Manuscript prepared for J. Name
with version 3.2 of the LaTeX class copernicus.cls.
Date: 18 June 2010

A linear mixed model, with non-stationary mean and covariance, for soil potassium based on gamma radiometry

K.A. Haskard[1], **B.G. Rawlins**[2], **and R.M. Lark**[1]

[1]Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ, U.K.
[2]British Geological Survey, Keyworth, Nottingham NG12 5GG, U.K.

Abstract. In this paper we present a linear mixed model for the potassium content of soil across a large region of eastern England in which the mean is modelled as a linear function of the passive gamma-ray emissions of the earth surface in the energy interval commonly associated with potassium decay. Non-stationary models are proposed for the random effect, which is the variation not captured by this regression. Specifically, we assume that the local spectrum of the standardized random effect can be obtained by tempering a common (stationary) spectrum, that is to say raising its values to a power, the tempering parameter, which is itself modelled as a linear function of the radiometric data. This allows the 'smoothness' of the random effect to vary locally. In addition the local spatially correlated variance and 'nugget' variance (apparently uncorrelated given the resolution of the sampling) can also be modelled as a function of the radiometric data. Using the radiometric signal as a covariate gave some improvement in the precision of predictions of soil potassium at validation sites. In addition, there was evidence that non-stationary models for the random effect fitted the data better than stationary models, and this difference was statistically significant. Non-stationary models also appeared to describe the error variance of predictions at the validation sites better. Further work is needed on selection among alternative non-stationary models, since simple procedures used here, based on comparing log-likelihood ratios of nested models and the Akaike information criterion for non-nested models, did not identify the model which gave the best account of the prediction error variances at validation sites.

Correspondence to: R.M. Lark
(murray.lark@bbsrc.ac.uk)

1 Introduction

Soil information is costly and relatively sparse, so there is interest in methods to predict soil properties at unsampled sites from a set of sampled data. Generally the precision of such predictions can be improved if covariates, which reflect factors of soil formation, are incorporated into the predictor. There are various ways to do this. One of the most efficient, when direct observations of the target variable are distributed with reasonable coverage over the area of interest, is the empirical best linear unbiased predictor (E-BLUP) based on a linear mixed model (see below) in which the relationship between the target variable and the covariate is expressed as a linear function of the covariate, and the residual variation is expressed as a combination of a spatially correlated random effect (a random variable), and identically and independently distributed random error. The E-BLUP is, in effect, a combination of a regression-type prediction from the covariation and a kriging-type prediction from the random effect. This was discussed in the context of soil information by Lark et al. (2006), and the approach has been applied in various studies (e.g., Chai et al., 2008). In addition to the predictor, a prediction error variance is also computed which provides a measure of the uncertainty of the prediction.

In the standard LMM–E-BLUP approach to spatial prediction, as in all geostatistical methods, there is a necessary assumption of stationarity in the covariance of the random variation. For geostatistical prediction we require the covariance between the random effect at pairs of sites (including sites where we have data and sites where we require predictions). Since our data provide us, in effect, with a single realization of this random variable, the covariances cannot be obtained directly. Some generalization about the random variation is therefore necessary, so as to allow the parameters of the LMM to be estimated. Commonly we assume stationarity in the covariance (second-order stationarity). Under this assumption the variance of the random variable is the same

at any two locations, $\mathbf{s}_i$ and $\mathbf{s}_j$, and the correlation between the variable at the locations is a function of the lag vector between them,

$$\rho(\mathbf{s}_i - \mathbf{s}_j). \quad (1)$$

This assumption makes it possible to model the covariance of the random effect with some appropriate parametric function. The parameters of this function are best estimated by residual maximum likelihood (REML).

However, the assumption of stationarity in the covariance is not usually plausible when applied to properties of the soil, particularly across complex landscapes. It must first be noted that the assumption does not apply to data, but rather to a random function of which it is assumed that the data are a realization. However, behaviour of the data may indicate whether or not the assumption is plausible. When a variable appears substantially more heterogeneous in one part of the landscape than in another, or when the dominant spatial scale of variation in one part of the landscape differs from another, then this casts doubt on the plausibility of stationarity assumptions. This has been discussed by, for example, Voltz and Webster (1990) who examined variograms for soil properties over contrasting Jurassic strata in central England, included in a single data set, and found pronounced differences. Analysis of soil data sets using wavelet transforms (e.g., Lark and Webster, 2001) similarly cast doubt on the plausibility of stationarity assumptions.

Does the failure of this assumption matter? Lark (2009) compared two LMM for a soil data set. In one of these a conventional assumption of stationarity in the variance was made. In the second a non-stationary variance model was fitted (although the underlying autocorrelation was stationary). Lark (2009) showed that predictions made under these two models were very similar, but that the prediction error variances derived from the non-stationary model gave a much better description of the errors in predictions at validation sites. (If the autocorrelation appeared to be non-stationary, then this might affect the predictions themselves).

It would therefore seem worthwhile to attempt to model non-stationary covariation of soil properties as a basis for spatial prediction. For this reason Haskard and Lark (2009) presented a development and case study of the method of spectral tempering proposed by Pintore and Holmes (2004, 2005). This method is described in more detail below. Essentially it allows both the variance and autocorrelation of the property of interest to adapt locally in response to a set of covariates. Thus the variable might appear 'smoother' in some regions than another. Haskard and Lark (2009) also found that the prediction error variances based on this non-stationary model gave a better account of the uncertainty of predictions at validation sites than did a simpler stationary linear mixed model.

A covariate which has been widely used for the prediction of soil properties is passive gamma ray emissions from the earth's surface (McKenzie and Ryan, 1999; Wilford, 2008). These emissions arise from the decay of particular elements in the upper 35 cm or so of the soil. Gamma rays are emitted over an energy spectrum, which can be partitioned into bands dominated by particular decays. Three bands which are widely used in soil studies are those associated with potassium, uranium and thorium. Much of the application of gamma radiometry for prediction of soil properties has taken place in Australia. Here the total airborne radiometric signal – or the ratios between potassium and thorium concentrations – relates to the age of the weathered material at the land surface and provides information on soil texture properties (Taylor et al., 2002). Radiometry has therefore proved a useful tool to map complex patterns of soil variation that arise from erosion and deposition of material over long periods of time.

Rawlins et al. (2007) undertook a statistical analysis of the radiometric potassium signal from the soil surface in a part of eastern England, and showed that its variation could be attributed to a range of sources including the total potassium content of the soil determined from samples collected in the field. This suggested that gamma radiometry is a potentially useful source of soil information in the relatively young soils of the United Kingdom. This paper addresses the following question. If we have a data set on soil potassium as a basis for geostatistical prediction at unsampled sites, can we beneficially use airborne radiometric data to model both the mean and covariance of the target property in an appropriate LMM? By doing so, we may be able to improve both the precision of the predictions of soil potassium content and the validity of the prediction error variances.

