ΠΕΔΟΜΕΤΡΟΝ

The Newsletter of the Pedometrics Commission of the IUSS

Issue 26, March 2009

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

From the Chair

Dear Colleagues

Welcome to another issue of Pedometron. As usual we have a diverse contents for you to peruse. Although I am writing this on the first working day in March, this is the last Pedometron before the Pedometrics 2009 conference in Beijing at the end of August. April 30th is the deadline for abstracts; and I hope that all readers of Pedometron will consider submitting an item and participating in the first Pedometrics meeting on the continent of Asia.

In late 2001 Max Perutz, who won a Nobel Prize for working out the structure of the haemoglobin molecule, discovered that he had terminal cancer. One paper that he rushed to finish was on the molecular structure of the amyloid fibres that cause various brain diseases. According to his biographer, Georgina Ferry, his colleague Aaron Klug felt that he was drawing excessively strong conclusions, but Perutz stuck to his guns. The paper was published. After his death he was shown to be wrong, but the study which did so, inspired by his idea, contributed importantly to the understanding of the problem. Ferry writes that 'Klug now concedes the truth of philosopher A.N. Whitehead's dictum "It is more important that an idea is fruitful than that it is correct"' (Ferry, 2007). There is an irony in this. Perutz's wrong but fruitful idea got the airing that it deserved because of his eminence. Ferry's book shows that Perutz did not generally approach ideas he disapproved of, or criticism of his ideas, with Whiteheadian phlegmatism. This did not always help progress.

Debates in statistics generally fall into two camps. There are mathematical arguments where the right answer emerges sooner or later, and is only denied by the cranks. One such argument was that between Fisher and Pearson over the degrees of freedom for the contingency table. Fisher was right and Pearson was wrong, and for those who struggled with the theory, Fisher's simulation proved the point. The second kind of argument is philosophical. Statistics is concerned with how we make inferences about truth from data. Bayesians and Frequentists differ on fundamental questions such as what kinds of uncertainty can be validly described by probability distributions, or whether it is valid to express ignorance of some parameter by a uniform ('uniformative') prior distribution. These are not mathematical arguments, but rather arguments about how we use mathematics. Debate in good faith should clarify the assumptions of both sides, but it is unlikely to go away anytime soon.

Pedometrics has had its share of debates, and this Pedometron contains the first article in what I hope will be a series which formalizes them on paper. On page 4, Andreas Papritz argues that Indicator Kriging should be abandoned. We are expecting a rejoinder for the next issue, but anyone else who would like to contribute is invited to do so by sending a (short and focussed) argument to the me (murray.lark@bbsrc.ac.uk).

Let me make some observations about how this debate might proceed fruitfully.

Inside This Issue

From the Chair	1
The Richard Webster Medal	3
Why indicator kriging should be abandoned	4
Peter Alan Burrough	8
A new digital soil map of the world	10
The statistical aspects of national scale soil monitoring workshop	12
Report from the trenches	14
Alex's preferred papers	17
Did you miss…	19
Spatial coverage sampling	20
Mapping research hotspots	24
Profiles	27
Pedomathemagica	29

Many readers of Pedometron will have a personal investment in the debate. Perhaps you have published several papers that use IK, perhaps you have persuaded your local policy makers to fund a project in which you will use it. Now Andreas says 'abandon it'. What to do? I would suggest that, first time through, you read his article to understand it, not to find holes in the argument. Make sure that you really know what he is saying and why. Second, accept the technical points that you think he has established. Third, consider what you think the implications of these arguments really are for practice. That is not an invitation for the all-to-common argument along the lines that these statistical niceties have no bearing on the hard work in the real world done in my department, but you might ask whether the statistical case is a sufficient reason to abandon the technique, rather than to refine it.

I shall start the process by accepting, straight off, Andreas's criticism of a comment in a paper that I wrote comparing IK and DK (Lark & Ferguson 2004 in his reference list). There I suggested that, since DK explicitly uses simple kriging of the Hermite polynomials, second-order stationarity is required, but this is not the case for IK. I should have thought deeper about the indicator variable and recognized Andreas's point that if it shows drift in the mean then stationarity in the variance also fails since the mean and variance of an indicator are not independent.

I hope that this debate will be fruitful. Please contact Budi or me if there is a topic on which you would like to write a polemic for future issues.

I would like to finish with a personal word about Peter Burrough, who's obituary appears in this issue. I met Peter on several occasions, including his visits to Oxford during my student days. I was always impressed by the range of his ideas, his generosity with them, and his enthusiasm. He was also a man with wide interests. When we last spoke it was on a terrace in Montpellier during the first workshop on Digital Soil Mapping, and we discussed hypotheses to explain human experiences of 'the numinous.' Peter was an important communicator of that strand of pedometrics that emerged in Oxford in the late 1960s, most particularly through his many students who have been influential pedometricians. He will be greatly missed.

With best regards

Murray

Ferry, G. 2007. Max Perutz and the Secret of Life. Pimlico, London. The quotation is from page 248.

Best Paper in Pedometrics 2008

The deadline for nominations, announced in the last issue, expired with just two nominations received. These have been passed to a senior and experienced soil scientist. A final list will be circulated as a Pedometron Special Flier before the end of April, with details on voting. Watch this space.

The Richard Webster Medal:

an award by the Pedometrics Commission of the International Union of Soil Sciences

The Richard Webster Medal: an award by the Pedometrics Commission of the International Union of Soil Sciences

The Richard Webster medal was established before the last World Congress of the International Union of Soil Sciences (IUSS). The award is for the best body of work that has advanced pedometrics (the subject) in the period between the IUSS World Congress of 2006 and the next one in 2010. However, achievements before that period will also form part of the evaluation (see more detail below). The award will be made at the next meeting of the IUSS World Congress. The first award was made to Professor Alex McBratney (University of Sydney) at the World Congress in Philadelphia (USA).

Guidelines for the award of the Richard Webster Medal

The official rules are also at http://www.iuss.org/popup/Webster_medal.htm

Requirements and eligibility for the award of the Richard Webster Medal

1. Soil scientists eligible for the award will have shown:

    a)a distinction in the application of mathematics or statistics in soil science through their published works,
    b)innovative research in the field of pedometrics,
    c)leadership qualities in pedometrics research, for example, by leading a strong research team,
    d)contributions to various aspects of education in pedometrics (e.g. supervision of doctoral students, teaching of pedometrics courses in higher education, the development of courses for broader professional needs),
    e)and service to pedometrics (e.g. by serving on a committee of the Pedometrics Commission or promoting pedometrics to the IUSS).

2) A nominee should be a member of the IUSS at the time of the nomination and have been involved in activities associated with pedometrics, in particular.

3) The nominee must be living at the time of the selection; retired pedometricians still active in pedometrics research will be eligible for the award. The nominee should be willing to receive the medal at the time and place designated by the IUSS World Congress, and be a keynote speaker at the next conference of the Pedometrics Commission (held biannually) following the presentation of the medal.

4) The Pedometrics Commission will pay for the recipient's travel expenses to attend the Pedometrics meeting where the keynote address will be given.

5) Members of the Awards and Prizes Committee shall be ineligible to receive the medal while serving on the Committee.

6) The award of the Richard Webster Medal shall not be presented to any one individual more than once.

Nominations procedure

1) Nominations for the Richard Webster Medal should be made by a colleague or colleagues who know the person's work well. The nomination should include a résumé and a short statement (a maximum of 750 words) summarizing the relevant qualifications of the nominee with respect to the conditions outlined in the section, requirements and eligibility, above.

2) The proposer(s) should submit the following on behalf of their nominee two months before the next IUSS conference (August 2010), i.e. before the 1st of June 2010:

    a)their published work for the four-year period between consecutive IUSS meetings,
    b)a suitable curriculum vitae that gives:
    - all previous publications,
    - positions held,
    - research undertaken,
    - education of others,
    - teaching courses developed,
    - and leadership and management of research projects .

This material should be sent to the Pedometrics Awards Committee chair, Professor Margaret Oliver at m.a.oliver@reading.ac.uk"

Why indicator kriging should be abandoned

Andreas Papritz

Institute of Terrestrial Ecosystems,
Department of Environmental Sciences,
ETH Zurich, 8092 Zurich, Switzerland

According to ISI Web of Science®, about 210 journal articles and 110 contributions to conference proceedings have been published about *indicator kriging* (IK for short) to date. Figure 1 shows that the number of publications has increased steadily since A. Journel published his paper on IK in 1983. It is likely that the proponents of IK would see this increase as proof of the merits of the method. For me, it is a rather alarming example of how an apparent lack of understanding by many scientists can support a methodology that lacks theoretically consistent foundations.

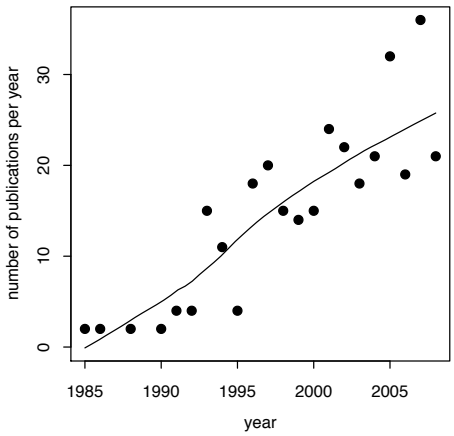


Figure 1: Number of publications per year about indicator kriging (data taken from ISI Web of Science®).

To substantiate my contention, I highlight and discuss some limitations of IK that arise mostly from basic probability theory. Notwithstanding their elementary nature, these limitations are frequently ignored. To address these I consider the spatial prediction of attributes with real values, and I focus on theory. For the time being, I leave until later the question of whether the use of IK can be "sanctioned by practice" as I assume that this might be a major argument in expected replies by advocates of IK.

E-mail address: papritz[at]env.ethz.ch

Notation

Let $Z(\boldsymbol{s})$ denote a real valued random variable used to model the value, $z(\boldsymbol{s})$, of an attribute measured at location $\boldsymbol{s} \in \mathbb{R}^d, d = 1, 2, 3$, and let $I(\boldsymbol{s}; z')$ for a specific cut-off z' be the indicator transform $I(\boldsymbol{s}; z') = 1$ if $Z(\boldsymbol{s}) \leq z'$ and $I(\boldsymbol{s}; z') = 0$ otherwise. The indicator transform of a measurement is denoted by $i(\boldsymbol{s}; z')$. From probability theory about Bernoulli random variables we have

$$\begin{aligned} \mathrm{E}\,[I(\boldsymbol{s}; z')] &= \mathrm{Prob}[Z(\boldsymbol{s}) \leq z'] = F(\boldsymbol{s}; z'), \quad (1) \\ \mathrm{Var}\,[I(\boldsymbol{s}; z')] &= F(\boldsymbol{s}; z') \cdot (1 - F(\boldsymbol{s}; z')), \quad (2) \end{aligned}$$

where $F(\boldsymbol{s}; z)$ is the cumulative distribution function (cdf) of $Z(\boldsymbol{s})$, $\mathrm{E}\,[.]$ and $\mathrm{Var}\,[.]$ are the expectation and variance operators, and $\mathrm{Prob}[A]$ denotes the probability of event A. Further, let $\mathrm{Cov}\,[.]$ and $\mathrm{Cor}\,[.]$ denote the covariance and correlation operators. The (cross-)covariance function of the indicators for two cut-offs z' and z'',

$$C_I(\boldsymbol{s}, \boldsymbol{s} + \boldsymbol{h}; z', z'') = \mathrm{Cov}\,[I(\boldsymbol{s}; z'), I(\boldsymbol{s} + \boldsymbol{h}; z'')],$$

is related to the bivariate cumulative distribution function,

$$F(\boldsymbol{s}, \boldsymbol{s} + \boldsymbol{h}; z', z'') = \mathrm{Prob}[Z(\boldsymbol{s}) \leq z', Z(\boldsymbol{s} + \boldsymbol{h}) \leq z''],$$

of the two random variables $Z(\boldsymbol{s})$ and $Z(\boldsymbol{s} + \boldsymbol{h})$ by

$$\begin{aligned} &C_I(\boldsymbol{s}, \boldsymbol{s} + \boldsymbol{h}; z', z'') = \\ &\quad F(\boldsymbol{s}, \boldsymbol{s} + \boldsymbol{h}; z', z'') - F(\boldsymbol{s}; z') \cdot F(\boldsymbol{s} + \boldsymbol{h}; z''). \quad (3) \end{aligned}$$

See, for examples, Journel and Posa (1990).

Two limitations of IK

Now, I am ready to discuss two major limitations of IK that restrict its use seriously in practice. By "in practice" I mean the case where we consider our measurements as a sample from a *single* realisation of a random process $\{Z(\boldsymbol{s})\}$:

(A) Indicator kriging requires, in practice, a random process model with stationary bivariate distributions and thus precludes the modelling of data that show a spatial trend or unbounded variation. The method proposed by Goovaerts and Journel (1995) to merge nominal (e.g. a soil map) or real valued auxiliary information (e.g. terrain attributes derived from a digital elevation model) with "hard" indicator data is an *ad-hoc* procedure that lacks optimality.

