Data Fusion in a Spatial Multivariate Framework: Trading off Hypotheses Against Information

D. Fasbender and P. Bogaert

Abstract Due to the exponential growth in the amount and diversity of data that one may expect to provide greater modeling and predictions opportunities, there is a real need for methods that aim at reconciling them inside a flexible and sound theoretical framework. In a geostatistical prediction context, beside more or less straighforward variations around univariate kriging (e.g. kriging with external drift), the most classical method (i.e. cokriging) is based on a multivariate random field approach of the problem, at the price of strong modeling hypotheses. However, there are expected practical situations where these hypotheses may be hard to fulfill or do not make sense from a modeling viewpoint.

This paper proposes an alternative way of using secondary information for spatial prediction. Based on a data fusion perspective, a general theoretical procedure is proposed. Simple cokriging and Bayesian data fusion are compared both from theoretical and practical viewpoints. Theoretical differences are first emphasized based on the corresponding modeling hypotheses. A case study based on synthetic data subsequently allows to compare both methods from a practical viewpoint. It is shown that, in spite of some simplifying hypotheses required by data fusion, the method is offering quite comparable performances in situations where simple cokriging is expected to be the best possible predictor. Moreover, it offers a much greater flexibility and opens new avenues for incorporating a wide panel of very different and possibly numerous secondary information that, by nature, would not easily fit into a multivariate random field framework, as required by cokriging.

1 Introduction

As the classical (co)kriging predictor relies on the knowledge of first and second-order moments for a given set of random fields (RF's), we will assume here that first and second-order stationarities can be assumed for all variables, and that the

D. Fasbender
Department of Environmental Sciences and Land Use Planning,
Université catholique de Louvain, Belgium
e-mail: fasbender@enge.ucl.ac.be

A. Soares et al. (eds.), *geoENV VI – Geostatistics for Environmental Applications*, 457–466

corresponding functions (i.e. the mean functions and the (cross-)covariance functions) can reasonably be inferred from the data. Without loss of generality and for the sake of simplicity, we will restrict here the discussion to variables with known mean functions so that simple cokriging can be used, though the methodology includes of course the case of non stationary mean and intrinsic stationarity as well (see e.g. Christakos, 1992).

In a first section, a short recall of simple cokriging (SCoK) formulation is presented. The hypotheses involved in this model are emphasized and modeling issues are pointed out with respect to both theoretical and practical viewpoints. Subsequently, the Bayesian Data Fusion (BDF) methodology is presented. For the sake of conciseness and without loss of generality, presentation will be restricted here to the particular case of bivariate Gaussian RFs. Pros and cons of the method are discussed too, and a synthetic case study is presented. The corresponding results indicate that BDF is an interesting alternative in the context of spatial prediction that need to account for additional secondary information which may not fit very well into a multivariate stationary RF framework.

2 Simple Cokriging

In the case of several correlated RFs that are sampled over the same spatial domain, assuming that means are known everywhere, the classical geostatistical method used for conducting multivariate prediction is SCoK (Cressie, 1991). If $\mathcal{Z} = \{Z(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^d, Z(\mathbf{x}) \in \mathbb{R}^1\}$ and $\mathcal{Y} = \{Y(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^d, Y(\mathbf{x}) \in \mathbb{R}^1\}$ are two zero-mean second-order stationary RFs and $\mathbf{Z}_\alpha = Z_1, \ldots, Z_n$ and $\mathbf{Y}_\gamma = Y_{n+1}, \ldots, Y_{n+m}$ are corresponding random vectors sampled from them at locations $\mathbf{x}_i$ $(i = 1, \ldots, m+n)$, the SCoK predictor for Z_0 is then

$$Z_0^p = \begin{pmatrix} \boldsymbol{\sigma}'_\alpha & \boldsymbol{\sigma}'_\gamma \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{\alpha\alpha} & \boldsymbol{\Sigma}_{\alpha\gamma} \\ \boldsymbol{\Sigma}_{\gamma\alpha} & \boldsymbol{\Sigma}_{\gamma\gamma} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{z}_\alpha \\ \mathbf{y}_\gamma \end{pmatrix} \tag{1}$$

where the $\boldsymbol{\sigma}$s and $\boldsymbol{\Sigma}$s are obtained from the partitioning of the covariance matrix $\boldsymbol{\Sigma}$ for the whole vector $(Z_0, \mathbf{Z}'_\alpha, \mathbf{Y}'_\gamma)$, with

