BEC Subzones.	This variable was used in one particular area to define narrow buffers around non-forested meadiws in frosty, coarse textured areas.	MacMillan et al., 2007	Sand_Fringe
15	formfile	PLAN	Across-slope or plan curvature (given in °/100 m)	Used infrequently as an additional indication of convergence or divergence of surface flow.	MacMillan et al., 2007	Convex_a
16	formfile	PROF	Down-slope or profile curvature (given in °/100 m)	Used infrequently as an indication of surface concavity or convexity.	MacMillan et al., 2000	Concave
17	geofile	ELEV	Elevation above sea level as reported by the 25 m GRID DEM.	Several Site Series were defined in terms of a range in elevation over which they occured.	Not Necessary to document	GT1800
18	geofile	B3	Digital number for Band 3 extracted from the false color satellite image	The Band 3 contract stretched data was level sliced to indicate relative type of land cover.	This report.	Treed

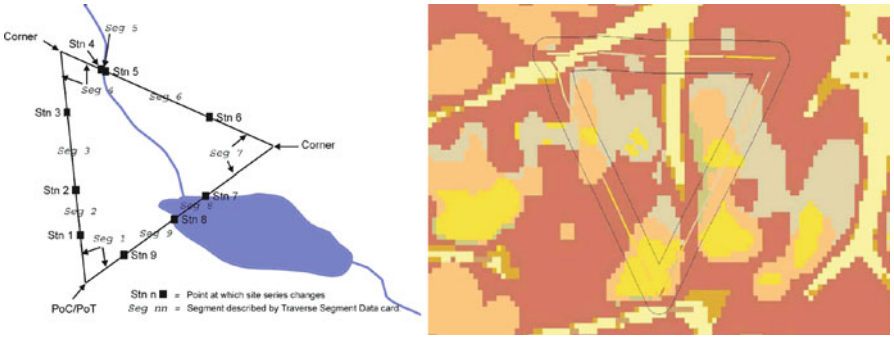


Fig. 27.2 Illustration of closed triangular traverses used to collect and analyze accuracy data

determined the bearing of the initial leg of the traverse and the traverse proceeded in a clockwise direction. Traverse stations, depicted by black squares, identified changes in ecological sites (site series). The traverse was executed clockwise from the Point of Commencement (PoC). Data recorded included descriptive site data, the primary classification, and an alternate classification for the segment. Confidence indexes were recorded for the primary classification and an alternate classification if the classification of the segment was ambiguous. All stations were located using high resolution, correctable GPS readings. All segment lengths and bearings were confirmed with compass and hip chain measurements. These closed traverses were used not for perceived statistical efficiencies but rather to minimize travel time by returning the field crew to its starting point.

A staged sampling protocol was used to ensure that a sufficiently large sample was obtained for each project area while at the same time minimizing the sampling cost necessary to achieve the required level of confidence. Sample locations (Fig. 27.3) were randomly selected points from the systematic sampling grid used by the B.C. Ministry of Forests vegetation resources inventory.

27.2.4 Inference Models

The Cariboo PEM used a knowledge-based approach to prepare and apply predictive rules. The process of creating knowledge-based, or heuristic, rule bases is relatively straightforward and has been described in detail by MacMillan et al. (2007). Each class to be predicted is defined using a fuzzy semantic import (SI) model, as proposed by Burrough (1989) and applied by MacMillan et al. (2000, 2007). Each class of ecological entity is defined as a weighted linear average of a series of defining attributes; where attribute values are computed in terms of fuzzy membership functions that relate the value of an input parameter (e.g. slope gradient) to the likelihood of that value matching the concept of the class used to define the attribute (e.g. steep slopes).

The procedures did make extensive use of existing knowledge and local ecological expertise to develop rules for classifying ecological entities. In particular,

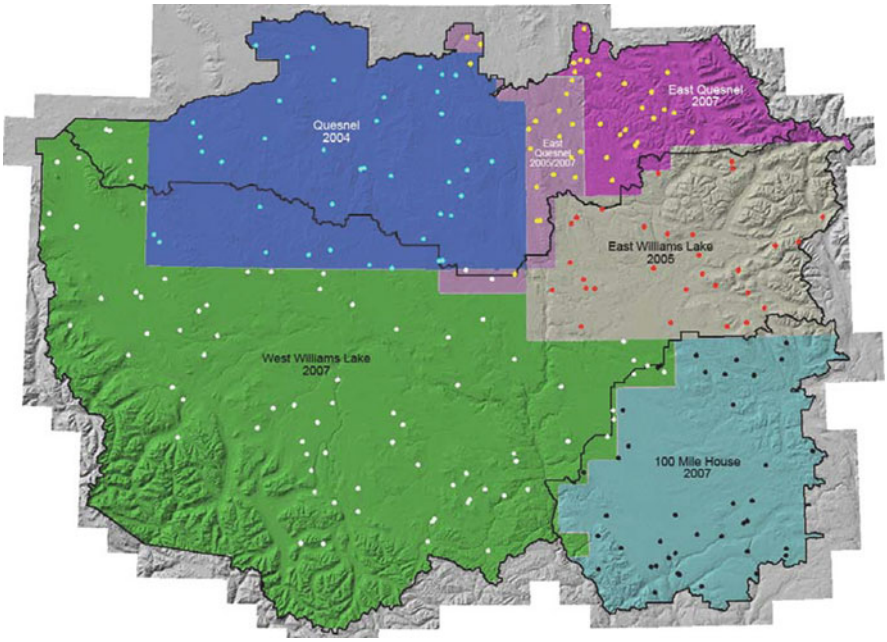


Fig. 27.3 Location of field accuracy assessment traverses relative to project and TSA boundaries

the procedures were strongly based on the ability to consult, and make use of, the extensive local experience and knowledge of the Regional Research Ecologist (see Section 14.1). In this and many other respects, the predictive procedures used here exhibit strong similarities to the expert-knowledge based approach of SoLIM (Zhu et al., 2001) and in particular to the prototype category theory implementation of SoLIM described by Qi et al., (2006) (see also Bayesian Belief Networks, Section 16.3.2).

27.2.5 Accuracy Assessment Calculations

Calculations of map accuracy adhered to a provincially published and accepted protocol (Moon et al., 2005). The intent of these procedures was not to assess the degree of exact categorical match between predicted and observed ecological classes at exact point locations. Rather, the intent was to assess how closely the proportions of ecological classes predicted to occur within a small area, equivalent in size to the area for which management decisions are typically made for forested areas, matched the proportions of those classes observed within such an area in the field (a similar intent is evident in the estimates of proportions of predicted soil classes in associations and consociations discussed in Chapter 15).

For analysis purposes, each closed triangular traverse is considered to define a corridor extending approximately 30 m on either side of the traverse. The proportion

Table 27.2 Example of calculation of overlap between predicted and observed ecological classes

			Site Series % Composition										
Trans	Classif'n	Data Source	01	02	03	04	05	06	07	08	09	99	Traverse Overlap
T21SB	Primary Call	Map	57	0	7	21	1	2	11	0	0	1	
		Traverse%	36	0	0	21	0	28	1	4	4	6	
		% Overlap	36	0	0	21	0	2	1	0	0	1	61
	Alternate Call	Map	57	0	7	21	0	2	11	0	0	1	
		Traverse%	50	0	0	21	0	14	1	3	4	6	
		% Overlap	50	0	0	21	0	2	1	0	0	1	75

of each ecological site series for each traverse was computed as the sum of intercept distances of each site series along the traverse divided by the total traverse length. The map's prediction of site series occurring along the traverse was determined by generating a 30-m buffer around the traverse (see Fig. 27.2). This created a 60-m corridor centered on the traverse. Summation of the area of each site series predicted to fall within the corridor, divided by the total area of the corridor, yielded the map prediction of ecological site series composition.

Thematic accuracy is based on a measure of classification overlap (Meidinger, 2003). Two measures of overlap were used. The first measure was primary overlap based on comparing the primary site series call to the observed site series. The second measure was alternate overlap and uses the alternate Site Series call for the observed site series rather than the primary call, if such use increases the degree of overlap. Alternate calls were only used if the confidence attached to the primary call was moderate or low.

Table 27.2 presents an example overlap table for one traverse. Primary call overlap was 61%. The map over-predicted site series 01 (57 versus 36%) and under-predicted site series 06 (2 versus 28%) relative to the traverse. Fourteen percent of the traverse classified as 06 by the primary call was considered moderate to low confidence and had an alternate site series call of 01. The traverse proportion of site series 01 was therefore increased by 14% and the traverse proportion for site series 06 was decreased by 14%. This re-allocation and recalculation increases the overlap for site series 01–50% without reducing the overlap of the 06 site series. The alternate call overlap was therefore 75%.

27.3 Results

27.3.1 Results from the PEM Modeling

PEM maps, at a scale of 1:20,000, were completed and delivered for an area of over 8.2 million ha in the former Cariboo Forest Region of B.C (Fig. 27.4). The mapping was planned and executed as a number of separately contracted projects beginning in October, 2003 and ending in April, 2008 (see Fig. 27.4 and Table 27.3).

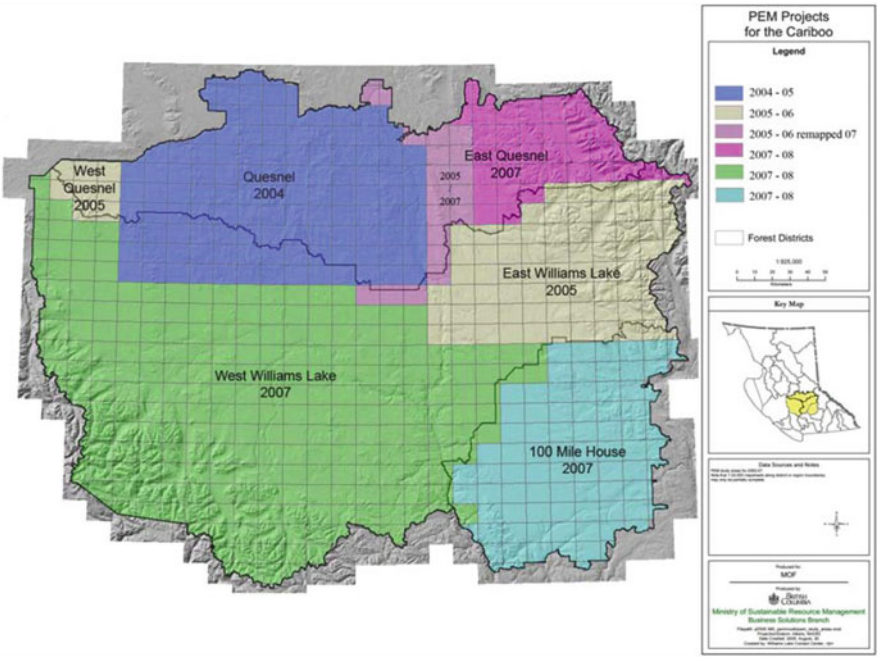


Fig. 27.4 Location of individual annually-staged project areas relative to TSA boundaries

The project was carried out by a relatively small team that consisted of a knowledge engineer and PEM modeler (the first author), a project technical monitor (the second author), a local ecological expert (the third author), and a project manager (the fourth author) (see Fig. 27.6). Specific contributions were also made by a number of contractors hired to prepare input data layers, collect field accuracy data and produce final cartographic map products.

The total cost for all project activities was Can $ 2.8 million or 34 cents per ha. The rate of progress was approximately 2 million ha per year for the two person modeling team. The final maps were prepared and delivered in both raster and labelled vector format.

27.3.2 Accuracy Assessment of the PEM Model

Initially, assessments of accuracy were computed and reported for individual project areas, at the end of each staged project, as illustrated in Fig. 27.4. Upon final completion of all six individual projects, a single, seamless PEM map was prepared for the entire extent of the former Cariboo Forest Region. The accuracy assessment data were then re-evaluated to re-compute the accuracy by Timber Supply Area (TSA) and for the region as a whole (Table 27.3). The lower accuracy computed for the full Quesnel TSA reflected a particular problem with difficulties encountered in

Table 27.3 Schedule, accuracy assessment results and number of accuracy transects by TSA

No.	Project Name	Area (Million ha)	Project Dates Start - End	Accuracy Assessment Results				
				Exact	Alternate	Acceptable	(Overlap) Dominant	No. Transects
1	Quesnel TSA	2.075	Oct 2003 - April 2006	62.5	69.4	77.7	78.1	76
2	Williams Lake TSA	4.930	Oct 2006 - Dec 2007	65.3	69.0	77.7	75.2	114
3	100 Mile House TSA	1.235	Dec 2007 - April 2008	68.8	73.7	81.7	87.5	40
	Cariboo Region Total	8.240		65.2	69.8	77.9	77.4	230

Table 27.4 Complete itemization of all costs for all components of the Cariboo PEM project

Project component	Cost ($)	Percent (%)	$/ha
Project management	$295, 000.00	10.40	$0.036
BGC localization (big BEC)	$437, 278.00	15.41	$0.053
Manual materials & exceptions mapping	$186, 326.00	6.57	$0.023
Mandated structural stage modeling	$46, 963.00	1.66	$0.006
DEM preparation	$7, 200.00	0.25	$0.001
PEM knowledge base development and modeling	$900, 000.00	31.72	$0.109
Accuracy assessment data collection and analysis	$773, 314.00	27.26	$0.094
Final submission to the spatial data warehouse	$191, 050.00	6.73	$0.023
	$2, 837, 131.00	100.00	$0.344

differentiating two very similar ecological classes in one particular BGC unit within the eastern portion of the Quesnel TSA.

27.3.3 Complete Costs for all Activities

Table 27.4 lists and itemizes all final costs for all components of the Cariboo PEM project for the entire 8.2 million ha. Costs for accuracy assessment were increased due to a fire that destroyed a full field season's worth of field observations that then had to be re-collected. Costs for final submission of the digital products to the provincial spatial data warehouse include work to generalize the raw raster data and convert it to smoother vector polygons that were large enough to receive and display textual labels. All costs are fully accounted for with no free, donated labor or in-kind contributions.

27.4 Discussion

To summarize and generalize our experiences we recast them in the form of several principles adapted from pedometrics, modeling and other literature. We offer these principles in the hope that they may prove useful to others with a need to undertake operational predictive mapping. These principles (see McBratney et al., 2002 for 1 and 2) are:

1. *Principle No. 1: Efficiency* – Do not predict something that is easier to measure or map than the predictor. Mathematically, Efficiency of Predicted/ Efficiency of Predictor > 1.
2. *Principle No 2: Uncertainty* – Do not make DSM predictions unless you can evaluate their uncertainty, and for a given problem, if a set of alternative DSM predictions is available, use the one with minimum variance.
3. *Principle No 3: Parsimony* – Use as few input variables as possible to make DSM predictions but equally, use as many as are absolutely necessary to achieve success.

4. *Principle No 4: Simplicity* – Keep it simple. Increased precision in description of predicted entities leads to decreased accuracy of the predictions. Splitting increases error.
5. *Principle No 5: Rigor* – Question everything. Test all assumptions. Don't assume that your assumptions are correct. Devise and apply rigorous, objective tests to evaluate all assumptions.
6. *Principle No 6: Continuity* – Embrace the future but value the past. Respect and use existing knowledge and data to its maximum but don't let it impede progress.
7. *Principle No 7: Ambition* – Go big or go home. The bigger the area mapped the greater the potential economies of scale. But do the work in small, incremental stages.
8. *Principle No 8: Stratification* – Divide and conquer. Turn one big, complex problem into many smaller, simpler problems. Define homogeneous domains where rules can apply.
9. *Principle No 9: Teamwork* – Organize for success. Don't try to have one person do everything and do not give control of schedule and budget to the main implementer.
10. *Principle No 10: Client supremacy* – Give the client what they want and need. If the client tells you they want something then find a way to give it to them.

27.4.1 Efficiency

We believe that DSM practitioners have a tendency to lose track of principle No 1; this being that one should not try to predict anything that is easier, faster or more accurate to map directly and manually. It is common to get overly preoccupied with the elegance of predictive models and to put unwarranted effort into trying to model something that could be mapped directly more rapidly and accurately. In this project, we unashamedly used manual visual interpretation and digitizing to identify and capture information on the spatial distribution of broad patterns of variation in parent material texture and depth and of non-forested exceptions areas.

Table 27.4 indicates that the exercise to manually interpret and digitize information about the spatial patterns of variation in parent material texture and depth and of the locations of major non-forested areas cost less than 2.5 cents per ha. Any effort to model these distributions using digital data sets would have to cost less than this and produce results that were of equal or greater accuracy in order to represent a viable alternative to the manual mapping procedures (see Chapters 4, 13, 21, 22, 28, 31, 32 for other studies that reported on advantages of using manually prepared input maps).

27.4.2 Uncertainty

We contend that principle No 2, the requirement to measure and report on the uncertainty of any predictions, is also frequently under-represented in DSM projects (see

also Lagacherie, 2008). Perhaps the most important aspect of this project was that, from the onset, it had a clear definition of what the clients agreed would constitute success and also had an agreed-upon, relatively objective method for measuring that success. Assessment of uncertainty is considered by many other chapters in this book using either separate, independent validation data sets (Chapters 14, 18, 19) or, more frequently, some form of leave one out or out of the bag cross validation (Chapters 15, 17, 20, 22, 27).

The amount of effort and expense directed towards measuring and quantifying the uncertainty (or accuracy) of the maps was almost as great as the effort that went into producing the maps in the first place (Table 27.4). This was a highly valuable exercise that provided much opportunity to identify and speculate upon reasons for differences between predicted and observed patterns of spatial variation in ecological classes.

27.4.3 Parsimony

With respect to principle No. 3, the project consistently tried to adhere to the principle of parsimony. The construction of knowledge-based rules proceeded in a stepwise fashion, starting with very simple rules that used as few predictor variables as possible. New input variables were only added to the initial KB rules if the results produced by the initial rules were deemed to be incorrect or insufficient. Carefully adding one new predictor variable at a time to a specific KB rule provided an opportunity to assess the impact of the new consideration on the resulting output and to maintain control over, and understanding of, the interactions at work in the rules. Stepwise selection of predictor variables or other similar measures adopted to ensure parsimony are reported on in Chapters 3, 4, 16, 21, 30.

27.4.4 Simplicity

Our experiences also confirmed a pattern in which increases in precision were seen to result in decreases in accuracy. The more precise the definition and description of a particular map entity, the less likely that entity was to be correctly mapped. In the language of soil survey, the more splits that a mapper makes and the finer these splits are, the more difficult it becomes to successfully predict these precisely defined entities using predictive modeling. We refer to principal No. 4 and suggest that predictive modelers make a point of not being overly ambitious in terms of defining precise entities.

27.4.5 Rigor

We learned early on not to trust our assumptions blindly or base our decisions automatically on acceptance of expert opinions or judgment. All major decisions were

made after devising and implementing measures designed to test assumptions using objective criteria. We provide here one example of testing an assumption that field accuracy determinations represented truth.

Assessments of class map error using field observations made by a local expert assume that the local expert is always correct and the predictive map incorrect in all cases of non-agreement. We undertook some efforts to evaluate the degree to which different local experts could agree upon the correct classification for exactly the same locations in the field. We considered this to provide an assessment of measurement error. We had four different experts traverse and classify exactly the same four accuracy transects at different times and with no opportunity to converse or discuss their respective assessments. The results of this comparison are listed in Table 27.5.

Local experts could not agree with each other on the correct proportion of classes along the tested traverses at more than 64% average agreement, 71% if credit was given for alternate calls. The highest level of overlap agreement between any two experts was 86% and the lowest was 37%. In terms of exact agreement of classifications at exact locations (spatially congruent accuracy) average agreement between any two experts was only 42% with the lowest exact agreement between any two experts being 23% and the highest exact agreement being 73%. All four experts agreed with each other at exact locations on average 21% of the time with a low of 11% exact agreement for one traverse and a high of 30% for the best. This small test appears to suggest that it would be unrealistic to expect any predictive map to agree with the proportions of classes observed by local experts along transects in the field at any better than 64–71%.

Table 27.5 Tabulation of results of an evaluation of measurement error among four local experts

Type of Agreement or Error Assessment	Exact Class at Exact Locations Spatially Congruent Agreement	Proportions of Classes in Transect Compositional Agreement
Average Primary Call Agreement (any 2 experts)	42%	64%
Minimum Primary Call Agreement (any 2 experts)	23%	37%
Maximum Primary Call Agreement (any 2 experts)	73%	86%
Average Primary Call Agreement (all 4 experts)	21%	Not Reported
Minimum Primary Call Agreement (all 4 experts)	11%	Not Reported
Maximum Primary Call Agreement (all 4 experts)	30%	Not Reported
Average Alternate Call Agreement (any 2 experts)	Not Reported	71%
Minimum Alternate Call Agreement (any 2 experts)	Not Reported	43%
Maximum Alternate Call Agreement (any 2 experts)	Not Reported	95%

27.4.6 Continuity

Notwithstanding our recognition of the need to rigorously test all assumptions and question all expert opinions, we came to recognize the benefits of valuing and using existing expert knowledge and data. It is important to respect and value existing expert knowledge and data while simultaneously not letting it impede adoption of new or improved methods.

Our experience led us to conclude that the knowledge-based approach to predictive mapping used here is probably the only viable approach for mapping very large areas efficiently. It does not require the collection of vast numbers of field observations to support data mining to build rule bases. It is clearly parsimonious in that expert knowledge can be used to select and limit the type and number of predictor data sets. It is efficient in that it makes maximum possible use of previously acquired local expert knowledge and ecological understanding. Similar conclusions about the value of using local expert knowledge are given in Chapters 4, 13, 14, 24, 25, 28, 31.

We provide here one example of continuity between existing and new methods. Portions of the Cariboo map area were covered by traditional, manually-prepared, Terrestrial Ecosystem Maps (TEM) as well as the new PEM. The TEM and PEM maps were independent and neither had been considered in the preparation of the other. Overlaying the vector polygon boundaries of the TEM maps on top of colored hillshaded PEM maps provided an opportunity to compare and contrast the similarities and differences between manually prepared TEM maps and automated PEM maps (Fig. 27.5).

The most striking aspect of any comparison of the TEM and PEM maps is how closely similar they were in terms of the identification and location of boundaries between the main ecological entities. At a gross level, both sets of maps were obviously trying to partition the landscape in a similar fashion, with major boundaries tied to slope breaks which were interpreted as being the locations for major changes in moisture regimes and other significant site features. Both sets of maps clearly attempted to delineate similar toposequences beginning at the top with thin dry ridges and progressing down-slope through dry to mesic steep to moderate upper to mid slopes, moister, moderate to gentle mid to lower slopes, imperfectly drained,

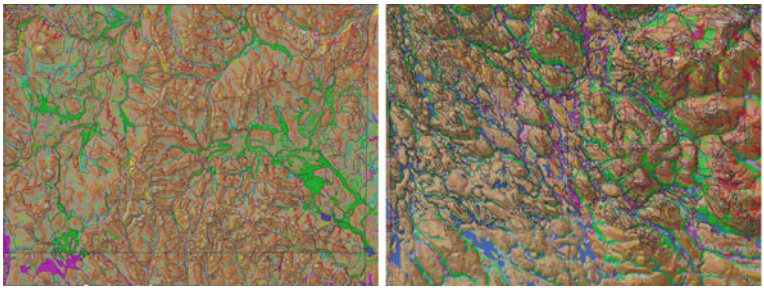


Fig. 27.5 Comparison of 1:50,000 (*left*) and 1:20,000 (*right*) vector TEM vs. gridded PEM maps

gentle to level toe-slopes, and more poorly drained level or sloping valley bottoms. The most obvious difference between the TEM and PEM maps was that, for the most part, the TEM maps did not provide as much spatially-explicit detail about within-polygon variation in ecological classes or site conditions. See Chapters 27 and 28 for discussions of similarities and differences between conventional manual maps and digital soil maps. Comparison of the two sets of maps reveals the strong degree of continuity in conceptualization of ecological spatial entities between the conventional TEM and new PEM maps.

27.4.7 Ambition (Go Big or Go Home)

Our experiences strongly support the notion that thinking big and daring to undertake mapping of very large areas yields benefits related to achieving economies of scale. Both rates of progress and measured map accuracy increased as this project progressed and knowledge-based rules were defined and iteratively improved. A library of existing rules that had been applied and tested in previous areas proved useful to have to use as a starting point for developing new rules in similar areas. It was evident that the way to complete large projects successfully was to break them up into a number of smaller areas and complete one smaller area at a time.

27.4.8 Stratification

Our experiences led us to recognize the vital importance of establishing regional physiographic and geomorphic context to define meaningful classification domains within which particular sets of rules could be developed and applied. Stratification of landscapes to improve predictive modeling is also discussed in Chapters 4, 16, 24, 25 and 28. The ability to sub-divide an entire map area into smaller, and more environmentally homogeneous, sub-areas makes it much easier to identify and apply classification rules that produce consistent and reasonable results. It also means that one big problem can be addressed as a series of smaller, and more tractable, problems. We developed and applied approaches to define and delineate classification domains automatically. However, we also recognized that there was considerable room for improvement in defining these domains.

27.4.9 Teamwork

We felt that there were advantages gained in partitioning responsibilities and building a team of individual contributors with clearly delimited responsibilities and tasks apportioned according to skills and capabilities (see Fig. 27.6). In contrast to many traditional soil survey map projects, this project did not have a single individual charged with doing the majority of the project tasks. Responsibilities were assigned

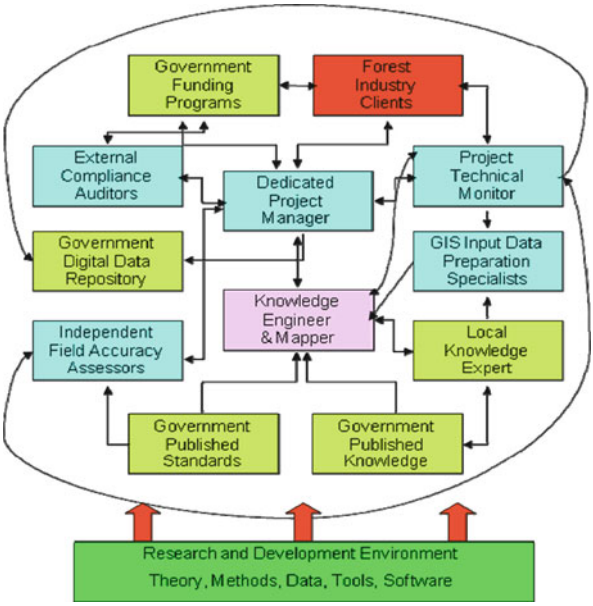


Fig. 27.6 Schematic illustration of project organization and division of responsibilities

based on identifying individuals with both the capabilities and the time to fully address each task. No one person was expected to possess all required expertise or to have the time to complete all tasks in the most efficient way. Ours was a true team approach with each participant contributing specific deliverables according to their skills and experience.

