Stoch Environ Res Risk Assess (2007) 21:695–709
DOI 10.1007/s00477-006-0080-3

ORIGINAL PAPER

Bayesian data fusion in a spatial prediction context: a general formulation

P. Bogaert · D. Fasbender

Published online: 8 November 2006

Abstract In spite of the exponential growth in the amount of data that one may expect to provide greater modeling and predictions opportunities, the number and diversity of sources over which this information is fragmented is growing at an even faster rate. As a consequence, there is real need for methods that aim at reconciling them inside an epistemically sound theoretical framework. In a statistical spatial prediction framework, classical methods are based on a multivariate approach of the problem, at the price of strong modeling hypotheses. Though new avenues have been recently opened by focusing on the integration of uncertain data sources, to the best of our knowledges there have been no systematic attemps to explicitly account for information redundancy through a data fusion procedure. Starting from the simple concept of measurement errors, this paper proposes an approach for integrating multiple information processing as a part of the prediction process itself through a Bayesian approach. A general formulation is first proposed for deriving the prediction distribution of a continuous variable of interest at unsampled locations using on more or less uncertain (soft) information at neighboring locations. The case of multiple information is then considered, with a Bayesian solution to the problem of fusing multiple information that are provided as separate conditional probability distributions. Well-known methods and results are derived as limit cases. The convenient hypothesis of conditional independence is discussed by the light of information theory and maximum entropy principle, and a methodology is suggested for the optimal selection of the most informative subset of information, if needed. Based on a synthetic case study, an application of the methodology is presented and discussed.

Keywords Data merging · Measurement errors · Soft information · Kriging · Entropy

1 Introduction

As emphasized by van der Putten et al. (2002), in spite of the exponential growth in the amount of data that one may naively expect to provide greater opportunities for improving modeling and predictions, reality is somewhat different. In many real world applications, the number and diversity of sources, over which this information is fragmented, grows at an even faster rate, thus making these information more and more difficult to be used jointly. As a consequence, there is a growing need for methods that aim at reconciling them inside a unique and sound theoretical framework.

Data fusion is a generic expression widely found in the literature, that conveys the idea of combining at best information coming from different sources in order to achieve improved accuracy and better inferences. In principle, fusion of multiple sources of information provides significant advantages over single source data, as improved estimate of the physical phenomena should be obtained via redundant information. Although the general idea is rather simple, the way this goal is precisely achieved and its objectives

P. Bogaert (✉) · D. Fasbender
UCL/AGR0/MILA/ENGE, Université Catholique de Louvain, Croix du Sud 2/16,
1348 Louvain-la-Neuve, Belgium
e-mail: bogaert@enge.ucl.ac.be

D. Fasbender
e-mail: fasbender@enge.ucl.ac.be

are however diverse as well as widely dependent on the field of application.

In data mining applications for marketing purposes, data fusion techniques typically aim at achieving a complete data file from different sources/databases which do not contain the same units (Cho et al. 2003; Rässler 2004), a method also called statistical matching in this context. In the field of applied computer sciences that involves multiple data sources such as sensor readings or model decision, data fusion is a main concern due to the widely different nature of the outputs that are generated. Among others, this involves for example biometrics verification systems (e.g., Duc et al. 1997; Ross and Jain 2003), surveillance systems (e.g., Jones et al. 2003), robotics (e.g., Cremer et al. 2001; Pradalier et al. 2003), medical imagery (e.g., Song et al. 2003) or military/civil engineering (e.g., Gros et al., 1999; Sohn and Lee 2003). Data fusion has also important applications in classification of remote sensing images (e.g., Costantini et al. 1997; Melgani and Serpico 2002; Simone et al. 2002) and in environmental modeling.

Besides this diversity of objectives and applications, the way data fusion is precisely achieved covered a wide spectrum of methods and topics as diverse as neural networks, *k*-means clustering, fuzzy logic, Kalman filtering, Dempster-Shafer theory of evidence, or traditional Bayesian theory (just to quote few of them) that may involve information fusion at different levels (data-level, feature-level or decision-level); see Sohn and Lee (2003) or Ross and Jain (2003) for a description of these concepts. However, as emphasized by Fassinut-Mombot and Choquel (2004), the most classical techniques of information fusion are based on probability theory associated with Bayesian decision theory. In a spatiotemporal study of wind speed over ocean surface, Wikle et al. (2001) show how, for example, combining in a Bayesian procedure two different set of measurements (namely satellite data and output from models based on synoptic measurements) that are indirect measurements of the same underlying true (but unknown) values may lead to serious improvements for prediction.

Although Bayesian methods—or variations around them—have been widely used for fusing collocated information (i.e., coregistered information in the remote sensing terminology), there have been little attempts to integrate multiple redundant information through a data fusion process in a spatial prediction framework, where what is typically sought for is (a distribution of) predicted values at unsampled locations, to be obtained from the knowledge of the same or different sampled variables at a set of spatially close locations. Standard methods rather aim at using these variables into a classical multivariate framework, at the price of strong hypotheses for modeling their joint dependence (see, e.g., Cressie 1993, p. 141; Goovaerts, 1997, p. 113; Wackernagel 1995, p. 152; Chilès and Delfiner 1999, p. 339 for detailed description and discussion about these methods) that may even lead to the use of dimensionality reduction techniques when number of variables is too high.

New avenues have been recently opened by Christakos (2000) and Christakos et al. (2002), who proposed the so-called set of Bayesian maximum entropy (BME) methods that aim at rigorously accounting for data uncertainty in the prediction process. A comparison of the BME methods with non Bayesian approaches can also be found in Christakos (2002). Although these methods have proved to be successful in a wide range of applications (see, e.g., Bogaert and D'Or 2003; D'Or and Bogaert 2004; Savelievaa et al. 2005), data fusion as a way of accounting for information redundancy is not a part of the process.

Using the classical concept of measurement errors embedded into a Bayesian framework, this paper is an attempt to show how an efficient general formulation of the spatial prediction problem can easily be achieved. The first part of the paper mainly focus on a presentation of general theoretical results. Starting from an additive error model, it is first shown how uncertainty can be accounted for at the price of mild independence hypotheses. Connections with well-known methods and results that are obtained as singular cases are emphasized. With the help of information theory, the rational behind some of these independence hypotheses is explained, and a procedure for selecting the most useful subset of information to be fused is suggested too. Using afterward the concept of dual notation for information, simpler equations involving only prior and conditional distributions are obtained. In the second part of the paper, the concept of Bayesian data fusion is presented. It is shown how multiple collocated information can be merged into a single conditional distribution. Specific results for the Gaussian case are presented, as well as the possibility of defining an informativeness index, measuring the amount of information brought by any conditional distribution. Finally, the third part of the paper aims at presenting a simplified illustration of some of these concepts based on a synthetic case study.

2 A prediction model with uncertain information

Assume that $\mathbf{Z} = (Z_0,...,Z_n)'$ is a random vector sampled from a spatial random field (RF) $\mathfrak{Z} = \{Z(\mathbf{x}) : \mathbf{x} \subset D \subseteq \mathbb{R}^d, Z(\mathbf{x}) \in \mathbb{R}\}$ and that we know the joint multivariate probability distribution function (pdf) $f(\mathbf{z})$, where each Z_i is associated with a different spatial location $\mathbf{x}_i$ in our context. We will consider that it is not possible to directly observe z_i values for these variables, but that instead observed y_i values are available, where the Y_i's are functionally related to the Z_i's up to random errors E_i's. The Y_i's may possibly (but not necessarily) correspond partially or totally to different physical variables, so that the random vector $\mathbf{Y} = (Y_0,...,Y_n)'$ should be merely regarded as a collection of random variables rather than as resulting from the sampling of a same spatial RF $\mathfrak{Y}$ at locations $\mathbf{x}_0,...,\mathbf{x}_n$. It will be considered here that both Z_i and E_i (and consequently Y_i) are of the continuous type, though all results that will be presented can be extended to discrete or mixed random variables without problem.

2.1 The case of simple additive errors

Let us consider first hereafter the simple case where $Y_i = Z_i + E_i$, so that in vector notations $\mathbf{Y} = \mathbf{Z} + \mathbf{E}$ (the general case of an arbitrary relationship $\mathbf{Y}=\mathbf{g}(\mathbf{Z}) + \mathbf{E}$ will be presented later on). Given the fact that values $\mathbf{y} = (y_0,...,y_n)'$ are observed, what is sought for is the conditional pdf $f(\mathbf{z}|\mathbf{y})$, where

$$f(\mathbf{z}|\mathbf{y}) = f(z_0,\ldots,z_n|z_0+e_0,\ldots,z_n+e_n)$$

For deriving this pdf, let us make the hypothesis that $\mathbf{E}\perp\mathbf{Z}$, i.e. $\mathbf{E}$ can be considered as stochastically independent from $\mathbf{Z}$. Using Bayes theorem,

$$f(\mathbf{z}|\mathbf{y}) = \frac{f(\mathbf{y}|\mathbf{z})f(\mathbf{z})}{\int\limits_{\mathbb{R}^n} f(\mathbf{y}|\mathbf{z})f(\mathbf{z})\mathrm{d}\mathbf{z}} = \frac{1}{A}f(\mathbf{y}|\mathbf{z})f(\mathbf{z}) \tag{1}$$

where A plays the role of a normalization constant. Because we assumed that $\mathbf{Y} = \mathbf{Z} + \mathbf{E}$ with $\mathbf{E}\perp\mathbf{Z}$, it is easy to see that

$$F(\mathbf{y}|\mathbf{z}) = P(\mathbf{Z}+\mathbf{E}<\mathbf{y}|\mathbf{z}) = P(\mathbf{E}<\mathbf{y}-\mathbf{z}) = F_{\mathbf{E}}(\mathbf{y}-\mathbf{z})$$

so that we have

$$f(\mathbf{y}|\mathbf{z}) = \frac{\partial}{\partial \mathbf{y}}F(\mathbf{y}|\mathbf{z}) = f_{\mathbf{E}}(\mathbf{y}-\mathbf{z}) \tag{2}$$

It is worth noting that Eq. 2 is directly linked to the classical convolution theorem for obtaining the joint distribution of a sum of random variables (see e.g. Papoulis 1991, p. 136), extended here in the multivariate case, so that

$$f(\mathbf{y}) = \int\limits_{\mathbb{R}}\cdots\int\limits_{\mathbb{R}} f(\mathbf{y}|\mathbf{z})f(\mathbf{z})\mathrm{d}\mathbf{z} = \int\limits_{\mathbb{R}}\cdots\int\limits_{\mathbb{R}} f_{\mathbf{E}}(\mathbf{y}-\mathbf{z})f(\mathbf{z})\mathrm{d}\mathbf{z}$$

However, in our case we are interested in $f(\mathbf{z}|\mathbf{y})$ instead, so plugging Eq. 2 into Eq. 1 gives

$$f(\mathbf{z}|\mathbf{y}) = \frac{1}{A}f_{\mathbf{E}}(\mathbf{y}-\mathbf{z})f(\mathbf{z}) \tag{3}$$

As prediction is generally sought for the single z_0, we will focus on the marginal conditional pdf at location $\mathbf{x}_0$, that can be obtained by integrating Eq. 3 over other variables, so that

$$f(z_0|\mathbf{y}) = \frac{1}{A}\int\limits_{\mathbb{R}}\cdots\int\limits_{\mathbb{R}} f_{\mathbf{E}}(\mathbf{y}-\mathbf{z})f(\mathbf{z})\mathrm{d}z_1\cdots\mathrm{d}z_n \tag{4}$$

Further simplifications can be obtained by assuming the mutual independence $E_0\perp\cdots\perp E_n$ for the errors, so that

$$f_{\mathbf{E}}(\mathbf{y}-\mathbf{z}) = \prod_{i=0}^{n} f_{E_i}(y_i-z_i) \iff f(\mathbf{y}|\mathbf{z}) = \prod_{i=0}^{n} f(y_i|\mathbf{z}) \tag{5}$$

showing that $(Y_0\perp\cdots\perp Y_n)|\mathbf{z}$, and we finally obtain

$$f(z_0|\mathbf{y}) = \frac{1}{A}f_{E_0}(y_0-z_0)\int\limits_{\mathbb{R}}\cdots\int\limits_{\mathbb{R}} f(\mathbf{z})\prod_{i=1}^{n} f_{E_i}(y_i-z_i)\mathrm{d}z_1\cdots\mathrm{d}z_n \tag{6}$$

In summary, Eq. 4 allows to derive the conditional pdf of Z_0 at an arbitrary location $\mathbf{x}_0$ if we have observed values for variables $\mathbf{Y}$ that are linearly related to $\mathbf{Z}$ up to random errors $\mathbf{E}$ assuming that these errors are independent from $\mathbf{Z}$, where Eq. 6 assumes additionally that these errors are mutually independent.

