Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-018-1578-1

ORIGINAL PAPER

A comparison of search strategies to design the cokriging neighborhood for predicting coregionalized variables

Nasser Madani[1] · Xavier Emery[2,3]

Abstract
Cokriging allows predicting coregionalized variables from sampling information, by considering their spatial joint dependence structure. When secondary covariates are available exhaustively, solving the cokriging equations may become prohibitive, which motivates the use of a moving search neighborhood to select a subset of data, based on their closeness to the target location and the screen effect approximation. This paper investigates the efficiency of different strategies for designing a sub-optimal neighborhood wherein the simplification of the cokriging equations is challenging. To do so, five alternatives (single search, multiple search, strictly collocated search, multi-collocated search and isotopic search) are tested and compared with the reference unique neighborhood, through synthetic examples with different data configurations and spatial joint correlation models. The results indicate that the multi-collocated and multiple searches bear the highest resemblance to the reference case under the analyzed spatial structure models, while the single and the isotopic searches, which do not differentiate the primary and secondary sampling designs, yield the poorest results in terms of cokriging error variance.

Keywords Screening effect · Multi-collocated cokriging · Strictly collocated cokriging · Markov-type models · Intrinsic correlation · Cokriging neighborhood · Heterotopic sampling

1 Introduction

Cokriging is used in the earth sciences for predicting coregionalized variables at locations where no observation is available. Application fields include mineral resource assessment (Journel and Huijbregts 1978; Pan et al. 1993; Gálvez and Emery 2011; Emery 2012; da Silva and Costa 2014; Minnitt and Deutsch 2014; Uygucgil and Konuk 2015; Cornah and Machaka 2015), petroleum reservoir modeling (Xu et al. 1992; Hohn 1999; Masihi and Zarei 2010; Schwab et al. 2011; Cao et al. 2014; Jalalalhosseini et al. 2014), groundwater hydrology (Ahmed and de Marsily 1987; D'Agostino et al. 1997, 1998; Kitanidis 1997; Boezio et al. 2006; Dalla Libera et al. 2017; Olea et al. 2018), geochemistry (Wackernagel 1988; Roberts and McKenna 2009; Tolosana-Delgado and van den Boogaart 2013; Lark et al. 2014; Pawlowsky-Glahn et al. 2015; Fabijańczyk et al. 2016; Fouedjio 2018), soil sciences (Yates and Warrick 1987; Stein et al. 1988), and environmental sciences (Goovaerts 1997; Bohorquez et al. 2017; Borkowski and Kwiatkowska-Malina 2017).

Cokriging is of particular importance when the variable of main interest (hereafter called primary variable) is sparsely sampled and is correlated with one or several secondary variables that are available extensively at the locations where the primary variable must be predicted (Vargas-Guzmán and Jim Yeh 1999; Wackernagel 2003). However, in such a case, applying cokriging may be problematic due to the computational requirements caused by the large number of data to process (Emery 2009; Gálvez and Emery 2011; Chilès and Delfiner 2012). This situation motivates the need to reduce the number of data to be used in the cokriging system, by considering the data located in a neighborhood of the target location and dropping out all the remaining data. In this respect, several strategies have been proposed to select the neighboring data, such as the strictly

✉ Nasser Madani
nasser.madani@nu.edu.kz

[1] Department of Mining Engineering, School of Mining and Geosciences, Nazarbayev University, Astana, Kazakhstan

[2] Department of Mining Engineering, University of Chile, Santiago, Chile

[3] Advanced Mining Technology Center, University of Chile, Santiago, Chile

Published online: 06 July 2018

collocated cokriging approximation (Xu et al. 1992), where a single data of each secondary variable (the one situated at the target location) is retained, or the multi-collocated approximation (Rivoirard 2001), which also incorporates the secondary data that are collocated with the primary data.

The abovementioned strategies are based on the concept of screening effect, according to which the information of the selected neighboring data screens out the influence of the other data, which would have a small (ideally, a zero) weight in the full cokriging implementation (Goovaerts 1997; Chilès and Delfiner 2012). Rivoirard (2001, 2004) and Subramanyam and Pandalai (2004, 2008) showed that the screening of either primary or secondary data actually depends on the multivariate data configuration and also on the spatial correlation structure of the coregionalized variables. A situation of interest arises when the cross-covariance functions between the secondary and primary variables are proportional to the direct covariance (auto-covariance) of the primary variable, in which case the secondary data are totally screened out by the collocated primary data (Rivoirard 2004; Subramanyam and Pandalai 2004). However, other authors claim that, in such a case, only the secondary data located at the target location is worthwhile being selected (strictly collocated cokriging), a practice that is still widespread in application fields related to natural resources assessment.

The goal of this paper is twofold. First, it is of interest to show how the spatial correlation structure of the coregionalized variables relates to the screening effect property. Second, it aims at providing guidelines to define a suitable search strategy (design of a moving neighborhood) that yields optimal or sub-optimal cokriging results, hence minimizing the loss of information caused by the discarded data when cokriging in a unique neighborhood (keeping all the primary and secondary data) is impractical. The outline is as follows: Sect. 2 recalls the main concepts about cokriging, data selection and coregionalization modeling that will be used in the paper; Sect. 3 investigates, through a synthetic example, the relationships between screening effect, strictly collocated and multi-collocated cokriging under specific coregionalization models, while Sect. 4 addresses the problem of comparing five neighborhood designs (in terms of prediction accuracy) under different coregionalization models, to determine which design yields the results closest to, or farthest from, that of the unique neighborhood. Conclusions follow in Sect. 5.

2 Recall on geostatistical multivariate modeling and prediction

2.1 Cokriging

2.1.1 Conventional simple cokriging

Simple cokriging is a generalization of simple kriging, i.e., kriging with a known mean value, and aims to predict primary and secondary variables by taking into account their joint spatial correlation structure (Journel and Huijbregts 1978; Goovaerts 1997; Wackernagel 2003; Chilès and Delfiner 2012). Provided that these variables are represented by second-order stationary random fields, the cokriging predictor and the variance of the prediction error (known as the simple cokriging variance) for the primary variable (hereafter denoted with index 1) given one secondary variable (denoted with index 2) are defined as (Myers 1982):

$$Z^*_{SCK}(x_0) = m_1 + \sum_{\alpha=1}^{n_1} \omega^1_\alpha \left(Z_1\left(x_{1,\alpha}\right) - m_1\right) + \sum_{\alpha=1}^{n_2} \omega^2_\alpha \left(Z_2\left(x_{2,\alpha}\right) - m_2\right) \quad (1)$$

$$\sigma^2_{SCK}(x_0) = C_{11}(x_0 - x_0) - \sum_{\alpha=1}^{n_1} \omega^1_\alpha C_{11}\left(x_{1,\alpha} - x_0\right) - \sum_{\alpha=1}^{n_2} \omega^2_\alpha C_{21}\left(x_{2,\alpha} - x_0\right) \quad (2)$$

where ω^i_α (i = 1, 2) is the weight assigned to the data $Z_i\left(x_{i,\alpha}\right)$ of the i-th variable Z_i at the α-th data location $x_{i,\alpha}$ (α = 1,… n_i) of this variable, x_0 is the location targeted for prediction; m_i is the mean value of the i-th variable Z_i; C_{ij} is the direct $(i = j)$ or cross $(i \neq j)$ covariance between variables Z_i and Z_j (i, j = 1, 2). The previous equations can be generalized to the case with more than one secondary variable, at the price of heavier notation, which will not be considered in this work. Note that the numbers of data are not necessarily the same for the primary and secondary variables, a case known as a heterotopic sampling design (Wackernagel 2003) in opposition to the isotopic (equally-sampled) case. The weights ω^i_α required in Eqs. (1) and (2) are obtained by solving the following system of linear equations:

$$\begin{cases} \sum_{\alpha=1}^{n_1} \omega^1_\alpha C_{11}\left(x_{1,\beta} - x_{1,\alpha}\right) + \sum_{\alpha=1}^{n_2} \omega^2_\alpha C_{12}\left(x_{1,\beta} - x_{2,\alpha}\right) = C_{11}\left(x_{1,\beta} - x_0\right), \beta = 1, \ldots n_1 \\ \sum_{\alpha=1}^{n_1} \omega^1_\alpha C_{21}\left(x_{2,\beta} - x_{1,\alpha}\right) + \sum_{\alpha=1}^{n_2} \omega^2_\alpha C_{22}\left(x_{2,\beta} - x_{2,\alpha}\right) = C_{21}\left(x_{2,\beta} - x_0\right), \beta = 1, \ldots n_2 \end{cases} \quad (3)$$

Different neighborhood strategies can be used to reduce the number of data for cokriging. For instance, a single search strategy selects the data locations that are geographically the closest to the target location x_0, irrespective of which variables are known at those locations, whereas a multiple search strategy consists in selecting the closest data of each (primary or secondary) variable.

