ΠΕΔΟΜΕΤΡΟΝ

The Newsletter of the Pedometrics Commission of the IUSS

Issue 25, December 2008

Chair: Murray Lark *Vice Chair & Editor: Budiman Minasny*

From the Chair

Dear Pedometricians,

Another year has passed, and an active one for the Commission as witness the reports of various meetings contained in this issue of Pedometron.

The year 2008 also marked two centenaries that are significant for those of us who are interested in statistics and the soil. Take the statistics first. In 1908 W.S. Gosset published a paper (under the name Student) entitled 'The Probable Error of a Mean' (Student, 1908). We know that the average value, , of an independent random sample of size *n*, drawn from a population of mean μ and variance σ^2 can be treated as a random variable with an expected value μ and variance σ^2/n, and, furthermore, as the sample size increases so the central limit theorem permits us to assume that the sample average is a normal random variable regardless of the distribution of the variable itself. In principle this allows us to obtain a confidence interval for , there is a 95% probability that m lies in the interval $\pm 1.96 \times \sigma / \sqrt{n}$. But there is a problem, we do not know σ^2, we have only the estimate, σ^2, which has its own sample error. Gosset solved the problem, and the distribution of the average of a sample, standardized by its own sample standard error, is known as Student's *t* distribution to this day. The *t*-test, and confidence intervals computed from the t distribution are so well-known that it is easy to forget that Gosset's paper was such a breakthrough. Hacking (2006) points out that even the great Laplace had previously failed to find a wholly satisfactory way to set confidence intervals on estimates from small samples, and says that "W.S. Gosset's famous statistic, '*t*', was probably the first device to overcome this kind of inexactness." In this sense the year 1908 marked the beginning of applied statistics on samples of realistic size.

Gosset's paper is pragmatic. He uses long derivations, that he himself says are 'tedious' to find moment coefficients for sample variances, but admits that these appear to arise from a simpler law that he cannot find. He reports a simulation study to test his findings. He obtained data from a biometrical study of 3000 criminals, and wrote the observations for each criminal onto a card. The 3000 cards were now his population, and he sampled them randomly many times over. On 750 occasions he drew a random sample of four cards, and so he obtained 750 sample averages and standard deviations to test his formulae. Four years on R.A. Fisher, still then a student, corresponded with Gosset about his paper, pointing out generalizations, and an error: Gosset calculated the sample variance as the sum of squares divided by the sample size n, Fisher proved that the divisor should be *n*-1. The two men did not meet until the early 1920s when they collaborated at Rothamsted to tabulate the integrals of the *t* distribution. By now Fisher had shown that the *t* statistic could be applied to a wider range of problems, and that it pointed to the existence of a more generally interesting statistic, the ratio of two sample variances, *F*.

The second anniversary is Fritz Haber's discovery of the Haber (or Haber-Bosch) process by which nitrogen

Inside This Issue

is reduced to ammonia at high temperature and pressure over a metal catalyst. Until then the reduction of atmospheric nitrogen to forms in which living organisms could use it was done by nitrogen fixing prokaryotes, like the famous *Rhizobium* in symbiosis with legumes, with a small amount also being fixed during lightning strikes. Most nitrate-bearing fertilizers came from natural sources of biological origin, especially guano, and potassium nitrate deposits in the fossil lake beds of Chile. In 2007, by contrast, it is estimated that 55% of the nitrogen fixed globally was fixed by the Haber process, and that this underpinned 80% of world agricultural production. The Haber process produces 'bread from air', but you need not accept all the superstitions of the organic lobby to see that it is not without problems. Most importantly, a process that requires high temperatures and pressures consumes a lot of energy. This contributes to the carbon budget of agriculture, and to farmers' costs. Increases and variations in fertilizer costs in the UK since 2005 have partly reflected energy costs, particularly costs of natural gas.

So we have a dual legacy from 1908. The ingenuity of Gosset, and later Fisher, transformed the practice of science. Haber's ingenuity made possible the increase in per capita food production that accompanied the massive increase in global population in the 20th century. Pedometricians are privileged to sit at the interface between these two areas of science and technology, but with that comes a responsibility to engage with some pressing problems.

Best wishes for 2009, and I hope that as many of you as possible will be able to attend Pedometrics 2009 in Beijing.

Murray

Hacking, I. 2006. The Emergence of Probability. 2nd Edition. Cambridge University Press.

'Student' 1908. The probable error of a mean. Biometrika 6, 1–25.

Best paper in Pedometrics 2007

A late surge in voting made for a good turn-out this year with 39 votes cast. We have also seen a wider participation in the vote than previous years, with many IUSS members not usually involved in the Commission's activities contributing. Budiman, the invigilator, used all the latest data mining techniques to detect any suspicious voting patterns, and can declare the result thoroughly free and fair.

The winning paper was:

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

Congratulations to Raphael and colleagues for an excellent paper.

The award will be presented at Pedometrics 2009 in Beijing, when we shall also announce Best Paper in Pedometrics 2008.

Once again, please consider which papers published in 2008 you consider merit the title best paper. Send your nominations to Murray (murray.lark@bbsrc.ac.uk) by the end of February 2009. We shall then ask a senior pedometrician to complete a set of five nominations.

☆Pedometrics 09☆

One World One Soil

Beijing, 26-28 August 2009

The Biennial Meeting of Commission 1.5 Pedometrics, Division 1 of the International Union of Soil Science (IUSS) will be held at the International Conference Centre - China Agricultural University Beijing.

Calling for Abstracts Now

Important Dates:

Deadline for submission of abstract: 31 March 2009

Notification of paper acceptance: 1 May 2009

Deadline for Registration: July 1, 2009

Visit http://2009.pedometrics.org to submit your abstract on all aspects of pedometrics research.

同一世界同一土壤

以前
只有一种土
在我们的世界上

对我来说
是山西
黄土高原黄绵土

无需地图
没有怀疑
没有争论

我经常问
为什么我必须
找到另一种
再另一种

大卫万德 林登

One World One Soil by David Van Der Linden

Report from the 3rd Global Workshop on Digital Soil Mapping

By Janis Boettinger

The 3rd Global Workshop on Digital Soil Mapping, "Digital Soil Mapping: Bridging Research, Production, and Environmental Application" took place 30 September through 3 October 2008 on the campus of Utah State University in Logan, Utah, USA (DSM 2008; http://dsmusa.org). Digital soil mapping is the generation of georeferenced soil databases by quantitative modeling of field, laboratory, and environmental data. The IUSS Working Group on Digital Soil Mapping (http://www.digitalsoilmapping.org/) organizes biennial Global Workshops on Digital Soil Mapping in alternate years with the conferences of IUSS commission 1.5 Pedometrics. DSM 2008 in Utah, USA, followed DSM 2006 in Rio de Janeiro, Brazil, and DSM 2004 in Montpellier, France.

The goal of DSM 2008 was to explore the state-of-the-art in digital soil mapping and to develop strategies for bridging cutting-edge research, production soil mapping, and environmental applications of digital soil data. DSM 2008 attracted 99 participants from 20 countries representing all populated continents.

The workshop was kicked off by the field trip on 30 September. About 50 participants traveled from the bottom of Cache Valley to the top of the Bear River Range, exploring the origins of the Basin and Range, the Great Basin, Pleistocene pluvial Lake Bonneville, orographic climate effect and alpine glaciation, the Sevier thrust belt, and the Middle Rocky Mountains. We observed and discussed clayey sodic soils in deep-water deposits of Lake Bonneville, the use of Landsat spectral data for digital soil mapping of saline and wet soils on the shores of the Great Salt Lake, soils illustrating the dramatic effects of aspect on organic matter accumulation and carbonate translocation on steep slopes, the factors influencing rapid cementation of soil horizons by calcium carbonate, and the impacts of aspen (broadleaf) vs. conifer vegetation on soil development. (French visitors were surprised to see Paris and Montpelier from the overlook at the final stop – Paris and Montpelier, Idaho.)

The workshop program spanning 1-3 October was organized into seven session themes (http://dsmusa.org/programm.pdf): 1) Evaluating and using legacy data in digital soil mapping; 2) Exploring new sampling schemes and environmental covariates in digital soil mapping; 3) Using integrated sensors or other new technologies for inferring soil properties or status; 4)Innovative inference systems (new methodologies for predicting soil classes and properties, and estimating uncertainties); 5) Global Digital Soil Mapping; 6) Using digital soil mapping products and their uncertainties for soil assessment and environmental applications; and 7) Protocol and capacity building for making digital soil mapping operational. The importance of our work during DSM 2008 was framed by the inspiring keynote presentation by Dr. Alfred Hartemink (ISRIC World Soil Information, Netherlands), “Soils are Back on the Global Agenda.” Each session theme was launched by a keynote presentation, followed by short presentations and 30 minutes of fo-

cused discussion. Keynote speakers for the sessions 1 through 7, respectively, were Dr. Alan Hewitt (Landcare Research, New Zealand), Dr. Janis Boettinger (Utah State University, USA), Dr. Raphael Viscarra-Rossel (CSIRO, Australia), Dr. Thorsten Behrens (University of Tübingen, Germany), Dr. Alex McBratney (The University of Sydney, Australia), Dr. John Triantafilis (University of New South Wales, Australia), and Dr. Sabine Grunwald (University of Florida, USA). The workshop inspired participants to manifest the future of DSM from diverse perspectives, using innovative approaches, and addressing the ultimate challenge of creating and delivering a global digital soil map.

The organizers sincerely thank participants for traveling to Logan, Utah, USA, to make DSM 2008 a successful and productive endeavor. We look forward to seeing you at DSM 2010!

View of the syncline illustrating the Sevier thrust belt in the Bear River Range at Stop #6.

Group at Field Trip Stop #1 exploring the origins of the Basin and Range, the Great Basin, and pluvial Lake Bonneville

Participants examining the sodic soil formed in silty clay deposits of Lake Bonneville.

Bob MacMillan, Florence Carré, Thorsten Behrens, and Raphael Viscarra-Rossel showing off their tokens for a free scoop of Utah State University's famous Aggie Ice Cream.

Alfred Hartemink and Amada Moore observing soils under aspen and conifers at Beaver Mountain Ski Area, Stop #8.

Photos taken by David Howell

Photos by David Howell

Photos by John T.

Report from Eurosoil 2008

The Eurosoil Congress 2008 was held in Vienna, Austria, 25 - 29 August 2008, under the general theme "Soil - Society – Environment". The principal organizer was Professor Winfried E.H. Blum from the University of Natural Resources and Applied Life Sciences (BOKU) Vienna, in cooperation with national soil science societies of Austria, Croatia, Czech Republic, Hungary, Slovakia, Slovenia and Switzerland. Mirjam Uebelhoer from Vienna Medical Academy was responsible for technical organization. A convenient venue was provided by the University of Technology, Vienna.

The Congress was attended by as much as about 1500 participants from 77 countries. One symposium was devoted to Pedometrics and Digital Soil Mapping. It took place on Tuesday, August 26, 2008. The morning keynote talk was given by Murray Lark. He showed some basic assumptions and approaches of pedometrics that "combines strands of speculative science and practical art". Selected methods of pedometrics were presented on real datasets. The keynote speech given in a very comprehensive way even for non-pedometricians was a perfect beginning of the symposium.

The symposium then consisted of 18 regular talks divided into four topics: Methods and Approaches of Pedometrics, DEM/DTM Exploitation in Digital Soil Mapping, Geophysics and Other Auxiliary Data Exploitation in DSM, and Pedometrical Applications. The presenters were from Germany, France, Belgium, Italy, Austria, the Netherlands, Spain, and the United Kingdom. Moreover, 24 posters of this symposium were presented by participants from different European countries and even from outside Europe, namely from Iran and Egypt. The number of presentations on this symposium is not very big compared to the total number of about 650 oral and about 750 poster presentations of the whole conference. Nevertheless, the meeting room was pretty full, so that the symposium attracted quite a lot of people. Moreover, there were numerous presentations on other symposia that used methods of pedometrics and digital soil mapping and showed their various applications.

The topics of both oral and poster presentations on the symposium indicated the principal task for digital soil mapping and soil spatial variation assessment nowadays. This task is to obtain sufficient amounts of data with sufficient density. Using DEM and other auxiliary data, particularly from emerging technologies like ground penetrating radar, gamma-ray spectrometry, electrical resistivity, or EMI, is a way how to get the information.

The symposium did not start probably any revolution in pedometrics, the presentations showed rather different applications of pedometrics and DSM than some new numerical methods and models. Nevertheless, we believe that the symposium was not important only for the community of pedometricians, but that its importance lies mainly in the fact that people from other fields of soils science could see what is the state-of-the-art in this field or just what pedometrics and digital soil mapping are about. Opening pedometrics to wider soil science public on such a large event might be the most important moment of this symposium.

We as conveners of the symposium would like to thank the scientific and organizing committees of the EUROSOIL Congress 2008 for the possibility to organize this symposium, all the presenters for their valuable contributions, and Murray Lark and Thorsten Behrens who helped us to chair the sessions.

We can look forward to the next EUROSOIL Congress that will be held in Bari, Italy, in 2012. Let's hope that pedometrics and digital soil mapping will be presented there even more.

Lubos Boruvka, Czech Republic

Endre Dobos, Hungary

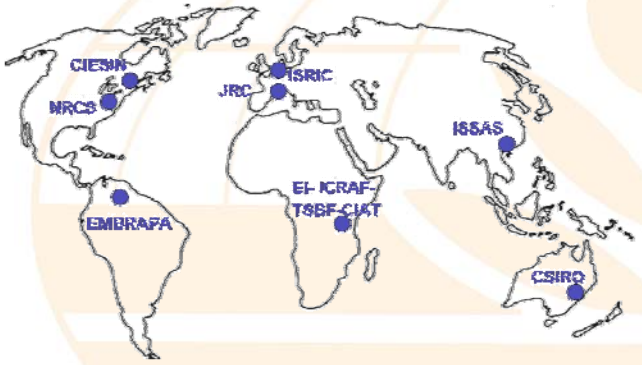
GlobalSoilMap.net

[glö'bəl soll mæp dot nɛt]

First planning meeting GlobalSoilMap.net

27-29th September 2008 Logan, Utah, USA

A two day planning meeting was held at the University of Utah, USA, preceding the Third Global Workshop on Digital Soil Mapping. Representatives of all nodes of the global consortium were present. The meeting focused on activities in all nodes followed by a technical workshop on key variables, and data protocols, and a workshop on web delivery. The last day was spent on the launches, timetables and all other business. Some of the main issues resolved during this meeting were minimum data set, spline functions + uncertainty for each cell, writing of 4 manuals, the set-up of SharePoint, regular teleconferences, and some ideas for funding for each of the nodes.



Each node representative gave an overview of activities in the region/country. Key issues were then countries included in each of the nodes and the approach for funding. Overall, there is limited awareness of the project in all nodes – regional workshops advocating the concept, bringing together the national institutes as well as developing funding strategies were considered attractive but funds are lacking to organise such workshops. Once the project has been launched awareness may grow which will facilitate the activities to set up the nodes.

