ΠΕΔΟ METRON

The Newsletter of the Pedometrics Commission of the IUSS

Issue 31, June 2012

Chair: A-Xing Zhu **Vice Chair: Dick. J. Brus**
Coordinator: Murray Lark **Layout: Jing Liu**

From the Chair

Dear Fellow Pedometricians,

This is the first issue since I take over as the Chair of the Pedometrics Commission. The challenge for keeping Pedometron on time is the collection of materials. Murray Lark generously offered himself to help with the collection of materials for Pedometron. Jing Liu, from University of Wisconsin-Madison has volunteered to do the layout of the materials. I think we have a good team but we still need your active contribution to make Pedometron alive, interesting and stay on time.

While this issue of Pedometron was delayed but pedometricians have been very busy. We had a very successful Pedometrics 2011 in Trest. In this issue we have two reports on this. At this meeting Jaap de Gruiter received the 2010 Richard Webster Medal for his outstanding achievements in pedometrics. The medal also put him in par with his grandson in terms of being a owner of metals. In February we commerated in Nanjing the Launching of the Global Soil Partnership's Asia Soil Science Network and GlobalSoilMap.net East Asia Node. In March the Global Soil Partnership met in the UN FAO headquarters (Rome, Italy). Barry Rawlins provides in this issue a very nice summary of this meeting. In April, the 5th Global Workshop on DSM was held in Sydney and the pictures provided in Jenette Goodman's report are hoped to give a lively taste of another successful event.

For those of you who just cannot get enough of statistics we have in this issue three major pieces to meet your craving: "From Inverse Modelling to Soil Geostatistics" from Budiman, "Dealing with Below Quantification Limit Data in Geostatistical analyses" by Thomas and his colleagues, and that by Dick on sampling.

The "On Trying to Bridge a Gap " by Jaap iterated the importance of teaching and the impacts teachers on individuals and in turn on the advancement of science. It is a refreshing article to read and to enjoy.

Speaking of enjoy, you should take a bit of time to solve the Pedomathemagica problems. Be sure to remember that the answer to the Problem 2 of Pedomathemagica is not ZHU.

Best wishes,

A-Xing Zhu

Inside this Issue

Reports From Pedometrics 2011: Innovations in Pedometrics: Part I

From Eef Meerschman
Department of Soil Management
Ghent University, Belgium

Once upon a time 81 pedometricians from 27 different countries met in a castle to participate in a roundtable meeting on the future of pedometrics. For three days they tackled new ideas and discussed the holy grail of soil modelling...

The conference 'Pedometrics 2011: Innovations in Pedometrics' took place from 30 August to 2 September in Třešť, Czech Republic. The castle was situated in a picturesque forest park in the middle of Třešťské Vrchy, a part of the Czech-Moravian Higlands. This meant the participants had to travel for more than 150 km after arriving at Prague's airport. Luckily, this transportation was perfectly organized by the Organizing Committee –better known as Ondřej– of the Czech University of Life Sciences Prague, lead by Luboš Borůvka.

During the opening ceremony Thorsten Behrens already revealed that Pedometrics 2013 will take place in Kenya, thereby visiting the only continent where Pedometrics never took place before. Then, the conference took off with the first keynote of Jasper Vrugt, who introduced several of his –freely available– algorithms (a.o. MCMC-DREAM) for the uncertainty quantification of environmental models. The second morning Ben Marchant delivered an enriching keynote lecture on the challenges of soil monitoring, which is the main interest of the recently created working group on Soil Monitoring. A highlight of the conference was the keynote by Jaap de Gruiter, nicely entitled 'Gapping the bridge'. After a 43 years long career in pedometrics he is the proud owner –not least to impress his grandson– of the Richard Webster Medal 2010. In the programme of the conference were enclosed 46 oral presentations, divided into eight sessions going from 'Pedometrical methods for soil assessment' to 'Signal processing of remote and proximal sensing applied to soils'. The participants took part in lively discussions about the no less than 38 posters during two poster sessions. The best poster contest was won by J. Balkovič, followed by A. Akramkhanov and P. Roudier.

A dozen of participants also attended to the pre-conference workshop (28 – 29 August) 'Bayesian Inverse Modelling in the Earth Sciences: Theory, Concepts and Applications' taught by Jasper Vrugt and Sander Huisman. This enthusiastic duo introduced us to the world of Bayesian parameter and state estimation methods and taught us how to work with the DREAM (DiffeRential Evolution Adaptive Metropolis) algorithm.

Not only the high scientific level, but also the smooth organization made Pedometrics 2011 –my first Pedometrics– so successful. We were spoiled with delicious Czech food and drinks and two wonderful Czech music performances. The fact that everyone stayed in the same hotel, created a friendly atmosphere and made it easy for young scientists to become acquainted with the Pedometrics community.

-end-

Reports From Pedometrics 2011: Innovations in Pedometrics: Part II

From Thomas Bishop, The University of Sydney

Pedometrics 2011 was held in the Czech Republic and for me was my second Pedometrics Conference, the first being in 1999 during my PhD. My conference began with the pre-conference workshop in Prague on Bayesian Inverse Modelling presented by Jasper Vrugt and Sander Huisman. This was an inspired choice by Lubos and the Organising Committee as Bayesian statistics is gaining in popularity but can be difficult to implement. Based on my experience at the workshop the DREAM algorithm and associated R and MATLAB code could solve this problem, assuming an objective function can be encoded by the Pedometrician…… From there we were first on the bus for the 2 hour trip to Trest Castle via the airport where the first wave of participants had arrived. The next day was the field trip which ended in the beautiful town of Telc followed by the late night arrival of the second wave of participants.

Over the next two and half days the 81 participants from 27 countries listened to 46 oral presentations and 3 keynote presentations; Jasper Vrugt on Bayesian Modelling, Ben Marchant on Monitoring and finally Jaap de Gruitjer who presented the keynote on the last day to commemorate receiving the 2nd Richard Webster Medal. In addition there were 38 poster presentations. The talks were diverse covering a range of topics reflecting the diverse field that Pedometrics has become. These ranged from geostatistics (Meerscham, Horta), scaling (Roudier, Biswas), 3D modeling (Lacoste, Wheeler), Bayesian methods (Orton), monitoring (Brus), sampling (Lark), uncertainty (Zhu), modeling fine-scale variation (Gerke) and sensing of soil (Lagacherie). A conference is a personal experience and for me the highlight was the intimate nature of the workshop which made it easier to meet new and old friends. This was partly due to the number of participants but also due to the choice of venue where for 3 days we were all together in Trest Castle with nothing to do but interact! In recent times I have been going to too many conferences in large cities with 1000+ participants. This was a nice change.

All things can be improved and I do wonder if the 15 minute talk + 5 for questions is too long for such a specialized conference. For example, I don't need to be told about the importance of soil carbon or digital soil mapping. Shorter talks may have been punchier and given participants with posters a chance to present. The other concern is the number of people that gave two talks, many of these were for late withdrawals but I do wonder if some of the poster presenters could have been placed on a reserve list. I know some conferences only allow one talk per person. This is something to think about for the future.

So has Pedometrics changed in the 12 years between my 2 Pedometrics Conferences? The methods have seemed to become more complex and diverse. I seem to remember a lot of geostatistics in 1999 but now we have so many more tools. The diversity of tools is good but one of the side effects of the complexity is that it seems the gap is widening between those at leading edge of using new methods (copulas, Bayesian methods) versus the practitioner. In 1999 a numerate soil scientist could be somewhere near the leading edge in Pedometrics but this seems to be more difficult today where vast experience or formal training in statistics is required. This may not be a bad thing but if the gap between the leading edge of Pedometrics and the practitioner does become too large then how will relevant is Pedometrics to the typical soil scientist. One way to avoid this is to keep ensuring we are using Pedometrics to solve real problems, of concern to soil scientists and the wider community.

Finally I would like to thank Lubos and his helpful and good natured team of students who helped organise and run the conference.

-end-

A report from the Global Soil Partnership Meeting (Rome: 20-23rd March 2012)

From Barry Rawlins
British Geological Survey (UK)

I attended the recent Global Soil Partnership (GSP) meeting in Rome. Rather than write this as a free text article, I have prepared a set of Q&A's so the text is a bit easier to navigate.

What is the aim of the Global Soil Partnership?

To support and facilitate joint efforts towards sustainable management of soil resources for food security and climate change adaptation and mitigation. Its second pillar is to "strengthen soil data and information: data collection, validation, reporting, monitoring and integration of data with other disciplines". The aim of the meeting in Rome was to develop an agreed approach to this second pillar.

Who leads or manages the GSP?

The FAO, with input from a broad set of stakeholders.

What was the aim of the Rome meeting?

To review current status of soil information at global and regional level, establish an improved knowledge of the current soil mapping initiatives and state-of-the art tools and methods for soil mapping and information dissemination.

How was the GSP meeting structured?

In four sessions: 1) status and needs of global soil info., 2) tools for polygon based mapping – much of this from the recent EU project e-SOTER. 3) tools for point based mapping, 4) a way forward for Global Soil information.

What were the main points from the FAO keynote address (by Parviz Koohafkan**) ?**

Soil activities and soil mapping have declined in recent years – with negative impacts on decision making. Soil has been dormant in FAO. But soils are now back on the international agenda. We need to make good use of technologies and methods to provide quality soil information.

Was Intellectual Property (IP) often associated with soil data seen as a problem / impediment to wider development of soil information?

Yes, this issue was raised on several occasions. Specifically, during discussions at which EU representatives were present. Although the EU funded e-soter project had delivered the methods to produce an EU wide soil map, there is currently no way to deliver this because of IP restrictions. It was clear that IP restrictions limited the use and development of soil data in many parts of the World, not just in Europe. There was no clear strategy for addressing this issue which seems to have been an impediment for many years. One of the objectives of Globalsoilmap.net is to overcome some of the impediments associate with IP relating to historical soil data.

What global soil information currently exists (Freddy Nachtergaele (FAO/IIASA) **)?**

The Harmonized World Soil Database (HWSD) maintained by the FAO. This is in urgent need of updating with 1) improved geographical coverage, 2) improved Quality of Soil Property predictions, and 3) improved Harmonization in cooperation with e-SOTER and Globsoilmap.net.

What points were raised after the presentations relating to the e-soter project?

It was suggested that in addition to the methods of digital soil mapping relying on terrain information, there were other sources of data such as geological information which might be usefully utilised to further develop soil maps. An update of the 1:1 million scale soil map of Europe is the current goal for using the methods developed by e-soter and this would feed into the HWSD.

What were identified as the ways of moving forward (Luca Montanarella-JRC) ?

