A stochastic–geometric model of the variability of soil formed in Pleistocene patterned ground.

R.M. Lark[1*], E. Meerschman[2], M. Van Meirvenne[2].

[1] *British Geological Survey, Keyworth, Nottinghamshire NG12 5GG, U.K.,* [2] *Research Group Soil Spatial Inventory Techniques, Department of Soil Management, Faculty of Bioscience Engineering, Ghent University, Coupure 653, 9000 Gent, Belgium.*

Abstract

In this paper we develop a model for the spatial variability of apparent electrical conductivity, $\mathrm{EC_a}$, of soil formed in relict patterned ground. The model is based on the continuous local trend (CLT) random processes introduced by Lark (2012) (Geoderma, 189–190, 661–670). These models are non-Gaussian and so their parameters cannot be estimated just by fitting a variogram model. We show how a plausible CLT model, and parameters for this model, can be found by the structured use of soil knowledge about the pedogenic processes in the particular environment and the physical properties of the soil material, along with some limited descriptive statistics on the target variable. This approach is attractive to soil scientists in that it makes the geostatistical analysis of soil properties an explicitly pedological procedure, and not simply a numerical exercise. We use this approach to develop a CLT model for $\mathrm{EC_a}$ at our target site. We then develop a test statistic which measures the extent to which soils on this site with small values of $\mathrm{EC_a}$, which are coarser and so more permeable, tend to be spatially connected in the landscape. When we apply this statistic to our data we get results which indicate that the CLT model is more appropriate for the variable than is a Gaussian model, even after transformation of the data. The CLT model could be used to generate training images of soil processes to be used for computing conditional distributions of variables at unsampled sites by multiple point geostatistical algorithms.

Keywords: Patterned ground; Apparent electrical conductivity; Electromagnetic induction; Stochastic geometry; Voronoi tessellation; Multiple point geostatistics; Pedometrics.

*Corresponding author: *E-mail address*: mlark@nerc.ac.uk (R.M. Lark).

1. Introduction

'Mais surtout nous insisterons sur la nécessité d'incorporer au maximum la physique du problème et le contexte géologique de la zone étudiée'. Chilès and Guillen (1984).

In most geostatistical analyses of soil the data are assumed to be a realization of a multi-Gaussian random function, perhaps after they have been transformed so that their histogram represents a Gaussian distribution. Furthermore, the random function commonly has a spatial covariance function drawn from a limited subset of models (Webster and Oliver, 2007), which are used because of their convenient mathematical properties. In some of the earth sciences there has been progress in the development of random functions with parameters that are determined, or at least constrained, by parameters of underlying processes which have a physical meaning (e.g. Kolvos et al., 2004; Chilès and Guillen, 1984). This has advantages (Lark, 2012a), for example, the efficiency of spatial sampling to model the spatial covariance function could be improved if prior distributions for covariance parameters could be specified from process knowledge. However, this has not been achieved in soil science. Lark (2012a) suggested that this is probably because the variables that soil scientists study are commonly influenced by a more complex set of factors at more diverse spatial scales than is the case for the variables where it has proved possible to specify the covariance function from process information. For example, the covariance function for diffusion processes is well-established (Whittle, 1954; 1962), and diffusion is a source of spatial variation in the concentration of nutrients in soil, but it is just one of many sources of spatial variation, and is of limited importance at the spatial scales most generally studied for practical purposes.

Lark (2012a,2012b) suggested that progress might be made by recognizing a number of distinct *modes* of soil variation, simple and generalizable rules that capture how the effects of factors of soil variation vary laterally, and which map naturally on to particular spatial random functions. For example, in conditions where soil variation is strongly determined by differences between discrete domains in the landscape (such as geological units, topographic units, fields etc.) then a subdivision of space into random sets such as Poisson Voronoi polygons may be appropriate (Lark, 2009) and properties of the spatial model (such as the mean chord length of the polygons) may be given a physical meaning.

Lark (2012b) proposed a mode of soil variation: continuous local trends. Under this mode of variation soil varies laterally in space, changing continuously rather than in a step-wise fashion; and these trends are local and repeating, so that they are essentially unpredictable (in contrast to a large-scale trend in a variable that might be observed across a study area). Examples of continuous local trends would be concentration gradients around the rhizosphere, or around individual plants, and catenary variation at landscape scale. Lark (2012b) proposed a general family of random functions to describe continuous local trends (CLT random functions). The value of a CLT variable at some location is given by a distance function, whose argument is the distance from the location of interest to the nearest event in a realization of a spatial point process. This makes the CLT a random function. The CLT variables considered by Lark (2012b), and in this paper, are Poisson CLT (PCLT) variables because the spatial point process is completely spatially random. Lark (2012b) estimated parameters of a PCLT process from data on a soil variable. It was also pointed out that the PCLT process might differ from a comparable Gaussian random function with respect to its multiple point statistics (Strebelle, 2002). This raises the possibility that PCLT models, as well as mapping closely on to a particular mode of soil variation, might be practically useful for applications where spatial connectivity plays a major role controlling processes in soil and so the multiple point statistics of the variable are important.

In this paper we use a PCLT random function to model the variation of apparent electrical conductivity, EC_a, of soil at a site where this variable is strongly influenced by spatial patterns in the parent material. These patterns arose from the development of ice wedges in Eocene clay under permafrost conditions, and subsequent infilling by coarser material which leads to strong textural contrasts in the soil. The objective is to show how soil knowledge: general knowledge about soil formation in the particular environment and its relationship to EC_a, and some simple descriptive statistics of the data (summary statistics and empirical variograms), allow us to select and fit a PCLT model. We then compare the PCLT model with a trans-Gaussian (TG) model of the data, i.e. a model fitted by conventional geostatistical analysis after the data have been transformed to approximate normality. Specifically we compare the models with respect to a statistic that summarizes the spatial connectivity of the coarser material, which might be relevant to simulations of transport processes in the soil. We then evaluate which model appears best to represent the spatial pattern in the data.

2. Case Study

2.1 The study area and data collection.

We surveyed an area of Pleistocene patterned ground in the sandy silt region of Belgium. The patterned ground was identified by polygonal crop marks on an aerial photograph and interpreted to be the result of ice wedge formation during the last glacial period. The study area and data collection were discussed in detail by Meerschman et al. (2011), therefore we limit ourselves here to a brief presentation of it. More general information on ice-wedge polygons constitutes part of the soil-knowledge base that we use in this study, and is presented in section 2.3.2 below as it is required.

The study area (0.6 ha) was located in an agricultural field in Deinze, Belgium (central coordinates: 51° 01′16″N, 3°29′41″E). Excavation of a small part of the study area (6×6-m) to a depth of 0.9 m uncovered an ice-wedge pseudomorph with a diameter of about 6 m. The wedges were formed in clay-rich Tertiary marine sediments that were covered with a 0.6 m layer of silty-sand Quaternary deposits. Texture analysis on 94 subsoil samples (0.6 - 0.8 m) showed a clear contrast between the Eocene host material (on average 21% clay) and the superficial material (on average 6% clay).

Previous studies (Saey et al., 2009; Cockx et al., 2006) have shown that $\mathrm{EC_a}$ is a useful covariate to study textural variability at profile and polygon-scale in soils formed in these conditions. The study area was surveyed with a mobile proximal soil sensor measuring $\mathrm{EC_a}$($\mathrm{mS\ m^{-1}}$) of an underlying soil volume down to approximately 1.5 m. The sensor was mounted on a sled pulled by an all terrain vehicle. The vehicle drove along parallel lines with an in-between distance of on average 0.75 m. The within-line distance between sensor response registrations was 0.15 m.

2.2 Initial data analysis.

Meerschman et al. (2011) noted that the $\mathrm{EC_a}$measurements clearly reflected the polygonal patterns: small $\mathrm{EC_a}$values indicated the former ice wedges filled with lighter material. In addition to the short-range variation in $\mathrm{EC_a}$, there were large values of $\mathrm{EC_a}$

near an old field track in the north-east of the surveyed region. To avoid any assumptions about the form of this trend we decided to restrict our analyses to the lower left quadrant of the surveyed area, a region of approximately 40×40-m, with 17 792 observations, which excludes this area with elevated $\mathrm{EC_a}$. Figure 1 shows a post-plot of these data.

Figure 2 shows the histogram of the data. Summary statistics are presented in Table 1. Note that the data are mildly skewed. In the analyses reported below the PCLT model was fitted in all cases to the raw data, and all analyses with the TG model were done with the data after a transformation which is described in section 2.3.1 below.

2.3 Spatial analysis.

In this section we describe the analysis of the $\mathrm{EC_a}$ data to fit a TG model and a PCLT model. The first task (section 2.3.1) was straightforward after a data transformation, which is described. In section 2.3.2 we describe how soil knowledge was used to fit the PCLT model.

2.3.1. Trans-Gaussian model

The objective of the case study is to compare a continuous local trend (PCLT) model of the data with a trans-Gaussian (TG) model, as might be used in standard geostatistical analysis. Although the data are only mildly skew, since the objective of this exercise is to compare a Gaussian or Trans-Gaussian model with a stochastic geometric alternative, it was decided to transform the data so that the histogram and summary statistics were as close as possible to those expected for data drawn from a Gaussian random variable. We therefore used a Box-Cox transformation of the data to normality for the TG modelling:

$$\begin{aligned} y &= \frac{z^{\zeta}-1}{\zeta} \quad \zeta \neq 1, \\ &= \log_e(z) \quad \zeta = 1, \end{aligned} \tag{1}$$

where z is a value on the original scale and y is a transformed value. We used the BOXCOX procedure from the MASS package (Venables and Ripley, 2002) for the R platform (R Development Core Team, 2012) to find the likelihood profile of the ζ parameter, and selected the value with maximum likelihood. The data were then transformed with the maximum likelihood estimate of ζ, substituted into Eq. (1) and then standardized to zero mean and unit variance. The estimate of ζ and summary statistics for the data after transformation, and standardization, are presented in Table 2.

An isotropic empirical variogram of the transformed and standardized data was then computed using the method of moments estimator due to Matheron (1962) as implemented in the FVARIOGRAM directive in GenStat (Payne, 2010). An authorized model was then fitted to the estimated variogram by weighted least squares (Cressie, 1985) using the MVARIOGRAM procedure in GenStat (Harding et al., 2010). Alternative models were considered and the stable or powered exponential model was selected on the basis of the Akaike information criterion (McBratney and Webster, 1986). This variogram model takes the form

$$\gamma(r) \;=\; c_0 + c_1\left(1 - \exp(-\{r/a\}^{\kappa}\right), \tag{2}$$

where c_0 and c_1 are, respectively, the variances of the nugget and spatially correlated components of the variable, r is lag distance, a is a distance parameter and κ is a shape parameter where $0 < \kappa \leq 2$. The estimates of these parameters are presented in Table 2, and the estimates of the variogram of the TG variable, and the fitted model are shown in Figure 3.

