Available online at www.sciencedirect.com

SCIENCE DIRECT®

Geoderma 128 (2005) 192–207

GEODERMA

www.elsevier.com/locate/geoderma

The Matérn function as a general model for soil variograms

Budiman Minasny*, Alex. B. McBratney

Faculty of Agriculture, Food and Natural Resources, The University of Sydney, NSW 2006, Australia

Available online 23 May 2005

Abstract

The variogram is important in pedometrics for describing and quantifying soil spatial variability. Therefore, it is essential to have a model that can describe various spatial processes and to use appropriate techniques for estimating its parameters. The Matérn model is a generalization of several theoretical variogram models that incorporates a smoothness parameter. We show the flexibility of the Matérn model using simulation and apply the Matérn model to some soil data in Australia. Parameters of the Matérn model were determined by restricted maximum likelihood (REML), and weighted nonlinear least-squares (WNLS) on the empirical variogram. The Matérn model is shown to be flexible and can be used to describe many isotropic spatial soil processes. The REML method fits the local spatial process correctly, however the drawback is the lengthy computation. Meanwhile WNLS fits only the shape of the calculated empirical variogram, and parameters estimated from WNLS can be misleading. From this study, the smoothness parameter of soil data from point measurement appears to be in the range of 0.25–0.50 and can be considered to be a rough spatial process.

1. Introduction

The (semi)variogram is the keystone of geostatistics. Knowledge of the explicit mathematical form of the variogram enables quantitative description of spatial variation (McBratney and Pringle, 1999), prediction of local and regional soil properties by kriging, and the design of optimal sampling schemes. Therefore, a good model and appropriate techniques for estimating the variogram parameters are essential. The variogram is usually obtained from spatial data by the method of moments and subsequent fitting of a theoretical model to the empirical variogram using nonlinear least-squares (Webster and Oliver, 2001). Stein (1999) criticized this procedure because the method of moments for estimating the variogram can be misleading, and commonly used variogram models (spherical, exponential, Gaussian) lack flexibility. Common models assume a certain shape for the local spatial process; in other words the variogram has a predetermined behaviour near its origin. Furthermore, Stein (1999) disapproved of the use of nested models as there are many inherent uncertainties in the spatial data that makes it difficult to estimate all the parameters of a nested model accurately. As most models assume the same local behaviour there is little advantage in using nested models for kriging. As an alternative, Stein (1999) promoted the use of the

* Corresponding author. Tel.: +61 2 9036 9043; fax: +61 2 9351 3706.

E-mail address: budiman@acss.usyd.edu.au (B. Minasny).

0016-7061/$ - see front matter
doi:10.1016/j.geoderma.2005.04.003

Matérn class of models which he named after the Swedish forestry statistician, Bertil Matérn (Matérn, 1960).

The Matérn model has great flexibility for modelling the spatial covariance, and it can model many local spatial processes. Thus, it can be used as a general model of soil variation This will be useful for modelling the local variogram automatically for spatial prediction (Haas, 1990; Walter et al., 2001) as the shape of local variation can be adapted to different parts of a field. Our interest in using it is as a generalized function for automatic local variogram modelling in our software Vesper (Minasny et al., 2002). As far as we know at present the full Matérn model has not been used for soil data. This paper aims to assess the value of the Matérn model for describing spatial variation in soil. We shall describe the Matérn model, exemplify it with the simulation of random fields, explain how to fit it to spatial data, and illustrate its application to soil data.

2. The Matérn model

The Matérn isotropic covariance function is given as (Handcock and Stein, 1993; Stein, 1999):

$$F(h) = \frac{1}{2^{v-1}\Gamma(v)}\left(\frac{h}{r}\right)^{v} K_v\left(\frac{h}{r}\right) \tag{1}$$

where h is the separation distance, K_v is a modified Bessel function of the second kind[1] of order v (Abramowitz and Stegun, 1972), Γ is the gamma function, r is the range or distance parameter ($r > 0$) which measures how quickly the correlations decay with distance, and v is the smoothness parameter ($v > 0$). The model was introduced by Matérn (1960), but was deduced earlier by Whittle (1954, Eq. 65) (constrained to $v = 1$). Ripley (1981, p. 56) attributed this function to Whittle, while Stein (1999, p. 49) believed Matérn (1960) is the first to have recommended this function. Consequently we think this function should be probably called Whittle–Matérn, however we call it the Matérn function in this paper. An alternative parameterization of Eq. (1) has been suggested by Handcock and Wallis (1994):

$$F(h) = \frac{1}{2^{v-1}\Gamma(v)}\left(\frac{2v^{1/2}h}{r}\right)^{v} K_v\left(\frac{2v^{1/2}h}{r}\right), \tag{2}$$

which allows r to be less dependent on v (Stein, 1999).

The covariance matrix $\mathbf{C}$ of spatial data has elements C_{ij}, the covariance between samples at locations x_i and x_j:

$$C_{ij} = c_0\delta_{ij} + c_1 F(h_{ij}). \tag{3}$$

where δ_{ij} is the Kronecker delta: $\delta_{ij}\begin{cases}0 \text{ for } i \neq j\\ 1 \text{ for } i = j\end{cases}$, $F(h)$ is the Matérn model (Eq. (1)), c_0 is the nugget variance, and $c_0 + c_1$ is the sill variance. The corresponding semivariance is

$$\gamma(h) = c_0 + c_1\left(1 - \frac{1}{2^{v-1}\Gamma(v)}\left(\frac{h}{r}\right)^{v} K_v\left(\frac{h}{r}\right)\right). \tag{4}$$

The Matérn model has great flexibility for modelling the spatial covariance compared with the standard models because of its smoothness parameter v. When v is small ($v \to 0$) it implies that the spatial process is rough, and when it is large ($v \to \infty$) that the process is smooth (see Fig. 1). For a finite value of the range parameter, r, the Matérn function represents several bounded models. If v is of the form $(m+1/2)$, where m is a nonnegative integer, then $F(h)$ is the product of a polynomial of degree m in (h/r) and $\exp(-h/r)$:

$v = 1/2$, $F(h) = \exp(-h/r)$ or exponential model,
$v = 3/2$, $F(h) = [(h/r)+1]\exp(-h/r)$,
$v = 5/2$, $F(h) = [(h/r)^2 + 3(h/r) + 3]\exp(-h/r)$

(http://functions.wolfram.com/BesselAiryStruveFunctions/BesselK).

If $v \to \infty$ the Matérn function corresponds to a Gaussian model. The case of $v = 1$ is Whittle's function (Whittle, 1954; Webster and Oliver, 2001). If the range parameter is large ($r \to \infty$ or so-called unbounded), it approximates the power function when $v > 0$, and a log function or de Wijs function (de Wijs, 1951, 1953) when $v \to 0$. Thus the Matérn model can be thought of as a generalization of several theoretical

[1] The modified Bessel functions of the second kind are sometimes called the Basset functions, modified Bessel functions of the third kind, or MacDonald functions (Spanier and Oldham, 1987, p. 499; Weisstein, 1999).

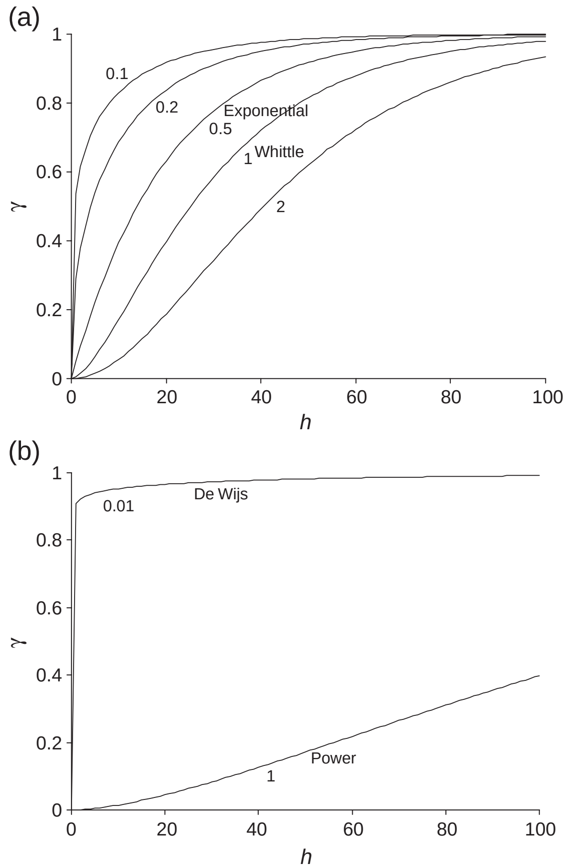

Fig. 1. Plots of the Matérn variogram with varying smoothness parameter v and constant range and sill: (a) $c_0=0$, $c_1=1$ and range $r=20$; (b) $c_0=0$, $c_1=1$ and $r=100$. Numbers beside the curves are the value of the smoothness parameter.

variogram models. These generalizations are summarized in Table 1.