Though the previous developments have been conducted out of the context of classical methods in spatial statistics or geostatistics, it is easy to show that they are intimately related to well-known approaches. A particular interpretation of Eq. 6 is to consider that each Y_i variable can be viewed as an indirect measurements of the true Z_i through, e.g., a same imperfect measuring device, so that each measurement Y_i includes an additive measurement error, leading to the relation $Y_i = Z_i + E_i$, where it is reasonable in general to assume that errors are mutually independent and do not depend on the true value. This is similar, e.g., to the hypothesis by e.g. Wikle et al. (2001) and Pradalier et al. (2003), who assume that, conditionally to the true values, multiple measurements of the same underlying unknown quantity can be conveniently viewed as independent from each others, thus leading to simple expressions involving the product of the corresponding various pdf's. Assuming now additionally that these errors are $N(0, \sigma_i^2)$ and that $\mathbf{Z} \sim N(\boldsymbol{\mu}, \Sigma)$, it can be shown that the conditional mean $\mathbb{E}[Z_0|\mathbf{y}]$ and the conditional variance $\mathbb{V}ar[Z_0|\mathbf{y}]$ are identical to the kriging predictor and variance as obtained from the so-called simple kriging with measurement errors (Cressie 1993; see Appendices for the proof).

2.2 Generalization to arbitrary functionals

Equation 6 can be slightly generalized if one consider that the observable Y_i are linked to the Z_i's through various arbitrary functionals (see Fig. 1), so that $Y_i = g_i(Z_i) + E_i$ (we will restrict here the developments to the case were each Y_i is only functionally dependent of a single corresponding Z_i, though a similar approach can be applied when Y_i possibly depends on several variables in $\mathbf{Z}$). We thus have $\mathbf{Y} = \mathbf{g}(\mathbf{Z}) + \mathbf{E}$, where $\mathbf{g}(\mathbf{z}) = (g_0(z_0),...,g_n(z_n))'$, and what is sought for is

$$f(\mathbf{z}|\mathbf{y}) = f(z_0, \ldots, z_n | g_0(z_0) + e_0, \ldots, g_n(z_n) + e_n)$$

Reasoning along the same lines as before with $\mathbf{Y} = \mathbf{g}(\mathbf{Z}) + \mathbf{E}$ and $\mathbf{E} \perp \mathbf{Z}$,

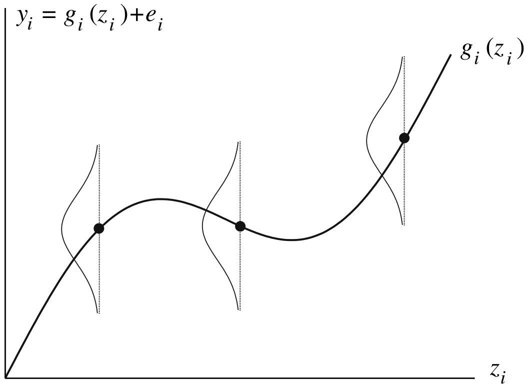

Fig. 1 Arbitrary functional $y_i = g_i(z_i) + e_i$

$$\begin{aligned} F(\mathbf{y}|\mathbf{z}) &= P(\mathbf{g}(\mathbf{Z}) + \mathbf{E} < \mathbf{y}|\mathbf{z}) = P(\mathbf{E} < \mathbf{y} - \mathbf{g}(\mathbf{z})) \\ &= F_{\mathbf{E}}(\mathbf{y} - \mathbf{g}(\mathbf{z})) \end{aligned} \tag{7}$$

so that we obtain similarly

$$f(\mathbf{y}|\mathbf{z}) = \frac{\partial}{\partial \mathbf{y}} F(\mathbf{y}|\mathbf{z}) = f_{\mathbf{E}}(\mathbf{y} - \mathbf{g}(\mathbf{z})) \tag{8}$$

that can be plugged into Eq. 1. If the mutual independence hypothesis for the errors is assumed, this simplifies again to

$$f_{\mathbf{E}}(\mathbf{y} - \mathbf{g}(\mathbf{z})) = \prod_{i=0}^{n} f_{E_i}(y_i - g_i(z_i)) \iff f(\mathbf{y}|\mathbf{z}) = \prod_{i=0}^{n} f(y_i|\mathbf{z})$$

and we finally obtain the equivalent of Eq. 6, with

$$\begin{aligned} f(z_0|\mathbf{y}) = &\frac{1}{A} f_{E_0}(y_0 - g_0(z_0)) \\ &\times \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(\mathbf{z}) \prod_{i=1}^{n} f_{E_i}(y_i - g_i(z_i)) \mathrm{d}z_1 \cdots \mathrm{d}z_n \end{aligned} \tag{9}$$

As an illustration, Fig. 2 shows the graphs of the functions $f_{E_i}(y_i - g_i(z_i))$ with respect to z_i, computed for different observed values y_i and for the arbitrary functional $g_i(\cdot)$ given in Fig. 1. One can remark that, because of the possibly non linear and non monotonic relationship between Y_i and Z_i, these functions may exhibit complex shapes.

It is worth noting that the basic assumption for the model above is that the random vector $\mathbf{E}$ is considered as independent from the vector of interest $\mathbf{Z}$, or stated in other words the pdf of the errors $\mathbf{E}$ is the same whatever the choice for the conditioning $\mathbf{z}$ values. Though this hypothesis proved to be convenient for deriving the previous results, it can be limiting as, e.g., in the case of heteroskedasticity where the variance of errors depends on the $\mathbf{z}$ values. One can remark however that arbitrary monotonic transformations can be applied on the Y_i's without affecting the generality of the results, as traditionally done in regression analysis.

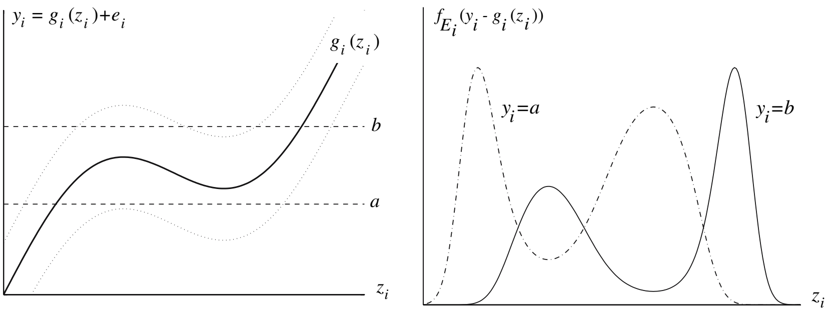

Fig. 2 *Left part* shows the graph for the arbitrary functional $y_i = g_i(z_i) + e_i$ as in Fig. 1, along with the bounds of the ± 2σ probability interval for E_i assumed to be $N(0,\sigma^2)$. *Right part* shows the expression for $f_E(y_i - g_i(z_i))$ as a function of z_i when $y_i = a$ and $y_i = b$, with same scale for horizontal axis as on *left part*

2.3 Mixing soft and not so soft (hard) information

One can consider various situations with respect to the comparative amount of error attached with each observable Y_i. In some situations, the errors for a subset of these variables can be so high that they do not convey much information about the corresponding Z_i's, whereas in other situations these errors are so small that they can be neglected in the computations. We will consider both cases hereafter.

For the first case, let us write $\mathbf{Y} = (\mathbf{Y}_a^{'},\mathbf{Y}_b^{'})'$ where $\mathbf{Y}_a = (Y_0,...,Y_m)'$ and $\mathbf{Y}_b = (Y_{m+1},...,Y_n)'$, where we assume that $\mathbf{E}_a \perp \mathbf{E}_b$. If the errors $\mathbf{E}_a$ tend to be very high compared to $\mathbf{g}(\mathbf{Z}_a)$, we can make the approximation $\mathbf{Y}_a = \mathbf{g}(\mathbf{Z}_a) + \mathbf{E}_a \simeq \mathbf{E}_a$ so that using Eq. 7 leads to

$$F(\mathbf{y}_a|\mathbf{z}_a) \simeq P(\mathbf{E}_a < \mathbf{y}_a|\mathbf{z}_a) = P(\mathbf{E}_a < \mathbf{y}_a) = F_{\mathbf{E}_a}(\mathbf{y}_a)$$

and we obtain $f(\mathbf{y}_a|\mathbf{z}_a) \simeq f_{\mathbf{E}a}(\mathbf{y}_a)$. Plugging this result into Eq. 1 and using the fact that $(\mathbf{Y}_a \perp \mathbf{Y}_b)|\mathbf{z}$ because $\mathbf{E}_a \perp \mathbf{E}_b$, we obtain

$$\begin{aligned} f(\mathbf{z}|\mathbf{y}_a,\mathbf{y}_b) &= \frac{1}{A} f(\mathbf{y}_a,\mathbf{y}_b|\mathbf{z}) f(\mathbf{z}) \\ &\simeq \frac{1}{A} f_{\mathbf{E}_a}(\mathbf{y}_a) f_{\mathbf{E}_b}(\mathbf{y}_b - \mathbf{g}(\mathbf{z}_b)) f(\mathbf{z}) \end{aligned}$$

Integrating now this result over $z_1,...,z_n$, we get

$$\begin{aligned} f(z_0|\mathbf{y}_a,\mathbf{y}_b) &\simeq \frac{1}{A} f_{\mathbf{E}_a}(\mathbf{y}_a) \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{\mathbf{E}_b}(\mathbf{y}_b - \mathbf{g}(\mathbf{z}_b)) \\ &\quad \times f(\mathbf{z}) \mathrm{d}z_1 \cdots \mathrm{d}z_n = \frac{1}{B} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{\mathbf{E}_b}(\mathbf{y}_b - \mathbf{g}(\mathbf{z}_b)) \\ &\quad \times f(\mathbf{z}_b) \mathrm{d}\mathbf{z}_b = f(z_0|\mathbf{y}_b) \end{aligned} \tag{10}$$

where $B = A/f_{\mathbf{E}a}(\mathbf{y}_a)$ is the new normalization constant. As seen from Eq. 10, if the error $\mathbf{E}_a$ is assumed to be very high compared to $\mathbf{g}(\mathbf{Z}_a)$, we can assume that $\mathbf{Y}_a$ does not bring any significant information on Z_0 and can be discarded for the prediction, as $f(z_0|\mathbf{y}_a,\mathbf{y}_b) \simeq f(z_0|\mathbf{y}_b)$.