2.1.2 Strictly collocated cokriging

Strictly collocated cokriging only retains the secondary data located at x_0 along with the primary data $Z_1(x_{1,\alpha}), \alpha = 1, \ldots n_1$. This secondary data is assumed to screen out the influence of the secondary data that are located farther away (Journel 1999). In the case of a single secondary variable (Z_2), the predictor and the error variance are built up with (Xu et al. 1992; Almeida and Journel 1994):

$$Z^*_{SCCK}(x_0) = m_1 + \sum_{\alpha=1}^{n_1} \omega^1_\alpha \left(Z_1(x_{1,\alpha}) - m_1\right) + \omega^2_0 (Z_2(x_0) - m_2) \quad (4)$$

$$\sigma^2_{SCCK}(x_0) = C_{11}(x_0 - x_0) - \sum_{\alpha=1}^{n_1} \omega^1_\alpha C_{11}(x_{1,\alpha} - x_0) - \omega^2_0 C_{21}(x_0 - x_0) \quad (5)$$

and the strictly collocated cokriging system for such a neighborhood is:

$$\begin{cases} \sum_{\alpha=1}^{n_1} \omega^1_\alpha C_{11}(x_{1,\beta} - x_{1,\alpha}) + \omega^2_0 C_{12}(x_{1,\beta} - x_0) = C_{11}(x_{1,\beta} - x_0), \beta = 1, \ldots n_1 \\ \sum_{\alpha=1}^{n_1} \omega^1_\alpha C_{21}(x_0 - x_{1,\alpha}) + \omega^2_0 C_{22}(x_0 - x_0) = C_{21}(x_0 - x_0) \end{cases} \quad (6)$$

with the same notations as in the previous subsection, except for the index 0 used to numerate the location ($x_{2,0} = x_0$) and the weight (ω^2_0) assigned to the collocated secondary data $Z_2(x_0)$.

2.1.3 Multi-collocated cokriging

In multi-collocated cokriging, the retained secondary data are the ones available at the target location x_0 and at the locations of the primary data $x_{1,\alpha}, \alpha = 1, \ldots n_1$. In the case of a single secondary variable, the cokriging predictor and the error variance are given by (Rivoirard 2001; Wackernagel 2003; Chilès and Delfiner 2012):

$$Z^*_{MCCK}(x_0) = m_1 + \sum_{\alpha=1}^{n_1} \omega^1_\alpha \left(Z_1(x_{1,\alpha}) - m_1\right) + \sum_{\alpha=0}^{n_1} \omega^2_\alpha \left(Z_2(x_{2,\alpha}) - m_2\right) \quad (7)$$

$$\sigma^2_{MCCK}(x_0) = C_{11}(x_0 - x_0) - \sum_{\alpha=1}^{n_1} \omega^1_\alpha C_{11}(x_{1,\alpha} - x_0) - \sum_{\alpha=0}^{n_1} \omega^2_\alpha C_{21}(x_{2,\alpha} - x_0) \quad (8)$$

with $x_{2,\alpha} = x_{1,\alpha}$ for $\alpha = 1, \ldots n_1$ and $x_{2,0} = x_0$. The cokriging weights are obtained by solving the following equations:

$$\begin{cases} \sum_{\alpha=1}^{n_1} \omega^1_\alpha C_{11}(x_{1,\beta} - x_{1,\alpha}) + \sum_{\alpha=0}^{n_1} \omega^2_\alpha C_{12}(x_{1,\beta} - x_{2,\alpha}) = C_{11}(x_{1,\beta} - x_0), \beta = 1, \ldots n_1 \\ \sum_{\alpha=1}^{n_1} \omega^1_\alpha C_{21}(x_{2,\beta} - x_{1,\alpha}) + \sum_{\alpha=0}^{n_1} \omega^2_\alpha C_{22}(x_{2,\beta} - x_{2,\alpha}) = C_{21}(x_{2,\beta} - x_0), \beta = 0, \ldots n_1 \end{cases} \quad (9)$$

2.2 Coregionalization modeling

2.2.1 Linear model of coregionalization (LMC)

Solving the cokriging system requires the knowledge of the direct and cross-covariances between the primary and secondary variables. In this respect, the linear model of coregionalization is widely used to fit such covariances, owing to its mathematical simplicity and tractability (Journel and Huijbregts 1978; Goovaerts 1997; Wackernagel 2003). In this model, the direct and cross-covariances $C_{ij}(h)$ $(i,j = 1,2)$ are defined as weighted sums of L basic covariances, also called basic nested structures:

$$C_{ij}(h) = \sum_{l=1}^{L} b^l_{ij} c_l(h) \quad (10)$$

where, for each structure ($l = 1, \ldots L$), $\left(b^l_{ij}\right)_{i,j=1,2}$ is a 2×2 real-valued, symmetric, positive semi-definite matrix (coregionalization matrix) and $c_l(h)$ is a permissible stationary covariance model (basic nested structure). In practice, such a model can be fitted to a set of experimental direct and cross-covariances by means of semi-automated algorithms (Goulard and Voltz 1992; Emery 2010).

2.2.2 Markov-type models

Other models for describing the joint spatial correlation structure of coregionalized variables are the Markov-type models, denoted as MM1 and MM2 in the literature. MM1 needs to model the primary covariance function $C_{11}(h)$; the cross-covariance functions $C_{12}(h)$ and $C_{21}(h)$ are then

inferred by the following approximation (Almeida and Journel 1994):

$$C_{12}(h) = C_{21}(h) \cong \frac{C_{12}(0)}{C_{11}(0)} C_{11}(h), \quad (11)$$

where $C_{12}(0)$ is the covariance between primary and secondary collocated data, while $C_{11}(0)$ is the variance of the primary data. The resulting cross-covariances (Eq. 11) are proportional to the primary direct covariance $C_{11}(h)$ and share its characteristics (shape, correlation range, relative nugget effect).

If the cross-covariances $C_{12}(h)$ and $C_{21}(h)$ share the characteristics of the secondary covariance $C_{22}(h)$, one may use the MM2 model instead (Journel 1999):

$$C_{12}(h) = C_{21}(h) \cong \frac{C_{12}(0)}{C_{22}(0)} C_{22}(h), \quad (12)$$

so that the cross-covariances are now proportional to the secondary direct covariance.

2.2.3 Intrinsic correlation model

The intrinsic correlation model is the simplest model, as it assumes that all the direct and cross-covariances are proportional to the same spatial correlation function (Wackernagel 2003):

$$C_{ij}(h) = b_{ij} c(h) \quad (13)$$

where $\left(b_{ij}\right)_{i,j=1,2}$ is a 2×2 symmetric, positive semi-definite matrix (coregionalization matrix) and $c(h)$ is a permissible covariance model. This is a particular case of both MM1 and MM2 models, and also of the linear model of coregionalization (with $L = 1$ basic covariance).

3 Investigating the screening effect in strictly and multi-collocated cokriging

Several authors argue that, under the assumption of a Markov-type model (Journel 1999; Babak and Deutsch 2009) or an intrinsic correlation model (Rivoirard 2001; Wackernagel 2003), the collocated secondary data totally screen out the influence of any other secondary data.

In particular, there is a wide belief that it is enough, for the prediction of a primary variable, to retain the secondary data situated at the target location and that the remaining secondary data do not add substantial knowledge. In other words, strictly collocated cokriging would be equivalent to full cokriging. One consequence of this result is that the error variance should not be affected by adding more secondary data.

To demonstrate that this belief is erroneous, we will show a few examples in a two-dimensional Euclidean space, in which non-collocated secondary data receive non-zero weights in the cokriging predictor. The following cases are considered.

Case I (strictly collocated cokriging) Primary data are available at the four vertices of a square $\{x_1, x_2, x_3, x_4\}$ and a single secondary data is available at the target location x_0 that coincides with the center of the square (Fig. 1a).
Case II (multi-collocated cokriging) Primary data are available at locations $\{x_1, x_2, x_3, x_4\}$ and secondary data are available at locations $\{x_0, x_1, x_2, x_3, x_4\}$ (Fig. 1b).
Case III (full cokriging) Primary data are available at locations $\{x_1, x_2, x_3, x_4\}$ and secondary data are available at the target location x_0, at the primary data locations $\{x_1, x_2, x_3, x_4\}$ and at other four locations in the square $\{x_5, x_6, x_7, x_8\}$ (Fig. 1c).

In each case, three coregionalization models are tested, in which the direct and cross-covariances are isotropic exponential (*Exp*) structures:

- MM1:

$$C_{11}(h) = 1.0Exp_{14}(h), C_{12}(h) = 0.7Exp_{14}(h), C_{22}(h) = 1.0Exp_{10}(h)$$

- MM2:

$$C_{11}(h) = 1.0Exp_{10}(h), C_{12}(h) = 0.7Exp_{14}(h), C_{22}(h) = 1.0Exp_{14}(h)$$

- Intrinsic correlation:

$$C_{11}(h) = C_{22}(h) = 1.0Exp_{10}(h), C_{12}(h) = 0.7Exp_{10}(h).$$

The coefficients preceding each exponential structure indicate the sill of this structure, while the values in subscript indicate the practical ranges of correlation (distances beyond which the correlation is less than 5% of the sill value). All these models can be seen as particular cases of the parsimonious bivariate Matérn model proposed by Gneiting et al. (2010); the conditions of mathematical validity are fulfilled in each case.