2.3.2. Stochastic Geometric model

Estimates of the isotropic variogram of the raw data on $\mathrm{EC_a}$were obtained using the method of moments estimator due to Matheron (1962) as described for the transformed data in section 2.3.1. (these are the solid symbols in Figure 6). The identification and fitting of an appropriate stochastic geometric model for the soil variable will allow us to plot a continuous variogram function for these estimates.

When a TG model is fitted it is assumed that, after any transformation, the data $\mathbf{y} = \{y(\mathbf{x}_1), y(\mathbf{x}_2), \ldots, y(\mathbf{x}_n)\}$ from the n locations $\mathbf{x}_1, \mathbf{x}_1, \ldots \mathbf{x}_n$ can be regarded as a realization of an n-variate Gaussian random variable, $\mathbf{Y}$. Under this assumption the variogram of $\mathbf{Y}$ entirely summarizes the information that the data contains about its spatial variability, and the task of estimating model parameters, under the assumption of a stationary mean, reduces to the task of estimating variogram parameters. This is not the case with models for random variables, such as the PCLT models, which have non-zero odd moments of order three or larger, and therefore are not Gaussian. The fitting of a PCLT model cannot, therefore, simply reduce to the computation of parameters which minimize the weighted sum of squared residuals between the empirical and fitted variogram.

In this study our approach to the selection and estimation of a PCLT model is to constrain it by soil knowledge. Soil knowledge consists of general understanding of the underlying processes that influence soil formation and so the variation of the target variable, and also of general quantitative information about the variable in the study site or a homologous site, represented by summary statistics, empirical variograms or similar information. In the following sections we go through a semi-formal process of model identification based on inferences from soil knowledge and culminating in the estimation of parameters for an appropriate model. Each subsection is headed with a question, and with the general source of soil knowledge used to address it. The individual elements of soil knowledge are then summarized in brief labelled sentences, expanded in a short paragraph. Inferences from this soil knowledge are then set out.

2.3.2.1. Question: 'What mode of soil variation?' Soil knowledge about the underlying pedogenetic process.

The identification of a general mode of soil variation is based on two items of soil knowledge which are listed below.

SK1. *The dominant source of soil variation at metre scales in this landscape is the presence of Pleistocene ice-wedge polygons.* These are described in more detail by Meerschman et al (2011). Ice-wedge polygons form in periglacial conditions on surfaces with slopes less than a critical value. Over much of central Europe ice-wedge polygons formed in periglacial conditions during the Quaternary, they are detectable at the study site from airphotography. It has been shown (Cresto Aleina et al., 2012) that the comparable polygonal patterns in ground of contemporary tundra can be modelled as a Poisson Voronoi Tessellation (PVT), that is to say one may postulate an underlying homogeneous spatial point process of completely spatially random seed points, and any one polygon consists of all locations nearest to one associated seed point than to any of the others. See Lark (2009) for a summary of some of the properties of PVT spatial processes and Okabe et al. (2000) for a more complete account. Note, in particular, that the polygons generated by this process are not of uniform size or shape. By analogy we infer that a PVT model would be a plausible descriptor of the ice-wedge polygons at the study site.

SK2. *We may expect more or less continuous variation in depth-integrated soil properties from the centre to the edge of any polygon.* Much of the polygonal patterned ground formed in Europe and North America during the Quaternary was covered by aeolian or glacio-fluvial sand or silty deposits. These have an important role in subsequent pedogenesis (Catt, 1979; Walters, 1994) imposing local lateral trends. At the centre of a polygon there is typically a relatively thin layer of sandy or silty superficial material over the host material in which the ice wedges originally formed. After thawing, the space previously occupied by ice in the wedges that delineate the polygons was typically filled with the superficial material. Any depth-integrated soil property, such as $\mathrm{EC_a}$, can therefore be expected to vary laterally (although not necessarily linearly) from the centre of the polygon to its edge if there is a texture contrast between the host material and the superficial material. There is such a contrast at the Deinze study site where the overlying material is silty-sand Quaternary deposits, and the host material is Eocene sandy clay (Meerschman et al., 2011).

From these two elements of soil knowledge we may infer that the spatial variation of a depth integrated soil property such as $\mathrm{EC_a}$, in these conditions, can plausibly be regarded as a Poisson Continuous Local Trend random process as defined by Lark (2012b). In the next section we consider what distance function might be proposed.

2.3.2.2 Question: 'What type of distance function is plausible?' Soil knowledge about pedogenetic processes and summary statistics.

SK3. *We may expect* $\mathrm{EC_a}$ *to decline from the polygon centre to the rim.* It is generally found that measurements of $\mathrm{EC_a}$ made by electromagnetic induction are positively correlated with the clay content of the soil (e.g. Kachanoski et al., 2002; Saey et al., 2009). For this reason we should expect $\mathrm{EC_a}$, as a depth-integrated variable, to decline from the polygon centre, where the thickness of sandy and silty material over the heavier host material is thinner, to the edge of the polygon where the former ice wedge is filled with the lighter material. This was found to be the case by Meerschman et al. (2011).

SK4. *The data on* $\mathrm{EC_a}$ *are mildly positively skewed.* This can be seen in Table 1.

The simplest PCLT model, as used by Lark (2012b), has a linear distance function $\mathcal{D}(k) \propto k$. If the distance function has a positive slope, i.e. $\{k' > k\} \rightarrow \{\mathcal{D}(k') > \mathcal{D}(k)\}$, then it can be seen that the corresponding PCLT random function has a moderate positive skewness (about 0.65). A linear distance function with a negative slope, needed for consistency with **SK3**, would therefore give rise to a random function with a moderately negative skewness. This is not compatible with **SK4**.

Of the distance functions examined by Lark (2012b) one in which the distance function is proportional to the reciprocal of distance is compatible with **SK3** and **SK4**. The reciprocal of distance declines with distance (**SK3**), and the example of such a random function given by Lark (2012b) has mild positive skewness (**SK4**). On this basis it was decided to proceed with further analysis on the assumption that the data on $\mathrm{EC_a}$ could be regarded as realizations of a PCLT process with a distance function linearly proportional to

$$\mathcal{D}(k) = \frac{1}{k + \alpha}, \tag{3}$$

where k is distance to the nearest event of the underlying spatial point process, and α is a parameter which must take some value $\alpha > 0$ to ensure that the distance function is defined for all positive k. We refer to this PCLT as the inverse-distance PCLT in the remainder of this paper. Note that the distance function in Eq. (3) defines what we shall call the standard PCLT variable. The random variable that models the target soil variable is linearly proportional to the standard PCLT variable, so fitting the model entails the estimation of parameters of the standard PCLT along with a scale parameter which is the a priori variance of the random variable.

The inverse-distance function was selected because it was seen to be a simple function, at least potentially compatible with available soil knowledge. In due course its parameters are estimated and this gives some further indication of its plausibility, and in section 2.3.3 we evaluate statistics to compare its plausibility with the TG model.

We call the standard inverse-distance PCLT random function Z_{id}. We shall model the $\mathrm{EC_a}$ data as a realization of a random function Y where

$$Y = \beta Z = \beta \left(Z_{\mathrm{n}} + Z_{\mathrm{id}} \right), \tag{4}$$

where β is a constant of proportionality and Z_{n} is an independently and identically distributed Gaussian nugget component of mean zero. This nugget component is included in the random model for the target variable to account for any variation spatially correlated at scales finer than the sampling interval. This is common practice in geostatistical modelling with standard covariance models such as the spherical, exponential or Matérn.

We now obtain the cumulative distribution and density functions of Z_{id}. We first define the inverse of the distance function in Eq.(3), $\dot{\mathcal{D}}(z_{id})$, such that

$$\left\{ z_{\mathrm{id}} \;=\; \mathcal{D}(k) \;=\; \frac{1}{k+\alpha} \right\} \rightleftarrows \left\{ \dot{\mathcal{D}}(z_{id}) \;=\; k \right\}.$$

Then

$$\dot{\mathcal{D}}(z_{\mathrm{id}}) \;=\; \frac{1}{z_{\mathrm{id}}} - \alpha. \tag{5}$$

Since $\mathcal{D}(k)$ is monotonic and decreasing with increasing k for admissible (non-negative) values of k, the marginal cumulative distribution function of Z_{id}, $F_{\mathrm{id}}(z)$ can be written as

$$F_{\mathrm{id}}(z_{\mathrm{id}}) \;=\; 1 \;-\; F_k(\dot{\mathcal{D}}(z)), \tag{6}$$

where $F_k(k)$ is the marginal cumulative distribution function of k. In Eq. (14) of Lark (2012b) it is shown that, for a Poisson point process in 2-D with intensity λ,

$$F(k) \;=\; 1 \;-\; \exp\left\{-\lambda\pi k^2\right\}, \tag{7}$$

and so

$$F_{\mathrm{id}}(z_{\mathrm{id}}) \;=\; \exp\left\{-\lambda\pi\left(\frac{1}{z_{\mathrm{id}}} - \alpha\right)^2\right\}, \tag{8}$$

which is defined for $0 \leq z_{\mathrm{id}} \leq 1/\alpha$, which shows that the random function Z_{id} has an upper and a lower bound.

By differentiation of $F_{\mathrm{id}}(z_{\mathrm{id}})$ with respect to z_{id} we can obtain a probability density function (PDF):

$$\begin{aligned} f_{\mathrm{id}}(z_{\mathrm{id}}) \;&=\; \frac{2\lambda\pi\left(\frac{1}{z_{\mathrm{id}}} - \alpha\right)}{z_{\mathrm{id}}^2} \exp\left\{-\lambda\pi\left(\frac{1}{z_{\mathrm{id}}} - \alpha\right)^2\right\}, \quad 0 < z_{\mathrm{id}} \leq \frac{1}{\alpha} \\ &=\; 0, \qquad \text{otherwise.} \end{aligned} \tag{9}$$

A soil variable modelled as an inverse-distance PCLT random function is assumed to have a spatially correlated component that is linearly proportional to z_{id} for some values of the parameters α and λ. As noted above, the soil variable is assumed to be a realization of a random function Z that includes an independent Gaussian nugget component of mean zero. If the PDF of the nugget component is denoted by $f_{\mathrm{n}}(z_{\mathrm{n}}$, then the PDF of Z, $f(Z)$, can be obtained by the convolution operation

$$f(z) \; = \; \int_{-\infty}^{\infty} f_{\mathrm{n}}(x) f_{\mathrm{id}}(z-x) \; \mathrm{d}x, \tag{10}$$

since Z_{id} and Z_{n} are independent random variables (Dudewicz and Mishra, 1988).

The next question that we consider is a plausible range of values for the α parameter.