For the second case, assume that the errors $\mathbf{E}_b$ tend to be very small compared to $\mathbf{g}(\mathbf{Z}_b)$, so that we can make the approximation $\mathbf{Y}_b = \mathbf{g}(\mathbf{Z}_b) + \mathbf{E}_b \simeq \mathbf{g}(\mathbf{Z}_b)$. Using Eq. 7 again leads to

$$\begin{aligned} F(\mathbf{y}_b|\mathbf{z}_b) &\simeq P(\mathbf{g}(\mathbf{Z}_b) < \mathbf{y}_b|\mathbf{z}_b) \\ &= P(\mathbf{y}_b - \mathbf{g}(\mathbf{z}_b) > 0) = \begin{cases} 0 & \text{if } \mathbf{y}_b < \mathbf{g}(\mathbf{z}_b) \\ 1 & \text{if } \mathbf{y}_b \geq \mathbf{g}(\mathbf{z}_b) \end{cases} \end{aligned}$$

so that $F(\mathbf{y}_b|\mathbf{z}_b)$ is the Heaviside step function, whose derivative is $f(\mathbf{y}_b|\mathbf{z}_b) = \delta(\mathbf{y}_b - \mathbf{g}(\mathbf{z}_b))$, the Dirac delta function. Plugging this result into Eq. 1 and knowing that $(\mathbf{Y}_a \perp \mathbf{Y}_b)|\mathbf{z}$, we obtain

$$f(\mathbf{z}|\mathbf{y}_a,\mathbf{y}_b) \simeq \frac{1}{A} f_{\mathbf{E}_a}(\mathbf{y}_a - \mathbf{g}(\mathbf{z}_a)) \delta(\mathbf{y}_b - \mathbf{g}(\mathbf{z}_b)) f(\mathbf{z})$$

and integrating now this result over $z_1,...,z_n$, we get

$$\begin{aligned} f(z_0|\mathbf{y}_a,\mathbf{y}_b) &\simeq \frac{1}{A} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{\mathbf{E}_a}(\mathbf{y}_a - \mathbf{g}(\mathbf{z}_a)) \\ &\quad \times \left[\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \delta(\mathbf{y}_b - \mathbf{g}(\mathbf{z}_b)) f(\mathbf{z}) \mathrm{d}\mathbf{z}_b \right] \mathrm{d}\mathbf{z}_a \end{aligned} \tag{11}$$

If the system of equations $\mathbf{y}_b - \mathbf{g}(\mathbf{z}_b) = 0$ can be solved with respect to $\mathbf{z}_b$ so that $\mathbf{z}_b = \mathbf{g}^{-1}(\mathbf{y}_b)$ is the unique solution (this will happen, e.g., if for each Y_i we have $Y_i = g_i(Z_i)$ where the $g_i(\cdot)$'s are monotonic so that $Z_i = g_i^{-1}(Y_i)\ \forall\ i = m+1,...,n$), we then have from the properties of the Dirac delta function

$$\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \delta(\mathbf{y}_b - \mathbf{g}(\mathbf{z}_b)) f(\mathbf{z}) \mathrm{d}\mathbf{z}_b = f(\mathbf{z}_a,\mathbf{z}_b) \quad \text{with}$$
$$\mathbf{z}_b = \mathbf{g}^{-1}(\mathbf{y}_b) \text{ known}$$

Plugging this result into Eq. 11 and writing $f(\mathbf{z}_a,\mathbf{z}_b) = f(\mathbf{z}_a|\mathbf{z}_b)f(\mathbf{z}_b)$,

$$\begin{aligned} f(z_0|\mathbf{y}_a,\mathbf{y}_b) &\simeq \frac{1}{B} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{\mathbf{E}_a}(\mathbf{y}_a - \mathbf{g}(\mathbf{z}_a)) f(\mathbf{z}_a|\mathbf{z}_b) \mathrm{d}\mathbf{z}_a \\ &= f(z_0|\mathbf{y}_a,\mathbf{z}_b) \end{aligned} \tag{12}$$

where $B = A/f(\mathbf{z}_b)$. If the system of equations $\mathbf{y}_b - \mathbf{g}(\mathbf{z}_b) = 0$ admits several solutions $\mathbf{z}_{b,1},...,\mathbf{z}_{b,k}$, the event $\mathbf{Y}_b = \mathbf{y}_b$ is equivalent to $\mathbf{Z}_b = (\mathbf{Z}_b = \mathbf{z}_{b,1}) \cup \cdots \cup (\mathbf{Z}_b = \mathbf{z}_{b,k})$ where the elementary events $\mathbf{Z}_b = \mathbf{z}_{b,j}$ $(j = 1,...,k)$ are mutually exclusive, so that

$$\begin{aligned} f(z_0|\mathbf{y}_a,\mathbf{y}_b) &= f(z_0|\mathbf{y}_a,(\mathbf{Z}_b = \mathbf{z}_{b,1}) \cup \cdots \cup (\mathbf{Z}_b = \mathbf{z}_{b,k})) \\ &= \frac{1}{\sum_{j=1}^{k} f(\mathbf{z}_{b,j})} \sum_{j=1}^{k} f(z_0|\mathbf{y}_a,\mathbf{z}_{b,j}))f(\mathbf{z}_{b,j}) \end{aligned}$$

From Eq. 12, one can see that if we assume the errors $\mathbf{E}_b$ to be negligible, this amounts to incorporating the knowledge brought by $\mathbf{Y}_a$ directly from the conditional pdf $f(\mathbf{z}_a|\mathbf{z}_b)$, with $f(\mathbf{z}_a|\mathbf{z}_b) = f(\mathbf{z}_a,\mathbf{z}_b)/f(\mathbf{z}_b)$, where the $\mathbf{z}_b$ values are computed from the observed $\mathbf{y}_b$ through the inverse relationship $\mathbf{z}_b = \mathbf{g}^{-1}(\mathbf{y}_b)$. If several solutions $\mathbf{z}_{b,j}$ exist, the procedure is repeated for each $\mathbf{z}_{b,j}$ and the results are weighted by the corresponding $f(\mathbf{z}_{b,j})$ values. It is worth noting that, by extension, Eq. 12 includes of course the more classical case where, for at least a subset of locations, values $\mathbf{z}_b$ are directly observed for the variable of interest. Finally, though both situations have been presented separately, they can of course be combined in an arbitrary way, and further simplifications are obtained by assuming the mutual independence for the errors.

2.4 Conditional independence and entropy

From Eqs. 5 and 9, one can see that simple analytical results are obtained by assuming $E_0 \perp \cdots \perp E_n$, this corresponding to the very convenient conditional independence hypothesis $(Y_0 \perp \cdots \perp Y_n)|\mathbf{z}$, as it alleviates the need of inference for the joint pdf of errors in Eqs. 2 and 8. Thus, when only a limited subset of Y_i is at hand, estimating this joint pdf may prove to be difficult. Clearly, substituting the joint pdf of errors by the product of the corresponding marginal pdf's will induce a loss of potentially valuable information. However, in the framework of information theory, one can prove that in absence of clear joint information, the best choice is precisely to assume this conditional independence. Indeed, from the maximum entropy principle, what is sought for is the joint pdf $(\mathbf{Y},\mathbf{Z})$ for which the corresponding entropy $H(\mathbf{Y},\mathbf{Z})$ is maximum, with

$$H(\mathbf{Y},\mathbf{Z}) = -\int\limits_{\mathbb{R}^n}\int\limits_{\mathbb{R}^n} f(\mathbf{y},\mathbf{z})\ln f(\mathbf{y},\mathbf{z})\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{z}$$

We also have the relation $H(\mathbf{Y},\mathbf{Z}) = H(\mathbf{Z}) + H(\mathbf{Y},\mathbf{Z})$ where the marginal entropy $H(\mathbf{Z})$ is known as we have specified $f(\mathbf{z})$, with

$$H(\mathbf{Z}) = -\int\limits_{\mathbb{R}^n} f(\mathbf{z})\ln f(\mathbf{z})\mathrm{d}\mathbf{z}$$

Since $H(\mathbf{Z})$ is known, the maximum of $H(\mathbf{Y},\mathbf{Z})$ is reached when the conditional entropy $H(\mathbf{Y}|\mathbf{Z})$ is maximum, where

$$H(\mathbf{Y}|\mathbf{Z}) = -\int\limits_{\mathbb{R}^n}\int\limits_{\mathbb{R}^n} f(\mathbf{y},\mathbf{z})\ln f(\mathbf{y}|\mathbf{z})\mathrm{d}\mathbf{y}\,\mathrm{d}\mathbf{z}$$

We also know that $H(\mathbf{Y}|\mathbf{Z}) \leq H(Y_0|\mathbf{Z}) + \ldots + H(Y_n|\mathbf{Z})$ (see e.g., Papoulis 1991), where the marginal conditional entropies $H(Y_i|\mathbf{Z})$ are given by

$$H(Y_i|\mathbf{Z}) = -\int\limits_{\mathbb{R}^n}\int\limits_{\mathbb{R}} f(y_i|\mathbf{z})\ln f(y_i|\mathbf{z})\mathrm{d}y_i\,\mathrm{d}\mathbf{z} \quad \forall i = 1,\ldots,n \tag{13}$$

If each Y_i depends on a unique Z_i so that $Y_i = g_i(Z_i) + e_i$, we can write that

$$f(y_i|\mathbf{z}) = f(y_i|z_i) = f_{E_i}(y_i - g_i(z_i)) \tag{14}$$

or stated in other words, knowing the complete set of realized values $\{z_0,\ldots,z_n\}$ does not bring any additional information on Y_i if we already know the single realized value z_i. As the error pdf's $f_{E_i}(e_i)$ are specified, one can evaluate $f_{E_i}(y_i - g_i(z_i))$ for any (y_i,z_i) so that plugging Eq. 14 into Eq. 13 shows that the various $H(Y_i|\mathbf{Z})$ are known too, with

$$H(Y_i|\mathbf{Z}) = H(Y_i|Z_i) = -\int\limits_{\mathbb{R}}\int\limits_{\mathbb{R}} f_{E_i}(y_i - g_i(z_i)) \times \ln f_{E_i}(y_i - g_i(z_i))\mathrm{d}y_i\,\mathrm{d}z_i$$

As we finally know that

$$H(\mathbf{Y}|\mathbf{Z}) = H(Y_0|\mathbf{Z}) + \cdots + H(Y_n|\mathbf{Z}) \iff (Y_0 \perp \cdots \perp Y_n)|\mathbf{z}$$

assuming the conditional independence for the Y_i's with respect to $\mathbf{z}$ is thus equivalent to maximizing $H(\mathbf{Y},\mathbf{Z})$ when the pdf's $f(\mathbf{z})$ and $f_{E_i}(e_i)$ as well as the relationships $Y_i = g_i(Z_i) + E_i$ are specified.