The data model that will be developed for GlobalSoilMap.net should be based on the key questions on global soil information. There are generic and specific questions and that should be reflected in both the model and minimum data set. The design of the model should be for the future, compatible with other global efforts, and it is therefore important to link up with GEOSS (Global Earth Observation System of Systems). After some discussion on the model and discussing some variants (notional 5 layer model, idealised profiles) the group opted for the function or spline model[1].

The minimum data set was discussed: there are primary properties that are much needed, can be predicted and should be available across the globe, and a range of secondary properties that will probably be different in different parts of the world. The primary properties are:

- organic C (affects fertility and physical properties)
- clay content
- bulk density.

From these attributes, C density and available water capacity can be inferred. Secondary properties are pH, ECEC (cations + exchangeable acidity) and EC, which are available from many soil survey reports and have been routinely measured. Then there is range of other properties (available P, DCB Fe, N etc) that are not part of the core variates. It was agreed that the SRTM 90m data will be used. GlobalSoilMap.net will harvest from the nodes that will remain solely responsible that the national data fit into the global set up. The question on updates and interpretations remain to be discussed but the metadata standards have to be set by the project.

The project will be officially launched in 2009: on 12-16th January in Nairobi, Kenya, and on 17th February, New York, USA. The global launch will take place at the Columbia University. The whole consortium recognised the importance of a timely and big-bang launch – it will create much needed attention for fund raising and get national institutions involved.

Left to Right: Amanda Moore, Janis Boettinger, Alex McBratney, Jon Hempel, Alfred Hartemink, Mike Grundy, Lou Mendonça Santos, Florence Carre, Neil McKenzie, Pedro Sanchez, Vincent van Engelen, Ganlin Zhang

[1]See: Bishop, T.F.A., McBratney, A.B. and Laslett, G.M., 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, 91: 27-45

Finding the Boundary

Alice Milne & Murray Lark

Pedometricians are soil scientists first and statisticians second. This means that our models should not just be statistically convenient, they should also make sense when we think about the soil. Consider a simple problem: how can we express, as a model, the relationship between clay content of the soil and the organic carbon content? In some conditions a linear model of some sort is reasonable. As we consider soils of increasing clay content so the available water capacity may increase, and so the net primary production of the vegetation it supports, and so the litter inputs and so on. The expected organic carbon content of the soil increases with clay content, but at any given clay content there will be variations around this mean due to other factors. But it is not always so simple. Some soil scientists have suggested that a key factor in soil organic matter dynamics is the protection of a fraction of the organic carbon against bacterial action by its close association with clay particles. If we consider agricultural land, where the equilibrium carbon content is lower than under natural vegetation, it is therefore possible that the clay content determines a minimum organic carbon content (the residual amount inherited from past vegetation and locked up in complexes with clay particles). At a given clay content, then, there is a minimum organic content, but we might see more (e.g. in soils where large inputs of organic matter have been made over time). What we need to model is a lower boundary on the plot of organic carbon content against clay content.

Similar thinking has been followed by some soil scientists concerned to model nitrous oxide emissions from the soil. It has its origin in the 'law of the minimum' propounded by Justus von Liebig (1863), a founding father of agricultural chemistry. At a given concentration of nitrate in the soil, for example, there is a maximum rate of denitrification determined by the chemical kinetics. However, a particular soil might not express the maximum rate given its nitrate content — perhaps because the pH is too low, or the oxygen partial pressure is too high. In this case the response to nitrate, in a plot of measured emission rates against nitrate concentration, will be expressed by an upper boundary.

Interest in such boundaries goes back a long time. Figure 1 shows measurements of strawberry weight against achene number. Achenes are the 'pips' on the

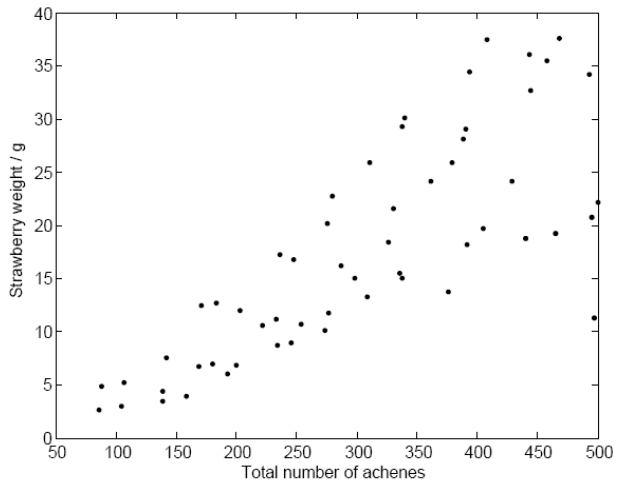


Figure 1. Simulated measurements of strawberry weight (g) against total achene number (based on Webb, 1972).

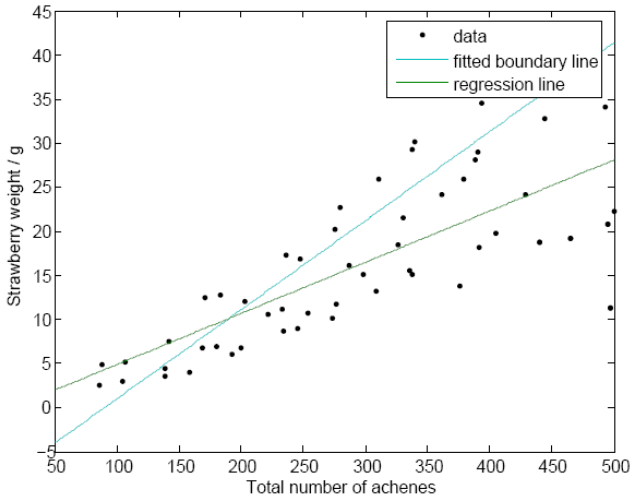


Figure 2. Simulated measurements of strawberry weight (g) against total achene number, with a regression line and an upper boundary line fitted. The boundary line was fitted using the method described in Milne et al. (2006b).

surface of the strawberry. The achenes are the true fruits in the botanical sense, and the fleshy part of the berry is the receptacle of the original flower which swells in response to auxins that the achenes release after fertilization. The number of achenes that are fertilized therefore determines the total release of auxin and the size of the berry. However, a strawberry on a water-stressed plant might not be able to grow as large as the number of fertilize achenes would permit. For this strawberry it is water and not auxins that are limiting. The biological effect of achene number on strawberry weight might therefore be expressed by an upper boundary on the plot. A regression line would give us a prediction of berry weight from achene numbers, but it would not express the underlying biological relationship. Figure 2 shows a regression and an upper boundary line fitted to the data.

The boundary line concept is attributed to Webb (1972), who suggested that the upper boundary (or in some circumstances lower boundary) may be of more biological interest than the line of best fit after examining data on strawberry weight and achene number provided by his colleagues at Long Ashton Research Station in western England. However, as we note above, the concept of limiting factor models predates this significantly (von Liebig, 1863). The boundary line model has been of practical interest to biologists in various contexts. For example, in horticulture (Abbott et al., 1970), to model trace gas emissions from soil (Elliot & de Jong,1993; Schmidt et al., 2000), to determine limiting ranges of the values of soil nutrient tests (Schnug et al., 1996), to interpret soil sensor data (Kitchen et al., 1999) and for modelling within-field yield variation as a basis for recommendations on site-specific management (Shatar & McBratney, 2004). Unfortunately, in many applications the methods used to estimate the boundary line model have been somewhat ad hoc lacking a theoretical basis. One of us recently attended a meeting on soil modelling, held under the auspices of the European Science Foundation. A well-known soil scientist reported some boundary line results, and when asked how the lines had been modelled, admitted that he had drawn them with a ruler. That is a fair starting point, but we need to do better. In particular we need some way to test the hypothesis that one variable determines a boundary for the response. Boundary line models are plausible in some circumstances, but not all. If auxin release and water content interact to determine strawberry weight, for example, then a boundary model might be inappropriate.

In our papers Milne et al. (2006a&b) we addressed these two issues by developing a statistically robust method for fitting a boundary line model to data, as well as presenting methods to assess the suitability of the boundary model for the data.

To fit the boundary line model we assume that the data are from a censored probability distribution where the censor is the boundary line. An example of a censored probably distribution is shown in Figure 3, illustrating the suitability of the distribution for this application. Maximum likelihood is used to fit the parameters of the distribution and confidence intervals are deduced. The model also includes a parameter which accounts for measurement error in the dependent variable. This influences the position of the boundary line in the envelope of the data. The smaller the measurement error the closer the boundary line will be to the data envelope. The form of the boundary line (i.e. linear, quadratic) is selected prior to the parameter fitting and is based on biological knowledge and/or visual inspection of the data. The boundary line can take any number of forms but as complexity increases we are less likely to find a sensible fit with the optimization algorithm.

So what do we mean when we ask if the boundary line model is a suitable description of our data? Firstly does the boundary line model make biological sense or is the mechanism relating the data better represented by some other model such as regression. Secondly, are there enough data in the region of the true boundary line for us to estimate its form? Consider a case where nutrients and available water determine crop yield in accordance with a limiting factor model.

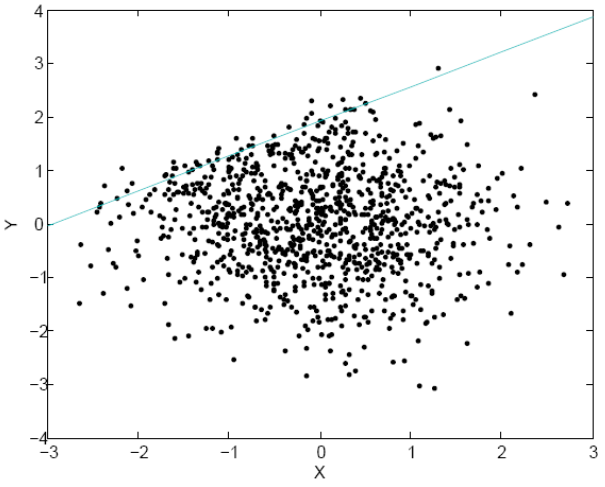


Figure 3. An example of a censored probably distribution with upper boundary line fitted using the method described in Milne et al. (2006b).

If crop yield is limited by nutrients at all sites we have observed, then available water is nowhere limiting, and the envelope of a plot of yield against water will not be a realization of the boundary but will fall within it. We developed a method based on convex hulls which compares the density of points in the neighbourhood of the perceived boundary with the number of points we would expect if the data were simply a realization of a bivariate normal distribution. If the data are limited by the independent variable then we would expect to see a higher density of points in the region of the boundary than from a bivariate distribution. We test the null hypotheses that the data are from a bivariate distribution. The p-value gives us a measure of the strength of evidence we have that a boundary line is represented in the data.

A second method of assessing the fitted boundary line model is to compare its goodness of fit with a bivariate normal model. This is done using Akaike's information criterion (AIC) (Akaike, 1973), which allows us to compare model performance based on a compromise between parsimony and goodness of fit (which usually improve with increasing numbers of parameters). The smaller the AIC value the more appropriate the model is.

Case study: Soil organic carbon and clay content

Figure 4 shows measurements of clay content and soil organic carbon (SOC) made on the Broadbalk wheat experiment at Rothamsted, Harpenden, UK (Figure 5) (Watts et al., 2006). As noted above, we expect to see a lower boundary in the data. As an aside, it is interesting to note that the Broadbalk experiment was

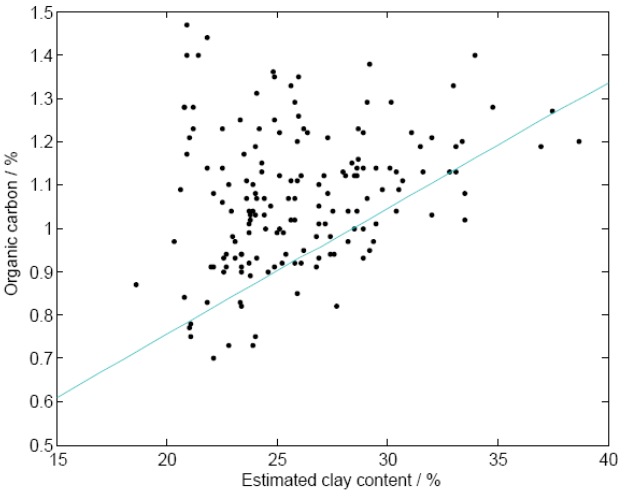


Figure 4. Measurements of clay content and soil organic carbon (SOC) made on the Broadbalk wheat experiment at Rothamsted, Harpenden, UK (Watts et al., 2006). The straight line is the lower boundary fitted using the method described in Milne et al. (2006b).

Figure 5. Broadbalk field at Rothamsted, a long-term experiment on the soil under a wheat crop started in 1843. ©Rothamsted Research, used with permission.

set up in 1843 by John Lawes to counter Justus von Liebig's over-simple application of his own law of the minimum to the nitrogen nutrition of plants. Liebig held that, since 80% of the atmosphere is nitrogen, nitrogen cannot limit plant growth. We know now, of course, that what matters to the plant is the oxidation state of the nitrogen, and the plant requires nitrate or ammonium from the soil, which may well be in limiting supply.

A straight line model was chosen for the boundary line (y=ax+b, where y is organic carbon and x is clay content) . This made sense conceptually and visual inspection suggests it is appropriate. The fitted parameter values are a = 0.029 and b = 0.176 with confidence intervals (0.0213, 0.037) and (-0.062, 0.415) respectively. The measurement error was estimated to be 0.095 with confidence interval (0.069, 0.121). The fitted line is shown on Figure 4.

The convex hull test showed no significant evidence of a boundary at p = 0.05 but the AIC proved the boundary model (AIC = 727) to be better than a regression type model (AIC = 746). The fact that we sometimes get different inferences from the two tests is not surprising as they are testing different (though

biologically linked) hypothesis. Visually, there is a clear increasing trend on the lower part of the data suggesting a limiting relationship, and we can biologically explain this observation. However, the data are sparse in the neighbourhood of the boundary which is reflected by the results of the convex hull test. A few additional points near the lower bound could change the positioning of the boundary estimate by a reasonable amount so it would be unwise to have too much confidence in it. Ideally more data should be sought. We could conclude that biological knowledge of the system, visual inspection and the results of the AIC suggest a boundary model is appropriate, but the convex hull test result makes us cautious about our ability to estimate the location of the boundary given the limited data set.

The boundary line methods discussed here improves on other methods in the literature. To our knowledge, no other procedures have been proposed to detect the presence of a boundary — although such analysis should accompany any attempts to fit a boundary. The boundary line model we propose has a clear theoretical basis, and parameters can be meaningfully interpreted.

The programs developed to do the analysis can be accessed on the web at

http://www.rothamsted.bbsrc.ac.uk/bab/index.php?folder=environmetrics&page=alice1 where there is also a more detailed description of the methods and more examples.

If you would like to discuss applying our boundary line methods to your data please email alice.milne@bbsrc.ac.uk.

Acknowledgement We thank Chris Watts, Lawrence Clark, Paul Poulton, David Powlson and Andy Whitmore for letting us use their data in this study.

References

Abbott, A. J., Best, G. R., Webb, R. A. 1970. The relation of achene number to berry weight in strawberry fruit. Journal of Horticultural Science 45, 215-222.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In Petov, B.N. and Csaki, F., (Eds). 2nd International Symposium on Information Theory, pp. 267-281. Akademia Kiado, Budapest.