(B) Even in the stationary case, it is difficult to infer in practice a consistent model for the stationary bivariate distributions from data. Many applications of IK rely, therefore, on inconsistent probabilistic models.

IK and non-stationary models

Let us first consider limitation (A). For a random process with stationary bivariate distributions equations (1)–(3) simplify to

$$\begin{aligned} \mathrm{E}\left[I(\boldsymbol{s};z')\right] &= F(z'), && (4)\\ \mathrm{Var}\left[I(\boldsymbol{s};z')\right] &= F(z')\cdot(1-F(z')), && (5)\\ C_I(\boldsymbol{h};z',z'') &= F(\boldsymbol{h};z',z'')-F(z')\cdot F(z''). && (6) \end{aligned}$$

Clearly, the right-hand sides of these equations do not depend on $\boldsymbol{s}$, so that the (cross-)covariance of the indicators is a function of the lag $\boldsymbol{h}$ only. Notice that equation (4) means that the expectations of the random variables, say $\mathrm{E}\left[Z(\boldsymbol{s})\right]=\mu(\boldsymbol{s})$, may not vary in space because if they did the cdf would not be constant. Furthermore, equations (4)–(6) show that we may (at least hope to) infer the first two moments of the indicators when we have data from only one realisation of $\{Z(\boldsymbol{s})\}$. To estimate the expectations and (cross-)covariances of the indicators we replace the averaging of multiple realisations by averaging over space. Spatial averaging, however, is inappropriate in the general case of non-stationary distributions, i.e. for models with moments given by equations (1)–(3).

In spite of the above, Goovaerts and Journel (1995) proposed to extend the IK methodology to random processes with spatially varying $\mu(\boldsymbol{s})$. They called their method "simple IK with varying local means". The terms "simple IK with local prior means", "soft IK" or "IK with trend" have since been used to denote the approach also. Apparently, the authors realized that the indicators have non-stationary (co-)variances if $\mu(\boldsymbol{s})$ varies spatially. Given an estimate, $\widehat{F}(\boldsymbol{s};z')$, of the cdf, they proposed to estimate the variogram of $I(\boldsymbol{s};z')$ by fitting model functions to the sample variogram

$$\widehat{\gamma}_R(\boldsymbol{s}_i;\boldsymbol{h}_k;z',z') = \frac{1}{2\,N(\boldsymbol{h}_k)}\sum_{i=1}^{N(\boldsymbol{h}_k)}\left\{r(\boldsymbol{s}_i;z')-r(\boldsymbol{s}_i+\boldsymbol{h}_k;z')\right\}^2, \quad (7)$$

of the indicator residuals

$$r(\boldsymbol{s};z') = i(\boldsymbol{s};z') - \widehat{F}(\boldsymbol{s};z')$$

In equation (7) $N(\boldsymbol{h}_k)$ is the number of data pairs in lag-class $\boldsymbol{h}_k$).

Unfortunately, they failed to recognize that half the expected squared difference of the indicator residuals, i.e. their semivariance, is *not* independent of $\boldsymbol{s}$ (Papritz *et al.*, 2005), even if (unrealistically) the true cdf (i.e. if $\widehat{F}(\boldsymbol{s};z')=F(\boldsymbol{s};z')$) is assumed to be known:

$$\begin{aligned} \frac{1}{2}\mathrm{E}\left[\{R(\boldsymbol{s};z')-R(\boldsymbol{s}+\boldsymbol{h};z')\}^2\right] = \\ \frac{1}{2}\mathrm{Var}\left[R(\boldsymbol{s};z')-R(\boldsymbol{s}+\boldsymbol{h};z')\right] = \frac{1}{2}\{ \\ F(\boldsymbol{s};z')\cdot(1-F(\boldsymbol{s};z'))+ \\ F(\boldsymbol{s}+\boldsymbol{h};z')\cdot(1-F(\boldsymbol{s}+\boldsymbol{h};z')) \\ \}-\{ \\ F(\boldsymbol{s},\boldsymbol{s}+\boldsymbol{h};z',z')- \\ F(\boldsymbol{s};z')\cdot F(\boldsymbol{s}+\boldsymbol{h};z') \\ \}. \end{aligned} \quad (8)$$

As above, $F(\boldsymbol{s};z')$ and $F(\boldsymbol{s},\boldsymbol{s}+\boldsymbol{h};z',z')$ are functions of $\boldsymbol{s}$ in the non-stationary case. Hence, the right-hand side of equation (8) still depends on $\boldsymbol{s}$. Grouping the observed indicator residuals into lag classes and computing a sample variogram by the customary method-of-moments estimator render it meaningless in this instance. The simulations in the supplement to this article on http://www.pedometrics.org/paper/ik1_appendix.pdf illustrate that simple IK loses its mean square optimality when the variogram of the indicator residuals is estimated according to equation (7). We can then merely *hope* that kriging provides better predictions than other *ad-hoc* procedures such as inverse distance weighting of the indicators.

The indicator transforms of $\{Z(\boldsymbol{s})\}$ with constant $\mu(\boldsymbol{s})$ but unbounded variogram have non-stationary covariances, too. To see this, we consider Gaussian, zero order intrinsic $\{Z(s)\}$, $s\in\mathbb{R}$, with a linear variogram, $\gamma(h)=h$. Two increments, say $\tilde{Z}(s)=Z(s)-Z(0)$ and $\tilde{Z}(t)=Z(t)-Z(0)$, are then normally distributed with variances $\mathrm{Var}\left[\tilde{Z}(s)\right]=2s$, $\mathrm{Var}\left[\tilde{Z}(t)\right]=2t$ and correlation

$$\mathrm{Cor}\left[\tilde{Z}(s),\tilde{Z}(t)\right]=\rho=\frac{\min(s,t)}{\sqrt{s\,t}}. \quad (9)$$

Thus, their bivariate density function is equal to (Abramowitz and Stegun, 1965, p. 936)

$$g(z_s,z_t;s,t,\rho) = \frac{1}{4\pi\sqrt{s\,t(1-\rho^2)}} \cdot\exp\left(-\frac{z_s^2/s^2-2\rho z_s z_t/\sqrt{s\,t}+z_t^2/t^2}{4(1-\rho^2)}\right), \quad (10)$$

where ρ is of course given by equation (9). The covariance of the indicator transforms of the increments is

related to $g(z_s, z_t; s, t, \rho)$ by (Chilès and Delfiner, 1999, p. 400)

$$C_I(s,t;z',z'') = \int_0^{\frac{\min(s,t)}{\sqrt{s\,t}}} g(z',z'';s,t,\rho)\,\mathrm{d}\rho. \qquad (11)$$

Clearly, $C_I(s,t;z',z'')$ depends on s and t not only through the lag $h = |s-t|$, and the covariance is non-stationary. Although the loss of efficiency of simple IK with variograms estimated by equation (7) was quite small in simulations of Gaussian $\{Z(\boldsymbol{s})\}$ with constant $\mu(\boldsymbol{s})$ but unbounded variogram, there are no grounds to claim, as for example Lark and Ferguson (2004) did, that IK offers an advantage over disjunctive kriging (DK) for random processes with unbounded variograms.

I conclude by stating that any attempt to model a random process by IK with spatially varying first and second moments —either explicitly or implicitly by using ordinary IK within a local neighbourhood of support points—requires the modelling of non-stationary indicator variograms to preserve the mean square optimality of kriging. As we cannot estimate non-stationary variograms from only one realization of $\{Z(\boldsymbol{s})\}$, IK is in practice limited to geostatistical analyses of data without an apparent trend and a bounded variogram.

Inconsistent modelling of indicator variograms

Let us now consider limitation (B). In a typical application of IK, the indicator transforms of $\{Z(\boldsymbol{s})\}$ are computed, not just for one, but for a series of increasing cut-offs $z', z'', z''', \ldots$, that give rise to several sets of indicator variables $\{I(\boldsymbol{s};z')\}, \{I(\boldsymbol{s};z'')\}, \{I(\boldsymbol{s};z''')\}, \ldots$. Sample variograms are then computed for each set, and permissible variogram models are fitted to them. However, unlike the variogram of $\{Z(\boldsymbol{s})\}$, (cross-)variogram models for indicators must—in addition to the usual positive definiteness— meet further constraints:

i. The sill of an indicator variogram should be no larger than 0.25 as this is the upper bound for the variance of a Bernoulli random variable (cf. equation 5). In practice, this condition is frequently violated (e.g. Walker *et al.*, 2008; Lee *et al.*, 2007; Goovaerts, 1994).

ii. Journel and Posa (1990) list two more constraints that the second moments of indicators must satisfy. For the sake of simplicity, the relations are given here for the (cross-)covariances, but corresponding conditions apply to the (cross-)variograms. For $z' \le z''$, $z''' \le z''''$ and all $\boldsymbol{h}$ we must have

$$C_I(\boldsymbol{h};z'',z'''') - C_I(\boldsymbol{h};z',z''') \ge F(z')\cdot F(z''') - F(z'')\cdot F(z''''), \qquad (12)$$

$$C_I(\boldsymbol{h};z',z'''') + C_I(\boldsymbol{h};z'',z''') - C_I(\boldsymbol{h};z',z''') - C_I(\boldsymbol{h};z'',z'''') \le (F(z'')-F(z'))\cdot(F(z'''')-F(z''')). \qquad (13)$$

It is common practice to fit model functions to indicator sample variograms without checking whether the above conditions are met. It is then doubtful whether a set of fitted indicator variogram models codes the bivariate distributions of a stationary random process consistently.

iii. Matheron (1989) showed further that an indicator variogram must not be positively curved near the origin because any valid indicator variogram must satisfy the "triangle inequality"

$$\gamma_I(\boldsymbol{h}_1+\boldsymbol{h}_2;z',z') \le \gamma_I(\boldsymbol{h}_1;z',z') + \gamma_I(\boldsymbol{h}_2;z',z').$$

Thus, it is a mistake to use the Gaussian model without a nugget constant for indicator variograms (cf. Lloyd and Atkinson, 2001). Matheron also showed (cited in Armstrong, 1992) that there are some restrictions for the parameters of a "hole-effect" model to be a valid indicator variogram.

Model functions fitted to indicator sample variograms for a series of cut-offs can neither be fitted independently from $F(z)$ nor independently from one another. The above discussion demonstrates that positive definiteness is a necessary, but not sufficient, condition for consistent coding of the bivariate distributions of a stationary random process. Thus, the common practice, namely of ignoring these constraints when modelling indicator variograms, could lead to inconsistent probabilistic models. Although one could improve on this by using algorithms that fit the indicator variograms subject to all known constraints, this is hardly ever done in practice.

Summary and Conclusions

From a theoretical point of view, many applications of IK either are unsoundly based and as a result lack optimality (they ignore the non-stationary second moments

of the indicator for data with trend or unbounded variograms) or rely on inconsistent probabilistic models (studies that violate the constraints i–iii). Whether or not IK offers some empirical advantages over theoretically sound approaches should become clear from the expected continuation of this debate.

References

Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*. Dover, New York.

Armstrong, M. (1992). Positive definiteness is not enough. *Mathematical Geology*, **24**, 135–144.

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, New York.

Goovaerts, P. (1994). Comparative performance of indicator algorithms for modeling conditional probability distribution functions. *Mathematical Geology*, **26**, 389–411.

Goovaerts, P. and Journel, A. G. (1995). Integrating soil map information in modelling the spatial variation of continuous soil properties. *European Journal of Soil Science*, **46**, 397–414.

Journel, A. G. (1983). Nonparametric estimation of spatial distributions. *Mathematical Geology*, **15**, 445–468.

Journel, A. G. and Posa, D. (1990). Characteristic behavior and order relations for indicator variograms. *Mathematical Geology*, **22**, 1011–1025.

Lark, M. and Ferguson, R. B. (2004). Mapping risk of soil nutrient deficiency or excess by disjunctive and indicator kriging. *Geoderma*, **118**, 39–53.

Lee, J.-J., Jang, C.-S., Wang, S.-H., and Liu, C.-W. (2007). Evaluation of potential health risk of arsenic-affected groundwater using indicator kriging and dose response model. *Science of the Total Environment*, **384**, 151–162.

Lloyd, C. D. and Atkinson, P. M. (2001). Assessing uncertainty in estimates with ordinary and indicator kriging. *Computers and Geosciences*, **27**, 929–937.

Matheron, G. (1989). The internal consistency of models in geostatistics. In M. Armstrong, editor, *Geostatistics*, volume 1, pages 21–38, Dordrecht. Kluwer.

Papritz, A., Herzig, C., Borer, F., and Bono, R. (2005). Modelling the spatial distribution of copper in the soils around the metal smelter in northwestern Switzerland. In P. Renard, H. Demougeot-Renard, and R. Froidevaux, editors, *Geostatistics for Environmental Applications*, pages 343–354. Springer Verlag.

