

19th World Congress of Soil Science

Working Group 1.3 **Digital soil assessment**

Soil Solutions for a Changing World,

Brisbane, Australia

1 – 6 August 2010

Table of Contents

	Page
Table of Contents	ii
1 A conceptual framework for a Victorian soil-landscape inference system	1
2 A hybrid soil mapping approach using SOTER, SoLIM and Classification trees	5
3 Assessment of Digital Soil Mapping products: independent ground-truthing is essential	9
4 Digital mapping of soil carbon	13
5 Digital mapping of soil waterlogging as a support to wetland delineation at regional scale: learning strategies and accuracy assessment	17
6 Disaggregation of landform components within land systems of the Victorian Mallee using a Digital Elevation Model	20
7 Generalized linear models and multivariate analysis applied to predict soil spatial distribution in south Brazil	24
8 Implementation of a new algorithm for Clapas in R language in order to improve efficiency of soil survey	28
9 Local regression kriging approach for analysing high density data	32
10 Methods for updating the drainage class map in Flanders, Belgium	35
11 Predictive soil mapping as a means to aggregate and improve existing soil databases using classification trees and knowledge integration	38
12 Quantifying the uncertainty in digital soil class maps developed using model-based approaches	42
13 Soil suitability and crop versatility assessment using fuzzy analysis at a farm scale	46
14 Spatial frameworks to support digital soil mapping in Victoria	50
15 Synthesis of knowledge on soil carbon spatial patterns across a large subtropical soil-landscape in Southern U.S	54
16 Uncertainty assessment of mapping topsoil DDT spatial distribution using multiple indicator kriging	58
17 What is the spatial representation of digital soil maps? An issue of the spatial entity	62

A conceptual framework for a Victorian soil-landscape inference system

Nathan Robinson^A, Steve Williams^B, Richard MacEwan^C

^ADepartment of Primary Industries, Epsom, Victoria, Australia, Email Nathan.Robinson@dpi.vic.gov.au

^BFuture Farming Systems Research, Department of Primary Industries, Epsom, VIC, Australia, Email Steve.Williams@dpi.vic.gov.au

^CFuture Farming Systems Research, Department of Primary Industries, Epsom, VIC, Australia, Email Richard.MacEwan@dpi.vic.gov.au

Abstract

The demand for soil information over varying spatial and temporal extents to supply a range of modelling, planning and reporting needs is increasing. A soil-landscape inference system (derived from soil inference systems – SINFERS) provides a system for managing the evolution of digital soil mapping products including spatially continuous or classified soil properties in a logical and ordered manner. The Victorian Soil-Landscape Inference System (VSLIS) concept is a three component model with input, inference and output (or product) phases that resemble a basic and logical workflow. VSLIS will deliver estimates of uncertainty for soil parameters in high demand by modellers, and deliver information products to users via the Victorian Resources Online (VRO) web portal using the Victorian Soil Information System (VSIS) as its data engine. The VSLIS operation will require multiple inference methods, even for the same soil parameter, to be spatially assigned and constrained to satisfy Victoria's physiographic and pedological diversity and input data qualities.

Key Words

Soils, inference systems, modelling, data.

Introduction

Many biophysical and socioeconomic models require soil data to predict agricultural production and potential environmental impacts. Increasingly the demand for soil information has shifted from point measurements (e.g. soil site) to spatially continuous soil parameter surfaces at varying scales. A major limitation for modelling farming system impacts to the soil resource is the paucity of soil site data in the Victorian Soil Information System (VSIS) (MacEwan 2007). The utility of soil information collected in previous government projects may be limited depending upon the scale and purpose of the original survey (e.g. land resource assessment, land capability assessment, irrigation survey, research investigation) and therefore potentially inappropriate for new applications.

Soil inference systems (SINFERS) (Dale *et al.* 1989) were refined by McBratney *et al.* (2002) when exploring the use of Pedo-Transfer Functions (PTF's) as the knowledge rules for inference systems. Soil-landscape inference systems are derived from soil spatial inference (Lagacherie and McBratney 2007). They consist of procedures that use soil and landscape ancillary datasets to derive spatially continuous or classified soil properties. These provide the basis for inferring soil variables including soil functions. The VSLIS is largely supported through the provision of ancillary datasets or predictive environment covariates (e.g. radiometrics, digital elevation models, satellite imagery, soil maps etc.) in addition to geo-referenced soil observation data of various forms (e.g. laboratory analysis, morphology descriptions, spectroscopy, hydraulic properties and time-sequence measurements). The provision of this data must be serviced by quality systems to properly support the VSLIS. Inadequate ability to access, integrate and manipulate soil data has associated impacts on derivative information products including modelling parameters. These limitations are being addressed through refinement of the VSIS (Williams *et al.* 2009a) and information integration of the broader spectrum of soil information resources from associated systems. A soil-landscape inference system can be self-updating and provide users with the latest system predictions from most recent soil observations stored in the VSIS and ancillary datasets. This will ensure modellers will always receive the most up-to-date inference system predictions.

Victoria's landscapes and current soil information status

Victoria has a diversity of landscapes, reflecting many different processes acting on the land within the earth's crust over a considerable span of time (Jenkin 1982). Landscapes are shaped by geomorphic and pedogenic processes operating at microscopic to megascopic scales'. Approaches to predict the occurrence and uncertainty of soil-landscape parameters will need to be sensitive to both Victoria's physiographic and pedological diversity, and the availability of input data and its inherent qualities (McBratney *et al.* 2002).

Current status of soil and land data (sites and maps)

In Victoria there are over 350 documented soil and land surveys and studies that have been undertaken during the last 80 years. These surveys range in scale from 1:10 000 (large scale soil survey) to 1:250 000 (small scale soil/landform and land systems).

Soil site data from past surveys is being continually entered into the VSIS. Advances in the design of the Australian Soil Resource Information System (ASRIS) have been integrated into the VSIS. The resultant data model shares much in common with the thinking and concepts underpinning the OpenGIS implementation standard for Observations and Measurements (O&M) (OGC 2007). Currently there are 2800 higher quality soil sites in VSIS with a process to further enter soil site data across the state.

The established need for soils data in Victoria

A workshop 'Soil landscape parameters for modelling' held in Bendigo in March 2009 identified the critical (sensitive) soil parameters needed for models used by the Victorian government (Table 1). The key soil parameters for models relative to their sensitivity or dependence in these models are linked to modelling domains (hydrological, growth, carbon and other). Many of the hydrological and physical parameters are common to most model domains. The VSLIS will be used to estimate uncertainty for soil parameters in high demand by modellers, and deliver these data to users.

Table 1. Landscape model domain sensitivity to soil parameters (source Victorian DPI)

Hydrological							Growth							Carbon							Other	
Air-dry moisture content	Critical Lower Limit / Permanent Wilting Point	Drained Upper Limit / Field Capacity	K_{sat}	Infiltration rate	Rooting depth	Moisture characteristic	Clay %	Sand %	Silt %	Stones %	Bulk density	Soil structure	Soil texture	C/N ratio	CEC	NH_4	NO_3	Organic carbon	P	Total N	Carbon fractions	Soil depth

Note: the highly sensitive parameters are shaded

Overview of the Victorian Soil-Landscape Inference System (VSLIS) concept

Although the primary purpose of the VSLIS is to deliver soil-landscape parameters for modellers, it also delivers a conceptual framework to assist the management of the associated activities and processes. It will become the system that records and applies (over time where possible) the inference approaches that are deemed appropriate for the different Victorian landscapes (based on expert knowledge of the parameters being inferred, landscape and data limitations). This will provide the basis for managing the evolution of these inference systems and their associated products in a logical and ordered manner. The VSLIS will integrate with the existing DPI Modelling Information and Knowledge Environment (MIKE) (Williams *et al.* 2009b) and extend it from a passive register of modelling activity and model metadata to an active engine guiding modelling and inference activities.

The VSLIS in its simplest conceptualisation is a three component model with input, inference and output phases (Figure 1). In this form it also resembles a basic workflow. During the input phase there will be workflows associated with data packaging, delivery, and quality assurance and other "filtering" (ie spatial and temporal) and processing of data for use. The inference phase contains a complex of workflows involving analysis and modelling. These are equivalent in concept to the 'kepler' workflows described by Barsheghian *et al.* 2008. In the output phase workflows are involved in taking the output of the inference processes and assembling these into usable products. Where possible workflows will be automated enabling some products to 'live' (ie. evolve as new data or improved inference models are added). Initially the system will operate to provide a conceptual framework guiding projects and manual inference activities. This need will always be present as some forms of inference will always have a cognitive or tacit knowledge component and resist automation.

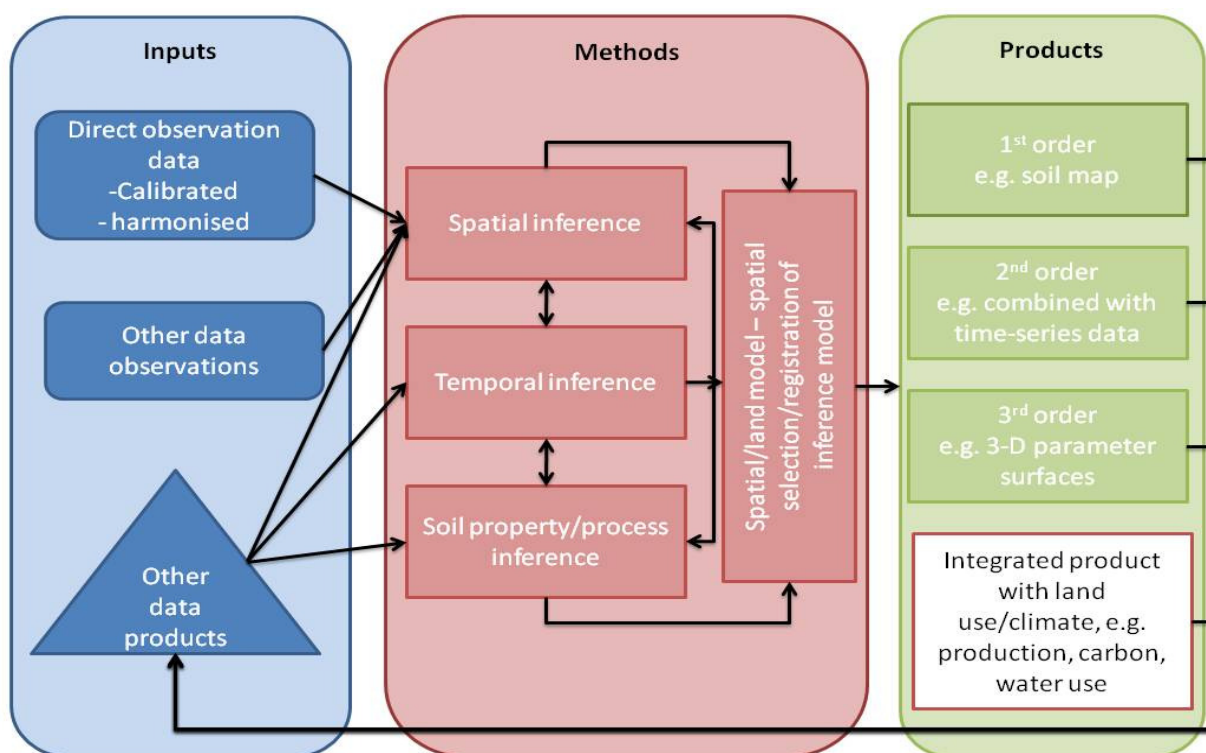


Figure 1. Proposed conceptual framework for the Victorian Soil-Landscape Inference System

Key framework elements

The key elements of the VSLIS are described below according to the component phases in the system.

Input Phase: The phase is characterised by a spectrum of services associated with input data and its metadata. These services support workflows associated with feeding data into inference processes and need to allow data filtering functions including spatial and temporal selection. Mature standards for some functionality will require further evolution and development to meet the VSLIS needs. In some cases outputs from inference processes will be fed back into the data store for use by other inference processes. It will take significant effort to create the requisite metadata to support workflows within this phase.

Inference Phase: VSLIS Inference is soil parameter centric with three domains recognised; (1) soil property, (2) temporal change and (3) soil spatial inference. Inference models and rules associated with these domains may vary in their level of workflow integration and in some cases may be quite discrete elements. Each model or rule will have individual spatial assignments to the Victorian landscape. The establishment of a systems environment to register and manage linkages between of both this spatial alignment and the available and evolving data inputs for inference elements is a significant requirement and challenge.

Output Phase: The VSLIS has three distinct orders of products based on product dimensionality (see figure 1). First order products have two or less dimensions (ie x,y or lat,long) and are represented by traditional graphs and maps. The addition of temporal or depth (z) to these products elevates it to the second order and including both progresses it to third order. Integrated products result when VSLIS products (of any order) are combined with other data and data products, often via analytical or modeling processes. Typical integrated products include land suitability and land impact mapping. Primary workflows in this phase consolidate the outputs from different inference approaches across the Victorian landscape for specific soil parameters.

Inference system selection design concept

VSLIS operation must allow multiple inference methods, even for the same soil parameter, to apply in different areas within the Victorian landscape. This requires that individual inference models be spatially assigned and constrained. This concept is used in modelling systems based on regular spatial grids such as the Catchment Assessment Toolkit (CAT) (Hocking *et al.* 2009) and Platform for Environmental Modelling Support (PEMS) (Chan *et al.* 2008) and those based on irregular spatial units such as the MIKE (Williams *et al.* 2009b). A workflow engine will be required where the inference process is fully automated. This will not only need to manage the workflows and processes associated with the inference step but also workflows to harmonise and integrate the various results. Open source tools, particularly for the latter purpose have been developed in Germany by the Humboldt project (HUMBOLDT 2009) and are freely available.

Current system status and implementation

Although many of the parts of the VSLIS have current stand-alone equivalents in the research community the system should be regarded as still in specification and design. Development is intended to be incremental and will take advantage of existing investments and initiatives where possible. That said a substantial part of the input phase is well under-way with the VSIS and MIKE systems intended as major component systems supporting the VSLIS. There are many efforts underway globally that are improving soil inference models and approaches. The interoperability and system linkages within the VSLIS have yet to be fully designed. On the output side most current products are first order, a few are second order with no known third order products.

Conclusions

Co-ordination of future developments in Victoria should be facilitated with support and engagement of the National Soil and Terrain Committee to also inform other states of Victorian progress, but also potential opportunities to collaborate with these states on areas of interest in development of a Digital Soil Map of Victoria. The approach embodied by the VSLIS should be regarded as essential in practically assisting the realisation of synergies between current soil related research and development activities within DPI. Creating the VSLIS will streamline and improve both the responsiveness and quality of the knowledge chain for users of soil and land information. This will benefit other research and associated policy formulation and natural resource management. To achieve this, the most significant future area of work for the VSLIS will be the design, development and implementation of the supporting standards, infrastructures and metadata for system workflows and interoperability. Additionally in support of users it is believed that although first order products can be produced and disseminated outside the system (ie paper or file based), in contrast the higher product orders will increasingly require interactive approaches and support within the system from visualization technologies of rising sophistication.

References

- Barsheghian D, Altintas I, Jones MB (2008) Accessing and Using Sensor Data within the Kepler Scientific Workflow System. Environmental Information Management Conference, September 10-11, 2008, pp 26-32. (Albuquerque, New Mexico).
- Chan TO, Beverley C, Ebert S, Garnett N, Lewis A, Pettit C, Thankappan M, Williams SR (2008) Platform for Environmental Modelling Support: A Grid Cell Data Infrastructure for Modellers. In 'Landscape Analysis and Visualisation: Earth and Environmental Science'. (Eds C Pettit, W Cartright, I Bishop, K Lowell, D Pullar, D Duncan) pp. 97-117. (Springer Link)
- Dale MB, McBratney AB, Russell JS (1989) On the role of expert systems and numerical taxonomy in soil classification. *Journal of Soil Science* **40**, 223-234.
- Hocking MJ, Beverley C (2009) Prioritising landscape investment using a catchment-based process model. 18th World IMACS / MODSIM Congress, Cairns, Australia, 13-17th July 2009, pp 4015-4021.
- HUMBOLDT (2009) HUMBOLDT Towards the harmonisation of spatial information in Europe. Web: <http://www.esdi-humboldt.eu/home.html>
- MacEwan RJ (2007) New Dimensions for Agricultural Landscapes, Final Report for ORL 1.1 (MIS 05188). Department of Primary Industries.
- McBratney AB, Minasny B, Cattle SR, Vervoort RW (2002) From pedotransfer functions to soil inference systems. *Geoderma* **109**, 41-73.
- OGC (2007) Open Geospatial Consortium Implementation Standard - Observations & Measurements, OGC Document 07-022 Version 1.0 (Eds S Cox)
- Williams SR, Hunter DC, Robinson N (2009a) Victorian Soil Information System, Conference proceedings of Surveying and Spatial Sciences Institute Biennial International Conference, Adelaide, Aust. Oct 2009, pp
- Williams SR, Pettit C, Hunter DC, Cherry D (2009b) A Prototype Metadata Tool for Land Use Change and Impact Models – A Case Study in Regional Victoria, Australia. 11th GSDI International Conference: SDI Convergence – Research, Emerging Trends and Critical Assessment, (Eds: Loenen, Besemer & Zevenbergen) June 2009, Delft, pp 137-150.

A hybrid soil mapping approach using SOTER, SoLIM and Classification trees

Ulrich Schuler^A, Petra Erbe^B, Karl Stahr^C and Ludger Herrmann^C

^AFederal Institute for Geosciences and Natural Resources (BGR), B2.2 Spatial Information Soil and Water, Stilleweg 2, D-30655 Hanover, Germany, Email ulrich.schuler@bgr.de

^BThe Uplands Program, Hohenheim Office, Faculty of Agriculture, Chiang Mai University, Chiang Mai 50200, Thailand, Email petra.erbe@gmail.com

^CUniversity of Hohenheim, Institute of Soil Science and Land Evaluation (310), D-70593 Stuttgart, Germany, Email kstahr@uni-hohenheim.de, herrmann@uni-hohenheim.de

Abstract

The investigated mapping approach combines different (soil) mapping concepts including SOTER (FAO 2005), SoLIM (Zhu *et al.* 2001), expert knowledge and classification trees (Breiman *et al.* 1984). North-western Thailand was stratified into SOTER terrains using DEM data and geological information. Detailed soil mapping was tested in three pilot areas within different SOTER terrains in north-western Thailand. Therefore, classification trees were calibrated with soil data from transect-based sampling points. Maps of the pilot areas were created by implementing classification rules derived from Classification Tree (CART algorithm) and expert knowledge in SoLIM. Validation was performed using soil reference maps and independent sampling points. The reference soil maps contain information from transects, sampling points along local trails, sampling points for areas of low point density or high soil variability, reference profiles, LANDSAT and SPOT images, and topographic information. The mean sampling distance was around 200 m. The reference soil maps were manually created based on expert knowledge using ArcGIS 9.2 software. The hybrid mapping approach based soil maps showed high correspondence with the respective reference soil map and a very high degree of matching with independent sampling points. This hybrid mapping approach seems to be very useful for reconnaissance soil mapping.

Key Words

Transect mapping, reconnaissance soil mapping, and expert knowledge, northern Thailand.

Introduction

Many developing and emerging countries face unsustainable agricultural land use. The obligatory land use planning requires regional soil information. However, the data density of soil information is commonly rather low for such areas. Intensive soil mapping is very costly and time consuming. Therefore, the development of alternative quick, cheap, but sufficiently accurate mapping methods is indispensable. The main objective of this study was to develop an efficient and transparent soil mapping approach, facilitating the regionalization of soil information detailed enough to be used as baseline information for land use planning and modelling. The approach tested blends different mapping concepts including SOTER (FAO 2005), Classification Tree (CART algorithm) (Breiman *et al.* 1984), and SoLIM (Zhu *et al.* 2001). This mapping approach was tested in three pilot areas within different landscape units, reflecting together the major landscape variability in north-western Thailand.

Methods

Research area

The entire research area is located in north-western Thailand covering the area between the border to Myanmar in the north and west and the basin of Chiang Mai in the west and a line south of Doi Inthanon (Figure 1). Detailed soil mapping was carried out in three different pilot areas representing together the major landscapes of the region. The resulting maps (as well as all other available data) were used as reference for the studied mapping approach. The Mae Sa Mai pilot area (10.5 km²) represents a SOTER terrain with high gradient mountains mainly consisting of granite and gneiss. The Huay Bong pilot area (6.8 km²) has high gradient mountains with sandstone as major lithology. The Bor Krai pilot area (8.5 km²) shows strong karst features and hence limestone as major lithology.

Materials

Digitized topographic and geological maps, aerial photographs as well as LANDSAT and SPOT images were available as baseline information for the field surveys. A topographic map with a scale of 1:50,000 compiled by the Royal Thai Survey Department (1976) provided contour lines with 20 m intervals. The

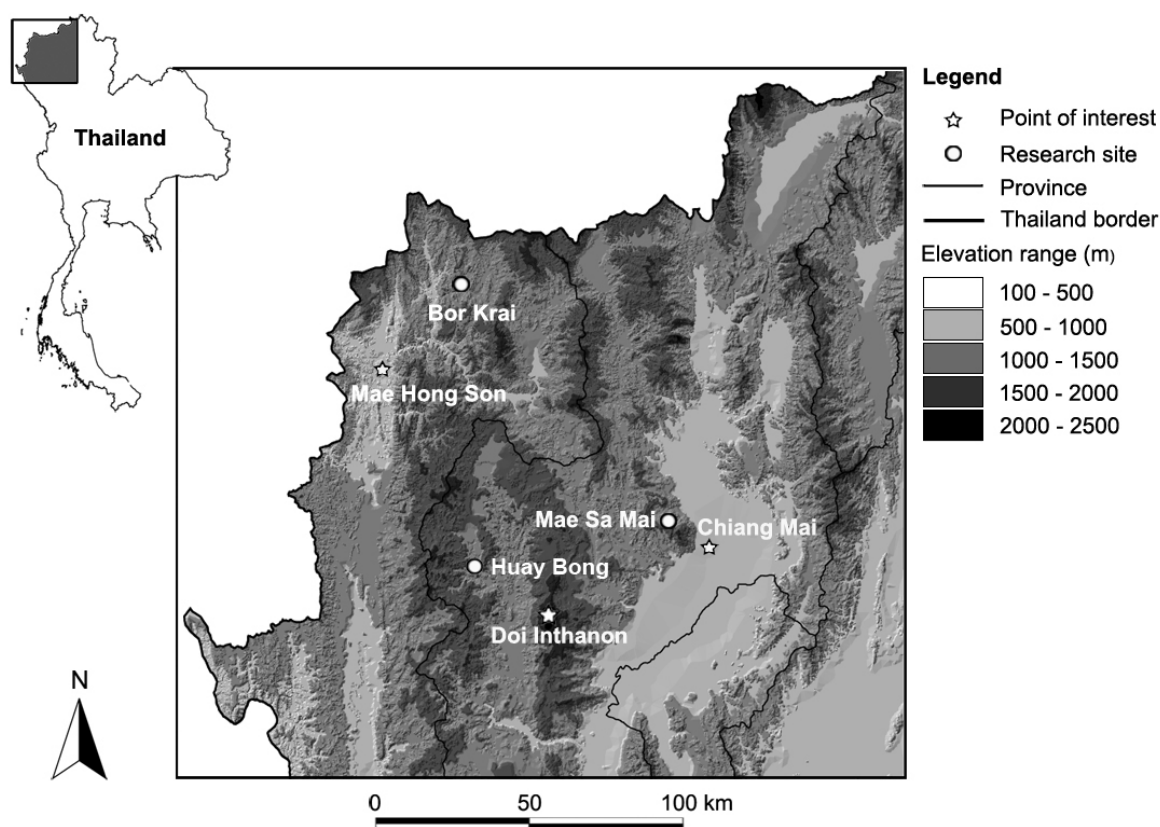


Figure 1. The research area including the three pilot areas in northwestern Thailand.

geological map has a scale of 1:250,000 and was compiled by the German Geological Mission (1979). The aerial photographs (taken in November 1999) for the Mae Sa watershed have a scale of 1:15,000 and were provided by the Royal Thai Survey Department (1999). The LANDSAT 7 ETM+ image was provided by 'Global Land Cover Facility - GLCF' (GLFC 2007). This image was taken on 5 March 2000. The SPOT 5 images were provided by 'Geo-Informatic and Space Technology Agency – GISTDA' (GISTDA 2007). The SPOT 5 image covering the Mae Sa Mai area was taken on 6 November 2006. The SPOT 5 images for Huay Bong and Bor Krai were taken on 22 February 2007, and 1 December 2006, respectively. During the field trips, a hand-held Garmin GPS III (Garmin, USA) was used to obtain coordinates of the observation points. The evaluation of the field and laboratory data was carried out using MS Access 2003 (Microsoft, USA), Past 1.81 (Paleontological Museum Oslo, Norway) and ArcGIS 9.2 (ESRI, USA) software.

Mapping

The tested hybrid mapping approach consists of the following steps:

- 1.) Delineation of SOTER terrains in order to stratify different landscapes as a baseline for Classification Tree and expert knowledge based soil mapping. The SOTER terrains were delineated blending geological information with stratified terrain classes. These terrain classes were beforehand generated blending classes of slope, relief intensity, dissection and hypsometry (Dobos *et al.* 2005).
- 2.) Transect based soil mapping in three pilot areas within different SOTER terrains. The transect lines, which cover different geomorphic units, parent materials and land use types, facilitate the detection of rules for local soil distribution (Schlichting *et al.* 1995). The number of sampling points used for transect mapping was 199 in Mae Sa Mai, 170 in Huay Bong and 322 in Bor Krai. Due to the rugged, steep, and partly inaccessible terrain the sampling was mainly performed along local trails.
- 3.) Using CART algorithm based classification trees (Breiman *et al.* 1984) to derive soil distribution rules for each SOTER terrain. The classification trees were generated using the SPSS 16.0 software package. The used input data were elevation, slope, curvature, aspect, petrography, LANDSAT 7 (bands 1-8), and SPOT 5 (bands 1-4) extracted at the locations of all transect-based training points.
- 4.) Implementation of soil distribution rules into the SoLIM model.
- 5.) Modifying and adding soil distribution rules based on expert knowledge, where soil distribution rules with the Classification Tree could not be generated due to insufficient soil information.
- 6.) Generating a soil map by running the SoLIM model.

Validation

Soil maps based on the hybrid mapping approach were validated in comparison with reference soil maps and with independent sampling points. The reference maps included the maximum amount of information available. All reference soil maps contain information from soil catenas, sampling points along local trails, and for areas of low point density or high soil variability, reference profiles, LANDSAT and SPOT images, and topographic information. The mean sampling distance was around approximately 200 m. Finally, the reference soil maps were manually created based on expert knowledge using ArcGIS 9.2 software. Additionally, for each area 15% of all sample points were randomly selected as validation points. These points were exclusively used for validation. In Mae Sa Mai 36, in Huay Bong 30, and in Bor Krai 55 validation points were used.

Results

For more than 75% of the soils clay illuviation was identified as the major soil forming process. Accordingly, soils were mainly classified as Alisols and Acrisols. Less frequent soil types were Cambisols, Umbrisols and Regosols. The remaining soil types mapped (Anthrosols, Chernozems, Ferralsols, Fluvisols, Gleysols, Leptosols, and Technosols) represent less than 2% of all soils. Reference soil maps showed that the soil cover in the Mae Sa Mai area is dominated by Acrisols (84%) followed by Cambisols (9%), Umbrisols (4%), and Technosols (2%). Anthrosols, Chernozems, Gleysols, Leptosols, and water bodies were present in the remaining area. In contrast to Mae Sa Mai, Alisols prevail in Huay Bong (77%), followed by Cambisols (13%), Regosols (9%), Leptosols (2%), and Fluvisols (0.1%). In Bor Krai the predominant mapping units were Alisols (64%), Acrisols (27%), and limestone outcrops (5%), while Cambisols, Chernozems, Ferralsols, Fluvisols, Gleysols, Leptosols, Luvisols, and Umbrisols comprise less than 1%. In all three areas, the classification tree-based maps corresponded to at least 70 % with the respective reference soil map. The correspondence to the reference map was 79 % at Mae Sa Mai, 70 % at Huay Bong, and 84 % at Bor Krai. The validation with independent sampling points showed matches of 77 % at Mae Sa, 73 % at Huay Bong and 74 % at Bor Krai.

Discussion

The tested hybrid mapping approach combines the advantages of different mapping concepts. The SOTER concept (FAO 1995) provides a guideline to stratify an area into homogenous landscape units or so called SOTER terrains. Each SOTER terrain has its characteristic topography, major lithology, climate, and soil distribution. The rules of soil distribution within the SOTER terrain units can be best explored with classification trees (Breiman *et al.* 1984). Finally, the SoLIM model (Zhu *et al.* 2001) enables the realization of classification rules and the combination of soil information of the different SOTER terrains. The major advantage of the tested hybrid mapping approach in comparison with other digital soil mapping approaches like artificial neural networks (Behrens *et al.* 2005), nominal logistic regression (Debell-Gilo and Etzelmüller 2009) or random forest (Breiman 2001) is its transparency. Each soil distribution rule is well documented and can be modified later. Further it is possible to map an area stepwise or to harmonize different map sheets.

Conclusions

The tested hybrid mapping approach is highly suitable for soil mapping especially in developing and emerging countries. The documentation of soil distribution rules facilitates a continuous improvement of soil maps by different projects or editors. Soil maps based on this hybrid mapping approach can be easily harmonized.

Acknowledgements

This research, which was carried out within the framework of the Special Research Program 564, was funded by the Deutsche Forschungsgemeinschaft (DFG) whose support is gratefully acknowledged. Thankfully acknowledged is Dirk Euler for his expert advice regarding statistical questions. Thanks go to Dr. J.F. Maxwell for critical reviewing of the manuscript and for giving his expert advice. We are very grateful to the villagers of Bor Krai, Huay Bong, and Mae Sa Mai for their support.