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_0^2 & \boldsymbol{\sigma}'_\alpha & \boldsymbol{\sigma}'_\gamma \\ \boldsymbol{\sigma}_\alpha & \boldsymbol{\Sigma}_{\alpha\alpha} & \boldsymbol{\Sigma}_{\alpha\gamma} \\ \boldsymbol{\sigma}_\gamma & \boldsymbol{\Sigma}_{\gamma\alpha} & \boldsymbol{\Sigma}_{\gamma\gamma} \end{pmatrix} \tag{2}$$

whereas the corresponding variance of prediction is given by

$$\sigma_0^p = \sigma_0^2 - \begin{pmatrix} \boldsymbol{\sigma}'_\alpha & \boldsymbol{\sigma}'_\gamma \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{\alpha\alpha} & \boldsymbol{\Sigma}_{\alpha\gamma} \\ \boldsymbol{\Sigma}_{\gamma\alpha} & \boldsymbol{\Sigma}_{\gamma\gamma} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\sigma}_\alpha \\ \boldsymbol{\sigma}_\gamma \end{pmatrix} \tag{3}$$

Clearly, SCoK is the best linear predictor, but not necessarily the best predictor. However, assuming multivariate gaussianity for $(Z_0, \mathbf{Z}'_\alpha, \mathbf{Y}'_\gamma)$, the best predictor is linear, and SCoK then corresponds to the result of a linear regression, with

$$Z_0^p = \mathbb{E}\left[Z_0|\mathbf{z}_\alpha, \mathbf{y}_\gamma\right] \quad ; \quad \sigma_0^p = \mathbb{V}ar\left[Z_0|\mathbf{z}_\alpha, \mathbf{y}_\gamma\right] \tag{4}$$

and where $Z_0|\mathbf{z}_\alpha, \mathbf{y}_\gamma \sim N(Z_0^p, \sigma_0^p)$. In other words, SCoK can only be considered as the best predictor when full gaussianity holds. If this is not the case, SCoK is still a valuable predictor, but its prediction variance is not necessarily the smallest possible one and the true conditional probability distribution function (pdf) for $Z_0|\mathbf{z}_\alpha, \mathbf{y}_\gamma$ is not Gaussian in general.

Though SCoK is a straighforward and well-known method that can be easily numerically implemented without difficulties, there are some practical issues that need to be addressed. Clearly, obtaining the covariance matrix $\boldsymbol{\Sigma}$ for arbitrary locations rely on a multivariate modeling of the covariance functions, in order to ensure positive-definiteness of the final results. The most frequently used method is based on the so-called Linear Model of Coreginalization (see e.g. Chilès and Delfiner, 1999 or Goovaerts, 1997), which for a bivariate case is written as

$$\begin{pmatrix} C_{\alpha\alpha}(\mathbf{h}) & C_{\alpha\beta}(\mathbf{h}) \\ C_{\alpha\beta}(\mathbf{h}) & C_{\beta\beta}(\mathbf{h}) \end{pmatrix} = \sum_i \begin{pmatrix} a_{i,\alpha\alpha} & a_{i,\alpha\beta} \\ a_{i,\alpha\beta} & a_{i,\beta\beta} \end{pmatrix} c_i(\mathbf{h}) = \sum_i \mathbf{A}_i c_i(\mathbf{h}) \tag{5}$$

where all matrices $\mathbf{A}_i$ are positive definite and where the same elementary positive definite covariance models $c_i(\mathbf{h})$ must be used for modeling all (cross-)covariance functions. Clearly, this imposes an important (and rarely discussed) conceptual constraint on the method, as any secondary variable that need to be accounted for when predicting Z_0 must fit into the second-order stationary RF paradigm, i.e. the random vector $\mathbf{Y}_\gamma$ must be considered as a sample from a second-order stationary RF that can be characterized by a covariance function $C_{\beta\beta}(\mathbf{x}_i - \mathbf{x}_j)$ that only depends on $\mathbf{x}_i - \mathbf{x}_j$ but neither on $\mathbf{x}_i$ nor on $\mathbf{x}_j$. Unfortunately, it happens frequently that quite useful information does not fit well into this paradigm. As a simple example, in an environmental pollution context, the distance $\mathbf{x} - \mathbf{x}_j$ to a chemical industry located at $\mathbf{x}_j$ is likely to be a quite relevant measure for quantifying atmospheric deposition $Y(\mathbf{x})$, but $Y(\mathbf{x})$ cannot be considered as coming from a RF whose covariance function would be translation invariant, i.e. depending only on $\mathbf{h}$, of course. This in turn may seriously impair the prediction of a related variable of interest $Z(\mathbf{x})$ (e.g. soil pollution) using a cokriging approach, as corresponding covariance function estimates $\widehat{C}_{\beta\beta}(\mathbf{h})$ and $\widehat{C}_{\alpha\beta}(\mathbf{h})$ are meaningless.

