ELSEVIER

Journal of Statistical Planning and Inference 112 (2003) 3–41

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Prediction of nonlinear spatial functionals

Jeremy Aldworth[a], Noel Cressie[b,*]

[a] *IDEC Pharmaceuticals Corporation, San Diego, CA 92121, USA*
[b] *Department of Statistics, The Ohio State University, 1958 Neil Avenue 404 Cockins Hall, Columbus, OH 43210-1247, USA*

Abstract

Spatial statistical methodology can be useful in the arena of environmental regulation. Some regulatory questions may be addressed by predicting linear functionals of the underlying signal, but other questions may require the prediction of nonlinear functionals of the signal. For example, in order to be in regulatory compliance, air-pollution levels have to fall within specified limits over some geographic region; whether or not they are in compliance and where they are out of compliance are nonlinear functionals. We propose a spatial empirical Bayes model for environmental data collected over a region.

Further, we propose a predictor, based on the kriging methodology with extra constraints, that implies useful unbiasedness properties in predicting nonlinear spatial functionals. This predictor, called covariance-matching constrained kriging, is an optimal linear predictor that matches not only first moments but second moments (including specified covariances) as well.
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Empirical Bayes; Covariance-matching constrained kriging; Geostatistics; Measurement error; Particulate matter

1. Introduction

The motivation for the spatial statistical methodology presented in this paper comes from problems commonly encountered in the arena of environmental regulation. For example, the US Environmental Protection Agency (EPA) has a federal mandate to identify air pollutants injurious to public health and then to propose air pollution "standards" (i.e., threshold levels of the pollutants) that should not be exceeded. A major component of air pollution is particulate matter (PM), and this is categorized by size. Consequently, PM_{10} refers to the category of particulate matter in which

* Corresponding author. Tel.: +1-614-292-5194; fax: +1-614-292-2096.
E-mail address: ncressie@stat.ohio-state.edu (N. Cressie).

0378-3758/03/$ - see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S0378-3758(02)00321-X

particles do not exceed 10 μm in aerodynamic diameter. The PM_{10} measurement is an indicator for particles that penetrate to the tracheobronchial and gas-exchange regions of the lung (US EPA, 1996, p. IV-3a). The EPA has adopted an annual average daily PM_{10} standard in the range 40–50 $\mu g/m^3$ and a 24-hour standard of 150 $\mu g/m^3$. This means that if the PM_{10} level of some region (e.g., the city of Pittsburgh, PA) exceeds the standard over the appropriate time period, then that region is deemed to be out of compliance with federal regulations. The problem here is how to predict reliably whether or not the PM_{10} level exceeds the standard, based on observations obtained from a small number of monitoring stations in the region of interest. Now, it is entirely possible that, even if the region of study is in PM_{10} compliance, there may be local neighborhoods that are not (e.g., neighborhoods downwind of pollution-emitting factories). Ideally, one would like to identify such neighborhoods based on the data available. Other quantities to predict might be the *proportion* of the entire region not in compliance or the average amount of PM_{10} in the neighborhoods out of compliance, as compared to the average amount in the neighborhoods in compliance. Such predictands are clearly *nonlinear* functionals of the study variable, here the PM_{10} level throughout the region of study.

Let us now return to a generic spatial statistical set-up, where we consider some study variable that varies over a given spatial domain (e.g., PM_{10} levels over Pittsburgh), and suppose that measurements associated with it are taken at selected locations in the domain. We are interested in the prediction of some measurable functional of the study variable, based on the observed data, but it is unlikely that the study variable can be measured exactly (i.e., without measurement error). Notice that in much of the geostatistics literature there is an implicit assumption that observations are exact measurements of the study variable in question (e.g., Journel and Huijbregts, 1978; Rivoirard, 1994); a geostatistical model in which a measurement-error component is explicitly included is given by Cressie (1988) and, as we show in this paper, it can be viewed as a spatial empirical Bayes model.

Consider the prediction of *linear* functionals of the study variable, such as its prediction at some unsampled location, or the prediction of its spatial mean over some subregion in the domain (e.g., Matheron, 1963; Journel and Huijbregts, 1978; Cressie, 1993a). Ordinary-kriging and universal-kriging predictors are natural choices for predicting such functionals. They are linear in the data, simple to construct (e.g., Journel and Huijbregts, 1978, Chapter V; Cressie, 1993a, Chapter 3), and they can be extended easily to the case where the data are multivariate (e.g., Myers, 1982; Ver Hoef and Cressie, 1993). In addition, these linear kriging predictors easily handle additive measurement error; the unbiasedness of the linear predictors remains unchanged in the presence of measurement error and the mean-squared prediction error simply includes the measurement-error variance as an additive component (Cressie, 1988).

When predicting *nonlinear* functionals of the study variable (or linear functionals of a *non-Gaussian* study variable), the major thrust in geostatistics has been to use nonlinear predictors, such as indicator kriging (Journel, 1983), indicator cokriging (Lajaunie, 1990), and disjunctive kriging (Matheron, 1976). However, when a measurement-error component with positive variance is modeled (here using empirical Bayes), these nonlinear predictors are no longer unbiased and can perform poorly (Aldworth, 1998).

Furthermore, they are not well suited to prediction of the spatial mean over some subregion of the domain (Cressie, 1993a, Section 5.2).

Another approach to the problem of prediction of nonlinear spatial functionals is *conditional simulation*, which is a simulation algorithm originally developed by Journel (1974) that yields more realistic realizations of the spatial process of interest than the smoothed predictions obtained from the linear kriging predictors. Later manifestations of conditional simulation have shown up in the statistical literature in the form of Markov chain Monte Carlo simulation of spatial processes (e.g., Besag and Green, 1993; Diggle et al., 1998). Predictors based upon a single realization of a conditional simulation may have poor optimality properties but, by the law of large numbers, these properties will improve for predictors based on averages of many realizations of a conditional simulation (provided that all relevant parameters have been correctly specified). It should be noted that conditional simulation can be viewed as an empirical Bayes Monte Carlo procedure. Such a procedure requires the joint distribution of all random components to be (implicitly or explicitly) specified, and known values of, or estimates of, deterministic parameters to be specified. It also requires large data storage and computer resources.

In this paper, we propose a new kriging predictor, in the spirit of the linear and second-order moment nature of other kriging predictors, but with constraints forcing elements of the variance matrix of a vector of linear predictors to match those from the corresponding predictands. This predictor, which we term covariance-matching constrained kriging (CM), has useful unbiasedness properties, approximate second-order optimality properties, handles additive measurement error straightforwardly, and can predict spatial means over subregions just as easily as at points. It is a multivariate generalization of the *constrained kriging predictor* (Cressie, 1993b), where the variance of a scalar-valued linear predictor is constrained to match that of its corresponding scalar-valued predictand.

The underlying spatial empirical Bayes model (consisting of a spatial signal contaminated with measurement error) is presented in Section 2, and Section 3 discusses prediction of nonlinear functionals of the signal. The CM methodology is presented in Section 4. In Section 5, a computer-simulation experiment is conducted to explore the properties of the different predictors, and Section 6 illustrates the CM methodology on PM_{10} data in the Pittsburgh area. Finally, Section 7 contains discussion and conclusions.

2. Spatial empirical Bayes Model

A *spatial process* is a real- (or vector-) valued stochastic process

$$Z(\cdot) \equiv \{Z(\mathbf{s}): \mathbf{s} \in D\},$$

where $D \subset \mathbb{R}^d$. In this paper we consider only real-valued $Z(\cdot)$.

Suppose that *data* from the spatial process $Z(\cdot)$ consist of the observations $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$ taken at the locations in $A \equiv \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset D$, where A is called a *sample* (of spatial locations). We assume that the sample has been specified somehow and so we

treat the locations as fixed. Readers interested in how randomness in A can be considered within the context of geostatistics might consult the review given in Aldworth and Cressie (1999).

The observations are assumed to have been contaminated with additive measurement error, according to the following model:

$$Z(\mathbf{s}) = S(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad \mathbf{s} \in D, \tag{2.1}$$

where $S(\cdot)$ represents the spatial process of fundamental interest, namely the "signal", and $\varepsilon(\cdot)$ is a zero-mean, white-noise process representing the measurement error. The signal and noise components are assumed independent. This is an empirical Bayes model because, conditional on $\{S(\mathbf{s})\colon \mathbf{s} \in D\}$, (2.1) can equivalently be written as $\Pr(Z(\mathbf{s}) \leqslant z | S(\mathbf{s})) = \Pr(\varepsilon(\mathbf{s}) \leqslant z - S(\mathbf{s}))$; independent for $\mathbf{s} \in D$. Any parameters associated with specification of the distribution of $S(\cdot)$ and $\varepsilon(\cdot)$ are to be estimated.

Now we specify the distribution of the signal process $S(\cdot)$. This can be viewed as a prior distribution; we write

$$S(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \quad \mathbf{s} \in D, \tag{2.2}$$

where $\mu(\cdot)$ is the large-scale, deterministic, mean structure of the signal (i.e., $\mu(\cdot) \equiv E[S(\cdot)]$ is the trend), and $\delta(\cdot)$ is the small-scale stochastic structure that models the spatial dependence. That is

$$E[\delta(\mathbf{s})] = 0, \quad \mathbf{s} \in D, \tag{2.3}$$

$$\operatorname{cov}[\delta(\mathbf{s}), \delta(\mathbf{u})] \equiv C(\mathbf{s}, \mathbf{u}), \quad \mathbf{s}, \mathbf{u} \in D, \tag{2.4}$$

where $C(\cdot,\cdot)$ is the *covariance function* of $\delta(\cdot)$ (equivalently, of $S(\cdot)$). If we denote

$$\operatorname{var}[\varepsilon(\mathbf{s})] \equiv \tau^2, \quad \mathbf{s} \in D, \tag{2.5}$$

then we can easily calculate the spatial covariance function of the observed "noisy" process,

$$C_z(\mathbf{s}, \mathbf{u}) \equiv \operatorname{cov}[Z(\mathbf{s}), Z(\mathbf{u})], \quad \mathbf{s}, \mathbf{u} \in D$$

to be

$$C_z(\mathbf{s}, \mathbf{u}) = \begin{cases} C(\mathbf{s}, \mathbf{u}) + \tau^2, & \mathbf{s} = \mathbf{u}, \\ C(\mathbf{s}, \mathbf{u}), & \mathbf{s} \neq \mathbf{u}. \end{cases}$$

Another useful measure of the spatial dependence among the data is the *variogram*, defined as

$$2\gamma_z(\mathbf{s}, \mathbf{u}) \equiv \operatorname{var}[Z(\mathbf{s}) - Z(\mathbf{u})] = C_z(\mathbf{s}, \mathbf{s}) + C_z(\mathbf{u}, \mathbf{u}) - 2C_z(\mathbf{s}, \mathbf{u}), \quad \mathbf{s}, \mathbf{u} \in D.$$

When $\mathbf{s} \neq \mathbf{u}$,

$$2\gamma_z(\mathbf{s}, \mathbf{u}) = 2(\gamma(\mathbf{s}, \mathbf{u}) + \tau^2),$$

where $2\gamma(\cdot,\cdot)$ is the variogram of $\delta(\cdot)$ (equivalently, of $S(\cdot)$). When $\mathbf{s} = \mathbf{u}$, $2\gamma_z(\mathbf{s}, \mathbf{u}) = 2\gamma(\mathbf{s}, \mathbf{u}) = 0$. Note that if $\tau^2 = 0$, then $C(\cdot,\cdot) \equiv C_z(\cdot,\cdot)$ and $2\gamma(\cdot,\cdot) \equiv 2\gamma_z(\cdot,\cdot)$. The quantities $\gamma_z(\cdot,\cdot)$ and $\gamma(\cdot,\cdot)$ are called the *semivariograms* of $Z(\cdot)$ and $S(\cdot)$, respectively.

Suppose that replicate observations are taken at one or more of the sites in A. Then the spatial model (2.1) can be generalized to accommodate the replicate observations as follows:

$$Z_i(\mathbf{s}_j) = S(\mathbf{s}_j) + \varepsilon_i(\mathbf{s}_j), \quad i = 1,\ldots,n_j,\ j = 1,\ldots,J,\ \mathbf{s}_1,\ldots,\mathbf{s}_J \in D, \tag{2.6}$$

where $\{\varepsilon_i(\cdot)\colon i = 1,\ldots,\max_j(n_j)\}$ are independent, zero-mean, white-noise processes with $\mathrm{var}[\varepsilon_i(\mathbf{s})] \equiv \tau^2$; $\mathbf{s} \in D$, $i = 1,\ldots,\max_j(n_j)$. Then,

$$\mathrm{cov}[Z_i(\mathbf{s}_j), Z_{i'}(\mathbf{s}_{j'})] = \begin{cases} C(\mathbf{s}_j,\mathbf{s}_j) + \tau^2, & i = i',\ j = j', \\ C(\mathbf{s}_j,\mathbf{s}_{j'}), & \text{otherwise.} \end{cases}$$

We conclude this section by restating our goal in technical terms: Assuming the measurement-error model (2.1) (or (2.6)) and the prior model (2.2)–(2.4), we are interested in predicting some functional of the "noiseless" (prior) S-process, not that of the "noisy" Z-process. Of course, what we actually observe are "noisy" data $\{Z(\mathbf{s}_1),\ldots,Z(\mathbf{s}_n)\}$, which are used in making the prediction. This Bayesian problem becomes *empirical* Bayesian because we assume that the noise and spatial-covariance parameters are unknown and require estimation.

