Using third-order cumulants to investigate spatial variation: a case study on the porosity of the Bunter Sandstone

R.M. Lark*

British Geological Survey, Keyworth, Nottinghamshire NG12 5GG, U.K.

Abstract

The multivariate cumulants characterize aspects of the spatial variability of a regionalized variable. A centred multivariate Gaussian random variable, for example, has zero third-order cumulants. In this paper it is shown how the third-order cumulants can be used to test the plausibility of the assumption of multivariate normality for the porosity of an important formation, the Bunter Sandstone in the North Sea. The results suggest that the spatial variability of this variable deviates from multivariate normality, and that this assumption may lead to misleading inferences about, for example, the uncertainty attached to kriging predictions.

---

1. Introduction

Geostatistical analysis of spatially variable geological data allows us to quantify the uncertainties in inferences made from partial samples by treating data as realizations of a random field. In most cases the underlying model is multivariate Gaussian, and the plausibility of this assumption is usually judged from the marginal distribution of observations (e.g. Webster and Oliver, 2007). Where necessary the data may be transformed, for example to logarithms or, more generally, by the Box-Cox transformation. However, it is recognized that the assumption of a Gaussian or trans-Gaussian (Gaussian after transformation) distribution is not always safe, and, particularly, that it might not hold even when it seems plausible for the marginal distribution of the data. Of particular concern is the recognition that, under the multivariate Gaussian model, the first and second order

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

moments entirely characterize the spatial distribution of a variable since all odd moments larger than the first are zero and all even moments larger than the second can be written in terms of it. However, it is known that the complex geometries that may be encountered in geological data, the strongly-connected patterns of coarse-textured alluvium in former braided streams are a *locus classicus*, might not be fully characterized by the first and second moments, and more complex spatial distributions are necessary (e.g. Guardiano and Srivastava, 1993).

It is therefore necessary to develop exploratory methods to examine the higher-order behaviour of spatially variable data. Dimitrakopoulos et al. (2010) have shown how higher order spatial cumulants of random variables can capture features of dense training images that are not compatible with the assumption of an underlying multivariate-Gaussian variable. The objective of the present paper is to show how such a cumulant can be used in an inferential framework to test the strength of evidence against the null hypothesis that, possibly relatively sparse, observations are drawn from a variable in which these cumulants take values expected in the Gaussian case; and to identify exploratory statistics that might be used to judge whether a Gaussian assumption is plausible. The approach is illustrated using data on porosity of an important sedimentary formation under the North Sea. A sound spatial stochastic model for this variable is necessary because the pore-space in this unit may be important as a site for future carbon capture and storage (Holloway, 2009).

2. Cumulants

A real-valued random variable, Z, with a probability density function $f_Z(z)$, has a moment-generating function:

$$M(v) \;=\; \mathrm{E}\left[\exp\{vZ\}\right] \;=\; \int_{-\infty}^{\infty} \exp\{vz\} f_Z(z) \mathrm{d}z. \tag{1}$$

If $M(v)$ has a Taylor series expansion about the origin then it may be written as

$$M(v) \;=\; \mathrm{E}\left[\exp\{vZ\}\right] \;=\; \mathrm{E}\left[1 + vZ + \frac{v^2}{2!}Z^2 + \ldots + \frac{v^r}{r!}Z^r + \ldots\right]. \tag{2}$$

Note that the rth non-centred moment of Z,

$$\mu'_r \; = \; \mathrm{E}\left[Z^r\right],$$

is the coefficient of $\frac{v^r}{r!}$ in the rth term in this expansion, hence the name of the function.

Cumulants of the random variable may be defined in a similar and related way. The cumulant generating function is

$$K(v) \; = \; \ln\left(\mathrm{E}\left[\exp\{vZ\}\right]\right),$$

so we may write

$$1 + \mu'_1 \frac{v}{1!} + \mu'_2 \frac{v^2}{2!} + \ldots + \mu'_r \frac{v^r}{r!} + \ldots = \; \exp\left\{\kappa_1 \frac{v}{1!} + \kappa_2 \frac{v^2}{2!} + \ldots + \kappa_r \frac{v^r}{r!} + \ldots\right\}, \tag{3}$$

where κ_r is the rth cumulant of Z.

The cumulants and moments of a distribution are related, for example (Kendall and Stuart, 1977)

$$\begin{aligned} \mu'_1 &= \kappa_1, \\ \mu'_2 &= \kappa_1^2 + \kappa_2, \\ \mu'_3 &= \kappa_1^3 + 3\kappa_1\kappa_2 + \kappa_3. \end{aligned} \tag{4}$$

However, cumulants have certain properties which can make them more useful than moments. In particular they generalize simply to the multivariate case (McCullagh and Kolassa, 2009). Consider an n-variate random variate $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_n\}$. One may define entries in the mean vector of Z, the matrix of second non-centred moments and the array of non-centred third moments as

$$\begin{aligned} \mathrm{E}_r &= \mathrm{E}[Z_r] \\ \mathrm{E}_{rs} &= \mathrm{E}[Z_r Z_s] \\ \mathrm{E}_{rst} &= \mathrm{E}[Z_r Z_s Z_t] \end{aligned} \tag{5}$$

We denote linear combinations of the variables in $\mathbf{Z}$, and the powers of this term using Einstein's simplified convention for notation of multiple summations (Kuptsov, 2001):

$$v_r Z_r \equiv \sum_{r=1}^{n} v_r Z_r \tag{6}$$

$$v_r v_s Z_r Z_s \equiv \sum_{r=1}^{n}\sum_{s=1}^{n} v_r v_s Z_r Z_s = (v_r Z_r)^2 \tag{7}$$

where the term $v_r Z_r$ on the right is defined in Eq [6],

$$v_r v_s v_t Z_r Z_s Z_t \equiv \sum_{r=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n} v_r v_s v_t Z_r Z_s Z_t = (v_r Z_r)^3 \text{ etc.} \tag{8}$$

Given this notation, the multivariate moment-generating function can be expanded as

$$M(v) = 1 + v_r E_r + \frac{v_r v_s E_{rs}}{2!} + \ldots \tag{9}$$

and, similarly,

$$K(v) = \ln(M(v)) = \kappa^r \frac{v_r}{1!} + \kappa^{r,s} \frac{v_r v_s}{2!} \ldots. \tag{10}$$

As in the univariate case, the cumulants of increasing order, $\kappa^r, \kappa^{r,s}, \ldots$ appear as coefficients in the expansion. The moments and cumulants in the multivariate case are found to be related in a simple way, the moments of some order are given by the sum of products of cumulants over partitions of the superscripts so, for moments and cumulants of order up to three:

$$\mathrm{E}_r = \kappa^r, \tag{11}$$

$$\mathrm{E}_{rs} = \kappa^{r,s} + \kappa^r \kappa^s, \tag{12}$$

and

$$\mathrm{E}_{rst} = \kappa^{r,s,t} + \kappa^{r,s}\kappa^t + \kappa^{r,t}\kappa^s + \kappa^{s,t}\kappa^r + \kappa^r \kappa^s \kappa^t. \tag{13}$$

The expressions above can be rearranged to express the cumulant of order k as functions of moments of order $m \leq k$ and cumulants of order $< k$:

$$\kappa^r = \kappa^r, \tag{14}$$

$$\kappa^{r,s} = \mathrm{E}_{rs} - \kappa^r \kappa^s, \tag{15}$$

and, rearranging Eq [13] and substituting Eq [15] for the second-order cumulants,

$$\begin{aligned}
\kappa^{r,s,t} &= \mathrm{E}_{rst} - \kappa^{r,s}\kappa^t - \kappa^{r,t}\kappa^s - \kappa^{s,t}\kappa^r - \kappa^r\kappa^s\kappa^t, \\
&= \mathrm{E}_{rst} - \left[\mathrm{E}_{rs} - \kappa^r\kappa^s\right]\kappa^t - \left[\mathrm{E}_{rt} - \kappa^r\kappa^t\right]\kappa^s - \left[\mathrm{E}_{st} - \kappa^s\kappa^t\right]\kappa^r - \kappa^r\kappa^s\kappa^t, \\
&= \mathrm{E}_{rst} - \mathrm{E}_{rs}\kappa^t - \mathrm{E}_{rt}\kappa^s - \mathrm{E}_{st}\kappa^r + 2\kappa^r\kappa^s\kappa^t, \\
&= \mathrm{E}_{rst} - \mathrm{E}_{rs}\mathrm{E}_t - \mathrm{E}_{rt}\mathrm{E}_s - \mathrm{E}_{st}\mathrm{E}_r + 2\mathrm{E}_r\mathrm{E}_s\mathrm{E}_t.
\end{aligned} \tag{16}$$

For zero mean $\mathbf{Z}$ Eq[15] and Eq[16] simplify to

$$\kappa^{r,s} = \mathrm{E}_{rs} = \mathrm{Cov}\left[Z_r, Z_s\right], \tag{17}$$

where $\mathrm{Cov}\left[\cdot,\cdot\right]$ denotes the covariance of the terms in the brackets, and

$$\kappa^{r,s,t} = \mathrm{E}_{rst}, \tag{18}$$

i.e. the third cumulant is equal to the third moment. This is zero for multivariate Gaussian $\mathbf{Z}$. In fact all multivariate cumulants of order $m > 2$ are zero for the Gaussian case (Bilodeau and Brenner, 1999). This is demonstrated for the fourth cumulant in the appendix.