Elliot, J. A., de Jong, E. 1993. Prediction of field denitrification rates: a boundary-line approach. Soil Science Society of America Journal 57, 82-87.

Kitchen, N. R., Sudduth, K. A., Drummond, S. T. 1999. Soil electrical conductivity as a crop productivity measure for claypan soils. Journal of Production Agriculture 12, 607-617.

Milne, A. E., Wheeler, H. C., Lark, R. M. 2006a. On testing biological data set for the presence of a boundary. Annals of Applied Biology 149, 213-222.

Milne, A. E., Ferguson, R. B., Lark, R. M. 2006b. Estimating a boundary line model for a biological response by maximum likelihood. Annals of Applied Biology 149, 223-234.

Schmidt, U., Thöni, H., Kaupenjohannm, M. 2000. Using a boundary line approach to analyze N2O flux data from agricultural soils. Nutrient Cycling in Agroecosystems 57, 119-129.

Schnug, E., Heym, J., Murphy, D. P. 1996. Establishing critical values for soil and plant analysis by means of the Boundary Line Development System (BOLIDES). Communications in Soil Science and Plant Analysis 27, 2739-2748.

Shatar, T. M., McBratney, A. B. 2004. Boundary-line analysis of field-scale yield response to soil properties. Journal of Agricultural Science 142, 1-7.

Von Liebig, J. 1863. The Natural Laws of Husbandry. London: Walton and Maberly.

Watts, C. W., Clark, L. J., Poulton, P. R., Powlson, D. S., Whitmore, A. P. 2006. The role of clay, organic carbon and cropping on mouldboard plough draught measured on the Broadbalk Wheat Experiment at Rothamsted. Soil Use and Management 22, 334-341.

Webb, R. A. 1972. Use of the boundary line in analysis of biological data. Journal of Horticultural Science 47:309-319.

Some experiments on using data-mining techniques for digital soil mapping

Budi & Alex

There is an increasing use of data mining techniques for soil prediction, especially in DSM as seen in the last workshop. Combined or ensemble models, such as boosting and bagging, have become very popular.

Leo Breiman in a paper in Statistical Science (2001) compared the two cultures in statistical modelling: data modelling and data mining. Breiman argued that with increasing number of data, sometimes it is impossible to draw a mechanism from the data, and if our goal is to use data to solve problems, then we should adopt the black-box approach. Breiman contended that nature's mechanisms are generally complex and cannot be summarised by simple models (such as linear or logistic regression). Thus, inferring a mechanism can be risky, a deceptively simple picture of the inside. Breiman promoted the Rashomon effect (generating a multiplicity of models) rather than Occam's razor (simplicity). He stated that accuracy and simplicity (interpretability) are in conflict. Dimensionality is a blessing rather than a curse. [See Box 1]

Following his argument, Breiman also introduced Random Forests, a data-mining algorithm with software that is freely available in Fortran code or in R. It is a classification and also a regression model, where the algorithm build many trees based on bootsrapping and aggregating the results. Random forests combine the tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The model can handle lots of inputs, has a high accuracy and also does not overfit! This is achieved by using different subsets of the training data (with bootstrap) and using different subsets of the predictors for training the tree (determined randomly). Thus only patterns that are present in the data would be detected consistently by a majority of the trees. See http://www.stat.berkeley.edu/~breiman/RandomForests/ for more details.

This is certainly a great feature, that means you can put all your covariates (or *scorpan* factors) in the model, don't need to worry about the form of the model, let Random Forests figure out the model and which variables are important for prediction. And you don't have to worry about overfitting.

Is there such a thing as a panacea? Let's give it a try.

First, we try prediction for generating pedotransfer functions to predict water content at wilting point (-1500 kPa). We used the US-SCS soil database contained in the ISRIC global database. After removing outliers, we obtain 15000 data points for prediction and 10000 points for validation. First we generate a model predicting WP from 4 well-known and obvious predictors (clay, sand, organic C, CEC). We obtained R^2 = 0.84 (RMSE = 2.9%) for the validation set, higher than just using simple linear model R^2 = 0.79 (RMSE = 3.3%) . Next, we use 22 predictors (including depths, 10 sizes of particle distribution, exchangeable cations, pH, and bulk density). Using 22 predictors we obtain a higher R^2 of 0.89 (RMSE = 2.4%) (Fig. 1). So it seems it is true that RF does not overfit, and including more predictors can result in better prediction. Maybe RF can find some hidden relationship that we cannot see? RF identified clay and CEC, as the most important predictors in the 22 variables model (Fig. 2).

From a practical point, perhaps we should consider is the model with 22 predictors really useful. The difference in RMSE for the 22 and 4 predictors is 0.5%. Can we get an accuracy of less than 0.5% when measuring of soil moisture at wilting point? I don't think so.

Common sense and basic soil science are important, we should be wary when we make such soil predictions!

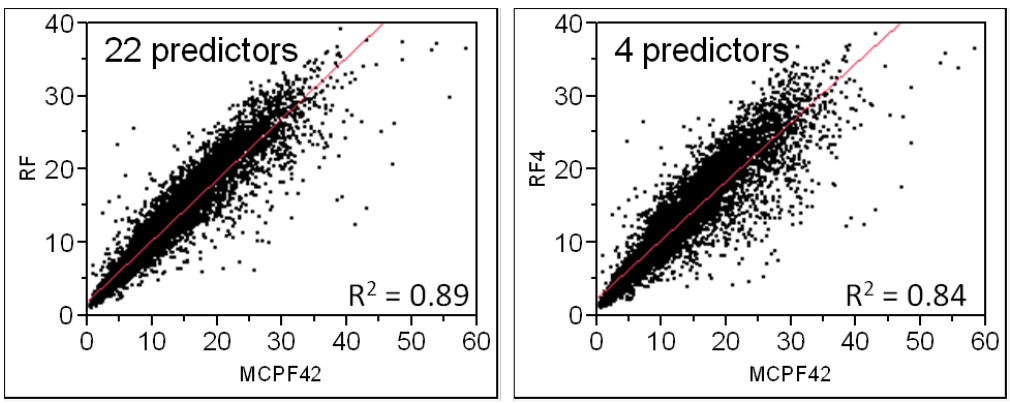


Fig. 1. Prediction of water content at wilting point (-1500 kPa) using Random forests based on 4 predictors and 22 predictors

In the next example, we test RF for digital soil mapping. We have a dataset of subsoil pH from the Hunter Valley in New South Wales, Australia. We split the data into 550 samples for prediction and 295 samples for validation.

We have covariates from a DEM (and derivatives), Landsat images (7 bands + derivatives), landuse (from aerial photo), soil-landscape units (from a soil map) and physiography. In total we obtain 32 covariates. We used Random Forests with 500 trees (as recommended).

From the validation results (Fig. 3), we can conclude that random forests is better than a simple linear model and using all covariates (including x,y) produce the best fit. More interestingly only using x and y (and elevation) give us a good fit. So we don't even need to purchase or download Landsat images.

But let's look at the maps generated in Fig. 4. Maps in Fig. 4c and d shows the most realistic prediction. Map 4b looks realistic but appears blocky. Maps 4c and d that use all covariates, show very detailed information from Landsat images and terrain attributes. It is interesting that the map generated using linear model has similar pattern to maps 4c and 4d. So, it might be simple model already captures the pattern of the properties, and using more predictors and possible

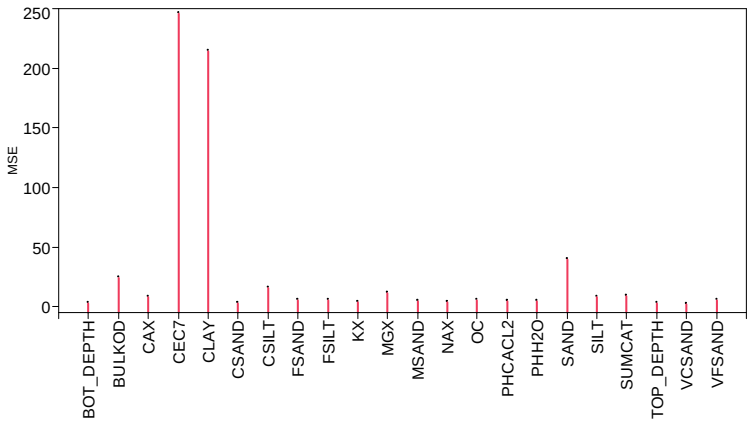


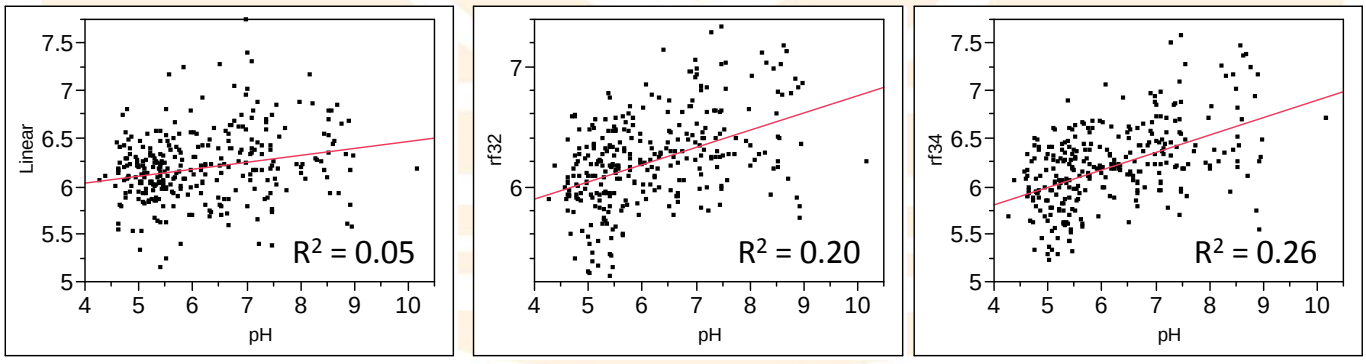
Figure 2. Variable importance for prediction of water content at wilting point with random forests

interactions, the resulting maps appear smoother.

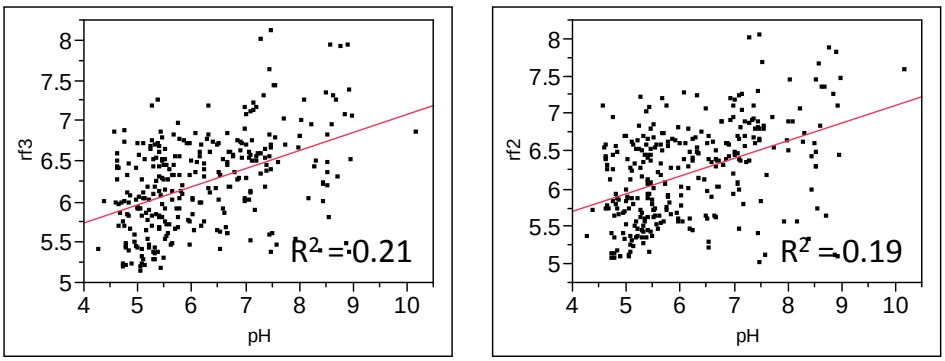
Maps that only used x,y coordinates as predictors are not realistic, they give artificial splits just based on location of the samples and spatial coordinates (Fig. 4e). Including elevation (Fig. 4f) we just see a map with artificial split of spatial location and contour.

Let us consider the problem of overfitting. It is true RF give us a higher accuracy compared to linear model, even just using x and y coordinates as predictors. But is this real? The map appears unrealistic.

This is a danger of using spatial coordinates in any



Linear regression: 7 cov. RF: 32 covariates RF: 32 covariates+ X,Y



RF: Elevation + X,Y RF: X, Y

Figure 3. Prediction of subsoil PH in the Hunter Valley based on environmental covariates.

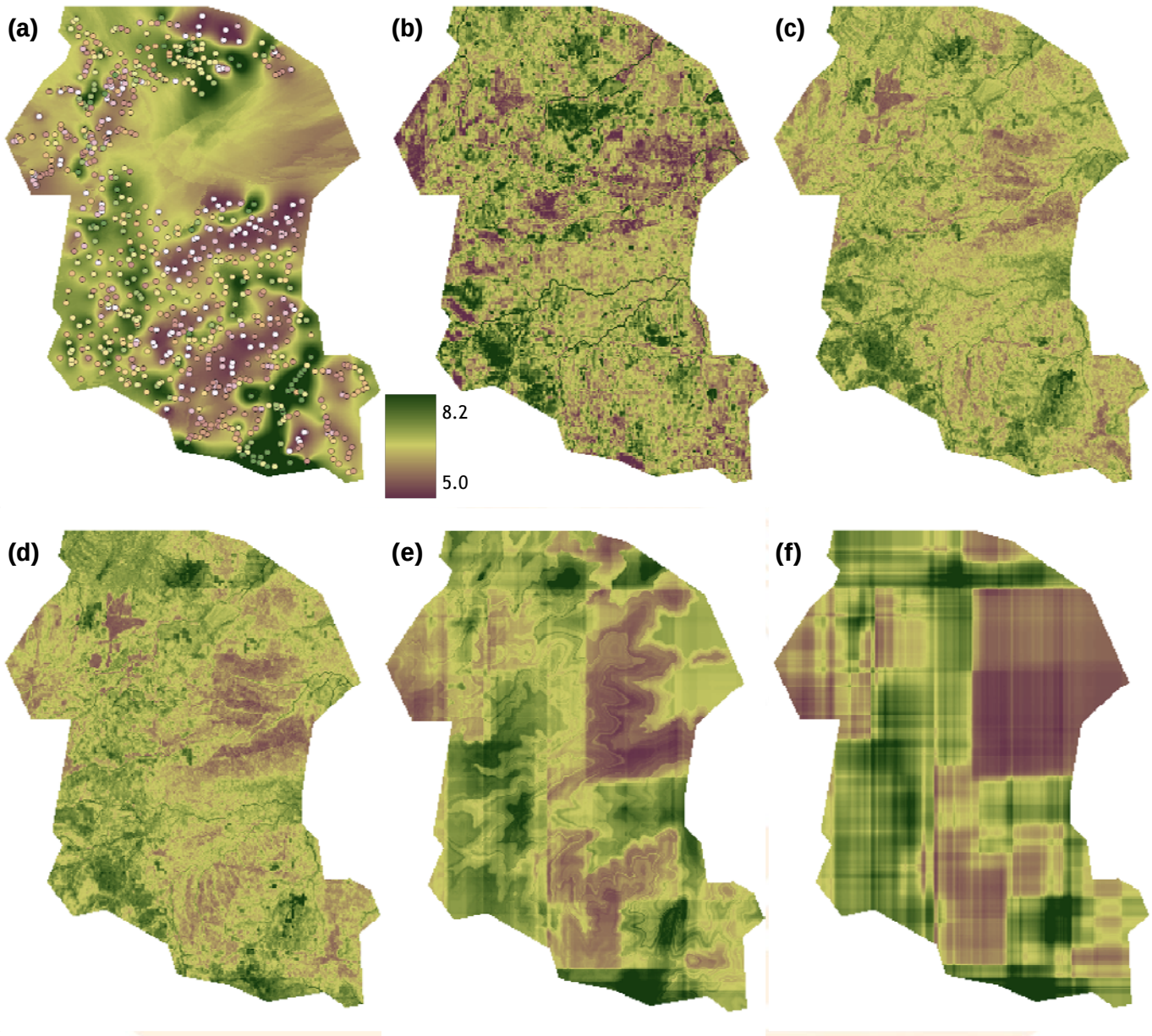


Figure 4. Prediction of subsoil pH in the Hunter Valley. (a) Points represent measurement and the surface is the ordinary kriged value. (b) prediction using a linear model with 7 covariates, (c) prediction using Random Forests (RF) with 32 covariates, (d) prediction using RF with 32 covariates and x,y, (e) prediction using prediction using RF with x,y, and elevation, (f) prediction using prediction using RF with x,y .

regression or tree models. It just partitions the data to find regions that give the lowest error. How do we measure the "reality" of the predicted map.