It is also interesting to remark that when one only knows the expectation vectors $\boldsymbol{\mu}_{\mathbf{E}}$, $\boldsymbol{\mu}_{\mathbf{Z}}$ and the covariance matrices $\Sigma_{\mathbf{E}}$, $\Sigma_{\mathbf{Z}}$, the maximum for $H(\mathbf{Y},\mathbf{Z})$ is reached when $\mathbf{Z} \sim N(\boldsymbol{\mu}_{\mathbf{Z}}, \Sigma_{\mathbf{Z}})$ and $\mathbf{E} \sim N(\boldsymbol{\mu}_{\mathbf{E}}, \Sigma_{\mathbf{E}})$ with $\mathbf{E} \perp \mathbf{Z}$, thus leading again to the conditional independence $(Y_0 \perp \cdots \perp Y_n)|\mathbf{Z}$. If only the variances $\sigma^2_{E_i}$ for the errors are known instead of $\Sigma_{\mathbf{E}}$, then the maximum entropy is obtained with $E_i \sim N(\mu_{E_i}, \sigma^2_{E_i})$ and $E_0 \perp \cdots \perp E_n$ (see Appendices for these proofs). In the case of linear relationships $Y_i = Z_i + E_i$ the conditional pdf $f(z_0|\mathbf{y})$ will always be Gaussian, with conditional mean and variance as given by simple kriging with measurement errors.

2.5 Dual notation for information

Up to now, all notations have been given in terms of the error pdf $f_{E_i}(\cdot)$ along with the functionals $g_i(\cdot)$. Both for practical and theoretical reasons, it is sometimes more convenient to reformulate these error pdf's as functions of conditional pdf's on the Z_i's. Starting from Eq. 14 and using Bayes theorem again, we have

$$f_{E_i}(y_i - g_i(z_i)) = f(y_i|z_i) = \frac{f(z_i|y_i)}{f(z_i)} f(y_i) \tag{15}$$

Using Eq. 15 allows to express each $f_{E_i}(y_i - g_i(z_i))$ in Eq. 9 in terms of a conditional pdf on z_i, so that

$$f(z_0|\mathbf{y}) = \frac{1}{B}\frac{f(z_0|y_0)}{f(z_0)}\int\limits_{\mathbb{R}}\cdots\int\limits_{\mathbb{R}} f(\mathbf{z})\prod_{i=1}^{n}\frac{f(z_i|y_i)}{f(z_i)}\mathrm{d}z_1\cdots\mathrm{d}z_n \tag{16}$$

where $B = A/\prod_{i=0}^{n} f(y_i)$ is the new normalization constant, as all $f(y_i)$'s can be assimilated to constants in Eq. 16 due to the conditioning on the set $\{y_0,...,y_n\}$. Equation 16 is an alternative but equivalent way of incorporating the uncertainty that has the advantage of being more intuitive. Information conveyed by any Y_i is directly expressed through a conditional pdf for the corresponding Z_i. In this sense, Eq. 16 can be viewed as a complete recoding of the information brought by all the Y_i's, these information being now directly expressed through their influence on the Z_i variables.

From Eq. 16, it is also worth noting that the influence of a given y_i value on the final result is directly linked to how much $f(z_i|y_i)/f(z_i)$ is different from one as z_i varies. Stated in other words, the influence of knowing y_i can be measured by the discrepancies between the conditional pdf $f(z_i|y_i)$ and the marginal pdf $f(z_i)$. If a specific y_i is non informative with respect to Z_i, i.e. when $f(z_i|y_i)/f(z_i) = 1$ for some z_i, the corresponding factor will be filtered out from Eq. 16 (note that, for a fixed y_i value, the equality $f(z_i|y_i) = f(z_i)$ does not necessarily implies $Y_i \perp Z_i$, as there could be specific values of y_i for which this is true, whereas the equality is not verified in general). As a consequence, combining any non informative y_i values together with other informative values y_j ($i \neq j$) will not cause any harm, as irrelevant information will be automatically filtered out during the prediction process. By the light of this last remark, it is thus useful to quantify the amount of information brought by any Y_i variable through some kind of informativeness index. This will be discussed in detail in Sect. 3.4.

3 Data fusion

Up to this point, the use of Eq. 9 is restricted to situations where a single measured value is available at each one of the $\mathbf{x}_0,...,\mathbf{x}_n$ locations . However, situations where multiple information are made available occurs commonly and need to be accounted for. This encompass, e.g., the cases of (1) repeated measurements at location $\mathbf{x}_i$ of the same physical variable related to Z_i (with the same or with different measuring devices), (2) different physical variables jointly measured at location $\mathbf{x}_i$, all of them being related in a different way to Z_i. As using more relevant information is expected to lead to more accurate predictions, it is thus of some concern to derive rules for fusing these information. Considerable efforts and literature have been devoted to this issue in spatial statistics. Most traditional approaches rely on a multivariate framework, where a joint distribution for the whole set of variables has to be inferred first and conditioned on observed values afterwards. However, due to the possibly highly different nature and properties of these variables (e.g., categorical vs. continuous-valued variables) and potentially high number of variables, this can prove to be a very complex or even impossible task. We propose hereafter a different approach, that relies first on a full recoding of the information expressed as a unique conditional distribution on the variable of interest at every point where information is initially available, so that Eq. 16 can be used afterwards for combining them in order to get posterior distributions at arbitrary locations. A simple fusion rule is suggested and is discussed by the light of the Naive Bayes fusion (NBF) rule.

3.1 Multiple collocated information

Let us assume without loss of generality that, for the last location $\mathbf{x}_n$, a set of m realized values $\{y_{n,1},...,y_{n,m}\}$ are available, all of them being related to the same z_n through the relations

$$Y_{n,j} = g_j(Z_n) + E_{n,j} \quad \forall j = 1,\ldots,m \tag{17}$$

Let us denote $\mathbf{Y}_n = (Y_{n,1},...,Y_{n,m})'$ and $\mathbf{y}_a = (y_0,...,y_{n-1})'$ (where the a subscript will always refer to the first $n-1$ elements of the corresponding vector), and let us write more simply Eq. 17 as $\mathbf{Y}_n = \mathbf{g}(Z_n) + \mathbf{E}_n$. What we are interested in is the conditional pdf $f(\mathbf{z}|\mathbf{y}_a,\mathbf{y}_n)$. Using Bayes' theorem,

$$f(\mathbf{z}|\mathbf{y}_a,\mathbf{y}_n) = \frac{f(\mathbf{y}_a,\mathbf{y}_n|\mathbf{z})f(\mathbf{z})}{\int_{\mathbb{R}^n} f(\mathbf{y}_a,\mathbf{y}_n|\mathbf{z})f(\mathbf{z})\mathrm{d}\mathbf{z}} = \frac{1}{A}f(\mathbf{y}_a,\mathbf{y}_n|\mathbf{z})f(\mathbf{z})$$

We will assume as before that $\mathbf{E}_a \perp \mathbf{Z}$ and $\mathbf{E}_n \perp \mathbf{Z}$, so that again

$$f(\mathbf{y}_a|\mathbf{z}) = f_{\mathbf{E}_a}(\mathbf{y}_a - \mathbf{g}(\mathbf{z}_a)); \quad f(\mathbf{y}_n|\mathbf{z}) = f_{\mathbf{E}_n}(\mathbf{y}_n - \mathbf{g}(z_n))$$

Assuming moreover that $\mathbf{E}_a \perp \mathbf{E}_n$, we have the result

$$f(\mathbf{y}_a,\mathbf{y}_n|\mathbf{z}) = f_{\mathbf{E}_a}(\mathbf{y}_a - \mathbf{g}(\mathbf{z}_a))f_{\mathbf{E}_n}(\mathbf{y}_n - \mathbf{g}(z_n))$$

so that the formula that combines all the information for prediction at location $\mathbf{x}_0$ becomes

$$f(z_0|\mathbf{y}_a,\mathbf{y}_n) = \frac{1}{A}\int\limits_{\mathbb{R}}\cdots\int\limits_{\mathbb{R}} f(\mathbf{z})f_{\mathbf{E}_a}(\mathbf{y}_a-\mathbf{g}(\mathbf{z}_a))f_{\mathbf{E}_n}(\mathbf{y}_n-\mathbf{g}(z_n))\mathrm{d}z_1\cdots\mathrm{d}z_n$$

Of course, this procedure can be easily generalized for any other set of locations where repeated measurements are made available. Further simplifications can be obtained by assuming the mutual independence for $\mathbf{E}_a$ and $\mathbf{E}_n$, so that along with the dual notation for information, where

$$f_{E_{n,j}}(y_{n,j}-g_j(z_n)) \propto \frac{f(z_n|y_{n,j})}{f(z_n)} \tag{18}$$

we get similarly to Eq. 16 the simpler result

$$\begin{aligned} f(z_0|\mathbf{y}) &\propto \frac{f(z_0|y_0)}{f(z_0)}\int\limits_{\mathbb{R}}\cdots\int\limits_{\mathbb{R}} f(\mathbf{z})f(z_n)^{-m}\prod_{i=1}^{n-1}\frac{f(z_i|y_i)}{f(z_i)} \\ &\quad\times\prod_{j=1}^{m}f(z_n|y_{n,j})\mathrm{d}z_1\cdots\mathrm{d}z_n \\ &\propto \frac{f(z_0|y_0)}{f(z_0)}\int\limits_{\mathbb{R}}\cdots\int\limits_{\mathbb{R}} f(\mathbf{z})\phi(z_n|\mathbf{y}_n)\prod_{i=1}^{n-1}\frac{f(z_i|y_i)}{f(z_i)}\mathrm{d}z_1\cdots\mathrm{d}z_n \end{aligned} \tag{19}$$

where $\phi(z_n|\mathbf{y}_n)$ denotes the posterior pdf resulting from the fusion operation for collocated information, i.e.

$$\phi(z_n|\mathbf{y}_n) \propto f(z_n)^{-m}\prod_{j=1}^{m}f(z_n|y_{n,j})$$

Through this operation, the information brought separately by each observed $y_{n,j}$ with respect to a same Z_n are thus fused into a single pdf $\phi(z_n|\mathbf{y}_n)$.

As a last remark, it is worth noting that in various spatial applications like, e.g., remote sensing, image analysis or cartography, a common situation that occurs is the need for fusing collocated information at the prediction location itself, discarding any other possible information at other locations (so that the spatial prediction part of the problem does not appear). If we define $\mathbf{Y}_0 = (Y_{0,1},\ldots,Y_{0,m})'$ with $Y_{0,j} = Z_0 + E_{0,j}\ \forall\ j = 1,\ldots,m$ and reasoning along the same lines as above by assuming mutual independence for the $E_{0,j}$'s, it is then easy to see that the posterior pdf simply becomes

$$f(z_0|\mathbf{y}_0) \propto \frac{1}{(f(z_0))^{m-1}}\prod_{j=1}^{m}f(z_0|y_{0,j}) \propto \phi(z_0|\mathbf{y}_0) \tag{20}$$

which corresponds to the so-called naive Bayes' (or Idiot's Bayes's) fusion rule of the individual posterior predictors as given by the $f(z_0|y_{0,j})$'s.