2.3.2.3 Question: 'What is a plausible range of values for, λ, the intensity of the process and for the parameter α of the distance function?' Soil knowledge from field observations and an estimate of the proportion of variation of $\mathrm{EC_a}$ *that is attributable to the nugget component*

SK5. *Meerschman et al. (2011) report a detailed excavation of a polygonal cell with diameter about 6 m, which they regard as typical from airphoto evidence.* If all cells have a diameter of d m then the average intensity of an underlying spatial point process is the reciprocal of the cell area which may be approximated (treating the cells as circular) by $4/\pi d^2$. On the basis of the observation of Meerschman et al. (2011) it was decided to consider a range of possible values of λ for the spatial point process in the interval $\left[0.02\mathrm{m}^{-2}, 0.08\mathrm{m}^{-2}\right]$ which corresponds to a range of polygon diameters from 4 to 8 m (i.e. 2 m either side of the value proposed as representative).

SK6. *The nugget variance of the (untransformed)* $\mathrm{EC_a}$ *data is about 10% of the correlated variance.* This information is used to calculate moments for the variable Z, given values of α and λ, by evaluation of the PDF in Eq.(10). It should be noted that in the final model the nugget variance is estimated separately, and is not constrained by this assumption. To obtain this proportion we fitted a powered exponential model, Eq.(2), to the empirical variogram of the $\mathrm{EC_a}$data(not shown here) using the MVARIOGRAM procedure in GenStat (Harding et al, 2010).

The mean and variance of an inverse-distance PCLT random function, Z_{id}, for some values of the parameters α and λ was obtained from the PDF in Eq.(9), the QDAG algorithm in the IMSL library (Visual Numerics, 2006) was used for numerical integration. It was then possible to compute the variance of an independent Gaussian nugget component, Z_{n}, such that the variances of Z_{id} and Z_{n} were in the same ratio as **SK6** suggests pertains for the $\mathrm{EC_a}$ data. The coefficient of skewness for the sum of these two random variables could then be calculated from moments obtained by numerical integration of the convolution of the distributions of Z_{id} and Z_{n}, as described in Eq.(10).

Figure 4 is a plot of values of the skewness coefficient of variable Z_{id} for values of the parameters α and λ, the range for λ obtained from **SK3**. Note that over much of the range of values of λ it is α that has the strongest effect on the skewness. The two contours drawn on the Figure bound a region within which the skewness is in the interval $[0.25, 0.5]$. We regard this as mild positive skewness, compatible with **SK4**, and so we assume that jointly plausible values of α and λ lie within these limits. The Figure shows, for example, that values of α less than 2 m seem unlikely to be compatible with **SK4** since coefficients of skewness for such variables are larger than 0.5. Similarly, if $\lambda = 0.05$ then a plausible range of values of α indicated by the Figure is 2.5–3.8 m.

2.3.2.4 Model fitting given the soil knowledge

Estimates of the isotropic variogram of the raw data on $\mathrm{EC_a}$were obtained using the method of moments estimator due to Matheron (1962) as implemented in the FVARIOGRAM directive in GenStat (Payne, 2010). An inverse-distance PCLT model was then fitted to the estimates by weighted least squares, but subject to the condition that α and λ fall jointly within the range defined by the two contours shown in Figure 4. The variogram for the standard PCLT variable Z_{id} variable depends only on the parameters α and λ. In order to fit the PCLT model to the empirical variogram of the soil process it is also necessary to estimate the proportionality constant β which scales the standard PCLT variable to the variable assumed to be realized in the soil data, as shown in Eq (4). This is done indirectly here by direct estimation of the *a priori* (sill) variance of the correlated component of the variogram of Y (defined in Eq (4))

$$c_1 = \beta^2 \mathrm{var}\left[Z_{\mathrm{id}}\right]$$

along with a nugget component

$$c_0 \;=\; \beta^2 \mathrm{var}\left[Z_\mathrm{n}\right]$$

where $\mathrm{var}[Z]$ denotes the *a priori* variance of random variable Z. The fitted variogram for the target random variable, Y, was specified by:

$$\gamma(r) \;=\; c_0 + c_1 g_\mathrm{id}(r|\alpha,\lambda), \tag{11}$$

where $g_\mathrm{id}(r|\alpha,\lambda)$ is the variogram of the PCLT process with parameters α and β and the *a priori* variance scaled to 1.0 thus:

$$g_{\mathrm{id}|\alpha,\lambda}(r) \;=\; 1 - \frac{C_\mathrm{id}(r|\alpha,\lambda)}{C_\mathrm{id}(0|\alpha,\lambda)}, \tag{12}$$

where $C_\mathrm{id}(r|\alpha,\lambda)$ is the covariance function for lag r for the standard inverse-distance PCLT process with parameters α and λ. The covariance function for a variable in 2-D is given by

$$C_\mathrm{id}(r|\alpha,\lambda) \;=\; \int_{\mathbb{R}^2} \left\{S(k,k_r) + F(k) + F(k_r) - F(k)F(k_r) - 1\right\} \left\{-\frac{1}{(k+\alpha)^2}\right\} \mathrm{d}k \; \left\{-\frac{1}{(k_r+\alpha)^2}\right\} \mathrm{d}k_r, \tag{13}$$

where $S(k,k_r)$ is the joint survival function for the underlying spatial point process, as defined by Lark (2012b). This equation is obtained directly from Eq.(20) of Lark (2012b) and the reader is referred to that paper for details.

The inverse distance model was fitted as follows.

i). The value of the parameter α was set at a fixed value, in turn α =2.0 m, 2.25 m, 2.50 m . . .

ii). The parameter λ was then set at values over some range $[\lambda_{\alpha,\mathrm{min}}, \lambda_{\alpha,\mathrm{max}}]$ where $0.02 \leq \lambda_{\alpha,\mathrm{min}} < \lambda_{\alpha,\mathrm{max}} \leq 0.08$ such that for specified α and any $\lambda \in [\lambda_{\alpha,\mathrm{min}}, \lambda_{\alpha,\mathrm{max}}]$ the expected value of the skewness coefficient, as read off Figure 4, was within the interval [0.25, 0.5].

iii). For the set values of α and λ values of c_o and c_1 were found so that the weighted sum of squared deviations of the variogram function in Eq (11) and the empirical variogram (Cressie, 1985) were minimized. These values were found with the IMSL optimization subroutine BCPOL (Visual Numerics, 2006).

iv). Repetition of step (iii) for successive values of $\lambda \in [\lambda_{\alpha,\min}, \lambda_{\alpha,\max}]$ produced a 'profile plot' of the weighted sum of squares, WSS, against λ. Such plots were produced for successive values of α, as designated in step (i). Estimates of α, λ, c_0 and c_1 were found from the profile plot for which the minimum WSS was the smallest of all observed values.

The resulting estimates of α and λ were 2.5 m and 0.07 m^{-2} respectively. The estimates of c_0 and c_1 were 0.49 and 4.03 respectively. Figure 5 shows the profile plot of the weighted sum of squares with $\alpha = 2.5$ m and Figure 6 shows the empirical variogram for the untransformed data and the fitted inverse-distance PCLT model. In Figure 7 is shown (line) the corresponding distribution function for the random function $Z = Z_{\mathrm{id}} + z_{\mathrm{n}}$ standardized to zero mean and unit variance according to the values of the mean and standard deviation obtained from the PDF in Eq (10). Also plotted on Figure 7 are points from the empirical CDF of the standardized $\mathrm{EC_a}$ data. The theoretical and empirical distribution functions are in reasonable agreement, although the median of the former seems to be rather smaller than the latter.

2.3.3. Comparing the TG and PCLT models

It is well known that Gaussian (and trans-Gaussian) models of spatial variation, in which all information on variability is expressed by two-point statistics such as the covariance function, are not able to reproduce all important features of natural spatial fields, which must be represented by higher-order moments (e.g. Guardiano and Srivastava, 1993). This has been the motivation for the development of multiple point statistics. In this section we investigate whether the PCLT model allows better characterization of the spatial structure of the $\mathrm{EC_a}$ data than does the TG model.

One feature of the Gaussian and trans-Gaussian random variables that often limits their applicability is the fact that large values of the variable tend to be spatially isolated from other large values, the same holds for small values (e.g. Gómez-Hernández and Wen, 1997; Strebelle, 2002). In this case study we may consider locations with small values of $\mathrm{EC_a}$. These locations are likely to be dominated by lighter sandy and silty Quaternary material, rather than the heavier-textured Eocene host material, and so will have larger porosity and hydraulic conductivity, than sites where the $\mathrm{EC_a}$ is larger. If the TG model does not adequately represent the connectivity of such areas then any modelling based on TG simulation will fail to represent processes where this lateral connectivity matters. This could include processes such as lateral movement of a pollutant plume in saturated conditions, the response of the water table to drainage schemes or the lateral spread of root pathogens. Figure 8 shows sets of realization of each of the fitted PCLT and TG models for $\mathrm{EC_a}$. The inverse-distance PCLT realizations were generated directly following the procedure used by Lark (2012b). The TG realizations were obtained by Sequential Gaussian Simulation using the SGSIM subroutine from the GSLIB library (Deutsch and Journel, 1997) modified to use the powered exponential variogram function. On visual inspection it can be seen that, while some large patches with smaller $\mathrm{EC_a}$values are seen in the TG realization, there are fewer isolated small patches with small $\mathrm{EC_a}$values in the inverse-distance PCLT realization, which has large and connected regions with small conductivity around the boundaries of the Voronoi cells of the underlying point process. However, this visual inspection is of limited usefulness and a more objective measure is needed.

To this end we consider a simple test statistic, which can be readily evaluated on the $\mathrm{EC_a}$ data which are more or less regularly sampled but which do not constitute a comprehensively observed 'image'. We define the statistic $P(\tau, \Delta)$ as the expected proportion of observations within a square window of width Δ, centred at a randomly selected location $\mathbf{x}$ which are $\leq \tau$, conditional on the value at $\mathbf{x}$ being $\leq \tau$. We may expect these values to be smaller for a TG random function than for a function which better-represents the spatial structure of a variable in which small values tend to be spatially connected.

We estimated $P(\tau, \Delta)$ for the TG and PCLT random functions fitted to the $\mathrm{EC_a}$ data by simulation. These are denoted by $P_{\mathrm{TG}}(\tau, \Delta)$ and $P_{\mathrm{PCLT}}(\tau, \Delta)$ respectively. We considered windows of width 2 m or larger (because approximately 40 $\mathrm{EC_a}$ observations occur within a 2-m window). Each simulation program generated a single independent realization of the random function at 25 equally-spaced locations in a window of width Δ one of which was at the centre of the window. If the simulated value at the centre was $\leq \tau$, the conditioning criterion, then the realization was retained and $P(\tau, \Delta)$ was estimated as the proportion of the observations in the window for which $\leq \tau$. This was repeated until 10 000 independent realizations which met the criterion that the central value was $\leq \tau$ had been obtained. The PCLT realizations were generated using the procedure described by Lark (2012b). The TG realizations were obtained by LU decomposition (Goovaerts, 1997). The mean value of $P_{\mathrm{TG}}(\tau, \Delta)$ and the standard deviation of the 10 000 independent values, were computed for different values of Δ and for τ set to the median, first quartile and first octile of the $\mathrm{EC_a}$data. This was also done for $P_{\mathrm{PCLT}}(\tau, \Delta)$. The difference between the mean values of $P_{\mathrm{PCLT}}(\tau, \Delta)$ and $P_{\mathrm{TG}}(\tau, \Delta)$ for these different thresholds and for windows of different size, are plotted in Figure 9.