The parameter, v, which controls the smoothness of the spatial process, should be determined from the spatial data. McCullagh and Clifford (2003) investigated the nature of spatial dependence of agricultural crops using the Matérn model. They analysed sixteen sets of crop yield data from field trials at several places, and found that the range parameter, r, tends to be large and the smoothness parameter, v, is small. They concluded from these results that agricultural processes are rough and suggested that the *loi du terroir* model (essentially the de Wijsian model of geostatistics) with $r \to \infty$ and $v \to 0$ should be used for all crops in all seasons.

Heuvelink (2000) pointed out that the examples of Stein (1999) are based on artificial data sets, and that the applicability of the Matérn model to soil data is 'not an unconditional yes'. Heuvelink (2000) argued that the assumptions by Stein pertain to random fields that are differentiable and measurement errors that are negligible. Heuvelink implied that the Matérn model must have a zero nugget variance which is difficult to achieve with real soil data. Nevertheless, Stein (1999) and McCullagh and Clifford (2003) showed that the Matérn model is applicable to observations with error.

Common practice in pedometrics and the earth, agricultural, environmental and biological sciences for variogram estimation is first to calculate the empirical (so-called experimental) variogram by the method of moments (Matheron, 1965), and then to fit a model to the empirical variogram by (weighted) nonlinear least-squares. An alternative method uses maximum likelihood (ML) or restricted maximum likelihood (REML) which estimates parameters of the model directly from the data, on the assumption that it is a multivariate normal distribution. The ML and REML methods are advocated by Stein (1999) but are described as blind (black box) by Chilès and Delfiner (1999). Zimmerman and Zimmerman (1991) compared different methods for estimating the variogram and its use for kriging. They found that for prediction by kriging, the method of moments performed as well as the more computationally demanding ML and REML methods. Lark (2000b) compared the method of moments and ML for estimating variograms; his simulation studies showed that in some circumstances the ML method might be advantageous. However, his comparison of the methods with the data on heavy metal concentrations in soil showed no significant

Table 1
The equivalent of the Matérn covariance function

	Equivalent function	Approximate function
	$r>0$, Bounded	$r \to \infty$, Unbounded
$v \to 0$		De Wijs: $-\log(h)$
$v>0$, integer		$-h^{2v}\log(h)$
$v>0$, non-integer		Power: $-h^{2v}$
$v=0.5$	Exponential: $\exp(-h/\mathrm{r})$	
$v=1$	Whittle: $(h/r)\,K_1(h/r)$	$-h^2\log(h)$
$v \to \infty$	Gaussian: $\exp(-h^2/r^2)$	

difference between them. Stein (1999) demonstrated with simulated data that the method of moments is poor at describing the smooth process. Lophaven et al. (2002) compared several methods of variogram estimation with simulated data and showed that ML and REML performed better than the method of moments. Results from the prediction of salinity levels in the sea showed that the choice of method of variogram estimation does not affect the kriged predictions significantly, but the maximum likelihood method reduces the uncertainty of the predictions.

3. Methods

3.1. Simulation of random fields

To illustrate the behaviour of the Matérn model, we simulated random fields along transects with different smoothness parameters. In addition, we took into account the effect of a nugget variance. A random field with covariance **C** can be simulated using the lower and upper triangular matrix (LU) decomposition method (Deutsch and Journel, 1997). The covariance matrix **C** is decomposed by Cholesky factorization:

$$\mathbf{C} = \mathbf{L}\mathbf{L}^T, \tag{5}$$

where **L** is the lower triangular matrix. To generate a realization of random field **z** with mean μ, matrix **L** is multiplied by the vector, $\boldsymbol{r}_\mathrm{n}$, of independent normal random numbers with zero mean and unit variance:

$$\mathbf{z} = \mathbf{L}\boldsymbol{r}_\mathrm{n} + \mu. \tag{6}$$

3.2. Estimating the Matérn parameter of random fields

We simulated random fields on a square grid of 10 m × 10 m, with nodes separated by 1 m and with a range parameter, r = 5 m. Two types of field were simulated with and without a nugget effect (c_0 = 0.1), both having a sill variance of 1.0:

(1) a rough field with v = 0.1, and
(2) a smooth field with v = 1.0.

We simulated 400 random fields by LU decomposition and estimated the Matérn parameters for each realization.

We compare and discuss two methods for estimating parameters of the Matérn model. The first is the conventional method, where the empirical variogram is calculated and then fitted with the Matérn model by weighted nonlinear least-squares (WNLS). The second method estimates the parameters directly from the data by restricted maximum likelihood (REML). The data and Matlab codes used in this study are available from the authors' website: www.usyd.edu.au/su/agric/acpa/software.

3.3. Empirical variogram and weighted nonlinear least-squares

We calculated the empirical variogram by the method of moments (Matheron, 1965):

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left[z(x_i) - z(x_i + h)\right]^2, \tag{7}$$

where $N(h)$ is the number of pairs of observations separated by distance h. The Matérn model (Eq. (4)) is fitted to the empirical variogram by weighted nonlinear least-squares (WNLS) (Jian et al., 1996) by minimising the objective function:

$$O(\boldsymbol{\theta}) = \sum_i \frac{N_i}{[\sigma(\hat{\gamma}_i)]^2} [\hat{\gamma}_i - \gamma_i(\boldsymbol{\theta})]^2 \tag{8}$$

where N is the number of pairs for lag i, $\hat{\gamma}$ is the empirical variogram, $\gamma(\boldsymbol{\theta})$ is the Matérn model with parameter vector $\boldsymbol{\theta}$ and $\sigma(\hat{\gamma})$ is the standard deviation of the empirical semivariance. The minimization of Eq. (8) is achieved by the Marquardt nonlinear least-squares algorithm as implemented in Matlab by Nielsen (1999).

3.4. Maximum and restricted maximum likelihood method

Another method for estimating the variogram is by the Maximum Likelihood (ML) method which was introduced by Kitanidis (1983, 1985) and Mardia and Marshall (1984). The basis of the method is the assumption that the data follow a multivariate normal distribution. The probability density function of **z** with mean **μ** and covariance matrix **Σ** is given by:

$$p(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{z}-\mu)^T\Sigma^{-1}(\mathbf{z}-\mu)\right]. \tag{9}$$

If we define a linear spatial model:

$$\mathbf{z}(\mathbf{x}) = \mathbf{M}\boldsymbol{\beta} + \varepsilon(\mathbf{x}), \tag{10}$$

where **x** is the vector of spatial coordinates, **z**(**x**) are the observed values, **M** is the trend function, **β** is parameter for the trend, and ε is the error with a mean of zero and a covariance, cov{ε($\mathbf{x}_i$), ε($\mathbf{x}_j$)} = **C**($\mathbf{x}_i$, $\mathbf{x}_j$). Based on Eq. (9), the log-likelihood is:

$$\ell(\theta,\boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}(\mathbf{z}-\mathbf{M}\boldsymbol{\beta})^T\mathbf{C}^{-1}(\mathbf{z}-\mathbf{M}\boldsymbol{\beta}). \tag{11}$$

The vector **θ** can be found by maximizing Eq. (11); however, it requires the estimation of **β**. To simplify the solution, **β** can be estimated from a weighted least-squares solution:

$$\hat{\boldsymbol{\beta}} = \mathbf{W}^{-1}\mathbf{M}^T\mathbf{C}^{-1}\mathbf{z}, \tag{12}$$

where **W** as the inverse of the covariance matrix of estimation errors (Kitanidis, 1987):

$$\mathbf{W} = \mathbf{M}^T\mathbf{C}^{-1}\mathbf{M}. \tag{13}$$

For estimating the variogram parameters only, **M** is set to 1 which means $\hat{\beta}$ is the estimated mean (Lark, 2000b). Thus **θ** can be found by substituting $\hat{\boldsymbol{\beta}}$ into Eq. (11) and maximizing it with an optimization algorithm. This method is described fully in Lark (2000b).