Walker, J. S., Balling Jr., R. C., Briggs, J. M., Katti, M., Warren, P. S., and Wentz, E. A. (2008). Birds of feather: Interpolating distribution patterns of urban birds. *Computers, Environment and Urban Systems*, **32**, 19–28.

Peter Alan Burrough

26 August 1944 - 9 January 2009

taken by Nicholas Burrough, and used with kind permission

Many pedologists and geographers will know of Peter Burrough from his book Principles of Geographical Information Systems for Land Resource Assessment in which he described how to organize, analyse and present spatial data on soil and land. The first edition was the outcome of his experience in soil survey, landscape classification and quantitative pedology.

Peter Burrough took his first degree in chemistry at the University of Sussex. He won a scholarship to pursue research in organic chemistry at Oxford. Once there, however, he discovered that soil was more interesting, and he joined Philip Beckett's small band of heretics who were questioning the orthodoxy of soil survey and seeking to place survey and classification on a proper quantitative basis. He was awarded his doctorate for his contribution.

In the last year of his doctoral studies he was appointed junior lecturer in the university's Geography Department. There his interest in geography, a subject he had not studied at school, grew. He successfully applied to join the British Overseas Development Administration and was appointed to serve as soil surveyor in Sabah, Malaysia. In Sabah he maintained his interest in statistical pedology while doing `bread-and-butter' survey for rural development. He then spent three years as lecturer in geography and soil science in the University of New South Wales. It was barren time, with a heavy teaching load and no time for research. So in 1976 he moved to the Netherlands, initially in the Soil Survey Institute in Wageningen and later in Wageningen University where he threw himself into Dutch life and culture. There his research career took off. He developed computer-based methods for landscape classification and display, leading to numerous publications on a variety of topics including fractals, geostatistics, error propagation and fuzzy classification. In 1984 his prowess, achievement and enthusiasm were recognized by the University of Utrecht which appointed him as professor of physical geography and geographical information systems.

He finished his book in between these two jobs. The book was an instant success in a time that GIS was rapidly developing and there were no authoritative texts yet. Peter became a GIS celebrity and travelled the globe to give keynote addresses and to promote his work and that of his students. Although his interests widened to encompass topics well outside of soil science, he continued to publish in journals of soil science and may be regarded as one of the founders of the pedometrics community. The new methods from mathematics, statistics and computer science that he introduced to soil science have helped shape the way we do quantitative soil science today. The British Soil Science Society recognized this when it made him an honorary member in 2008.

Peter thrived in the dynamic environment at Utrecht University and loved teaching as much as research. Unfortunately in 2005 the university's shortage of money forced him into early retirement, but it gave him the opportunity to accept an honorary research professorship at Oxford University. Sadly, illness soon took hold and prevented him from implementing his plans for research, and he returned to the Netherlands in October 2008.

Many will remember Peter for his charismatic presentations and influential publications, but what characterized Peter most was his unbounded enthusiasm and excitement for research. While in charge of a large research group with many responsibilities, he would still find time to develop tools for spatial analysis for his students, to make new discoveries and to share these with whoever passed his room. It is his passion for science that we shall remember most.

Richard Webster and Gerard Heuvelink

☆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: 30th April 2009

Notification of paper acceptance: 1st June 2009

Deadline for Registration: 1st July 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

A New Digital Soil Map of the World

GlobalSoilMap.net is a new global project that aims to make a new digital soil map.It is funded by a 18 million US$ from the Bill & Melinda Gates Foundation and AGRA to map the soils in Africa and to establish the global consortium. The *GlobalSoilMap.net* consortium, which is led by ISRIC - World Soil Information (Wageningen, Netherlands), includes the Joint Research Centre of the European Commission (Ispra, Italy), CSIRO (Canberra, Australia), the University of Sydney (Sydney, Australia), Institute of Soil Science of the Chinese Academy of Sciences (Nanjing, China), the Earth Institute at Columbia University (New York, USA), the US Department of Agriculture - Natural Resources Conservation Service (Morgantown, USA), IRD (Montpellier, France), the Brazilian Agricultural Research Corporation (Embrapa, Rio de Janeiro) and CIAT-TSBF (Nairobi, Kenya). The African leg of the project was launched in Nairobi in January and the global project was launched in New York in February 2009. And the following quotes are from press releases that preceded the Nairobi and New York launch.

Neil McKenzie and Jon Hempel in the eroded fields near Kisumu, Kenya, January 2009

Alex McBratney from the University of Sydney in Australia enthuses about the new map, "The global digital soil map will use advances in technologies including remote sensing, data mining and spatial databases, and our improved scientific understanding of soil, for accurate prediction and sampling of soil properties. The new maps will replace the beautiful coloured paper soil maps developed in the last century which depicted soil types and which were largely qualitative and somewhat fixed depictions of soil distribution. Digital soil maps, with their infinity of shades and colours and ways of presentation are essentially spatial information systems of soil properties key to the soil's sustainable productivity and ecosystem function. Digital soil maps are quantitative and dynamic and are in tune with the needs of scientists, policy makers and government officials. In a sense their use is only limited by the imagination of potential users. It is truly thrilling to be part of such a global enterprise."

Work has started in sub-Saharan Africa, to create the Africa Soil Information Service (AfSIS). "The best science and technology available must be deployed immediately if Africa's soils are to be managed in a sustainable manner. Let there be no mistake about the significance of this wonderful project," said Kofi Annan, chairman of AGRA and former UN Secretary-General, in a recent statement. "This initiative will provide farmers, policy makers, and scientists crucial information on how to address declining soil fertility in regions such as sub-Saharan Africa," explains Pedro Sanchez, director of AfSIS. "Soil mapping can help with that because it is one of the pillars to the challenge of sustainable development," according to Jeffrey Sachs, director of the Earth Institute at Columbia University (USA) and special advisor to the UN Secretary-General.

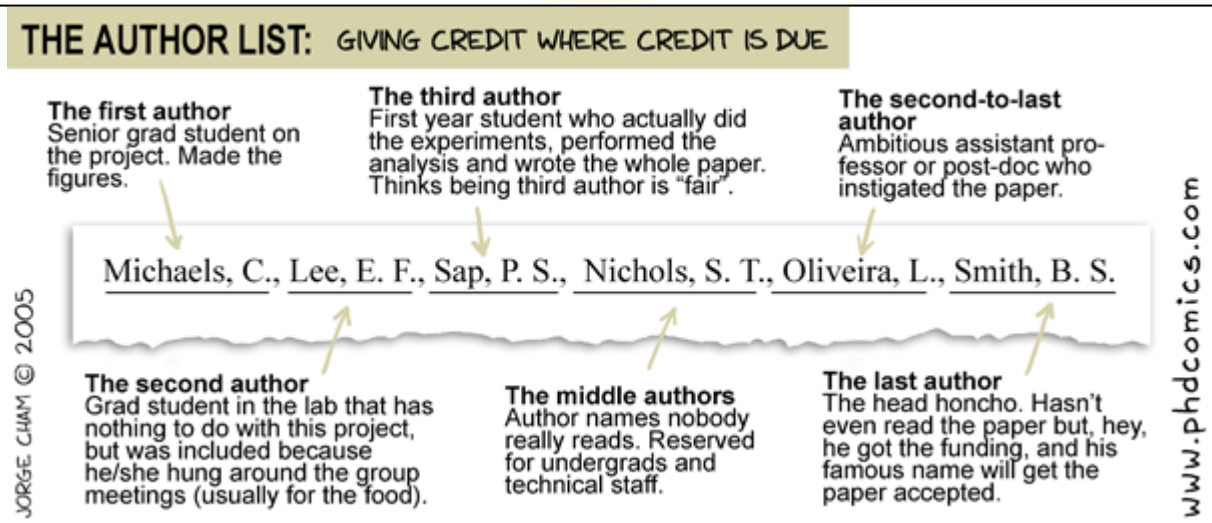
The map will have many uses in different parts of the world. Neil McKenzie, the Chief Land and Water of CSIRO in Australia states that: "In Oceania, reliable soil information is needed to assess and improve the efficiency of rain-fed and irrigated agriculture. The

Consortium meeting in New York. L to R: Luca Montanarella, Sonya Ahamed, Jon Hempel, Alfred Hartemink, Neil McKenzie, Lou Mendonca-Santos, Pedro Sanchez, Prem Bindraban, Young Hong, Peter Okoth, Gan-lin Zhang. 18th Feb 2009. Snow.

challenge of food security and human nutrition is a major issue and there is an urgent need to minimise exploitative land uses and soil degradation (especially through erosion and acidification). The region is very vulnerable to climate change and soil information is essential for planning major shifts in land-use, for example, in southern Australia where water scarcity is already a problem. As with other parts of the world, the best soils for biosequestration of carbon have to be located."

The GlobalSoilMap.net project will foster collaboration between institutions in Canada, Mexico and the USA to produce soil property data that is trans-national in nature, according to Jon Hempel, Co-Director-National Geospatial Development Center of the National Resource Conservation Service in the USA. Jon Hempel: "Legacy and heritage soil survey data holdings across North America that have been produced at different scales and under different taxonomic systems will be harmonized into a common, consistent and geographically contiguous dataset of soil properties. It will allow scientists and officials to more easily make application of the data for many interpretive uses across the North American continent."

www.globalsoilmap.net



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

Report from the Statistical Aspects of National Scale Soil Monitoring Workshop.

11th–12th December 2008, Rothamsted Research, UK.

This workshop brought together 40 or so scientists to discuss mutual interests in soil monitoring. We met in the bleak mid-winter damp of Rothamsted, but were inspired by the presiding genius of Ronald Fisher (although it was the wrong time of year to observe any variations in the 167th season of the Broadbalk wheat experiment). Most participants were from Europe, but the Americas were represented by Henry Lin from Penn State and Beate Zimmermann from the Smithsonian Tropical Research Institute. The proceedings of the meeting were structured around three keynote talks. The first, by Dick Brus (Alterra, Wageningen) was on the design of monitoring networks. The second, by Peter Loveland, (Rothamsted, formerly Cranfield University), was on interactions between scientists and policy makers on matters concerned with soil monitoring, with particular reference to the European Union. The third, by Dominique Arrouays (INRA, Orléans) considered some of the challenges of monitoring soil at national scale, and how approaches to analysis could be structured to meet the objectives.

In between the keynotes and individual presentations there were three breakout sessions, which tackled questions related to the themes of the keynote talks. Rather than leading you, patient reader, through a linear breakdown of proceedings, I shall present some of these questions to ponder, and some of our responses. If you want to have your say, then I suggest that you post your views on the Pedometrics Google Group.

How do we decide upon the temporal frequency of surveys when we have little information about the temporal correlations of soil properties?

- There may be scope to use process models to predict the changes the monitoring scheme has to detect, but this only takes us so far.
- In practice constraints of resources rather than clever statistics will probably constrain the sampling interval. In short, sample as often as the policy makers will pay for.
- But with appropriate analysis, it should be possible to refine the sampling interval over time, as information is obtained. We are likely to arrive at different answers for different properties in different places.

How do we ensure that sampling designs are robust, given that the actual variations of soil will be more complex than our assumptions?

- Design-based sampling is generally more robust than model-based to the existence of structures in the data that we don't know about in advance, keep it simple.
- Strategies can be compared over a space of model parameters to show which is most robust (Dick Brus illustrated this in his keynote).
- Sponsors must be aware of these limitations, a scheme is not necessarily the best possible, it is the best we can manage given the constraints on resources and inevitable uncertainty.

Murray Lark

Attentive audience

Workshop dinner (well it was nearly Christmas)

How do we translate statistical statements about the precision of a monitoring scheme into terms understood by policy-makers?

- Outline the costs that may arise from a wrong decision, and where possible quantify the risk. For example, a proposed scheme would allow the onset of a trend of magnitude x to be detected earlier than an alternative.
- Attempt cost-benefit analyses for a range of scenarios, focussed on the big issues (pollutants, risk of flooding etc)
- Policy makers want to see progress, and will not respond positively to an account of the difficulties in delivering a scheme.

Peter Loveland's keynote

Is the choice between model-based and design-based methods always straightforward for monitoring?

- No, there is always a trade-off.
  - Design-based sampling ensures that our results lack bias, and our confidence intervals should be meaningful, without any complex assumptions.
  - On the other hand, a model-based design (such as a grid) ensures good spatial coverage, and is flexible, allowing us to adapt to other requirements (such as producing local predictions).
- We should aim for simplicity, which design-based approaches generally deliver when there are not strong reasons to use model-based (e.g. where we need to map local variations).
- D-B samples can be analysed, post hoc, by model-based methods, and benefits are sometimes achieved by this (improved precision)
- M-B samples cannot be analysed post hoc as if they were DB, some modelling, or other assumptions are needed.

Thanks to all participants for making this a lively and interesting workshop, and particular thanks to Ben Marchant from Rothamsted who organized the scientific programme and the logistics with commendable efficiency. Thanks to Kathy Haskard of Rothamsted for the photographs.