The private sector business model adopted by the project established clearly defined contracted deliverables with clearly defined and enforceable deadlines and defined measures of acceptable quality. An important element of this organization was the fact that the main implementer (the knowledge engineer) was not given control over the project schedule or budget. This rested with the project manager and the external auditors. This approach kept the project on time and on budget and ensured that the clients received what they required at a cost they could count on.

27.4.10 Client Supremacy

Finally, it should be self evident that mapping projects are undertaken for clients and that the clients should receive what they ask for and what they need when they need it. The principal responsibility of the project technical monitor in Fig. 27.6 was to ensure that the clients got what they paid for and that all deliverables represented the best effort to get them what they needed at the best possible cost and within the shortest time frame that was feasible. We describe below one example of generation of products that the client requested that were not originally envisaged or produced.

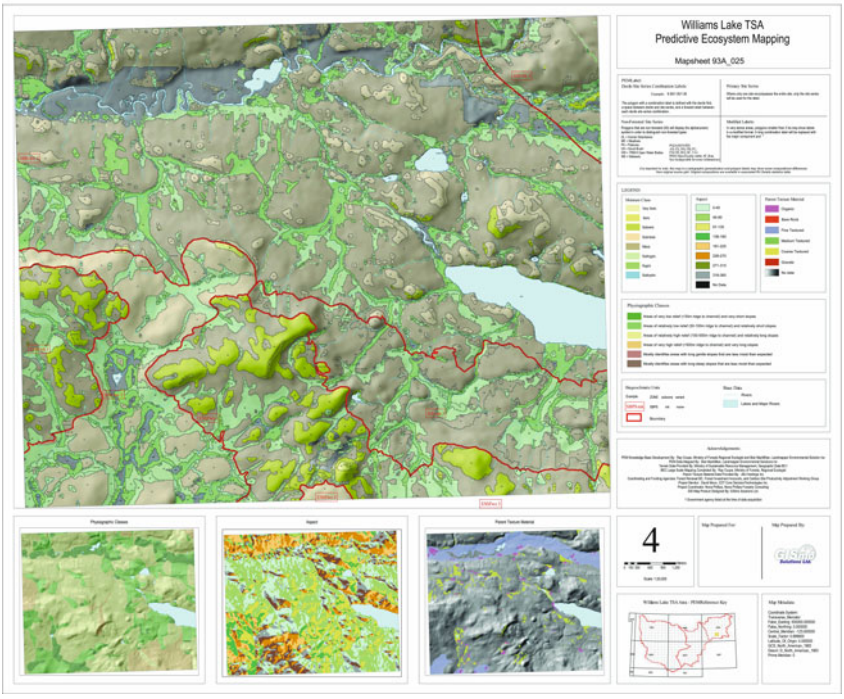


Fig. 27.7 Illustration of a final, cartographically-enhanced 1:20,000 scale vector map product

Upon completion of the initial PEM modeling, it became very apparent that there was a high demand, among some of the forest industry clients, for relatively traditional vector map products with neat polygonal boundaries and informative cartographic labels. A considerable amount of time, effort and money was consequently expended in smoothing and generalizing the original raster PEM, in converting these generalized raster maps to vector polygons that did not have excessive amounts of small polygons or jagged lines and finally in assigning and placing neat cartographic labels in each vector polygon (see Fig. 27.7). It goes to show that there is still some need for traditional cartographic products and that not all users are comfortable with, and want to use, digital maps in raster format. It also shows that the clients should always get what they want.

27.5 Conclusions

27.5.1 Successes and Accomplishments

We completed operational predictive mapping for a very large area (8.2 million ha) successfully, achieving acceptable levels of predictive accuracy (69%) while

simultaneously dramatically lowering costs (from $1.50-$3.50 to $0.34 per ha) and increasing rates of map production.

We collected and analyzed a very large and statistically valid data set of field observations that conclusively demonstrated the level of predictive accuracy achieved by the maps and revealed many of the conditions and probable causes linked to errors and inaccuracies in the maps.

We achieved enthusiastic acceptance and use of the predictive maps by both the private sector forestry clients and the relevant government ministries. At the beginning of the project, many people were skeptical about the usefulness and reliability of the predictive maps produced by this method and had a preference for the more traditional terrestrial ecosystem maps (TEM) that they were familiar with. This preference changed dramatically to favor the final PEM maps.

The PEM maps were sufficiently general in design that they were able to respond to a dramatic change in major intended use from providing information to support efforts to increase the annual allowable harvest to providing information useful for planning salvage and remediation activities to address the catastrophic effects of the mountain pine beetle infestation.

27.5.2 Opportunities for Improvement

Obviously, we would have liked to have achieved higher levels of accuracy in predicting both the proportions of ecological classes within small areas and the exact ecological class at exact point locations. Our studies showed, however, that there is an upper limit (65–71%) to the levels of accuracy that we can hope to achieve when predicting ecological classes.

We conclude that it would be desirable to investigate and adopt new methods for improving the input layers that provide geomorphic, physiographic and lithological context to classification rules. If you know that a point is located on a flood plain or fan or that the parent material texture is sand or gravel and not medium textured till, you greatly improve the ability to successfully predict the correct spatial entity at that location. Context is critical in predictive mapping.

We are less sure of our ability to improve prediction of exact classes at exact point locations. In theory, improved spatial resolution provided by new technologies for acquiring and producing fine resolution digital elevation models (LiDAR, Radar, and digital imagery), should lead to an improved ability to predict exact classes at exact point locations. Our efforts to use finer resolution DEMs (5 and 10 m) did not lead to improvements in predictive accuracy. Finer resolution DEM data present challenges for separating information into short range and longer range signals. Predictive procedures will have to adopt explicit procedures to analyze fine resolution data at multiple scales in order to profit from the more detailed information.

Acknowledgments We acknowledge with gratitude the contributions and support of the forest industry client representatives Tim Harding and John Stace-Smith of Tolko Industries Ltd. and Earl Spielman, Al Hicks and Guy Burdikin of West Fraser Mills Ltd. Contractors Timberline Forest

Industry Consultants and JMJ Holdings Ltd. produced the manually interpreted exception maps and collected the accuracy assessment field data. Meridian Mapping Ltd. prepared the DEM for the entire area. GISmo Solutions Ltd. produced the final cartographically-enhanced vector map products. Funding was received from the BC Forest Investment Account (FIA) and from the, now-superseded, Forest Renewal BC (FRBC) account. All image data copyright the province of British Columbia.

References

Bockheim, J.G., Gennadiyev, A.N., Hammer, R.D., and Tandarich, J.P., 2005. Historical development of key concepts in pedology. Geoderma 124:23–36.

Burrough, P.A., 1989. Fuzzy mathematical methods for soil survey and land evaluation. Journal of Soil Science 40:477–492.

Eyton, J.R., 1991. Rate-of-change maps. Cartography and Geographic Information Systems 18:87–103.

Hudson, B.D., 1992. The soil survey as a paradigm-based science. Soil Science Society of America Journal 56:836–841.

Lagacherie, P., 2008. Digital soil mapping: A state of the art, pp. 3–14. In: Hartemink, A.E., McBratney, A.E., and Mendonça-Santos, M.L. (eds.), Digital Soil Mapping with Limited Data. Springer, Amsterdam.

MacMillan, R.A., Moon, D.E., and Coupé, R.A., 2007. Automated predictive ecological mapping in a forest region of B.C., Canada, 2001–2005. Geoderma 140:353–373.

MacMillan, R.A., Pettapiece, W.W., Nolan, S.C., and Goddard, T.W., 2000. A generic procedure for automatically segmenting landforms into landform elements using DEMs, heuristic rules and fuzzy logic. Fuzzy Sets and Systems 113(1):81–109.

McBratney, A.B., Minasny, B., Cattle, S.R., and Vervoort, R.W., 2002. From pedotransfer functions to soil inference systems. Geoderma 109:41–73.

Meidinger, D.V., 2003. Protocol for accuracy assessment of ecosystem maps. Research Branch – British Columbia Ministry of Forests . Victoria, B.C. Technical Report 011.

Moon, D., Dunlop, D., Iles, K., and Phillips, N., 2005. A protocol for assessing thematic map accuracy using small-area sampling. Research Branch – British Columbia Ministry of Forests Technical Report 023.

Pojar, J., Klinka, K., and Meidinger, D.V., 1987. Biogeoclimatic ecosystem classification in British Columbia. Forest Ecological Management 22:119–154.

Quinn, P., Beven, K., Chevallier, P., and Planchon, O., 1991. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. Hydrological Processes 5:59–79.

Qi, F., Zhu, A.-X., Harrower, M., and Burt, J.E., 2006. Fuzzy soil mapping based on prototype category theory. Geoderma 136:774–787.

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., and Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of America Journal 65:1463–1472.

Chapter 28
Building Digital Soil Mapping Capacity in the Natural Resources Conservation Service: Mojave Desert Operational Initiative

A.C. Moore, D.W. Howell, C. Haydu-Houdeshell, C. Blinn, J. Hempel, and D. Smith

Abstract The Natural Resources Conservation Service (NRCS), within the context of the U.S. National Cooperative Soil Survey (NCSS), is working to integrate digital soil mapping methods with existing soil survey procedures. As this effort moves forward, it must address technological, managerial, and political challenges. To better understand these challenges and potential solutions, NRCS is establishing Digital Soil Mapping Operational Initiatives. These projects aim to demonstrate the utility of digital soil mapping in a production setting, provide training to soil scientists in digital soil mapping methods, contribute to completion of the initial soil survey or update of existing surveys, develop detailed instructions for implementing digital soil mapping methods, provide useful soil information products to complement existing soil survey data, and document methods and results. The first Operational Initiative was initiated at the Victorville, California Major Land Resource Area (MLRA) Soil Survey Office (SSO), which is responsible for the soil survey of Mojave Desert region. The immediate focus of this office is completing the initial soil survey for Joshua Tree National Park and adjacent private lands. Under the operational initiative umbrella, detailed digital data sets including IFSAR digital elevation models and an ASTER mosaic have been compiled. Derivatives from these and other data sets are being used to stratify the project area for sampling and modeling, and as inputs into continuous soil property predictive models. Model outputs will be used to develop Soil Survey Geographic (SSURGO) data products. Technical support for this project is provided by digital soil mapping soil scientists at the MLRA SSO the California State Office, and the National Geospatial Development Center, as well as other NRCS staff and NCSS cooperators.

A.C. Moore (✉)
USDA Natural Resources Conservation Service, 339 Busch's Frontage Road, Suite 301, Annapolis, MD 21409, USA
e-mail: amanda.moore@md.usda.gov

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_28,

Keywords Digital soil mapping · Soil survey · Mojave Desert · Organizational change

28.1 Introduction

In spite of significant advances in the development and application of digital soil mapping methods within the academic community, implementation of digital soil mapping in large and small organizations remains a challenging prospect. The Natural Resources Conservation Service (NRCS), within the context of the U.S. National Cooperative Soil Survey (NCSS), is working to integrate digital soil mapping methods with existing soil survey procedures. We define digital soil mapping as the use of digital representations of soil-forming factors, soil field data, soil maps, and a variety of statistical modeling and classification methods to produce maps of soil features. At this time, NRCS uses digital soil mapping products as pre-maps of soil variation over the landscape to guide field work and improve efficiency. In the future, digital soil mapping products may also provide potential new soil survey products showing the distribution of meaningful soil properties or interpretations of soil functions.

In order to successfully integrate digital soil mapping methods into the soil survey program, NRCS and the NCSS must address technological, managerial, and political challenges. To better understand these challenges and potential solutions, NRCS is establishing Digital Soil Mapping Operational Initiatives. These projects aim to demonstrate the utility of digital soil mapping in a production setting, provide training to soil scientists in digital soil mapping methods, develop detailed instructions for implementing digital soil mapping methods, provide useful soil information products to complement existing soil survey data, and document methods and results while contributing to the completion of initial soil surveys or update of existing soil surveys. In particular, our goal is to use these initiatives to make digital soil mapping methods part of NCSS's normal operating procedures.

Each operational initiative will be housed in a Major Land Resource Area (MLRA) Soil Survey Office (SSO) and will be integrated with the activities of that office. Soil scientists with skills in digital soil mapping, GIS, remote sensing, and/or statistics will be placed in MLRA SSOs to support operational initiative efforts. These soil scientists may be current employees with additional training or new hires.

The first formal NCSS Digital Soil Mapping Operational Initiative in the United States has been located at the Victorville Soil Survey Office in southern California. However, there are many ongoing digital soil mapping activities within the National Cooperative Soil Survey. The personnel involved in the Mojave Desert Operational Initiative project will attempt to incorporate lessons learned from these activities and from other digital soil mapping efforts around the world. The framework developed for this initiative may be used in future Operational Initiatives.

A founding principle of these projects is collaboration with the digital soil mapping community. This initiative depends on the active involvement and innovation of the research community. International collaboration and communication is also

an important part of this initiative. The collaboration network and communication infrastructure developed for this initiative will provide a foundation for a permanent international clearing house for digital soil mapping data, methods, and results.

This paper focuses on the physical characteristics of the Mojave Desert Operational Initiative study area, the operational initiative framework, and the progress to date. See Chapter 29 for a discussion of sampling and modeling procedures implemented in the Mojave Desert Operational Initiative, as well as for a comparison of results from traditional and digital soil mapping procedures.

28.2 Materials and Methods

28.2.1 Study Area

Our initial study area includes Joshua Tree National Park (JOTR) and the adjacent private lands, which comprise approximately 388,500 ha in the southern portion of the Mojave Desert Ecosystem (Fig. 28.1). Subsequent work will focus on the Mojave Desert National Preserve in the central portion of the Mojave Desert Ecosystem. The JOTR area is characteristic of the basin and range province with broad coalescing alluvial fans, internally-drained playas, and uplifted mountain ridges. Vegetation is dominated by low, widely spaced shrubs (NPS, 2008). Mean annual air temperature is about 22°C and mean annual precipitation is about 110 mm, though temperature decreases and precipitation increases as elevation increases in the western portion of the study area. The soil temperature regime ranges from hyperthermic to mesic and the soil moisture regime ranges from typic-aridic to xeric-aridic. The parent materials are mainly granitoid and metamorphic rocks. A typical JOTR landscape is depicted in Fig. 28.2.

Fig. 28.1 Location of the Mojave desert national operational initiative project within the contiguous United States

Fig. 28.2 Typical landscape within Joshua tree national park, including the park's namesake species, Joshua tree (*Yucca brevifolia* Engelm.) (Blinn, 2008, Personal communication)

28.2.2 Project Plan

Soil scientists in NRCS-California have a long tradition of using digital soil mapping concepts to develop digital products that support soil survey activities (Howell et al., 2007, 2008). This project builds on that work, and aims to fully develop those methods while integrating the process into existing soil survey operations. A detailed project plan outlining key steps in a "digital soil mapping soil survey" project was developed jointly by project participants (NRCS Staff, 2007). This project plan may serve as a template for future operational initiatives.

The original project plan spans approximately 2 years, and coincides with the timeline for completion of the initial soil survey of JOTR. Key phases of the project include:

- *Knowledge Acquisition*: This phase involves documenting existing knowledge of soils and landscapes for the project area and identifying key soil properties. A set of soil properties considered important for soil management and classification in JOTR were identified and a hierarchical decision tree documenting the environmental conditions under which particular soil series or map units occur is being developed.
- *Geospatial Database Development*: This phase involves the acquisition and processing of geospatial data. Key data layers and processing steps for the Mojave Desert initiative are identified in Section 28.2.3.
- *Landscape Stratification*: This phase involves the subdivision of the project area into physiographically similar regions in order to determine major landscape breaks and model domains, find representative areas for model development, and guide sampling procedures. See Section 24.4 for a discussion of an analogous concept (soilscapes).
- *Sampling – Phase I*: This phase involves sampling to build digital soil mapping models and document soil map units. The JOTR sampling plan was designed to utilize existing observations collected over the course of the soil survey. Gaps in existing observations were identified (see Section 28.2.3.3) and additional sam-

ples were taken. A minimum data set was identified based on soil properties slated for modeling as well as the needs of the soil survey.

- *Statistical Modeling*: This phase involves the selection and implementation of appropriate models to predict selected soil properties. We will evaluate current and past statistical modeling efforts in California and select modeling methods based on landscape, available data, and other factors. The focus will be on development of soil class and property models from terrain and spectral data using multivariate statistics.
- *Sampling – Phase II*: This phase involves iterative sampling to improve model performance and accuracy as well as collection of an independent dataset for accuracy assessment.
- *Accuracy Assessment and Documentation*: This phase involves the selection and application of quantitative accuracy assessment methods to the digital soil mapping product(s) and the integration of these methods into NCSS quality assurance standards. Accuracy assessment and documentation phases will be iterative with each annual mapping cycle.
- *Product Development*: This phase involves the production of soil information products. In accordance with National Cooperative Soil Survey guidelines, one product that will be produced for JOTR will be area-class soil maps that meet the current Soil Survey Geographic Database (SSURGO) standard. However, staff will also work to develop additional soil information products for the project area including maps of selected soil properties (e.g., depth to and presence/absence of carbonates, argillic horizon, and duripans), a decision tree model or knowledge base to capture soil-landscape relationships, sampling protocols for assessing existing soil observations and determining where additional observations may be needed, quantitative landscape models for key soil properties, and accuracy assessment procedures for continuous raster datasets. Detailed written instructions for completing successful methods will also be developed.

28.2.3 Geospatial Database Development

Key digital data layers for the area include a 5 m resolution IFSAR-derived digital elevation model (DEM), 10 and 30 m resolution DEMs from the National Elevation Dataset, a spring season ASTER Mosaic, and leaf off, leaf on, and spring season Landsat 7 ETM+ mosaics. The goal with these data is to produce derivatives that are significant model covariates and that cover the entire Operational Initiative area. See Sections 29.2.2 and 29.2.3.2 for more information about the selection and development digital data layers used in this project.

28.2.3.1 Digital Elevation Models

The IFSAR DEM data were resampled to 10 m resolution to reduce file size. Two example terrain derivatives are slope position (Fig. 28.3) and fuzzy land element classification (Fig. 28.4).

Slope position was computed from the IFSAR DEM using an Arc Macro Language (AML) script developed by Hatfield (1996) (using the following settings:

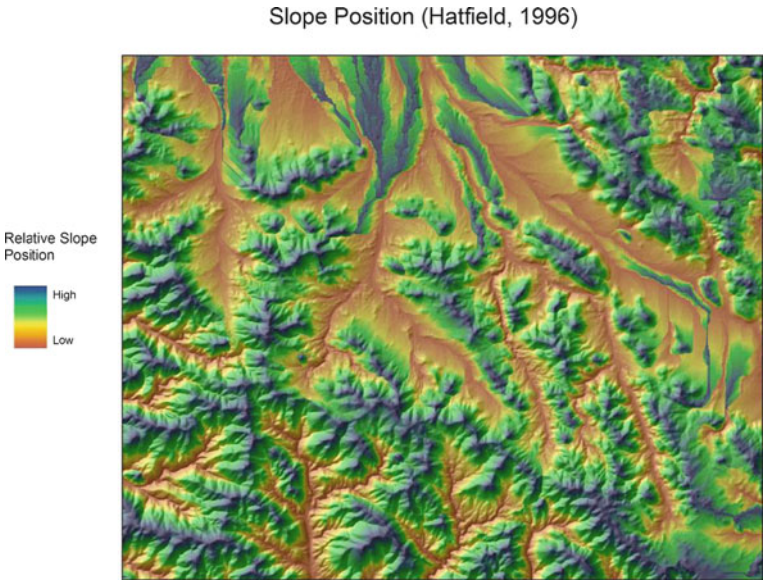


Fig. 28.3 Slope position computed from Hatfield's (1996) slopeposition AML

sink and peak fill z limit – 10 m, valley and ridge minimum accumulation – 4,000 cells). Slope position is valuable because it is calculated on localized base levels and shows relative position within nested soil catenas (Fig. 28.3). Other useful terrain derivatives in this area include multiresolution valley bottom flatness index (Gallant and Dowling, 2003), terrain ruggedness index (Riley et al., 1999), and compound topographic index (Beven and Kirkby, 1979). Other correlates are identified in Table 29.1.

Fuzzy land elements were computed from the IFSAR DEM using methods developed by Schmidt and Hewitt (2004). A 100 m^2 (1 ha) neighborhood provided the best results. Fuzzy land element classification was used to classify landform elements in an objective, explicit, and reproducible manner. These landform elements provide a new and useful set of mountain landform terms and depict elements that can be related to soil components within map units. Names for the elements within mountain landforms have been overlooked in most landform terminologies. An example classification shows the 15 landform elements (Fig. 28.4).

These data along with the satellite imagery and field data described in Sections 28.2.3.2 and 28.2.3.3 are available for download from https://sharepoint.ngdc.wvu.edu/sites/digital_soils.

28.2.3.2 Satellite Imagery

Preliminary ASTER mosaics of the study area were created from the US Geological Survey AST14DMO product, which consists of a DEM and orthorectified imagery, is terrain-corrected and contains calibrated radiance for each of the 14

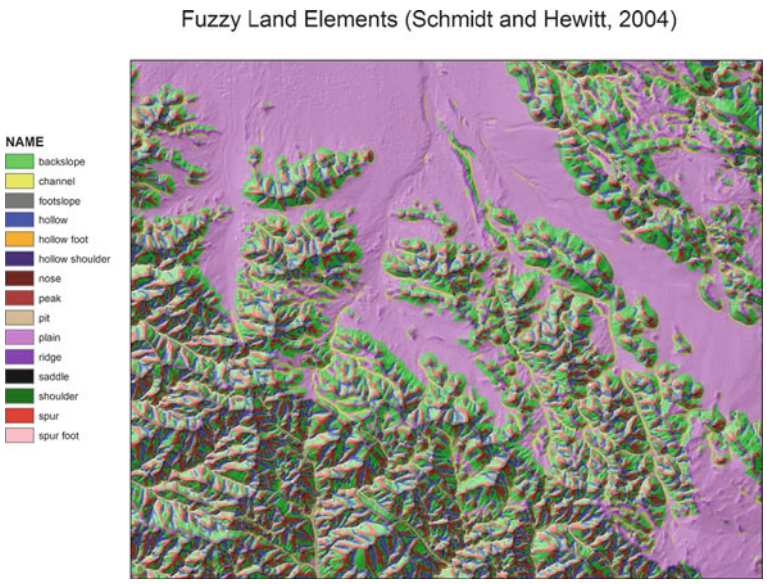


Fig. 28.4 Fuzzy land elements computed from Schmidt and Hewitt's (2004) AMLs merged into a single landform layer

bands. Ten ASTER scenes from five dates were required to cover the entire study area (Fig. 28.5). Separate images were created for visible near-infrared (VNIR), shortwave infrared (SWIR), and thermal infrared (TIR) wavelengths. Prior to mosaicking, three VNIR and six SWIR bands were converted to top of atmosphere reflectance and five TIR bands were converted to true radiance. Variation within the mosaic due primarily to phenological stage of vegetation is evident; however,

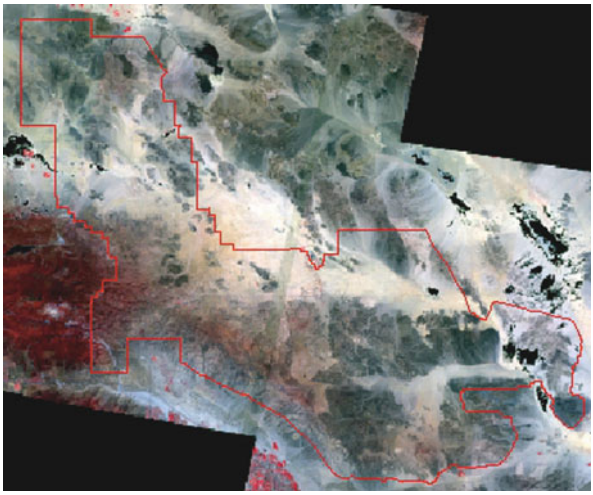


Fig. 28.5 ASTER VNIR mosaic with study area outline

future work will focus on minimizing these variations which can affect subsequent classifications. (Blinn, 2008).

Landsat mosaics for three seasons (leaf-off, leaf-on, and spring) were also developed for the study area from the Multi-Resolution Land Characterization (MRLC) consortium's 2001 National Land Cover Database (NLCD). For each season, several images were produced including normalized difference vegetation index (NDVI); tasseled-cap (TC) transformation; gypsic band ratio (Neild et al., 2007); and natric band ratio (Blinn, 2008; Neild et al., 2007). See Chapter 2 for further discussion of uses of remotely sensed imagery in digital soil mapping and Chapter 10 for additional discussion of ASTER imagery in digital soil mapping.

28.2.3.3 Field Data

Field data are being collected concurrently with the soil survey of JOTR in accordance to NRCS field data collection protocols. Approximately 487 soil observations for JOTR were collected prior to the beginning of this project. Additional observations were located using stratified random sampling (see Section 29.2.2). These data include depth to and presence/absence of carbonates, depth to and presence/absence of argillic horizon, depth to and presence/absence of duripan, depth to paralithic contact, depth to lithic contact, particle-size class, particle-size class, surface texture, and other attributes.

Initial field verification of digital data layers developed for JOTR indicate the potential for good correlations between environmental covariates and soil distributions. Fuzzy land element data layers will be used to evaluate and model individual soil components and to plan field sampling. The ASTER SWIR image may be useful for locating pediment areas. Other relationships are being evaluated.

28.2.4 Staffing and Training

Perhaps the most critical component of this project has been the development of digital soil mapping-savvy Major Land Resource Area (MLRA) Soil Survey Office teams. This development occurs by training existing soil scientists in digital soil mapping concepts and methods and by hiring new soil survey staff with skills in GIS, remote sensing, and statistics. Our goal is for digital soil mapping methods combined with strong field soil mapping knowledge to become standard practice in MLRA Soil Survey Offices.

As part of this initiative, a new soil scientist with a strong academic background in GIS, remote sensing, terrain analysis, and statistics as well as field skills in describing and mapping soils was hired by NRCS-California for the Victorville Soil Survey Office.

Collaboration with NCSS and other partners is another important way to build skills within soil survey office teams. To supplement in-house expertise, a post-doctoral research associate with expertise in remote sensing was contracted to assist

with the acquisition of satellite imagery and development of mosaics and derivative products as well as to advise survey members on issues related to remote sensing.

28.3 Results and Discussion

Results from this initial sampling and modeling cycle are discussed in detail in Section 29.3. We will focus our discussion in this paper on the process of institutionalizing new procedures.