2 The statistical model

2.1 The stationary model, estimation and prediction

The analyses in this paper are based on the linear mixed model (LMM) in which our data are assumed to be a realization of a random variate, with a single value at any location,

$$\mathbf{z} = \mathbf{X}\boldsymbol{\tau} + \mathbf{u} + \mathbf{e}, \quad (2)$$

where $\mathbf{z}$ is the $n \times 1$ vector of observed values at locations $\mathbf{s}_1, \mathbf{s}_2, \ldots \mathbf{s}_n$ respectively, $\boldsymbol{\tau}$ is a $t \times 1$ vector of fixed effects, such as to allow for a smooth spatial trend or other external effects, with corresponding $n \times t$ design matrix $\mathbf{X}$, $\mathbf{u}$ is a $n \times 1$ vector of zero-mean random effects and $\mathbf{e}$ is an $n \times 1$ vector of independent random errors, one element for each observation. In the application of the LMM to spatial data the random component $\mathbf{u}$ models the spatially-correlated component of variation and $\mathbf{e}$ is the so-called 'nugget effect' which incorporates uncorrelated measurement error and sources of variation in the target property that are uncorrelated over the shortest interval between observations. We assume that the random components $\mathbf{u}$ and $\mathbf{e}$ are mutually independent. The

variance matrix of $\mathbf{u}$ is $\mathbf{G}$ $(n \times n)$ whose elements depend only on the sample locations under the assumption of a stationary covariance function, with a few parameters. These parameters are variances, and spatial parameters that characterize the spatial correlation function introduced in Equation (1) above. The random vector $\mathbf{e}$ has zero mean and covariance matrix $\mathbf{R}$ $(n \times n)$. The matrix $\mathbf{R}$ is diagonal when $\mathbf{e}$ denotes a nugget effect and $\mathbf{R} = \sigma^2_{\text{nugget}}\mathbf{I}$ when it is assumed that the nugget effect has stationary variance σ^2_{nugget}. The variance matrix of the random vector $\mathbf{z}$ is denoted $\mathbf{H} = \mathbf{G} + \mathbf{R}$.

Because the mean of $\mathbf{z}$ in Equation (2) is given by $\mathbf{X}\boldsymbol{\tau}$ it is not necessarily assumed to be stationary, and the fixed effects could describe pronounced spatial trends. However, in standard applications of LMM to spatial data it is assumed that the covariance of $\mathbf{z}$ is stationary, as described above. In this study we used an isotropic exponential correlation function for the variable $\mathbf{u}$. If the variance of $\mathbf{u}$ is σ^2, then, with a stationary exponential model, element $\{i,j\}$ of the covariance matrix $\mathbf{G}$, which is the covariance between the values of z at the two locations $\mathbf{s}_i$ and $\mathbf{s}_j$, is given by

$$\sigma^2\rho(\| \mathbf{s}_i - \mathbf{s}_j \|;\phi) = \sigma^2 \exp\{(- \| \mathbf{s}_i - \mathbf{s}_j \|)/\phi\}, \quad (3)$$

where ϕ is a distance parameter. Note that in this isotropic model the separation between the two locations is expressed as a (scalar) distance, but the model can be extended to an anisotropic case in two or more dimensions (Haskard et al., 2007). When such a stationary model has been specified then the variance parameters (σ^2_{nugget}, σ^2 and ϕ in this case) can be estimated by residual maximum likelihood (REML), and these estimates can then be used to obtain the best linear unbiased estimator (BLUE) of the fixed effect coefficients — $\boldsymbol{\tau}$ in Equation (2) — by generalized least squares (details are given by Lark et al., 2006).

Once the parameters of the LMM are estimated then the empirical best linear unbiased predictor (E-BLUP) can be computed at unsampled sites where the fixed effects are known. Let there be p such sites, for which the fixed effects are contained in the $p \times t$ design matrix $\mathbf{X}_\text{p}$. We require the $p \times 1$ vector of E-BLUPs which is

$$\tilde{\mathbf{z}}_\text{p} = \mathbf{X}_\text{p}\hat{\boldsymbol{\tau}} + \tilde{\mathbf{u}}_\text{p} + \tilde{\mathbf{e}}_\text{p} \quad (4)$$

where $\hat{\boldsymbol{\tau}}$ is the BLUE of the fixed effects vector $\boldsymbol{\tau}$, $\tilde{\mathbf{u}}_\text{p}$ is the E-BLUP of the spatially-correlated random effects vector $\mathbf{u}_\text{p}$ at the prediction locations, and $\tilde{\mathbf{e}}_\text{p}$ is the E-BLUP of the nugget effect, which is zero at unsampled locations. The E-BLUP $\tilde{\mathbf{u}}_\text{p}$ is given by

$$\begin{aligned}\tilde{\mathbf{u}}_\text{p} &= \mathbf{G}_\text{po}\mathbf{G}^{-1}\tilde{\mathbf{u}} \\ &= \mathbf{G}_\text{po}\mathbf{z}^\text{T}\mathbf{P}\mathbf{z},\end{aligned}$$

where $\mathbf{G}_\text{po} = \text{Cov}\{\mathbf{u}_\text{p}, \mathbf{u}\}$, the elements of which can be computed given the REML estimates of the variance parameters, and $\mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}(\mathbf{X}^\text{T}\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}^\text{T}\mathbf{H}^{-1}$. This shows that the E-BLUP at an unsampled location consists of a regression-type component ($\mathbf{X}_\text{p}\hat{\boldsymbol{\tau}}$) and a kriging-type component ($\tilde{\mathbf{u}}_\text{p}$).

The PEV is the variance of the prediction error, $\text{Var}\{\tilde{\mathbf{z}}_\text{p} - \mathbf{z}_\text{p}\}$, which can be obtained from

$$\begin{aligned}&[\mathbf{X}_\text{p}\,, \mathbf{G}_\text{po}\mathbf{G}^{-1}]\,\mathbf{A}^{-1}[\mathbf{X}_\text{p}\,, \mathbf{G}_\text{po}\mathbf{G}^{-1}]^\text{T} \\ &\quad + \mathbf{G}_\text{pp} - \mathbf{G}_\text{po}\mathbf{G}^{-1}\mathbf{G}_\text{po}^\text{T} + \mathbf{R}_\text{pp}, \end{aligned} \quad (5)$$

where $\mathbf{G}_\text{pp} = \text{Var}\{\mathbf{u}_\text{p}\}$, $\mathbf{R}_\text{pp} = \text{Var}\{\mathbf{e}_\text{p}\}$ (again, obtained from the REML estimates of the variance parameters) and $\mathbf{A}$ is the coefficient matrix from the mixed model equations

$$\mathbf{A} = \begin{bmatrix} \mathbf{X}^\text{T}\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^\text{T}\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^\text{T}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^\text{T}\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}. \quad (6)$$

Again, more detail is provided by Lark et al. (2006), and the reader looking for a more extensive treatment is referred to Stein (1999)

2.2 The non-stationary model

In this paper we use the modified and extended version of the spatially adaptive spectral tempering procedure (Pintore and Holmes, 2004, 2005) that we introduced and described in full elsewhere (Haskard and Lark, 2009). We outline this procedure below.

