A new approach to spatial data interpolation using higher-order statistics

Shen Liu[a,c], Vo Anh[a,c], James McGree[a,c], Erhan Kozan[1a,c], Rodney C. Wolff[b,c]

[a] Mathematical Sciences School, Queensland University of Technology, Brisbane
[b] WH Bryan Mining and Geology Research Centre, the University of Queensland, Brisbane
[c] Cooperative Research Centre for Optimised Resource Extraction (CRC-ORE)

Abstract: Interpolation techniques for spatial data have been applied frequently in various fields of geosciences. Although most conventional interpolation methods assume that it is sufficient to use first- and second-order statistics to characterize random fields, researchers have now realized that these methods cannot always provide reliable interpolation results, since geological and environmental phenomena tend to be very complex, presenting non-Gaussian distribution and/or non-linear inter-variable relationship. This paper proposes a new approach to the interpolation of spatial data, which can be applied with great flexibility. Suitable cross-variable higher-order spatial statistics are developed to measure the spatial relationship between the random variable at an unsampled location and those in its neighbourhood. Given the computed cross-variable higher-order spatial statistics, the conditional probability density function (CPDF) is approximated via polynomial expansions, which is then utilized to determine the interpolated value at the unsampled location as an expectation. In addition, the uncertainty associated with the interpolation is quantified by constructing prediction intervals of interpolated values. The proposed method is applied to a mineral deposit dataset, and the results demonstrate that it outperforms kriging methods in uncertainty quantification. The introduction of the cross-variable higher-order spatial statistics noticeably improves the quality of the interpolation since it enriches the information that can be extracted from the observed data, and this benefit is substantial when working with data that are sparse or have non-trivial dependence structures.

Keywords: Geostatistics, interpolation, uncertainty quantification, mineral deposit

[1] Correspondence to: Room P843, P Block, Gardens Point Campus, Queensland University of Technology, Brisbane, Qld 4001, Australia. Tel.: +61 7 31381371; fax: +61 7 31381029. E-mail: e.kozan@qut.edu.au.

1 Introduction

In the analysis of spatially distributed phenomena, samples are collected at a finite number of locations, and it is often necessary to obtain predictions at various unsampled locations. In mineral deposit evaluation, for example, due to the high cost of drilling, drill sampling of orebodies can be of the order of one part in a million. Clearly, estimation of orebody properties is required at un-drilled locations. To carry out such tasks, interpolation techniques are of much interest. A predicted value at an unsampled location is generated as a function of observations at sampled locations, by exploiting measurements of spatial relationships within a neighbourhood of the target location.

The cornerstone of a high-quality interpolation is to examine carefully the distributional properties of the underlying random field, as these properties determine the spatial dependence of the value at the unsampled location on values at its surrounding locations. Distributional properties provide information about the uncertainty of the interpolated values, which means that estimation is not limited to point estimation. As claimed by Schelin and Luna (2010), interpolated values are of interest only if combined with information about their accuracy. This is especially important in geological or environmental applications. For example, in mineral exploration an estimated value of the potential ore resource does not have much meaning if the associated estimation uncertainty is unknown. Furthermore, estimating uncertainty precisely is very important in practice, especially for decision makers, so that strategy is not made upon misleading information about risk. Motivated by this context, the aim of this paper is to propose a new interpolation method to provide information on both point estimates and their associated uncertainty, using robust quantification of spatial features.

In the literature, various approaches have been introduced to study the distributional properties of random fields. Most conventional methods assume that it is sufficient to use the first- and second-order statistics (e.g. variograms) to characterize random fields. Although a portion of past studies showed that some of these methods, such as ordinary kriging and inverse distance weighted interpolation, may achieve desirable performance (Babak, 2013; Hwang et al., 2012; Babak and Deutsch, 2009; Rojas-Avellaneda and Silvan-Cardenas, 2006; Saito et al., 2005), a growing number of researchers have realized that conventional methods have noticeable drawbacks. As claimed by Gaetan and Guyon (2010), Remy et al. (2009) and Chilès and Delfiner (1999), these methods usually rely on linear models and Gaussian

distribution assumptions. However, many natural phenomena exhibit non-trivial spatial features, such as non-linear inter-variable dependence or non-Gaussian distributions. Consequently, conventional methods are often ineffective in spatial modelling and characterization, providing imprecise information about spatial dependence structure. Such ineffectiveness is especially severe in geological or environmental studies, as discussed by Mustapha and Dimitrakopoulos (2011, 2010a, 2010b), Strebelle (2002) and Tjelmeland and Besag (1998). In the past decade, various techniques have been developed to improve the reliability of characterizing random fields, such as bootstrapping (Kleijnen et al., 2012; Schelin and Luna, 2010; Loh and Stein, 2008, 2004; Mukul et al., 2004), copula-based methods (Kazianka, 2013; Pilz et al., 2012; Kazianka and Pilz, 2010a, 2010b; Bárdossy and Li, 2008; Bárdossy, 2006), kernel-based methods (Honarkhah and Caers, 2010; Scheidt and Caers, 2010, 2009a, 2009b), Bayesian (Nieto-Barajas and Sinha, 2014; Troldborg et al., 2012; Pilz et al., 2012; Kazianka and Pilz, 2011, 2012), multi-point simulation methods (De Iaco, 2013; Boucher, 2009; Chugunova and Hu, 2008; Wu et al., 2008; Mirowski et al., 2008; Arpat and Caers, 2007; Zhang et al., 2006; Strebelle, 2002), multi-scale simulations using wavelets (Chatterjee and Dimitrakopoulos, 2012; Gloaguen and Dimitrakopoulos, 2009, 2008), and spatial-cumulant-based simulation methods (Goodfellow et al., 2012; Machuca-Mory and Dimitrakopoulos, 2012; Mustapha et al., 2011; Mustapha and Dimitrakopoulos, 2011, 2010a, 2010b, Dimitrakopoulos et al., 2010).

Although there exists no evidence that methods using higher-order statistics (higher than the first and second order) consistently outperform the conventional ones, past studies, as stated above, have reached the conclusion that higher-order statistics can contribute to the characterization of spatial features to some appreciable extent, especially with respect to describing uncertainty. Higher-order spatial statistics were first introduced by Dimitrakopoulos et al. (2010), followed by a number of studies including Mustapha and Dimitrakopoulos (2010a, 2010b, 2011), Mustapha et al. (2011), Machuca-Mory and Dimitrakopoulos (2012), and Goodfellow et al. (2012). In these studies, the particular form of the higher-order spatial statistics are spatial cumulants, which are used to evaluate the dependence structure of a spatially distributed random variable at an unsampled location x_0, denoted $Z(x_0)$, based on values at some sample locations in the neighbourhood. Dimitrakopoulos et al. (2010) concluded from case studies that spatial cumulants up to and including fifth-order are efficient in reflecting characteristics of various geological patterns, while Mustapha and Dimitrakopoulos (2010a) stated that the fourth and fifth order cumulants

can describe well complex patterns with either sparse or dense data. Nevertheless, spatial cumulants are restricted to a single variable, which cannot be considered for examining the multivariate dependence structure. As a result, such single-variable cumulants lead to potential loss of information about multivariate relationships, and the possibility of enhancing interpolation performance by observing multiple variables is excluded. This is especially undesirable in mining, since processing performance is a function of several grade, geochemical, and geometallurgical variables. For instance, considering both copper and sulphur (which are often co-present) is more meaningful than considering copper only, since sulphur impacts on throughput at several processing stages. To improve on the current state of knowledge, we introduce suitable cross-variable higher-order spatial statistics to examine multivariate dependence structure.