3. Spatial prediction of functionals of the S-process

Prediction presupposes some target. If the measurement-error spatial model (2.1) is assumed, then the target, or predictand, will be some functional of $S(\cdot)$ (i.e., not some functional of $Z(\cdot)$). Consider the *linear* predictand given by

$$h(S(\cdot)) = S(B), \quad B \subset D.$$

If $B = \{\mathbf{s}_0\}$, then we call $S(\mathbf{s}_0)$ a *point predictand*. If B is a block of nonzero volume, namely $|B| = \int_B 1\,\mathrm{d}\mathbf{u} > 0$, then

$$S(B) \equiv \frac{1}{|B|}\int_B S(\mathbf{u})\,\mathrm{d}\mathbf{u}, \tag{3.1}$$

which we shall call a *block predictand* (also known as the *spatial mean* over the subregion B). Such predictands can easily be made nonlinear by considering

$$h(S(\cdot)) = g(S(B)), \quad B \subset D,$$

where g is some nonlinear measurable function and B is any measurable subset of D, including a countable number of points. Examples of nonlinear g might include the "smooth" function, $g(S(B)) = \log(S(B))$, or the "nonsmooth" function, $g(S(B)) = I(S(B) \leqslant t)$, where $I(\cdot)$ is the indicator function (i.e., $I(A) = 1$ if A is true, and is zero otherwise), and t is some specified threshold.

3.1. Block prediction

From Bayesian decision theory applied in the context of spatial statistics (e.g., Cressie, 1993a, p. 107), the *best predictor* of $g(S(B))$ (with respect to squared-error loss) is

$$E[g(S(B))|\mathbf{Z}],$$

where g is some measurable function. Consider, for the moment, linear functions g, such as that given by (3.1). Gotway and Cressie (1993) show that the *best* (*heterogeneous*) *linear predictor*, also known as the *simple kriging predictor*, of $S(B)$ is given by

$$\hat{S}_{\mathrm{sk}}(B) \equiv \mu(B) + \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \boldsymbol{\mu}), \tag{3.2}$$

where $\mu(B) \equiv E[S(B)]$, $\boldsymbol{\mu} \equiv E[\mathbf{Z}]$, $\boldsymbol{\Sigma} \equiv \mathrm{var}[\mathbf{Z}] = \mathrm{var}[(S(\mathbf{s}_1), \ldots, S(\mathbf{s}_n))'] + \tau^2\mathbf{I}$, $\mathbf{I}$ is the $n \times n$ identity matrix, $\mathbf{c}(B) \equiv \mathrm{cov}[\mathbf{Z}, S(B)] = (C(\mathbf{s}_1, B), \ldots, C(\mathbf{s}_n, B))'$, and $C(\mathbf{s}, B) \equiv \frac{1}{|B|}\int_B C(\mathbf{s}, \mathbf{u})\,\mathrm{d}\mathbf{u}$; if $B = \{\mathbf{s}_0\}$ then $\mathbf{c}(\mathbf{s}_0) \equiv \mathrm{var}[\mathbf{Z}, S(\mathbf{s}_0)] = (C(\mathbf{s}_1, \mathbf{s}_0), \ldots, C(\mathbf{s}_n, \mathbf{s}_0))'$. The mean-squared prediction error (MSPE) of $\hat{S}_{\mathrm{sk}}(B)$ is given by

$$\mathrm{MSPE}[\hat{S}_{\mathrm{sk}}(B)] = C(B, B) - \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B), \tag{3.3}$$

where $C(B, B) \equiv \mathrm{var}[S(B)] = |B|^{-2}\int_B\int_B C(\mathbf{s}, \mathbf{u})\,\mathrm{d}\mathbf{s}\,\mathrm{d}\mathbf{u}$, if $|B| > 0$; otherwise, $C(\mathbf{s}_0, \mathbf{s}_0) \equiv \mathrm{var}[S(\mathbf{s}_0)]$. By definition, the best predictor of $S(B)$, $E[S(B)|\mathbf{Z}]$, has smaller MSPE than (3.3), but if $(S(\cdot), \varepsilon(\cdot))$ is a bivariate Gaussian process, then $\hat{S}_{\mathrm{sk}}(B) = E[S(B)|\mathbf{Z}]$.

Suppose that $E[S(\cdot)] \equiv \mu(\cdot; \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$, which implies that

$$\mu(B; \boldsymbol{\beta}) = \mathbf{x}(B)'\boldsymbol{\beta}, \quad B \subset D, \tag{3.4}$$

where $\mathbf{x}(B) \equiv (x_0(B), \ldots, x_p(B))'$, $x_j(B) \equiv \frac{1}{|B|}\int_B x_j(\mathbf{u})\,\mathrm{d}\mathbf{u}$; $j = 0, 1, \ldots, p$, and $\{x_0(\cdot), \ldots, x_p(\cdot)\}$ are known functions. If $\boldsymbol{\beta}$ is unknown, then the *best linear unbiased estimator* (BLUE) of $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}} \equiv (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}, \tag{3.5}$$

where $\mathbf{X} \equiv (\mathbf{x}(\mathbf{s}_1), \ldots, \mathbf{x}(\mathbf{s}_n))'$, $\mathbf{x}(\mathbf{s}_i) \equiv (x_0(\mathbf{s}_i), \ldots, x_p(\mathbf{s}_i))'$; $i = 1, \ldots, n$, and $\boldsymbol{\Sigma} \equiv \mathrm{var}[\mathbf{Z}]$. From Goldberger (1962), it is easy to show that the *best linear unbiased predictor* (BLUP), also known as the *universal kriging* (*UK*) *predictor*, of $S(B)$ is given by

$$\hat{S}_{\mathrm{uk}}(B) \equiv \mathbf{x}(B)'\hat{\boldsymbol{\beta}} + \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{3.6}$$

The MSPE of $\hat{S}_{\mathrm{uk}}(B)$ is given by

$$\begin{aligned}\mathrm{MSPE}[\hat{S}_{\mathrm{uk}}(B)] = {} & C(B, B) - \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B) \\ & + (\mathbf{x}(B)' - \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}\mathbf{X})(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{x}(B) - \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B)).\end{aligned}$$

Now, consider prediction of $g(S(B))$, where g is a nonlinear function. Cressie (1993b) shows that the optimality properties of the linear predictors (3.2) and (3.6) do not carry over to prediction of $g(S(B))$, where g is nonlinear. In particular, Cressie (1993b) shows that $g(\hat{S}_{\mathrm{uk}}(B))$ is a biased predictor of $g(S(B))$, where the direction of the first-order bias depends on $g''(\cdot)$. If one is doing many such nonlinear predictions, as one might do when mapping small areas that are above or below a prespecified threshold, this bias can result in an inaccurate map. However, classical nonlinear predictors such as indicator kriging, indicator cokriging, and disjunctive kriging are problematic when applied to the prediction of $g(S(B))$, whenever B is not a single point; see the discussion in Cressie (1993a, Section 5.2) and Aldworth (1998). We require a predictor that adapts easily to linear/nonlinear predictands, absence/presence of measurement error, and point/block support in the predictand.

3.2. Conditional simulation

Universal kriging predictors are generally smoother than their corresponding predictands and consequently yield biased predictors of nonlinear functionals of $S(\cdot)$. Journel (1974) constructed a simulation algorithm that yielded more realistic looking realizations of the spatial process of interest (in his case, this was $Z(\cdot)$) than the surface defined by universal kriging. The algorithm (and later versions) is called *conditional simulation* because all realizations are conditional on having observed $\mathbf{Z}$. In the case where $Z(\cdot)$ is of interest, conditional simulation "honors" the data.

We shall now modify the development in Journel (1974) to obtain conditional-simulation algorithms for the signal $S(\cdot)$. Suppose that we can generate (nonconditional) simulation processes $S_N(\cdot)$ and $\varepsilon_N(\cdot)$, independent of one another (and of $S(\cdot)$ and $\varepsilon(\cdot)$), such that

$$E[S_N(\mathbf{s})] = E[S(\mathbf{s})] = \mu(\mathbf{s}), \quad \mathbf{s} \in D,$$

$$E[\varepsilon_N(\mathbf{s})] = E[\varepsilon(\mathbf{s})] = 0, \quad \mathbf{s} \in D,$$

$$\mathrm{cov}[S_N(\mathbf{s}), S_N(\mathbf{u})] = \mathrm{cov}[S(\mathbf{s}), S(\mathbf{u})], \quad \mathbf{s}, \mathbf{u} \in D,$$

$$\mathrm{cov}[\varepsilon_N(\mathbf{s}), \varepsilon_N(\mathbf{u})] = \mathrm{cov}[\varepsilon(\mathbf{s}), \varepsilon(\mathbf{u})], \quad \mathbf{s}, \mathbf{u} \in D.$$

Then consider the nonconditional simulation process, $Z_N(\mathbf{s}) \equiv S_N(\mathbf{s}) + \varepsilon_N(\mathbf{s})$; $\mathbf{s} \in D$.

Recall that the distribution of interest is the posterior distribution of $S(\cdot)|\mathbf{Z}$. What we require is a simulation process whose distribution at least approximately matches that of $S(\cdot)|\mathbf{Z}$. Define the conditional simulation of $S(\cdot)$, based on simple kriging, as

$$\begin{aligned} S_C(\mathbf{s}) &\equiv \hat{S}_{\mathrm{sk}}(\mathbf{s}) + (S_N(\mathbf{s}) - \hat{S}_{N,\mathrm{sk}}(\mathbf{s})), \quad \mathbf{s} \in D, \\ &= S_N(\mathbf{s}) + \mathbf{c}(\mathbf{s})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{Z}_N), \quad \mathbf{s} \in D, \end{aligned} \tag{3.7}$$

since $\hat{S}_{\mathrm{sk}}(\mathbf{s}) = \mu(\mathbf{s}) + \mathbf{c}(\mathbf{s})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \boldsymbol{\mu})$ and $\hat{S}_{N,\mathrm{sk}}(\mathbf{s}) \equiv \mu(\mathbf{s}) + \mathbf{c}(\mathbf{s})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z}_N - \boldsymbol{\mu})$, where $\mathbf{Z}_N = (Z_N(\mathbf{s}_1), \ldots, Z_N(\mathbf{s}_n))'$. Observe that $S_C(\mathbf{s})$ depends on $\mu(\cdot)$ only through $S_N(\mathbf{s})$ and $\mathbf{Z}_N$ (i.e., knowledge of $\mu(\cdot)$ is required to generate $S_N(\cdot)$).

From (3.7) and the orthogonality property of simple kriging, it can be seen that

$$E[S_C(\mathbf{s})] = E[S_N(\mathbf{s})] = E[S(\mathbf{s})], \quad \mathbf{s} \in D,$$

$$\mathrm{cov}[S_C(\mathbf{s}), S_C(\mathbf{u})] = \mathrm{cov}[S(\mathbf{s}), (S(\mathbf{u})], \quad \mathbf{s}, \mathbf{u} \in D$$

and

$$E[S_C(\mathbf{s})|\mathbf{Z}] = \hat{S}_{\mathrm{sk}}(\mathbf{s}), \quad \mathbf{s} \in D,$$

$$\mathrm{var}[S_C(\mathbf{s})|\mathbf{Z}] = \mathrm{var}[S(\mathbf{s}) - \hat{S}_{\mathrm{sk}}(\mathbf{s})] \equiv \mathrm{MSPE}[\hat{S}_{\mathrm{sk}}(\mathbf{s})], \quad \mathbf{s} \in D.$$

Furthermore, if $(S_N(\cdot), \varepsilon_N(\cdot))$ and $(S(\cdot), \varepsilon(\cdot))$ are both bivariate Gaussian processes, then

$$E[S_C(\mathbf{s})|\mathbf{Z}] = \hat{S}_{\mathrm{sk}}(\mathbf{s}) = E[S(\mathbf{s})|\mathbf{Z}], \quad \mathbf{s} \in D$$

and

$$\mathrm{cov}[S_C(\mathbf{s}), S_C(\mathbf{u})|\mathbf{Z}] = \mathrm{cov}[S(\mathbf{s}), S(\mathbf{u})|\mathbf{Z}].$$

Because the posterior distribution is Gaussian, it is seen in this case that conditional simulation given by (3.7) generates realizations from a Gaussian posterior distribution $S(\cdot)|\mathbf{Z}$, with mean and covariance functions as above. When $(S(\cdot), \varepsilon(\cdot))$ are no longer bivariate Gaussian, the realizations from (3.7) may not be from $S(\cdot)|\mathbf{Z}$, even though their first two (conditional) moments match.

It is easy to show that the MSPE of $S_C(\mathbf{s})$ satisfies

$$\mathrm{MSPE}[S_C(\mathbf{s})] = 2 \times \mathrm{MSPE}[\hat{S}_{\mathrm{sk}}(\mathbf{s})], \quad \mathbf{s} \in D,$$

which suggests that, for the purpose of *prediction* of $S(\mathbf{s})$, $\hat{S}_{\mathrm{sk}}(\mathbf{s})$ is always to be preferred to $S_C(\mathbf{s})$. Of course, one never predicts with just one conditional realization; conditional simulation allows many realizations to be drawn. Better predictors can then be obtained that depend on all realizations.

Let $S_{C,l}(\cdot)$ denote the lth conditional simulation, and suppose that its conditional distribution matches *exactly* that of $S(\cdot)|\mathbf{Z}$; then, by definition,

$$E[g(S(\cdot))|\mathbf{Z}] = E[g(S_{C,l}(\cdot))|\mathbf{Z}], \quad l = 1, \ldots, L \tag{3.8}$$

for any measurable function g. For example, if $g(S(\cdot)) = I(S(B) > t)$, then a conditionally unbiased estimator of the optimal predictor is

$$\hat{E}[I(S(B) > t)|\mathbf{Z}] = L^{-1} \sum_{l=1}^{L} I(S_{C,l}(B) > t), \tag{3.9}$$

where (3.9) converges a.s. to the best predictor $E[I(S(B) > t)|\mathbf{Z}]$, as $L \to \infty$.