Dimitrakopoulos et al. (2010) describe the extension of multivariate cumulants to the spatial random field $\mathbf{Z}(\mathbf{x})$. Consider the third-order cumulant. Given some location $\mathbf{x}$ we may define a set of three locations $\{\mathbf{x}, \mathbf{x}+\mathbf{h}_1, \mathbf{x}+\mathbf{h}_1+\mathbf{h}_2\}$ where $\mathbf{h}_1$ and $\mathbf{h}_2$ are lag vectors such that $\mathbf{h}_1 = h_1\mathbf{l}_1$ and $\mathbf{h}_2 = h_2\mathbf{l}_2$ where h_1 and h_2 are scalar lag distances and $\mathbf{l}_1$ and $\mathbf{l}_2$ are lag vectors of unit length. Note that this notation is somewhat different to that of Dimitrakopoulos et al. (2010). Given such a configuration, and making the ergodicity assumption that the distribution of $Z(\mathbf{x})$ is independent of $\mathbf{x}$, we may express the third-order cumulant for the random field at these locations as a function of lag only:

$$\begin{aligned}\kappa^3(\mathbf{h}_1,\mathbf{h}_2) &= \mathrm{E}\left[Z(\mathbf{x})Z(\mathbf{x}+\mathbf{h}_1)Z(\mathbf{x}+\mathbf{h}_1+\mathbf{h}_2)\right] \\ &\quad -\mathrm{E}\left[Z(\mathbf{x})\right]\mathrm{E}\left[Z(\mathbf{x}+\mathbf{h}_1)Z(\mathbf{x}+\mathbf{h}_1+\mathbf{h}_2)\right] \\ &\quad -\mathrm{E}\left[Z(\mathbf{x}+\mathbf{h}_1)\right]\mathrm{E}\left[Z(\mathbf{x})Z(\mathbf{x}+\mathbf{h}_1+\mathbf{h}_2)\right] \\ &\quad -\mathrm{E}\left[Z(\mathbf{x}+\mathbf{h}_1+\mathbf{h}_2)\right]\mathrm{E}\left[Z(\mathbf{x}+\mathbf{h}_1)Z(\mathbf{x})\right] \\ &\quad +2\mathrm{E}\left[Z(\mathbf{x})\right]\mathrm{E}\left[Z(\mathbf{x}+\mathbf{h}_1))\right]\mathrm{E}\left[Z(\mathbf{x}+\mathbf{h}_1+\mathbf{h}_2)\right], \end{aligned} \tag{19}$$

given Equation (16) When $\mathbf{Z}(\mathbf{x})$ is a zero mean spatial field this simplifies to

$$\kappa^3(\mathbf{h}_1,\mathbf{h}_2) = \mathrm{E}\left[Z(\mathbf{x})Z(\mathbf{x}+\mathbf{h}_1)Z(\mathbf{x}+\mathbf{h}_1+\mathbf{h}_2)\right]. \tag{20}$$

Note from the discussion above that, for a Gaussian random field, the cumulants $\kappa^r(\mathbf{h}_1,\mathbf{h}_2,\ldots,\mathbf{h}_{r-1})$ for any lags and for $r > 2$ are zero. This does not depend on assumptions of ergodicity.

As proposed by Dimitrakopoulos et al. (2010) cumulants may be estimated for specified lag combinations, such as $\mathbf{h}_1,\mathbf{h}_2$, by considering all sets of observations whose locations are translations of the basic template $[\{0,0\},\mathbf{h}_1,\mathbf{h}_1+\mathbf{h}_2]$. When observations are not regularly spaced it is necessary, as with estimation of the empirical variogram, to compute estimates for lag bins which allow for some variation or tolerance about a central lag. Under the assumption of ergodicity (at least up to the order of the cumulant of interest), the estimator for the third cumulant of a zero-mean random variable from a set of observations at locations X is therefore

$$\widehat{\kappa^3}(\mathbf{h}_1,\mathbf{h}_2) = \frac{1}{N(\mathbf{h}_1,\mathbf{h}_2)} \sum_{\{\mathbf{x},\mathbf{x}+\mathbf{h}_1,\mathbf{x}+\mathbf{h}_1+\mathbf{h}_2\}\in X} z(\mathbf{x})z(\mathbf{x}+\mathbf{h}_1)z(\mathbf{x}+\mathbf{h}_1+\mathbf{h}_2), \tag{21}$$

where there are $N(\mathbf{h}_1,\mathbf{h}_2)$ sets of observations whose locations are translations of the basic template $[\{0,0\},\mathbf{h}_1,\mathbf{h}_1+\mathbf{h}_2]$.

3. Materials and Methods

3.1. Data on the Bunter Sandstone porosity.

The data used in this study are all from the Bunter Sandstone formation. The Bunter Sandstone is a sheet-sand complex comprising mainly fine-grained but locally medium- or coarse-grained material (Cameron et al., 1992). It was deposited as fluvial channel sands in arid conditions in the lower Triassic. The Bunter Sandstone is a significant formation in the North Sea and corresponds to the Sherwood Sandstone group onshore. It is an important gas reservoir in the North Sea and is potentially important for carbon capture and storage (Holloway, 2009; Senior, 2010). For this reason the porosity of the Bunter Sandstone is of interest. The porosity of this material is affected by various factors including the structure of the original sediments, the depositional overburden, cementation of the material and subsequent diagenetic transformation (Bifani, 1986).

The data are derived from analysis of cores extracted from 32 wells across the North Sea. The cores were of variable length, and were sampled by extracting plug samples of one inch diameter, the diameter of the plug being in the vertical direction. The recorded depth of the plug was at its centre. The samples were not collected at absolutely regular intervals, the mean spacing down-core between successive samples was 0.6 m. Where coherent plugs could not be extracted a comparable volume of chipped material was removed. Each sampled specimen was washed to remove all hydrocarbons and oven-dried to a constant weight before porosity was determined by helium porosimetry. These are the best data available on the porosity of the Bunter Sandstone, but it is acknowledged that there may be some observational errors due to dissolution of halite cements during washing of the samples (Ketter, 1991). The analyses reported in this paper are limited to porosity data from plugs in water-filled sections of the cores, excluding results from gas-filled material. A total of 1282 measurements from the 32 cores were available.

3.2. Calculations.

3.2.1. Exploratory analysis and linear mixed model. The number of wells is too small to allow spatial modelling of the lateral variability of porosity in this formation. For this reason a linear mixed model of the following form was fitted for exploratory purposes

$$Z(i,x) \;=\; \mu + K_i + \eta(i,x), \tag{22}$$

where $Z(i,x)$ is a random variable: the porosity at depth x within the ith well. Note that we define locations within wells by scalar depths, effectively the data within any well are in one dimension. The mean porosity over all depths and wells is μ, K_i is a random effect drawn from a random variable with mean zero and variance σ^2_{B}; it represents the difference between the mean porosity for the ith well and the overall mean porosity. The term $\eta(i,x)$ is also a random effect of mean zero and variance, σ^2_{W}. This random effect accounts for the within-well variability. The covariance of the values of η at any two depths in the same borehole is

$$\begin{aligned}\mathrm{Cov}\left[\eta(i,x),\eta(i,x')\right] &= \sigma^2_{\mathrm{W}}, \qquad x = x' \\ &= (1-\xi)\,\sigma^2_{\mathrm{W}} R\left(|x-x'|;\boldsymbol{\psi}\right), \qquad x \neq x' \end{aligned} \tag{23}$$

where $R\left(\cdot;\boldsymbol{\psi}\right)$ is a correlation function with parameters in $\boldsymbol{\psi}$ and $\xi \in [0,1]$ is the nugget ratio, the proportion of the variance of η which is not correlated at spatial scales resolved by the sampling. This may include measurement error. Because the argument of the correlation function is the distance between two locations within a borehole rather than two absolute positions, the correlation structure is said to be second-order stationary (Journel and Huijbregts, 1977). Various correlation functions may be considered, provided that they guarantee a positive definite correlation matrix for η at any set of unique sites. One such function is the exponential:

$$R\left(|x-x'|;[r]\right) \;=\; \exp\left\{-|x-x'|/r\right\}, \tag{24}$$

with r, a distance parameter the only element in $\boldsymbol{\psi}$. An alternative is the spherical function:

$$\begin{aligned}R\left(|x-x'|;[a]\right) &= 0 \qquad a > |x-x'|, \\ &= 1 - \frac{3|x-x'|}{2a} + \frac{1}{2}\left(\frac{|x-x'|}{a}\right)^3 \qquad a \leq |x-x'|, \end{aligned} \tag{25}$$

for which a is the distance parameters, the range of the covariance function. Under these correlation models the term η at two locations in a borehole are expected to be more similar the closer they are in space.