The new interpolation method proposed in this paper can be outlined as follows. Firstly, we obtain the cross-variable higher-order spatial statistics, which are used to characterize the spatial relationship between $Z(x_0)$ and observed values at some sampled locations in the neighbourhood. To compute these statistics empirically, a spatial template is employed to search for data values by pre-determined spatial interrelationship. Given the computed cross-variable statistics, the conditional probability density function (CPDF) of $Z(x_0)$ is approximated using polynomial expansions. In particular, we consider expansions using Legendre polynomials, which have received much interest in spatial modelling (Machuca-Mory and Dimitrakopoulos, 2012; Dimitrakopoulos et al., 2010; Mustapha and Dimitrakopoulos, 2010b; Hosny, 2007). Once the approximation of the CPDF is obtained, the mathematical expectation $\mathrm{E}[Z(x_0)]$ is evaluated as the interpolated value at x_0. To evaluate the associated uncertainty, prediction intervals can be constructed from the approximated CPDF.

The rest of this paper is organized as follows. Section 2 proposes the cross-variable higher-order spatial statistics, which are used to characterize sophisticated spatial dependence structures. Details of the new interpolation method using the statistics from Section 2 are given in Section 3. Section 4 evaluates the performance of the proposed interpolation method by carrying out an application to a mineral deposit data set. Conclusions are drawn in Section 5.

2 Cross-variable higher-order spatial statistics

To achieve a better quantification of sophisticated spatial dependence structures, cross-variable higher-order spatial statistics are proposed in this section. Compared to ordinary single-variable spatial statistics, the cross-variable statistics enrich the information that can be extracted from data, and hence spatial features can be measured more accurately than single-variable statistics.

For a single variable, define the higher-order spatial moment as

$$m_{t_0,t_1,\dots,t_r} = E\left(Z^{t_0}(x)\, Z^{t_1}(x_1) \dots Z^{t_r}(x_r)\right), \tag{1}$$

where x is the reference spatial location, $x_1, \dots, x_r$ denote distinct locations in the neighbourhood of x, and $t_0, t_1, \dots, t_r$ denote the associated orders. For multiple variables, the cross-variable higher-order spatial moment is proposed by generalizing Equation (1). For the sake of simplicity, we only consider two variables in this study which can be generalized in a straightforward way. Let $Z_X(\cdot)$ and $Z_Y(\cdot)$ be two spatial random variables. The cross-variable higher-order spatial moment is defined as

$$m^X_{t_0,t_1,\dots,t_r} = E\left(Z_X^{t_0}(x)\, Z^{t_1}(x_1) \dots Z^{t_r}(x_r)\right), \tag{2}$$

where the superscript on the left hand side represents the variable $Z_X(\cdot)$ with respect to which the interpolation will be implemented. In this case we call $Z_X(\cdot)$ the principal variable and $Z_Y(\cdot)$ the contributing variable, as $Z_Y(\cdot)$ is expected to contribute to the interpolation of $Z_X(\cdot)$. It is important to note that $Z_X^{t_0}(x)$ has a subscript *X* while $Z^{t_1}(x_1) \dots Z^{t_r}(x_r)$ do not, implying that the former is confined to the principal variable but the latter can be either principal or contributing variables. Equation (2) is considered as a $(t_0, t_1, \dots, t_r)$-order cross-variable spatial moment. To illustrate, suppose $r = 3$ with x_1, x_3 associated with $Z_X(\cdot)$ and x_2 associated with $Z_Y(\cdot)$. The (2, 2, 1, 1)-order spatial moment is of the following form:

$$m^X_{2,2,1,1} = E\left(Z_X^2(x)\, Z_X^2(x_1)\, Z_Y(x_2)\, Z_X(x_3)\right).$$

To compute $m^X_{t_0,t_1,\dots,t_r}$ empirically, the first task is to quantify the spatial relationship between locations $x_1, \dots, x_r$ and the reference location x. That is, the relative locations of $x_1, \dots, x_r$ in relation to x need to be specified. For the one-dimensional (1D) cases, these relative locations can be determined by the distances from x to $x_1, \dots, x_r$. For 2D cases,

determining relative locations requires not only the distance but also the direction, which is specified by the azimuth, measured as the rotation clockwise from the north. For 3D cases, the dip, which is measured as the rotation vertically from the horizontal plane, is necessary also to measure relative locations. For the sake of generality, we refer to the "spatial template" technique developed by Dimitrakopoulos et al. (2010). Denote by h_i the distance from x to x_i, $i = 1,2,\dots,r$. Define unit vectors $\overrightarrow{d}_i$ as the directions from x to x_i, which can be expressed in terms of the azimuth and dip. Each distance, together with the corresponding direction, fully quantifies the relative location of an x_i in relation to the reference location. As a consequence, x_i can be re-written as $x + h_i\overrightarrow{d}_i$.

Denote by $T(h_1\overrightarrow{d}_1, \dots, h_r\overrightarrow{d}_r)$ the r-direction spatial template. A set of r+1 spatial locations $\{x_k,\ x_{k_1}, \dots, x_{k_r}\}$ is considered satisfying this template if the r relative locations $x_{k_1}, \dots, x_{k_r}$ in relation to x_k can be quantified by $T(h_1\overrightarrow{d}_1, \dots, h_r\overrightarrow{d}_r)$, i.e.,

$\{x_k,\ x_{k_1}, \dots, x_{k_r}\} \in T(h_1\overrightarrow{d}_1, \dots, h_r\overrightarrow{d}_r)$ if $x_{k_i} = x_k + h_i\overrightarrow{d}_i$ for $i = 1,2,\dots,r$.

Given a defined template, the estimated cross-variable higher-order spatial statistics can be computed empirically as

$$m^X_{t_0,t_1,\dots,t_r} = \frac{1}{N_T}\sum_{k=1}^{N_T} Z_X^{t_0}(x_k)\, Z^{t_1}\left(x_k + h_1\overrightarrow{d}_1\right) \dots Z^{t_r}\left(x_k + h_r\overrightarrow{d}_r\right), \quad (3)$$

where N_T denotes the number of sets of r locations that satisfy the template. Again, it should be stressed that $Z^{t_1}\left(x_k + h_1\overrightarrow{d}_1\right), \dots, Z^{t_r}\left(x_k + h_r\overrightarrow{d}_r\right)$ can be either principal or contributing variables. The computation of (3) involves carrying out an exhaustive searching process to discover all sets of locations that satisfy the template. The nearest neighbour search is incorporated in this study, which can be illustrated as follows. Fig. 1 displays an example in 2D space. Suppose a sample of data, as presented by Fig. 1a, is observed in 2D space. Assume that the interpolation is carried out at an unsampled location, denoted x_0 (Fig. 1b). The observed data points are searched within the neighbourhood of x_0, and the corresponding locations are denoted x_1, x_2 and x_3, respectively (Fig. 1c). As the number of observations in the neighbourhood is 3, a 3-direction linkage from x_0 to x_1, x_2 and x_3 is considered as a template (Fig. 1d). Since it is not reasonable to expect irregularly distributed data points falling exactly on the ends of the template, tolerance is taken into account for each template direction. The spatial template at the unsampled location x_0, with respect to the three sampled

locations in the neighbourhood, is then determined (Fig. 1e). Note that when determining the template, the unsampled location x_0 is known as the reference location. As long as the reference location does not change, the specifications of the template are fixed. That is, the length, azimuth and dip of each direction of the template remain the same throughout the searching process, which is with regard to a fixed reference location.

Once the template is determined, searching is carried out throughout the entire sample. To better understand how to determine if the template is satisfied, imagine that Fig. 1e is a transparent slide moving over Fig. 1a. At a sampled location, if there exist three observations that are within the three boxes of Fig. 1e respectively, the template is satisfied. For example, in Fig. 1f the template is not satisfied at locations x_1 and x_4 but at x_5, since observations at x_6, x_7 and x_8 are within the tolerance of the pre-defined template.