For each case of data configuration and each coregionalization model, cokriging is performed to predict the primary variable at the center of the square (x_0). Table 1 shows the resulting weights assigned to the primary and secondary data, as well as the error variance.

As can be seen in the table, one observes a reduction of the error variance when more secondary information is appended (i.e., from case I to case II, and from case II to case III), which indicates equal or better precision of the

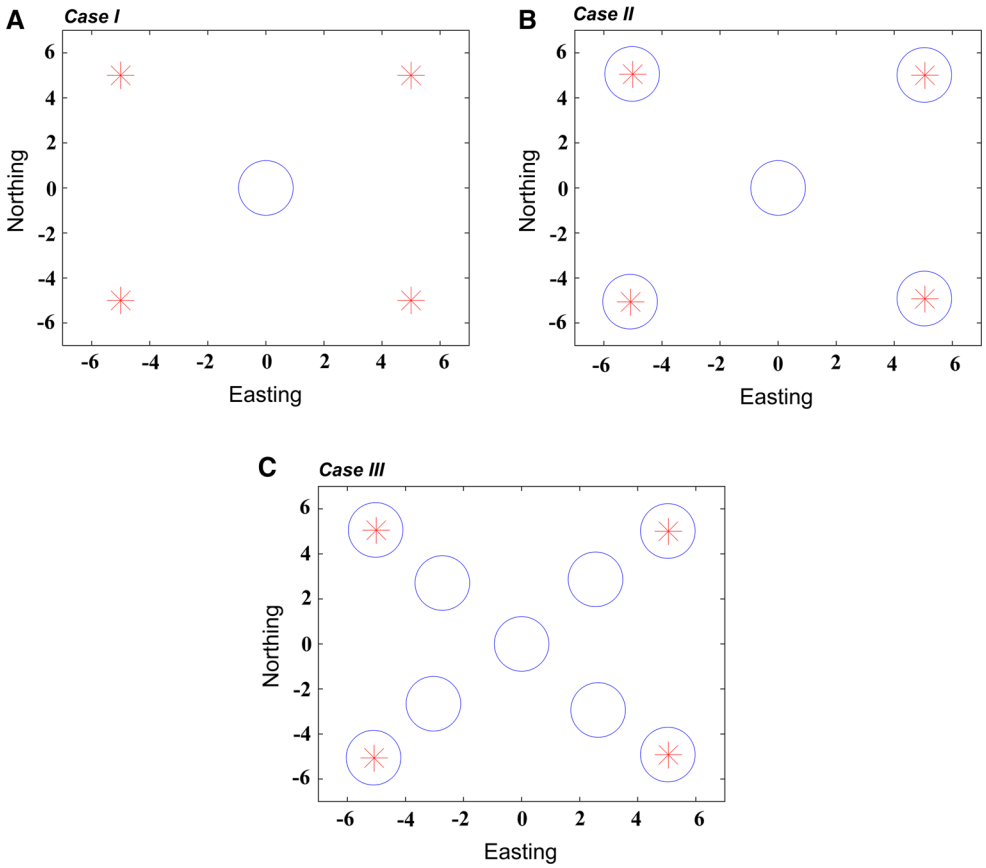

Fig. 1 Three different configurations for primary and secondary data locations (primary data: red crosses, secondary data: blue circles)

Table 1 Simple cokriging weights assigned to primary and secondary data, for each data configuration, coregionalization model and cokriging type

Locations	Coordinates		Data	MM1			MM2			Intrinsic correlation		
	East	North		Case I	Case II	Case III	Case I	Case II	Case III	Case I	Case II	Case III
x_1	− 5	− 5	Primary	0.0943	0.1055	0.0946	0.0120	0.0251	0.0251	0.0563	0.1076	0.1076
x_2	5	− 5	Primary	0.0943	0.1055	0.0946	0.0120	0.0251	0.0251	0.0563	0.1076	0.1076
x_3	5	5	Primary	0.0943	0.1055	0.0946	0.0120	0.0251	0.0251	0.0563	0.1076	0.1076
x_4	− 5	5	Primary	0.0943	0.1055	0.0946	0.0120	0.0251	0.0251	0.0563	0.1076	0.1076
x_1	− 5	− 5	Secondary		− 0.0161	− 0.0302		− 0.0176	− 0.0176		− 0.0753	− 0.0753
x_2	5	− 5	Secondary		− 0.0161	− 0.0302		− 0.0176	− 0.0176		− 0.0753	− 0.0753
x_3	5	5	Secondary		− 0.0161	− 0.0302		− 0.0176	− 0.0176		− 0.0753	− 0.0753
x_4	− 5	5	Secondary		− 0.0161	− 0.0302		− 0.0176	− 0.0176		− 0.0753	− 0.0753
x_5	− 3	− 3	Secondary			0.0481			0.0000			0.0000
x_6	3	− 3	Secondary			0.0481			0.0000			0.0000
x_7	3	3	Secondary			0.0481			0.0000			0.0000
x_8	− 3	3	Secondary			0.0481			0.0000			0.0000
x_0	0	0	Secondary	0.6420	0.6428	0.6024	0.6926	0.7000	0.7000	0.6811	0.7000	0.7000
Error variance				0.4677	0.4672	0.4595	0.5094	0.5088	0.5088	0.4962	0.4837	0.4837

predictor when more data is available (Emery 2009). In the full cokriging configuration (case III), all the secondary data, at either the locations in common with the primary data or the extra locations, receive non-zero weights under the MM1 spatial structure model, which indicates that no screening effect occurs with this MM1 model. In contrast,

under the MM2 model and intrinsic correlation model (a particular case of MM2), the secondary data at locations that do not coincide with the target location or with the primary data locations receive a zero weight, which corroborates that, in these spatial correlation models, full cokriging (case III) reduces to multi-collocated cokriging (case II), but never to strictly collocated cokriging (case I).

This result agrees with the findings of Rivoirard (2001) and goes against the argument of Journel (1999) according to which the collocated secondary data screen out the influence of all other secondary data when predicting a primary variable under a MM2 model. The proof given by Journel is actually valid in the absence of any primary data, but is erroneous when primary data are introduced. Several authors (Almeida and Journel 1994; Goovaerts 1997; Journel 1999; Babak and Deutsch 2009) have, mistakenly, suggested the presence of a screening effect and/or the equivalence between full cokriging and strictly collocated cokriging under a Markov-type (either MM1 or MM2) model.

To prove that, under the MM1, MM2 or intrinsic correlation model, strictly collocated cokriging cannot be equivalent to full cokriging (unless the specific cases of no spatial auto-correlation for the primary variable or no spatial cross-correlation between primary and secondary variables), let us consider the multi-collocated configuration (case II) and the intrinsic correlation model, which is a particular case of Markov-type model (both MM1 and MM2). When removing the collocated secondary data $Z_2(x_0)$, it is known (Emery 2009) that the weight of any retained data increases by the weight of the removed data (ω_0^2) times the cokriging weight assigned to the retained data when predicting the removed data. On the other hand, under the intrinsic correlation model and in an isotopic configuration (this situation holds when $Z_2(x_0)$ is removed), cokriging reduces to kriging each variable separately (Wackernagel 2003; Subramanyam and Pandalai 2004). Accordingly, the weights of the primary data remain unchanged (the removal of $Z_2(x_0)$ has no effect on the primary weights), while the weight of the secondary data $Z_2(x_{2,\alpha})$ ($\alpha = 1, \ldots, 4$) increases by $\omega_0^2\omega_\alpha^1$ and becomes equal to zero (secondary data receive zero weights under a isotopic configuration and intrinsic correlation model), that is: $\omega_\alpha^2 + \omega_0^2\omega_\alpha^1 = 0$ for $\alpha = 1, \ldots, 4$. Therefore, unless the primary data receive zero weights ($\omega_1^1 = \omega_2^1 = \omega_3^1 = \omega_4^1$), which happens with a pure nugget primary direct covariance model, or the collocated secondary data $Z_2(x_0)$ receives a zero weight ($\omega_0^2 = 0$), which happens when the cross-covariance between primary and secondary variables is identically zero, the secondary data weight ω_α^2 differs from zero. To sum up, in the intrinsic correlation model (therefore, also in the MM1 and MM2 models), the secondary data collocated with the primary data are likely to receive a non-zero weight and full cokriging is not the same as strictly collocated cokriging.