The maximum likelihood method requires the estimation of $\hat{\boldsymbol{\beta}}$ and it can lead to biased estimates of the variances. To circumvent this problem, linear combinations of **z** that do not depend on $\hat{\boldsymbol{\beta}}$ can be formed. This filters out the trend, and so the estimates of θ will not be distorted by an incorrect value for $\hat{\boldsymbol{\beta}}$. This is known as restricted maximum likelihood (REML) estimation, and it transforms the data **z** into stationary data increments, **y** (Kitanidis, 1983):

$$\mathbf{y} = \mathbf{T}\mathbf{z}. \tag{14}$$

The transformation matrix **T** is selected so that the trends defined in **M** can be filtered out:

$$\mathbf{T} = \mathbf{I} - \mathbf{M}\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\mathbf{M}^T \tag{15}$$

where **I** is the identity matrix.

The log-likelihood for REML is:

$$\ell(\theta,\mathbf{y}) = -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\log|\mathbf{W}| - \frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{Q}\mathbf{y} \tag{16}$$

where p is the number of parameters in **M,** and

$$\mathbf{Q} = \mathbf{I} - \mathbf{M}\mathbf{W}^{-1}\mathbf{M}^T\mathbf{C}^{-1} \tag{17}$$

and **W** is defined in Eq. (13). The REML estimate of **β** is given in Eq. (12).

The use of REML for estimating parameters of a spatial covariance function is demonstrated by Kitanidis (1983), and Kitanidis and Lane (1985). Laslett and McBratney (1990) used REML to estimate the variogram of soil pH. The theory is explained in depth in Kitanidis (1987) and Stein (1999).

3.5. REML parameter estimation and its uncertainty

To fit the Matérn model to data using REML, we seek a parameter vector **θ** = [c_0, c_1,r,v] that maximizes Eq. (16) subject to $c_0 \geq 0$ and c_1, r, $v > 0$. The calculations in this paper were implemented in Matlab with maximization of the REML log-likelihood (Eq. (16)) using the *fminsearch* function (MathWorks, 2004) which is a Nelder–Mead simplex algorithm (Nelder and Mead, 1965).

A common problem with the maximum likelihood and REML methods is that the solution may be multimodal (more than one set of parameter values can give a similar value for the log-likelihood) (Warnes and Ripley, 1987; Mardia and Watkins, 1989). Thus a good initial guess is required to obtain a global maximum. However, parameters at the global maximum might not be the optimal solution as it can produce a set of meaningless values (Warnes and Ripley, 1987). Mardia and Watkins (1989) suggested a profile likelihood method for estimating parameters of a spatial covariance model. The procedure of the profile likelihood method will be discussed in Section 4.3 when analysing the soil data.

To assess the uncertainty of the predicted parameters, we calculated the approximate standard error of the parameter estimates using the Fisher information matrix. The latter is the amount of information provided by data about an unknown parameter (Everitt, 2002). Mathematically, it is the expectation of the second partial derivatives of the log-likelihood for parameter $\boldsymbol{\theta}$:

$$\mathbf{F} = \mathbf{E}\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right)^2. \tag{18}$$

The (j,k)-th element of the Fisher information matrix for θ when $\mathbf{y}$ follows a normal distribution $\boldsymbol{N}(\mathbf{0}, \mathbf{C}_\theta)$ (Stein, 1999), can be evaluated as:

$$f_{jk}(\theta) = \frac{1}{2}\mathbf{tr}\left[\mathbf{C}_\theta^{-1}\frac{\partial \mathbf{C}_\theta}{\partial \theta_j}\mathbf{C}_\theta^{-1}\frac{\partial \mathbf{C}_\theta}{\partial \theta_k}\right]. \tag{19}$$

The variance–covariance matrix of the REML estimates of $\boldsymbol{\theta}$ is the inverse of the Fisher information matrix. The derivative of the Matérn covariance matrix $\mathbf{C}_\theta$ with respect to its parameter vector $\boldsymbol{\theta}$ is evaluated numerically using finite differences.

3.6. Soil data

Soil data from Australia were used to test the application of the Matérn model. The data are from the following soil survey studies:

- Thickness of soil A horizon (Pettitt and McBratney, 1993). These data are from a field near Forbes, New South Wales. The sampling was based on a nested sampling design where the field was divided into 18 equal-sized blocks of 160 m^2. In each block there were six sampling points along transect 156 m long. The points were separated by distances of 125 m, 25 m, 5 m, 1 m, and 25 cm that formed a five-fold geometric progression.
- Soil pH in H_2O and $CaCl_2$ of 1 ha area at a research station in Samford, Queensland (Laslett et al., 1987). The samples were collected with the following design:
  - −121 grid samples from the nodes of a square grid of 11 × 11 with a spacing of 10 m,
  - −11 samples adjacent (20 cm separation) to those at the grid nodes,
  - −11 samples located half way between randomly selected pairs of grid sites,
- Soil organic carbon content of the surface soil, from the survey in Edgeroi, New South Wales (McGarry et al., 1989). Samples were obtained from 210 sites on a systematic equilateral triangular grid with an approximate distance of 2.8 km between sites.
- Soil pH from a survey in the Pokolbin area of the lower Hunter Valley, New South Wales, where 128 samples of soil were taken at a depth of 40–50 cm by purposive random sampling with an average separation distance of 200 m.
- Apparent electrical conductivity data in the lower Namoi valley of northern New South Wales. Data from a farm at the northern edge of the Pilliga State Forest were obtained by the mobile EM-31 instrument along a transect of 200 m, with measurements about 1 m apart.

4. Results

4.1. Simulation

One hundred data points were simulated along a transect of 100 m and are shown in Fig. 2. The simulation used the same pseudorandom numbers with varying parameters for the Matérn model. For the bounded model (Fig. 2a–d), the range, r, was set to 20 m, the smoothness parameter v was varied from 0.1 to 2, and the sill (c_0+c_1) was set to 1. For the unbounded model (Fig. 2f,g) r was set to 100 m. The shape of the variogram is similar to the one shown in Fig. 1a. For each simulation a nugget effect of $c_0=0$ is shown as the solid line, and that of $c_0=0.2$ as a dashed line. This figure shows that the Matérn model can model random fields with different degrees of smoothness from rough to smooth. The nugget effect, which represents discontinuous variation, has a marked effect on the simulated field. Adding a nugget component to a rough field does not change the shape of the random field, but doing so to a smoothly varying field makes it appear rough. This raises an important point that needs to be investigated: can the nugget effect and smoothness parameter be estimated simultaneously?

4.2. Estimating the Matérn parameter of random fields

Rough and smooth random fields were simulated with and without a nugget effect of 0.1 and a total sill variance of 1.0; for each field there were 400 realizations. For each realization, the Matérn parameters (c_0, c_1, r, v) were estimated using REML and WNLS using its 'true' value from the simulation as an initial guess. The mean and standard deviation of the Matérn parameters estimated using both methods are given in Table 2. The results indicate that the means of the estimated parameters are quite different from their true values and all the estimated parameters have large standard deviations, especially c_1 and r. On the other hand, if we take the average of the variogram (variance with distance) as calculated from the estimated parameters using REML and WNLS (Figs. 3 and 4), the mean of the estimated variogram corresponds well with the true variogram. This indicates that we cannot simply average the parameter values, as they might not be normally distributed and they are highly correlated with one another. It also implies that the parameters have to be interpreted empirically.