One of the advantages of geostatistics and kriging has been its dependence only on the second-order properties of $S(\cdot)$ and $Z(\cdot)$. This nonparametric feature has had a great deal to do with its popularity in the earth, atmospheric (where it is called objective analysis), and environmental sciences. However, conditional simulation is highly parametric in that its definition implies all possible joint, marginal, and conditional distributions (which could all eventually be determined using estimates like the right-hand side of (3.9)). As has already been noted, these (conditional) distributions of $S_{C,l}(\cdot)$ may not match those of $S(\cdot)$, beyond their first two moments. So, although one obtains a very precise estimate of the right-hand side of (3.8) with a large number L of conditional simulations, it may be a very biased estimate of the quantity of interest, namely $E[g(S(\cdot))|\mathbf{Z}]$. This is borne out in the simulation experiment presented in Section 5 and is also illustrated by Gotway (1994).

4. Kriging with constraints

Suppose that

$$g_1(S(\cdot)) \equiv (S(\mathbf{s}_0))^2$$

a nonlinear function that is quadratic in a point value. Consider the simple-kriging predictor $\hat{S}_{\mathrm{sk}}(\mathbf{s}_0)$, which is a best linear predictor for $S(\mathbf{s}_0)$. It is tempting to use

$$\hat{g}_1(S(\cdot)) \equiv (\hat{S}_{\mathrm{sk}}(\mathbf{s}_0))^2$$

as a predictor for $(S(\mathbf{s}_0))^2$, but a quick calculation of means shows this to be an inherently biased predictor. For example, suppose that $S(\cdot)$ is second-order stationary with $E[S(\mathbf{s})] \equiv \mu$ and $\mathrm{var}[S(\mathbf{s})] \equiv \sigma^2$. Then,

$$E[(S(\mathbf{s}_0))^2] = \mu^2 + \sigma^2,$$

which is larger than or equal to

$$E[(\hat{S}_{\mathrm{sk}}(\mathbf{s}_0))^2] = \mu^2 + \mathbf{c}(\mathbf{s}_0)'\boldsymbol{\Sigma}^{-1}\mathbf{c}(\mathbf{s}_0);$$

see (3.3). The bias, $\mathbf{c}(\mathbf{s}_0)'\boldsymbol{\Sigma}^{-1}\mathbf{c}(\mathbf{s}_0) - \sigma^2$, can be substantial.

Intuitively, the simple-kriging predictor "oversmooths" for some purposes. Although the first moments of $S(\mathbf{s}_0)$ and $\hat{S}_{\mathrm{sk}}(\mathbf{s}_0)$ are matched, $\hat{S}_{\mathrm{sk}}(\mathbf{s}_0)$ is obtained by minimizing the mean-squared prediction error without any regard for other moments of $S(\mathbf{s}_0)$. Typically, a histogram of simple-kriging predictors shows a much tighter distribution than that of its predictands. In view of this, Cressie (1993b) proposed the *constrained kriging* (CK) *predictor* $\hat{S}_{\mathrm{ck}}(\mathbf{s}_0)$, which is the best predictor that matches not only first moments but second moments (excluding covariances) as well. He shows that for prediction of $g(S(B))$, where g is smooth, the predictor $g(\hat{S}_{\mathrm{ck}}(B))$ is approximately unbiased and approximately best and, if $(S(\cdot), \varepsilon(\cdot))$ is bivariate Gaussian, then $g(\hat{S}_{\mathrm{ck}}(B))$ is exactly unbiased. We extend this methodology to the case where covariances are also matched.

Suppose we wish to predict $g(\mathbf{S})$, where g is some scalar-valued function that is nonlinear in its vector-valued argument,

$$\mathbf{S} \equiv (S(B_1), \ldots, S(B_m))', \quad B_1, \ldots, B_m \subset D.$$

Consider predictors of the form $g(\hat{\mathbf{S}})$, where $\hat{\mathbf{S}} \equiv \mathbf{A}'\mathbf{Z}$ is a linear predictor of $\mathbf{S}$, and $\mathbf{A} \equiv (\mathbf{a}_1, \ldots, \mathbf{a}_m)$ is an $n \times m$ matrix.

Assume that the coefficients $\mathbf{A}$ satisfy the constraints,

$$\text{(C1)}: E[\mathbf{A}'\mathbf{Z}] = E[\mathbf{S}],$$

$$\text{(C2)}: \ \mathrm{var}[\mathbf{A}'\mathbf{Z}] = \mathrm{var}[\mathbf{S}].$$

Notice that (C1) is standard in geostatistics and (C2) matches the variances and covariances of random variables in the prediction set $\{S(B_1), \ldots, S(B_m)\}$; cf. conditional simulation, for which *all* first and second moments of the process are matched (Section 3.2). For future reference, we write $\boldsymbol{\mu}_m \equiv E[\mathbf{S}]$ and $\boldsymbol{\Sigma}_m \equiv \mathrm{var}[\mathbf{S}]$, where recall that $\mathbf{S}$ is an $m \times 1$ vector. If $(S(\cdot), \varepsilon(\cdot))$ is bivariate Gaussian, and if (C1) and (C2) hold, then $\mathbf{A}'\mathbf{Z}$ and $\mathbf{S}$ are equal in distribution. As a consequence,

$$E[g(\mathbf{A}'\mathbf{Z})] = E[g(\mathbf{S})]$$

for any integrable function g, and hence $g(\mathbf{A}'\mathbf{Z})$ is an unbiased predictor of $g(\mathbf{S})$.

Suppose that $(S(\cdot), \varepsilon(\cdot))$ is not bivariate Gaussian, but suppose that g is smooth enough to possess two continuous derivatives; that is, the $m \times 1$ vector of first derivatives, $g'(\mathbf{x}) \equiv \partial g(\mathbf{x})/\partial \mathbf{x}$, and the $m \times m$ matrix of second derivatives, $g''(\mathbf{x}) \equiv \partial^2 g(\mathbf{x})/\partial \mathbf{x} \partial \mathbf{x}'$, both exist and are continuous. In what is to follow, we use the δ-method

(e.g., Schott, 1997, Chapter 8) to motivate why constraints (C1) and (C2) are desirable when looking for optimal predictors of $g(\mathbf{S})$. We have,

$$E[g(\mathbf{A}'\mathbf{Z})] \simeq g(E[\mathbf{A}'\mathbf{Z}]) + E[(\mathbf{A}'\mathbf{Z} - E[\mathbf{A}'\mathbf{Z}])' g''(E[\mathbf{A}'\mathbf{Z}])(\mathbf{A}'\mathbf{Z} - E[\mathbf{A}'\mathbf{Z}])]/2$$
$$= g(E[\mathbf{A}'\mathbf{Z}]) + \mathrm{tr}\{g''(E[\mathbf{A}'\mathbf{Z}])\mathrm{var}[\mathbf{A}'\mathbf{Z}]\}/2$$

and similarly,

$$E[g(\mathbf{S})] \simeq g(E[\mathbf{S}]) + \mathrm{tr}\{g''(E[\mathbf{S}])\mathrm{var}[\mathbf{S}]\}/2,$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator, and $g''(\mathbf{w}) = \partial^2 g(\mathbf{x})/\partial\mathbf{x}\partial\mathbf{x}'|_{\mathbf{x}=\mathbf{w}}$. Consequently, if (C1) and (C2) hold, we see that

$$E[g(\mathbf{A}'\mathbf{Z})] \simeq E[g(\mathbf{S})]$$

for g smooth enough to possess two continuous derivatives. Thus, we could look for an optimal predictor by choosing the $n \times m$ matrix $\mathbf{A}$ that satisfies constraints (C1) and (C2) and minimizes the scalar quantity, $\mathrm{MSPE}[g(\mathbf{A}'\mathbf{Z})]$. The problem is that this latter quantity cannot be calculated from the model specified in (2.1)–(2.5). But, if $\mathbf{A}$ satisfies constraints (C1) and (C2) and g is smooth enough to possess two continuous derivatives, then by the δ-method,

$$\mathrm{MSPE}[g(\mathbf{A}'\mathbf{Z})] \equiv E[g(\mathbf{A}'\mathbf{Z}) - g(\mathbf{S})]^2$$
$$\simeq [g'(\boldsymbol{\mu}_m)]'\mathrm{var}[\mathbf{A}'\mathbf{Z}][g'(\boldsymbol{\mu}_m)] + [g'(\boldsymbol{\mu}_m)]'\mathrm{var}[\mathbf{S}][g'(\boldsymbol{\mu}_m)]$$
$$- 2[g'(\boldsymbol{\mu}_m)]'\mathrm{cov}[\mathbf{A}'\mathbf{Z}, \mathbf{S}][g'(\boldsymbol{\mu}_m)],$$

where $g'(\boldsymbol{\mu}_m) = \partial g(\mathbf{x})/\partial\mathbf{x}|_{\mathbf{x}=\boldsymbol{\mu}_m}$ and recall that $\boldsymbol{\mu}_m \equiv E[\mathbf{S}] = E[\mathbf{A}'\mathbf{Z}]$. Therefore,

$$\mathrm{MSPE}[g(\mathbf{A}'\mathbf{Z})] \simeq \mathbf{y}'\mathbf{M}_A\mathbf{y},$$

where $\mathbf{y} \equiv g'(\boldsymbol{\mu}_m)$ and $\mathbf{M}_A \equiv \mathrm{MSPE}[\mathbf{A}'\mathbf{Z}] = \mathrm{var}[\mathbf{A}'\mathbf{Z} - \mathbf{S}]$, by (C1).

These calculations motivate looking for a predictor of $g(\mathbf{S})$ given by $g(\mathbf{A}_0'\mathbf{Z})$, where $\mathbf{A}_0$ minimizes $\mathbf{y}'\mathbf{M}_A\mathbf{y}$ subject to (C1) and (C2) being satisfied. In empirical Bayes terminology, we are seeking the best predictor from the class of homogeneously linear predictors that satisfy a mean constraint and a covariance-matrix constraint, where the loss function is assumed to be squared-error loss; this results in an approximately unbiased and minimum MSPE predictor $g(\mathbf{A}_0'\mathbf{Z})$. If we further assume Gaussianity, the predictor is exactly unbiased. In what is to follow, we shall investigate this optimization problem, including existence, uniqueness, and properties of the solution.

Assume that the linear model (3.4) holds. Then, $\boldsymbol{\mu}_m \equiv (\mu(B_1), \ldots, \mu(B_m))' = \mathbf{X}_m'\boldsymbol{\beta}$, where $\mathbf{X}_m \equiv (\mathbf{x}(B_1), \ldots, \mathbf{x}(B_m))$ is a $(p+1) \times m$ matrix. Consequently, conditions (C1) and (C2) can be written as

(C1): $\mathbf{A}'\mathbf{X} = \mathbf{X}_m'$,

(C2): $\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} = \boldsymbol{\Sigma}_m$.

These constraints can also be written as

$$\text{(C1): } \mathbf{a}_i'\mathbf{X} = \mathbf{x}(B_i)', \quad i = 1, \ldots, m,$$

$$\text{(C2): } \mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_j = C(B_i, B_j), \quad 1 \leqslant i \leqslant j \leqslant m, \tag{4.1}$$

where recall that $\mathbf{A} \equiv (\mathbf{a}_1, \ldots, \mathbf{a}_m)$.

Now,

$$\mathbf{y}'\mathbf{M}_A\mathbf{y} = \mathbf{y}'E[(\mathbf{A}'\mathbf{Z} - \mathbf{S})(\mathbf{A}'\mathbf{Z} - \mathbf{S})']\mathbf{y}$$
$$= \mathbf{y}'\{\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} - \mathbf{A}'\mathbf{C} - \mathbf{C}'\mathbf{A} + \boldsymbol{\Sigma}_m\}\mathbf{y},$$

where $\mathbf{C} \equiv \operatorname{cov}(\mathbf{Z}, \mathbf{S}) = (\mathbf{c}(B_1), \ldots, \mathbf{c}(B_m))$ and $\mathbf{c}(B_i)$ is defined just after (3.2). If $\mathbf{A}$ satisfies (C2), then

$$\mathbf{y}'\mathbf{M}_A\mathbf{y} = 2\mathbf{y}'\{\boldsymbol{\Sigma}_m - \mathbf{A}'\mathbf{C}\}\mathbf{y},$$

which we want to minimize with respect to $\mathbf{A}$, subject to (C1) and (C2). Therefore, the objective function to maximize, with respect to $\mathbf{A}$, $\{\boldsymbol{\lambda}_{1i}\}$, and $\{\lambda_{2ij}\}$, is

$$f(\mathbf{A}) \equiv 2\mathbf{y}'\mathbf{A}'\mathbf{C}\mathbf{y} + 2\sum_{i=1}^{m}(\mathbf{a}_i'\mathbf{X} - \mathbf{x}(B_i)')\boldsymbol{\lambda}_{1i} - \sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_{2ij}(\mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_j - C(B_i, B_j)),$$

where $\lambda_{2ji} \equiv \lambda_{2ij}$, for $i \leqslant j$.

Because we are assuming the linear model (3.4), it is useful to consider the vector of universal kriging (UK) predictors, which we denote as $\hat{\mathbf{S}}_{\text{uk}}$. From (3.6),

$$\hat{\mathbf{S}}_{\text{uk}} \equiv \mathbf{X}_m'\hat{\boldsymbol{\beta}} + \mathbf{C}'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{4.2}$$

Notice that $\hat{\mathbf{S}}_{\text{uk}}$ can be written as $\mathbf{A}_{\text{uk}}'\mathbf{Z}$, a linear predictor in $\mathbf{Z}$. From the second-order properties of universal kriging, $E[\hat{\mathbf{S}}_{\text{uk}}] = E[\mathbf{S}]$, but typically $\operatorname{var}[\hat{\mathbf{S}}_{\text{uk}}] \neq \operatorname{var}[\mathbf{S}]$. Thus, although $\hat{\mathbf{S}}_{\text{uk}}$ satisfies (C1), typically it does not satisfy (C2).

The following propositions suggest how to formulate a predictor $g(\mathbf{A}_0'\mathbf{Z})$ that possesses certain optimality properties subject to constraints (C1) and (C2).