Figure 9 shows three things. First, the mean value of $P_{\mathrm{PCLT}}(\tau, \Delta)$ is larger than that of $P_{\mathrm{TG}}(\tau, \Delta)$ for given τ and δ. That is to say, given that a value falls below a threshold, there is a larger proportion of neighbouring values which do so for the PCLT process than for the TG process. Second, the effect depends on the threshold, and increases as the threshold becomes more extreme relative to the overall distribution. Third, the effect depends on the window size. It is small for a large window, but it is also notable that the difference is larger for the window width 4 m than the window width 2 m. This reflects the spatial scale of the random function.

The $P(\tau, \Delta)$ statistic was then estimated from the $\mathrm{EC_a}$ data for the same three threshold values used in the simulations, and for $\Delta = 4$m given that this window showed the largest differences between the two processes in the simulation. An independent random subsample of 250 observations for which $\mathrm{EC_a} \leq \tau$ was obtained, the proportion of $\mathrm{EC_a}$ observations within a square window, width Δ about each of these observations was computed. The results are shown in Figure 10. The mean value of $P_{\mathrm{TG}}(\tau, \Delta)$ and $P_{\mathrm{PCLT}}(\tau, \Delta)$ from the simulations are plotted, and for each of these the 95% confidence interval for the mean of a sample of 250 independent observations is also shown, based on the variances of the values obtained by simulation. The estimates from the $\mathrm{EC_a}$ data are also plotted. Note that for all three thresholds the values of $P(\tau, \Delta)$ for the data are larger than the upper limit of the confidence interval for the TG process. For τ equal to the median and the first quartile the values from the data are within the confidence interval for the PCLT process, for the first octile the estimate is slightly smaller than the confidence interval for the PCLT process, but closer to the expected value for the PCLT process than it is for the TG process.

3. Discussion

The overall objective of this study was to identify a stochastic model for a soil property that varies according to some mode, and to base this identification as far as possible on knowledge of the underlying soil process and, at most, some simple descriptive statistics of the variable such as the empirical variogram and summary statistics. This was achieved in this study by employing general soil knowledge in a structured way. This is proposed as a framework for similar studies on soil variation in contrasting modes.

The PCLT model used here is a stochastic model of soil variability selected because it represents a particular model of soil variation. This places it in between the most common approach to stochastic modelling, where a Gaussian or TG model is selected for convenience, and approaches based on direct specification of the form of the covariance function from a mechanistic model of the process. The latter has been achieved only for a limited set of processes over a limited range of spatial scales, e.g. Whittle (1954, 1962), Kolvos et al. (2004). Essentially the PCLT model is selected because it is in some sense an analogue of the soil process of interest. A similar approach has been used elsewhere. For example, Smith et al. (200) selected a 'blur' process based on convolution to model the space-time covariance of atmospheric pollutants, the convolution process was an analogue of pollutant dispersal. Similarly, Brochu and Marcotte (1993) selected a generalized Cauchy variogram model for observations on hydraulic head on the grounds that this process had physical analogies with a gravimetric field, which is mechanistically linked to the Cauchy model.

The use of stochastic geometric analogues of soil processes to generate stochastic models is attractive. It remains to be seen how wide a range of soil processes can be represented that way, and it is accepted that lateral textural variations in patterned ground are at once likely to be represented by simple geometric models and rather unrepresentative of soil variation in most conditions. None the less, the approach to the identification of models based on finding operators that are analogues for processes in the soil is likely to be more successful than the search for stochastic models based on strictly mechanistic models. It must also be noted that the stochastic geometric approach naturally reproduces non-Gaussian variation which must be characterized by moments of order higher than two, whereas the mechanistic approaches to spatial modelling are often explicitly based on two-point statistics, the covariance function (e.g. Whittle, 1954; 1962).

The particular advantage of the stochastic geometric approach in this case study is how the inverse-distance PCLT model was better than the TG model in terms of the test statistic on the connectivity of values with small $\mathrm{EC_a}$. If one wanted to generate conditional simulations of the soil in this environment as a basis for computing, for example, distributions of upscaled processes such as pollutant transport across a block of land, then the inverse-distance PCLT model would produce superior representations of the connectivity of material with large conductivities, and so of preferential flow pathways.

There is considerable scope for further development of this approach. Other distance functions could be considered for this variable, and for others. In this study we looked for the simplest distance function that seemed to be compatible with soil knowledge, and there may be scope further to refine a framework for selecting a function. More specific soil knowledge could be used. For example, in the case study considered here, one could generate a simple conceptual 3D model of a polygon, with material with different dielectrical properties, and compute the expected trend function from models of the EM properties of the soil. While the objective of this particular study was to restrict the use of direct observations on the target variable to simple descriptive statistics, one might also conduct specific surveys at fine scale on transects across polygons in order to identify plausible distance functions for further studies. It should also be noted that the homogeneous Poisson process, while a default spatial model, is not the only one available and might not be generally appropriate. While it was selected in this case on the basis of recent work on patterned ground (Cresto Aleina et al., 2012), it is likely that, at the limit, a more overdispersed spatial process would be more appropriate for this problem, with fewer close-spaced points than in the homogeneous Poisson case.

The model-fitting framework in this study made combined use of point estimates of the variogram, and a weighted least squares criterion for parameter estimation, subject to constraints identified from soil knowledge which imposed constraints on the modelled parameters based, in this case, on the coefficient of skewness. This remains a somewhat arbitrary procedure for parameter estimation. Ideally a likelihood-based estimator should be derived. This is unlikely to be straightforward, not least because the joint distribution function of any PCLT process is complex and requires geometrical functions for which analytical expressions are not known. In other settings, when the likelihood function is expensive to evaluate, parameters may be estimated by an extension of the method of moments to include higher order moments than the usual first and second. An example of this is given by Iskander and Zoubir (1999), and it is suggested that a method of higher-order moments is most likely to be a tractable solution to fitting stochastic geometric models.

There is scope for further work on the comparison of realizations of the PCLT and TG processes with respect to multiple point statistics and for weighing the evidence that one model rather than the other best represents particular data. We used a relatively simple statistic in this paper, given that our data are not-quite regularly sampled and so do not constitute an image. However, it would be interesting to see how statistics developed for images (e.g. De Iaco and Maggio, 2011) might be used to evaluate alternative stochastic models. That said, the statistic which we used in this paper was not a general measure of spatial structure but rather was focussed on a particular problem of direct interest (i.e. the connectedness of areas likely to have larger hydraulic conductivities). This is arguably more relevant than a generalized measure. It would be interesting to develop methods to quantify the spatial structure of random fields as this affects particular processes. For example, one might compare the outcomes of a process model for the dispersal of contaminant plumes when it is run with input data on conductivity or similar model parameters which are realizations of contrasting random processes.

Any PCLT model could be used in conventional spatial prediction by kriging since the variogram or, equivalently, the covariance function can be specified. However, since the PCLT covariance function is not available in closed form, it would generally be more efficient to use a standard variogram function for kriging; and since kriging uses only the two-point statistics of a variable there is unlikely to be any benefit in using the PCLT model rather than a standard spatial model for this purpose. The value of the PCLT model is not to provide an alternative form of the covariance function, but rather for spatial prediction of non-Gaussian variables whose multivariate distribution is not entirely characterized by the covariance function. Spatial prediction in such cases may be may be done by codes such as SNESIM (Strebelle, 2002) or the direct sampling (DS) algorithm of Mariethoz et al. (2010) which allow one to obtain conditional distributions at unsampled sites from multiple realizations of a non-Gaussian process. These procedures require training images of the variables of interest, and the availability of sufficient training images of adequate quality is a potential limitation on the use of multiple point geostatistical methods in soil science. For this reason Pyrcz et al. (2008) developed a library of training images for a particular geological setting (fluvial and deepwater reservoirs) by a combination of stochastic and object-based simulation methods. If an appropriate PCLT process could be identified for a particular soil variable, then it might be used similarly to generate training images, either for a library or as required for a multiple point conditional simulation. It is easy to generate multiple training images from a PCLT model. This would be particularly advantageous for the DS algorithm, because it has been noted (e.g. Meerschman et al., 2013) that multiple realization generated by the DS algorithm sometimes all include exact copies of significant patches of the (single) training image. This could be avoided by modifying the DS algorithm to sample multiple training images in random order, when these can readily be generated.

4. Conclusions

We have shown how a structured use of soil knowledge allows us to fit an appropriate stochastic geometric model to data on a soil property in a particular environment. Furthermore, we have shown that this model appears to capture features of the spatial variation of our target variable better than the standard Gaussian model, even after transformation of the data to marginal normality. There is more work to be done in the development of this approach, and exploring its practical implications but we believe this case study shows that there is considerable potential. In particular, realizations of PCLT processes may be better than standard TG simulations for predicting outcomes of non-linear processes such as contaminant transport, and for quantifying the uncertainty of such predictions. If PCLT models succeed in capturing the multiple point behaviour of soil variables, then PCLT simulation can be used to provide an inexhaustible supply of training images for existing multiple point prediction code. This removes one major limitation on the application of this emerging geostatistical methodology.

Acknowledgements

This paper is published with the permission of the Director of the British Geological Survey (Natural Environment Research Council). The second author was funded by the Fund for Scientific Research-Flanders (FWO-Vlaanderen). We are grateful to Professor Alex McBratney and to two anonymous referees for helpful comments on this paper.

References

Brochu, Y., Marcotte, D. 2003. A simple approach to account for radial flow and boundary conditions when kriging hydraulic head fields for confined aquifers. Mathematical Geology, 35, 111–139.

Brown, P.E., Roberts, G.O., Kåresen, K.F., Tonellato, S. 2000. Blur-generated non-separable space-time models. Journal of the Royal Statistical Society B, 62, 847–860.

Catt, J.A. 1979. Soils and Quaternary Geology in Britain. Journal of Soil Science 30, 607–642.

Chilès, J.P., Guillen, A., 1984. Variogrammes et krigeages pour la gravimétrie et le magnétisme. Sciences de la Terre – Série Informatique Géologique, 20, 455–468.

Cockx, L., Ghysels, G., Van Meirvenne, M., Heyse, I. 2006. Prospecting frost-wedge pseudomorphs and their polygonal network using the electromagnetic induction sensor EM38DD. Permafrost and Periglacial Processes, 17, 163–168.

Cressie, N. 1985. Fitting variogram models by weighted least squares. Mathematical Geology, 17, 563–586.

Cresto Aleina, F., Brovkin, V., Muster, S., Boike, J., Kutzbach, L., Zuyev, S. 2012. Poisson-Voronoi diagrams and the polygonal tundra. Geophysical Research Abstracts, 14, EGU2012-1963-1.