For a rough field, the WNLS method tends to overestimate the variance at short distances whereas REML can estimate the variogram reasonably well (Fig. 3). If we note the variance at short distances (Fig. 3b) we can see that REML also overestimates the nugget variance and the estimate has a large uncertainty. This suggests a potential problem for soil data which are usually 'noisy' and have a large nugget

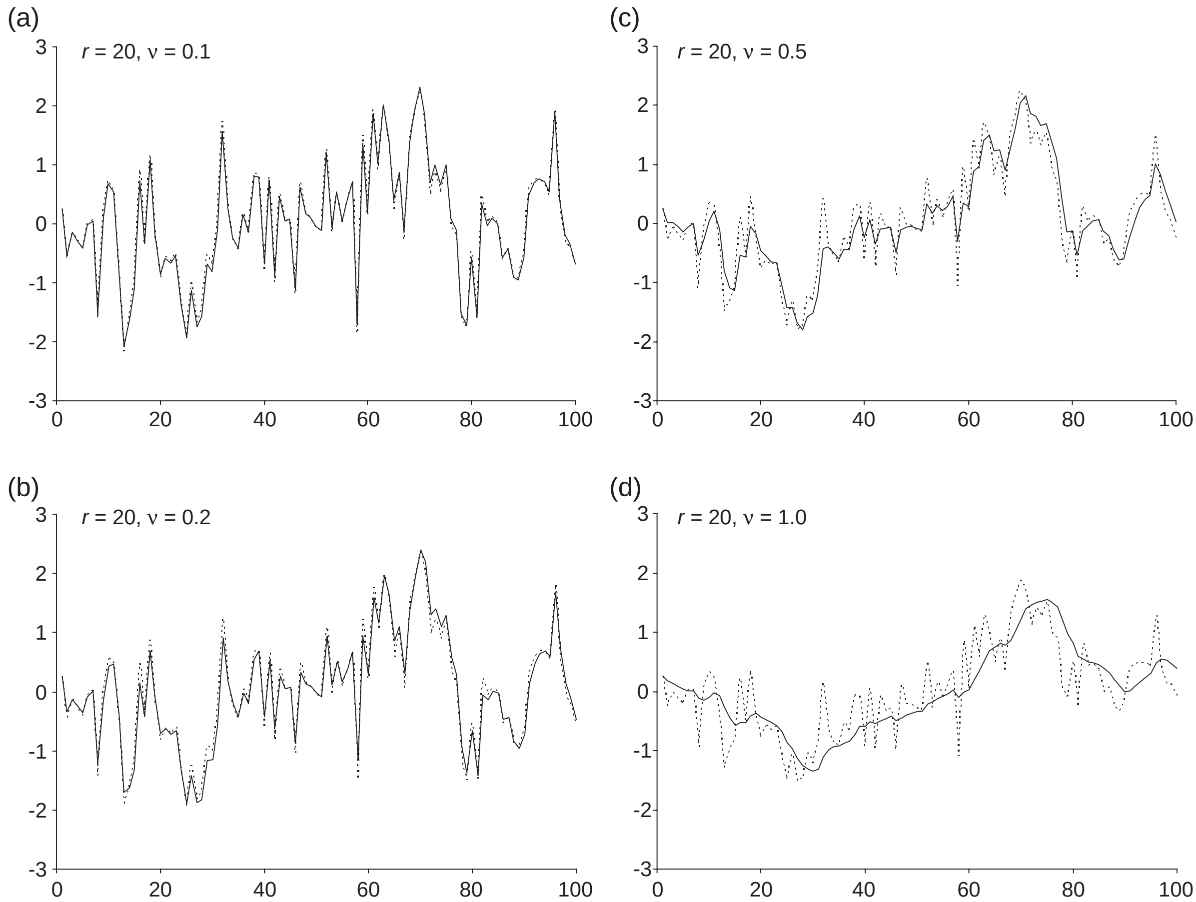

Fig. 2. Values on a transect of 100 m simulated from variograms with varying smoothness parameter v and constant nugget and sill. The solid line is based on $c_0=0$ and $c_1=1$, and the dotted line on $c_0=0.1$, $c_1=0.9$.

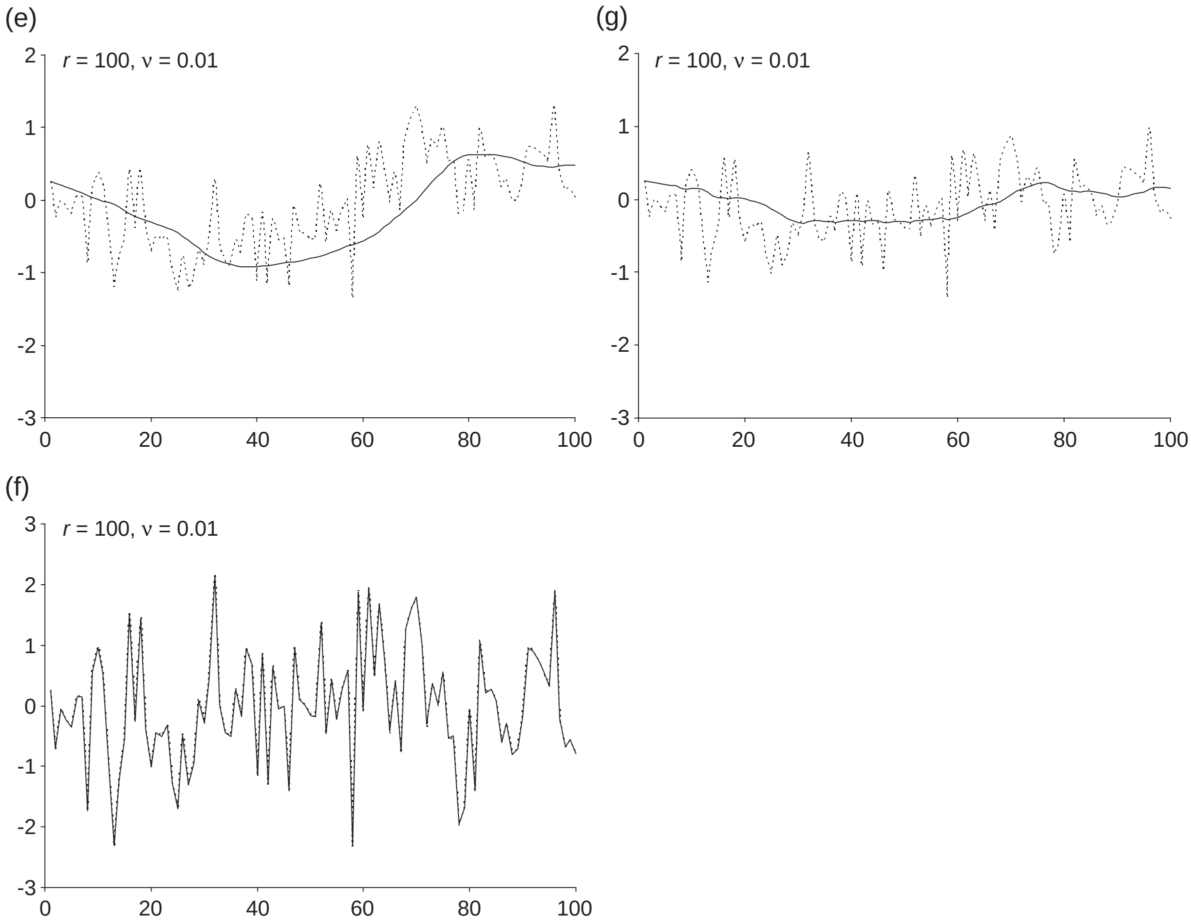

Fig. 2 (*conitnued*).

variance. As illustrated in Fig. 2a, for a rough field it is quite difficult to discern the effect of a nugget variance.

For a field that varies smoothly (Fig. 4) the mean of the predicted variogram curve using REML and WNLS matches reasonably well with the true variogram. Restricted maximum likelihood can estimate the nugget effect quite well (as seen by the relatively small uncertainty). The uncertainty is greater with increasing separation distance. This may also be due to the uncertainty in the LU simulation.

4.3. Analysis of soil data

Analysis of soil data is more complicated than that of simulated data. The optimization needs a good starting point; usually we only can guess the values of r and v. In general, we do not have information on the short-scale variation and the measurement uncertainty to be able to estimate the nugget variance.

We used the profile likelihood method for REML. The procedure adopted here was as follows:

- Choose a set of values for v, (e.g. 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1.0, 1.5, 2.0).
- For each value of v, maximize the log-likelihood over $[c_0, c_1, r]$.
- Plot the log-likelihood, ℓ, values as a function of v, and find the value of v that has the largest ℓ.
- Choose the values of $[v, c_0, c_1, r]$ where ℓ is maximum; these can be used as the estimated REML parameters. To refine the estimates further,

Table 2
Parameters of the Matérn model estimated using REML and weighted nonlinear least-squares (WNLS) from empirical variograms of the simulated data

	c_0	c_1	r	v
True	**0.10**	**0.90**	**5.0**	**0.10**
REML mean	0.44	4.51×10^7	9.4×10^9	2.57
REML standard deviation	0.32	5.74×10^8	1.5×10^{11}	2.46
Mean REML curve	0.44	0.57	4.9	0.20
WNLS mean	0.32	2.98	18.4	3.17
WNLS standard deviation	0.36	5.65	55.7	4.44
Mean WNLS curve	0.32	1.30	384.5	0.09
True	**0.10**	**0.90**	**5.0**	**1.00**
REML mean	0.08	2.19×10^6	3.16×10^7	2.18
REML standard deviation	0.04	3.30×10^7	5.0×10^8	1.98
Mean REML curve	0.08	1.27	8.7	0.75
WNLS mean	0.06	11.08	41.3	2.72
WNLS standard deviation	0.07	34.24	113.8	3.67
Mean WNLS curve	0.06	1.80	13.2	0.64

use these parameters as an initial guess and maximize the log-likelihood over all parameters $[c_0, c_1, r, v]$.

