ΠΕΔΟ METRON

The Newsletter of the Pedometrics Commission of the IUSS

Issue 34, April 2014 **Chair: A-Xing Zhu** **Vice Chair: Dick. J. Brus**

Layout: Jing Liu

Inside this Issue

From the Chair

Dear Fellow Pedometricians and Friends,

I would like to start this issue of Pedometron with congratulations to Dr. Budiman Minasny of University of Sydney and Dr. Lin Yang of Chinese Academy of Sciences for being elected as the next chair and vice chair of Pedometrics Commission, respectively. I would also like to thank them for their willingness to serve.

I would like to take the liberty of using this issue as the beginning of transition to their term by documenting the discussions which started when Dick Brus and I took and continue up to this point. Most of these issues were discussed and solutions were recommended at the business meeting at Pedometrics 2013 in Nairobi, Kenya. Here I share these discussions with you.

The first issue is the financial accounting for the Commission. I remember that Thorsten Behrens had much difficulty to setup an account for the Commission in Germany due to regulations on banking in Germany. It took me sometime to figure out how to set it up in USA which is much more lenient toward this type of accounts. In addition, to transfer the fund every four years is a waste not only in terms of energy but also in terms of costs because of the wiring fees. The suggested solution from the business of meeting at Pedometrics 2013 is to set up a treasurer position for the Commission and ask someone who is very stable to serve as the treasure. This treasurer is not an official position of IUSS which only elects the chair and vice-chair, but a position within the commission and for the convenience of the commission and under the direction of the chair and vice chair.

The second issue is the hosting of the Pedometrics website. The website is an important outlet and face of the Commission but it requires quite a bit of technical knowledge to maintain it. We had a series of issues with the current provider and have investigated other venues or companies for hosting the sites. Eventually, we did not move upon a request from Budiman under the impression that he is designing a new solution to it. Nevertheless, the maintenance of this website should not be the duty of the chair, nor of the vice chair, as it asks for specific expertise not required for these two positions and most people elected into these positions would not have this specific technical expertise. The recommendation from the business meeting is to set up a webmaster position with similar status in the Commission as the above mentioned treasure.

The third issue is the Pedometrics conference series.

From the Chair

There are two specific aspects to this issue. The first aspect is "too many conferences" for the pedometrics/digital soil mapping professionals. Currently, we have Pedometrics and Global Workshop on Digital Soil Mapping (from the Working Group on digital soil mapping) alternate every year. In addition, the other two working groups (the Working Group on proximal sensing and the Working Group on soil monitoring) also hold regular or semi-regular meetings. Furthermore, the regional soil science societies also have sessions on pedometrics. All of these compete for research outcomes and travel resources. As a result attendance to some of these conferences was low. It has been a challenge to attract high quality papers for the special issue of Geoderma from the Pedometrics conference due to this "thin-spread" of the research outcomes over so many similar conferences. What was suggested at the business meeting is to merge the conferences organized by the three Working Groups (DSM, proximal sensing, and soil monitoring) with the main biennial Pedometrics conferences to form a bigger conference with general pedometrics sessions and working group specific sessions.

The second aspect is the financial contribution from the Pedometrics conference series to the Commission. Currently, the Commission does not have any channels of revenue generation and its activities are very much limited. One suggestion was to have the Pedometrics conference generate a surplus to be transferred to the Commission. This can be achieved through a slight increase in conference registration fee and/or the proceeds from the pre-conference workshops. Through the latter Pedometrics 2013 generated about $US1000 for the Commission. This was the result of the efforts by Leigh, Keith and the workshop instructors (Gerard Heuvelink, Tomislav Hengl, myself). I urge future conference organizers and workshop instructors to make this effort.

I hope that I have not bored you all with these details but I felt, through my experience as the chair of the Commission, that these are important issues for the Commission to continue as one of the most vibrant and nurturing organizations in IUSS.

Best wishes,

A-Xing Zhu

News and Updates

❖ **Elected Officials of Pedometrics Commission**

Chair Elected: Dr. Budiman Minasny

Budiman Minasny is an associate professor in soil modelling. He was awarded the Future Fellowship from the Australian Research Council to develop dynamic soil-landscape models. He is interested in finding how soil change in space and time. He has an undergraduate degree from Universitas Sumatera Utara in Indonesia and a MAgr and PhD degrees in soil science from the University of Sydney.

Vice Chair Elected: Dr. Lin Yang

Lin Yang is Associate Professor of Geography, at the Institute of Geographical Sciences and Natural Resources, Chinese Academy of Sciences. Ph.D., 2009, Institute of Geographical Sciences and Natural Resources, Chinese Academy of Sciences. Her specialism is spatial sampling design and digital soil mapping; and she has been working in digital soil mapping for the past 10 years, producing over 20 research articles.

News and Updates

❖ Election of the Working Group on Proximal Soil Sensing

Dr. Raphael Viscarra-Rossel and Dr. Viacheslav Adamchuk as chair and vice chair have been leading the Working Group on Proximal Soil Sensing (PSS) since its conception in 2008. The working group is one of the most active working groups in IUSS. The WG attracted and nurtured a cohort of people who are very enthusiastic and passionate about PSS. We thank Dr. Raphael Viscarra-Rossel and Dr. Viacheslav Adamchuk for their great contribution during the early stage of this working group.

The election was held in September last year by email. There were three candidates for the chair and two for the vice chair. The candidates for chair were Marc van Meirvenne, Bo Stenberg and Zhou Shi. The candidates for vice-chair were Robin Gebbers and Abdul Mouazen. Forty people voted and Dr. Marc van Meirvenne and Dr. Robin Gebbers were elected as the chair and the vice chair of the working group, respectively. Congratulations to both of them!

Chair Elected: Dr. Marc Van Meirvenne

Dr. Marc Van Meirvenne is a professor in Soil Science, Department of Soil Management, Ghent University, Belgium. He is a former chairman of the Working Goup on Pedometrics (1998-2002) with interests and activities related to Proximal Soil Sensing. His first soil sensor was an EM38DD (first buyer in Europe, in 2000) which was used mainly to characterize soil variability for soil mapping and (precision) agriculture. Increasingly Dr. Van Mairvenne included archaeological and environmental targets into his mapping goals. He also expanded to the Dualem 21S (first buyer worldwide in 2007) and since then he specialized into the processing of multireceiver EMI images, both electrical conductivity and magnetic susceptibility. More recently (2011) he included ground penetrating radar (3D-Radar) to fully characterize the soil medium in all its EM properties (i.e. including the dielectric permittivity). Recently he purchased the Dualem 421S sensor for deeper exploration. Dr. Van Meirvenne was one of the initiators of the creation of the WG on Proximal Soil Sensing. He and his collaborators have actively participated at the biennial workshops of the WG.

Vice Chair Elected: Dr. Robin Gebbers

Dr. Robin Gebbers has studied agronomy (agro-ecology) in Rostock, Germany. His work on soils started with his Bachelor thesis on earthworm abundance in organic and conventional farming. In his Master thesis he used geostatistics, DEM and aerial imagery to predict soil properties at a very fine scale. At his first job as a scientist he took photos from a small aircraft and processed them together with DEMs in order to assess soil erosion at the field scale. In his way Dr. Gebbers became involved with precision agriculture. He did research on site-specific fertilization for four years within Germany's largest precision agriculture project "preagro". After that he was working on the comparison of geo-electrical sensors for soil mapping at the University of Potsdam with Dr.Erika Lück. He started at the Leibniz-Institute of Agricultural Engineering in Potsdam, Germany in 2006. There he finished his PhD thesis on „Accuracy Assessment in the System of Site-Specific Base Fertilization". At the same institute he became a senior scientist and the coordinator of research on "precision farming and precision livestock farming". The team comprises about 20 scientists. His research interests are in proximal soil sensing, crop sensing and spatial data analysis for precision agriculture. Among his main publications were book chapters for Margaret Oliver's "Geostatistical Applications for Precision Agriculture" and Martin Trauth's "MATLAB Recipes for Earth Sciences" as well as an invited review paper on "Precision Agriculture and Food Security" in Science, coauthored by Dr. Viacheslav Adamchuk.

News and Updates

❖ New Additions to the Advisory Board

One suggestion from the business meeting at Pedometrics' 2013 is to inject new blood into the advisory board. Below are the new additions since the conference:

Bas Kempen (Bas.Kempen@wur.nl)

Bas (1980) holds MSc degrees in Soil inventory & Land Evaluation and Geo-Information science & Remote Sensing from Wageningen University. He obtained his PhD from Wageningen University in 2011 with a thesis on digital soil mapping.