References

Behrens T, Förster H, Scholten T, Steinrücken U, Spies E.-D, Goldschmitt M (2005) Digital soil mapping using artificial neural networks. *Plant. Nutr. Soil Sci.* **168**, 21-33.

- Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth International group, California, United States of America 358 pp.
- Breiman L (2001) Random forests. *Machine Learning* **45**, 5-32.
- FAO (1995) 'Global and national soils and terrain digital databases (SOTER)'. World Soil Resources Reports 74. (FAO: Rome).
- IUSS Working Group WRB (2006) World reference base for soil resources 2006. World Soil Resources Reports 103, Rome.
- German Geological Mission (1979) 'Geological map of northern Thailand 1:250.000'. (Federal Institute for Geosciences and Natural Resources: Germany).
- Debella-Gilo M, Etzelmüller B (2009) Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: Examples from Vestfold County, Norway. *Catena* **77**, 8-18.
- Dobos E, Daroussin J, Montanarella L (2005) 'An SRTM-based procedure to delineate SOTER Terrain Units on 1:1 and 1:5 million scales'. (EUR 21571 EN, Institute for Environment and Sustainability, Joint Research Centre: Ispra).
- GISTDA (2007a) 'SPOT 5 image; SPOT-5 K-J: 255-312, Date: 06-11-2006'.
<http://www.gistda.or.th/Gistda/HtmlGistda/Html/index2.htm>
- GISTDA (2007b) 'SPOT 5 image; SPOT-5 K-J: 255-311, Date: 01-12-2006'.
<http://www.gistda.or.th/Gistda/HtmlGistda/Html/index2.htm>
- GISTDA (2007c) 'SPOT 5 image; SPOT-5 K-J: 255-313, Date: 22-12-2007'.
<http://www.gistda.or.th/Gistda/HtmlGistda/Html/index2.htm>
- GLFC (2007) 'Landsat ETM+, WRS-2, Path 131, Row 047, Date: 2000-03-05, EarthSat, Ortho, GeoCover Myanmar (Burma), Thailand'. <http://glcf.umiacs.umd.edu>
- Herrmann L (Editor) (2005) ,Das kleine Bodenkochbuch (Version 2005)'. (Institute of Soil Science and Land Evaluation, University of Hohenheim: Stuttgart).
- Royal Thai Survey Department (1976) 'Topographic Maps, Scale 1:50.000'. (Bangkok).
- Royal Thai Survey Department (1999) 'Aerial photographs of the Mae Sa watershed probably taken in November 1999, Scale 1:15.000'. (Bangkok).
- Schlichting E, Blume HP, Stahr K (1995) ,Bodenkundliches Praktikum'. Pareys Studentexte, Berlin 81, 1-295.
- Zhu AX, Hudson B, Burt J, Lubich K, Simonson, D, (2001) Soil mapping using GIS, expert knowledge, and fuzzy logic. *Science Society of America Journal* **65**, 1463-1472.

Assessment of Digital Soil Mapping products: independent ground-truthing is essential

Elisabeth Bui^A

^ACSIRO Land and Water, GPO Box 1666, Canberra ACT 2601, Australia

Abstract

Models and maps for predicted soil properties produced over agricultural areas of Australia using legacy soil survey data have been viewed with suspicion by many, yet these are key to the success of projects such as the GlobalSoilMap.net. While the modelling procedure encompassed a statistical model uncertainty assessment, that assessment is short of an accuracy assessment of the predicted maps. Here models and predicted maps for topsoil (0-30 cm) soil organic C, total N, and total P are presented and assessed against new independent data that serves as ground-truth for an accuracy assessment of the maps. The map of predicted SOC is credible, more so than the map of total N that consistently over-estimates N.

Keywords

Accuracy, error propagation, digital soil mapping, legacy data, soil organic C, total N, total P.

Introduction

During the first phase of the National Land and Water Resources Audit (NLWRA) the Australian Soil Resources Information Systems (ASRIS) project in 2001, a relatively large point database of soil properties was created by collating various legacy databases into a single Oracle database (Johnston *et al.* 2003). Depending on the soil property, 5,000 to 24,000 points had useful data. Using this point database linked to national environmental data for climate (19 continuous variables), geology (23 discrete classes), land use (14 discrete classes), 4 Landsat MSS bands, and topography (14 continuous terrain variables), rule induction using Cubist (<http://www.rulequest.com>) decision trees was used to predict the spatial distribution of soil properties across the intensively used agricultural areas of Australia (Henderson *et al.* 2001). In this talk, models and predicted maps for soil organic C (SOC), total N and P in the 0-30 cm depth interval will be presented and assessed. The maps have been produced at 0.01° resolution (~1.1 km) and are part of the Australian Natural Resources Atlas, available from: <http://www.nlwra.gov.au/national-land-and-water-resources-audit/atlas>.

Modelling and statistical diagnostic assessment

Cubist models are presented as a series of rules, each with starting with conditional if statement that subsets the data. Continuous predictor variables can feature as splitting criteria in conditional statements and in the linear regressions at each leaf of the piecewise linear decision trees but categorical predictors can only be used to subset the data. Models were constructed with a 70:30 training to test data split: 70% of the observations were randomly selected to construct the model in the model development stage; 30% were held back in order to assess the performance of each model. Once the strongest possible model according to performance on the test data was identified, it was refitted using all the data to maximize the use of the relatively sparse data over Australia, with the same model form and options. The performance of the model on the full data set was assessed by 10-fold cross validation. The data were randomly split into 10 partitions or folds; at each step, nine of these partitions were used to fit the model and the performance assessed on the remaining partition held back as the test data. This procedure was repeated for each partition sequentially. The performance, averaged over all 10 partitions held back, delivers the cross-validated performance assessment. The performance of models was also assessed in terms of a number of key indicators: the number of points used in the model, the R^2 between measured and predicted values, the (rank) correlation, the RMSE (root mean square error), which gives an estimate of the standard deviation of the errors, the average error, and the relative error. The average error gives the average absolute difference between the observed and predicted values, i.e

$$\text{average error} = \frac{1}{m} \sum_{j=1}^m |y_j - \hat{y}_j|.$$

Lower average errors imply that the predicted values are closer to the observed values more often. The

average error is also known as the mean absolute deviation. The relative error is defined as the ratio of the average absolute error magnitude to the average error magnitude that would result from predicting the mean value:

$$\text{relative error} = \frac{\frac{1}{m} \sum_{j=1}^m |y_j - \hat{y}_j|}{\frac{1}{m} \sum_{j=1}^m |y_j - \bar{y}|}$$

If there is little improvement on the mean, the environmental variables have little predictive capacity and the relative error is close to 1. Generally, the smaller the relative error, the better the model. These model diagnostic statistics were reported for the model test subset and for discrete regions of Australia in (Henderson *et al.* 2001). They are summarized in Table 1 for the Cubist models used to map SOC and P.

Table 1. Performance of final Cubist models used to make predictions and map the soil properties

Model diagnostics	SOC	Total P
Number of points used	11483	7377
R ² (predicted vs observed)	0.49	0.83
Average error	0.38	0.61
Relative error	0.64	0.49

Because of a strong correlation between SOC and total N and the poor spatial distribution of points with N measurements, N was predicted as a function of SOC: $\log N = -2.6589 + 0.8761 \log \text{SOC}$ (Henderson *et al.* 2001). The simple linear regression model for total N appeared to over-predict N at the low end but its statistical performance assessment was good: RMSE was relatively low (0.42 on the log scale, $R^2 = 0.75$) (Henderson *et al.* 2001).

Knowledge-based assessment

The modelling is not explicitly spatial, i.e., it does not use geographical coordinates as predictors, rather, spatial structure is introduced implicitly by reliance on predictors that are available spatially extensively. In the absence of newly collected, independent ground-truth data, evaluation of the models in a spatial context can proceed via an evaluation of the spatial distribution of the predictors in the context of model structure (Bui *et al.* 2006). ASRIS maps were thus assessed against expert knowledge in natural sciences using visualization of model rules and of patterns of usage of predictor variables—what variables were important in models, whether consistent patterns emerged in their thresholds, and the spatial pattern defined by these thresholds (Bui *et al.* 2006).

The Cubist model for soil organic C had 29 rules whereas the model for total P had 18 rules—the smaller number of rules for the P model suggests that the environmental correlation patterns are more evident in that dataset. The model for total P performed better than that for SOC in terms of model evaluation statistics (Table 1) however both appeared reasonable in terms of their predicted spatial patterns and what is known about the soil processes driving these soil nutrient patterns (Bui *et al.* 2006). Climatic variables alone were the most important predictors in the SOC model whereas lithology was also important in the total P model. Visualization of model rules showed a spatial correspondence between extent of rules and bioregions of Australia, as independently determined by the Interim BioRegionalization of Australia expert committee. A major spatial pattern in climatic thresholds seemed to correspond to soils with SOC > 2% and to the distribution of rainforests and *Eucalyptus* forests along the Australian coast.

Independent assessment/Validation

The diagnostic performance evaluation gives an estimate of the uncertainty associated with the models. However the accuracy of the predictions from the models still needs to be assessed—in remote sensing research, this is referred to as ‘validation’ and is usually performed by collecting independent ground-truth data. The dataset reported in the Auxiliary Material of Wynn *et al.* (2006) has been used as an independent dataset for validation of the predictions in ASRIS. The data of Wynn *et al.* (2006) were collected over 1999–2002 using a sampling design spatially stratified across the range of Australian native vegetation formations, and analysed by a single laboratory procedure (LECO furnace) for SOC and total N for depth 0–30 cm, near and away from trees. Unfortunately, no P data are reported. A total of 25 points overlap with the ASRIS extent.

While it appears that the SOC model for the topsoil layer (0-30 cm) is relatively poor based on the Cubist statistical model evaluation (Table 1), validation against independent data collected by Wynn *et al.* (2006) suggests that the predictions are better than suggested by the Cubist model diagnostic statistics (Figure 1). This discrepancy is likely due to the laboratory measurement errors associated with different SOC determination procedures pooled together in the ASRIS database (Henderson *et al.* 2001; Johnston *et al.* 2003): the ASRIS point database used to build and test the Cubist models contains a lot of errors. Nevertheless the Cubist algorithm was able to identify meaningful structure under fairly low signal to noise conditions to generate a credible model for topsoil SOC.

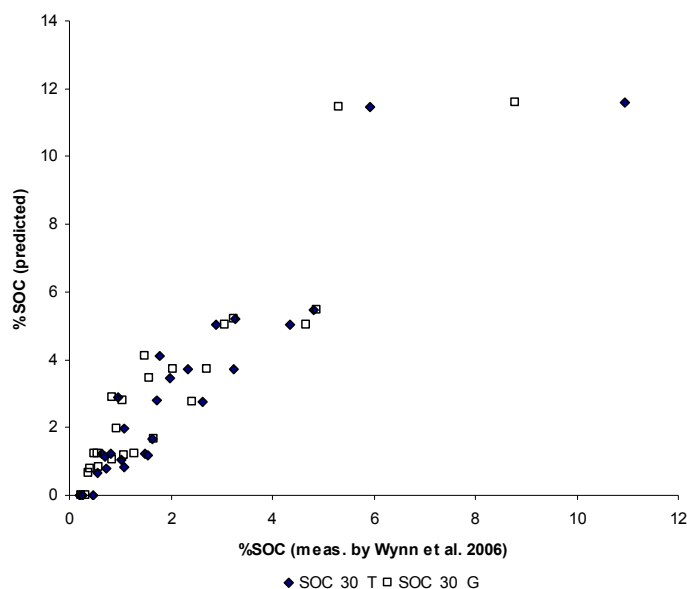


Figure 1. Relationship between SOC predicted with ASRIS data and data reported by Wynn *et al.* (2006). R^2 between predictions and SOC_30_T (near trees) is 0.84 and R^2 between predictions and SOC_30_G (away from trees, in grass) is 0.84. There is a tendency toward over-estimation of SOC.

Because of its reliance on the linear regression relationship with SOC, the total N map incorporates errors in the underlying SOC map. Validated against the data of Wynn *et al.* (2006), the modelled map for total N was found to be consistently over-estimating N throughout the range of N values (Figure 2), not only at the low end as suggested by Henderson *et al.* (2001). The likely error at high N content is much larger than at low N.

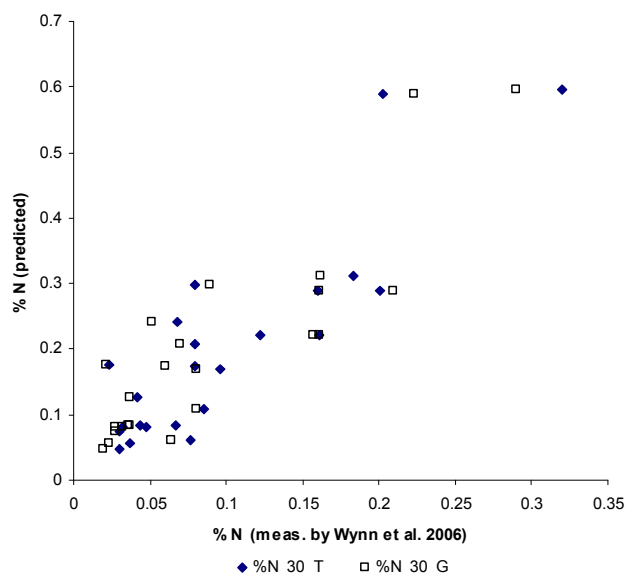


Figure 2. Relationship between total N predicted with ASRIS data and data reported by Wynn *et al.* (2006). R^2 between predictions and N_30_T (near trees) is 0.76 and R^2 between predictions and N_30_G (away from trees, in grass) is 0.77.

This problem starts with the tendency toward over-estimation in the SOC map but is also partially due to the

logarithmic transformation of N and SOC data in the linear regression model used to produce the N map. Plotting SOC against total N on a linear graph shows that there are two sub-populations, a large one associated with a C:N ratio of ~12 and another smaller one associated with C:N ratio of > 12 (Figure 3); these are not so evident on a log-log graph. Soils with a high C:N ratio have a low N content and their N level may be over-estimated by the model used to make the total N map.

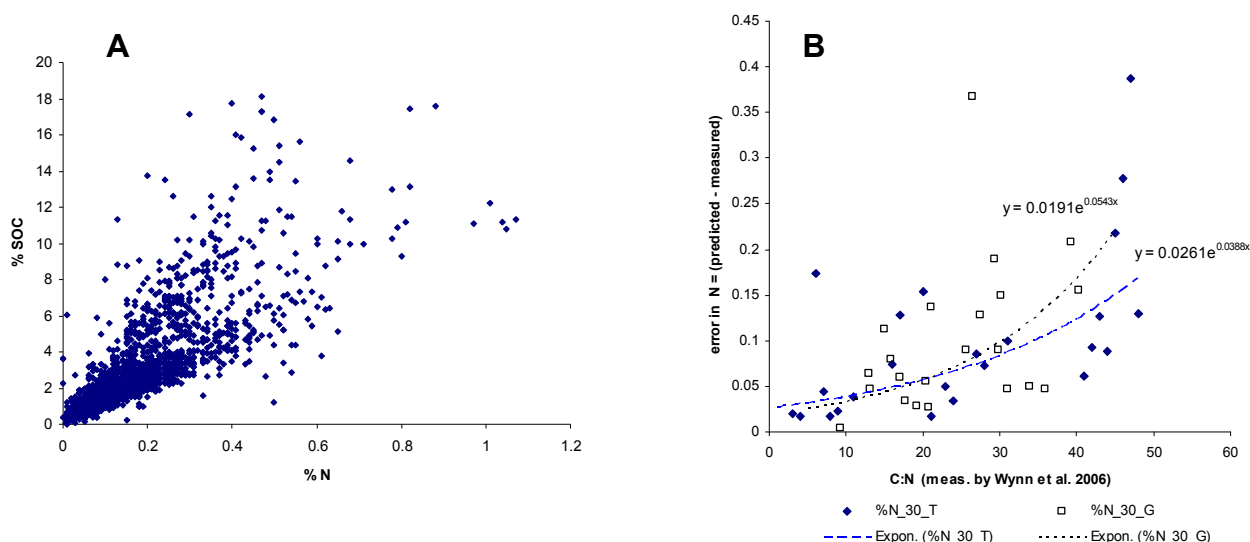


Figure 3. A) Relationship between topsoil SOC and total N on linear axes; B) Error in N predicted becomes exponentially larger as C:N increases.

Conclusion

Whereas uncertainty assessment of the models using statistical diagnostics appeared to suggest that the SOC map was not likely to be reliable, accuracy assessment against newly collected independent data suggests that the map is credible, although it shows a slight tendency toward over-estimation. The map of total N over-estimates N content consistently, especially at the upper end of the range—this shows how errors can be propagated and amplified during modelling. The P map could not be independently assessed for its accuracy. Although the independent dataset is small it demonstrates that ground-truth is essential for accuracy assessment of digital soil mapping predictions and that even a limited number of ground-truth points can be informative.

References

- Johnston RM, Barry SJ, Bleys E, Bui EN, Moran CJ, Simon DAP, Carlile P, McKenzie NJ, Henderson B, Chapman G, Imhoff M, Maschmedt D, Howe D, Grose C, Schoknecht N, Powell B, Grundy M (2003) ASRIS: The database. *Australian Journal of Soil Research* **41**, 1021-1036.
- Henderson BL, Bui EN, Moran CJ, Simon DAP, Carlile P (2001) ASRIS: Continental-scale soil property predictions from point data. *Technical Report 28/01*, CSIRO Land and Water, Canberra.
- Bui EN, Henderson BL, Viergever K (2006) Knowledge discovery from models of soil properties developed through data mining. *Ecological Modelling* **191**, 431-446.
- Wynn JG, Bird MI, Vellen L, Grand-Clement E, Carter J, Berry SL (2006) Continental-scale measurement of the soil organic carbon pool with climatic, edaphic, and biotic controls. *Global Biogeochemical Cycles* **20**, GB1007, doi:10.1029/2005GB2576.

Digital mapping of soil carbon

Budiman Minasny^A, Alex. B. McBratney^A, Brendan Malone^A, Yiyi Sulaeman^B

^AFaculty of Agriculture, Food and Natural Resources, The University of Sydney, NSW, 2006 Australia. Email budiman.minasny@sydney.edu.au, alex.mcbratney@sydney.edu.au, brendan.malone@sydney.edu.au

^B Indonesian Center for Agricultural Land Resources & Development (BBSDL P). Jl. Ir. H. Juanda 98 Bogor 16123. Indonesia. Email: yies2001@yahoo.com

Abstract

Digital soil mapping is the creation of a spatial soil information system using field and laboratory observation methods coupled with quantitative spatial prediction techniques. Digital soil mapping follows the advancement in soil and environmental observations using proximal and remote sensing. This paper discusses the methods in digital soil mapping and shows the application for mapping & monitoring soil carbon using two examples. The first example shows the mapping of whole profile soil carbon in Edgeroi, Australia. We combined equal-area spline depth functions with digital soil mapping techniques to predict the vertical and lateral variation of carbon storage across the area. We also show the uncertainty of the prediction using a new technique. The second example is the use of legacy soil data to detect the spatio-temporal changes in topsoil organic carbon in the island of Java, Indonesia.

Key words

Food security, water security, energy security, climate security, soil science, soil carbon, soil assessment, digital soil mapping

Introduction

There is a global demand for soil data and information for food security and global environmental management. This is also a large interest in recognizing the soil system as a significant terrestrial sink of carbon. The reliable assessment and monitoring of soil carbon stocks is of key importance for soil conservation and in mitigation strategies for increased atmospheric carbon. In this paper we discuss the recent advances in digital soil mapping, and show the application for mapping & monitoring soil carbon.

Digital soil mapping is defined as: the creation and population of spatial soil information systems by the use of field and laboratory observational methods coupled with spatial and non-spatial soil inference systems (Lagacherie *et al.* 2007). Digital soil mapping does not just produce a paper map; it is a dynamic process in which geographically referenced databases are created at a given spatial resolution. A digital soil map is essentially a spatial database of soil properties, based on a sample of landscape at known locations. Field sampling is used to determine spatial distribution of soil properties, which are mostly measured in the laboratory. These data are then used to predict soil properties in areas not sampled. Digital soil maps describe the uncertainties associated with such predictions and, when based on longitudinal data, can provide information on dynamic soil properties. The process is summarized in Figure 1.

There are three main steps in digital soil mapping.

Step 1, that of data input, starts with the production of base maps, assembling and calibrating full coverage of covariates from available data [e.g., the 90 × 90 m resolution digital terrain models from Shuttle Radar Topography Mission (SRTM v.3)] for the region of interest. Covariates, reflecting state factors of soil formation, include terrain attributes, gamma radiometric imagery, multi- and hyper-spectral imagery, landuse, geology and prior soil maps.

Step 2 the spatial soil inference system, which involves estimation of soil properties, expressed as estimates and their uncertainties. They are derived by using quantitative relations between point soil measurements and the spatially covered covariates. The model for digital soil mapping can be written as:

$$S = f(s, c, o, r, p, a, n) + e$$

where S is soil properties of interest; s soil and other properties of the soil; c climatic properties of the environment; o organisms; r topography; p parent material; a age (the time factor); n space (spatial position absolute and relative); e : autocorrelated random spatial variation, predicted with a variogram and kriging. This is called the *scorpan* model (McBratney *et al.* 2003).

In **step 3**, spatially inferred soil properties are used to predict more difficult-to-measure soil functions, such as available soil water storage, carbon density, and phosphorus fixation. This is achieved using pedotransfer functions built into a soil inference system (McBratney *et al.* 2002). These soil functions largely determine the capacity of soils to deliver various provisioning and regulating ecosystem services. The overall uncertainty of the prediction is assessed by combining uncertainties of input data, spatial inference, and soil functions.

There is a fourth step which leads to assessment. **Step 4** was recently elucidated by Carré *et al.* (2007). This recognizes that the information should be used to provide information to policy-makers as well as land managers.

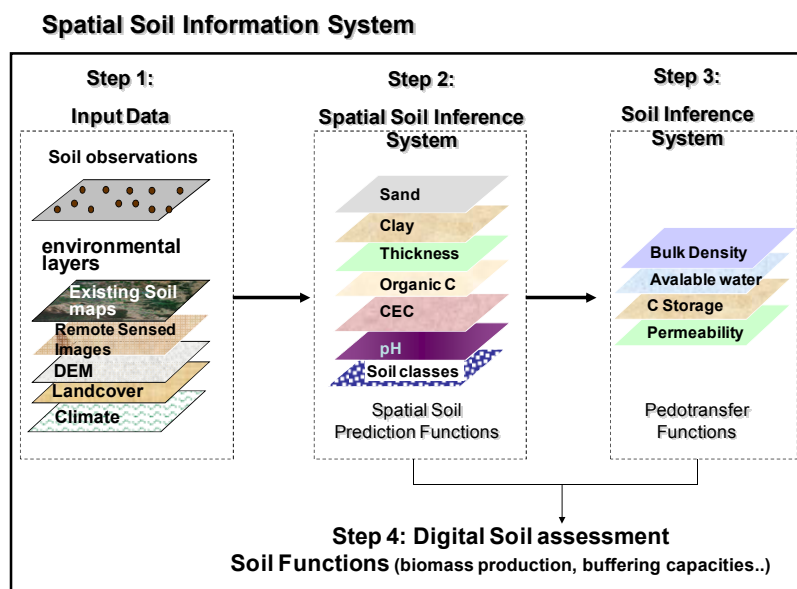


Figure 1. Digital soil mapping

Mapping and detecting changes in soil carbon

Here we provide two examples on the use of digital soil mapping and spatial inference system, for mapping and detecting the changes in soil carbon.

Mapping continuous depth functions of soil organic carbon

The *scorpan* model was expanded to include full-profile prediction at every point, by fitting the covariates to the depth parameters of an equal-area quadratic spline. The so-called continuous layer model provides much more detailed predictions. This results in predictions or maps of soil properties at potentially all depths (Malone *et al.* 2010). Using the Edgeroi district in north-western NSW as the test site, we combined equal-area spline depth functions with digital soil mapping techniques to predict the vertical and lateral variation of carbon storage across the 1500 km² area. Neural network models were constructed for soil carbon to model their relationship with a suite of *scorpan* factors derived from a digital elevation model (*r*), radiometric data (*s,p*) and Landsat imagery (*o*). The resulting geo-database of quantitative soil information describing its spatial and vertical variation is an example of what can be generated with this proposed methodology (Figure 2).

We also derived an uncertainty estimates based on a new empirical approach. Uncertainty in this case is treated as the probability distribution of the output model errors, which comprises all types of uncertainty (model structure, model parameters and data). Our approach is based on fuzzy k-means with extragrades (McBratney and De Gruijter 1992), an extension of the method by Shrestha and Solomatine (2006). The concept is to partition the model input (covariates) space into different clusters having similar values of model errors. The covariates used for prediction is partitioned into several classes using fuzzy k-means with extragrades. Each class is then represented by a prediction interval determined from the empirical distribution. The fuzzy k-means with extragrades method is also used to identify and sufficiently penalize those observations outside the domain of the calibration data. Using the class centroids, a new observation can be allocated memberships to each of the established classes. Prediction limits for new observations then can be calculated as a weighted average of the membership values.

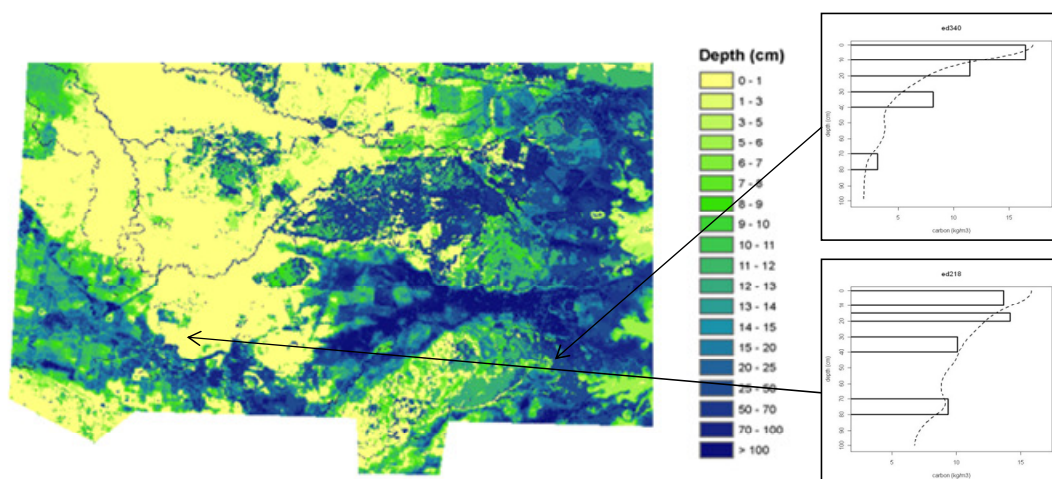


Figure 2. The depth at which soil C < 1%

Using legacy soil data to detect temporal changes in soil organic matter.

The second example deals with the use of legacy data to detect spatio-temporal changes in organic matter at a regional scale. Legacy soil data is mostly used to provide information on the spatial distribution of soil. However legacy soil data can also be used to detect the temporal changes in soil properties (Saby *et al.* 2008). The model here is based on the soil (*s*) and time (*s*) factors. A dataset of soil properties in Indonesia from 1930-1990 was compiled by Lindert (2000). The database contained 2,200 best-detailed soil profiles from Java which has uneven coverage in time, space, and land use. We extended the Lindert database to include new data of 235 profiles from surveys post 1990 conducted by the Soil Research Institute in Bogor. Here we are looking at the soil carbon content from the topsoil (Figure 3).

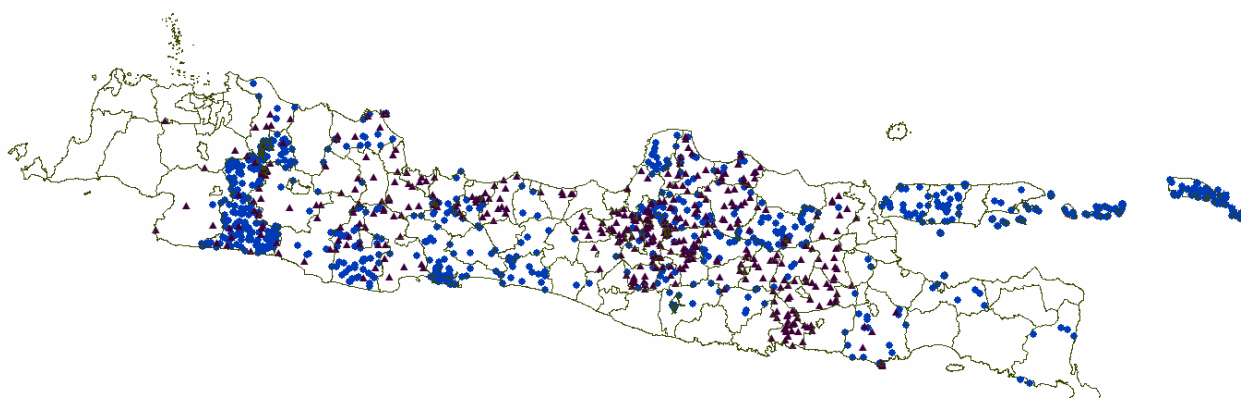


Figure 3. Data density for the period of 1930-1940 (blue dots) and 1990-2007 (red triangles).