Significant improvement of global soil information depend on certain conditions: 1)neutral leadership, 2) intellectual property rights on data remain with the data producers, 3) participatory working practices (Networking), 4) national capacities exist, 5) clear user needs are established. 6) Take a long term

perspective. This may help to develop a version 2.0 of the HWSD at 1km resolution and version 1 of GlobalSoilMap.net.

Is sufficient attention paid to the users of the data – is there agreement on who they are?

It was suggested that more attention should be paid to end users such as farmers, and the food security situation. Others suggested that the main users of the data were climate change and food security modellers. Another view as that large commercial firms have a great interest in the derived products from soil information. It was concluded that both kind of end users would benefit directly or indirectly of enhanced soil information. Members from ISRIC said they would continue to serve the international community as a world soil data centre and the vital role in the organization of the data structure in the GSP pillar, but this should be seen as a collective effort.

What were seen as the next steps?

Generation of a working paper covering the following issues:

- ❖Governance and Structural Organization
- ❖The links between Global soil information and end-users
- ❖Primary soil data and spatial data products including accuracy issues.
- ❖Reporting on global soil health: soil capacity and functions.
- ❖Technical monitoring
- ❖Global monitoring network
- ❖Archives, References and standards

A drafting committee was appointed to deliver this document by May 2012! This would then go out to consultation in June 2012 with the intention that the report would be available in time for the Rio+20 meeting.

……and what presentation did you give, Barry?

I gave a presentation on the new Soil App which the British Geological Survey and our partner institute (the Centre for Ecology and Hydrology) have developed initially for the UK but which in future will be developed to provide soil information directly to users with tablets and smartphones so that soil information can be used by everyone, anywhere. The App, called mySoil, has just been submitted to the App Store. I learnt whilst at the meeting that there are already such Apps available in the USA. One of the major potential advances is for users to supply their own soil data from the field based on their own observations or use of sensors. In addition to the soil App, my colleagues at the British Geological Survey recently developed an Augmented Reality approach to displaying geological map information superimposed on the image in the viewfinder of a smartphone. They are planning to develop this for soil-related information as well. I think the educational potential of this technology is amazing particularly when during fieldtrips this could be used as an aid to landscape and soil variation.

Caption: Icon for the new soil App which has been jointly developed by BGS and CEH in the UK

Caption: Augmented reality visualization – here geological boundaries are superimposed on the landscape and the same approach is planned for showing soil units

-end-

Report from the 5th Global Workshop on Digital Soil Mapping

From Jenette Goodman
Purdue University

Held from April 10th to the 13th of 2012, the Fifth Global Workshop on Digital Soil Mapping took place in the famed Australian city, Sydney. Participants from around the world took part in four days of learning, presentation, exchange and discussion; including a full day field trip to Australia's internationally known wine-producing region, the Hunter Valley. The University of Sydney hosted nearly 150 researchers from a number of disciplines, focused on furthering the science of Digital Soil Mapping (DSM) and Assessment. The 2012 meeting of this biennial workshop, themed "Digital Soil Assessments and Beyond", called attention to the increasing need for exploration of digital methods to interpret and evaluate the present state of world soils, as well as a proposed shift in focus from digital mapping to assessment in the application of predictive soil modeling.

A total of 82 papers, including 8 keynotes, were presented in a series of 10 sessions over the course of the four day workshop. Each session included a number of five minute presentations followed by a brisk 30 minute discussion in which participants and presenters were allotted time for question and debate. The sessions focused on a variety of topics including: Digital Soil Assessment, DSM in the environment, soil maps, legacy data & covariates, Digital Soil Modeling, digital mapping of soil classes, sampling and monitoring in DSM, cyber infrastructure & expert system in DSM, operational DSM, proximal, remote sensing and spectroscopy of soil, and GlobalSoilMap.net. In addition, 20 posters were presented in two afternoon sessions.

Keynotes speakers included Robert Hill, a distinguished former member of the Australian Senate, Lee Belbin, currently working for the Atlas of Living Australia, Jeffery Walker, a Professor in the Department of Civil Engineering at Monash University, Dr. Garry Willgoose, a Professor at the University of Newcastle in the field of geomorphology and hydrology, Dr. David Clifford, senior research scientist in the Division of Mathematics, Informatics and Statistics at CSIRO, Dr. Ian C. Lau of CSIRO's Exploration and Environmental Sensing Group, Bruce Simons, an SDI Information Modeler at CSIRO who is also involved in the OneGeology Project, Dr. Neil McKenzie, Chief of CSIRO Land and Water, and Dr. Janis Boettinger, Chair of the Digital Soil Mapping working group and Professor and Vice Provost at Utah State University.

Among the many presentations, organic carbon mapping and carbon stock assessment were recurrent themes; bringing into focus the importance placed on understanding and quantifying soil carbon relations by the international users of digital soils information.

The highlight of the conference, held on the third day, consisted of a field excursion to the Hunter Wine Country Private Irrigation District, an area of approximately 220 km^2 located in the Lower Hunter Valley. This "New World" wine producing region has a rich history of viticulture dating back to the 1820's and is characterized by its marl soils, which are highly valued for their lime content and superior grape cultivating ability. Participants were treated to spectacular views of neatly rowed vineyards while exploring two distinct and colorful soil pits. The field

trip was appropriately concluded with a small sampling of the famous Hunter Valley wines.

The Fifth Global Workshop on Digital Soil Mapping provided an opportunity for the international Digital Soil Mapping and Assessment community to meet and exchange ideas, research, and perspectives. The city of Sydney served as an ideal host, promoting the international research and comradery indicative of the globally oriented Digital Soil Mapping community. Expectations are high for the 2014 Global Workshop on Digital Soil Mapping in Nanjing.

Editor's note:

At the DSM working group meeting, Mogens Greve from Aarhus University, Denmark was appointed as the new vice-chair. There were 5 proposals to host the next DSM workshop in 2014: Wageningen (the Netherlands), Quebec (Canada), Seoul (South Korea), Nanjing (China), and Bali (Indonesia). Each proponent presented their proposal, and the vote goes to Nanjing.

The workshop also gave several awards:

❖Budiman Minasny was voted for the Peter Burrough award for the best idea in DSM for his paper entitled *Mapping the occurrence and thickness of soil horizon within soil profile*.

❖Best oral presentation by student was awarded to Jenette Goodman from Purdue University, US for her paper *Application predicting soil organic carbon using mixed conceptual and geostatistical models in glaciated landscapes*.

❖Best oral presentation was awarded to Julian Caudeville from INERIS, France for *Spatial modelling of human exposure to soil contamination*.

❖Best poster was awarded to Monjoon, Rossiter, Jetten and Udomsri for their poster *Implementing DSM in the Thai soil survey*.

-end-

From inverse modelling to soil geostatistics

From Budiman Minasny
The University of Sydney

I started my PhD in soil physics, and one of the topics was to estimate soil hydraulic parameters from a disc permeameter infiltration experiment using inverse modelling. Inverse modelling involves matching field observations (infiltration data) with prediction using models. The model is iteratively rerun to adjust the parameters so it can provide the best description of the real-life situation (the data from the infiltration experiment).

While not obvious, there is an analogy between inverse modelling in soil physics and geostatistics. Both disciplines have the same problem of estimating the parameters of a model (hydraulic functions in soil physics and the spatial covariance function in pedometrics).

The field of inverse modelling in soil physics began to flourish in the late 1980s and 1990s with the availability of personal computers and increasing computer power. The US Salinity Lab. at Riverside provided various free Fortran program and codes for the inverse procedure for soil water and solute transport. The authors (Rien van Genuchten) at Riverside made use of a Levenberg-Marquardt nonlinear least-squares algorithm (Marquardt, 1963) developed by Duane Meeter (1964) to match the modelled and observed water flow pattern by adjusting the parameters of hydraulic properties. The nonlinear least-squares approach assumes that it is possible to calculate the first derivative of the function to be minimised, the starting parameter values are near the desired function, and the function is reasonably smooth. In practice, these assumptions do not apply, however it is still successfully applied to complex models such as Richards equation which describes water flow in the soil.

In the early development, inverse modelling required a well-posed problem, meaning that a solution should exist, unique: there is only one solution, and stable: the solution does not change with slight modification of the data. The major problem in inverse modelling is that different parameter sets can lead to a similar response. It is common to have many local minima, and optimization techniques can easily fall into the local minima depending on the starting condition. More robust optimisation techniques have been proposed, i.e. global optimization algorithms that attempt to search the whole parameter domain to find global minimum, e.g. genetic algorithm, simulated annealing, and downhill simplex.

In geostatistics, the problem is to estimate parameters of a variogram model. Empirical variograms were calculated by the method of moments and a variogram model was fitted to the empirical variogram using a (weighted) nonlinear least squares algorithm. Many argued about how to choose the best model, some suggested best by manually adjusting the nugget, sill and range parameters. Alex and Dick (1986) recommended the best practice for fitting a variogram model using weighted nonlinear least squares and used the AIC criterion to select the best model when several models are available. However, there is also a great amount of literature that debated which type of weighting to use in the nonlinear fit.

In the late 1990s with the availability of proximal soil sensors and harvester that collect on-the-go soil and yield data, massive data became available. The computational time to calculate an empirical variogram of the whole field (an area can have more than 100,000 observations) now became a problem, and if we were to sample only parts of the data for variogram calculation we would lose lots of information by assuming a single variogram model for the whole field, resulting in a very smooth map. To accommodate this, we developed the VESPER program which used kriging with local variograms (Haas, 1990). This involves searching for the closest neighbourhood for each prediction site, estimating the empirical variogram from the neighbourhood, fitting a variogram model to the data automatically using a non-linear least squares approach, kriging using a local neighbourhood and local variogram parameters, and calculating the uncertainty of kriging prediction. The first prototype was developed in the S programming language (precursor to R), but the application was too slow (it is still now). We then decided to develop a standalone program at the Australian Centre for Precision Agriculture. A right type of nonlinear least-squares algorithm is also necessary, some codes (e.g. Jian et al. 1996 used the routine from Numerical Recipes) does not converge readily and close initial estimates are essential. My experience in inverse modelling allowed me to quickly adapt the Levenberg Marquardt nonlinear

least squares algorithm for automated variogram model fitting. Combined with Alex's Fortran kriging code written in 1980s resulted in a robust system that calculates empirical variograms, fits a model automatically, and performs kriging. After more than 10 years, the program still works very well and is able to handle a large amount of data.