Proposition 1. *Assume that* $\mathbf{P} \equiv \operatorname{var}[\mathbf{S}] - \operatorname{var}[\mathbf{X}_m'\hat{\boldsymbol{\beta}}]$ *and* $Q \equiv \operatorname{var}[\hat{\mathbf{S}}_{\text{uk}}] - \operatorname{var}[\mathbf{X}_m'\hat{\boldsymbol{\beta}}]$ *are positive definite (p.d.), where* $\hat{\boldsymbol{\beta}}$ *is the BLUE of* $\boldsymbol{\beta}$ *given by* (3.5) *and* $\hat{\mathbf{S}}_{\text{uk}}$ *is the UK predictor given by* (4.2). *Define the symmetric p.d. matrices* $\mathbf{P}_1 \equiv \mathbf{P}^{1/2}$ *and* $\mathbf{Q}_1 \equiv \mathbf{Q}^{1/2}$. *Then there exists an invertible* $m \times m$ *matrix* $\mathbf{T}$ *and a diagonal matrix* $\mathbf{N} \equiv \operatorname{diag}(v_1, \ldots, v_m)$ *with positive entries* $\{v_i\}$ *such that*

$$\mathbf{T}\mathbf{Q}_1^{-1}\mathbf{T}' = \mathbf{N} = (\mathbf{T}')^{-1}\mathbf{P}_1\mathbf{T}^{-1}.$$

Furthermore, if we define

$$\mathbf{K} \equiv \mathbf{Q}_1^{-1}\mathbf{P}_1,$$

then

$$\mathbf{K} = \mathbf{T}^{-1}\mathbf{N}^2\mathbf{T} = \sum_{i=1}^{m} v_i^2\mathbf{V}_i\mathbf{W}_i',$$

where $\mathbf{T}^{-1} \equiv (\mathbf{V}_1, \ldots, \mathbf{V}_m)$, *and hence* $\mathbf{T}' \equiv (\mathbf{W}_1, \ldots, \mathbf{W}_m)$ *and* $\mathbf{V}_i'\mathbf{W}_j = 1$, *if* $i = j$; $=0$, *otherwise.*

Proof. Define $\mathbf{R} \equiv \mathbf{Q}_1^{-1/2}\mathbf{P}_1\mathbf{Q}_1^{-1/2}$, which is symmetric and p.d. Let $\mathbf{U}$ be the matrix whose columns are $\mathbf{R}$'s normalized eigenvectors and write its eigenvalues as $\{v_i^2\}$. That is, $\mathbf{U}'\mathbf{U} = \mathbf{I}$ and $\mathbf{U}'\mathbf{R}\mathbf{U} = \mathrm{diag}(v_1^2, \ldots, v_m^2)$. Define

$$\mathbf{T} \equiv \mathrm{diag}(v_1^{1/2}, \ldots, v_m^{1/2})\mathbf{U}'\mathbf{Q}_1^{1/2}.$$

Then $\mathbf{T}\mathbf{Q}_1^{-1}\mathbf{T}' = \mathrm{diag}(v_1, \ldots, v_m) = (\mathbf{T}')^{-1}\mathbf{P}_1\mathbf{T}^{-1}$. Consequently,

$$\begin{aligned}\mathbf{K} = \mathbf{Q}_1^{-1}\mathbf{P}_1 &= \mathbf{T}^{-1}\,\mathrm{diag}(v_1^2, \ldots, v_m^2)\mathbf{T} \\ &= \sum_{i=1}^{m} v_i^2 \mathbf{V}_i \mathbf{W}_i'.\end{aligned}$$

Since $(\mathbf{T}^{-1})'\mathbf{T}' = \mathbf{I}$, we see that $\mathbf{V}_i'\mathbf{W}_j = 1$, if $i = j$; $=0$, otherwise. □

Proposition 2. *Assume that* $\mathbf{P}$, $\mathbf{Q}$, *and* $\boldsymbol{\Sigma}$ *are p.d. Recall from Proposition* 1 *that* $\mathbf{P}_1 \equiv \mathbf{P}^{1/2}$, $\mathbf{Q}_1 \equiv \mathbf{Q}^{1/2}$, *and* $\mathbf{K} = \mathbf{Q}_1^{-1}\mathbf{P}_1$. *Recall that the objective function to be maximized is*

$$f(\mathbf{A}) \equiv 2\mathbf{y}'\mathbf{A}'\mathbf{C}\mathbf{y} + 2\sum_{i=1}^{m}(\mathbf{a}_i'\mathbf{X} - \mathbf{x}(B_i)')\boldsymbol{\lambda}_{1i} - \sum_{i=1}^{m}\sum_{j=i}^{m}\lambda_{2ij}(\mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_j - C(B_i, B_j)),$$

where $\mathbf{y} = g'(\boldsymbol{\mu}_m)$. *Then* $f(\mathbf{A})$ *is maximized at*

$$\mathbf{A}_0 = \boldsymbol{\Sigma}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1})\mathbf{C}\mathbf{K} + \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m \tag{4.3}$$

for $\mathbf{y} = \mathbf{V}_1, \ldots, \mathbf{V}_m$, *defined in Proposition* 1.

Proof. See the appendix. □

Definition. Suppose that $\mathbf{P}$, $\mathbf{Q}$, and $\boldsymbol{\Sigma}$ are p.d. Then the *covariance-matching constrained kriging* (CM) *predictor* of $\mathbf{S}$ is

$$\hat{\mathbf{S}}_{\mathrm{cm}} \equiv \mathbf{A}_0'\mathbf{Z},$$

where $\mathbf{A}_0$ is given by (4.3).

It is not difficult to show that

$$\hat{\mathbf{S}}_{\mathrm{cm}} = \mathbf{X}_m'\hat{\boldsymbol{\beta}} + \mathbf{K}'\mathbf{C}'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{4.4}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}$, $\mathbf{K} = \mathbf{Q}_1^{-1}\mathbf{P}_1$, and recall that $\mathbf{P} = \mathbf{P}_1\mathbf{P}_1$ and $\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_1$. Proposition 2 establishes the optimality of the CM predictor (4.4) in m well-defined directions of $\mathbb{R}^m$. In Proposition 3 below, we verify that $\mathbf{A}_0$ does indeed satisfy constraints (C1) and (C2), as it should.

Proposition 3. *Assume that* $\mathbf{P}$, $\mathbf{Q}$, *and* $\boldsymbol{\Sigma}$ *are p.d. Then the coefficients* $\mathbf{A}_0$ *in* $\mathbf{A}_0'\mathbf{Z}$ *given by* (4.4), *satisfy constraints* (C1) *and* (C2).

Proof. See the appendix. □

Recall that if a predictor $\hat{\mathbf{S}}$ of $\mathbf{S}$ satisfies (C1) and (C2), then $g(\hat{\mathbf{S}})$ is an exactly unbiased predictor of $g(\mathbf{S})$ for any integrable function g if $(S(\cdot), \varepsilon(\cdot))$ is bivariate Gaussian. Further, even if $(S(\cdot), \varepsilon(\cdot))$ is not bivariate Gaussian, $g(\hat{\mathbf{S}})$ is an approximately unbiased predictor of $g(\mathbf{S})$ if g is smooth enough to possess two derivatives. Consequently, even if $\mathbf{y} \notin \{\mathbf{V}_1, \ldots, \mathbf{V}_m\}$ given in Proposition 2, $g(\hat{\mathbf{S}}_{\mathrm{cm}})$ is still approximately unbiased.

From (4.4), we obtain

$$\begin{aligned}\mathrm{MSPE}[\hat{\mathbf{S}}_{\mathrm{cm}}] &\equiv E[(\hat{\mathbf{S}}_{\mathrm{cm}} - \mathbf{S})(\hat{\mathbf{S}}_{\mathrm{cm}} - \mathbf{S})'] \\ &= 2\boldsymbol{\Sigma}_m - \mathbf{X}_m'(\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{C} - \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m \\ &\quad - (\mathbf{P}_1\mathbf{Q}_1 + \mathbf{Q}_1\mathbf{P}_1), \end{aligned} \tag{4.5}$$

since

$$\mathrm{cov}[\hat{\mathbf{S}}_{\mathrm{cm}}, \mathbf{S}] = \mathbf{X}_m'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{C} + \mathbf{P}_1\mathbf{Q}_1. \tag{4.6}$$

Expression (4.5) gives not only prediction variances, but also prediction covariances, and may be useful for defining design criteria.

Assuming that $\boldsymbol{\Sigma}$ is p.d., we see that the CM predictor (4.4) of $\mathbf{S}$ exists if,

(E1): $\mathbf{Q} \equiv \mathrm{var}[\hat{\mathbf{S}}_{\mathrm{uk}}] - \mathrm{var}[\mathbf{X}_m'\hat{\boldsymbol{\beta}}]$ is p.d.,

(E2): $\mathbf{P} \equiv \mathrm{var}[\mathbf{S}] - \mathrm{var}[\mathbf{X}_m'\hat{\boldsymbol{\beta}}]$ is p.d.

Consider the following result:

Proposition 4. *Suppose that $\boldsymbol{\Sigma}$ is p.d. Then the matrix $\mathbf{Q} \equiv \mathrm{var}[\hat{\mathbf{S}}_{\mathrm{uk}}] - \mathrm{var}[\mathbf{X}_m'\hat{\boldsymbol{\beta}}]$ is n.n.d* (i.e., *all of the eigenvalues of $\mathbf{Q}$ are nonnegative*).

Proof. See the appendix. □

Thus, existence condition (E1) can only be violated if $\mathbf{Q}$ possesses some zero-valued eigenvalues. Should this occur, a practical solution to this problem might be to add a small positive constant to the zero-valued eigenvalues of $\mathbf{Q}$. Clearly, the more stringent of the two conditions is (E2), since $\mathbf{P} = \boldsymbol{\Sigma}_m - \mathbf{X}_m'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m$ may not be n.n.d. (i.e., it may possess negative eigenvalues).

If (E2) is violated, a possible solution is to partition the prediction vector $\mathbf{S}$ into subvectors and relax the covariance-matching constraints so that they apply only to elements *within*, but not *between*, partitions. For example, suppose that $\mathbf{S}$ is partitioned into r parts,

$$\mathbf{S} \equiv [\mathbf{S}_1' | \ldots | \mathbf{S}_r']'.$$

Then, the "partitioned" constraints may be written as

(P1): $E[\mathbf{A}_i'\mathbf{Z}] = E[\mathbf{S}_i], \quad i = 1, \ldots, r,$

(P2): $\mathrm{var}[\mathbf{A}_i'\mathbf{Z}] = \mathrm{var}[\mathbf{S}_i], \quad i = 1, \ldots, r,$

where $\mathbf{A}$ is partitioned conformably as

$$\mathbf{A} \equiv [\mathbf{A}_1 | \ldots | \mathbf{A}_r].$$

Note that (P1) is exactly equivalent to (C1), but that (P2) represents weaker constraints than (C2) because there are fewer covariance conditions to be satisfied in (P2).

Proposition 2 can be applied to the prediction of each of the subvectors $\mathbf{S}_1, \ldots, \mathbf{S}_r$, and if in each case the predictors exist, then we obtain the "*partitioned*" *covariance-matching constrained kriging* (PC) *predictor* of $\mathbf{S}$, given by

$$\hat{\mathbf{S}}_{\mathrm{pc}} \equiv \mathbf{X}_m' \hat{\boldsymbol{\beta}} + \mathbf{K}_p' \mathbf{C}' \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{4.7}$$

where $\mathbf{K}_p \equiv \mathbf{Q}_{p1}^{-1} \mathbf{P}_{p1}$, and $\mathbf{P}_{p1}$ and $\mathbf{Q}_{p1}$ are defined below.

Define $\mathbf{P}_p$ as the positive-definite block-diagonal matrix whose ith block comprises the positive-definite submatrix,

$$\mathbf{P}_{p,i} \equiv \mathrm{var}[\mathbf{S}_i] - \mathrm{var}[\mathbf{X}_{m,i}' \hat{\boldsymbol{\beta}}] = \mathbf{P}_{p1,i} \mathbf{P}_{p1,i}, \quad i = 1, \ldots, r,$$

where $\{\mathbf{P}_{p1,i}\}$ are symmetric square roots and $\mathbf{X}_m \equiv [\mathbf{X}_{m,1} | \ldots | \mathbf{X}_{m,r}]$. Similarly, $\mathbf{Q}_p$ is the positive-definite block-diagonal matrix whose ith block comprises the positive-definite submatrix,

$$\mathbf{Q}_{p,i} \equiv \mathrm{var}[\hat{\mathbf{S}}_{\mathrm{uk},i}] - \mathrm{var}[\mathbf{X}_{m,i}' \hat{\boldsymbol{\beta}}] = \mathbf{Q}_{p1,i} \mathbf{Q}_{p1,i}, \quad i = 1, \ldots, r,$$

where $\{\mathbf{Q}_{p1,i}\}$ are symmetric square roots and $\hat{\mathbf{S}}_{\mathrm{uk}} \equiv [\hat{\mathbf{S}}_{\mathrm{uk},1}' | \ldots | \hat{\mathbf{S}}_{\mathrm{uk},r}']'$. That is, the elements in the diagonal blocks of $\mathbf{P}_p$ are identical to the corresponding elements of $\mathbf{P}$, but all other elements of $\mathbf{P}_p$ are zero; a similar structure holds for $\mathbf{Q}_p$.

If necessary, the partitioning of $\mathbf{S}$ could be refined to the situation where each partition contains only one element; that is,

$$\mathbf{S} \equiv [S(B_1) | \ldots | S(B_m)]'.$$

Then, $\mathbf{P}_p$ is a diagonal matrix with ith diagonal entry,

$$\mathbf{P}_{p,i} \equiv \mathbf{P}_{p1,i} \mathbf{P}_{p1,i} = \mathrm{var}[S(B_i)] - \mathrm{var}[\mathbf{x}(B_i)' \hat{\boldsymbol{\beta}}], \quad i = 1, \ldots, m,$$

and $\mathbf{Q}_p$ is a diagonal matrix with ith diagonal entry,

$$\mathbf{Q}_{p,i} \equiv \mathbf{Q}_{p1,i} \mathbf{Q}_{p1,i} = \mathrm{var}[\hat{S}_{\mathrm{uk}}(B_i)] - \mathrm{var}[\mathbf{x}(B_i)' \hat{\boldsymbol{\beta}}], \quad i = 1, \ldots, m.$$

In this case,

$$\hat{\mathbf{S}}_{\mathrm{pc}} \equiv (\hat{S}_{\mathrm{ck}}(B_1), \ldots, \hat{S}_{\mathrm{ck}}(B_m))',$$

where $\hat{S}_{\mathrm{ck}}(B)$ is the (univariate) constrained kriging predictor of $S(B)$, which we shall now describe briefly.

