Geoderma 185–186 (2012) 102–109

Contents lists available at SciVerse ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Distinguishing spatially correlated random variation in soil from a 'pure nugget' process

R.M. Lark *

British Geological Survey, Keyworth, Nottinghamshire NG12 5GG, UK

ARTICLE INFO

Article history:
Received 11 May 2011
Received in revised form 20 February 2012
Accepted 28 March 2012
Available online 25 May 2012

Keywords:
Linear mixed model
Nugget
Spatial dependence
Log likelihood-ratio

ABSTRACT

In most spatial analysis of soil variation it is assumed that the random variation not captured by fixed effects (class means or continuous covariates) is spatially dependent. It is proposed that this should be tested formally, both to justify the kriging component in subsequent spatial prediction and as evidence of the extent to which the included fixed effects have succeeded in accounting for soil variation that is spatially dependent at the scales resolved by the soil sampling. A formal test is possible by computing the log ratio of the likelihoods for a full model with spatially dependent random effects and a null model which is pure nugget. It is shown that the sampling distribution of the log likelihood-ratio under the null model is not $\chi^2(p)$ where p is the number of additional random effects parameters in the model with spatial dependence. This is because, while the null model is nested in the full model, parameters of the full model take bounding values in the null case. The sampling distribution may be computed by Monte Carlo simulations. It is shown that the power to reject the null model by the log likelihood-ratio test depends on the importance of the nugget effect in the underlying model, and on the sampling scheme. In many circumstances it may be hard to demonstrate spatial dependence. The recommended procedure was applied to some data on the organic carbon content of the topsoil and subsoil of a field in England. This was modelled either with the overall mean the only fixed effects, or with separate means for different soil map units as fixed effects. There was significant evidence for spatial dependence in the random effects at both depths when the overall mean was the only fixed effect. When map unit means were used as fixed effects there was significant, though weaker, spatial dependence in the topsoil, but the null model could not be rejected for the subsoil. This has implications for any further sampling to map organic carbon in the subsoil.

1. Introduction

In early statistical studies on soil variability and prediction from soil surveys (Webster and Beckett, 1968) a simple statistical model was used, implicitly or explicitly. Under this model the value of a soil property at a set of n locations, $\mathbf{S}$, is a random variate, $\mathbf{Y}(\mathbf{S})$, where

$$\mathbf{Y}(\mathbf{S}) = \mathbf{X}\tau + \varepsilon, \quad (1)$$

$\mathbf{X}$ is an $n \times p$ design matrix which associates each location in the set with one of p soil map units, τ is a $p \times 1$ vector of soil map unit means and ε is an independently and identically distributed (iid) random variate with mean zero and variance σ^2. Note that this is a fixed effects model, in which the soil map units are included because the scientist is interested in them; and, having identified them, designs an appropriate scheme on which to sample them. The resulting data are then analysed according to this model. The resulting estimated map unit means, $\hat{\tau}$, and estimate of the variance of ε, s^2 can then provide a prediction of the value of the soil property at an unsampled site, given the map unit that is delineated there, and an associated prediction error variance (e.g. Leenhardt et al., 1994).

This statistical model is entirely valid, provided that the assumption that ε is iid is justified by an appropriately randomised sampling scheme (de Gruijter et al., 2006). This is the design-based approach. However, there may be benefits for spatial prediction of soil properties if the spatial dependence of soil variation within map units is modelled statistically, and this is essential where the sample sites have not been selected by an appropriately randomised design (Lark and Cullis, 2004). The model-based approach, which encompasses geostatistical prediction, has been enthusiastically adopted by soil scientists since the seminal work of Burgess and Webster (1980).