De Iaco, S., Maggio, S. 2011. Validation techniques for geological patterns simulations based on variogram and multiple point statistics. Mathematical Geoscience 43, 483–500.

Deutsch, C.V., Journel, A.G. 1997. GSLIB: Geostatistical software library and user's guide. 2nd Edition. Oxford University Press, New York.

Dudewicz, E.J., Mishra, S.N. 1988. Modern mathematical statistics. John Wiley & Sons, New York.

Goovaerts, P. 1997. Geostatistics for natural resources evaluation. Oxford University Press, New York.

Gómez-Hernández, J.J., Wen, X-H. 1998. To be or not to be multi-Gaussian? A reflection on stochastic hydrology. Advances in Water Resources, 21, 47–61.

Guardiano, F., Srivastava, R.M., 1993. Multivariate geostatistics: beyond bivariate moments. In: Soares, A. (Ed.), Geostatistics-Troia, Vol. 1. Kluwer, Dordrecht, pp. 133–144.

Harding, S.A., Murray, D.A., Webster, R. 2010. MVARIOGRAM procedure in (ed.) R.W. Payne, GenStat Release 13 Reference Manual, Part 3 Procedure Library PL21. VSN International, Hemel Hempstead.

Kachanoski, R.G., Hendrickx, J.M.H., de Jong, E. 2002. Electromagnetic induction. In: (eds) J.H. Dane, G.C. Topp. Methods of Soil Analysis, Part 1, Physical Methods, Third Edition. Soil Science Society of America. (2002), pp.497–501.

Iskander, R., Zoubier, A.M. 1999. Estimation of the parameters of the K-distribution using higher order and fractional moments. IEEE Transactions on Aerospace and Electronic Systems, 35, 1453–1456.

Kolvos, A., Christakos, G., Hristopulos, D.T., Serre, M.L. 2004. Methods for generating non-separable spatiotemporal covariance models with potential environmental applications. Advances in Water Resources 27, 815–830.

Lark, R.M. 2009. A stochastic-geometric model of soil variation. European Journal of Soil Science, 60, 706–719.

Lark, R.M. 2012a. Towards soil geostatistics. Spatial Statistics 1, 92–99.

Lark, R.M. 2012b. A stochastic geometric model for continuous local trends in soil variation. Geoderma, 189–190, 661–670.

Mariethoz, G., Renard, P., Straubhaar, J. 2010. The Direct Sampling method to perform multiple-point geostatistical simulations. Water Resources Research 46, W11536.

Matheron, G. 1962. Traité de Géostatistique Appliqué, Tome 1. Memoires du Bureau de Recherches Géologiques et Minières, Paris.

McBratney, A.B., Webster, R. 1986. Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. Journal of Soil Science 37, 617–639.

Meerschman, E., Van Meirvenne, M., De Smedt, T., Islam, M.M., Meeuws, F., Van De Vijver, E., Ghysels, G. 2011. Imaging a polygonal network of ice-wedge casts with an electromagnetic induction sensor. Soil Science Society of America Journal 75, 2095–2100.

Meerschman, E., Pirot, G., Mariethoz, G., Straubhaar, J., Van Meirvenne, M., Renard, P. 2013. A practical guide to performing multiple-point statistical simulations with the Direct Sampling algorithm. Computers & Geosciences, 52, 307–324.

Okabe, A., Boots, B., Sugihara, K., Chiu, S.K. 2000. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams. 2nd Edition John Wiley & Sons, Chichester.

Payne, R.W. (ed) 2010. GenStat Release 13 Reference Manual, Part 2 Directives. VSN International, Hemel Hempstead.

Pyrcz, M.J., Boisvert, J.B., Deutsch, C.V. 2008. A library of training images for fluvial and deepwater reservoirs and associated code. Computers & Geosciences 43, 542–560.

R Development Core Team. 2012. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`.

Saey, T., Van Meirvenne, M., Vermeersch, H., Ameloot, N., Cockx, L. 2009. A pedotransfer function to evaluate the soil profile textural heterogeneity using proximally sensed apparent electrical conductivity. Geoderma, 150, 389–395.

Strebelle, S. 2002. Conditional simulation of complex geological structures using multiple-point statistics. Mathematical Geology, 34, 1–21.

Venables, W.N., Ripley, B.D. 2002. Modern Applied Statistics with S. Fourth Edition. Springer, New York.

Visual Numerics, 2006. IMSL Fortran Numerical Library Version 6.0. Visual Numerics, Houston, Texas.

Walters, J.C. 1994. Ice-wedge casts and relict polygonal patterned ground in North-East Iowa, USA. Permafrost and Periglacial Processes 5, 269–281.

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

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

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

Table 1. Summary statistics of the raw data on EC_a.

Statistic	mS m^{-1}
Average	31.37
Median	31.13
Standard deviation	2.2
Skewness	0.36
Quartile 1	29.9
Quartile 3	32.76
Octile 1	29.03
Octile 7	34.08

Table 2. Summary statistics of the data on EC_aafter the Box-Cox transformation and for the transformed data after standardization. Variogram parameters for the standardized data are also given.

Statistic	Transformed data	Transformed and standardized data
Average	1.508	0
Median	1.507	-0.056
Standard deviation	0.01	1
Skewness	0	0
Quartile 1	1.501	-0.646
Quartile 3	1.514	0.668
Octile 1	1.497	-1.085
Octile 7	1.52	1.216
ζ^*	-0.57	
Variogram parameters†		
c_0		0.12
c_1		0.84
a		1.91
κ		1.49

* Maximum likelihood estimate of the parameter of the Box-Cox transform, see Eq.(1)

† Powered (stable) exponential model, see Eq.(2).

Figure captions

1. EC_a data, coordinates are in metres relative to a local datum.

2. Histogram of EC_a data.

3. Empirical variogram of transformed and standardized EC_a data with a fitted model.

4. Values of the coefficient of skewness for an inverse-distance PCLT process with different values of the parameters λ and α. The two contours bound the region where we regard the variable as mildly positively skewed.

5. Profile plot of the weighted sum of squares for the fit of the inverse-distance PCLT variogram function against λ, with α fixed at 2.5 m.

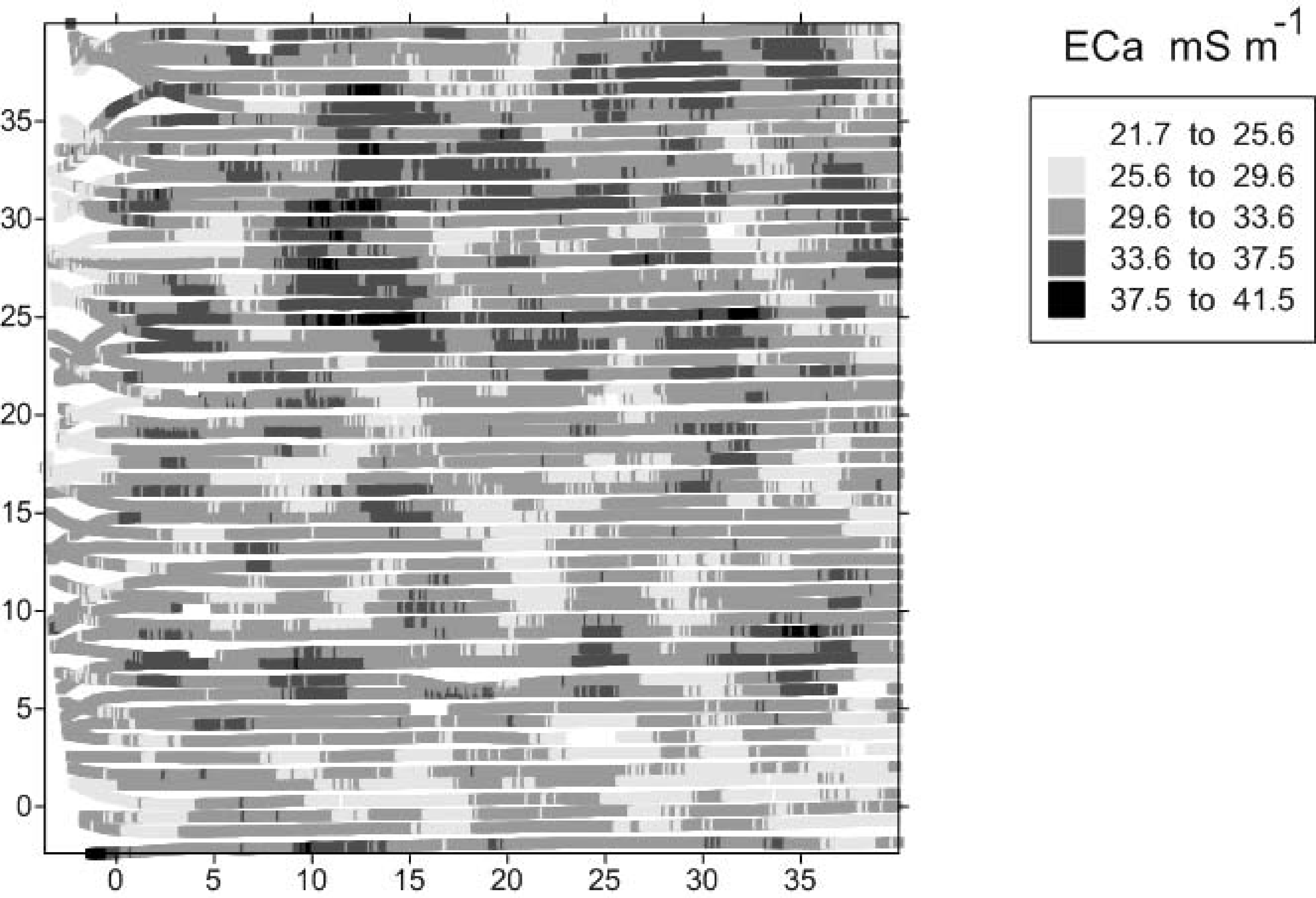
6. Empirical variogram of the untransformed EC_a data with the fitted inverse-distance PCLT variogram.

7. Marginal distribution function of the standardized inverse-distance PCLT random function with α=2.5 m and λ=0.07 m^{-2} (line). The points are from the empirical cumulative distribution function of the standardized EC_a data.

8. Realizations of (a) the inverse-distance PCLT random function and (b) the TG random function (back transformed to original units) on a 0.25-m square grid.

9. Plot of the difference between the mean of $P_{\mathrm{PCLT}}(\tau, \Delta)$ and that of $P_{\mathrm{TG}}(\tau, \Delta)$ for different window widths (Δ) and with τ set to the median, first quartile and first octile of the EC_adata. Mean for 10 000 realizations of each random function.