Consider the special case where $m = 1$. Then, $\mathbf{A}_0$ in (4.3) reduces to the $n \times 1$ vector,

$$\mathbf{a}_0 = \boldsymbol{\Sigma}^{-1} (\mathbf{I} - \mathbf{X}(\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Sigma}^{-1}) \mathbf{c}(B) k + \boldsymbol{\Sigma}^{-1} \mathbf{X}(\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{x}(B),$$

where

$$k \equiv \left\{ \frac{\mathrm{var}[S(B)] - \mathrm{var}[\mathbf{x}(B]' \hat{\boldsymbol{\beta}}]}{\mathrm{var}[\hat{S}_{\mathrm{uk}}(B)] - \mathrm{var}[\mathbf{x}(B)' \hat{\boldsymbol{\beta}}]} \right\}^{1/2},$$

and $\hat{S}_{\mathrm{uk}}(B)$, given by (3.6), is the universal kriging predictor of $S(B)$. Provided that the numerator and denominator in the previous expression are positive, then k is well

defined. In this case, we obtain the (univariate, variance-matching) *constrained kriging* (CK) *predictor* of $S(B)$, namely

$$\hat{S}_{\mathrm{ck}}(B) \equiv \mathbf{a}_0'\mathbf{Z} = \mathbf{x}(B)'\hat{\boldsymbol{\beta}} + k\mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{4.8}$$

Cressie (1993b) shows that (4.8) minimizes MSPE[$\mathbf{a}'\mathbf{Z}$] subject to $\mathbf{a}$ satisfying conditions (C1) and (C2) given by (4.1) for $m = 1$. After some algebra, the MSPE of $\hat{S}_{\mathrm{ck}}(B)$ is given by

$$\begin{aligned}\mathrm{MSPE}[\hat{S}_{\mathrm{ck}}(B)] = 2[&C(B,B) - \mathbf{x}(B)'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B)\\ &- \{(C(B,B) - \mathbf{x}(B)'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{x}(B))\\ &\times(\mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B) - \mathbf{c}(B)'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c}(B))\}^{1/2}].\end{aligned} \tag{4.9}$$

It will be seen in the simulation experiment (Section 5) and the example (Section 6) that the real benefits of extra constraints come when covariances are matched, in addition to matching variances.

Thus, we propose using CM in its purest form given by (4.4) and, if it does not exist, then we could use CM in its partitioned form given by (4.7). We use finer partitions until the constraint (P2) is satisfied. It is possible (although less desirable) that (P2) could reduce to just m univariate CKs given by (4.8). And, in particular cases where (4.8) does not exist (i.e., where k is not well defined), we may have to choose instead the universal kriging predictor (3.6) or a conditional-simulation predictor of the form described in (3.9). Cressie and Johannesson (2001) develop neighborhood-based constraining sets as a practical alternative to partitioning, avoiding the arbitrariness associated with defining the partitions.

5. Computer-simulation experiment

We designed a computer-simulation experiment to explore how covariance-matching constrained kriging compares with ordinary kriging, constrained kriging, and conditional simulation. The experiment yielded predictors of certain linear and nonlinear functionals of the state process $S(\cdot)$ defined over a given spatial domain, under different experimental conditions. The details of this experiment are now presented.

5.1. Spatial domain and spatial model

In this experiment, we consider a rectangular domain in $\mathbb{R}^2$, over which a triangular grid $D = \{(x,y)\colon x = 1,2,\ldots,9,\ \ y = 1 + 2k\sqrt{0.75},\ \ k = 0,1,\ldots,4\} \cup \{(x,y)\colon x = 1.5, 2.5, \ldots, 9.5,\ \ y = 1 + (2k+1)\sqrt{0.75},\ \ k = 0,1,\ldots,4\}$, is laid; see Fig. 1. The triangular grid was chosen for its superior properties to discretize a continuous domain (Serra, 1982), and because there is a natural partition of the grid into prediction sets that each consist of three near-neighbor grid nodes. Two triangular, discrete subdomains, $B_1 = \{(2, 1 + 2\sqrt{0.75}), (2.5, 1 + \sqrt{0.75}), (3, 1 + 2\sqrt{0.75})\}$ and $B_2 = \{(4.5, 1 + 9\sqrt{0.75}), (5, 1 + 8\sqrt{0.75}), (5.5, 9\sqrt{0.75})\}$, are given special emphasis in the simulation;

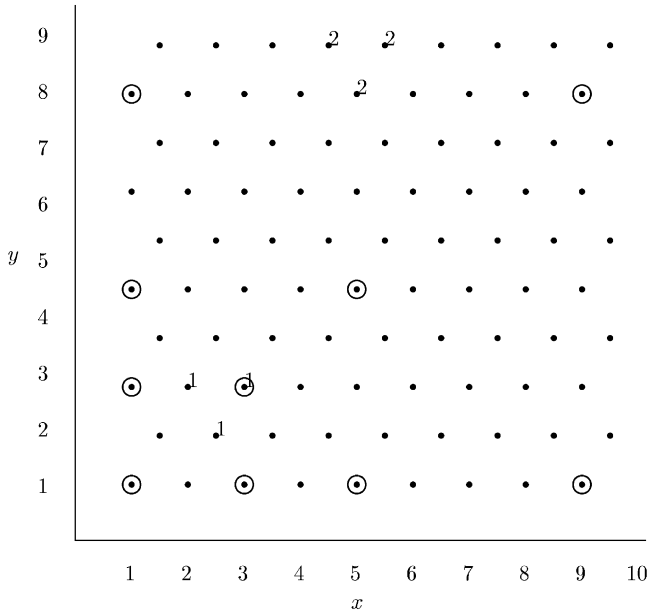

Fig. 1. Domain D for simulation comprises a triangular grid of 90 locations over a rectangular area. The subregions, B_1 and B_2, each comprise a triangular grid of 3 locations, and their locations are identified as "1" and "2", respectively. The 10 locations marked with a circle represent the sample A.

their locations are identified as "1" and "2", respectively, in Fig. 1. Observe that B_1 lies where the sample is most concentrated and that B_2 lies where the sample is least concentrated.

In addition, the 10 locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_{10}\}$ in D, marked with a circle in Fig. 1, represent the fixed set A of sampling locations. Data $\mathbf{Z} \equiv (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_{10}))'$ were generated on A according to the following spatial model:

$$Z(\mathbf{s}) \equiv S(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad \mathbf{s} = (x, y)' \in D,$$

where $\varepsilon(\cdot)$, independent of $S(\cdot)$, is a zero-mean, Gaussian, white-noise measurement-error process such that $\mathrm{var}[\varepsilon(\mathbf{s})] \equiv \tau^2$, and values for τ^2 are specified in Section 5.2 below. The state process $S(\cdot)$ is also specified in Section 5.2.

5.2. Factors of the experiment

Four factors were considered in this experiment: (i) distribution of the random field $S(\cdot)$, (ii) strength of spatial correlation, (iii) noise (i.e., measurement error), (iv) predictor.

Four random fields $\{S_i(\mathbf{s})\colon \mathbf{s} \in D\}$; $i = 1, \ldots, 4$, were considered. The rationale for their choice can be found in Cressie (1993b). In the first case,

$$S_1(\mathbf{s}) \equiv 10 + \delta(\mathbf{s}), \quad \mathbf{s} = (x, y)' \in D, \tag{5.1}$$

where $\delta(\cdot)$ is a zero-mean, Gaussian spatial process with "spherical" covariance function, given by

$$C_1(\mathbf{h}) \equiv \operatorname{cov}[\delta(\mathbf{s}), \delta(\mathbf{s}+\mathbf{h})] = \sigma^2 \begin{cases} 1 - \frac{3}{2}(\|\mathbf{h}\|/a) + \frac{1}{2}(\|\mathbf{h}\|/a)^3, & \|\mathbf{h}\| \leqslant a \\ 0, & \|\mathbf{h}\| > a, \end{cases}$$

$\sigma^2 = 1$, and the range parameter a (specified below) governs the strength of spatial correlation. The second random field is a log Gaussian process, defined as

$$S_2(\mathbf{s}) \equiv k_1 + k_2 \exp(S_1(\mathbf{s})), \quad \mathbf{s} = (x, y)' \in D,$$

where k_1 and k_2 are chosen so that the mean and variance of $S_2(\mathbf{s})$ are equal to those of $S_1(\mathbf{s})$; $\mathbf{s} \in D$. The third random field is defined as

$$S_3(\mathbf{s}) \equiv k_3 - k_4 \exp(S_1(\mathbf{s})), \quad \mathbf{s} = (x, y)' \in D.$$

That is, the distribution of $S_3(\mathbf{s})$ is a "mirror image" of that of $S_2(\mathbf{s})$, and the constants k_3 and k_4 are chosen so that the mean and variance of $S_3(\mathbf{s})$ are equal to those of $S_2(\mathbf{s})$ and $S_1(\mathbf{s})$; $\mathbf{s} \in D$.

Finally, the fourth random field is a mixed Gaussian process, defined as

$$S_4(\mathbf{s}) \equiv k_5 - k_6\{I(\mathbf{s})U(\mathbf{s}) + (1 - I(\mathbf{s}))V(\mathbf{s})\}, \quad \mathbf{s} = (x, y)' \in D,$$

where $I(\cdot)$ is a 0–1 random field such that for any $\mathbf{s} \neq \mathbf{u}$, $I(\mathbf{s}), I(\mathbf{u})$ are independent Bernoulli(0.75) random variables; $U(\cdot)$ is a Gaussian random field with mean equal to 16 and covariance function given by C_1; $V(\cdot)$ is a Gaussian random field with mean equal to 4 and covariance function given by $(1/\sqrt{10})^2 C_1$; $I(\cdot)$, $U(\cdot)$, and $V(\cdot)$ are mutually independent random fields; and finally, k_5 and k_6 are chosen so that the mean and variance of $S_4(\mathbf{s})$ are equal to those of $S_3(\mathbf{s})$, $S_2(\mathbf{s})$, and $S_1(\mathbf{s})$; $\mathbf{s} \in D$.

From Cressie (1993a, Section 3.2.2), it is easy to show that $k_1 = 10 - k_2 \mathrm{e}^{10.5}$, $k_2 = (\mathrm{e}^{21}(\mathrm{e}-1))^{-1/2}$, $k_3 = 10 + k_4 \mathrm{e}^{10.5}$, $k_4 = k_2$, $k_5 = 0.18975$, and $k_6 = 7.5333$. Observe that the marginal distribution of $S_1(\mathbf{s})$ is symmetric, that of $S_2(\mathbf{s})$ is skewed to the right, that of $S_3(\mathbf{s})$ is skewed to the left, and that of $S_4(\mathbf{s})$ is bimodal with asymmetric peaks.

The strength of spatial correlation in model (5.1) is governed by the range parameter a. Three levels of a were chosen: $a = 4.1$, 7.5, and 30. These values represent weak, moderate, and strong correlation structures, respectively.

Two levels of measurement error were considered: τ^2=0.1 and 2. These two levels will be referred to as *low* noise and *high* noise, respectively. The two values of τ^2 were chosen according to the square-root signal-to-noise ratio scale, σ/τ, from which we obtain the two ratios: $\sigma/\tau \in \{0.71, 3.16\}$.

Four types of predictor were considered, namely ordinary kriging (OK), ordinary constrained kriging (CK), ordinary covariance-matching constrained kriging (CM), and conditional simulation (CS).

In the case of the covariance-matching constrained kriging predictor, $\mathbf{S}$ was partitioned into 30 subvectors, each of length three, in a manner described in Section 4. The first subvector corresponds to the locations $\{(1,1), (1.5, 1+\sqrt{0.75}), (2,1)\}$, the second subvector corresponds to the locations $\{(2.5, 1+\sqrt{0.75}), (3,1), (3.5, 1+\sqrt{0.75})\}$, and

so forth. This yielded the predictor $\hat{\mathbf{S}}_{\mathrm{pc}}$ given by (4.7). Notice that we also could have partitioned $\mathbf{S}$ into one 90×1 vector (which yields $\hat{\mathbf{S}}_{\mathrm{cm}}$ given by (4.4)), two 45×1 vectors, three 30×1 vectors or, for that matter, forty five 2×1 vectors. The higher the dimension of the vectors, the less chance there is for covariance matching to be possible. Yet, if the dimension is too low, the full power of covariance matching is not felt. The 30 three-dimensional vectors chosen here represents a nice compromise for which covariance matching was possible for all 30 vectors.

The crucial component of (4.7) is the matrix $\mathbf{K}_p$ that allows matching of covariances. That matrix is obtained from two positive-definite symmetric matrices $\mathbf{P}_p$ and $\mathbf{Q}_p$, such that $\mathbf{K}_p = \mathbf{Q}_{p1}^{-1}\mathbf{P}_{p1}$. To obtain these matrices, we carry out a spectral decomposition of $\mathbf{P}_p$ and $\mathbf{Q}_p$. That is, we write

$$\mathbf{P}_p \equiv \mathbf{H}\boldsymbol{\Lambda}\mathbf{H}',$$

where $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$ is a diagonal matrix of positive eigenvalues, and $\mathbf{H}$ is a matrix of corresponding normalized eigenvectors. Then $\mathbf{P}_{p1}$ is the symmetric matrix,

$$\mathbf{P}_{p1} \equiv \mathbf{H}\boldsymbol{\Lambda}^{1/2}\mathbf{H}', \tag{5.2}$$

where $\boldsymbol{\Lambda}^{1/2} \equiv \mathrm{diag}\{\lambda_1^{1/2}, \ldots, \lambda_n^{1/2}\}$; $\mathbf{Q}_{p1}$ is chosen similarly. It is our intention in this simulation experiment to use empirical optimality criteria to determine how well $\hat{\mathbf{S}}_{\mathrm{pc}}$ performs.