3 Bayesian Data Fusion

By the light of the previous example, it appears that relying on a multivariate second-order stationary RF framework is quite arguable in some instances. In order to account for this issue, a new theoretical framework has recently been proposed. Its aim is to alleviate the need of second-order stationarity for secondary variables at the price of mild simplifying hypothesis. Due to space limitations, only main and most important results will be presented here. Additional details can be found and are disscussed at length in Bogaert and Fasbender (2007).

Let us assume that $\mathcal{Z}$ is a zero-mean second-order stationary RF of primary interest, where $\mathbf{Z}$ is random vector sampled from it. Let us define a mapping $g : \mathbb{R} \to \mathbb{R}$ such that $\mathbf{g}(\mathbf{Z}) = \{g(Z_i)\}$ is a new random vector, along with an arbitrary zero-mean random vector $\mathbf{E}$ of same size and independent from $\mathbf{Z}$. The basic assumption of BDF is to assume that, for any secondary variable, we have $\mathbf{Y} = \mathbf{g}(\mathbf{Z}) + \mathbf{E}$. Clearly, $\mathbf{Z}$ and $\mathbf{Y}$ are collocated random variables but $\mathbf{Y}$ is not longer second-order stationary in general, as its properties also depend on the arbitrary $\mathbf{E}$. According to the independence assumption $\mathbf{E} \perp \mathbf{Z}$, it is also clear that the conditional pdf for $\mathbf{Z}|\mathbf{y}$ is given by

$$f_{\mathbf{z}|\mathbf{y}}(\mathbf{z}|\mathbf{y}) \propto f_{\mathbf{z}} f_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z}) = f_{\mathbf{z}}(\mathbf{z}) f_{\mathbf{e}}(\mathbf{y} - \mathbf{g}(\mathbf{z})) \tag{6}$$

where $f_{\mathbf{z}}$ is the *a priori* pdf for $\mathbf{Z}$ and $f_{\mathbf{e}}$ is the pdf for $\mathbf{E}$. A more intuitive interpretation of Eq. (6) is to consider that each Y_i can be viewed as an indirect measurements of the true Z_i, as Y_i is a functional $g(Z_i)$ up to an additive error E_i, where it is reasonable in general to assume these errors as independent from the true $\mathbf{Z}$.

Starting from this last very general relation, a straightforward formulation can be proposed in a spatial prediction context. Let us consider that what is sought for is the conditional pdf for Z_0 given a set of observed values $\mathbf{z}_\alpha = z_1, \ldots, z_n$ for the RF of interest $\mathcal{Z}$ along with a set of observed values $\mathbf{y}_\gamma = y_{n+1}, \ldots, y_{n+m}$ for the auxiliary RF $\mathcal{Y}$. Defining additionally the vector $\mathbf{Z}_\beta = Z_{n+1}, \ldots, Z_{n+m}$ of unobserved variables at the same location as $\mathbf{Y}_\gamma$, the conditional pdf for $Z_0|\mathbf{z}_\alpha, \mathbf{y}_\gamma$ is then given by

$$f_{z_0|\mathbf{z}_\alpha,\mathbf{y}_\gamma}(z_0|\mathbf{z}_\alpha, \mathbf{y}_\gamma) \propto \int_{\mathbf{R}^m} f_{z_0,\mathbf{z}_\alpha,\mathbf{z}_\beta}(z_0, \mathbf{z}_\alpha, \mathbf{z}_\beta) f_{\mathbf{e}}(\mathbf{y}_\gamma - \mathbf{g}(\mathbf{z}_\beta)) d\mathbf{z}_\beta \tag{7}$$