How many parameters are needed to split the landscape into these predictions? Each tree in the forest has 110 nodes (yes, that many nodes, determined from number of data), and we used 500 trees. That means the model has 55000 parameters. Does it still not matter as long as we get a good prediction?

Imagine Figure 4e. being the variable or feature space, we can see that RF is not finding the trend in the data space, it just partitions it into areas which will give you a good fit based on the data.

While x,y represent scorpan's *n*, the space factor, it should be mentioned that in this context it should represents the spatial trend in the landscape, not some artificial split of the space based on data.

n not only represents spatial position such as easting, northing, or easting squared, northing squared (the familiar trend-surface parameters), but also more sophisticated [non-linear] and contextual representation of space as appropriate, i.e., distance from a watershed, along the ground distance from a stream, distance from roads, distance from a point source etc. The distance metrics don't need to be Euclidean. You can use your ingenuity here. It's worth while thinking about this.

From this example of using ensemble models, we should be thinking twice considering what need to be put in the model for prediction and whether a good R^2 and low RMSE is enough to justify how many variables we can put in the Random Forests.

Do not carry out soil data dredging (see Box 1). Be sensible and your common soil knowledge is important. See also Lark et al. (2007) on the use of expert knowledge and with control of false discovery rate to select predictors.

When you get lost in the Random Forests, there is no easy way to get out.

References

Breiman, L., 2001. Statistical modeling: The two cultures. With discussion. Statistical Science 16, 199-231.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5-32.

Lark, R.M., Bishop, T.F.A., Webster, R., 2007. Using expert knowledge with control of false discovery rate to select regressors for prediction of soil properties. Geoderma 138, 65-78.

Box 1

Data dredging

is the inappropriate search for 'statistically significant' relationships in large quantities of data with a large number of variables.

The Curse of dimensionality

is a term coined by Richard Bellman, an applied mathematician, known for his invention of dynamic programming. It describes the problem caused by the exponential increase in volume associated with adding extra dimensions to a (mathematical) space. One implication of the curse of dimensionality is that some methods for numerical solution of the Bellman equation require vastly more computer time when there are more state variables in the value function.

Ockham's razor

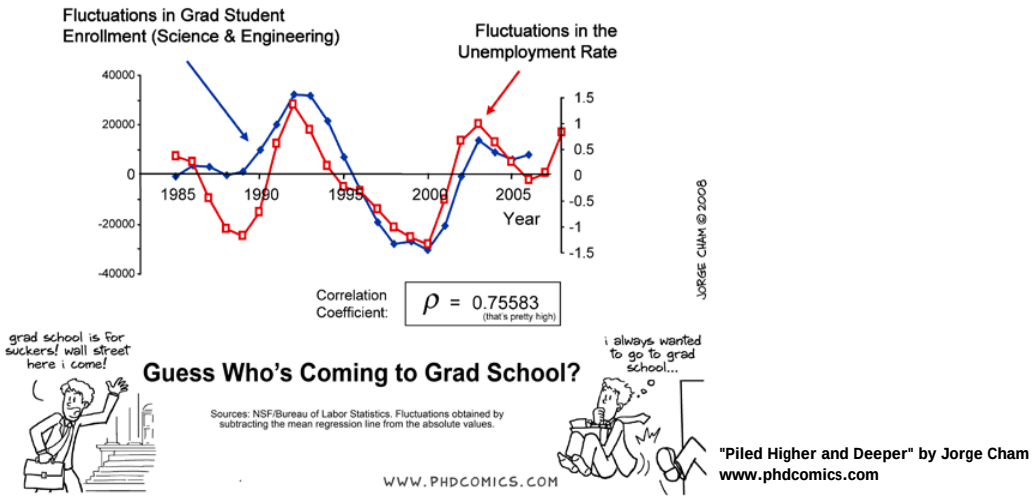
is a principle attributed to the 14th century English logician and Franciscan friar, William of Ockham. The principle states that the explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis or theory. The principle is often expressed in Latin as the lex parsimoniae ("law of parsimony"): "Pluralitas non est ponenda sine necessitate" which translates as "plurality should not be posited without necessity".

Rashomon effect

In psychology, this is the effect of the subjectivity of perception on recollection, by which observers of an event are able to produce substantially different but equally plausible accounts of it. It is named after a Japanese movie directed by Akira Kurosawa, in which a crime witnessed by four individuals from different vantage points. When they come to testify in court, they all report the same facts, but described in mutually contradictory ways.

Breiman (2001) called Rashomon Effect for data mining when a multitude of different descriptions or equations in a class of functions giving about the same minimum error rate.

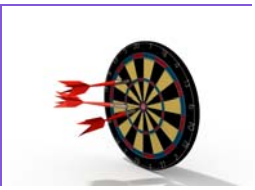
Source: http://en.wikipedia.org/wiki/



"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com

Did you miss this? ...

Murray



Not all papers on pedometrics are published in journals of soil science. The aim of this feature in Pedometron is to draw readers' attention to papers from non-soil science journals that tackle pedometrical problems. Send your submissions to the editor, with full publication details and a summary in your own words of the paper's message. To kick off, Murray Lark summarizes:

Breidt, F.J., Hsu, N-J & Ogle, S. (2007). Semi-parametric mixed models for increment-averaged data with application to carbon sequestration in agricultural soils. Journal of the American Statistical Association 102 (part 479), 803–812.

These workers, based at Colorado State University and National Tsing-Hua University, Taiwan, faced an interesting problem in meta-analysis. They had data from 63 paired studies across North America in which low-tillage and conventional cultivation were compared with respect to soil organic carbon content of the soil. They wanted to be able to answer questions such as 'at what depth is the effect of tillage practice most apparent after 15 years low-till?', or 'what is the total difference between the carbon content of the top 30cm of soil after H years?' The problem was that soil carbon was not recorded on the same depth increments in all studies. In fact there were no less than 211 different depth increments used in these studies.

The solution was to propose a semiparametric mixed model for the carbon content over a depth increment. The carbon content in the jth increment from the ith study is modelled as a site and increment-specific function of a set of covariates, and an integral, over the corresponding depth interval of an additive varying-coefficient model in which the site-specific coefficients combine smooth functions from a basis set such as polynomials or splines. There is an independent random error, and a fixed effect which is an integral of a correlated random variable over the depth increment of interest. The variance components can be estimated by REML, and the coefficients and random effects are then predicted.

The results of the analysis showed marked carbon storage under low-till relative to conventional cultivation in the top part of the plough-layer (to around 10cm depth), as expected, with the effect markedly greater in wet climatic conditions than dry. In the bottom part of the plough layer (10–18cm) there was more carbon storage under traditional cultivation. Below 18cm the difference in carbon content between the treatments was not significant.

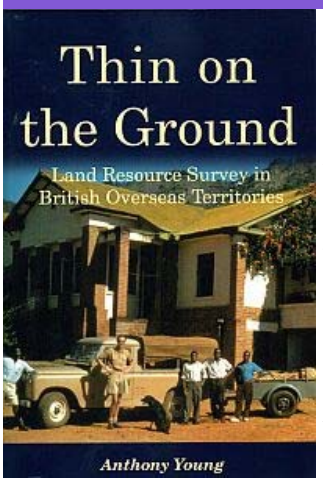
Semiparametric Mixed Models for Increment-Averaged Data With Application to Carbon Sequestration in Agricultural Soils

F. Jay BREIDT, Nan-Jung HSU, and Stephen OGLE

Adoption of conservation tillage practice in agriculture offers the potential to mitigate greenhouse gas emissions. Studies comparing conservation tillage methods to traditional tillage pair fields under the two management systems and obtain soil core samples from each treatment. Cores are divided into multiple increments, and matching increments from one or more cores are aggregated and analyzed for carbon stock. These data represent not the actual value at a specific depth, but rather the total or average over a depth increment. A semiparametric mixed model is developed for such increment-averaged data. The model uses parametric fixed effects to represent covariate effects, random effects to capture correlation within studies, and an integrated smooth function to describe effects of depth. The depth function is specified as an additive model, estimated with penalized splines using standard mixed model software. Smoothing parameters are automatically selected using restricted maximum likelihood. The methodology is applied to the problem of estimating a change in carbon stock due to a change in tillage practice.

KEY WORDS: Core sample; Greenhouse gas; Nonparametric regression; Ornstein–Uhlenbeck process; Penalized spline; Restricted maximum likelihood; Varying-coefficient model.

Book Review



Thin on the Ground Land Resource Surveys in British Overseas Territories

by Anthony Young. The Memoir Club, Stanhope. ISBN 978-1-84104-175-9. 230 pp.

Anthony Young begins his book with a quotation from Curtis Marbut (whom he later tactfully but firmly divests of any claim to be the founder of soil survey). In 1923 Marbut observed that there was little systematic information available on the soils of Africa, beyond what could be deduced in broad terms from knowledge of climate (with the exception of parts of South Africa). However, by 1964 D'Hoore had been able to compile his famous Soil Map of Africa at 1:5 000 000 from information provided by local experts, and in 1961 FAO produced its Soil Map of the World from similar sources. How had the availability of soil information changed so dramatically in something under 40 years?

The answer is that from the 1920s through to the Second World War visionaries in various parts of the world began to explore the soil landscapes of large tracts of land. This laid the foundations for a period of intensive reconnaissance soil survey from the 1950s to the mid 1970s. It is this story that Young tells, at least for the British overseas territories of the time. Canada, Australia, New Zealand and South Africa had Dominion status at the very beginning of his period, but they are excluded from the book.

The story starts with the pioneers. Among these are some names to conjure with. Geoffrey Milne should be known to all pedometricians as the originator of the *catena* concept. This made possible small-scale mapping of large areas within which individual soil classes could not be delineated, but particular repeating patterns, linked to topography, could be described and then interpreted on the ground by a user of the map. Milne surveyed the territories of Uganda, Tanganyika (now Tanzania), Kenya and Zanzibar (now part of Tanzania). Another pioneer was Colin Trapnell who, equipped with an Oxford education in Latin and Greek literature, and some limited experience of field ecology from university expeditions, set out to map the vegetation and soils of Northern Rhodesia (now Zambia). Both Trapnell and Milne epitomized the best of field science, not merely describing and cataloguing, but deriving original and fruitful insights into the origins of soil and vegetation systems in the landscape. Their work was to provide a basis for subsequent surveys, not only because of the information that their surveys provided on large areas of Africa, but also because of the widespread value of the approaches to survey that they had developed.

In the aftermath of the Second World War came the period that Young calls the 'Golden Age' of reconnaissance survey. This built on the insights of the pioneers, but also on other developments, notably the use of air photography that was demonstrated in forest surveys of Burma the 1920s and subsequently used in Northern Rhodesia (as it then was) by Trapnell and his colleagues. Also important was the development of land system survey by the CSIRO, the first survey of the Katherine–Darwin region was completed in 1946.

The 'Golden Age' was also an age of change for the British overseas territories. In 1947 India and Pakistan achieved independence, and by the 1960s most of Britain's territories in Africa were on the way to independence. So Young's Golden Age started with a drive to develop the agricultural potential of the colonies, and ended with work done to support the agricultural development of newly independent countries. The political context of the work by the soil surveyors is outside the scope of Young's book, although it is clear that they often faced suspicion in the field from local people wondering exactly why they were taking an interest in the soil. Given the often explosive role that land and land tenure was to play in the post-colonial history of many former British territories (consider, for example, Zimbabwe) it would be interesting to see the politics of soil information and land evaluation explored in more detail.

So what did the soil surveyors of the Golden Age do? Well one of them has described his work in the pages of *Pedometron* (21, 5–8). Richard Webster, whose later career furthered the development of statistical methods in soil survey and nurtured the origin of pedometrics, worked as a surveyor in the late 1950s and early 1960s in Northern Rhodesia, shortly to attain independence as Zambia. His work was driven by the need to improve food production in areas where settlement was increasing due to the demand for labour in the copper mines. During this period, as the winds of change prepared to usher in a new era of decolonization, other expatriate workers were labouring to survey soils with potential for cocoa production (in Nigeria, Trinidad and elsewhere), for irrigated cotton production (the Gezira in Sudan) and for oil palm and rubber (Malaya) just to extract a few examples. Where such survey was not undertaken before attempts to encourage or introduce new crops the failure could be disastrous (the East African Groundnut Scheme's being a case in point). Some connection is made between soil survey and other soil science research, although it would be good to hear more about this, and the one example (the work by Nye and Greenland on shifting cultivation) is not very well summarized (they used a single-pool carbon model, for example, *pace* Young's account).

As well as describing the soil surveys, Young describes the people who did them. There were eccentrics, visionaries (whose grasp of the local production sys-

The Golden Age of reconnaissance survey: Richard Webster in a pit in Zambia. Courtesy of R. Webster.

tems as socio-economic and cultural wholes antedates recent interests in 'indigenous knowledge') and people whose lives became entirely immersed in the communities where they lived and worked. In these days when the bureaucrats rule the roost with their milestones and management systems it is hard not to envy those scientists who were often almost unencumbered by administration for the simple reason that they were working too far away to suffer from it. What is also clear is that these soil scientists were driven both by a passionate interest in their subject and by a desire to solve real problems and to promote the development of the land that they charted.

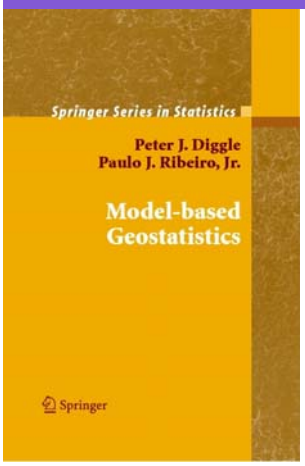
Some interesting themes arise in the book. One in particular is the question of what exactly the soil surveyor should be mapping. This was debated at various stages in Young's story, notably in the 1930s when Frederick Hardy (Trinidad) argued against general purpose surveys, which would typically attempt to delineate 'natural bodies' of soil on pedogenetic principles. Hardy saw little value in surveys that did not set out primarily to map constraints on cropping. His view fell out of favour for a while, but by the post-war period field surveyors, like Young himself in Nyasaland (now Malawi), were taking pains to use the soil map legend as a framework to present agronomic information and advice based on experimental results. There is an element of Anglo-Saxon empiricism versus idealism in this story, although, given the focus on British overseas territories both sides of the debate are not equally heard. However, Young does tell the tale of the first national-scale soil map of India, produced by a Russian professor in 1932 who saw no need actually to *visit* the country. Doubtless if she had visited, and found that Indian soils had not organized themselves on Dokuchayevian lines, it would, as with Hegel, have been so much the worse for the soil!