In the field of geostatistics, Diggle et al. (1998) introduced model-based geostatistics to describe an approach to geostatistical problems based on the application of formal statistical methods using an explicitly assumed stochastic model. This attracted applications in pedometrics, which defined a formal statistical model for prediction (linear-mixed model), where most of the parameters inference is based on maximum likelihood methods or unbiased estimates using REML (Lark and Cullis, 2004). The likelihood methods still rely on optimisation procedures, either based on nonlinear least squares or global optimization techniques such as annealing. These optimization techniques only attempt to find optimum parameters that give the best fit the observed data, ignoring the uncertainties in the models and the observed data.

In hydrology, the uncertainty of parameter estimates became more important, Keith Beven and Andy Binley from Lancaster University (Beven and Binley, 1992) proposed the concept of Generalized Likelihood Uncertainty Estimation (GLUE) for assessing the degree of uncertainty surrounding outputs from a simulation model. They recognised that many parameter sets within a model will give similar outputs to satisfy a given objective function and, it is only possible to assign the likelihood of each parameter set to be able to predict the system. The GLUE methodology uses Monte Carlo simulation, where the parameter space is sampled randomly and each parameter set is fed into the model, and a quantitative measure of performance is used to assess the parameters acceptability (the capability to reproduce observations).

Although the use of Bayesian approach for kriging is not new (Kitanidis, 1986), or called Bayesian Kriging by Omre (1987), Diggle et al. (1998) proposed the use of a Bayesian framework for parameter inference and prediction in model-based geostatistics. Instead of estimating the best set of parameters, a prior distribution of the parameters is assumed, and the challenge is to update the distribution using the observation. To do that, an efficient sampling procedure is required and thanks to the rapid development of Markov chain Monte Carlo (MCMC) method over the past 10 years, we are now able to do that efficiently.

However the application of MCMC method in model-based geostatistics was quite limited, partly because of the unavailability of an efficient MCMC procedure. Moyeed and Papritz (2002) among the first to employ MCMC to soil data, but they showed that MCMC sampling with the Metropolis Random Walk had no immediate advantages over other classical kriging methods. They concluded that MCMC required significantly more computational resources, and manual interaction to appropriately sample the posterior distribution. The available R library geoRglm by Ribeiro et al. used a more complicated Langevin-Metropolis (Don't ask me what it does) approach. This approach requires the gradient of the likelihood function to be evaluated, and also needs manual trial-and-error adjustment of the scale of the proposal distribution.

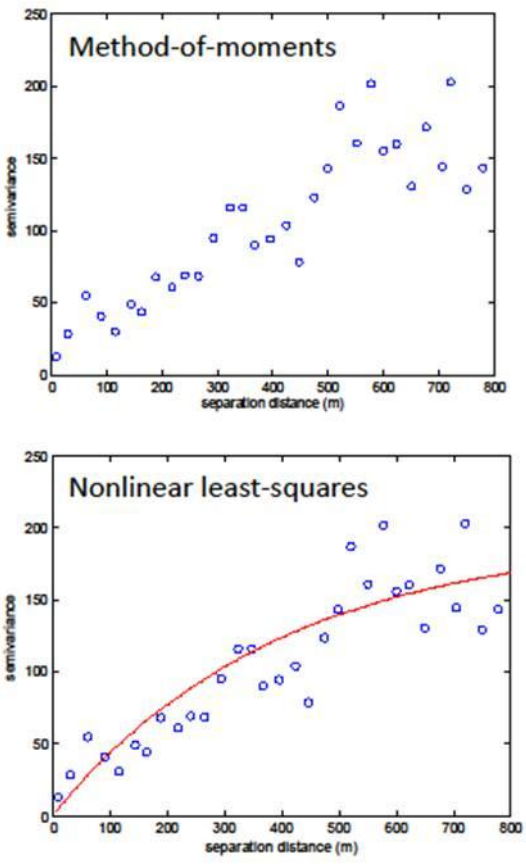
Similar problems also arose in hydrology, and this stimulated Jasper Vrugt (University of Amsterdam and UC Irvine) and co-workers to develop adaptive MCMC methods that run parallel Markov chains and can tune themselves to the posterior target distribution using information exchange between different trajectories. Vrugt et al. (2003) initially developed the Shuffled Complex Evolution Metropolis(SCEM-UA) algorithm which infers the most likely parameter set and its underlying posterior probability. We no longer seek the best solution, but attempt to find a distribution of parameters that best describes the response. An improvement of SCEM-UA was developed recently by Vrugt et al. (2008) called DREAM (DiffeRential Evolution Adaptive Metropolis); it allows for automated parameter searching, automatically tunes the scale and orientation of the proposal distribution during the search without any human intervention.

With the awareness of the development of MCMC method in hydrology, and realising parameter inference is partly an optimization problem, I decided to unite it with the linear-mixed model for parameter inference. It is futile to build a special inefficient MCMC algorithm to handle the geostatistical problem, while hydrologist have specialised in developing an efficient MCMC code to handle more complicated models with many more parameters. The resulted in collaboration with Jasper Vrugt in Minasny

et al. (2011). We were able to implement model-based soil geostatistics efficiently. We also showed that classical geostatistical parameter inference using empirical variogram is still valid, and constituted a realization of the parameter uncertainty. Regression kriging, although not optimum, is still part of the realizations of the parameter distribution.

The breakthroughs in MCMC parameter inference with cooperation between pedometrics and hydrology has opened up possibilities of applying non-linear and complex models for space-time models. Currently models with copulas (with Ben Marchant) have been successfully applied, a nonlinear spatial function using neural networks also been trialled. I had the chance to participate in Noel Cressie's workshop last year and learned that in his latest book Statistics for Spatio-Temporal Data, he proposed a Bayesian hierarchical model which allows the incorporation of mechanistic models and empirical data for dynamic modelling.

It is now an exciting time that we can incorporate complex, nonlinear models into space-time data! QED.



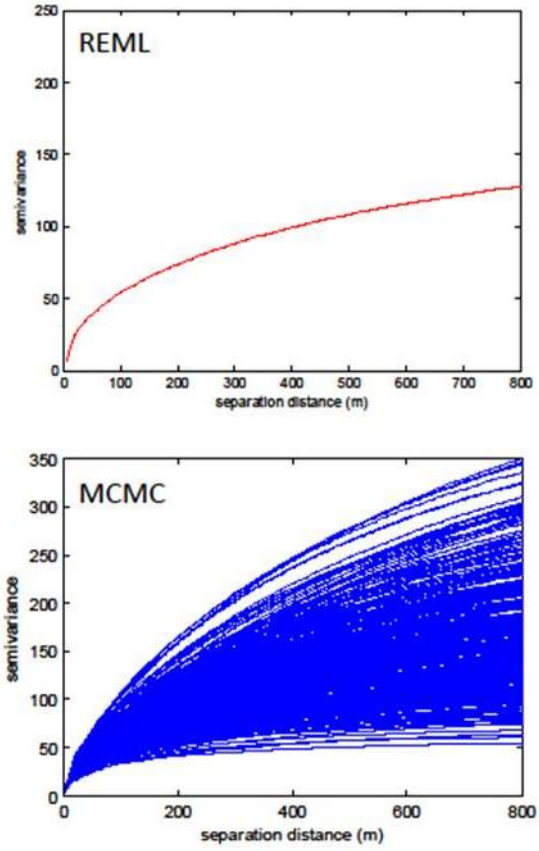


Figure 1. The evolution of variogram fitting, from method-of-moments to nonlinear least squares, REML and MCMC.

References:

Beven K, Binley A. 1992. The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes* 6, 279–298.

Diggle, P.J., Tawn, J.A., Moyeed, R.A., 1998. Model-based geostatistics. *Journal of the Royal Statistical Society Series C-Applied Statistics* 47, 299-326.

Haas, T.C., 1990. Kriging and automated variogram modeling within a moving window. *Atmospheric Environment* 24A, 1759-1769.

Jian, X., Olea, R.A., Yu, Y-S. 1996. Semivariogram modeling by weighted least squares, *Computers & Geosciences* 22, 387-397.

Kitanidis, P.K. 1986. Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water resources research* 22, 499-507.

Lark, R.M., Cullis, B.R., 2004. Model-based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science* 55, 799-813.

Marquardt, D.W. 1963. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* 11, 431-441.

Meeter, D. A. 1964. Non-linear least-squares(Gaushaus). Univ. of Wisconsin Computing Center. Program revised 1966.

McBratney, A.B., Webster, R., 1986. Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. *Journal of Soil Science* 37, 617-639.

Minasny, B., Vrugt, J.A., McBratney, A.B., 2011. Confronting uncertainty in model-based geostatistics using Markov Chain Monte Carlo simulation. *Geoderma* 163, 150-162.

Moyeed, R,A., Papritz,A., 2002. An empirical comparison of kriging methods for nonlinear spatial point prediction. *Mathematical Geology* 34, 365-386.

Omre, H., 1987. Bayesian kriging—merging observations and qualified guesses in kriging. *Mathematical Geology* 19, 25-39.

Vrugt, J.A., Gupta, H.V., Bouten, W., Sorooshian, S., 2003. A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters,. *Water Resources Research* 39, art. No. 1201.

Vrugt, J.A., ter Braak, C.J.F., Diks, C.G.H., Robinson, B.A., Hyman, J.M., Higdon, D., 2009. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation* 10, 273-290.

Dealing with below quantification limit data in geostatistical analyses

From Thomas Orton, Nicolas Saby, Dominique Arrouays, Claudy Jolivet, Estelle Villanneau, Ben Marchant, Giovanni Caria, Enrique Barriuso, Antonio Bispo, and Olivier Briand.

Data on soil properties are always subject to limitations of the measurement procedure. At INRA in France several spatial datasets contain some observations reported only as being below a limit of quantification (QL). Such a limit is used to report when a measurement's relative uncertainty is larger than acceptable. The question then is: how should we deal with such observations in a geostatistical analysis? We have applied methods based on censored data to deal with this issue in several case studies. Here, we present two such datasets, and compare the results from the censored data approach with those of an imputed data approach (replacing all below-QL data by half of the QL).

In our analyses, we have considered two steps: a model selection step, in which the best covariates and covariance and covariance models are selected, and a prediction step, for mapping the primary variable using the below-QL data. For model comparison, we have used likelihood ratio tests and the AIC model choice criterion, for which we fitted trend and covariance parameters by maximum ikelihood: in particular, Christensen's (2004) Monte Carlo maximum likelihood (MCML) method. This approach was originally developed for estimating parameters of generalized linear mixed models, but can be applied equally to estimate parameters based on censored data (Orton et al., 2012).