Fig. 5 shows the profile likelihood for the six sets of soil data. From the graphs we can identify the value of the smoothness parameter that maximizes the log-likelihood, ℓ. Most of the data show a large increase in ℓ as the value of v is increased before it reaches a maximum, then ℓ starts to decrease. A distinct optimum value for v can be seen for the soil thickness data at Forbes (Fig. 5a). However, for organic C (Fig. 5d) and the apparent electrical conductivity (ECa) transect data (Fig 5g) the maximum value of ℓ is indistinct, which indicates considerable uncertainty in the estimation of v.

Table 3 lists the parameters of the Matérn model estimated by REML and empirical variograms followed by weighted nonlinear least-squares (WNLS) and Fig. 6 shows plots of empirical variograms with their fitted models. Arguably the empirical variograms shown in Fig. 2 can be misleading; however these values merely enable a comparison with variograms computed using REML and WNLS. The presentation in Fig. 6 does not imply that the empirical value is the true variogram. A more appropriate measure of the goodness of fit is the log-likelihood value ℓ of REML, Eq. (16), as shown in Table 3.

For the A horizon thickness values at Forbes in Fig. 6a there is a clear difference between the model parameters estimated by WNLS and REML. It shows that the WNLS method has failed because the fitting by nonlinear least-squares is governed by the shape of the empirical variogram. In this example WNLS attempts to fit to the data at small separation distances (<200 m) and also at large distances (>600 m). The shape of the resulting model is close to Gaussian ($v=44$). This estimation of v is incorrect as the spatial process is quite irregular. If we model the empirical variogram to only half of its full distance we obtain a better estimate of v (=0.81). REML gives a more reasonable estimate of $v=0.24$. The log-likelihood value for REML is without doubt larger than for WNLS (Table 3). This highlights the discrepancy with

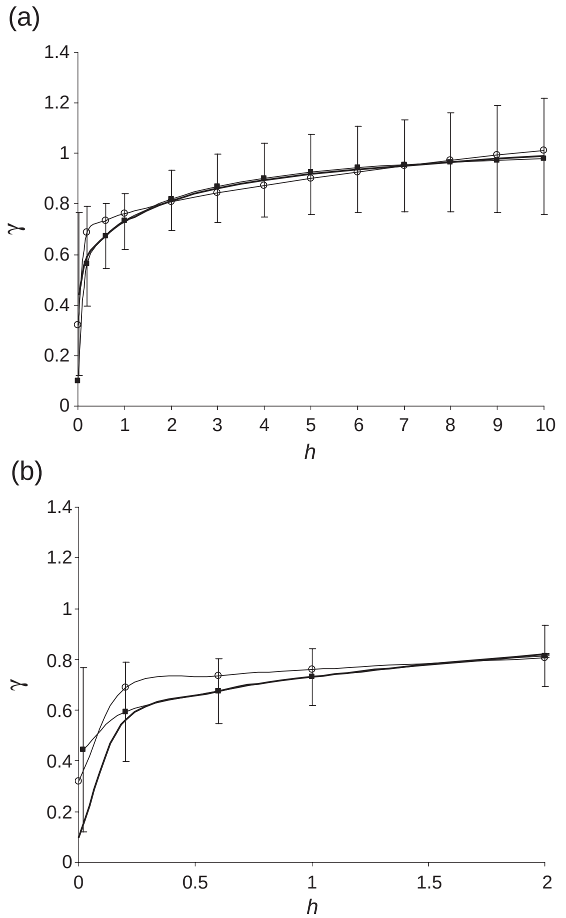

Fig. 3. The true variogram (bold line) and the mean of the estimated variogram using REML (solid line with black square) and WNLS (solid line with circle). The bars next to the black squares represent the standard deviations of the REML estimate.

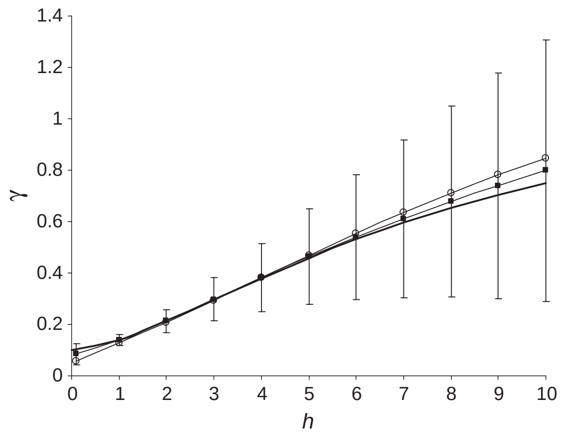

Fig. 4. The true variogram (bold line) and the mean of the estimated variogram using REML (solid line with black square) and WNLS (solid line with circle). The bar next to the black square represents the standard deviation of the REML estimate.

WNLS and indicates that this method cannot determine an appropriate value for v. When we have no information about the nugget component and estimate it using REML we get $c_0 = 0$. Pettitt and McBratney (1993) suggested that the variance at a short distance of 25 cm (which can be treated as nugget) is 1.14. When we fixed the nugget variance at 1.14, we obtained a value for v of 0.25, which is similar to the previous estimate. This highlights the importance of the sampling design, one that enables direct estimation of the nugget variance is valuable in this context. As shown in the simulation results, for a variable field it is difficult to obtain a good estimate of the nugget variance using either REML or WNLS.

Fig. 6b and c shows the variogram of soil pH in H_2O and $CaCl_2$ at Samford where model fitting by REML and WNLS gave similar results. The variogram shape is close to an exponential function, with a small value of v in the range 0.2–0.5, implying that the local spatial process of the soil is irregular. For the pH data from Samford, we calculated the measurement error from the replicates and used it as an estimate of the nugget variance. The part of the nugget effect arising from measurement error was quite small, 0.0005 and 0.0004 for pH in H_2O and $CaCl_2$, respectively. This suggests that the measurement error is much smaller than the variances at short distances.

Fig. 6d shows the variogram of topsoil (depth 0–10 cm) organic carbon content at Edgeroi. The spatial variation of organic carbon in this area is quite variable, however the WNLS method fitted a smooth curve through the empirical data giving a value of $v = 11$. Fitting the model to only half of the total distance did not improve the fit ($v = 34$). The REML method gave a value of $v = 0.3$ and quite a large value for r, indicating a power function.

Fig. 6e shows the subsoil pH at Pokolbin; again WNLS attempted to fit to the shape of the empirical variogram resulting in an estimate of $v = 17$. Meanwhile REML fits the local spatial variation as shown in the empirical variogram (Fig. 6e) up to a distance of 2000 m.

Fig. 6f shows the variogram of apparent EC along a transect on a farm in the lower Namoi valley and once more highlights the misleading fit of WNLS as it attempts to fit a curve that passes through the empirical variogram. By fitting the model to only half the total lag distance greatly improved the result. The REML estimate of v is 3 with a good fit to the empirical values to a distance of 50 m. The smooth process corresponds well with the data from a mobile sensing instrument which were intensive, where the physical volumes of the observations overlap.

4.4. General discussion

The results described above highlight the advantages and the drawbacks of using the Matérn model. The Matérn model is flexible and can be used to describe many isotropic soil spatial processes. It can fit a variety of variogram shapes ranging from bounded (exponential, Gaussian) to unbounded models (power functions). Nonlinear least-squares fitted closely to the empirical variogram, which can give misleading estimates of the smoothness parameter as it depends on the lag distance to which the model is fitted. Thus, it is difficult to estimate v, and in many cases it suggested an erroneously smooth process for soil data.