Criticisms about the use of a NBF rule can be raised on the ground that it implicitly relies on a conditional independence hypothesis of the variables. However, in a classification context, Kuncheva (2004) emphasized that NBF is experimentally observed to be surprisingly accurate and efficient, where the surprise comes from the fact that independence assumption is seldom true. Indeed, naive Bayes can often outperform more sophisticated classification methods (Altinçay 2005; Lewis 1998). To our opinion, by the light of the previously discussed relation between conditional independence and entropy maximization, this can be at least partially explained by the fact that NBF is precisely the choice that maximizes entropy if no additional information is incorporated about the joint dependence of these variables, so it will yield the posterior pdf which is the most likely to be observed (see Jaynes 2003, for an enlightening discussion about this principle). Note however that if a joint pdf $f(z_n,\mathbf{y}_n)$ can be reasonably well inferred from data at hand (thus alleviating the need of a conditional independence hypothesis), the corresponding conditional pdf $f(z_n|\mathbf{y}_n)$ can be used into Eq. 19 without any problem, instead of relying on the simpler NBF rule-based pdf $\phi(z_n|\mathbf{y}_n)$.

3.2 Weighting information for confidence

Let us assume in Eq. 19 that among the $f(z_n|y_{n,j})$'s, there is a subset $J\subset\{1,\ldots,m\}$ of them that are identical, so that we can write $f(z_n|y_{n,j}) = f(z_n|y_{n,J})\ \forall\ j\in J$, i.e., the same information on Z_n is brought by each $y_{n,j}$ when $j \in J$. We have now

$$\begin{aligned} \phi(z_n|\mathbf{y}_n) &\propto f(z_n)^{-m}f(z_n|y_{n,J})^{m_J}\prod_{j\notin J}f(z_n|y_{n,j}) \\ &= f(z_n)^{-m}f(z_n|y_{n,J})^{w_J m}\prod_{j\notin J}f(z_n|y_{n,j})^{w_j m} \end{aligned}$$

where m_J is the cardinality of J, with weights $w_J = m_J/m$ and $w_j = 1/m\ \forall j \notin J$ so that their sum is equal to 1. This is equivalent to a relative weighting of the pdf's. By generalization, it also suggests that pdf's can be fused by accounting for the confidence one may have in their respective accuracy (e.g., when they are inferred from data of varying quality or quantity, or when they come from experts that do not get the same credit), so that we could write

$$\phi(z_n|\mathbf{y}_n) \propto f(z_n)^{-m} \prod_{j=1}^{m} f(z_n|y_{n,j})^{w_j m} \quad \text{with} \sum_j w_j = 1 \tag{21}$$

Equation 21 can be interpreted easily by remarking that if $w_j > 1/m$ (or $w_j < 1/m$), the corresponding pdf will count for more (or for less) than a single pdf among the set of m fused ones, as $w_j m > 1$ (or as $w_j m < 1$), whereas Eq. 19 is the case where the same credit is given to each pdf.

3.3 The Gaussian case

For obvious inferential reasons, a quite common hypothesis is to assume that all $f(z_n|y_{n,j})$'s are the pdf's of $N(\mu_j,\sigma_j^2)$ variables, as inference for these pdf's only requires the estimation of the conditional expectations $\mathbb{E}[Z_n|y_{n,j}] = \mu_j$ and conditional variances $\mathbb{V}ar[Z_n|y_{n,j}] = \sigma_j^2$ with $j = 1,...,m$. It is then easy to prove that

$$\prod_{j=1}^{m} f(z_n|y_{n,j}) \propto \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{z_n-\mu}{\sigma}\right)^2\right)$$

so that the product of the $f(z_n|y_{n,j})$'s is proportional to the density of a Gaussian distribution with expectation μ and variance σ^2 that are given by

$$\mu = \sum_{j=1}^{m} \frac{1/\sigma_j^2}{\sum_{k=1}^{m}(1/\sigma_k^2)}\mu_j \qquad \sigma^2 = 1\Big/\sum_{j=1}^{m}\frac{1}{\sigma_j^2}$$

(see e.g. Papoulis 1991, p. 187 for obtaining this result based on a least squares criterion, and Duc et al. 1997, for the use of this result in a somewhat different context). It is worth noting that μ thus corresponds to a weighting of the μ_j's, where the weights are proportional to the inverse of the corresponding variances. This is consistent with the intuition, as the information brought by any particular $f(z_n|y_{n,j})$ decreases as its corresponding variance σ_j^2 increases. Moreover, one can also see that, as we have $\sigma_{m+1}^2 \geq 0\ \forall\ m = 1,2,...$, we have the inequality

$$\sum_{j=1}^{m+1}\sigma_j^2 \geq \sum_{j=1}^{m}\sigma_j^2 \iff 1\Big/\sum_{j=1}^{m+1}\frac{1}{\sigma_j^2} \leq 1\Big/\sum_{j=1}^{m}\frac{1}{\sigma_j^2}$$

thus showing that σ^2 can only decrease as the number of $f(z_n|y_{n,j})$'s increases. It is worth noting that the final fused pdf is not Gaussian in general, as it also depends on the a priori $f(z_n)$ in Eq. 19, and the influence of any specific $f(z_n|y_{n,j})$ on the final result will still depend on its distance from this *a priori* pdf. As an illustration, Fig. 3 shows the result of the fusion for three Gaussian pdf's. The left graph presents the Gaussian pdf's $f(z_n|y_{n,j})$ $(j = 1,2,3)$ to be fused as well as the corresponding product $\prod_{j=1}^{3} f(z_n|y_{n,j})$(up to the multiplicative normalization constant). The right graph shows the fusion pdf $\phi(z_n|\mathbf{y}_n)$ for two different *a priori* pdf's. One can easily see the additional effect of this pdf on the final result.

3.4 Informativeness index

As seen from Eq. 18, the amount of information brought by a specific $y_{n,j}$ can be quantified by a measuring of how much the ratio $f(z_n|y_{n,j})/f(z_n)$ is different from one. As z_n may vary, thus leading to different ratio values, one could be interested by how much, on the average, this ratio differs from one. Let us consider the natural logarithm of the ratio, so that its value is equal to 0 when the ratio is equal to one (as the logarithm is a monotonic function, the ordering of the

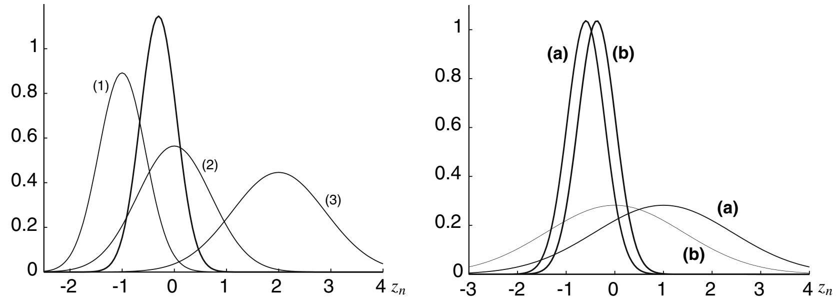

Fig. 3 *Left graph* shows Gaussian pdf's $f(z_n|y_{n,j})$ labeled as *1*, *2* and *3* with parameters $\mu_1 = -1$, $\mu_2 = 0$, $\mu_3 = 2$, $\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.5$, $\sigma_3^2 = 0.8$ along with the normalized product, which is Gaussian with $\mu = -0.30$ and $\sigma^2 = 0.12$. *Right graph* shows the fusion pdf's $\phi(z_n|\mathbf{y}_n)$ (*bold curves*) for two different choice *a* and *b* of the a priori pdf's $f(z_n)$ (*plain curves*) that are $N(1, 2)$ and $N(0, 2)$, respectively

ratios computed for $y_{n,1},...,y_{n,m}$ for the same z_n value will not be modified). The average discrepancy between $f(z_n|y_{n,j})$ and $f(z_n)$ with respect to Z_n and for a fixed $y_{n,j}$ value is thus given by

$$\mathbb{E}\left[\ln\frac{f(Z_n|y_{n,j})}{f(Z_n)}\right] = \int_{\mathbb{R}} f(z_n|y_{n,j})\ln\frac{f(z_n|y_{n,j})}{f(z_n)}\mathrm{d}z_n \quad \forall j=1,\ldots,m \tag{22}$$

where this equation is by definition the Kullback–Leibler (KL) distance or relative entropy between $f(z_n)$ and $f(z_n|y_{n,j})$ when the $y_{n,j}$ value is observed (Kullback and Leibler 1951). In practice, Eq. 22 can only be computed if $y_{n,j}$ is known. It is thus more convenient to define the expected KL distance with respect to $Y_{n,j}$, that can be used prior to any observation for $Y_{n,j}$. It is well known that this is by definition the mutual information $I(Z_n,Y_{n,j})$ (see Cover and Joy 2006), with

$$\begin{aligned} I(Y_{n,j},Z_n) &= \int_{\mathbb{R}}\int_{\mathbb{R}} f(y_{n,j},z_n)\ln\frac{f(y_{n,j},z_n)}{f(y_{n,j})f(z_n)}\mathrm{d}y_{n,j}\,\mathrm{d}z_n \\ &= \int_{\mathbb{R}} f(y_{n,j})\int_{\mathbb{R}} f(z_n|y_{n,j})\ln\frac{f(z_n|y_{n,j})}{f(z_n)}\mathrm{d}z_n\,\mathrm{d}y_{n,j} \\ &= \mathbb{E}\left[\int_{\mathbb{R}} f(z_n|Y_{n,j})\ln\frac{f(z_n|Y_{n,j})}{f(z_n)}\mathrm{d}z_n\right] \end{aligned} \tag{23}$$

Moreover, we also have the relation

$$I(Y_{n,j},Z_n) = H(Z_n) - H(Z_n|Y_{n,j}) \tag{24}$$

where $H(Z_n)$ is the entropy for Z_n and where $H(Z_n|Y_{n,j})$ is the conditional entropy, that are defined as

$$H(Z_n) = -\int_{\mathbb{R}} f(z_n)\ln f(z_n)\mathrm{d}z_n$$

$$H(Z_n|Y_{n,j}) = -\int_{\mathbb{R}} f(y_{n,j},z_n)\ln f(z_n|y_{n,j})\mathrm{d}y_{n,j}\,\mathrm{d}z_n$$

Comparing Eqs. 23 and 24 thus shows that measuring how much the logarithm of the ratio $f(z_n|y_{n,j})/f(z_n)$ differs from zero on the average is equivalent to measuring the information that $Y_{n,j}$ conveys on Z_n. As the various mutual information $I(Y_{n,j},Z_n)$ ($j = 1,...,m$) can be computed with respect to the same Z_n, they can be compared and sorted, thus allowing the selection of the "best" $Y_{n,j}$ variable, i.e., the variable for which the corresponding $I(Y_{n,j},Z_n)$ is maximum over the set $\{I(Y_{n,1},Z_n),...,I(Y_{n,m},Z_n)\}$. For example, in presence of abundant potential information (i.e., a high m value), for practical reasons one may consider using only a limited subset of $m' < m$ useful variables to be included.