The variance parameters of the linear mixed model in Eq. [22] — the variances σ_{B}^2 and σ_{W}^2, the nugget ratio ξ and the terms in $\boldsymbol{\psi}$ — are best estimated by residual maximum likelihood (REML) (Verbeke and Mohlenbergs, 2000). This entails the assumption that the random effects can plausibly be regarded as realizations of a normal random field. In the context of this study we examined the plausibility of this assumption (which we know cannot be strictly true because porosity is bounded in the interval [0,100]), by examining the marginal distribution of the residuals from an ordinary least squares fit of the LMM. Exploratory statistics were computed for the residuals, including the robust measure of skewness, the octile skew, proposed by Brys et al. (2003). Because porosity is a proportion, as noted above, we repeated this exploratory analysis after a logistic transformation of the porosities. Finally, the parameter of a Box-Cox transformation was estimated by maximum likelihood by means of the BOXCOX procedure in the MASS package for the R platform (Venables and Ripley, 2002) and exploratory analysis was undertaken on residuals after this transform. Results are presented below, but the following procedures may be followed on the basis either that the residuals appear to have a reasonably normal distribution or that this is plausible after an appropriate transformation.

The parameters of the linear mixed model were then estimated by REML. The lme procedure in the NLME library for R (Pinheiro et al., 2013; R Development Core Team, 2010) was used, and spherical and exponential correlation functions for $\boldsymbol{\eta}$ were considered. The variance parameters for $\boldsymbol{\eta}$ were tested by cross-validation. Each residual from the well mean was removed from the data set in turn and predicted by ordinary kriging from the remaining values in the same well. This was done using the XVOK2D algorithm in the GSLIB library (Deutsch and Journel, 1997). For each observation, $\eta(i, x)$ this provides a kriging estimate, $\tilde{\eta}(i, x)$, and the prediction error variance (kriging variance) $\sigma_{\mathrm{K}}^2(i, x)$. A useful diagnostic (Lark, 2009) is the standardized squared prediction error, with mean one and median 0.455 for normal kriging errors when the variance parameters are correct:

$$\theta(i,x) = \frac{\{\eta(i,x) - \tilde{\eta}(i,x)\}^2}{\sigma^2_{\mathrm{K}}(i,x)}. \tag{26}$$

The linear mixed modelling framework was used to test the hypothesis that porosity depends on depth down the well. Neither exploratory plots of the data nor these models provided any evidence for a trend in porosity with depth, and so I proceeded with the model in Equation [22] where the mean porosity is constant within any well.

3.2.2 Estimating κ^3 for particular templates. For a zero-mean ergodic random variable $\eta(i,x)$ on a set of one-dimensional wells, K, the third-order cumulant, defined for a random field in Eq. [20], is defined for scalar lag distances h_1 and h_2 by

$$\begin{aligned} \kappa^3_\eta(h_1,h_2) &= \\ &\mathrm{E}\left[\eta(i,x_1)\eta(i,x_2)\eta(i,x_3); |x_2-x_1| = h_1, |x_3-x_2| = h_2, (x_2-x_1)(x_3-x_2) > 0\right]_{i\in K}. \end{aligned} \tag{27}$$

Note that under this definition the locations are in order x_1, x_2, x_3 up or down the well, and the cumulant is symmetric in the sense that $\kappa^3_\eta(h_1,h_2) = \kappa^3_\eta(h_2,h_1)$.

In practice, when sampling is not on a regular array, it is necessary to allow some tolerance in the definition of the lag distances (Dimitrakopoulos et al., 2010). In this study we define a scalar-lag class $\tilde{h}$ as the interval $[h-\tau, h+\tau]$ where τ is the tolerance. We define the indicator variable

$$\begin{aligned} I(i,x_1,x_2,x_3;\tilde{h}_1,\tilde{h}_2) &= 1 \;\; i \in K, |x_2-x_1| \in \tilde{h}_1, |x_3-x_2| \in \tilde{h}_2, (x_2-x_1)(x_3-x_2) > 0 \\ &= 0 \;\; \text{otherwise}. \end{aligned} \tag{28}$$

We then define the estimate $\widehat{\kappa^3_\eta}(h_1,h_2)$ by

$$\begin{aligned} \widehat{\kappa^3_\eta}(h_1,h_2) &= \\ &\frac{1}{N_{h_1,h_2}} \sum_{i\in K} I(i,x_1,x_2,x_3;\tilde{h}_1,\tilde{h}_2)\{z(i,x_1)-\overline{z}_i\}\{z(i,x_2)-\overline{z}_i\}\{z(i,x_3)-\overline{z}_i\}, \end{aligned} \tag{29}$$

where $z(i,x_1)$ is the observed value of the variable at depth x_1 in the ith well, and $\overline{z}_i$ is the average value of the variable over all observations in the ith well. The summation is over all sets of three observations within all wells in the set K and N_{h_1,h_2} is the sum of the indicator over all these observations.

In this study the estimate $\widehat{\kappa^3_\eta}(h_1,h_2)$ was computed for lag distances 25 cm, 50 cm, ..., 500 cm with lag tolerance $\tau = 12.5$ cm.

3.2.3. Testing $\widehat{\kappa^3}$ against a null hypothesis of normality. As noted above the expected value of the third cumulant for a multivariate normal random variable is zero. Values of $\widehat{\kappa^3}$ for some h_1, h_2 provide evidence against this null hypothesis, but this evidence must be assessed accounting for the sample variance of the estimates. This is complicated by the lack of independence of the observations from which the estimate is obtained, so a Monte Carlo simulation procedure was developed.

Under the null hypothesis of multivariate normality the variability of the data is entirely accounted for by the variances and associated parameters of the random effects in the linear mixed model, Eq [22]. The Monte Carlo procedure requires that we can generate realizations of the random term η from the linear mixed model. We denote the set of values of this random variable by the $N \times 1$ vector $\boldsymbol{\eta}$ which corresponds to the full set of N observations. The covariance matrix of the random variate $\boldsymbol{\eta}$ is denoted by $\mathbf{V}$ where

$$\mathbf{V} = \xi\sigma^2_{\mathrm{W}}\mathbf{I} + (1-\xi)\sigma^2_{\mathrm{W}}\mathbf{R}, \tag{30}$$

where $\mathbf{I}$ is a $N \times N$ identity matrix and $\mathbf{R}$ is an $N \times N$ correlation matrix such that the entry $\mathbf{R}\{k,l\}$ for the lth observation $\eta(i,d)$ and the kth $\eta(j,d')$ is:

$$\begin{aligned}\mathbf{R}\{k,l\} &= 0, \quad \forall i \neq j \\ &= R\left(|d-d'|;\boldsymbol{\psi}\right), \quad \forall i = j,\end{aligned} \tag{31}$$

where R is a correlation function with parameters in $\boldsymbol{\psi}$. In this study the correlation function fitted by REML, and the estimated parameters were used. Once $\mathbf{V}$ has been computed it is possible to find its Cholesky factorization:

$$\mathbf{V} = \mathbf{L}\mathbf{L}^*, \tag{32}$$

where $\mathbf{L}$ is a lower-triangular matrix with real and positive diagonal elements and $\mathbf{L}^*$ is its conjugate transpose. This factorization is guaranteed to exist because the matrix $\mathbf{R}$, as a covariance matrix computed from an authorized correlation function, is positive-definite and symmetric with real values. It is then possible to generate a realization of $\boldsymbol{\eta}$ by computing

$$\boldsymbol{\eta} \;=\; \mathbf{Lg}, \tag{33}$$

where the elements of $\mathbf{g}$ are independent values with a standard normal distribution.

In this study the IMSL subroutine CHFAC was used to compute the Cholesky factorization. One may then substitute the elements of $\boldsymbol{\eta}$ for the values of z in Eq. [29] to compute $\widehat{\kappa_\eta^3}(h_1, h_2)$ for the same lag distances for which this was computed for the original data. It is immaterial that the between-well random effect is not simulated here since the mean value for each well is subtracted from each observation in Eq. [29]. Since $\boldsymbol{\eta}$ is simulated for the same locations as the data, the value of $\widehat{\kappa_\eta^3}(h_1, h_2)$ for some lag distances computed from the simulated data can be regarded as a realization of the sampling distribution of our observed statistic under the null hypothesis of a multivariate normal distribution. Note also that the sample error of each well mean, which contributes to the error of the estimation of $\widehat{\kappa_\eta^3}$ which is estimated on the assumption of zero mean, also appears in the simulation procedure and so is included in the Monte Carlo approximation to the sampling distribution of $\widehat{\kappa_\eta^3}$. In this study 100 000 realizations of $\boldsymbol{\eta}$ were generated and used to compute the sampling distribution of $\widehat{\kappa_\eta^3}(h_1, h_2)$ for the specified lags under the null hypothesis.