This is a very general and nonlinear formula for prediction that requires the knowledge of the joint pdf $f_{z_0,\mathbf{z}_\alpha,\mathbf{z}_\beta}$ as well as the joint pdf of errors $f_{\mathbf{e}}$. A classical approach would be to consider $\mathcal{Z}$ as a Gaussian RF along with a mutual independence hypothesis for the vector $\mathbf{E}$, so that $f_{z_0,\mathbf{z}_\alpha,\mathbf{z}_\beta}$ is multivariate Gaussian and $f_{\mathbf{e}} = \prod_i f_{e_i}$. Using again Baye's theorem, we then have $f_{e_i}(y_i - g(z_i)) \propto f_{z_i|y_i}(z_i|y_i)/f_{z_i}(z_i)$, so that Eq. (7) simplifies to

$$f_{z_0|\mathbf{z}_\alpha,\mathbf{y}_\gamma}(z_0|\mathbf{z}_\alpha, \mathbf{y}_\gamma) \propto \int_{\mathbf{R}^m} f_{z_0,\mathbf{z}_\alpha,\mathbf{z}_\beta}(z_0, \mathbf{z}_\alpha, \mathbf{z}_\beta) \prod_i \frac{f_{z_i|y_i}(z_i|y_i)}{f_{z_i}(z_i)} d\mathbf{z}_\beta \tag{8}$$

As a consequence, Eq. (8) is still a nonlinear expression that requires multivariate integration, but it is now possible to express $f_{z_0,\mathbf{z}_\alpha,\mathbf{z}_\beta}$ from the covariance function $C_{\alpha\alpha}(\mathbf{h})$ as well as to infer the pdf's $f_{z_i|y_i}$ from the data set. A synthetical illustration of this can be found in Bogaert and Fasbender (2007). The real advantage of this formulation is that it is not longer needed to have any stationarity hypothesis about $\mathcal{Y}$ compared to SCoK. Moreover, it can be shown that a similar reasonsing can be used in order to account for multiple secondary information without any difficulties.

4 Comparing Data Fusion and Cokriging

Though BDF and SCoK may appear at the first sight as completely different (and thus difficult to compare) approaches for accounting for secondary information, it can however be shown that close analytical linear relations for expressing the conditional mean and variances can be obtained for BDF if an additional multivariate Gaussian hypothesis is assumed to hold.

It has been reminded that, for jointly Gaussian RF's $\mathcal{Y}$ and $\mathcal{Z}$, the corresponding conditional pdf $f_{z_0|\mathbf{z}_\alpha,\mathbf{y}_\gamma}$ is Gaussian with mean and variance given by Eq. (4), so that SCoK is the best predictor. On the other side, for BDF, one can also remark that assuming that $\mathcal{Y}$ and $\mathcal{Z}$ are jointly zero-mean Gaussian RF's leads to a zero-mean Gaussian vector $(Z_0, \mathbf{Z}_\alpha, \mathbf{Z}_\beta, \mathbf{Y}_\gamma)$ with covariance matrix $\boldsymbol{\Sigma}$ as given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_0^2 & \boldsymbol{\sigma}'_\alpha & \boldsymbol{\sigma}'_\beta & \boldsymbol{\sigma}'_\gamma \\ \boldsymbol{\sigma}_\alpha & \boldsymbol{\Sigma}_{\alpha\alpha} & \boldsymbol{\Sigma}_{\alpha\beta} & \boldsymbol{\Sigma}_{\alpha\gamma} \\ \boldsymbol{\sigma}_\beta & \boldsymbol{\Sigma}_{\beta\alpha} & \boldsymbol{\Sigma}_{\beta\beta} & \boldsymbol{\Sigma}_{\beta\gamma} \\ \boldsymbol{\sigma}_\gamma & \boldsymbol{\Sigma}_{\gamma\alpha} & \boldsymbol{\Sigma}_{\gamma\beta} & \boldsymbol{\Sigma}_{\gamma\gamma} \end{pmatrix} \tag{9}$$