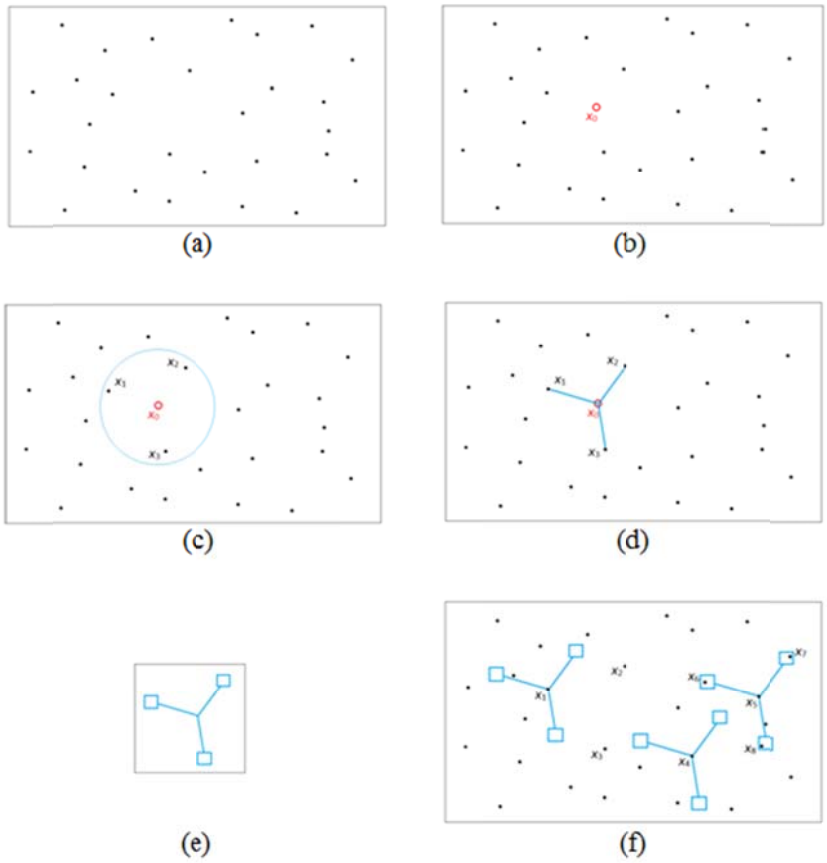

Fig. 1 An illustrative example of the searching template

Once the exhaustive search is completed, the cross-variable higher-order spatial statistics can be computed using data points that satisfy the template.

3 A new interpolation technique

In this section a new approach to interpolating spatial data is developed. The aim is to estimate a local conditional probability density function (CPDF) at each unsampled location using the observed data. Let Z be a real-valued stationary and ergodic random field in $\mathbb{R}^s$. In mining applications, we consider $s = 3$. Assume that $Z_X(\cdot)$ and $Z_Y(\cdot)$ are two random variables and $\{Z_X(x_1), \ldots, Z_X(x_n)\}$ and $\{Z_Y(y_1), \ldots, Z_Y(y_m)\}$ are observations on $Z_X(\cdot)$ and $Z_Y(\cdot)$ at locations $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$, respectively. Note that $x_i \neq x_j$ and $y_i \neq y_j$ for $i \neq j$, but we allow $x_i = y_j$ for some i and j, i.e., it is possible to observe both variables at one location.

Denote by x_0 and y_0 the unsampled locations, at which the interpolation of $Z_X(\cdot)$ and $Z_Y(\cdot)$ is carried out by estimating the following local density functions, conditioning on the hard data:

$f(Z_X(x_0)|Z_X(x_1), \ldots, Z_X(x_n), Z_Y(y_1), \ldots, Z_Y(y_m))$, and

$f(Z_Y(y_0)|Z_X(x_1), \ldots, Z_X(x_n), Z_Y(y_1), \ldots, Z_Y(y_m))$.

Without loss of generality, only the CPDF of $Z_X(x_0)$ is studied hereafter. By the Bayes rule, the CPDF can be expressed as

$$f(Z_X(x_0)|Z_X(x_1), \ldots, Z_X(x_n), Z_Y(y_1), \ldots, Z_Y(y_m)) = \frac{f(Z_X(x_0), Z_X(x_1), \ldots, Z_X(x_n), Z_Y(y_1), \ldots, Z_Y(y_m))}{f(Z_X(x_1), \ldots, Z_X(x_n), Z_Y(y_1), \ldots, Z_Y(y_m))}. \quad (4)$$

The estimation of (1) only involves the approximation of the numerator of the right hand side of (4), namely $f(Z_X(x_0), Z_X(x_1), \ldots, Z_X(x_n), Z_Y(y_1), \ldots, Z_Y(y_m))$, referred to as $f(\boldsymbol{Z}_X)$ hereafter. This is because the denominator can be obtained from $f(\boldsymbol{Z}_X)$ by integrating out $Z_X(x_0)$ over its support. To achieve a high-quality approximation of $f(\boldsymbol{Z}_X)$, we utilize the cross-variable higher-order spatial statistics to quantify the dependence structure of $Z_X(x_0)$ on the observed values in the neighbourhood of x_0. By doing so, the approximation of the CPDF is expected to be more accurate than that produced from ordinary single-variable spatial statistics, as the information associated with contributing variable(s) can be extracted and contribute to the approximation.

Given the computed cross-variable higher-order spatial statistics, $f(\boldsymbol{Z}_X)$ can be expressed as an infinite series expansion in terms of the Legendre polynomials, following Mustapha and Dimitrakopoulos (2010a, b). Legendre functions are solutions to the Legendre's differential equation (Liu and Spiegel, 1999), which is of the form

$$\frac{d}{dx}\left[(1-x^2)\frac{d}{dx}P_p(x)\right]+p(p+1)P_p(x)=0,$$

and for $p=0,1,2,\ldots$ the solutions form the Legendre polynomial sequence. The p^{th}-order Legendre polynomial is given by

$$P_p(x)=\frac{1}{2^p p!}\left(\frac{d}{dx}\right)^p[(x^2-1)^p]=\sum_{k=0}^{p}a_{k,p}x^k, \tag{5}$$

and the following recursive relation holds:

$$P_{p+1}(x)=\frac{(2p+1)}{(p+1)}xP_p(x)-\frac{p}{(p+1)}P_{p-1}(x).$$

It is obvious that $P_0(x)=1$ and $P_1(x)=x$, and using the recursive relation above one may obtain that $P_2(x)=(3x^2-1)/2$, $P_3(x)=(5x^3-3x)/2$, and so on. See Liu and Spiegel (1999) for more details. An important property of the Legendre polynomials is that they are orthogonal in the interval [-1, 1].

To express $f(\boldsymbol{Z}_X)$ in terms of the Legendre polynomials, the normalized version of (5) is required, which is given by

$$\bar{P}_p(x)=\sqrt{\frac{2p+1}{2}}P_p(x),$$

and $f(\boldsymbol{Z}_X)$ is then expressed as

$$f(\boldsymbol{Z}_X)=\sum_{i_0=0}^{\infty}\sum_{i_1=0}^{\infty}\ldots\sum_{i_n=0}^{\infty}\sum_{i_{n+1}=0}^{\infty}\ldots\sum_{i_{n+m}=0}^{\infty}L_{\,i_0,i_1,\ldots,i_n,\ldots,i_{n+m}}\times\bar{P}_{i_1}\big(Z_X(x_1)\big)\times\ldots\times\bar{P}_{i_n}\big(Z_X(x_n)\big)$$

$$\times\bar{P}_{i_{n+1}}\big(Z_Y(y_1)\big)\times\ldots\times\bar{P}_{i_{n+m}}(Z_Y(y_m))\times\bar{P}_{i_0}\big(Z_X(x_0)\big),$$

where $L_{\,i_0,i_1,\ldots,i_n,\ldots,i_{n+m}}$ are the polynomial coefficients. In practice, only a finite order, denoted ω, is considered such that $f(\boldsymbol{Z}_X)$ has the following truncation approximation:

$$f(\boldsymbol{Z}_X)\approx\sum_{i_0=0}^{\omega}\sum_{i_1=0}^{i_0}\ldots\sum_{i_n=0}^{i_{n-1}}\sum_{i_{n+1}=0}^{i_n}\ldots\sum_{i_{n+m}=0}^{i_{n+m-1}}L_{\,\overline{i_0},\overline{i_1},\ldots,\overline{i_n},\ldots,i_{n+m}}\times\bar{P}_{\overline{i_1}}\big(Z_X(x_1)\big)\times\ldots\times\bar{P}_{\overline{i_n}}\big(Z_X(x_n)\big)$$

$$\times\bar{P}_{\overline{i_{n+1}}}\big(Z_Y(y_1)\big)\times\ldots\times\bar{P}_{i_{n+m}}(Z_Y(y_m))\times\bar{P}_{\overline{i_0}}\big(Z_X(x_0)\big)=\hat{f}(\boldsymbol{Z}_X), \tag{6}$$

where $\bar{\imath_k} = i_k - i_{k+1}$, $\hat{f}(\boldsymbol{Z}_X)$ denotes the estimated CPDF, and $L_{\bar{\imath_0},\bar{\imath_1},\ldots,\overline{\imath_n},\ldots,i_{n+m}}$ are the coefficients of the Legendre polynomials of different orders. Mustapha and Dimitrakopoulos (2010a, b) state that based on the orthogonality property of the Legendre polynomials, these coefficients can be expressed as follows:

$$L_{\bar{\imath_0},\bar{\imath_1},\ldots,\overline{\imath_n},\ldots,i_{n+m}}$$

$$= \int \bar{P}_{\bar{\imath_0}}\big(Z_X(x_0)\big)\bar{P}_{\bar{\imath_1}}\big(Z_X(x_1)\big) \ldots \bar{P}_{i_{n+m}}\big(Z_Y(y_m)\big) f(\boldsymbol{Z}_X) dZ_X(x_0) dZ_X(x_1) \ldots Z_Y(y_m), \quad (7)$$

and following (5), (7) becomes

$$L_{\bar{\imath_0},\bar{\imath_1},\ldots,\overline{\imath_n},\ldots,i_{n+m}} = \sqrt{\frac{2\bar{\imath_0}+1}{2}} \ldots \sqrt{\frac{2i_{n+m}+1}{2}}$$

$$\times \sum_{t_0=0}^{\bar{\imath_0}} a_{t_0,\bar{\imath_0}} \ldots \sum_{t_{n+m}=0}^{\overline{\imath_{n+m}}} a_{t_{n+m},i_{n+m}} \int Z_X^{t_0}(x_0) \ldots Z_Y^{t_{n+m}}(y_m) f(\boldsymbol{Z}_X) dZ_X(x_0) dZ_X(x_1) \ldots dZ_Y(y_m)$$

$$= \sqrt{\frac{2\bar{\imath_0}+1}{2}} \ldots \sqrt{\frac{2i_{n+m}+1}{2}} \sum_{t_0=0}^{\bar{\imath_0}} a_{t_0,\bar{\imath_0}} \ldots \sum_{t_{n+m}=0}^{\overline{\imath_{n+m}}} a_{t_{n+m},i_{n+m}} E\left(Z_X^{t_0}(x_0) \ldots Z_Y^{t_{n+m}}(y_m)\right), \quad (8)$$

where $E\left(Z_X^{t_0}(x_0) \ldots Z_Y^{t_{n+m}}(y_m)\right) := m^X_{t_0,\ldots,t_{n+m}}$ is the cross-variable higher-order spatial moment as discussed in the previous section. Given that $m^X_{t_0,\ldots,t_{n+m}}$ can be computed using (3) and $a_{k,p}$'s are constants, $L_{\bar{\imath_0},\bar{\imath_1},\ldots,\overline{\imath_n},\ldots,i_{n+m}}$ is then computable and hence the approximation of $f(\boldsymbol{Z}_X)$ using (6) is feasible.

In summary, $f(\boldsymbol{Z}_X)$ is approximated by the following steps:

i. Within the neighbourhood of the unsampled location x_0, determine r observed values (usually the closest ones to x_0) that the CPDF is conditional on, and then construct the r-direction spatial template;
ii. Carry out the searching process as explained in Section 2;
iii. Compute the cross-variable higher-order spatial statistics using (3);
iv. Compute the Legendre polynomial coefficients using (8);
v. Compute the Legendre polynomials of different orders using (5); and
vi. Using the outcome from Steps iv and v, obtain the approximated CPDF $\hat{f}(\boldsymbol{Z}_X)$ by (6).

Once $f(\boldsymbol{Z}_X)$ is approximated, the CPDF at the unsampled location x_0, denoted $\hat{f}_{\boldsymbol{Z}_X}\big(Z_X(x_0)\big)$, is estimated by

$$\hat{f}_{\boldsymbol{Z}_X}\big(Z_X(x_0)\big) = \hat{f}\big(Z_X(x_0)\big|Z_X(x_1),\dots,Z_X(x_n),Z_Y(y_1),\dots,Z_Y(y_m)\big) = \frac{\hat{f}(\boldsymbol{Z}_X)}{\int \hat{f}(\boldsymbol{Z}_X)\,\mathrm{d}Z_X(x_0)},$$

and then the interpolated value at x_0, given the observed data in its neighbourhood, is

$$E\big(Z_X(x_0)\big|Z_X(x_1),\dots,Z_X(x_n),Z_Y(y_1),\dots,Z_Y(y_m)\big) = \int Z_X(x_0)\hat{f}_{\boldsymbol{Z}_X}\big(Z_X(x_0)\big)\mathrm{d}Z_X(x_0). \quad (9)$$

For a pre-determined nominal level of significance α, the $100(1-\alpha)\%$ prediction interval can be constructed as follows:

$$\left[\hat{F}^{-1}_{\boldsymbol{Z}_X}\left(\frac{\alpha}{2}\right),\ \hat{F}^{-1}_{\boldsymbol{Z}_X}\left(1-\frac{\alpha}{2}\right)\right], \quad (10)$$

where $\hat{F}_{\boldsymbol{Z}_X}(\cdot)$ denotes the cumulative distribution function that is obtained from $\hat{f}_{\boldsymbol{Z}_X}(\cdot)$, and $\hat{F}^{-1}_{\boldsymbol{Z}_X}(\cdot)$ is the inverse function of $\hat{F}_{\boldsymbol{Z}_X}(\cdot)$.

It should be stressed that the validity of approximating $f(\boldsymbol{Z}_X)$ using (4) builds on the property that the Legendre polynomials are orthogonal on the interval [-1, 1]. As a result, to apply the proposed method one should scale all data values to $[-1, 1]^s$ (s = 1, 2 or 3) beforehand, and back-transform the interpolated values afterwards.

Although the methodology presented above considers only two variables (one principal and one contributing), it can be extended in a straightforward manner to work with more than two variables. Note that, as the number of the contributing variables increases, the computational cost increases dramatically. In practice, one may select a small number of contributing variables that have the greatest explanatory power over the principal variable for the sake of computational efficiency.

4 An empirical study

In this section, the proposed interpolation method is applied to drill-hole data extracted from a mineral deposit located in Australia. A total of 1453 observations were obtained from the drill-holes, each of which contains the observed copper grade (Cu) and sulphur content (S).

Table 1 provides descriptive statistics, while Fig. 2 displays the map and histograms of drill-hole samples[2]. It is observed that both Cu and S show severe skewness to the right and large excess kurtosis compared to a Gaussian distribution, indicating that considering only the first- and second-order statistics is not appropriate.