Two approaches were used to examine the extent to which the empirical cumulants of the data are consistent or otherwise with a null hypothesis of normality. The first was to find the maximum absolute value of the estimated cumulants over all lag distances,

$$\widehat{\kappa^3}_{\eta,\max} \;=\; \max\left\{\left|\widehat{\kappa_\eta^3}(h_1,h_2)\right|; h_1 = 25, 50, \ldots, 500\text{cm}; h_2 = 25, 50, \ldots, 500\text{cm}\right\}. \tag{34}$$

This statistic was evaluated for the empirical residuals from the well means, and then for each of a set of 100 000 realizations of $\boldsymbol{\eta}$, generated as described above. Since the expected value of the cumulant under the null hypothesis of a multivariate Gaussian random variable is zero a large value of $\widehat{\kappa^3}_{\eta,\max}$ provides evidence against this null hypothesis. The strength of evidence is measured by a p-value which can be approximated by ordering the values of $\widehat{\kappa^3}_{\eta,\max}$ from the simulations and computing the proportion of these which exceed the observed value.

The second approach was to test the separate cumulants for each lag pair h_1, h_2. For some observed lag pair at which the observed cumulant is $\widehat{\kappa_\eta^3}(h_1, h_2)$ the p-value for the null hypothesis of a zero cumulant is computed by finding the proportion of the 100 000 realizations of $\boldsymbol{\eta}$ for which the cumulant fall outwith the interval $\left[-\left|\widehat{\kappa_\eta^3}(h_1, h_2)\right|, +\left|\widehat{\kappa_\eta^3}(h_1, h_2)\right|\right]$. These p-values were inspected for a set of lag combinations, excluding those with fewer than 600 supporting triplets of observations. This is a multiple hypothesis test, in which we examine a family of null hypotheses which are not mutually independent. For that reason it is necessary to control the family-wise error rate (FWER), α_{rmFW}, which is the probability of one or more of the family of null hypotheses' being rejected although all of them are true. The simplest way to control the family-wise error rate for a set of m hypotheses is to reject only those for which $p < \alpha_{\mathrm{FW}}/m$. This is the Bonferroni control of FWER, and is valid for non-independent hypotheses (Snedecor and Cochran, 1980). However, it is relatively lacking in power. An alternative, also valid for non-independent hypotheses, is the procedure due to Holm (1979). In Holm's procedure one orders the null hypotheses $H_1, H_2, \ldots, H_m$ in order of ascending p-value, $p_1, p_2, \ldots, p_m$. One then evaluates for successive $k = 1, 2, \ldots, m$ whether

$$p_k > \frac{\alpha_{\mathrm{FW}}}{m + 1 - k}.$$

Let k_r be the smallest value of k for which this expression is true. One may then reject, with FWER α_{FW}, the null hypotheses $H_1, H_2, \ldots, H_{k_r-1}$. This procedure was followed to find the subset of lag pairs for which the null hypothesis that the cumulant is zero could be rejected.

3.2.4 Exploring the implications of a non-zero cumulant. In order to gain insight into the nature of the variability of a variable with non-zero third order cumulants for lag-pairs h_1, h_2 I examined 3-D plots of the triplets of observations $\{z(d), z(d+h_1), z(d+h_2)\}$ using the SCATTERPLOT3D package in R. This is comparable to the examination of two-dimensional scatterplots of $\{z(\mathbf{x}), z(\mathbf{x}+\mathbf{h})\}$ which is sometimes advocated as an exploratory technique in geostatistics (Goovaerts, 1997).

4. Results

Table 1 presents summary statistics for residuals for porosity from the well mean, and the same residuals for data after logistic or Box-Cox transformation. Note that there is little appreciable effect of the Box-Cox transformation, and the 95% confidence interval of the Box-Cox parameter included the value 1, under which the transform is equivalent to adding a constant to the variable and has no effect on the shape of the distribution. The residuals after a logistic transform are more skewed than in the other two cases. All of these exploratory statistics suggest that an assumption of normality of the residuals with no transformation seemed plausible. Figure 1 shows the histogram of these residuals and their empirical Quantile-Quantile plot which should lie on the bisector.

Table 2 shows the results of the REML estimation of the variance parameters for the linear mixed model for porosity set out in Eq[22]. Figure 2 shows the histogram of cross-validation errors for the selected model (exponential) and the Q-Q plot. These show that the errors are close to normal in their distribution. The mean and median standard square cross validation errors are in Table 2. Note that the mean is close to 1.0, but the median is rather smaller than is expected.

It was found that the numbers of triplets of observations from which to estimate the cumulant for particular lag pairs N_{h_1,h_2} varied. For most pairs of lags there were between 600 and 1600 triplets, so those lags supported by fewer observations were discarded. Figure 3 shows the estimated values $\widehat{\kappa^3_\eta}(h_1, h_2)$ which are plotted only in the lower half of the plot (where $h_1 > h_2$). The dots in the upper half of the plot indicate the lags at which the number of supporting triplets of observations was fewer than 600.

The largest absolute value of the third cumulant over the lags considered was 57.8 for lag-pair {50 cm, 250 cm}. Table 3 shows the percentiles of the maximum absolute value of the third cumulant over 100 000 realizations of the Gaussian model, and also percentiles of the third cumulant for lags {50 cm, 250 cm}. Figure 4 shows the approximate density functions for (a) the maximum absolute value of the third cumulant over all lags and (b) the third cumulant for lags {50 cm, 250 cm} from the 100 000 realizations. The density was obtained by the KERNELDENSITY procedure in GenStat (Goedhart, 2009). This, and the percentiles in Table 3, indicate that the cumulant is distributed more or less symmetrically about zero under the null hypothesis of a multivariate Gaussian distribution. The percentiles of the maximum absolute value of the third cumulant over all lags in Table 3 shows that the approximate p-value for the evidence provided by the absolute maximum third cumulant for these data against a null hypothesis of normality is less than 0.01, but larger than 0.001.

In the upper half of Figure 3 are plotted those cumulants which were significantly different from zero as judged by the p-values computed for each lag pair from the 100 000 realizations, with FWER controlled at 0.05. There are six lag pairs at which the cumulants are significantly non-zero. Note that the significant cumulants are negative for smaller lags — {50 cm, 250 cm},{50 cm, 225 cm} and {100 cm, 225 cm} — and positive for the longer lags, {50 cm, 425 cm}, {150 cm, 275 cm} and {275 cm, 500 cm}.

Three-dimensional scatter-plots were examined for data triplets (residuals from the well mean) with the smallest (most negative) and largest (most positive) cumulant, corresponding to lags {50 cm, 250 cm} and {150 cm, 275 cm} respectively. I do not attempt to reproduce them here but the effects that they show can be illustrated by two two-dimensional plots of residuals for two locations, x_1 and $x_2 = x_1 + h_1$ with, respectively $\eta(x_3) > 0$ and $\eta(x_3) \leq 0$ where $x_3 = x_2 + h_2$. These plots are shown in Figure 5, along with the correlations between the variables on the plots. Note that the 'positive quadrants' of the plot, where $\eta(x_1)\eta(x_2) > 0$, have been given a grey background. It is apparent that the correlation between $\eta(x_1)$ and $\eta(x_2)$ differ between the cases where $\eta(x_3) > 0$ and $\eta(x_3) \leq 0$, and these differences are significant in each case with $p < 0.001$. This difference in correlation is an expression of the non-zero third cumulant of $\boldsymbol{\eta}$ which has been found for these lag pairs, since it means that distribution of observations between the positive and negative quadrants of these plots is different for the case where $\eta(x_3) > 0$ and where $\eta(x_3) \leq 0$. Furthermore, this difference in correlation is inconsistent with the assumption of second-order stationarity under which the correlation between $\eta(x_1)$ and $\eta(x_2)$ should depend only on h_1.

5. Discussion

In the work above five general results were obtained from the exploratory analysis of the porosity data.

1. Summary statistics and histograms on the marginal distribution of the data, including after transformation. (Figure 1, Table 1).
2. A plot of the third cumulant of the centred data for a range of lags (Figure 3).
3. P-values for tests of the null hypothesis of an underlying multivariate Gaussian process based on the third cumulants.
4. Scatter plots of data triplets and associated correlations (Figure 4).
5. Results from the cross validation of the fitted linear mixed model (Figure 2, Table 2).

The significance tests on the cumulants — item (3) in the list above — allow us to reject the null hypothesis of an underlying multivariate Gaussian random variable. Whether this is, of itself, of direct practical relevance is open to debate. Webster and Oliver (2007) suggest that significance tests for conformity to distributions are not particularly valuable for the purpose of assessing the plausibility of distributional assumptions. We know in most cases that a variable is not strictly normally distributed, and, particularly with large data sets, we do not expect the null hypothesis of normality to be accepted. For example, with the data in this paper, we know that they cannot have a Gaussian distribution at the limit since porosity is bounded on the interval $[0, 1]$. However, the exploratory statistics of these data indicated that they are close to symmetrically distributed with a bell-shaped histogram, and that neither the logistic nor the Box-Cox transformation improved this. Following the guidelines of Webster and Oliver (2007) one would normally proceed on the basis that a normality assumption is plausible.