Finally, the conditional simulation predictor of $g(S(B))$ is defined as

$$\hat{p}_{\mathrm{cs}}(g(S(B))) \equiv L^{-1} \sum_{l=1}^{L} g(S_{C,l}(B)),$$

where $S_{C,l}(B)$ is defined in Eq. (3.9) and, in this experiment, $L = 1600$. We chose the value $L = 1600$, because the width of the *relative* confidence interval (i.e, the confidence interval one would obtain if the standard error were equal to one) is $4L^{-1/2} = 0.1$.

5.3. Predictands

The predictands $g(S(\cdot))$ under consideration fall into three broad classes. In the first class, the following predictands were included,

$$\{S(D), S(B_1), S(B_2)\},$$

where $S(B)$, for a generic subset B of D, is defined as

$$S(B) \equiv |B|^{-1} \sum_{\mathbf{s} \in B} S(\mathbf{s}), \quad B \subset D,$$

and B_1 and B_2 are shown in Fig. 1. Note that these three predictands are linear functionals of $S(\cdot)$, and ordinary kriging is expected to predict these well, particularly if $S(\cdot)$ is Gaussian.

The second class of predictands consists of

$$\{F_S(q(\alpha)) \colon \alpha = 0.50, 0.75, 0.90\},$$

where

$$F_S(t) \equiv |D|^{-1} \sum_{\mathbf{s}\in D} I(S(\mathbf{s}) \leqslant t), \quad t \in \mathbb{R}$$

and $q(\alpha) \equiv \inf\{t\colon F_S(t) \geqslant \alpha\}$. Observe that $F_S(t)$ is a linear function of univariate nonlinear functionals of $S(\cdot)$. Cressie (1993b) proposed constrained kriging (CK) for predictands of this type, where variances only are matched for each of the summands. Indeed, it is for predictands of this type (i.e., an average of nonlinear functionals) that the approximate unbiasedness property of CK can be very useful in reducing the experimental MSPE. Clearly, CM will also be approximately unbiased but with larger theoretical MSPE than that of CK, since it matches more second-order moments (namely, covariances). It remains to be seen whether these theoretical MSPE inequalities hold in the experiment.

Nonlinear functions of univariate nonlinear functionals of $S(\cdot)$ make up the third class of predictands, and this class includes

$$\{Q_S(q(\alpha))\colon \alpha = 0.50, 0.75, 0.90\},$$

where

$$Q_S(t) \equiv \sum_{\mathbf{s}\in D} S(\mathbf{s}) I(S(\mathbf{s}) > t)/|D|(1 - F_S(t)), \quad t \in \mathbb{R}.$$

In environmental applications, this quantity represents the average value of the pollutant above a threshold. Predictands of this type can only be expressed as some functional of the vector-valued argument $\mathbf{S}$, and hence seem well suited for CM.

5.4. Responses of the experiment

We considered two responses in this experiment: The experimental mean-squared prediction error (MSPE) and the experimental bias of the predictors. Suppose that $\hat{p}(g(S(\cdot)); \mathbf{Z})$ is a generic predictor of the predictand $g(S(\cdot))$. Then the MSPE of $\hat{p}(g(S(\cdot)); \mathbf{Z})$ is

$$\begin{aligned}\mathrm{MSPE}[\hat{p}(g(S(\cdot)); \mathbf{Z})] &\equiv E[\hat{p}(g(S(\cdot)); \mathbf{Z}) - g(S(\cdot))]^2 \\ &= \mathrm{var}[\hat{p}(g(S(\cdot)); \mathbf{Z}) - g(S(\cdot))] \\ &\quad + \{E[\hat{p}(g(S(\cdot)); \mathbf{Z})] - g(S(\cdot))\}^2.\end{aligned}$$

The MSPE and bias of each predictor for all combinations of the factors and predictands of the experiment were estimated by generating 1600 realizations of the appropriate S-processes and ε-processes, yielding the experimental MSPE and the experimental bias:

$$\widehat{\mathrm{MSPE}}[\hat{p}(g(S(\cdot)); \mathbf{Z})] \equiv \frac{1}{1600} \sum_{i=1}^{1600} (\hat{p}(g(S^{(i)}(\cdot)), \mathbf{Z}^{(i)}) - g(S^{(i)}(\cdot)))^2,$$

$$\widehat{\mathrm{bias}}[\hat{p}(g(S(\cdot)); \mathbf{Z})] \equiv \frac{1}{1600} \sum_{i=1}^{1600} (\hat{p}(g(S^{(i)}(\cdot)), \mathbf{Z}^{(i)}) - g(S^{(i)}(\cdot))),$$

where $S^{(i)}(\cdot)$ represents the ith simulation of the random field $S(\cdot)$; $i = 1, \ldots, 1600$, $\mathbf{Z}^{(i)} \equiv (S^{(i)}(\mathbf{s}_1)+\varepsilon^{(i)}(\mathbf{s}_1), \ldots, S^{(i)}(\mathbf{s}_{10})+\varepsilon^{(i)}(\mathbf{s}_{10}))'$ represents the ith simulated data vector; $i=1, \ldots, 1600$, and $\varepsilon^{(i)}(\cdot)$ represents the ith simulation of the measurement-error process; $i = 1, \ldots, 1600$.

5.5. Results of the experiment

In this computer-simulation experiment, all the covariance parameters of $Z(\cdot)$ (i.e., this includes the parameter τ^2) were completely specified. An advantage of specifying these parameters is that a "clean" comparison can be made among the predictors in question, without any confounding effects that may be due to parameter estimation. Since estimated covariance parameters are used in the same empirical-Bayes, "plug-in" fashion for each of the four predictors, we are in effect assuming their influence is the same in each case. Estimation of such parameters using geostatistical methodology is given in Section 6.2.

Prediction of $S(D)$, $S(B_1)$, *and* $S(B_2)$: With respect to the *bias criterion*, all predictors performed well in predicting $S(D)$, $S(B_1)$, and $S(B_2)$, in all cases.

With respect to the *MSPE criterion*, the differences among the predictors were more pronounced. Ordinary kriging (OK) performed best (or nearly best) in every case, as expected. Conditional simulation (CS) compared well with OK, except when the random field was $S_4(\cdot)$ (i.e., bimodal, mixed Gaussian), where it performed worst. Covariance-matching constrained kriging (CM) outperformed constrained kriging (CK) in every case, a welcome but slightly surprising outcome. The differences among the predictors diminished markedly as the strength of spatial dependence increased, to the point where all predictors were comparable when the range was large, namely $a = 30$.

In Table 1 we present, as an example, bias and MSPE results for the prediction of $S(B_2)$, recalling that B_2 is located where the sample is most sparse. We see that even in these cases (where the absolute bias and MSPE were generally larger than in the prediction of $S(D)$ and $S(B_1)$), bias contributed little towards the MSPE.

Prediction of $F_S(q(0.50))$, $F_S(q(0.75))$, *and* $F_S(q(0.90))$: With respect to the *bias criterion*, CK and CM performed best in the Gaussian case, and reasonably well in the non-Gaussian cases. Surprisingly, OK performed well (sometimes best) in some non-Gaussian cases, but CS consistently performed worst.

With respect to the *MSPE criterion*, OK consistently performed worst and CS usually performed best, and the difference between the two predictors was pronounced when the measurement error was large, namely $\tau^2 = 2$. The predictors CM and CK consistently performed very similarly, but with CM usually slightly better. The performance of all predictors often worsened as the strength of spatial dependence, a, increased.

In Table 2 we present, as an example, bias and MSPE results for the prediction of $F_S(q(0.75))$, and these results reflect those of $F_S(q(0.50))$ and $F_S(q(0.90))$ reasonably closely.

Prediction of $Q_S(q(0.50))$, $Q_S(q(0.75))$, *and* $Q_S(q(0.90))$: Observe that $Q_S(t)$ is not defined if $1 - F_S(t) = 0$. Among all predictors with $t = q(0.75)$ and $q(0.90)$, there were many occasions among the 1600 simulations generated where predictions

Table 1
The MSPE and bias of predictors of $S(B_2)$, for all combinations of the experimental factors in the simulation experiment

		$S(B_2)$							
		RF $= S_1(\cdot)$		RF $= S_2(\cdot)$		RF $= S_3(\cdot)$		RF $= S_4(\cdot)$	
		MSPE	Bias	MSPE	Bias	MSPE	Bias	MSPE	Bias
$a = 4.1$	OK	0.587	0.011	0.596	−0.022	0.666	−0.016	0.027	0.001
$\tau^2 = 0.1$	CK	1.886	0.020	2.002	−0.031	2.088	−0.030	0.146	−0.007
	CM	1.588	0.022	1.652	−0.036	1.775	−0.028	0.113	−0.005
	CS	0.730	0.033	0.654	−0.021	0.765	−0.002	0.960	−0.016
$a = 7.5$	OK	0.909	0.009	0.921	−0.001	0.892	−0.028	0.052	0.007
$\tau^2 = 0.1$	CK	1.843	−0.011	2.002	−0.011	1.917	−0.008	0.143	0.006
	CM	1.713	−0.007	1.856	−0.008	1.776	−0.010	0.127	0.006
	CS	0.962	0.004	0.858	0.003	0.838	−0.028	0.220	0.034
$a = 30$	OK	0.378	0.006	0.532	−0.006	0.563	0.011	0.043	−0.004
$\tau^2 = 0.1$	CK	0.548	0.014	0.724	−0.010	0.812	0.007	0.077	−0.009
	CM	0.510	0.013	0.678	−0.010	0.757	0.008	0.070	−0.008
	CS	0.378	0.006	0.530	−0.005	0.561	0.011	0.044	−0.005
$a = 4.1$	OK	0.731	−0.015	0.769	0.020	0.814	0.032	0.230	0.019
$\tau^2 = 2$	CK	1.832	0.005	1.801	0.059	1.807	0.021	0.691	0.041
	CM	1.511	−0.005	1.495	0.049	1.512	0.021	0.524	0.032
	CS	0.805	−0.023	0.780	0.016	0.783	0.041	1.039	−0.055
$a = 7.5$	OK	0.786	−0.014	0.732	−0.025	0.787	0.010	0.285	−0.006
$\tau^2 = 2$	CK	1.508	−0.006	1.405	−0.033	1.521	−0.028	0.648	−0.019
	CM	1.354	−0.009	1.258	−0.029	1.363	−0.020	0.561	−0.016
	CS	0.870	−0.017	0.756	−0.022	0.771	0.008	0.635	−0.016
$a = 30$	OK	0.391	0.008	0.380	−0.002	0.687	0.034	0.234	−0.012
$\tau^2 = 2$	CK	0.424	0.008	0.408	−0.002	0.737	0.035	0.252	−0.013
	CM	0.372	0.008	0.364	−0.002	0.657	0.033	0.224	−0.011
	CS	0.365	0.002	0.343	−0.011	0.776	0.037	0.204	−0.006

Ordinary kriging is denoted "OK", constrained kriging "CK", covariance-matching constrained kriging "CM", and conditional simulation "CS"; the abbreviation "RF" refers to random field.

of $1 - F_S(t)$ were found to be zero; these simulations were excluded from the analysis. Consequently, estimates of bias and MSPE of predictors of $Q_S(t)$ may be optimistic in those cases where predictions of $1 - F_S(t)$ were zero for a non-negligible proportion of simulations. In all cases, the proportion of all simulations for which the predictions of $1 - F_S(t)$ were found to be zero, was consistently highest for OK, next highest for CK, lower for CM, and lowest for CS.

With respect to the *bias criterion*, OK performed worst when the random field was $S_1(\cdot)$ or $S_2(\cdot)$, and best for $S_3(\cdot)$ or $S_4(\cdot)$. Conversely, CS performed well (sometimes best) when the random field was $S_1(\cdot)$ or $S_2(\cdot)$, and worst for $S_3(\cdot)$ or $S_4(\cdot)$. The predictors CM and CK consistently performed very similarly; they performed well for

Table 2
Similar to Table 1, except the MSPE and bias of predictors of $F_S(q(0.75))$ for all combinations of the experimental factors, are given

		$F_S(q(0.75))$							
		RF $= S_1(\cdot)$		RF $= S_2(\cdot)$		RF $= S_3(\cdot)$		RF $= S_4(\cdot)$	
		MSPE	Bias	MSPE	Bias	MSPE	Bias	MSPE	Bias
$a = 4.1$	OK	0.031	−0.126	0.059	0.032	0.034	−0.136	0.060	0.049
$\tau^2 = 0.1$	CK	0.022	0.006	0.032	0.065	0.025	0.037	0.052	0.138
	CM	0.020	−0.004	0.035	0.075	0.021	0.025	0.050	0.134
	CS	0.008	0.014	0.038	0.165	0.006	0.060	0.078	0.198
$a = 7.5$	OK	0.024	−0.078	0.046	0.050	0.031	−0.062	0.068	0.103
$\tau^2 = 0.1$	CK	0.020	0.002	0.039	0.079	0.031	0.041	0.063	0.141
	CM	0.018	0.003	0.039	0.084	0.028	0.042	0.062	0.142
	CS	0.010	0.017	0.038	0.145	0.012	0.066	0.052	0.194
$a = 30$	OK	0.028	−0.062	0.057	0.053	0.053	0.010	0.096	0.148
$\tau^2 = 0.1$	CK	0.026	−0.002	0.052	0.082	0.053	0.068	0.094	0.174
	CM	0.025	−0.001	0.050	0.084	0.051	0.068	0.092	0.175
	CS	0.014	0.024	0.038	0.122	0.028	0.087	0.056	0.204
$a = 4.1$	OK	0.081	−0.086	0.153	0.114	0.098	−0.032	0.179	0.169
$\tau^2 = 2$	CK	0.048	0.011	0.085	0.144	0.063	0.060	0.106	0.196
	CM	0.046	0.001	0.086	0.144	0.061	0.054	0.108	0.194
	CS	0.011	0.027	0.047	0.185	0.006	0.061	0.119	0.207
$a = 7.5$	OK	0.077	−0.053	0.153	0.144	0.100	0.013	0.173	0.194
$\tau^2 = 2$	CK	0.055	0.010	0.107	0.166	0.077	0.078	0.129	0.210
	CM	0.053	0.010	0.105	0.167	0.075	0.079	0.126	0.210
	CS	0.024	0.043	0.062	0.188	0.015	0.078	0.098	0.209
$a = 30$	OK	0.144	0.034	0.198	0.146	0.169	0.081	0.230	0.220
$\tau^2 = 2$	CK	0.138	0.037	0.193	0.148	0.164	0.084	0.222	0.221
	CM	0.132	0.038	0.183	0.148	0.155	0.083	0.214	0.222
	CS	0.056	0.100	0.088	0.185	0.052	0.117	0.102	0.229

$S_1(\cdot)$ or $S_2(\cdot)$, and adequately or well (certainly superior to CS) for $S_3(\cdot)$ and $S_4(\cdot)$. The differences among the predictors diminished markedly as a increased, to the point where all predictors were comparable when the strength of spatial dependence was large, namely $a = 30$.