Balancing the adoption of new methods, including digital soil mapping methods, with high production goals is necessary if those methods are to be integrated into the workflow of a typical soil survey office. Gradual implementation of digital soil mapping methods allows soil survey staff time to learn new methods and determine how they can best be applied, while still permitting them to meet challenging production goals. As always, the long term goal of implementing digital soil mapping methods in a coordinated manner is to improve the quality of soil survey data and increase the efficiency of soil survey operations. However, these improvements may not be seen until new methods are well integrated into current workflows. An excellent example of the development of a digital soil mapping process tied directly to an organization's workflow is presented in Chapter 31 (Sections 31.1 and 31.3 in particular).

Support from agency employees with expertise in digital soil mapping, including state, regional, and national GIS and soil specialists as well as other MLRA soil scientists, facilitates technology transfer within the soil survey program. Collaboration with NCSS and other partners ensures access to individuals with skills we may not yet have within NRCS and exposes us to new ideas that may benefit the soil survey program in the future. However, we believe that digital soil mapping methods will be more quickly and effectively integrated into day to day soil survey operations when MLRA Soil Survey Offices have individuals with expertise in digital soil mapping on site.

Development of a detailed project plan, while time-consuming, provides an opportunity for each collaborator to contribute to the development of the project, and ensures that all project participants understand the project goals and timeline. In addition, these plans meet NRCS Soil Survey Division requirements for project plans and can be incorporated into an MLRA Soil Survey Office's annual work plan and long range plan.

To become accepted, new methods must enhance a soil surveyor's understanding of soil-landscape relationships and improve the ways these relationships can be extrapolated and represented. In the context of this operational initiative, the soil survey team is documenting their understanding of soil-landscape relationships in JOTR via a detailed mapping key. Digital data layers such as fuzzy land elements and ASTER VNIR and SWIR images are being used to assist with the extrapolation of these relationships across the study area.

Further discussion of factors in the successful implementation of operational digital soil mapping is presented in Section 27.4, where 10 guidelines are proposed.

28.4 Conclusions

Integration of digital soil mapping methods into existing NCSS soil survey protocols is a challenging task. Digital soil mapping activities ranging from simple GIS exercises to sophisticated statistical models are occurring in soil survey offices around the US on an *ad hoc* basis; our goal with this and future operational initiatives is to formalize the application of digital soil mapping methods and provide a framework within which digital soil mapping can grow.

After this first year of implementation, model outputs, sampling protocols, and accuracy assessments from the Mojave Desert Operational Initiative are being evaluated and publicized. With each cycle of the soil survey, the MLRA SSO will gradually incorporate more digital soil mapping methods into their workflow. Soil survey in the U.S. has always developed in these annual cycles: soil scientists are trained in new methods during the office season and apply these methods during the field season. As digital soil mapping methods are added to the annual sequence of soil survey operations, the soil survey program changes. The operational initiatives will provide a framework to coordinate and communicate these advancements.

References

Beven, K.J., and Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrological Sciences Bulletin 24:43–69.

Blinn, C., 2008. 3rd Quarter, FY 2008 Progress Report. Digital Soil Mapping Operational Initiative – Mojave Desert Region. Submitted to NRCS-National Geospatial Development Center in July 2008.

Gallant, J.C., and Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research 39:ESG41–ESG413.

Hatfield, D., 1996. Slopeposition AML. USDA Forest Service, Portland, OR.

Howell, D., Kim, Y., Haydu-Houdeshell, C., Clemmer, P., Almaraz, R., and Ballmer, M., 2007. Fitting soil property spatial distribution models in the Mojave Desert for digital soil mapping, pp. 465–475. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science, Vol. 31. Elsevier, Amsterdam.

Howell, D., Kim, Y., and Haydu-Houdeshell, C., 2008. Development and application of digital soil mapping within traditional soil survey: What will it grow into? pp. 43–51. In: Hartemink, A., McBratney, A.B., and Mendonça-Santos, L. (eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Neild, S.J., Boettinger, J.L., and Ramsey, R.D. 2007. Digitally mapping gypsic and natric soil areas using Landsat ETM data. Soil Science of America Journal 71:245–252.

NPS (U.S. National Park Service). 2008. Joshua Tree National Park- Mojave Desert, http://www.nps.gov/jotr/naturescience/ (last verified 2 September 2008).

NRCS Staff, 2007. Mojave Desert Digital Soil Mapping Operational Initiative Project Plan, https://sharepoint.ngdc.wvu.edu/sites/digital_soils/mojave/Shared%20Documents/Mojave_DSM_plan_v2%20091207%20California.doc (last verified 8 September 2008).

Riley, S.J., DeGloria, S.D., and Elliot, R., 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Intermountain Journal of Sciences 5:23–27.

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

Chapter 29
A Qualitative Comparison of Conventional Soil Survey and Digital Soil Mapping Approaches

S.M. Roecker, D.W. Howell, C.A. Haydu-Houdeshell, and C. Blinn

Abstract Research in digital soil mapping has indicated that its methodologies could be successfully extended to a field setting where it could enhance the quality and scientific foundation of soil surveys, as well as save time and money. These assumptions are being put to the test in the Mojave Desert of California as part of a continuing effort within ongoing soil surveys. One of the questions being posed in this study is how comparable is a third-order soil map created by the conventional soil survey approach to one created via digital soil mapping. To compare the two mapping approaches the subgroup level of Soil Taxonomy was chosen as the response variable and the map unit as the unit of comparison. Within each map unit the proportion and number of subgroups predicted by each mapping approach was qualitatively compared. Within the intermontane basins the predictive model used for digital soil mapping, estimated a smaller proportion and number of soil subgroups associated with fan remnants. Whereas within the mountains the conventional approach predicted fewer soil subgroups, and a different composition of soil subgroups.

Keywords Soil survey · Digital soil mapping · Soil Taxonomy · Random forest · Mojave Desert

29.1 Introduction

Research in digital soil mapping has indicated that its methodologies (i.e., pedometric techniques) could be successfully extended to a field setting, where they could enhance the quality and scientific foundation of soil surveys, as well as save time and money. As part of a continuing effort within ongoing soil surveys in the Mojave Desert of California, USA, pedometric techniques are being explored so as to hopefully realize such benefits (Howell et al., 2007, 2008). One of the questions posed in this study, is how comparable are maps of soil types created by conventional soil

S.M. Roecker (✉)
USDA Natural Resources Conservation Service, Victorville, CA 92392, USA
e-mail: stephen.roecker@ca.usda.gov

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_29,

survey to digital soil mapping? Fundamentally both approaches to soil mapping are similar in that they make use of relationships between the soil and more readily observable land surface properties, such as shape, position, and reflectance. Beyond this, both can take different approaches to all aspects of soil mapping, including project planning and preparation, sampling design, field operation, soil measurement, predictive modeling, and geographic representation. Ultimately these differences mean that any comparison between these two approaches cannot be wholly quantitative, unless based on an estimate of the amount of variance explained by each of these methods for predicting soil properties. No such dataset existed for this study. Still a qualitative comparison of mapping approaches is warranted so as to determine what benefits or challenges might be realized from adopting pedometric techniques within a production soil survey setting.

Although much has been written on the benefits of digital soil mapping relative to conventional soil survey, few studies have directly compared the accuracy of both approaches for predicting soil classes. One of the earlier studies done by Skidmore et al. (1996) reported comparable levels of forest soil map accuracy between the conventional approach (74%) and a Bayesian expert system coupled with a geographic information system (GIS) (70%). In two case studies in differing terrain, Zhu et al. (2001) reported that soil maps produced by conventional soil survey were less accurate (61 and 67%) than soil maps produced by the soil-land inference model (SoLIM) (81 and 84%), which is a similarity representation based on GIS and expert-knowledge. When predicting ecological classes for the Cariboo Forest Region (8.2 million ha) within British Columbia, Canada, MacMillan et al. (see Table 3 in Chapter 27) achieved accuracies of 65–70%, by formalizing the existing expert knowledge using a fuzzy Semantic Import model and a series of terrain attributes, multispectral imagery, and manually interpreted layers. In contrast, MacMillan et al. (see Section 27.4.5) reported only 64% agreement among the assessors when examining the same transects. Thompson and Kolka (2005) found that topsoil organic carbon estimated from published soil surveys were about half that estimated by linear regression using terrain attributes. For predicting soil drainage class, Bell et al. (1994) found general agreement (67%) between a soil survey and a quantitative soil-landscape model based on parent material, terrain attributes, and surface drainage feature proximity.

First, these studies demonstrate that regardless of the soil mapping approach a large amount of unexplained variability remains, which is to be expected given the limitations of our pedogenic models (Wilding, 1994). Second, many authors assume that they would expect greater accuracies with better spatial data for model development. However in our experience (Howell et al., 2008), simply increasing the spatial and spectral resolution of our spatial data has not dramatically improved our results. Third, digital soil mapping provides more spatially explicit and detailed estimates of soil spatial variability. While digital soil mapping is limited by the resolution of the spatial data, the conventional approach is limited by the scale of the base map and the inability to represent continuous spatial variation. It is worth noting that most approaches to digital soil mapping are limited by the availability of soil

data, and therefore can not readily incorporate the existing pedogenic knowledge of soil scientists (McKenzie et al., 2000). In light of these considerations, the objective of this study was to compare third-order soil maps produced by conventional soil survey to digital soil mapping in a study area in the Mojave Desert, USA.

29.2 Materials and Methods

29.2.1 Study Area

The 150-km^2 study area was the 7.5' Joshua Tree South quadrangle in the Western Mojave Desert of Southern California (Fig. 29.1), which overlaps the northern boundary of Joshua Tree National Park. Access to both the national park and private lands was limited due to the area's rough terrain and trespass issues on private lands. The study area is part of two ongoing soil surveys that are mandated to be finished in less than 5 years. The survey areas together encompass approximately 710,000 ha of land. These areas occur within the Basin and Range Physiographic Province, which is characterized by small isolated mountain ranges that protrude from large alluvial-filled intermontane basins (Peterson, 1981). Within the Mojave Desert sub-province the soil temperature regime is predominantly thermic, while the soil moisture regime is predominantly aridic.

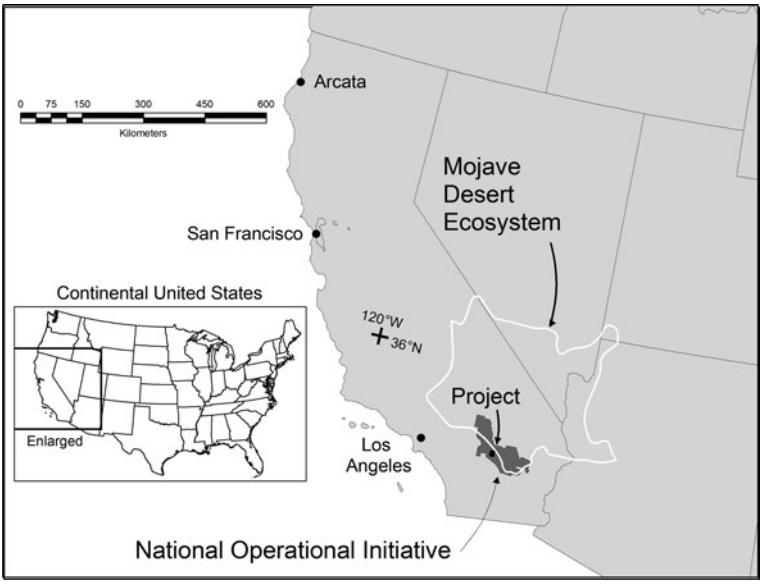


Fig. 29.1 Location of the project area in the southwestern United States

29.2.2 Sampling

Following the example of McKenzie et al. (2000), a combination of stratified-random and purposive sites were sampled (Fig. 29.2). This approach was chosen to reduce sampling bias while ensuring that areas of known pedogenic significance were not overlooked.

A limited set of stratifying variables were chosen that conformed to the existing pedogenic knowledge of the area. The foremost stratifying variable was the multi-resolution valley bottom flatness (MRVBF) index (Gallant and Dowling, 2003), which was used to separate the mountains and intermontane basins within the study area. This index provides a suitable separation of both mountains and intermontane basins, as it estimates both flatness and lowness by combining measures of slope gradient and elevation percentile (Gallant and Wilson, 2000) computed over multiple digital elevation model (DEM) resolutions.

Upon separating the mountains and intermontane basins, each of these domains was stratified further by a combination of terrain attributes and band ratios. Specifically within the mountains, tangential curvature (Kt) (Gallant and Wilson, 2000) was separated into three classes using Jenks optimization method (Jenks, 1967), also known as natural breaks within ArcGIS. This local measure of flow convergence and divergence was meant to roughly separate out land elements such as spurs, backslopes and hollows.

Within the intermontane basins, catchment slope (CS) (Gallant and Wilson, 2000) was separated into four quantile classes. Here catchment slope was interpreted as distinguishing between stable and active land elements.

For both domains the band ratios used for stratification were derived from Landsat Enhanced Thematic Mapper Plus (ETM+), and designed to enhance iron (3/7), carbonates (3/2), and clay (5/7). To harness the three band ratios simultaneously

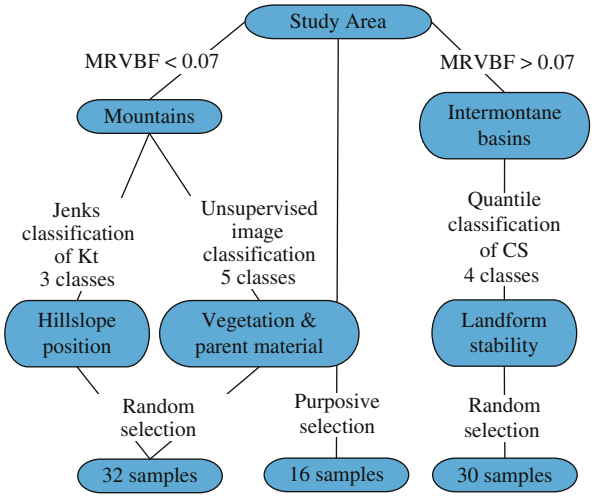


Fig. 29.2 Illustration of sampling plan. (MRVBF = Multiresolution valley bottom flatness; Kt = tangential curvature; CS = Catchment slope)

a hierarchical unsupervised classification was derived from a thousand point random sample of each domain for the entire National Park using Ward's method (Ward, 1963). This classification suggested five classes for each domain, which were then used to create a signature file and classify the remaining study area using maximum likelihood classification. When visually examined the classification appeared to distinguish between distinctly different parent materials and vegetation communities. When examined within the study area's intermontane basins, only one class was interpreted as meaningful, therefore the five image classes for this domain were merged into one.

The intersection of these variables produced sixteen strata. After a series of exclusion rules were used to mask unrepresentative and inaccessible areas, a total of 62 random sites were generated using Hawth's Analysis Tools for ArcGIS (Hawthorne, 2009), and apportioned amongst the strata proportional to area. To mask unrepresentative areas, a series of exclusion rules were designed to eliminate areas within <10 m of a road, and individual strata that were $<$ 1 acre (<0.4 ha) within the mountains and <10 acres (<4 ha) with the intermontane basins. To mask inaccessible areas, elusion rules were designed to eliminate areas >2 miles (>3.2 km) from a road or a cost distance (based on slope) $>27{,}000$. The Soil Survey Project Leader determined that several distinct landforms were under-represented, thus 16 additional purposive sites was also sampled. At all sites, a soil pit was excavated to 150 cm or a limiting layer, described according to Schoeneberger et al. (2002), and classified according to Soil Taxonomy (Soil Survey Staff, 2006).

29.2.3 Soil Mapping Approaches

29.2.3.1 Conventional Soil Survey

The 1-m National Agriculture Imagery Program (NAIP) aerial photography of the study area was examined using a stereoscope. Map units were delineated mainly on observable similarities in landforms or groups of landform components (USDA-NRCS, 2008). Once delineated, the percentage of major and minor ($<15\%$ of unit) landform components of each soil mapping unit were estimated by visually studying the imagery or by field reconnaissance. After field descriptions were completed for all locations within all mapping units, profiles were classified (Soil Survey Staff, 2006), soil information correlated and compiled and assembled into map unit descriptions. All map units were derived prior to observing the results from the random forest to avoid biasing the conventional mapping approach.

29.2.3.2 Digital Soil Mapping

The statistical model used for soil spatial prediction was random forests (Breiman, 2001), which was implemented in the statistical computing environment R (R Development Core Team, 2009) using the randomForest package. (Liaw and Wiener, 2002). To train the model, 500 trees were grown each with a minimum

terminal node size of seven. To construct each individual tree, a bootstrap sample of approximately two thirds of the observations and only 6 variables were used. Using the observations omitted from the bootstrap sample within each tree, termed the "out of bag" data, an average error rate was computed to assess the goodness of fit of the model. Also using the "out of bag" data, estimates of variable importance were generated to determine the most influential variables within the model.

The environmental covariates used in this study included a wide range of predictive variables, listed in Table 29.1. The primary data sources for these covariates were a 5-m DEM derived from interferometric synethic aperature radar (IFSAR), multispectral imagery acquired from Landsat ETM+ on the 6th of July 2000 from path 39 and row 36, and climate data interpolated by Michaelsen (2002a, b, c, d). Prior to computing the terrain attributes, the DEM was resampled from its original horizontal resolution of 5–15-m using the nearest neighbor method and hydrologically corrected using the impact reduction technique implemented within the Terrain Analysis System (TAS) (Lindsay, 2005). To compute the geometric terrain attributes Landserf (Wood, 2008) was used, while for the hydrologic and topo-climatic terrain attributes SAGA (Bock et al., 2008) was used. For the purposes of spatial analysis all of the environmental covariates, other than the terrain attributes were resampled to common grid size of 15-m by cubic convolution.

29.2.4 Qualitative Comparison Between Conventional and Digital Soil Mapping Approaches

To assess the two soil mapping approaches a qualitative comparison was made of the proportion and number of soil types estimated by each mapping approach within the map units delineated by the conventional approach. For the digital soil mapping approach this was estimated by intersecting the output of the random forest with the map unit polygons. The soil type estimated within each map unit was the subgroup level of Soil Taxonomy. This taxonomic level was chosen because it has a good balance between its overall number of classes and the amount of information contained within them.

29.3 Results and Discussion

29.3.1 Results of the Conventional Soil Survey and Digital Soil Mapping Approaches

Within the study area a total of sixteen soil subgroups were observed (Table 29.2). While sites were sampled across the entire study area only seven map units (Table 29.4) were developed for approximately 75% of the study area, using 51 observations. More map units have yet to be developed, but it was decided that the

Table 29.1 Environmental covariates used for spatial prediction by the random forest

Environmental covariates		Definition	References
Geometric terrain attributes	Slope gradient (SG)	Downward gradient (percent)	Wood (1996)
	Profile curvature (KP)	Curvature in the direction of the maximum downward gradient (radians)	Wood (1996)
	Tangential curvature (KT)	Curvature perpendicular to the direction maximum downward gradient (radians)	Wood (1996)
	Contour curvature (KC)	Curvature along the contour (radians)	Wood (1996)
	Mean curvature (KM)	Mean of profile and tangential curvature (radians)	Wood (1996)
Hydrologic terrain attributes	Flow accumulation (FA)	Upslope catchment area (meters^2)	Quinn et al. (1991)
	Flow path length (FPL)	Flow distance from ridge (meters)	Gallant and Wilson (2000)
	Catchment height (CH)	Mean elevation of upslope catchment area (meters)	Gallant and Wilson (2000)
	Catchment slope (CS)	Average gradient of upslope catchment area (percent)	Gallant and Wilson (2000)
	Topographic wetness index (TWI)	ln(FA/SG)	Wilson and Gallant (2000)
	Stream power index (SPI)	ln(FA*SG)	Wilson and Gallant (2000)
	Multiresolution valley bottom flatness (MRVBF)	Measure of flatness and lowness	Gallant and Dowling (2003)
	Multiresolution ridgetop flatness index (MRRTF)	Measure of flatness and upness	Gallant and Dowling (2003)
Topo-climatic terrain attributes	Altitude above channel network (ZC)	Height above interpolated channel surface (meters)	Bohner and Antonic (2009)
	Morphometric protection index (MPI)	Sum of maximum gradient over a 2000m radius	
	SAGA flow accumulation (FASG)	Iteratively modified upslope catchment area (meters^2)	Bohner and Antonic (2009)
	Duration of insolation (DRAD)	Amount of time under illumination (hours)	Wilson and Gallant (2000)
	Solar radiation (SRAD)	Amount of incoming solar radiation (kWh/mA^2)	Wilson and Gallant (2000)

Table 29.1 (continued)

Environmental covariates		Definition	References
Multispectral imagery	Bands 1-7	Path 39 row 36, L7, July 6th 2000	
	Normalized vegetation index (NDVI)	(b4-b3)/(b4+b3)	
	Natric index	(b5-b7)/(b5+b7)	Nield et al. (2007)
	Gypsic index	(b5-b4)/(b5+b4)	Nield et al. (2007)
	Tasseled-cap components	soil brightness (TC1), greenness (TC2), yellow stuff (TC3)	
	Carbonates (BR32)	b3/b2	Clemmen (2003, Band-ratio Landsat 5 Thematic Mapper Imagery, Personal communication, USDI Bureau of Land Management)
	Ferrous iron (BR37)	b3/b7	Clemmen (2003, Band-ratio Landsat 5 Thematic Mapper Imagery, Personal communication, USDI Bureau of Land Management)
	Clay (BR57)	b5/b7	Clemmen (2003, Band-ratio Landsat 5 Thematic Mapper Imagery, Personal communication, USDI Bureau of Land Management)
Geology	Earth material	Granitoid and felsic metamorphic rocks	MDEP (2000)
Climate	January average minimum temperature (JANTEMP) July average maximum temperature (JULTEMP) Summer precipitation (SUMPPT) Winter precipitation (WINPPT)	Interpolated from data spanning 1961–1990.	Michaelsen (2002a, b, c, d)

Table 29.2 Contingency table of the number of soil observations by soil subgroup and map unit

Map units									
Physiographic domain	Intermontane basins					Mountains			
Soil subgroup / component	CH06	DHCH1	DHCH2	DW	JS5	JS4	SMR3	Other	Total
Arenic Haplargids	0	2	0	0	0	0	0	0	**2**
Arenic Paleargids	0	0	0	0	1	0	0	0	**1**
Cambidic Haplodurids	0	0	0	0	0	0	0	1	**1**
Lithic Xeric Haplocambids	0	0	0	0	0	0	1	0	**1**
Lithic Torriorthents	0	0	0	0	0	1	2	2	**5**
Lithic Torripsamments	0	0	0	0	0	1	3	1	**5**
Torriorthentic Haploxerolls	0	0	0	0	0	3	1	2	**6**
Torripsammentic Haploxerolls	0	0	0	0	1	2	1	0	**4**
Typic Haplargids	0	1	1	0	1	0	0	3	**6**
Typic Haplocalcids	1	0	0	0	0	0	0	1	**2**
Typic Haplocambids	1	0	0	0	0	0	0	0	**1**
Typic Torriorthents	0	0	0	0	0	0	2	2	**4**
Typic Torripsamments	4	1	2	2	7	1	3	7	**27**
Xeric Torriorthents	0	0	0	0	0	1	0	2	**3**
Xeric Torripsamments	0	0	0	0	0	3	1	6	**10**
Total	**6**	**4**	**3**	**2**	**10**	**12**	**14**	**27**	**78**

sampling strategy did not provide adequate documentation to develop map units for some areas. In comparison the random forest was trained on all 78 observations, which would seem to suggest that digital soil mapping can make more efficient use of soil observations. This would be a naïve interpretation though, as soil scientists often take into account soil-landscape relationships observed elsewhere, assuming that similar soil-forming factors exist. Extrapolation of soil-landscapes relationships within soil survey is practiced when dealing with inaccessible lands in remote areas, which is a common problem within the western United States. For this reason many within soil survey perceive digital soil mapping to be of potential value.

The results of the random forest using all 78 observations gave an OOB error estimate of 51% or conversely an OOB overall accuracy of 49% (Table 29.3). Due to the low frequency of most classes, those with less than five observations were incorrectly classified. However in many cases the incorrect class was a similar soil or occurred in a similar landscape position. Stum et al. (see Section 15.3) has interpreted this to indicate the occurrence of areas that might best be mapped as associations or complexes, and therefore best be handled by attaching multiple membership to such classes. According to the variable importance measure, mean decrease in accuracy, the most important ten variables listed in order of importance were as follows FASG, MRVBF, FA, TWI, SPI, CH SG, JANTEMP, JULTEMP, and MRRTF. Amongst the top ten variables over half were hydrologic terrain attributes.

Table 29.3 Confusion matrix from the random forest

Soil subgroup / component		Predicted																
Observed		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total	Producer's accuracy
Arenic Haplargids	**1**	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	**2**	**0**
Arenic Paleargids	**2**	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	**1**	**0**
Cambidic Haplodurids	**3**	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	**1**	**0**
Lithic Xeric Haplocambids	**4**	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	**1**	**0**
Lithic Torriorthents	**5**	0	0	0	0	3	0	0	0	0	0	0	0	1	0	1	**5**	**0.6**
Lithic Torripsamments	**6**	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	**5**	**0**
Torriorthentic Haploxerolls	**7**	0	0	0	0	0	2	2	0	0	0	0	0	0	0	2	**6**	**0.33**
Torripsammentic Haploxerolls	**8**	0	0	0	0	0	0	0	0	1	0	0	0	1	0	2	**4**	**0**
Typic Haplargids	**9**	0	0	0	0	0	0	0	1	2	0	0	0	3	0	0	**6**	**0.33**
Typic Haplocalcids	**10**	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	**2**	**0**
Typic Haplocambids	**11**	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	**1**	**0**
Typic Torriorthents	**12**	0	0	0	0	1	0	0	0	1	0	0	0	2	0	0	**4**	**0**
Typic Torripsamments	**13**	0	0	0	0	1	0	0	0	2	0	0	0	24	0	0	**27**	**0.89**
Xeric Torriorthents	**14**	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	**3**	**0**
Xeric Torripsamments	**15**	0	0	0	0	0	0	2	0	0	0	0	0	1	0	7	**10**	**0.7**
Total		**0**	**0**	**0**	**0**	**5**	**2**	**6**	**1**	**9**	**0**	**0**	**0**	**38**	**0**	**17**	**78**	**Overall accuracy**
User's accuracy		**0**	**0**	**0**	**0**	**0.6**	**0**	**0.33**	**0**	**0.22**	**0**	**0**	**0**	**0.63**	**0**	**0.41**		**0.49**

29.3.2 Comparison Between Mapping Approaches

Within the intermontane basins six map units were delineated (Table 29.4; Fig. 29.3). The majority of soil subgroups within this physiographic domain are either characterized as having a sandy particle-size throughout (i.e., Typic Torripsamments), or as having an argillic horizon (i.e., Typic Haplargids) (Soil Survey Staff, 2006). The presence of an argillic or other diagnostic subsurface horizon within this environment is inferred as the result of an older land surface, which is referred to as a fan remnant. In comparing the results of the two mapping approaches within Table 29.4, it is apparent that both methods predicted predominately Typic Torripsamments and Typic Haplargids. Discrepancies between the two approaches exist in the absolute composition and number of components predicted. For the random forest all soil subgroups associated within fan remnants were classified as Typic Haplargids. Meanwhile, the conventional approach recognized an additional number of soil subgroups associated with fan remnants, but each were estimated to be less than 5% for all but one map unit (DHCH1). The conventional approach consistently estimated more soil subgroups and soil components than did the random forest, particularly in those map units with good visible expression of fan remnants such as DHCH1 and DHCH2. Given that the correspondence between the random forest and soil observations within these map units appears no less accurate than the overall map, it is unclear whether the random forest underestimates the true proportion of soil subgroups associated with fan remnants. It may be that the spatial resolution of the environmental covariates failed to accurately portray the fan remnants, but our experience has shown that increasing the spatial resolution, of terrain attributes in particularly, creates the alternative problem of increasing the amount of noise present in the spatial data (see Chapter 5).