However, we do not present results from this first step here. In this article, we present just the results from the second stage of analysis, spatial prediction, for which we applied a Bayesian approach, implemented by Markov chain Monte Carlo (MCMC). De Oliveira (2005) and Fridley and Dixon (2007) have implemented similar Bayesian approaches to incorporate censored observations for calculating the predictions in geostatistical case studies, the latter using a simulation study to demonstrate better parameter estimates and more accurate predictions compared to the imputation approach. In both the MCML parameter estimation method and the Bayesian prediction method, a likelihood function is defined by integrating the usual likelihood with respect to the unknown values of the censored data. Since this integral cannot be calculated in closed form, it is approximated using a Monte Carlo method: MCML for parameter estimation in the model comparison stage, or MCMC for approximating the predictive distribution in the prediction stage.

Validation of predictions based on censored data is not straightforward, because the usual cross-validation statistics — bias, mean squared error, and mean and median of the standardized squared prediction error — cannot be calculated using the censored data. We have therefore considered an approach to assess the predicted probabilities, $\hat{p}_{>Z_T}(x_i)$, of exceeding various 'contamination' thresholds, Z_T. If the predicted probabilities at the *N* data locations provide a fair assessment of uncertainty, then the expected total number of contaminated locations, $N_{>Z_T}$, would be

$$\hat{m}_{>Z_T} = \sum_{i=1}^{N} \hat{p}_{>Z_T}(x_i)$$

The variance for this total would be

$$\hat{v}_{>Z_T} = \sum_{i=1}^{N} \hat{p}_{>Z_T}(x_i)[1 - \hat{p}_{>Z_T}(x_i)]$$

(this is based on the assumption that data provide *N* independent validation locations, giving *N* Bernoulli trials, each with a different probability, $\hat{p}_{>Z_T}(x_i)$, of 'success'). Assuming that $\hat{m}_{>Z_T}$ and $\hat{v}_{>Z_T}$ parameterize a normal distribution for the total number of contaminated sites, the set of predicted probabilities, $\hat{p}_{>Z_T}(x_i)$, suggest a 95 % confidence interval for the number of contaminated sites of:

$$\hat{CI}_{>Z_T} = [\hat{m}_{>Z_T} - 1.96\sqrt{\hat{v}_{>Z_T}}, \hat{m}_{>Z_T} + 1.96\sqrt{\hat{v}_{>Z_T}}]$$

We compare this to the actual number of data above the contamination threshold to determine whether the method is providing a fair assessment of uncertainty. We use this approach to validate the predictions for several thresholds, Z_T, above the QL.

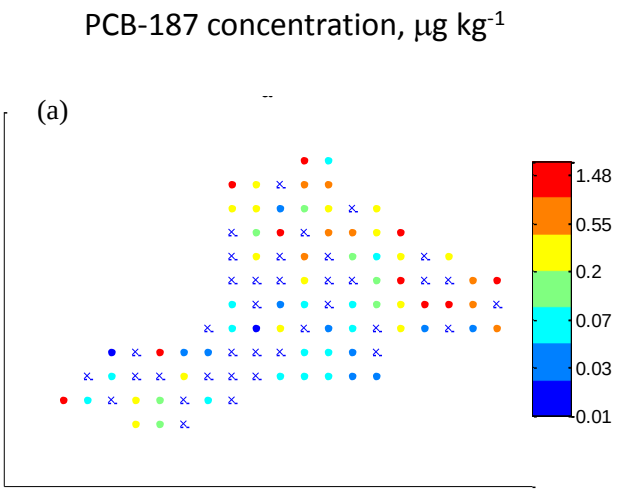
We now present the cross-validation results and prediction maps from two case studies, comparing

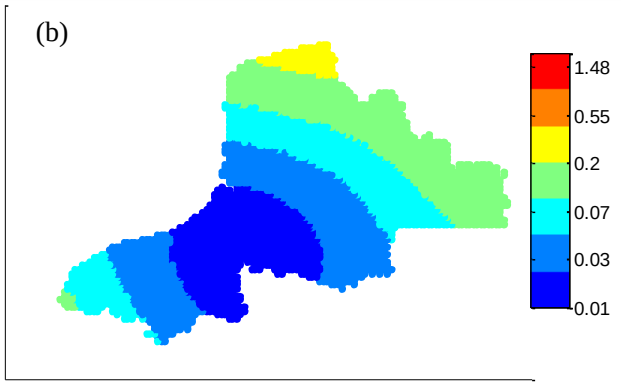
Dealing with below quantification limit data in geostatistical analyses

results from the censored and imputed data approaches. We present predictions on the original scale using the median back-transform (in the first study the log transform was applied to the original data, and in the second a Box-Cox transform was used). In practice, some other statistic of the predictive distribution (mean, or probability of exceeding a contamination threshold) might be more appropriate than the median.

Case study 1 – PCB-187 concentrations in soil in a region of northern France.

Our main aim in this case study (Orton et al., 2012) was to demonstrate the methodology for dealing with censored data. Polychlorinated biphenyls (PCBs) are anthropogenic persistent organic pollutants, mainly derived from industrial activities (i.e., transformers, capacitors, paints). They have been detected in various kinds of environmental compartments, where they can be bioaccumulated, particularly in soils rich in organic matter. The PCB-187 concentration was measured for 105 soil samples from a region in northern France: 37 of the observations gave below-QL data, as shown in Figure 1 a (crosses show the observations below the QL of 0.02 µg kg^{-1}).





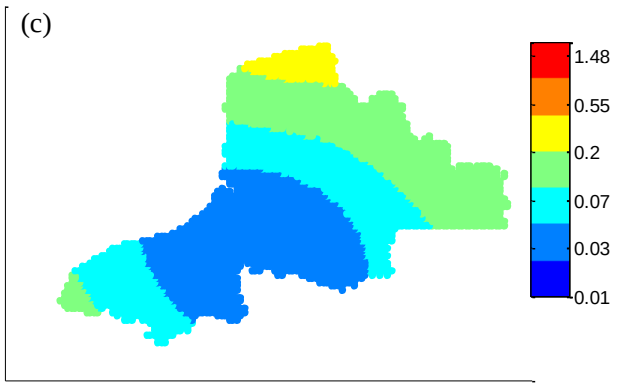


Figure 1 – (a) PCB-187 data with crosses for below-QL (0.02 µg kg^{-1}) observations, (b) predictions using censored data approach, and (c) using imputed data approach.

The predictions from the censored and imputed data approaches are shown in Figures 1b and 1c, respectively. The predictions shown are without including fixed effects, so that the mean is assumed constant. The censored data approach gives an area of predictions below the QL, whereas the imputed data approach did not. The validation results are shown in Figure 2, in which the dashed lines show the predicted 95 % CI for $N_{>z_T}$, and the solid lines the actual number of contaminated data locations. The results in Figure 2 a demonstrate that the uncertainty was well characterized by the censored data method for all values of z_T whereas the imputed data method (Figure 2 b) tended to underestimate the probability of exceeding some of the lower thresholds.

Dealing with below quantification limit data in geostatistical analyses

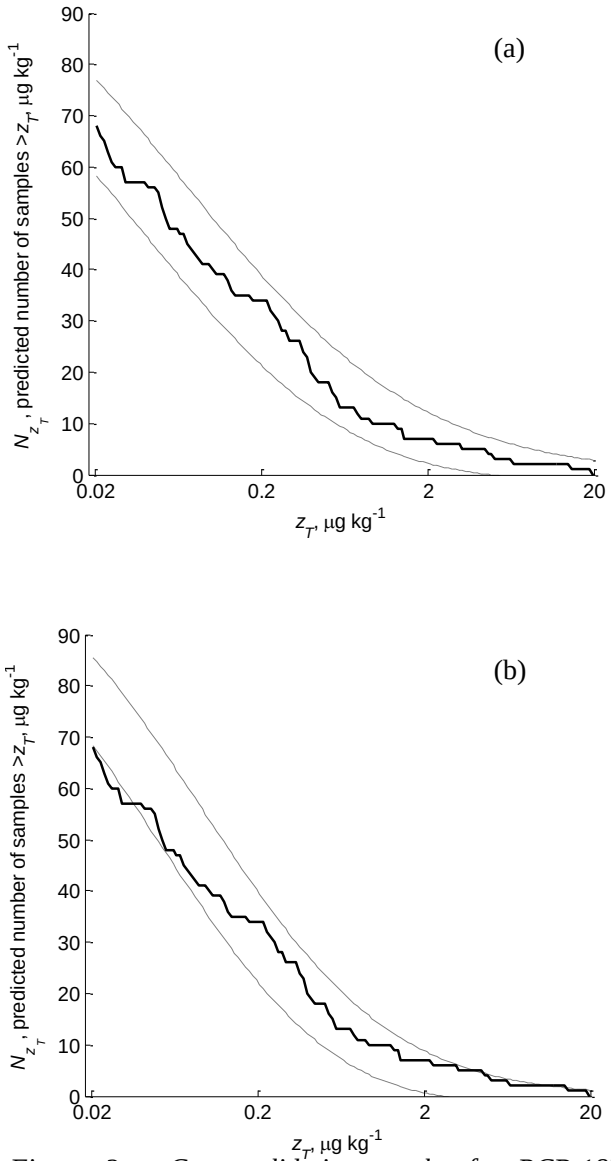


Figure 2 – Cross-validation results for PCB-187 predictions from (a) censored data and (b) imputed data approaches. Dashed lines show the predicted 95 % confidence intervals for the actual number of contaminated locations (solid lines).

Case study 2 – Dioxin concentrations in soil close to a disused incinerator

The second case study is mapping the dioxin concentration at a field scale in the vicinity of a disused incinerator (operational between approximately 1970 and 2000). Data were collected at 72 locations (Figure 3 a) close to the site to evaluate the level of pollution and assess the need for decontamination.