If we denote $Y_{n,[1]},...,Y_{n,[m]}$ as the $Y_{n,j}$ variables sorted in decreasing order of mutual information values so that $I(Y_{n,[1]},Z_n) \geq I(Y_{n,[2]},Z_n) > \cdots \geq I(Y_{n,[m]},Z_n)$, a naive solution would be to use the subset of these m' first variables. Unfortunately, such a simple approach does not take into account the possible redundancy of information between the $Y_{n,[j]}$ variables (e.g., taken separately, $Y_{n,[1]}$ and $Y_{n,[2]}$ are the two most informative variables, but it is possible that $Y_{n,[2]}$ does not convey much more additional information on Z_n compared to what is already included when only using $Y_{n,[1]}$). A more satisfactory solution is obtained using the definition of the mutual information $I(\mathbf{Y}_n,Z_n)$ for the whole set of variables $Y_{n,1},...,Y_{n,m}$ (see e.g. Papoulis 1991, p. 562), with

$$I(\mathbf{Y}_n,Z_n) = \int_{\mathbb{R}}\int_{\mathbb{R}}\cdots\int_{\mathbb{R}} f(\mathbf{y}_n,z_n) \times \ln\frac{f(\mathbf{y}_n,z_n)}{f(\mathbf{y}_n)f(z_n)}\mathrm{d}y_{n,1}\cdots\mathrm{d}y_{n,m}\,\mathrm{d}z_n \tag{25}$$

Using the fact that

$$\begin{aligned} f(\mathbf{y}_n,z_n) &= f(y_{n,1},z_n)f(y_{n,2},z_n|y_{n,1})\cdots \\ &\quad\times f(y_{n,m},z_n|\{y_{n,1},\ldots,y_{n,m-1}\}) \\ f(\mathbf{y}_n) &= f(y_{n,1})f(y_{n,2}|y_{n,1})\cdots f(y_{n,m}|\{y_{n,1},\ldots,y_{n,m-1}\}) \end{aligned}$$

and substituting these results into Eq. 25, it is easy to show after some elementary manipulations that

$$\begin{aligned} I(\mathbf{Y}_n,Z_n) &= I(Y_{n,1},Z_n) + I(Y_{n,2},Z_n|Y_{n,1}) + \cdots \\ &\quad + I(Y_{n,m},Z_n|\{Y_{n,1},\ldots,Y_{n,m-1}\}) \end{aligned} \tag{26}$$

where the various (conditional) mutual information in the right-hand side of Eq. 26 are defined as

$$\begin{aligned} I(Y_{n,1},Z_n) &= H(Z_n) - H(Z_n|Y_{n,1}) \\ I(Y_{n,2},Z_n|Y_{n,1}) &= H(Z_n|Y_{n,1}) - H(Z_n|\{Y_{n,1},Y_{n,2}\}) \\ &\vdots \\ I(Y_{n,m},Z_n|\{Y_{n,1},\ldots,Y_{n,m-1}\}) &= H(Z_n|\{Y_{n,1},\ldots,Y_{n,m-1}\}) - H(Z_n|\mathbf{Y}_n) \end{aligned}$$

From a strict point of view, selecting the best (optimal) subset $\mathbf{Y}_{\mathrm{opt}}\subset\mathbf{Y}_n$ of m' variables would thus correspond to look for the m' variables such that, with respect to all other possible subsets of same size, the corresponding mutual information $I(\mathbf{Y}_{\mathrm{opt}},Z_n)$ is maximum. (see e.g. Fassinut-Mombot and Choquel 2004, for a similar idea). Unfortunately, this is an optimization problem that involves combinatorics, with number of possible combinations given by $C_m^{m'}$, the binomial coefficient. However, mutual information decomposition as given by Eq. 26 suggests a simpler forward selection procedure for selecting sequentially this subset:

1. select $Y_{n,[1]}$ so that $I(Y_{n,[1]},Z_n) \geq I(Y_{n,[j]},Z_n)\ \forall\ j$
2. select $Y_{n,[2]}$ so that $I(Y_{n,[2]},Z_n|Y_{n,[1]}) \geq I(Y_{n,[j]},Z_n|Y_{n,[1]})\ \forall\ j \neq 1$,
3. repeat this for $i = 3,\ldots,m'$, where at the ith step the selected $Y_{n,[i]}$ is such that $I(Y_{n,[i]},Z_n|\{Y_{n,[1]},\ldots,Y_{n,[i-1]}\}) \geq I(Y_{n,[j]},Z_n|\{Y_{n,[1]},\ldots,Y_{n,[i-1]}\})\ \forall\ j \neq 1,\ldots,i-1$.

this procedure being a direct analogy with the forward procedure in multiple regression, where the aim is the sequential selection of the most explanatory subset of variables to be included into a regression model (Neter et al. 1996).

4 A synthetic case study

In order to illustrate some of the general principle of the methodology, a simple synthetic case study will be presented. Clearly, due to space limitations, only some of the concepts presented in this paper can be illustrated here. We will thus mainly focus on a situation where what is sought for is to fuse information that do not easily fit jointly into a classical multivariate RF framework.

Let us assume that we are interested in the spatial prediction of a continuous random fields $\mathfrak{Z}$, with a typical realization as given by Fig. 4a. This corresponds to a smooth RF realization obtained by taking the absolute value of a zero-mean unit-variance Gaussian RF $\mathfrak{Y}$ with Gaussian covariance function, i.e. $Z_i = |Y_i|\ \forall\ \mathbf{x}_i\in D$ where $Y_i \sim N(0,1)$. The aim of smoothness and absolute value is to create a network-like appearance on the map, the location of this network being part of the available information. Let us define this network as a binary RF $\mathfrak{N}$ for which $(N_i = 1) \equiv (|Y_i| < y_{0.55})$, with $y_{0.55}$ the 0.55 quantile of a $N(0,1)$ distribution, so that $P(N_i = 1) = 0.1$ and $P(N_i = 0) = 0.9$ (see Fig. 4b). Aside from this RF, we will consider another independent realization of $\mathfrak{Y}$ that will be used to define the binary RF $\mathfrak{C}$ where $(C_i = 1) \equiv (Y_i > 0)$ so that $P(C_i = 0) = P(C_i = 1) = 0.5$ (see Fig. 4c). This information is irrelevant for predicting values of Fig. 4a as $C_i\perp Z_j\ \forall\ i,j$, but we will force its use in order to illustrate the automatic filtering properties of the method.

In order to rebuild Fig. 4a at best, we will consider that the only available information are a set of sparsely sampled values $\mathbf{Z} = (Z_1,\ldots,Z_{75})'$, along with the spatially exhaustive realization for $\mathfrak{N}$ and $\mathfrak{C}$. Clearly, this whole set of information does not lead to a nice and straightforward formulation in a traditional continuous multivariate RF approach if the aim is to predict Z_0 at unsampled locations $\mathbf{x}_0$, as (1) it requires to combine both continuous and categorical RF's, (2) the information brought by the realized event $n_0 = 0$ or 1 at prediction location is considerably poorer than the information brought by the distance d_{0i} between $\mathbf{x}_0$ and the closest location $\mathbf{x}_i$ for which $n_i = 1$, (3) unfortunately, this set of new random variables D_{0i} does not correspond to a second-order stationary RF.

Instead of relying on a rather complex multivariate approach, the idea is thus to recode the set of indirect information about $\mathfrak{Z}$ in terms of conditional distributions $f(z_0|d_{0i})$ and $f(z_0|c_0)$. Figure 4c shows the estimated regression $\mathbb{E}[Z_0|d_{0i}]$ along with confidence bounds from the estimated $\mathbb{V}ar[Z_0|d_{0i}]$, where it will be assumed that $Z_0|d_{0i}$ is Gaussian distributed. Assuming (wrongly) that $(Z_0,\mathbf{Z})'$ is Gaussian distributed too, the conditional $Z_0|\mathbf{z}$ is also Gaussian. Finally, assuming again normality, the corresponding $f(z_0|c_0)$ are estimated too when $c_0 = 1$ or when $c_0 = 0$ (see Fig. 6). Using now a NBF rule at the prediction location itself, we obtain according to Eq. 20 the result

$$f(z_0|\mathbf{z},d_{0i},c_0) \propto \frac{f(z_0|\mathbf{z})f(z_0|d_{0i})f(z_0|c_0)}{f^2(z_0)} \tag{27}$$

where all $f(z_0|\cdot)$'s on the right part of equality have been assumed Gaussian and where it is expected that $f(z_0|\mathbf{z},d_{0i},c_0) = f(z_0|\mathbf{z},d_{0i})$ as knowing C_0 is irrelevant, and thus $f(z_0|c_0) = f(z_0)$ either when $c_0 = 1$ or when $c_0 = 0$.

The resulting map of predicted values $\mathbb{E}[Z_0|\mathbf{z},d_{0i}]$ obtained using Eq. 27 is shown in Fig. 5b, along with the predicted map $\mathbb{E}[Z_0|\mathbf{z}]$ obtained by neglecting the network information. Clearly, accounting for the network yields significantly better results. A comparison between the maps for $\mathbb{E}[Z_0|\mathbf{z},d_{0i},c_0]$ and $\mathbb{E}[Z_0|\mathbf{z},d_{0i}]$ (not shown for the sake of brevity) also confirms that making use or not of the C_0 variable does not have any significant impact on the results, as expected.

More specifically, Fig. 6 illustrates the results obtained for two different prediction locations $\mathbf{x}_0$ that are respectively close or remote from the network. As

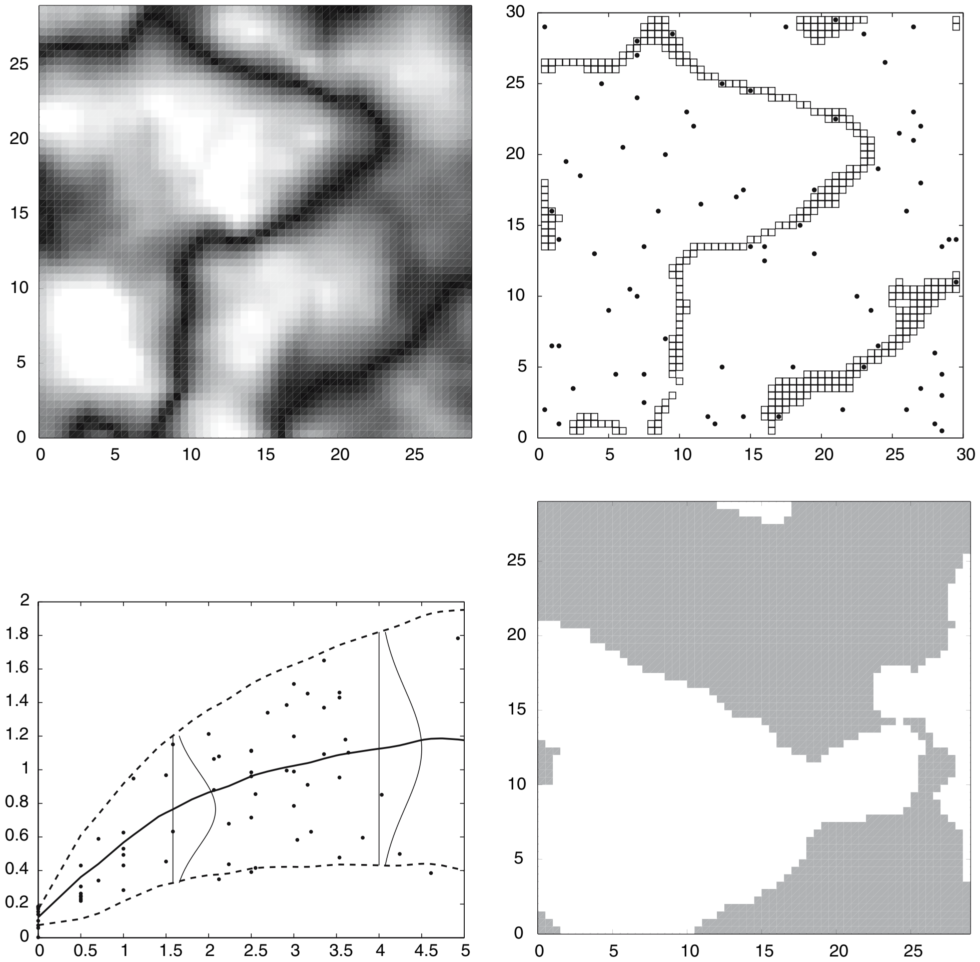

Fig. 4 Synthetic case study. **a** Shows a realization on a 60×60 grid of the $\mathfrak{Z}$ RF, where lower values appearing in *dark* have the meaning of a network. **b** Shows locations for 75 sampled values z_i (*black dots*) as a subset of the realization, along with locations where the network binary RF $\mathfrak{N}$ takes value $n_j = 1$ (*white squares*). **c** Shows the estimated regression $\mathbb{E}[Z_i|d_{ij}]$ (plain line) estimated from the (Z_i,d_{ij})'s (*black dots*) where d_{ij} is the Euclidean distance to the closest location where $n_j = 1$, along with the symmetric 0.95 confidence bounds (*dashed lines*) assuming normality. **d** Shows the realization of the $\mathfrak{C}$ RF ($c_i = 1$ in *white* and $c_i = 0$ in *gray*)