Figure 4 shows the topsoil soil carbon content in Java over time, showing a rapid decline of soil carbon from the early 1930s to 1970. Java is the most crowded island in Indonesia, with richest soil from volcanic activities (Inceptisols, Andisols) and large floodplains (Entisols). Its land is most intensively farmed, and thus the organic C trends reflect human activities over time. The median value of C during 1930-1940 is 2.1%, while the median value in 1960-1970 is 0.7%. This rapid drop is due to the high conversion of forests into plantations and food crops. In the early development during the Dutch colonialism, most land development is towards plantations such as tea, rubber, coffee, tobacco etc. This is followed by rapid conversion to food crops in the 1950s, leading to a massive rice production in the 1960s. This resulted in a rapid decline of 1.5% of soil organic C. Following the decline, there is a slight increase of C after 1970s. This is the result of the farmers' effort to remediate the soil fertility by adding fertilizers, returning crop residues, and applying green compost and manures. In the 1990s also there is a large interest in organic farming. We can see the increase of organic C to a level of 1.1% in the 2000s. This trend is also observed in the quality or organic matter in the soil as measured by C/N values (Figure 4).

Legacy soil data come from traditional soil survey with no statistical criteria for sampling. The surveys can be selective and may be purposive and changes with space and time (Figure 3). This may lead to biases in the areas being sampled over space and time. The consistency and accuracy of laboratory methods used is also unknown. However, our empirical analysis is able to show the dynamics of soil organic C over the Java island. We argue that because we have large enough samples, we can represent the average soil C level for each period.

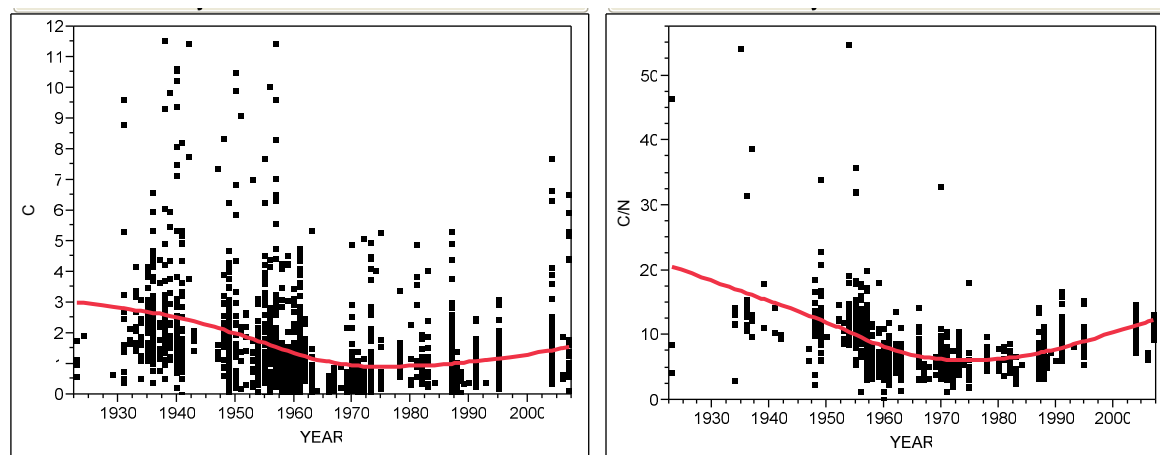


Figure 4. Soil organic C content and C/N ratio over time for top soils in Java.

Conclusions

Digital soil mapping allows the mapping and monitoring the changes in soil carbon. mapping the depth function allows the quantification of soil carbon across large areas. Legacy soil data also allow us to evaluate the dynamics of soil carbon over a large area, as affected by human activities. Although the rates of decomposition and accumulations are affected by various environmental conditions, we are able to detect the trend in Java. There is a lack of data on the accumulation of carbon over large areas; in this study we are able to estimate the average C decomposition rate in the island of Java (topsoil 0-10 cm) during 1930-1950s is 37 kg/m²/year while the accumulation rate during 1990 to 2000s is 27 kg/m²/year.

References

- Carré F, M^cBratney AB, Mayr T, Montanarella L (2007) Digital soil assessments: beyond DSM. *Geoderma* **142**, 69-79.
- Lagacherie P, M^cBratney AB, Voltz M (2007) Digital Soil Mapping: An Introductory Perspective. *Developments in Soil Science* 31 (Elsevier: Amsterdam).
- Lindert PH (2000) *Shifting Ground, The Changing Agricultural Soils of China and Indonesia*. (MIT Press: Cambridge).
- Malone BP, M^cBratney AB, Minasny B (2010). Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma* **154**, 138-152.
- M^cBratney A B, De Gruijter JJ (1992) A continuum approach to soil classification by modified fuzzy k-means with extragrades. *Journal of Soil Science* **43**, 159-175.
- M^cBratney AB, Minasny B, Cattle SR, Vervoort RW (2002) From pedotransfer functions to soil inference systems. *Geoderma* **109**, 41-73.
- M^cBratney AB, Mendonça Santos ML, Minasny B (2003). On digital soil mapping. *Geoderma* **117**, 3-52.
- Saby NPA, Arrouays D, Antoni V, Lemerrier B, Follain S, Walter C, Schwartz C (2008) Changes in soil organic carbon in a mountainous French region, 1990–2004. *Soil Use and Management* **24**, 254 – 262.
- Shrestha DL, Solomatine DP (2006). Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks* **19**, 225-235.

Digital mapping of soil waterlogging as a support to wetland delineation at regional scale: learning strategies and accuracy assessment

Christian Walter^{A,B}, Marine Lacoste^{A,B}, Blandine Lemerrier^{A,B}

^A INRA, UMR 1069, Soil Agro and hydrosystem, Spatialization, Rennes, France, Christian.Walter@agrocampus-ouest.fr

^B University Européenne de Bretagne, Agrocampus Ouest, 65 rue de St Brieuc, Rennes, France

Abstract

Environmental regulation at European level compels member states to a better protection of wetland areas. Systematic inventory and delineations of wetlands have therefore been recently undertaken in France and consider soil redoximorphic features using criteria fixed by decree. Digital soil mapping is tested as a support of wetland delineation at the regional scale of Brittany (France) and the study focuses on two objectives: (i) to compare the efficiency of learning strategies based on punctual observations or existing detailed soil maps to extrapolate spatial models of soil waterlogging; (ii) to assess the influence of prediction uncertainties of soil waterlogging on the final wetland delineation. A classification tree is used: MART (Multiple Additive and Regression Trees). MART allows the creation of predictive model based on point data and predictive variables: topography, landscape map obtain by remote sensing, airborne gamma-ray spectrometry, etc. Two kinds of models are inferred: P-Model is based on 5129 punctual soil descriptions; M-Model is based on existing soil maps at 1:25 000 scale covering 11% of the study area. Model validation by independent data set indicates an overall better performance of P-Model explained by a better spatial coverage. M-Model nevertheless performed slightly better in areas similar to its learning conditions. Probability estimates of soil waterlogging classes were finally combined to estimate the probability of occurrence of wetland conditions meeting the regulation criteria.

Key Words

Stochastic gradient boosting, classification trees, learning machine, soil waterlogging, wetland delineation.

Introduction

Enhanced access to attributes describing the physical environment and recent advances in digital soil mapping, GIS and statistics areas offer new perspectives for spatialization of soil properties. Detailed existing soil maps constitute an interesting information source on soil spatial distribution, but have a generally limited spatial extension. An alternative approach is to use punctual observations which may be more evenly distributed. The knowledge embedded into soil maps by soil surveyors or in punctual descriptions can be retrieved and explicitly formulated using environmental data (Moran and Bui 2002; McBratney *et al.* 2003).

This study considers the ability of Digital Soil mapping to assist wetland delineation at regional scale, with two objectives: (i) to compare the efficiency of learning strategies based on punctual observations or existing detailed soil maps to extrapolate spatial models of soil waterlogging; (ii) to assess the influence of prediction uncertainties of soil waterlogging on the final wetland delineation.

Material and methods

Study area

The study area is related to Brittany, a west French region of 27 020 900 ha. The geology of the region, influenced by several orogenies and transgressions, is complex. North and South Brittany mainly presents igneous and metamorphic rocks, whereas the centre of Brittany shows sedimentary rocks. Brittany is also covered by superficial deposits, particularly by Aeolian loam in the North of the area. Topography, parent material and climate gradients are considered to be the main factors of regional soil waterlogging variability.

Model creation and datasets

The MART (Multiple Adaptive Regression Tree) method was used to create predictive models. This method allows solving predictive learning problems building classification or regression trees and using “stochastic gradient boosting” (Fiedman 1999). This particular boosting method is known to significantly improve accuracy compared with simple regression and classification trees.

The model requires two kinds of input data: training data, which correspond to the response variable to predict, and environmental predictors. Two models were created:

- P-Model was based on 5 129 profiles (punctual observations of soil), and 4/5 of them were used as training data to build the models;
- M-Model was based on existing soil maps at 1:25 000 scale covering 11% of the study area. These maps were resampled and integrated as training dataset into the MART learning machine.

17 environmental predictors were used, compound by: (i) terrain attributes derived from a 50 m resolution DEM (elevation, slope.), (ii) emissions of K, Th and U derived from airborne gamma-ray spectrometry, (iii) geological variables at 1:250 000 scale (bedrock lithology and superficial deposits) and (iv) landscape map obtained by remote sensing.

Models outputs were extrapolated at regional scale with a 50 m x 50 m resolution, enabling the prediction of the occurrence probability of each soil waterlogging class.

Validation procedure

Models results were validated in four ways: (i) comparing the model predictions to all the profiles (internal validation); (ii) comparing the model predictions to 1/5 of the profiles, not used to create the models (cross-validation); (iii) comparing the model to existing soil maps at 1:25 000 scale covering 4% of the study area (external validation); (iv) estimation of the quality of the prediction by experts.

Wetland delineation accuracy

Soil waterlogging classes were interpreted considering wetland delineation criteria fixed by decree. Probabilities of the different soil waterlogging classes available were thereafter combined to estimate the probability of occurrence of a wetland.

Results

P-Model prediction of soil waterlogging

Internal and cross-validation of the model showed an accuracy of 72% and 68%. The most important predictors were respectively parent material, land use, elevation from stream, and Compound Topographic Index. Overall accuracy for external validation was 57%. According to the experts, the prediction was globally satisfactory and appeared of homogeneous quality over the region.

M-Model prediction of soil waterlogging

Internal validation accuracy was 66 %. The most important predictors were the previously parent material, Compound Topographic Index, deviation from mean K emissions and bedrock lithology. Overall accuracy for external validation was 55%. According to experts, the prediction was of good quality in situations analog to the training areas, but of poor quality in other situations.

Wetland delineation

Probabilities of occurrence of soil waterlogging classes were combined to estimate the probability of each pixel to be considered as a wetland following regulations rules. This on-going work focuses on situations with high certainty to be included or excluded in wetland zones and also on situations of high uncertainty.

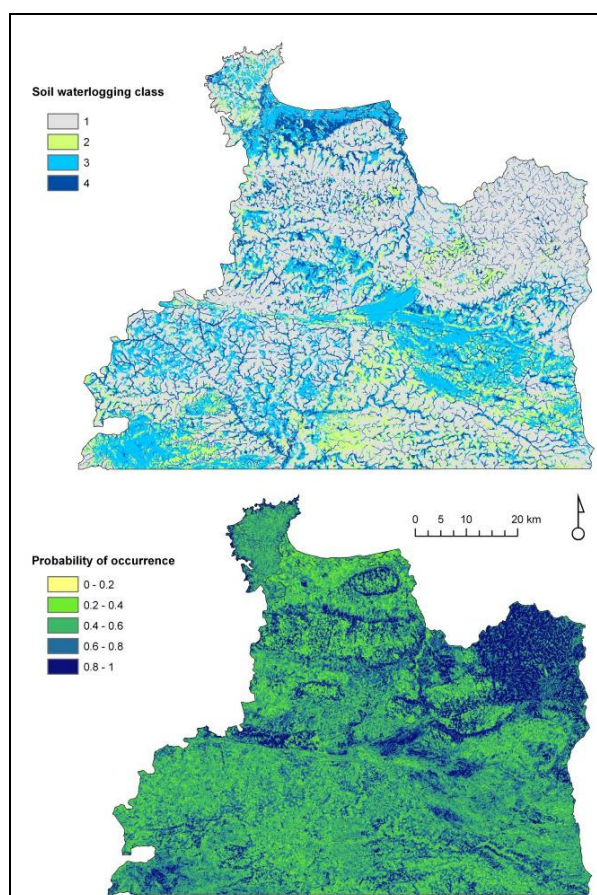


Figure 1. Prediction of soil waterlogging class and associated probability using the M-Model prediction on a subarea of Brittany

Discussion - Conclusion

The results of this work indicate an overall good performance of the MART predictions for soil waterlogging: 66 to 70% of pixels with existing soil information through profiles were correctly predicted. Soil waterlogging prediction was mostly influenced by the predicted soil parent material and topographic conditions for both models. The prediction of soil waterlogging appeared of good quality for the localization of the most redoximorphic areas, but less precise for intermediate classes. This means that digital soil mapping may be used to delineate with high reliability the most redoximorphic zones and that additional field work is needed to disentangle intermediate situations: we show that the probability of occurrence of a wetland zone as derived from a P-model is a relevant indicator to identify situations where additional information has to be gathered.

P-Model and M-Model predictions agreed at 66%. Model validation indicates an overall better performance of P-Model explained by a better spatial coverage. M-Model performed better in areas similar to their learning conditions. The external validation of the model using soil map training data showed an accuracy of 60%. So the two methods showed overall similarity, and the use of punctual information as training data in learning methods appears to be a promising approach.

References

- Friedman JH (1999) Greedy function approximation: A gradient boosting machine. *Annals of Statistics* **29**(5), 1189-1232.
- Friedman JH, Meulman JJ (2003) Multiple additive regression trees with application in epidemiology. *Statistics in Medicine* **22**, 1365-1381.
- Grinand C, Arrouays D, Laroche B, Martin MP (2008) Extrapolating regional landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma* **143**, 180-190.
- Jenny H (1941) 'Factors of soil formation.' (McGraw-Hill Publishing: New-York)
- Moran CJ, Bui EN (2002) Spatial data mining for enhanced soil map modelling. *International Journal of Geographical Information Science* **16**(6), 533-549.

Disaggregation of landform components within land systems of the Victorian Mallee using a Digital Elevation Model

Jonathan Hopley^A and Nathan Robinson^B

^AFuture Farming Systems Research, Department of Primary Industries, Epsom, VIC, Australia,
Email jonathan.hopley@dpi.vic.gov.au,

^BFuture Farming Systems Research, Department of Primary Industries, Epsom, VIC, Australia,
Email nathan.robinson@dpi.vic.gov.au

Abstract

This paper presents an approach to mapping landforms within the Victorian Mallee, a landscape primarily moulded by aeolian processes. The approach combines a variety of spatial modelling techniques based on an assessment of their ability to map target landform components identified within existing broad-scale land systems of the region. Expert knowledge of the distribution and topographic profile of these landforms, as described by Rowan and Downes (1963), has guided the approach. The success of individual modelling techniques in predicting target landforms is largely dependent upon the topography and relief of the landscape. Combining outputs from the various spatial modelling techniques, including rule-sets to combine DEM derivative surfaces (such as relative elevation, aspect, slope and curvature), the FLAG, Fuzzy Landscape Analysis GIS, model and MrVBF, Multi-resolution Valley Bottom Floor, index has proven reliable in the delineation of various landform components in the Central Mallee and Hopetoun land systems. A statistical validation undertaken to assess the quality of model outputs showed an accuracy of 84% was achieved for seven of the ten landform components (comprising 97% of the study area). An increase in the resolution of landform mapping resulting from this work will: improve the accuracy and precision of modelling; monitor land degradation with greater certainty and provide a valuable input into the creation of digital soil maps through soil inference systems.

Key Words

Terrain analysis, landform disaggregation, Victorian Mallee, FLAG, MrVBF, DEM.

Introduction

There exists a need to increase the resolution of terrain mapping in the Victorian Mallee. While soils types across the aeolian landscapes of this region are diverse and mixed, the landform components within the landscape serve as a basis for defining 'likely' soil occurrence (Murphy *et al.* 2005). Detailed landform maps will therefore improve the accuracy and precision of land capability and degradation modelling and will facilitate the creation of digital soil maps through soil inference systems.

The description of Mallee land systems provided by Rowan and Downes in 1963 still serves as the most used terrain interpretation of the Victorian Mallee. Whilst providing broad-scale information about these idealised landscape sections, the land systems at a nominal scale of 1:250,000 do not provide the spatial detail required to identify individual landforms in the landscape. This effectively reduces the resolution at which effective modelling and assessment of land management issues can occur.

The mapping of landforms utilising secondary surface derivatives, such as slope and aspect, generated from a Digital Elevation Model (DEM), has been successfully trialled within a single land system of the Mallee (MacEwan *et al.* 2007). This study seeks to further develop this methodology by incorporating the FLAG (Fuzzy Landscape Analysis GIS) model and MrVBF (Multi-resolution Valley Bottom Floor) index. The combination of the terrain modelling techniques guided by landform information contained within existing land system descriptions recognises the operational limitations of automated landform modelling in diverse, landscape-scale terrains, especially in low relief landscapes (MacMillan *et al.* 2004).

The study area comprises the Central Mallee and Hopetoun land systems of the Mallee, the terrain of which is relatively subdued in amplitude. Therefore a digital elevation model (DEM) with sufficient detail in resolution and accuracy (vertical and horizontal) was required to distinguish the component landforms.

Methods

The analysis involved the following key steps which are summarised in Figure 1.

1. *Field reconnaissance with project team, pedologists and regional experts.* Field trips were conducted where expert opinion regarding land and soil formation (including that of the land system originator, Jim Rowan) was obtained and field observations were made. Information gained from this work was used to direct the modelling approach.
2. *Preparing the DEM and generating derivative topographic surfaces including slope, aspect, curvature and relative elevation.* A DEM with a spatial resolution of 10 metres and a vertical accuracy of +/- 5 metres was clipped to the study area. Surfaces generated from the DEM using a GIS provide a range of local topographic attributes for each grid cell.
3. *Developing and applying rule-sets (value thresholds) to combine the derivative surfaces.* Map algebra rules were trialled to combine various derivative surfaces to present position in the landscape of the target landforms. For example, linear east-west dunes present in the Central Mallee land system were identified by selecting grid cells that met a relative elevation threshold and an aspect orientation.
4. *Applying the MrVBF index and FLAG model to selected sections of the land systems.* Sections were chosen based on the DEM's suitability and on regional topography. Sections with a relatively high relief, such as those containing ridges, were identified as being areas more likely to be appropriate for these models (FLAG and MrVBF).

The MrVBF index (Gallant & Dowling 2003) defines valley bottoms from hillslopes at a range of scales and combines landscape values into a single index. FLAG is a topo-sequence model that is useful in landscape delineation and identifying position in the landscape relative to other points in the terrain. An UPNESS index, the 'fraction of the total landscape monotonically uphill from each pixel', together with concave and convex break-of-slope inflection points is used to assign grid cells to different landform components of the landscape continuum based on their position in the sequence (Roberts *et al.* 1997, Summerell *et al.* 2004, Summerell *et al.* 2005).

Integration of the terrain model applications (MrVBF and FLAG) as described by Murphy *et al.* (2005) provides 'an overall better landform delineation procedure' capturing the strengths of both models. Here the MrVBF index is especially useful in mapping depositional areas within the landscape by focussing on valley floors at multiple scales, while the FLAG landforms derived from the UPNESS index attempts to represent landforms associated with hillslopes.

5. *Combining MrVBF, FLAG and rule-set model outputs to produce a quality single landform raster dataset.* The selection, combination and classification of model outputs varied across the study area depending on absolute elevation and regional terrain. The process was guided by: visual assessment of the DEM; other datasets such as aerial and geomorphological data; Rowan and Downes' (1963) descriptions of the landforms and their distribution; field visits and expert opinion. During development, visual assessment of a randomised selection of points for each modelled landform component was conducted and model outputs and combinations were refined to improve the accuracy of the mapping.
6. *Field validation of model outputs.* A field validation exercise was undertaken to measure the accuracy of the model outputs. The validation methodology involved a stratified sampling approach with 30 sample points being randomly generated for each landform component. Each sample point was visually assessed for its likely membership to the modelled landform class. Both the location of the sample point and the context of the surrounding landscape were used to inform the assessment. Sample points that were incorrectly modelled were assigned to the most likely landform class. Results, presented in Table 1, led to a final refinement of the model outputs.
7. *Cleaning the dataset to remove background noise.* Filtering algorithms, such as majority filter, and ArcScan functions were used to remove noise and fill 'holes' (unclassified cells) in the final map.

8. *Delineating land units.* The landform map provided an opportunity to delineate homogenous areas at a finer scale than the original land system mapping. These areas have been referred to as land units. Mapping these land units recognises that within the existing extents of the land systems there is variability in landform patterns. Land units also assist in distinguishing the morphological variation in land formations that have been classified as the same landform component, for example convergent and linear dune fields. The delineation has been largely based on the spatial pattern of mapped landforms.

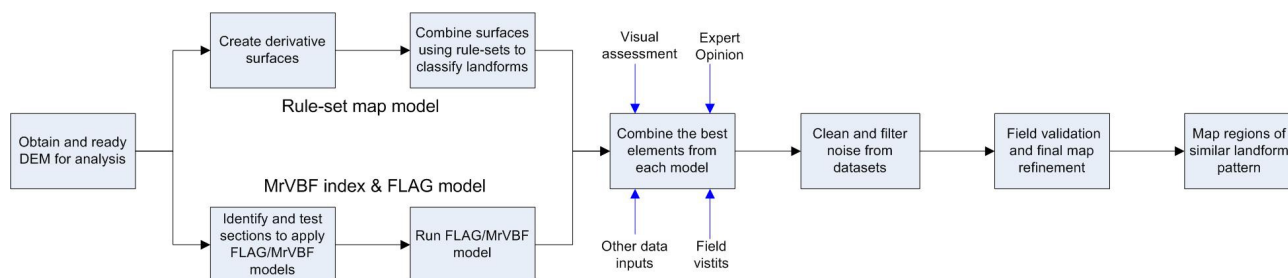


Figure 1. A chart summarising the key steps in the land system disaggregation methodology.

Results

The combination of spatial modelling techniques has proven reliable in the delineation of various landform components for the Central Mallee and Hopetoun land systems. A statistical validation undertaken to assess the quality of model outputs showed that seven of the ten landform classes, comprising 97% of the study area, achieved an accuracy of 84.6%. The three other landform classes, comprising 3% of the study area, achieved an accuracy of 50%. A re-classification of some of the mapped landforms in these three classes to the more generic landforms of ‘Undulating Plains’ and ‘Rises on Undulating Plains’ brought the overall classification accuracy to over 80%. Table 1 identifies each mapped landform class, its percentage cover of the study area and its field validation results, including an error matrix showing the most likely misclassification.

Table 1. Validation results for each landform class. Incorrect samples were assigned by visual assessment to the most likely landform class.

Target landform class	%	%	% of target sample re-assigned to a different landform class										
	Study area	Correct	EWDP	EWDR	R	RR	LRS	RUP	UP	PL	LAR	PFL	
East-West dunes on plains (EWDP)	9.5	89.7	-	-	-	-	-	3.4	6.9	-	-	-	
East-West Dunes on ridges (EWDR)	0.7	80.0	-	-	7.0	10.0	-	3.0	-	-	-	-	
Ridges (R)	4.6	80.0	-	-	-	-	-	-	20.0	-	-	-	
Rises on ridges (RR)	3.7	76.7	-	3.3	20.0	-	-	-	-	-	-	-	
Lower ridge slopes (LRS)	3.4	83.3	-	-	-	-	-	-	16.7	-	-	-	
Rises on undulating plains (RUP)	23.1	89.7	3.4	-	3.4	-	-	-	3.4	-	-	-	
Undulating plains (UP)	51.9	92.9	7.1	-	-	-	-	-	-	-	-	-	
Prominent Lunettes (PL)	1.0	56.7	6.7	-	-	-	-	30.0	3.3	-	3.3	-	
Lunettes associated with ridge (LAR)	0.3	46.7	-	-	-	-	-	-	53.3	-	-	-	
Prominent former lakebeds (PFL)	1.9	46.7	-	-	-	-	-	-	53.3	-	-	-	
Total (Average)	100.0	74.2	1.7	0.3	3.0	1.0	0.0	3.6	15.7	0.0	0.3	0.0	

Analysis of the map output indicates areas where landform components have been over- and under classified (Figure 2). This can in part be attributed to small scale variations in the morphological characteristics of the target landforms. However, as the mapping has been applied to landscape scale systems these errors of omission and commission may be deemed acceptable.

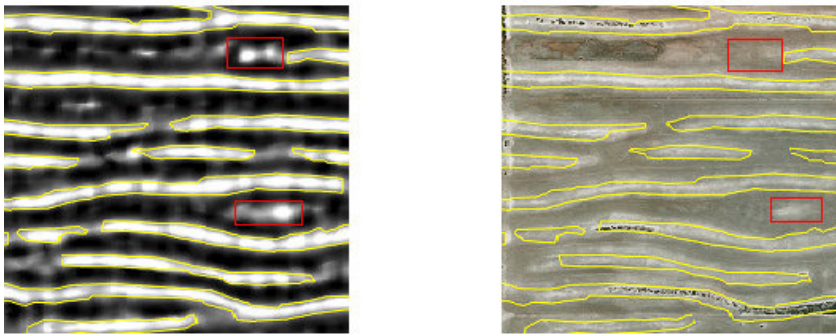


Figure 2. East–west dunes in the Central Mallee Land System. The left image shows the classified dunes (yellow polygons) superimposed over a relative elevation (70 m) surface, the right image shows the same line work superimposed over the aerial photography. The north–south extents of the dunes match well with the imagery; the areas identified by the red boxes show an under-classification of dunes in this region.

Conclusion

The ability of the MrVBF, FLAG and the rule-sets (slope, relative elevation, aspect etc) to map target landforms depends on the terrain (in particular sufficiently pronounced relief) and the topographic profiles of the landforms. Combining each of the model outputs using expert opinion, field observations, complimentary data and visual assessment has produced a landform dataset of sub-paddock resolution that will provide a sound basis for land use impact modelling and a valuable input into soil inference systems.

References

- Gallant JC, Dowling TI (2003) A multi-resolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research* **39**, 1347–1359.
- MacEwan R, Clark R, Robinson N, McNeill J (2007) ‘Assessing Wind Erosion Risk in the Mallee.’ (Department of Primary Industries: Victoria).
- MacMillan RA, Jones KR, McNabb DG (2004) Defining a hierarchy of spatial entities for environmental analysis and modelling using digital elevation models (DEMs). *Computers, Environment and Urban Systems* **28**, 175 – 200.
- Murphy B, Vaze J, Teng J, Tuteja NK, Gallant J, Summerell G, Young J, Wild J (2005) Modelling landscapes using terrain analysis to delineate landforms and predict soil depths – examples from catchments in NSW. In ‘Proceedings of the MODSIM 2005 International Congress on Modelling and Simulation’. (Eds A Zerger, RM Argent) (Modelling and Simulation Society of Australia and New Zealand).
- Roberts DW, Dowling TI, Walker J (1997) ‘FLAG: A Fuzzy Landscape Analysis GIS Method for Dryland Salinity Assessment.’ CSIRO, Land and Water Technical Report 8/97. (CSIRO: Canberra). www.clw.csiro.au/publications/technical/technical97.html
- Rowan JN, Downes RG (1963) ‘A study of the Land in North-Western Victoria.’ Soil Conservation Authority, Technical Communication No. 2, Victoria.
- Summerell GK, Dowling TI, Wild JA, Beale G (2004) FLAG UPNESS and its application for mapping seasonally wet to waterlogged soils. *Australian Journal of Soil Research* **42**, 155–162.
- Summerell GK, Vaze J, Tuteja NK, Grayson RB, Beale G (2005) Delineating the major landforms using an objective hydrological terrain analysis method. *Water resources Research* **41**, W12416, doi:10.1029/2005WR004013.

Generalized linear models and multivariate analysis applied to predict soil spatial distribution in south Brazil.

Alexandre ten Caten^A, Ricardo Simão Diniz Dalmolin^B, Fabrício de Araújo Pedron^C

^AInstituto Federal Farroupilha Campus Júlio de Castilhos, Júlio de Castilhos, RS, Brasil, Email acaten@yahoo.com.br

^BUniversidade Federal de Santa Maria – Departamento de Solos, Santa Maria, RS, Brasil, Email dalmolinrsd@gmail.com

^CUniversidade Federal de Santa Maria – Departamento de Solos, Santa Maria, RS, Brasil, Email fapedron@ymail.com

Abstract

Digital Soil Mapping (DSM) is an interdisciplinary science involving soil, statistics, mathematics, and geomatics knowledge applied to generate soil spatial information. This study aimed to use Principal Component (PC) as covariates in logistic models for the prediction of soil classes in the south of Brazil. Principal Component Analysis (PCA) was applied to nine terrain attributes: elevation, slope, distance to the nearest stream; planar curvature, profile curvature, radiation index, natural logarithm of contributing area, topographic wetness index and sediment transport capacity. The retained PC was used as explanatory covariates in Multiple Logistic Regressions (MLR), which were trained with soil information provided by an available soil map on 1:50.000 scale. The three retained components explained 65.57% of the variability in the original data. The logistic models reproduce the original map in 58.20% (kappa index), and the predictive ability of the models was 48.53%. Soil units with the smallest areas were not properly spatialized, and logistic models were not able to distinguish the soil classes too close on the landscape. MLR need further investigation, since they have a huge applying potential in the immense not mapped areas of the Brazilian territory.

Key Words

Soil map, polytomic, Shuttle Radar Topography Mission.

Introduction

Digital Soil Map (DSM) aims the creation an 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 (Lagacherie and McBratney 2007). The availability of technologies which generate data related to processes and factors of soil formation, allied to the use of statistical and mathematical techniques, makes possible the use of DSM to cater the rising demand for soil spatial information.