Other Measures of Academic Productivity: INVITED TALKS

Invited Talks (adjusted) = # of Invited Talks you've given - # of times you just repeated the same old spiel

Range:	1 pretty much everybody

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

REPORT FROM THE TRENCHES:

Preparing developing-country students for pedometrics

D G Rossiter

International Institute for Geo-Information Science & Earth Observation (ITC)
Enschede (NL)
http://www.itc.nl/personal/rossiter/

Pedometron and journals such as Geoderma are full of exciting and sophisticated developments in the application of math and statistics to soil science. In most less-developed countries, however, these are hardly known or mechanically and often inappropriately applied. There is a serious disconnect between these two worlds, which ITC is charged with bridging as part of its mission: "capacity building and institutional development of professional and academic organizations and individuals ... in countries that are economically and/or technologically less developed." The "capacity" in this context is the ability to understand and apply pedometric techniques for a deeper understanding of the soil resource and to make better decisions.

All ITC students are post-graduate and most are supposed to have some working experience in their professional field. They come to the Netherlands to upgrade their skills and apply them in an MSc thesis. In fields such as earth sciences they are assumed to have an appropriate university degree, which should include the relevant domain background (e.g. geology) and also relevant methods (e.g. statistics and university-level maths). Unfortunately almost all our students are deficient in one or both of these areas. Yet, we want to educate them and thus contribute to development. Good examples are an agricultural statistician from Malawi and an urban planner from China, neither who has taken a soils or even earth science course at university and with no soils field experience, who have been assigned by their respective ministries to learn about soils to apply in their jobs. Other students have some soils background and work experience but almost no statistics, let alone calculus or linear algebra; this is typical of agriculture college graduates in many developing countries.

ITC has not had a separate soil survey course for several years; in common with many universities soils are now included somewhat vaguely in earth sciences, natural resources, water resources, and even urban planning courses. All of these require sound statistical thinking, especially for MSc thesis research. In our modular system all MSc students are exposed to statistical thinking in a Research Skills module, and are offered optional advanced topics in data analysis strategy, geostatistics, and quantitative modelling. Domain knowledge such as soil science is insinuated when possible, mostly via directed readings assigned by the student's tutor and thesis coach.

(ITC also offers distance education courses, for example my "Geostatistics and Open-Source Statistical Computing", six weeks half-time; here I only deal with the MSc course.)

What do I do with these students, in the limited time, and given the impossibility of a semester course or sequence?

Above all, I want them to learn how to learn:

(1) They should be able to read and understand statistics textbooks in order to apply the right techniques for each situation, and meet the assumptions of each technique.

I expose them to a variety of texts available in our library, and show how to pick one at the level appropriate for them. For most earth science students the text of Davis is at the perfect level, and contains a wide variety of relevant topics:

Davis, J.C., 2002. Statistics and data analysis in geology. John Wiley & Sons, New York, xvi, 638 pp.

For soils students that are bit more sophisticated I recommend:

Webster, R. and Oliver, M.A., 2008. Geostatistics for environmental scientists. John Wiley & Sons Ltd., 332 pp.

although I think the earlier text is more useful for beginners; too bad it's out of print and the publisher won't let us photo-copy it:

Webster, R. and Oliver, M.A., 1990. Statistical methods in soil and land resource survey. Oxford University Press, Oxford.

The book of Goovaerts is an excellent and comprehensive reference but too detailed for most beginners:

Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Applied Geostatistics. Oxford University Press, New York; Oxford, 483 pp.

A problem with texts for our clients is their price. I have tried without success to negotiate with the publishers of Webster and Goovaerts to either buy books at substantial discount or photocopy them and send the royalties to the publisher; one publisher offered a 10% discount and the other never answered. Publishing on-line or as e-books may be a solution: the UseR! series is somewhat more reasonable, e.g. $60 for:

Bivand, R.S., Pebesma, E.J. and Gómez-Rubio, V., 2008. Applied Spatial Data Analysis with R. UseR! Springer, 378 pp.

(2) Students should be able to understand journal articles and repeat the methods on other datasets.

Often the students must review concepts presented in the paper that they do not know. Here the reference list is quite important, as well as a clear expository style.

A good example of an accessible paper (among many I could have chosen) is:

Minasny, B. and McBratney, A.B., 2007. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma, 142(3-4): 285-293.

Their section on Theory is a clear exposition of the choices they made, and why, with references appropriate for a student without the necessary background. For example: "Training in supervised classification involves minimising some error measure (Hastie et al., 2001)", the cited reference is a good text:

Hastie, T., Tibshirani, R. and Friedman, J.H., 2001. The elements of statistical learning : data mining, inference, and prediction. Springer series in statistics. Springer, New York, xvi, 533 p. pp.

This is followed by a clear derivation.

Many developing-country workers do not have good library access, either physical or internet. Thus references should be as accessible as possible. Papers from conferences or obscure journals (unlikely to be available) should be avoided if possible.

(3) It is more important that students understand statistical thinking, rather than specific methods.

All statistical models have assumptions: what are these? how can you tell if they're met? what are the consequences of violating them?

For example, kriging interpolation is applicable in the presence of stationary spatial dependence which can be modelled, but if the geographic phenomenon is due to a regional trend, it is certainly not appropriate. So I spend considerable effort in comparing approaches and when each may be applicable.

(4) Some fundamental methods must be understood in some detail; the most important is linear modelling (single and multiple predictors) in feature space (also, trend surfaces in geographic space, although these are less useful).

For geostatistics, the fundamental methods remains trend identification and removal, variogram analysis and ordinary kriging.

Computer programs

For our client group I insist on free computer programs. Fortunately one of the best is not only free but open-source: the R environment for statistical computing (http://www.r-project.org/). I prepare all my exercises with R, Sweave and LaTeX so that the executable code is provided along with verified output. An outstanding feature of R is the wide variety of contributed packages, so the student soon sees "there's more than one way to do it". Also, methods typically have many options, all of which are applicable in some situations. The student learns that "press the button" or "accept the defaults in a dialog box" is not acceptable practice. Life is complicated, accept it!

R is also fairly easy to program, and is based on a modern programming language. I have prepared some technical notes (available via my ITC home page) using R, e.g. implementing Webster's split-moving-window approach:

Webster, R., 1973. Automatic soil-boundary location from transect data. Mathematical Geology, 5: 27-37.

The more ambitious students are able to write simple programs or modify existing ones.

I avoid Excel (or open-source equivalents) for anything beyond initial data entry; far better to get the data into R and develop analysis scripts which allow reproducible analysis and professional graphics.

Commercial programs such as SPSS and ArcGIS Spatial Analyst have three strikes against them for the group I am trying to teach: cost, push-the-button ease of use, and poor programmability. Spatial Analyst is very poorly documented; despite repeated attempts I have not been able to discover how the empirical variogram display is computed nor how a variogram is fit.

Papers

Finally, here is a list of some of my favourite journal articles for teaching. I have an extensive list of specialised papers from my favourite pedometric authors (e.g. Lark, Minasny, Viscarra Rossel, Brus) which I recommend to students as they enter their thesis phase; these are more general and used in teaching.

(1) Statistical thinking and elementary methods

Webster, R., 2001. Statistics to support soil research and their presentation. European Journal of Soil Science, 52(2): 331-340.

This one is simple but so many students benefit from just such an approach. This is supplemented by the aide-memoire from the Webster & Oliver text listed above.

Webster, R., 1997. Regression and functional relations. European Journal of Soil Science, 48(3): 557-566.

Far too many students jump into regression when it's structural relations they really want. I find Webster's expository style a good model for the students.

(2) Geostatistics

Oliver, M.A. and Webster, R., 1991. How geostatistics can help you. Soil Use & Management, 7(4): 206-217.

This is the most gentle introduction to "why should I learn this complicated stuff?".

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma, 103(1-2): 3--26.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma, 89(1-2): 1-45.

Both of these are comprehensive comparisons of approaches.

Webster, R., Welham, S.J., Potts, J.M. and Oliver, M.A., 2006. Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. Computers & Geosciences, 32(9): 1320-1333.

I have a soft spot for this one, since I grew up in Youden and Mehlich territory (upstate New York) and have visited their study area.

(3) Case studies

Goovaerts, P., 2000. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology, 228(1-2): 113-129.

Dubois, G., Malczewski, J. and Cort, M.D., 2003. Mapping radioactivity in the environment - Spatial Interpolation Comparison 97. EUR 20667 EN, Office for Official Publications of the European Communities, Luxembourg.

This serves as a model of an intelligent approach to solve a problem with a variety of techniques, pointing out the (dis)advantages of each. Papers have been collected in an EU publication free for download.

SICC '97 http://www.ai-geostats.org/index.php?id=45

(4) Digital soil mapping

Anyone getting into DSM is given this one, of course:

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

Conclusion

Pedometricians, keep on inventing the latest sophisticated methods! Keep on publishing excellent papers and writing reviews and texts. However, spare a thought for those who are far below the level needed to appreciate your cutting-edge work, and provide a stepped approach to bring them into the community.

Alex's Most Preferred Pedometrics Papers IV

Hole, F.D., Hironaka, M., 1960. An experiment in ordination of some soil profiles. Proceedings of the Soil Science Society of America 24, 309-312.

In this paper some prescient scientists first tried to reveal the multidimensional structure of the universe of soil profiles, and when I read the paper some fifteen years after it was written, it was that structure that fascinated me, not how they got it.

This first attempt was modest. They took some representative data from 25 different major groups of soil representing many great soil groups of the world and calculated dissimilarity coefficients using a simple metric. They went on to show the dissimilarity matrix graphically, which doesn't really reveal the structure, unless you have a special relativistic mind.

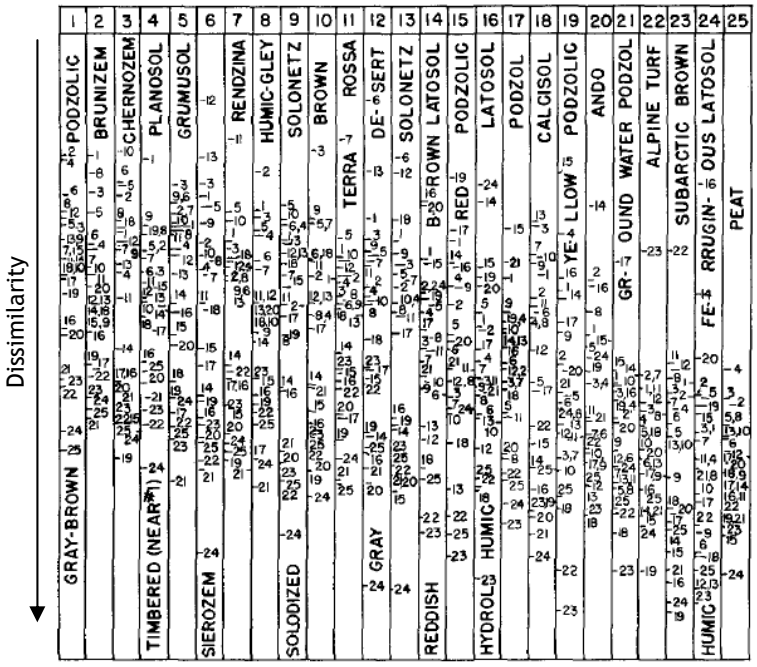
That was the easy bit. Although multi-dimensional scaling (Toregerson, 1958) had been invented they didn't use it. In fact one suspects that no digital computer was used for this study. They literally used carpentry. (The Goons might have complained about not being able to get the wood – but this was prosperous America – and Wisconsin has a lot of wood.) They made a best fit three-dimensional configuration by cutting lengths of wood in relation to the distance (dissimilarity) between distance between profiles.

We have taken the data from the diagram and present it as a shaded similarity matrix for the first time. We took the dissimilarity data and calculated the minimum spanning tree (Gower and Ross, 1969) and presented the first two dimensions of a non-linear map (Sammon, 1969). You can see the same thing on Hole and Hironaka's 3-D diagram.

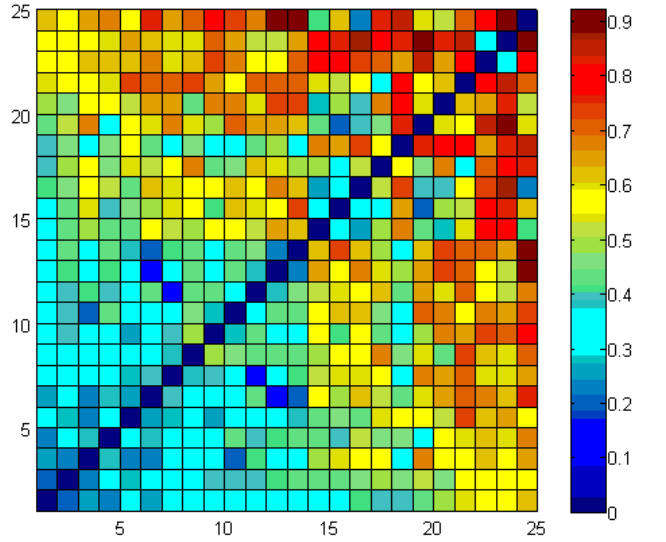
Francis Hole was a famous pedologist who had lots of good quantitative ideas and a passion for soil. It seems that Hironaka (and I apologise for not being able to discover his first name) was from the Botany and Soil Science departments and this suggests an important link. The botanists at UW Madison seem to have been at the forefront of the quantitative revolution in the late 1950's. Hole and Hironaka (1960) shows a large degree of similarity with techniques presented in Bray and Curtis' (1957) longer and more sophisticated paper in Ecological Monographs.