Consider a case where we are interested in a total of $n_t = n + n_p$ locations, at n of which we have direct observations of our target variable as well as covariates for the fixed effects, and at the other n_p of which we know the values of the covariates for the fixed effects and require predictions of the target variable because it has not been measured there. We may propose an initial stationary variance model for our variable, and from the parameters of this compute a $n_t \times n_t$ covariance matrix, $\mathbf{C}$. A principal components analysis of this matrix yields what is known as its spectral decomposition

$$\mathbf{C} = \sum_{k=1}^{n_t} \mathbf{v}_k \lambda_k \mathbf{v}_k^\text{T}, \quad (7)$$

where $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{n_t}$ are the n_t eigenvectors, and $\lambda_1, \lambda_2, \ldots, \lambda_{n_t}$ are the corresponding eigenvalues. These latter constitute an empirical spectrum of the data. The eigenvectors can be regarded as a *basis* for the data, that is to say we can represent the variable as a weighted sum of the eigenvectors. If we examine the eigenvectors it can be seen that some correspond to high-spatial frequency (short-range) components of the variable where as others account for low frequency components, broader trends. The eigenvalues, as an empirical spectrum, describe how the overall variance of the variable is partitioned between these components. Tempering is a method to modify the form of the spectrum, by raising its components to a power η. Tempering can therefore change both the overall variance of a variable, and its distribution between spatial scales. The latter is equivalent to a modification of the spatial autocorrelation of

a variable. In spectral tempering the power η is modelled as a function of location $\eta(\mathbf{s})$. This allows the covariance of the variable to change with location, to give a non-stationary covariance matrix

$$\mathbf{C}^{\mathrm{NS}}_{i,j} = \sum_{k=1}^{n} [\mathbf{v}_k]_i \lambda_k^{\eta(\mathbf{s}_i,\mathbf{s}_j)} [\mathbf{v}_k]_j, \tag{8}$$

where

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

In our modified spectral tempering procedure the function $\eta(\mathbf{s})$ is used to adapt the local autocorrelation of the variable (we refer to this as the 'smoothness' of the variable), and the spatially correlated variance (we refer to this as the spatial variance) and the nugget variance are modelled separately, also as functions of spatial coordinates. The variance of the random effect at location $\mathbf{s}_i$ under this non-stationary model is given by $\sigma^2\kappa(\mathbf{s}_i)$, where σ^2 is the variance in the initial stationary model. Similarly the nugget variance at location $\mathbf{s}_i$ is given by $\gamma^2(\mathbf{s}_i)$. Clearly these functions must return positive values for all n_t locations. Once some general parametric forms for the functions $\eta(\mathbf{s})$, $\kappa(\mathbf{s})$ and $\gamma^2(\mathbf{s})$ have been proposed the parameters are estimated by REML. The residual log-likelihood can be computed for any proposed set of parameters. First we obtain a provisional non-stationary covariance matrix $\mathbf{C}^{\mathrm{NS}}$, by tempering the empirical spectrum obtained from initial stationary covariance matrix $\mathbf{C}$. This is done with Equation (8). This preliminary matrix is then rescaled to a non-stationary correlation matrix, $\mathbf{B}$, by dividing each element by the square-root of the product of the corresponding elements on the main diagonal, and then the non-stationary matrix of the random effect $\mathbf{G}^{\mathrm{NS}}$ is obtained by

$$\mathbf{G}^{\mathrm{NS}} = \sigma^2 \,\mathrm{diag}(\boldsymbol{\kappa})^{\frac{1}{2}} \mathbf{B}\, \mathrm{diag}(\boldsymbol{\kappa})^{\frac{1}{2}}, \tag{9}$$

where $\boldsymbol{\kappa} = [\kappa(\mathbf{s}_1),\kappa(\mathbf{s}_2),\ldots]^{\mathrm{T}}$ and σ^2 is the initial stationary variance. The matrix $\mathbf{G}^{\mathrm{NS}}$, along with the non-stationary nugget variances in $\boldsymbol{\gamma}^2$, constitutes the non-stationary variance component of a linear mixed model.

The resulting non-stationary variance parameters can be used to compute all the variance matrices required to compute the E-BLUP at the n_p sites where we require predictions of the target variable, and its associated prediction error variance.

3 Soil data and their analysis

3.1 Stationary variance models

Our data are obtained from 890 locations in a region of approximately 2400 km^2 in north-east England, collected by the G-BASE project (Johnson et al., 2005) and described in detail by Rawlins et al. (2007). At each location we had data on total potassium content of the soil, $\mathrm{K}_{\mathrm{soil}}$, (depth 35–50cm). We also had a corresponding value of the radiometric potassium variable, K_{Γ}, extracted from the gamma emission spectrum measured by a 256-channel Picodas PGAM 1000 (Model 6.11) airborne spectroradiometer. The ground footprint corresponding to each measurement was an ellipse with a long axis length of approximately 200 m.

Five cores were taken at the centre and vertices of a square, length 20 m, centred at the nominal location of the data point. The cores were combined and the bulk sample then dried and disaggregated, sieved to pass 150 μm then coned and quartered and subsampled to produce a 50-g subsample which was ground in an agate planetary ball mill. A further subsample was then analysed for total content of a range of elements, including potassium, by wavelength dispersive X-ray fluorescence spectrometry (XRFS), described in detail by Johnson et al. (2005).

We subdivided this data set at random into a set of 222 observations for modelling and 661 at which predictions would be obtained for validation. We restricted the size of the modelling data set to about a quarter of the available values since, if the data from which the predictions are computed were too dense, the prediction error variances would only be sensitive to the spatial covariance of the random effects over the shortest lags. Seven observations were removed because they took extreme values (large and zero). After exploratory analysis we decided to transform both the total potassium content and the gamma potassium values to natural logarithms, since under these transforms an ordinary least squares regression of $\log\mathrm{K}_{\mathrm{soil}}$ on $\log\mathrm{K}_{\Gamma}$ gave residuals with a histogram (see Figure 1) which appears consistent with an assumption of a normally distributed random variable. The exploratory statistics of the residuals are also consistent with an assumption of normality. The coefficient of skewness is 0.13 and the excess kurtosis is only 0.18. Furthermore the mean and median (0.001 and -0.002 respectively) are very similar, and the first and third quartiles (-0.694 and 0.744 respectively) are symmetrical about the median. Figure 2 shows the scatter plot of the log-transformed variables, and Figure 3 shows a classified post-plot of the residuals, i.e. a plot of the coordinates of the sample points with the value of the residuals indicated by the size of the symbols. While an assumption of normality of the residuals is plausible from the histogram (Figure 1), it is apparent from Figure 2 that the scatter of the residuals appears larger when the radiometric potassium variable is small. This is not compatible with the assumption of a stationary variance, although it must be remembered that stationarity is not a property of data, but of the model that we postulate as underlying the data. Such exploratory analysis can therefore only be indicative. Figure 3 is suggestive of spatial dependence in the residuals, since those of similar size appear to be clustered.