The Principal Component Analysis (PCA) is a multivariate method that allows the change of a set of initial variables, correlated among them, into another set of non-correlated variables, the so called, Principal Components (PC) (Johnson and Wichern 1992). It is a mathematical procedure, not statistical. Does not require the assumption of normality distribution and leads to no statistical test of significance (Webster 2001). The PCA, being based on a linear model, has a positive applicability on studies that relate soil and environmental predictors, once that, rarely exists a non-linear relation between them (Gaussian) (Odeh *et al.* 1991).

In cases where the result of an inference can be given considering many categories or polytomic (soil classes), an alternative is to work with the probability of occurrence in each one of the categories. To do so, it is applied the Multiple Logistic Regression (MLR), which is a flexible technique, because, it does not present any requirement for its application concerning the explicative variable distribution. There is no need for a normal distribution, linear correlation and measures using the same scale or homogeneity of variance. The explicative variables can be a mixture of binary data with discrete and continuous data (Chatterjee and Hadi 2006).

This study aimed at generating a soil map from a training area using predictive principal components as covariates in multiple logistic regressions.

Methods

Characterization of the study area

The study area is in São Pedro do Sul, located in the central region of Rio Grande do Sul – Brazil. This area has a surface of 873 km², being comprehended between the coordinates 29°46' to 29°26' south latitude and 54°30' to 53°56' west longitude. It comprehends a transitory region between the physiographic regions of

Medium Plateau and Central Depression in Rio Grande do Sul State – Brazil. This area was chosen because of a available semi-detailed soil map in 1:50000 scale (Klamt *et al.* 2001).

Extraction of the principal components

The terrain attributes ELEV (Elevation), SLOP (Slope), DIST (Distance to Nearest Stream), PLNC (Plan Curvature), PRFC (Profile Curvature), RADI (Radiation Index), LNCA (Natural Logarithm of Contributing Area), TWI (Topographic Wetness Index) e SPI (Stream Power Index), were generated according to Wilson & Gallant (2000) from a DEM / SRTM. A set of 70000 points were randomly created to sample the Information Plans (IP) [layers] of the terrain attributes, these charted information in the form of a text (ASCII) were processed to PCA. It was verified the adequacy of the samples through the individual test Measure of Sample Adequacy (MAS) and general Kaiser Meyer Olkin (KMO) aiming at verifying the correlation degree among the variables and the PCA justification. The eigenvalue numbers retained were conditioned to the ones that had the minimum value equal one. The rotated eigenvector (VARIMAX), resulting from PCA, was used to calculate the new variables, which are not correlated.

Predicted Map from the existing soil map

The MLR were generated using the PC as explicative variables and the soil classes in the existing soil map, to the level of order (1st Level of the Brazilian System of soil classification), as predicted variables. For adjusting the MLR models was just considered the significant parameters to the level of 5% (Wald test). Each logit function generated a probability map about the existence of a certain soil class in the landscape. These values were placed together in only one IP, with higher value among the plans defining the predicted class in that pixel. The quality of the generated maps was evaluated concerning its capacity to reproduce the original map. The capacity to predict the soil classes in an area where not data were used to generate the models was evaluated as well.

Results

Principal Components

For the PC analysis application in the correlation matrix of the attributes, it was first verified the data adequacy by the individual test MAS and general KMO. MAS values under 0.5 indicate that the variable is not proper for the PCA application, as it was mentioned in the literature. Among the terrain attributes PRFC and LNCA obtained values respectively of 0.58 and 0.56, which can be considered a low value for the application of these variables in PCA. However, being the number of attributes only nine; it was chosen to keep all the variables. The KMO value of the quality set was only 0.66, being considered a low value for the PCA application, but, the application is still possible in this case.

After the PCA application to the nine terrain attributes it was generated nine PC, each one concentrating a decreasing percentage of variability of the initial data (Figure 1). The three first PC have an eigenvalue higher than one, and they were kept in this study. Keeping just the three first components mean the loss of one-third of the data variability piled up in the new variables of the fourth and ninth component. Although there is a significant loss on the data variability pattern, there is a gain with the simplification in the number of variables.

Predicted Soil Map

The generated soil map did not displayed the spatially among the Leptosols, Plinthosols and Nitisols classes. However, the Cambisols, Lixisols and the Planosols were visually displayed in a similar way to the one already shown in the existing map and in the relation soil-landscape of the study area. The Planosols were displayed in the lower parts of the landscape and the Cambisols in the declivities and hill tops, at the same time, the Lixisols were distributed in the small hills and the Cambisols were attributed to the regions where the Leptosols were found. The reason for not displaying the Leptosols, Plinthosols and Nitisols classes was their small representativeness in the total of the used samples in the logistic models. These classes correspond only to 5%, 3% and 1% of the total in the 70.000 samples selected at random for the generation of the models.

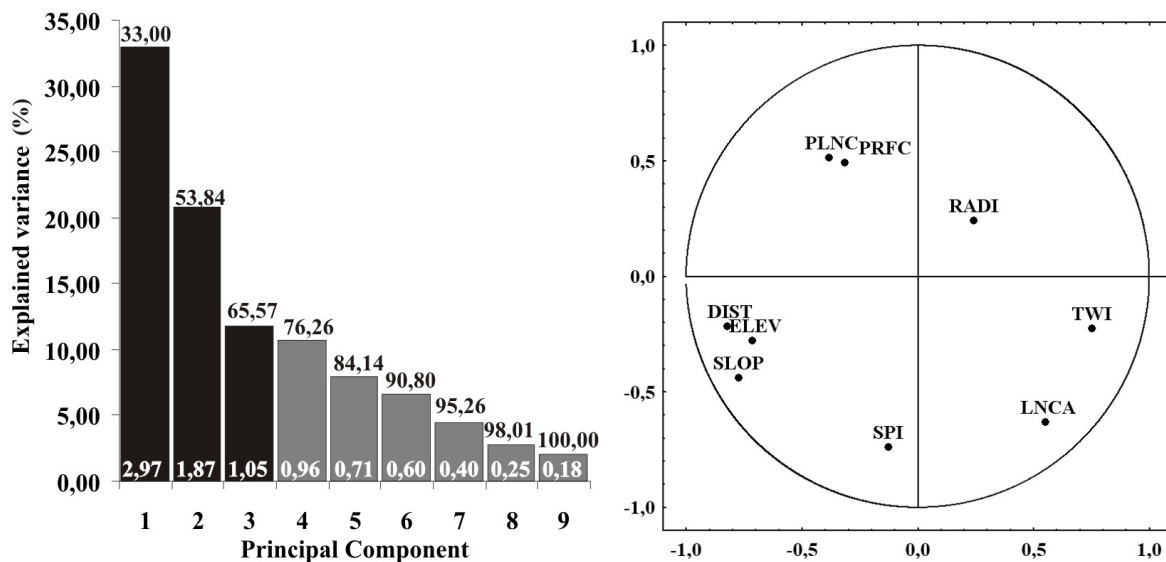


Figure1. (Left) Variability explained by each one of the nine principal components (PC) generated. The eigenvalue of each component are in blank space on the base and in the interior of each bar. The three first bars in bold have eigenvalue higher than the unit. The value over the bars indicates the accumulated percentage variability. (Right) Circular Dispersion Diagram indicates the correlations between the first (1) and second (2) PC variables. ELEV (Elevation), SLOP (Slope), DIST (Distance to Nearest Stream), PLNC (Plan Curvature), PRFC (Profile Curvature), RADI (Radiation index), LNCA (Natural Logarithm of Contributing Area), TWI (Topographic Wetness Index) e SPI (Stream Power Index).

The most common mistakes when mapping were the ones among the classes spatially close related regarding the map outlining. The soils of the wetland, near to river border were mistaken for Lixisols, these ones were mistakenly displayed in the positions of the Cambisols which were mistaken for Leptosols. These errors can have their origin in the borders of each soil class. The inference of the true class, from the analysis of the existing map, can be really difficult for the models due to problems in the outlining of the coroplectic map that served as training. The slight difference among the terrain attributes, which may not present any type of gradient in the borders of the soil class polygons, was another difficulty.

Boruvka e Penizek (2007) used neural networks for the soil class spatialization and verified that classes, which are very similar considering the constitution processes, tend to be mistaken by the models. The classes need to be well defined and distinct, to an efficient spatialization to be possible. According to the authors, the use of any methodology should consider the categorical level to be predicted due to the local heterogeneity. The available data to generate the models (number of profiles or area) should be considered as well.

The general accuracy (GA) of the predicted map was 74.3%. This value can be considered positive even that does not consider the correct mapped points by chance. To change this, it is considered the kappa index (K) 58.20% as a more realistic measure for the predicted map quality.

To verify the predictive capacity of the models, it was carried out an accuracy test in an area where no data was used to generate the model. This procedure validates the model, or, in another way, tests its real inference capacity or predictive capacity. The GA was higher than the general accuracy of the reference area, where data were used to train the models, reaching the value of 79.4%. However, a more realistic measure about the map quality in this region, shows that the map accuracy was smaller, with the K index of 48.53%, ten points less than the area where data were used to train the models.

The results of this study show the PCA potential to reduce the number of variables to be applied on the models. It also allows the visualization of the correlations among the original variables and produces new non-correlated ones. Although the PCA application implies loss of original data variability, using the three first PC allowed the logistic models to reproduce the soil classes with an accuracy of 60%, in relation to the original map. This accuracy is compatible with literature data. The use of PCA will probably be increased, with a higher number of predictive variables and with better values of MAS and KMO.

The MLR can be more effective for the soil class spatialization if these classes have a higher relative representativeness among them. Sustaining the Hengl *et al.* (2007) data, this study showed that the soils classes with smaller relative area are not spatially adequate by the logistic models.

Conclusion

The use of principal components in multiple logistic regressions implies on simplified models, comparatively, in the use of all original covariates. However, when explaining the new variables, this simplification can be associated to power reduction because of a smaller retained variance, and also, the new variables cannot have a physical, chemical or biological meaning to constitute the soil.

The logistic models presented lower quality when used outside the training area. This way, the predictive capacity of the models is associated to the model generation in similar areas; areas where the models will be applied considering the processes and factors of soil formation.

Soil classes spatially close to each other in the landscape are mistaken by predictive models.

Soil classes with smaller relative proportion data, used for model training, tend to be inappropriately predicted.

References

- Boruvka L, Penizek V (2007) A test of an artificial neural network allocation procedure using the Czech Soil Survey of Agricultural Land data. In 'Digital soil mapping: an introductory perspective'. (Eds P Lagacherie, A Mcbratney, M Voltz) pp. 415-424. (Elsevier)
- Chatterjee S, Hadi AS (2006) 'Regression analysis by example'. (John Willey & Sons)
- Hengl T (2007) Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. *Geoderma*, **140**, 417-427.
- Johnson RA, Wichern DW (1992) 'Applied multivariate statistical analysis. (New Jarsey: Prentice-Hall)
- Klamt E, Flores CA, Cabral DR (2001) Solos do Município de São Pedro do Sul. Departamento. de Solos/CCR/UFSM. Santa Maria, 96 p.
- Lagacherie P, McBratney A (2007) Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. In 'Digital soil mapping: an introductory perspective.' (Eds. P Lagacherie, A Mcbratney, M Voltz). pp. 3-22. (Elsevier).
- Odeh IOA, Chittleborough DJ, McBratney A (1991) Elucidation of soil-landform interrelationships by canonical ordination analysis. *Geoderma* **49**, 1-32.
- Webster R (2001) Statistics to support soil research and their presentation. *European Journal of Soil Science*, **52**, 331-340.
- Wilson JP, Gallant JC (2000) Digital terrain analysis. In 'Terrain analysis: principles and applications.' (Eds JP Wilson, JC Gallant) pp. 1-27. (Wiley & Sons)

Implementation of a new algorithm for Clapas in R language in order to improve efficiency of soil survey

Sébastien Lehmann^A, Olivier Vitry^B, Dominique Arrouays^A.

^AINRA INFOSOL, Orléans, France, Email sebastien.lehmann@orleans.inra.fr, dominique.arrouays@orleans.inra.fr

^B Université d'Orléans, UFR de Mathématiques, Orléans, France, Email vitryolivier@hotmail.fr

Abstract

The growing development of information technologies and the breakthroughs made in signal and image analysis allow automating many protocols. Digital Soil Mapping is one of the numerous applications illustrating this ability to engineer and innovate computed, automated and repeatable treatment and analysis methods. The clapas algorithm developed in 1994 by J.-M Robbez-Masson is within this framework. It allows the segmentation of an image using reference areas and ancillary variables chosen to discriminate different soil mapping units. The map resulting from this segmentation can then be used as a basis for a traditional soil mapping as it simplifies and eases the field work. We have implemented this algorithm in the 'R' language in order to generalize its use and to enhance its functionalities.

Key Words

Clapas, soilscape, Digital Soil Mapping, spatial segmentation, spatial analysis.

Introduction

Digital soil mapping is computer-assisted production of a digital map of soil type and/or soil properties. The generic framework of Digital Soil Mapping has been defined by McBratney *et al.* (2003) as scorpan-SSPFe (soil spatial prediction function with spatially autocorrelated errors) method. Most classification algorithms do not integrate spatial relationships within the model. However, environmental attributes at neighbouring locations can be of great interest in predicting soil pattern. Therefore, some authors investigated the potential of incorporating local neighbourhood information into the training pixels using convolution filtering operations (Grinand *et al.* 2008). Another way to integrate spatial relationships is to analyse the global patterns of soil forming factors over a window centred on the pixel of interest (Lagacherie *et al.* 2001). This kind of method has also been used in remote sensing for segmentation and landscape delineation. A method for calculating distance between soilscares, named Clapas, was initially designed for describing quantitatively, comparing and classifying soilscares in small-scale soil surveys (). It has then been adapted by Lagacherie *et al.* (2001) to map the representativity of a reference area.

In this study, the Clapas method is implemented in a free software environment and with recent spatial exploration advances, namely implementing moving windows that can be oriented according to the main relief characteristics.

Methods

Calculating distances between soilscares with Clapas

Clapas was initially designed for describing quantitatively, comparing and classifying soilscares in small-scale soil surveys (Robbez-Masson 1994). The procedure has three steps: (i) defining a soilscape variable at a given point from available soil forming factors, (ii) using this soilscape variable for a quantitative synthetic description of the site soilscape, and (iii) comparing sites regarding their soilscares.

In the following we detail these steps in succession.

Let $v(x)$ be a variable describing the elementary landscape at each site x , i.e. the point-to-point combination of soil forming factors. $v(x)$ is a categorical variable taking its values in $\{v_1, v_2, v_i, v_p\}$, the set of the p elementary landscape classes of the region. Each elementary landscape class v_i is defined by a unique combination of soil forming factor classes on a point-to-point basis. Soil forming factor classes are either mapping units (e.g. geological units or land use classes) or derived from a pre-classification of quantitative variables (e.g. elevation, slope gradient of a DEM). p classes can be identified, p being less or equal to the product of the numbers of classes of each soil forming factor.

The soilscape of site x is then defined by a “cover-frequency vector” (Wharton 1982)

$I(x) = (f_1(x), f_2(x), \dots, f_i(x), \dots, f_p(x))$ where $f_i(x)$ is the relative frequency of class v_i within an area

delineated around x to include the set of neighbouring sites which must be taken into account for describing the soilscape at x . The size and the shape of this area are user-defined. In this new version of Clapas, that we have developed in R language, the neighbourhood area is elliptic, which requires the setting of the two main radii of the ellipse, and can be automatically oriented according to the slope aspect.

The soilscape of site x can then be compared quantitatively with the one of a reference site x_0 by computing the distance $d(x, x_0)$ between the vectors $I(x)$ and $I(x_0)$. The following Manhattan distance was preferred among distances dealing with qualitative variables because of its robustness:

$$d(x, x_0) = \sum_{i=1}^p |f_i(x) - f_i(x_0)| \quad (1)$$

The distances calculated with Eq. (1) range between 0, i.e. same composition of classes within the explored areas, and 2, i.e. disjointed cover-frequency vectors with no classes in common. These distances can then be used for allocating individuals to pre-defined reference soilscales. In Clapas, allocation is performed on a nearest neighbour basis. Figure 1 provides a simple example of a Clapas application. First, four elementary soilscape classes are defined by combinations of soil forming factor maps (Figure 1a).

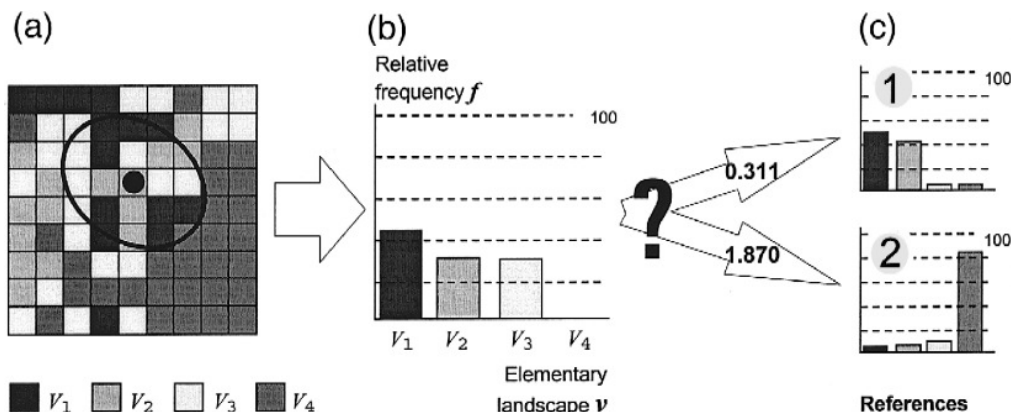


Figure 1. An example of application of the Clapas Procedure.

Then, the soilscape of the black dotted site of Figure 1a is quantitatively described by a cover frequency vector (histogram Figure 1b), which is computed from an elliptic neighbourhood of the site. Finally, the soilscape description of the black dotted site is compared with both soilscape descriptions of reference sites (1 and 2, Figure 1c). In this example, the computed Manhattan distances reveal that the studied soilscape is closer to reference site 1 (distance = 0.311) than reference site 2 (distance = 1.870).

In practice, the Clapas procedure deals with raster cells of a landscape scale image, each cell representing a site of the region. Prior to Clapas processing, the input maps of soil forming factor are thus converted into a grid structure (Lagacherie *et al.* 2001). Preparation and postprocessing of the data are done within a free statistic software (R) and Geographical Information Systems (ArcGisTM).

The region studied and the data

The region studied is situated in the south west of France, at the piedmont plain of the Pyrenees, in the Gers department. It corresponds to a soil map of Mirande at 1:50 000 scale (Figure 2).

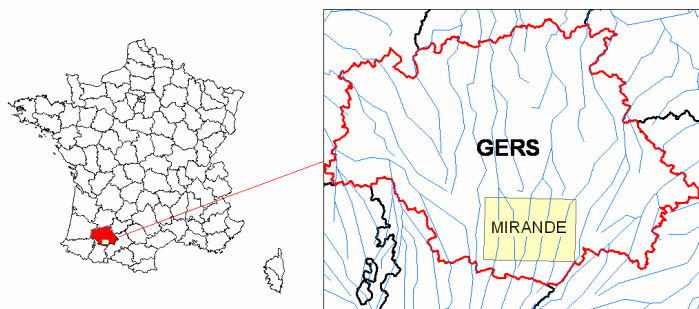


Figure 2. Study area: The map of mirande is located in south west of France, in the Gers department.

Between the dissymmetric valleys, the soft hills are made of a stack of layers corresponding to sedimentary cycles. The rhythmic sedimentation begins with pudding stone or conglomerate, then molasse and limestone

beds covered by a thick layer of marl.

The pedological context is characterized by a general phenomenon of decarbonation and eluviation which leads quickly to a reduction phasis followed by planosolization.

Those conditions have produced soils like Fluvisols, luvic, eutric, calcaric and hypereutric Cambisols, Luvisols and Albeluvisols.

Our prior experience in surveying the region, and statistical tests (Lehmann *et al.* 2007) led us to select three landscape parameters for identifying soil associations: parent material classes, Beven index (Beven *et al.* 1979) and Relief Index. Those two last indices were calculated with ArcGis™ tools from a 50x50 m Digital Elevation Model distributed by the French National Geographic Inventory. The Beven and Relief Index were discretized into 3 and 4 classes respectively. They have been further combined with the parent material map containing 8 classes, into a new image (Figure 2) as required by the procedure for calculating soilscape distances.

The soil map of Mirande was used for running and validating the procedure of soil prediction from the reference area to the rest of the map.

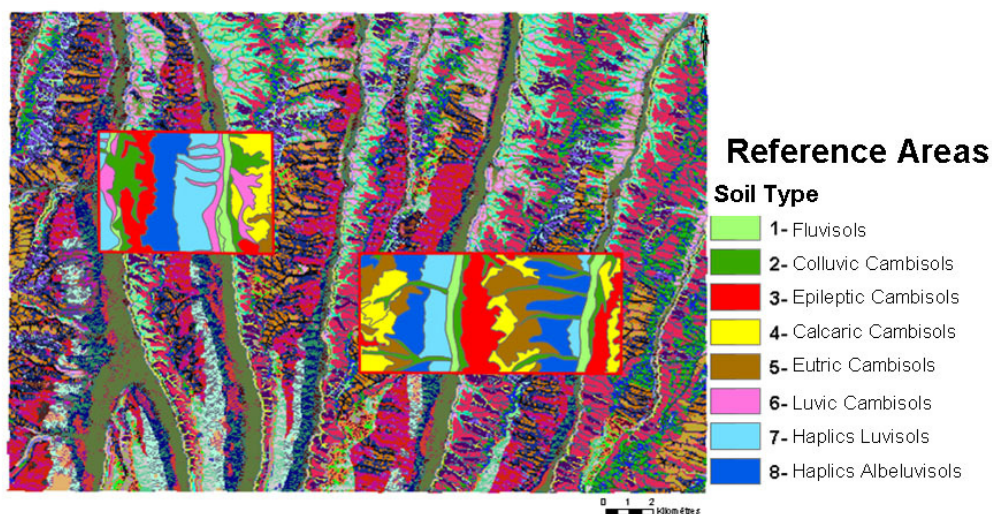


Figure 3. Image of the combined parameters covered by the reference areas.

Results

A simple visual comparison of the simplified soil map with the one generated by Clapas (Figure 4) is enough to judge the strong resemblance of the two maps.

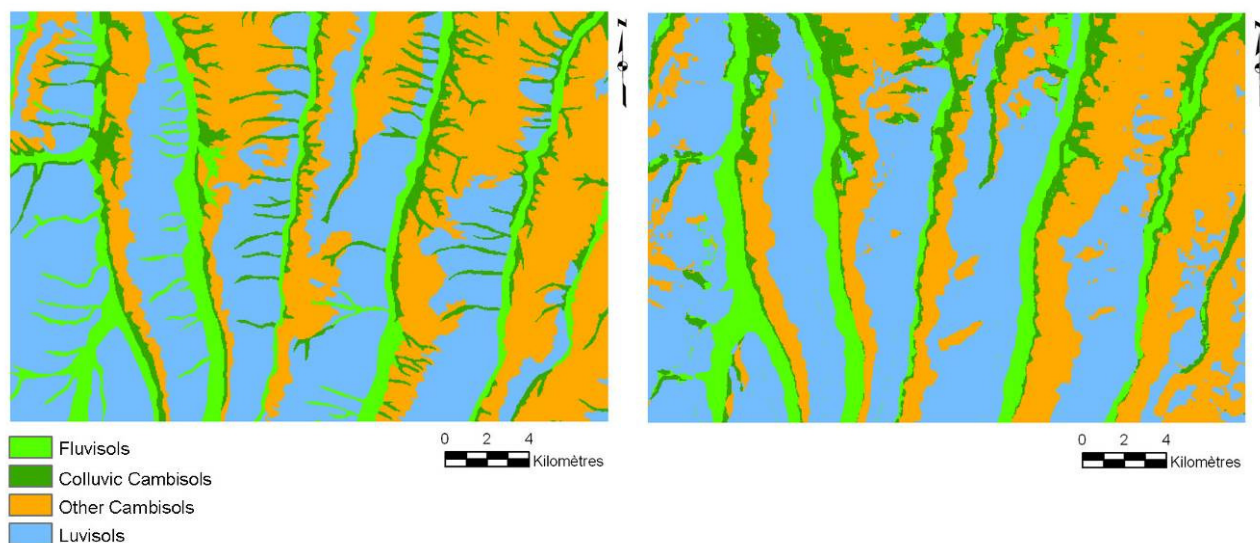


Figure 4. Comparison of the simplified soil map (left), with the one generated by Clapas (right).

The confusion matrix (Table 1) confirms it. The classifier has a 73% global recognition rate. It is also ensured by an acceptable kappa rate of 59%. Looking in detail at the results of the classification, we can see from the matrix that the separation between Colluvic Cambisols and the other Cambisols is difficult. Fluvisols have a 64% recognition rate despite a relative strong confusion with the 4th class: the Luvisols, which represent 1/5th of the prediction for that class. The Luvisols and the Cambisols are quite well predicted with the same rate of 78%.

Predicted	Actual				pixel number	commission error	accuracy of the prediction in %
	1	2	3	4			
1	18175	3135	662	6235	28207	0.36	0.64
2	4214	8181	6868	2094	21357	0.62	0.38
3	875	5996	60949	9877	77697	0.22	0.78
4	4154	3926	14957	81820	104857	0.22	0.78
pixel number	27418	21238	83436	100026	232118	total number of pixels	
omission error	0.34	0.61	0.27	0.18			

Global classification rate	Indice Kappa
0.73	0.59

Table 1. Confusion matrix of the prediction with 4 classes.

Conclusion

The implementation of Clapas in R shows exactly the same results than the results produced with the old version of Clapas (Lehmann 2007). However it has now two new improvements that must be tested in the future and compared with other procedures like Mart Classification (Grinand *et al.* 2008). The first is that the number of combination in the image of entry is no more limited to 255 which enable to create more classes and will likely improve the accuracy of the prediction. The second is that the orientation of the elliptic window that considers the neighbourhood of a pixel can now be automatically oriented according to slope aspect.

References

- Beven K, Kirkby M (1979) A physically-based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin* **24**, 43-69.
- Grinand C, Arrouays D, Laroche B, Martin MP (2008) Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma* **143**, 180-190.
- Lagacherie P, Robbez-Masson JM, Nguyen-The N, Barthès JP (2001) Mapping of reference area representativity using a mathematical soilscape distance. *Geoderma* **101**, 105-118.
- Lehmann S, Bégon JC, Eimberck M, Daroussin J, Wynns R, Arrouays D (2007) Utilisation du logiciel Clapas pour l'aide à la délimitation de pédopaysage. *Etude et gestion des sols* **14**, 135-151.
- McBratney AB, Mendonça-Santos ML, Minasny B (2003) On digital soil Mapping. *Geoderma* **117**, 3-52.
- Robbez-Masson JM (1994) Reconnaissance et délimitation de motifs d'organisation spatiale. Application à la cartographie des pédopaysages. PhD Thesis, Ecole Nationale Supérieure Agronomique de Montpellier.
- Wharton SW (1982) A contextual classification method for recognizing land use patterns in high resolution remotely sensed data. *Pattern Recognition* **15**, 317-324.

Local regression kriging approach for analysing high density data

Wei Sun^{A,B,C}, Budiman Minasny^B and Alex McBratney^B

^AChina Agricultural University, East Campus, Research Center for Precision Agriculture, Qing Hua Dong Lu 17, Beijing, 100083, P.R. China.

^BAustralian Centre for Precision Agriculture, Faculty of Agriculture, Food and Natural Resources, The University of Sydney, NSW 2006, Australia.

^CCorresponding author. Email sunwei1028@126.com

Abstract

This paper demonstrates the development and an application of a local regression-kriging (RK) program. It illustrates the performance of the local RK method for prediction of soil properties using high density crop yield data in time series. We developed a software named RKGuider to carry out a serial steps of RK, especially local RK automatically and using the program, we produce a high resolution soil electrical conductivity map from a coarse survey. The result demonstrated there is spatial correlation between yield data in time series and soil electrical conductivity. High density yield data can be taken as auxiliary variables to predict soil EC using RK.

Key Words

Local regression kriging, variogram model, linear model, yield, electrical conductivity.

Introduction

Regression kriging is a spatial prediction technique which adds the regression value of exhaustive variables and the kriging value of residuals together. The local RK algorithm was developed especially to take into account the local correlation between environmental variables and the unsatisfactory goodness of fit of the spatial variance model for the entire data set. Beside that, local RK is developed to deal with rich data, using the data in full capacity or to improve prediction from a global model. Local RK result is the sum of a local regression of auxiliary variables and local kriging of the regression residuals. This is called RK in a moving window and was considered as the “next step” for RK development (Hengl *et al.* 2007). This algorithm will play more and more important role in geostatistics because lots of covariates are available dramatically now by advancement technology (Pebesma 2006). However there is a serious constraint to wider use of RK, especially local RK, which is that a user must carry out various steps in a variety of software packages (Hengl *et al.* 2007). Furthermore, there is no software that can do local RK efficiently. Therefore it is necessary to develop a framework for spatial interpolation based on local RK.

Methods

Algorithm

Local RK involves the following steps:

- 1) Searching for the closest neighbourhood for each prediction site,
- 2) Fitting a linear regression model predicting the attribute from the covariates from the neighbourhood data,
- 3) Calculating a residual of the regression model for each neighbourhood point,
- 4) Estimating the variogram from the residuals,
- 5) Fitting a variogram model to them
- 6) Predicting the residual value and its uncertainty for each prediction site with kriging,
- 7) Summing up the regression value and the kriging of residuals together, and calculating the uncertainty.