Table 1. Descriptive statistics of Cu and S

	Cu	**S**
Mean	0.7603	0.9071
Median	0.6010	0.7000
Maximum	5.9000	6.0290
Minimum	0.0080	0.0100
Coefficient of Variation	0.9034	0.8976
Skewness	1.8765	1.5888
Kurtosis	8.8913	5.8855

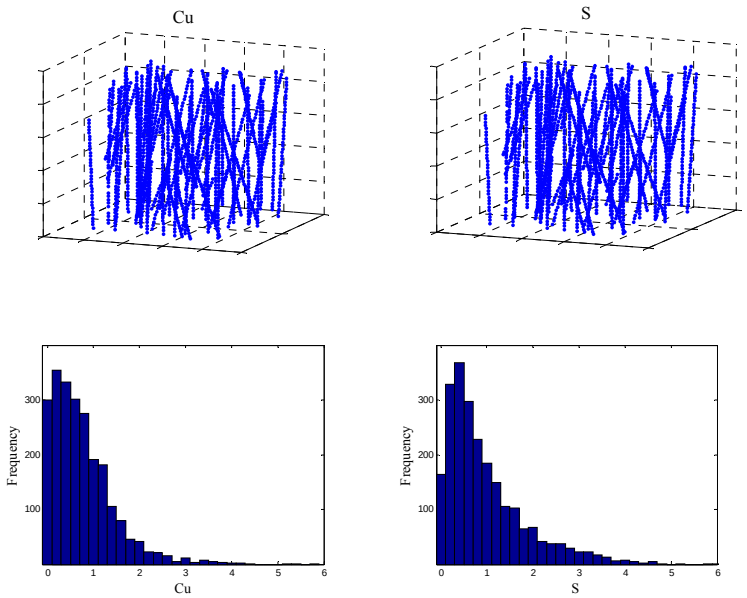

Fig. 2. Drill-hole maps (first row) and histograms of Cu and S (second row)

To evaluate the performance of the newly proposed interpolation method, we examine both the interpolation accuracy and the quality of uncertainty quantification. The former is

[2] Due to confidentiality issues, drill-hole samples are uniformly coloured in the map, i.e. specific values of Cu and S are not displayed.

achieved by computing interpolation errors, while the latter is carried out by inspecting prediction intervals. In particular, the "leave-one-out" cross-validation process is implemented. We delete one observation at a time and treat it as if it were unknown, and then the newly proposed interpolation method is applied to obtain an empirical distribution function.

To examine the interpolation accuracy, the interpolated value is obtained by (9) and the interpolation error is calculated as the difference between the interpolated value and the real value. The mean absolute error (MAE) is computed as the average of all individual interpolation errors, and it is used as an overall measure of interpolation accuracy. To quantify the uncertainty associated with the interpolation, we generate a prediction interval by (10) with the nominal coverage rate equal to 0.95. The generated prediction interval is expected to cover the true value for 95% of the time over repeated sampling. To examine this, the mean coverage rate is calculated as

$$C = \frac{\sum_{i=1}^{N} 1(\hat{L}_i \leq Z_i \leq \hat{U}_i)}{N},$$

where N is the total number of constructed prediction intervals, Z_i is the true value of the i^{th} prediction interval, $\hat{L}_i$ and $\hat{U}_i$ are the lower and upper bounds of the i^{th} prediction interval respectively, and the function $1(\cdot)$ is the indicator function which returns to 1 if the condition inside the parentheses is satisfied and 0 otherwise. If a prediction interval provides an accurate quantification of uncertainty, the calculated mean coverage rate C should be close to the nominal 95% level. To test whether C is statistically different from the nominal level, we follow Kim et al. (2011) and construct the following confidence interval:

$$\left[p - 1.96\sqrt{\frac{p(1-p)}{N}}, p + 1.96\sqrt{\frac{p(1-p)}{N}}\right], \tag{11}$$

where $p = 0.95$ is the nominal coverage rate, and 1.96 is the two-tail critical value of the standard normal distribution at the 5% level of significance. If C is within this interval, we do not reject the null hypothesis that the coverage rate is equal to the nominal rate at the 95% level. Furthermore, we examine the mean width of the estimated prediction intervals, which is calculated as $W = N^{-1}\sum_{i=1}^{N}(\hat{U}_i - \hat{L}_i)$. A higher value of W implies more uncertainty associated with the interpolation, and hence the prediction intervals are less informative.

Note that there are two parameters that need to be specified before implementing the proposed method, namely, the number of directions in the searching template (r in Eq. 1, 2 and 3), and the order to which the CPDF is approximated (ω in Eq. 6). One would expect more precise interpolation results as r and ω increase, since larger r and ω values imply better approximation to the CPDF. In this study, we consider r = 2, 3, 4, 5, 6 or 7. For the approximation order, preliminary results showed that the proposed interpolation method failed to achieve good performance when $\omega \leq 3$. As a consequence, we consider ω = 4, 6, 8 or 10 in this study. Note that an increase in either r or ω will cause extra computational complexity. It is observed from the computation below that holding other conditions constant, every unit increased in ω results in a bigger amount of extra computing time than that of every unit increased in r.

Once r is specified, the searching template is determined as the spatial linkage from a reference location to the r nearest data points in its neighbourhood. This implies that for irregularly spaced data the searching template is not universal but location-dependent, i.e. it may vary from one reference location to another. To determine if the cross-variable higher-order spatial statistics can improve the interpolation performance, results generated from cross-variable higher-order spatial statistics are compared to those from single-variable higher-order spatial statistics. Since both Cu and S are observed in each of the drill-hole samples, the spatial template for the contributing variable is the same as that of the principal variable. To determine if the newly proposed interpolation method can outperform conventional interpolation techniques, we consider the widely used kriging method as the benchmark. Both the interpolation accuracy and uncertainty quantification performance of the newly proposed method are evaluated against those of the kriging method. When single-variable higher-order spatial statistics are used, the produced results are compared to those from ordinary kriging[3]. When cross-variable higher-order spatial statistics are considered, the performance is evaluated against that of the co-kriging. The spherical variogram model is employed to describe spatial dependence when the kriging is carried out, and the parameters are estimated by the weighted least squares method (Cressie, 1985). The isotropy assumption is imposed, as directional semivariograms do not exhibit obvious anisotropy. MAE's are used to determine which interpolation method is more accurate, while the 95% prediction intervals are used to determine which method performs better in uncertainty quantification. We prefer

[3] As pointed out by Li et al. (2010), ordinary kriging is the most widely used geostatistical method, producing the best linear unbiased predictions.

the prediction interval whose mean coverage rate is within the bounds stated by (11). If such a condition is achieved by a number of prediction intervals, we prefer the one with the smaller value of the mean width.

