SOIL PROPERTIES AND PEDOMETRICS

R. Murray Lark
Biomathematics and Bioinformatics Division, Rothamsted Research, Harpenden, Hertfordshire, United Kingdom.

Keywords: Bayesian maximum entropy, biometrics, digital soil mapping, fuzzy sets, geostatistics, kriging, multivariate analysis, pedodiversity, pedometrics, precision agriculture, soil inventory, soil monitoring, soil survey, statistics, wavelet transforms.

Contents

Summary

Pedometrics is a branch of soil science, concerned with the quantitative analysis of the variations of soil by mathematical and statistical methods. In this chapter some of the key themes of pedometrics are described, both in terms of its origin as a discipline, the methods used by its practitioners for the core problems of estimating and predicting soil properties, some emerging techniques, and aspects of the discipline that are likely to prove to be important in the future.

1. Introduction

Any properties of the soil that we might measure vary substantially. This is because the soil is a complex material that arises from physical, chemical and biological processes, all of which interact and many of which are non-linear in their effects. Furthermore, these processes operate at different spatial scales, e.g. at continental-scale variation in climate, or at field scale where soils vary in texture over a few tens or hundreds of meters. In the laboratory, soil processes are studied that occur over millimeters, such as the large differences in trace gas emissions driven by variations in the organic content of the soil.

This variation poses problems. If we wish to advise farmers on how best to manage the soil, to advise governments on the scale and extent of soil pollution, or to study the soil so as to understand it better, then it is necessary to characterize the soil quantitatively. This must be done by measuring the soil's properties, but because these vary intensely over different spatial scales the collection and interpretation of soil data is not simple. Pedometrics has arisen as a branch of soil science concerned with the development of mathematical and statistical methods to handle the variation of the soil and to ensure that robust and reliable conclusions can be drawn from complex soil data.

The term *pedometrics* was coined by one of its foremost practitioners, Professor Alex McBratney of the University of Sydney, Australia. The term is formed by analogy with well-established quantitative disciplines such as biometrics, econometrics and environmetrics. An international body of researchers in this area now constitutes the Pedometrics Commission of the International Union of Soil Scientists. The Commission defines pedometrics as *the application of mathematical and statistical methods for the study of the distribution and genesis of soils*. As with all definitions this is subject to criticism, and it is not the intention here to dwell on the definition of pedometrics. The meaning of a word is acquired by learning how it is used, and the goal in this chapter is to illustrate this from the origins, current practice and emerging themes in the discipline. The reader can also learn more from the website of the Commission (http://www.pedometrics.org/) and from special issues of the Journal *Geoderma* which present papers from the Commission's biennial meetings — Vol 62 (1–3), Vol 89 (1–2), Vol 97 (3–4), Vol 103 (1–2), Vol 112 (3–4), Vol 128, (3–4).

The goal of pedometrics is to achieve a better understanding of the soil as a phenomenon that varies over different scales in space and time. This understanding is important, both for improved soil management and for our scientific appreciation of the soil and the systems (agronomic, ecological and hydrological) of which it is a part. For this reason much of pedometrics is concerned with predicting the properties of the soil in space and time, with sampling and monitoring the soil and with modeling the soil's behavior.

2. Origins of Pedometrics

Early in the twentieth century soil scientists and agronomists were concerned to improve the experimental basis of their science, and it was recognized that the extreme variability of the basic material was a significant challenge. Many early studies were focused on the spatial variations of the crop (e.g. Mercer and Hall, 1911), but direct investigations of the soil were also conducted at Davis, e.g. Waynick and Sharp (1919). The soil of two fields was sampled, each on a square grid (80 points) with 20 points added to give finer spaces. The soil was analyzed for total carbon and nitrogen contents. It was concluded that one or a few samples will never be adequate to characterize the soil, different soils will vary to different extents, composite or bulk samples present one way of handling the extreme variability of soils and that the design of experimental work could be made more certain and efficient by proper investigation of the relation between sample size and the precision of resulting estimates. While we might approach some of these problems differently today, these conclusions remain sound and pertinent nearly 90 years on. The sampling scheme for the latter paper is also of interest, since it shows a grasp of the importance of capturing different scales of variation when sampling.

The most important advances in field experimentation and sampling in the early 20th Century were those made by Fisher (1925). One of his central contributions was to put the principle of randomization onto a sound theoretical footing. Randomization means that we select sample points (or allocate experimental treatments) in a random way, so that we can only stipulate in advance the probability that a particular site is sampled. While the idea of randomization was not new, Fisher showed how it can provide rigorous grounds for treating data as random variables, and so for analyzing data, and making inferences on a statistical basis.