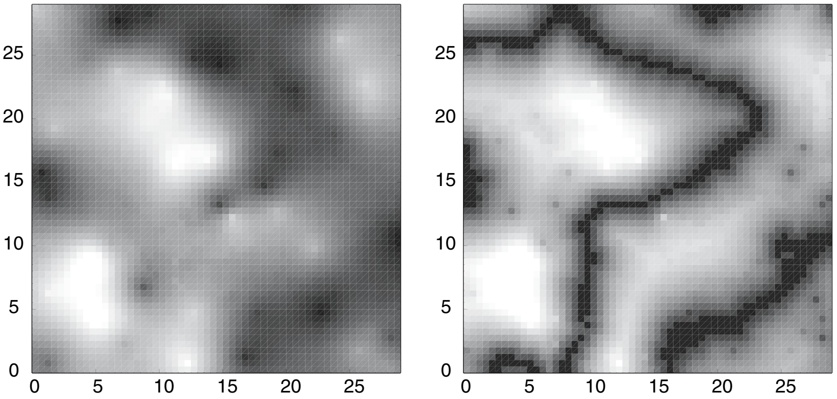

Fig. 5 Prediction results. **a** Shows the map of the $\mathbb{E}[Z_i|\mathbf{z}]$'s on the same 60×60 grid as for realization, thus neglecting network information (simple kriging). **b** Shows the map of the $\mathbb{E}[Z_i|\mathbf{z}, d_{ij}]$'s as obtained from the NBF rule neglecting information about C_0. Map of the $\mathbb{E}[Z_i|\mathbf{z}, d_{ij}, c_0]$'s (not shown here) is similar

we expect that $\lim_{d_{0i}\to\infty} f(z_0|d_{0i}) = f(z_0)$, the network does not have significant influence on the result when $\mathbf{x}_0$ is far from it, so the result is solely influenced by surrounding measurements $\mathbf{z}$, with $f(z_0|\mathbf{z},d_{0i}) \simeq f(z_0|\mathbf{z})$. On the contrary, for $\mathbf{x}_0$ close to or inside the network, $f(z_0|d_{0i})$ may become much more informative than $f(z_0|\mathbf{z})$, especially if there are no close $\mathbf{z}$ surrounding measurements, so that $f(z_0|\mathbf{z},d_{0i}) \simeq f(z_0|d_{0i})$, thus yielding a very good restitution of the network. From these figures, one can also see that, for the estimated distributions $f(z_0|c_0)$, we both have $f(z_0|c_0) \simeq f(z_0)$ when $c_0 = 1$ and when $c_0 = 0$, thus showing that these distributions will have no sensible effect on the final result for the fusion, i.e. $f(z_0|\mathbf{z},d_{0i},c_0) \simeq f(z_0|\mathbf{z},d_{0i})$.

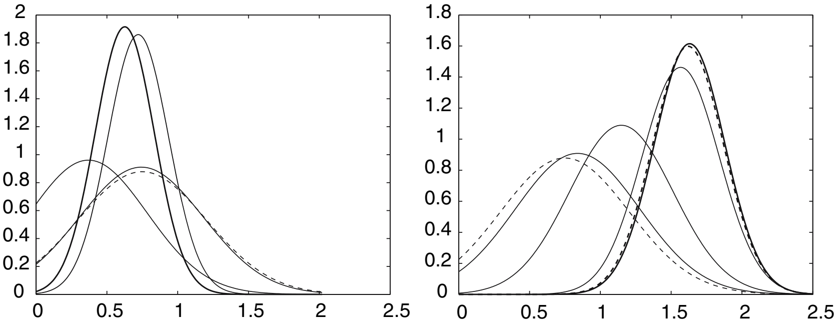

Fig. 6 Influence of the information in the fusion process. **a** Represents the pdf's to be fused along with the result using the NBF rule for a prediction location close to the network. **b** is the same for a remote location. *Symbols* for the curves are *dashed line* for a priori pdf $f(z_0)$, *plain lines* for (1) $f(z_0|d_{0i})$, (2) $f(z_0|\mathbf{z})$ and (3) $f(z_0|c_0)$, with $c_0 = 0$ in **a** and $c_0 = 1$ in **b**. *Bold lines* correspond to $f(z_0|\mathbf{z},d_{0i},c_0)$ as obtained from NBF rule. Additionally, *dashed bold lines* correspond to $f(z_0|\mathbf{z},d_{0i})$, i.e. neglecting the useless information about C_0. The two last curves cannot be distinguished from each other on the *left graph*

5 Discussion and conclusions

In this paper, starting from a simple and classical measurement errors context, it has been shown that a general formulation can be obtained for the prediction of a spatial RF at unsampled locations using various sources of direct and/or indirect measurements. Several well-known situations that arise as limit cases of the method have been presented, and connections between the convenient conditional independence hypothesis and maximum entropy properties have been emphasized too. Finally, with the help of information theory, it has been shown that a useful informativeness index can be derived with the aim of selecting the ‘‘best’’ subset of information to be used.

Clearly, the major interest of the proposed procedure is to allow the user to incorporate in a flexible way a potentially high number of secondary information sources that may exhibit high diversity, by avoiding at the same time major complications with respect to modeling hypotheses and computing time requirements. As all secondary information can be recoded (and fused if needed) as conditional distributions with respect to the variable of interest, this leads to a same generic formulation for prediction, that relies on subsequent multivariate integrations over a joint distribution.

Of course, this paper should not be viewed and is by no way a prosecution against the use of sound multivariate methods, that aim at fully accounting for the joint spatial dependence of RF's. However, it is stressed that the need for such models is not as critical as it may appear at a first sight. Moreover, it is not uncommon situation that multivariate models may bring more problems than solutions. Indeed, these models typically rely on more or less demanding hypotheses that may prove to be impossible to fulfill in a reasonable way with data at hand, especially when number and diversity of information sources that need to be accounted for is high. As a consequence, the user is commonly facing non-obvious choices like, e.g. (1) fooling the model by forcing it to use data that do not naturally match the required hypotheses, with potentially serious (and often poorly assessed) effects on the final results, (2) using dimensionality reduction techniques or data transformation for improving the adequacy between data properties and model hypotheses, or (3) discarding potentially valuable information based on the single fact that they do not nicely fit into the model framework. Even for situations were the need for such choices can be avoided, in most of the cases difficult practical issues remains to be addressed, as computing burden and modeling complexity is typically expected to quickly increases with number of information sources.

With respect to these aspects, the proposed methodology has nice features, as all efforts are concentrated on the initial optimal recoding of the information, thus cutting down the need for subsequent heavy computations and complex modeling issues. Clearly, the performance of the method is still to be evaluated and compared to more complex modeling approaches in specific circumstances. Among others, further work should be devoted to quantifying the impact on the final results of the information loss which is induced by neglecting the spatial correlation between RF's . However, based on the simple synthetic case study that has been presented, it can already be seen as a quite feasible and realistic alternative.

6 Appendices

6.1 Kriging with measurement errors

Assume that $\mathbf{Z}$ is a random vector sampled from a Gaussian second-order stationary spatial RF $\mathfrak{Z} = \{Z(\mathbf{x}) : \mathbf{x} \in D \subseteq \mathbb{R}^d, Z(\mathbf{x}) \in \mathbb{R}\}$ so that $\mathfrak{Z}$ is fully characterized by its mean function $\mu(\mathbf{x})$ and covariance function $C(\mathbf{h})$. Let us denote $\Sigma_\mathbf{Z} = \mathbb{C}ov[\mathbf{Z}]$, $\boldsymbol{\mu}_\mathbf{Z} = \mathbb{E}[\mathbf{Z}]$. Let us consider that $\mathbb{E}[\mathbf{E}] = \mathbf{0}$ and $\mathbb{C}ov[\mathbf{E}] = \sigma_E^2\mathbf{I}$, where $\mathbf{I}$ is an identity matrix of appropriate size. For $\mathbf{Y} = \mathbf{Z} + \mathbf{E}$, we thus have $\boldsymbol{\mu}_\mathbf{Y} = \boldsymbol{\mu}_\mathbf{Z}$ and $\Sigma_\mathbf{Y} = \Sigma_\mathbf{Z} + \sigma_E^2\mathbf{I}$ because $\mathbf{E}\perp\mathbf{Z}$. Let us now consider the following partitions

$$\mathbf{Y} = \begin{pmatrix} Y_0 \\ \mathbf{Y}_b \end{pmatrix}; \quad \boldsymbol{\mu}_\mathbf{Z} = \begin{pmatrix} \mu_0 \\ \boldsymbol{\mu}_b \end{pmatrix}; \quad \Sigma_\mathbf{Z} = \begin{pmatrix} \sigma_0^2 & \boldsymbol{\sigma}' \\ \boldsymbol{\sigma} & \Sigma_b \end{pmatrix}$$

The simple kriging with measurement errors predictor Z_0^p and the corresponding prediction variance $\sigma_0^{2,p}$ are then given by the expressions

$$Z_0^p = \boldsymbol{\lambda}'(\mathbf{Y}_b - \boldsymbol{\mu_b}) + \mu_0; \quad \sigma_0^{2,p} = \sigma_0^2 - \boldsymbol{\lambda}'(\Sigma_b + \sigma_E^2\mathbf{I})\boldsymbol{\lambda} \tag{A.1}$$

where the vector of weights $\boldsymbol{\lambda}$ is the solution of a linear system of equations

$$(\Sigma_b + \sigma_E^2\mathbf{I})\boldsymbol{\lambda} = \boldsymbol{\sigma} \iff \boldsymbol{\lambda}' = \boldsymbol{\sigma}'(\Sigma_b + \sigma_E^2\mathbf{I})^{-1}$$

Using Eq. 6 under the same hypotheses, it can then be shown that

$$Z_0^p = \mathbb{E}[Z_0|\mathbf{y}]; \quad \sigma_0^{2,p} = \mathbb{V}ar[Z_0|\mathbf{y}]$$