Exploratory analysis also showed that there was no pronounced anisotropy in the residuals, so we elected to use an isotropic variance function. We also found that an exponen-

tial stationary covariance function gave the best fit in a linear mixed model in which the fixed effect structure was a linear function of $\log K_\Gamma$. We used this stationary covariance function to obtain the E-BLUP and corresponding PEVs for $\log K_{soil}$ at all the 661 validation sites. This stationary function was also used to provide an initial stationary empirical spectrum for spectral tempering to give a non-stationary variance model.

In addition we fitted a linear mixed model in which the only fixed effect was an overall mean, specifying a stationary exponential covariance function for the random effect. This was also used to compute the corresponding E-BLUP and PEVs of $\log K_{soil}$ at the validation sites. This provides a basis for assessing the utility of K_Γ as a covariate for spatial prediction of K_{soil}.

3.2 Non-stationary variance models

In this study we hypothesized that the covariance of transformed K_{soil}, as well as its local mean, could be predicted from K_Γ. This hypothesis is compatible with our comments on Figure 2 above. We therefore proposed linear functions of $\log K_\Gamma$ to obtain the expressions required for non-stationary variance models: $\eta(\mathbf{s})$, $\kappa(\mathbf{s})$ and $\gamma^2(\mathbf{s})$ so that, for example

$$\eta(\mathbf{s}_i) = a_0 + a_1 \log K_\Gamma(\mathbf{s}_i).$$

The overall variance models are labeled by a three-character code, with the three characters denoting, respectively, smoothness, spatial variance and nugget variance. In each position, 'L' denotes a linear function in $\log K_\Gamma$ (two parameters to estimate), '1' denotes a constant only (one parameter to estimate), and '0' denotes fixing the function at the value of the initial stationary variance model (no parameters to estimate for this term). The option '0' was never considered for the spatial and nugget variance components. When selected for the smoothness it forces the variance model to be exponential with the same distance parameter throughout the region. The spatial variance and nugget variance may still be varied.

Note that the initial stationary model requires an estimate for the nugget variance and for the variance of the spatial variance, so can be seen to require estimation of two variance parameters, corresponding to model-code 011, conditional on the initial stationary variance model. (If we allow the method to estimate a common spatial variance and common nugget variance, with $\eta(\mathbf{s})$ fixed at 1, it should yield those optimal values we obtained when fitting the initial stationary model. However, if $\eta(\mathbf{s})$ is different from 1, whether constant or otherwise, different variance estimates would be optimal).

A fully non-stationary variance model (LLL) would have six parameters, conditional on the initial stationary empirical spectrum, by comparison to a stationary model with two (011). Models of intermediate complexity are also possible, so we require a method for model selection. In our previous paper (Haskard and Lark, 2009) we proposed a procedure to decide whether or not additional variance parameters are justified by the improved fit that they achieve. It is important to note that this is a procedure for comparing alternative variance models, it is assumed that the fixed effects structures of all the models are identical, for example a linear function of $\log K_\Gamma$. In this procedure we start with the full non-stationary variance model (six parameters). The stationary model can be regarded as a particular case of the full non-stationary model, nested within it, in which the parameters take particular values. There are various pathways from the full model to the stationary model through successively simpler models each of which is nested within the (more complex) models above it on the pathway. This is represented in a lattice diagram — see Figure 4, Haskard and Lark (2009). Having estimated the residual log-likelihood for all the models on the lattice, we may then consider in turn each alternative to the full model that is one step down each of the possible pathways. Whether the more complex model is justified can be decided by testing twice the log-ratio of the likelihoods of the two models against χ^2 with N degrees of freedom where the complex model has N more parameters than the simpler. If the null hypothesis is accepted the more complex form of the model is not significantly better than the one a step below it, and the step down the path is justified. By repeating this procedure down all the pathways until the null hypothesis for a particular comparison is rejected ($P < 0.05$), we may obtain a set of candidate models. These are not necessarily nested, if not they can be compared with respect to the Akaike information criterion, twice the number of parameters minus twice the log-residual likelihood, which will be smaller for the more parsimonious model (Akaike, 1973).

At each validation location we therefore had a prediction and PEV from each of: (i) a stationary model in which the only fixed effect is the overall mean, (ii) a stationary model in which the overall mean and coefficients of K_Γ are fixed effects, and (iii) a set of non-stationary models in which the overall mean and coefficients of K_Γ are fixed effects and the variance model has differing degrees of complexity. We computed a set of validation statistics from observations and predictions at these sites. In general if $z(\mathbf{s}_i)$ is the observed value of $\log K_{soil}$ at the ith validation location out of n_p, and $\tilde{Z}(\mathbf{s}_i)$ is the E-BLUP, given some particular LMM with σ_p^2 the corresponding PEV, then the prediction error (PE) is

$$PE = z(\mathbf{s}_i) - \tilde{Z}(\mathbf{s}_i),$$

and the standardized squared prediction error ($SSPE$) is

$$SSPE = \frac{\left\{z(\mathbf{s}_i) - \tilde{Z}(\mathbf{s}_i)\right\}^2}{\sigma_p^2}.$$

We computed the mean values of PE and of PE^2 over the 661 validation sites. As a measure of the quality of predictions we computed what we call the prediction adjusted

coefficient of determination, R^2_{PA} where

$$R^2_{\mathrm{PA}} = 1 - \frac{\overline{PE^2}}{s^2}, \quad (10)$$

where $\overline{PE^2}$ and s^2 are, respectively the mean value of PE^2 and the sample variance of the validation data set. If the predictions accounted for all the variation in the validation data (i.e. they are all exact) then $R^2_{\mathrm{PA}} = 1$. If the predictor is no better than the sample mean then $R^2_{\mathrm{PA}} = 0$, and a very poor predictor (e.g. a biased one) could have $R^2_{\mathrm{PA}} < 0$.

The $SSPE$ is a measure of the validity of the PEV. The expected value is 1 when the PEVs are reliable. However, it has been found (e.g., Lark, 2009) that the median $SSPE$ which is a more robust statistic may be more useful than the mean for assessing PEVs, since it is less affected by a few validation points at which the $SSPE$ is very large or small. Under an assumption of normal prediction errors the expected value of the median $SSPE$ is 0.455.

We computed an empirical distribution of the mean and median $SSPE$ under conditions where the same distribution of observation and validation sites are used and the data are a realization of a known spatial model in which the fixed effects are a linear function of the overall mean and $\log\mathrm{K}_\Gamma$ and the random effects have a stationary distribution with parameters equal to those fitted in our model (0,1,1). We computed 1001 realizations of this process. The central 95% of the distribution of the mean $SSPE$ was 0.881 to 1.126, with median 0.996, while the central 95% of the distribution of the median $SSPE$ was 0.375 to 0.545, with median 0.454.