Bas currently works for ISRIC - World Soil Information where he is in charge of the SOTER programme. His research interests are in using quantitative methods for updating soil maps and sampling for validation of (soil) maps. He lives in Wageningen, The Netherlands.

Jing Liu (jliu93@wisc.edu)

Jing is a PhD candidate in the Department of Geography, University of Wisconsin-Madison. Her research interests include digital soil mapping, machine learning and data mining applied to spatial analysis and spatial high-performance/throughput computing technologies.

Brendan Malone (brendan.malone@sydney.edu.au)

Brendan is a post-doctoral researcher within the Soil Security Laboratory at the University of Sydney specializing in pedometric, chemometric and digital soil mapping and assessment research. Brendan is innately passionate about soil in general, and believes sound innovations, particularly in pedometrics and digital soil mapping, can and will contribute largely to solving many of the environmental and natural resource issues we are experiencing around the world today.

Joulia Meshalkina (jlmesh@list.ru)

Joulia is working as a senior researcher on department of Agriculture and Agroecology of Soil Science faculty of Moscow Lomonosov State University (Moscow, Russia). Fields of interest are Soil science, Ecology, Data Management, Geostatistics, Precision Agriculture, Digital Soil Mapping. I am secretary of the Pedometrics commission of Dokuchaev Soil Science Society (Russia).

Bui Le Vinh (bui_le_vinh@yahoo.com)

Bui Le Vinh is a Soil Science lecturer at Faculty of Land Management, Hanoi University of Agriculture, PhD candidate at Hohenheim University, PhD research topic: Soil mapping using fuzzy logic method for a region having strong relief variations of northwestern Vietnam.

Pierre Roudier (roudierp@landcareresearch.co.nz)

Pierre Roudier is working as a scientist at Landcare Research - Manaaki Whenua, and is based in Palmerston North, New Zealand. His research focuses on digital soil mapping, visible near-infrared spectroscopy and wireless sensor networks for the study of soils in New Zealand and in the Dry Valleys of Antarctica.

Bertin Takoutsing (B.Takoutsing@cgiar.org)

Bertin Takoutsing is a Land and Water Management Scientist at the World agroforestry Centre (ICRAF). He oversees and takes a lead role in the implementation of Land Health program and the application of Infrared spectroscopy in assessing soil quality in West and Central Africa region. Prior to joining ICRAF, Bertin Takoutsing held numerous leadership positions in development organisations and has expertise in sustainable land management.

Lin Yang (yanglin@lreis.ac.cn)

Lin Yang is Associate Professor of Geography, at the Institute of Geographical Sciences and Natural Resources, Chinese Academy of Sciences. Ph.D., 2009, Institute of Geographical Sciences and Natural Resources, Chinese Academy of Sciences. Her specialism is spatial sampling design and digital soil mapping; and she has been working in digital soil mapping for the past 10 years, producing over 20 research articles.

❖ The formation of the award committee and its charges

Based on the business meeting at Pedometrics 2013 in Nairobi, Kenya, the Award Committee for Pedometrics was reformulated. The Committee is now charged with the handling of the following two awards: the once every four year Webster Medal Award and the annual Best Paper Award.

The membership on the committee was nominated and voted by the advisory board. The makeup of the current award committee is:

Chair David Rossiter

Members (in alphabetical order of last name**)**:

- Sabine Grunwald
- Alex McBratney
- Margaret Oliver
- Lin Yang

We congratulate them for earning the trust and thank them for their contribution!

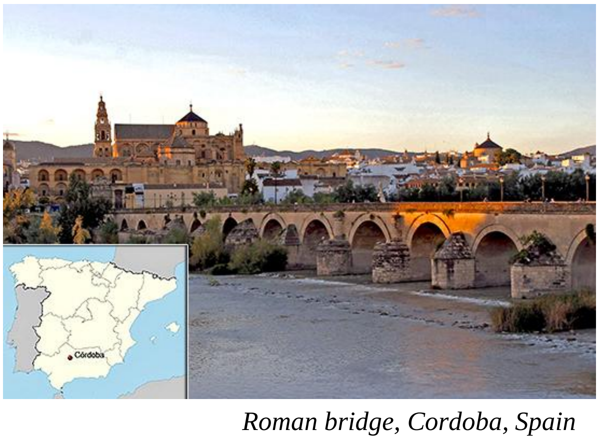
❖ The best paper awards

At the Pedometrics conference in Nairobi the best paper awards in Pedometrics were announced. These were:

- 2010: B.P. Marchant, N.P.A. Saby, R.M. Lark, P.H. Bellamy, C.C. Jolivet & D. Arrouays: ‘Robust prediction of soil properties at the national scale: Cadmium content of French soils’. *European Journal of Soil Science*, 61,144–152.
- 2011: D.J. Brus and J.J. de Gruijter: ‘Design-based Generalized Least Squares estimation of status and trend of soil properties from monitoring data’. *Geoderma*,164,172–180.
- 2012: R.M. Lark: ‘Towards soil geostatistics’. *Spatial Statistics*, 1,92–98

❖ Pedometrics 2015

Pedometrics 2015 will be held in Cordoba, Spain, September 2015, with dates to be announced soon.



Roman bridge, Cordoba, Spain

Prismatic soil aggregates from Posadas (Cordoba, SW Spain). Photo courtesy of José M. Recio (University of Cordoba, Spain).

News and Updates

❖ Recent development from SoLIM

SoLIM Solutions 2013 is the latest version of software for digital soil mapping using the SoLIM concept. If you are not familiar with the SoLIM concept, please visit http://solim.goegraphy.wisc.edu. Compared to the previous versions, SoLIM Solutions 2013 has quite a number of significant improvements and new features based on the recent research achievements from the SoLIM group both in U.S. and in China. Some highlights are listed below:

- **More flexible ways to encode expert knowledge**

  Soil mapping based on expert knowledge has always been an important feature in SoLIM. Experts encode their knowledge into fuzzy rules for soil inference. The new version provides interactive visual interface for fuzzy rule definition. It also added more methods for fuzzy rule definition, such as fitting fuzzy membership curves using key points. Those enhancements greatly facilitate the elicitation of knowledge from soil experts.

- **Knowledge extraction (data mining) from soil maps**

  Legacy soil maps serve as a valuable knowledge source for digital soil mapping. The new version of SoLIM supports the extraction of fuzzy rules (knowledge) from soil maps. Those extracted rules can be easily incorporated into knowledge based soil mapping.

- **Knowledge extraction from purposive sampling**

  For areas with no local soil surveys and no legacy maps, field sampling is needed to make soil maps. This new version supports the design of efficient sampling scheme (purposive sampling) so that it can capture soil spatial variation with a few field samples. Once the field samples are collected, knowledge can be extracted from those samples and fed into knowledge based soil mapping.

- **Soil covariate extraction from remote sensing**

  For areas with less terrain variation, commonly-used predictor variables (e.g. terrain attributes) are insufficient to predict soil distributions. The new version supports the extraction of effective soil spatial covariates from remotely-sensed surface feedback dynamics. Those spatial covariates can help to differentiate different soils and work together with other predictor variables.

- **Data-driven (sample-based) soil mapping**

  With the increasingly available soil samples from different sources (e.g. citizen science data, legacy sample achieves), data-driven approach is increasingly important in digital soil mapping. In addition to knowledge based soil mapping, this new version of SoLIM introduces data-driven soil mapping. It supports soil mapping using ad-hoc field samples. It also provides prediction uncertainty for every location a predication is made.

❖ Champagne for Jasper Vrugt

On December 20, 2013, Jasper Vrugt visited Wageningen to give a lecture and educate Wageningen hydrologists. Gerard Heuvelink took the opportunity to hand over the bottle of champagne that Jasper had won by being the first to solve the ‘Nearest Neighbour Interpolation Pedomathemagica Inverse Modelling Problem’ (see Pedometron 31). Jasper said “Just in time for the New Year celebrations!”.

Pedometrics Conference 2013 - Report
26 - 31 August 2013; co hosted by CIAT & ICRAF in Nairobi, Kenya

Showcasing innovative research and application of the mathematical, spatial and temporal modeling of soil through interactive discussions and technical sessions

I. Pre-conference Data Analysis Workshop (26-27 August 2013)

Fifty-six participants were exposed to a variety of analytical techniques for mapping soil properties and understanding spatial dependencies of soil variables. Gerard Heuvelink from ISRIC led a workshop highlighting geostatistical techniques, Tomislav Hengl from ISRIC showed products using the GSIF package, and A-Xing Zhu from the University of Wisconsin (and his team of Jing Liu, Lin Yang and Fei Du) led a hands-on tutorial on using SOLIM. This was the largest data analysis workshop hosted by the Pedometrics Division of the IUSS!