With respect to the *MSPE criterion*, the predictors performed almost as they did in terms of bias. That is, OK performed worst for $S_1(\cdot)$ or $S_2(\cdot)$, and best for $S_3(\cdot)$ or $S_4(\cdot)$; CS performed best for $S_1(\cdot)$ and some cases of $S_2(\cdot)$, and worst for $S_3(\cdot)$ or $S_4(\cdot)$. The predictor CM, which usually outperformed CK, performed well in most cases (sometimes best) and never performed worst.

In Table 3 we present, as an example, bias and MSPE results for the prediction of $Q_S(q(0.75))$, and these results reflect those of $Q_S(q(0.50))$ and $Q_S(q(0.90))$ well. In

Table 3
Similar to Table 1, except the MSPE and bias of predictors of $Q_S(q(0.75))$ conditional on $1 - F_S(q(0.75)) > 0$, for all combinations of the experimental factors, are given

		$Q_S(q(0.75))$, conditional on $1 - F_S(q(0.75)) > 0$							
		RF $= S_1(\cdot)$		RF $= S_2(\cdot)$		RF $= S_3(\cdot)$		RF $= S_4(\cdot)$	
		MSPE	Bias	MSPE	Bias	MSPE	Bias	MSPE	Bias
$a = 4.1$	OK	0.089	0.241	0.489	0.533	0.008	−0.049	0.005	−0.038
$\tau^2 = 0.1$	CK	0.080	0.042	0.488	0.177	0.039	−0.153	0.028	−0.131
	CM	0.063	0.073	0.408	0.260	0.055	−0.179	0.024	−0.122
	CS	0.021	−0.004	0.287	0.205	0.190	−0.434	0.416	−0.577
$a = 7.5$	OK	0.063	0.188	0.366	0.345	0.009	−0.042	0.009	−0.058
$\tau^2 = 0.1$	CK	0.051	0.055	0.406	0.181	0.041	−0.139	0.022	−0.113
	CM	0.047	0.053	0.368	0.196	0.052	−0.166	0.023	−0.117
	CS	0.022	−0.009	0.334	0.144	0.122	−0.340	0.188	−0.416
$a = 30$	OK	0.020	0.100	0.085	0.115	0.011	−0.020	0.009	−0.070
$\tau^2 = 0.1$	CK	0.016	0.043	0.059	0.055	0.019	−0.072	0.018	−0.104
	CM	0.015	0.037	0.060	0.056	0.019	−0.080	0.018	−0.110
	CS	0.008	−0.010	0.094	0.003	0.038	−0.162	0.060	−0.238
$a = 4.1$	OK	0.148	0.331	0.663	0.596	0.038	−0.125	0.076	−0.185
$\tau^2 = 2$	CK	0.102	0.079	0.420	0.293	0.204	−0.361	0.252	−0.426
	CM	0.089	0.107	0.429	0.338	0.172	−0.337	0.220	−0.401
	CS	0.029	−0.015	0.391	0.206	0.252	−0.500	0.718	−0.730
$a = 7.5$	OK	0.115	0.253	0.483	0.374	0.061	−0.153	0.099	−0.222
$\tau^2 = 2$	CK	0.088	0.088	0.382	0.200	0.163	−0.302	0.212	−0.376
	CM	0.079	0.081	0.382	0.204	0.156	−0.306	0.211	−0.381
	CS	0.041	−0.037	0.480	0.103	0.202	−0.441	0.488	−0.619
$a = 30$	OK	0.051	0.047	0.142	0.007	0.084	−0.150	0.114	−0.233
$\tau^2 = 2$	CK	0.051	0.034	0.139	−0.012	0.088	−0.167	0.119	−0.244
	CM	0.047	0.042	0.132	−0.003	0.079	−0.152	0.112	−0.236
	CS	0.043	−0.107	0.195	−0.128	0.105	−0.283	0.207	−0.421

addition, as a visual aid to fix ideas, plots in terms of the squared-bias and MSPE results for the prediction of $Q_S(q(0.75))$ are given in Fig. 2.

5.6. Summary

Several general conclusions can be made from this computer-simulation experiment:

1. For the prediction of linear predictands, all predictors had excellent unbiasedness properties, but ordinary kriging (OK) is the preferred predictor in all cases, because of its superior performance in terms of the MSPE. The MSPE performance of covariance-matching constrained kriging (CM) was always quite good, and

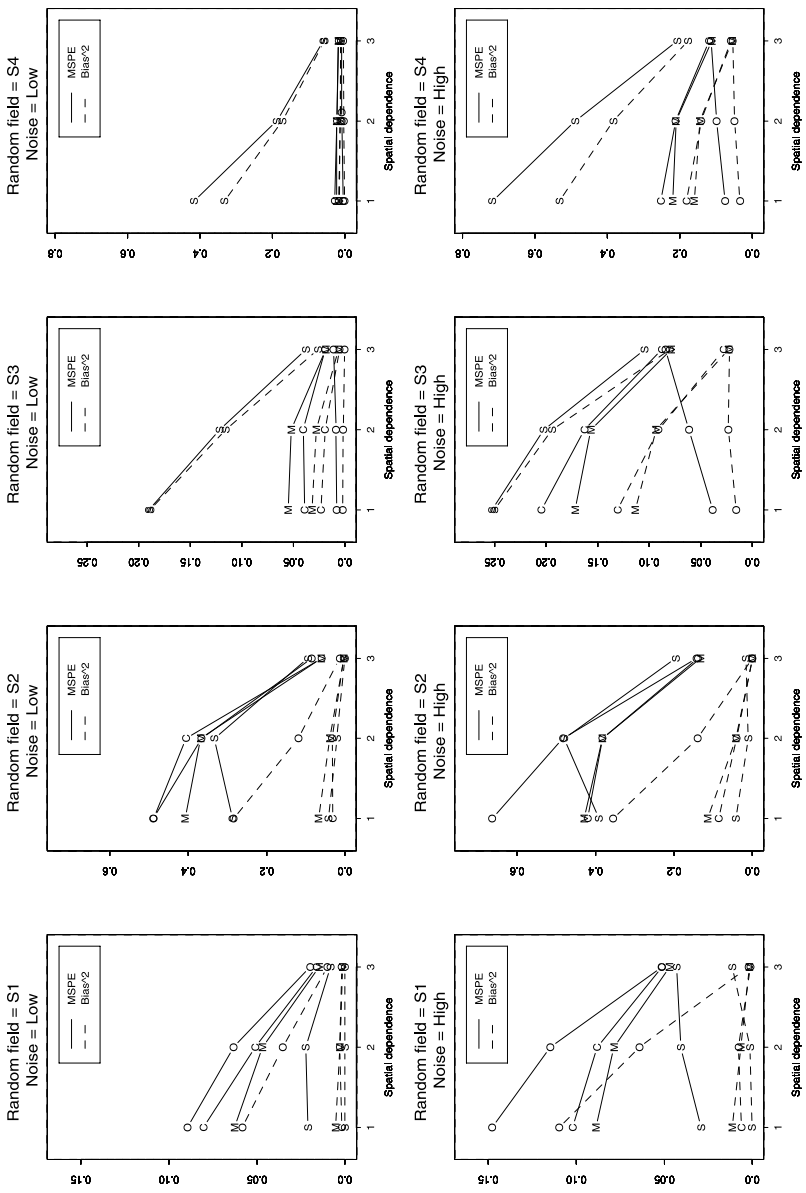

Fig. 2. The MSPE and squared bias of predictors of $Q_S(q(0.75))$ conditional on $1 - F_S(q(0.75)) > 0$, for all combinations of the experimental factors. Shown are Ordinary kriging (O), Constrained kriging (C), covariance-Matching constrained kriging (M), and conditional Simulation (S). The notation "1", "2", and "3", represents the three values of the range a, in increasing order.

consistently better than constrained kriging (CK); on the other hand, conditional simulation (CS) had MSPEs that were sometimes the best and sometimes the worst.

2. For the prediction of $F_S(t)$, an average of univariate nonlinear predictands, the unbiasedness properties of CK and CM were best in the Gaussian case and good in the non-Gaussian cases. Although OK showed surprisingly good unbiasedness properties in some cases, its relatively high MSPEs in all cases rule it out as a potential predictor here. And, while CS usually displayed very good MSPE properties, it usually also displayed the worst bias.
3. For the prediction of $Q_S(t)$, a nonlinear function of nonlinear predictands, CM and CK performed well in terms of bias, except in some non-Gaussian cases with noisy data. CM usually outperformed CK with respect to the MSPE. The performance of OK and CS was highly dependent on the nature of the random field; in some cases, OK performed best and CS performed worst with respect to bias and the MSPE, and in other cases the converse was true.
4. CM and CK usually performed fairly similarly, irrespective of predictand, but CM usually had the smaller MSPE, which was a welcome bonus.
5. In the prediction of $S(B)$ or $Q_S(t)$, the four predictors performed similarly when the spatial-dependence parameter a was large.

In conclusion, covariance-matching constrained kriging (CM) offers an attractive alternative to ordinary kriging (OK) and conditional simulation (CS), especially when the distribution of the underlying process is far from Gaussian (e.g., bimodal mixed Gaussian) and the predictand possesses some degree of nonlinearity. Moreover, while it is tempting to treat the realizations from CS as "typical" spatial surfaces, the performance of CS in this experiment shows that it cannot be trusted in a number of important cases (see also Gotway, 1994).

6. Spatial analysis of particulate matter in the Pittsburgh area

In this section, we shall apply the four predictors, encountered in Section 5, to the analysis of PM_{10} (i.e., particulate matter with particles smaller than 10 μm, in aerodynamic diameter, per cubic meter of air; see Section 1) over the Pittsburgh area. In 1996, PM_{10} was monitored at 27 sites in the Pittsburgh area. The monitoring stations were scattered throughout the Pittsburgh Metropolitan Area and environs. At the 27 sites, four different monitoring devices (probe types) were used, at several sites more than one probe type was used, and at two sites the same probe type was duplicated. In addition, the monitoring intervals varied from daily to 3 to 6-day (synchronous) intervals.

6.1. Exploratory data analysis

The exploratory data analysis, which is described in more detail in Cressie et al. (1999), consisted of two phases. In the first phase, the analysis focused on differences among the probe types and on appropriate transformations of the data to stabilize the

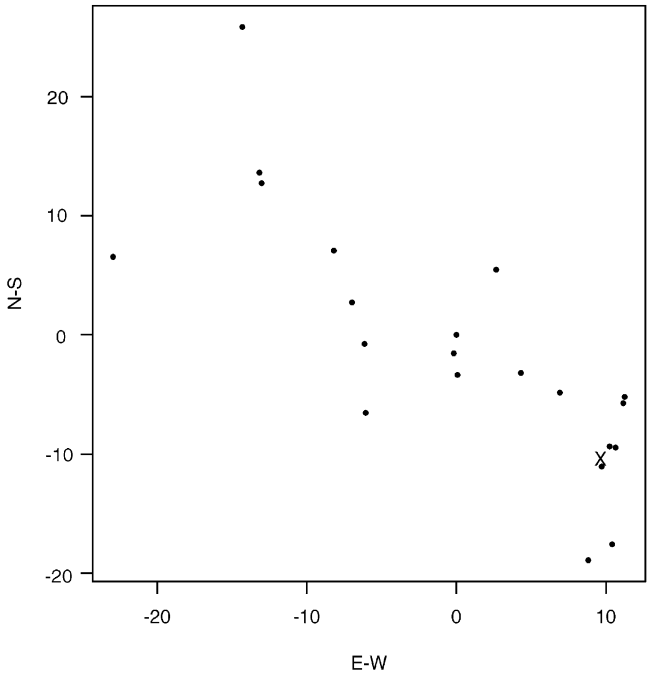

Fig. 3. Map of configuration of 22 sites in the Pittsburgh area on day 226 (August 13), 1996. The site marked "x" corresponds to the location of two spatial outliers. The horizontal axis is in the east–west direction, the vertical axis is in the north–south direction, and the units of both axes are in miles.

variance and obtain an approximately symmetric distribution; this analysis was based on all the data recorded in 1996.

Observations corresponding to the two most frequent probe types were compared (the other two probe types were too poorly represented for any meaningful comparisons), from which it was concluded that all probe types were essentially measuring the same quantity with the same measurement error, resulting in more monitoring sites with essentially replicate observations. After considering various transformations from among the power family of transformations, a log transformation was found to perform well in stabilizing the variance and allowing a Gaussian distribution as a model for the data.