Several simplifications will then occur. First, the functional g is now a linear mapping with $\mathbf{g}(\mathbf{z}_\beta) = \frac{\sigma_{yz}}{\sigma_0^2}\mathbf{z}_\beta$, where $\sigma_{yz} = \sigma_{zy} = Cov(Z(\mathbf{x}), Y(\mathbf{x}))$, $\forall \mathbf{x} \in \mathbb{R}^d$. Second, $\mathbf{E}$ is now Gaussian with covariance matrix equal to $\sigma^2_{\gamma|\beta}\mathbf{I}$ where $\sigma^2_{\gamma|\beta}$ is equal to $\sigma_y^2 - \frac{\sigma^2_{yz}}{\sigma_0^2}$ with $\sigma_y^2 = \mathbb{V}ar[Y(\mathbf{x})]$, $\forall \mathbf{x} \in \mathbb{R}^d$. Third, since $(Z'_0, \mathbf{Z}'_\alpha, \mathbf{Z}'_\beta)'$ and $\mathbf{E}$ are Gaussian vectors, the product of pdf's in Eq. (8) is proportional to a Gaussian pdf with mean vector $\mathbf{M}$ and covariance matrix $\mathbf{S}$ given by

$$\begin{cases} \mathbf{S}^{-1} = \begin{pmatrix} \sigma_0^2 & \boldsymbol{\sigma}'_\alpha & \boldsymbol{\sigma}'_\beta \\ \boldsymbol{\sigma}_\alpha & \boldsymbol{\Sigma}_{\alpha\alpha} & \boldsymbol{\Sigma}_{\alpha\beta} \\ \boldsymbol{\sigma}_\beta & \boldsymbol{\Sigma}_{\beta\alpha} & \boldsymbol{\Sigma}_{\beta\beta} \end{pmatrix}^{-1} + \begin{pmatrix} 0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \\ \mathbf{0} & \mathbf{0} & \frac{1}{\sigma^2_{\gamma|\beta}}\frac{\sigma^2_{yz}}{\sigma_0^4}\mathbf{I} \end{pmatrix} \\ \mathbf{M} = \frac{1}{\sigma^2_{\gamma|\beta}}\mathbf{S}\begin{pmatrix} 0 \\ \mathbf{0} \\ \frac{\sigma_{yz}}{\sigma_0^2}\mathbf{y}_\gamma \end{pmatrix} \end{cases} \tag{10}$$

Finally, since the integrand of Eq. (8) is proportional to a Gaussian pdf, integrating over $\mathbf{z}_\beta$ and conditioning on $\mathbf{z}_\alpha$ leads to the conclusion that the conditional pdf $Z_0|\mathbf{z}_\alpha, \mathbf{y}_\gamma$ is univariate Gaussian, thus completely characterized by its mean and variance, with

$$\begin{cases} \mathbb{E}[Z_0|\mathbf{z}_\alpha, \mathbf{y}_\gamma] = \frac{1}{\sigma^2_{\gamma|\beta}}\frac{\sigma_{yz}}{\sigma_0^2}\mathbf{s}'_\beta\mathbf{y}_\gamma + \mathbf{s}'_\alpha\mathbf{S}^{-1}_{\alpha\alpha}\left(\mathbf{z}_\alpha - \frac{1}{\sigma^2_{\gamma|\beta}}\frac{\sigma_{yz}}{\sigma_0^2}\mathbf{S}_{\alpha\beta}\mathbf{y}_\gamma\right) \\ \mathbb{V}ar[Z_0|\mathbf{z}_\alpha, \mathbf{y}_\gamma] = s_0 - \mathbf{s}'_\alpha\mathbf{S}^{-1}_{\alpha\alpha}\mathbf{s}_\alpha \end{cases} \tag{11}$$

where

$$\mathbf{S} = \begin{pmatrix} s_0 & \mathbf{s}'_\alpha & \mathbf{s}'_\beta \\ \mathbf{s}_\alpha & \mathbf{S}_{\alpha\alpha} & \mathbf{S}_{\alpha\beta} \\ \mathbf{s}_\beta & \mathbf{S}_{\beta\alpha} & \mathbf{S}_{\beta\beta} \end{pmatrix} \tag{12}$$

The gain of BDF on SCok in terms of inference is non negligible. Indeed, instead of inferring the multivariate covariance model (with LMC namely), we only need to estimate three simple quantities: i) $C_{\alpha\alpha}(\mathbf{h})$ the covariance function of $\mathcal{Z}$, ii) σ_y^2 the variance of $Y(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^d$ and iii) σ_{yz} the pointwise covariance of $Z(\mathbf{x})$ and $Y(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^d$.