4 Investigating the efficiency of cokriging search strategies

4.1 Definition of search strategies and coregionalization models

In this section, it is of interest to compare different strategies for choosing the cokriging neighborhood, i.e., for selecting the relevant data for cokriging, and to determine to what extent multi-collocated cokriging bears a resemblance to full cokriging with a unique neighborhood. To do so, a two-dimensional regular grid with 60×60 nodes is created and 200 out of the 3600 nodes are randomly selected as sampling locations. The primary variable is allocated to the 200 sampling locations, whereas the secondary variable is exhaustively allocated to all the 3600 grid nodes (Fig. 2a). Simple cokriging is then applied to derive the variances of the prediction errors for the primary variable at the 3400 grid nodes where this variable has not been sampled. These variances are used as a criterion for comparing the following neighborhood strategies:

1. *Single search (SS)* This strategy searches for the data at the 20 closest locations, irrespective of whether the primary and/or secondary variables are known at these locations.
2. *Multiple search (MS)* This strategy is implemented into two parts: the first part searches for the 20 closest data of the primary variable and the second part searches for the 20 closest data of the secondary variable, independently of the first part.
3. *Isotopic search (IS)* The 20 closest sampling locations that convey both the primary and secondary variables are selected.
4. *Strictly collocated search (SCS)* The primary data at the 20 closest sampling locations are selected, together with the secondary data at the target location.
5. *Multi-collocated search (MCS)* The 20 closest sampling locations that convey both the primary and secondary variables are selected, together with the secondary data at the target location.
6. *Unique search (US)* All the available primary (200) and secondary (3600) data are selected.

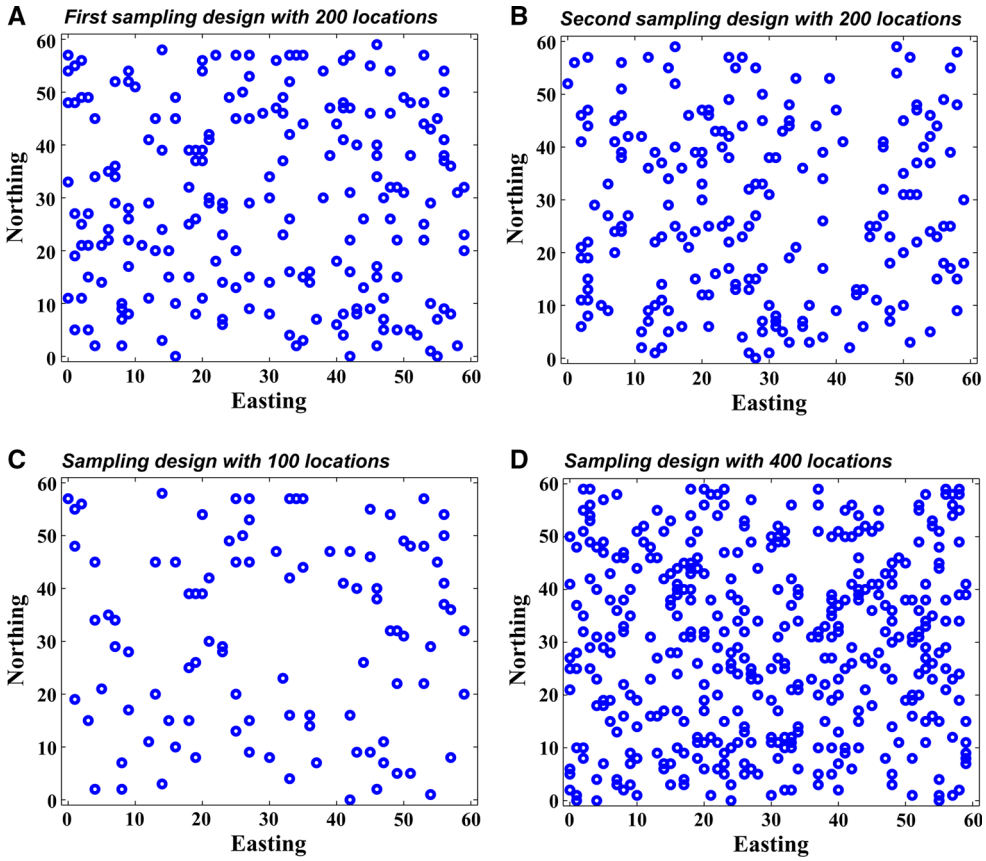

Fig. 2 Four sampling designs with 200 (**a**, **b**), 100 (**c**) and 400 (**d**) primary data locations

Six spatial structure models, involving isotropic exponential structures (*Exp*) and nugget effects (*Nug*), are considered:

- MM1-A:

$$C_{11}(h) = Exp_{56}(h), C_{12}(h) = 0.7Exp_{56}(h), C_{22}(h) = Exp_{40}(h)$$

- MM1-B:

$$C_{11}(h) = 0.3Nug(h) + Exp_{56}(h), C_{12}(h) = 0.21Nug(h) + 0.7Exp_{56}(h), C_{22}(h) = 0.3nug(h) + Exp_{40}(h)$$

- MM2-A:

$$C_{11}(h) = Exp_{40}(h), C_{12}(h) = 0.7Exp_{56}(h), C_{22}(h) = Exp_{56}(h)$$

- MM2-B:

$$C_{11}(h) = 0.3nug(h) + Exp_{40}(h), C_{12}(h) = 0.21nug(h) + 0.7Exp_{56}(h), C_{22}(h) = 0.3nug(h) + Exp_{56}(h)$$

- Complex case-A:

$$C_{11}(h) = 0.3nug(h) + Exp_{56}(h), C_{12}(h) = 0.7Exp_{56}(h), C_{22}(h) = Exp_{40}(h)$$

- Complex case-B:

$$C_{11}(h) = Exp_{56}(h), C_{12}(h) = 0.7Exp_{40}(h), C_{22}(h) = 0.3nug(h) + Exp_{40}(h).$$

In the Markov-type models, the cross-covariance is proportional to the direct covariance of the primary variable (MM1-A and MM1-B) or of the secondary variable (MM2-A and MM2-B). This is no longer the case in the last two models (complex cases), where the cross-covariance is continuous (no nugget effect), while the direct covariances have a different correlation range or the same range and a nugget effect.

4.2 Results

Figure 3 shows the distributions (through box plots) of the variances of the prediction errors over the 3400 target grid nodes for the aforementioned six models and six search

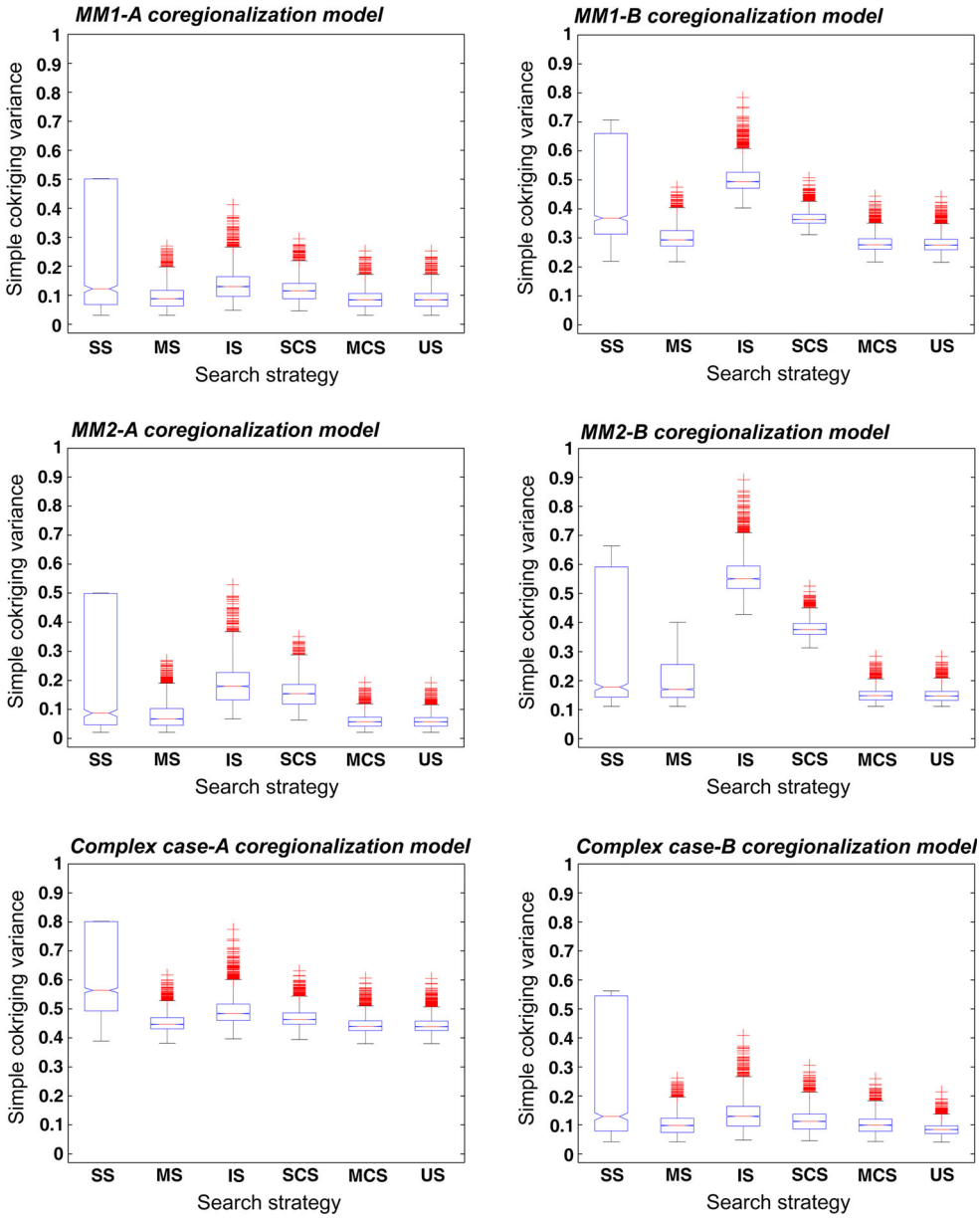

Fig. 3 Box plots representing the distribution of the simple cokriging error variances at grid nodes with no primary data, for different coregionalization models and search strategies (primary data from the sampling design in Fig. 2a)