In most early soil geostatistics all the soil variation was treated as an autocorrelated random process, but it has been recognised that categorical information, such as conventional soil surveys, and continuous covariates, such as remote sensor measurements, can be combined with geostatistical modelling of the remaining spatial variation. This is the basis of much contemporary work on digital soil mapping (McBratney et al., 2003). Lark et al. (2006) showed how the model in Eq. (1), extended to a linear mixed model, generalised

* Tel.: +44 115 9363026.
E-mail address: mlark@nerc.ac.uk.

0016-7061/$ – see front matter
doi:10.1016/j.geoderma.2012.03.029

classical geostatistics accordingly. Now the soil property is modelled by

$$\mathbf{Y}(\mathbf{S}) = \mathbf{X}\tau + \mathbf{Z}\mathbf{u} + \varepsilon, \tag{2}$$

where $\mathbf{Z}$ is a design matrix which associates each observation with a random variable in $\mathbf{u}$. (Note that $\mathbf{Z}$ is $n \times n$ in the usual case where there is no more than one observation at any location in space). The random variable $\mathbf{u}$ is spatially correlated, so it has a correlation matrix $\mathbf{G}$ with the elements on the main diagonal all equal to 1, and off-diagonal elements $\{i, j\}$ taking, in general, non-zero values that depend, under assumptions of second order stationarity, on a parametric function $C(\mathbf{x_i} - \mathbf{x_j}|\psi)$, where the vector $\mathbf{x_i}$ is the location of the ith observation and ψ is a vector of autocorrelation parameters, such as the spatial parameter a of the well-known exponential function

$$C_{\exp}(\mathbf{x_i} - \mathbf{x_j}|\mathrm{a}) = \exp\left\{-\frac{\left|\mathbf{x_i} - \mathbf{x_j}\right|}{\mathrm{a}}\right\}. \tag{3}$$

It is assumed that the random components have the multivariate normal joint distribution

$$\begin{bmatrix}\mathbf{u}\\ \varepsilon\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2\xi\mathbf{G} & 0\\ 0 & \sigma^2\mathbf{I}\end{bmatrix}\right), \tag{4}$$

where σ^2 is the variance of ε and ξ is the ratio of the variance of $\mathbf{u}$ to that of ε. The assumption of an underlying multivariate normal random function is implicit in all standard likelihood estimators, such as the one used in this paper. The data analyst should examine the data to ensure that this is a plausible assumption, perhaps after transformation, but it cannot be absolutely verified. However, it has been observed that likelihood estimators are robust to departures from normality, and that they are optimal estimators by an entropy criterion even in cases where strict normality does not hold (Lark, 2000).

It is worth reflecting on the persistence of the iid component in the linear mixed model, ε in Eqs. (2) and (4). This 'nugget' variability represents all variation that does not appear spatially correlated over the intervals $\mathbf{x_i} - \mathbf{x_j}$ represented in the data set. This may include measurement error, but the nugget variance can be larger than the known measurement error variance (e.g. Rawlins et al., 2003), indicating that there are substantial sources of variation in the soil operating at larger spatial frequencies (finer scales) than the sampling scheme can resolve.

It is a common assumption in soil geostatistics that the soil exhibits spatially dependent variation at scales bounded at the top of the frequency range (fine scales) by the resolution of the sampling network and not accounted for by the fixed effects in Eq. (2). This assumption is generally reasonable, but it is the contention of this paper that it should be formally examined as a matter of course. We can think of the fixed effects in Eq. (2) as representing our soil science knowledge about variable y. This may be the *generalised soil knowledge* of a soil survey if the fixed effects include map units, *specific soil knowledge* if the fixed effect is the prediction from, for example, a process model (e.g. Stacey et al., 2006) or *tacit soil knowledge* that some covariate should be correlated with the variable —Rawlins et al. (2009) used elevation and gamma ray emissions as covariates in a model for soil organic carbon. As we increase the knowledge content of our statistical model so the random effects will become relatively less important. We may also expect that, as our soil knowledge becomes increasingly comprehensive, and as proximal remote sensors become increasingly well-tailored to measuring soil properties, and increase in resolution, so the extent to which we can explain the spatially correlated component of the variation of soil properties should increase. This is illustrated, for example, by Rawlins et al. (2009) who showed the variogram function for the random effects in a model of soil organic carbon decreasing both in sill variance and in the range as more terms were added to the fixed effects part of the model. In short, as we increase the content and sophistication of the soil knowledge of our statistical models so we should examine the possibility that the unexplained variation will not show spatial dependence within the frequency range resolved by the sampling.