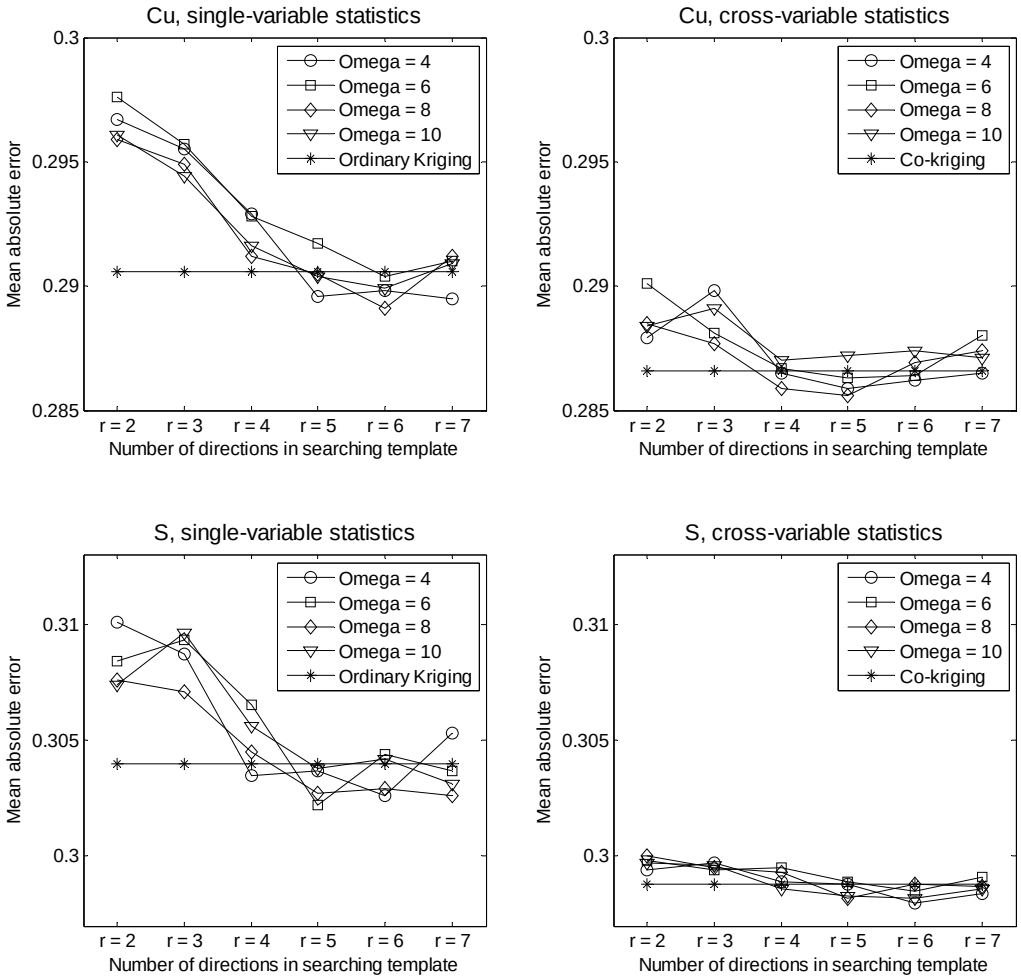

Fig. 3. Mean absolute error comparison

Fig. 3 displays the comparison of the MAE values obtained from single-variable higher-order statistics, cross-variable higher-order statistics and the kriging methods. The first and second columns correspond respectively to the interpolation methods using single- and cross-variable higher-order statistics, while the upper and lower rows exhibit in turn the errors of the interpolated Cu and S values. In each subplot, five sets of MAE values are displayed. The first four sets correspond to the newly proposed interpolation method using four different approximation orders (ω = 4, 6, 8, 10), while the other is associated to the kriging method which is represented by a solid line with asterisk markers. All the MAE values are plotted

against various numbers of searching template directions (r = 2, …,7) which are labelled on the horizontal axis.

In terms of the interpolation accuracy, it is observed from Fig. 3 that the newly proposed method performs similarly to the kriging methods when the number of template directions is larger than or equal to 4. No matter whether the single- or cross-variable higher-order statistics are employed, the MAE values of the proposed interpolation method are fairly close to those of the kriging methods when $r \geq 4$. When $r = 2$ or 3, the proposed interpolation method produces higher MAE's than the benchmark, but the differences in values are not very large, especially when the cross-variable higher-order spatial statistics are employed.

Comparing the first and second columns of Fig. 3, it is clear that incorporating the cross-variable higher-order spatial statistics substantially improves the accuracy of the proposed interpolation method. The interpolation errors produced from the cross-variable statistics are noticeably smaller than those from the single-variable statistics. This is in accordance with our hypothesis that the cross-variable higher-order spatial statistics enrich the information that can be extracted from the observed data, and hence can lead to more accurate interpolation. Such a conclusion is supported by the comparison between ordinary kriging and co-kriging: the latter produces lower MAE's than the former.

Having seen that the proposed method and the kriging methods achieve similar interpolation accuracy when $r \geq 4$, the uncertainty quantification performance is then examined. Fig. 4 and 5 display the mean coverage rates and mean widths of the produced prediction intervals at the 95% nominal level.