In some respects this debate stimulated the origins of pedometrics. The key idea behind the work on the utility and costs and benefits of soil survey that took place in Oxford in the 1960s and 1970s was that the value of a map could be judged by how well it allowed the user to predict soil conditions at unsampled sites. These ideas were first discussed by Philip Beckett and Richard Webster when the former was on an Oxford University Expedition to Zambia, which illustrates another of Young's points — that many of the soil scientists who cut their teeth in survey work in overseas territories went on to make important contributions to soil science.

Young sees the period after 1975 as one of consolidation, in which information collected by colonial surveyors, and their successors who worked for Britain's Overseas Development Administration, was used and applied, for example in the FAO's framework for land evaluation. He is also clear that the need for soil information remains pressing, but does the story he tells have any lessons for how the problem might be solved?

In some respects Young's view is conservative. It is a great mistake, he says, to see satellite imagery as an alternative to field work; and although he acknowledges a role for GIS, he states that 'Conventional air-photograph interpretation retains its primary value for more detailed work.' The problem, according to Young, is that governments are simply not prepared to make the required resources available. The digital soil mapping community might see things rather differently. After all, the very fruitful catena model of soil distribution was invented in response to a practical problem of resource (a large and complex landscape had to be mapped by a small team with little time); and so, provided that they are based on understanding of the soil, numerical models to predict soil properties from satellite imagery or other data might be similarly valuable. However, I do think that this unashamed call for soil survey in the field should provoke us to reflect. We should never offer DSM as a panacea but rather should make it clear that the authorities should adequately fund surveys which mix field work and DSM in order to predict soil conditions at the necessary scales with necessary precision. I do think that this books offers a shot across the bows of naïve expectations of what mining digital data can achieve. Are point-wise statistics of derivatives of digital elevation models (DEM) a real substitute for a skilled scientist's interpretation of a stereoscopic image of terrain? I suspect that the DEM is the future, but I hope that this will not be accepted at least until we have either found some way of emulating the delineation of land facets and land systems, or showing that a terrain index really does summarize all that and more. The pioneers of land resources survey were people with flair and vision. Without comparable commitment, and insight into the soil, we shall fare poorly at tackling the same problems of rational land use and development in the uncertain times ahead.

Murray Lark. Rothamsted



Model-based Geostatistics

By Peter J. Diggle and Paulo J. Ribeiro Jr. Springer, New York, 228p. (2007).

Geostatistics books come in many flavours. Some are very applied, make frequent use of examples and avoid equations where they can. Others are highly theoretical and situated in an abstract mathematical world that seems to have no connection with the real world. There are books that breathe the geostatistical tradition and treat geostatistics as a branch of the earth sciences. These books often seem to regard geostatistics more as an art than a science. Others consider geostatistics as part of spatial statistics and emphasise that geostatistical theory and methods can benefit much from placing it in the context of modern statistics. This book by Diggle and Ribeiro clearly belongs to the latter category, but unlike many other books in this category, it is remarkably able to show applied geostatisticians how the theory presented in the book can be put into practice. It is the first book that brings advanced spatial statistical subjects, such as generalised linear geostatistical models, likelihood-based inference and Bayesian inference of spatial stochastic models, within reach of practitioners. It is an excellent book that fills an important gap.

The book has eight chapters. Chapter 1 introduces the four example data sets that are used throughout the book (one of which is a soil data set of 178 observations on topsoil calcium and magnesium content from an experimental site in Brazil), gives an overview of the contents of the book and concludes with a section on computational aspects. Chapter 2 uses an elevation data example to outline the entire model-based geostatistics chain that is worked out in detail in subsequent chapters. It describes sampling design, model formulation, exploratory data analysis, parameter estimation and spatial prediction and simulation. Chapters 3 and 4 address model formulation, the first deals with (possibly transformed) Gaussian models, the second with generalised linear geostatistical models. Chapter 5 explains how trend and variogram parameters of the model may be estimated from the data. It treats ordinary and weighted least squares methods, but the emphasis is on maximum likelihood estimation. Spatial prediction is presented in chapter 6. This chapter not only derives the kriging equations but also uses simple examples to illustrate how kriging works. Chapter 7, the longest and most difficult, takes a Bayesian approach to geostatistical modelling and prediction. Here, the authors draw on much of their own contributions to geostatistics. The essence of the Bayesian approach is that the parameters of the stochastic model are no longer treated as fixed or deterministic, but become random variables and unknown as well. The authors nicely point out on page 214 that this is a perfectly sensible approach, because *parameter* literally means beyond measurement (hence unknown and unknowable). Prior distributions for parameters are updated to a posterior by incorporating information derived from the data. Analytical solutions to this problem are rare and Markov Chain Monte Carlo (MCMC) methods are used instead. The final chapter discusses the design problem of where to locate the sample points that define the data set. This short and somewhat isolated chapter reads like an introduction to what could become a book in itself.

Diggle and Ribeiro make clear from the outset that what they have against classical geostatistics is that it usually does not specify the stochastic model that is assumed to have generated the data. They advocate that an explicit stochastic model must always be formulated, because only then can ad hoc methods of inference be replaced by formal statistical methods. They name their approach 'model-based' geostatistics. This makes sense but when I came across this term the first time it confused me. It is also associated with the distinction between 'design-based' and 'model-based' approaches to spatial sampling and estimation (Brus and De Gruijter 1997), in which case both classical geostatistics and model-based geostatistics would be model-based.

The book discusses three main models in depth. In increasing order of complexity, these are the Gaussian linear model, the transformed Gaussian model and the generalised linear geostatistical model. The third model, which is a natural extension of the generalised linear model from linear regression, is a very rich model that can handle a variety of cases, among others skewed data, count data and presence-absence data. However, one important restriction of the model is that it assumes that conditional to the signal (mean), the responses (measurements) are mutually independent. In other words, spatial dependence is only included in the expected value of the response. This is fine when the difference between signal and

response is truly independent, such as for the examples of radioactive caesium photon emission and prevalence of malaria parasites in blood samples described in the book, but perhaps not for the majority of pedometric applications, such as the spatial distribution of soil type or earth worms over a region.

It is not only the first chapter but in fact all chapters except the last that close with a section on computation. This is one of the strengths of the book. Readers can reproduce the results presented in the book because the example data can be obtained from the book's website and the computation sections spell out the R-code used. The R statistical language and environment (http://www.r-project.org/) is free software and increasingly used in academia and research. It is perfectly suited for geostatistical analyses thanks to the many functions contained in the base package and numerous libraries, which also include functions for exploratory data analysis, graphics and GIS analyses. R is command line driven, which gives great flexibility and automatic archiving but it must be said that infrequent users struggle with the syntax each time they pick it up again. The many example command files provided in this book are therefore very helpful. The book also has a website http://www.leg.ufpr.br/mbgbook/ which contains the codes and datasets.

The book is excellently written and much care has been taken to create a high-quality product (only the index is too short to be really useful). The writing style is direct and succinct, which may explain that the book is remarkably comprehensive for a book of only 228 pages. The authors are clearly trained in explaining complex issues in an understandable way. The many examples and useful tips show that they have ample experience with modelling real-world data. The book does require a fair bit of statistical background. Readers must be familiar with matrix algebra, Gaussian processes, expectation operators and integration. Brief introductions to statistical models, classical and Bayesian inference, and prediction are provided in the appendix, but these are short and of a high level of abstraction too.

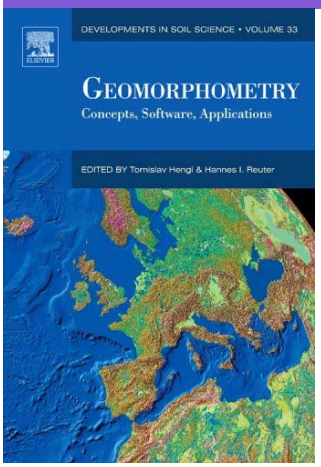
Without ignoring the basics, this excellent book goes much beyond conventional geostatistics by providing an in depth treatment of maximum likelihood inference, trans-Gaussian kriging, non-linear spatial aggregation, generalised geostatistical models and Bayesian inference. These advanced approaches often are not analytically tractable, and resort is therefore taken to numerical methods such as MCMC simulation. One thing that became clear to me while reading how these methods are applied in practice, is that it involves many experience-based decisions. Each new case requires answers to questions like: how many runs are needed, how to select an appropriate non-informative prior, how to choose the transition kernel that has a major impact on computational efficiency? Perhaps even model-based geostatistics is more an art than a science?

Reference

Brus, D.J., De Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). Geoderma 80, 1-59.

Gerard B.M. Heuvelink
Environmental Sciences Group
Wageningen University and Research Centre
P.O. Box 147, 6700 AA Wageningen, Netherlands
E-mail address: gerard.heuvelink@wur.nl

This is a modified version of a book review that was published in Geoderma 146, 489-490.



Geomorphometry. Concepts, Software, Applications

Tomislav Hengl and Hannes Reuter (Editor)

Developments in Soil Science Vol. 33. Elsevier (2008). 765pp.

Geomorphometry is the science of quantitative land surface analysis. This book captures it nicely. With increasing application and research in digital soil mapping and the availability of global DEM, this book comes just at the right time. The book, edited by celebrated pedometricians Tomi Hengl and Hannes Reuter, provides the basics and most the things you need to know about geomorphometry. It is written from a practical user's point of view, which attempts to answer most questions on how to use DEMs for analysis and prediction. The book is written by 19 authors from various fields: environmental and soil science, geography, geomorphology, hydrology, ecology, meteorology, and oceanography. This book can be seen as a revision and update of Wilson and Gallant 's Terrain Analysis (2000), which is out of print. Although each chapter is written by different authors, it is written as a book, not just a collection of papers. Each chapter is formatted in a uniform style, and the same DEM is used as examples throughout. Tomi once showed me his online reviewing system, where each author can go through each chapter and exchange views and provide feedback. The book also has a website http://www.geomorphometry.org/ which contains the DEM examples, codes, and links to software.

The book is divided into 3 parts: Concepts, Software and Applications. The concepts section begins with setting the terminology straight; no more argument on the difference between digital elevation model and digital terrain model. The history is clearly recorded, spanning from ancient Greece to James Clark Maxwell to SRTM. The following chapters include models of land surface, how to produce a DEM, how to prepare for geomorphometric analysis, uncertainty analysis, definition of land-surface parameters, parameters related to hydrology and topo-climatology, and finally landforms.

The software section covers most widely used program for geomorphometry analysis, from commercial ArcGIS to freeware and/or open source software: ILWIS, Landserf, MicroDEM, RiverTools, SAGA, and TAS, GRASS. Examples of command line used in each of the software are instructive so that readers can perform their own analysis.

Section 3 shows a range of applications from digital soil mapping, vegetation mapping, geomorphology, ecological mapping, hydrologic modelling, meteorology, and precision agriculture. The final chapter outlines the future of geomorphometry, with particular emphasis is getting a higher resolution DEM (LIDAR).

This book demonstrates the authors' expertise, knowledge, and passion which blend nicely in a user-friendly text. For pedometricians, and digital soil mappers, this is a great resource; it captures the essential background context and also demonstrates specific applications. Chapters 19 illustrates the concept and modelling of landscape with examples in soil prediction. Chapter 20 shows applications in digital soil mapping. For beginners, Tomi has formulated a recipe in a case study for prediction of soil properties using R. Tomi reminds us several times: 80% of the digital soil mapping projects have used DEMs as important covariates. To understand how it works and what do the parameters mean is an essential key to generate useful soil-landscape models.

This book is written for a general audience, and will definitely attract geographers, hydrologists, environmental modellers and other disciplines interested in using DEMs. I think because Tomi is often very critical of other books, he has tried to perfect this one. There are few snags and typos that can be ironed out. The explanation on Dokuchaev and Jenny's state factor equation is not entirely correct, and needs a bit of revision. I also have a wish list, but the book is already 765 pages. Overall it is a great piece of work, I really appreciate all the work, and the effort behind it.

Budiman Minasny
The University of Sydney
Australia

The Soil Formation Equation

Alex & Budi

We are all familiar with Jenny's factors of soil formation and the state-factor equation:

$$S = f(c, o, r, p, t, ...).$$

We also know that the first to publish a soil-forming factor equation was Dokuchaiev in 1898:

$$S = f(c, o, p).$$

We are not sure how the original equation actually looked, but according to Jenny (1961), this is the formula in "Western notation".

Here we want to show another contribution prior to Jenny by Charles Frederick Shaw, which we hypothesized led to the development of Jenny's state factor equation.

Jenny (1961) wrote "A number of later investigators, e.g., Shaw, also have published formulas, unaware of Dokuchaiev's early work."

Chas Shaw was a professor of soil technology at the University of California at Berkeley during that time. In 1927 on the First International Congress of Soil Science in Washington DC, he identified several factors of soil formation:
"Throughout the world the characteristics of any particular soil are determined by the Time during which Climate and Vegetative Cover have been exerting their combined influence upon the Parent Material."

Shaw published a paper titled "Potent Factors in Soil Formation" in Ecology April 1930. In this paper he wrote:
"Soils are formed by the modification and partial destruction of the parent material by the action of water, air, temperature changes, and organic life. Expressed as a formula, soil formation becomes:

$$S = M (C + V)^T$$

or Soil (S) is formed from the parent material (M) by the work of climatic factors (C) and Vegetation (V) through a period of time."

And later in the same paper he added another factor:
"There is another activity, that must be recognized. This is the modification of the surface by erosion and deposition (D):

$$S = M (C + V)^T + D$$ "

Shaw presented his model at the Second International Congress of Soil Science in Leningrad in 1930, his formula is in the form:

$$S = M (C+V) T + D$$

Shaw remarked that "Most of the older classifications of soils fail to recognize all of these factors, and as a result, can be applied locally only, where the included factors are dominant."

Shaw, in apparent contradiction to Jenny's later claim (see above) stated in his paper that he was aware of the earlier work by Dokuchaiev:
"The 'climatic' classification, so ably set forth by Dokuchaiev and his followers, recognizes the factors of climate (C) and organic life (V) but omits parent material (M) and erosion and deposition (D) and includes only the mature stage under the factor time (T). "

The discussion following the paper is an interesting one. Here are some of the quotes:

"Dr. Joseph said he was under the impression that Prof. Shaw did not intend his formula to be taken as an equation in a mathematical sense but only as an expression to show the factors involved in soil formation. "

"Prof. Romell suggested to Dr. Shaw in order not to hurt the feelings of the mathematicians to simply put S equal to a general function of the other symbols:

$S = f(M, C, V, T, D)$. "

"Prof. Sentius suggested that the work of man be included in the formula. Prof. Shaw pointed out that the term V included all organic life."

"Prof. Zakharov: The formula of Prof. Shaw is very interesting as a new attempt to express the connection existing between soil and soil forming factors. However **the formula does not take into account the relief, the animals and the activity of man**, but it is very important that denudation is not forgotten."

From the above discussion, it seems that following the the discussion in the conference, Shaw could have reformulated his equation into:

$$S = f(M, C, V, T, R).$$

But it seems that Shaw did not continue to develop his model.