Within the mountains two map units were delineated (Table 29.4; Fig. 29.3). There are no dominant soil subgroups within either map unit. To simplify the map unit design in response to the amount of taxonomic diversity observed, similar soil subgroups were correlated in the conventional approach. In comparison the random forest predicted a different composition and greater number of soil subgroups. Between the two map units, JS4 has the most agreement between the two mapping approaches. In both map units the conventional approach predicted Lithic Torripsamments and Torripsammentic Haploxerolls (sandy particle-size throughout), whereas the random forest predicted predominantly Lithic Torriorthents and Torriorthentic Haploxerolls (no sandy particle-size throughout). This discrepancy between the two mapping approaches may be the result of the sampling plan, which from the outset evenly distributed the sampling locations across the study area, but not the eventual map units. Although we expected to capture any uneven sampling with our additional purposive samples. Therefore the random forest may have exploited soil-landscape relationships that cut across map unit lines, by pooling the additional soil observations from across the study area. This suggests that the map units developed by the conventional approach, maybe over-specified to the soil observations made within the map unit.

Table 29.4 Contingency table of the proportion of each soil subgroup by map unit

	Conventional component percentages							Random forest component percentages						
Physiographic domain	Intermontane basins					Mountains		Intermontane basins					Mountains	
Soil subgroup / component	CH06	DHCH1	DHCH2	DW	JS5	JS4	SMR3	CH06	DHCH1	DHCH2	DW	JS5	JS4	SMR3
Arenic Haplargids	0	85	0	0	4	0	0	0	0	0	0	0	0	0
Arenic Paleargids	0	0	0	0	3	0	0	0	0	0	0	0	0	0
Cambidic Haplodurids	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lithic xeric Haplocambids	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lithic Torriorthents	0	0	0	0	0	0	0	0	0	2	0	0	2	10
Lithic Torripsamments	0	0	0	0	0	0	65	0	0	0	0	0	2	2
Torriorthentic Haploxerolls	0	0	0	0	0	0	0	0	0	0	0	0	33	17
Torripsammentic Haploxerolls	0	0	0	0	0	25	0	0	0	0	0	1	3	3
Typic Haplargids	5	0	32	0	20	0	0	12	14	8	4	14	1	2
Typic Haplocalcids	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Typic Haplocambids	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Typic Torriorthents	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Typic Torripsamments	90	14	58	98	73	8	15	88	86	90	96	84	19	31
Xeric Torriorthents	0	0	0	0	0	5	0	0	0	0	0	0	0	0
Xeric Torripsamments	0	0	0	0	0	47	0	0	0	0	0	1	40	34
Rock Outcrop	0	0	10	0	0	15	20	NA	NA	NA	NA	NA	NA	NA
Riverwash	0	1	0	2	0	0	0	NA	NA	NA	NA	NA	NA	NA

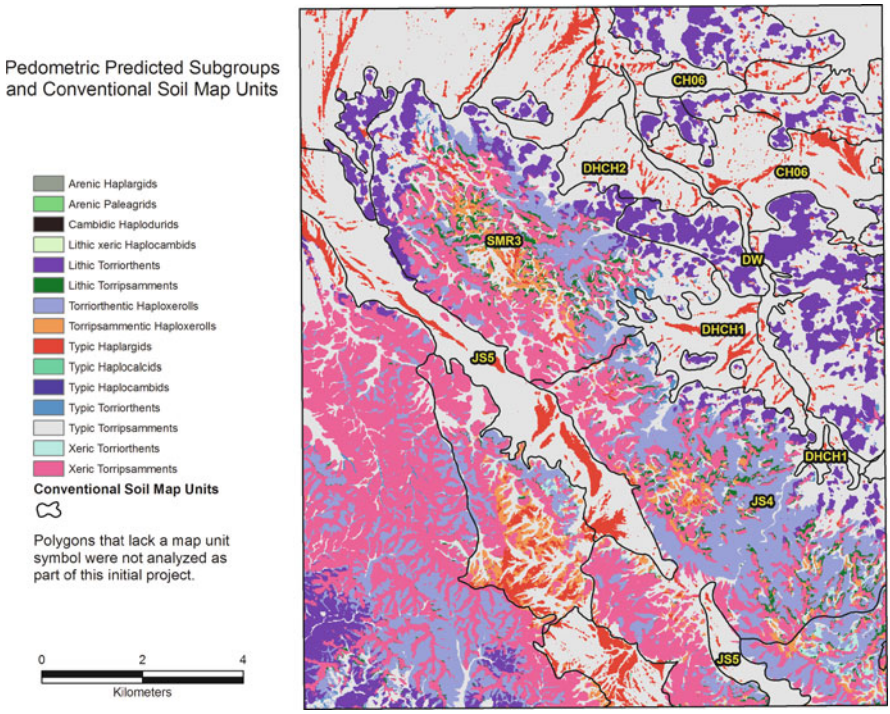


Fig. 29.3 Subgroup prediction from the random forest overlaid with the soil map unit lines. Those polygons which are not labeled were predelineated, but not finalized

Further complicating the interpretation of these results is the inclusion of rock outcrop and riverwash within the composition of the conventional map unit estimates. Estimates of miscellaneous areas, such as rock outcrop, are often included as components within map units, but are visually estimated and not sampled directly. As the size of rock outcrops within the study area fall well below the spatial resolution of the environmental covariates, their proportion within each pixel might best be estimated separately.

Last but not least, the difference between the two soil mapping approaches geographic representation is unmistakable. Both approaches resulted in a crisp separation between the mountains and intermontane basins within the study area, but the random forest provided much additional detail by predicting an individual soil component for each pixel. The conventional approach also provides spatial information on the location of its soil components, but this information is descriptive only and requires the map user to infer the location of the components within each map unit. Although the random forest neglected to predict the occurrence of some soil components, this does not imply they are absent, but rather improbable within any given pixel.

29.4 Conclusions

This study presents a qualitative comparison between third-order soil maps produced by conventional soil survey and digital soil mapping methodologies. On the one hand the digital soil mapping methodology used – random forest – predicted an individual soil subgroup for each pixel, with which also came an estimate of the uncertainty associated with that prediction. The conventional approach on the other hand delineated areas of land within which exist a complex of dissimilar soil subgroups. The location of soil subgroups within a map unit were described by their landscape position. No estimate of uncertainty was given, but the map unit purity was reflected by the number and proportion of soil components. In comparison both approaches produced not only differing amounts of spatial detail, but also differing compositions and numbers of soil components within map units. The difference between these two approaches suggests that further investigation is required in this study area.

For the time being the question remains: how should soil scientists use the outputs of digital soil mapping methodologies? Given the current set of soil survey standards within the USA we see the results of digital soil mapping methodologies as an instrument to help estimate map unit compositions and guide delineations. In addition they could also be used to provide a new form of map unit documentation and as a tool to educate soil information users as to the variable composition of map units. In the future the question will likely be determined by whatever mapping approach is called for by the map user. MacMillan et al. (see Section 27.5.1) reported that in his experience map users were initially skeptical of digital soil mapping methodologies, but after some introduction displayed enthusiastic acceptance. A similar example of operational digital soil mapping in the wilderness areas of Washington State by Rodgers (2000), Briggs (2004), and Frazier et al. (2009) further demonstrate the success of geographic information technologies for predicting soil series.

References

Bell, J.C., Cunningham, R.L., and Havens, M.W., 1994. Soil drainage class probability mapping using a soil-landscape model. Soil Science Society of America Journal 58:464–470.

Bock, M., Bohner, J., Conrad, O., Kothe, R., and Ringeler A., 2008. SAGA: System for the Automated Geoscientific Analysis. Dept. of Physical Geography, Hamburg, Germany. URL http://www.saga-gis.org/en/index.html/ (last verified 19 April 2009).

Bohner, J., and Antonic, O., 2009. Land-surface parameters specific to topo-climatology, pp. 195–226. In: Hengl,T., and Reuter, H.I., (eds.), Geomorphometry: Concepts, software, applications. Elsevier, Amsterdam.

Breiman, L., 2001. Random forests. Machine Learning 45:5–32.

Briggs, C.A.D., 2004. GIS-based mapping of soil distribution in Thunder Creek watershed, North Cascades National Park, Washington. MS thesis Washington State Univ., Pullman, WA.

Frazier, B.E., Rodgers, T.M., Briggs, C.A.D., and Rupp, R.A., 2009. Remote area soil proxy modeling technique. Soil Survey Horizons 50:62–67.

Gallant, J.C., and Wilson, J.P., 2000. Primary topographic attributes, pp. 51–86. In: Wilson, J.C.,and Gallant, J.C. (eds.), Terrain Analysis: Principles and Applications. Wiley, New York.

Gallant, J.C., and Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research 39(12):1347, doi:10.1029/2002WR001426, 2003.

Hawthorne, B., 2009. Hawth's Analysis Tools for ArcGIS. http://www.spatialecology.com/htools/ (last verified 19 April 2009).

Howell, D.W., Kim, Y., Haydu-Houdeshell, C.A., Clemmer, P., Almaraz, R., and Ballmer, M., 2007. Fitting soil property spatial distribution models in the Mojave Desert for digital soil mapping, pp. 465–476. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), Digital Soil Mapping: An Introductory Perspective. Elsevier, Amsterdam.

Howell, D.W., Kim, Y.G., and Haydu-Houdeshell, C.A., 2008. Development and application of digital soil mapping within traditional soil survey: What will it grow into? pp. 43–52. In: Hartemink, A.E., McBratney, A., and Mendonca-Santos, M.L. (eds.), Digital Soil Mapping With Limited Data. Springer, Berlin.

Jenks, G.F., 1967. The data model concept of statistical mapping. The International Yearbook of Cartography 7:186–190.

Lindsay, J.B., 2005. The terrain analysis system: a tool for hydro-geomorphic applications. Hydrological Processes 19:1123–1130.

Liaw, A., and Wiener, M., 2002. Classification and regression by randomForest. R News 2(3):18–22.

McKenzie, N.J., Gessler, P.E., Ryan, P.J., and O'Connell, D., 2000. The role of terrain analysis in soil mapping, pp. 245–265. In: Wilson, J.P., and Gallant, J.C. (eds.), Terrain analysis: Principles and Applications.Wiley, New York.

Michaelsen, J., 2002a. Mojave January Average Temperature. A digital spatial database (ArcInfo). U.S. Geological Survey. URL http://www.mojavedata.gov/ (last verified 19 July 2009).

Michaelsen, J., 2002b. Mojave July Average Temperature. A digital spatial database (ArcInfo). U.S. Geological Survey. URL http://www.mojavedata.gov/ (last verified 19 July 2009).

Michaelsen, J., 2002c. Mojave Summer Precipitation. A digital spatial database (ArcInfo). U.S. Geological Survey. URL http://www.mojavedata.gov/ (last verified 19 July 2009).

Michaelsen, J., 2002d. Mojave Winter Precipitation. A digital spatial database (ArcInfo). U.S. Geological Survey. URL http://www.mojavedata.gov/ (last verified 19 July 2009).

Mojave Desert Ecosystem Program (MDEP)., 2000. Geomorphic landforms and surface composition of the California Mojave Desert. http://www.mojavedata.gov/ (last verified 19 July 2009).

Nield, S.J., Boettinger, J.L., and Ramsey, R.D., 2007. Digitally mapping gypsic and natric soil areas using Landsat ETM data. Soil Science Society of America Journal 71:245–252.

Peterson, F.F., 1981. Landforms of the basin and range province: Defined for soil survey. Nevada Agricultural Experiment Station Technical Bulletin 28, University of Nevada – Reno, NV. 52 p.

Quinn, P., Beven, K., Chevallier, P., and Planchon, O., 1991. The prediction of hillslope flow paths for distributed hydrological modeling using digital elevation terrain models. Hydrological Processes 5:59–79.

R Development Core Team. (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/ (last verified 19 April 2009).

Rodgers, T.M., 2000. Modeling soils of the Sawtooth and Pasayten wilderness areas with a GIS. MS thesis. Washington State Univ., Pullman, WA.

Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., and Broderson, W.D. (eds.), 2002. Field Book for Describing and Sampling Soils, Version 2.0. Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE.

Skidmore, A.K., Watford, F., Luckananurug, P., and Ryan, P.J., 1996. An operational GIS expert system for mapping forest soils. Photogrammatic Engineering and Remote Sensing 62:501–511.

Soil Survey Staff., 2006. Keys to Soil Taxonomy, USDA 10th ed. Natural Resources Conservation Service, Washington, DC.

Thompson, J.A., and Kolka, R.K., 2005. Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modeling. Soil Science Society of America Journal 69:1086–1093.

U.S. Department of Agriculture, Natural Resources Conservation Service., 2008. National Soil Survey Handbook, title 430-VI. [Online] Available: http://soils.usda.gov/technical/handbook/

Zhu, A.X., Hudson, B., Burt, J., Lubich, K., and Simonson, D., 2001. Soil mapping using GIS, expert knowledge, and fuzzy logic. Soil Science Society of America Journal 65:1463–1472.

Ward, J.H., 1963. Hierarchical groupings to optimize an objective function. Journal of the American Statistical Association 58:236–244.

Wilding, L.P., 1994. Factors of soil formation: Contributions to pedology, pp. 15–30. In: Amundson, R., Harden, J., and Singer, M. (eds.), Factors of Soil Formation: A Fiftieth Anniversary Perspective. SSSA Special Publication 33, Soil Sci. Soc. of Am., Madison, WI.

Wilson, J.P., and Gallant, J.C., 2000. Secondary topographic attributes, pp. 87–131. In: Wilson, J.P., and Gallant, J.C. (eds.), Terrain Analysis: Principles and Applications. Wiley, New York.

Wood, J.D., 1996. The Geomorphological Characterization of Digital Elevation Models. Ph.D. thesis, University of Leicester.

Wood, J.D., 2008. Lanserf version 2.2. URL http://www.soi.city.ac.uk/∼jwo/landserf/ (last verified 19 July 2009).

Chapter 30
Applying the Optimum Index Factor to Multiple Data Types in Soil Survey

S. Kienast-Brown and J.L. Boettinger

Abstract Digital soil mapping requires simple, straight-forward methods that can be easily implemented into daily activities of soil survey. The Optimum Index Factor (OIF) was developed by Chavez et al. (1982, 1984) as a method for determining the three-band combination that maximizes the variability in a particular multispectral scene. The OIF is based on the amount of total variance and correlation within and between all possible band combinations in the dataset. Although the OIF method was developed for Landsat TM data, the concept and methodology are applicable to any multilayer dataset. We used the OIF method in a subset area of the initial soil survey of the Duchesne Area, Utah, USA, to help determine which combination of data layers would be most useful for modeling soil distribution. Unique multiband images created from layers of multiple data types (elevation and remote sensing derivatives) were evaluated using the OIF method to determine which data layers would maximize the biophysical variability in the study area. A multiband image was created from the optimum combinations of data layers and used for classification and modeling in ERDAS Imagine. The output from the classification and modeling are being evaluated as pre-maps for soil mapping activities in the study area.

Keywords Correlation · Remote sensing · Digital elevation model · Pre-mapping · Unsupervised classification

30.1 Introduction

Over the past decade, digital soil mapping techniques have been incorporated into soil survey activities and used in combination with traditional soil survey methods. As technology, data availability, and knowledge of digital soil mapping techniques continually improve, digital soil mapping is more prevalent in the daily activities of

S. Kienast-Brown (✉)
USDA Natural Resources Conservation Service, Department of Plants, Soils, and Climate, Utah State University, 4820 Old Main Hill, Logan, UT 84322-4820, USA
e-mail: suzann.kienast@ut.usda.gov

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_30,

production soil survey where a soil map must be produced and delivered in a relatively short period of time. Digital soil mapping has become operational in several NRCS soil survey offices, and several NRCS soil survey projects have implemented digital soil mapping to complete soil surveys (see also Chapter 28).

Cole and Boettinger (2007) developed a raster-based classification methodology for mapping soils in the Powder River Basin, Wyoming, USA. This methodology applied Jenny's (1941) model for soil development to digital data layers for classification and modeling of soil map unit distribution. This methodology was refined by Saunders and Boettinger (2007) in the Green River Basin of Wyoming, USA, with the addition of classification and regression trees to predict soil map unit distribution. Kienast-Brown and Boettinger (2007) used Landsat classification to refine wet and saline soil map units along the eastern margin of the Great Salt Lake, Utah, USA, as an update to existing soil surveys. Howell et al. (2007) used digital soil mapping to model soil genetic features to increase understanding of soil-landscape relationships and guide field data collection in the Mojave Desert, USA, soil survey area. Howell et al. (2008) refined this work by implementing higher spatial resolution data for predicting soil genetic features, and found that the higher resolution data did not significantly improve the understanding of soil-landscape relationships in the Mojave Desert, USA, soil survey area.

Derivatives of remotely-sensed spectral data and digital elevation models are widely used as environmental covariates in modeling soil-landscape relationships (McBratney et al., 2003). Landsat spectral band ratios 3/2, 3/7, and 5/7 have been interpreted to enhance carbonate radicals, ferrous iron, and hydroxyl radicals, respectively, in surface soil and geologic materials (Amen and Blaszczynski, 2001). Cole and Boettinger (2007) and Saunders and Boettinger (2007) combined these soil enhancement ratios with elevation-derived data to predict soil map unit distribution in the Powder River Basin and Green River Basin of Wyoming, respectively. The normalized difference ratio of Landsat bands 5 and 2 can be diagnostic for calcareous sedimentary rocks: $(5-2)/(5+2)$, and is useful for distinguishing sedimentary from igneous parent material (Boettinger et al., 2008).

Elevation derivatives commonly considered for modeling soil-landscape relationships are slope steepness, slope curvature, and terrain ruggedness index, which measures average elevation change between any point on a grid and its surrounding area (Riley et al., 1999). Howell et al. (2008) considered these variables for modeling soil-landscape relationships in the Mojave Desert, USA, soil survey area. Compound topographic index (Beven and Kirby, 1979) primarily reflects hydrologic accumulation processes and is commonly used in soil-landscape modeling (McBratney et al., 2003).

The optimum index factor (OIF) is a method for determining the three band combination that maximizes the variability in a particular scene and was developed by Chavez et al. (1982, 1984). The optimum three band combination is useful for visualization purposes, but may also be useful in analysis. An OIF value is calculated for each of the 20 possible three-band combinations that can be created from the six bands of Landsat TM data (excluding the thermal band). Chavez et al. (1982) indicates that although the OIF method was developed for Landsat TM data, it is

applicable to any multispectral remote sensing dataset or geophysical/geochemical dataset. OIF is based on the amount of total variance and correlation within and between the possible band combinations. The three band combinations are ranked, and those with the largest OIF values contain the most information (measured by variance) with the least amount of duplication (measured by correlation) and are the best choices for maximizing the variability in a particular scene (Jensen, 2005). Nield et al. (2007) used the OIF method to select Landsat TM band combinations of 1, 5, 7 for visually analyzing areas with gypsic soils, and 4, 5, 7 for areas with natric soils in the Emery County, Utah, USA, soil survey area. This analysis was used to refine distribution of soil map units containing gypsic or natric soils.

Successful implementation of digital soil mapping requires simple, straightforward methods that can be easily implemented into daily activities of soil mapping. Howell et al. (2008) stated "Tools for analyzing soil-landscape relationships need to be developed for easy application by field soil scientists using standard soil survey office software." Pre-mapping an area before beginning the field work is commonly completed using digital soil mapping techniques. Pre-mapping involves selecting digital data layers that represent environmental covariates, and using the layers to stratify the landscape and aid in soil map unit development.

We evaluated the OIF method as a simple, straight-forward method for determining which data layers, derived from elevation data and remote-sensing images, would best represent the full range of biophysical characteristics in a diverse study area with semi-desert to high-mountain climate conditions and broad alluvial and glacial landforms to steep mountain landforms. The optimum data layers were combined into a multiband image used for classification and modeling, and ultimately to create a pre-map for the study area.

30.2 Materials and Methods

30.2.1 Study Area

The study area is located in north-eastern Utah, USA, in Duchesne County on the south side of the Uinta mountain range (Fig. 30.1). The study area is approximately 44,515 ha and ranges in elevation from 1,635 to 3,150 m from south to north. The area receives approximately 230 mm of precipitation at the lower elevations and up to 710 mm at the higher elevations. The vegetation ranges from semi-desert shrub land in the southern portion to high-mountain alpine forest in the northern portion.

The geology is a mix of Quaternary alluvial, colluvial, and glacial deposits and Tertiary sedimentary deposits. The alluvial and colluvial deposits are dominantly found on alluvial fans and terraces, and the glacial deposits are dominantly found on broad gently sloping outwash terraces, moraines, and steep escarpments. The Tertiary sedimentary deposits are dominantly found on hill slopes and steep mountain slopes. The mountains are located in the northern portion of the study area and give way to alluvial fans, glacial outwash terraces, and moraines at their base to the

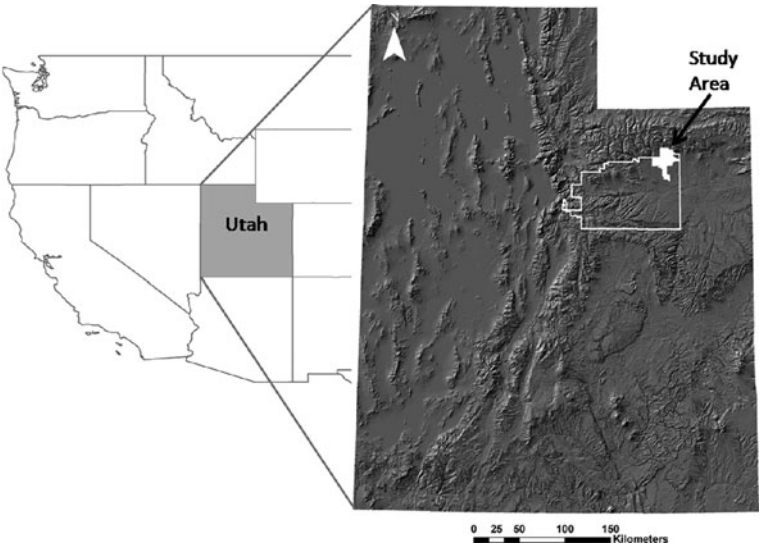


Fig. 30.1 Study area location in Duchesne County Soil Survey Area, Utah, USA

south. Glacial and alluvial landforms intermix from the base of the mountains and extend to the southern boundary of the study area (Fig. 30.2).

30.2.2 Digital Data

Elevation and remote-sensing derivatives were created and combined into multiband images, and input to the OIF calculation. Elevation derivatives were calculated from 10 m National Elevation Dataset (NED) data that had been clipped to the study area. Pre-processing to create a hydrologically correct DEM was completed in ESRI ArcGIS 9.2 ArcInfo before calculating the following derivatives: profile curvature, plan curvature, percent slope, solar radiation, terrain ruggedness index, and compound topographic index. This set of elevation derivatives was chosen as a standard set of variables that are easily calculated in ArcGIS and widely used in digital soil mapping. Table 30.1 shows details on the calculation method, covariate represented, and value range for each elevation derivative.

Landsat 7 imagery from Path 37, Row 32 June 6, 2006 was used for all remote-sensing derivatives. All image processing was completed in ERDAS Imagine 9.2. The Landsat image was standardized using the COST atmospheric correction method (Chavez, 1996) and then resampled to 10 m spatial resolution for use with the elevation derivatives using bilinear interpolation resample method. The resulting 10 m resolution image was then subset to the study area.

A series of normalized band ratios targeting mineralogical properties of surface materials were calculated from the processed Landsat imagery (see Sections 15.2.2 and 29.2.2). All normalized ratios were calculated using this basic formula:

Fig. 30.2 Geology and 10 m hillshade of the study area showing major geologic and landform patterns. Qg = Quaternary surficial glacial deposits; Qa = Quaternary surficial alluvium and colluvium; Qao = Quaternary surficial older alluvium and colluvium; T3 = Tertiary Duchesne River, Uinta, Bridger, Crazy Hollow and other formations; T4 = Tertiary Salt Lake Formation and other valley-filling alluvial, lacustrine, and volcanic units

Normalized Difference Ratio = (Band A − Band B)/(Band A + Band B). The Normalized Difference Vegetation Index (NDVI) was also calculated. Table 30.1 shows details on specific band ratios used, calculation method, covariate represented, and value range for each remote-sensing derivative.