10. $P(\tau, \Delta)$ with $\Delta = 4$ m plotted against τ set to the median, first quartile or first octile. The solid disc, •, is the mean value from 10 000 realizations of the PCLT random function, the solid square, ■, is the mean value from 10 000 realizations of the TG random function. The horizontal bars show the 95% confidence interval for the mean of based on 250 independently and randomly selected locations that mean the conditioning criteria. The crosses, × show the mean values for 250 independently and randomly selected sites in the EC_a data set.



ECa mS m^{-1}
21.7 to 25.6
25.6 to 29.6
29.6 to 33.6
33.6 to 37.5
37.5 to 41.5
0
5
10
15
20
25
30
35

Figure 2

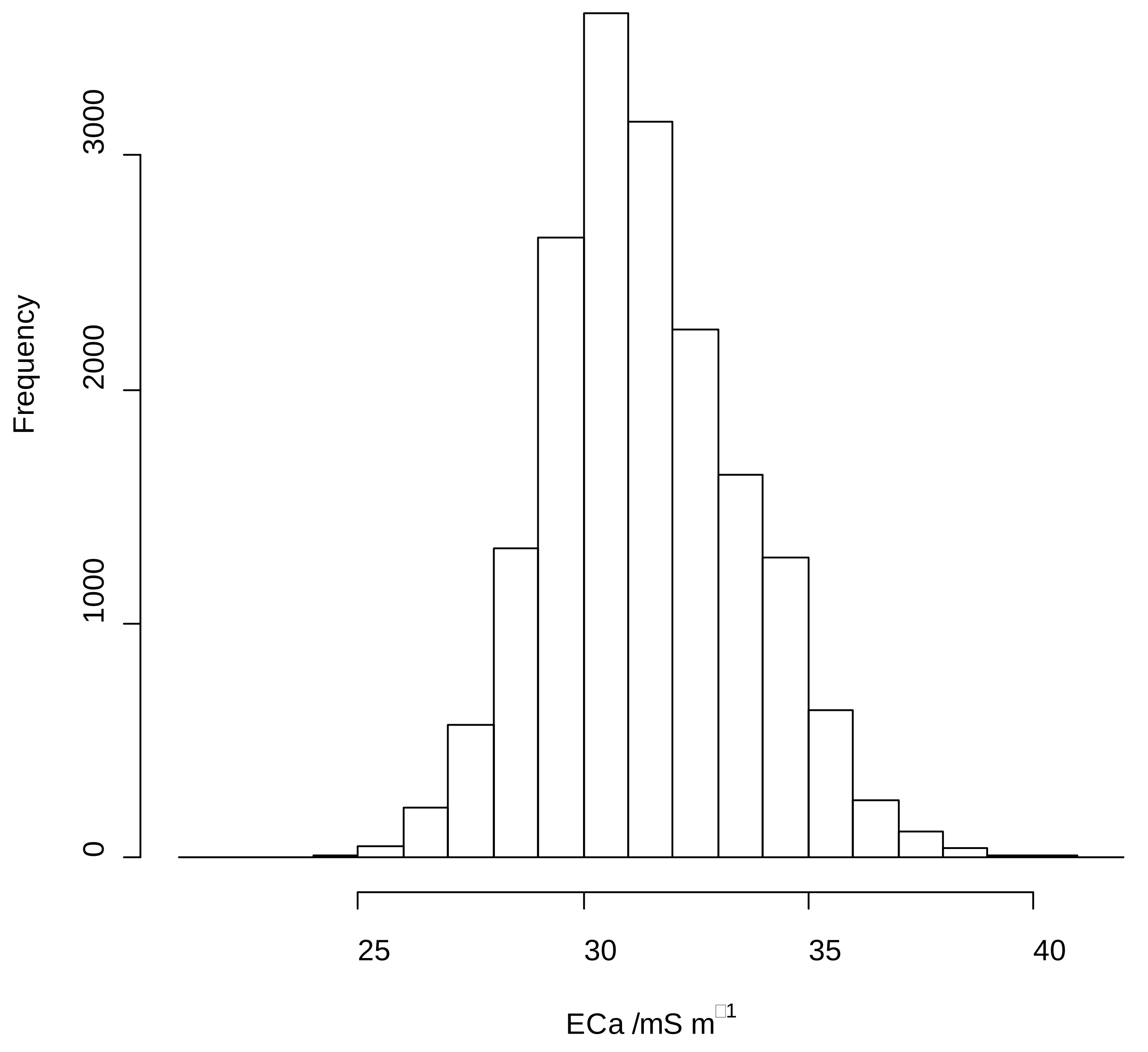
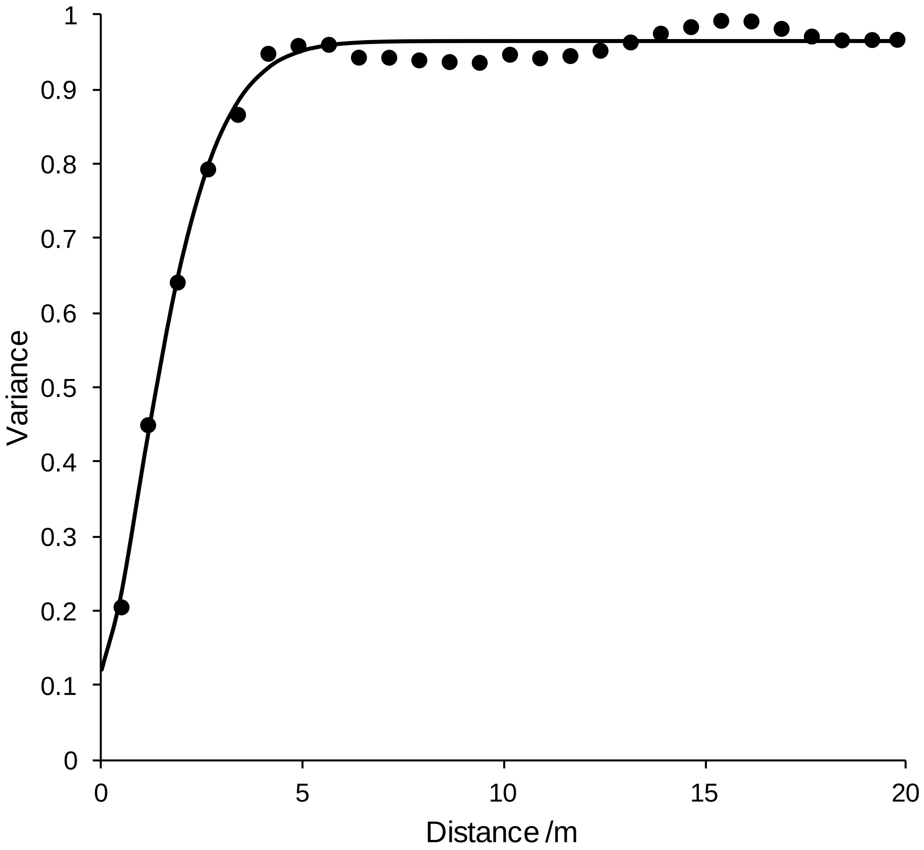
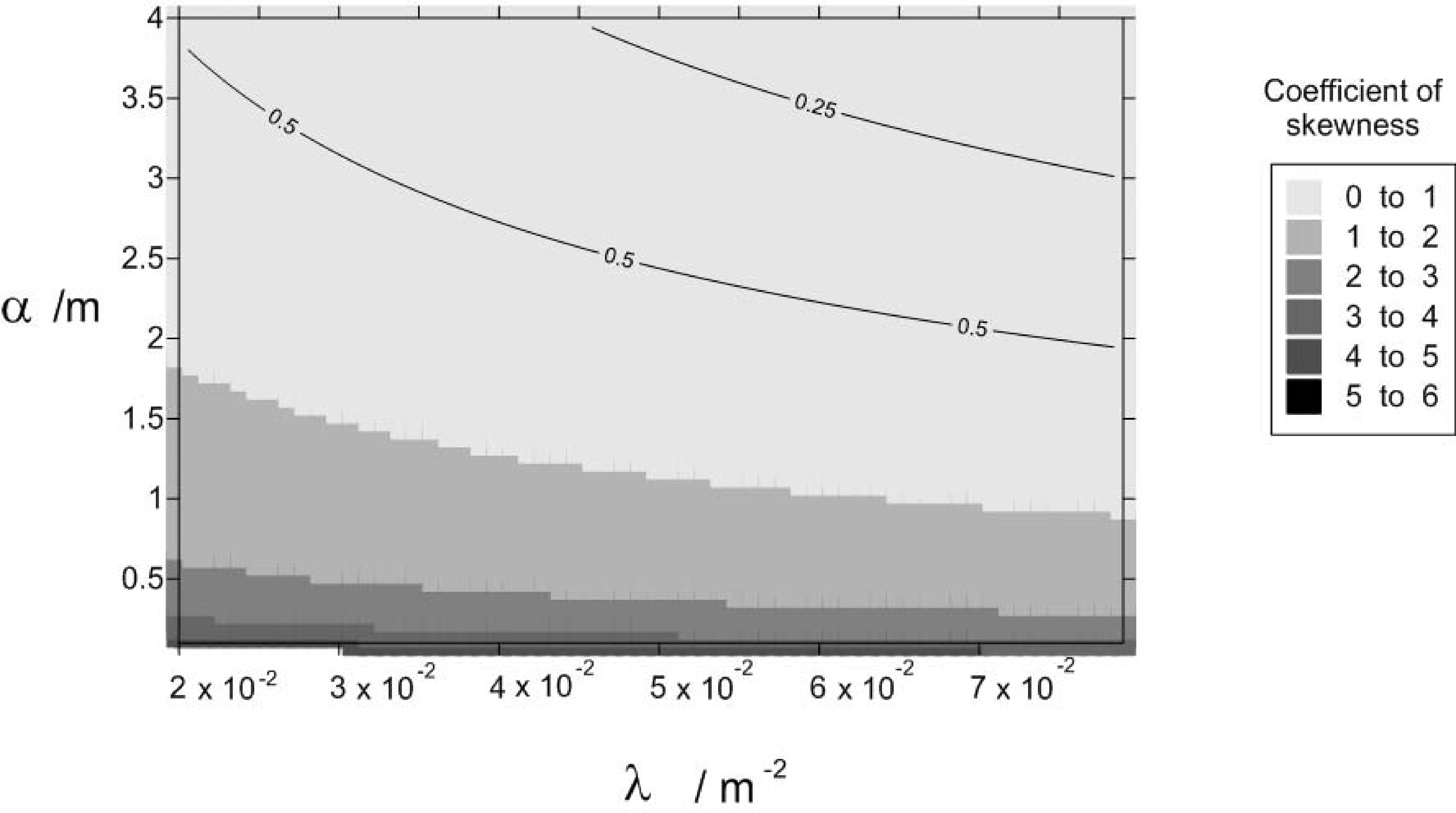


Figure 3





4
3.5
3
2.5
2
1.5
1
0.5
α /m
0.5
0.25
2 x 10^-2
3 x 10^-2
4 x 10^-2
5 x 10^-2
6 x 10^-2
7 x 10^-2
λ / m^-2
Coefficient of skewness
0 to 1
1 to 2
2 to 3
3 to 4
4 to 5
5 to 6