II. Pedometrics Main Conference (29-30 August 2013)

Over 65 participants, from fifteen countries were welcomed to Pedometrics 2013 by Director General of ICRAF, Dr. Tony Simons; Director of Soils Research Area at CIAT, Dr. Deborah Bossio; Dr. Anthony Esilaba, Principle Scientist at Kenyan Agricultural Research Institute (KARI), and Dr. A-Xing Zhu, Chair of Division 1.5: Pedometrics of the International Union of Soil Science. This was the first Pedometrics Conference hosted by a CGIAR centre and the first time held in the Tropics!

National media coverage included:
Kenya Broadcasting Corporation: http://www.youtube.com/watch?v=10RBTFrQz44&feature=youtu.be&a; www.scienceafrica.co.ke and our CIAT blog: http://ciatblogs.cgiar.org/soils/pedometrics-comes-to-the-tropics

Keynote addresses were also delivered by Dr. Tor Vågen, Senior Scientist at ICRAF and leader of the GeoScience Lab (gsl.worldagroforestry.org) and Marco Nocita on behalf of SOIL ACTION group at the Joint Research Centre of the European Commission.

Main topics discussed at the conference included: new approaches in digital soil mapping; uncertainty analysis; sampling design and scale; advances in proximal and remote sensing; new pedotransfer functions for tropical soils; and analytical techniques for assessing soil organic carbon stocks. At the cocktail party the best paper awards for 2010, 2011, and 2012 were given. The references to these papers are posted at www.pedometrics.org. Alexey Sorokin was awarded best poster presentation for Pedometrics 2013, presented by Dick Brus of Alterra.

Participants visited the Mpala Research Centre (MRC) on the Laikipia Plateau in north-central Kenya. En route, soils under remnant Mt. Kenya forest were viewed and different soil classification systems were discussed (e.g., Russian, Portuguese, WRB and US).

Photographed on the top right, Gerard Heuvelink is reading the WRB definition of an Acrisol, while Colby Brungard (below right) of Utah State University and David Brown of Washington State University classified the soil as a Typic Kandiusult using US Soil Taxonomy.

The Laikipia Plateau - located northwest of Mount Kenya (Africa's second highest mountain at 5,199 m) spans 10,000 km^2 and forms the core of the wider 56,000 km^2 Ewaso ecosystem. The semi-arid rangelands are important grazing lands and encompass privately owned ranches and conservancies as well as community grazing lands for the Maasai and Samburu tribes.

Maasai pastoralists explained current community grazing projects and the importance of maintaining land health due to the fragility of these soils (photo right at the gully erosion site).

At Mpala we viewed landscapes dominated by Acacia drepanolobium on vertic soils (below left). Vince Lang is photographed bottom right using HCl to identify carbonates in the soil matrix. Lunch was enjoyed at MRC on the Ewaso Nyiro river (photo bottom right).

The final profile was classified as a Calcisol in the SOTER map, but the group identified it as a Lixisols due to clay illuviation, lack of high amounts of $CaCO_3$, and a pH of 6.5.

Emeritus Russian colleague, Nataliya Belousova, is photographed in this profile (below center). Nataliya was eager to enter each soil pit and explore the properties of tropical soils!

Pre-processing, sampling and modelling (soil) vis-IR data using the 'prospectr' and 'resemble' packages

Leonardo Ramirez-Lopez[1,2] and Antoine Stevens[3]
[1]Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Switzerland.
[2]Institute of Terrestrial Ecosystems, Swiss Federal Institute of Technology (ETH) Zurich, Switzerland
[3]Georges Lemaître Centre for Earth and Climate Research, Université Catholique de Louvain, Belgium

1. Introduction

Visible and infrared diffuse reflectance (vis-IR) spectroscopy has rapidly become a popular and valuable tool among the pedometric community due to its widely known and well documented advantages over conventional methods of soil analysis (e.g. non-destructive, cheaper and faster). Soil vis-IR spectral libraries containing large amounts of soil data are being developed by different people around the world. Some of these libraries such as the (large-scale) ones developed by the World Agroforestry Centre and by the Joint Research Centre of the European commission are already freely available. In this respect, we decide to release two R packages for analyzing soil vis-IR spectra. The first package includes functions for spectral pre-processing and calibration sampling, while the second one includes functions for modeling complex spectral data. The two packages presented here were developed over the past two years gathering functions implemented for our regular work needs, courses and functions implemented just for fun during our spare time. Both packages can be downloaded the CRAN repository (http://cran.r-project.org/web/packages/prospectr/index.html; http://cran.r-project.org/web/packages/resemble/index.html) and from the package development websites (http://antoinestevens.github.io/prospectr/; http://l-ramirez-lopez.github.io/resemble/)

1.1 Why in R?

While Matlab remains, by far, the programming language of choice in the chemometric community, the use of R is rapidly expanding and seems to have already overtaken Octave and SAS (Fig.1). This increasing popularity is probably driven by the rise of R as a programming language itself (rstats.com). This trend participates to the increasing need for more reliability and reproducibility in published research. R is a free and open source software and hence R-based codes facilitates the understanding and sharing of research results (Eaton, 2012). Quoting Ince et al (2012), "… we have reached the point that, with some exceptions, anything less than release of actual source code is an indefensible approach for any scientific results that depend on computation, because not releasing such code raises needless, and needlessly confusing, roadblocks to reproducibility.". Using open source codes gives also to the researchers a greater control on their research, as they are able to inspect the code and learn about the algorithms they are using. Besides, it is somehow easier to modify and improve open source programs, therefore fostering innovation, as demonstrated by the popularity of version control systems and the concept of social coding. Even the limitations related to the lack of a graphical user interface is progressively breaking with the development programs capable of wrapping R scripts up in Graphical User Interfaces or interactive web applications.

There is now a broad array of R packages that can be used for chemometrics and spectroscopic analysis. Two CRAN task views dedicated to "Chemometrics and Computational Physics" and "Multivariate Statistics" suggest already more than 190 R packages. It is clear, however, that there are still many functions and algorithms that are not yet implemented in the R language are commonly available in Matlab or proprietary software such as WinISI (FOSS NIRSystems/Tecator Infrasoft International, LLC, Silver Spring, MD, USA) or Unscrambler (CAMO, PROCESS, AS, OSLO, Norway). Here, we make a brief overview of the capabilities of two new R packages called "prospectr" and "resemble" that implement algorithms for (i) pre-processing, (ii) sample selection and (iii) memory-based learning (a.k.a local regression).

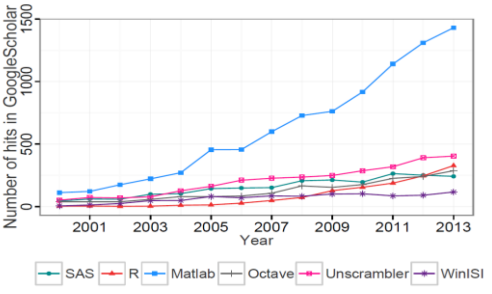


Figure 1. Number of Google Scholar hits by year using the search string: NIR + spectroscopy + name_of_software, where name_of_software is a string with the following values: "SAS Institute"-JMP; "R software" OR "R project" OR" r-project" OR "R Core Development Team" OR "R Development Core Team" OR "R package"; "Matlab" OR "the Mathworks"; "Octave"; "Unscrambler"; "WinISI".

2. The prospectr package

The prospectr package gathers algorithms commonly-used in spectroscopy for pre-treating spectra and select calibration samples. Some of the pre-processing functions are already available in other package. However, functions in prospectr package are optimized for speed using C++ code. This is especially useful when working on large databases because we often do not know which pre-processing or combination of pre-processing options will actually improve the quality of the spectra (for further chemometric analyses) so that many of them are usually tested.

The pre-processing functions that are currently available in the package are listed in Table 1. The aim of signal pre-treatment is to improve data quality before modeling and remove physical information from the spectra. Applying a pre-treatment can increase the repeatability/reproducibility of the method, model robustness and accuracy.

Table 1. Pre-processing functions available in the prospectr package

Function name	Description
`Movav`	simple moving (or running) average filter
`savitzkyGolay`	Savitzky-Golay smoothing and derivative
`gapDer`	gap-segment derivative
`continuumRemoval`	compute continuum-removed values
`Detrend`	detrend normalization
`standardNormalVariate`	standard normal variate transformation
`binning`	average a signal in column bins
`resample`	resample a signal to new band positions
`resample2`	resample a signal using new FWHM values
`blockScale`	block scaling
`blockNorm`	sum of squares block weighting
`spliceCorrection`	Correct spectra for steps at the splice of detectors in an ASD FieldSpec Pro

The prospectr package also implements several functions for selecting samples in spectral datasets. Spectroscopic models are usually developed on a representative portion of the data (training/calibration set) and validated on the remaining set of samples (test/validation set) (Table 2). There are many solutions for selecting calibration samples, for instance: (i) random selection, (ii) stratified random sampling on percentiles of a given response variable and (iii) purposive sampling based on information contained in the spectral data. The prospectr package focuses on the third solution which can ensure a good prediction performance of spectroscopic models. Generally these algorithms will create a training set having a flat distribution over the spectral space.