Clearly REML is a better method for estimating spatial variance. The effect of trend in the data can be accounted for and it fits the local variation well. However using REML to model the variogram is not straightforward; a profile likelihood method is required to obtain reasonable estimates of the parameters. The drawbacks of using REML are the heavy

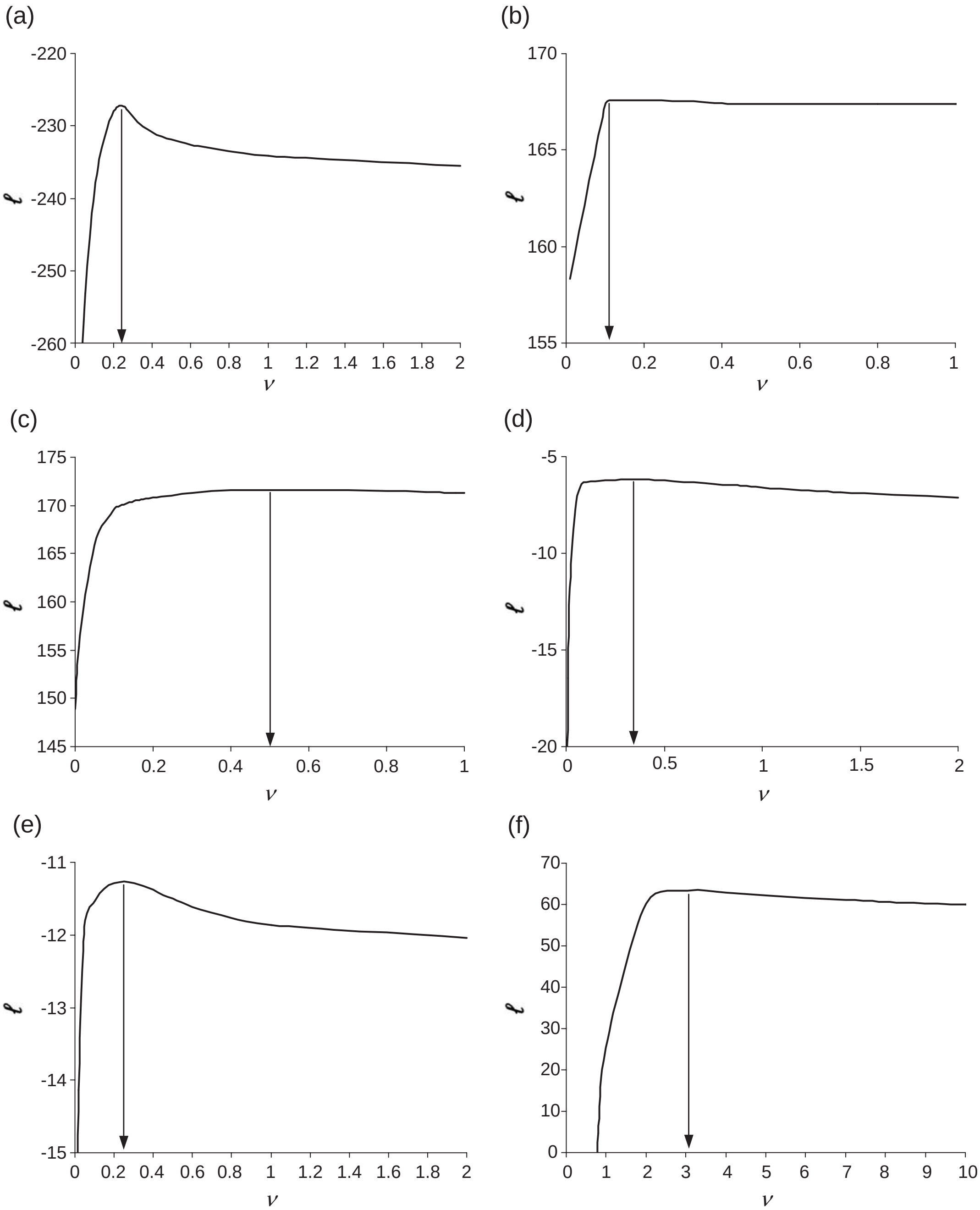

Fig. 5. Profile of smoothness parameter v and the log-likelihood values ℓ for: (a) a horizon thickness at Forbes, (b) pH in H_2O at Samford, (c) pH in $CaCl_2$ at Samford, (d) organic C content at Edgeroi, (e) pH in H_2O at Pokolbin, and (f) ECa along a transect. The arrow points to the value of v where ℓ is maximized.

computation, and the assumption of a multivariate normal distribution. It requires matrix inversion of the covariance matrix **C** at each iteration to maximize Eq. (16), for large sets of soil data (number of observations > 1000) obtained by remote or proximal sensing, REML will be difficult to apply. Nevertheless, some

Table 3
Parameters of the Matérn model for soil data estimated using REML and weighted nonlinear least squares (WNLS)

Variable	Method	c_0	c_1	r	υ	ℓ[a]
A hor. thickness,	REML	0.00	257.93 (221.43)[b]	2515.0 (5162.2)	0.24 (0.07)	−227.2
Forbes	REML[c]	1.14	194.32 (139.42)	1296.1 (2158.0)	0.25 (0.07)	−227.9
	WNLS[d]	22.28	138.07	28.4	43.99	−252.0
	WNLS[e]	4.63	131.48	217.6	0.81	−254.7
pH H_2O,	REML	0.006 (0.019)	0.039 (0.021)	35.0 (41.8)	0.17 (0.21)	167.6
Samford	WNLS[d]	0.004	0.038	17.16	0.17	167.1
	WNLS[e]	0.010	0.032	20.25	0.30	167.2
pH $CaCl_2$,	REML	0.015 (0.007)	0.033 (0.012)	19.5 (16.3)	0.52 (0.45)	171.6
Samford	WNLS[d]	0.014	0.030	19.3	0.37	171.1
	WNLS[e]	0.015	0.072	78.0	0.48	169.9
Organic C,	REML	0.27 (0.13)	0.70 (0.66)	10377.0 (27714.0)	0.32 (0.42)	− 6.2
Edgeroi	WNLS[d]	0.36	1.40	1049.9	10.72	− 8.3
	WNLS[e]	0.39	3.52	1369.4	34.12	−11.9
Subsoil pH,	REML	0.31 (0.10)	2.67 (3.54)	493270.0 (3296513)	0.25 (0.31)	−11.3
Pokolbin	WNLS[d]	0.36	0.39	405.4	16.92	−12.6
	WNLS[e]	0.38	0.51	1251.5	4.77	−13.3
ECa,	REML	0.05 (0.01)	90.71 (36.70)	10.2 (2.7)	3.04 (0.58)	63.4
transect	WNLS[d]	4.96	11348.00	26774.0	0.44	−285.2
	WNLS[e]	0.26	103.68	14.1	1.89	−15.9

[a] ℓ=log−likelihood for REML, Eq. (16).
[b] Numbers in the brackets represent standard error of the parameter estimates using REML.
[c] REML estimated parameters with c_0 being fixed at its 'true' value.
[d] Parameters estimated by fitting the Matérn model to the whole distance of the empirical variogram.
[e] Parameters estimated by fitting the Matérn model to half of the maximum distance of the empirical variogram.

more computationally efficient methods have been proposed (e.g. Stein et al., 2004). Furthermore, Stein (1999) suggested the use of a local variogram, as introduced by Haas (1990), and applied to soil data by Walter et al. (2001). Kriging with a local variogram has been used extensively with intensive spatial data, such as yield and electromagnetic induction data. Spatial interpolation could be done at specified locations by calculating a local variogram first from the 100 closest points in the neighbourhood using REML and then kriging using the estimated variogram parameters. At present this method is unsuitable for automatic fitting of the variogram model, however this is not impossible in the future as computer processing power increases and more efficient algorithms become available.

The assumption of a multivariate normal distribution is also a condition for the method-of-moments variogram, although there are many robust estimators that can deal with skewed distributions and outliers. Lark (2000a) discussed and compared several of these. However, Stein (1999) pointed out that these robust estimators do not take the dependencies in the data into account fully, and they might be less precise than the likelihood method.

Estimation of the nugget variance is quite uncertain in a rough field. The nugget variance represents discontinuous variation and its estimate is a combination of measurement error and short-range spatial variation. As we increase the smoothness, it is less likely that there is short-range variation, thus the apparent nugget effect will be most likely due to measurement error. Most soil data will be rough, thus an appropriate sampling design is desirable to estimate the fine-scale variation (e.g. Pettitt and McBratney, 1993). If replicate measurements are taken for some samples, and enough samples at small separations are included, the nugget effect could be split into measurement and short-distance spatial components.