strategies. The unique search neighborhood can be treated as the reference against which to compare the results, insofar as it corresponds to the best possible prediction (no data discarded). In all the cases, the multi-collocated search yields almost the same variance distribution as the reference distribution, while the isotopic and single searches provide the poorest results (highest variances), followed by the strictly collocated and the multiple searches, the latter being the one that delivers results closest to the multi-collocated search. Accordingly, in this example, the search strategies can be ordered from the best to the worst, based upon their closeness to the reference (unique neighborhood), as follows: multi-collocated, multiple, strictly collocated, isotopic and single searches. In order to assess the significance of the gain or loss in the error variance, Table 2 gives the average variance over the 3400 non-sampled grid nodes for the different coregionalization models and search strategies under consideration. It is seen that, with respect to the unique search (US), the average variance increases between 36.00 and 226.59% with the single search (SS) (i.e., the average variance obtained with SS is between 136.00 and 326.59% times the average variance obtained with US), between 2.05 and 31.32% with the multiple search (MS), between 11.00 and 273.68% with the isotopic search (IS), between 5.60 and 161.57% with the strictly collocated search (SCS), and between 0.08 and 19.74% with the multi-collocated search (MCS).

Table 2 Average error variance over 3400 target grid nodes, for each coregionalization model and search strategy (200 sampling locations, simple cokriging)

Coregionalization model	Search strategy	Average error variance	Percentage of average error variance obtained with unique search
MM1-A	SS	0.2147	246.54
	MS	0.0941	108.05
	IS	0.1351	155.22
	SCS	0.1180	135.51
	MCS	0.0873	100.23
	US	0.0871	100.00
MM1-B	SS	0.4291	153.35
	MS	0.3002	107.28
	IS	0.5030	179.74
	SCS	0.3676	131.37
	MCS	0.2817	100.66
	US	0.2798	100.00
MM2-A	SS	0.1949	326.59
	MS	0.0784	131.32
	IS	0.1872	313.76
	SCS	0.1561	261.57
	MCS	0.0600	100.55
	US	0.0597	100.00
MM2-B	SS	0.2899	193.24
	MS	0.1936	129.03
	IS	0.5606	373.68
	SCS	0.3798	253.20
	MCS	0.1508	100.54
	US	0.1500	100.00
Complex case-A	SS	0.6036	136.00
	MS	0.4529	102.05
	IS	0.4927	111.00
	SCS	0.4687	105.60
	MCS	0.4442	100.08
	US	0.4439	100.00
Complex case-B	SS	0.2341	275.68
	MS	0.1020	120.10
	IS	0.1351	159.06
	SCS	0.1153	135.74
	MCS	0.1017	119.74
	US	0.0849	100.00

4.3 Sensitization

To investigate whether or not the previous ordering is sensitive to the chosen cokriging type and sampling design, the same experiment is repeated with the following modifications:

1. choosing another design of 200 randomly selected sampling locations (Fig. 2b);
2. choosing a design of 100 randomly selected sampling locations (Fig. 2c);
3. choosing a design of 400 randomly selected sampling locations (Fig. 2d);
4. keeping the original design of 200 sampling locations (Fig. 2a) and substituting ordinary cokriging (cokriging with unknown mean values) for simple cokriging.

In all the cases, the ordering of the search strategies from best to worst remain unchanged: multi-collocated, multiple, strictly collocated, isotopic and single searches (Figs. 4, 5, 6, 7). Globally, the error variance increases when fewer data are available (case of 100 sampling locations) and decreases when more data are available

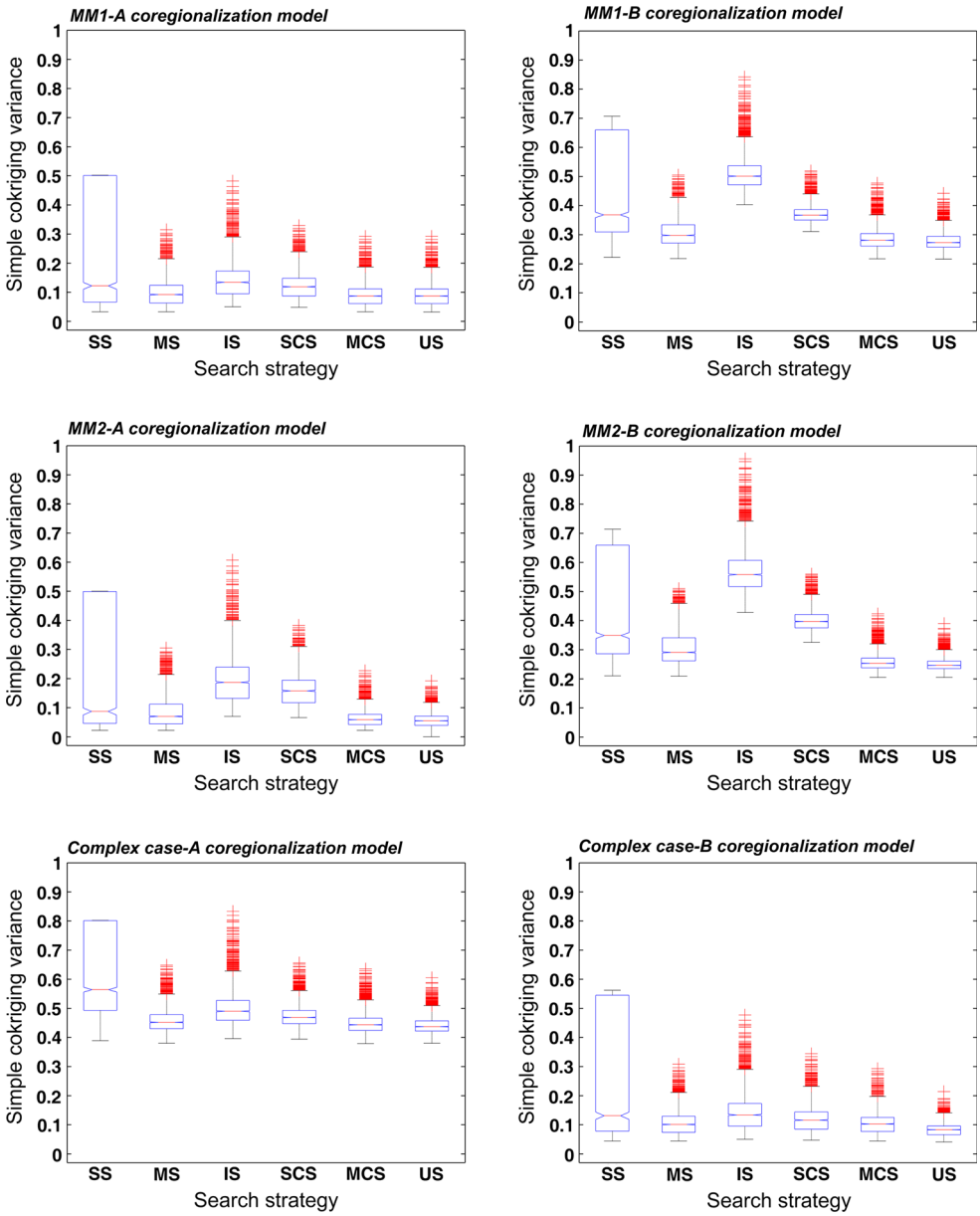

Fig. 4 Box plots representing the distribution of the simple cokriging error variances at grid nodes with no primary data, for different coregionalization models and search strategies (primary data from the sampling design in Fig. 2b)

(case of 400 sampling locations). However, in both cases, the orders of magnitude of the increase with respect to the unique search (US) (Tables 3, 4) are comparable to the case of 200 sampling locations (Table 2). The single search (SS) provides the distribution of error variance with the highest spread, including an unbounded distribution in the case of ordinary cokriging. This is explained because ordinary cokriging fails when no primary data is selected in the cokriging neighborhood, which frequently happens with SS (formally, the cokriging variance is infinite in such a situation).

4.4 Discussion

The previous subsections tested six search strategies under six coregionalization models, four sampling designs and two cokriging types (simple and ordinary), providing some generality to the classification of the search strategies in terms of efficiency (how much decreases or increases the variance of the cokriging error by selecting one or another search strategy).