4 Results

Figure 4 displays a hierarchy of non-stationary variance models that were fitted, with lines connecting models that are nested and therefore for which likelihood-ratio tests can be carried out. P-values are displayed on the connecting lines when they are less than 0.05. On the basis of residual log-likelihood, two candidate models were identified. The first was 01L with constant variance, with correlation fixed at exponential (i.e. not spatially adapting) and, with the nugget variance changing spatially, depending linearly on $\log\mathrm{K}_\Gamma$. The second candidate model was L11, which keeps both the spatial variance and the nugget variance constant but allows the smoothness to adapt spatially as a linear function of $\log\mathrm{K}_\Gamma$. However this required four parameters to be estimated, and while the two models L11 and 01L cannot be compared by a log-likelihood test because they are not nested, on the basis of the Akaike Information Criterion the latter was preferred.

Table 1 shows validation results for the E-BLUPs based on LMMs with stationary variance models, and a selection of cases with non-stationary models. Note first that there was a reduction in the mean PE^2 on using $\log\mathrm{K}_\Gamma$ as a fixed effect for prediction (R^2_{PA} increased from 0.67) but very little difference among the models with this fixed effect with respect to PEV and PEV^2, for most of the models $R^2_{\mathrm{PA}} = 0.74$. In short there was little difference, with respect to the precision of the predictions, between E-BLUPs with $\log\mathrm{K}_\Gamma$ as a fixed effect and stationary and non-stationary variance models, even in non-stationary models where the smoothness (and so the autocorrelation) was adapted. However, there was an effect on the quality of the computed PEV. Note that the median $SSPE$ for the BLUPs from the stationary model (with $\log\mathrm{K}_\Gamma$ as a fixed effect) was smaller than the 2.5 percentile, suggesting that the error variance was overestimated. This is consistent with results of Lark (2009) and Haskard and Lark (2009), who found that the median $SSPE$ of E-BLUPs from stationary models indicated that the PEVs were overestimated. However, some of the non-stationary models for which validation results of the corresponding E-BLUPs are presented, show mean and median $SSPE$ much closer to the expected values.

5 Discussion and conclusions

These results show that improvements in the precision of spatial prediction of soil potassium (log-transformed) of about 20% (mean square error) were achieved by using the radiometric potassium signal as a covariate. It is also seen that the PEV of E-BLUPs based on a model with a stationary variance structure, in which this covariate is used, appear to overestimate the uncertainty of the predictions. When E-BLUPs were computed using LMMs based on non-stationary covariance models the PEVs were better as judged by the mean and median $SSPE$. However, it must be noted that the LMM selected on the likelihood criteria and AIC (model 01L), only achieved a small improvement over the stationary model, as judged by the median $SSPE$. The best results on $SSPE$ were obtained with model L11, in which the smoothness was adapted, but this model would not be selected on the fitting criteria alone, since its advantages over 01L with respect to the likelihood were not large enough to justify the inclusion of an additional parameter.

Our results therefore indicate that across this region some model for non-stationarity in the variance is appropriate, if we want to put reliance on the PEV of E-BLUPs. They also suggest that spatially adaptive tempering, with K_Γ as a predictor for the tempering parameter η and local variances, has potential to improve modelling of the variance of soil potassium. Application of our approach to other landscapes would help to determine whether more complex, non-stationary variance models based on suitable covariates can improve prediction of soil properties at unsampled sites.

There are clearly questions for further research, prompted by the fact that the best non-stationary variance model (selected on likelihood criteria) did not give the best PEVs when these were validated on independent data. It may be that the dependence of the non-stationary parameters on the predictor

could be better modelled than by the simple linear functions that we used here. Improving this modelling would require the development of some appropriate method for exploratory analysis of the data. While we might hope that the model selection on the likelihood criteria for the modelling data subset would identify the model which subsequently proves best on the validation data, it isby no means guaranteed. An alternative procedure for model selection might be developed, based on prediction efficacy with a separate validation data set. Alternatively a cross-validation or generalized cross-validation procedure (Marcotte, 1995) could be used for model selection.

To conclude, we have shown that the adapted and extended spectral tempering method that was presented by Haskard and Lark (2009)) can be applied to large two-dimensional data sets. We have shown that the use of gamma radiometric data can improve spatial prediction of a soil property, and that in principle using spectral tempering can improve modelling of the spatially-dependent variance of this property, and so the PEVs of the predictions. Further work remains to be done on the problem of model selection for non-stationary variance structures.

Acknowledgements. This research was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) of the United Kingdom, (grant BB/E001599/1). We are grateful to Sue Welham and Brian Cullis for helpful discussions of the spectral tempering methodology. The data were collected under core grant research funding to the British Geological Survey from the Natural Environment Research Council. This paper is published with the permission of the Executive Director of the British Geological Survey (Natural Environment Research Council). We thank all the staff of the British Geological Survey and volunteers involved in the collection and analysis of soil samples in the G-BASE project, and staff involved in the collection and processing of the radiometric survey data.

References

Akaike, H.: Information theory and an extension of the maximum likelihood principle, in: Petrov, B.N. and Csaki F. (Eds), Second International Symposium on Information Theory, Akademai Kiado, Budapest, pp. 267–281, 1973.

Chai, X., Shen, C., Yuan, X. and Huang, Y.: Spatial prediction of soil organic matter in the presence of different external trends with REML–EBLUP, Geoderma, 148, 159–166, 2008.

Haskard, K.A., Cullis, B.R. and Verbyla, A.P.: Anisotropic Matérn correlation and spatial prediction using REML, J. Agric., Biol. and Env. Stat., 12, 147–160, 2007.

Haskard, K.A. and Lark, R.M.: Modelling non-stationary variance of soil properties by tempering an empirical spectrum, Geoderma, 153, 18–28, 2009.

Johnson, C.C., Breward, N., Ander, E. L. and Ault, L. 2005. G-BASE: Baseline geochemical mapping of Great Britain and Northern Ireland. Geochem-Explor. Env A., 5, 1–13.

Lark, R.M.: Kriging a soil variable with a simple non-stationary variance model, J. Agric. Biol. and Env. Stat., 14, 301–321, 2009.

Lark, R.M. and Webster, R.: Changes in variance and correlation of soil properties with scale and location: analysis using an adapted maximal overlap discrete wavelet transform. Eur. J. Soil Sci., 52, 547–562, 2001.

Lark, R.M., Cullis, B.R. and Welham, S.J.: On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur. J. Soil Sci., 57, 787–799, 2006.

Lark, R.M., Milne, A.E., Addiscott, T.M., Goulding, K.W.T., Webster, C.P. and OFlaherty, S.: Scale- and location-dependent correlation of nitrous oxide emissions with soil properties: an analysis using wavelets, Eur. J. Soil Sci, 55, 611–627, 2004.