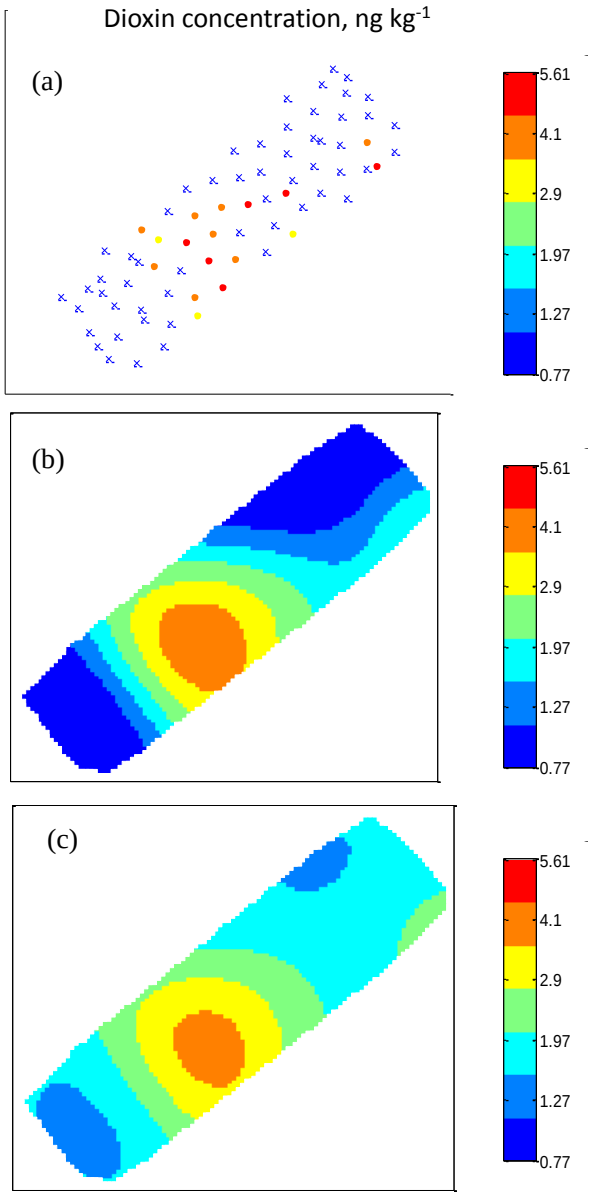


Figure 3 – (a) Dioxin data with crosses for below-QL (3 ng kg^{-1}), (b) predictions using censored data approach, and (c) using imputed data approach.

The map of predictions from the imputed data approach (Figures 3 c) is smoother than that for the censored data approach (Figure 3 b), which gives higher peaks and lower troughs. The cross-validation results (Figures 4 a and b) demonstrate that the censored data approach produces a more reliable assessment of uncertainty.

Dealing with below quantification limit data in geostatistical analyses

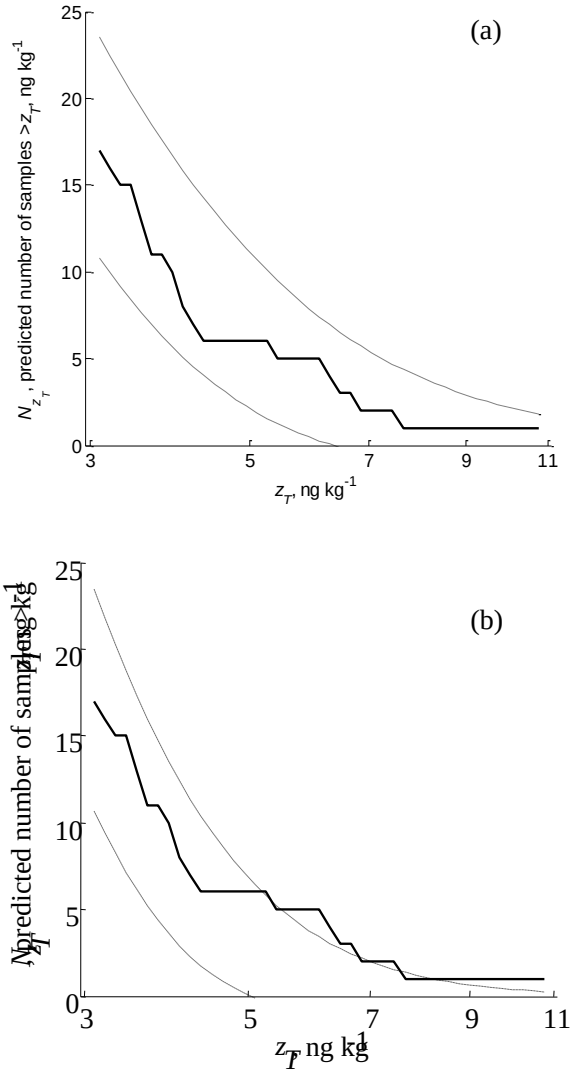


Figure 4 – Cross-validation results for dioxin concentration predictions from (a) censored data and (b) imputed data approaches. Dashed lines show the predicted 95 % confidence intervals for the actual number of contaminated locations (solid lines).

Summary

We have compared the censored and imputed data approaches using two case studies. In each case, cross-validation showed that the censored data approach gave a fair assessment of prediction uncertainty: the imputed data approach did not provide such a fair assessment. Certainly, if we do follow a stringent approach based on a sensible representation of below-QL observations (such as the censored data approach), then we can be confident that conclusions reached are not the result of an *ad hoc* representation of uncertainty. In further work, we will be looking at the effects of below-QL data on cokriging predictions, when there are censored data on two correlated spatial variables; this is the case with the dioxin case study, which offers a good chance to study this issue.

References

De Oliveira, V. 2005. Bayesian inference and prediction of Gaussian random fields based on censored data. *Journal of Computational and Graphical Statistics*, 14:95-115.

Fridley, B.L., and P. Dixon. 2007. Data augmentation for a Bayesian spatial model involving censored observations. *Environmetrics*, 18:107-123.

Orton, T.G., Saby, N.P.A., Arrouays, D., Jolivet, C.C., Villanneau, E.J., Paroissien, J.-B., Marchant, B.P., Caria, G., Barriuso, E., Bispo, A. and Briand, O. 2012. Analyzing the spatial distribution of PCB concentrations in soils using below quantification limit data. *Journal of Environmental Quality*. In press. doi:10.2135/jeq2011.0478

Villanneau, E.J., Saby, N.P.A., Arrouays, D., Jolivet, C.C., Boulonne, L., Caria, G., Barriuso, E., Bispo, A. and Briand, O. 2009. Spatial distribution of lindane in topsoil of Northern France. *Chemosphere*, 77:1249-1255.

How to define, sample for and estimate the regional trend in soil monitoring?

From D.J. Brus
Soil Science Centre,
Wageningen University and Research Centre

1 Introduction

In soil monitoring we are often interested in whether the soil property of interest has been changed. Think for instance of changes in the soil carbon stock. With more than two sampling times we may be interested in the average change per time unit (for instance decade), which is equivalent to the *linear* trend. The average change per time unit generally will vary in space, some soil profiles respond quickly, others slowly. When we do not have enough budget for *mapping* the linear trend at point-locations, an alternative aim is to estimate the regional trend, defined as the linear trend of the spatial mean of the soil property of interest. In recent papers we have shown that this linear trend can be defined in different ways (Brus and de Gruijter, 2011, 2012). In this short paper I will elaborate on these definitions and illustrate sampling strategies for the trend with a simulated space–time field of soil organic matter (SOM) content (Figure 1)

2 Trend defined as population parameter

The linear trend can be defined as as a linear combination of the spatial means at the sampling times:

$$b = \frac{\sum_{j=1}^{r}(t_j - \bar{t})(\bar{z}_j - \bar{\bar{z}})}{\sum_{j=1}^{r}(t_j - \bar{t})^2} \tag{1}$$

with r the number of sampling times, $\bar{t}$ the mean of the sampling times, and $\bar{\bar{z}}$ the mean of the spatial means. You may recognize this as the Ordinary Least Squares (OLS) estimator of the slope of a linear model for $\bar{z}$ (dependent or response variable) and t as predictor. However, here the trend is not a model parameter, but a population parameter. The population or universe of interest consists of a finite set of (infinite or finite) spatial populations, $\mathcal{U} = \{\mathcal{S}_1, \mathcal{S}_2 \cdots \mathcal{S}_r\}$, with $\mathcal{S}_1$ the spatial population at sampling time t_1, *et cetera*. This universe is a subset only of the

How to define, sample for and estimate the regional trend in soil monitoring?

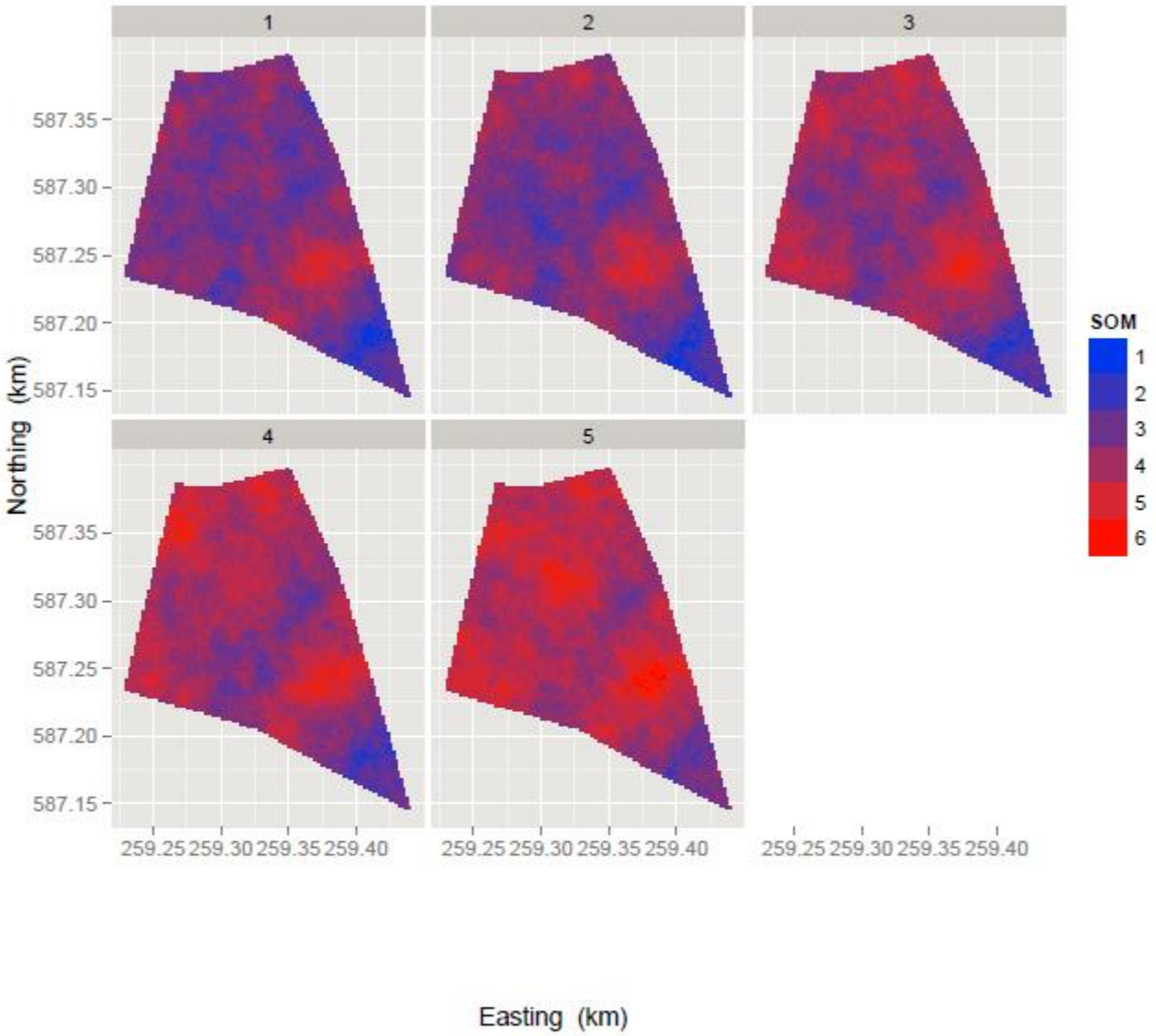


Figure 1: Simulated space–time field of soil organic matter content. The five panels show the spatial fields at the sampling times. The sampling interval is constant (e.g. 10 years).