Figure 5

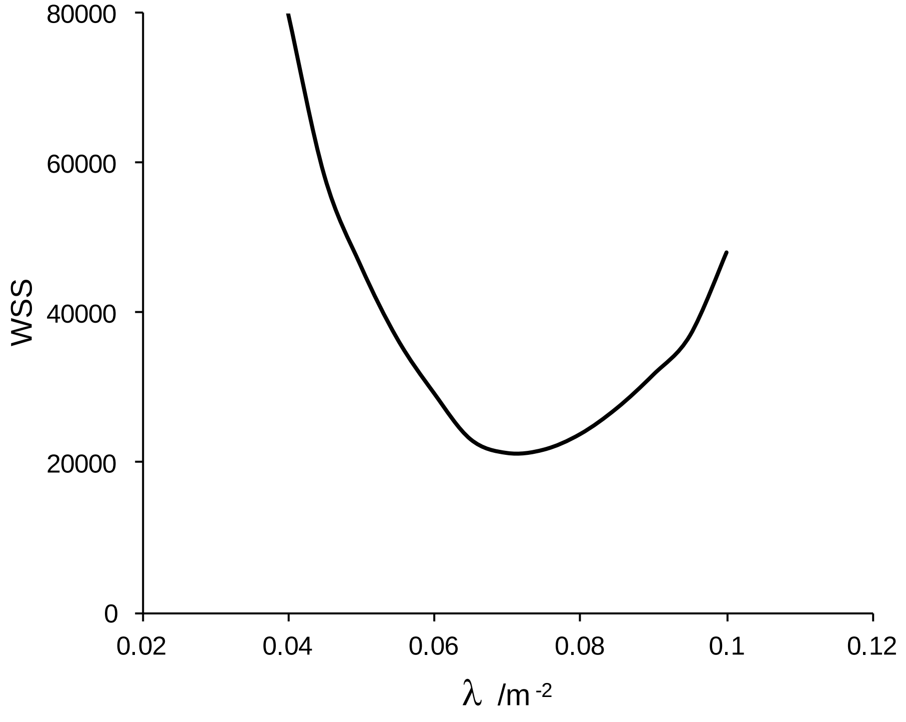


Figure 6

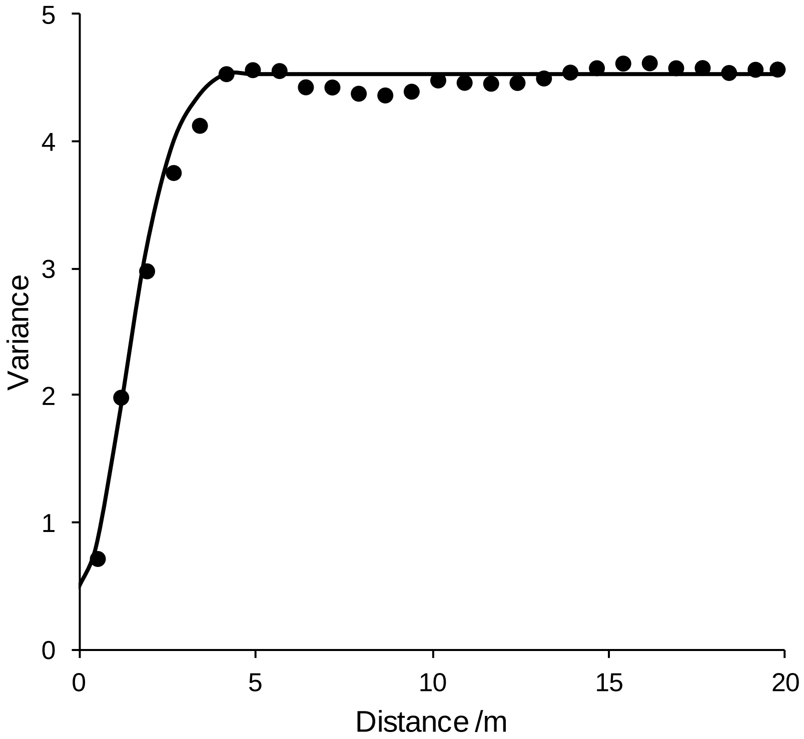
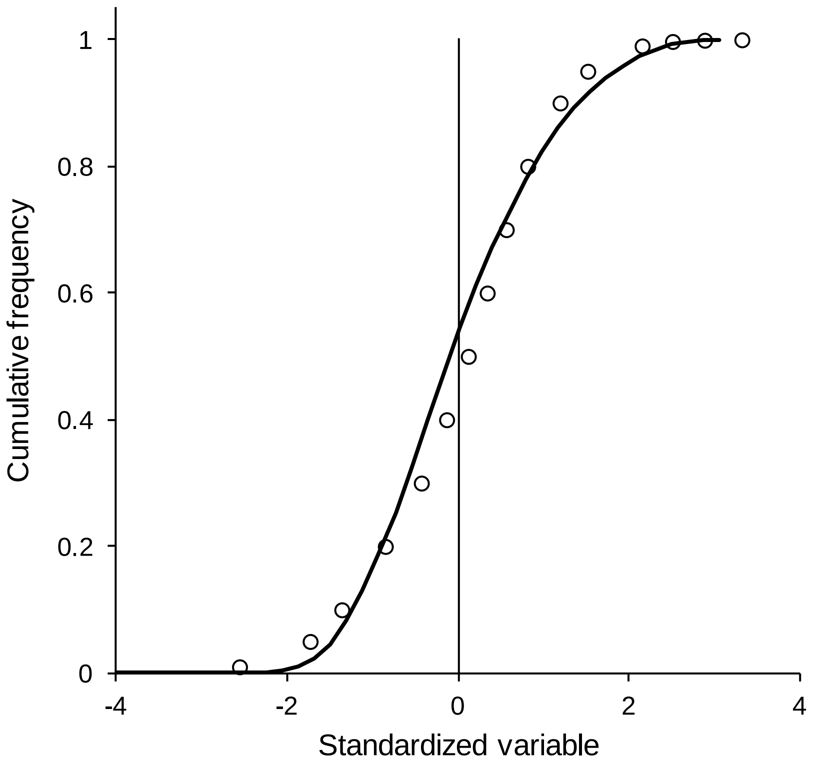
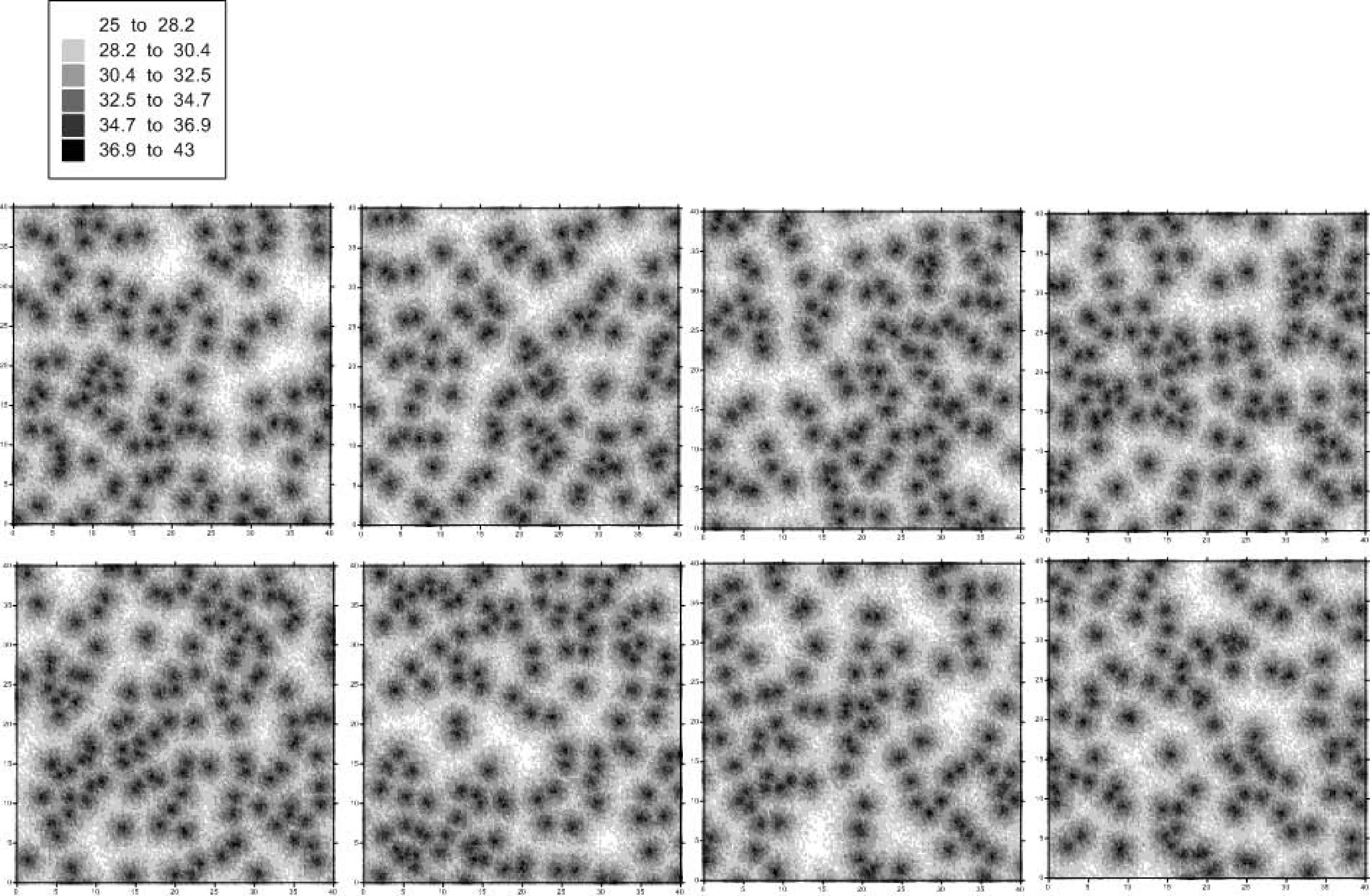
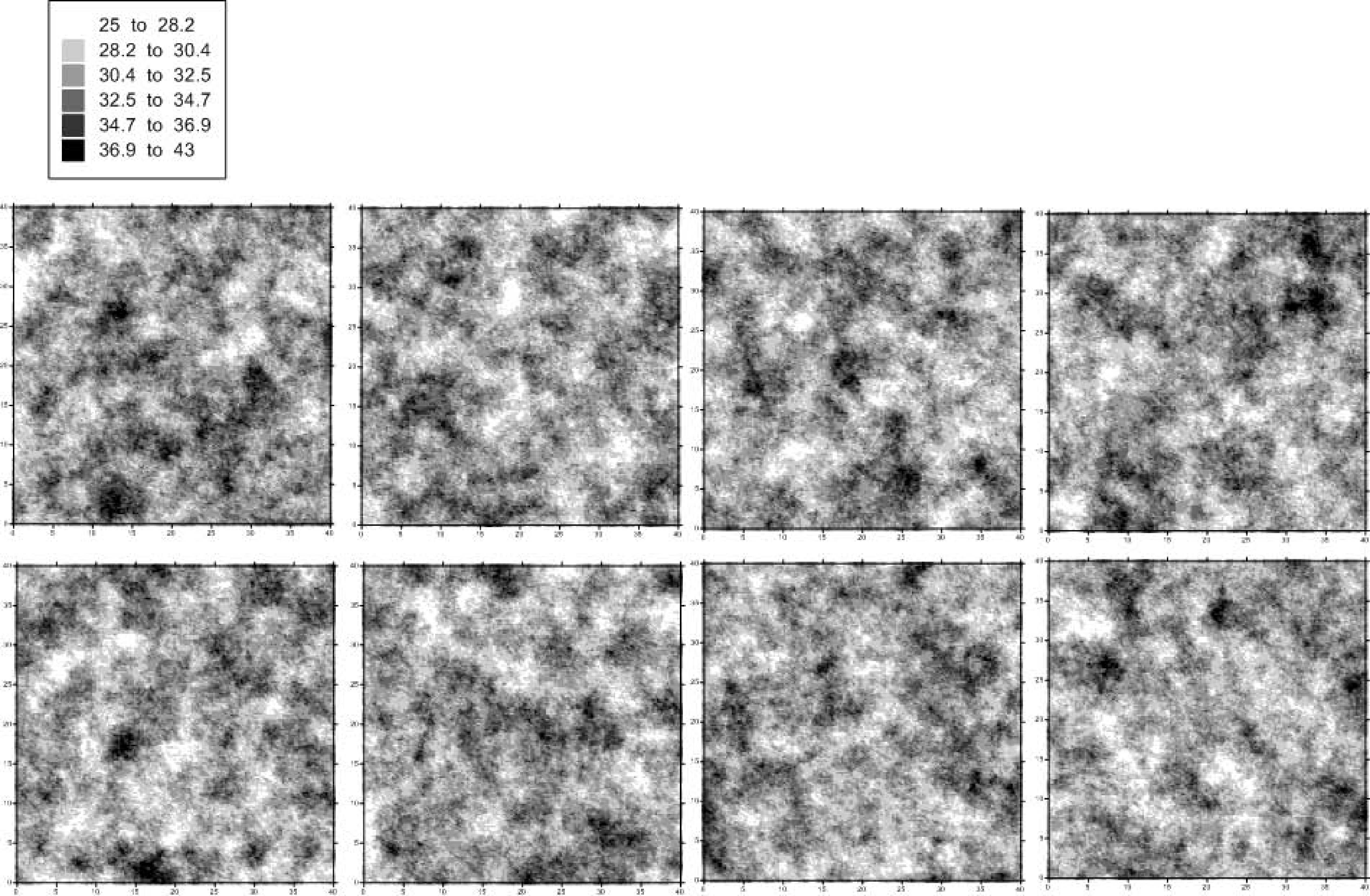