The first attempt at ordination in soil science may well have been Cox and Martin's (1937) paper on discriminant analysis of soil bacteria. There have been many ordination studies since then, but most of them have been local. I really think we've failed these two guys by not developing this further to a global view. They were pedologists with a quantitative bent – they knew what they were after.

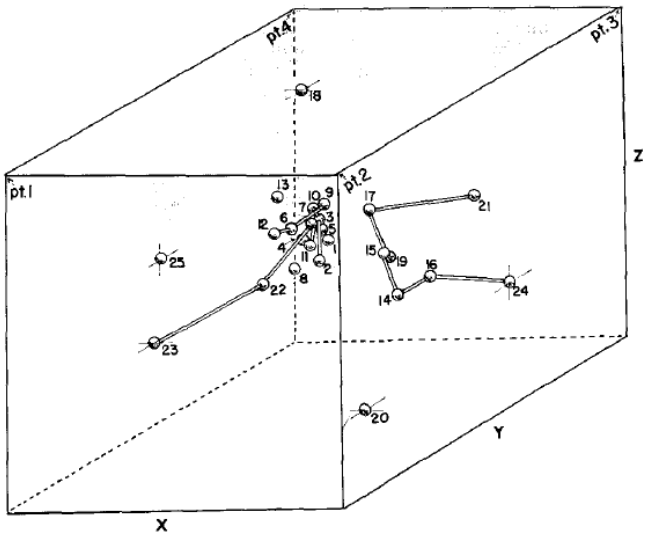
By now we should have a good idea of the structure of



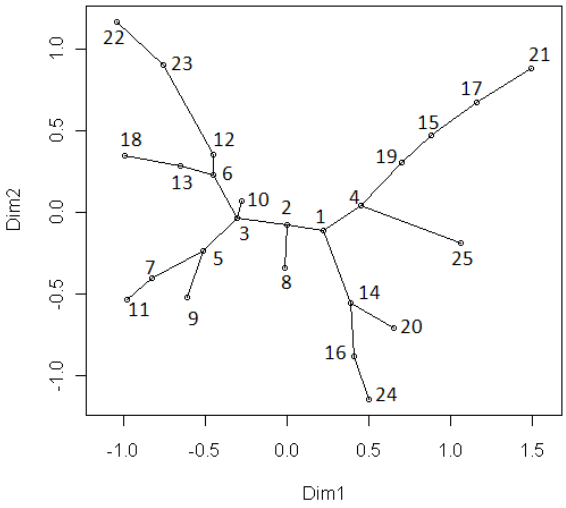
Degree of similarity between soil groups (Hole & Hironaka, 1960)



Relative distance matrix between soil groups. 1 Gray-brown Podzolic. 2 Brunizem, 3 Chernozem, 4 Timbered Planosol, 5 Grumosol, 6 Sierozem, 7 Rendzina, 8 Humic Gley, 9 Solonetz, 10 Brown 11 Terra Rossa, 12 Gray Desert, 13 Solonetz, 14 Redidish-brown Latosol, 15 Red Podzolic, 16 Hydrol Humic Latosol, 17 Podzol, 18 Calcisol, 19 Yellow Podzolic, 20 Ando, 21 Groundwater Podzol, 22 Alpine Turf, 23 Subarctic Brown, 24 Humic Ferruginous Latosol, 25 Peat.



Soil groups represented in "carpentry" techniques. (Hole & Hironaka, 1960)



Plot of soil groups in a minimum spanning tree.

the multidimensional space of soil horizons and classes derived from them and of soil profiles and classes derived from them. We simply don't – it seems we have lost fifty years. There are several careers in this. Recently we made a guess (Minasny, McBratney Hartemink, On global diversity, Geoderma, Submitted). We calculated the distance between the WRB soil groups and take principal coordinates (a technique invented for this very purpose by Gower (1966) which seems to be similar to Torgerson's scaling and first used for soil by Rayner (1966) and show the two most important orthogonal axes.

We've too often got lost in techniques and forget the goal about understanding soil and I'm probably as culpable as anyone.

Bray, J.R., Curtis, J.T., 1957. An ordination of the upland forest communities of southern Wisconsin Ecological Monographs 27, 326-349.

Cox, G.M., Martin, W.P., 1937. Use of a discriminant function for differentiating soils with different Azotobacter populations. Iowa State College Journal of Science 11, 323-332.

Goons, 1959. The Goon Show, Volume 10 You Can't Get The Wood, You Know! BBC

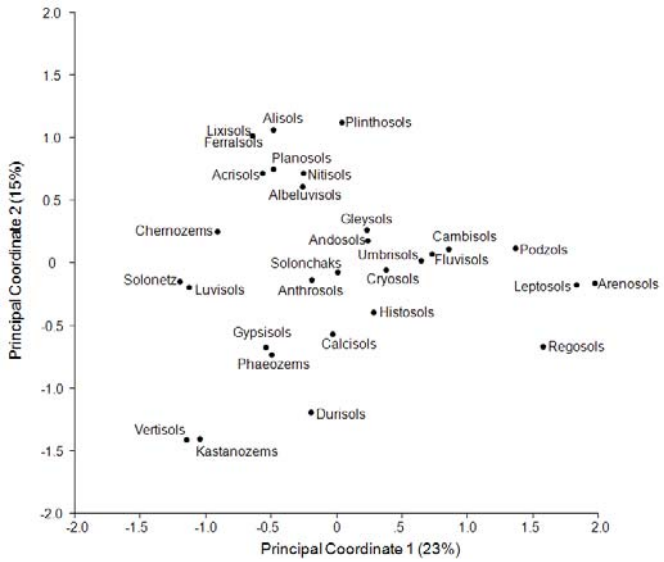
Gower, J. C., 1966. Some distance properties of latent root and vector methods used in multivariate analysis: Biometrika 53, 325-338.

Gower, J.C., Ross, G.J.S., 1969, Minimum spanning trees and single linkage cluster analysis. Applied Statistics 18, 54-64.

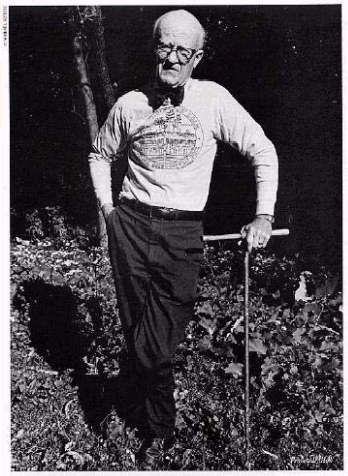
Rayner, J. H., 1966. Classification of soils by numerical methods. Journal of Soil Science 17, 79-92.

Sammon, J. W., 1969. A non-linear mapping for data structure analysis. IEEE Transactions on Computers C-18, 401-409.

Torgerson, W. S., 1958. Theory & Methods of Scaling. Wiley, New York.



Ordination of WRB soil groups (Minasny, McBratney & Hartemink).



I remember taking a hike led by Francis Hole during a "Prairies Jubilee" festival held at Goose Pond Sanctuary in Arlington, Wisconsin. We walked down the road a bit, with Dr. Hole in the lead, playing his fiddle and singing songs extolling the glories and mysteries of Soil. Suddenly he stopped playing, halting the march. He had us take off our shoes and socks and step barefooted out onto the prairie soil. "No talking now," he said. "Just walk quietly through the grasses and contemplate the complex and beautiful, yet unseen, world beneath your feet." He led on, playing a soft tune on his fiddle. I had a feeling that I was in a wonderful church.
--Martha C. Anderson (From http://www.soils.wisc.edu/~barak/fdh/index.html)

Did you miss this? ...

Murray

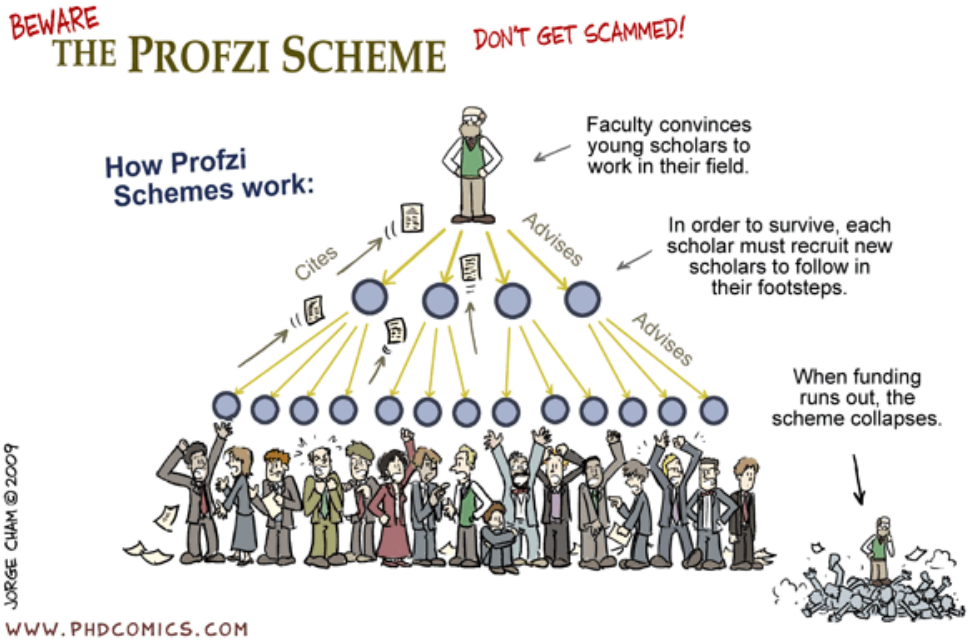
Gilles Guillot , Denis Kan-King-Yu , Joël Michelin and Philippe Huet. Inference of a hidden spatial tessellation from multivariate data: application to the delineation of homogeneous regions in an agricultural field. Journal of the Royal Statistical Society: Series C (Applied Statistics) 55, 407-430.

Abstract. In a precision farming context, differentiated management decisions regarding fertilization, application of lime and other cultivation activities may require the subdivision of the field into homogeneous regions with respect to the soil variables of main agronomic significance. The paper develops an approach that is aimed at delineating homogeneous regions on the basis of measurements of a categorical and quantitative nature, namely soil type and resistivity measurements at different soil layers. We propose a Bayesian multivariate spatial model and embed it in a Markov chain Monte Carlo inference scheme. Implementation is discussed using real data from a 15-ha field. Although applied to soil data, this model could be relevant in areas of spatial modelling as diverse as epidemiology, ecology or meteorology.

Much pedometrical analysis proceeds on the basis that the soil varies continuously from place to place. But we know that sometimes there really are boundaries in the soil landscape: at a geological fault, at a field boundary, at a break of slope. When there is information on these boundaries we can incorporate them into a geostatistical analysis as fixed effects in the linear mixed model. But what if we do not have this information. Some years ago Richard Webster showed how boundaries can be detected on a transect by analysing the variation in a moving window, but this does not generalize readily to two dimensions.

These authors propose a solution. They assume an underlying coloured Voronoi tessellation. That is to say, they assume an underlying completely spatially random distribution of seed points across the study area, and that the landscape is divided into polygons, each of which consists of all the locations that are nearer to one of the seed points than to any other seed point. Each cell is then randomly allocated to a class. Within any class of cells there is a uniform mean value for a target soil property. Variation about this mean has a distinctive variogram.

No mean task for inference, you may say; and you would be right. So, perhaps inevitably, the problem is put in Bayesian terms, and the inference is done by Monte Carlo Markov Chains (MCMC). The result is interesting, and reasonably convincing. Is it a workable methodology for practical purposes? Read the paper and decide for yourself.



Spatial Coverage Sampling on Various Spatial Scales

Dennis Walvoort, Dick Brus and Jaap de Gruijter

'What shall we use to fill the empty spaces...?'
Pink Floyd - The Wall (1979)

Introduction

De Gruijter et al. (2006) describe three sampling methods that result in sampling patterns suitable for mapping, i.e., centered grid sampling, geostatistical sampling and spatial coverage sampling. These methods have in common that they spread the sampling locations evenly over the study area in order to maximise the precision of geostatistical predictions.