How to define, sample for and estimate the regional trend in soil monitoring?

$\mathcal{U} = \mathcal{S} \times \mathcal{T}$ with $\mathcal{T}$ the temporal universe (ter Braak et al., 2008). I will not go into sampling approaches for this definition of the trend.

Parameter b as defined in Eq. 2 can also be seen as the slope parameter that is obtained when the response variable is known for all population units (exhaustive fit). Here the population 'units' are not sampling units (points) but populations themselves, viz. the spatial populations at the r sampling times. The response variable is the *spatial mean* of SOM. When the spatial means are known for all population units, i.e. at at all sampling times, then parameter b is also known without error, see hereafter.

Eq. 2 can be rewritten as a linear combination of the spatial means at the sampling times:

$$b = \frac{\sum_{j=1}^{r}(t_j - \bar{t})\bar{z}_j}{\sum_{j=1}^{r}(t_j - \bar{t})^2} = \sum_{j=1}^{r} w_j \bar{z}_j \tag{2}$$

with the weights w_j equal to

$$w_j = \frac{t_j - \bar{t}}{\sum_{j=1}^{r}(t_j - \bar{t})^2} \tag{3}$$

This shows that the trend can be estimated via estimation of the spatial means at the sampling times, and as a consequence a design-based sampling approach is recommendable. I will elaborate now on estimation for space–time designs with no or complete overlap (static-synchronous, independent synchronous, serially alternating) and for space-time designs with partial overlap (supplemented panel, rotating panel).

2.1 Space–time designs with no or complete overlap

With space–time designs in which the spatial samples at the sampling times $t_1 \cdots t_r$ have no overlap, i.e. no locations are revisited, or complete overlap, i.e. all locations are revisited, the spatial mean at a given time is estimated on the basis of the measurements at that time only, using the well-known design-based estimators. Given these estimated means the linear trend can be estimated as a linear combination of the estimated means

$$\hat{b} = \sum_{j=1}^{r} w_j \hat{\bar{z}}_j = \mathbf{w}' \, \hat{\mathbf{z}} \tag{4}$$

2.2 Space–time designs with partial overlap

For space–time designs with partial overlap such as the supplemented and the rotating panel, the precision of the estimated mean at a given sampling time can be increased by using the measurements at the other times as covariates. This can be

achieved by Generalized Least Squares (GLS) estimation of the spatial means. First panel-specific estimates of the spatial means are computed, referred to as 'elementary estimates'. A panel is a group of locations observed at the same set of sampling times. These elementary estimates are then combined into one estimate of the mean per time t_j by

$$\hat{\mathbf{z}}_{\mathrm{GLS}} = (\mathbf{X}'\widehat{\mathbf{C}}_{\mathrm{e}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{C}}_{\mathrm{e}}^{-1}\hat{\mathbf{z}}_{\mathrm{e}} \tag{5}$$

with $\hat{\mathbf{z}}_{\mathrm{e}}$ the vector of elementary estimates of the spatial means, $\mathbf{X}$ the design matrix with 0's and 1's, and $\widehat{\mathbf{C}}_{\mathrm{e}}$ the estimated covariance matrix of the elementary estimates. Hopefully you recognize this equation from your statistics courses on regression analysis as the GLS-estimator of the regression coefficients. In linear regression analysis we have observations on a target variable and one or more predictors, covariates. In ordinary linear regression it is assumed that the observations are independent. Correlation between the observations can be accounted for by estimating the variance-covariance matrix of the observations, and using this matrix in GLS fitting of the linear model. Here the observations of the target variable are the elementary estimates of the spatial means at $t_1 \cdots t_r$. The predictors are indicators for the sampling times. There are as many predictors as there are sampling times.

Once the means are estimated by GLS, the trend can be estimated as a linear combination of these estimated means:

$$\hat{b}_{\mathrm{GLS}} = \mathbf{w}'\hat{\mathbf{z}}_{\mathrm{GLS}} \tag{6}$$

with $\mathbf{w}$ as before (Eq. 3). With small spatial sample sizes the estimated sampling covariance matrix $\widehat{\mathbf{C}}_{\mathrm{e}}$ can be poorly defined, leading to extreme values for the estimated trend. In such cases I recommend to estimate the trend with Eq. 4.

2.3 Effect of number of sampling locations on variance of estimated trend

The sampling variance of the estimated trend can be reduced by increasing the number of sampling times and the number of sampling locations per time. Besides, there is a clear effect of the type of space–time design (Fig. 2) and of the spatial design. Fig. 3 shows the standard error of the estimated trend as a function of the number of sampling locations per time, for a static-synchronous space–time design and simple random sampling in space. If the entire study area would be sampled all five times, the standard error would be 0. There is no uncertainty left about the trend. Figure 3 (subfigure in the middle) shows the true spatial means at the five times plotted against the sampling time and the estimated linear trend. As can be seen the true spatial means are not located precisely on the fitted line. In regression analysis we would say that there is a residual variance. As a consequence in regression analysis the variance of the estimated regression coefficients (intercept and slope) is not 0 but a positive value.

How to define, sample for and estimate the regional trend in soil monitoring?

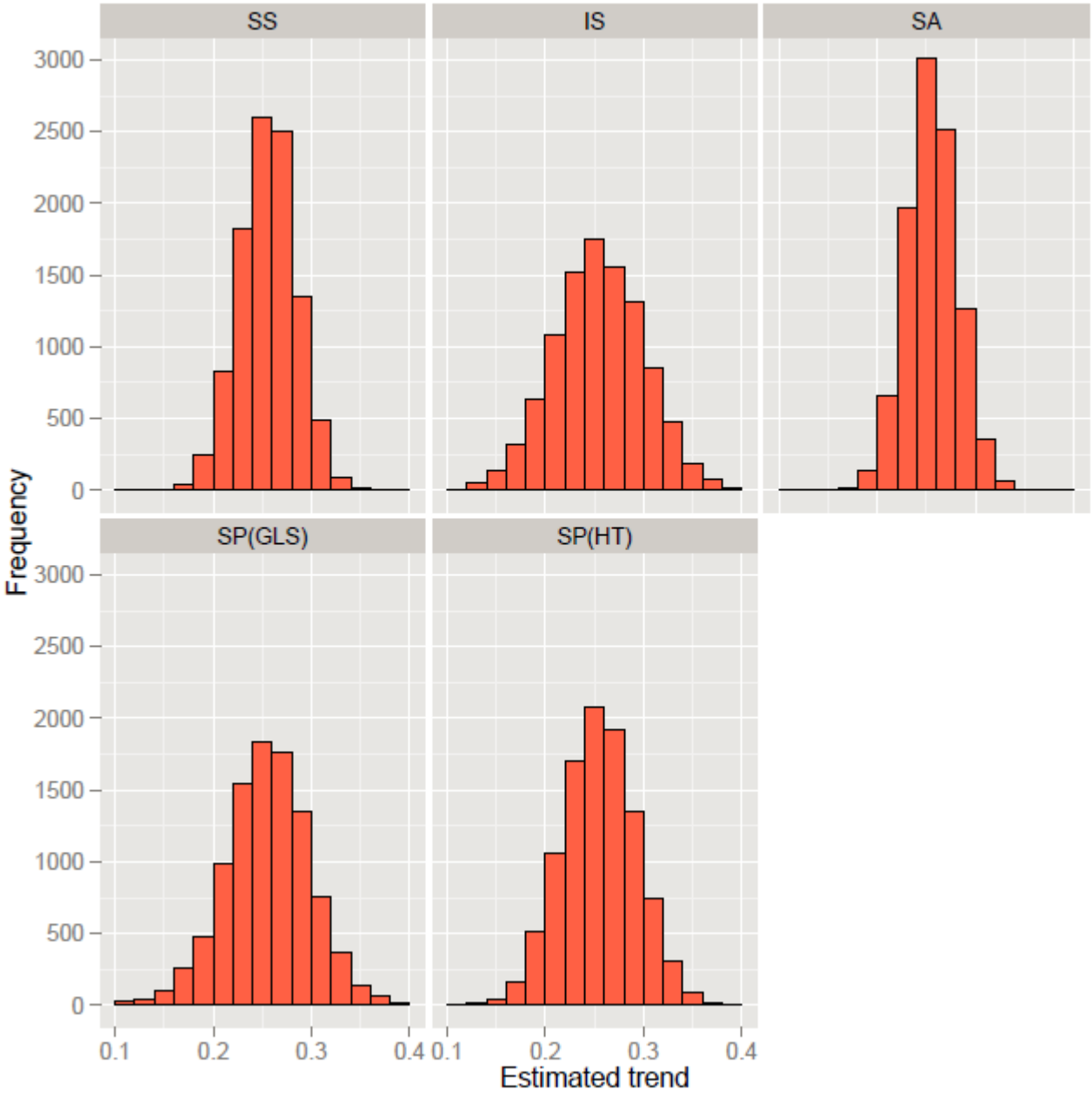


Figure 2: Histograms of 10,000 repeated estimates of the trend of the mean defined as population parameter, for static-synchronous (SS), independent-synchronous (IS), serially alternating (SA) and supplemented panel (SP) sampling, five sampling times and 20 locations per time selected by simple random sampling (sampled from the space–time field of Figure 1). In supplemented panel sampling 10 locations are revisited. For SP the trend is estimated both by Eq. 4 (SP(HT)) and by Eq. 6 (SP(GLS)). Note the long tails of the sampling distribution of the estimated trend with SP(GLS), caused by the poorly defined covariance matrix. The serially alternating design had the smallest sampling variance of the estimated trend