Prior to calculating OIF, the elevation derivative layers and the remote-sensing derivative layers were stacked into two separate six layer multiband images using the Layer Stack utility in ERDAS Imagine 9.2. OIF can be calculated on an image with any number of layers, but there were six elevation derivatives and six remote-sensing

Table 30.1 Layer name, data source, calculation method, covariate represented, and value range for each data layer considered for the OIF calculations

Layer	Data Source	Calculation Method	Covariate	Minimum Value	Maximum Value
Profile Curvature	DEM	ESRI ArcGIS 9.2 ArcToolbox	relief	−7.53	9.55
Plan Curvature	DEM	ESRI ArcGIS 9.2 ArcToolbox	relief	−4.66	8.4
Percent Slope	DEM	ESRI ArcGIS 9.2 ArcToolbox	relief	0	214
Solar Radiation (WH/m^2)	DEM	ESRI ArcGIS 9.2 ArcToolbox	relief, climate	1538	5596
Terrain Ruggedness Index	DEM	ESRI ArcGIS 9.2 ArcInfo; AML script[a]	relief, parent material, time	0	30
Compound Topographic Index	DEM	ESRI ArcGIS 9.2 ArcInfo; AML script[b]	relief	2.31	24.2
Normalized Band 3/Band 2	Landsat	ERDAS Imagine 9.2 Model Maker	parent material (carbonate radicals)	−0.4	1
Normalized Band 3/Band 7	Landsat	ERDAS Imagine 9.2 Model Maker	parent material (ferrous iron)	−0.67	1
Normalized Band 5/Band 2	Landsat	ERDAS Imagine 9.2 Model Maker	parent material (calcareous sedimentary rocks)	−1	0.78
Normalized Band 5/Band 4	Landsat	ERDAS Imagine 9.2 Model Maker	parent material (ferrous minerals)	−1	1
Normalized Band 5/Band 7	Landsat	ERDAS Imagine 9.2 Model Maker	parent material (hydroxyl radicals)	−1	0.55
NDVI	Landsat	ERDAS Imagine 9.2 Spectral Indices Utility	vegetation	−1	0.91

[a](Evans, 2004b; Riley et al., 1999)
[b](Evans, 2004a; Beven and Kirby, 1979)

derivates we wished to consider for the study area. The six-band elevation derivative image consisted of the following layers: profile curvature (layer 1), plan curvature (layer 2), percent slope (layer 3), solar radiation (layer 4), terrain ruggedness index (layer 5), and compound topographic index (layer 6). The six-band remote-sensing derivative image consisted of the following layers: band 3/band 2 (layer 1), band 3/band 7 (layer2), band 5/band 2 (layer 3), band 5/band 4 (layer 4), band 5/band 7 (layer 5), and NDVI (layer 6).

30.2.3 Optimum Index Factor (OIF) Calculation

OIF is based on the amount of total variance and correlation within and between the possible three-band combinations in an image. The elevation derivative image and the remote-sensing derivative image each contain six layers; therefore, twenty possible three-band combinations were evaluated for each image using OIF. Variance was measured by the standard deviation of each layer, and correlation was measured by the correlation matrix for each image.

Once the two multiband images were created, the following steps were completed for each image: The correlation matrix was calculated for the image using ERDAS Imagine 9.2 Model Maker. The correlation coefficients from the correlation matrix and the standard deviation values for each layer were imported into a spreadsheet where the OIF values were calculated for the twenty possible three-band combinations. The following equation was used to calculate OIF:

$$\mathrm{OIF} = \frac{\sum_{k=1}^{3} s_k}{\sum_{j=1}^{3} \mathrm{Abs}(r_j)}$$

where s_k is the standard deviation for band k, and r_j is the absolute value of the correlation coefficient between any two of the three bands being evaluated (Jensen, 2005). The OIF values were then ranked to show which three-band combination had the highest OIF value, and therefore, should contain the most information with the least amount of duplication.

30.2.4 Classification

After the optimum three-band combination for the elevation derivative image and the remote-sensing derivative image were calculated using OIF, a six layer image was created from the layers in the top ranked three-band combination for each image. Before creating the multiband image, the solar radiation and three remote-sensing derivative data layers were transformed by either dividing or multiplying the layer by a constant value (i.e., 10 or 100) to bring the data ranges for the layers closer together. This was done in anticipation of normalizing the values over a common data range before the classification step, so each layer has equal influence

in the classification. After transforming the necessary layers, the six layers were stacked into a multiband image. The image was then normalized over the common data range of minimum value = 1 and maximum value = 40 using the Rescale utility in ERDAS Imagine 9.2. No data ranges were compressed in this process, only stretched to fit the range of 1–40. The resulting normalized six layer image was used as input for unsupervised classification.

Unsupervised classification was completed in ERDAS Imagine 9.2 using the ISODATA algorithm. Based on variability of the study area determined from visual interpretation of the digital geology map and the 2004 Southwest Regional GAP land cover data (2008), seven to ten final classes were expected. Therefore, a starting number of 15 classes was chosen for the unsupervised classification based on a standard rule to start with twice the number of desired classes. Transformed divergence and spectral signatures were used to evaluate separability between classes. Classes that had low separability were combined or eliminated to create the final set of class signatures. The final class signatures were used for supervised classification of the study area using the minimum distance to means classification algorithm. Clump and eliminate processes were applied to the final classified image to smooth the classification for pre-mapping activities. Patterns identified between classes in the final classification were used as a guide for digitizing pre-map units in the study area. All pre-map unit polygons were attributed with geology, landform, and dominant vegetation. A digital geology map and the 2004 Southwest Regional GAP land cover data (2008) (see Section 6.2.2) were used to assign geology and dominant vegetation to the pre-map units. Landform was assigned by the analyst by evaluating a classified slope map and 1 m spatial resolution color aerial photography.

30.3 Results and Discussion

30.3.1 Optimum Index Factor (OIF) Calculation

Table 30.2 contains the correlation matrix, standard deviation values, and top six ranked OIF values and three-band combinations for both the elevation derivative and remote-sensing derivative images. The top ranked band combination for the elevation derivative image was layer 1 (profile curvature), layer 4 (solar radiation), and layer 5 (terrain ruggedness index). The second ranked band combination contained profile curvature and solar radiation, but percent slope instead of terrain ruggedness index. The top two band combinations were very similar; especially considering terrain ruggedness index is an expression of slope. The top ranked band combination for the elevation derivative image met our expectations, containing solar radiation, a curvature variable, and either slope or terrain ruggedness index. This set of elevation derivatives are often used to represent environmental covariates for digital soil mapping activities in landscapes of the western United States. Solar radiation seems to be an important digital layer for capturing climate variability due to aspect.

Table 30.2 Correlation matrix, standard deviation values, and top six ranked OIF values for the elevation derivative image and remote-sensing derivative image

Elevation Derivative Image Correlation Matrix

Layer	1	2	3	4	5	6
1	1.0000	−0.2777	0.0100	−0.0219	0.0133	0.1227
2	−0.2777	1.0000	0.0536	0.0121	0.0451	−0.3443
3	0.0100	0.0536	1.0000	0.0010	0.9993	−0.5404
4	−0.0219	0.0121	0.0010	1.0000	0.0003	−0.0151
5	0.0133	0.0451	0.9993	0.0003	1.0000	−0.5372
6	0.1227	−0.3443	−0.5404	−0.0151	−0.5372	1.0000
Standard Deviation	0.295	0.224	12.226	306.609	3.005	2.014

Ranked OIF Values

Band Combination	OIF
1. 1,4,5	37310.13
2. 1,3,4	29196.89
3. 2,4,5	5388.17
4. 2,3,4	4781.44
5. 1,4,6	3606.45
6. 1,2,4	1068.27

Remote-Sensing Derivative Image Correlation Matrix

Layer	1	2	3	4	5	6
1	1.0000	−0.0723	0.0352	0.4810	−0.6343	−0.5426
2	−0.0723	1.0000	−0.8940	−0.3515	−0.2265	−0.1319
3	0.0352	−0.8940	1.0000	0.0882	0.5424	0.4088
4	0.4810	−0.3515	0.0882	1.0000	−0.5667	−0.8504
5	−0.6343	−0.2265	0.5424	−0.5667	1.0000	0.8432
6	−0.5426	−0.1319	0.4088	−0.8504	0.8432	1.0000
Standard Deviation	0.036	0.075	0.089	0.147	0.062	0.18

Ranked OIF Values

Band Combination	OIF
1. 3,4,5	4.66
2. 1,2,4	4.51
3. 1,3,5	3.29
4. 1,3,6	3.09
5. 3,4,6	1.18
6. 1,5,6	0.83

The effects of aspect on soil development can be quite pronounced in the semi-arid, high-elevation landscapes of Utah.

The top ranked band combination for the remote-sensing derivative image was layer 3 (band 5/band2, calcareous sedimentary rocks), layer 4 (band 5/band 4, ferrous materials), and layer 5 (band 5/band 7, hydroxyl radicals). The second ranked band combination was layer 1 (band 3/band2, carbonate radicals), layer 2 (band 3/band 7, ferrous iron), and layer 4 (band 5/band 4, ferrous materials). The top two ranked band combinations indicated sediments containing ferrous and calcareous or carbonate minerals are common in the study area which was expected due to the mineralogy of the sedimentary deposits. Hydroxyls also appeared to be an important variable, and can indicate the presence of clay minerals. This characteristic could be linked to the glacial deposits in the study area. NDVI was not included in the optimum band combinations. Variables related to mineralogy dominated the optimum band combinations, showing the influence of the parent material covariate in the study area. From this, we conclude parent material is influencing variability in the study area to a greater degree than vegetation. The Landsat middle infrared band 5 was present in all three bands of the optimum band combination and appears to be very important in characterizing parent material in the study area. This is consistent with findings in other studies in semi-arid landscapes (Boettinger et al., 2008).

30.3.2 Classification

A six layer image was created from the layers in the top ranked three-band combination for the elevation derivative image and the remote-sensing derivative image for classification. The multiband image contained profile curvature (layer 1), solar radiation (layer 2), terrain ruggedness index (layer 3), normalized band 5/band 2; calcareous sedimentary rocks (layer 4), normalized band 5/band 4; ferrous minerals (layer 5), and normalized band 5/band 7; hydroxyl radicals (layer 6). The six layers included in the multiband image, the original minimum and maximum values, the transformed minimum and maximum values, the original data range, and the normalized data range are all listed in Table 30.3. The minimum and maximum values for profile curvature (layer 1) and terrain ruggedness index (layer 3) were not transformed, and the data range for solar radiation (layer 2) determined the common data range (1–40) over which to normalize all other data layers.

The classification process started with a 15-class unsupervised classification. The resulting signature set was evaluated using transformed divergence. The original fifteen classes had an average transformed divergence value of 1930 with a minimum of 1,406. After merging or eliminating classes with low transformed divergence values, the resulting 10 classes had an average transformed divergence value of 1,944 with a minimum of 1,499. The final 10 classes seemed to capture the variability in the study area to the desired degree for pre-mapping activities. The 10-class signature set was used in supervised classification with the minimum distance to means algorithm to produce a classified image of the study area. After applying

Table 30.3 The six layers included in the final multiband image used for classification, the original minimum and maximum values, the transformed minimum and maximum values, the original data range, and the normalized data range

Layer	Covariate	Minimum Value	Maximum Value	Transformed Minimum Value	Transformed Maximum Value	Range	Normalized Range
1. Profile Curvature	relief	−7.53	9.55	−7.53	9.55	17.1	40
2. Solar Radiation (WH/m^2)	relief, climate	1538	5596	15	55	40	40
3. Terrain Ruggedness Index	relief, parent material, time	0	30	0	30	30	40
4. Normalized Band 5/Band 2	parent material (calcareous sedimentary rocks)	−1	0.78	−10	7.8	17.8	40
5. Normalized Band 5/Band 4	parent material (ferrous minerals)	−1	1	−10	10	20	40
6. Normalized Band 5/Band 7	parent material (hydroxyl radicals)	−1	0.55	−10	5.5	15.5	40

clump and eliminate processes to the resulting classification, it was used as a guide for developing pre-map units based on geology, landform, and dominant vegetation.

The classification was useful for identifying major patterns on the landscape throughout the study area. Some of the variability captured by the classification was generalized in the pre-map units because the intent of the pre-map units was to capture major changes on the landscape. However, the full variability of the classification will be useful for refining final map units and determining map unit composition during the actual field work. The final pre-map and classification for the study area are shown in Fig. 30.3. The pre-map will be used by the Bureau of Indian Affairs (BIA) for range studies during the 2008 field season. During the 2009 field season, NRCS soil scientists will use the pre-map to develop a sampling

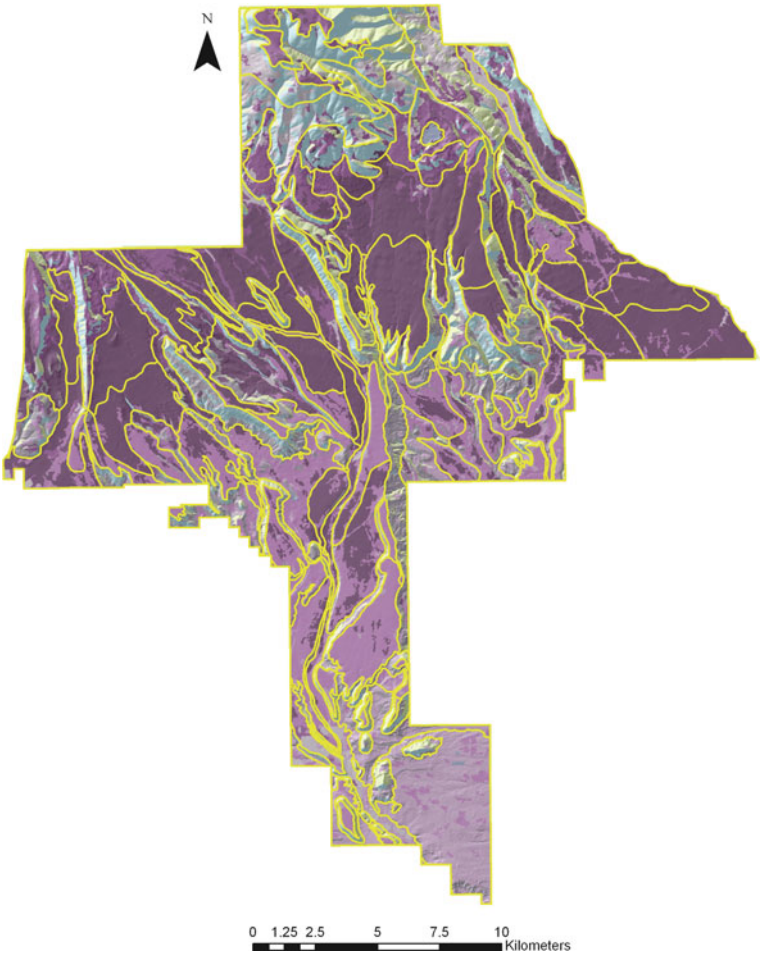


Fig. 30.3 Final classification (*background*) and pre-map units for the study area (*yellow polygons*)

plan and refine final map units for the study area. The resulting mapping will be published as part of the Duchesne Area, Utah Soil Survey.

30.4 Conclusions

Although the OIF method was originally developed and tested on remotely-sensed spectral data, we successfully applied it to both remote-sensing and digital elevation model-derived data to determine which data layers would maximize the biophysical variability in the study area. The OIF method was useful for narrowing the many choices of data layers into a manageable set of meaningful environmental covariates to use in classification, and ultimately to create a pre-map of the study area. Using the OIF method for selecting a set of environmental covariate data layers expedited the process of stratifying the landscape and developing pre-map units. A method that maximizes the biophysical variability in a particular area based on statistical relationships between data layers is desirable for soil survey activities, and lends validity to the final map product. A pre-map was the desired result in this case study, but the OIF method for selecting data layers could be applied to many other situations that involve choosing a set of data layers that represent meaningful environmental covariates.

Successful implementation of digital soil mapping in soil survey requires simple, straight-forward methods that can be easily implemented into daily activities of soil mapping using readily available resources. The OIF method was easily incorporated into the pre-mapping process and proved to be a simple, straight-forward method for determining which data layers, derived from elevation data and remote-sensing images, would maximize the biophysical variability in the study area.

References

Amen, A., and Blaszczynski, J., 2001. Integrated Landscape Analysis, pp. 2–20. U.S. Department of the Interior, Bureau of Land Management, National Science and Technology Center, Denver, CO.

Beven, K.J., and Kirby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrological Sciences Bulletin 24(1):43–69.

Boettinger, J.L., Ramsey, R.D., Bodily, J.M., Cole, N.J., Kienast-Brown, S., Nield, S.J., Saunders, A.M., and Stum, A.K., 2008. Landsat spectral data for digital soil mapping. pp. 193–202. In: Haremink, A., McBratney, A.B., and de Lourdes Mendonça-Santos, M., (eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Chavez, P.S., 1996. Image-based atmospheric corrections—revisited and revised. Photogrammetric Engineering and Remote Sensing 62(9):1025–1036.

Chavez, P.S., Berlin, G.L., and Sowers, L.B., 1982. Statistical method for selecting Landsat MSS ratios. Journal of Applied Photographic Engineering 8(1):23–30.

Chavez, P.S., Guptill, S.C., and Bowell, J.A., 1984. Image processing techniques for Thematic Mapper data. Proceedings, ASPRS Technical Papers 2:728–742.

Cole, N.J., and Boettinger, J.L., 2007. Pedogenic understanding raster classification methodology for mapping soils, Powder River Basin, Wyoming, USA. pp. 377–388. In: Lagacherie, P.,

McBratney, A.B., and Voltz, M., (eds.), Digital Soil Mapping: An introductory perspective. Developments in Soil Science Vol. 31, Elsevier, Amsterdam.

Evans, J., 2004a. Compound Topographic Index AML. http://arcscripts.esri.com/details.asp?dbid= 11863 (last verified September 8 2008).

Evans, J., 2004b. Topographic Ruggedness Index AML. http://arcscripts.esri.com/details.asp?dbid=12435 (last verified September 8 2008).

Howell, D., Kim, Y.G., and Haydu-Houdeshell, C.A., 2008. Development and application of digital soil mapping within traditional soil survey: What will it grow into? pp. 43–52. In: Hartemink, A., McBratney, A.B., and de Lourdes Mendonça-Santos, M. (eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Howell, D., Kim, Y.G., Haydu-Houdeshell, C.A, Clemmer, P., Almaraz, R., and Ballmer, M., 2007. Fitting soil property spatial distribution models in the Mojave Desert for digital soil mapping. pp. 465–476. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), Digital Soil Mapping: An introductory perspective. Developments in Soil Science Vol. 31, Elsevier, Amsterdam.

Intermountain Region Digital Image Archive Center., 2008. Southwest Regional Gap Analysis Project. http://earth.gis.usu.edu/swgap/index.html (last verified 22 September 2008).

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

Jensen, J.R., 2005. Introductory Digital Image Processing. Prentice Hall, Upper Saddle River, NJ.

Kienast-Brown, S., and Boettinger, J.L., 2007. Land cover classification from Landsat imagery for mapping dynamic wet and saline soils. pp. 235–244. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), Digital Soil Mapping: An introductory perspective. Developments in Soil Science Vol. 31, Elsevier, Amsterdam.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

Nield, S.J., Boettinger, J.L., and Ramsey, R.D., 2007. Digitally mapping gypsic and natric soil areas using landsat ETM data. Soil Science Society of America Journal 71(1):245–252.

Riley, S.J., DeGloria, S.D., and Elliot, R., 1999. A terrain ruggedness index that quantifies topographic heterogeneity. Intermountain Journal of Sciences 5:1–4.

Saunders, A.M., and Boettinger, J.L., 2007. Incorporating classification trees into a pedogenic understanding raster classification methodology, Green River Basin, Wyoming, USA. pp. 389–400. In: Lagacherie, P., McBratney, A.B., and Voltz, M. (eds.), Digital Soil Mapping: An Introductory Perspective. Developments in Soil Science Vol. 31, Elsevier, Amsterdam.

Chapter 31
U.S. Department of Agriculture (USDA) TEUI Geospatial Toolkit: An Operational Ecosystem Inventory Application

Haans Fisk, Robert Benton, Corey Unger, Timothy King, and Sharie Williamson

Abstract The TEUI-Geospatial Toolkit (Toolkit) is an operational ecological inventory application used by the U.S. Department of Agriculture (USDA) Forest Service and other land management agencies. This resource mapping tool complements traditional inventory methods by streamlining the collection and analysis of inventory information in a digital environment. The Toolkit is based on the USDA Forest Service Terrestrial Ecological Unit Inventory Technical Guide, which complies with the National Cooperative Soil Survey (NCSS) standards. It is intended for resource specialists with a strong background in terrestrial mapping and intermediate geographic information system (GIS) skills. This Environmental Systems Research Institute, Inc. (ESRI) ArcGIS^{TM}mapping tool guides the user through the TEUI mapping process, helps stratify landscapes and analyze environmental characteristics with geospatial data. Products derived with this operational mapping application comply with corporate data standards and are stored in corporate database systems. Design, development and operational support are performed by the USDA Forest Service Remote Sensing Application Center (RSAC). RSAC provides technical assistance to field units and increases application awareness at various meetings, workshops and conferences. Staying connected with the TEUI soil mapping community establishes an essential feedback loop for gathering new ideas and enhancing application functionality. This paper provides an overview of the current application and highlights specific mapping and data analysis functionality. It also identifies specific benefits that are realized through geospatial technologies as the Toolkit is implemented in the pre-map phase of a TEUI project.

Keywords Operational ecological inventory · Resource mapping tool · Landscape stratification

H. Fisk (✉)
USDA Forest Service, Remote Sensing Applications Center, 2222 W. 2300 South, Salt Lake City, UT 84119, USA
e-mail: hfisk@fs.fed.us

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_31,

31.1 Introduction

Terrestrial Ecological Unit Inventory (TEUI) is the land survey system used by the U.S. Department of Agriculture (USDA) Forest Service for classifying and mapping ecological types. Ecological types are defined by abiotic and biotic environmental factors that incorporate combinations of climate, physiography, geology, soil and vegetation. The purpose of TEUI is to classify ecosystem types and map land areas that have similar management capabilities (Cleland et al., 1997).

The TEUI Technical Guide (Winthers et al., 2005) provides methods and procedures for inventorying lands administered by the agency. The process involves three primary activities that include classification, mapping and sampling (Fig. 31.1). Ideally, this process begins with identifying taxonomic units within the survey area (classification), then determining where those types occur spatially (mapping) and finally estimating individual components to determine the number of acres comprising ecological types (sampling). In practice, the inventory progresses in cycles, so that the inventory product becomes refined through successive iterations.

TEUI products include maps, spatial and tabular databases, map unit descriptions, ecological-type descriptions and interpretations. These data are stored and managed by the USDA Forest Service Natural Resource Information System Terrestrial Module (NRIS-Terra). This information provides basic land-unit information for land planners to assess ecosystem capabilities, determine sustainable production levels and make informed and practical management decisions. For example, it supports ecological and watershed assessments, burned-area emergency rehabilitation (BAER), range-allotment updates, forest-plan revisions and project-level planning and analysis.

Traditional TEUI or Soil Resource Inventory (SRI) methods rely heavily on interpretation of aerial photography using a stereoscope (USDA Soil Conservation Service, 1993). Although this platform has been used for decades, there are drawbacks

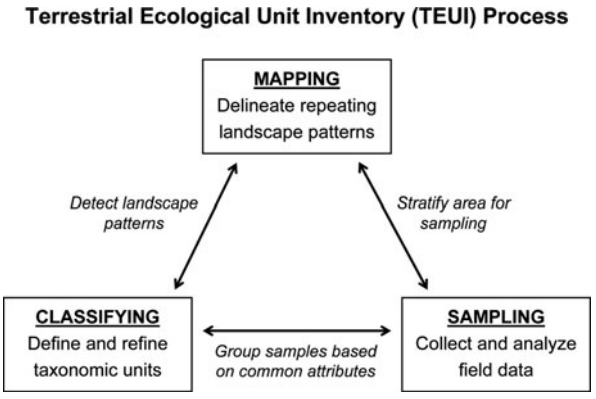


Fig. 31.1 The Terrestrial Ecological Unit Inventory process is iterative and involves three primary activities that include classification, mapping and sampling. As provided by the USFS *TEUI Technical Guide,* these elements provide the foundation for identifying taxonomic units, determining their location and estimating total acres of ecological types

when applying it across large survey areas. Assembling and preparing hundreds of photographs is time consuming and cumbersome. Characterizing land unit areas and consistently applying map unit delineation criteria may reflect the ability and bias of the individual photo interpreter which compromises the value of the TEUI products. Moreover, once the landscape delineations have been created on hard-copy photographs, additional effort is necessary to transfer line work into a GIS.

As of 2007, 55 million acres of USDA Forest Service land lacked modern TEUI or SRI. An additional 18 million acres, although mapped, did not meet standards of the National Cooperative Soil Survey (NCSS). Given the high cost of traditional TEUI surveys (US$2 to US$3 an acre); the USDA Forest Service is leveraging new technology to complete ecological-unit inventory faster and more economically (Fallon et al., 1994; Lane and Fisk, 2002). Remote sensing, geospatial technologies and raw computing power have dramatically improved over the last few years affecting our ability to visualize entire landscapes and stratify repeating patterns more consistently and efficiently. In searching for less expensive ways to conduct TEUI, the USDA Forest Service designed and developed the TEUI-Geospatial Toolkit application to streamline the mapping process and provide resource management a cost-effective alternative to traditional methods.

31.2 TEUI-Geospatial Toolkit

The TEUI-Geospatial Toolkit (Toolkit) implements the mapping standards prescribed in the TEUI Technical Guide and directly supports field units. The application is designed for resource specialists with a strong background in terrestrial mapping and intermediate GIS skills. The results of research and available geospatial technology are integrated into this digital mapping application as push-button solutions that simplify complex data processing procedures. Example procedures include multi-layer trait analysis (*Map Unit Statistics*), polygon attribution (*Reassign Map Unit Polygons*) and field map generation. Thus, end-users can accomplish their tasks more efficiently and increase overall workforce productivity. Built as an ESRI ArcMap extension, the Toolkit gives non-technical resource specialists the ability to access geospatial data, visualize landscapes, design and delineate ecological map units, validate map criteria and generate field maps (Fig. 31.2). Resulting output products comply with USDA Forest Service GIS Data Dictionary standards and can be transferred to NRIS-Terra.

31.2.1 Geospatial Data Acquisition

The Toolkit provides a solution for creating consistent and continuous geospatial data layers and delivering standard products to resource specialists.[1] The standard

[1] It is not required that one request data via the data provisioning system in order to take advantage of the Toolkit application.

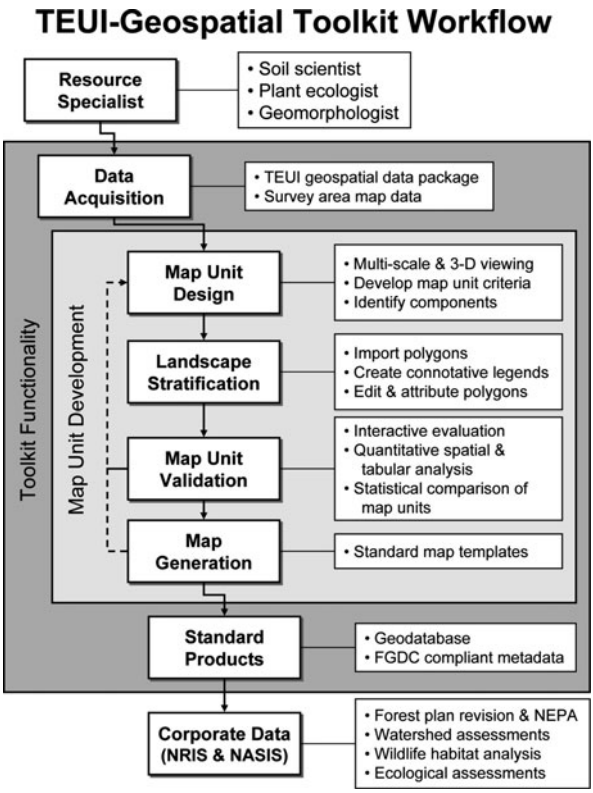


Fig. 31.2 The TEUI-Geospatial Toolkit implements the mapping protocols provided by the USFS *TEUI Technical Guide*. The integrated mapping activities consist of data acquisition, map unit design, landscape stratification, map unit validation, map generation, and standard products. This process uses iterative workflows to provide and refine standard products that can be retained in enterprise corporate databases

TEUI geospatial data package (TEUI GDP) contains 32 raster (pixel-based) and 16 vector (point, line and polygon) layers, which provide a foundation for conducting TEUI in a digital environment (Fig. 31.3).