How is this approach extended to the consideration of multivariate normality? The plot of the cumulants (Figure 3) may indicate possible systematic deviations from the expected value (zero), e.g. clustering of small (large negative) or large positive values at particular lags, and the significance test indicates whether or not the general pattern is compatible with sampling error from an underlying Gaussian process. The cumulant plot also leads us to the particular data triplet plots which merit further investigation. These triplet plots are visualizable projections of the data which allow us to see the particular deviation from normality which the corresponding cumulant represents. In this case we can identify notable differences between the correlation of $\eta(x_1)$ with $\eta(x_2)$ conditional on the value of $\eta(x_3)$. This is not consistent with an assumption of stationarity in the covariance. This is consistent with the cross-validation results, presented in Table 2. Note that the median squared standard prediction error is rather less than the expected value of 0.455. This may be due to the non-stationarity of the underlying variable, as found by Lark (2009) in the comparison of kriging results from stationary and non-stationary variance models. It is also possible that outlying data values could influence both the squared standard prediction errors and estimates of the cumulants. The exploratory data analysis did not indicate any marginal outliers in the data, but spatial outliers, values unusual in their local context, may be present. One possible area for future work is to develop robust estimators of the cumulants, but it would be necessary to find estimators that do not import distributional assumptions through the use of particular consistency corrections (Lark, 2000) while remaining reasonably efficient.

In short, the analysis of the third cumulants of the variable provides us with a basis for identifying particular plots of the data which allow us to examine its deviation from a stationary normal process directly, and to interpret other results such as those from the cross-validation. I would agree with Webster and Oliver (2007) that, in general, we should not base decisions about the validity of distributional assumptions on tests of conformity to the particular distribution. Further work is required to develop exploratory statistics based on the cumulants, which allow us to make an informed pragmatic judgement about the plausibility of the distributional assumption. We require, for example, rules of thumb such as that enunciated by Webster and Oliver (2007) that some transformation of data is required if the coefficient of skewness exceeds 0.5. Such rules of thumb might be based on plots of the cumulant such as Figure 3, and must be based on experience of a range of data sets and the robustness of the Gaussian assumption when predicting or simulating the measured variable.

Note that in the case study there was no evidence for any trend in porosity with depth, and so it was assumed that the mean porosity in any well was constant. If a trend was found then this would be subtracted from the observations before computation of the cumulants, and the Monte Carlo procedure to approximate the sample distribution of the cumulant under the null hypothesis would have to be extended to include the contribution of the uncertainty in the estimation of the trend just as the reported procedure accounted for the uncertainty in the estimation of the well means.

Given the sparsity of wells, and the distances between them, the current study was limited to cumulants in one dimension, attention was also focussed on the third cumulants. Any third cumulant in one dimension is defined for a lag pair, and so can easily be displayed in 2-D plots. The extension of this method to higher-order cumulants, to two or more dimensions, or both would make it harder to use visualization in the analysis of data. However, the general principles used in this paper, for the estimation of empirical cumulants and the use of multiple hypothesis testing methods to find lag-combinations at which the data provide evidence against a multivariate Gaussian model, could be extended to sets of more than two lag combinations in a straightforward way, and so to higher-order cumulants and more than one dimension. Plots for visual interpretation could then be generated as appropriate projections, in the same spirit of the triplet plots used in this paper. Those considerations aside, the one-dimensional case illustrated here remains of considerable relevance since many porosity or conductivity fields in geology can only be examined intensively down-core. This is because of the relative sparsity of cores, particularly offshore, and the fact that they are often widely spaced which limits the scope to examine lateral variability.

These results give reason for concern about the suitability of prediction error variances and other measures of uncertainty based on the multivariate Gaussian model of porosity in the Bunter Sandstone. It should also be recalled that regionalized variables with non-Gaussian distributions may have more complex geometrical structure than Gaussian variables, particularly with respect to the connectivity of extreme values (e.g. Guardiano and Srivastava, 1993). This means that simulations of porosity fields from multivariate Gaussian random variables, even if these well-reproduce the marginal statistics of porosity, may fail to represent all aspects of the spatial structure of the variable (such as the volumes of regions of continuous large or small porosity) which may be relevant to questions of fluid flow or potential gas storage in the field.

One way to deal with this may be by copula methods (e.g. Haslauer et al, 2012), although the development of appropriate spatial copula models other than the Gaussian which can be fitted to sizeable data sets is at an early stage. An alternative is to use the methods of multiple point geostatistical modelling,(e.g. Strebelle, 2001), but these require large data sets for training. One solution would be to find a non-Gaussian stochastic model which reproduces the cumulants of interest. A possible general form of the model would be one in which a well is divided into intervals by randomly located boundaries (occurring as a Poisson process, so that the boundaries have an exponential distribution). The resulting segments of the well could be regarded as distinct geological facies. In the simplest such model all observations within any one of the segments thus-formed take a value drawn from a centred Gaussian random variable, Y. It is known (Lark, 2010) that this random field is not multivariate Gaussian (although its marginal distribution is). However, one can see that its third cumulant is zero since:

$$\kappa^3(\mathbf{h}_1, \mathbf{h}_2) \;=\; p_1(\mathbf{h}_1, \mathbf{h}_2)\mathrm{E}[Y^3] + p_2(\mathbf{h}_1, \mathbf{h}_2)\mathrm{E}[Y]\mathrm{E}[Y^2] + p_3(\mathbf{h}_1, \mathbf{h}_2)\mathrm{E}[Y]^3, \tag{35}$$

where $p_1(\mathbf{h}_1, \mathbf{h}_2)$ is the probability that all three locations in the template fall in different segments, $p_2(\mathbf{h}_1, \mathbf{h}_2)$ is the probability that two sites fall in one segment and one in another and $p_3(\mathbf{h}_1, \mathbf{h}_2)$ is the probability that all three locations fall into the same segment. These probabilities need not be evaluated since it is clear, from the fact that the variable is centred and Gaussian, so $\mathrm{E}[Y] = \mathrm{E}[Y^3] = 0$, that all three terms are zero. In a more complex version of this model one might postulate, for example, a correlation between the thickness of the segment and its expected porosity. In some preliminary simulations it was found that the resulting random variable may have a marginal distribution which appears Gaussian when the correlation between segment thickness and mean porosity is not too strong, but that the third cumulants were systematically smaller than zero for pairs of short lags (Figure 6). This is not offered as an alternative model for the Bunter Sandstone porosity, but simply as an indicator that the kind of spatial variation that has been found in reality might be reproduced by an appropriate stochastic model. This is a topic for further work, and should account for known general properties of the geological units. For example, while one might postulate relationships between grain size and facies thickness in depositional environments, porosity is also affected by overburden, diagenetic transformations of the sandstone and other processes which may be spatially dependent but are not obviously reproducible by a stochastic geometry.

6. Conclusions

It has been shown how the third cumulant of a spatial variable observed in linear data sets (wells) can be used in an inferential context to test the null hypothesis that the underlying distribution of the variable is multivariate Gaussian, and to guide exploratory analysis to test the plausibility of this distributional assumption. This approach was applied to data on porosity of the Bunter Sandstone and showed that there were features of its distribution which appear incompatible with the assumption of stationarity and multivariate Gaussian variation. This has potential implications for the use of standard geostatistical methods to characterize the uncertainty that attends inferences about this variable. This might require that multiple point geostatistics are used for this variable. Alternatively some non-Gaussian random variable might be postulated as a model, and an example of one which has some common features with the data is discussed. In practice it might be possible to develop such a model for porosity of the Bunter Sandstone; such a model should take account of our understanding of the depositional and diagenetic processes that control this variable.

Acknowledgements

I am grateful to colleagues at BGS (Sam Holloway, Sarah Hannis) for guidance about the Bunter Sandstone and its properties, and to IHS who are owners of a subset of the porosity data for permission to use them. This paper is published with the permission of the Executive Director of the British Geological Survey (NERC).

References

Bifani, R. 1986. Esmond Gas Complex. In: Brooks, J., Goff, J.C. & Van Hoorn, B. (eds) Habitat of Paleozoic Gas in NW Europe. Geological Society, London, Special Publications, 23, 209–221.

Bilodeau, M., Brenner, D. 1999. Theory of Multivariate Statistics. Springer, New York.

Brys, G., Hubert, M., Struyf, A. 2003. A comparison of some new measures of skewness. In: Developments in Robust Statistics (eds R. Dutter, P. Filzmoser, U. Gather and P.J. Rousseeuw), pp. 98–113. Physica-Verlag, Heidelberg.

Cameron, T.D.J., Crosby, A., Balson, P.S., Jeffrey, D.H., Lott, G.K., Bulat, J., Harrison, D.J. 1992. United Kingdom offshore report: the geology of the southern North Sea. London, HMSO for the British Geological Survey.

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

Dimitrakopoulos, R., Mustapha, H., Gloaguen, E. 2010. High-order statistics of spatial random fields: exploring spatial cumulants for modeling complex non-Gaussian and non-linear phenomena. Mathematical Geoscience, 42, 65–99.