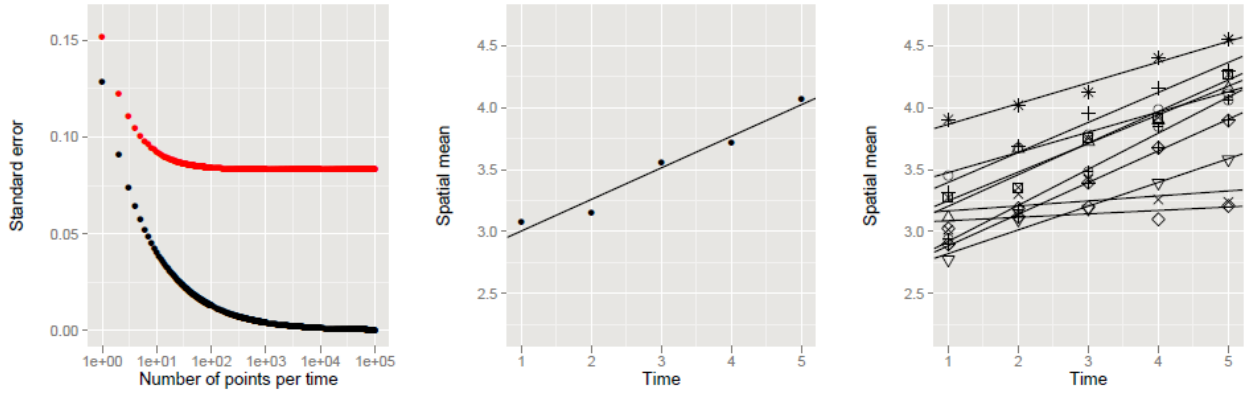


Figure 3: Left: standard error of the estimated trend, defined as a population parameter (black) or as a model parameter (red), as a function of the number of points per time. Sampling design: static-synchronous with simple random sampling in space. Middle: true spatial means of simulated space–time field (Fig. 1), plotted against the sampling time, and the linear trend of the spatial means. Right: true spatial means and linear trend fitted by OLS for 10 realizations of the space–time model used in simulating Fig. 1

3 Trend defined as model parameter

Fitting the straight line of Fig. 3 by OLS with standard statistical software results in an estimated trend of 0.255 which is equal to the estimated trend defined as a population parameter. However, the standard deviation of the estimated trend equals 0.027, which is small, but definitely larger than 0. The reason that in standard regression analysis the variance is not 0 is that the true spatial means are considered as realizations of random variables. In OLS fitting of the simple linear model the spatial means at the sampling times are assumed to be identically and independently distributed with expectation $\beta_1 + \beta_2 \cdot t$ and constant variance (the variance of the residuals). The coefficients β_1 and β_2 are model parameters, the intercept and the slope, respectively. The parameter β_2 describes the average change of the spatial mean per time unit, the linear temporal trend. This is the target parameter to be estimated. So, contrary to the previous section a time-series model is introduced for the spatial means

$$\bar{Z}(t_j) = \beta_1 + \beta_2 \cdot t_j + \eta(t_j) \quad j = 1 \cdots r \tag{7}$$

where $\eta(t_j)$ is the model residual (model error) of the spatial mean at time t_j. The spatial mean at time t_j is now in capital, indicating that it is a random variable. With the trend defined as a model parameter the sampled space–time field of Fig. 1 is treated as just one realization of a stochastic space–time process. I simulated 10 of these space–time fields, computed for each simulated space–time field the true

spatial means at the five times, and fitted the model by OLS. The result is presented in Fig. 3 (subfigure at the right). The fitted trend clearly varies between the model-realizations. The variation is even much larger than expected from the estimated variance of the trend as obtained with OLS fitting (standard deviation 0.027). This can be explained by the correlation of the spatial means. In OLS it is assumed that these spatial means are uncorrelated (identically independently distributed, iid), however this assumption is clearly violated by the space–time model used in simulating the space–time fields. The spatial means are correlated in time, amplifying the variance of the trend between model realizations. Fig. 3 (subfigure on the left) shows that the standard error of the estimated trend, defined as a model parameter, with exhaustive spatial sampling is about 0.083.

In practice the spatial means are unknown, and must be estimated from a sample. When these spatial means are estimated from probability samples and design-based estimators, then the space–time sampling approach becomes a hybrid, design- and model-based approach. To explain this approach I will first consider the simple situation where we have only one estimate of the mean per time, and then proceed with the situation with more than one elementary estimate per time, as obtained with space–time designs with partial overlap.

3.1 Space–time designs with no or complete overlap

In the hybrid approach it is assumed that the spatial means can be described by a linear mixed model

$$\mathbf{Z} = \mathbf{D}\boldsymbol{\beta} + \boldsymbol{\eta} \,, \tag{8}$$

with $\mathbf{Z}$ the r-vector with true spatial means at the sampling times, $\mathbf{D}$ the $r \times p$ design-matrix, $\boldsymbol{\beta}$ the p-vector with regression coefficients and $\boldsymbol{\eta}$ the r-vector with model errors. This matrix equation is equivalent to the model of Eq. 7 for a design-matrix $\mathbf{D}$ with the first column a vector of ones and the second column a vector with the sampling times. The model errors $\boldsymbol{\eta}$ have zero mean and an $r \times r$ covariance matrix $\mathbf{C}_\xi$. This is the matrix with the variances and covariances of the spatial means between realisations of the space–time model. In practice the spatial means are unknown, and in the hybrid approach these means are estimated from spatial probability samples. With no or complete overlap these spatial means are estimated by design-based estimators. The sampling introduces an additional error component in the model:

$$\widehat{\mathbf{Z}} = \mathbf{D}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\epsilon} \,, \tag{9}$$

with $\boldsymbol{\epsilon}$ the r-vector with sampling errors. The sampling errors have zero mean and an $r \times r$ covariance matrix $\mathbf{C}_p$, the sampling covariance matrix of the estimated spatial means that we have seen many times before. The model errors and sampling errors are independent, as they originate from independent stochastic processes. The

overall covariance matrix of the estimated spatial means equals

$$\mathbf{C}_{\xi p} = \mathbf{C}_{\xi} + \mathbf{C}_{p} \; . \tag{10}$$

With known covariance matrix $\mathbf{C}_{\xi p}$, the regression coefficients can be estimated by

$$\hat{\boldsymbol{\beta}} = (\mathbf{D}'\mathbf{C}_{\xi p}^{-1}\mathbf{D})^{-1}\mathbf{D}'\mathbf{C}_{\xi p}^{-1}\hat{\mathbf{Z}} \tag{11}$$

3.2 Space–time designs with partial overlap

Model 9 is reformulated so that multiple estimates of the spatial mean at a given time are accounted for:

$$\widehat{\mathbf{Z}}_{\mathrm{e}} = \mathbf{D}_{\mathrm{e}}\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\eta} + \boldsymbol{\epsilon}_{\mathrm{e}} \; . \tag{12}$$

Design matrix $\mathbf{D}_{\mathrm{e}}$ now has dimension $E \times p$ with E the total number of elementary estimates. $\mathbf{X}$ is a random-effect design matrix (dimension $E \times r$) with zeroes and ones selecting the appropriate element of $\boldsymbol{\eta}$. Vector $\boldsymbol{\eta}$ is as before, but vector $\boldsymbol{\epsilon}_{\mathrm{e}}$ now has length E as we have multiple sampling errors per sampling time, one per elementary estimate. The overall covariance matrix of the estimated spatial means equals

$$\mathbf{C}_{\xi p} = \mathbf{X}\mathbf{C}_{\xi}\mathbf{X}' + \mathbf{C}_{ep} \; . \tag{13}$$

with covariance matrix $\mathbf{C}_{\xi}$ as before (dimension $r \times r$) and $\mathbf{C}_{ep}$ the sampling covariance matrix of the elementary estimates (dimension $E \times E$). With known covariance matrix $\mathbf{C}_{\xi p}$, the regression coefficients can be estimated by

$$\hat{\boldsymbol{\beta}} = (\mathbf{D}_{\mathrm{e}}'\mathbf{C}_{\xi p}^{-1}\mathbf{D}_{\mathrm{e}})^{-1}\mathbf{D}_{\mathrm{e}}'\mathbf{C}_{\xi p}^{-1}\hat{\mathbf{Z}}_{\mathrm{e}} \tag{14}$$

With small spatial sample sizes the estimated covariance matrix can be not positive definite or poorly defined, leading to missing values or extreme estimates. In this case a simple alternative is to estimate the spatial means at the sampling times by the design-based estimators, as well as as their sampling variances and covariances, and then proceed as in the previous section for samples with no or complete overlap.

4 Which definition?

The question remains what definition can best be chosen. I think the definition is at least partly determined by the aim of the monitoring project. If the aim is to describe the trend during the monitoring period, then a definition in terms of a population parameter is more appropriate than as a model parameter. A definition of the trend as a model parameter comes into scope if we want to use the results for forecasting, i.e. predicting the status in the future. If we use the estimated trend of the mean defined as a population parameter and its standard error for this, then this may lead

to too optimistic estimates of the precision. Clearly, for forecasting the structure of the trend is extremely important. In the case study on SOM a linear trend might not be very realistic when forecasting over long terms. It is more likely that the trend is asymptotically towards a maximum (or minimum in case of a negative trend), which can be modelled, for instance, by an exponential decay (in increasing or decreasing form). In this case the aim would be to estimate the parameters of this exponential model.

Another factor that may help in choosing a definition can be the feasibility of the statistical sampling approach. The definition of the trend has implications for the statistical sampling approach. When defined as a model parameter, a hybrid approach is needed. This sampling approach requires the calibration of a time-series model for the spatial means. which can be difficult. The more sampling times, the more information is obtained on the model. With a few sampling times only, the building of the model can become unfeasible. Strong assumptions are then needed, for instance on stationarity of the spatial mean and on the covariogram model. Besides, the model parameter estimates may become very unreliable. The quality of the estimates, especially the variance of the estimated regional trend, depends on the quality of these assumptions and estimates. With a few sampling times only, we might prefer a model-free sampling approach. Judging a hybrid sampling approach as unfeasible entails that we must abandon the trend defined as a model parameter, and embrace the trend defined as a population parameter as in Eq. 2 as the space–time parameter to be estimated.

References

Brus, D. J. and de Gruijter, J. J. (2011). Design-based Generalized Least Squares estimation of status and trend of soil properties from monitoring data. *Geoderma*, 164:172–180.

Brus, D. J. and de Gruijter, J. J. (2012). A hybrid design-based and model-based sampling approach to estimate the temporal trend of spatial means. *Geoderma*, 173-174:241–248.

ter Braak, C. J. F., Brus, D. J., and Pebesma, E. J. (2008). Comparing sampling patterns for kriging the spatial mean temporal trend. *Journal of Agricultural, Biological and Environmental Statistics*, 13:159–176.