In order to make the local regression-kriging application simpler, a software named RKguider was developed to carry out these steps automatically. It uses automatic linear and nonlinear regression routines, in which a search algorithm was applied for increasing speed. The program also incorporates the maximum likelihood method (Mardia and Marshall 1984) which refines iteratively the prediction of the variogram parameters and linear model regression. The program is an advancement of the Vesper program which only performs kriging with local variogram (Minasny *et al.* 1999). RKguider was developed using Microsoft Visual C++ 6.0 development system. The program offers a friendly interface with a range of options to users to deal with dataset, and provide the flexibility to calculate global RK and local RK. Weighting options could be changed before fitting semivariogram modeling. During processing of local RK, it can provide a real-time graphical

display of the local regression and semivariogram modeling and the searching neighbours could be set as analyst wants. The output is an ASCII text including the prediction point coordinates, the predicted value, the regression variance, the kriging variance, the total variance and the regression parameters. One limit of the program is that the covariates would better be continuous. For binary variables, it can sometimes fail when the local neighborhood only has a single value.

Application and testing

The program was applied to high density crop yield data gathered from yield monitors. The objective is to produce a high resolution soil electrical conductivity map from a coarse survey. Continuous four years yield data are available from 2003 to 2006 in 4 meters resolution, which are taken as the target variables used in linear regression, because high spatial relationship between soil EC and yield has been reported.

Results

Digital maps at a resolution of 4m x 4m for soil ECa were predicted using local kriging and local RK. Linear regression map and residuals map were also presented in Figure 2. Data-out validation are used to compare local kriging with 100 neighbours, local RK with 100, 200 and 300 neighbours separately and linear regression. The data set was divided into interpolation (about 7500) and validation set (about 2500), then R^2 and Standardized squared deviation $\theta(x)$ were taken as index to measure the prediction efficiency and the goodness of theoretical estimates (Table 1)

Table 1. Statistics results.

Prediction Method	Median $\theta(x)$	Mean $\theta(x)$	RMSE	R^2
Linear regression	-	-	9.67	0.251
Local RK with 100 neighbors	0.280	1.097	3.23	0.916
Local RK with 200 neighbors	0.230	0.999	3.13	0.921
Local RK with 300 neighbors	0.226	0.937	3.16	0.920
Local kriging with 100 neighbours	0.700	3.328	2.77	0.939

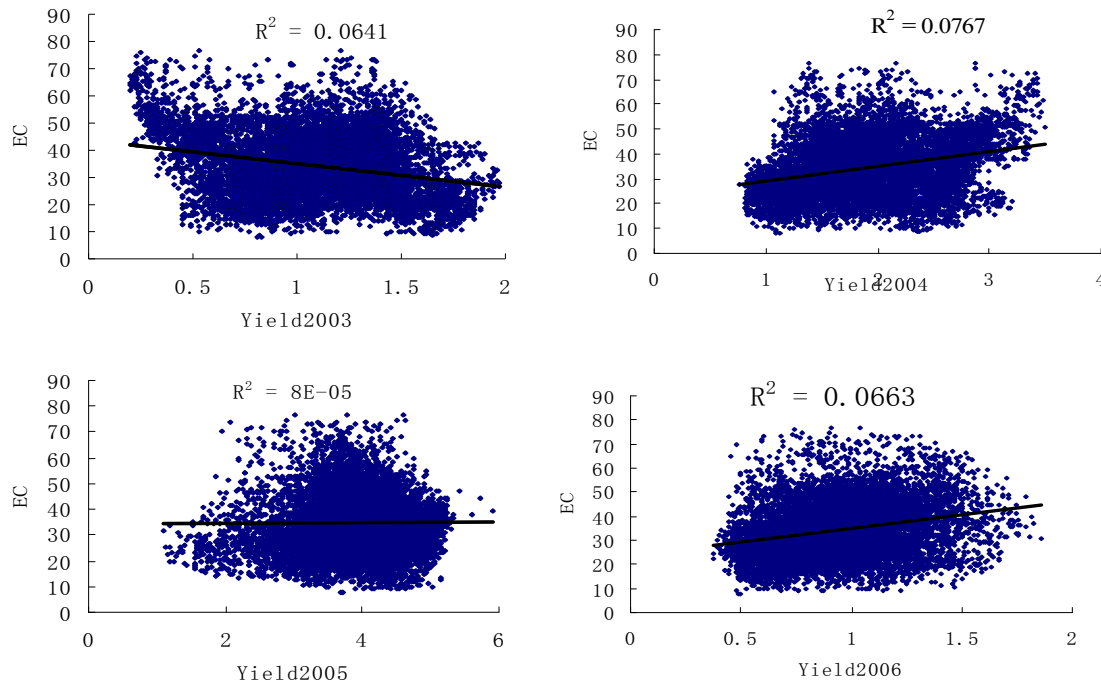


Figure 1. Scatter plot of EC and yield data.

Although the results show that the local RK method does not present much better results than local kriging; however we are able to understand the pattern of soil ECa and its relationship with crop yield. General linear regression (Figure 1) shows that there are variable response between ECa and yield, however the spatial pattern reveal areas where yield for each year has a correlation with soil ECa.

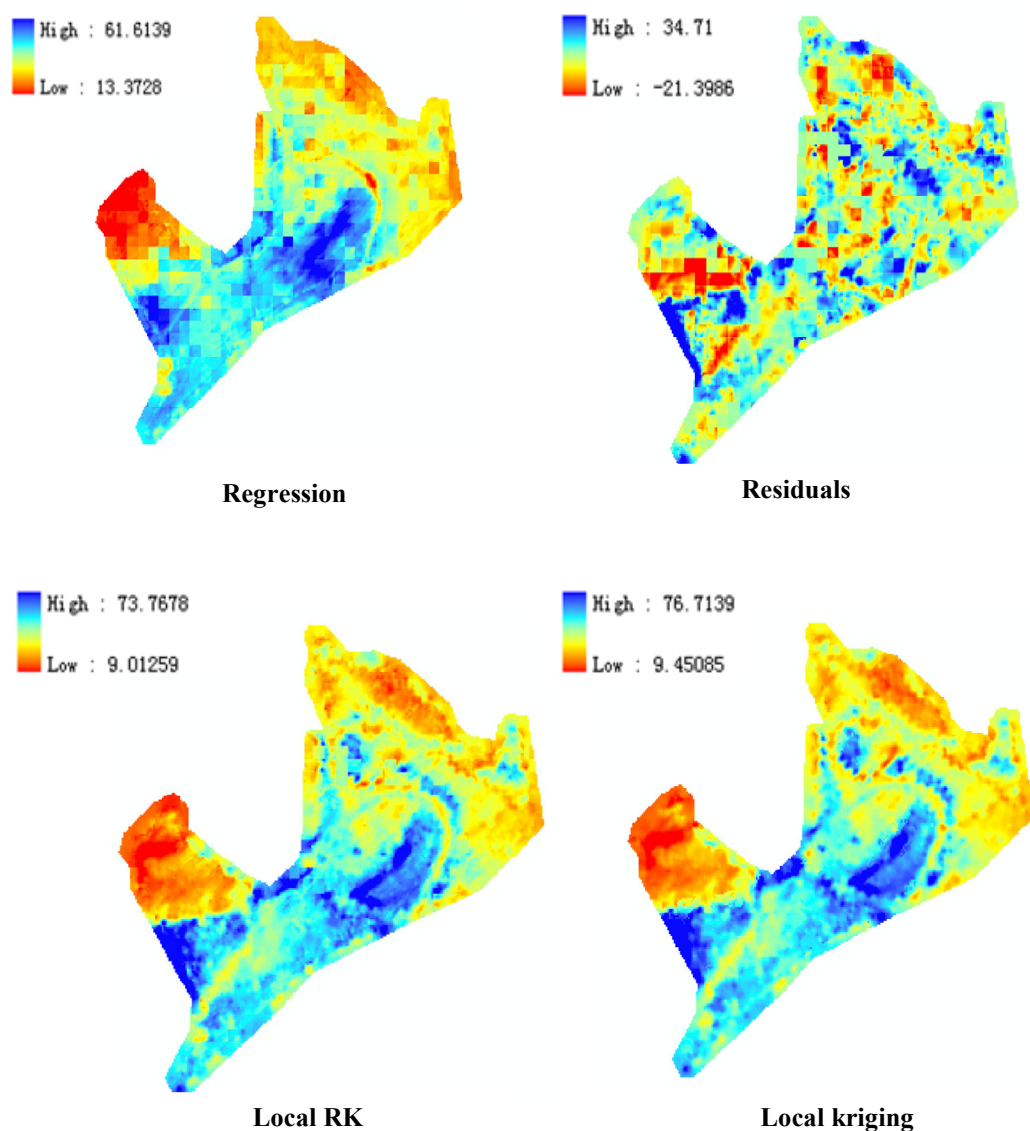


Figure 2. Digital maps at a resolution of 4 m x 4 m for soil ECa obtained using local regression kriging and local kriging.

Conclusion

We have developed a program for conducting local regression kriging. There is correlation between yield data in time series and soil electrical conductivity. High density yield data can be taken as auxiliary variables to predict soil EC using RK.

References

- Mardia KV, Marshall RJ (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135-146.
- Minasny B, McBratney AB, Whelan BM (2002) VESPER, Variogram Estimation and Spatial Prediction plus Error. Australian Center for Precision Agriculture, The University Sydney, NSW 2006, Australia.
- Hengl T, Heuvelink GBM, Rossiter D (2007) About regression-kriging: From equations to case studies. *Computers and Geosciences* **33**, 1301-1315
- Pebesma E (2006) The Role of External Variables and GIS Databases in Geostatistical Analysis. *Transactions in GIS* **10**, 615-632.

Methods for updating the drainage class map in Flanders, Belgium

Johan Van de Wauw^A and Peter Finke^B

^ADepartment of Geology and Soil Science, Universiteit Gent, Krijgslaan 281/S8, B-9000 Gent, Belgium, Email johan.vandewauw@ugent.be

^BDepartment of Geology and Soil Science, Universiteit Gent, Krijgslaan 281/S8, B-9000 Gent, Belgium, Email peter.finke@ugent.be

Abstract

Phreatic groundwater dynamics are one of the most important land characteristics for agriculture, nature development and other land uses. In Belgium, these dynamics are usually estimated from the natural drainage classes, indicated on the Belgian soil map. This information is however partly outdated, due to human intervention (artificial drainage, levelling, groundwater extraction) and –possibly- climate change. Moreover, these morphological classes were not based on actual groundwater measurements. Two groups of methods to update the old map using measured groundwater levels were applied at two locations in Flanders. A first group are 'relabeling' methods. These methods preserve the spatial structure of the old map, but assign new classes to it based on the new groundwater level observations. A second method 'remapping' uses area-wide high-resolution digital auxiliary information to remap the area and create new mapping boundaries. These methods were applied to two different locations in Flanders: the valley of river Dijle (800 ha, south of Leuven) and an area close to the village of Kluizen (300 ha, east of Ghent). Validation shows that remapping provides better results than relabeling methods, although both groups of methods improve the quality of the original map

Key Words

Digital soil mapping, water tables, uncertainty.

Introduction

Phreatic groundwater dynamics are one of the most important land characteristics for agriculture, nature development and other land uses. These dynamics are usually estimated from the natural drainage classes that are indicated on the Belgian soil map (1/20.000), based on data collected during the national soil survey (1947-1971). This natural drainage condition on the soil maps was derived from the depth of gley mottles and a reduction horizon and their position in the landscape. They are indicated using combined classes of the depth of reduction and the depth of mottling (table 1). These morphogenetic features do not always reflect recent changes in the hydrology, and their expression is strongly related to other soil properties like pH and organic carbon content.

Table 1. Original definitions of natural drainage classes in Belgium.

Code	Depth (cm) where mottling and permanent reduction features start			
	Clay/silt textures		Sandy Textures	
	mottling	reduction	Mottling	reduction
a	-	-	-	-
b	>120	-	90-120	-
c	80-120	-	60-90	-
d	50-80	-	40-60	-
e	30-50	>80	20-40	>80
f	0-30	40-80	0-20	40-80
h	30-50	-	20-40	-
i	0-30	-	0-20	-
g	-	<40	-	<40

A common interpretation (ie Van Damme 1969, Boucneau 1996) of these drainage classes as a set of mean highest and mean lowest groundwater asks for a different estimation, based on measured groundwater levels. The mean highest water level and mean lowest water level are defined as the mean value of the three highest and lowest groundwater levels measured biweekly for at least 8 years, preferably longer (30 years) for climate representativeness. Different methodologies have been proposed to update maps of groundwater dynamics using different techniques: relabeling methods (Finke 2000) and remapping methods (Finke *et al.* 2004), but these methods have not yet been applied to and compared within the same area.

Methodology

Study Areas

These methods were applied to two different locations in Flanders: the valley of river Dijle (800 ha, south of Leuven), where a large number of groundwater observations was present in 123 locations, and an area close to the village of Kluizen (300 ha, east of Ghent) where very few observations were present and 100 new observations were sampled.

Time series analysis of long measurement series

Only few locations have sufficient data for derivation of MHW and MLW, and the estimations are based on different climatic periods. However, a larger number of locations has sufficient information for fitting a time series model (von Asmuth *et al.* 2002). This model is calibrated using precipitation surplus and biweekly observations and can be used to expand the measurement serie, and to make it climate representative. Different stochastic simulations of the time series model are used to estimate MHW and MLW, and can be used to derive other parameters such as frequency of exceedence and regime curves.

Well-timed observations (winter-summer)

A number of well timed observations is used to increase the number of points where MHW and MLW can be estimated. In these points, two observations in winter/summer are made of the groundwater depth, when the groundwater is expected to be close to the mean highest/lowest water table. Linear regression between the observations in the long measurement series on the same day as the well-timed observations is used.

Relabeling methods

In relabeling methods (Finke 2000) the point data of MHW and MLW are used to assign new values to polygons of the existing soil map. This was done per polygon and per stratum.

Remapping method

In the remapping method (Finke *et al.* 2004), a set of non-correlated auxiliary variables like DEM, slope, distance (horizontal and vertical) to drainage network and the old drainage class are chosen using Mallows Cp and used with regression kriging to create a new drainage class map.

Validation

A set of 33 (Dijle Valley) and 31 (Kluizen) well timed observations is present to validate the mapping results. Additional observations in 15 locations in Dijle Valley allow a further validation including the conversion of well-timed observations to mean highest/lowest water table.

Results

In both study areas, remapping appears to be the best method, followed by relabeling using one value per stratum.

Table 2. Results of the different map upgrading methods: Remapping is the best method, followed by relabeling by stratum.

	Kluizen (MAE in cm)		Dijle (MAE in cm)	
	MHW	MLW	MHW	MLW
Old	8.8	9.1	9.6	17.7
Relabeling: polygonwise	3.3	4.8	6.9	10.9
Relabeling: Per stratum	3.0	5.5	5.0	6.7
Remapping	1.1	3.8	1.3	3.4

References

- Boucneau G, Van Meirvenne M, Desmet J, Hofman G (1996) A methodology to evaluate the reliability of the Belgian soil map for predicting the actual water table characteristics. *Geoderma* **69**, 193-207.
- Finke PA (2000) Updating the (1 : 50,000) Dutch groundwater table class map by statistical methods: an analysis of quality versus cost. *Geoderma* **97**, 329-350
- Finke PA, Brus DJ, Bierkens MFP, Hoogland T, Knotters M, de Vries F (2004) Mapping groundwater dynamics using multiple sources of exhaustive high resolution data. *Geoderma* **123**, 23-39.

- Vandamme J, De Leenheer L (1969) Fluctuations of the ground water tabel over a period of 5 years (1963-68) in the sandy soils of the Campine in the province of Antwerp (Belgium) (in French). *Pedologie* **19**, 275-320.
- von Asmuth JR, Bierkens MFP, Maas K (2002) Transfer function-noise modeling in continuous time using predefined impulse response functions. *Water Resources Research* **38**, doi:10.1029/2001WR001136

Predictive soil mapping as a means to aggregate and improve existing soil databases using classification trees and knowledge integration

Jan Willer^A, Rainer Baritz^B, Einar Eberhardt^B and Reinhold Jahn^A

^AMartin Luther University of Halle-Wittenberg, Halle-Wittenberg, Germany, Email Jan.Willer@bgr.de

^BFederal Institute for Geosciences and Natural Resources, Hannover, Germany, Email Rainer.Baritz@bgr.de

Abstract

A methodology using digital soil mapping has been developed to improve the reconnaissance mapping 1:200.000. It combines soil data from different sources in order to develop a seamless, nation-wide, consistent soil geometric and semantic database. The use of legacy data involves some challenges coming from differing datasets, soil descriptions, mapping strategies and data gaps which are typical for mapping campaigns which last over few decades. These problems have to be solved by applying and developing extensive semantic harmonization and quality control procedures.

Digital soil mapping techniques are considered to be a powerful tool to harmonise data from different data sources, filling gaps in existing soil maps and cross-validation. The model and methodical pathway presented here has been developed as a hybrid approach, combining (a) classification tree analysis of existing soil maps and auxiliary data (elevation models, geological maps, climatic data, etc.) as regionalisation method for discrete soil classes, and (b) the integration of knowledge about regional soil forming processes through expert-based rules. As the final product, predictive conceptual soil maps are generated. The methodology can be used to harmonize soil data from different sources and to speed up the mapping process in areas with a lack of soil mapping data.

Key Words

Digital soil mapping, classification tree, Soil data harmonizing, Project SIAM.

Introduction

The project SIAM (Soil Inference and Mapping Project) aims at putting together a set of digital soil mapping techniques that allow an integration of soil and auxiliary data from different sources and scale to form a consistent level of soil information at a fixed target scale (here: 1:200.000-1:250.000, Germany). The resulting methodology is then applied to gap filling, quality checking and harmonizing existing reconnaissance mapping sheets throughout the study area (Germany, stratified into soilscales). Even though the national soil project builds on a harmonized assessment schemes, different data sources were used in different parts of the country. In some areas, digital high-resolution mapping data are only quite patchy and not area covering. These differences and data gaps print through into the individual sheets 1:200.000 in different ways. Since qualitative experiences and legacy data sources are to be utilized as much as possible, a hybrid approach was selected. If successfully calibrated, the system is also able to predict conceptual maps in areas not covered with such legacy data.

Methods

The approach presented here combines the analysis of existing soil maps with classification trees and predictive modelling with knowledge integration. Classification trees are an established method in digital soil mapping approaches (Behrens and Scholten 2007), and in comparison with other data mining techniques, the derived models permit a better interpretability by soil scientists. As a pilot area, the map sheet Cologne with a total area of 6400² was used (Figure 1). Inside this test sheet, a training area for calibration was selected based on the quality of 1:50,000 mapping and data density (Figure 2), which is representative for most parts of the Rhenish Slate Mountains. Classification trees were used to extract a concept model from the pilot area soil map. The rules of the classification tree were put into the software SolimSolution (Zuh *et al.* 2004) for recording and modifying the model by knowledge integration and producing predictive soil maps (Figure 3).

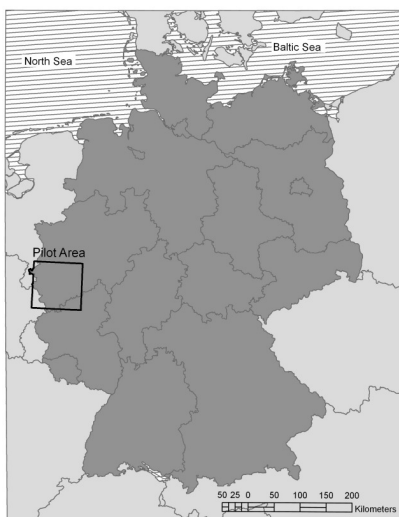


Figure 1. Location of the Pilot Area

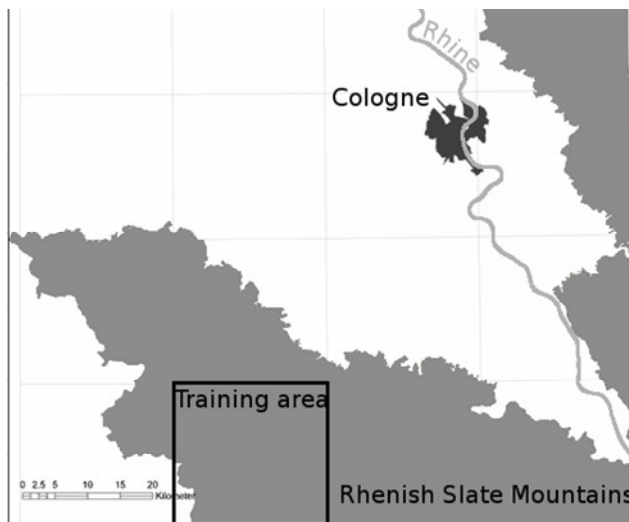


Figure2: Location of the training area within the Pilot area

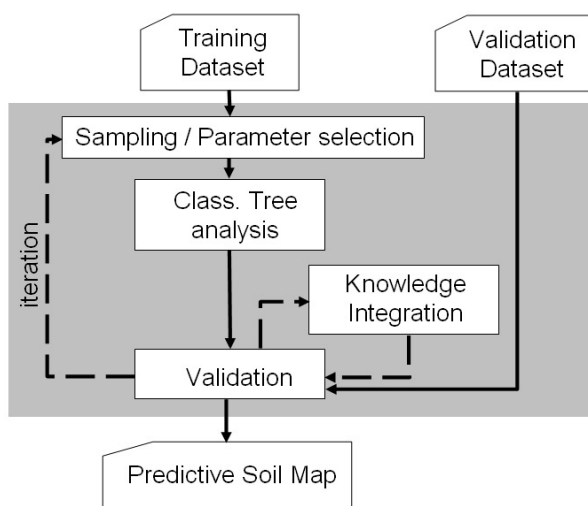


Figure 3. Overview of method

Data selection and preparation

Selection of the data is a crucial step within the model development. Accuracy of the results depends on the accuracy of the input data, but as well on the parameter selection appropriate for the target scale. At a scale of 1:200k, environmental factors relevant to describe the general spatial patterns of soil associations differ from larger scales. As we use data from various scales and sources, reduction of noise originating from previous generalisation steps is essential. Soil information was derived from the soil map 1:50,000 of North Rhine-Westphalia. The legend units of this soil map describe discrete associations of soil units that combine genetic soil types with parent material classes. For the target scale, these map units need to be aggregated and spatially generalised to form even more complex units.

Classification Tree Analysis

For the Classification Tree Analysis, we used the Software GUIDE (Loh 2008, 2009). It has some advantages over the better known CART algorithm (Breimann *et al.* 1983), as it allows unbiased split selection at each node and also interaction detection within the nodes. For the split selection we used the simple linear model. More complex models using the kernel or nearest-neighbour method did not significantly improve the results while making the tree more complex and less interpretable.

Data selection included four main steps:

- (1) Soil scapes: delineation of mountainous regions with similar geological environment
Cluster analysis for grid datasets (SAGA Software package) has combined the following data:
 - standard deviation of altitude (250 m radius)
 - relative altitude (50 km radius)
 - geological units (geolog. map 1:100k)
- (2) Apriori definition of target map units
For orientation, map units of the soil map 1:50,000 were aggregated to a target scale of 1:200k by the geological service. The aggregation was based on soil morphogenetic types, depth to bedrock, texture, base saturation and parent material. The 34 Units of the map 1:50,000 in the test area were aggregated into 14 units suitable for the 1:200k map.
- (3) Development of ancillary data base
 - Terrain analysis using a local DEM 25 m (SAGA GIS, Köthe 2006)
 - Geological map 1:100kClimate data showed no significant correlation with the soil distribution, but tended to produce artificial over-fitting in the classification tree analysis. The same held true for land use data derived from satellite images.
- (4) Noise reduction
 - a: DEM filtering:
 - Gaussian filter to achieve a moderate smoothing of the DGM
 - multidirectional filter (Selige *et al.* 2006)
 - b: Buffer functions to settle misfits between the soil map and geologic map polygons with their respective topographies and degree of generalisation
 - c: Identification of outlier soil polygons following the approach of (Qi 2006).

Knowledge Integration

The rules obtained during the classification tree analysis transferred to the inference modelling software SolimSolution (Zhu *et al.* 2001) in order to explicitly record the classification rules. In a subsequent step, the model was further modified, based on regional expert knowledge or by means of other statistical approaches, because fuzzy membership functions can deliberately be changed by the user. For each pixel of the predictive map, the rules applied and the decisive parameter for its assignment to a soil class can be extracted. Classification misfits can be traced back to the deciding rule, even if highly complex classification rules are used. In this way, it is possible to integrate expert knowledge or adapt it with complementary datasets. In the current state of the model, we use a knowledge-based rule for one soil association (dominated by Haplic Luvisols, see Figure 4) that is underrepresented in the training area.

Validation

As a validation, we compared the model results with existing soil maps in the training area and in a separate validation area (neighbouring sheet 1:50,000) with similar environmental conditions (Figure 4). User's, producer's and overall accuracy were calculated following Congalton and Green (1999). Validation with auxiliary data is planned but not concluded yet.

Results and Conclusions

We were able to predict all 14 apriori map units (Figure 4) that were defined in the training area with an overall accuracy of 0.53 in the training area and 0.51 in the validation area, as compared to the existing soil map 1:50,000. The model developed in the training area shows visually plausible results for a large area of the Rhineland Slate Mountains with similar geological setting. Hence, our approach provided an independent tool for comparing soil maps from different origin and scale. Due to the data selection and the parameter settings, a reasonable generalization for our target scale of 1:200k of the soil map has been achieved. For certain areas, a lack of input data, e.g. incomplete data on the parent material or historical land use, produced hardly assessable inaccuracies. As a conclusion, the comparative low congruence with the existing soil map is not only a result of model inaccuracy, but can also be attributed to the required level of generalization.

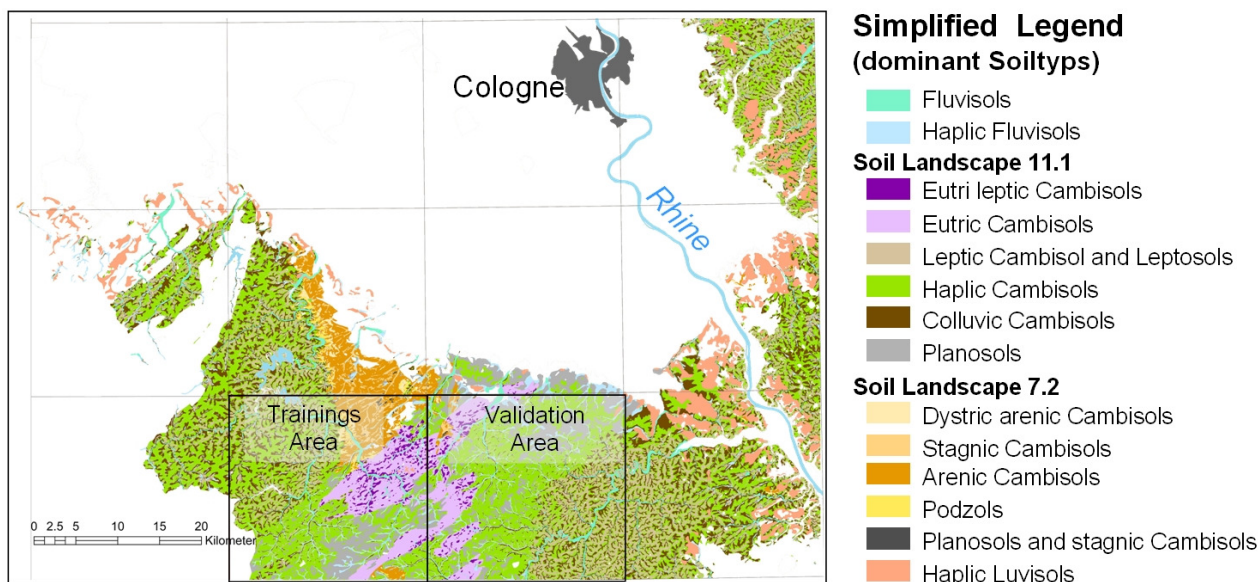


Figure 4. Predictive soil map for the Rhenish Slate Mountains

Outlook

For further improvement of the model, more accurate parent material information shall be obtained by a combined approach of relief analysis, parent material information from the soil map and spatial distance to the respective source rock units of the geological map. Knowledge integration should be improved, in a way that allows uncertainty assessment, as it is considered to be an important advantage over models using only geostatistical methods. Quantifying the effect of the generalisation on the model accuracy shall be done by a systematic comparison with unpruned trees highly adapted to the training area.

References

- Behrens T, Scholten, T (2007) A comparison of data-mining techniques in predictive soil mapping. In 'Digital soil mapping. An introductory perspective'. (Eds P Lagacherie, AB McBratney, M Voltz) pp. 353-364. vol 31. (Elsevier).
- Breiman L, Friedman JH, Olshen RA, Stone CJ (1983) CART: Classification and Regression Trees. Wadsworth: Belmont, CA.
- Congalton RG, Green K (1999) Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Lewis Publishers, pp. 45 – 47. (Boca Raton).
- Köthe R, Bock M (2006) Development and use in practice of SAGA modules for high quality analysis of geodata. *Göttinger Geographische Abhandlungen*, **115**, 85-96.
- Loh W (2008) Classification and Regression Tree Methods. In 'Encyclopedia of Statistics in Quality and Reliability' (Eds Ruggeri, Kenett, Faltin) pp. 315-323. (Wiley).
- Loh, W (2009) Improving the precision of classification trees. *Annals of Applied Statistics* **3**, 13 pp.
- Lee J (1980) Digital image enhancement and noise filtering by use of local statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2**, 165-168.
- Schmidt K, Behrens T, Scholten T (2008) Instance selection and classification tree analysis for large spatial datasets in digital soil mapping. *Geoderma*, **146**, 138-146.
- Selige T, Böhner J, Bock M (2006) Processing of SRTM X-SAR Data to correct interferometric elevation models for land surface process applications. *Göttinger Geographische Abhandlungen*, **115**, 97-104.
- Qi F (2004) Knowledge Discovery from Area-Class Resource Maps: Data Preprocessing for Noise Reduction. *Transactions in GIS* **8**, 297-308.
- Zhu AX, Hudson B, Burt V, Lubich K, Simonson D (2001) Soil Mapping Using GIS, Expert Knowledge, and Fuzzy Logic. *Soil sci. Soc. Am. J.* **65**, 1463-1472.
- Zhu A, Moore AC, Smith MP, Liu J, Burt J, Qi F, Simonson D, Hempel J, Lubich K (2004) Advances In Information Technology For Soil Surveys: The Solim Effort. In 'Innovative techniques in soil survey: "Developing the foundation for a new generation of soil resource inventories and their utilization"' (Eds Eswaran H, Vijarnsorn P, Vearasilp T, Padmanabhan E) pp. 25-42.