Raster data include topographic, climate and spectral indices, as well as multi-spectral and multi-resolution backdrop imagery (see Chapter 2 for examples of DEM-derived attributes used to represent topography and climate in DSM applications). Vector layers include USFS cartographic feature files, standard USGS 7.5 and 3.75-min quadrangles and natural segments.[2] To access these standard products,

[2] Natural segments are custom polygon layers that divide the landscape into relatively homogeneous polygons, based on topographic products (e.g., fully-illuminated hillshade, slope, elevation) and spectral imagery (e.g., Landsat Thematic Mapper and digital orthophoto quadrangle resolution merge imagery). The production routine involves the use of Definiens Developer and a standard multi-resolution segmentation process to create seven hierarchical polygon levels. In the absence

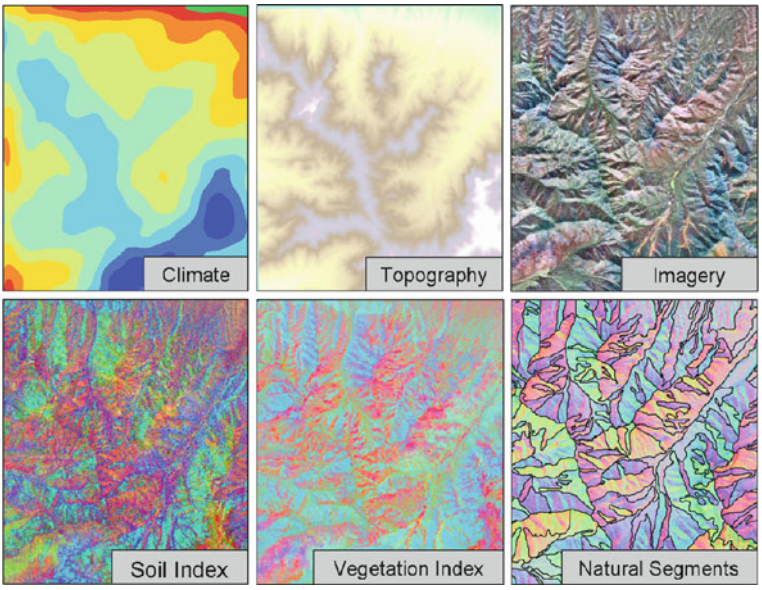


Fig. 31.3 The TEUI GDP provides a foundation for conducting resource inventory in a digital environment. Shown above are samples of analytic, vector and backdrop layers in the TEUI GDP

the requestor simply submits a survey area (polygon file) to the TEUI Data Coordinator, who in turn generates the TEUI GDP, compresses it and delivers it to the requestor.

31.2.2 Map Unit Design

Map unit design involves assessing and conceptualizing repeating landscape patterns. As defined in the TEUI Technical Guide (Winthers et al., 2005), map unit design is the "process establishing the relationship between classifications and the products depicting them." It further states that classification is the "grouping of similar types according to criteria considered significant for this purpose. The rules for classification must be clarified before identifying the types within the classification standard. The classification methods should be clear, precise, quantitative (where possible) and based on objective criteria so that the outcome would be the same whoever performs the definition (or description)." (See Section 27.2.4 for a discussion of attributes used to represent ecological classes. See also Section 24.4 for a discussion of an analogous concept (soilscapes)).

---

of existing line work (such as no previous soil survey for the area), the natural segments are used as a starter set of line work. Observation suggests that it is easier to modify this line work than it is to create new line work from scratch.

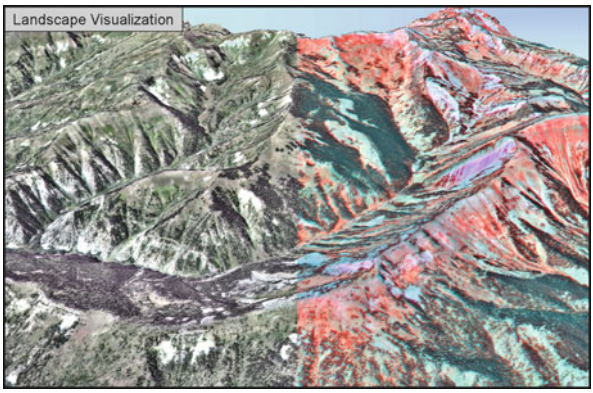


Fig. 31.4 The Toolkit uses ArcScene interactive visualization capabilities to display landscapes using a variety of data at multiple scales (a blended ETMDOQ merge. These capabilities enable resource specialists to see landscape patterns and develop mapping criteria. Shown above is a 3-m ETMDOQ merge draped over a 10-m DEM surface: *Left* – natural color composite (RGB – band 3, band 2, band 1); *Right* – false color infrared composite (RGB – band 4, band 5, band 3)

The Toolkit offers resource scientists three interactive capabilities to develop and capture preliminary map unit concepts: viewing landscapes at multiple scales, viewing landscapes in 3-D and defining map unit properties. Viewing the landscape at multiple scales and in 3-D allows users to develop mapping criteria, identify individual components and establish a strategy for delineating ecological resources (Fig. 31.4). Once map unit criteria are identified, map unit characteristics are documented using the *Define Map Unit* utility. This interface stores and organizes key information such as map unit symbol, map unit name, as well as lengthier map unit descriptions.

31.2.3 Landscape Stratification

Landscape stratification is the process of dividing a survey area into repeating ecological landscape units with similar abiotic and biotic soil forming properties. The Toolkit provides two ways to accelerate and enhance the stratification process: (1) digitizing polygons directly into GIS and (2) attributing those polygons with appropriate map unit symbols. Polygons that segment the survey area can be generated through basic heads-up digitizing with the aid of backdrop imagery which provides spatial and contextual reference. The resulting polygons are an expression of the map unit design (classification scheme). Also, map unit polygons can be imported from existing data layers using the *Import Map Unit Polygons* utility and refined using standard ArcMap^{TM}editing utilities to reflect map unit concepts. Using this method, polygon attributes are documented manually. Alternatively, the *Connotative Legend* tool may be used. This tool uses polygons as zones for calculating statistics (zonal majority) for selected layers (e.g., slope, aspect, elevation). The

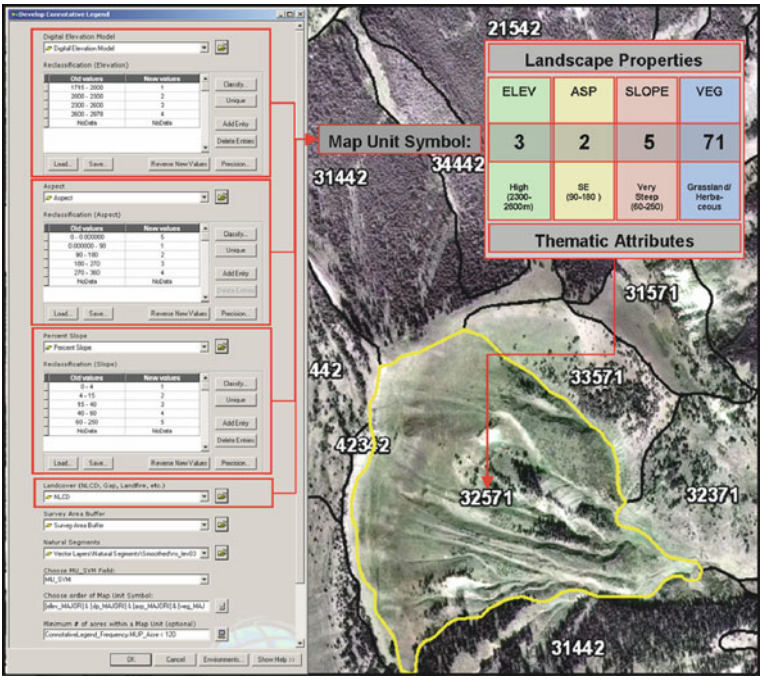


Fig. 31.5 The connotative legend feature helps organize landscape properties and applies user-defined criteria to predict spatial landscape patterns. The backdrop is ETMDOQ merge imagery (RGB – band 3, band 2, band 1)

concatenated statistics are used to label each polygon (Fig. 31.5). This compact label is a symbol that repeats over the landscape giving the resource specialist an appreciation for the representation of the specific map unit across the study area.

31.2.4 Map Unit Validation

Map unit validation is the process of evaluating attributed polygons and assessing the homogeneity of the characterization criteria. This activity is part of a process in which (a) outliers in the classification are identified; (b) adjustments to the classification scheme or landscape stratification are implemented; and (c) the effects of the adjustments are observed and reevaluated. The resource specialist iterates through this process until s/he is satisfied that all polygons adhere to the classification scheme.

The Toolkit provides three features to help evaluate map units: computing tabular statistics, analyzing unit properties and comparing map units. Together, these features provide a quantitative assessment of how closely the landscape delineation corresponds to ideal concept underpinning the classification scheme.

Unique to the Toolkit are its analytical charting utilities that allow users to generate, display and summarize a variety of statistical measures used to delineate

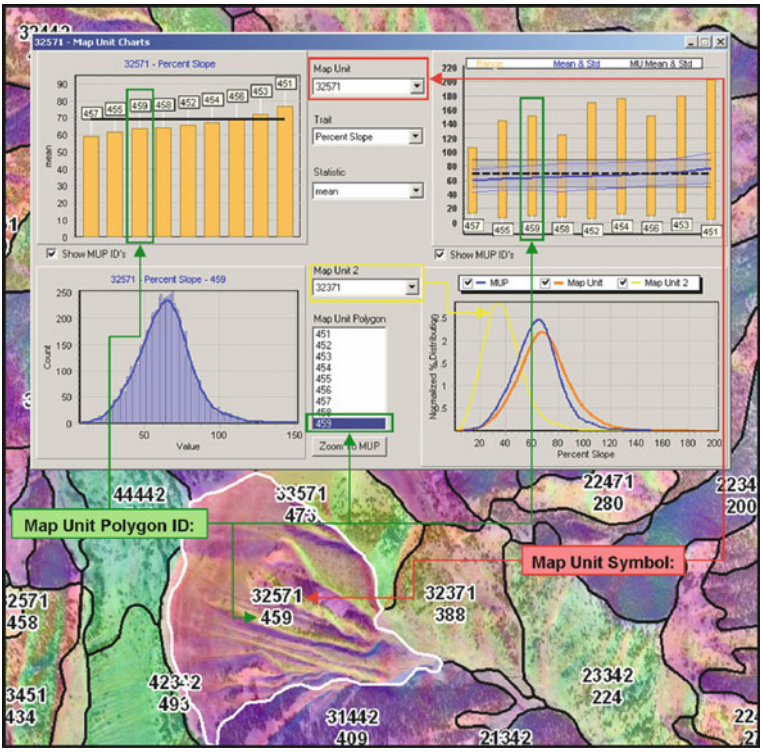


Fig. 31.6 The analytical charting tools help validate delineation and characterization criteria as well as assess overall map consistency. In the background above, preliminary map units are overlaid on a blended ETMDOQ merge and fully illuminated hillshade. In the foreground, the Map Unit Chart shows: (in the *upper left corner*) that map unit 32571 is comprised of nine polygons, 451 through 459; (in the *lower left*) that one of those polygons, 459, has a Gaussian distribution of values for percent slope; (in the *upper right*) that the range of values and the mean value for percent slope in polygon 459 are consistent with values for other polygons in the map unit; and (in the *lower right*) that the distribution of values for this map unit differs as compared to the distribution of another map unit, in this case 32371

terrestrial ecological units (Fig. 31.6). These interactive charts provide a new way to integrate environmental raster data into the validation process and better understand the relationships between spatial and tabular data. Overall these utilities help specialists consistently stratify landscapes, quantify landscape properties and improve the resulting map products.

31.2.5 Map Generation

Hard-copy field maps are important tools for collecting field documentation and validating map unit delineation. The Toolkit streamlines the process for generating field maps and provides four standard map templates that simplify and increase the

efficiency of map production. Map templates provide the capability to create maps at scales of 1:9,000, 1:12,000, 1:24,000 and a variably scaled map that includes the extent of the study area.

Map layers are automatically symbolized, but can be updated to reflect project specific conditions. Users can also specify an assortment of backdrop imagery and vector layers, as well as include map unit boundaries and representative field sample locations. These maps can be printed as hardcopy or exported to other electronic formats such as PDF. The map template technology is based on Map Books, an ESRI application.

31.3 Discussion

Since the TEUI-Geospatial Toolkit application is founded on a GIS, a number of advantages are available immediately to the resource specialist. First, the GIS can enforce standards assuring the needs of both the project and the organization are met. Second, attribution is assured. In this way, the value of the corporate database will increase with each project undertaken. Third, data integrity rules are explicit within the geodatabase so that topology and attribution rules are reinforced. This greatly increases the utility of the data to researchers and other third-party users of the products produced from the project.

The Toolkit provides a convenient link to ESRI's ArcScene™ application[3] that supports and enhances initial reconnaissance efforts of pre-mapping. The *Navigate* and *Flyover* tools provide insight regarding the logistics facing the field teams. By draping assorted backdrop imagery (e.g., multispectral Landsat ETM or high-resolution NAIP) over a digital elevation model (DEM), the resource specialist can have a virtual tour of those areas that may present hazards or challenges to a ground crew. If these areas must be visited, adequate provision can be made to support the team and ensure its safety. Alternatively, it may be that landscape delineation revealed that such areas are well represented in areas that present fewer challenges to access. Under such a situation, the priority to visit difficult terrain can be lowered and perhaps avoided altogether.

In the pre-mapping phase of the project, it is expected that the soil scientist will define some preliminary delineation. S/he has a number of options for creating the initial set of polygons – a) digitize polygons from scratch, b) import existing line work, or c) make use of the generated natural segments. Without regard to the method chosen, one of the benefits of using a GIS is immediately apparent. Landscape delineations can be visually inspected and assessed right at the workstation. For example, by draping the preliminary polygons over high resolution backdrop imagery, the soil scientist can observe whether image patterns align with soil delineations. A great degree of alignment suggests that the classification scheme

[3] ArcScene is licensed separately from ArcMap. In order to access ArcScene functionality, a license must be secured.

used to delineate the landscape is also expressed by other measures that serve as proxies for soils. The degree to which there is disagreement or conflict between the preliminary polygons and other soil proxies can cause the resource specialist to reflect on the tentative classification scheme and consider changes that may bring about better alignment. This refinement can be completed in a GIS environment in a fraction of the time it would take to replicate the process using traditional aerial photography.

Finally, the prospect exists to develop more than one classification scheme and associated delineations during the part of the season that is not suitable for field work. When the field conditions improve, early field samples can quickly reveal which of the classifications is likely to best represent the ground conditions. Should sampling requirements differ significantly between the schemes, field teams could be deployed with greater efficiency resulting in a savings in both salary and travel expense.

In a typical TEUI project, the methods and techniques described above are sequenced to define a workflow. The most obvious characteristic of the workflow is the iterative or cyclic pattern in which the methods are used for various phases of the project. For example, the sequence of *Map Unit Design* → *Landscape Stratification* → *Map Unit Validation* applies equally well whether in the pre-mapping phase or in the data reduction phase following a field season. What needs to be recognized is that nothing material has changed regarding the workflow as compared to traditional surveys that do not use the Toolkit. The application simply brings technology to bear on certain activities within the existing conceptual workflow. In addition to improving the efficiency and effectiveness of the resource specialist, there is an increased likelihood that the resource management organizational structure will support a transition to automated methods.

Development of the TEUI-Geospatial Toolkit and technical transfer/deployment is handled by the Remote Sensing Applications Center (RSAC) which is guided by the TEUI soil mapping community. RSAC is a detached Washington Office staff unit of the USDA Forest Service located in Salt Lake City, Utah. A variety of support materials exist to promote organizational awareness including short illustrative publications, lengthier reports, posters and presentations. A user guide describes application functionality in a step-by-step fashion that is intended to support the resource specialist in the day-to-day use of the Toolkit. Finally, a USDA Forest Service internal website provides key awareness information and training materials on demand as well as a number of useful links to related websites. These support materials provide the backbone and infrastructure to service customer needs.

Institutionalizing the Toolkit involves sharing awareness materials with different audiences. RSAC increases awareness about the application to national and regional TEUI program managers by giving presentations at key meetings, workshops and conferences. The USDA Forest Service also shares the application and support materials with interagency partners, cooperates in a number of interagency activities and works directly with other national technical centers to make the application more widely available.

To prepare the technical resource specialists for the transition to automated methods, RSAC develops and provides instructor-led and self-led trainings complete with

lecture presentations, hands-on exercises and exercise data sets. RSAC conducts on-site training in regions where major projects are about to launch; delivers formal classroom training for individuals from different administrative units; and hosts web-enabled training for specialists that may not be able to schedule time away from the office. RSAC also provides the use of a web-based self-led tutorial. This tutorial provides a great alternative to the instructor-led formats for managers and specialists who need to understand the application beyond a conceptual level and want to do so at their convenience.

Now that the Toolkit is deployed to the field, RSAC maintains the application and its functionality on certified USDA Forest Service computer operating systems. RSAC uses a software maintenance philosophy called continuous integration which helps isolate and track the software defects. It allows them to effectively manage changes to the application and adapt to evolving enterprise and commercial architectures. This maintenance approach requires RSAC to document application defects (e.g., software "bugs") and track potential enhancements received through customer feedback. As the Toolkit application periodically evolves, they update user documentation and redistribute materials using the dedicated website.

In addition, RSAC coordinates with program management and directly supports end users by building geospatial data packages and providing real time support to individual projects (see Section 31.2.1). Customers can call or e-mail RSAC staff to discuss workflow strategies or resolve application issues. If necessary, RSAC may link directly to the customer's computer and troubleshoot the problem together.

To better inform the reader on how the Toolkit is being used by soil scientists, the following example is provided. In the winter of 2006 and spring of 2007, the Caribou-Targhee National Forest (NF) used the Toolkit to complete TEUI at the land type level in a previously unmapped portion of the forest. This was completed by an interdisciplinary team who collaboratively borrowed concepts from adjacent soil surveys and used the Toolkit to implement the TEUI pre-mapping process. A local soil scientist on the Caribou Targhee NF, a corporate data steward from the Intermountain Regional Office and staff from the RSAC comprised the interdisciplinary team.

In this project, the team integrated the core activities of the TEUI pre-mapping process: Map Unit Design, Landscape Stratification and Map Unit Validation. They defined map unit concepts for the study area based on fundamental soil-forming factors: climate, organisms, relief, parent material and time that originated from completed and adjacent soil surveys. Geospatial data streamlined the delineation of initial landscape stratification and labeling of polygons according to the map unit definitions. The resulting map unit polygons were visually inspected against topographic and multi-spectral imagery and evaluated more rigorously against geospatial data layers to assure consistency within the study area. The local soil scientist associated connotative legend map units (map unit names derived from the values of the trait data used to delineate the units) to adjacent soil survey map units (map units named using established conventions). These products of the TEUI pre-map were used to plan the field work and to direct collection of necessary field inventory data. The costs to develop and execute this study were about one-third the costs of traditional pre-mapping methods.

31.4 Conclusions

The USDA Forest Service needs soil or basic terrestrial resource information to practice sustainable resource management. The TEUI-Geospatial Toolkit supports resource management by integrating geospatial technology with USDA Forest Service TEUI protocols and is used by the agency and its partners as a cost-effective and credible alternative for consistently collecting natural-resource information. It bridges an important technology gap that exists for many resource specialists and streamlines their workflow by enabling them to access geospatial data, design ecological map units, delineate landscape patterns, analyze map unit properties and generate standard field maps. Products derived using the Toolkit comply with corporate information-system standards.

References

Cleland, D.T., Avers, P.E., McCabe, W.H., Jensen, M.E., Bailey, R.G., King, T., and Russell, W.E., 1997. National hierarchical framework of ecological units, pp. 181–200. In: Boyce, M.S., and Haney, A., (ed.), Ecosystem management applications for sustainable forest and wildlife resources. Yale University Press, New Haven, CT.

Fallon, D., Svalberg, T., Tart, D., Maus, P., Wirth, T., and Lachowski, H., 1994. Use of satellite imagery and digital elevation models in the Bridger-Teton integrated resource inventory, pp. 77–83. In Jerry Greer (ed.) Remote Sensing and Ecosystem Management: Proceedings of the Fifth Forest Service Remote Sensing Applications Conference; 1994 April 11, American Society for Photogrammety and Remote Sensing, Bethesda, MA.

Lane, J., and Fisk, H., 2002. Reducing non-map error in assessing map accuracy. In Jerry Dean Greer (ed.) Proceedings of the Ninth Forest Service Remote Sensing Applications Conference, The American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland.

Winthers, E., Fallon, D., Haglund, J. DeMeo, T., Nowacki, G., Tart, D., Ferwerda, M., Robertson, G., Gallegos, A., Rorick, A., Cleland, D.T., and Robbie, W., 2005. Terrestrial Ecological Unit Inventory technical guide: U.S. Department of Agriculture, Forest Service, Washington Office, Ecosystem Management Coordination Staff. Washington, DC, 245 p.

Soil Conservation Service. (1993). Soil Survey Manual. (Handbook 18). U.S. Department of Agriculture, Washington, DC.

Chapter 32
Predictive Soil Maps Based on Geomorphic Mapping, Remote Sensing, and Soil Databases in the Desert Southwest

S.N. Bacon, E.V. McDonald, G.K. Dalldorf, S.E. Baker, D.E. Sabol Jr, T.B. Minor, S.D. Bassett, S.R. MacCabe, and T.F. Bullard

Abstract We present an expert based system to rapidly predict the shallow soil attributes that control dust emissions in the arid southwest U.S. Our system's framework integrates geomorphic mapping, remote sensing, and the assignment of soil properties to geomorphic map units using a soil database within a geographic information systems (GIS) framework. This expert based system is based on soil state factor-forming model parameters that include: (1) climate data, (2) landform, (3) parent material, and (4) soil age. The four soil-forming data layers are integrated together to query the soil database. To validate the accuracy of the expert based model and resultant predictive soil map, a blind test was performed at Cadiz Valley in the Mojave Desert, California. The desert terrain in Cadiz Valley consists of alluvial fans, fan remnants, sand dunes, and playa features. The test began with three users independently mapping an area of over 335 km^2 using 1:40,000-scale base maps to rapidly create geomorphic and age class layers, and then integrating these with climate and parent material layers. The results of the four data layers were then queried in the soil data base and soil attributes assigned to map unit layers. The soil-forming model presented here is geomorphic-based, and considers soil age as a significant factor in accurately predicting soil conditions in hyper arid to mildly arid regions. This work comprises a successful first step in the development of an expert-based system to map shallow soil conditions in support of dust emission models in remote desert regions.

Keywords Predictive soil map · Terrain hazard map · Arid soils · Geomorphology · Mojave Desert

S.N. Bacon (✉)
Division of Earth and Ecosystem Sciences, Desert Research Institute, Reno, NV 89512, USA
e-mail: sbacon@dri.edu

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_32,

32.1 Introduction

Knowledge of terrain elements with high dust emission potential has numerous beneficial applications from supporting a variety of military operations in desert environments to mitigating construction impacts near urban areas in the desert southwest USA to maintaining or improving local air quality. Avoiding locations susceptible to extreme dust emissions and other terrain-related hazards requires the ability to predict soil and terrain conditions, often from limited information (Scull et al., 2005). The goal of this paper is to show one example of the application of an integrated, predictive tool for forecasting desert terrain conditions, such as those that contribute to the dust emission potential of particular landscape elements.

The technical approach of this application relies on the systematic integration of desert landform parameters in geomorphic models for predicting terrain conditions in a GIS framework (e.g., Carré and Girard, 2002; McBratney et al., 2003). Advances in earth science research have established that unique, predictable relations exist among landscape position, soils, vegetation, and geology (e.g., Birkeland, 1999). Further, new instrumentation allows the collection of a wide range of environmental information to characterize surface and subsurface conditions in arid regions, for example tension infiltrometers, ground penetration radar, portable wind tunnels, and hyperspectral and multispectral imagery (e.g., Caldwell et al., 2008; Meadows et al., 2006; Sweeney et al., 2008; see section 8.1 and section 9.2).

By integrating models and methods from geomorphology, soil science, climatology, and atmospheric science with remote sensing and other technologies, a predictive model can be developed. The intent of this predictive model is to rapidly generate the boundaries of discrete landforms and the assignment of soil attributes in order to classify the landscape into specific terrain hazards. The soil-forming model presented here is geomorphic-based, and considers soil age as a significant factor in accurately predicting soil conditions in hyper arid to mildly arid regions (see Section 4.3). Dependent on particular input criteria, a variety of terrain hazards can be predicted, ranging from dust emission potential to vehicle trafficability. The following objectives need to be executed to ensure the performance of the predictive model: (1) application of advance knowledge of surface processes, assessment of remote sensing technologies, and compilation of existing data for predictive assessment of desert terrain conditions, (2) integration of this knowledge to develop a dynamic GIS platform for predicting desert terrain conditions and potential hazards by assigning soil attributes to terrain features from a soil database, and (3) validation of the predictive model across diverse desert terrains. Implementation of this new tool represents a first attempt to rapidly describe surface soil characteristics simultaneously with landform identification, thereby creating a delivery format readily usable to predict a variety of terrain hazards.

32.2 Materials and Methods

32.2.1 Study Area

To validate the accuracy of the expert based model and integration of a soil database to generate a predictive soil map, we performed a blind test at Cadiz Valley in the Mojave Desert, California (Fig. 32.1 inset). The desert terrain in Cadiz Valley consists of Basin and Range physiography, which includes mountain highlands and pediments, composed principally of Mesozoic granitic rocks that are separated by large areas of alluvial fans, sand dunes, and playa features (Fig. 32.1). The climate of Cadiz Valley is primarily semi-arid desert with areas at the valley floor as arid desert (PRISM Group, 2007).