On trying to bridge a gap

From Jaap de Gruijter
Wageningen University and Research Centre

I have been asked to write something down about how I became involved in pedometrics. Thinking of that, many thoughts came into my mind. Here is how I finally organized some of them, in the hope to make a readable story out of it. Let me start with what I now see as the very beginning.

At my primary school in Eindhoven I had the luck of having a good old-fashioned teacher who encouraged his pupils to collect wild plants and make a herbarium. I liked doing that, and this was the starting point of a life-long interest in field biology. Once at the high school a brother-in-law took me to a public lecture about 'physics in the field', given by a well-known physicist. This made me aware of the fact that the physical world around us shows a host of interesting phenomena that can be observed and studied in a scientific manner, and that mathematical formulas are handy tools for doing that. So I liked doing biology, physics, algebra and geometry.

Clearly the logical options for an academic study were biology, physics and mathematics. So I decided to go for agriculture in Wageningen. I realized only much later that the important choices in my life were never based on logic. Nevertheless, studying agriculture wasn't too far fetched. My mother was a farmer's daughter and I had spent several holidays on her ancestral farm in Zeeland.

After entering the university as a 17-year old student there appeared to be a host of things that I could do and seemingly had to do, but getting my degree as soon as possible would not have been at the top of my list, if I would have had a list. However I had to make a choice, and my internal votes went to tropical soil science as the main subject. That seemed to promise me a rather romantic life, but of course I rationalized this by claiming that soil is mankind's most important natural resource, and to study it is a noble enterprise as well as an intellectual challenge. However, my inclination to quantitative approaches soon made me study as much mathematics and statistics as possible in Wageningen. Finally I chose soil physics as my main subject, with mathematics, agro-hydrology and tropical soil science as secondary subjects.

In 1967 I did a 6-month apprenticeship at the soil physics laboratory in Paramaribo, Surinam. My experiences there were more related to statistics than to soil physics. In Wageningen I had taken classes on sampling technique, and here I got the opportunity to bring theory into practice. The head of the laboratory was skeptical about the idea of random sampling, but I got permission to do my first stratified random sampling. I measured throughfall during sprinkling on banana palms, and I was hot, wet, muddy and excited. Other important experiences were the amazingly large errors in measuring the surface area of a rain gauge, and an analysis of error propagation in test results of a new extraction method for CEC determination.

Piet Buringh, the Professor of tropical soil science, asked me to write a report on numerical soil taxonomy, an exciting new research area. Buringh gave a copy of my report to Jaap Schelling, who was the leader of the research section of the Dutch Soil Survey Institute. Soon thereafter I finished my MSc study and schelling offered me a position as PhD student in his institute to study the potential of cluster analysis for soil classification. At the same time I got the option to go to the FAO to work as a tropical soil scientist, or to the USA for a PhD study in soil physics. The conditions of these two options seemed rather uncertain. Furthermore, I was just married, my wife and I had a baby, and my mother-in-law was seriously ill. We decided to stay in Wageningen. I didn't realize then that this decision would prove crucial for my whole career. I had put myself on the track of what we now call pedometrics, and I ever stayed on that track. I never wanted to embark on a new course, primarily because research in this field never stopped fascinating me and the research environment generally suited me well.

In the early sixties it was as if an iron wall separated pedology from statistics. Pedologists and statisticians seems to live on different planets. As a young student I was disappointed by the way pedology was taught to us; mostly descriptive, narrative and classification-oriented. Hypotheses on soil genesis and map accuracy were abundantly generated, but no one bothered about testing them. I felt that statistical methodology could be quite valuable in pedology as many of these methods were especially developed for dealing with uncertainty. So in my mind that gap was to be bridged as fast as possible. It didn't go that fast. After some years, however, I found myself working on a kind of bridge between pedology and statistics.

An exciting rewarding position, sometimes as a statistical consultant for pedologists, and sometimes as a pedologist with my own research projects.

In those early years I was, as far as I know, the only one in my country who actively worked on this bridge, and I must admit that in the first years I sometimes felt lonely. In my institute, my boss was the only person who was interested and supported me. Classical pedology, with free survey as its core and the whole paradigm around it, was *not* the main stream in our survey institute. It was just the only stream! For instance, I had to face explicit opposition in the first staff meeting that I attended, and an anonymous satire circulated in the institute. It was therefore a relief to discover that internationally I was not quite a maverick. The group created by Philip Beckett in Oxford, with Webster, Burrough and Bie, had been working hard already, with very interesting results. Then, during the last conference of the former Working Group on Soil Information Systems (also an initiative of Schelling) I met Alex McBratney, who later visited me in Wageningen and made me enthusiastic about the theory of fuzzy sets developed by Zadeh. Together we worked on fuzzy soil classification and mapping, which was the beginning of a long lasting cooperation.

Meanwhile Schelling had taken another initiative: research on the accuracy of soil maps made by his institute. This kind of work had already been started by Beckett's group. This made me dive deeper into the statistical theory of sampling. The books by Cochran and Särndal opened for me a world on its own, separate and different from the main body of statistics. Soon I was lucky again. Dick Brus of our institute joined me in work on soil sampling, and he initiated the broadening from sampling in space to sampling in space and time, in other words: monitoring.

During all those years I have had the luck of being surrounded by many good colleagues. Generous enough to support me, and brave enough to ask me for help when needed. Patient enough to wait when it took me long to find an answer to their question. Confident enough to tell me if my answer was unsatisfactory, and honest enough to have no hidden agenda. There are too many of them to name them all, but I feel it would be inappropriate if I didn't mention three of them: Jaap Schelling, my first boss, Alex McBratney, all those years a very stimulating colleague and friend, and Dick Brus, an excellent colleague and good friend as well. I am grateful to all of them, for I couldn't have worked so long on that bridge if they wouldn't have been with me. Thank you all!

Caption: Jaap with his grandson

-end-

Pedomathemagica

Problem1:

As readers of Pedomathemagica will recall, Alf and Bert are two soil surveyors. Alf is more mathematically competent than his colleague, a fact which he exploits mercilessly. One evening in the pub Bert says "I am fed up with you, Alf. You are always winning bets off me, so I end up digging all the pits and buying all the beer." "I am sorry about that, Bert," says Alf. "Tell you what, I'll give you a really simple problem, and if you get it right I will dig all the pits and buy all the beer until the next Pedometrics conference, but if you get it wrong then pit-digging and beer-buying will be your job until that same date." Bert says, "OK, but let me see the problem before I agree." "Certainly", says Alf, and he takes a piece of chalk and writes on the table:

$$0.99\dot{9} \neq 1, \text{ True or False?}$$

Bert is delighted, "True, of course!" he laughs. Is Alf about to dig his first pit and buy his first pint, or has Bert been duped yet again?

(From Murray Lark)

Problem2:

Let us assume that the Pedometrics commission has established a 'Young Pedometrics' commission (this may actually be a good idea!) that is led by three people: Phil, Jack and Sheila. The Pedometrics commission itself is also led by three people (a chairman, secretary and treasurer), whose names happen to be the same as those of the Young Pedometrics commission. The only difference is that these three all have a PhD degree whereas the young ones don't, so we have Dr. Phil, Dr. Jack and Dr. Sheila to lead the PM commission. We can also reveal:

1. Sheila lives in Chicago.
2. Jack earns $40,000 per year.
3. The secretary of the PM commission lives half-way between Chicago and New York.
4. His neighbour, a member of the Young PM commission, earns exactly three times as much as he.
5. The namesake of the secretary of the PM commission lives in New York.
6. Dr. Phil plays better chess than the chairman of the PM commission.

What is the name of the chairman of the PM commission?

(From Gerard Heuvelink)

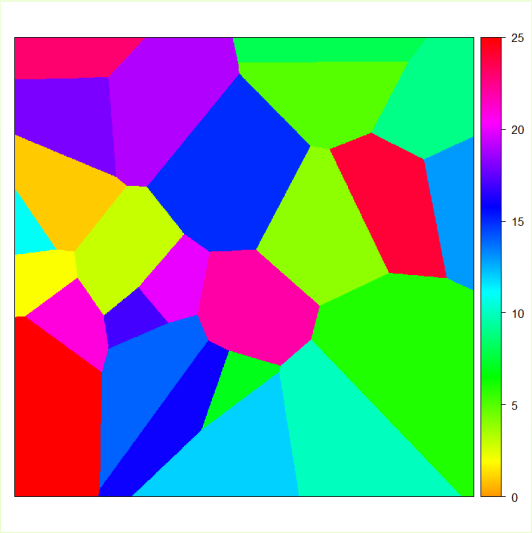
Problem3:

Last year I was at a meeting of the GlobalSoilMap.net consortium and at some stage we entered an interesting discussion on the Intellectual Property rights of soil data and whether organisations that collect soil data should share their data freely with others or not. One opinion that was ventilated was that *primary* soil data (e.g. point observations of soil properties) cannot be shared whereas *derived* data (e.g. interpolated maps derived from the point observations) could be made public. I can understand this position, although it may not always be easy to clearly distinguish primary from derived soil data. More intriguing, however, is what to do if it were possible to derive the primary data from the derived data. In other words, what to do if it were possible to derive the original point data from the interpolated soil property map? If that were possible then we would run into problems because releasing the derived data would effectively also release the primary data.

Pedomathemagica

Problem 3 (continue):

Consider the grid map below that was created from 25 point observations using nearest neighbour interpolation (i.e. using Thiessen or Voronoi polygons, see http://en.wikipedia.org/wiki/Voronoi_diagram). You can download this map as an ascii grid map from <http://www.pedometrics.org/data/*Pedomathemagica_June12_thiessen.asc*>. I think we all agree that in principle one can derive the geographic positions and values of the 25 original point observations that were used to create this map. But how to do this in practice?



The first pedometrician that develops and implements an algorithm that derives the original data set from the interpolated map (i.e. a table with the x, y and data value of the 25 points) such that the positional error of each point is less than twice the grid mesh wins a bottle of champagne. You may assume that all 25 points had unique data values. Please send your solution to gerard.heuvelink@wur.nl

Once this is solved (it may take a while!), we may consider doing the same thing with maps created with inverse distance interpolation or kriging. Here, the question that must be answered first is whether it is in principle possible to uniquely derive the original point observations from the interpolated maps. Those of you who find this a more interesting question can also win a bottle of champagne by providing a watertight proof of whether this can be done, is only possible in specific cases or is always unsolvable.

(From Gerard Heuvelink)