Geoderma 266 (2016) 84–97

Contents lists available at ScienceDirect

Geoderma

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

Changes in the variance of a soil property along a transect, a comparison of a non-stationary linear mixed model and a wavelet transform

R.M. Lark

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

ARTICLE INFO

Article history:
Received 14 October 2015
Received in revised form 25 November 2015
Accepted 7 December 2015
Available online 22 December 2015

Keywords:
Non-stationarity
Wavelets
Linear mixed model
Soil pH

ABSTRACT

The wavelet transform and the linear mixed model with spectral tempering are two methods which have been used to analyse soil data without assumptions of stationarity in the variance. In this paper both methods are compared on a single data set on soil pH where marked changes in parent material are expected to result in non-stationary variability. The two methods both identified the dominant feature of the data, a reduction in the variance of pH over Chalk parent material, and also identified less pronounced effects of other parent material contrasts. However, there were differences between the results which can be attributed to (i) the wavelet transform's analysis on discrete scales, for which local features are resolved with scale-dependent resolution; (ii) differences between the partition of variation into, respectively, smooth or detail components of the wavelet analysis and fixed or random effects of the linear mixed model; (iii) the fact that the identification of changes in the variance is done sequentially for the wavelet transform and simultaneously in the linear mixed model.

1. Introduction

The properties of soil depend on many factors, which vary at a range of spatial scales. As a result soil properties may show substantial spatial variability (Beckett and Webster, 1971), which requires statistical treatment. One statistical model of soil variation which has been widely used is the linear mixed model (LMM) (Lark et al., 2006) which is a generalization of the geostatistical model of regionalized random variables (Webster and Oliver, 2007). In the LMM we treat the variation of a soil property, z, in terms of fixed effects (categorical or continuous covariates), which represent factors that we can understand and measure. The LMM represents the remaining variation with random effects. There are two sets of random effects in a LMM, those which are spatially correlated because they are caused by factors which operate at spatial scales which can be resolved by the sampling used to obtain our data, and an uncorrelated white-noise component (called the nugget variation in geostatistics). The LMM is written

$$\mathbf{z} = \mathbf{X}\boldsymbol{\tau} + \mathbf{u} + \mathbf{e}, \tag{1}$$

where z is an n-vector containing n observations of variable z, $\mathbf{X}$ is a $n \times P$ matrix with n values of each of P fixed effects, $\boldsymbol{\tau}$ is a vector of fixed effects coefficients, $\mathbf{u}$ is a vector of correlated random variables and $\mathbf{e}$ is a vector of independent and identically distributed (iid) random variables, the nugget component. In the LMM we treat the random effects as multivariate normal (after appropriate transformation, if necessary) and of mean zero, which means they are characterized by their $n \times n$ covariance matrix. In the case of the iid random effect, e, the covariance matrix is given by $\sigma_e^2\mathbf{I}_n$ where σ_e^2 is the variance and $\mathbf{I}_n$ is an $n \times n$ identity matrix. In the case of $\mathbf{u}$ the covariance matrix, $\mathbf{C}$, has a more complex structure reflecting the spatial dependence between observations.

Parameters of the LMM must be estimated from data. This provides something of a challenge for the random effects because, treating them as a realization of a multivariate gaussian process means that we have just one realization from which to estimate the covariance parameters. This is solved by stationarity assumptions. In the LMM a common assumption is second order stationarity, whereby the covariance of u at locations $\mathbf{s}_i$ and $\mathbf{s}_j$ is a function only of the separation in space between them: $\mathbf{s}_i - \mathbf{s}_j$. In this study we simplify the model by considering only the distance between the locations, $h = |\mathbf{s}_i - \mathbf{s}_j|$, but directional dependence can be modelled too. One can therefore express $\mathbf{C}$ as the product of a correlation matrix, the entries of which depend only on the distances between the corresponding observations, and a constant variance, σ_u^2. The entries in the correlation matrix can be modelled most generally with a Matérn correlation function (Diggle and Ribeiro, 2007).