Quantifying the uncertainty in digital soil class maps developed using model-based approaches

M. A. Nelson^{A,C}, I. O. A. Odeh^A, T. F. A. Bishop^A and N. Weber^B

^AFaculty of Agriculture, Food and Natural Resources, The University of Sydney, Sydney, NSW, Australia.

^BSchool of Mathematics and Statistics, The University of Sydney, Sydney, NSW, Australia.

^CCorresponding author. Email michael.n@usyd.edu.au

Abstract

Digital soil assessments use digital soil class maps as inputs for modelling soil processes. The accuracy of the digital soil class maps is required for the outputs digital soil assessments to be most useful. Few studies have considered the quantification of uncertainty in digital soil class maps. The use of model based approaches means the uncertainty in parameter estimates may be quantified. This study considers the effect of this model parameter uncertainty on digital soil class map, and provides a methodology to perform conditional simulation for a digital soil class map when model based approaches are used. We consider a multinomial logistic regression model and a Poisson multinomial generalised linear spatial model. The results of both the logistic regression and generalised linear spatial models show uncertainty is greatest near the boundary of soil classes. The simulated soil class maps can be used as input for digital soil assessments. The computational burden of deriving the required Fisher information matrices from the generalized linear spatial models is high, with further study required into methods of approximation.

Key Words

Geostatistics, accuracy, maximum likelihood, Monte Carlo Markov Chain.

Introduction

Digital soil class maps provide an important source of soil information from which harder to measure soil properties can be inferred. As such digital soil class maps form an essential component of the *soil attribute space inference system* thus forms the foundations of the digital soil assessment framework (Carrè *et al.* 2007). A significant advantage of the digital soil mapping framework is that predictions made should be accompanied by estimates of associated uncertainty (McBratney *et al.* 2003). For digital soil assessment, both the predictions contained within a digital soil map, and the associated uncertainty, are vital (Carrè *et al.* 2007).

The four main sources of uncertainty in a digital soil maps are input, model, positional and analytical (Bishop *et al.* under review). In practice, the enumeration and quantification of the uncertainty has lagged behind the development of digital soil mapping techniques, in particular for digital soil class mapping. While numerous papers have considered each of the four main sources of uncertainty four soil property mapping and Bishop *et al.* (under review) have considered relative contributions when all sources are considered simultaneously, few papers have investigated the uncertainty associated with digital soil class mapping predictions.

Model-based geostatistics (Diggle *et al.* 1998) provides a unified generic approach that merges geostatistical methods with generalized linear mixed models (GLMMs). The generalized linear spatial model (GLSM, Diggle *et al.* 1998; Zhang 2002) is a GLMM where the random effects are spatially correlated. Nelson *et al.* (2009) developed a GLSM for digital soil class mapping. The use of model-based geostatistics for digital soil class mapping allows the evaluation of prediction variance whilst also providing estimates of parameter variance which can be used to quantify uncertainty from the model parameters (Dowd and Pardo-Igúzquiza 2002).

The assessment of model uncertainty for a digital soil map has usually taken the form of conditional simulation by sequential Gaussian simulation given a set of model parameters (Goovaerts 1997), in this study, however, we will use the alternative approach suggested by Dowd and Pardo-Igúzquiza (2002). We investigate the effect of parameter estimate uncertainty on predictions in a digital soil class map developed using model-based approaches – both a Poisson multinomial generalized linear spatial model and a multinomial logistic regression model.

Methods

Parameter estimation uncertainty under model-based approaches

Under a model-based approach to digital soil class mapping, inference by maximum likelihood allows the estimation of the uncertainty in the model parameters. The main premise of this approach is the asymptotic multivariate normality of the parameter estimates, whose variance-covariance matrix is approximated by the Fisher information matrix (Dowd and Pardo-Igúzquiza 2002, McCulloch and Searle 2001)

$$\theta: AN\left(\hat{\theta}, I(\hat{\theta})^{-1}\right)$$

$$I(\theta) = -E\left[\frac{\partial^2 \ln L_c}{\partial \theta \partial \theta'}\right] \quad (1)$$

where θ are the model parameters and $\ln L_c$ is the likelihood function. For generalized linear models (GLMs) and GLMMs the estimation of the Fisher information matrix (Equation 1) is computationally burdensome, however numerical methods can be used. For a GLSM, the variance parameters and trend parameters are independent, thus the information matrix is block diagonal (Zhang 2002).

By taking numerous realizations of the model parameters given these distributional assumptions, simulations of a digital soil class map can be produced.

Study area

As a case study, we map a small area of the Upper Namoi river catchment in NSW, Australia. The soil data is from a small section of the Curlewis Soil-Landscape Map (Banks 1994) with four mapping units present. Four environmental covariates were considered for use in the predictive model, SRTM DEM, γ radiometric %K, NDVI and the multi-resolution valley bottom flatness index (MRVBF, Gallant and Dowling 2003).

Model specifications

For the multinomial logistic regression model, a stepwise procedure using Akaike Information Criteria (AIC) was used to select the most parsimonious model, for which the observed information matrix for the parameters was estimated, and 200 realizations drawn. The Poisson multinomial GLSM models each soil class as an independent Poisson variable, with spatial random effects introduced for each soil class (Nelson *et al.* 2009). As with the multinomial logistic regression model, the observed information matrix was approximated for the AIC selected most parsimonious model, and from which 50 realizations of the parameters were drawn. We consider the effect of this uncertainty on the most likely soil class at each location on a digital soil class map. All statistical analyses were performed in R (R Development Core Team 2008), with the GLSM implemented using the geoRglm package (Christensen and Ribeiro 2002).

Results

Multinomial logistic regression

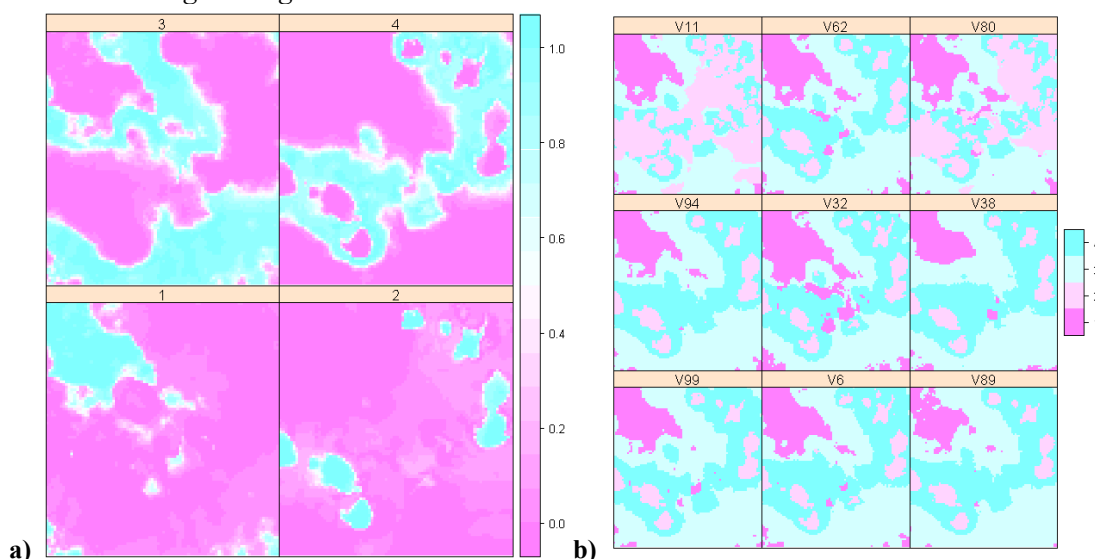


Figure 1. a) Proportion of simulations at each prediction location where each soil class is most likely for a digital soil class map produced using a multinomial logistic regression model; b) Nine simulations of the digital soil class map using realizations drawn from the estimated distribution of the parameter estimates.

Figure 1 (a) shows the effect of parameter estimate uncertainty on the most likely soil class at each location of the digital soil class map. This shows the greatest effect nearest the boundary of the soil classes. Figure 1 (b) shows 9 simulations of the digital soil class map using different realizations of the parameter estimates. The main feature of Figure 1 (b) is the variation in the extents of classes 2 and 4 in the middle section of the map, highlighting the difficulty in differentiating between these classes using this model, and the uncertainty in the resultant digital soil class map.

Poisson multinomial GLSM

Figure 2 (a) shows the effect of parameter estimate uncertainty on the most likely soil class at each location of the digital soil class map. The main feature of this is the lack of prediction of class 1. Figure 2 (b) shows 9 simulations of the digital soil class map using different realizations of the parameter estimates. Again the lack of class 1 predictions is evident. These are a result of the difficulty in approximating the Fisher information matrix under a GLSM. Figures 1 (a) and 2(a) show similar results for all classes except class 1.

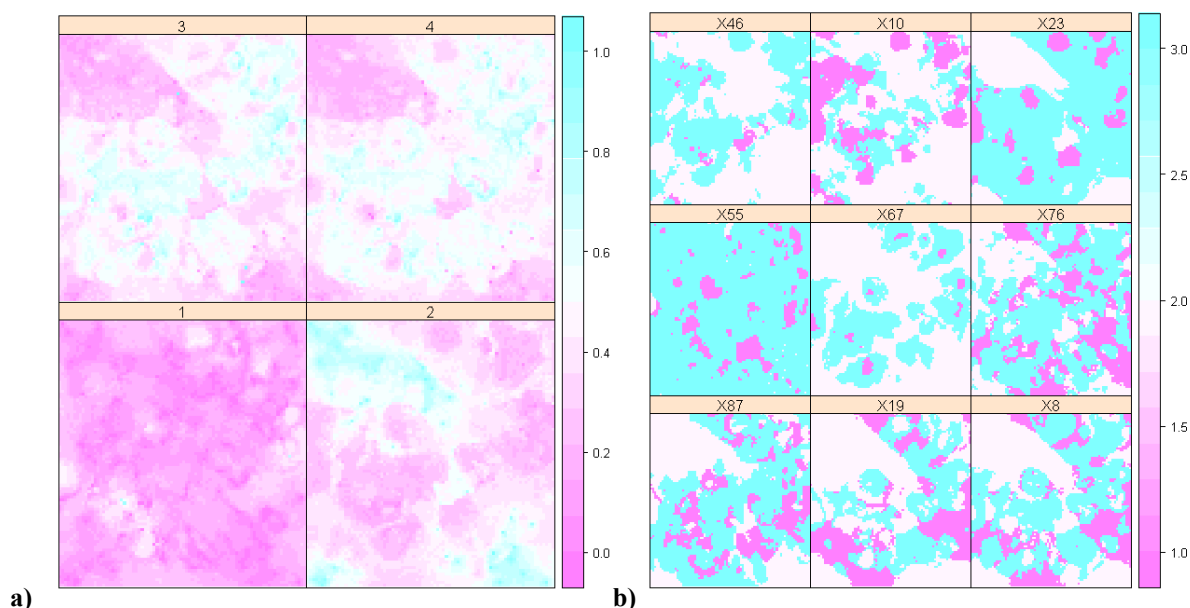


Figure 1. a) Proportion of simulations at each prediction location where each soil class is most likely for a digital soil class map produced using Poisson multinomial GLSM; b) Nine simulations of the digital soil class map using realizations drawn from the estimated distribution of the parameter estimates.

Discussion and Conclusion

The distributional assumptions of parameter estimates are valid as large sample asymptotes (McCulloch and Searle 2001). The sample size for this study (100) may not be large enough. Parameter estimation and the derivation of the Fisher information matrix under a GLMM are not trivial computations. Further study on the various approaches and methods is required to find a method that is both accurate and computationally reasonable. We have shown that for a digital soil class map produced using model based approaches, conditional simulation using the uncertainty in the model parameter estimation provides a useful assessment of the map uncertainty that may be used in digital soil assessments.

References

- Banks R (1994b) 'Soil Landscapes of the Curlew 1:100 000 Sheet: Map'. (Department of Conservation and Land Management: Sydney).
- Bishop TFA, Nelson MA, Triantifilis J, Odeh IOA (2010) Enumeration of all sources of uncertainty in a digital soil map. *European Journal of Soil Science* (under review)
- Carré F, McBratney AB, Mayr T, Montanarella L (2007) Digital soil assessments: Beyond DSM. *Geoderma* **142**, 69 - 79.
- Christensen OF, Ribeiro Jr PJ (2002) geoRglm: A package for generalized linear spatial models. *R-News* **2**, 26-28.
- Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics. *Journal of the Royal Statistical Society Series C-Applied Statistics* **47**, 299-326.
- Dowd PA, Pardo Igúzquiza E (2002) The Incorporation of Model Uncertainty in Geostatistical Simulation.

- Geographical and Environmental Modelling* **6**, 147-169.
- Gallant JC, Dowling TI (2003) A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research* **39**, ESG41-ESG413.
- Goovaerts P (1997) 'Geostatistics for natural resource evaluation'. (Oxford University Press: New York).
- McBratney AB, Santos MLM, Minasny B (2003) On digital soil mapping. *Geoderma* **117**, 3- 52.
- McCulloch CE, Searle SR (2001) 'Generalized, linear and mixed models' (John Wiley and Sons: New York)
- Nelson MA, Bishop TFA, Odeh IOA, Weber N (2009) A generalized linear spatial model for digital soil class mapping [Digital soil class mapping using model-based geostatistics]. In 'Proceedings of Pedometrics 2009 Conference, Beijing, 60-61'.
- R Development Core Team (2008) 'R: A Language and Environment for Statistical Computing'. (R Foundation for Statistical Computing).
- Zhang H (2002) On estimation and prediction for spatial generalized linear mixed models. *Biometrics* **58**, 129-136.

Soil suitability and crop versatility assessment using fuzzy analysis at a farm scale

Inakwu O.A. Odeh and Mark Crawford

Faculty of Agriculture, Food & Natural Resources, The University of Sydney, Sydney NSW 2006, Email i.odeh@usyd.edu.au

Abstract

There is the need to explore simultaneous soil suitability analysis for several viable crops to determine the overall suitability or versatility of a given area or a farm. The main objectives of this paper are to: fit fuzzy membership functions (FMFs) to determine soil suitability for multiple crops; determine the diversity of multiple suitability at each location using the Shannon's Index and equitability equation; and determine the overall versatility at the farm scale, and establish a new versatility equation by combining the previous versatility methodology with a diversity index. To understand the overall versatility of the study area suitability analysis was carried out for each of the several crops- canola, barley, field pea, Lucerne and wheat. The multiple suitability analysis demonstrated subtle differences in the trend or patterns of the individual crop maps. The development of the improved versatility map, incorporating Shannon's index, yielded important information for management decisions. The results indicate that the Northern and Southern paddocks of the study area exhibited higher versatility than the rest of the farm and would be highly suited for multiple crop rotations. Another pertinent point is that the areas of low versatility could be studied further to determine which of the crops are better suited to the soil.

Key Words

Soil suitability analysis; soil versatility; crop suitability, soil evaluation, land evaluation.

Introduction

With fluctuating rainfall patterns or higher frequency of dry periods, efficient land utilisation for agricultural production systems is required for the survival of most farms in Australia. Therefore land evaluation techniques and their resulting soil or land suitability maps must address the economic viability and provide information for management decisions at field or farm scale. Modern precision agricultural practices require across-farm and/or within-field soil variability which should be accounted for in the suitability assessment for it to be an effective tool for management decisions. Soil function under a number of land uses should also be assessed as it provides different options to the farmer and thus reduces the farmer's dependency on a single land use. There is therefore the need to explore simultaneous soil suitability analysis for several viable crops to determine the overall suitability or versatility. The main objectives of this paper are to i) fit fuzzy membership functions (FMFs) to determine soil suitability for multiple crops; ii) determine the diversity of multiple suitability at each location using the Shannon's Index and equitability equation; iii) determine the overall versatility at the farm scale, and establish a new versatility equation by combining the previous versatility methodology with a diversity index.

Methods

Derivation of soil suitability for crops

The study area is located in the southern region of New South Wales (NSW), Australia, within the Riverina bioregion in SW NSW. It is about 20kms NW of Corowa, which is situated on the Murray River. In deriving the soil suitability for a number of crops a set of rules and suitability scores were based on range of values of each soil quality indicator for the different crop varieties (Table 4.1). Literature and expert opinions on each crop were the main guides for devising the rules. The crops considered in this study were wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), canola (*Brassicaceae* family), peas (*Pisum sativum*), and lucerne (*Medicago sativa*) were predominantly differentially sensitive to soil pH ranges and salt tolerance before loss of yield. They were included in the versatility analysis as they represented the range of crops grown in the district. A few of the soil quality indicators such as CEC, ESP, OC and Ca/Mg ratio, are broadly non-limiting to cropping production across regions.

Creation of continuous soil suitability map for multiple crops for the derivation of soil versatility

In creating the continuous soil suitability maps for each crop, we followed the more efficient procedure of soil suitability analysis prior to interpolation (Figure 1). We then used the resulting crop suitability maps to derive the overall versatility for the multiple crops at each grid location of the continuous maps.



Figure 1. Flow diagram illustrating the method used for determining the suitability maps for individual crop.

In previous studies (Triantafilis and McBratney 1993; Triantafilis *et al.* 2001) versatility was derived through some form of aggregation of soil suitability for multiple crops were derived by arithmetic mean. However, this approach can be limiting because soil versatility is not efficiently accounted for. It also lacks the ability to take into consideration the interaction among the suitability criteria, i.e., the soil quality indicators. By incorporating the diversity measure the overarching interactions among different uses reinforce the overall versatility. For example when diversity measure is incorporated into versatility analysis based on suitability at two points for each crop are 0.9, 0.8, 0.7, 0.6, 0.5, and 0.7, 0.7, 0.7, 0.7, 0.7 respectively both would have a mean of 0.7 but are characterised by vastly different diversity in terms of suitability for different uses. This may not be appropriate for matching management decisions, as the full understanding versatility is not accounted for. For this reason we combined the classical arithmetic mean versatility (Eq 1) and diversity index or Shannon Index ((Eq 2) at each grid point to create a new versatility.

$$Z = \frac{X_i}{\sum_{i=1}^n X_i} \quad (1)$$

where Z is the standardized proportion score for aggregated suitability for each individual crop, X_i is the suitability score of each individual crop;

$$H = -\sum_{i=1}^n z_i * \ln z_i \quad (2)$$

where H is Shannon's diversity index, n is the total number of crops, z_i is the proportion of n made up of the i th crops.

Results

Overall the patterns of the soil suitability maps for each crop are quite similar, which was to be expected, as the farm lies in the Wheat Belt of Eastern Australia, mostly invariably suitable for most grain crops.

However, there are subtle differences among the various crops. By focusing on a paddock in Figure 2 the subtle differences are obvious. This is demonstrated in the Middle Eastern section of the Figure 2a which map depicts a higher range of suitability for Lucerne than that of Field Pea (Figure 2b), even though the differences are smaller for Lucerne than Field Pea, with more area falling within the 0.78-0.84. Other areas depicting subtle differences in suitability range was the Northern section of the paddock with Lucerne again showing higher values. These subtle differences are further reinforced by statistical analysis. Table 1 presents the overall mean, maximum, minimum and standard deviation of suitability grades for each crop.

Table 1. Statistical breakdown of suitability grades for each crop

Crop	Max	Min	Mean	S.D
Wheat	0.97	0.64	0.83	0.06
Canola	0.95	0.60	0.81	0.06
Barley	0.99	0.59	0.82	0.07
Field Pea	0.98	0.61	0.82	0.06
Lucerne	0.97	0.62	0.83	0.05

The statistical parameters are remarkably similar, indicating little difference among the suitability for the selected crops. However, what the statistical breakdown demonstrates is that all the crops selected for suitability analysis were found to be highly amenable to the soils in the farm with suitability mean values ranging from 0.81-0.83. Even though the statistical analysis does not pick up the subtle differences in the spatial patterns of suitability values (Figure 2), even though the methodology developed could be part of a larger scale analysis such as regional or national to provide a sufficient base for managerial decisions.

Another key objective of this paper was to compare classical versatility analysis reported by Triantafilis and McBratney (1993) and Triantafilis *et al.* (2001) at the farm scale. The result is illustrated in Figure 3a. As this versatility was derived by the summation of the mean suitability values for the five crops, we would

expect the versatility to mirror the suitability maps. As such, and like the suitability maps for each crop, the map shows relatively low variability across the study area. Thus the spatial variation in the versatility map epitomizes those of individual suitability maps.

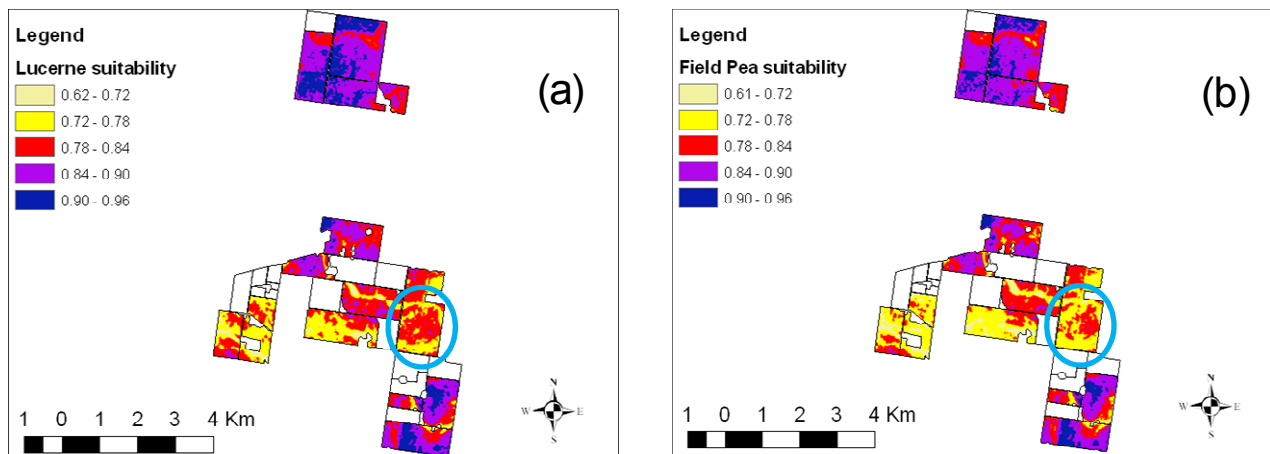


Figure 2. Suitability map for a) Lucerne and b) Field Pea (circle paddock illustrates where subtle differences)

An obvious finding is that relatively high versatility coincides with incidence of large suitability for most of the crops, and vice versa. This is illustrated by the areas with high versatility values >0.83 located in the Northern and Southern paddocks and the mid to low versatility values depicted in the middle paddocks (compare Figure 3b). We opined that this approach does not wholly account for versatility and diversity of suitability for different crops at a given location with respect to adaptability to a variety of crops. Therefore to improve the output and to account for the diversity of suitability, the Shannon diversity index was introduced and the results are discussed in the next section.

As the diversity index was derived using a function equivalent to the one used for species (bio)diversity analysis it is prudent to interpret the results carefully. Firstly, traditionally Shannon's index accounts for both different species and their evenness within a community. It must also be noted that traditionally the results of diversity indices are compared with other communities; for example, diversity of species present in rainforests and grasslands communities could be compared.

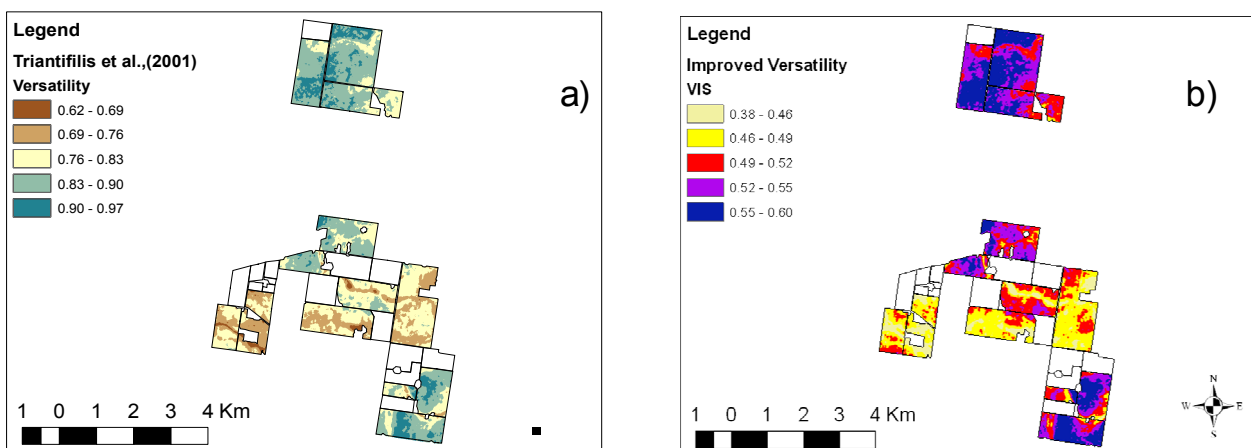


Figure 3. a) Classical versatility map versus b) Improver versatility map

The resulting versatility based on suitability scores, incorporating Shannon's index, assumes a value between 0 and 1, with 1 being the soil is completely versatile; i.e. handle any crop rotation, and 0 meaning the soil is unsuited for multiple crop rotations. The result of this combined versatility-Shannon index is shown in Figure 3b. The map illustrates versatility values that are more diverse than the traditional versatility (Figure 3a) of Triantafilis and McBratney (1993). Both the northern and southern areas depict a mid to versatility range high (0.49-0.60) while the mid areas are portrayed as unsuitable for multiple crop rotations. This information can be vastly beneficial to the landholder when deciding long-term crop rotation and crop

selection; it also provides information as to which areas that can be subject to possibly land use change, i.e. revert from cropping to grazing pasture. This point is also emphasised in Triantafilis *et al.* (2001) in which they showed distinct areas as having high potential for different land use.

Conclusion

In this paper the development of an improved versatility analysis for farm scale management decisions was the main objective. Previous methodology was extended by the inclusion of a diversity index. The initial results of soil suitability for individual crops, as illustrated by suitability maps, exhibit equitable means ranging between 0.81 - 0.83. However, the multiple suitability analysis demonstrated subtle differences in the trend or patterns of the individual crop maps. The development of the improved versatility map yielded important information for management decisions. The results indicate that the Northern and Southern paddocks of the study area exhibited higher versatility range than the rest of the farm and would be highly suited for multiple crop rotations. Another pertinent point is that the areas of low versatility had been identified through this process and could then be subjected to further detailed land use assessment to determine alternative land uses best suited for the areas. Both of these points are vastly important in an era of economic uncertainty when efficiency in relation to productivity is essential to the survival of the farm business. As stated earlier the methodology for versatility analysis has not been greatly covered in the literature since the early work of Triantafilis and McBratney (1993) and not at the farm extent and this research would reignite further research in this area.

References

- Triantafilis J, McBratney AB (1993) Application of continuous methods of soil classification and land suitability assessment in the lower Namoi Valley. p. 172. CSIRO Division of Soils Report No. 121.
- Triantafilis J, Ward WT, McBratney AB (2001) Land suitability in the Namoi Valley of Australia, using a continuous model. *Australian Journal of Soil Research* **39**, 273-290.

Spatial frameworks to support digital soil mapping in Victoria

David Rees^A, Nathan Robinson^B, Mark Imhof^C, Bernie Joyce^D and Richard MacEwan^B

^AFuture Farming Systems Research Division, Department of Primary Industries – Parkville, Victoria, Email David.Rees@dpi.vic.gov.au,

^BFuture Farming Systems Research Division, Department of Primary Industries - Epsom, Victoria, Email Nathan.Robinson@dpi.vic.gov.au, Richard.MacEwan@dpi.vic.gov.au

^CFuture Farming Systems Research Division, Department of Primary Industries – Ellinbank, Victoria, Email Mark.Imhof@dpi.vic.gov.au

^DSchool of Earth Sciences, University of Melbourne, Email ebj@unimelb.edu.au

Abstract

The development of a digital soil mapping (DSM) program for Victoria, Australia is envisaged. The digital soil layer will consist of a specified grid populated by attributes with calculated error terms. A spatial framework is regarded as a useful and efficient means of development of the DSM and would allow for spatial unconformities (i.e. spatial stratified sampling). The new Victorian Geomorphological Framework (VGF) has been used as a refinement for the boundaries used in the national ASRIS scheme and for the development of Primary Production Landscapes (PPLs) in Victoria. The three spatial frameworks provide reference areas for point-derived datasets.

Key Words

Geomorphology, spatial hierarchy, soil survey, land systems.