When using an inappropriate search strategy such as SS or IS, the loss of precision is considerable in some configurations of the target and sampling locations, yielding an error variance that can be twice or three times greater than

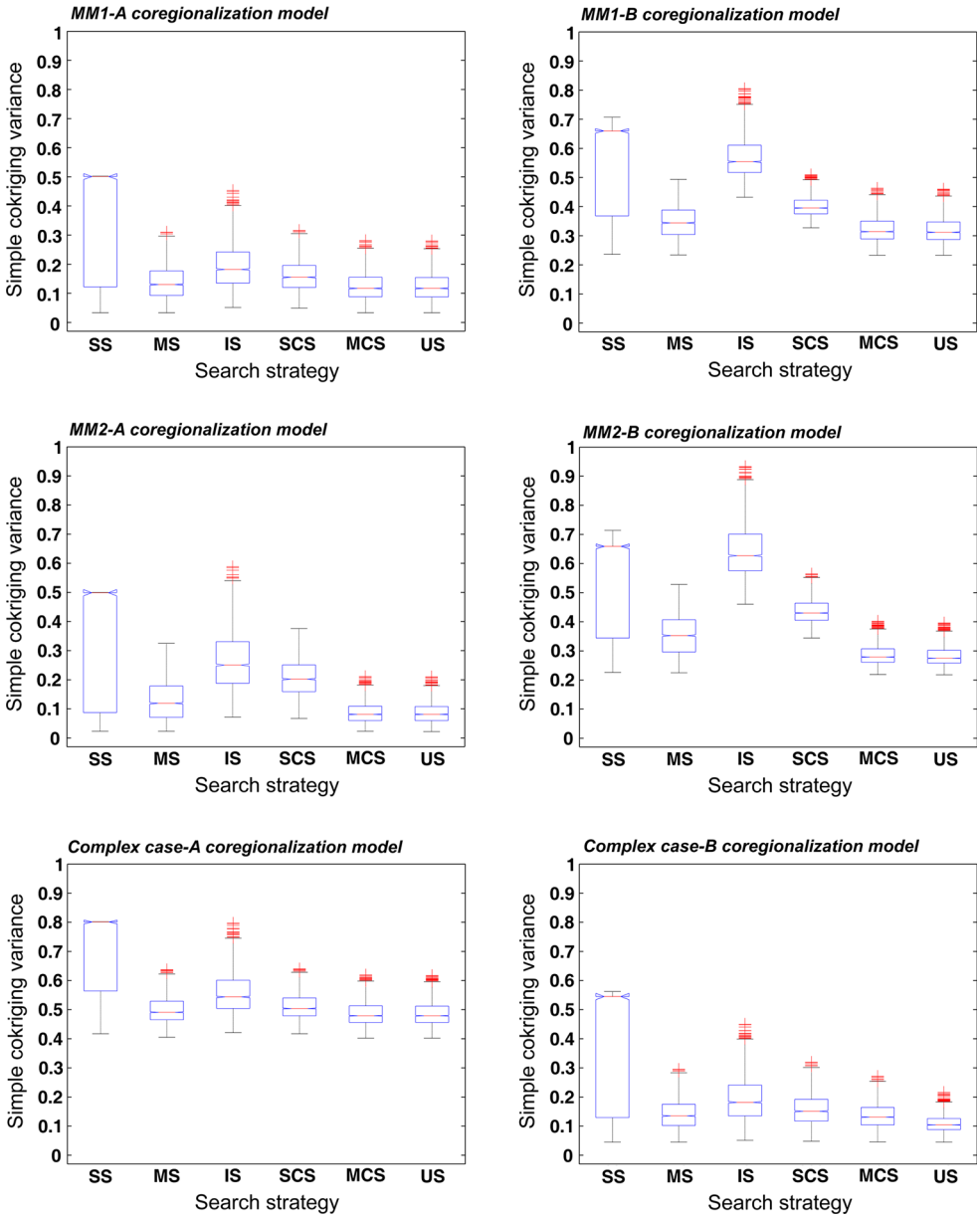

Fig. 5 Box plots representing the distribution of the simple cokriging error variances at grid nodes with no primary data, for different coregionalization models and search strategies (primary data from the sampling design in Fig. 2c)

the error variance obtained with optimal or sub-optimal search strategies such as MCS or MS. It is noteworthy that MCS consistently yields a better precision (lower error variance) than MS, although the secondary data selected with MS are closer to the target location than the secondary data selected with MCS. This suggests the importance of selecting secondary data located at (or around) the same positions as the primary data, in order to better "calibrate" the secondary information to the primary one.

In practice, in the presence of an exhaustively known secondary variable, the implementation of MCS takes as much computational time as that of IS (except for the target location, the selected secondary data are located at the same points as the selected primary data), while MS is more demanding, insofar as two searches are needed, one for the primary data (similar to MCS or IS) and the other one for the secondary data (similar to SS). However, MS is still applicable when the secondary variable is not exhaustively known and therefore turns out to be particularly interesting in cases of heterotopic sampling designs with an under-sampled primary variable.

Given the current computational capacities, the extra time needed in using an improved search strategy (MS or MCS) is generally not a bottleneck in the application of cokriging. In contrast, it is often critical to obtain the lowest possible error variance. Indeed, due to the

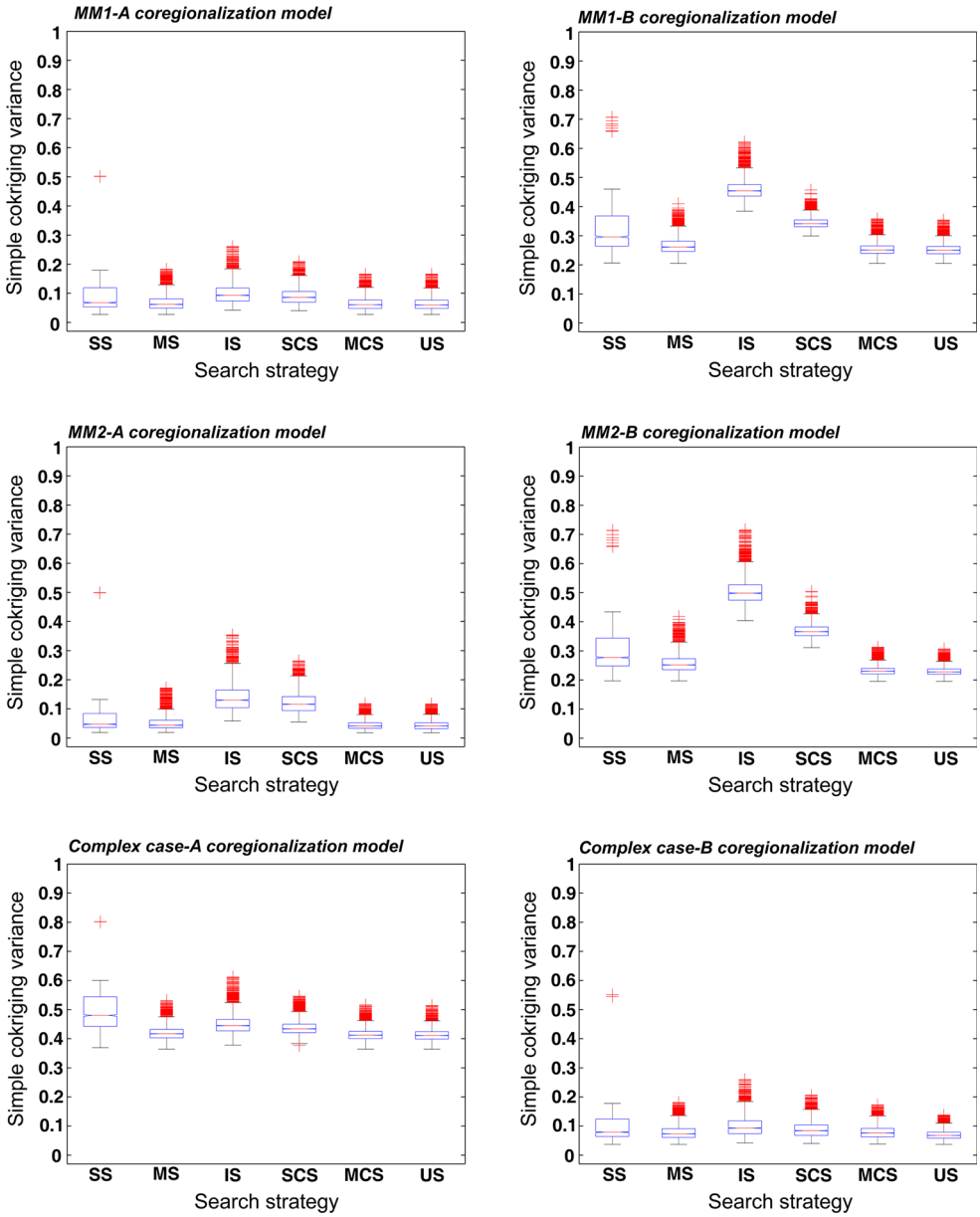

Fig. 6 Box plots representing the distribution of the simple cokriging error variances at grid nodes with no primary data, for different coregionalization models and search strategies (primary data from the sampling design in Fig. 2d)

orthogonality relationship between the simple cokriging predictor and the simple cokriging error (Chilès and Delfiner 2012), the variance of the primary variable is the sum of the variance of the predictor and the variance of the prediction error. Accordingly, in addition to an increase of the predictor precision, a reduction of the error variance implies an increase in the variance of the predictor, i.e., a decrease of the smoothing effect of cokriging. Furthermore, due to error propagations, cokriging with a moving neighborhood can be problematic when it is used in iterative simulation algorithms, such as sequential Gaussian cosimulation (Emery and Peláez 2011) or Gibbs sampling (Emery et al. 2014), reason for which the design of an efficient (optimal or sub-optimal) search strategy is essential.