In centered grid sampling, some kind of grid (usually a square grid, but sometimes a triangular or hexagonal grid) is placed over the study area, and samples are taken at the grid nodes. If a variogram is available, then the grid spacing can be optimised by procedures like OSSFIM (McBratney & Webster, 1981). Notwithstanding its appealing simplicity, centred grid sampling may be sub-optimal when the study area has an irregular shape, or when it contains areas that cannot be sampled (such as built-up areas, bird breeding areas, etc.). In addition, it is hard to take existing sampling locations into account.

Geostatistical sampling is much more flexible in these respects. In geostatistical sampling, a sampling pattern is optimised by minimising the variances of the prediction errors (Sacks & Schiller, 1988; van Groenigen et al., 1999). Geostatistical sampling heavily depends on a model of the spatial structure, like a variogram model. Unfortunately, a variogram model is not always available. In addition, the optimisation procedures for geostatistical sampling (e.g., spatial simulated annealing) are usually computationally demanding and need some prior tuning.

Spatial coverage sampling, on the other hand, does not need a variogram model. Instead, it uses a geometric criterion to optimise the sampling pattern. For example, Brus et al. (2003) proposed that one minimises the mean of the squared shortest distances (MSSD) between the sampling locations and an imaginary fine grid covering the study area. This criterion can be efficiently minimised by *k*-means (Hartigan & Wong, 1979). Brus et al. (2003) showed that a sampling pattern based on the MSSD has a mean ordinary kriging variance (MOKV) only marginally larger than that of a sampling pattern obtained by directly minimising the MOKV. The *k*-means algorithm has been used before for sampling by Brus et al. (1999) for estimating spatial means and by Walvoort et al. (2000) for mapping.

Unfortunately, software on spatial coverage sampling is not generally available. Therefore, researchers often have to resort to centered grid sampling instead. The aim of this article is to present new software for spatial coverage sampling. First the software will be briefly described, followed by some illustrative examples.

The spcosa-package

The software is implemented as an R package (R Development Core Team, 2008). R is a free programming environment for data analysis and graphics that has become extremely popular during the last decade. It offers many add-on packages for spatial data analysis and visualisation and is highly extensible. Our package is called 'spcosa' and implements the following sampling methods:

- spatial coverage sampling;
- spatial coverage sampling with prior points ('spatial infill sampling');
- random sampling from compact geographical strata;
- random sampling from compact geographical strata for composites.

Each method uses a variant of *k*-means to optimise the sampling pattern. The basic idea is to distribute sampling points evenly over the study area by selecting these points in compact geographical strata. Compact strata can be obtained by *k*-means clustering of the cells making up a fine grid representing the study area of interest. Two *k*-means algorithms have been implemented in the spcosa-package: a transfer algorithm and a swapping algorithm. The transfer algorithm obtains compact clusters (geographical strata) by transferring cells from one cluster to the other, whereas the swapping algorithm achieves this by

swapping cells between clusters. The first algorithm results in compact clusters (which are not necessarily of equal size), whereas the second algorithm results in compact clusters of equal size. In this article, the focus is on spatial coverage sampling. That is, the centroids of the strata are taken as the sampling locations. For examples on random sampling, the reader is referred to the package documentation (Walvoort et al., 2009).

Examples

In this section, we will present some examples on spatial coverage sampling on various spatial scales. We will start at the field scale and zoom out to the global scale. The first two examples have been adopted from actual research projects, but have been simplified for didactical reasons.

Field scale

The first example is about spatial coverage sampling of an agricultural field in the South-West of the Netherlands. The aim is to create a sampling pattern for mapping soil nitrogen and soil organic matter contents. Spatial coverage sampling has been applied to yield the sampling pattern in Figure 1. The sampling points are evenly distributed over the field. Note that the sampling pattern bears some resemblance to a triangular grid.

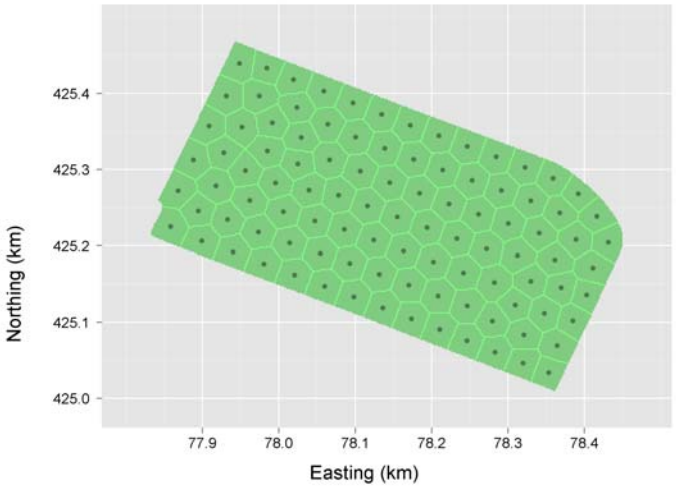


Figure 1: Spatial coverage sampling for an arable field in the South-West of the Netherlands.

Regional scale

Going from the field scale to the regional scale, it will be more likely that enclosures occur that should not be sampled (e.g., buildings, roads, and water courses). That is the case in our next example, which is about

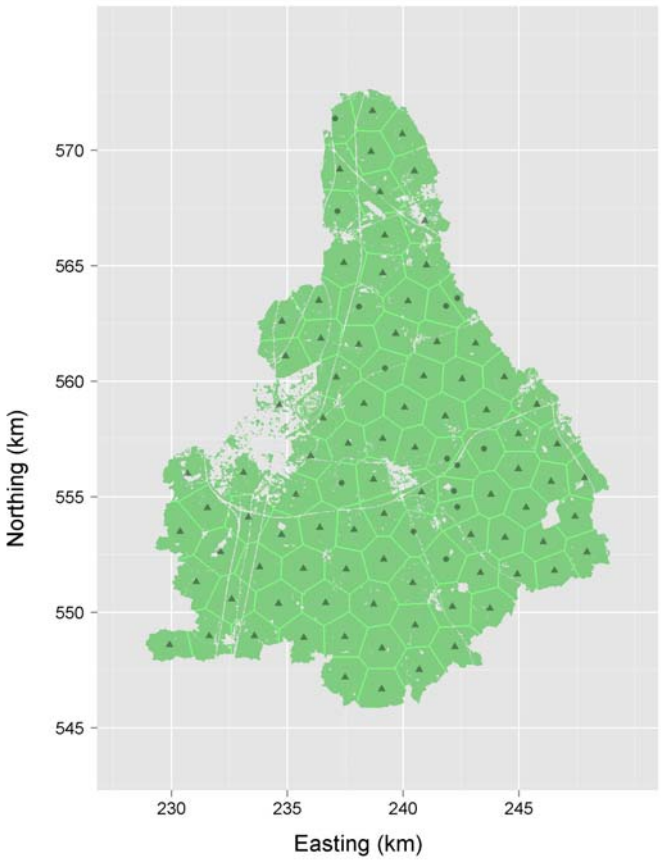


Figure 2: Spatial coverage sampling with prior points for the Drentsche Aa catchment. Prior points are given as dots, new points as triangles. The gray enclosures are built-up areas, roads and water courses.

mapping phosphorus related soil properties in the catchment of the Drentsche Aa River. This catchment is located in the North-East of the Netherlands. At fourteen locations in this catchment the phosphorus status is known from a previous soil inventory. The aim is to add eighty-six new locations taking these prior locations into account. Figure 2 shows the resulting sampling pattern. The fourteen prior locations are given as circles, the new locations are given as triangles. Note that the new locations are evenly spread over the catchment and keep some distance from the prior locations. Also note that new locations are not placed in built-up areas.

Global scale

Taking a giant leap from the regional scale to the global scale, an additional complication comes into scope: the curvature of the Earth's surface. In the examples above, the *k*-means algorithms use squared Euclidean distances in the objective function. However, at continental and global scales, squared Euclidean distances are not appropriate, and squared great circle distances should be used instead. In addition, also the way in which centroids have to be computed

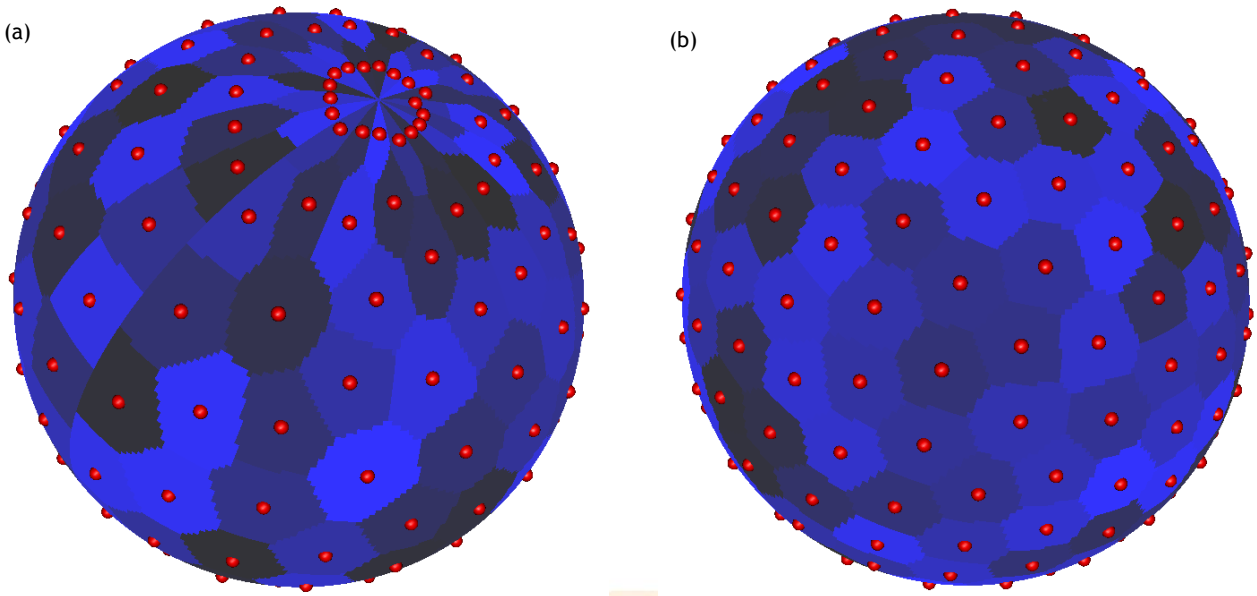


Figure 3: Spatial coverage sampling for the entire Earth: a) based on squared Euclidean distances, b) based on squared great circle distances. The sampling locations are given in red, the strata in different shades of blue. The shades of blue do not have a special meaning other than to make the strata more distinctive.

is more complicated for a sphere. These issues can be illustrated by applying spatial coverage sampling to evenly distribute 200 points over the surface of the Earth (Figure 3). In Figure 3a, a *k*-means algorithm that uses squared Euclidean distances has been used to compute strata and centroids. Note that the sampling density is greater near the poles than near the equator. The algorithm clearly failed to distribute the sampling locations evenly over the sphere. In addition, the strata in Figure 3a suffer from pronounced edge effects near the poles and at 180 degrees longitude (runs from the lower-left to the upper-right in Figure 3a). The strata are discontinuous at this meridian, i.e., two points on opposite sides of the meridian are treated as very distant when squared Euclidean distances are used. Figure 3b shows the sampling pattern in case a variant of *k*-means has been used based on squared great circle distances. These sampling locations are more evenly distributed and don't suffer from edge effects.

Availability

R and the spcosa-package can be downloaded from the Comprehensive R Archive Network (cran.r-project.org). More examples and details on the implemented algorithms can be found in Walvoort et al., (submitted) and in the package itself (Walvoort et al., 2009). The package also contains a tutorial.

References

Brus, D. J., L. E. E. M. Spatjens, J. J. de Gruijter, 1999. A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. Geoderma 89: 129-148.

Brus, D. J., J. J. de Gruijter and J. W. van Groenigen, 2003. Designing spatial coverage samples by the k-means clustering algorithm. Proceedings of 8th International FZK/TNO conference on contaminated soil. ConSoil 2003, Gent Belgium, p. 504-509

de Gruijter, J., D. Brus, M. Bierkens, M. Knotters, 2006. Sampling for Natural Resource Monitoring. Springer, Berlin, 332 pp.

Hartigan, J. A. and M. A. Wong, 1979. A k-Means Clustering Algorithm. Applied Statistics 28: 100-108

McBratney, A.B., and R. Webster, 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables: 2 program and examples. Computers & Geosciences 7: 335-365.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sacks, J. and S. Schiller, 1988. Spatial Designs. In: Gupta, S. S. and J. O. Berger (eds.). Statistical Decision Theory and Related Topics IV. Springer Verlag, New York.

van Groenigen, J. W., W. Siderius and A. Stein, 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87: 239-259

Walvoort, D. J. J., J. Bouma, J. J. de Gruijter, P. D. Peters, 2000. Using ISFETs for Proximal Sensing in Precision Agriculture. In: P. C. Robert, 2000. Proceedings 5th International Conference on Precision Agriculture

Walvoort, D. J. J., D. J. Brus, J. J. de Gruijter, 2009. spcosa: Spatial Coverage Sampling. Available at cran.r-project.org

Walvoort, D. J. J., D. J. Brus, J. J. de Gruijter, 2009. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Submitted to Computers & Geosciences.