Indeed, from Eq. 5, it is clear that we have

$$f(\mathbf{y},\mathbf{z}) = f(\mathbf{y}|\mathbf{z})f(\mathbf{z}) = f(\mathbf{z})\prod_{i=0}^{n} f_{E_i}(y_i - z_i) \tag{A.2}$$

Plugging the expression of the Gaussian pdf's into Eq. A.2 gives

$$\begin{aligned}
f(\mathbf{y},\mathbf{z}) &\propto \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_\mathbf{Z})'\Sigma_\mathbf{Z}^{-1}(\mathbf{z}-\boldsymbol{\mu}_\mathbf{Z}) - \frac{1}{2\sigma_E^2}(\mathbf{y}-\mathbf{z})'(\mathbf{y}-\mathbf{z})\right) \\
&= \exp(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_\mathbf{Z})'\left(\Sigma_\mathbf{Z}^{-1} + \frac{1}{\sigma_E^2}\mathbf{I}\right)(\mathbf{z}-\boldsymbol{\mu}_\mathbf{Z}) \\
&\quad + \frac{1}{2\sigma_E^2}(\mathbf{y}-\boldsymbol{\mu}_\mathbf{Z})'(\mathbf{z}-\boldsymbol{\mu}_\mathbf{Z}) - \frac{1}{2\sigma_E^2}(\mathbf{y}-\boldsymbol{\mu}_\mathbf{Z})'(\mathbf{y}-\boldsymbol{\mu}_\mathbf{Z}) \\
&= \exp\left(-\tfrac{1}{2}\begin{pmatrix} \mathbf{z}-\boldsymbol{\mu}_\mathbf{Z} \\ \mathbf{y}-\boldsymbol{\mu}_\mathbf{Z} \end{pmatrix}'\begin{pmatrix} \Sigma_\mathbf{Z}^{-1} + \frac{1}{\sigma_E^2}\mathbf{I} & -\frac{1}{\sigma_E^2}\mathbf{I} \\ -\frac{1}{\sigma_E^2}\mathbf{I} & \frac{1}{\sigma_E^2}\mathbf{I} \end{pmatrix}\begin{pmatrix} \mathbf{z}-\boldsymbol{\mu}_\mathbf{Z} \\ \mathbf{y}-\boldsymbol{\mu}_\mathbf{Z} \end{pmatrix}\right)
\end{aligned}$$

and this final expression proves clearly that $(\mathbf{Y},\mathbf{Z})$ is multivariate Gaussian with $\mathbb{E}[\mathbf{Y}] = \mathbb{E}[\mathbf{Z}] = \boldsymbol{\mu}_\mathbf{Z}$, where the inner matrix in the last line is the inverse of the joint covariance matrix $\Sigma_{(\mathbf{Y},\mathbf{Z})}$, so that taking its inverse gives

$$\Sigma_{(\mathbf{Y},\mathbf{Z})} = \begin{pmatrix} \Sigma_\mathbf{Z} & \Sigma_\mathbf{Z} \\ \Sigma_\mathbf{Z} & \Sigma_\mathbf{Z} + \sigma_E^2\mathbf{I} \end{pmatrix}$$

Therefore, we know from multivariate Gaussian distribution theory that $Z_0|\mathbf{y}$ follows a Gaussian distribution with conditional mean and variance given by

$$\mathbb{E}[Z_0|\mathbf{y}] = \boldsymbol{\sigma}'\left(\Sigma_b + \sigma_E^2\mathbf{I}\right)^{-1}(\mathbf{Y}_b - \boldsymbol{\mu}_b) + \mu_0 = \boldsymbol{\lambda}'(\mathbf{Y}_b - \boldsymbol{\mu}_b) + \mu_0$$

$$\mathbb{V}ar[Z_0|\mathbf{y}] = \sigma_0^2 - \boldsymbol{\sigma}'\left(\Sigma_b + \sigma_E^2\mathbf{I}\right)^{-1}\boldsymbol{\sigma}$$

these being precisely the expressions for Z_0^p and $\sigma_0^{2,p}$ as given by Eq. A.1.

6.2 Maximum entropy with known moments

If we assume that we only know the vector of means $\boldsymbol{\mu}$ and covariance matrix Σ for $(\mathbf{E}', \mathbf{Z}')'$, with

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_\mathbf{E} \\ \boldsymbol{\mu}_\mathbf{Z} \end{pmatrix}; \quad \Sigma = \begin{pmatrix} \Sigma_\mathbf{E} & V \\ V & \Sigma_\mathbf{Z} \end{pmatrix}$$

then we know that the maximum entropy solution for its joint pdf will be $(\mathbf{E}', \mathbf{Z}')' \sim N(\boldsymbol{\mu}, \Sigma)$, with entropy $H(\mathbf{E},\mathbf{Z})$ given by

$$H(\mathbf{E},\mathbf{Z}) = \ln\sqrt{(2\pi)^{2n}} + \ln\sqrt{\det(\Sigma)}$$

so that $H(\mathbf{E},\mathbf{Z})$ is maximum if $\det(\Sigma)$ is maximum. We also know that

$$\det(\Sigma) = \det(\Sigma_\mathbf{Z})\det(\Sigma_\mathbf{E} - V\Sigma_\mathbf{Z}^{-1}V)$$

where $\det(\Sigma_\mathbf{Z})$ is known and where $\Sigma_\mathbf{E} - V\Sigma_\mathbf{Z}^{-1}V$ is the Schur's complement. From Fisher's inequality, we have

$$\det(\Sigma) \le \det(\Sigma_\mathbf{Z})\det(\Sigma_\mathbf{E})$$

with equality occurring for $V = \mathbf{0}$, thus showing that assuming the independence $\mathbf{E}\perp\mathbf{Z}$ leads to the maximum value if V is unknown. The maximum entropy solution is thus obtained with $\mathbf{Z} \sim N(\boldsymbol{\mu}_\mathbf{Z}, \Sigma_\mathbf{Z})$, $\mathbf{E} \sim N(\boldsymbol{\mu}_\mathbf{E}, \Sigma_\mathbf{E})$ and $\mathbf{E}\perp\mathbf{Z}$. From Hadamard's inequality, we also have

$$\det(\Sigma_\mathbf{E}) \le \prod_{i=0}^{n} \sigma_{E_i}^2$$

and as we finally know that

$$\det(\mathbf{D}) = \prod_{i=0}^{n} \sigma_{E_i}^2 \quad \text{with} \quad \mathbf{D} = \begin{pmatrix} \sigma_{E_0}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{E_1}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{E_n}^2 \end{pmatrix}$$

this shows that $\det(\Sigma) \leq \det(\Sigma_{\mathbf{Z}})\det(\mathbf{D})$. If only variances σ_{Ei}^2 are known instead of $\Sigma_{\mathbf{E}}$, the maximum entropy is thus reached by assuming additionally $E_0 \perp \cdots \perp E_n$.

References

Altinçay H (2005) On naive Bayesian fusion of dependent classifiers. Pattern Recognit Lett 26:2463–2473

Bogaert P, D'Or D (2003) Estimating soil properties from thematic soil maps: the Bayesian maximum entropy approach. Soil Sci Soc Am J 66:1492–1500

Chilès J-P, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York, 720 p

Cho S, Beak S, Kim JS (2003) Exploring artificial intelligence-based data fusion for conjoint analysis. Expert Syst Appl 24:287–294

Christakos G (2000) Modern spatiotemporal geostatistics. Oxford University Press, New York, 304 p. (3nd Reprint, 2001)

Christakos G (2002) On the assimilation of uncertain physical knowledge bases: Bayesian and non-Bayesian techniques. Adv Water Resour 25:1257–1274

Christakos G, Bogaert P, Serre ML (2002) Temporal GIS (with CD-ROM). Springer, Berlin Heidelberg New York, NY, 220 p

Costantini M, Farina A, Zirilli F (1997) The fusion of different resolution SAR images. In: Proceedings of the IEEE 85, pp 139–146

Cover T, Joy A (2006) Elements of information theory, 2nd edn. Wiley, New York, 748 pp

Cremer F, Schutte K, Schavemaker JGM, den Breejen E (2001) A comparison of decision-level sensor-fusion methods for anti-personnel land mine detection. Inf Fusion 2:187–208

Cressie N (1993) Statistics for spatial data, revised edition. Wiley, New York, 928 p

D'Or D, Bogaert P (2004) Spatial prediction of categorical variables with the Bayesian maximum entropy approach: the Ooypolder case study. Eur J Soil Sci 55:763-776

Duc B, Bigün ES, Bigün J, Maître G, Fischer S (1997) Fusion of audio and video information for multi modal person authentification. Pattern Recognit Lett 18:835–843

Fassinut-Mombot B, Choquel J-B (2004) A new probabilistic and entropy fusion approach for management of information sources. Inf Fusion 5:35–47

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York, 483 p

Gros XE, Bousigue J, Takahashi K (1999) NDT data fusion at the pixel level. NDT E Int 32:283–292

Jaynes ET (2003) Probability theory: the logic of science. Cambridge University Press, Cambridge, 758 p

Jones GD, Allsop RE, Gilby JH (2003) Bayesian analysis for fusion of data from disparate imaging systems for surveillance. Image Vis Comput 21:843-849

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86

Kuncheva LI (2004) Combining pattern classifiers methods and algorithms. Wiley, New York, 350 p

Lewis D (1998) Naive (Bayes) at forty: the independence assumption in information retrieval. Conference proceedings of the European Conference on Machine Learning, Springer, Berlin Heidelberg New York, pp 4–15

Melgani F, Serpico SB (2002) A statistical approach to the fusion of spectral and spatio-temporal contextual information for the classification of remote-sensing images. Patter Recognit Lett 23:1053–1061

Neter J, Kutner MH, Wasserman W (1996) Applied linear statistical models. McGraw-Hill/Irwin, 1408 p

Papoulis A (1991) Probability, random variables, and stochastic processes. McGraw-Hill, 3rd edn, 666 p

Pradalier C, Colas F, Bessiere P (2003) Expressing Bayesian fusion as a product of distributions: application in robotics. In: Proceedings IEEE-RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, USA

van der Putten P, Kok JN, Gupta A (2002) Why the information explosion can be bad for data mining, and how data fusion provides a way out. In: Proceedings of the Second SIAM International Conference on Data Mining, Arlington, USA

Rässler S (2004) Data fusion: identification problems, validity, and multiple imputation. Aust J Stat 33:153–171

Ross A, Jain A (2003) Information fusion in biometrics. Pattern Recognit Lett 24:2115–2125

Savelievaa E, Demyanova V, Kanevski M, Serre M, Christakos G (2005) BME-based uncertainty assessment of the Chernobyl fallout. Geoderma 128:312–324

Simone G, Farina A, Morabito FC, Serpico SB, Bruzzone L (2002) Image fusion techniques for remote sensing applications. Inf Fusion 3:3–15

Sohn SY, Lee SH (2003) Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. Saf Sci 41:1–14

Song XB, Abu-Mostafa Y, Sill J, Kasdan H, ad Pavel M (2003) Robust image recognition by fusion of contextual information. Inf Fusion 3:277–287

Wackernagel H (1995) Multivariate Geostatistics. Springer, Berlin Heidelberg New York, 291 p

Wikle CK, Milliff RF, Nychka D, Berliner LM (2001) Spatial-temporal hierarchical Bayesian modeling: tropical ocean surface winds. J Am Stat Assoc 96:382–397