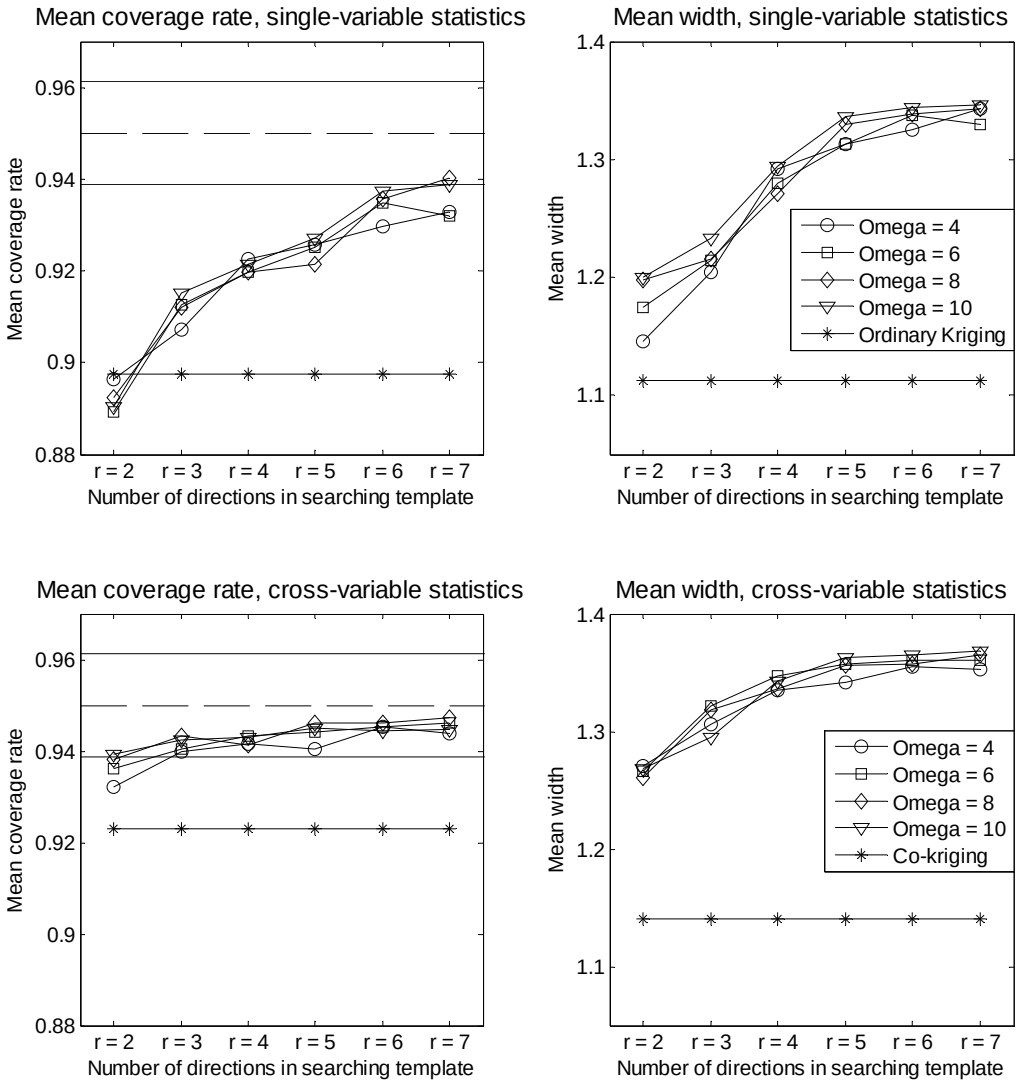

Fig. 4. Mean coverage rate and mean width comparison, Cu

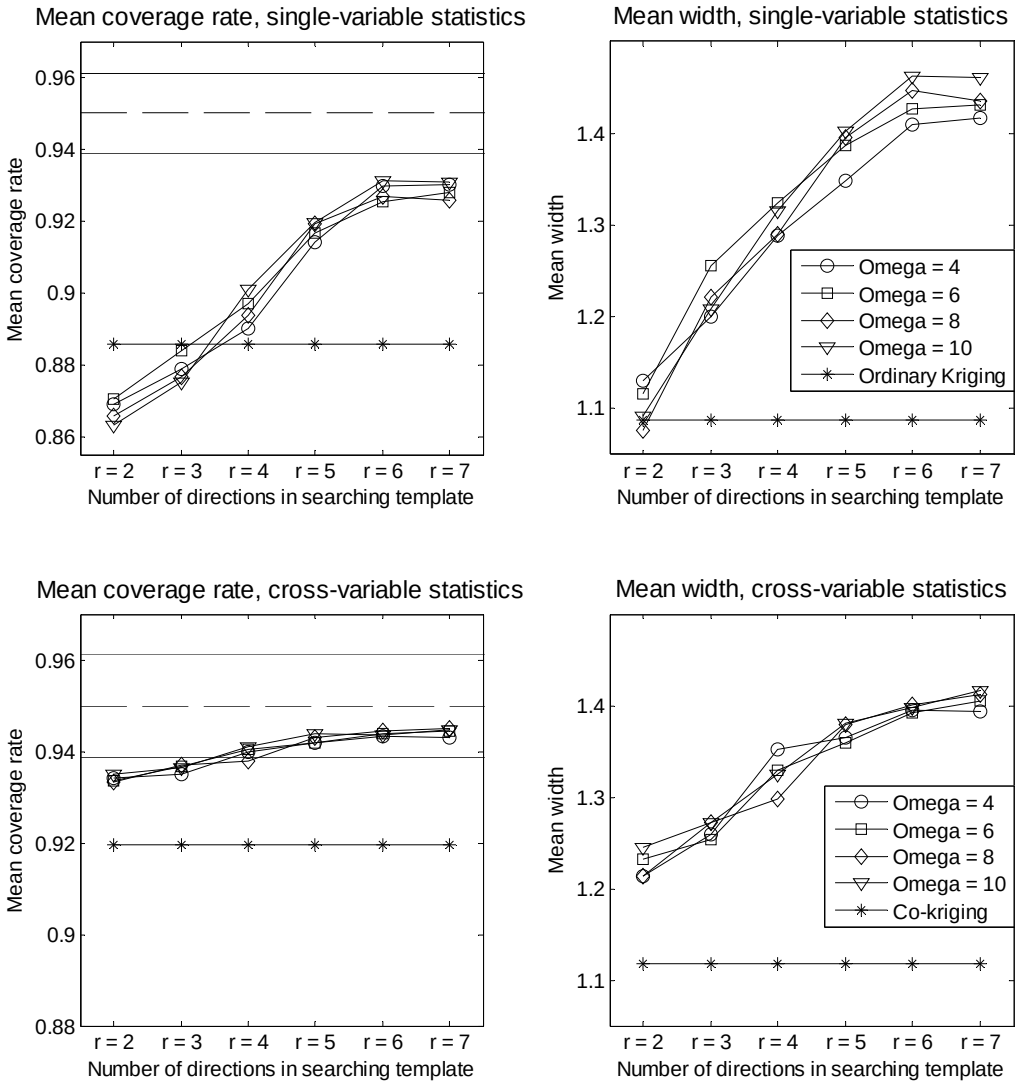

Fig. 5. Mean coverage rate and mean width comparison, S

Fig. 4 and Fig. 5 exhibit the comparison of the mean coverage rates (first column) and mean widths (second column) of the prediction intervals that are constructed from single- and cross-variable higher-order statistics as well as the kriging methods. The first and second rows in each figure correspond respectively to the interpolation methods using single- and cross-variable higher-order statistics. As before, various numbers of searching template directions are labelled on the horizontal axis, and the results associated to the kriging methods are represented by a solid line with asterisk markers. The solid and dashed

horizontal lines in the coverage rate graphs indicate 95% confidence intervals around the nominal coverage rate of 0.95. Such confidence intervals are constructed according to (11).

Fig. 4 reports the results when the copper grade is treated as the principal variable. It is obvious that as the number of directions in searching template increases, both the mean coverage rate and mean width of the proposed interpolation method tend to increase. When the single-variable higher-order spatial statistics are used, most of the observed mean coverage rates are outside the confidence interval of the nominal 95% level, implying that they are statistically different from the nominal rate. However, if the cross-variable higher-order spatial statistics are employed, the newly proposed interpolation method has mean coverage rates within the confidence interval in most cases, implying that the nominal coverage rate has been achieved. In contrast, the kriging methods appear to under-cover the true values severely, as the corresponding coverage rates are far below the lower limit of the confidence interval. This feature is also evident from the mean width graphs, where the kriging prediction intervals are much narrower than the others. Similar conclusions can be drawn from Fig. 5 which displays the results when the sulphur content is considered as the principal variable. When the single-variable higher-order statistics are used, none of the coverage rates is within the confidence interval. Nonetheless, for $r \geq 4$ the coverage rates of the proposed interpolation method are closer to the nominal rate. When the cross-variable higher-order statistics are employed, for $r \geq 4$ most coverage rates of the proposed method are within the confidence band, implying desirable performance in assessing uncertainty. The good performance remains for all the selected approximation orders, as those ω values have produced very similar results. In contrast, the kriging methods grossly underestimate the uncertainty, producing low coverage rates and relatively narrow prediction intervals.

It is summarized from Fig. 4 and 5 that the proposed interpolation method appears to achieve better performance in uncertainty quantification than the kriging methods, as its mean coverage rates are much closer to the nominal rate when the number of searching template directions is not too small. The kriging methods tend to grossly underestimate the uncertainty, which are evidenced by lower coverage rates and smaller interval widths. Moreover, the cross-variable higher-order spatial statistics enhances the uncertainty quantification substantially, as a much greater proportion of coverage rates are within the confidence interval around the nominal rate.

5 Concluding remarks

In reality, geological or environmental phenomena often present sophisticated spatial features which cannot be adequately modelled by the methods that only consider first- and second-order statistics. In this study, we relaxed the Gaussianity and linearity assumptions and proposed a new interpolation method. We developed cross-variable higher-order spatial statistics to characterize non-Gaussian distributions and nonlinear dependence patterns associated with the data, and on this basis the conditional probability density function at each of the unsampled locations was approximated using polynomial expansions. Other than the Legendre polynomials that were considered in this study, other types, such as the Laguerre polynomials, might improve the interpolation performance. Further research is being undertaken to address this issue.

The application to a mineral deposit data set showed that the proposed technique is superior to the kriging method if the number of searching template directions is not less than 4, Such superiority was demonstrated by considering jointly the interpolation error and uncertainty quantification performance. While the interpolation errors of the newly proposed method are close to those of the kriging methods, the prediction intervals constructed based on higher-order spatial statistics showed better performance in uncertainty quantification as the coverage rates are much closer to the nominal level than those from the kriging methods. In addition, it was demonstrated that incorporating cross-variable statistics can enhance both the interpolation accuracy and the uncertainty quantification.

We believe that this research has values for industrial applications. Incorporating the cross-variable higher-order statistics, rather than the single-variable statistics, enriches the information that can be extracted from data. In addition to an interpolated value, the density approximation provides information about its associated uncertainty, and further statistical inferences are available as well. Therefore, decision makers can benefit largely from the proposed method. Future research will be carried out with respect to blocks of ore bodies, including the interpolation of unsampled blocks, and the quantification of the uncertainty associated with the interpolation.

Acknowledgements

The authors would like to acknowledge the support of CRC-ORE, established and supported by the Australian Government's Cooperative Research Centres Programme. The authors are also grateful to the two anonymous referees for their valuable comments and suggestions that have contributed to improving the quality and presentation of this paper.