Table 2. Calibration sampling functions available in the prospectr package

Function name	Description
`kenStone`	Kennard–Stone algorithm (Kennard and Stone, 1969) selects the two most spectrally distant samples and then iteratively select the rest of the samples by selecting (at each iteration) the farthest sample to the points already selected
`duplex`	DUPLEX algorithm of Snee (1977) is similar to Kennard-Stone but selects also a set of validation samples that have similar properties to the training set

Table 2 (cont.). Calibration sampling functions available in the prospectr package

Function name	Description
puchwein	Algorithm of Puchwein (1988) selects samples based on their mahalanobis distance to the centre of the data
shenkWest	SELECT algorithm of Shenk and Westerhaus (1991) is an iterative method which selects samples having the maximum number of neighbours within a given distance and remove the neighours samples from the list of samples
naes	Performs a k-means clustering with the number of cluster equal to the number of samples to select and select one sample in each cluster either randomly or by another decision rule (Naes et al., 2002)
honigs	Select samples based on the size of their absorption features (Honigs et al., 1985)

In addition, prospectr provides two other functions to read binary or text files from an ASD instrument (`readASD`) and detect replicate outliers with the Cochran C test (`cochranTest`).

3. The resemble package

The initial idea of developing the resemble package was to implement a function dedicated to non-linear modelling of complex soil visible and infrared spectral data based on memory-based learning (MBL, a.k.a instance-based learning or local modeling in the chemometrics literature). Several authors have shown that MBL algorithms usually outperform other algorithms used for modeling soil spectral data (e.g. Genot et al., 2011; Ramirez-Lopez et al., 2013a).

The package also includes functions for: computing and evaluate spectral dissimilarity matrices; projecting the spectra onto low dimensional orthogonal variables; removing irrelevant spectra from a reference set; plotting results, etc. The main functions of the package are summarized in Table 3. As in the prospectr package, several of the functions included in the package use C++ code. Furthermore, some of them offer the possibility to be executed in parallel (i.e. using multiple CPU or processor cores).

Table 3. Summary of the main functions included in the resemble package

Function name	Description
Computing and evaluate spectral dissimilarity matrices	
fDiss	Euclidean, Mahalanobis and cosine (a.k.a spectral angle mapper) dissimilarities
corDiss	Correlation and moving window correlation dissimilarities
sid	Spectral information divergence between spectra or between the probability distributions of spectra
orthoDiss	Principal components and partial least squares dissimilarity (including several options)
simEval	Evaluates a given similarity/dissimilarity matrix based on the concept of side information (Ramirez-Lopez et al., 2013ab)
Functions to perform orthogonal projections	
pcProjection	Projects the spectra onto a principal component space.
plsProjection	Projects the spectra onto a partial least squares component space
orthoProjection	Reproduces either the pcProjection or the plsProjection functions
Functions for performing local modeling	
mblControl	controls some modeling aspects of the mbl function
mbl	models the spectra and predicts a given response variable by using memory-based learning

In order to expand a little bit more on the `mbl` function, let's define first the basic input datasets:

- Reference (training) set: Dataset with n reference samples (e.g. spectral library) to be used in the calibration of spectral models. Xr represents the matrix of samples (containing the spectral predictor variables) and Yr represents a given response variable corresponding to Xr.
- Prediction set: Data set with m samples where the response variable (Yu) is unknown. However, it can be predicted by applying a spectral model (calibrated by using Xr and Yr) on the spectra of these samples (Xu).

In order to predict each value in Yu, the `mbl` function takes each sample in Xu and searches in Xr for its k-nearest neighbors (most spectrally similar samples). Then a (local) model is calibrated with these (reference) neighbors and it immediately predicts the correspondent value in Yu from Xu. In the function, the k-nearest neighbor search is performed by computing spectral dissimilarity matrices between samples. The `mbl` function offers the following regression options for calibrating the (local) models: Gaussian process, Partial least squares, and two different version of weighted average partial least squares regression.

4. Short R code examples of some of the functions

For the following examples, the data of the Chimiométrie 2006 challenge (Fernandez Pierna and Dardenne, 2008) will be used. These are basic examples, therefore detailed information about the algorithms are not given here. We encourage the interested reader to have a look on the packages documentation. First, let’s install and import the packages and the data:

```
install.packages("prospectr")
install.packages("resemble")
library(prospectr)
library(resemble)
data(NIRsoil)
```

The spectral data can be smoothed by using the Savitzky and Golay filter. In this case we use a window size of 11 spectral variables and a polynomial order of 3 (no differentiation). After smoothing, the original data is then replaced with the filtered spectra:

```
sg <- savitzkyGolay(NIRsoil$spc, p = 3, w = 11, m = 0)
NIRsoil$spc <- sg
```

Different calibration sampling algorithms can be used, here we ilustrate the use of the respective functions for selecting samples based on kennard-stone and k-means (naes) algorithms. In the following example 40 samples are selected based on the first 17 principal components of the spectra:

```
kss <- kenStone(X = NIRsoil$spc, k = 40, pc = 17)
kms <- naes(X = NIRsoil$spc, k = 40, pc = 17)
```

The indices of the selected calibration samples can be accessed by:

```
kss$model
kms$model
```

The Chimiométrie 2006 challenge data already contained a column which determine the samples that should be used for training the spectral models. In the following examples we show how to predict the "unknown" values of soil total carbon (Ciso) by using the memory-based learning function included in the resemble package. First let’s remove the samples with missing values of Ciso and then split the data into train and "unknown" sets:

```
NIRsoil <- NIRsoil[!is.na(NIRsoil$Ciso),]
X.unknown <- NIRsoil$spc[!NIRsoil$train,]
Y.unknown <- NIRsoil$Ciso[!NIRsoil$train]
Y.train <- NIRsoil$Ciso[!!NIRsoil$train]
X.train <- NIRsoil$spc[!!NIRsoil$train,]
```

Now the `mblControl` and the `mbl` functions can be used to perform the local modeling and prediction process. In this example we will customize these functions to reproduce the LOCAL (Shenk et al., 1998) and the spectrum based-learner (`sbl`, Ramirez-Lopez et al., 2013a) algorithms. The correlation dissimilarity method which is used in LOCAL for nearest neighbor selection can be defined through the `sm` argument of the `mblControl` function. The validation method can be also specified through `mblControl`. In this case, the cross-validation method used is the nearest neighbor validation (NNv, Ramirez-Lopez et al., 2013a):

```
local.ctrl <- mblControl(sm = "cor", valMethod = "NNv")
```

Once the control object is defined, we can proceed with the modeling and prediction steps. In this case, we will test a sequence of ten different neighbours (from 60 to 150 in steps of 10) that we want to test for predicting the Ciso values of the unknown samples. In LOCAL the local models are fitted by using a weighted average partial least squares algorithm (`wapls`). This regression method uses multiple models generated by multiple PLS components (i.e. between a minimum and a maximum number of PLS components). For each local regression the final predicted value is a weighted average of all the predicted values generated by the multiple PLS models. In the `mbl` function, this specific regression algorithm is named "`wapls1`" and can be specified thorougth the method argument. The maximum and minimum number of PLS components must be defined in the `pls.c` argument (for this example we use the 9 and 25 as the minimum and maximum number of components).

```
eval.neighbours <- seq(60, 150, by = 10)
local <- mbl(Yr = Y.train, Xr = X.train,
Xu = X.unknown,
             mblCtrl = local.ctrl,
             dissUsage = "none",
             k = eval.neighbours,
             method = "wapls1",
             pls.c = c(9, 25))
plot(local, g = "validation", main = "LOCAL")
```

From the above results we conclude that the optimal number of neighbors is 80 (Fig. 2). Now the predictions of Ciso can be accessed by using the `getPredictions` function. The predicted values can be compared against the actual Ciso values and cross-validation statistics can be computed:

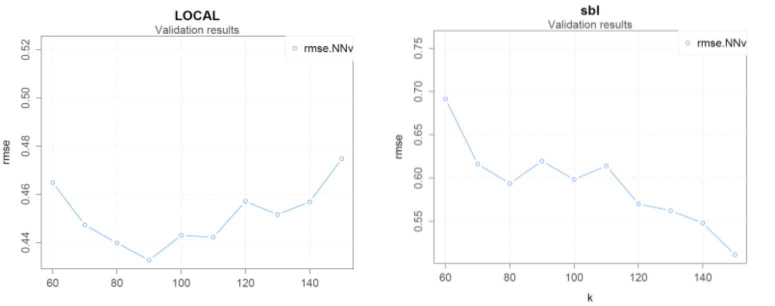


Figure 2. Results of the nearest-neighbor cross-validation for the two examples presented here. These graphs can be obtained by applying the plot function on the obtained `mbl` object. They can be used to select an adequate number of neighbors. Left: results for the LOCAL algorithm, Right: results for the `sbl` algorithm.

```
predicted.local <- getPredictions(local)$Nearest_neighbours_80
rmse.local <- sqrt(mean((Y.unknown - predicted.local)^2))
R2.local <- cor(Y.unknown, predicted.local)^2
```

Similar to the above code example for LOCAL, the functions can be customized to reproduce the `sbl` algorithm. In this case, a PC dissimilarity is used with an optimized PC selection method. The algorithm used to fit the local models is the gaussian process regression (gpr) and the local distance matrices are used as a source of additional predictors in addition to the spectral variables. The code for performing sbl predictions is as follows:

```
sbl.ctrl <- mblControl(sm = "pc", pcSelection = list("opc", 40), valMethod = "NNv")
sbl <- mbl(Yr = Y.train, Xr = X.train, Xu = X.unknown,
          mblCtrl = sbl.ctrl,
          dissUsage = "predictors",
          k = eval.neighbours,
          method = "gpr")
plot(sbl, g = "validation", main = "sbl")
predicted.sbl <- getPredictions(sbl)$Nearest_neighbours_150
R2.sbl <- cor(Y.unknown, predicted.sbl)^2
rmse.sbl <- sqrt(mean((Y.unknown - predicted.sbl)^2))
```

References

Eaton, J. W. 2012. GNU Octave and reproducible research. *Journal of Process Control*, 22(8), 1433-1438.

Fernandez Pierna, J.A., Dardenne, P. 2008. Soil parameter quantification by NIRS as a Chemometric challenge at 'Chimiométrie 2006'. *Chemometrics and Intelligent Laboratory Systems* 91, 94–98.

Genot, V., Colinet, G., Bock, L., Vanvyve, D., Reusen, Y., Dardenne, P. 2011. Near infrared reflectance spectroscopy for estimating soil characteristics valuable in the diagnosis of soil fertility. *Journal of Near Infrared Spectroscopy*, 19(2), 117.

Honigs, D. E., Hieftje, G. M., Mark, H. L., Hirschfeld, T. B. 1985. Unique-sample selection via near-infrared spectral subtraction. *Analytical Chemistry*, 57(12), 2299-2303.

Ince, D. C., Hatton, L., Graham-Cumming, J. 2012. The case for open computer programs. *Nature*, 482(7386), 485-488.

Kennard, R. W., Stone, L. A. 1969. Computer aided design of experiments. *Technometrics*, 11(1), 137-148.

Tormod, N., Tomas, I., Fearn, T., Tony, D. 2002. *A user-friendly guide to multivariate calibration and classification. NIR*, Chichester.

Puchwein, G. 1988. Selection of calibration samples for near-infrared spectrometry by factor analysis of spectra. *Analytical Chemistry*, 60(6), 569-573.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Rossel, R. A., Demattê, J. A. M., Scholten, T. 2013b. Distance and similarity-search metrics for use with soil vis–NIR spectra. *Geoderma*, 199, 43-53.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J. A. M., Scholten, T. 2013a. The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets. *Geoderma*, 195, 268-279.

rstats.com. "The Popularity of Data Analysis Software." r4stats.com. Accessed March 09, 2014. http://r4stats.com/articles/popularity/

Shenk, J. S., Westerhaus, M. O. 1991. Population definition, sample selection, and calibration procedures for near infrared reflectance spectroscopy. *Crop science*, 31(2), 469-474.

Shenk, J. S., Westerhaus, M. O., Berzaghi, P. 1998. Investigation of a LOCAL calibration procedure for near infrared instruments. *Journal of Near Infrared Spectroscopy*, 5(4), 223-232.

Snee, R. D. 1977. Validation of regression models: methods and examples. *Technometrics*, 19(4), 415-428.

Pedometricians can and should make more use of Expert Information

Gerard Heuvelink[1]
[1]Soil Geography and Landscape Group, Wageningen University,Wageningen, Netherlands.

On June 30, 13:30 hours, Phuong Truong from Wageningen University will (likely) defend her PhD-thesis "Expert knowledge in geostatistical inference and prediction". The aula of the university easily seats 500 people, so why don't all pedometricians from around the world come and witness this event? Please consider yourself invited! You will need to make your own travel arrangements and unfortunately we cannot support you financially, but surely that will not stop you from making this exciting trip. Wageningen in June is even more 'gezellig' than usual and a very nice place to be. On top of that, you will learn about how pedometricians can make more use of expert knowledge in geostatistical modelling and prediction.

Now if for some reason you won't be able to make it, let me briefly explain what Phuong did over the past four years. She worked on four main topics, and I describe all four below. If you are interested you can also send an email to Phuong (phuong.truong@wur.nl) or me (gerard.heuvelink@wur.nl) so that we can send you a digital copy of her thesis.

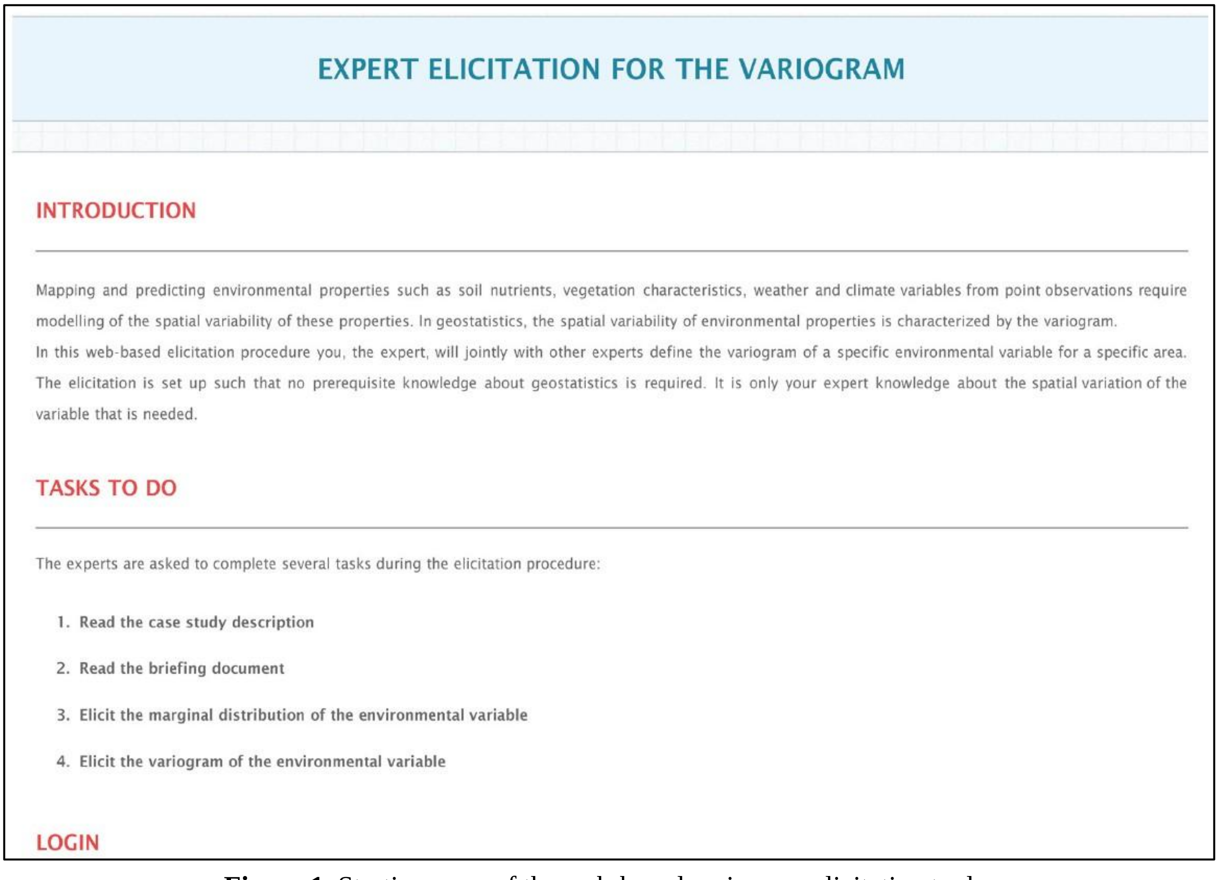
1. Web-based tool for expert elicitation of the variogram