Jenny was also present at that conference and presented his paper “Relation between soil humus and climate.” Jenny was noted to have interacted with Shaw in Berkeley (Amundson, 2004), Shaw told Jenny he (Jenny) ‘‘was going to be the new Dokuchaiev.’’ Shaw suddenly died in 1939.

So we hypothesize that Jenny’s factors of soil formation and state equation were developed following the discussion of Shaw’s paper in 1930.

We’d like to thank Dr. R.D. Hammer, US EPA, for suggesting a possible link between Shaw’s work, the 1930 conference, and Jenny’s formulation.

References

R. Amundson. History of Soil Science: Hans Jenny.

H. Jenny. Derivation of State Factor Equations of Soils and Ecosystems. Soil Sci Soc Am J 25:385-388 (1961)

C.F. Shaw. The basis of classification and key to the soils of California. Proceedings and Papers of the First International Congress of Soil Science. Vol. IV, 291-317.

C.F. Shaw. Is "PEDOLOGY" Soil Science? J. Am. Soc. Agron. (1930) 22: 235-238.

C.F. Shaw. Potent factors in soil formation. Ecology 11:239-245 (1930).

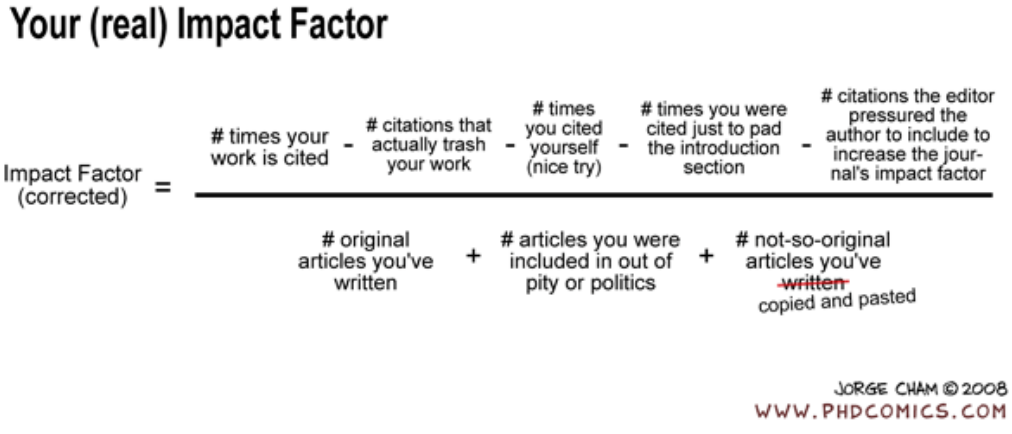
C.F. Shaw, A soil formation formula, Proceedings and papers of the second international congress of soil science, Comm. V vol. 5 (1930), pp. 7–14. http://www.usyd.edu.au/su/agric/acpa/paper/Shaw1930.pdf

Charles Fredrick Shaw (1881-1939) was born at West Henrietta, New York, May 2, 1881. After completing his college preparatory courses at Starkey Seminary, he entered Cornell University in 1902 and graduated in 1906 with the degree of B.S. in Agriculture. Charles Shaw was a student of Dr. Jay Bonsteel.

Upon graduation Shaw was appointed Scientific Assistant in the United States Bureau of Soils. In 1907 he went to the Pennsylvania State College as Instructor in Agronomy, and became Assistant Professor of Agronomy in 1909. In 1913 he came to the University of California as Professor of Soil Technology. Shaw was also in charge of the soil survey work in California. Probably no one had as complete knowledge of the soils of California as Professor Shaw. Because of his special knowledge in this field, his advice was often sought by officials of the United States Conservation Service. Shaw traveled extensively, visiting Hawaii, Australia, New Zealand, China, Russia, Germany, France, England, and Mexico. On leave from the University in 1930, he served as professor in the University of Nan-Jing, and helped to inaugurate a systematic survey of the soils of China. He wrote the book “The Soils of China”.

In 1926 Shaw, who was serving as chairman of the ASSA Committee on Terminology, proposed a terminology glossary (Shaw, 1927). Here, for the first time, definitions were compiled, including soil layer, soil horizon, and soil profile.

Source:
University of California
Professional Soil Scientists Association of California.
http://www.pssac.org/



EGU 2009

19 – 24th April 2009, Vienna, Austria.

Digital Soil Mapping

We are glad to announce the DSM session (Soil System Science Group / Session SSS12-Digital soil mapping: novel approaches to the prediction of key soil properties for modelling physical processes) in the European Geosciences Union General Assembly 2009.

The EGU DSM Session focuses on the mapping of key parameters and input variables for modelling soil processes. As we consider more complex models, applied to larger geographical regions, the demand for information on these inputs becomes harder to meet. Digital soil mapping is concerned with the provision of spatial information on soil properties on the basis of ancillary variables, such as proxy and remote sensor data, and limited direct measurements.

For more information see: http://meetingorganizer.copernicus.org/EGU2009/session/907

For abstract submission: http://meetings.copernicus.org/egu2009/abstract_management/index.html

or email: Florence.Carre@jrc.it

Complexity and nonlinearity in soils

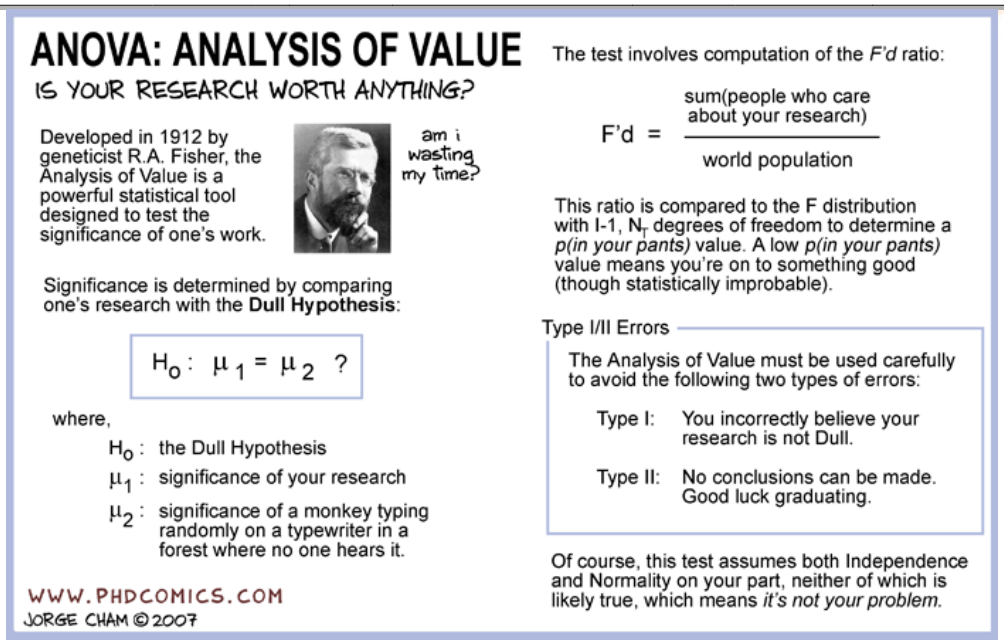
Session NP3.9/SSS39 at the European Geosciences Union in Vienna next April is a joint venture between the Nonlinear Processes in Geophysics and Soil System Sciences groups of EGU.

The objective of this session is to examine quantitative methods, including statistical approaches to scaling behaviour and mechanistic models, that shed light on the complex behaviour of soil systems over a wide range of scales of spatial organization. The non-linear processes group of EGU includes geophysicists and mathematicians, but until recently has had few if any interactions with the pedometrics community. We hope that pedometricians will consider attending this meeting to encourage a two-way flow of ideas and insights.

For more information see: http://meetingorganizer.copernicus.org/EGU2009/session/1292

To submit an abstract: http://meetingorganizer.copernicus.org/EGU2009/abstractsubmission/1292

or email anamaria.tarquis@upm.es



"Piled Higher and Deeper" by Jorge Cham
www.phdcomics.com

A Working Group on Proximal Soil Sensing (WG-PSS)

Raphael VISCARRA ROSSEL

The establishment of a working group on proximal soil sensing (PSS) was first discussed during the Pedometrics 2005 meeting in Naples, Florida, and then again at the World Congress of Soil Science in Philadelphia in 2006. However, the idea did not gain traction until the 1st Global Workshop on High Resolution Digital Soil Sensing and Mapping held at The University of Sydney in February 2008 (see Pedometron, Issue 24, May 2008 for reports on this meeting). It may be that this was exactly what it needed!

The outcome from the various discussions during and following the workshop was that a working group on PSS (WG-PSS) would definitely be valuable and timely and that a proposal for its establishment should be made to the International Union of Soil Sciences (IUSS). The proposal was prepared by the end of June and submitted to the IUSS Inter-Congress Council Meeting, 30 June–4 July 2008, in Brisbane, Queensland.

During the IUSS meeting, the council considered a number of proposals for working groups. Four were approved. The outcome of the vote for the WG-PSS was 24 ayes, 0 nays and 2 abstentions (IUSS Bulletin 113, November 2008). The council then approved the establishment of a WG-PSS and the divisional chairs were asked to consider the appointment of WG officers.

The WG-PSS is to be set up within a Commission of Division D1: Soils in Space and Time (C1.5: Pedometrics) and a Commission of Division D2: Soil Properties and Processes (C2.1 Soil Physics).

The rationales for the WG-PSS are that PSS is developing into a vibrant area of multidisciplinary research and that a working group will enable greater interaction and collaboration between scientists and engineers with a common interest in applying state-of-the-art sensing technologies to the study of soil processes and spatiotemporal soil variability.

The goal of the WG-PSS is to provide a community of practice, consisting of soil scientists and engineers, to stimulate and focus research and development of PSS worldwide by: (i) holding biannual multidisciplinary meetings or workshops, (ii) providing training workshops to build capacity in PSS, and (iii) developing a set of guidelines, norms and quality standards for PSS.

The WG-PSS would aim to address questions such as:

- Where, when and how do we sample using proximal soil sensors?
- Which techniques are most suitable for different soil properties?
- How do we fuse the data and information from multiple sensors?
- How do we analyse large (and possibly high spatial resolution) datasets?
- What information can we get from spatiotemporal measurement?
- Can we synthesise all aspects of PSS into a sound methodology?
- What are economically justified and practically feasible levels of investment in PSS technology for specific applications?
- What combinations of PSS technology are appropriate, targeted and economically feasible to meet the soil information requirements of developing countries in particular environmental circumstances (e.g. dry tropics)?
- What contributions could PSS make to astropedology, the remote study of soil or soil-like material by "Rover" systems on other planets?

The proposal was strengthened by the fact that a WG-PSS would bring together the various groups around the world working on the development of PSS techniques and that it links strongly to all the subdisciplines in soil science, but particularly to soil physics, soil chemistry, pedometrics and digital soil mapping, as well as other science and engineering disciplines (Figure 1).

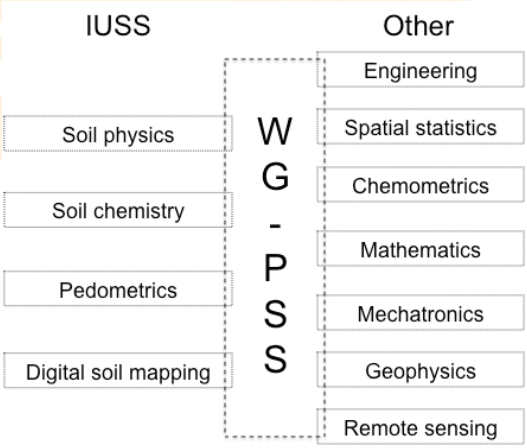


Figure 1 Linkages of the proposed Working Group on Proximal Soil Sensing (WG-PSS) to soil science and other disciplines

We are very excited about the establishment of this WG and we hope to count with your interest and support.

The Soil Spectroscopy Group and the development of a global soil spectral library

Raphael VISCARRA ROSSEL

on behalf of The Soil Spectroscopy Group

Introduction

This collaboration aims to develop a global soil spectral library and to establish a community of practice for soil spectroscopy. This will help progress soil spectroscopy from an almost purely research tool to a more widely adopted and useful technique for soil analysis, proximal soil sensing, soil monitoring and digital soil mapping.

The initiative started in April 2008 with a proposal for the project to be conducted in a number of stages to investigate the following topics:

- Global soil diversity and variation can be characterised using diffuse reflectance spectra.
- Soil spectral calibrations can be used to predict soil properties globally.
- Soil spectroscopy can be a useful tool for digital soil mapping.

Currently, the soil spectral library is being developed using legacy soil organic carbon (OC) and clay content data and vis-NIR (350–2500 nm) spectra, but in future we aim to include other soil properties and mid-IR (2500–25000 nm) spectra.

The group already has more than 40 collaborators from six continents and 20 countries (Table 1) and the library consists of 5223 spectra from 43 countries.

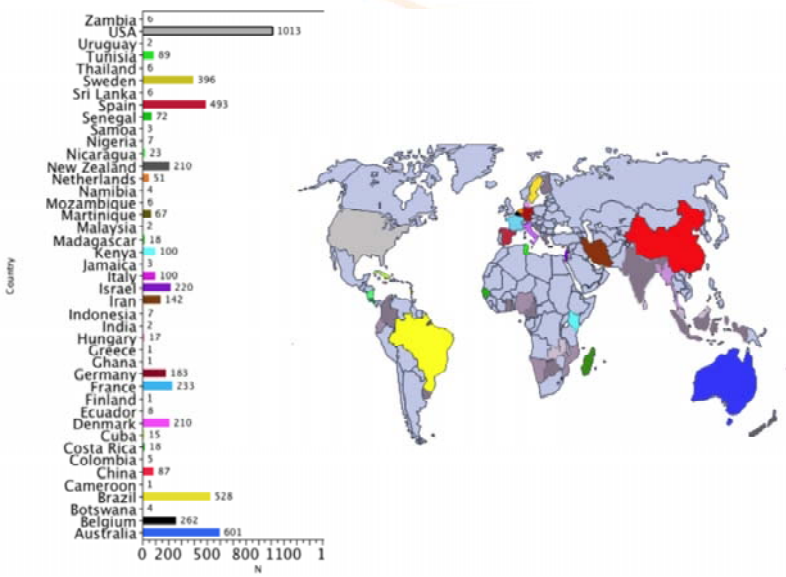


Figure 1. The number of spectra in the global spectral library from each of the 43 countries.

The library accounts for spectra from approximately only 22% of the world's countries, some of which are poorly represented with only very few spectra (Figure 1).

We would like to encourage participation from as many countries as possible, particularly, we would like contributions from counties in Central and South America, Mexico, Canada, Russia and countries in Eastern Europe, Africa and Asia. As you can see from Figure 1, we are missing a lot of countries and for some, e.g. China we have only very few data!

Do you want to join the group and contribute spectra to the global library?

The requirements for contributing spectra to the global library are as follows:

- Spectra collected in the 350–2500 nm range every 1 nm.
- At least soil OC and clay content data and also any other soil chemical, physical, biological and mineralogical data, noting which analytical techniques were used.
- Coordinates (in WGS84 format) for each sample.