Goedhart, P.W. 2009. KERNELDENSITY procedure in (ed.) R.W. Payne, GenStat Release 12 Reference Manual, Part 3 Procedure Library PL20. VSN International, Hemel Hempstead.

Goovaerts, P. 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York.

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

Haslauer, C.P., Guthke, P., Bárdossy, A., Sudicky, E.A. 2012. Effects of non-Gaussian copula-based hydraulic conductivity fields on macrodispersion. Water Resources Research, 48, W07507.

Holloway, S. 2009 Storage capacity and containment issues for carbon dioxide capture and geological storage on the UK continental shelf. Proceedings of the Institution of Mechanical Engineers. Part A, Journal of Power and Energy, 223 (A3), 239–248.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65–70.

Jansen, P.H.M., Stoica, P. 1988. On the expectation of the product of four matrix-values Gaussian random variables. IEEE Transactions on Automation and Control, 33, 867–870.

Journel, A.G., Huijbregts, C.J. 1978. Mining Geostatistics. Academic Press, London.

Ketter, F.J. 1991. The Esmond, Forbes and Gordon Fields, Blocks 43/8a, 43/13a, 43/15a, 43/20a, UK North Sea. In: Abbotts, I.L. (ed.) United Kingdom Oil and Gas Fields, 25 Years Commemorative Volume. Geological Society, London, Memoirs, 14, 425–432.

Kuptsov, L.P. 2001. Einstein's Rule. In: Hazewinkel, M. (Ed.) Encyclopaedia of Mathematics, Springer, Berlin.

Lark, R.M. 2000. 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. 2009. Kriging a soil variable with a simple non-stationary variance model. Journal of Agricultural, Biological and Environmental Statistics, 14, 301–321.

Lark, R.M. 2010. Two contrasting spatial processes with a common variogram: inference about spatial models from higher-order statistics. *European Journal of Soil Science*, 61, 479–492.

Lark, R.M, Bellamy, P.H. Rawlins, B.G. 2006. Spatio-temporal variability of some metal concentrations in the soil of eastern England, and implications for soil monitoring. Geoderma 133, 363–379.

McCullagh, P., Kolassa, J. 2009. Cumulants. Scholarpedia 4, 4699.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., The R Development Core Team, 2013. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-110.

R Development Core Team 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Senior, B. 2010. CO_2 Storage in the UK — Industry Potential. Department Department for Energy and Climate Change (DECC) Report URN 10D/512, DECC, London.

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.

Verbeke, G., Molenberghs, G. 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.

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

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

Appendix. The fourth cumulant of a multivariate-Gaussian random variable is zero.

Using the notation from section 2, and considering the zero-mean case for brevity of notation, the fourth multivariate cumulant can be written as

$$\kappa^{r,s,t,u} = \mathrm{E}_{rstu} - \{\mathrm{E}_{rs}\mathrm{E}_{tu} + \mathrm{E}_{rt}\mathrm{E}_{su} + \mathrm{E}_{ru}\mathrm{E}_{st}\}\,, \tag{36}$$

see McCullagh and Kolassa, 2009. Now, for multivariate-Gaussian $\mathbf{Z} \equiv [Z_1, Z_2, Z_3, Z_4]$ the expected product $Z_1Z_2Z_3Z_4$ is

$$\begin{aligned} \mathrm{E}\left[Z_1Z_2Z_3Z_4\right] = & \quad \mathrm{Cov}\left[Z_1, Z_2\right]\mathrm{Cov}\left[Z_3, Z_4\right] \\ & + \mathrm{Cov}\left[Z_1, Z_3\right]\mathrm{Cov}\left[Z_2, Z_4\right] \\ & + \mathrm{Cov}\left[Z_1, Z_4\right]\mathrm{Cov}\left[Z_2, Z_3\right], \end{aligned} \tag{37}$$

because of the disappearance of odd-order moments, see, for example, Jansen and Stoica (1988). Note that the term in braces on the RHS of Eq. [36] is equivalent to the RHS of Eq. [37], from which it follows immediately that $\kappa^{r,s,t,u} = 0$.

Table 1. Summary statistics of residuals from mean well porosity using the original data, data after a logistic transformation and data after a Box-Cox transformation.

	Original data	After Box-Cox* transformation	After logistic transformation
Mean	0.00	0.00	0.00
Median	0.28	0.26	0.09
Skewness	−0.04	−0.09	−1.36
Standard deviation	6.61	5.25	0.64
Quartile 1	−4.44	−3.51	−0.69
Quartile 3	4.27	3.32	0.59
Octile skewness	−0.07	−0.08	−0.20

*The maximum-likelihood estimate of the Box-Cox transformation parameter was 0.92 with 95% confidence interval [0.83,1.01].

Table 2. Results from REML estimation of random effects parameters, and cross-validation.

Model	log-Likelihood	AIC	
Exponential	−4166.1	8342.3	
Spherical	−4176.8	8363.7	
Selected model			
Model	Random effects parameters		
	σ^2_{B}	σ^2_{W}	ξ
Exponential	23.59	45.23	18×10^{-9}
Cross-validation results			
Mean error	0.004		
Mean standardized squared error	1.06		
Median standardized squared error	0.32		

Table 3. Quantiles of (a) Maximum value of the third cumulant over all lags, $\widehat{\kappa^3}_{\eta,\max}$; and (b) Value of the third cumulant for lag pair {50 cm, 250 cm}, $\widehat{\kappa^3}_{\eta}(50 \text{ cm}, 250 \text{ cm})$; computed from 100 000 realizations of the random model for $\boldsymbol{\eta}$.

$\widehat{\kappa^3}_{\eta,\max}$	
Quantile	Value
0.5	29.5
0.9	40
0.95	43.9
0.99	53.3
0.999	68.7

$\widehat{\kappa^3}_{\eta}(50 \text{ cm}, 250 \text{ cm})$	
Quantile	Value
0.001	-26.3
0.01	-19.4
0.05	-13.6
0.1	-10.5
0.5	0.0
0.9	10.5
0.95	13.7
0.99	19.8
0.999	27.1

Figure Captions

1. (a) Histogram of residuals from well mean porosity and (b) Gaussian Q-Q plot with bisector.

2. (a) Histogram of residuals cross-validation kriging errors and (b) Gaussian Q-Q plot with bisector.

3. Map of estimates, $\widehat{\kappa_\eta^3}(h_1, h_2)$ (below the diagonal). Symbols appear above the diagonal where the estimate was judged significantly different from zero. Small grey circles indicate where the estimate is supported by fewer than 600 triplets.

4. Estimated density functions for (top) the maximum absolute value of the third cumulant over all lag pairs under a null hypothesis of a multivariate Gaussian distribution and (bottom) the third cumulant for lag pair {50 cm, 250 cm}.

5. Scatter plots of data triplets for observations at x_1 and $x_2 = x_1 + h_1$ for (left) $\eta(x_3) < 0$ and (right) $\eta(x_3) > 0$, $\quad x_3 = x_2 + h_2$. Top row, $h_1 = 50\,\mathrm{cm}$, $h_2 = 250\,\mathrm{cm}$; bottom row, $h_1 = 150\,\mathrm{cm}$, $h_2 = 275\,\mathrm{cm}$.

6. Map of estimates $\widehat{\kappa_\eta^3}(h_1, h_2)$ for a simulated random variable in which wells are divided randomly into segments and segment porosity is weakly correlated with segment thickness.

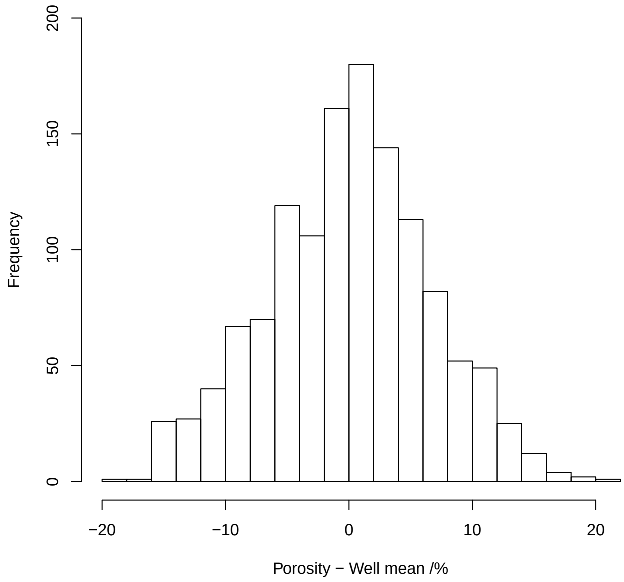


Figure 1(a)