We all know that the variogram is the keystone of geostatistics and a prerequisite for kriging. We also know that quite a few observations are needed to estimate the variogram reliably. Dick Webster and Margaret Oliver write in their book that a minimum of about 200 observations are required, and perhaps this number can be reduced if we replace Matheron's Method of Moments estimator by (Restricted) Maximum Likelihood estimation, but still there will be many practical cases where there simply are not enough observations to estimate the variogram from only point observations. Alex, Budi and Brendan (surnames not needed) contributed a very interesting item to Pedometron 32, in which they describe a method to guess the variogram. Phuong basically did the same, but rather than guessing, she used a formal statistical expert elicitation approach. As it happens, this is a research field in its own right, and we can learn a lot from the expert elicitation research community. Formal expert elicitation provides a sound scientific basis to reliably and consistently extract knowledge from experts. Phuong developed an elicitation protocol for the variogram of an environmental variable and implemented it as a web-based tool. The protocol has two main rounds: elicitation of the marginal probability distribution and elicitation of the variogram. The first round extracts from experts quantiles of the marginal probability distribution by asking questions such as "Which is the lowest possible value of Z?" and ""What is the value Z_{med} such that there is a 50% probability that the value of Z is less than or equal to this value?". Of course, experts have first been prepared for their task using a briefing document that explains probabilistic terms and aims to avoid various types of judgement bias. The second round faces the problem of extracting the variogram from experts without geostatistical background. How to do this? The approach used by Phuong is to elicit the experts' judgement about the amount of variation that can be expected at various spatial lags. This involves questions such as: "Could you specify a value T such that there is a 50% probability that the spatial increment is less than or equal to this value for each of the following distances?". Here, it was first explained to experts what 'spatial increment' means: it is the absolute value of the difference between the values of Z at two locations separated by the given distance. That is easy, don't you agree? Once all experts had completed the elicitation task (in the case study Phuong used five experts), their judgements had to be pooled. For this, Phuong used 'mathematical aggregation', which basically boils down to taking a weighed average. The alternative is 'behavioural aggregation', which may be interpreted as locking up all experts together in a room and not letting them out until they agree. Since a PhD must be completed in a reasonable amount of time, Phuong chose mathematical aggregation.

2. Uncertainty quantification of soil property maps with statistical expert elicitation

This part of Phuong's research applied the tool developed in the first part to elicit from experts the accuracy in a given, deterministic soil property map. We all know that soil property maps are not perfect, and that it is important to know how close they are to reality to be able to tell for which purposes we can use them and for which not. Now we also know that it is not enough to characterise the uncertainty associated with a soil property map by that single, 'magical' number, known as the Root Mean Squared Error. To fully characterise the uncertainty, we need to know more, among others the variogram of the map error. It can be derived from a sufficiently large number of point observations of the error in the map, but what if there are no observations?

Surprise, surprise: ask the expert! And so this is what Phuong did, using a NATMAP map of the volumetric soil water content at field capacity (%) of the East Anglia Chalk area as a case. Six experts were asked to elicit the marginal distribution and the variogram of the error in the map. This was more difficult than the elicitation in the first study, because ‘error’ is less tangible than a real environmental variable, such as ‘temperature’ or ‘clay content’. Experts had quite different opinions about the magnitude of the uncertainty in the soil property, their interquartile ranges varied between 4 and 30%. It was also odd that two out of six experts judged the median of the error bigger than zero. This effectively meant that they reasoned that the NATMAP was positively biased, because the error was defined as the true value minus the mapped value. The elicited nugget-to-sill ratios were small and fairly similar for five out of six experts, but one expert was of the opinion that the nugget was 90% of the sill. This was the only Dutch expert (all others were British), it may reveal the ‘noisy’ character of Dutch soil scientists, but as yet this is only a hypothesis that needs further research.



EXPERT ELICITATION FOR THE VARIOGRAM

INTRODUCTION

Mapping and predicting environmental properties such as soil nutrients, vegetation characteristics, weather and climate variables from point observations require modelling of the spatial variability of these properties. In geostatistics, the spatial variability of environmental properties is characterized by the variogram.
In this web-based elicitation procedure you, the expert, will jointly with other experts define the variogram of a specific environmental variable for a specific area. The elicitation is set up such that no prerequisite knowledge about geostatistics is required. It is only your expert knowledge about the spatial variation of the variable that is needed.

TASKS TO DO

The experts are asked to complete several tasks during the elicitation procedure:

1. Read the case study description
2. Read the briefing document
3. Elicit the marginal distribution of the environmental variable
4. Elicit the variogram of the environmental variable

LOGIN

Figure 1. Starting page of the web-based variogram elicitation tool.

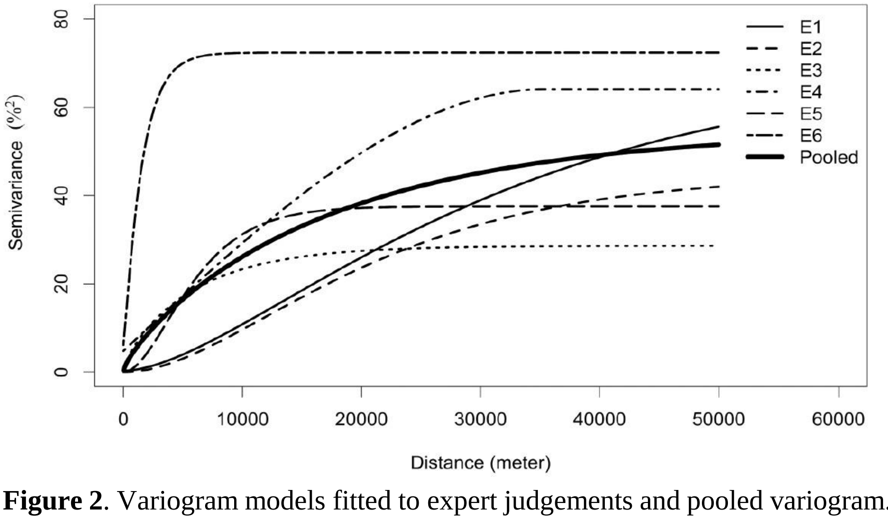
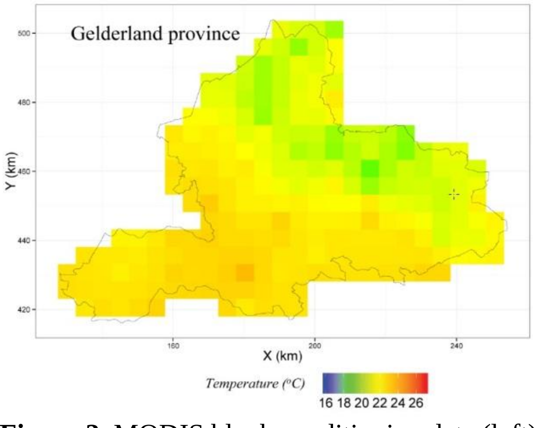


Figure 2. Variogram models fitted to expert judgements and pooled variogram.

3. Bayesian area-to-point kriging using expert knowledge as informative priors

The third part of Phuong's research focused on elicitation of the nugget variance and was done in the context of area-to-point kriging. This is a recently developed technique (among others work by Phaedon Kyriakidis, Carol Gotway, Pierre Goovaerts), that was introduced to soil science by Ruth Kerry and co-workers in a 2012 Geoderma paper. In area-to-point kriging we do the opposite of block kriging: we predict at points from observations at blocks. Most relevant applications are in remote sensing and climatology, where the output of Global Circulation Models may need to be scaled down to smaller geographical units. Area-to-point kriging works very fine (albeit slow), but the problem is that in order to apply it, one needs to point support variogram. There are people (PG even wrote code that can do the job) who claim that the point support variogram can be derived from block support data, but this is not true. The point support nugget cannot be inferred because the micro-scale variability cancels out when block averaging is done, and so one can never tell from block support observations how large the micro-scale variability was. So, what to do? Phuong proposed, and this will be no surprise, to use expert elicitation. In this case, the block support data also provide information (notably about the variogram sill and range), which was included by taking a Bayesian approach. The expert-elicited variogram was taken as a prior, and next a posterior was calculated by multiplying the prior with the likelihood of the block observations. The results nicely show that the block data provide no information about the nugget, because the prior and posterior of the nugget parameter were nearly identical. However, block data could reduce the prior uncertainty about the range and sill. All this was applied to downscaling MODIS temperature air data, but there should be interesting soil applications too, such as downscaling of remotely sensed base soil moisture data.