Marcotte, D.: Generalized cross-validation for covariance model selection, Math. Geol., 27, 659–672, 1995.

McKenzie, N. J. and Ryan, P. J.: Spatial prediction of soil properties using environmental correlation, Geoderma, 89, 67–94, 1999.

Minasny, B. and McBratney, A.B.: The Matérn function as a general model for soil variograms, Geoderma, 128, 192–207, 2005.

Pintore, A. and Holmes, C.C.: Spatially adaptive non-stationary covariance functions via spatially adaptive spectra, Technical report, Department of Statistics, University of Oxford, Oxford, UK, 2004 Available at: http://www.stats.ox.ac.uk/˜cholmes/Reports/spectral_tempering.pdf

Pintore, A. and Holmes, C.C.: A dimension-reducing approach for spectral tempering using empirical orthogonal functions, in: Leuangthong, O. and Deutsch, C.V. (Eds), Geostatistics Banff 2004. Springer, Dordrecht, pp 1007–1015, 2005.

Rawlins, B.G., Lark, R.M. and Webster, R.: Understanding airborne radiometric survey signals across part of Eastern England, Earth Surface Proc. and Landforms, 32, 1503–1515, 2007.

Stein, M.L.:Interpolation of Spatial Data: Some Theory for Kriging. Springer-Verlag, New York, 1999.

Taylor, M.J., Smettem, K., Pracilio, G. and Verboom, W.: Relationships between soil properties and high-resolution radiometrics, central eastern Wheatbelt, Western Australia, Exploration Geophysics, 33, 95–102, 2002.

Voltz, M. and Webster, R.: A comparison of kriging, cubic splines and classification for predicting soil properties from sample information, J. Soil Sci, 41, 473–490, 1990.

Webster, R. and Oliver, M.A.: Geostatistics for Environmental Scientists, 2nd Edition, John Wiley & Sons, Chichester, 2007.

Whittle, P.: Topographic correlations, power-law covariance functions and diffusion, Biometrika, 49, 305–314, 1962.

Wilford J.: Remote sensing with gamma-ray spectrometry, in: McKenzie N.J., Webster R., Grundy M.J. and Ringrose-Voase A.J. (Eds), Australian Soil and Land Resource Handbook: Guidelines for Surveying Soil and Land Resources, 2nd ed., CSIRO Publishing, Melbourne, 2008.

Figure captions

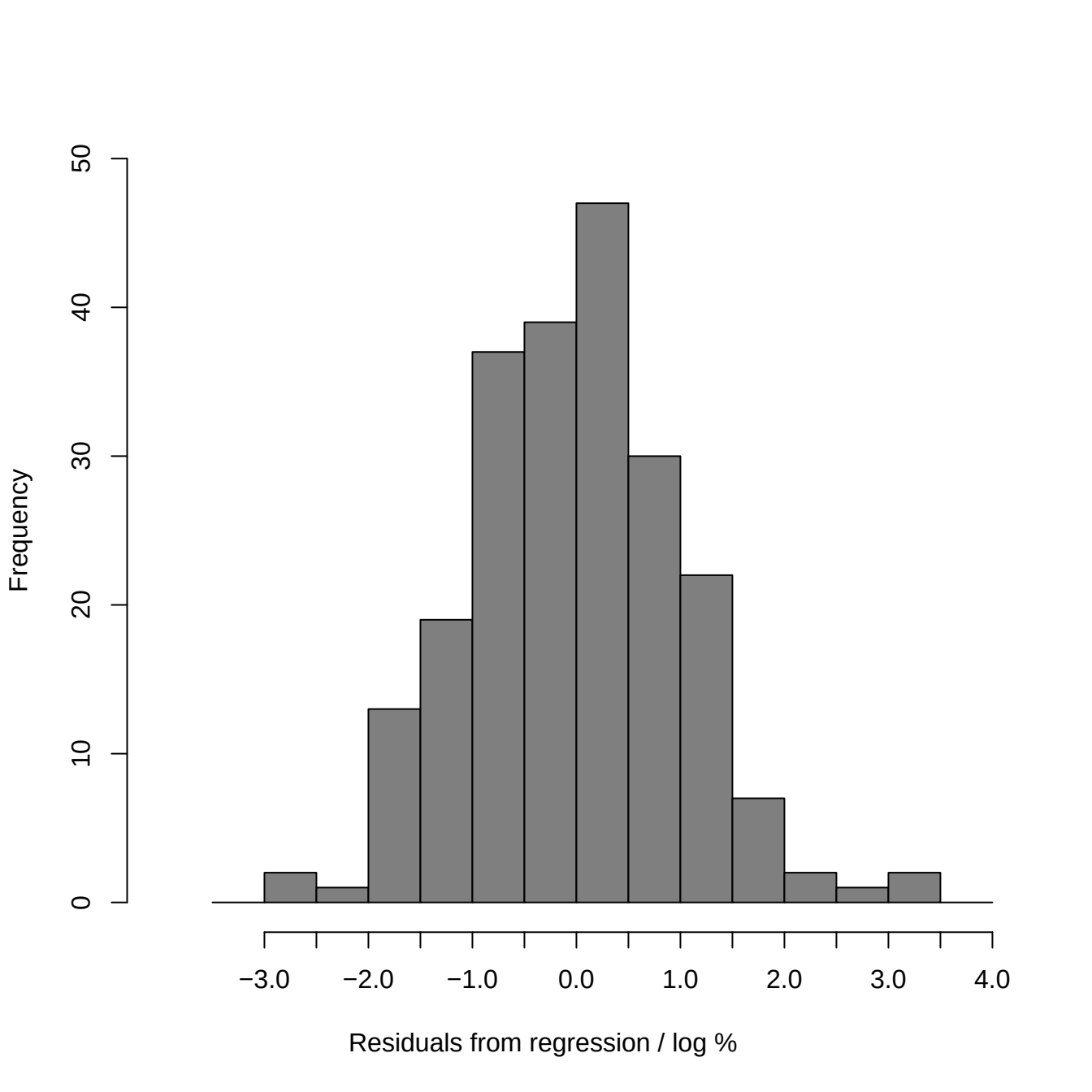
1. Histogram of the residuals of an exploratory regression of $\log K_{soil}$ on $\log K_{\Gamma}$.
2. Scatterplot of $\log K_{soil}$ against $\log K_{\Gamma}$.
3. Postplot of the residuals of the exploratory regression of $\log K_{soil}$ on $\log K_{\Gamma}$.
4. Lattice diagram of alternative non-stationary covariance models.