Comparing Eqs. (1) and (3) with Eq. (11) is not straightforward, but it is worth noting that (i) BDF now provides expressions for the conditional expectation and variance that are linear with respect to the observed values $\mathbf{z}_\alpha$ and $\mathbf{y}_\gamma$ as for SCoK, and (ii) results for BDF and SCoK can be compared too on a common basis, as we fulfill the optimal conditions for using SCoK as the best possible predictor. It is worth remembering however that, in general, BDF and SCoK will obey distinct optimality properties: SCoK is optimal under the hypothesis of jointly Gaussian RF's $\mathcal{Y}$ and $\mathcal{Z}$, whereas BDF only rely on a Gaussian hypothesis for $\mathcal{Z}$ and an independence hypothesis for $\mathbf{E}$, which is a somewhat milder hypothesis than assuming joint Gaussianity for both RF's $\mathcal{Y}$ and $\mathcal{Z}$.

5 A Synthetic Case Study

In order to illustrate the similitudes of the results that are obtained using both approaches in a situation where SCoK is expected to give the best possible results, a synthetic case study is presented. The aim here is to show that, even under optimal conditions for SCoK, using SCoK instead of BDF does not significantly increase the quality of predictions. Stated in other words, the loss of information due to the use of BDF instead of SCoK does not dramatically affect the quality of the predictions.

Let assume a smooth zero-mean unit-variance Gaussian RF $\mathcal{Z}$ for which a realization over a 100×100 regular grid is given in Fig. 1a (covariance function is exponential with sill equal to 1 and range equal to 30). Let assume also a second Gaussian RF $\mathcal{Y}$ (Fig. 1b) defined as a linear combination of the RF $\mathcal{Z}$ and another zero-mean unit-variance Gaussian RF $\mathcal{E}$ with the same covariance function. By taking $\mathcal{Y} = 2\mathcal{Z} + \mathcal{E}$, the resulting RF $\mathcal{Y}$ is thus a zero-mean Gaussian RF with variance equal to 5, and both RF's are jointly Gaussian with pointwise correlation equal to 0.894. Under these conditions, SCoK is thus the best possible predictor. For conducting predictions, two random samples $\mathbf{z}$ and $\mathbf{y}$ are extracted from these simulated grids. In order to be in a situation when SCoK would be interesting compared to simple kriging (i.e. the auxiliary $\mathbf{y}$ conveys valuable extra information compared to what is already known from $\mathbf{z}$), samples size have been chosen equal to 200 and 400 for $\mathbf{z}$ and $\mathbf{y}$, respectively. Predictions of $\mathcal{Z}$ is then conducted at the nodes of the grid using $\mathbf{z}$ and $\mathbf{y}$ as observed values. It is worth noting too that sampling has been conducted so that there are no locations for which $\mathcal{Z}$ and $\mathcal{Y}$ are jointly observed.

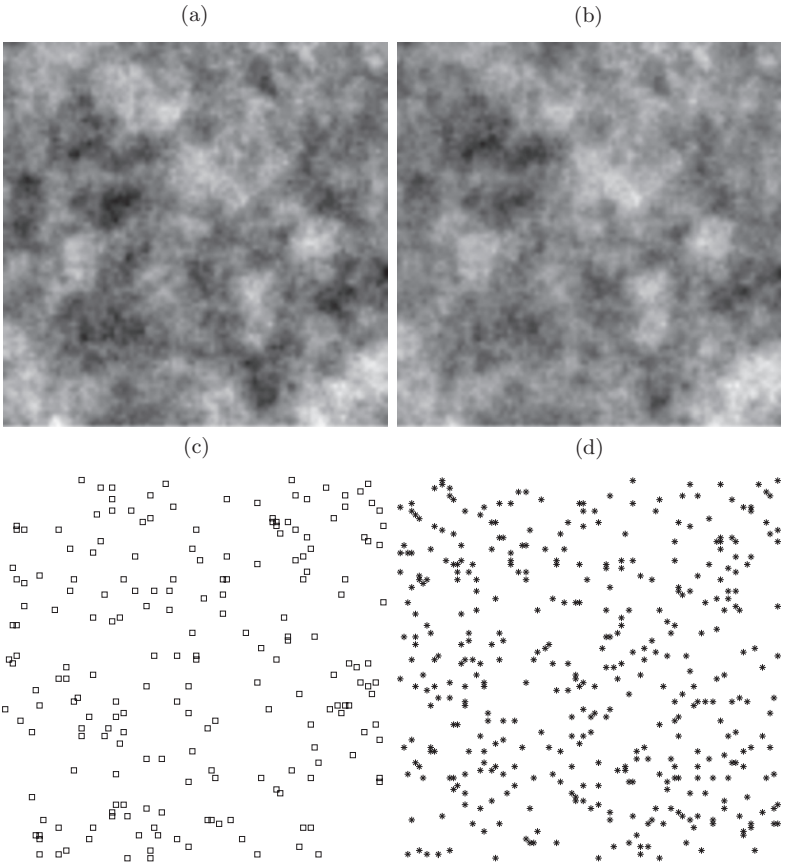

Fig. 1 Simulations over a 100×100 regular grid for RF $\mathcal{Z}$ **(a)** and RF $\mathcal{Y}$ **(b)**, along with the random sample $\mathbf{z}$ at 200 locations **(c)** and the random sample $\mathbf{y}$ at 400 locations **(d)**