Introduction

The land and soils that support our agricultural industries are managed in many different ways, reflecting the diversity of Victoria's landscape, and their condition and inherent capability provides a basis for sustainable land use. Spatial frameworks that collate information on soil and land resources support land management and land use decision-making through the provision of spatial data for modelling and reporting purposes. In Victoria, the previous most commonly used spatial hierarchy was based on land systems; derived by integrating environmental features including geology, landform, climate, soils and native vegetation using an ecological approach (Christian and Stewart 1946) and used by Rowan and Downes (1963) and Gibbons and Downes (1964) in Victoria. A number of land system and soil/landscape surveys were subsequently combined to form a land systems map of the state. The advent of Digital Elevation Models (DEM), Geographic Information Systems (GIS), Airborne Gamma-Ray Spectrometry (AGRS), Light Detection and Ranging (LiDAR) and satellite imagery have provided opportunities to re-assess landscapes, refine boundaries and define critical land attributes, improving the overall quality of this information. Geomorphology is the study of landforms, their origin and evolution, the investigation of relationships between landform development and processes that shape and configure these landforms such as tectonic movement, volcanism, erosion and deposition cycles (Hills 1975). Importantly, geomorphology provides a 'fundamental template on which landscape processes and human interactions with those processes take place' (Conacher 2002). Geomorphology has been integrated as a spatial hierarchy of land unit descriptions known as the Victorian Geomorphology Framework (VGF). Recently the VGF, land use maps, climatic records, and regional experience of agronomists and land managers have been integrated to define and determine Victoria's major 'Primary Production Landscapes' (PPL).

In parallel to development of the VGF, a national project has been collating soil and terrain data and representing this spatially as the Australian Soil Resource Information System (ASRIS). This spatial hierarchy for land units provides on-line access to primary soil and land data on behalf of the National Committee on Soil and Terrain (NCST). The VGF and units sit within and complement this national approach. The three spatial hierarchies (VGF, PPL and ASRIS) will be discussed in this paper with reference to their current and future application to support digital soil mapping in Victoria. The recent collation and review of soil and land surveys will also be described for future integration in a Victorian soil condition monitoring program and the generation of soil parameters for systems-based landscape modelling.

Spatial frameworks

Victorian Geomorphology Framework

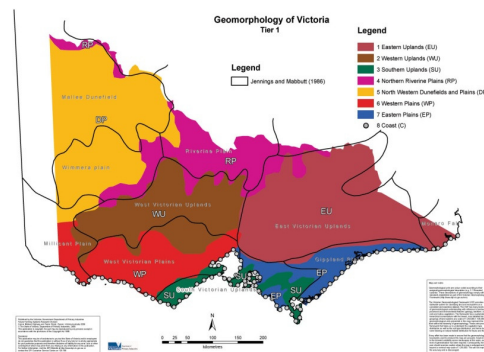
The VGF describes and defines details of Victoria's landscapes and provides a hierarchy to align past and future soil and land information. The classification system followed an existing framework established by

Hills (1975) using landscape morphology as the major factor in the determination of geomorphological areas. The 'tiered' system incorporates geomorphological, pedological and ecological information, enabling users to gain an understanding of both soil and vegetation distribution. The most recent version of statewide land systems derived from the Jenkin and Rowan (1987) two-tier framework included nine 'level one' divisions and 29 'level-two' divisions. The revised VGF was expanded to a three-tier system, incorporating eight 'level-one' divisions (Figure 1), 34 'level-two' divisions and 95 'level-three' divisions. A comparison of the level-one divisions is provided in Table 1. Updates of the VGF are available on the Victorian Resources Online (VRO) website http://www.dpi.vic.gov.au/dpi/vro/vrosite.nsf/pages/landform_geomorphology.

Table 1. Broadest division (Tier 1) of Geomorphological units for Victoria

Victorian Geomorphology Framework
1. Eastern Uplands (EU)
2. Western Uplands (WU)
3. Southern Uplands (SU)
4. Northern Riverine Plains (RP)
5. North west Dunefields and Plains (DP)
6. Western Plains (WP)
7. Eastern Plains (EP)
8. Coastal Features (CF)

Figure 1. Victorian Geomorphological Framework (Tier 1) overlain by Jennings and Mabbitt (1986) map units



Primary Production Landscapes

The Primary Production Landscapes (PPLs) of Victoria is the spatial hierarchy used to identify major agricultural divisions across the state. The PPL hierarchy is a two-tier system that comprises six regional units and twenty two sub-regions. Each PPL has been described according to the agricultural industries and associated farming systems that operate within major geomorphological regions (as determined by the VGF) and inherent soil management issues and associated major soil types are identified. Regional location, physiographic divisions (incorporating terrain and geomorphology) and major soil types were used to delineate PPLs. These PPLs were characterised for dominant soil types and associated inherent management issues. Landform descriptions have been collated from those of the VGF. PPLs were also described by major agricultural industries and practices that occur within them.

Australian Soil Resource Information System (ASRIS)

ASRIS (McKenzie *et al.* 2005: http://www.asris.csiro.au/index_ie.html) provides access to a suite of primary and interpreted soil and land parameters. As an online (web interface) system, the ASRIS structure is a spatial hierarchy of land-unit mapping integrated with individual field sites (e.g. soil site). Land qualities are described along with a functional soil database that informs the spatial context throughout the seven-levels of ASRIS (see comparison with the VGF in Table 2).

Alignment between spatial frameworks


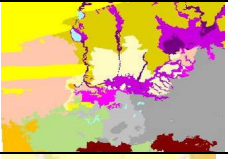

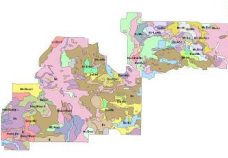
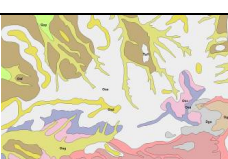
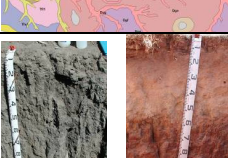
Victoria is well positioned to provide these high level stratifications for ASRIS, as they are being developed in parallel. Within ASRIS, land qualities are defined in greater detail at lower levels (4 to 6) and summaries are provided at higher levels (1 to 3). Level 7 represents an individual soil site in the landscape. The land qualities provided in the system relate to fundamental soil parameters relevant to soil health including soil thickness, water storage, permeability, fertility, salinity, and erodibility. The PPL framework is being promoted as a hierarchy that can be used to stratify soil health monitoring across Victoria and which could offer a practical approach to soil health/condition monitoring across regions.

Soil and land surveys/studies to inform a Victorian digital soil mapping program

In Victoria there are over 350 documented soil and land surveys/studies, undertaken by various state and federal government organisations during the last 80 years. Many of these had little or no associated soil and land mapping, so fewer than half of these were considered suitable for digital soil mapping purposes. The surveys range in scale from 1:10 000 (soil survey) to 1:250 000 (land systems) (Figure 2). Within surveys, the range of soils, soil site density, and scale of mapping and intended purpose of the survey all strongly influence the quality of the final mapping product. The criteria used in a recent assessment and stratification

of surveys included year, undertaken, spatial extent, potential to link mapping units with the VGF, potential to refine mapping, number of soil sites and or quality of site information, purpose of survey, and a subjective estimation of survey quality (incorporating an assessment of surveyor experience, number of reference sites assessed and map scale).

Table 2. Relationship between VGF hierarchy and ASRIS levels

VGF tier illustration	VGF tier level	VGF nominal mapping scale	VGF divisions	ASRIS level and tract name
Not applicable				1 Division
	1	1:5 000 000 to 1:1 000 000	8	2 Province
	2	1:1 000 000 to 1:500 000	34	3 Zone
	3	1:500 000 to 1:100 000	95	4 District
	4	1:100 000 to 1:25 000	not defined	5 System
	5	1:25 000 to 1:100	not defined	6 Facet
	6	> 1:100	not defined	7 Site

Stratification and spatial alignment results

From the stratification process, 106 surveys were considered valuable from a future digital soil mapping perspective for Victoria. The assessment of soil and land surveys assumed an ability to disaggregate the broader-scale surveys or provide existing components/land element descriptions; usually at less than 1:25 000 scale (less than 1:35 000 scale is used here). In total there are 51 surveys at a scale finer than 1:35 000 that equate to level 5 of the VGF and level 6 (or ‘facet’) of ASRIS (these are unlikely to require further refinement given the resolution of survey information); 37 surveys at a scale between 35 000 and 100 000 (level 4 of the VGF and level 5 (or ‘system’) of ASRIS), and 18 surveys at a scale greater than 100 000 that equate to level 3 of the VGF and level 4 (or ‘district’) of ASRIS (Figure 2a to 2c). Significant overlap occurs between the surveys at the same scale and across scales (Figure 2d).

Conclusion

The spatial frameworks of the VGF, PPL and ASRIS are complementary and will provide a valuable hierarchy as part of state and national stratification (for assessing wind erosion, water erosion, soil carbon and acidification) and to guide future development of a digital soil mapping program for Victoria. These frameworks will enable data and information products to be integrated and derived at a range of scales. The continued enhancement of these frameworks and provision of digital soil products to users will be reliant upon new analytical approaches and ancillary data sources to inform these mapping programs.

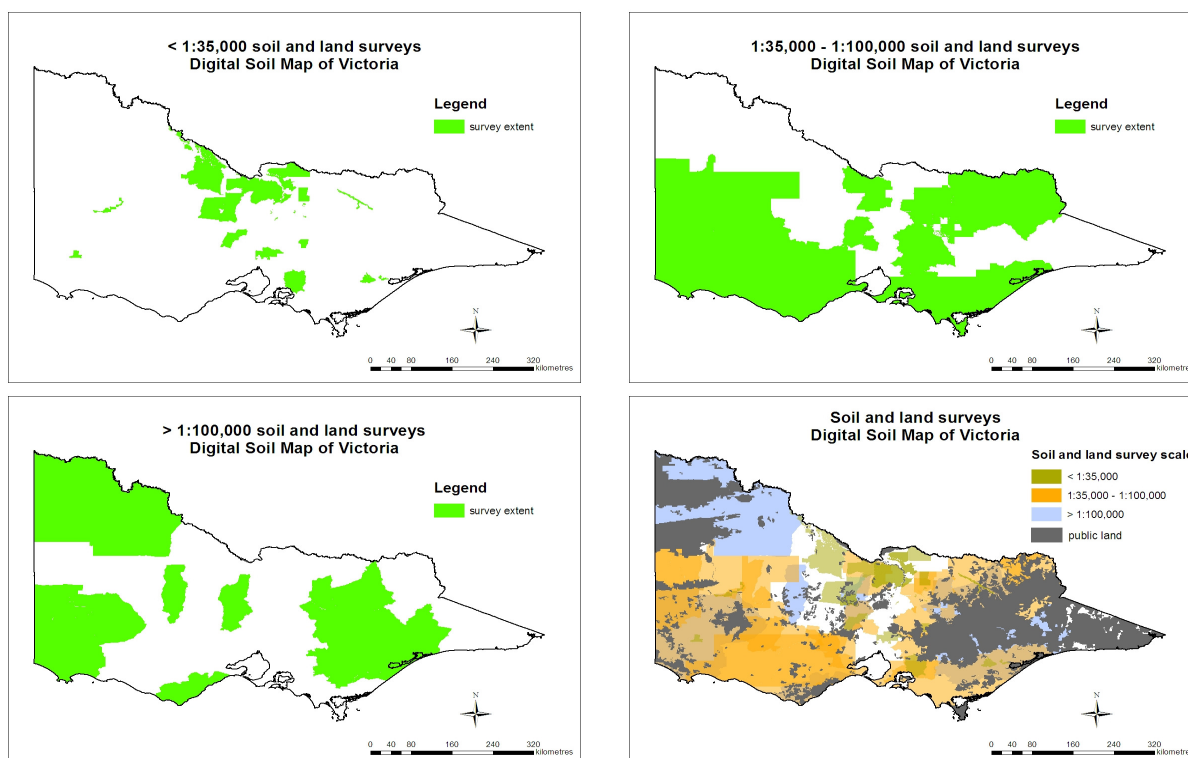


Figure 2. Soil and land surveys: a) finer than 1:35 000 scale; b) between 1:35 000 and 1:100 000 scale; c) coarser than 1:100 000 scale; d) various scales with broad surveys overlain by finer scale surveys (note surveys are partially transparent in order to portray the overlap between survey scales).

References

- Christian CS, Stewart GA (1946) 'General Report on Survey of the Katherine-Darwin Region'. Land Research Series No. 1. (CSIRO: Australia).
- Conacher A (2002) A Role for Geomorphology in Integrated Catchment Management. *Australian Geographical Studies* **40**, 179–195.
- Gibbons FR, Downes RG (1964) 'A Study of the Land in South-western Victoria'. Technical Communication No. 3. Soil Conservation Authority, Victoria.
- Hills ES (1975) *Physiography of Victoria* (First edition). (Whitcombe & Tombs: Melbourne).
- Isbell RF (2002) 'The Australian Soil Classification'. Revised Edition. (CSIRO Publishing: Collingwood, Victoria).
- Jenkin JJ, Rowan JN (1987) Physical Resources. In 'Agriculture in Victoria'. (Eds DJ Connor, DF Smith) (Australian Institute of Agricultural Science: Melbourne).
- Jennings JN, Mabbutt JA (1986) Physiographic outlines and regions. In 'Australia, a geography. Volume 1. The natural environment'. (Ed DN Jeans) (Sydney University Press: Sydney)
- McKenzie NJ, Jacquier DW, Maschmedt DJ, Griffin AE, Brough DM (2005) 'The Australian Soil Resource Information System, Technical Specifications'. Australian Collaborative Land Evaluation Program.
- Rees DB (2000) 'Land Systems of Victoria'. Edition 3. Technical Report No. 56. (Centre for Land Protection Research: Bendigo).
- Rowan JN, Downes RG (1963) 'A Study of the Land in North-western Victoria'. Technical Communication No. 2. (Soil Conservation Authority: Victoria).

Synthesis of knowledge on soil carbon spatial patterns across a large subtropical soil-landscape in Southern U.S.

Sabine Grunwald and Gustavo M. Vasques

Soil and Water Science Department, University of Florida, Gainesville, FL, USA, Email sabgru@ufl.edu

Abstract

The global soil carbon (C) pool is about five times the biotic pool and about four times the atmospheric pool. Landscapes that sequester large amounts of soil C have potential to mitigate global climate change. However, spatially-explicit assessment of soil C across large regions is limited by the number and density of soil observations to capture the underlying variability across soil-landscape continua. The objectives of this study were to (i) synthesize current knowledge on spatially-explicit soil organic carbon (SOC) assessment using different point and polygon soil datasets collected in Florida, U.S. (~140,680 km²) and a large mixed-use watershed nested within Florida (~3,580 km²), and (ii) compare different digital soil mapping methods (aggregation, geostatistical interpolation, and pedo-transfer functions) with different spatial resolutions. The mean SOC across Florida ranged from 13.95 to 47.80 kg/m² and total SOC stocks from 1.99 to 6.82 Pg. Total SOC stock in Florida obtained using different data/methods was 4.110 ± 1.01 Pg (mean \pm std. error) accounting for approximately 0.13% of soil C on earth assuming that the global pool is 3,250 Pg C. Average SOC in the watershed was 17.49 ± 6.89 kg/m², and total SOC was 61.18 ± 24.08 Tg. At both extents, Florida and the watershed, magnitude of differences were found in SOC stocks (means, ranges and absolute values) using different point and polygon soil datasets and aggregation/upscaling methods. Fusing of different soil datasets and methods can help to better capture SOC variability in soil-landscapes.

Key Words

Soil organic carbon, digital soil mapping, aggregation, upscaling, geostatistics, soil carbon assessment

Introduction

It has been estimated that the total global soil carbon (C) pool including wetlands and permafrost (3,250 Pg C) is about five times the biotic pool (650 Pg C) and about four times the atmospheric pool (780 Pg C) (Field *et al.* 2007). Carbon fluxes between soil, biotic and atmospheric pools are dynamic in space and through time and dependent on a multi-factorial system of environmental and anthropogenic drivers. Quantifying C sources, sinks and ecosystem processes that modulate the global C system is critical to identify imbalances and counteract global climate warming. But spatially-explicit assessments of soil C across large landscapes are crude at best. Global soil organic C (SOC) assessment differs widely among soil types, ecosystem types and land uses and has been estimated to vary between 3 to 250 kg/m² (after Jacobson *et al.* 2004). Guo *et al.* (2006) assessed soil C storage across the U.S. using polygon-based legacy data from the U.S. State Soil Geographic (STATSGO) database (currently U.S. General Soil Map) at map scale of 1:250,000. They found that Florida (U.S.) ranks highest in SOC on a per unit area basis among all U.S. states, with 35.3 kg/m² up to 2 m over an area of 136,490 km². Spatially-explicit point measurements (n: 244) were used to assess SOC in Spodosols in Florida observing concentrations in the range of 10.4 ± 0.8 kg/m² from 0 to 1 m, and 18.3 ± 0.8 kg/m² from 0 to varying profile depths, of which 9.2 ± 0.6 kg/m² were stored in spodic horizons (Stone *et al.* 1993). Conditions in Florida's subtropical landscape are favourable to accumulate large amounts of soil C due to flat topography (0 – 105 m amsl), high water table, extensive freshwater marshes, and high biomass production, which have fostered formation of C-rich soils including Histosols with 11% and Spodosols with 31% soil areal coverage.

The objectives of this study were to (i) synthesize current knowledge on spatially-explicit SOC assessment using different point and polygon soil datasets collected in Florida, U.S. and a large mixed-use watershed nested within Florida, and (ii) compare different digital soil mapping methods (aggregation, geostatistical interpolation, and pedo-transfer functions – PTFs) with different spatial resolutions.

Methods

Datasets

We used two polygon-based soil datasets: STATSGO (scale: 1:250,000, time period: 1994) and Soil Survey Geographic (SSURGO) database (scale: 1:12,000 to 1:31,680, time period: 1961 to 2004) from Soil Data Mart (Natural Resource Conservation Service – NRCS, <http://soildatamart.nrcs.usda.gov>). Both Soil Data Mart sets contain soil taxonomic, bulk density (BD), and SOC data associated to soil map units (polygons), which consist of several horizons.

In addition we used horizon-based point observations of SOC from the Florida Soil Characterization Dataset (FSCD, Soil and Water Science Department, University of Florida and NRCS) which entails 1,099 georeferenced BD and SOC observations up to 2 m covering a mapped area of ~140,000 km² (time period: 1965 to 1996). The Santa Fe River Watershed (SFRW) (size: 3,580 km²) was mapped at 141 sites at four fixed depth intervals: 0-30, 30-60, 60-120, and 120-180 cm (time period: 2003 to 2005).

Methodology

All methods produced SOC estimates at the depth from 0 to 100 cm. Method 1 (Florida and SFRW): SOC contents of Soil Data Mart (STATSGO and SSURGO) were calculated by map unit by multiplying the area-weighted average of SOC concentration (in %) by the area-weighted average bulk density (in g/cm³) of the components within the map unit. Method 2 (Florida): SOC contents were calculated by multiplying the SOC concentration (in %) by the soil bulk density (in g/cm³) using point FSCD observations. Method 3 (Florida): SOC contents were estimated by block kriging (BK) of ln-transformed SOC observations (FSCD) using a 250-m block size with 5 x 5 averaged estimations within each block. Method 4 (Florida): Average SOC contents by soil order obtained from FSCD observations were applied to STATSGO soil orders (7 in total). Method 5 (SFRW): Ordinary kriging (OK) of ln-transformed SOC observations using a 100-m grid size. Method 6 (SFRW): BK of ln-transformed SOC observations using a 30-m block size with 5 x 5 averaged estimations within each block (Vasques *et al.*, 2010). Method 7 (SFRW): Class PTF – Average SOC contents by SSURGO soil series from 139 observations were applied to SSURGO soil series. Method 8 (SFRW): Class PTF – Average SOC contents by soil order/land use (LU) combinations from 139 observations were applied to the respective areas.

Results and Discussion

Spatially-explicit soil organic carbon assessment across a large subtropical region in U.S. (Florida)

The SOC derived by different methods are summarized in Table 1. The mean SOC ranged from 13.85 to 47.80 kg/m² and total SOC stocks from 1.99 to 6.82 Pg. STATSGO (Method 1) estimated the upper bound of SOC, whereas FSCD (Method 2) described the lower bound, providing conservative estimates. Total SOC stock in Florida obtained using different data/methods was 4.110 ± 1.01 Pg (mean \pm std. error) accounting for approximately 0.13% of soil C on earth assuming that the global pool is 3,250 Pg C (Field *et al.* 2007). According to the soil order class PTF (Method 4), Histosols constitute 11% of Florida soils, but store 53% of the total SOC stock; and Spodosols occupy 31% of Florida soils and store 21% of the SOC. Entisols occupy 24% of the area and store 11% of the total SOC stock. Histosols store the largest amount of SOC with 51.82 ± 23.62 kg/m² (mean \pm std. dev.) followed by Mollisols (13.98 ± 10.97), Inceptisols (13.20 ± 10.46), Spodosols (8.86 ± 5.81), Alfisols (5.58 ± 4.61), Entisols (4.83 ± 8.58), and Ultisols (4.10 ± 3.56) kg/m² (Grunwald 2008).

Currently, Florida's wetlands cover about an area of 15,098 km² which has been steadily declining. In the area of the Gulf of Mexico 150,138 ha of wetlands have been lost (1998 to 2004) (Stedman and Dahl 2008) and drainage of the Everglades changed south Florida from a subtropical wetland (~1880) to a human dominated landscape with a strong tourism, retirement, and agricultural economy. As a result, the Greater Everglades ecosystem is half of its original size with current extent of only about 8,250 km² which would translate into loss in SOC of about 0.43 Pg C, according to Method 4, in the period of ~1880 to current. Carbon credits and registries promote restoration of wetlands that accumulate large amounts of soil C but need to be cautiously assessed. Considering the formation of a 1-m Histosol soil profile at an accretion rate of 1.1 cm/yr in Florida nutrient-enriched wetlands (Reddy *et al.* 1993) and assuming average methane (CH₄) emissions of 0.85 g/m²/d (Schipper and Reddy 1994), this would translate into total CH₄ emission of 0.095 Pg CO₂eq. (over a period of 90.9 yrs.) and a total soil C net gain of 0.194 Pg CO₂eq. However, in non-enriched wetlands the soil accretion rate is less with about 0.25 cm/yr (Reddy *et al.* 1993) and contrasts with CH₄ emissions of 0.418 Pg CO₂eq., which would lead to a total soil C net loss of 0.129 PgCO₂eq. (over a period of 400 yrs.) (Grunwald 2008). These calculations have not yet accounted for the Global Warming Potential factor of 25 for CH₄ (Intergovernmental Panel on Climate Change 2007). Many land use practices – some involving land use changes – have shown to increase SOC and thus received considerable attention for their possible role in climate change mitigation. Fransluebbers (2005) assessed a SOC sequestration rate for the southeastern U.S. at 153.7 Mg CO₂eq/km²/yr. If 50% of the agricultural area in Florida would be converted from conventional to no-tillage, a total net gain of 1,723,077 Mg CO₂eq/yr could be achieved.

Table 1. Estimates of SOC stocks to 1 m in Florida.

Soil Data	Method	n	Map unit	Min. (kg/m ²)	Max. (kg/m ²)	Mean (kg/m ²)	Std. Dev. (kg/m ²)	Total (Florida) (Pg)	Mean stock (Florida) (kg/m ²)
SSURGO	1	655,155	map unit	0.67	291.77	24.17	39.31	3.52	27.32
STATSGO	1	2,823	polygons						
			map unit	4.01	264.32	58.44	62.67	6.82	47.80
			polygons						
FSCD points	2	1,099	points	0.13	207.98	12.85	23.69	N/A	N/A
FSCD points	3 (BK)	2.28 x 10 ⁶	250-m	2.82	116.19	13.95	12.28	1.99	13.95
			pixels						
FSCD by STATSGO	4	7	soil orders	7.70	144.17	32.84	45.63	4.11	28.83

Spatially-explicit soil organic carbon assessment across the Santa Fe River Watershed in north-central Florida

The SOC derived by different methods are summarized in Table 2. STATSGO overestimated SOC relative to other methods. Overall, best agreement between SOC estimates across the watershed was found in areas of low SOC stock, whereas areas of medium to high SOC (esp. river valleys, wetlands, Histosols, and Spodosols) had higher coefficients of variations (CV) (Methods: 1 and 5-8); thus, contributing to a highly uneven distribution of SOC differences over the watershed (map not shown). The mean CV (Methods: 1 and 5-8) was 42.54% indicating the high variability among different aggregation/upscaling methods to estimate SOC. Average SOC in the watershed was 17.49 ± 6.89 kg/m², and total SOC was 61.18 ± 24.08 million tons.

Table 2. Estimates of SOC stocks to 1 m in the Santa Fe River Watershed.

Soil Data	Method	n	Map-unit	Min. (kg/m ²)	Max. (kg/m ²)	Mean (kg/m ²)	Std. Dev. (kg/m ²)	Total (SFRW) (Mg)	Mean stock (SFRW) (kg/m ²)
SSURGO	1	193	map unit	2.93	138.83	21.82	24.42	53,350,771	15.25
			polygons						
STATSGO	1	36	map unit	5.06	173.89	32.09	61.68	105,459,947	30.15
			polygons						
SFRW points	5 (OK)	3.59 x 10 ⁵	100-m	2.62	160.50	10.95	3.67	38,376,698	10.95
			pixels						
SFRW points	6 (BK) [†]	3.98 x 10 ⁶	30-m	3.20	199.37	19.08	6.01	68,389,193	19.08
			pixels						
SFRW by SSURGO	7	174	soil series	2.66	108.04	13.69	18.50	40,515,841	11.58
SFRW by SSURGO/LU	8	24	soil order/LU	5.51	143.52	18.36	28.29	68,220,134	19.50

[†] Vasques *et al.* (2010)

Conclusion

At both extents, Florida and the SFRW, magnitude of differences were found in SOC stocks (means, ranges and absolute values) using different point and polygon soil datasets and aggregation/upscaling methods. Although these subtropical landscapes store huge amounts of SOC, regardless of soil data/methods used, it is difficult to assess which accounting method performs best. Validation of point estimates of SOC (OK) suffer from the effect of different supports between validation soil samples (points) and output pixel sizes, which are assumed to be represented by the point estimate. Block kriging estimates of SOC are difficult to validate since a validation sample would need to represent the variability in SOC within each block. And soil map units are assumed to be internally homogenous and represented by one assigned SOC value, which often does not match variability of SOC across landscapes or validation sample supports. To resolve this dilemma will require joint effort and more research to further explain the variation of SOC and reduce the uncertainty in SOC estimates. Fusing of different soil datasets and methods can help to address these shortcomings as shown in this study.

Acknowledgement

All staff and scientists who contributed to the development of the Florida Soil Characterization Database are

thanked to make this project possible. Soil Data Mart created and maintained by NRCS is given full credit. We also like to acknowledge Christine M. Bliss, Gregory L. Bruland and Nicholas B. Comerford, who helped with soil sample collection in the Santa Fe River Watershed; Chunhao Xu and Yu Wang for the laboratory analysis of samples from the watershed; and N. DiGruttolo for GIS editing of layers. We like to thank the Cooperative Ecosystem Service Unit (NRCS) and Nutrient Science for Improved Watershed Management Program (U.S. Department of Agriculture) and the USDA-AFRI-NIFA project 2007-35107-18368 “Rapid Assessment and Trajectory Modeling of Changes in Soil Carbon across a Southeastern Landscape” (Core Project of the North American Carbon Program) for providing partial funding of this project.

References

- Field CB, Sarmiento J, Hales B (2007) The carbon cycle of North America in a global context. In ‘The first state of the carbon cycle report (SOCCR) – Synthesis and assessment product 2.2’ (Eds. AW King, L Dilling, GP Zimmerman, DM Fairman, RA Houghton, G Marland, AZ Rose, TJ Wilbanks), pp. 21-28 (Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research).
- Franzluebbers AJ (2005) Soil organic carbon sequestration and agricultural greenhouse gas emissions in the southeastern U.S. *Soil and Tillage Research* **83**, 120-147.
- Grunwald S (2008) Role of soils to sequester carbon. In ‘*Opportunities for greenhouse gas reduction by agriculture and forestry in Florida*’ (Eds. S Mulkey, J Alavalapati, A Hodges, AC Wilkie, S Grunwald), pp. 1-61 (University of Florida, School of Natural Resources and Environment – Department of Environmental Defense: Washington DC).
- Guo Y., Amundson R, Gong P., Yu Q (2006) Quantity and spatial variability of soil carbon in the conterminous United States. *Soil Science Society of America Journal* **70**, 590-600.
- Intergovernmental Panel on Climate Change (2007) Fourth assessment report on climate change. Available at: <http://www.ipcc.ch/>. Last verified: 30 Oct 2009.
- Jacobson MC, Charlson RJ, Rodhe H, Orians GH (2004) *Earth system science – From biogeochemical cycles to global change*. International Geophysics Series Vol. 72, Elsevier Academic Press: New York NY.
- Reddy KR, DeLaune RD, DeBusk WF, Koch MS (1993) Long-term nutrient accumulation rates in the Everglades. *Soil Science Society of America Journal* **57**, 1147-1155.
- Schipper LA, Reddy KR (1994) Methane production and emissions from four reclaimed and pristine wetlands of southeastern U.S. *Soil Science Society of America Journal* **58**, 1270-1275.
- Stedman S, Dahl TE (2008) *Status and trends of wetlands in the coastal watersheds of the Eastern United States 1998 to 2004* (National Oceanic and Atmospheric Administration, National Marine Fisheries Service and U.S. Department of the Interior, Fish and Wildlife Service).
- Stone EL, Harris WG, Brown RB, Kuehl RJ (1993) Carbon storage in Florida Spodosols. *Soil Science Society of America Journal* **57**, 179-182.
- Vasques GM, Grunwald S, Comerford NB, Sickman JO (2010) Regional modeling of soil carbon at multiple depths within a subtropical watershed. *Geoderma* (in press); doi:10.1016/j.geoderma.2010.03.002 .