One such analysis, which Fisher developed, was the analysis of variance, by which we may treat the mean square deviation of a set of data from their mean (the variance) as the sum of different components, some of which may be experimental effects, and others of which are intrinsic sources of variation in the environment (such as blocks in the experimental design: another of Fisher's innovations). The analysis of variance was soon put to good effect in the study of the soil by Youden and Mehlich (1937). These workers introduced spatially nested sampling designs, which permit an analysis of the variance of data into components associated with variations over different distances in space.

In practice soil scientists often set out to account for soil variation by classifying the soil. Soil profile classes (such as soil series) may then provide the basis for the legend of a soil map, the legend units of which may correspond (primarily) to one class or to an association of classes. When working at smaller cartographic scales the landscape may be classified directly, perhaps from air photographs or satellite images. In both cases field sampling allows to characterize legend units with respect to soil properties, and if the legend is soundly based and reflects sufficient important sources of variation, then these data provide a basis for predicting soil conditions at unvisited sites.

Much important work in pedometrics originated in the attempt to put the above model for spatial prediction of soil properties onto a quantitative basis. This made use of Fisher's analysis of variance. The model is as follows. If $z(\mathbf{x})$ is the value of a soil property at some location that we denote by a vector $\mathbf{x}$, then under the classification model:

$$z(\mathbf{x}) = \mu_i + \varepsilon(\mathbf{x}), \tag{1}$$

where μ_i is the mean value of the variable z for the class i identified at location $\mathbf{x}$ and $\varepsilon(\mathbf{x})$ is a random variable. This has a mean of zero and its variance quantifies the variability of z within the classes, that is to say the amount of variation that the classification fails to explain. Typically, we assume that observations of $\varepsilon(\mathbf{x})$ at different location are independent of each other. This is justified in the analysis of data for which the sample sites are selected according to some randomized design, but not otherwise.

Under random sampling such that the appearance of a particular class i at the nth sample position can be regarded as a random event, the contribution of differences between the soil class means to the overall variance of z is a random effect, and it is associated with a variance of intrinsic interest. The variance of the random term $\varepsilon(\mathbf{x})$ is also of interest. We may use these variance components to test the hypothesis that the class means do not in fact differ from each other, to measure the proportion of variation in z that is explained by the classification and to quantify the expected squared error of a class mean as a predictor of the value of z at a site which is not sampled (Webster and Beckett, 1968).

Another aspect of classification that proved a fruitful area for research in the early days of pedometrics is the automated classification of the soil. Cluster analysis methods have been developed primarily for the analysis of biological data, to try to emulate with an automatic algorithm the process by which the expert taxonomist classifies biological species, and allocates individuals to classes. Soil scientists applied these methods to their data, and found them to be useful for summarizing important relationships between variables and identifying important groupings of observations. Such methods are still used for the exploratory analysis of complex data sets, but it might be questioned how far they have influenced the practice of soil classification.

The disadvantage of classification as a basis for studies of soil variability is immediately obvious. When we use a classification to express soil variation we do this by drawing boundaries on the landscape. Soil surveyors may become very skilled at identifying boundaries on air photographs or in the field, and some research has been devoted to the analysis of patterns of soil boundaries and the uniformity of the units that they define (see also *Soil Geography and Classification*). However, soil boundaries, while informative interpretations of the soil-landscape, do not constitute a realistic model of soil variation. Our expected value of the soil property under Eq. (1) is the class mean, which changes abruptly at a boundary, but soil variation is almost always continuous, not abrupt. We require continuous models of soil variation.

How are we to understand continuous soil variation statistically? One approach is to make a regression of a soil property on spatial co-ordinates, or perhaps a regression on some polynomial function of co-ordinates. This is trend surface analysis, and it has been applied by soil scientists. Its disadvantage historically was the computation of appropriate measures of the uncertainty of its predictions at unsampled sites. In part this was because the classical approach to fitting such trend models required that the error variable be treated as independent. This is implausible. There was a need for new developments.