Upcoming Events

EGU 2009 19 – 24th April 2009, Vienna, Austria. http://meetings.copernicus.org/egu2009/

Digital Soil Mapping

Session SSS12-Digital soil mapping: novel approaches to the prediction of key soil properties for modelling physical processes

Complexity and nonlinearity in soils

Session NP3.9/SSS39, a joint venture between the Nonlinear Processes in Geophysics and Soil System Sciences groups of EGU.

Diffuse Reflectance Spectroscopy

Session SSS25 Diffuse reflectance spectroscopy in soil science and land resource assessment.

Pedometrics 2009

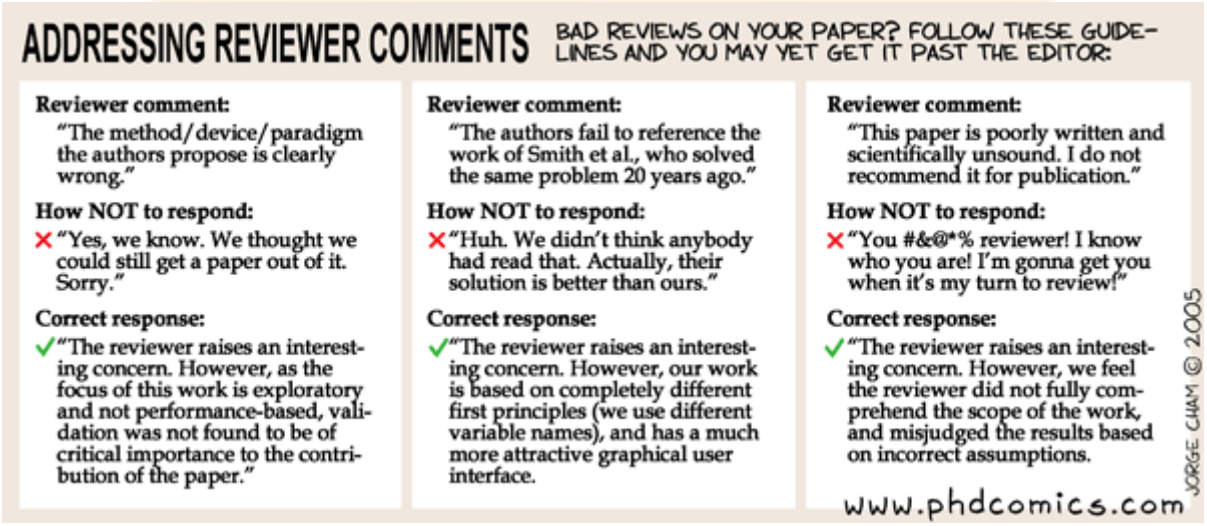
International Conference Centre, China Agricultural University, Beijing. 26-28 August 2009. http://www.pedometrics.org/2009/

International Conference on Geomorphometry

University of Zurich, Department of Geography (Irchel campus). Workshops 29 August & 30 August 2009. Conference 31 August - 2 September 2009. http://2009.geomorphometry.org/

International Conference on Soil Geography: New Horizons

Huatulco Santa Cruz, Oaxaca, Mexico, 16-20 November 2009. http://www.soilgeography09.fciencias.unam.mx/



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

Mapping research hot-spots using Citation Rate and Google geocoding service

Tomi Hengl

Any researcher or research organization can be successfully evaluated nowadays using web services such as Web of Science, SCOPUS, Google Scholar or similar (Meho and Yang, 2007). Objective measures such as Citation Rate (number of citations an author or a library item receives in average per year) can be used to depict the most influential authors/publications and research institutes/organizations in the world. If the library items are linked to geographical location, such data can also be used to generate scientific productivity and excellence maps.

We have recently analyzed the publications in the field of geostatistics and produced global maps of research excellence (Hengl et al., 2009). We will now guide you through all steps taken so some of you might try to run the similar analysis for any given pedometrical field .

You first need to obtain publications and their citation statistics from the Web of Science, Scopus and/or Google Scholar, and focused on the citation rates (CR). For each publication, you need to have also the contact author addressees. Then, you can attach geographic coordinates to each publication by using the contact author's address and the Google's API service. Once you attach the coordinates to each article, you can analyze this dataset using some point pattern analysis (or geostatistical) algorithm, e.g. to derive the global density maps of citations, which can be used to detect areas of scientific excellence for a given field.

STEP 1: Obtain the publication records for a given scientific field

In our case (geostatistics), we started by defining geostatistics by listing a number of keywords that are unique for the field and can be associated only with a limited number of authors. After we have determined those keywords, we can run queries on various databases to obtain all references belonging to that group. In the case of WoS, the query was:

```
topic=(kriging OR variogram OR "spatial statistic" OR "spatial interpolation" OR "spatial predict" OR "spatial sampling" OR geostatistic*)
```

and in the case of SCOPUS:

```
TITLE-ABS-KEY(kriging OR variogram OR "spatial statistic" OR "spatial interpolation" OR "spatial predict" OR "spatial sampling" OR geostatistic*)
```

Once we retrieve the results of query, we can sort them by relevance (number of times specified words appear in the text) and then export the first e.g. 2000 from the list. This way we are sure that we will be really processing representative articles. In the case of Google Scholar, we are not able to sort the results based on the relevance so we searched citations with ANY of the words: kriging, interpolation, and sampling, and with all of the words: spatial, statistic* and variogram. This can be efficiently run using the "Publish or Perish" software provided by Anne-Wil Harzing (Harzing and van der Wal, 2007).

These queries gave us 6,393 publications from WoS, 10,491 from SCOPUS and 5,389 publications from GS (compare with the results of Zhou et al. 2007). The WoS and SCOPUS publications were first sorted by relevance and then the first 4,000 entries were exported, filtered and reorganized to allow for further statistical analysis and processing. The GS database, which is noisy, requires filtering before it can be used. We often found duplicate or triplicate publications in the systems, but there are also many publications with misspelling (special symbols) of authors' names. However, most of these can be easily filtered out, either by visually examining the results or by running operations in R environment for statistical computing.

STEP 2: Attach geographic coordinates to each publication

In the following step, we need to attach geographic coordinates to the extracted articles by using the address of the contact author (we will focus on the results from WoS only). Here we use the Google's geo-

graphic service, which allows us to get geographic coordinates given a street + city + country address (see also coverage detail of Google maps). First, register your own Google API key. Now, to geocode an address, you can run in R:

```
> readLines(url("http://maps.google.com/maps/geo?q=1600+Amphitheatre+Parkway,+Mountain+View,+CA&output=csv&key=abcdefg"), n=1, warn=FALSE)
```

which will give four numbers: 1. HTTP status code, 2. accuracy, 3. latitude, and 4. longitude. In the case from above:

```
[1] 200.00000 8.00000 37.42197 -122.08414
```

the status code is 200 (meaning "No errors occurred; the address was successfully parsed and its geocode has been returned"; see also the status code table), the geocoding accuracy is 8 (meaning highly accurate; see also the accuracy constants), longitude is 37.42197 and the latitude is -122.08414.

Note that the address of a location needs to be provided in the following format:

```
"StreetNumber+Street,+City,+Country"
```

We can now loop this operation for a vector of addresses (contact authors):

```
> library(spatstat)

> library(rgdal)

> library(maps)

> googlekey <- "abcd"  # please obtain the correct Google API key!

>
```

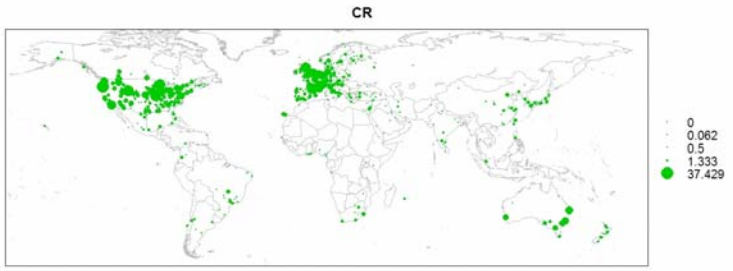
Obtaining longitude/latitudes from the Google API service can be problematic for slower internet connections and a long list of addresses. In fact, Google limits the number of geocode requests to 15,000 in a 24 hour period (read more). Also note that many articles have multiple addresses, so it might be a good idea to split the CR values among authors, e.g. using some decay function: e.g. if there are four authors, the first authors gets 50% of credit, the second 25%, third 15% and the last 10%.

STEP 3: Run spatial analysis and produce density maps

After we have obtained coordinates for each article, we can convert the table into a point map using:

```
> wosmap <- subset(wos, !is.na(wos$lat))

# insert a small location error to reduce duplicate points;

> wosmap$rlat <- round(wosmap$lat + rnorm(1, mean=0, sd=0.001), 4)

> wosmap$rlon <- round(wosmap$lon + rnorm(1, mean=0, sd=0.001), 4)

> coordinates(wosmap) <-~rlon+rlat

> proj4string(wosmap) <- CRS("+proj=longlat +datum=WGS84")

> worldmap <- map2SpatialLines(map("world", fill=TRUE, col="transparent", plot=FALSE),

proj4string=CRS("+proj=longlat +ellps=WGS84"))

> bubble(wosmap[!is.na(wosmap$CR),"CR"], sp.layout=list("sp.lines", worldmap,

col="grey"), maxsize=2)
```

which produces the following plot:



A bubble plot showing the citation rates in the field of geostatistics. Based on results in January 2008. (See http://spatial-analyst.net/wiki/images/a/a8/Fig_CR_geostatistics_worldmap.jpg for a bigger image).

Once we had attached the geographic location to a selection of articles, we can use the isotropic Gaussian kernel smoother (weighted by the CR) to map scientific excellence around the world. This can be run in e.g. the spatstat package (Baddeley, 2008). First, we will import a 20 arcminutes mask map of the world with all land areas:

```
> worldmaps20 <- readGDAL("mask20.asc")

> names(worldmaps20) <- "mask"

> wowin <- as(worldmaps20, "owin")
```

Next, we can convert the point map to a point pattern (spatstat data format) and run an isotropic Gaussian kernel with a bandwidth of 0.5 arcdegrees:

```
> wosCR.ppp <- ppp(wosmap@coords[,1], wosmap@coords[,2], marks=wosmap$CR, window=wowin)

> densCR <- density.ppp(wosCR.ppp, 0.5, weights=wosmap$CR, edge=TRUE)

> plot(densCR)

# export to a GIS format:

> dens.CR=as(densCR, "SpatialGridDataFrame")
```

World maps of bibliometric parameters for geostatistics estimated using a sample of 4000 articles: (a) density of published research articles generated using the isotropic Gaussian kernel with a standard deviation of 0.5 arcdegrees; (b) the same but weighted using the CRs for each article. Based on results in January 2008.(See http://spatial-analyst.net/wiki/images/6/6b/Fig_geostatworld.jpg for higher resolution image)

The figure above shows locations of both high productivity and high CR. This revealed clusters of scientific excellence around European locations such as Barcelona, London, Louvain, Norwich, Paris, Utrecht, Wageningen and Zürich; US locations such as Stanford, Ann Arbor, Tucson, Corvallis, Seattle, Boulder, Montreal, Baltimore, Durham, Santa Barbara and Los Angeles; and also around Canberra, Melbourne, Sydney, Santiago, Taipei, and Beijing. So if you plan to study or do top-geostatistics, these are the places where you should go!

Interested to run similar analysis for pedometrics or its subfield? Take a look at the R script we used to analyze geostatistics and let us know if you experience problems.

http://spatial-analyst.net/wiki/index.php?title=Mapping_research_hot-spots

References

Baddeley, A., 2008. Analysing spatial point patterns in R. CSIRO, Canberra, Australia.

Gotway Crawford, C.A., Young, L.J. 2008. Geostatistics: What's Hot, What's Not, and Other Food for Thought. In: Wan, Y. et al. (eds) Proceeding of the 8th international symposium on spatial accuracy assessment in natural resources and environmental sciences, World Academic Union (Press), pp. 8-16.

Harzing, A.W.K., Wal, R. van der 2008. Google Scholar as a new source for citation analysis? Ethics in Science and Environmental Politics, vol. 8, no. 1, pp. 62-71.

Hengl, T., Minasny, B., Gould, M., 2009? A geostatistical analysis of geostatistics. Scientometrics, in press.

Meho L.I., Yang K. 2007. A New Era in Citation and Bibliometric Analyses: Web of Science, Scopus, and Google Scholar. Journal of the American Society for Information Science and Technology 58:1-21.

Zhou F., Huai-Cheng G., Yun-Shan H., Chao-Zhong W., 2007. Scientometric analysis of geostatistics using multivariate methods. Scientometrics 73:265-279

PostScript Note

A website called Author Mapper http://authormapper.com/ from Springer searches journal articles (and plots the location of the authors on a map. However, it only for Springer publications and only on the corresponding authors.