$$\rho(h|\nu,\phi) = \frac{(h/\phi)^\nu K_\nu(h/\phi)}{2^{\nu-1}\Gamma(\nu)} \tag{2}$$

E-mail address: mlark@nerc.ac.uk.

http://dx.doi.org/10.1016/j.geoderma.2015.12.002

where $K_\nu(\cdot)$ denotes the modified Bessel function of the second kind of order ν, $\Gamma(\cdot)$ is the gamma function, ϕ is a distance parameter, and ν is a parameter which determines the smoothness of the spatial process. If $\nu = 0.5$ then the Matèrn function is equivalent to the widely-used exponential correlation, larger ν give smoother variation and smaller ν rougher variation. The random effects of the LMM are therefore fully characterized by the set of parameters $\theta = \{\phi, \nu, \sigma_u^2, \sigma_e^2\}$. Estimates of these parameters are best obtained by residual maximum likelihood (REML) as described by Lark et al. (2006).

Under the model outlined above the covariance of the random effects at any two locations depends only on the distance in space between them. This is a necessary assumption to make the model estimable, but it has a cost for pedological plausibility. Consider a transect from the levées of a small river, across braided sediment deposits onto gentle slopes covered with soliflucted material and onto local hilltops with loess over the underlying sandstone. If we examine the clay content of the soil on this transect we will observe trends in the mean, which may be accounted for with appropriate fixed effects in the LMM. However, the LMM also requires the assumption that the variation about this mean is homogeneous across the transect. This seems implausible, given the different processes (alluvial deposition, solifluction, aeolian deposition) causing textural variation, and the different scales at which they operate. The implausibility of the stationarity assumption may undermine the prediction error variances calculated for interpolated values of the clay content (Lark, 2009), more generally the parameters of the random effects do not represent soil variability anywhere on the transect.

The stationarity assumption is one reason why the LMM, as commonly implemented, may often give limited insight into soil variation. This is one reason why soil scientists considered an alternative analysis to examine scale-dependent variation in soil. This is the wavelet transform. The discrete wavelet transform (DWT) is discussed in more detail elsewhere (e.g. Lark and Webster, 1999) and in the theory section below. In short the DWT represents data by a set of coefficients that represent local variability at different spatial scales (discrete intervals of spatial frequency). In the context of the example transect above, wavelet coefficients at some scale may differ in magnitude from one part of the landscape to another, representing changes in the variability of the property of interest.

Recently, attention has been directed to the extension of the LMM to cases where the random effects have a non-stationary covariance. Of particular interest here is the development by Haskard and Lark (2009) of the spectral tempering method proposed by Pintore and Holmes (2004, 2005). This method allows one to model changes in the variance and autocorrelation of a variable as a function of location in space or some other covariate. This is done by considering the empirical spectrum of the data of interest, and modifying it locally to adjust the distribution of variance between spatial frequencies, as well as the absolute variance.

The LMM with spectral tempering and the DWT are different but complementary ways to represent spatial variation without assuming homogeneity of the variability. However, the two analyses have yet to be compared on a common data set. Such a comparison would be of interest. First, variations in the tempering parameter of the LMM with spectral tempering and variations in the relative magnitude of wavelet coefficients for different spatial scales must both reflect the particular spatial heterogeneity of variation of some soil property, and a direct comparison should be instructive about the methods and the insight that they can give into soil variation. Second, if the spectral tempering random effects model in the LMM can, at least in some circumstances, provide information on changes in soil variation comparable with those from the DWT then this could be useful for the interpretative analysis of irregularly sampled data on the soil, where the scope for wavelet analysis is limited (Milne and Lark, 2009).