Figure 7

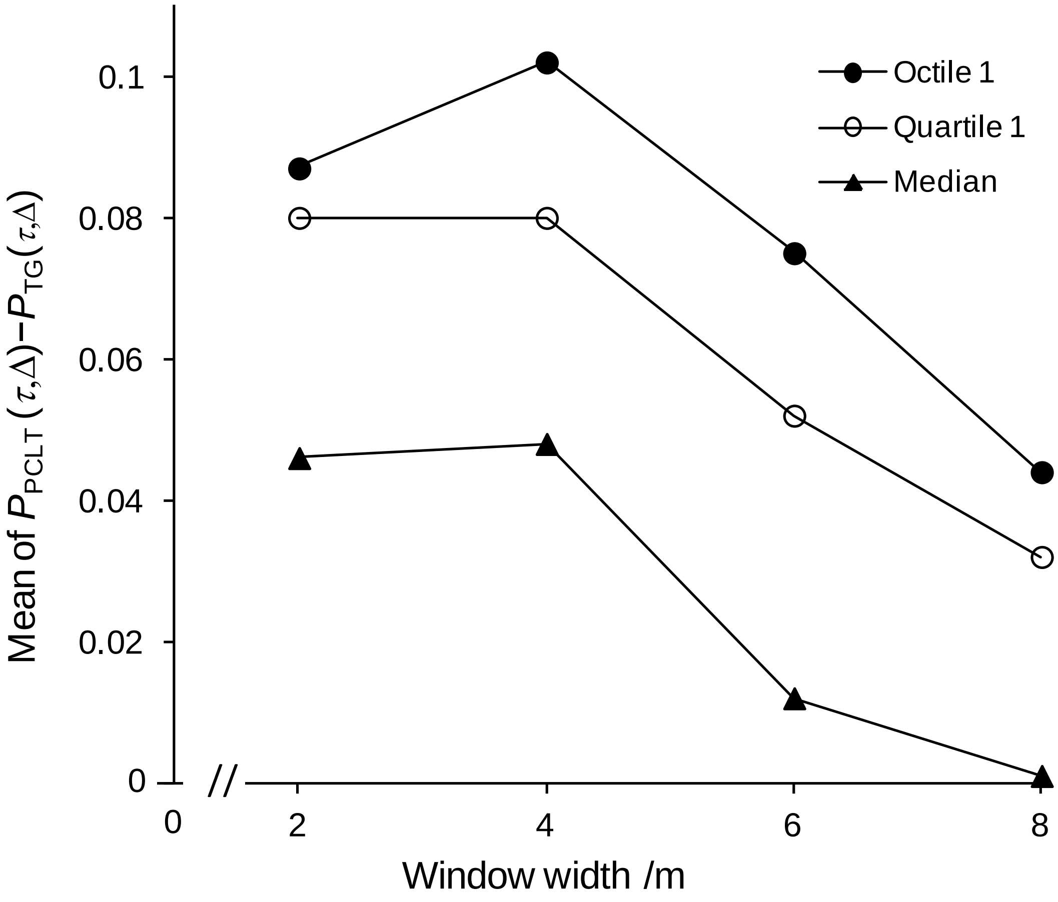

Cumulative frequency
1
0.8
0.6
0.4
0.2
0
-4
-2
0
2
4
Standardized variable


25 to 28.2
28.2 to 30.4
30.4 to 32.5
32.5 to 34.7
34.7 to 36.9
36.9 to 43



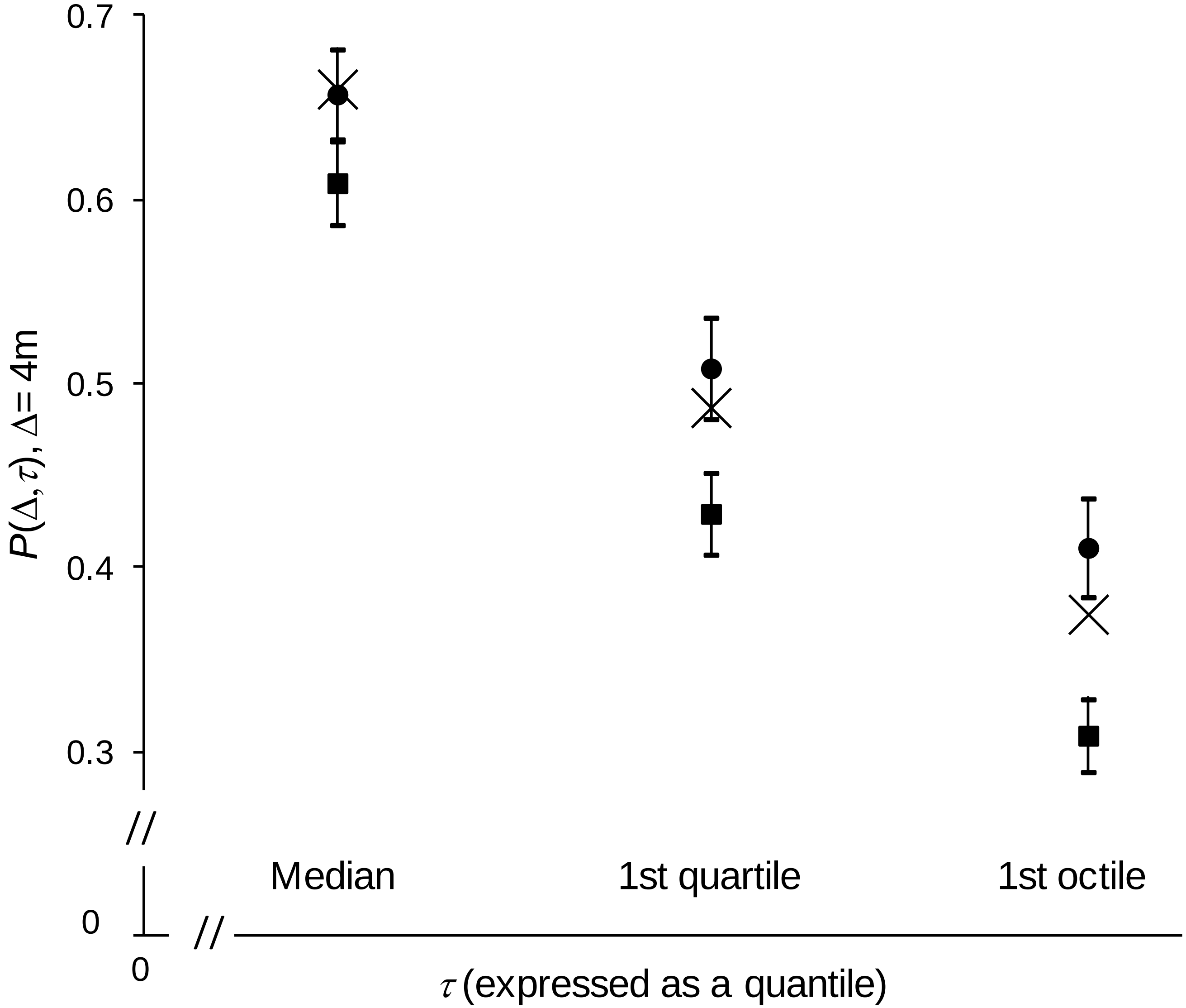
25 to 28.2
28.2 to 30.4
30.4 to 32.5
32.5 to 34.7
34.7 to 36.9
36.9 to 43

Figure 9



Octile 1
Quartile 1
Median
0.1
0.08
0.06
0.04
0.02
0
0
2
4
6
8
Window width /m
Mean of $P_{PCLT}(\tau,\Delta)-P_{TG}(\tau,\Delta)$

Figure 10


0.7
0.6
0.5
0.4
0.3
0
$P(\Delta,\tau)$, Δ= 4m
Median
1st quartile
1st octile
0
τ (expressed as a quantile)