This study emphasizes the importance of samples spatial layout in determining the Matérn parameters. From the examples, it appears that samples taken at regular intervals (pH at Stamford, Organic C at Edgeroi, and ECa transect) are problematic for estimating an optimum v value, as shown by the

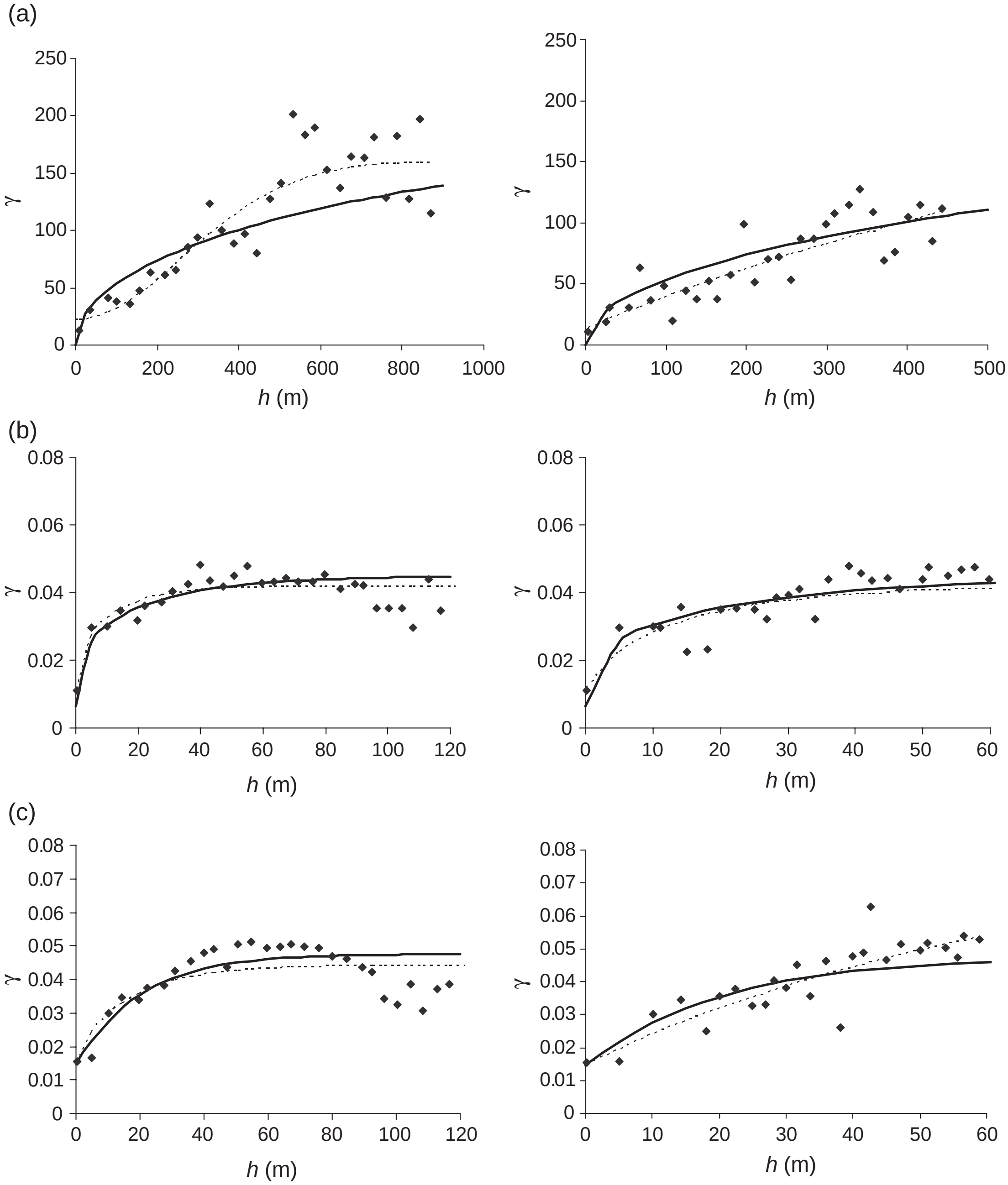

Fig. 6. Variograms of soil properties; symbols represent the empirical variogram, the solid lines are Matérn functions estimated by REML, and the dashed lines are estimated using weighted nonlinear least-squares on the empirical variogram. Figures on the left show the empirical variogram modelled to the full extent of the spatial data, and those on the right show the empirical variogram to half the full extent. The soil properties are: (a) a horizon thickness at Forbes, (b) pH in H_2O at Samford, (c) pH in $CaCl_2$ at Samford, (d) organic C content at Edgeroi, (e) pH in H_2O at Pokolbin, and (f) ECa along a transect.

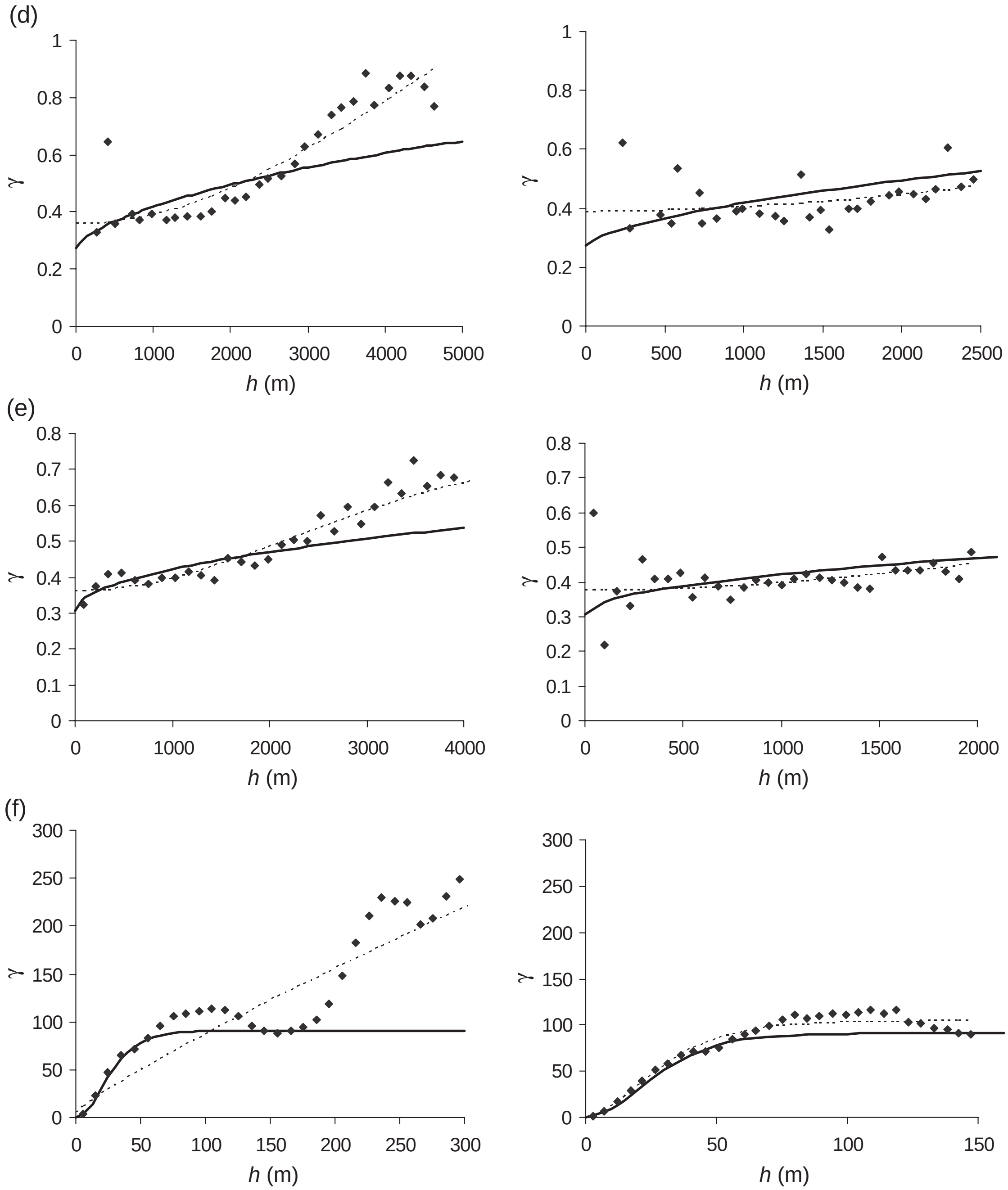

Fig. 6 (*continued*).

indistinct maximum value of the log-likelihood (Fig. 5) and large uncertainty in the estimates (Table 3). Meanwhile samples taken with a range of spatial separation distances (soil thickness at Forbes and pH at Pokolbin) give distinct values of v where the log-likelihood is maximized. Further study will be needed to elicit the spatial configuration for optimal determination of the Matérn parameters.