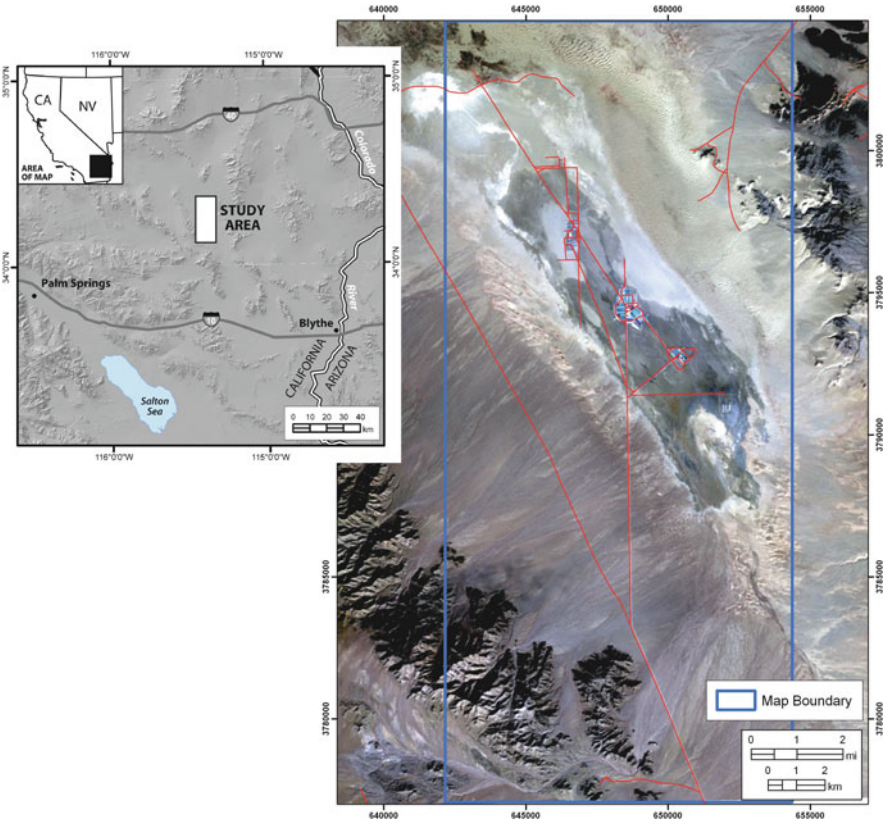


Fig. 32.1 Inset map showing the study area within Cadiz Valley, California (*white box*). LANDSAT ETM+ imagery showing the distribution of mountain highlands, alluvial fans, sand dunes, and playa features within the map area

32.2.2 Rationale and Significance of Landform-based Terrain Predictions

Existing geographic data sources (e.g., digital elevation models, satellite and aerial imagery, and geologic maps) can provide some information regarding general terrain conditions, but are insufficient for predicting dust emission or identifying other terrain hazards because they lack detailed spatial information on soil conditions. By systematically integrating the observational knowledge about the distribution, age, and geology of desert landforms and associated soil and surface conditions, geomorphic-based models provide an essential platform for predicting terrain conditions. Using these models, knowledge of how desert landscapes evolve, and of the principal surficial processes that drive surface evolution and soil formation, can provide a powerful means to assess dust emission or trafficability.

The concept of this geomorphic model simply refers to the integration of surface observations with models of landscape evolution to predict what lies below the immediate surface. In other words, years of research have demonstrated that systematic relationships exist between landscape position and soil formation processes that account for the observed distribution of soils across desert terrains. This knowledge can be essentially applied in reverse: knowledge of soil forming processes can now be applied to identify landforms to develop predictive soil maps. Prediction of critical surface and subsurface conditions can be further augmented (in terms of time and quality) by linking the inversion of geomorphic models with surface characteristics, most of which are identified from remotely sensed imagery from satellites.

32.2.3 Model for Predicting Terrain Conditions

The conceptual procedure for predicting soil and terrain conditions is based on two concepts. First, a soil evolves from a preexisting parent material (i.e., lithology of rock) into a well-defined soil closely related to discrete and identifiable landforms. This relation allows the ability to reasonably predict soil types if general knowledge of overall climate (especially precipitation), landform age and type, and soil parent material can be identified or surmised. This approach follows the soil-forming factor concept of Jenny (1941) that includes: Time, Parent Material, Topography, Climate, and Organisms. Second, most landforms can be easily identified using available aerial imagery (Jayko et al., 2005).

Application of this approach represents the idea of geomorphic model inversion. Predictions of subsurface conditions are based on the integration of conceptual models involving how the desert surface and soils evolve, with observation of existing surface characteristics (e.g., landform morphology, surface cover, microtopography). The process of generating the predictive soil attributes in Cadiz Valley was based on the following geomorphic-based model parameters and the use of a soil database (see Fig. 32.2) as follows:

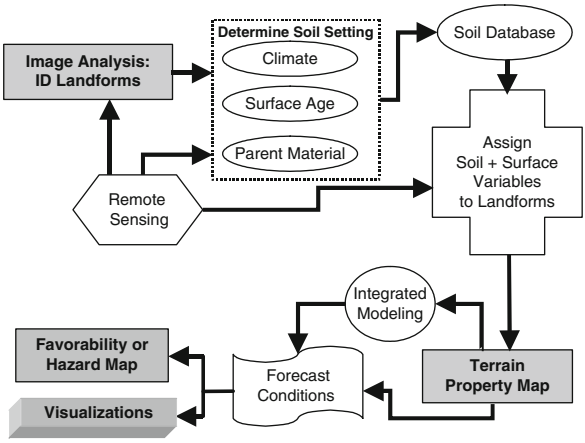


Fig. 32.2 Flowchart summarizing GIS platform designed for dynamic predictions of desert terrain conditions. Diagram shows relative position of three fundamental components or products (image analysis, soil or terrain property mapping, and output in maps or visualizations) and associated degrees of data integration and application

(1) Image Analysis: The expert-based identification of landform or geomorphic surfaces was made using 15-m resolution LANDSAT ETM+ imagery combined with a 10-m resolution digital elevation model (DEM). Line work was performed at a fixed scale of 1:40,000 to rapidly create geomorphic map units within a GIS framework.

(2) Determine Soil Setting: The soil setting was estimated by integrating the following soil-forming factors, excluding organisms:

Climatic Data: Annual precipitation was used to determine the aridity of a region to establish the climate parameter. To do this, the $\sim$ 4 km horizontal resolution digital grid estimates of Parameter-elevation Regressions on Independent Slopes Model (PRISM) data were applied to the map area (Prism Group: http://www.prism.oregonstate.edu/).

Parent Material: The parent material of the soil of each landform map unit was determined by integrating published geologic maps (Howard, 2002) and image classification techniques within remote sensing applications of a subset of ASTER multispectral data. The classification was based on the analysis of reflectance (AST07XT) and emissivity (AST05) of rock/soil compositions (Fig. 32.3). The resulting product consists of a GIS layer representing parent material.

Landform Surface Age (Time and Topography): The assignment of relative age classes to each landform or geomorphic surface map unit was based on cross-cutting relations, surface morphology and roughness, and topographic relief observable on the multispectral imagery. Soil age was assigned to Quaternary aged (less than 1.8 million years old) deposits only, and is represented by labels, for example, Quaternary alluvial landforms are labeled Qf1 to Qf5, from oldest to youngest (Fig. 32.4).

(3) Soil Database and Predictive Soil Map: Principal soil and surface attributes were assigned to geomorphic map units using soil data supplied through a linked

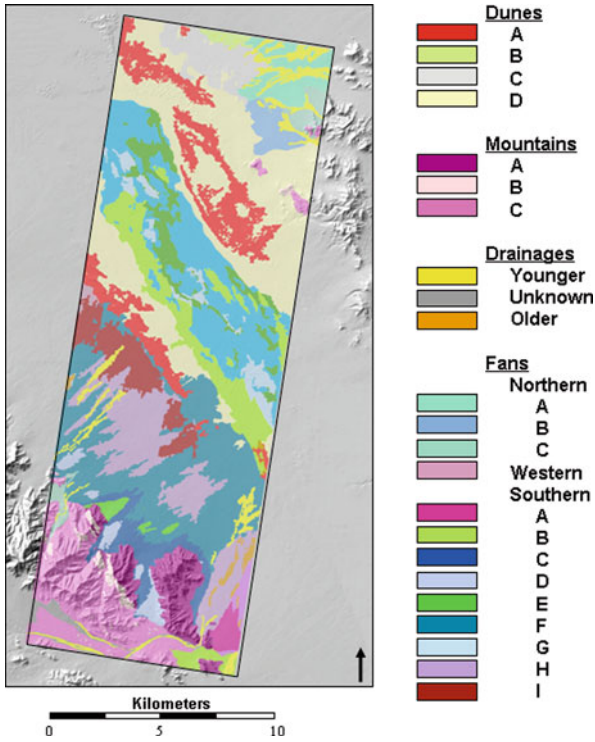


Fig. 32.3 Remote sensing mapping results of mountain highlands, alluvial fans, as well as sand dunes and playa parent material map units within Cadiz Valley based on ASTER multispectral data

database using the four soil-forming data layers. These four layers were integrated together to query the soil database through the use of an interactive GIS tool with pull-down menus that search the database for the most representative soil attributes. The database, created and maintained by the Desert Research Institute (DRI), currently contains descriptions of 813 georeferenced soil observation sites, amounting to 4,116 pedological horizon descriptions from the Mojave and Sonoran Deserts, U.S., Negev Desert, Israel, and selected sites in southwest Asia. The sources of information are primarily from published peer-reviewed journal articles and the U.S. Department of Agriculture (USDA) Natural Resource Conservation Service soil databases or published soil surveys in the southwest U.S.

(4) Terrain Hazard Map: The spatial distribution of terrain hazards across the landscape was based on soil attributes from the predictive soil map. The soil attributes of each soil map unit can be numerically modeled or qualitatively represented dependent on a specific terrain hazard of interest, and shown on a map by a six-fold hazard class system (None, Very Low, Low, Moderate, High, and Very High).

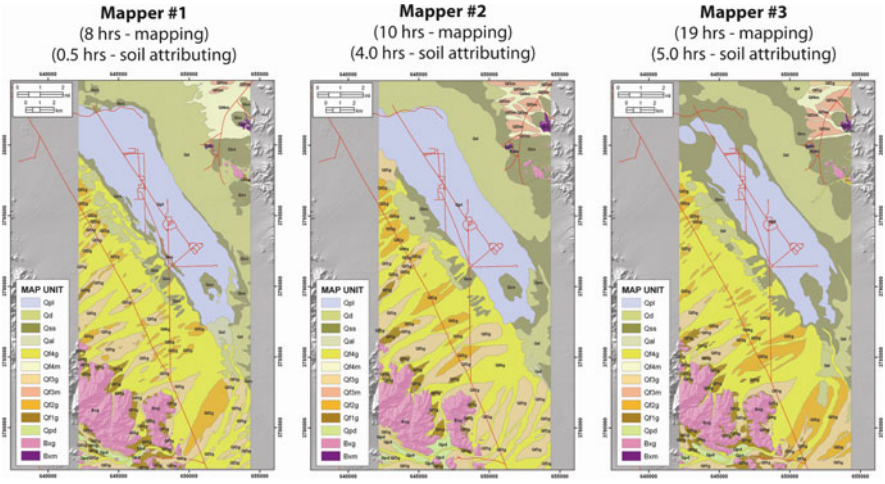


Fig. 32.4 Geomorphic mapping was conducted independently by three users (mappers) through the analysis of several digital data sources (i.e., imagery and DEM). Thirteen geomorphic map units representing the relative age of landform surfaces were delineated from satellite image interpretation by all three users. The assignment of a relative age for each map unit is based on further interpretation of cross-cutting relations and surface morphology characteristics observable on the imagery. Quaternary landform descriptors are the following: Qpl – playa; Qd – dune; Qss – sand sheet; Qal – alluvium; Qf1 to Qf4 – alluvial fan, older to younger, respectively; Qpd – pediment; Bx – bedrock. Each map unit includes a parent material/lithology descriptor derived from a USGS geologic map (e.g., g = granitic; m = meta-sedimentary), and correlation with the landform and parent material map generated from ASTER imagery (see Fig. 32.3)

32.3 Results and Discussion

32.3.1 Results of the Predictive Soil Maps and Model

We performed a test at Cadiz Valley in the Mojave Desert, California to assess the viability of the expert based model to generate a predictive soil map and resultant dust emission potential map. The test began with three users independently mapping an area of over 335 km^2 using 1:40,000-scale base maps to rapidly create geomorphic and age class layers, and then integrating these with climatic and parent material data layers. The results of the four data layers were then queried in the soil data base and soil attributes assigned to the geomorphic map unit layers, which resulted in a predictive soil map. The time taken by each of the three users to complete the predictive soil map was 8.5, 14, and 24 h respectively (Fig. 32.4).

Rapid identification and mapping of the soils and their parent bedrock sources continues to be an important step in developing a predictive model of soil and terrain attributes due to the fact that soil properties are highly dependent on the properties of the soil parent material, whether it is bedrock or sediment. In support of the

test of the predictive mapping model, the parent material in the Cadiz Valley area of California was mapped, at the same time as the geomorphic mapping was being performed, based on combinations of VIS/SWIR/TIR bands collected on 4 July 2005 by ASTER spectral imagery. The area is challenging as it is composed primarily of granitic and meta-granitic rocks, lithologies that can appear spectrally similar. The map was produced "blind," meaning those producing the map had no previous knowledge of the rock composition in the area nor did they consult any outside sources, such as geologic maps. We continue to refine our parent material mapping techniques and plan to apply them to areas with even greater parent material and landform diversity than Cadiz Valley.

32.3.2 Accuracy Assessment of Model

Current project results include success in establishing linkages between common desert landforms and soil properties, and developing a GIS-based platform linking expert-based analysis of desert terrain imagery with derived terrain property maps that can be used in predicting dust emission. To test the accuracy of the predictive mapping, we collected field data by describing and sampling test pits on eight geomorphic surfaces, and compared the image-based delineation of map unit boundaries with field observations thereof. Soil pits were dug to a depth of ~0.5 m to describe the shallow subsoil. Soil profiles were described using the conventions of the USDA, per Soil Survey Staff (1998) as modified by Birkeland (1999), which divides the soil profile into genetic horizons. Age determinations were made in the field based on correlating the degree of soil development and surface morphology of landforms in Cadiz Valley to other landforms in the Mojave Desert that have numerical age control.

The accuracy assessment of the soil predictions was achieved by comparing the field data with the predicted soil attributes (see Section 29.2) from the soil database

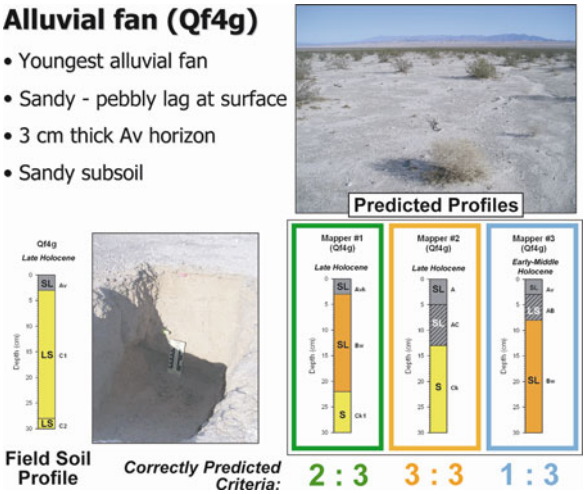


Fig. 32.5 Eight soil pits were described as part of the field validation of each mapper's soil predictions. The soil pit from alluvial fan unit (Qf4g) is included as an example. Here, the late Holocene soil was correctly predicted by mapper #2 (3 of 3), whereas mapper #1 was 2 of 3, and incorrectly predicted the master horizons. Mapper #3 was 1 of 3, and incorrectly predicted the master horizons and soil age

using three validation criteria (Fig. 32.5) as follows: (1) correct prediction of master horizon name and horizon boundary depth (to within 10 cm), (2) correct prediction of particle size to within one textural class shown on the USDA textural triangle, and (3) correct prediction of soil age. Of the 24 total possible validation criteria combinations, each of the three users had prediction accuracies of 79, 83, and 64% respectively (Fig. 32.6).

32.3.3 Resultant Dust Emission Potential Map

The dust emission potential map of Cadiz Valley was generated by assigning a dust potential rating class based on (1) a worst-case scenario where a combination of dry soil conditions and anthropogenically disturbed surfaces would allow for the greatest degree of dust emission, (2) the general terrain characteristics of each geomorphic map unit, including relief and surface morphology and their ability to limit dust emission, and (3) the estimated potential dust concentration (clay, silt, and fine sand) within a typical soil profile of each predicted soil map unit. A six-fold rating class system ranging from Very High to None was developed to categorize the disturbed (anthropogenic) dust potential during the dry season (Fig. 32.7). The degree of dust emission potential of specific geomorphic surfaces is evident across the valley, with playa and sand sheets being Very High, sand dunes as High, alluvial fans ranging from High to Low, which is primarily dependent on surface age and parent material, pediments being Very Low, and bedrock as None. Although there were some differences with the accuracy of predicting soil attributes, as well as the

Soil Prediction Results

	Mapper #1			Mapper #2			Mapper #3		
	Validation Criteria			*Validation Criteria*			*Validation Criteria*		
Geomorphic Map Unit	*Horizon*	*Particle Size*	*Soil Age*	*Horizon*	*Particle Size*	*Soil Age*	*Horizon*	*Particle Size*	*Soil Age*
Qf1g	√	√	X	√	√	X	√	√	X
Qf2g	√	√	X	√	√	X	√	√	X
Qf3g	√	√	√	√	√	√	√	√	√
Qf3m	√	√	√	√	√	√	X	X	√
Qf4g	X	√	√	√	√	√	X	√	X
Qf4m	√	X	√	√	X	√	X	√	X
Qss	√	√	√	√	√	√	√	√	√
Qpl	√	X	√	√	X	√	√	X	√
Prediction Accuracy	19 of 24 (79%)			20 of 24 (83%)			15 of 24 (63%)		
Time Taken	8 hrs-mapping 0.5 hrs-soil attributing			10 hrs-mapping 4 hrs-soilattributing			19 hrs-mapping 5 hrs-soilattributing		

Fig. 32.6 Results of the blind test following the assignment of soil attributes from DRI's soil database, and field validation of these attributes based on the validation of master horizon, particle size (textural class), and soil age. The time involved to make the predictive soil maps is also shown. $Gold = 3$ of 3 criteria were correctly predicted; $Green = 2$ of 3 criteria were correctly predicted; and $Blue = 1$ of 3 criteria was correctly predicted

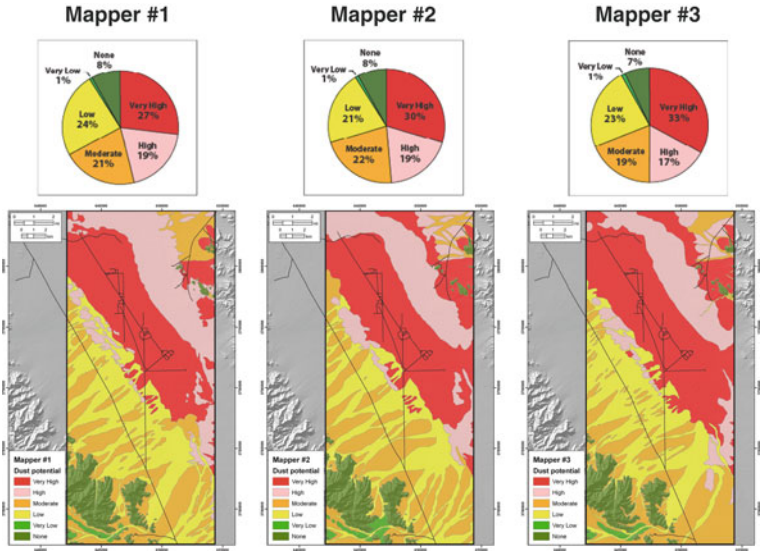


Fig. 32.7 Predictive terrain hazard map showing dust emission potential of the upper 50 cm of the soil profile under dry conditions. Results are based on the integration of geomorphic mapping and assignment of soil attributes from DRI's soil database. The spatial distribution and percent area of the hazard rating classes on individual maps are generally in good agreement between the three mappers

boundaries of geomorphic units between the three independent mappers, the spatial distribution and percent area of the hazard rating classes are similar.

32.4 Conclusions

Data integration through use of the GIS framework during the past year has resulted in significant advances in desert terrain characterization and predictive soil mapping. New tools developed to link the DRI-generated soil database with existing GIS technologies created a means whereby geomorphologists can efficiently and rapidly identify landforms and assign soil characteristics. Implementation of this new tool represents a first attempt to describe surface soil characteristics simultaneously with landform identification. The new tool creates a delivery format readily usable for the prediction of terrain hazards. The soil-forming model presented here is geomorphic-based, and considers soil age as a significant factor in accurately predicting soil conditions, which differs from most pedometric derived predictive soil mapping techniques (e.g., McBratney et al., 2000; Scull et al., 2003).

This work comprises a successful initial step in the development of a rapid and efficient expert based system to map shallow soil conditions associated with distinct geomorphic features, which will be capable of producing cost-effective and high resolution predictive soil maps to support dust emission models in remote and

poorly characterized desert regions. Future efforts will incorporate information from unmanned airborne vehicles operating remote sensing instruments to quickly link surface observables to subsurface conditions over a wider scale in near real-time. Integration of climatic information and weather forecasts with knowledge of soil hydrologic properties will provide additional predictive capability, thereby refining our approach and the capability of integrating seasonal information into the model's framework.

Acknowledgements We thank Suzann Kienast-Brown for a thoughtful and constructive review that strengthened the manuscript. Research was funded by Army Research Office (DAAD19-03-1-0159) and contributions from U.S. Army Yuma Proving Ground (W9124R-07-C-0028/CLIN 0001-ACRN-AA).

References

Birkeland, P.W., 1999. Soils and Geomorphology. 3rd ed. Oxford University Press, New York, 433 p.

Carré, F., and Girard, M.C. ,2002. Quantitative mapping of soil types based on regression kriging of taxonomic distances with landform and land cover attributes. Geoderma 110:241–263.

Caldwell, T.G., Young, M.H., Zhu, J., and McDonald, E.V., 2008. Spatial structure of hydraulic properties from canopy to interspace in the Mojave Desert. Geophysical Research Letters 35:L19406.

Howard, K.A., 2002. Geologic Map of the Sheep Hole Moutains 30'x60' Quadrangle, San Bernardino and Riverside Counties, California. U.S. Geological Survey, Miscellaneous Field Investigations MF-2344 (GIS digital database).

Jayko, A.S., Menges, C.M., and Thompson, R.A., 2005. Digital method for regional mapping of surficial deposits in arid regions, example from central Death Valley, Inyo County, CA. U.S. Geological Survey, Open-File Report 2005-1445, 31 p.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw Hill Book Company, New York. 281 p.

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., Mendonça Santos, M.L., and Minasny, B., 2003. On digital soil mapping. Geoderma 117:3–52.

Meadows, D.G., Young, M.H., and McDonald, E.V., 2006. Estimating the fine soil fraction of desert pavements using ground penetration radar. Vadose Zone Journal 5:720–730.

PRISM Group, 2007. Oregon State University. http://www.prismclimate.org (created 5 September 2007).

Scull, P., Franklin, J., Chadwick, O.A., and McArthur, D., 2003. Predictive soil mapping: A review. Progress in Physical Geography 27(2):171–197.

Scull, P., Okin, G., Chadwick, O.A., and Franklin, J., 2005. A comparison of methods to predict soil surface texture in an alluvial basin. The Professional Geographer 57(3):423–437.

Soil Survey Staff., 1998. Soil Survey Manual. USDA handbook 8th ed., Soil Conservation Service, US Government Printing Office, Washington, DC.

Sweeney, M., Etyemezian, V., Macpherson, T., Nickling, W., Gillies, J., Nikolich, G., and McDonald, E., 2008. Comparison of PI-SWERL with dust emission measurements from a straight-line field wind tunnel. Journal of Geophysical Research 113:F01012.

Chapter 33
GlobalSoilMap.net – A New Digital Soil Map of the World

Alfred E. Hartemink, Jon Hempel, Philippe Lagacherie, Alex McBratney, Neil McKenzie, Robert A. MacMillan, Budiman Minasny, Luca Montanarella, Maria de Lourdes Mendonça Santos, Pedro Sanchez, Markus Walsh, and Gan-Lin Zhang

Abstract Knowledge of the world soil resources is fragmented and dated. There is a need for accurate, up-to-date and spatially referenced soil information as frequently expressed by the modelling community, farmers and land users, and policy and decision makers. This need coincides with an enormous leap in technologies that allow for accurately collecting and predicting soil properties. We work on a new digital soil map of the world using state-of-the-art and emerging technologies for soil mapping and predicting soil properties. The global land surface will be mapped in 5 years and the map consists of the primary functional soil properties at a grid resolution of 90 by 90 m. It will be freely available, web-accessible and widely distributed and used. The maps will be produced by a global consortium with centres in each of the continents: NRCS for North America, Embrapa for Latin America, JRC for Europe, TSBF-CIAT for Africa, ISSAS for parts of Asia and CSIRO for Oceania. This new global soil map will be supplemented by interpretation and functionality options that aim to assist better decisions in a range of global issues like food production and hunger eradication, climate change, and environmental degradation. In November 2008, a grant has of US$ 18 million has been obtained from the Bill & Melinda Gates foundation to map most parts in Sub-Sahara Africa, and make all Sub-Saharan Africa data available. From this grant there are funds for coordinating efforts in the global consortium.

Keywords Soil map · Global consortium · Soil information system

A.E. Hartemink (✉)
ISRIC – World Soil Information, Wageningen, The Netherlands
e-mail: alfred.hartemink@wur.nl

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_33,

33.1 Introduction

There is no comprehensive, digital map of the world's soils. This lack of easy-to-use information results in adverse or insufficient policies at the district, national, and continental levels that affect food production and result in unnecessary land degradation and increased greenhouse gas emissions (McBratney et al., 2003). Organisations involved in research at the global scale need accurate and geo-referenced soil information to assess and/or predict the environmental impact of different land-use scenarios. A digital fine-resolution global soil map would enable climatologists, hydrologists, crop modelers, foresters and agricultural scientists, among others, to better predict the effects of climate change or new technologies on food production and environmental health. In some instances, crop failure due to drought or excessive wetness could be identified in early-warning systems in ways that are relevant to policymakers.