Table 1. Names and countries of participants in the global soil spectral library project

Collaborators	Continent/Country
EUROPE	
Bosse Stenberg, Jan Eriksson	Sweden
Anton Thomsen, Maria Knadel	Denmark
Harm Bartholomeus	The Netherlands ISRIC samples
Antoine Stevens, Valerie Genot, Bas van Wesemael	Belgium
Youssef Fouad, Christian Walter	France (Brittany)
Cecile Gomez, Philippe Lagacherie	France (Herault)
Cesar Guerrero	Spain
Thorsten Behrens	Germany
Kristin Boetcher, Thomas Kemper	Italy
NORTH AMERICA	
David Brown, Ken Sudduth, Newell Kitchen, Brent Myers, Sabine Grunwald	USA
Martial Bernoux, Didier Brunet, Bernard Barthes	Martinique
SOUTH AMERICA	
Alexandre Demattе	Brazil
AFRICA	
Keith Shepherd, Andrew Sila	Kenya ISRIC samples
Aichi Hamouda	Tunisia
Martial Bernoux, Didier Brunet, Bernard Barthes	Madagascar Senegal
ASIA	
Zhou Shi	China
Eyal Ben-Dor	Israel
Hakim Absolou	Iran
OCEANIA	
Raphael Viscarra Rossel, Alex McBratney, Ted Griffin	Australia
Carolyn Hedley, Bambang Kusumo, Mike Hedley, Mike Tuhoy	New Zealand

- Soil classification for each sample, preferably using FAO-WRB (FAO, 1998).
- Future access to soil samples for mid-IR scanning.

If you do not have access to a spectrometer and would like to join the group, we can arrange to have the soils scanned at CSIRO in Australia or in a collaborating institution nearer to you. We have done this with a number of countries already.

Some preliminary results

The data

Currently, the vis-NIR (350–2500 nm) spectral library contains 5223 spectra. Of these, 4859 samples have soil OC data and 4345 samples have clay content data. Although we request that contributions are made with soil classification and coordinates, only a small proportion of samples possess these data. A principal components analysis (PCA) of the spectra showed that more than 95% of the variance in the spectra could be accounted for by the first four components. A scatter plot of the first three scores is shown in Figure 2a.

Except for a few outliers and the group of spectra with larger, positive scores (represented by the brown and light blue coloured samples in Figure 2a), the spectra were coherently distributed in multivariate space (Figure 2a). These different samples are those from Iran, the Herault region of southern France and some from Tunisia, which contain large amounts of carbonate. The average soil OC of the data was 1.8% with a standard deviation of 2.1% and the distribution was positively skewed with OC values ranging from 0 to

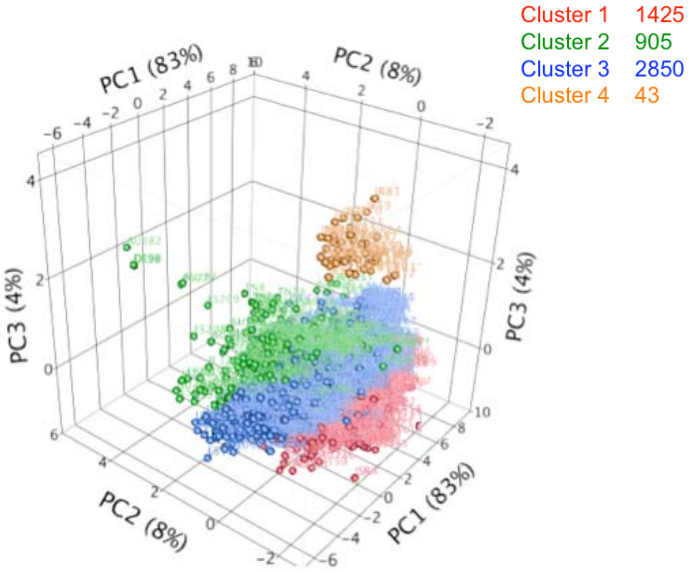


Figure 3. K-means clustering of the vis-NIR spectra using the first five principal component scores.

29%. The average clay content was 25%, the standard deviation 19.4% and clay content ranged from 0 to 92% (Figure 2b).

What can we say about the soils from their spectra?

A k-means clustering of the PCA scores showed that the spectra could be grouped into four clusters. The first cluster contained 1425 spectra (red), the second had 905 spectra (green), the third was the largest with 2850 spectra (blue) and cluster four (orange) contained only 43 samples but was the most distinct (Figure 3).

Cluster 1 soils were the more organic soils. Their average spectrum (Figure 4a red) had low reflectance values and an evenly increasing slope in the visible and shortwave NIR (400-1100 nm), which is characteristic of soil containing considerable amounts of organic materials.

The absorption features near 1400 nm, 1900 nm, 2200 nm and 2350 nm are important features for mineralogical interpretation. On average, soils in cluster 1, 3 and 4 (Figure 4a red, blue and orange) had a more prominent 1900 nm

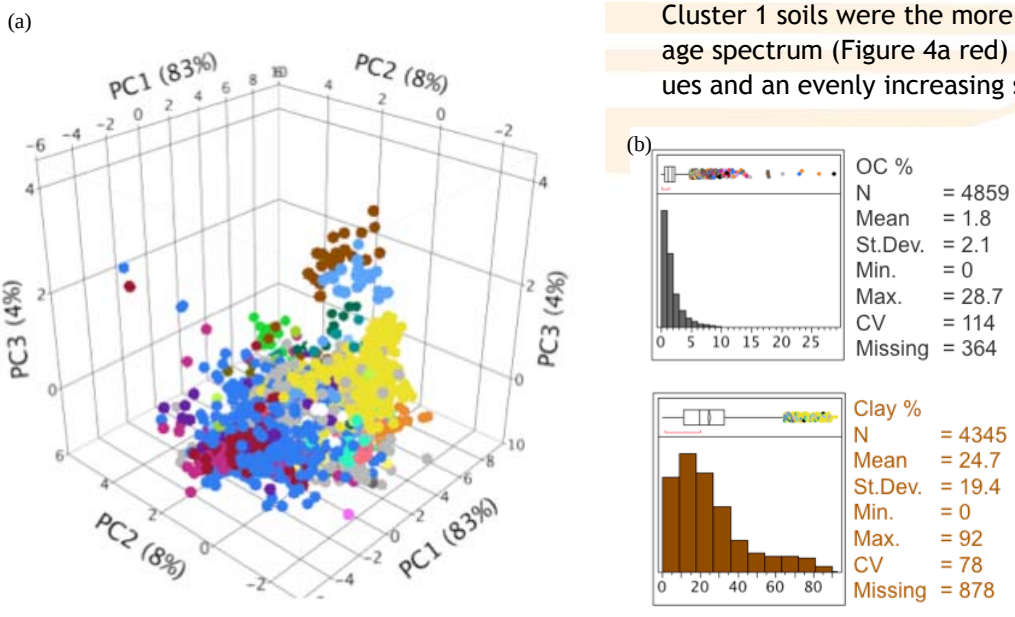


Figure 2. (a) Principal component analysis scores plot for the first three scores derived from the spectra and (b) histograms and distribution of soil OC and clay content. Colours in (a) represent different countries as per Figure 1.

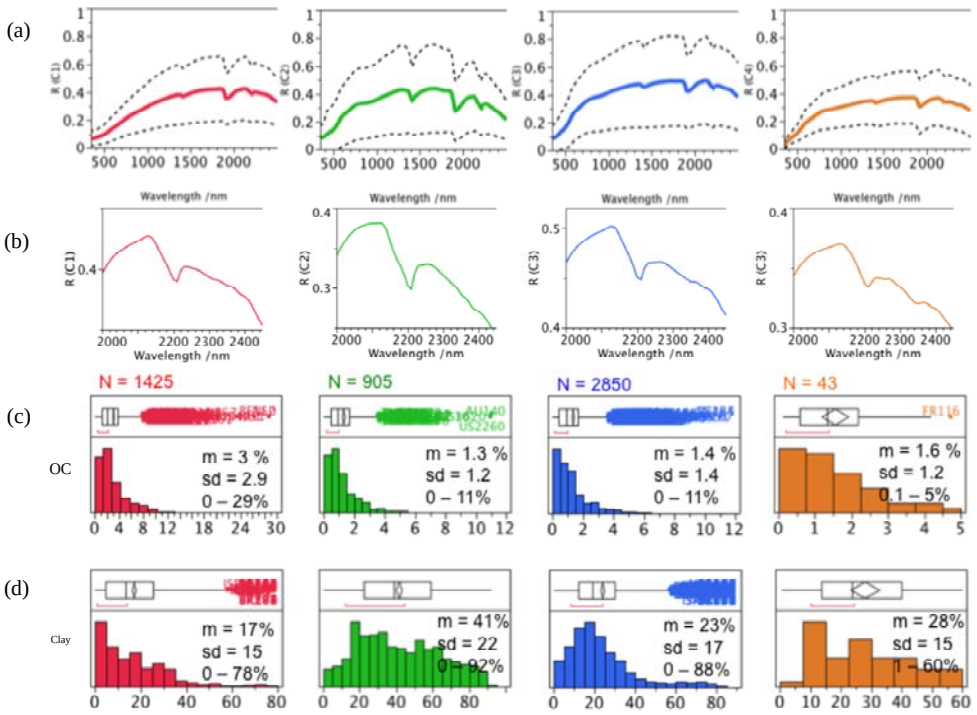


Figure 4. For each cluster: row (a) average spectra with 95% confidence intervals, row (b) magnification of 2000 nm-2500 nm wavelength region showing mineral absorption bands, row (c) distribution and descriptive statistics for soil OC and row (d) distribution and descriptive statistics for clay content.

band compared to the shallower 1400 nm and 2200 nm bands. The 1900 nm band indicates the presence of molecular water in the sample. Minerals that can absorb water in their structure, like smectite, show intense absorptions near 1900 nm, with only weaker absorptions near 1400 nm and 2200 nm. On average, soils in cluster 1 contained the largest amount of OC and the smallest amounts of clay (Figure 4c and d red).

The average spectra of clusters 2, 3 and 4 are representative of more mineral soils (Figure 4a green, blue and orange, respectively). They have broad but prominent absorption feature near 450 nm and 900 nm, which indicate the presence of the iron oxides. The characteristic absorption of haematite occurs near 890 nm and goethite near 960 nm. In cluster 2, the features near 1400 nm, 1900 nm and 2200 nm are also more prominent than those of the clusters 1, 3 and 4 (Figure 4 a and b). The metal-hydroxyl stretch combination vibration near 2200 nm is particularly important for mineralogical identification. The presence of an absorption doublet in cluster 2 soils (Figure 4b green) is indication of the presence of kaolin. This doublet is also prominent in the 1400 nm hydroxyl overtone vibration of cluster 2 soils. On average, cluster 2 soils contain more clay and less OC than soils in the other clusters (Figure 4c and d). Soils in clusters 3 and 4 are alike and contain similar amounts of OC and clay (Figure 4 blue and orange). The main difference is that cluster 4 soils have a prominent absorption feature near 2350 nm, which is due to a combination vibration of a carbonate fundamental (Figure 4b orange). As you can see, soil vis-NIR spectra can tell you a lot about the mineral-organic composition of soils, which can then be used to infer their properties!

Correspondence analysis - which continents/countries have similar spectra/soils?

The proportion of spectra, represented by each of the four clusters, currently in the global library by country and continent is given Table 2.

For example, there are 601 Australian soil spectra in the library (Figure 1), and from Table 2, 24.3% of these spectra are represented by cluster 1, the higher organic content soils with predominantly smectitic mineralogy; 41.4% by cluster 2, which are soils with high clay content, iron oxides and predominantly kaolinitic mineralogy and 34.3% by cluster 3, which are soils with generally low OC, contain iron oxides and their mineralogy is predominantly smectitic. There are 142 spectra from Iran in the library (Figure 1), and from Table 2, 12% of these soils are represented by cluster 1, 16.2%, 57% by cluster 3 and 14.8% by cluster 4, which represent soils with high carbonate content.

Similarly, the data can be looked at on a continental basis (Table 2). Africa contains the largest proportion of cluster 1 soils, Asia, Europe, South and North America contain a significantly larger proportion of cluster 3 soils, while Oceania contains even proportions of cluster 1, cluster 2 and cluster 3 soils. Only Africa, Asia and Europe contain cluster 4 spectra (Table 2). Of course, because in a lot of cases we do not have many or comprehensive spectra for each country, at this stage, the results of this analysis need to be interpreted with care.

Global multivariate calibrations for predictions of soil OC and clay content

The partial least squares (PLS) factors were clustered for each soil property separately. These factors are similar to the more commonly used principal component scores. However, unlike the PCA scores, the first few PLS factors have maximum covariance with the response variable — in this case soil OC and clay con-

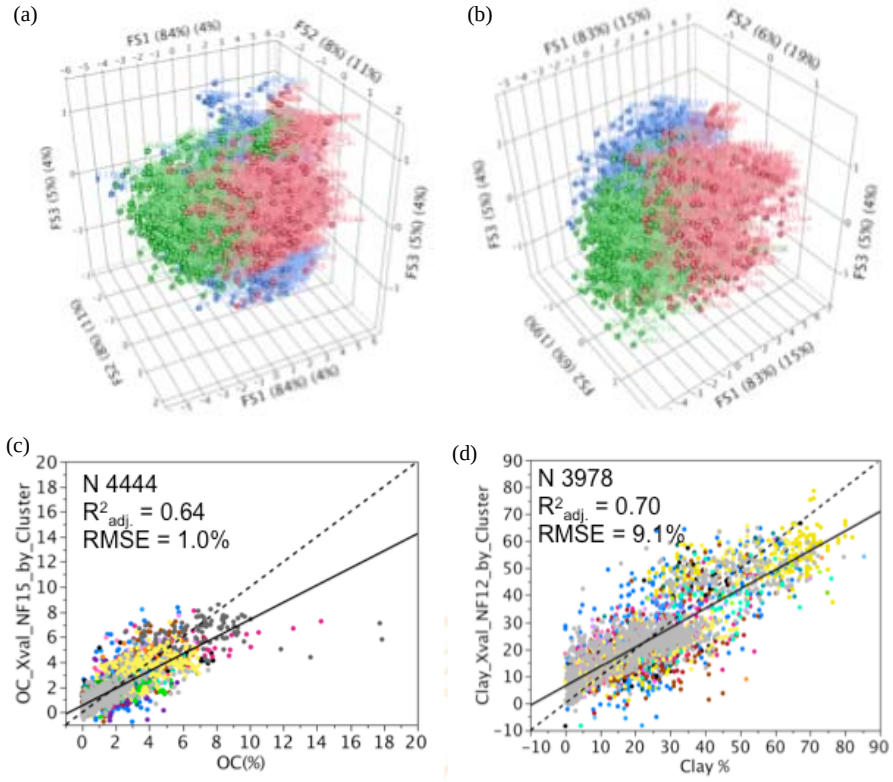


Figure 5. (a) and (b) PLS factor scores for soil OC and clay content, respectively, and (c) and (d) leave-10-out cross-validation predictions of soil OC and clay content, respectively.

Table 2. Spectral cluster by country and continent. The average spectra and soil OC and clay content for each cluster is given in Figure 4. The number of spectra from each country is given in Figure 1.