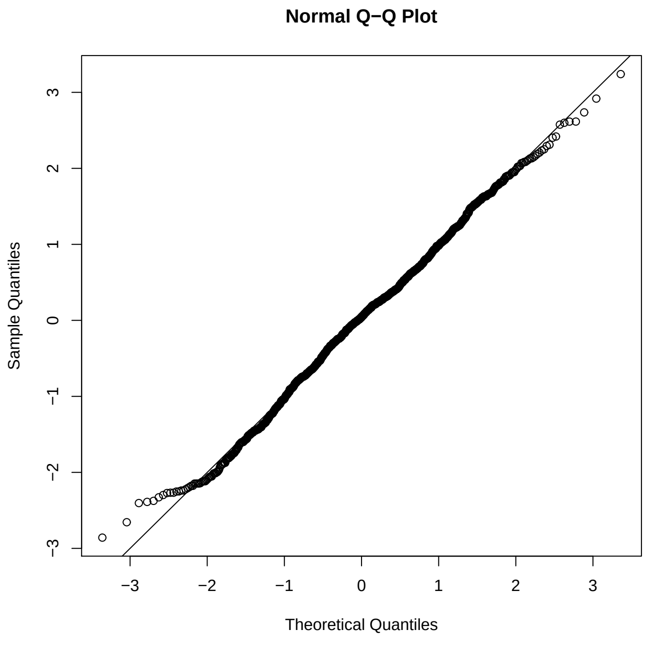


Figure 1(b)

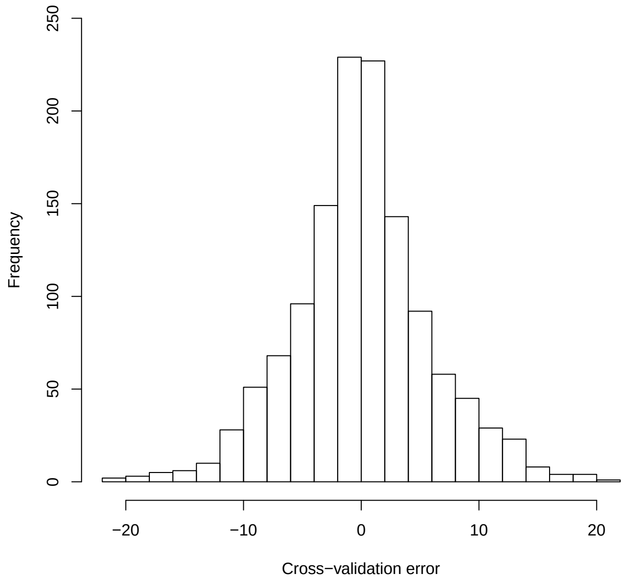


Figure 2(a)

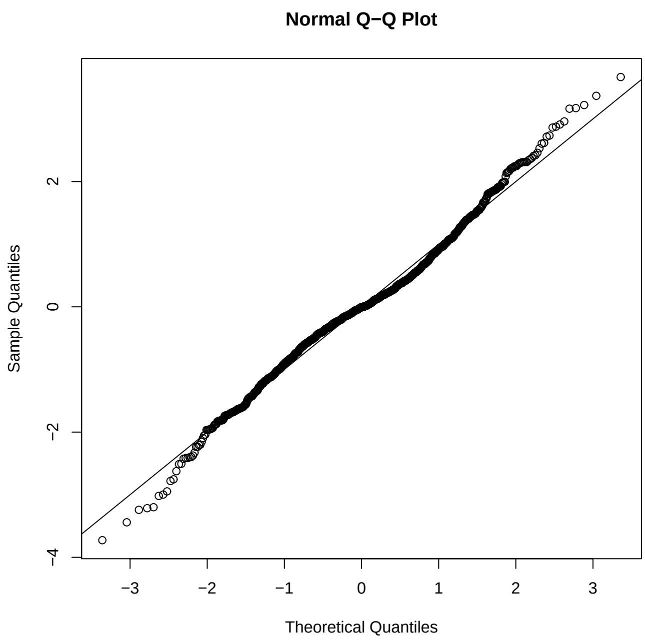
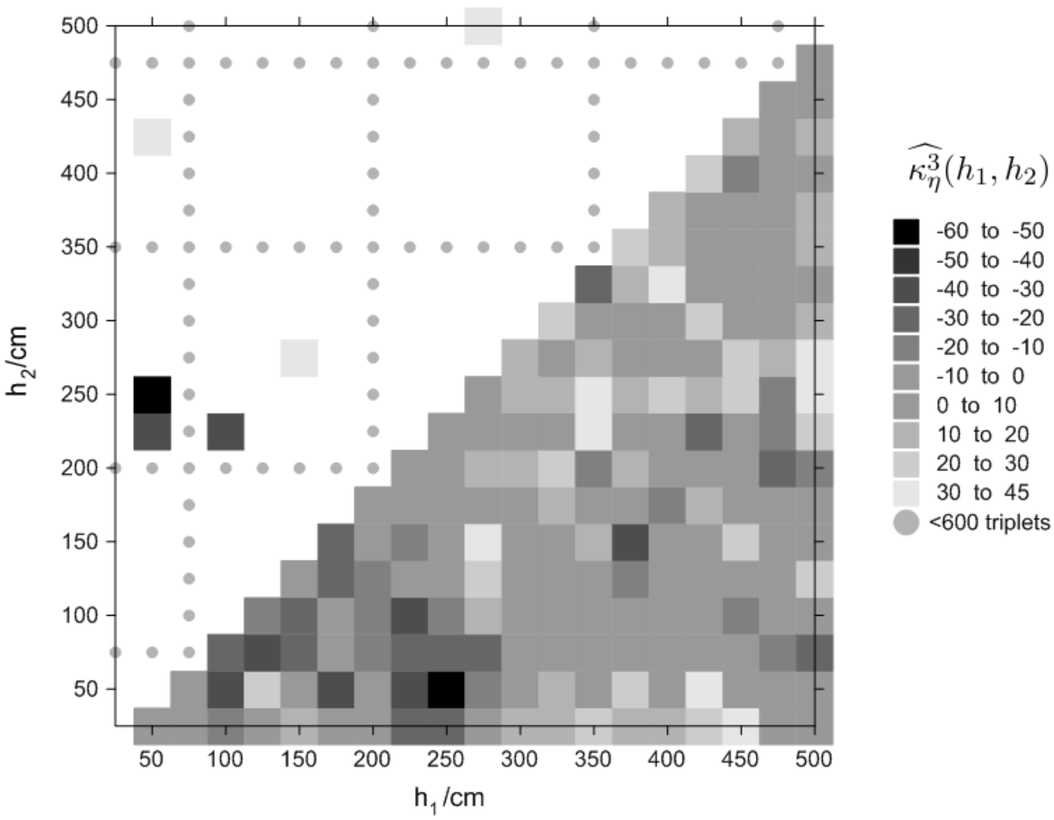
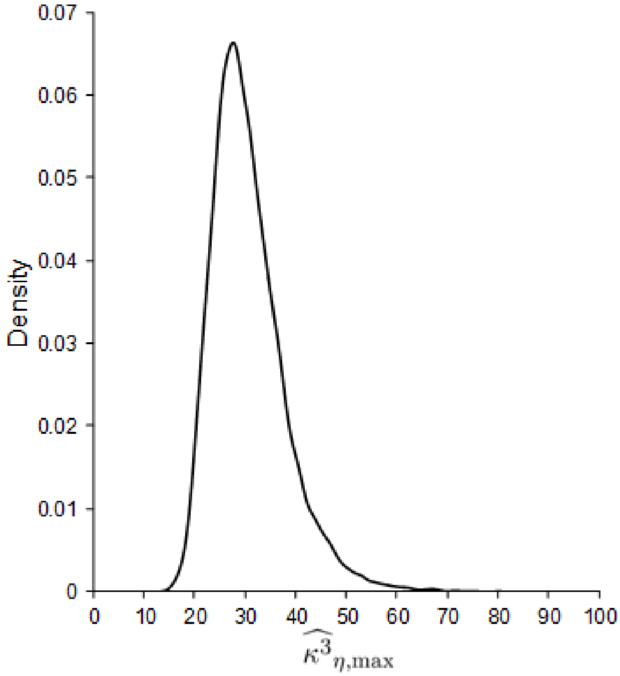


Figure 2(b)

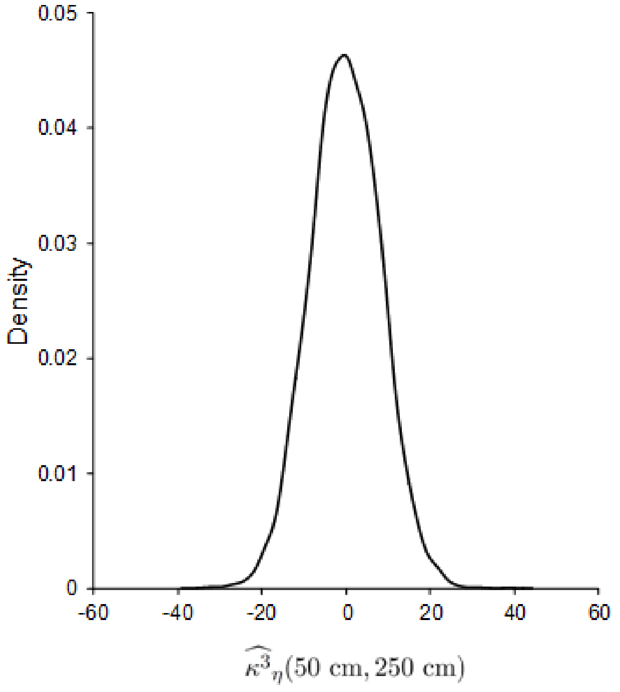


$\widehat{\kappa^3_\eta}(h_1, h_2)$
-60 to -50
-50 to -40
-40 to -30
-30 to -20
-20 to -10
-10 to 0
0 to 10
10 to 20
20 to 30
30 to 45
<600 triplets
h_2/cm
h_1/cm
50
100
150
200
250
300
350
400
450
500