Within the last decade, we have witnessed remarkable progress in Earth observation techniques and prediction of soil properties from data generated by remote and on-the-ground sensors (e.g. Ben-Dor et al., 2006; Gomez et al., 2008; Hengl et al., 2007). At the same time, statistical techniques have been developed that allow for prediction of soil properties in areas with little or no information (See Chapter 12) as well as indicating the uncertainty of such predictions (see for example Chapter 19). Soil spectral analysis is becoming a robust and cheap tool for identifying soil functional attributes (Shepherd and Walsh, 2002; Vasques et al., 2008; and Chapter 7). The digital revolution that has transformed many of the sciences is waiting to be applied to create global soil maps (Sanchez et al., 2009).

33.2 The Existing Soil Maps

Soils have been mapped by systematic field observations complimented with laboratory analysis and analogue cartographic methods in many countries but in less detail in poorer countries (Hartemink, 2008). Most soil maps are polygon maps showing the characteristics of *soil types*. These maps have a long history of being useful for generalized land use planning and management. However, they have some major drawbacks:

- They are static. The maps do not provide direct information on the dynamics of soil condition (e.g. rates of nutrient depletion) whereas such information is of great interest to farmers and policymakers.
- They are inflexible for quantitative studies. Such studies (e.g. food production, erosion hazard, carbon balance) generally require information on the *functional* properties of a soil (e.g. available water capacity, permeability, nutrient supply) rather than a soil name (e.g. a polygon labeled "Kikuyu red loam" tells very little about how much nitrogen it can supply to a maize crop for the following season).
- Information is lost. The traditional map and report presents a highly summarized account of the soils of a region. The loss occurs because the reporting format requires information to be condensed and data to be classified

- The information is often presented at a scale that is seldom useful for the particular question.
- The data model implicit in polygon maps is difficult to integrate with most other forms of natural resource data that are grid based (e.g. satellite imagery, digital elevation models, climate data).

Polygon mapping was a sensible solution in the pre-digital era but digital methods allow for completely new ways of acquiring, disseminating and using soil information. Online geographic information systems, in particular, allow us to dramatically improve access to information on the functional properties of soils (McKenzie and Austin, 1993).

The current state of affairs is dire. There is one global soil map at a scale of 1:5 million (FAO UNESCO) that was produced between 1971 and 1981 (Dudal and Batisse, 1978). Based on soil surveys conducted between the 1930s and the 1970s, these class maps show the distribution of soil types using FAO terminology. Inevitably, the information behind the map is uneven and little-to-no information was available from several important regions. The FAO-UNESCO maps were digitized in vector format in 1984. In the mid 1990s, a 9-km raster version was produced by FAO, but with no improvement in the content. A digitized map does not constitute a digital map. However, these maps are an important data layer in digital soil mapping (see for example the Chapters 16 and 24).

33.3 Objectives of *GlobalSoilMap.net*

GlobalSoilMap.net is being developed to provide primary soil data in a form that will meet the demands of a broad range of users including governments, natural resource managers, educational institutions, planners, researchers and agriculturalists. The online system will provide access to the best available soil and land resource information in a consistent format across the globe – the level of detail and reliability will depend on the survey coverage and field data available in each region.

The *GlobalSoilMap.net* project has two principal objectives:

Objective 1: To compile the digital soil properties map
Objective 2: To provide a soil information system to the global scientific community that can be used for modelling and evaluation studies and that will aid in improved policy making at all levels of governance.

The process will also enable scientists, particularly from all parts of the world to exchange information and benefit from the rapid changes in technology.

A global consortium has been formed with coordinating institutes and centres in each continent (Fig. 33.1). The methodology for the global digital soil map is being developed (Chapter 34).

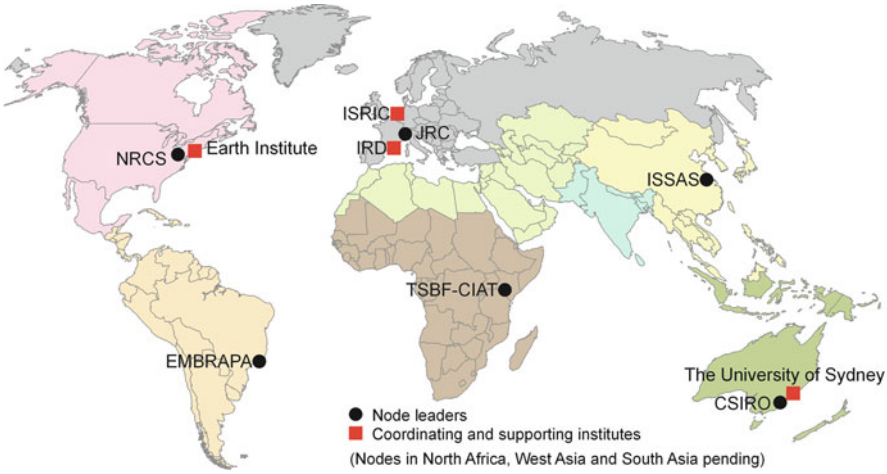


Fig. 33.1 Coordinating and supporting institutes of *GlobalSoilMap.net*

33.4 Some Outputs of *GlobalSoilMap.net*

GlobalSoilMap.net will provide access to fine grain data on a consistent set of soil functional properties that define soil depth, water storage, permeability, fertility, and carbon. Users will be able access and apply *GlobalSoilMap.net* in a variety of ways to suit their purposes. They may want to simply view and manipulate the data online (e.g. compare the soil patterns with satellite imagery or maps of land use). They may then compose and print local maps by combining several sources of online data (e.g. soil, climate, terrain and infrastructure). More sophisticated users may have portable computers with online geographic information systems that give field investigators access to useful information for their work. High-end users may take the outputs from *GlobalSoilMap.net* and supply them as inputs to sophisticated computer models for estimating food production or carbon dynamics.

GlobalSoilMap.net will provide users with an estimate of the uncertainty of each attribute for each grid cell. In the longer term, new sources of data will feed automatically into *GlobalSoilMap.net* and the uncertainties for attributes will decrease. While the information on uncertainty provides useful qualitative advice to a user, the real benefit will be for scientists, engineers and planners who need to translate their analyses of food security, impacts of climate change and so forth, into assessments of risk for decision makers. The final results may be expressed in a relatively simple form (e.g. the farming districts of Region A will fail to produce sufficient grain for local communities in 1 year out of 5) but the underlying computations will have been sophisticated.

Worldwide, millions of decisions are made every day on how to use soil and land. These range from specific on-site judgments with immediate actions, like for example, a farmer deciding to fertilize a field, or a forester establishing plantations for firewood – through decisions by governments on policy (e.g. carbon trading)

that may affect every part of their jurisdictions. Whatever the context, information is needed for sound decisions. Decisions and policies made in the absence of such information lead to inefficient use of resources and environmental degradation. *GlobalSoilMap.net* will help us assess what resources we have, their quality, and how to manage it to produce food and fiber, to secure water supplies and to conserve valuable assets (e.g. biodiversity within the soil).

References

Ben-Dor, E. et al., 2006. Quantitative mapping of the soil rubification process on sand dunes using an airborne hyperspectral sensor. Geoderma 131:1–21.

Dudal, R., and Batisse, M., 1978. The soil map of the world. Nature and Resources 14:2–6.

Gomez, C., Rossel, R.A.V., and McBratney, A.B., 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma 146:403–411.

Hartemink, A.E., 2008. Soil map density and nation's wealth and income, pp. 53–66. In: Hartemink, A.E., McBratney, A.B., and Mendonca-Santos, M.L. (eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Hengl, T., Toomanian, N., Reuter, H.I., and Malakouti, M.J., 2007. Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. Geoderma 140:417–427.

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

McKenzie, N.J., and Austin, M.P., 1993. A Quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57:329–355.

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.

Shepherd, K.D., and Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Science Society of America Journal 66:988–998.

Vasques, G.M., Grunwald, S., and Sickman, J.O., 2008. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma 146:14–25.

Chapter 34
Methodologies for Global Soil Mapping

B. Minasny and A.B. McBratney

I have here only made a bouquet of other people's flowers, having brought nothing of my own but the string that binds them together

Michel de Montaigne

Abstract The Global Digital Soil Properties Map consortium (www.GlobalSoilMap.net) has been formed with an objective to create a digital map of the world's soil properties. The methods for mapping soil properties globally are not straightforward as different parts of the world have varying data sources of varying qualities. This paper presents a set of methodologies for global digital soil mapping. The first stage involves a set of methodologies based on legacy soil data. The second stage comprises a set of methodologies to obtain new soil samples based on the available information or soil data. We present two decision trees for the methodologies and discuss each of the methods.

Keywords Regression kriging · Soil sampling · Legacy soil data · Map extrapolation · Data mining · Quality assessment

34.1 Introduction

The Global Digital Soil Properties Map consortium (www.globalsoilmap.net) has been formed and comprises representatives from universities, research centres, development organisations and private enterprises around the world (See Chapter 33). This is a response to the urgent need for accurate, up-to-date and spatially referenced soil information as expressed by the modelling community, farmers and land users, and policy and decision makers (e.g. European Commission, 2006; UNEP, 2007). The objective of this consortium is to create a digital map of the world's soil properties at a resolution of 90 m × 90 m to a depth of 1 m based

B. Minasny (✉)
Faculty of Agriculture, Food & Natural Resources, The University of Sydney, Sydney, NSW 2006, Australia
e-mail: budiman.minasny@sydney.edu.au

J.L. Boettinger et al. (eds.), *Digital Soil Mapping*, Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5_34,

on legacy and newly collected soil data. Pressures on land resources are fuelling the immediate need for soil information to make informed decisions about the soil resource. While the consortium has been established to begin the process, the methodology for global soil mapping needs to be developed to accomplish this ambitious plan.

Making a global soil map is not an easy task, and the methods for mapping soil properties globally are not straightforward as different parts of the world have varying data sources of varying qualities. The soil information can be from legacy soil profile data, existing soil maps, and data from reflectance spectra. The soil-landscape model will vary from place to place. Knowledge and techniques for regional soil mapping may not be applicable at a global scale.

This paper will present methodologies for digital soil mapping. We approach it based on a two stage process: stage 1 is mapping based on available soil information, and stage 2 focuses on how and where we should get new soil observations.

34.2 Stage 1 – A Set of Methodologies Based on Legacy Data

The methodology for global digital soil mapping based on legacy soil data is summarised in Fig. 34.1. For an area of interest, we assemble all the *scorpan* or environmental covariates and existing soil data. The second step is to check how the soil data cover the covariate space, and to select possible training areas.

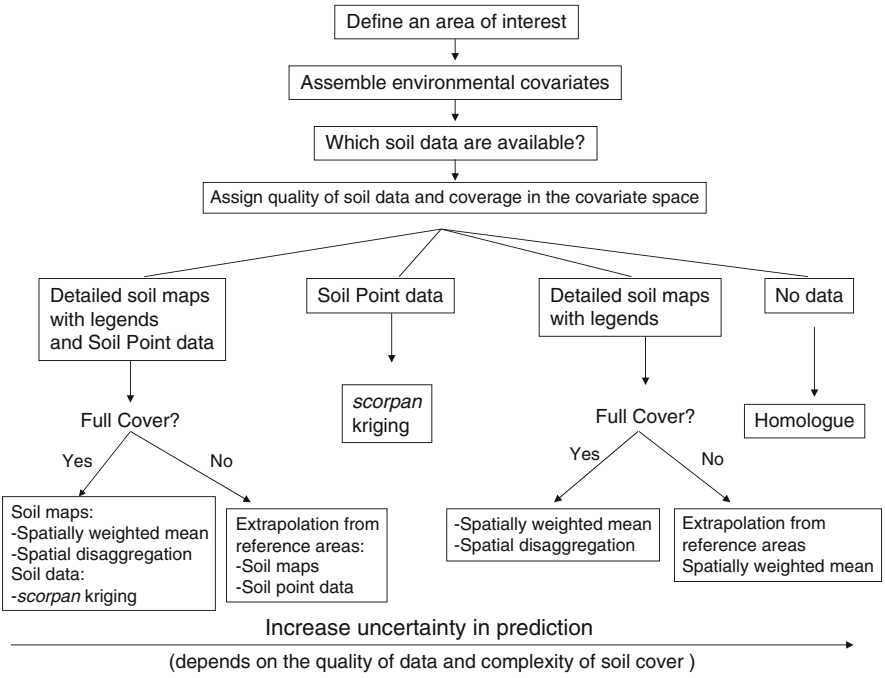


Fig. 34.1 A decision tree for digital soil mapping based on legacy soil data

The methods used for digital soil mapping depends on the availability of soil data. The possibilities in the order from the richest to the poorest soil information are:

1. *Detailed soil maps with legends and soil point data*
   This is the richest information that can give the best prediction of soil properties. Soil properties can be derived from both soil maps and soil point data. The available methods are: extracting soil properties from soil map using a spatially weighted measure of central tendency, e.g. the mean, spatial disaggregation of soil maps, *scorpan* kriging (see explanation below) and combination of these. An example of such an application is Henderson et al. (2001, 2005) in Australia and Chapter 24 in New Zealand.
2. *Soil point data*
   When soil point data are available, soil properties can be interpolated and extrapolated to the whole area by using a combination of empirical deterministic modelling and a stochastic spatial component. We have called this the *scorpan* kriging approach.
3. *Detailed soil maps with legends*
   When only soil maps are available, we need to extract soil properties from soil maps using some central and distributional concepts of soil mapping units.
4. *No data*
   When no data or soil maps exist in area, we will use an approach we call *homosoil*, which means that we need to estimate the likely soil properties under the observed soil-forming factors or *scorpan* factors.

The details of the methods are:

- *Extracting information from soil maps.*
  When a soil map is the only information available for an area, soil properties need to be derived from the map. The quality of the map depends on the scale and the sampling effect and variation of soil cover. A detailed map with scale 1:100,000 or better with a legend is assumed to be the most appropriate for use in global mapping. The information contained in soil maps is encapsulated in the definition of its soil map units, which collectively comprise the map legend, and in the spatial arrangement of the map units (Bui, 2004). The basis of soil prediction is the central concept of soil mapping unit. A soil property at a location is assumed to be the sum of the mean or median or mode mapping unit and a spatially independent noise term. The noise term accounts for the within-map variability. Bregt et al. (1987) showed that properties derived from soil survey maps have the same quality as maps obtained using spatial interpolation (kriging) of the properties. However studies also have shown that the accuracy or purity of conventional soil maps is around 60–65% (Marsman and de Gruijter, 1986).

  The current Australian Soil Resource Information System (ASRIS) estimated the soil properties based on percentage area covered by a mapping unit. Estimate of uncertainty is performed based on the proportion of the area covered and its assumed distribution (McKenzie et al., 2005).

Bui and Moran (2001) disaggregated soil associations mapped at reconnaissance scale (1:500,000–1:250,000) into pixels using the legend description and environmental variables to allocate the soils described for a particular unit into their respective landscape position. This adds detailed information to existing soil maps and can improve (increase) their scale. Their studies also showed that digital soil mapping techniques can reproduce an average of 73% of the soils series from the maps based on the environmental covariates at the resolution of 250 m × 250 m. Chapter 29 shows a comparison between conventional and pedometric approaches for mapping soil classes.

- *Extrapolation based on reference areas.*
  Lagacherie et al. (2001) proposed the computation of a taxonomic distance between the local soilscapes and those in a reference area. Soilscapes are represented by a cover frequency vector of elementary landscape classes (combinations of soil-forming factor classes) calculated within an area defined around the site under consideration. Another method which is commonly used is the modelling method of Bui and Moran (2001, 2003) where decision-tree rules are built in training areas where detailed soil maps are available, and the rules are extrapolated to larger areas where detailed mapping is unavailable. However both methods are used for extrapolation to areas within a given region and have not been tested on areas that are not geographically continuous and may be far apart. When a soil map is the only information available for an area, soil properties need to be derived from the map. The quality of the map depends on the scale of the map and the surveyor who prepared it. A detailed map with scale 1:100,000 or better with a legend is assumed to be the most appropriate. The information contained in soil maps is encapsulated in the definition of its soil map units, which collectively comprise the map legend, and in the spatial arrangement of the map units (Bui, 2004). Grinand et al. (2008) tested the idea of extrapolating regional soil landscapes from an existing soil map. They observed marked differences in accuracy between the training area and the extrapolated area. Sampling intensity did not appear to influence the accuracy of prediction. Spatial context integration by the use of a mean filtering algorithm on the covariates increased the accuracy of the prediction on the extrapolated area.

  Soil properties prediction from soil classes are usually based on the central concept of a soil mapping unit. A soil property at a location is assumed to be the sum of the mean or median or mode of the mapping unit to which it belongs and a spatially independent noise term. The noise term accounts for the within-mapping unit variability.
- *Scorpan kriging.*
  This method is used for interpolation and limited extrapolation of spatial soil point data. The assumption is that the spatial "trend" can be described by $f(s, c, o, r, p, a, n)$ and the residuals e modelled by variograms and a form of kriging. The final prediction is the sum of $f()$ and e. *Scorpan* or regression kriging (McBratney et al., 2003) allows incorporation of both deterministic and stochastic components in kriging:

$$S(\mathbf{x}) = f(\mathbf{Q}, \mathbf{x}) + e'(\mathbf{x})$$

where $f(\mathbf{Q}, \mathbf{x})$ is a function describing the structural component of S as a function of $\mathbf{Q}$ at $\mathbf{x}$, $e'(\mathbf{x})$ is the locally varying, spatially dependent residuals from $f(\mathbf{Q}, \mathbf{x})$.

In regression kriging the soil property S at unvisited site is first predicted by $f()$, and followed by kriging of the residuals of the model. A generalised least-squares approach can be adapted to refine the estimation of the parameters of the deterministic model and the spatial random process. See Chapters 18, 20, and 21 for application of regression kriging in digital soil mapping.

- *Combination of soil maps and soil point data*.
  Heuvelink and Bierkens (1992) showed that using information from soil maps and soil data together produces a more accurate map than using either of them separately. The procedure is done by taking a weighted average of soil map predictions and predictions obtained from spatial interpolation (kriging). The existing soil maps can be used as predictors as well; this usually increases the prediction accuracy (Henderson et al., 2001).
- *Homosoil*.
  This approach is required when no detailed map or soil observation is available in a region of interest. The method is based on the assumed homology of soil forming factors between a reference area and the region of interest. This includes: climate, physiography, parent materials of the area. This novel approach involves seeking the smallest taxonomic distance of the *scorpan* factors between the region of interest and other reference areas (with soil data) in the world. The rules calibrated in the reference area are applied in the region of interest realising its limitations and extrapolation consequences. This is discussed in detail in Chapter 12.

34.3 Quality Assessments

The maps produced can be assessed in terms of the uncertainty of prediction, and also the relative importance. The pixel-by-pixel adequacy of the map can be written as:

$$\text{Adequacy} = 1/(\text{Uncertainty} \times \text{relative importance})$$

The uncertainty of the prediction is quantified in terms of the input data, and prediction model. The quality of the input data depends on the laboratory analysis, surveyors experience, date of sampling, etc. Bui and Moran (2003) built a number of predictive tree models by randomly sampling over a training area, the models can be used to build up a map of consistency of predictions (or they called it as

"uncertainty"). This can be used to guide sampling in areas where the predictions are variable from tree to tree.

The greatest problem is the use of legacy soil data, as there are no statistical criteria for traditional soil sampling, and this may lead to biases in the areas being sampled. The challenge is to assess the reliability and quality of the legacy soil databases, and if there is a possibility of additional funding for sampling, to determine where new sampling units should be located. One solution is the use of hypercube sampling algorithm of Carré et al. (2007) to check the occupancy of the legacy sampling units in the hypercube of the probability distribution of the environmental covariates. This is to determine whether legacy soil survey data occupy the hypercube uniformly or if there is over- or under- observation in the partitions of the hypercube. The second part of the uncertainty is in the prediction models. Statistical measure needs to be incorporated in each of the method mentioned above; it is also able to handle the case of extrapolation. The other component of adequacy is the relative importance of an area. This is related to the target mapping; areas of the importance include identification of areas of food security problems, where soil information is urgently needed to increase food production, land degradation, and soil pollution. All of these are incorporated as a measure of adequacy of the soil property prediction. The adequacy measure will be important for targeting future sampling efforts.

34.4 Stage 2 –Methodologies to Obtain New Soil Data

When there is a possibility of obtaining funding for improving soil information, we need a good sampling design to determine where new sampling units should be located. The main issue is that existing legacy data have been sampled at different times in order to answer various questions. In this case, soil samples can have different soil variable descriptions and some areas of interest can be relatively over- or under- sampled. A general decision tree is given in Fig. 34.2, where the number and areas to be sampled depend on the type of available soil data and adequacy of the prediction model. Based on the availability of soil data, in the order from less to more soil samples required are:

- *Detailed soil maps with point soil observations*.
  The hypercube evaluation sampling method (Carré et al., 2007) can be applied to inspect the coverage of the *scorpan* variables. New sampling units are first placed in the strata with no sampling units and considering the density of covariates, to ensure that the hypercube is as maximally occupied as possible.
- *Soil point data.*
  The method of Brus and Heuvelink (2007) selects sites that minimises the variance of universal kriging using simulated annealing. This method assumes that the predictors are linearly related to the target variable, and that the variogram of the residuals is known. This method optimises sampling in both predictor

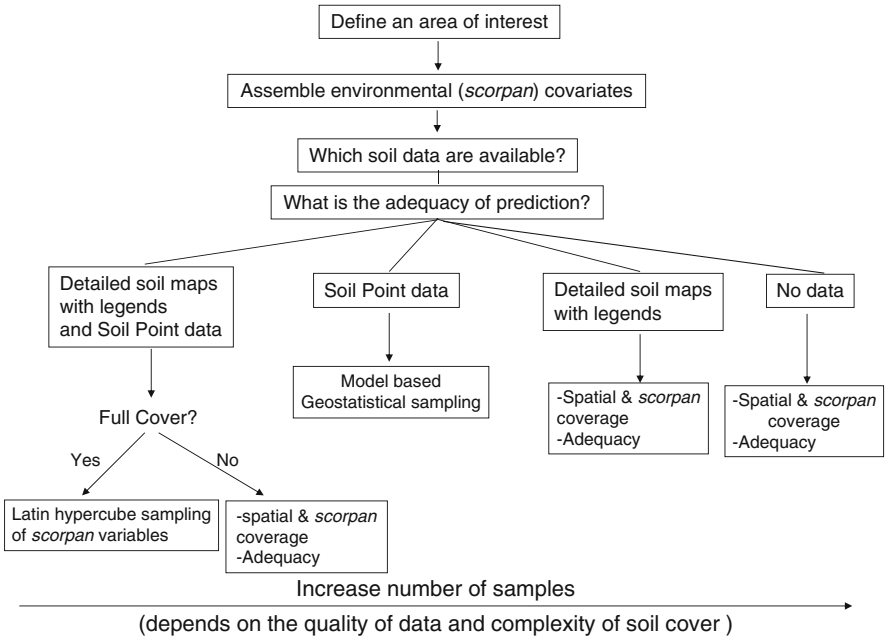


Fig. 34.2 A decision tree for soil sampling strategies

space and geographic space. However this method can be limited in practice as it assumed linearity of the prediction function and knowledge of the residual variogram.

– *Detailed soil maps*.
  Soil samples are needed to cover both representation of the soil mapping units and *scorpan* variables. The conditioned Latin hypercube sampling (cLHS) method (Minasny and McBratney, 2006) can be used in this instance. cLHS attempts to cover the range of values of each of the *scorpan* factors.
– *No soil data*
  The Latin hypercube sampling method (Minasny and McBratney, 2006) or fuzzy k-means clustering can be applied to cover both the spatial coverage and *scorpan* variables. The fuzzy k-means method (De Gruijter et al., 2008) classifies the *scorpan* variables into k classes, with k equals to the number of sampling units. The pixel with the largest membership for each class is selected as the sampling unit.

34.5 Conclusions

Here we present two sets of methodologies corresponding to two stages of global soil mapping. These are based on methods that have been successfully applied in large areas. A protocol and software need to be developed to perform this task. A working program can adapt itself into these environments resulting in the best

predictions and their uncertainties. There are many other aspects of global soil mapping that need to be addressed, such as: (1) Which method is the most robust in particular areas or situations? (2) Which are the crucial data layers? (3) What is the most appropriate way of modelling and depicting uncertainties?

References

Bregt, A.K., Bouma, J., and Jellinek. M., 1987. Comparison of thematic maps derived from a soil map and from kriging of point data. Geoderma 39:281–291.

Brus, D.J., and Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138:86–95.

Bui, E.N., 2004. Soil survey as a knowledge system. Geoderma 120:17–26.

Bui, E.N., and Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Geoderma 103:79–94.

Bui, E.N., and Moran, C.J., 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.

Carré, F., McBratney, A.B., and Minasny, B., 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma 141:1–14.

De Gruijter, J.J., McBratney, A.B., and Taylor, J., 2008. Sampling for high resolution soil mapping. Global Workshop on High Resolution Digital Soil Sensing & Mapping. The University of Sydney, Australia, 5–8 February 2008.

European Commission., 2006. Soil protection – the long story behind the strategy. Office for Official Publications of the European Communities, Luxembourg.

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 143:180–190.

Henderson, B., Bui, E., Moran, C., Simon, D., and Carlile, P., 2001. ASRIS: Continental-scale soil property predictions from point data. CSIRO Land and Water Technical Report 28/01. (http://www.clw.csiro.au/publications/technical2001/)

Henderson, B.L., Bui, E.N., Moran, C.J., and Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124:383–398.

Heuvelink, G.B.M., and Bierkens, M.F.P., 1992. Combining soil maps with interpolations from point observations to predict quantitative soil properties. Geoderma 55:1–15.

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

McKenzie, N.J., Jacquier, D.W., Maschmedt, D.J., Griffin, E.A., and Brough, D.M., 2005. ASRIS Australian Soil Resource Information System. Technical specifications. Version 1.5. October 2005. Australian Collaborative Evaluation Program.

Marsman, B., and De Gruijter, J.J., 1986. Quality of Soil Maps. In: A Comparison of Survey Methods in a Sandy AreaSoil Survey Papers No. 15, Netherlands Soil Survey Institute, Wageningen.

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

Lagacherie, P., Robbez-Masson, J.M., Nguyen-The, N., and Barthès, J.P., 2001. Mapping of reference area representativity using a mathematical soilscape distance. Geoderma 101:105–118.

UNEP, 2007. Global environmental outlook GEO4 – environment for development. United Nations Environment Programme, Nairobi.

Index

J.L. Boettinger et al. (eds.), *Digital Soil Mapping,* Progress in Soil Science 2,
DOI 10.1007/978-90-481-8863-5, © Springer Science+Business Media B.V. 2010