As a last remark, the effect of parameter misspecification has been ignored in all the previous experiments. In practice, estimated mean values for the primary and secondary variables and an estimated coregionalization model, naively assumed known without error, are used in cokriging. A misspecification of the mean values can strongly affect the cokriging predictions, but it has no impact on the

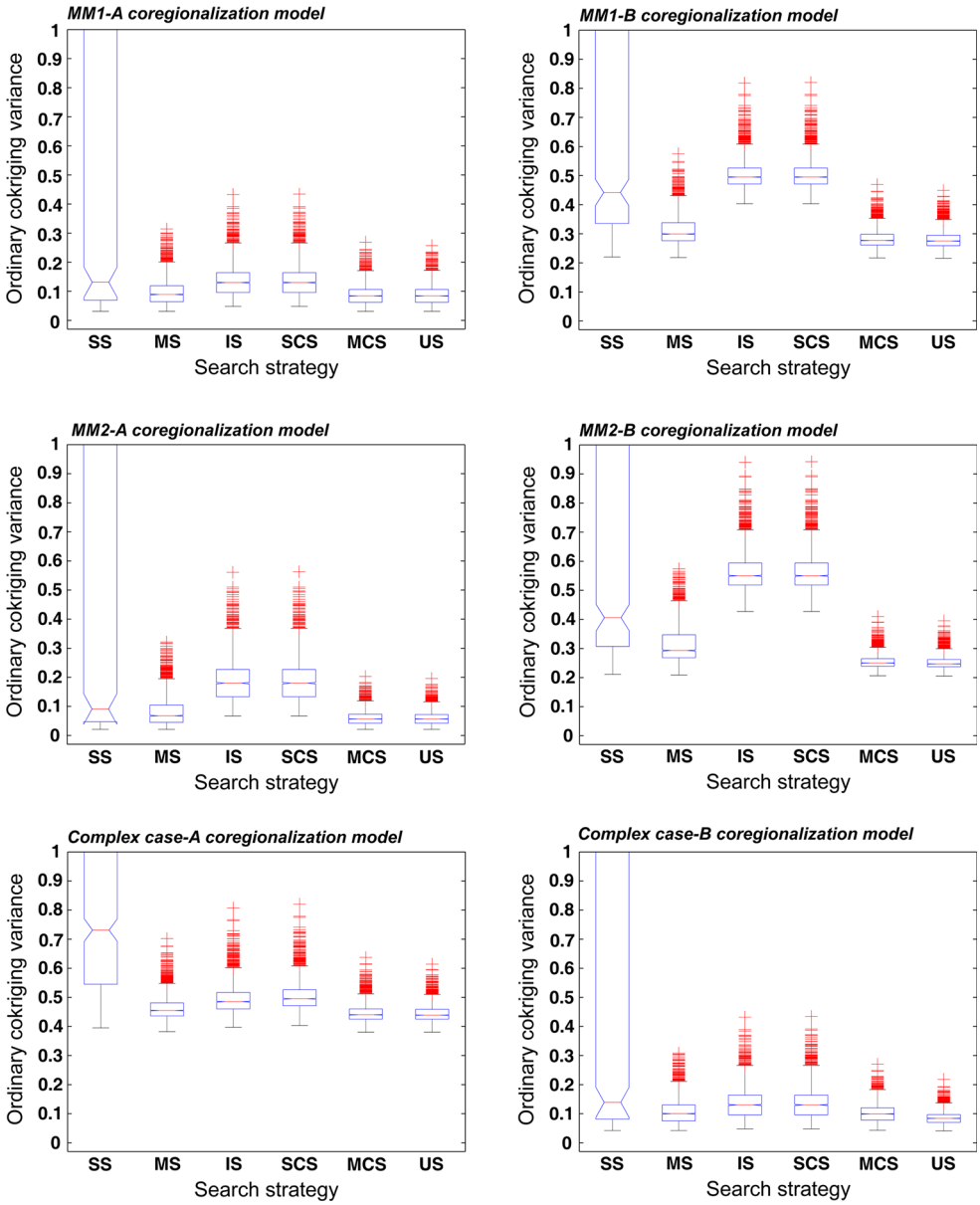

Fig. 7 Box plots representing the distribution of the ordinary cokriging error variances at grid nodes with no primary data, for different coregionalization models and search strategies (primary data from the sampling design in Fig. 2a)

calculated error variances (Eqs. 1–8), so that the design of the optimal or sub-optimal search strategy remains unchanged; to avoid biased predictions, simple cokriging should be substituted for ordinary cokriging. In contrast, a misspecification of the coregionalization model can have a significant impact on the calculated error variances. The reader is referred to Chilès and Delfiner (2012) and references therein for a discussion on some alternatives to traditional cokriging in the presence of uncertainty in the covariance parameters. Irrespective of the chosen alternative, the use of a moving neighborhood for local predictions yields an additional loss of precision (increase of the error variance) with respect to the unique search implementation, so that the results presented in the previous subsections are still of interest.

5 Conclusions

Cokriging is a widely used technique in spatial prediction problems. Its implementation becomes prohibitive when too many data are available, but the screening effect approximation may allow one to reduce the number of primary and/or secondary data without much loss of precision in the prediction. The best selection depends not only on the geometrical configuration of the data, but also

Table 3 Average error variance over 3500 target grid nodes, for each coregionalization model and search strategy (100 sampling locations, simple cokriging)

Coregionalization model	Search strategy	Average error variance	Percentage of average error variance obtained with unique search
MM1-A	SS	0.3270	265.43
	MS	0.1372	111.34
	IS	0.1923	156.11
	SCS	0.1590	129.07
	MCS	0.1235	100.22
	US	0.1232	100.00
MM1-B	SS	0.5243	164.31
	MS	0.3481	109.07
	IS	0.5675	177.84
	SCS	0.4003	125.45
	MCS	0.3217	100.81
	US	0.3191	100.00
MM2-A	SS	0.3129	364.82
	MS	0.1280	149.30
	IS	0.2638	307.54
	SCS	0.2055	239.61
	MCS	0.0862	100.48
	US	0.0858	100.00
MM2-B	SS	0.5149	182.73
	MS	0.3543	125.75
	IS	0.6424	227.97
	SCS	0.4356	154.59
	MCS	0.2856	101.35
	US	0.2818	100.00
Complex case-A	SS	0.6864	141.33
	MS	0.4982	102.59
	IS	0.5564	114.57
	SCS	0.5106	105.13
	MCS	0.4862	100.11
	US	0.4857	100.00
Complex case-B	SS	0.3557	331.51
	MS	0.1398	130.31
	IS	0.1921	179.05
	SCS	0.1555	144.88
	MCS	0.1351	125.94
	US	0.1073	100.00

on the spatial correlation structure of the primary and secondary variables.

Through the analyzed examples, it appears that multi-collocated cokriging coincides with full cokriging in the case of a Markov-type (MM2) model and does not deviate too much from it under the other spatial correlation models. An alternative, although providing slightly less precise predictions, is cokriging with a multiple search strategy, where the closest data of each variable are selected for the prediction. The good performances of these two strategies (multi-collocated and multiple) indicate that it is good practice to (1) select the primary data closest to the target location, (2) select the secondary data closest to the target location, and (3) select the secondary data located at (or around) the locations of the selected primary data. A strategy fulfilling these three criteria allows incorporating

Table 4 Average error variance over 3200 target grid nodes, for each coregionalization model and search strategy (400 sampling locations, simple cokriging)

Coregionalization model	Search strategy	Average error variance	Percentage of average error variance obtained with unique search
MM1-A	SS	0.1159	179.28
	MS	0.0677	104.70
	IS	0.1002	154.99
	SCS	0.0905	140.07
	MCS	0.0649	100.35
	US	0.0646	100.00
MM1-B	SS	0.3337	131.75
	MS	0.2661	105.09
	IS	0.4599	181.60
	SCS	0.3444	135.99
	MCS	0.2548	100.60
	US	0.2532	100.00
MM2-A	SS	0.0950	216.64
	MS	0.0511	116.57
	IS	0.1395	317.89
	SCS	0.1215	277.03
	MCS	0.0443	100.89
	US	0.0439	100.00
MM2-B	SS	0.3170	137.45
	MS	0.2580	111.86
	IS	0.5054	219.12
	SCS	0.3690	159.97
	MCS	0.2328	100.91
	US	0.2307	100.00
Complex case-A	SS	0.5130	123.85
	MS	0.4203	101.47
	IS	0.4504	108.74
	SCS	0.4379	105.73
	MCS	0.4149	100.15
	US	0.4142	100.00
Complex case-B	SS	0.1290	184.36
	MS	0.0780	111.52
	IS	0.1002	143.18
	SCS	0.0887	126.74
	MCS	0.0795	113.58
	US	0.0700	100.00

the most relevant information (according to geographical distance to the target location), while calibrating the secondary information to the primary one. In contrast, strictly collocated cokriging, which omits the secondary data except at the target location, is significantly poorer, indicating that the discarded secondary data (especially the ones at the primary data locations) have a strong influence on the prediction precision. Cokriging based on a single search or on an isotopic search strategy, which does not differentiate the primary and secondary sampling designs, yields the poorest results and should be avoided in case of a heterotopic sampling design.