Figure 3



0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Density
0
10
20
30
40
50
60
70
80
90
100
$\widehat{\kappa^3}_{\eta,\max}$



0.05
0.04
0.03
0.02
0.01
0
Density
-60
-40
-20
0
20
40
60
$\widehat{\kappa^3}_{\eta}(50\ \mathrm{cm}, 250\ \mathrm{cm})$

Figure 4

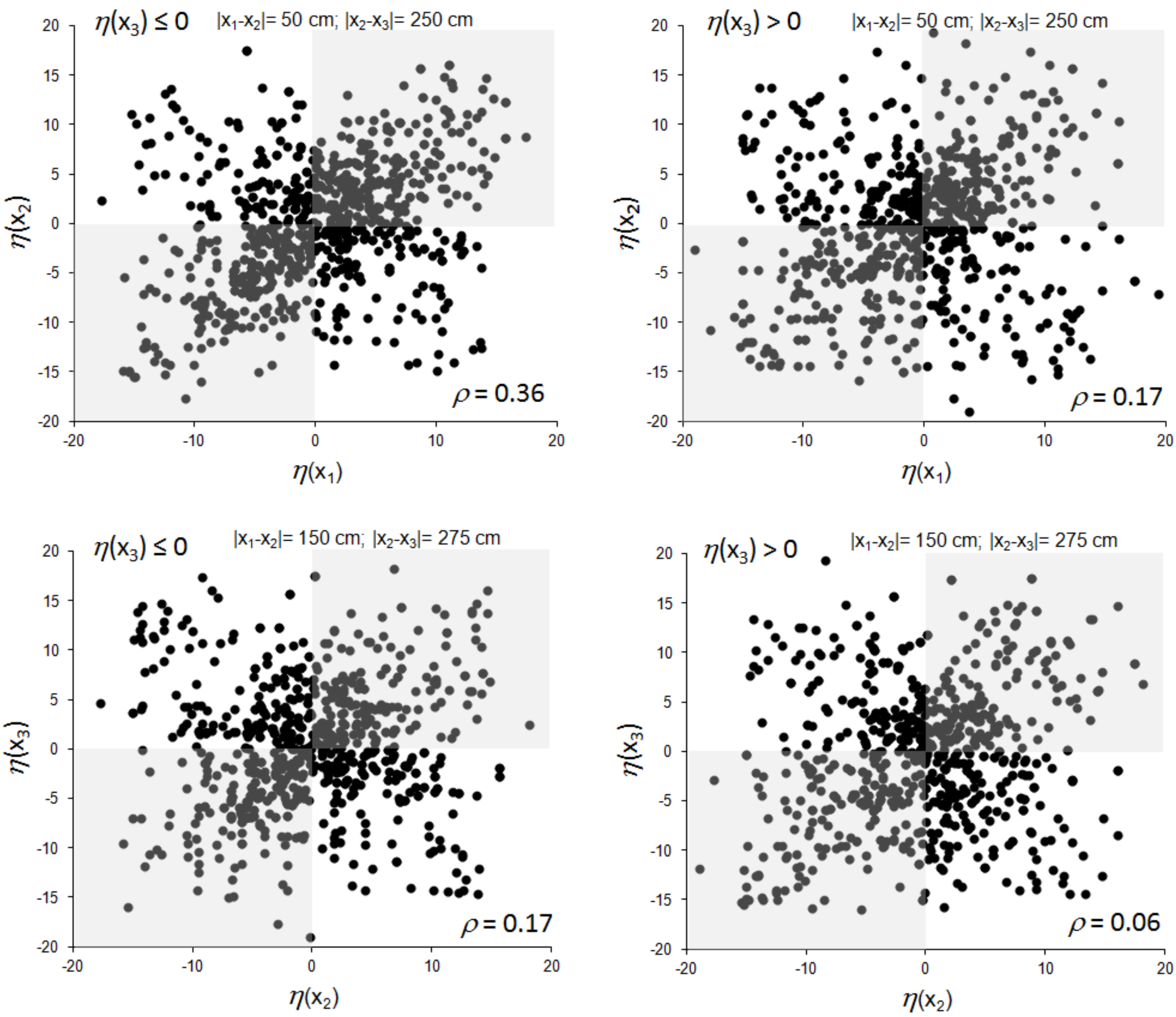
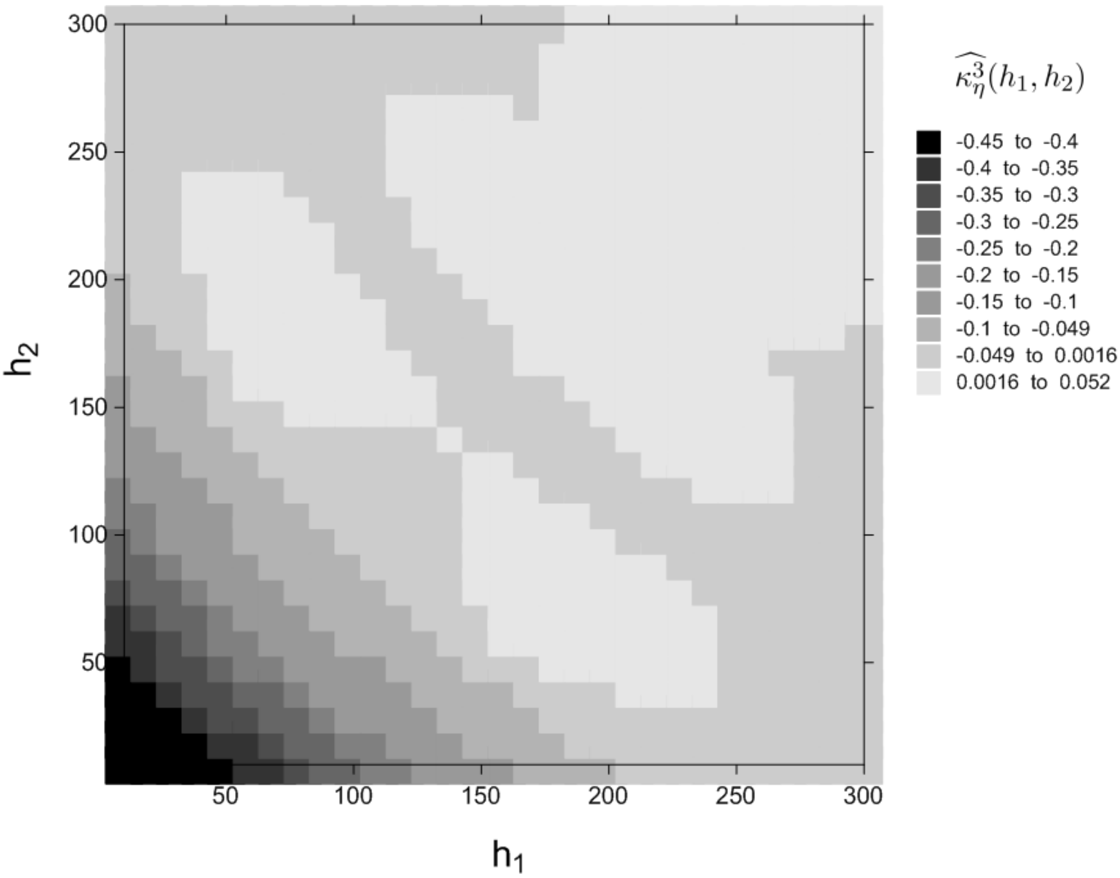

η(x3) ≤ 0
|x1-x2|= 50 cm; |x2-x3|= 250 cm
ρ = 0.36
η(x3) > 0
|x1-x2|= 50 cm; |x2-x3|= 250 cm
ρ = 0.17
η(x3) ≤ 0
|x1-x2|= 150 cm; |x2-x3|= 275 cm
ρ = 0.17
η(x3) > 0
|x1-x2|= 150 cm; |x2-x3|= 275 cm
ρ = 0.06
η(x1)
η(x2)
η(x3)

Figure 4



300
250
200
150
100
50
50
100
150
200
250
300
h_2
h_1
$\widehat{\kappa_\eta^3}(h_1, h_2)$
-0.45 to -0.4
-0.4 to -0.35
-0.35 to -0.3
-0.3 to -0.25
-0.25 to -0.2
-0.2 to -0.15
-0.15 to -0.1
-0.1 to -0.049
-0.049 to 0.0016
0.0016 to 0.052

Figure 6