In this paper I report the analysis of measurements of soil pH on a transect using both the LMM with spectral tempering and the DWT. The data set is selected as an example where distinct pedogenetic domains, with contrasting parent material, give rise to heterogeneous spatial variability. The spectral tempering model and the DWT-based analysis are compared.

2. Theory

2.1. Linear mixed model with spectral tempering

In the introduction I reviewed the LMM as commonly applied to soil variables. The non-stationary form of this model with spectral tempering starts from a stationary covariance matrix, $\mathbf{C}$, for the spatially correlated random term in the model, the random variable U. One may compute the n eigenvectors of $\mathbf{C}$, $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ and corresponding eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_n$, as in a principal components analysis. This provides the basis for what is called the spectral decomposition of the covariance matrix,

$$\begin{aligned}\mathbf{C} &= \sum_{k=1}^{n} \mathbf{v}_k \lambda_k \mathbf{v}_k^{\mathrm{T}} \\ &= \mathbf{V}\mathbf{D}\mathbf{V}^{\mathrm{T}},\end{aligned} \tag{3}$$

where the superscript 'T' denotes the transpose of a matrix or vector, the matrix $\mathbf{V}$ is $n \times n$ with the eigenvectors of $\mathbf{C}$ in its columns, and $\mathbf{D}$ is a matrix with zeros on all but the elements of the main diagonal, which contains the corresponding eigenvalues, ordered from the largest to the smallest.

The early and later eigenvalues correspond to low spatial frequencies (long-range variation) and to high frequencies (short-range variation) respectively (Haskard and Lark, 2009). The eigenvalues $\lambda_k, k = 1, 2, \ldots, n$, therefore constitute an empirical spectrum which describes how the variance of U is partitioned between the spatial frequencies. The empirical spectrum is not obtained directly from data but from a stationary covariance function, and is therefore itself stationary. I refer to this as the pre-tempering spectrum. Tempering is a method to adjust the spectrum locally; for example by a relative increase in the early eigenvalues (low spatial frequencies) in these areas where the variable appears to be smoother than it is elsewhere. Pintore and Holmes (2004) proposed that this is achieved by raising the terms of the pre-tempering spectrum to some positive power η. Where $\eta > 1$ the low-frequency terms in the spectrum are increased relative to the others, while setting a local value of $\eta < 1$ has the opposite effect, which enhances the short-range variation. Of course, if $\eta = 1$ the spectrum is unchanged. The spectrum can be adapted locally by allowing η to vary spatially. This is possible if we can express η as a function of location in space $\eta(\mathbf{s})$. The joint value of η for any two locations is obtained as

$$\eta(\mathbf{s}_i, \mathbf{s}_j) = 0.5\eta(\mathbf{s}_i) + 0.5\eta(\mathbf{s}_j). \tag{4}$$

A modified covariance matrix of U, which is in general non-stationary, $\mathbf{C}^{\mathrm{NS}}$, can then be obtained from the spectral decomposition of the pre-tempering covariance matrix. The (i,j)th element of this matrix is

$$\mathbf{C}_{\mathrm{ij}}^{\mathrm{NS}} = \sum_{k=1}^{n}, [\mathbf{v}_k] \lambda_k^{\eta_{\mathrm{ij}}} [\mathbf{v}_k]_j \tag{5}$$

where $[\mathbf{v}_k]_i$ denotes the ith element of $\mathbf{v}_k$, which corresponds to the ith location $\mathbf{s}_i$ and the term $\eta_{ij} = \eta(\mathbf{s}_i, \mathbf{s}_j)$ is obtained from Eq. (4). Haskard and Lark (2009) showed that $\mathbf{C}^{NS}$ is positive definite for positive values of η and positive definite $\mathbf{C}$. Given this, it is possible for some data set and a set of fixed effects, to compute the residual log-likelihood for some set of parameters that specify $\mathbf{C}^{NS}$, and so, by an appropriate numerical optimization, to find a set of parameter estimates that maximize this (or, equivalently, that minimize the negative residual log-likelihood).