In order to compare relative differences between BDF and SCoK results compared to simple kriging results, a relative comparison approach has been used here. Clearly, using simple kriging with the n closest observations of $\mathbf{z}$ (that will be denoted as SK_n) provides a lower bound in terms of prediction quality, as it only makes use of the primary variable. On the other side, using simple kriging with the n closest observations of $\mathbf{z}$ along with the $\mathbf{z}$ observations at the n closest locations for the observed $\mathbf{y}$ (that will be denoted as SK_{2n}) provides an upper bound, as it assumes that the true values for the primary variable are available at locations where only the auxiliary variable is observed. As a consequence, in terms of quality predictions, results for BDF and SCOK will be located somewhere in between these two bounds. This also provides a way to compare the results for BDF and SCoK on a relative scale ranging from SK_n to SK_{2n}.

Figure 2 shows the predictions results obtained using the four previously described methods, namely SCoK (Fig. 2d), BDF (Fig. 2c) and the two extreme simple kriging situations (Figs. 2a and 2b). One can notice that BDF and SCoK provide visually very similar results. This observation is confirmed from the Root Mean Squared Error (RMSE) as computed between simulated and predicted values (see Table 1). As expected, SCoK and BDF gives intermediate results in between the two kriging predictions, with only a slight advantage for SCoK when values are compared on a relative scale.

The above computations can been repeated by keeping everything identical except for the pointwise correlation between the two RF's $\mathcal{Z}$ and $\mathcal{Y}$. It can be seen from Fig. 3 that the relative difference between BDF and SCoK is null for high and

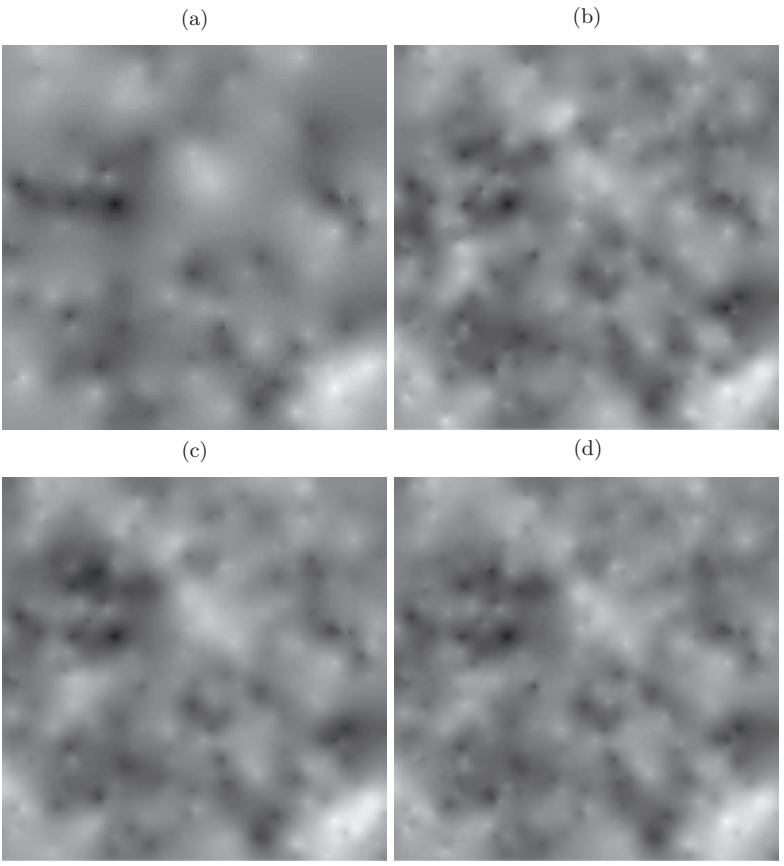

Fig. 2 Results for the different predictions methods. **(a)** is simple kriging with 10 closest points, **(b)** is simple Kriging with 2×10 closest points, **(c)** is BDF with 10 closest points for each RF and **(d)** is SCoK with 10 closest points for each RF