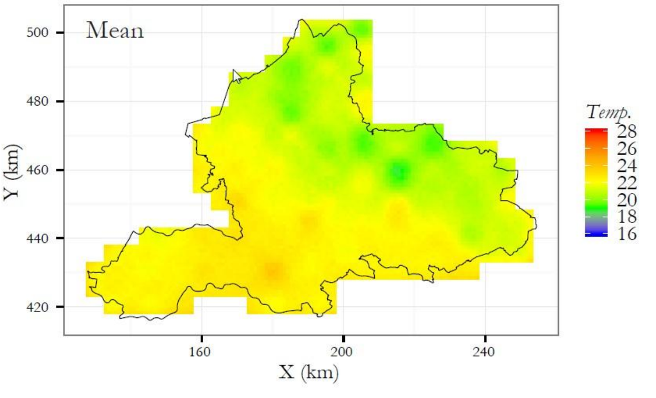


Figure 3. MODIS block conditioning data (left) and downscaled point predictions right) of air temperature.

4. Incorporating expert knowledge as observations in mapping biological soil quality indicators with regression cokriging

The fourth and final part of Phuong's thesis addresses the use of experts to generate new data. This is not a new idea, for instance Henning Omre already published a paper on merging observations and qualified guesses in kriging in 1987. Phuong used formal expert elicitation techniques as before, and used the expert-generated data not only in kriging but also for variogram estimation. The difficulty of course is in the fact that expert data are not the same as real data and that a naive merge would rather deteriorate results than improve them. Expert judgements can have large errors and so these should be treated as 'soft' data. Moreover, experts tend to 'smooth' reality, which affects the spatial correlation structure. Phuong came up with a model that takes all this into account, and applied it to mapping nematode indices for a Dutch nature area. In this case she used just one expert, and as it happens this expert was quite uncertain about the nematode index in the area, and so the added value of the expert information was only marginal.

So what did we learn from Phuong's research? It showed that expert information can be very useful in geostatistical and pedometric research, and that it should be done using formal expert elicitation procedures. But of course there is much more than that. If you want to know the rest, you will simply have to come to Wageningen on June 30. If you do, I will buy you a drink, that's for certain!

Updating the Soil Map of the Netherlands 1:50,000 by pedometric techniques

D.J. Brus*, F. de Vries and B. Kempen
Alterra, part of Wageningen University and Research Centre,
P.O. Box 47, 6700 AA Wageningen, The Netherlands
*Correspondence: D.J. Brus. E-mail: dick.brus@wur.nl

1. Introduction

In the 1960s the Soil Survey Institute of The Netherlands started the national soil survey at scale 1:50,000. In 1995 the last sheet of the map was published. The age of the map therefore varies from 20 to more than 40 years. According to this soil map more than half a million of ha of soils contain peat within 40 cm of the surface. These soils are mapped as peaty soils (thickness of peat < 40 cm) or as peat soils (thickness peat > 40 cm). Since the soil survey, part of the peat has disappeared through oxidation and ploughing. As a consequence peat soils may have changed into peaty soils, and peaty soils in to mineral soils. A reconnaissance survey in the eastern part of the Netherlands showed that the acreage of peat soils was reduced by about 50%. This clearly showed the need for an updated soil map. Actual information on the thickness of peat layers is of great importance for the evaluation of many soil functions such as agricultural production, carbon stock inventories and modelling studies on nutrient leaching.

For updating the soil map six soil-geographical subareas were distinguished. Until now the soil map has been updated in the two northernmost subareas, the northern till plateau of the provinces of Drenthe and Friesland (subarea 1), and the Fen peat soils in the transition zone of the till plateau to the marine clay soils in the west and north (subarea 2). These two areas cover 188,000 ha (Figure 1).

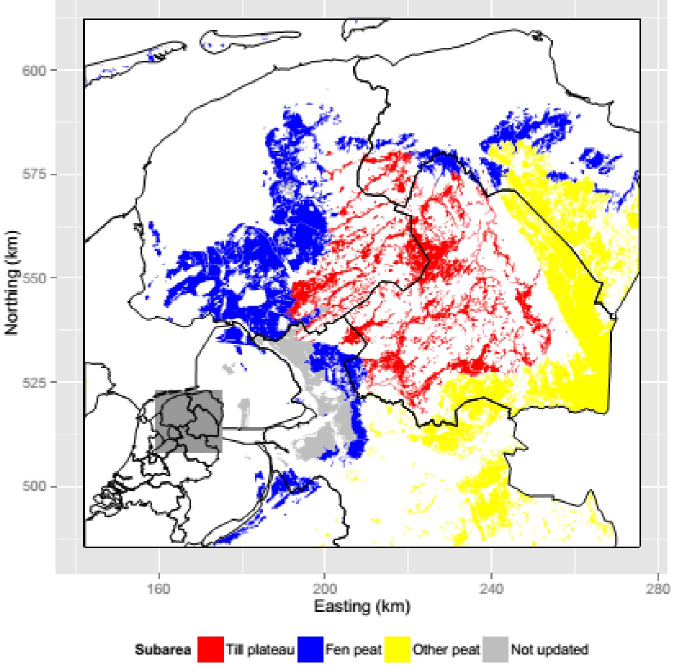


Figure 1. Peat areas in the north of the Netherlands.

2. Methodology

Figure 2 shows the various steps of the procedure used for mapping the actual peat thickness. We will describe the steps now shortly.

2.1. Constructing maps with covariates

The first step is the construction of maps with environmental covariates that are possibly related to the actual peat thickness. The covariate layers have been derived from a variety of data sources. Layers with information about peat type, peat thickness class, topsoil type, and sensitivity to oxidation were derived from the current 1:50 000 soil map. Seven groundwater table layers were used that indicate overall, summer and winter drainage conditions. Eleven relative elevation layers (5 continuous, 6 categorical) were derived from the 25-m resolution national DEM. Nine land cover layers were derived from the 25-m resolution national land use inventory. This inventory was simplified and reclassified into layers with different combinations of land cover classes. Eight layers representing historic land cover were derived from land cover maps from 1940, 1960, 1970, 1980 and 1990. Finally, these five historic land cover maps were used to estimate land reclamation age (conversion from nature to agricultural land), which resulted in seven layers with different combinations of reclamation age classes.

2.2. Collecting additional point data

The next step is the collection of additional soil profile descriptions at point locations. In subarea 1 one sampling location per 150 ha was selected. In subarea 2 the sampling density is one per 75 ha or, in the part of this subarea with thick peat soils, one per 50 ha. These additional locations were selected by spatial coverage sampling (Walvoort et al., 2010).

2.3. Updating peat thickness in legacy soil profile data

The thickness of peat layers in the soil profiles stored in the Soil Information System were updated with the model described by (Kempen et al., 2012):

$$z_{ti} = z_{0i} \cdot p_i^t$$

with z_{ti} the thickness t years after the soil profile i was described, z_{0i} the original peat thickness in soil profile i , p_i the proportion of the peat thickness in soil profile i remaining after one year, and t the time in years elapsed since the soil profile description. t equals 2011 (subarea 1) or 2012 (subarea 2) minus the year in which the soil profile was described. (Kempen et al., 2012) modelled the proportionality constant p by a non-spatial generalized linear model (GLM) with a logit link function, accounting for over- or under-dispersion:

$$p_i = \pi_i + \varepsilon_i$$

$$logit(\pi_i) = x_i^T \beta$$

$$E[\varepsilon_i] = 0$$

$$Var[\varepsilon_i] = \sigma^2 \pi_i (1 - \pi_i)$$

(Kempen et al., 2012) found no relation between $logit(\pi_i)$ and environmental covariates x. We used the values of $\hat{\pi}$ and $\hat{\sigma}^2$ as reported by (Kempen et al., 2012) to simulate 10,000 values for pi , using a beta(a,b) distribution. These simulated values were subsequently used to compute 10,000 updated peat thicknesses per legacy soil profile. These updated peat thickness were log-transformed, and the mean and variance of these log-transformed updated peat thicknesses were computed. The variance of the simulated peat thicknesses reflects our uncertainty about the actual peat thickness. This uncertainty was accounted for in spatial prediction of the peat thickness (see hereafter). Note that the larger t, the larger the variance of the simulated peat thicknesses, the smaller the weight attached to these data in spatial prediction.