References

Arpat B, Caers J (2007) Conditional simulation with patterns. Math Geosci 39:177–203

Babak O (2013) Inverse distance interpolation for facies modeling. Stoch Environ Res Risk Assess doi: 10.1007/s00477-013-0833-8

Babak O, Deutsch CV (2009) Statistical approach to inverse distance interpolation. Stoch Environ Res Risk Assess 23:543–553

Bárdossy A (2006) Copula-based geostatistical models for groundwater quality parameters. Water Resour Res 42:W11416

Bárdossy A, Li J (2008) Geostatistical interpolation using copulas. Water Resour Res 44:W07412

Boucher A (2009) Considering complex training images with search tree partitioning. Comput Geosci 35:1151–1158

Chatterjee S, Dimitrakopoulos R (2012) Multi-scale stochastic simulation with a wavelet-based approach. Comput Geosci 45:177–189

Chilès JP, Delfiner P (1999) Geostatistics—modeling spatial uncertainty. New York: Wiley

Chugunova TL, Hu, LY (2008) Multiple point simulations constrained by continuous auxiliary data. Math Geosci 40:133–146

Cressie N (1985) Fitting variogram models by weighted least squares. Math Geol 17:563–586

De Iaco S (2013) On the use of different metrics for assessing complex pattern reproductions. J Appl Stat doi: 10.1080/02664763.2012.754853

Dimitrakopoulos R, Mustapha H, Gloaguen E (2010) High-order statistics of spatial random fields: exploring spatial cumulants for modelling complex, non-Gaussian and non-linear phenomena. Math Geosci 42:65–99

Gaetan C, Guyon X (2010) Spatial Statistics and Modeling. New York: Springer

Gloaguen E, Dimitrakopoulos R (2008) Conditional wavelet based simulation of nonstationary geology using geophysical and model analogue information. Paper presented at the Proceedings of Geostats 2008—VIII International Geostatistics Congress, Santiago, Chile

Gloaguen E, Dimitrakopoulos R (2009) Two-dimensional conditional simulations based on the wavelet decomposition of training images. Math Geosci 41:679–701

Goodfellow R, Mustapha H, Dimitrakopoulos R (2012) Approximations of high-order spatial statistics through decomposition. Quantitative Geology and Geostatistics, Geostatistics Oslo 2012, 17, 91-102. doi: 10.1007/978-94-007-4153-9_8

Honarkhah M, Caers J (2010) Stochastic simulation of patterns using distance-based pattern modelling. Math Geosci 42:487–517

Hosny KM (2007) Exact Legendremoment computation for gray level images. Pattern Recogn 40:3597–3605

Hwang Y, Clark M, Rajagopalan B, Leavesley G (2012) Spatial interpolation schemes of daily precipitation for hydrologic modeling. Stoch Environ Res Risk Assess 26:295–320

Kazianka H (2013) Approximate copula-based estimation and prediction of discrete spatial data. Stoch Environ Res Risk Assess 27:2015–2026

Kazianka H, Pilz J (2010a) Copula-based geostatistical modeling of continuous and discrete data including covariates. Stoch Environ Res Risk Assess 24:661–673

Kazianka H, Pilz J (2010b) Spatial interpolation using copula-based geostatistical models. In P. Atkinson & C. Lloyd (Eds.), geoENV VII - Geostatistics for environmental applications. Berlin: Springer

Kazianka H, Pilz J (2011) Bayesian spatial modeling and interpolation using copulas. Comput Geosci 37:310–319

Kazianka H, Pilz J (2012) Objective Bayesian analysis of spatial data with uncertain nugget and range parameters. Can J Stat 40:304–327

Kim JH, Wong K, Athanasopoulos G, Liu S (2011). Beyond point forecasting: Evaluation of alternative prediction intervals for tourist arrivals. Int J Forecasting 27:887-901

Kleijnen JPC, Mehdad E, Beers WCM (2012) Convex and monotonic bootstrapped kriging. Proceedings of the 2012 Winter Simulation Conference

Li M, Shao Q, Renzullo L (2010) Estimation and spatial interpolaiton of rainfall intensity distribution form the effective rate of precipitation. Stoch Environ Res Risk Assess 24:117–130

Liu J, Spiegel MR (1999) Mathematical Handbook of Formulas and Tables (2nd ed.). New York: McGraw-Hill

Loh JM, Stein ML (2004) Bootstrapping a spatial point process. Stat Sinica 14:69-101

Loh JM, Stein ML (2008) A valid and fast spatial bootstrap for correlation functions. Astrophys J 681:726-734

Nieto-Barajas LE, Sinha T (2014) Bayesian interpolation of unequally spaced time series. Stoch Environ Res Risk Assess doi: 10.1007/s00477-014-0894-3

Machuca-Mory DF, Dimitrakopoulos R (2012) Simulation of a structurally-controlled gold deposit using high-order statistics. Paper presented at the Ninth International Geostatistics Congress, Oslo, Norway

Mirowski PW, Trtzlaff DM, Davies RC, McCormick DS, Williams N, Signer C (2008) Stationary scores on training images for multipoint geostatistics. Math Geosci 41:447–474

Mukul M, Roy D, Satpathy S, Kumar VA (2004) Bootstrapped spatial statistics: a more robust approach to the analysis of finite strain data. J Struct Geol 26:595–600

Mustapha H, Dimitrakopoulos R (2010a) A new approach for geological pattern recognition using high-order spatial cumulants. Comput Geosci 36:313–343

Mustapha H, Dimitrakopoulos R (2010b) High-order stochastic simulation of complex spatially distributed natural phenomena. Math Geosci 42:457–485

Mustapha H, Dimitrakopoulos R (2011) HOSIM: A high-order stochastic simulation algorithm for generating three-dimensional complex geological patterns. Comput Geosci 37:1242–1253

Mustapha H, Dimitrakopoulos R, Chatterjee S (2011) Geologic heterogeneity representation using high-order spatial cumulants for subsurface flow and transport simulations. Water Resour Res 47. doi: 10.1029/2010WR009515

Pilz J, Kazianka H, Spöck G (2012) Some advances in Bayesian spatial prediction and sampling design. Spatial Stat 1:65–81

Remy N, Boucher A, Wu J (2009) Applied geostatistics with SGeMs: a users's guide. Cambridge: Cambridge University Press

Rojas-Avellaneda D, Silvan-Cardenas JL (2006) Performance of geostatistical interpolation methods for modeling sampled data with non-stationary mean. Stoch Environ Res Risk Assess 20:455–467

Saito H, McKenna SA, Zimmerman DA, Coburn TC (2005) Geostatistical interpolation of object counts collected from multiple strip transects: Ordinary kriging versus finite domain kriging. Stoch Environ Res Risk Assess 19:71–85

Scheidt C, Caers J (2009a) Representing spatial uncertainty using distances and kernels. Math Geosci 41:397–419

Scheidt C, Caers J (2009b) Uncertainty quantification in reservoir performance using distances and kernel methods—application to a west africa deepwater turbidite reservoir. SPE J 14:680-692

Scheidt C, Caers J (2010) Bootstrap confidence intervals for reservoir model selection techniques. Comput Geosci 14:369–382

Schelin L, Luna SS (2010) Kriging prediction intervals based on semiparametric bootstrap. Math Geosci 42:985–1000

Strebelle S (2002) Conditional simulation of complex geological structures using multiple point statistics. Math Geosci 34:1–22

Tjelmeland H, Besag J (1998) Markov random fields with higher order interactions. Scand J Stat 25:415–433

Troldborg M, Nowak W, Lange IV, Santos MC, Binning PJ, Bjerg PL (2012) Application of Bayesian geostatistics for evaluation of mass discharge uncertainty at contaminated sites. Water Resour Res 48:W09535

Wu J, Boucher A, Zhang T (2008) SGeMS code for pattern simulation of continuous and categorical variables: FILTERSIM. Comput Geosci 34:1863–1876

Zhang T, Switzer P, Journel AG (2006) Filter-based classification of training image patterns for spatial simulation. Math Geosci 38:63–80