Uncertainty assessment of mapping topsoil DDT spatial distribution using multiple indicator kriging

Yong-Cun Zhao^A

^AInstitute of Soil Science, Chinese Academy of Sciences, Nanjing, 210008, China, Email yczhao@issas.ac.cn

Abstract

Topsoil samples (n=544) were collected from the Zhangjiagang county of China. MultiGaussian (MG) approach and multiple indicator kriging (MIK) were applied for mapping the spatial distribution of DDT concentrations and the probability that DDT concentration exceeded a critical threshold concentration. High variability of DDT concentrations was observed. Areas with higher DDT concentrations are mainly distributed in the northeastern parts of the county where fluvo-aquic soils are the dominated soil types and historical heavy applications of DDT pesticide for cottons were identified during investigation. The DDT spatial distributions obtained by MG and MIK methods showed similar spatial patterns except that estimate of MG method is smoother. The areas with DDT concentrations that exceeded the background value of the Chinese Environmental Quality Standard for Soils (CEQS) using MIK method is larger than those obtained by MG method. The estimated conditional cumulative distribution functions (ccdf) using MG and MIK methods are accurate and the goodness statistics G both close to 1. However the MIK is of higher precision for modelling DDT uncertainty because the widths of probability intervals are narrower while including the expected proportions of true values.

Key Words

Uncertainty, multiple indicator kriging, DDT, spatial distribution.

Introduction

Information and assessment of spatial patterns of DDT contaminated areas are important for risk assessment, soil remediation, as well as effective management recommendations. Spatial prediction of soil contaminant, however, usually involves uncertainties that need to be considered when making decisions for future management of contaminated areas (Goovaerts 2001), because such uncertainties can be propagated into subsequent environment modelling and fundamentally impacts the ultimate results of the model (Lark and Bolam 1997). In practice, however, it is difficult to accurately characterize the spatial patterns of pollutant because two common features are often involved in the pollution data: 1) highly positively skewed histograms (hot-spots), and 2) presence of data below the detection limit (censored observations) (Saito and Goovaerts 2000), i.e. DDT concentration in soils. The multiGaussian (MG) and indicator kriging (IK) approaches provide the estimate of the ccdf as opposed to the traditional kriging approach. And IK offers a way to deal with classes or populations of both high values and values below the detection limit. As pointed out by Journel (1983) the primary advantages of the IK approach are: 1) it makes no distributional assumptions; 2) the procedure is resistant to outliers; and 3) the procedure can accommodate high connectivity of extreme values. The specific objectives of this study are 1) to map the spatial distribution of DDT concentrations using MG and MIK methods; 2) to assess the performances of MG and MIK method for modelling the uncertainties associated with the mapping process.

Methods

Study area

The study area, Zhangjiagang County, is situated on a flat alluvial plain in Yangtze River Delta region of China (Figure 1). The total area is 999 km². With a north sub-tropical monsoon climate, Zhangjiagang has a mean annual temperature of 15.2°C and 1039.3 mm annual rainfall. The main soil types in the study area are fluvo-aquic soils (Aquic Cambosols) and paddy soils (Stagnic Anthrosols) (SSOSC 1984). There are different crop planting systems on each type of soils. On the fluvo-aquic soils, before the 1980's, the rotation of cotton as summer crop and wheat as winter crop was dominant, however, after the 1980's most of the cotton was increasingly substituted with rice. On paddy soils, the rotation of rice and wheat has always been the dominant planting system.

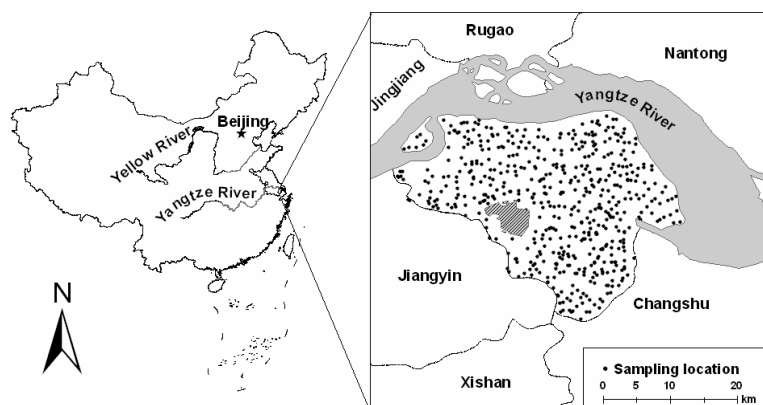


Figure 1. Location of Zhangjiagang soil sampling distribution patterns in 2004.

Soil sampling and chemical analysis

A total of 544 samples in the topsoil (0-15 cm) were collected in 2004 (Figure 1) according to soil types, land use, and thorough coverage of the study area. The collected soil samples were air-dried at room temperature and sieved to pass 70-mesh sieve for analysing DDD, DDE, and DDT using an Agilent gas chromatograph 6890 equipped with a Nickel 63 electron capture detector and a HP-5 column (30 m×0.25 mm inside diameter, 0.25 µm film thickness) (GB/T 14550-93) (SEPAC 1993).

MultiGaussian (MG) approach and multiple indicator kriging (MIK)

MG approach: If the random function (RF) $Z(\mathbf{u})$ is multivariate Gaussian, then the simple kriging estimate and variance identify the mean and variance of the posterior ccdf. Since that ccdf is Gaussian, it is fully determined by these two parameters (Deutsch and Journel 1998). MIK approach: The kriging algorithm applied to indicator data provides least-square estimates of the ccdf (Deutsch and Journel 1998).

Prediction precisions: The performances of the MG and MIK approaches were assessed using cross-validation (Goovaerts *et al.* 2004). The ability of different methods to estimate DDT concentration was quantified using the mean prediction error (MPE) and root mean squared prediction error (RMSPE). Local uncertainty: the p -probability intervals bounded by the $(1-p)/2$ and $(1+p)/2$ quartiles of the ccdf (PI) (Goovaerts 2001) and the “goodness” statistics G (Deutsch 1997) were used to assess the local uncertainty.

Results

Descriptive statistics

Table 1 showed that the range of Σ DDT concentrations is as large as 590 µg/kg, and p,p' -DDE is the major composition of DDT, the p,p' -DDE can account for 55% of the Σ DDT concentration averagely. Great variation exists in the DDT data, and the CV varied from 107 to 167%. Skewness of Σ DDT concentrations is 1.73, indicating that the distribution of Σ DDT data is strongly positively skewed, and suggested that higher Σ DDT concentration values existed and the spatial distribution of Σ DDT is not homogeneous.

Table 1. Descriptive statistics of DDT in soils of Zhangjiagang County, China (n=544).

	Mean µg/kg	Minimum µg/kg	Median µg/kg	Maximum µg/kg	Standard deviation µg/kg	Skewness	Kurtosis	CV %
p,p' -DDE	48.58	ND	23.11	344.30	61.08	1.98	6.76	126
p,p' -DDD	9.74	ND	6.93	86.81	11.31	3.35	18.30	116
p,p' -DDT	24.50	ND	14.74	384.66	35.41	4.93	38.60	145
o,p' -DDT	4.99	ND	0.78	61.17	8.35	2.45	10.01	167
Σ DDT	87.80	ND	50.60	590.10	93.86	1.73	6.31	107

Σ DDT represents the sum of four metabolites of DDT and the descriptive statistic were calculated assuming non-detect (ND) measurements were equal to one-half the detection limit.

Spatial distribution of DDT concentrations

Areas with higher DDT concentrations are mainly located in the northeastern parts of the county where fluvo-aquic soils are the dominated soil types (Figure 2). And areas with lower DDT concentrations are mainly distributed in southern parts where paddy soils are dominated. The DDT spatial patterns presented by the MIK method are more complex than those by MG method, especially in central and southern parts of the county. The local variability of DDT concentrations at central and southern parts can be intuitively observed

when using MIK. However such local variability was filtered out by MG method due to stronger smoothing effect of MG method. Figure 3 presented the scatterplot of observed Σ DDT concentrations versus estimates at each of the 544 individual sampling sites. The MG method resulted in an average underestimation of 54.93 $\mu\text{g/kg}$ and overestimation of 23.07 $\mu\text{g/kg}$. With respect to MIK method, the average underestimation and overestimation are 35.52 and 56.20 $\mu\text{g/kg}$, respectively. The MPE in MIK method is closer to zero than MG method, and the RMSPE in MIK method is lower than MG method, which highlights the MIK method is of higher precision for mapping the spatial distribution of DDT concentrations in the county.

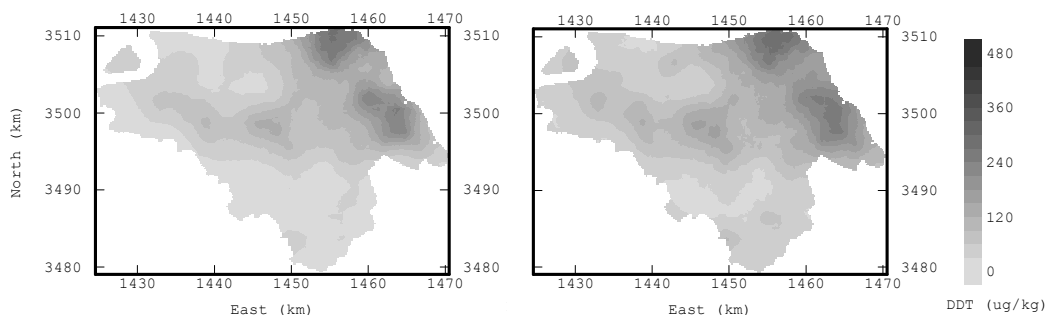


Figure 2. Estimates of DDT concentrations using multiGaussian (MG) method (left), and multiple indicator kriging (MIK) (right).

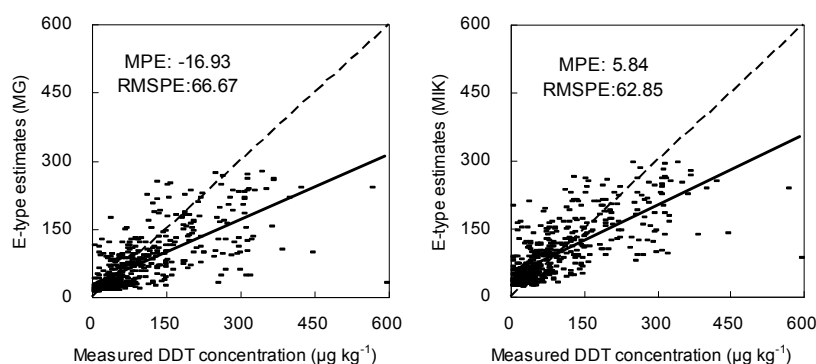


Figure 3. Cross-validation for E-type estimates of Σ DDT concentrations using multiGaussian (MG, left) method and multiple indicator kriging (MIK, right).

Performance for modelling uncertainty

The areas with DDT concentrations exceeded the background value of CEQS using MIK method is larger than those obtained by MG method (when critical probability 0.5 is used, areas with DDT concentrations exceeding 50 $\mu\text{g/kg}$ using MG and MIK methods covered 47% and 49%, respectively) (Figure 4). When the critical limit is set as 500 $\mu\text{g/kg}$, MIK and MG method are of probabilities lower than 0.5, indicating that although soils in the county are not polluted by DDT based on threshold 500 $\mu\text{g/kg}$ in CEQS, the potential risks of DDT residue in soils are still high because areas with DDT concentrations exceeded the background value 50 $\mu\text{g/kg}$ covered nearly half of the county.

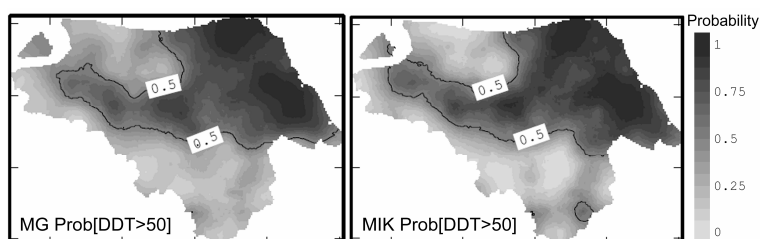


Figure 4. Probability for Σ DDT concentrations greater than the thresholds 50 $\mu\text{g/kg}$ and 500 $\mu\text{g/kg}$.

The accuracy plots (the proportions of the measured data falling into different probability intervals) in Figure 5 indicated that the probability intervals of MG and MIK models of uncertainty contain a higher than expected proportion of true values. The goodness statistics (G) for cross validation were 0.92 for MG method and 0.94 for MIK, which is closer to the ideal value of 1. This suggests that the ccdf was quite accurately estimated. However, Figure 5 (right) showed that the best model of uncertainty is obtained by MIK method

because the widths of PI are narrower (larger precision) while including the expected proportions of true values (large goodness statistics).

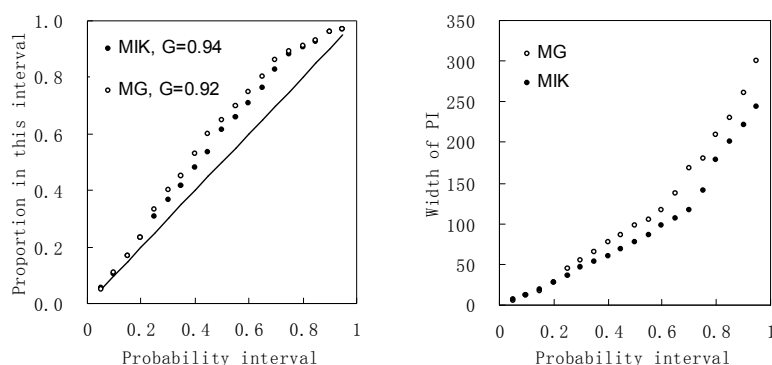


Figure 5. Accuracy plot (the proportion of measured Σ DDT values falling within theoretical probability intervals) and the width of these intervals versus the probability intervals.

Conclusion

Great variation exists in the DDTs data, and the CV varied from 107 to 167%. Areas with higher DDT concentrations are mainly located in the northeastern parts of Zhangjiagang county where fluvo-aquic soils are the dominated soil types, while areas with lower DDT concentrations are mainly distributed in southern parts. The DDT spatial distributions obtained by MG and MIK methods showed similar patterns except that estimate of MG method is smoother. The areas with DDT concentrations exceeded the background value of the CEQS using MIK method is larger than MG method. The estimated cdf using MG and MIK was accurate and the goodness statistics G both close to 1. However the MIK is of higher precision for modelling DDT uncertainty.

References

- Deutsch CV, Journel AG (1998) 'GSLIB: geostatistical software library and user's guide'. (Oxford University Press: New York)
- Deutsch CV (1997) Direct assessment of local accuracy and precision. In 'Geostatistics Wollongong '96' (Eds EY Baafi, NA Schofield) pp. 115-125. (Kluwer Academic Publishers: Dordrecht)
- Goovaerts P (2001) Geostatistical modelling of uncertainty in soil science. *Geoderma* **103**, 3-26.
- Goovaerts P, Avruskin G, Meliker J, Slotnick M, Jacquez GM, Nriagu J (2004) Modeling uncertainty about pollutant concentration and human exposure using geostatistics and a spacetime information system: Application to arsenic in groundwater of Southeast Michigan. In 'Accuracy 2004: Proceedings of the 6th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences' (Portland, Maine, USA)
- Journel AG (1983) Nonparametric estimation of spatial distributions. *Mathematical Geology* **15**, 445-468.
- Lark RM, Bolam HC (1997) Uncertainty in prediction and interpretation of spatially variable data on soils. *Geoderma* **77**, 263-282.
- Saito H, Goovaerts P (2000) Geostatistical interpolation of positively skewed and censored data in a dioxin contaminated site. *Environmental Science & Technology* **34**, 4228-4235.
- Soil Survey Office of Shazhou County (SSOSC) (1984) 'Description of Soils in Shazhou County.' (Agriculture Bureau of Suzhou City, Jiangsu Provincial Office for Soil Survey: Suzhou). (in Chinese)
- State Environmental Protection Administration of China (SEPAC) (1993) National Standard of People's Republic of China: Soil quality-Determination of BHC and DDT-Gas chromatography (GB/T 14550-1993). http://www.zhb.gov.cn/tech/hjbz/bzwb/trhj/trjcgfffbz/199401/t19940115_82035.htm

What is the spatial representation of digital soil maps? An issue of the spatial entity

Brendan Malone^A, Alex McBratney^A and Budiman Minasny^A

^AFaculty of Agriculture Food & Natural Resources, The University of Sydney, NSW 2006, Australia, Email b.malone@usyd.edu.au, a.mcbratney@usyd.edu.au, b.minasny@usyd.edu.au.

Abstract

The spatial entity used within a digital soil mapping framework can have profound implications for data reliability and interpretability. We compare the quality of results and maps of the variability of available water capacity (AWC) generated by both the classical point prediction approach and an innovative method that involves a block estimator. Rather than a single prediction point in the centre of a pixel, the block estimator method takes the weighted average of the soil attribute across the entire extent of a pixel through a block kriging approach. The results indicate AWC varies continuously across the landscape. However, maps produced by our block estimator method result in predictions and maps which are more fluid and smoothed reflecting a more realistic representation of the variability.

Key Words

Digital soil mapping, block kriging, available water capacity.

Introduction

The pixel model is an efficient method in which to display the spatial variability of soil properties and classes in a map format (Grunwald 2006). It is generally assumed that the value predicted for a soil property or class at a pixel is the same at every point within the extent of that pixel, no matter the resolution. In actuality, the predicted soil value is a single point located at the centre of the pixel (Figure 1). Thus the short range soil variability that may occur within a pixel will not be accurately addressed. A more appropriate method in which to assign a predicted soil value to a pixel is by way of a bulked mean within the entire extent of the pixel.

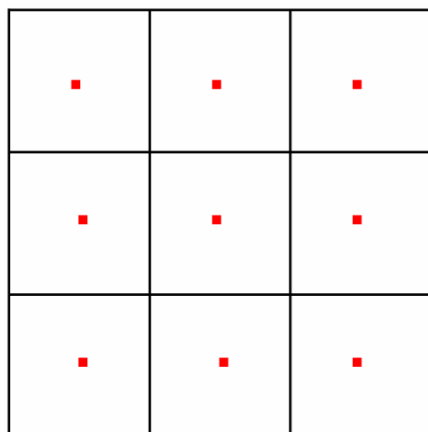


Figure 1. Representation of the pixel model and associated point prediction in the centre of the pixel.

For this study we demonstrate a method in which a mean prediction of soil properties within a pixel can be evaluated by using an existing point predicted digital soil map. This method involves the use of block kriging, where the kriged value represents a statistically weighted average of the entire extent of the pixel (Whelan *et al.* 2001). Using available water capacity (AWC) as an exemplar soil property, we compare and contrast maps produced by both methods i.e. point predictions vs. bulked mean predictions.

Methods

Study area

The study site (1500km²) is situated near Narrabri (30.32S 149.78E), 500km NNW of Sydney, NSW, Australia. Agricultural enterprises such as cropping and pastoral farming are predominant in the area (Figure2). The soil dataset consists of 341 soil profiles (Figure 2). The dataset describes and quantifies

various soil morphological, physical and chemical attributes at depth intervals of 0–0.1, 0.1–0.2, 0.3–0.4, 0.7–0.8, 1.2–1.3 and 2.5–2.6m (McGarry *et al.* 1989). The focus of this study was the prediction of AWC in the top 10cm of the soil profile; therefore we were only concerned with measurements in the 0–10cm depth increment. Using sand, clay and organic matter as inputs (McGarry *et al.* 1989) a pedo-transfer function generated estimates of AWC at each sample point (Minasny *et al.* 2006).

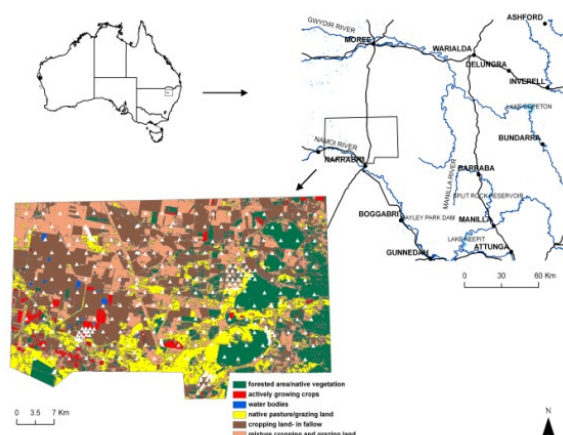


Figure 2. The Edgeroi study area.

Environmental data

A number of environmental covariates were sourced and interpolated onto a common grid of 90m resolution. These included:

- 3 arc-second (90m) digital elevation model (DEM). First and second derivatives, namely: slope, terrain wetness index (TWI), flow path length, altitude above channel network (AOCN) and multi-resolution index of valley bottom flatness (MRVBF) were determined.
- Landsat 7 ETM+ images from 2003. The Landsat bands were used for the approximation of land cover and land use. Vegetation cover and type was estimated using the Normalised Difference Vegetation Index (NDVI). Furthermore, the band ratios or more commonly, soil enhancement ratios of b3/b2, b3/b7 and b5/b7 were also derived.
- Gamma-radiometric survey data which provides a measure of the spatial distribution of three radioactive elements (potassium-K, thorium-Th and uranium-U) in the top 30–45 cm of the earth's crust. This data was used to approximate the distribution of various parent materials over the landscape.

Data analysis

We used a regression kriging approach to predict AWC at each sample point and across the study area. A neural network was constructed on a training data set of 261 points (leaving 80 for validation). The input variables were the various layers of environmental covariates. The target variable was the measured value of AWC at each training point. After the modeling process, residuals were evaluated for each point. A semi-variogram was used to assess the spatial distribution of residuals. We used an exponential model to krig the residuals onto the common 90m grid of the Edgeroi. For model validation, the profile formulae were applied to the 80 withheld data points. Residuals, estimated from the semi-variogram model of the training step residuals were added to the prediction resulting in a final prediction.

For mapping point estimates of AWC, profile formulae from the neural network were applied to the common 90m grid geo-database where only information relating to the environmental data existed to make predictions of AWC. A final prediction was determined by adding the kriged residual to the prediction at each point. A point predicted map of AWC was the resulting product of this procedure.

The final predictions of the point estimates of AWC were used as inputs for the block mean estimation. As a first step to the block estimation method is the realization uncertainty of the prediction at each point or pixel. The residual estimate is a good indication of this. As we could quantify this element of uncertainty, we incorporated it within the block kriging procedure by defining it as the σ^2 parameter. The σ^2 parameter was calculated by evaluating the variance of all the residuals across the study area. Secondly, the common grid was off-set by 1m to the left from the original grid. With the incorporation of the σ^2

parameter, block kriging was performed onto the offset grid. The block size used had equal-side dimensions of 90m. For each prediction via block, local exponential variograms were used. A block estimated predicted map of AWC was the resulting product of this procedure.

Results

For neural network training we used a 3 hidden node network. Lin's concordance correlation coefficient (CCC) between the observed and predicted AWC values was 0.61, indicating a substantial agreement along the 45° line (Lin 2000). Validation results were less impressive where a fair agreement (CCC = 0.27) was observed between the observed and predicted (Figure 3a). There was very little spatial autocorrelation of residuals beyond a separation distance of 83m (Figure 3b) which resulted in only a minor improvement in the final prediction where the CCC was 0.34 indicating still some significant deviations from the 45° line (Figure 3c).

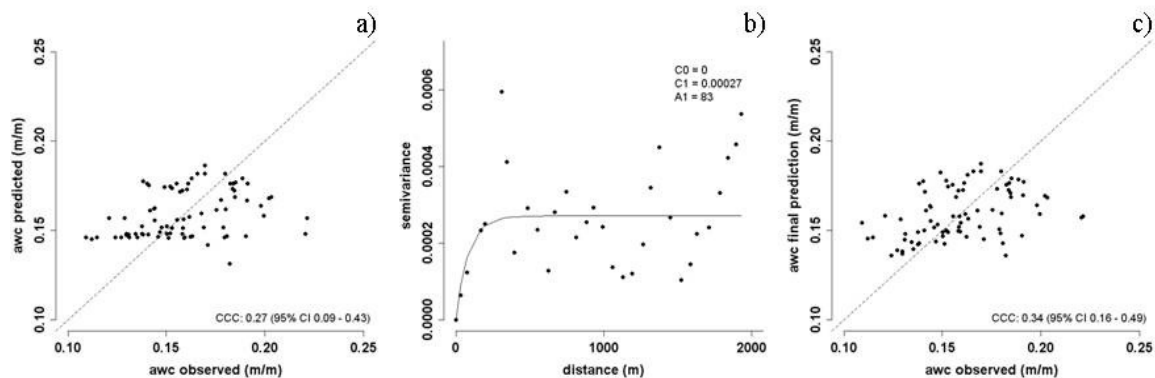


Figure 3. Validation results. (a) Observed vs. predicted plot (b) semi-variogram of residuals (c) Observed vs. final prediction plot.

The model results indicate there was some correlation between the target variable AWC and the available environmental data. This is shown in Figure 4a where AWC varies significantly across the study area. Some notable patterns exist such as a generally higher to lower gradient of AWC in the east to west direction. This coincides with an increasing proportion of land used for cropping as one moves in a westerly direction across the study area. Additionally the topography itself also becomes more open and flatter in the western area in comparison to the more undulating land features in the east.

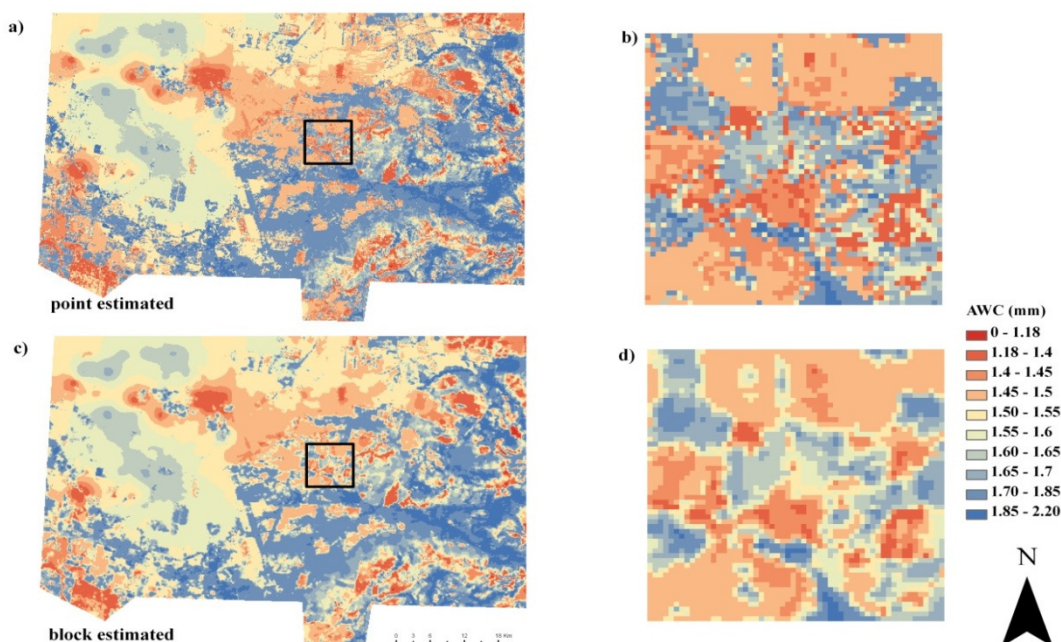


Figure 3. Comparisons of maps of AWC variability in the 0–10cm depth range generated from point predictions (a) and block estimated predictions (c). Zoomed in zone of 5km² of predictions made from points (b) or blocks (d).

The result of the block estimation method was overall very positive due to the fact that the visual appearance of the map is a lot smoother and clearer for interpretation than that of the point estimated map (Figure 3c). Broadly, the spatial variation of AWC changes very little when comparing both maps. The key difference is that in areas where there is significant local variation, the point predictions appear quite noisy in comparison to the block estimates where there is a more gradual transition in predicted values across an area. This phenomenon is illustrated where we selected a 5km² zone within the study area that displayed what appeared to be a high degree of spatial variability. Overall, it is difficult to separate both methods in terms of the general variation of AWC in the zone. However, for the point estimated predictions (Figure 3b), the zone is clearly pixilated or noisy. On the other hand, for the block estimations (Figure 3d), interpretation of AWC variability is easier to define as the noisiness has been smoothed resulting in a clearer map.

Discussion and conclusions

Block estimations of AWC across the Edgeroi significantly improved the visual quality and interpretability of the generated map. The procedure had the effect of smoothing out some of the noisiness that results from point predictions resulting in a more realistic depiction of AWC across the Edgeroi area. In a general context, the spatial entity of a soil map has significant implications for the interpretability of soil information. Essentially a pixel value of a soil property is a single observation located at the centre of that pixel. It is problematic to assume this value is the same for the entire extent of the pixel. Thus a more appropriate method for making predictions within a pixel is the incorporation of a block estimator in which makes a prediction that represents a statistically weighted average across the entire extent of a pixel. The result is that collectively, predictions are more continuous from one pixel to the next. This improves the quality of predictions, enhances the interpretability of the resulting soil map and realistically illustrates the spatial variation of soil properties across a defined area. As this method is relatively simple to implement, this method could easily be appended to the methodologies of other DSM projects.

References

- Grunwald S (2006) What do we really know about the space-time continuum of soil-landscapes. In 'Environmental Soil-Landscape Modeling'. (Ed. S Grunwald) pp. 3–36. (CRC Press: Boca Raton, FL).
- McGarry D, Ward WT, McBratney AB (1989) 'Soil studies in the Lower Namoi Valley: methods and data. The Edgeroi Dataset.' (CSIRO Division of Soils: Adelaide).
- Minasny B, McBratney AB, Mendonca-Santos ML, Odeh IOA, Guyon B (2006) Prediction and digital mapping of soil carbon storage in the Lower Namoi Valley. *Australian Journal of Soil Research* **44**, 233–244.
- Whelan BM, McBratney AB, Minasny B (2001) Vesper – Spatial prediction software for precision agriculture. In 'ECPA 2001. Third European Conference on Precision Agriculture.'. Agro Montpellier, Ecole Nationale Agronomique de Montpellier. (Eds G Grenier, S Blackmore) pp. 139–144.