Table 1 Quality assessment for the various prediction methods. RMSEs are computed as the root mean squared differences between simulated and predicted values. Relative RMSEs are RMSEs rescaled between 0 and 1 according to the bounds as provided by SK_n and SK_{2n} (value 1 is thus the best possible result whereas value 0 is the worst possible one)

-	SK_n	SCoK	Bayesian Data Fusion	SK_{2n}
RMSE	0.63	0.53	0.55	0.48
Relative RMSE	0	0.68	0.58	1

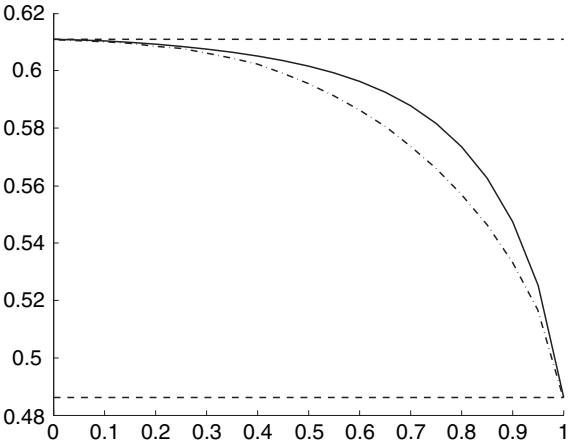

Fig. 3 Evolution of the RMSE with respect to the pointwise correlation between $\mathcal{Z}$ and $\mathcal{Y}$ RF's. Dashed lines correspond to SK_n (higher value) and SK_{2n} (lower value), whereas plain line and dotted line correspond to BDF and SCoK, respectively

low correlations (i.e. when secondary information is useless and when secondary information is equivalent to primary one, respectively), so that biggest differences will be observed for intermediate values.

6 Conclusions

In this paper, a BDF approach has been proposed as an interesting alternative way to account for various secondary information in a spatial prediction context. Indeed, the method does not rely on a classical multivariate second-order stationary RF framework, as it is required for cokriging. Because by nature BDF relies on a different set of assumptions, it is difficult to compare both methods from a general viewpoint in situations where both of them would be relevant. In order to overcome this difficulty, a comparison has been conducted in a situation where SCoK is known to be the best possible predictor, so that the loss of information caused by using BDF instead of SCoK can be assessed on an objective way. Results show that in terms of performances, differences are however quite limited.

Of course, there is no point in using data fusion instead of cokriging when one can reasonably assume that a second-order multivariate RF hypothesis holds, as SCoK is then by definition the right method to be used. However, the application field of BDF is quite more general, as it allows the user to account for secondary information that would not fit into this framework, permitting thus to deal with a much wider panel of situations that could not be meaningfully handled by SCoK. Hence, BDF can be viewed as a robust alternative to cokriging for multivariate prediction where cokriging hypotheses are known to be irrelevant or at least quite arguable.

Finally, it is worth noting that the aim of this paper was not to suggest that sound multivariate spatial prediction methods like cokriging, which have frequently proved to be quite useful, should be discarded or criticized when used under the appropriate hypotheses. It is rather the limited practical pertinency of these hypotheses which is at stake here. It is suggested that in situations where a multivariate modelling does not appear to be conceptually consistent with what is known from data, BDF is then a more reasonable choice by avoiding the need of using a multivariate model (e.g., as the LMC) at the price of mild simplifying hypotheses. Though rather simple in the case of a single auxiliary variable, this modelling problem is expected to become critical in situations where the number and diversity of auxiliary informations sources that need to be accounted for is increasing. This is a typical case where BDF is expected to be a much more flexible way of handling the problem than SCoK.

References

Bogaert P, Fasbender D (2007) Bayesian data fusion in a spatial prediction context: A general formulation. Stochastic Environmental Research and Risk Assessment. Published online

Chilès J.-P, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. John Wiley and Sons, Inc., New York, NY

Christakos G (1992) Random field models in earth sciences. Academic Press, San Diego, CA

Cressie N (1991) Statistics for spatial data. Wiley series in probability and mathematical statistics. John Wiley and Sons, Inc., New York, NY

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