Acknowledgements The first author acknowledges the Nazarbayev University for funding this work via "Faculty development competitive research Grants for 2018–2020" under Contract No. 090118FD5336. The second author acknowledges the Chilean Commission for Scientific and Technological Research (CONICYT), through Grant CONICYT PIA Anillo ACT1407.

References

Ahmed S, de Marsily G (1987) Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity. Water Resour Res 23(9):1717–1737

Almeida AS, Journel AG (1994) Joint simulation of multiple variables with a Markov-type coregionalization model. Math Geol 26(5):565–588

Babak O, Deutsch CV (2009) Collocated cokriging based on merged secondary attributes. Math Geosci 41:921–926

Boezio MNM, Costa JFCL, Koppe JC (2006) Kriging with an external drift versus collocated cokriging for water table mapping. Trans Inst Min Metall Sect B Appl Earth Sci 115(3):103–112

Bohorquez M, Giraldo R, Mateu J (2017) Multivariate functional random fields: prediction and optimal sampling. Stoch Env Res Risk Assess 31(1):53–70

Borkowski AS, Kwiatkowska-Malina J (2017) Geostatistical modelling as an assessment tool of soil pollution based on deposition from atmospheric air. Geosci J 21(4):645–653

Cao R, Ma YZ, Gomez E (2014) Geostatistical applications in petroleum reservoir modelling. J South Afr Inst Min Metall 114(8):625–629

Chilès JP, Delfiner P (2012) Geostatistics: modeling spatial uncertainty. Wiley, New York

Cornah A, Machaka E (2015) Integration of imprecise and biased data into mineral resource estimates. J South Afr Inst Min Metall 115(6):523–530

D'Agostino V, Greene EA, Passarella G, Vurro M (1998) Spatial and temporal study of nitrate concentration in groundwater by means of coregionalization. Environ Geol 36(3–4):285–295

da Silva CZ, Costa JF (2014) Minimum/maximum autocorrelation factors applied to grade estimation. Stoch Env Res Risk Assess 28(8):1929–1938

D'Agostino V, Passarella G, Vurro M (1997) Assessment of the optimal sampling arrangement based on cokriging estimation variance reduction approach. In: Holly FM, Alsoffar A (eds) Water for a changing global community, proceedings of the 27th congress of the international association for hydraulic research forrest. American Society of Civil Engineers, pp 246–252

Dalla Libera N, Fabbri P, Mason L, Piccinini L, Pola M (2017) Geostatistics as a tool to improve the natural background level definition: an application in groundwater. Sci Total Environ 598:330–340

Emery X (2009) The kriging update equations and their application to the selection of neighboring data. Comput Geosci 13(3):269–280

Emery X (2010) Iterative algorithms for fitting a linear model of coregionalization. Comput Geosci 36(9):1150–1160

Emery X (2012) Cokriging random fields with means related by known linear combinations. Comput Geosci 38(1):136–144

Emery X, Peláez M (2011) Assessing the accuracy of sequential Gaussian simulation and cosimulation. Comput Geosci 15(4):673–689

Emery X, Arroyo D, Peláez M (2014) Simulating large Gaussian random vectors subject to inequality constraints by Gibbs sampling. Math Geosci 46(3):265–283

Fabijańczyk P, Zawadzki J, Magiera T, Szuszkiewicz M (2016) A methodology of integration of magnetometric and geochemical soil contamination measurements. Geoderma 277:51–60

Fouedjio F (2018) A fully non-stationary linear coregionalization model for multivariate random fields. Stoch Env Res Risk Assess 32(6):1699–1721

Gálvez I, Emery X (2011) Multivariate resources modelling: which data are relevant for cokriging? In: Beniscelli J, Kuyvenhoven R, Hoal KO (eds) Proceedings of the 2nd international seminar on geology for the mining industry. Gecamin Ltda, Santiago, Chile, pp 10

Gneiting T, Kleiber W, Schlather M (2010) Matérn cross-covariance functions for multivariate random fields. J Am Stat Assoc 105(491):1167–1177

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

Goulard M, Voltz M (1992) Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. Math Geol 24(3):269–286

Hohn ME (1999) Geostatistics and petroleum geology, 2nd edn. Kluwer Academic, Dordrecht

Jalalalhosseini SM, Ali H, Mostafazadeh M (2014) Predicting porosity by using seismic multi-attributes and well data and combining these available data by geostatistical methods in a South Iranian oil field. Pet Sci Technol 32(1):29–37

Journel AG (1999) Markov models for cross-covariances. Math Geol 31(8):955–964

Journel AG, Huijbregts CJ (1978) Mining geostatistics. Academic Press, London

Kitanidis PK (1997) Introduction to geostatistics: applications to hydrology. Cambridge University Press, London

Lark RM, Ander EL, Cave MR, Knights KV, Glennon MM, Scanlon RP (2014) Mapping trace element deficiency by cokriging from regional geochemical soil data: a case study on cobalt for grazing sheep in Ireland. Geoderma 226–227(1):64–78

Masihi A, Zarei M (2010) Permeability modeling using ANN and collocated cokriging. In: 72nd European association of geoscientists and engineers conference and exhibition 2010: a new spring for geoscience. Incorporating SPE EUROPEC 2010. European Association of Geoscientists and Engineers (EAGE), vol 5, pp 3939–3943

Minnitt RCA, Deutsch CV (2014) Cokriging for optimal mineral resource estimates in mining operations. J South Afr Inst Min Metall 114(3):189–203

Myers DE (1982) Matrix formulation of cokriging. Math Geol 14(3):249–257

Olea RA, Raju NJ, Egozcue JJ, Pawlowsky-Glahn V, Singh S (2018) Advancements in hydrochemistry mapping: methods and application to groundwater arsenic and iron concentrations in Varanasi, Uttar Pradesh, India. Stoch Env Res Risk Assess 32(1):241–259

Pan G, Gaard D, Moss K, Heiner T (1993) A comparison between cokriging and ordinary kriging: case study with a polymetallic deposit. Math Geol 25(3):377–398

Pawlowsky-Glahn V, Egozcue JJ, Olea RA, Pardo-Igúzquiza E (2015) Cokriging of compositional balances including a dimension reduction and retrieval of original units. J South Afr Inst Min Metall 115(1):59–72

Rivoirard J (2001) Which models for collocated cokriging? Math Geol 33(2):117–131

Rivoirard J (2004) On some simplifications of the cokriging neighborhood. Math Geol 36(8):899–915

Roberts BL, McKenna SA (2009) The use of secondary information in geostatistical target area identification. Stoch Env Res Risk Assess 23(2):227–236

Schwab AM, Buckner S, Bramald JA, Cass J (2011) Improving reservoir modeling through integration of seismic data in eocene turbidites for West Brae field, central North Sea, United Kingdom. AAPG Mem 96:107–119

Stein A, Van Dooremolen W, Bouma J, Bregt AK (1988) Cokriging point data on moisture deficit. Soil Sci Soc Am J 52(5):1418–1423

Subramanyam A, Pandalai HS (2004) On the equivalence of the cokriging and kriging systems. Math Geol 36(4):507–523

Subramanyam A, Pandalai HS (2008) Data configurations and the cokriging system: simplification by screen effects. Math Geosci 40(4):425–443

Tolosana-Delgado R, van den Boogaart KG (2013) Joint consistent mapping of high-dimensional geochemical surveys. Math Geosci 45(8):983–1004

Uygucgil H, Konuk A (2015) Reserve estimation in multivariate mineral deposits using geostatistics and GIS. J Min Sci 51(5):993–1000

Vargas-Guzmán J, Jim Yeh TC (1999) Sequential kriging and cokriging: two powerful geostatistical approaches. Stoch Env Res Risk Assess 13(6):416–435

Wackernagel H (1988) Geostatistical techniques for interpreting multivariate spatial information. In: Chung CF, Fabbri AG, Sinding-Larsen R (eds) Quantitative analysis of mineral and energy resources. Reidel, Dordrecht, pp 393–409

Wackernagel H (2003) Multivariate Geostatistics: an introduction with applications. Springer, Berlin

Xu W, Tran TT, Srivastava RM, Journel AG (1992) Integrating seismic data in reservoir modeling: the collocated cokriging alternative. In: 67th SPE annual technical conference and exhibition. Society of Petroleum Engineers, SPE paper 24742, pp 833–842

Yates SR, Warrick AW (1987) Estimating soil water content using cokriging. Soil Sci Soc Am J 51:23–30