These came into pedometrics through two routes, first, through a reading of Yates's (1948) paper on systematic sampling. Yates, a colleague of Fisher, wrote a discussion on how one should analyze data which had not been collected by random sampling, but rather at regular intervals (as on a transect or grid). Without random sampling we may not treat our data as random variables. Yates suggested the computation of estimates of the variance from local segments of such a transect, and this idea inspired Beckett and co-workers to compute variances of soil properties within different-sized intervals of a transect, and to plot variance against the size of the interval. Such graphs will show something about the spatial variation of the soil properties studied, and may be related to different sources of soil variation. Beckett and Bie (1976) illustrated the ideas. Tellingly, this paper recognized explicitly the link between these graphs of variance and the variogram of regionalized variable theory. This may be the earliest reference to geostatistics in the soils literature.

Second, Webster and Cuanalo (1975) applied statistical ideas from time series analysis to the analysis of soil data sampled on a regular spatial transect. In particular they computed and plotted correlograms. In time series analysis the correlogram is a function that expresses the correlation of an observation of a variable at time t with an observation of the same variable at time $t + h$ where h is a lag interval. In general we expect the correlation to decay with increasing h expressing the intuition that the two values of the variable will generally be more similar the closer they are in time. Webster and Cuanalo (1975) discovered that the same result held for soil data sampled in space. This was important, because the correlogram and the semi-variogram of geostatistics are very closely related.

Given the extent of this *preparatio evangelica* it is not surprising that the geostatistical methods, still primarily used in mineral exploration, were discovered by soil scientists shortly afterwards. The first published study using geostatistical analysis to predict soil properties was by Burgess and Webster (1980). This ushered in the modern era of pedometrics, and it is to these geostatistical methods that the next section will refer.

3. Mathematical and Statistical Techniques used by Pedometricians

3.1. Well-established Workhorses

In this section three particular groups of techniques used by pedometricians are discussed. Geostatistics is perhaps the most widely used technology; multivariate methods are used to explore relationships between variables; fuzzy sets and fuzzy logic have been adopted for various problems where the conceptual interpretation of hard data values has a non-statistical uncertainty

-
-
-

Bibliography

Beckett, P.H.T. and Bie S.W. (1976). *Reconnaissance for Soil Survey. II. Pre-survey Estimates of the Intricacy of the Soil Pattern*. Journal of Soil Science, 26, 101–110. [An early study on soil variability in Australia using graphs that are precursors of the semivariogram].

Borůvka, L O., Vacek, O. and Jehlička, J. (2005). *Principal Component Analysis as a Tool to Indicate the Origin of Potentially Toxic Elements in Soils*. Geoderma, 128: 289–300. [A recent study on the distribution of toxic elements in soils using principal components to elucidate the effects of pollution and natural sources].

Burgess, T.M. and Webster, R., (1980). *Optimal Interpolation and Isarithmic Mapping of Soil Properties. I: The Semi-variogram and Punctual Kriging*. Journal of Soil Science, 31: 315-331. [The first paper in soil science literature to report an explicitly geostatistical analysis].

Burrough, P.A. (1989). *Fuzzy Mathematic Methods for Soil Survey and Land Evaluation*. Journal of Soil Science, 40: 447–492. [An early review of the scope for using fuzzy logic for problems in soil science].

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh. [Fisher's primary textbook on statistical methods, later supplemented by books on inference and experimental design].

Lark, R.M. (2005). *Analysis of Complex Soil Variation using Wavelets*. In: Grunwald, S. (ed). Environmental Soil Landscape Modeling: Geographic Information Technologies and Pedometrics. CRC Press, New York [Textbook chapter describing the use of wavelets in pedometrics].

Lark, R.M. and Cullis, B.R. (2004). *Model-based Analysis using REML for Inference from Systematically Sampled Data on Soil.* European Journal of Soil Science, 55: 799–813. [An article in the International Association for Mathematical Geology's Series for Students introducing linear mixed model methodology].

Lark, R.M, Bellamy, P.H. and Rawlins, B.G. (2006). *Spatio-temporal Variability of Some Metal Concentrations in the Soil of Eastern England, and Implications for Soil Monitoring.* Geoderma. 133, 363–379. [An application of methods proposed by Papritz and Flühler for the geostatistical estimation of change in soil properties].

McBratney, A.B. (1992). *On Variation, Uncertainty and Informatics in Environmental Soil Management.* Australian Journal of Soil Research, 30: 913–935. [A paper that introduced the concept of pedodiversity].

McBratney, A.B. and De Gruijter, J.J. (1992). *A Continuum Approach to Soil Classification by Modified Fuzzy k-means with Extragrades.* J. Soil Sci., 43: 159-175. [A fuzzy classification procedure for soil survey].