Pedometrician profile

Dick Brus
Alterra, Wageningen UR

How did you first become interested in soil science?

I can't remember, but my wife always says that I am an earthy kind of person. So, maybe this explains why I decided to start studying geology at Amsterdam. After my bachelor degree I switched to physical geography, and followed a course in soil science at Wageningen. This was my first experience of soil science. It was quite different from the courses on structural geology, sedimentology, palaeontology et cetera. I liked it because soil science paid more attention to practical issues such as land evaluation. But when thinking back to those courses on continental drift and seafloor spreading, climate change and sea level rise et cetera, I cannot avoid feeling a bit melancholic.

How were you introduced to pedometrics?

I got involved in pedometrics about twenty years ago. I started as a geomorphologist at the Soil Survey Institute, and then moved to the team working on the Soil Map of the Netherlands at scale 1:50 000. At that time, Ben Marsman and Jaap de Gruijter worked on sampling strategies for validation of soil maps, and on other statistical topics such as spatial interpolation and fuzzy classification. These quantitative methods were really new for me, and I was enthusiastic from the beginning.

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

My favorite pedometrics topic is sampling for survey and monitoring, and especially the fundamental differences between the design-based and model-based approach to sampling. This fundamental difference is also relevant to the design of experiments. Either you assign the treatments randomly to the experimental plots, or you model the spatial variation of soil factors that might have an effect. The conclusions that can be drawn from the experiment concern the experimental fields, which generally is rather a restricted area. C.D. Smith and D.E Johnson (2009) showed how design-based sampling and design-based experimental design can be combined. Experimental plots are randomly selected from a larger area, and treatments are randomly assigned to these randomly selected experimental fields. The variance estimator accounts for sampling variation of the treatment effects in the larger area, which strongly enhances the practical relevance of the experiment (C.D. Smith and D.E Johnson, 2009, Environmetrics 20: 86-100).

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

The most challenging topic, from a scientific point of view, which I am thinking about at the moment is the combination of design-based and model-based sampling approaches for monitoring. For monitoring we must select sampling locations and sampling times. I am thinking about a mixed sampling approach for estimating the temporal trend of the spatial mean. In this new approach sampling locations are selected by probability sampling but sampling times are selected non-randomly, at constant interval, with the first sampling round at the start and the final sampling at the end of the monitoring period. By selecting sampling locations randomly, calibration of a space-time model is not needed, a time-model for the spatial means is enough. This can be advantageous if we have sparse data for space-time modeling, and when the validity of the result is important.

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

Difficult question. I think an important issue is the use of soil legacy data in Digital Soil Mapping and Digital Soil Monitoring. These prior data contain a lot of information on the soil, but at the same time we may question the representativeness of the data. I have thought a bit on how these non-probability data can be combined with probability sample data for estimating spatial means for instance of soil map units (see Brus and de Gruijter, 2003, Environmental Monitoring and Assessment 83: 303-317), but I feel that we possibly need a Bayesian approach for this, which is fundamentally different from design-based and model-based approaches. I would like to encourage pedometricians to explore the potentials of this Bayesian approach in statistical DSM.

Non-Pedometrician profile

Johannes Lehmann
Cornell University

How did you first become interested in soil science?

When I started university, I did not know that I would graduate as a soil scientist. After all, who knows that something like a soil scientist even exists? I started studying environmental sciences for a degree called "Geoecology". After a couple of semesters at the University of Bayreuth, I was intrigued by the interdisciplinary nature of the science that is brought to bear on the investigation of soils. And it so happened that the soils program had several opportunities to work in the tropics, and the decision was made.

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

My program works on several aspects of soil organic matter and nutrient dynamics. One area of heightened activity is the investigation of mechanisms that lead to stable carbon in soils. Knowledge about carbon stabilization is important not only from a perspective of carbon sequestration and climate change mitigation, but also from a perspective of agricultural sustainability. Several avenues of increasing stable carbon in soil require more research including stabilization on mineral surfaces as well as transformation of biomass into black carbon. Follow-up questions include how such processes scale to the regional and global level.

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

Basic statistics ranging from t tests to analyses of variance and multiple regression, but also more advanced principle component analyses. Advancing our understanding of carbon stabilization mechanisms require development of spatial statistics on the scale of individual microaggregates that could allow radically new insight into the process. In addition, models for soil carbon and nitrogen turnover are used, both on the scale of an individual site and on regional or global scale. Increasingly life-cycle assessments and other budget approaches are being applied to carbon cycle science.

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

I increasingly straddle the area between empirical science and modeling, and have enjoyed tremendously fruitful collaboration with researchers that specialize in mathematical approaches to soil investigation. This is a very rewarding approach, and allows quantification of complex interactions between processes. Since soils are a very complex beast, such collaborations are almost a must, and require engagement on all parties.

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

In carbon cycle science, close collaboration between modeling and measurements are essential. Since disciplinary specialization is in many cases unavoidable to reach the scientific depth required to deal with complex methodological or analytical problems, working in groups is necessary, where individual members bring different knowledge and skills to the table. Finding and building a working relationship,- often beyond institutional and national boundaries – is not easy. And administrative hurdles don't make it any easier either. Communication is key to work across disciplines, but often the time is lacking to move collaboration forward. There are many areas of soil science that would benefit from such collaboration, if not all of them.

Answers to Pedomathemagica (issue 25)

with Dick Webster

In the previous issue of Pedometron we asked how you should respond to the situation if your paper were voted the best for the Pedometric prize. After guessing which of three boxes contains a prize, you are shown one of the remaining two which is empty. You are given the chance to switch your choice. Should you?

Answer: you should switch to the unopened box.

Explanation

In the first instance there is an equal probability that the prize is in any of the three cardboard boxes, so you have a 1 in 3 chance of choosing correctly.
Now consider the judge's reaction and the author's response to the second question.
If you choose correctly, with probability 1/3, then the judge may open either of the other two boxes. In these circumstances you will win if you stick to the original choice and lose if you switch.

If you choose an empty box, with probability 2/3, then the judge must open the only other empty box. The certificate is in the third box. Now you will win if you switch your choice to that box and lose if you stick.

Overall therefore, you have a 1 in 3 chance of winning by sticking to your original choice and a 2 in 3 chance of winning by switching.

Bayes's Theorem

As above, the chance of your choosing correctly the first time is 1/3; that is the probability that the certificate is in any particular box. Suppose that you choose box 3. As far as you are concerned the probability that the judge will open another box, say box 1, is

$$\begin{aligned}\Pr(\mathrm{O1}) &= \Pr(\mathrm{J1}) \times \Pr(\mathrm{O1}|\mathrm{J1}) \\ &+ \Pr(\mathrm{J2}) \times \Pr(\mathrm{O1}|\mathrm{J2}) \\ &+ \Pr(\mathrm{J3}) \times \Pr(\mathrm{O1}|\mathrm{J3}) \ ,\end{aligned}$$

in which O1 means open box 1, J1 means certificate in box 1, etc. When the certificate is not in box 1 this will be

$$\begin{aligned}\Pr(\mathrm{O1}) &= \frac{1}{3} \times 0 + \frac{1}{3} \times 1 + \frac{1}{3} \times \frac{1}{2} \\ &= \frac{1}{2} \ .\end{aligned}$$

When the certificate is in box 1 the judge will not open that box; when it is in box 2 he will certainly open box 1, and when it is in box 3 he may open either box 1 or box 2 with equal probabilities of 1/2. Hence the probability that the certificate is in box 2, given that the judge opens box 1, is

$$\begin{aligned}\Pr(\mathrm{J2}|\mathrm{O1}) &= \frac{\Pr(\mathrm{O1}|\mathrm{J2}) \times \Pr(\mathrm{J2})}{\Pr(\mathrm{O1})} \\ &= \frac{1 \times 1/3}{1/2} \\ &= \frac{2}{3} \ .\end{aligned}$$

In the same way you can find that Pr(J1|O2) is also 2/3.
The second question concerned some naïve engineers. They decided not to worry about the threat of an annual flood to a road that they were planning, because the expected frequency of the event was once in 100 years and the design life of the road was only 50 years. What is the probability of a flood happening during this 50-year period?

Answer: the probability is about 0.395, large enough to give pause for thought.

Explanation

The occurrence of annual flooding has a binomial distribution. Let *n* be the number of years of the road's life, *p* be the probability of a spring flood in each year and *x* be the number of spring floods during the *n* years. Then the probability distribution function we require is

$$y = \binom{n}{x} p^x (1-p)^{n-x} \ .$$

In our case $n = 50$ and $p = 0.01$, and we want to know first the probability of there being no floods, i.e. $x = 0$. Inserting these values into the above equation gives us

$$y = \binom{50}{0} 0.01^0 (1-0.01)^{50-0} =$$

$$1 \times 1 \times 0.99^{50} \approx 0.605 \ .$$

So the probability of a flood and therefore of failure of the road is $1-y \approx 0.395$.

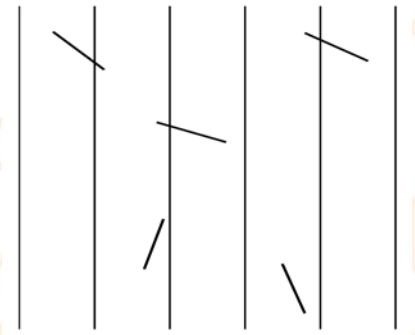
Pedomathemagica

with Gerard Heuvelink

Problem 1 (MEDIUM-HARD)

Pedometricians may sometimes be interested in calculating the probability that a hot spot is detected using a random sample of a given size, or the probability that a boundary between soil types is crossed with a randomly placed transect of a given length. A related famous problem is the following:

Suppose a needle of length L is dropped on a floor made of long wooden planks, whose width is L too (see the example figure below where five needles are randomly dropped on the floor). What is the probability that the needle crosses a boundary between planks?



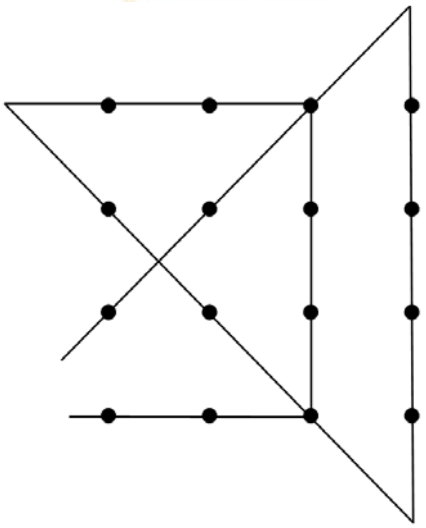
Problem 2 (MEDIUM-HARD)

Not all pedometricians speak always the truth. They are only human. In order to train yourself to be able to deal with these guys (and girls) and ask the right questions, solving the following problem may be useful. It is a somewhat more difficult version of another famous puzzle.

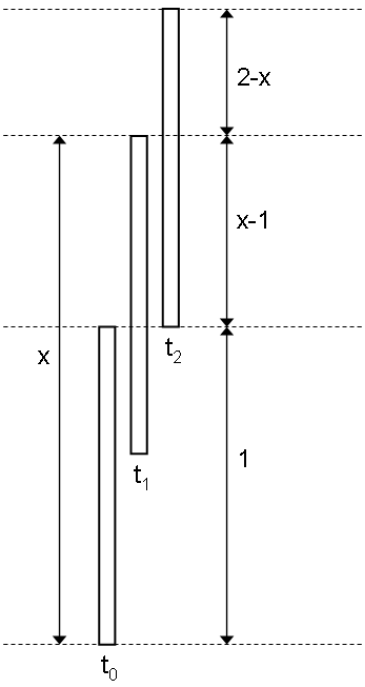
Suppose you walk along a road that splits into two directions at some point. You want to visit your friend who lives in town A nearby but you do not remember whether you should turn left or right. Three people stand at the crossing: one who always speaks the truth, one who always lies, and one who sometimes tells the truth and sometimes lies. You are allowed to ask two questions that must be answered by a yes or a no. You can ask the questions to whoever you choose and you may also ask both questions to the same person. The second question and to whom you pose it may be influenced by the answer to the first. Which questions do you ask, and to whom?

Answer to last issue's quiz

Problem 1 (EASY - MEDIUM)



Problem 2 (MEDIUM)



Let t_0 be the starting time, t_1 the time that the courier reaches the front of the cue and t_2 the final time when the courier has reached the back of the cue. Rectangles represent the position of the cue. The distance run by the courier equals x + x-1. Because both the courier and the cue have a constant speed, the following relationship holds:

$x/(x-1) = (x-1)/(2-x)$.

This yields:

$x(2-x) = (x-1)^2 \rightarrow$
$2x - x^2 = x^2 - 2x + 1 \rightarrow$
$2x^2 - 4x + 1 = 0 \rightarrow$
$x = 1 + \frac{1}{2}\sqrt{2}$

Hence, the distance run by the courier equals $1 + \sqrt{2} = 2.42$ km.