A further reason for considering the evidence for spatial correlation in the random component of a mixed model is the subsequent use of the model for spatial prediction. If we assume spatial dependence then the best linear unbiased predictor (BLUP) of the variable at an unsampled site includes a component that is a kriging prediction of the random term. Rather than implement this automatically, we would be best advised to weigh the evidence for spatial dependence, and select an approach to prediction accordingly.

This paper considers the problem of testing the significance of spatial dependence in the random effects of a linear mixed model. The problems are illustrated by simulation, and then the approach is demonstrated in a case study with some data on soil organic carbon.

2. Theory and simulations

2.1. The linear mixed model and model comparisons in the standard case

The linear mixed model in Eq. (2) is fitted to data, $\mathbf{y}$, by finding an estimate of the random effects parameters, $\theta = [\sigma^2, \xi, \psi]$ that maximises the residual likelihood:

$$\ell_{\mathrm{R}}\left(\theta|\mathbf{y}\right) = -\frac{1}{2}\{\log|\mathbf{H}| + \log|\mathbf{X}^{\mathrm{T}}\mathbf{H}\mathbf{X}| + (\mathrm{n}-\mathrm{p})\sigma^2 + \frac{1}{\sigma^2}\mathbf{y}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{W}\mathbf{C}^{-1}\mathbf{W}^{\mathrm{T}}\right)\mathbf{y}\}, \tag{5}$$

where $\mathbf{W} = [\mathbf{X}, \mathbf{Z}]$ and $\mathbf{H} = \xi\mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathrm{T}} + \mathbf{I}$.

The REML estimate is preferred because it reduces the bias in ordinary maximum likelihood estimation due to error in the fixed effects estimates. The residual likelihood is the likelihood of a variable that is a generalised filtering of the original data such that its maximisation provides consistent estimates of the random effects parameters that we require, and it is independent of the value of the unknown fixed effects coefficients (Patterson and Thompson, 1971). One consequence, however, is that the residual likelihood values for alternative models for the same data are comparable only if the models have the same fixed effects structure.

Consider a case where we wish to compare two alternative random effects models, with the fixed effects in common. The first model has the exponential covariance function defined in Eq. (3) above. The second is the stable model (Wackernagel, 2003)

$$C_{\mathrm{stable}}(\mathbf{x_i} - \mathbf{x_j}|\mathrm{a}) = \exp\left\{-\frac{\left|\mathbf{x_i} - \mathbf{x_j}\right|^{\alpha}}{\mathrm{a}^{\alpha}}\right\}, \tag{6}$$

where $0 < \alpha \leq 2$. Our question is whether it is appropriate to use the somewhat more complex stable model. The log-likelihood function for this model, $\ell_{\mathrm{R}}(\theta_{\mathrm{stable}}|\mathbf{y})$ is always larger than or equal to that for the simpler model, $\ell_{\mathrm{R}}(\theta_{\exp}|\mathbf{y})$ with one fewer parameter so some other criterion is needed. Formal inference can be based on the log-likelihood ratio statistic

$$L = 2\left(\left\{\ell_{\mathrm{R}}\left(\theta_{\mathrm{stable}}|\mathbf{y}\right) - \ell_{\mathrm{R}}\left(\theta_{\exp}|\mathbf{y}\right)\right\}. \tag{7}$$

Asymptotically under the null model (i.e. when the simpler model holds) this statistic is distributed as χ^2 with degrees of freedom equal