McBratney, A.B, Mendonça Santos, M.L. and Minasny, B. (2003). *On Digital Soil Mapping*. Geoderma, 117: 3-52.[A review of the problem of digital soil mapping and its methodologies].

McBratney, A.B., Webster, R. and Burgess, T.M. (1981). *The Design of Optimal Sampling Schemes for Local Estimation and Mapping of Regionalised Variables. I. Theory and Methods.* Computers and Geosciences, 7: 331–334. [A methodology for planning geostatistical surveys].

McBratney, A.B., Minasny, B., Cattle, S.R. and Vervoort, R.W. (2002). *From Pedotransfer Functions to Soil Inference Systems.* Geoderma, 109: 41–73. [A discussion and expansion of the pedotransfer function concept].

McKenzie, N. and Gallant, J. (2007). Digital Soil Mapping with Improved Environmental Predictors and Models of Pedogenesis. In: P. Lagacherie, A. B. McBratney, and M. Voltz (Eds.) 'Digital Soil Mapping: An Introductory Perspective'. Developments in Soil Science 31, Elsevier, Amsterdam. [An example of digital soil mapping based on a conceptual model of the soil landscape]

Mercer, W.B. and Hall, A.D. (1911). *The Experimental Error of Field Trials*. Journal of Agricultural Science, 4: 107–132. [The classical uniformity trial, an experiment to quantify the uncertainty in experimental measurements of crop responses].

Minasny, B and McBratney, A.B. (2001). *A Rudimentary Mechanistic Model for Soil Formation and*

Landscape Development: II. A Two-dimensional Model incorporating Chemical Weathering. Geoderma, 103: 161–179. [A pioneering study on the mechanistic modeling of soil landscapes].

Papritz, A. and Flühler, H. (1994). *Temporal Change of Spatially Autocorrelated Soil Properties: Optimal Estimation by Cokriging*. Geoderma, 62: 29–43. [A geostatistical method for soil monitoring].

Savelieva, E., Demyanov, V. Kanevski, M., Serre, M. and Christakos, G. (2005). *BME-based Uncertainty Assessment of the Chernobyl Fallout.* Geoderma, 128: 312–324. [The use of Bayesian maximum entropy to integrate different kinds of information].

Viscarra Rossel, R.A., Goovaerts, P. and McBratney, A.B., (2001). *Assessment of the Production and Economic Risks of Site-specific Liming using Geostatistical Uncertainty Modeling.* Environmetrics, 12: 699-711. [Paper explaining the use of pedometric methods to quantify the uncertainty underlying production decisions in precision agriculture].

Waynick, D.D. and Sharp, L.T. (1919). *Variability in Soils and its Significance to Past and Future Soil Investigations. II. Variations in Nitrogen and Carbon in Field Soils and Their Relation to the Accuracy of Field Trials.* University of California Publications in Agricultural Sciences, 4: 121–139. [An early example of detailed studies on soil variation].

Webster, R. and Beckett, P.H.T. (1968). *Quality and Usefulness of Soil Maps.* Nature, London, 219: 680–682. [Quantitative assessment of the utility of soil maps].

Webster, R. and Cuanalo de la C., H.E. (1975). *Soil Transect Correlograms of North Oxfordshire and Their Interpretation.* Journal of Soil Science, 26: 176–194. [Application of time series methods to identify spatial variation of the soil].

Webster, R. and Oliver, N. (1990). *Statistical Methods in Soil and Land Resource Surveys*. Oxford, Oxford University Press. [A standard text in pedometrics]

Yates F. (1948). *Systematic Sampling*. Phil. Trans. Roy. Soc. London, Ser. A.241, 345–377. [A classical paper on the quantification of variability from non-random observations].

Youden, W.J. and Mehlich, A. (1937). *Selection of Efficient Methods for Soil Sampling*. Contributions of the Boyce Thompson Institute for Plant Research, 9: 59–70. [A classical paper on the use of spatially nested sampling].

Biographical Sketch

Murray Lark is leader of the Environmetrics group at Rothamsted Research, England, and Visiting Professor in the Department of Soil Science at the University of Reading, England. He holds the degree of D.Phil. in soil science (1991) from the University of Oxford.

He is active in research in spatial analysis of environmental data with particular emphasis on the development of statistical and mathematical methodology. Most of his research is oriented towards problems in soil science. His group's research covers applied problems in agriculture and environmental management. He has worked on sampling methodologies for soil investigations and management at scales from the field to national. He also studies more basic problems to do with spatial analysis and prediction, particularly for complex variables.