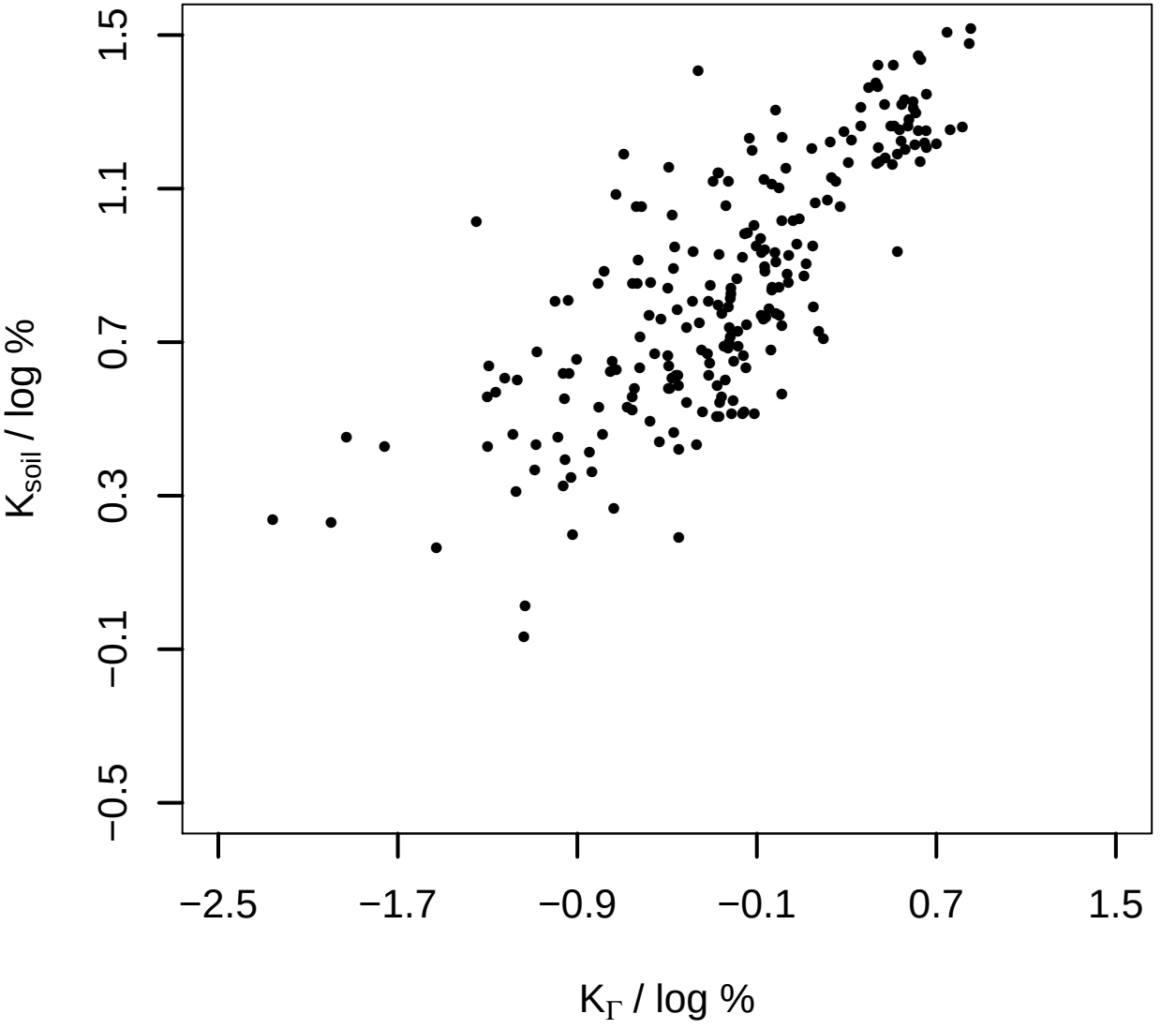
Table 1. Mean and median squared predictions errors and standardized squared prediction errors for selected non-stationary variance models for the BGS potassium data.

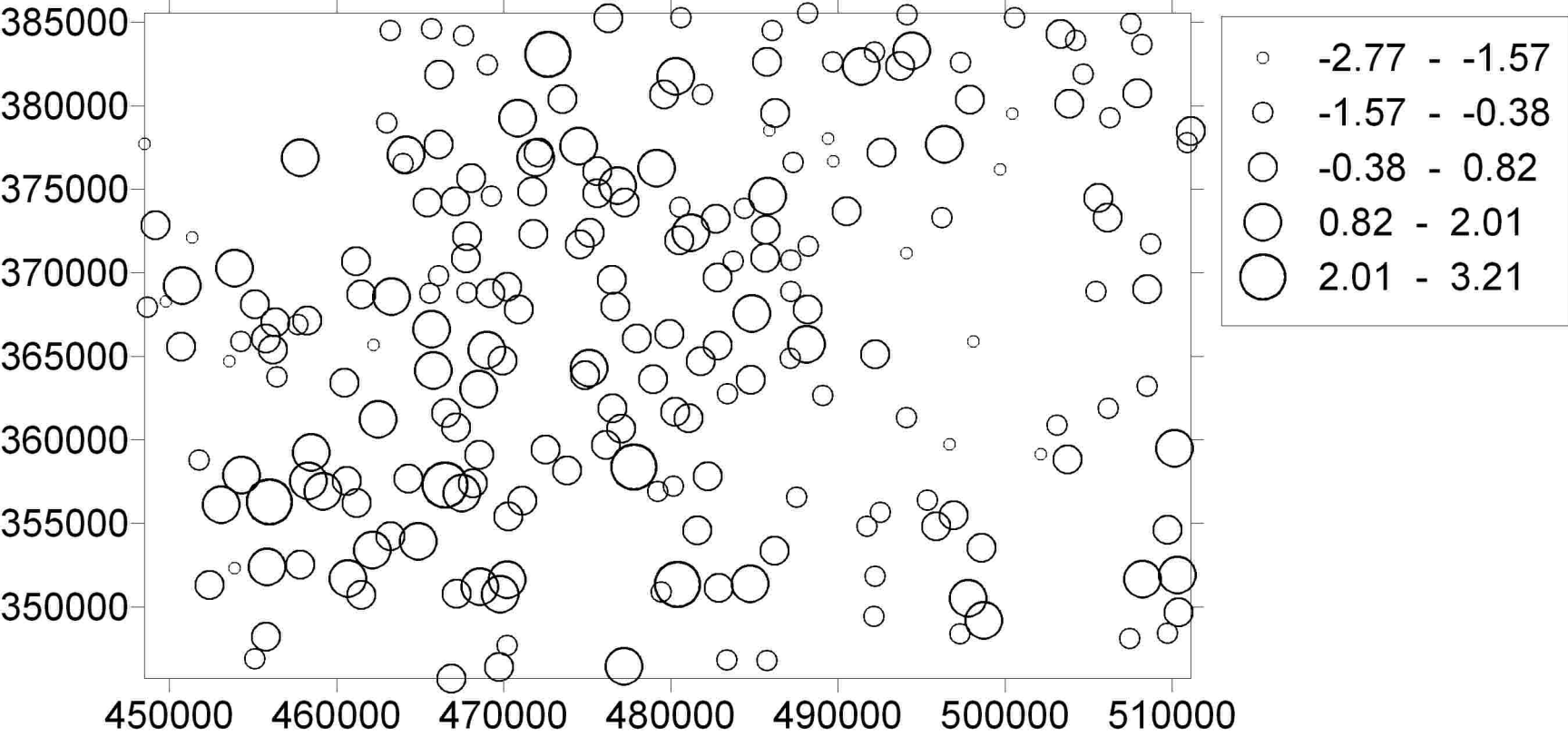
Model	log-residual likelihood	Number of parameters	AIC*	Mean PE	Mean PE^2	R^2_{PA}	Mean $SSPE$	Median $SSPE$
Fixed effect: mean only.								
Stationary covariance				0.144	0.035	0.676	0.930	0.347
Fixed effects: mean and K_Γ.								
011[a]	259.887	2	0	0.127	0.028	0.741	0.884	0.357
01L[b]	269.043	3	-16.31	0.128	0.029	0.731	0.920	0.364
L11[c]	269.838	4	-15.90	0.127	0.028	0.741	0.965	0.390
L1L[d]	270.641	5	-15.51	0.127	0.028	0.741	0.948	0.373
LL1[e]	269.887	5	-14.00	0.127	0.028	0.741	0.963	0.385
LLL[f]	271.877	6	-15.98	0.125	0.028	0.741	0.945	0.365