Cluster	1	2	3	4
Africa	42.2	30.5	27	0.3
Asia	6.9	22.3	66.3	4.5
Europe	28	10.1	60.8	1.1
N. America	31.7	16.2	52.2	
Oceania	33.8	30.7	35.5	
S. America	14.8	13.5	71.7	

Cluster	1	2	3	4
Australia	24.3	41.4	34.3	
Belgium	16.4		83.6	
Botswana	50		50	
Brazil	14.0	13.5	72.5	
Cameroon			100	
China	4.6		95.4	
Colombia		20	80	
Costa Rica	5.6	50	44.4	
Cuba	13.3	73.3	13.3	
Denmark	59.1	0.5	40.5	
Ecuador	75.0	12.5	12.5	
Finland			100	
France	27.0	11.6	52.4	
Germany	7.7	6.6	85.8	
Ghana		100		
Greece			100	
Hungary	29.4		70.6	
India		100		
Indonesia		85.7	14.3	
Iran	12	16.2	57	14.8
Israel	5	32.7	62.3	
Italy	22	30	48	
Jamaica		66.7	33.3	
Kenya	63	24	13	
Madagascar	22.2	22.2	55.6	
Malaysia	50	50		
Martinique	6	89.6	4.5	
Mozambique	16.7	16.7	66.7	
Namibia	25		75	
Netherlands	100			
New Zealand	61.4		38.6	
Nicaragua	8.7	78.3	13.0	
Nigeria	42.9	14.3	42.9	
Samoa		33.3	66.7	
Senegal	66.7	4.2	29.2	
Spain	20.3	22.1	57.6	
Sri Lanka		50	50	
Sweden	31.3	4.3	64.4	
Thailand		16.7	83.3	
Tunisia	4.5	67.4	27.0	1.1
Uruguay			100	
USA	34.8	8.3	57	
Zambia	66.7		33.3	

tent. The k-means clustered PLS factors for soil OC and clay contents are shown in Figure 5a and b.

PLSR models were developed for each cluster and leave-10-out cross-validation was used to assess the models. The predictions for each cluster were then combined and the results are shown in Figure 5c and d. Fifteen PLS factors produced the best predictions of soil OC content (R2 0.64 and RMSE 1%) (Figure 5c). For clay content, 12 factors produced the best cross-validation predictions (R2 0.70 and RMSE 9%) (Figure 5d).

The calibrations of the spectra to soil OC and clay content are encouraging (Figure 5), particularly when one considers that these are global calibrations and that the spectral library includes measurements of OC and clay that were made using different laboratory techniques, in different laboratories and by different people and the instruments used to collect the spectra were also different, although from the same manufacturer.

I would like to encourage anyone interested in this initiative to join us. In particular we need spectra from countries in South America, Central America, Asia, Africa, Eastern Europe the Pacific Islands.

To make this work we need participation from as many people around the world as possible.

If you are interested in contributing spectra to the global library please send me an email (raphael.viscarra-rossel@csiro.au) and join the group!

References

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

Pedometrician profile

Ron Corstanje

Cranfield University

How did you first become interested in soil science?

As a wastewater engineer I increasingly became interesting in understanding what was happening 'after the pipe' and typically this involved sediments. So my introduction to soil science became an immersion into the nutrient dynamics of the peat soils of the Everglades.

How were you introduced to pedometrics?

The Everglades is a vast and open expanse with patchy vegetation patterns and we were taking three cores at a small number of sites and using this information to characterize a vast area. I was introduced to pedometrics when the Dr Sabine Grunwald joined the Soil and Water Dept. at the University of Florida and started to work on the problem of spatial prediction in the Everglades.

What recent paper in pedometrics has caught your attention and why?

It is not as much recent as that it suddenly has become relevant to predicting ecosystem functions: Brown, Daniel, P. Goovaerts, A. Burnicki, and M.Y. Li. "Stochastic Simulation of Land-Cover Change Using Geostatistics and Generalized Additive Models." Phtogrammetric Engineering and Remote Sensing, 68 (10): 1051-1061. 2002.

What problem in pedometrics are you thinking about at the moment?

Scale dependencies in covariate data used for DSM predictions.

What big problem would you like pedometricians to tackle over the next 10 years?

A systems approach to soil variation that tackles the interaction between the soil biology, chemistry and physics. In what way can pedometrics contribute to our understanding of how the soil functions?

Ron Corstanje studied environmental engineering at Wageningen before moving to the University of Florida where he completed his Ph.D. in biogeochemistry under the supervision of Professor Ramesh Reddy. After postdoctoral work in Florida he joined the Environmetrics group at Rothamsted where he modelled ammonia volatilization from soil, studying variation and model performance at nested scales from pedon to catchment. He is now Senior Research Fellow in Pedometrics at the National Soil Resources Institute, Cranfield University in the UK.

Non-Pedometrician profile

Michael Vepraskas

North Carolina State University

How did you first become interested in soil science?

I was majoring in geology as an undergraduate at the University of Wisconsin in Madison in the 1970's, and found that the soils topics in a geomorphology class were the best part of the course. When I took an introductory soils class to learn more, I became even more impressed with how much I learned about earth processes from soil scientists. For example, the soils professor explained the structure of clay minerals much better than what I had just been taught in a mineralogy class in the geology department.

In addition, geology was a bit disappointing because it seemed that geologists were basically just guessing about what happened in the past. The geology professors were honest about this, but as a student it was still frustrating to never be sure about what was true in the information they were teaching us. On the other hand, when the geology professors talked about soils, they made it sound as if the soils professors at Wisconsin were the people who really knew what was happening below the surface. In a hydrogeology class, for example, the professor told us that the mathematics behind unsaturated flow were too complicated for him, and that we needed to go to the soil scientists for the answers. He even held up a soils map during one lecture and said that he wished geologic maps could be made as detailed as the soils map. After hearing all this for four years of undergraduate study in geology, I decided to do my graduate work in soil science in order to work with the people who weren't just guessing.

What are the most pressing questions at the moment in your area of soil science?

I study wetland soils. In the U.S., wetlands are protected by state and federal laws that prevent wetlands from being filled in or drained, unless they are replaced somewhere else. My research deals with finding better ways to identify wetlands and their boundaries, as well as finding environmentally sound ways to restore or create wetlands.

The most pressing problems related to pedometrics deal with wetland hydrology. In order for an area to be a wetland it must be saturated to within 30 cm of the surface, or inundated, for a period of 5% or more of the growing season, and this has to occur in at least 5 or more years out of every 10. We have to find better ways to determine where these requirements are met in landscapes if we want to identify wetlands accurately. Long-term monitoring is too expensive to be done at all sites. Hydrologic models offer the most promise, but the modeling can't be done at all locations where the results are needed. We need to develop simpler hydrologic models that compute water table fluctuations over time. We also need to find ways to extrapolate modeling results to other sites, in order for the modeling results to be useful. Our current work is using long-term (40-yr) modeling results to identify soil color patterns that develop where the minimum frequency and duration requirements for wetland hydrology occur. We believe that such color patterns could be used for more accurate assessments of where wetlands occur.

Hydrologic models could also be used in wetland design by predicting the hydrology of a site after it has been restored to a wetland.

Continues next page ...

Non-Pedometrician profile

...continued

Michael Vepraskas

At this time, restoring a drained land area back to a wetland is relatively easy from an engineering standpoint, if all that is needed is to plug ditches and plant trees. However, the hydrology of many restored sites is either too wet for the planted vegetation to survive, or it is too dry for the site to become a wetland. Modeling methods must be developed that help wetland designers determine the amount of site alteration needed to achieve a desired hydrology. This information can then be used to select appropriate vegetation to plant.

What statistical and mathematical methods are used in your area of soil science?

For our wetland studies we still use analysis of variance techniques to compare treatments. At this time, the other mathematical methods include hydrologic models. The model we use primarily is DRAINMOD, because it was developed at my university for the landscapes in the southeastern U.S. This model is relatively straightforward, and provides the same information as can be obtained from a well. This makes the simulation data easy to interpret.

Are you aware of any work by pedometricians that might be relevant to your science?

Pedometrics is outside my area of expertise, so I've asked two coworkers of mine at NC State, Josh Heitman, and Jeff White, to offer suggestions for this question and the next. We feel that predictions about the change in the spatial extent of wetlands under varying water table and climate conditions can be accomplished through incorporation of soil-landscape modeling using digital terrain analysis, soil pattern analysis, and analysis and modeling of spatial and temporal variation of soil properties, all of which fall within the realm of current work in Pedometrics.

What big problem would you like pedometricians to tackle over the next 10 years?

We believe that evaluating soil properties will remain key for delineating wetlands. Over the next 40 years, population growth will continue to place pressure on natural ecosystems, and climate will change. It will be important for pedometricians to develop tools that predict how wetland boundaries might spatially and temporally change in response to population and climate pressures. This can be done by combining results of hydrologic modeling with soil survey maps for example. Such results can be great aids to show land-use planners where climate change may impact soils in residential areas. To do this reliably, we need to develop more ways to extrapolate the modeling results from one soil landscape to another. That is something pedometricians are well positioned to do.

Another area that pedometricians could contribute, is to develop ways to educate more soil scientists on the value and uses of various computer-based models. Engineers are way ahead of us on this. Engineers use simulation models for much of their work, especially in design, and they trust the results they get from models. Modeling is still new to many soil scientists, and in my experience many still do not trust the results produced by models. This is especially true when modeling results are used to establish rules and regulations related to soil use and management. Pedometricians would seem to be well positioned to train students at the university level in the value of models, such as those used to evaluate hydrology. Models need to be developed that can be brought to the classroom and used effectively in teaching soil physics, fertility, soil classification, mapping, and so on.

Pedomathemagica

with Dick Webster

1. The Pedometrics Prize

The panel in charge of the annual Pedometrics prize recognizes that democracy is all very well; but, given that probability is such a prominent part of all that we do, the panel thinks that it should introduce some element of probability into the award itself. It therefore proposes to introduce a variant of a game that has caused and still causes controversy among professional mathematicians. Here it is.

The gilt-framed certificate is placed in one of three cardboard boxes by the presiding judge. The author who tops the popular vote is summoned by the judge to nominate the box that he or she believes to hold the certificate. The judge, knowing of course in which box the certificate was placed, opens one of the others revealing it to be empty. The judge then asks the author a second time which box holds the certificate. If the author nominates the box correctly the prize is his or hers; if not then he or she loses it.

Imagine now that you have topped the popular vote and the judge poses the question to you; how should you respond to maximize your chance of gaining the certificate and the prize? Should you stick to your initial choice of box? Or should you change it to the other unopened box?

2. Danger: Road Works!

Some engineers want to build a road over soft alluvial soil near a river. They test the soil and conclude from their tests that the soil would be strong enough to support the road provided that it never becomes saturated as a result of flooding. So they ask the river authority if it could tell them how likely a flood is, and they receive the answer: 'Every year in spring there is a surge in the river when the snow melts, and about once in a century it is so large as to cause flooding'. 'That is OK', say the engineers, 'because we design the road to last for only 50 years, and so there is no risk of the road's failing in that time'.

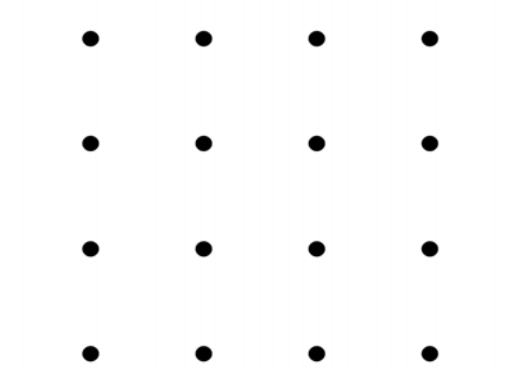
Clearly that is faulty reasoning. But if annual flooding is an independent random variable with expectation once per 100 years what is the risk that the road will fail during the 50 years of its planned lifetime?

Pedomathemagica

with Gerard Heuvelink

Problem 1 (EASY - MEDIUM)

Pedometricians familiar with transect sampling may find this one easy. Can you connect all 16 points with only six straight lines, without taking the pencil from the paper?

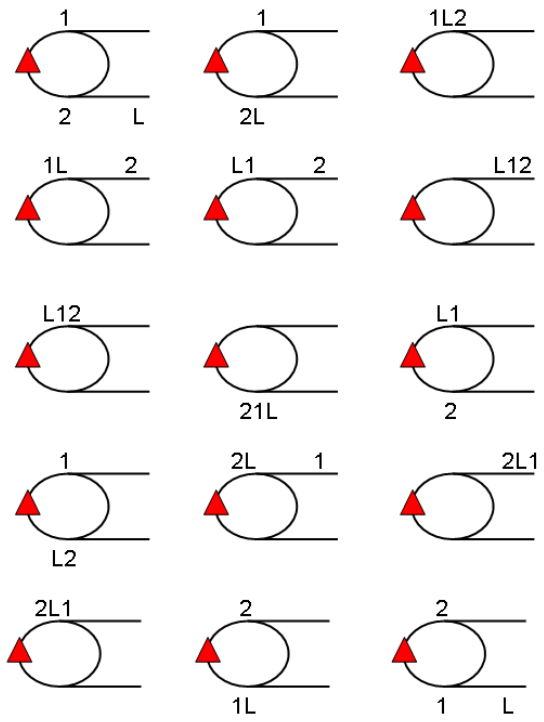


Problem 2 (MEDIUM)

A large group of pedometricians is enjoying an excursion. This time they do not travel by train (see Pedomathemagica Pedometron 24) but travel on foot. They form a long queue of 1 km that moves with a constant speed. There are excursion leaders at the front and back of the queue. The excursion leader at the back wants to send a message to the leader at the front and sends a courier who runs to the front with constant speed, passes on the message, turns and runs back immediately with the same speed. When he arrives at the back, it turns out that the queue has moved exactly one kilometre. How many kilometres did the courier run?

Answers to Last Issue's Problems

Problem 1 (EASY?)



Problem 2 (MEDIUM-HARD)

Pierre has the biggest chance of becoming first author, in spite of the fact that he has the worst shot. This is because each individual maximizes their chances. This means that Achim and Rosina will always aim for each other until one of the two is out. Pierre's optimal strategy (and this is the hard part) is to shoot in the air as long as Achim and Rosina are still both in the game.

The probability for Rosina to survive Achim is 0.5 (when she gets to fire before Achim) + 0.5 × 0.2 (when Achim gets to fire first and misses) = 0.6. If Rosina survives then Pierre gets the chance to fire at her. He has a 50% chance of success. If he fails, then Rosina will shoot him with certainty. This means that Rosina has a 0.6 × 0.5 = 3/10 probability of becoming first author.

If Achim survives over Rosina then there is also probability 0.5 that Pierre will miss so that Achim gets a chance at Pierre, but since he has only 0.8 probability of success, Pierre might get another go (and another, and another, and another,....). Achim's probability of becoming first author is therefore 0.4 × 0.5 × 0.8 + 0.4 × 0.5 × 0.2 × 0.5 × 0.8 + 0.4 × 0.5 × 0.2 × 0.5 × 0.2 × 0.5 × 0.8 + ...= 8/45.

Pierre has probability 1 - 3/10 - 8/45 = 47/90, the biggest of all three.