5. Conclusions

The Matérn model is flexible for modelling local spatial random processes. This paper shows the difficulty encountered by both WNLS and REML for estimating the parameters of the Matérn function. The WNLS estimation of v can be misleading and should be avoided. The REML method fits the local spatial process satisfactorily, however, the drawback is the heavy computation involved. Estimation of the nugget effect requires knowledge of the variation at short distances or the measurement variance.

From this study on soil data, the smoothness parameter is in the range of 0.25–0.5 (Table 3) and is considered to be rough (unsmooth). This explains why a model such as the exponential fits soil data reasonably well, assuming $v=0.5$ with the other parameters adjusted accordingly. For the ECa measurements from the mobile EM-31 instrument, a value of 3 for v suggests a smooth process; this is probably a result of the larger geometric support (around 3 m) and possible overlapping of the measurements. Thus assuming $v=0.5$ is reasonable for point (small geometric support) observations, it might not be equally true when the geometric support is large.

Acknowledgements

We would like to thank Dr. David Clifford for his insight on the profile likelihood method and helpful comments on this paper. Dr. Richard Webster and an anonymous reviewer are gratefully acknowledged for their valuable comments and suggestions. Finally, we thank Dr. Margaret Oliver for her great efforts in helping us turn our initial vague ideas and rough phrases into clearer concepts and more smoothly flowing sentences. We acknowledge the funding from ARC Digital Soil Mapping project.

References

Abramowitz, M., Stegun, I.E. (Eds.), 1972. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 10th Printing. U.S. Department of Commerce, National Bureau of Standards, Washington DC.

Chilès, J.-P., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley Interscience, New York.

de Wijs, H.J., 1951. Statistics of ore distribution: Part I. Frequency distribution of assay values. Journal of the Royal Netherlands Geological and Mining Society, New Series 13, 365–375.

de Wijs, H.J., 1953. Statistics of ore distribution: Part II. Theory of binomial distribution applied to sampling and engineering problems. Journal of the Royal Netherlands. Geological and Mining Society, New Series 15, 12–24.

Deutsch, C.V., Journel, A.G., 1997. GSLIB: Geostatistical Software Library and User's Guide, 2nd ed. Oxford University Press, New York.

Everitt, B.S., 2002. The Cambridge Dictionary of Statistics, 2nd ed. Cambridge University Press, Cambridge.

Haas, T.C., 1990. Kriging and automated variogram modeling within a moving window. Atmospheric Environment 24A, 1759–1769.

Handcock, M.S., Stein, M.L., 1993. A Bayesian analysis of kriging. Technometrics 35, 403–410.

Handcock, M.S., Wallis, J.R., 1994. An approach to statistical spatial–temporal modeling of meteorological fields (with discussion). Journal of the American Statistical Association 89, 368–390.

Heuvelink, G.B.M., 2000. Book Review: Interpolation of Spatial Data: Some Theory for Kriging: M.L. Stein, Springer, New York, 1999. Geoderma 96, 153–154.

Jian, X., Olea, R.A., Yu, Y.-S., 1996. Semivariogram modeling by weighted least-squares. Computers & Geosciences 22, 387–397.

Kitanidis, P.K., 1983. Statistical estimation of polynomial generalized covariance functions and hydrologic applications. Water Resources Research 19, 909–921.

Kitanidis, P.K., 1985. Minimum-variance unbiased quadratic estimation of covariances of regionalized variables. Journal of the International Association for Mathematical Geology 17, 195–208.

Kitanidis, P.K., 1987. Parametric estimation of covariances of regionalized variables. Water Resources Bulletin 23, 557–567.

Kitanidis, P.K., Lane, R.W., 1985. Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss–Newton method. Journal of Hydrology 79, 5–71.

Lark, R.M., 2000a. A comparison of some robust estimators of the variogram for use in soil survey. European Journal of Soil Science 51, 137–157.

Lark, R.M., 2000b. Estimating variograms of soil properties by the method-of-moments and maximum likelihood. European Journal of Soil Science 51, 717–728.

Laslett, G.M., McBratney, A.B., 1990. Further comparison of spatial methods for prediction of soil pH. Soil Science Society of America Journal 54, 1553–1558.

Laslett, G.M., McBratney, A.B., Pahl, P.J., Hutchinson, M.F., 1987. Comparison of several spatial prediction methods for soil pH. Journal of Soil Science 38, 325–341.

Lophaven, S., Carstensen, J., Rootzen, H., 2002. Methods for estimating the semivariogram. Symposium i Anvendt Statistik, pp. 128-144, Institut for Informationsbehandling, Handelshøjskolen i Århus.

Mardia, K.V., Marshall, R.J., 1984. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika 72, 135–146.

Mardia, K.V., Watkins, A.J., 1989. On multimodality of the likelihood in the spatial linear model. Biometrika 76, 289–295.

Matérn, B., 1960. Spatial variation. Meddelanden från Statens Skogsforskningsinstitut, 49, No. 5. [2nd Edition (1986), Lecture Notes in Statistics, No. 36, Springer, New York].

Matheron, G., 1965. Les Variables Régionalisées et leur Estimation. Masson, Paris.

MathWorks, 2004. Matlab Version 7.0, Release 14. The MathWorks Inc., Natick, MA.

McBratney, A.B., Pringle, M.J., 1999. Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. Precision Agriculture 1, 125–152.

McCullagh, P., Clifford, D., 2003. La loi du terroir: how oats and beans and barley grow. Technical Reports, Department of Statistics, University of Chicago.

McGarry, D., Ward, W.T., McBratney, A.B., 1989. Soil Studies in the Lower Namoi Valley: Methods and Data. The Edgeroi Data Set. CSIRO Division of Soils, Glen Osmond, South Australia.

Minasny, B., McBratney, A.B., Whelan, B.M., 2002. VESPER version 1.6. Australian Centre for Precision Agriculture, McMillan Building A05, The University of Sydney, NSW 2006.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Computer Journal 7, 308–313.

Nielsen, H.B., 1999. Damping parameter in Marquardt's method. Technical Report IMM-REP-1999-05, Department of Mathematical Modelling, Technical University of Denmark. http://www.imm.dtu.dk/~hbn/Software/.

Pettitt, A.N., McBratney, A.B., 1993. Sampling designs for estimating spatial variance components. Applied Statistics 42, 185–209.

Ripley, B.D., 1981. Spatial Statistics. John Wiley & Sons, New York.

Spanier, J., Oldham, K.B., 1987. Chapter 51. The Basset $K_v(x)$. An Atlas of Functions. Hemisphere, Washington, DC, pp. 499–507.

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

Stein, M.L., Chi, Z., Welty, L.J., 2004. Approximating likelihoods for large spatial data sets. Journal of the Royal Statistical Society. Series B, Statistical Methodology 66, 275–296.

Walter, C., McBratney, A.B., Douaoui, A., Minasny, B., 2001. Spatial prediction of topsoil salinity in the Chelif Valley, Algeria using kriging with local versus whole-area variograms. Australian Journal of Soil Research 39, 255–272.

Warnes, J.J., Ripley, B.D., 1987. Problems with likelihood estimation of covariance functions of spatial Gaussian processes. Biometrika 74, 640–642.

Webster, R., Oliver, M.A., 2001. Geostatistics for Environmental Scientists. John Wiley & Sons, Chichester.

Weisstein, E.W., 1999. Modified Bessel Function of the Second Kind. MathWorld—A Wolfram Web Resource. http://mathworld.wolfram.com/ModifiedBesselFunctionoftheSecond Kind.html.

Whittle, P., 1954. On stationary processes in the plane. Biometrika 41, 434–449.

Zimmerman, D.L., Zimmerman, M.B., 1991. A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. Technometrics 33, 77–91.