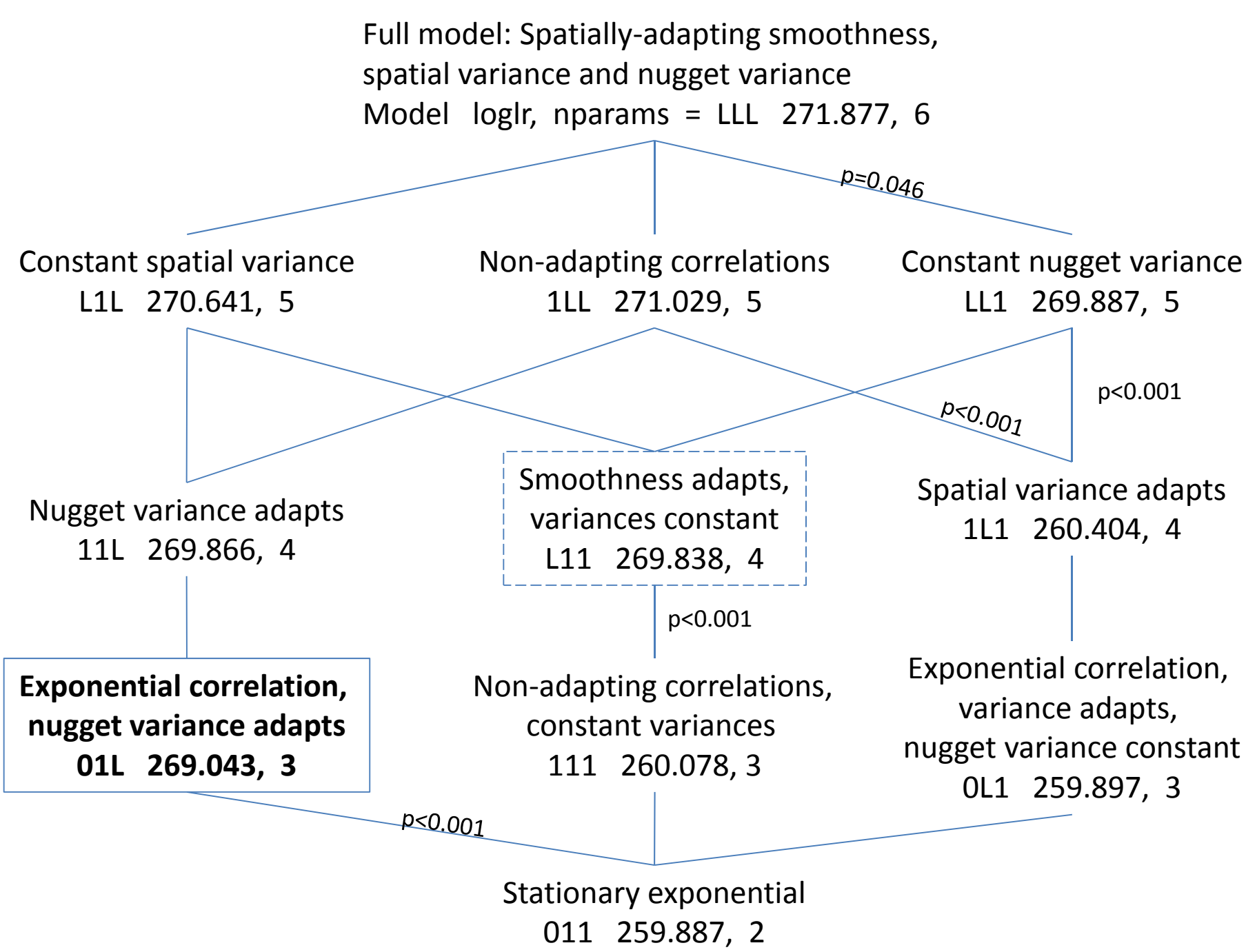
[a]Stationary covariance;[b]Non-stationary covariance, nugget variance adapts; [c]Non-stationary covariance, smoothness adapts; [d]Non-stationary covariance, nugget and smoothness adapt; [e] Non-stationary covariance, spatial variance and smoothness adapt; [f]Non-stationary covariance, nugget and spatial variance and smoothness adapt.

*The AIC is twice the number of parameters minus twice the log-residual likelihood, for ease of comparison the value of AIC for model 011 was subtracted from all values.


Frequency
0
10
20
30
40
50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
4.0
Residuals from regression / log %


1.5
1.1
0.7
0.3
−0.1
−0.5
K_soil / log %
−2.5
−1.7
−0.9
−0.1
0.7
1.5
K_Γ / log %


385000
380000
375000
370000
365000
360000
355000
350000
450000
460000
470000
480000
490000
500000
510000
-2.77 - -1.57
-1.57 - -0.38
-0.38 - 0.82
0.82 - 2.01
2.01 - 3.21


Full model: Spatially-adapting smoothness, spatial variance and nugget variance
Model loglr, nparams = LLL 271.877, 6
p=0.046
Constant spatial variance
L1L 270.641, 5
Non-adapting correlations
1LL 271.029, 5
Constant nugget variance
LL1 269.887, 5
p<0.001
p<0.001
Nugget variance adapts
11L 269.866, 4
Smoothness adapts, variances constant
L11 269.838, 4
Spatial variance adapts
1L1 260.404, 4
p<0.001
Exponential correlation, nugget variance adapts
01L 269.043, 3
Non-adapting correlations, constant variances
111 260.078, 3
Exponential correlation, variance adapts, nugget variance constant
0L1 259.897, 3
p<0.001
Stationary exponential
011 259.887, 2