2.4. Selection and calibration of a model for presence/absence of peat

Especially in subarea 1, a considerable proportion of the soil profiles contained no peat (peat thickness of 0 cm). Such zero-inflated distributions can be modelled by a mixture of two distributions, a Bernoulli distribution for the presene/absence of peat and a conditional distribution for the thickness of peat, conditional on the presence of peat. In building this model, also the legacy soil profile data with updated peat thicknesses were used in addition to the newly collected data. For these legacy

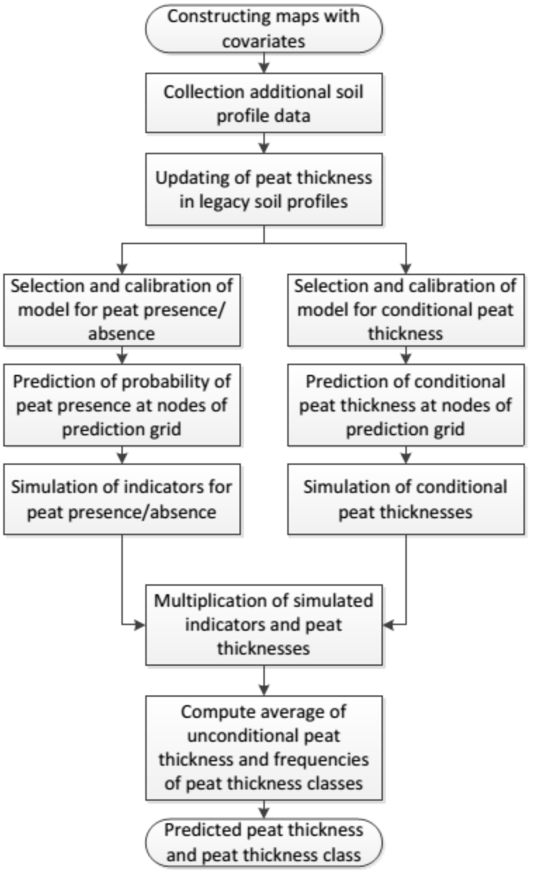


Figure 2. Steps in updating the soil map of peat thickness.

soil profile data a threshold value of 1 cm for the updated peat thickness was used: if the updated peat thickness (average of 10,000 simulated values) in a soil profile was less than 1 cm, then the peat indicator value for this soil profile was 0 (peat absent).

The presence/absence of peat was modelled by a non-spatial GLM model with a logit link function. The best model, i.e. the best combination of covariates, was selected on the basis of Akaike's Information Criterion (AIC). The most important covariate was the peat thickness in three classes as derived from the existing (not updated) soil map. Other selected covariates were, amongst others, groundwater table depth, relative elevation, reclamation period, and (historic) land use. Model residuals showed only very weak spatial correlation: in area 1 the relative nugget was very large, in area 2 the experimental variogram fluctuated around a horizontal line (see Figure 3).

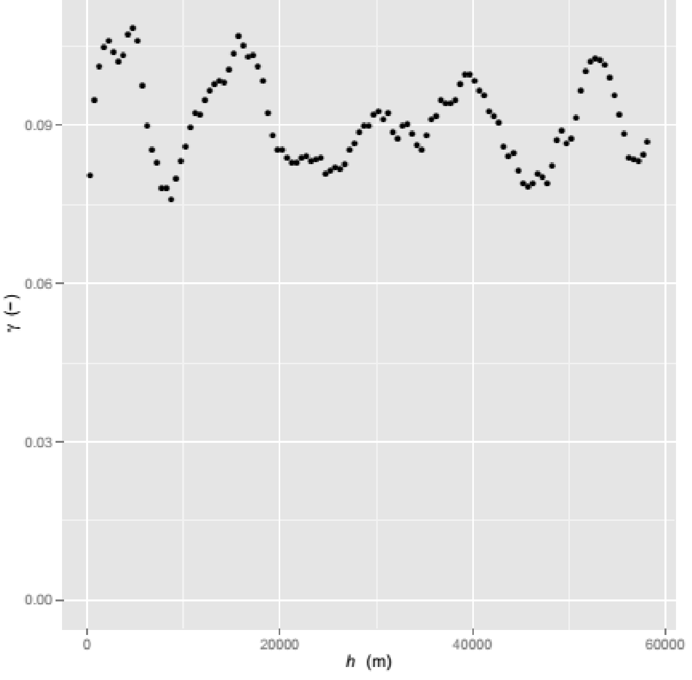


Figure 3. Experimental variogram of residuals of GLM for peat indicator in subarea 2.

2.5. Prediction of peat presence/absence and simulation of indicators

The calibrated GLM model was then used to predict the probability of peat presence at the nodes of a 50 *50 m grid. Notice that the probabilities as predicted by the GLM are automatically in the interval (0, 1). Next, at each grid node 1000 peat indicator values were simulated from a Bernoulli distribution, with the predicted probability as the "probability of success". The indicators at the grid nodes were simulated independently from each other, i.e. spatial correlation was not accounted for (no geostatistical simulation). Geostatistical simulation is not needed when results are not aggregated.

2.6. Selection and calibration of a model for conditional peat thickness

The next step was the modelling of the thickness of peat conditional on the presence of peat. So for building this model only soil profiles with peat were used. The updated peat thicknesses were log-transformed, so that the distribution became less skewed. The spatial distribution of these log-transformed peat thicknesses was modelled by a linear mixed model, i.e. the sum of a linear combination of covariates (linear trend) and a spatially correlated residual. The best trend model was selected using AIC, assuming uncorrelated residuals. The linear mixed model was then calibrated by iterative Generalized Least Squares.

2.7. Prediction and simulation of conditional peat thickness

The calibrated linear mixed model was then used to predict the conditional peat thickness at the nodes of the prediction grid. In geostatistical literature this is referred to as kriging with an external drift. The uncertainty about the updated peat thickness in the observed soil profiles was accounted for in kriging, by adding the variance of the 10,000 simulated peat thicknesses, see section 2.3, to the diagonal of the covariance matrix of the data used in kriging. The predicted log-transformed peat thickness and its prediction variance were then used to simulate 1000 values per node, assuming a normal distribution. These simulated values were then back-transformed by exponentiation.

2.8. Prediction of peat thickness and peat thickness class

For each grid node 1000 unconditional peat thicknesses were obtained by element-wise multiplication of the vectors with simulated indicators and conditional peat thicknesses. Next, the mean, median 5-percentile and 95-percentile of the resulting values were computed. Besides, the frequencies of three peat thickness classes, < 5 cm, 5 - 40 cm, and > 40 cm were computed. The thickness class with the largest probability was used as the predicted peat thickness class.

2.9. Validation

Predictions were validated by a probability sample not used in mapping (Brus et al., 2011). The probability sample was selected by stratified random sampling, using the predicted peat thickness for stratification. Within each stratum several blocks of 50*50 m were selected, and within these selected blocks several point locations (stratified two-stage random sampling).

3. Result

Figure 4 shows the predicted actual peat thickness for subarea 2. The square root of the mean and median squared errors of the predicted peat thickness at block-support for subarea 2 were 51 cm and 18 cm, respectively. The correlation between predicted peat thickness and observed peat thickness was 0.77. The overall purity of the three peat thickness classes was 72%.

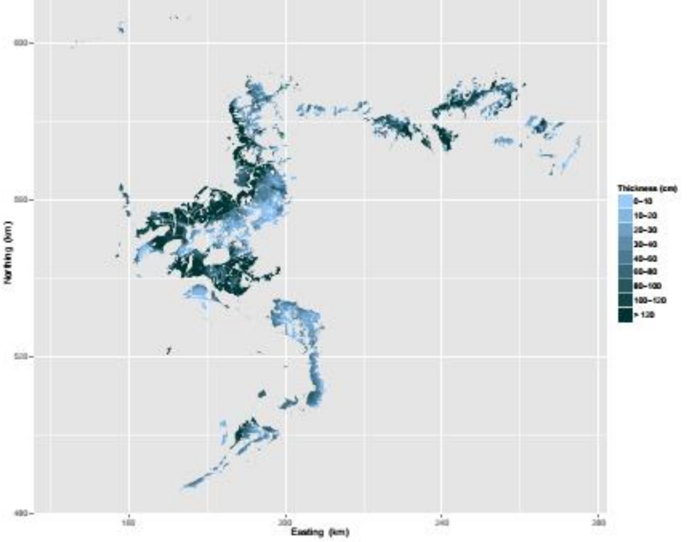


Figure 4. Predicted actual thickness of peat layer in subarea 2.

Reference

Brus, D. J., Kempen, B., and Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. European *Journal of Soil Science*, 62(3):394 - 407.

Kempen, B., Brus, D. J., de Vries, F., and Engel, B. (2012). Updating legacy soil data for digital soil mapping. In Minasny, B., Malone, B. P., and McBratney, A. B., editors, Digital Soil Assessments and Beyond. *Proceedings of the fifth Global Workshop on Digital Soil Mapping*. Sydney Australia, 10-13 April 2012.

Walvoort, D. J. J., Brus, D. J., and de Gruijter, J. J. (2010). An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers and Geosciences*, 36:1261{1267.