In the second phase, an exploratory *spatial* data analysis was performed on the data corresponding to a particular day, August 13, 1996 (day 226), upon which a detailed spatial analysis was done for illustrative purposes. Day 226 was chosen because a graphical analysis of the corresponding data exhibited a spatial-dependence structure sufficiently interesting to warrant a more detailed analysis. On this day, PM_{10} was monitored at 22 sites, eight of which had duplicate measurements. Fig. 3 shows a map of the configuration of these 22 sites, centered near Avalon, about 5 miles northwest of downtown Pittsburgh, and extending about 20 miles in each of the compass directions. At the scale of the problem, the earth's surface is approximately flat and the units on each axis are given in miles.

An analysis of semivariogram clouds detected two outlying values (i.e., unusual values relative to nearby values) that corresponded to a single site marked with an "x" in Fig. 3. This site (site 3004) was removed for variogram estimation (but not for prediction). Further analysis of semivariogram clouds and semivariogram estimates (with site 3004 removed) in several directions indicated that the spatial variability was *anisotropic* (i.e., *not* homogeneous in all directions), and that the spatial correlation appeared to be strongest in the ENE–WSW direction (i.e., 67.5° clockwise from N) and weakest in the NNE–SSW direction (i.e., 157.5° clockwise from N).

6.2. Modeling the semivariogram

Define

$$S(\mathbf{s}) \equiv \log \mathrm{PM}_{10}(\mathbf{s}), \quad \mathbf{s} \in D,$$

where D represents a two-dimensional convex region surrounding Pittsburgh and containing the 22 sampling locations. Assuming that the measurement error is normal and additive on the log-transformed scale, we obtain the measurement-error spatial model,

$$Z(\mathbf{s}) \equiv S(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad \mathbf{s} \in D,$$

where $Z(\cdot)$ represents the set of actual (and potential) observations on the log-transformed scale, and for each $\mathbf{s} \in D$, $\varepsilon(\mathbf{s})$ is distributed identically and independently as $N(0, \tau^2)$. In addition, $S(\cdot)$ and $\varepsilon(\cdot)$ are assumed to be independent. As we have seen in Section 2, this is a spatial empirical Bayes model. We now focus on the "empirical" part, namely estimation of the model parameters.

Nonparametric estimates of the semivariogram of the data (excluding site 3004) were calculated (see Cressie, 1993a, Section 2.4), yielding $\{\hat{\gamma}_z(\mathbf{h}_i)\colon i = 1, \ldots, H\}$, where $\mathbf{h}_i$; $i = 1, \ldots, H$, are lag vectors taken in only two directions, 67.5° and 157.5°, each with an angle tolerance of ±45° and a lag tolerance of 1.5 miles (i.e., any distance between pairs that lie within ±1.5 miles of $\|\mathbf{h}_i\|$ was considered to be of length $\|\mathbf{h}_i\|; = 1, \ldots, H$).

A semivariogram is said to possess *geometric anisotropy* if a simple rotation and scaling of the original axes exist such that under the transformed axes the semivariogram is isotropic (i.e., a function of the length $\|\mathbf{h}\|$ of the spatial lag $\mathbf{h}$). Assuming geometric anisotropy for the directional semivariograms, Journel and Huijbregts (1978, pp. 179–181) give the matrix that performs the appropriate rotation and scaling of the original axes as

$$\mathbf{A} = \begin{bmatrix} \cos^2\phi + \lambda \sin^2\phi & (1-\lambda)\sin\phi\cos\phi \\ (1-\lambda)\sin\phi\cos\phi & \sin^2\phi + \lambda\cos^2\phi \end{bmatrix}, \tag{6.1}$$

where ϕ represents the angle of rotation (clockwise) required to align the original N–S axis with the direction in which the spatial correlation is weakest. From the discussion above, $\phi = 157.5°$. The anisotropy-ratio parameter λ represents the scaling of the axes required for isotropy under the new axes, which is estimated by weighted least squares below.

Therefore, consider the following parametric semivariogram model:

$$\gamma_z(\mathbf{h}; \boldsymbol{\theta}) \equiv \gamma_z^o(\|\mathbf{A}\mathbf{h}\|; \boldsymbol{\theta}), \quad \mathbf{h} \in \mathbb{R}^2,$$

where $\mathbf{A}$ is given by (6.1). The parameters in $\boldsymbol{\theta}$ are fitted by the method of weighted least squares, proposed by Cressie (1985). That is, from nonparametric estimates $\{\hat{\gamma}_z^o(\|\mathbf{A}\mathbf{h}_i\|)\colon i=1,\ldots,H\}$, we choose $\hat{\boldsymbol{\theta}}$ in $\gamma_z^o(\|\mathbf{A}(\lambda)\mathbf{h}\|;\boldsymbol{\theta})$ to minimize

$$\sum_{i=1}^{H} N(\|\mathbf{A}(\lambda)\mathbf{h}_i\|)\left\{\frac{\hat{\gamma}_z^o(\|\mathbf{A}(\lambda)\mathbf{h}\|)}{\gamma_z^o(\|\mathbf{A}(\lambda)\mathbf{h}\|;\boldsymbol{\theta})}-1\right\}^2,$$

with respect to $\boldsymbol{\theta}\equiv(\lambda,\sigma_\eta^2,\sigma_W^2,a)'$, where we have chosen to feature the anisotropy-ratio parameter λ in $\mathbf{A}(\lambda)$; $N(\|\mathbf{A}(\lambda)\mathbf{h}_i\|)$ is the number of pairs of sites whose distance (after the axes have been transformed by $\mathbf{A}(\lambda)$) lies within a tolerance of 1.5 miles of $\|\mathbf{A}(\lambda)\mathbf{h}_i\|$; and the parameters σ_η^2, σ_W^2, and a are parameters of the spherical semivariogram model (see Cressie, 1993a, p. 61). This results in the estimated (geometrically anisotropic) semivariogram model given by

$$\gamma_z(\mathbf{h};\hat{\boldsymbol{\theta}})=\begin{cases}\hat{\sigma}_\eta^2+\hat{\sigma}_W^2, & \|\mathbf{A}(\hat{\lambda})\mathbf{h}\|>\hat{a},\\ \hat{\sigma}_\eta^2+\hat{\sigma}_W^2\{\frac{3}{2}(\|\mathbf{A}(\hat{\lambda})\mathbf{h}\|/\hat{a})-\frac{1}{2}(\|\mathbf{A}(\hat{\lambda})\mathbf{h}\|/\hat{a})^3\}, & 0<\|\mathbf{A}(\hat{\lambda})\mathbf{h}\|\leqslant\hat{a},\\ 0, & \mathbf{h}=\mathbf{0},\end{cases}$$

where $\hat{\lambda}=0.46923$, $\hat{\sigma}_\eta^2=0.00726$, $\hat{\sigma}_W^2=0.05188$, and $\hat{a}=22.720$. This semivariogram, and the nonparametric estimates to which it was fitted, is displayed in Fig. 4.

Finally, from the seven duplicates, we can estimate the measurement-error variance τ^2 as $\hat{\tau}^2=0.00706$, which allows us to construct $\boldsymbol{\Sigma}$ and $\mathbf{c}(B)$, defined in (3.2), for prediction. For example, we estimate $\mathrm{var}[Z(\mathbf{s}_i)]$ as $\hat{\sigma}_\eta^2+\hat{\sigma}_W^2=0.05914$, and $\mathrm{cov}[Z_1(\mathbf{s}_i),Z_2(\mathbf{s}_i)]$ as $\hat{\sigma}_\eta^2+\hat{\sigma}_W^2-\hat{\tau}^2\equiv\hat{\sigma}^2=0.05208$, where $Z_1(\mathbf{s}_i)$ and $Z_2(\mathbf{s}_i)$ represent duplicates from site $\mathbf{s}_i$, and recall that $\mathrm{var}[S(\mathbf{s})]=\sigma_\eta^2+\sigma_W^2-\tau^2\equiv\sigma^2$. It should be noted that should prior distributions be put on the parameters λ, σ_η^2, σ_W^2, a, and τ^2, the empirical Bayes model becomes a fully Bayesian model.

6.3. *Spatial prediction of* PM_{10}

Define

$$X(\mathbf{s})\equiv\exp(S(\mathbf{s})),\qquad \mathbf{s}\in D,$$

where $X(\mathbf{s})$ is the true PM_{10} value devoid of measurement error, and D represents a two-dimensional convex region surrounding Pittsburgh and containing the 22 sampling locations that reported on day 226, 1996.

In this section, we address several prediction objectives. A "smooth" surface of PM_{10} values is predicted to provide some idea of the nature of PM_{10} over D. Observed and predicted values of PM_{10} indicate that the region D is well within PM_{10} compliance (i.e., $\mathrm{PM}_{10}(\mathbf{s})\leqslant 150\ \mu\mathrm{g/m}^3$, for all $\mathbf{s}\in D$), so prediction of the proportion of D exceeding $150\ \mu\mathrm{g/m}^3$ is expected to be zero. However, for illustrative purposes, two arbitrary cutoff values (both within the range of observed values) were specified, and the proportion of D exceeding those two values is predicted. Finally, the average amount of PM_{10} that exceeds/does not exceed the two cutoff values are also predicted. The two

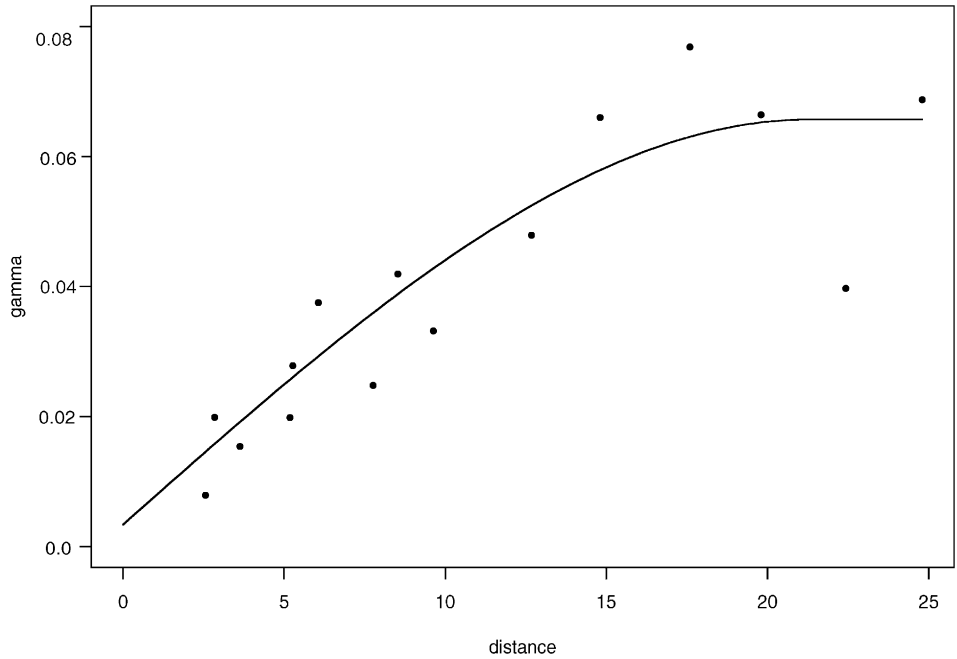

Fig. 4. Fitted isotropic spherical semivariogram. The lag distance is the scalar quantity $\|\mathbf{Ah}\|$, where $\mathbf{h}$ is the spatial lag vector between sites in the original co-ordinate system. The horizontal axis represents lag distance in miles, and the vertical axis represents fitted semivariogram values.

outliers discussed in Section 6.1 were included in the prediction part of this study since, although it is often justifiable to exclude unusual observations from over-influencing a global variogram model, local effects are usually of interest when predicting various features of $X(\cdot)$.

For mapping the PM_{10} values, we created a triangular grid of 279 prediction locations, $\{\mathbf{u}_i\colon i = 1,\ldots,279\}$, at which $X(\cdot)$ was predicted by ordinary kriging (OK), constrained kriging (CK), covariance-matching constrained kriging (CM), and conditional simulation (CS).

Define the OK predictor of $X(\mathbf{s}_0)$ as

$$\hat{X}_{\mathrm{ok}}(\mathbf{s}_0) \equiv \exp(\hat{S}_{\mathrm{ok}}(\mathbf{s}_0)), \qquad \mathbf{s}_0 \in D,$$

where $\hat{S}_{\mathrm{ok}}(\mathbf{s}_0)$ is the OK predictor of $S(\mathbf{s}_0)$. Observe that the ordinary kriging predictor of $X(\mathbf{s}_0)$ is biased, but since $X(\cdot)$ is a log Gaussian process, a bias correction can be applied to the predictor and the MSPE of the corrected predictor is easy to obtain (e.g., Cressie, 1993a, Section 3.2.2). Suppose that $E[S(\mathbf{s})] \equiv \mu$, $\mathrm{var}[S(\mathbf{s})] \equiv \sigma^2$, and $\mathrm{var}[\mathbf{Z}] \equiv \mathbf{\Sigma}$. Then, adjusting for measurement error, the bias-corrected OK predictor of $X(\mathbf{s}_0)$ is given as follows:

$$\check{X}_{\mathrm{ok}}(\mathbf{s}_0) \equiv \exp\{\hat{S}_{\mathrm{ok}}(\mathbf{s}_0) + (\sigma^2 - \mathrm{var}[\hat{S}_{\mathrm{ok}}(\mathbf{s}_0)])/2\}$$

and the MSPE of $\check{X}_{\mathrm{ok}}(\mathbf{s}_0)$ is given by

$$\mathrm{MSPE}[\check{X}_{\mathrm{ok}}(\mathbf{s}_0)] \equiv \{\exp(2\mu+\sigma^2)\} \times\{\exp(\sigma^2)+\exp(\mathrm{var}[\hat{S}_{\mathrm{ok}}(\mathbf{s}_0)])-2\exp(\mathrm{cov}[S(\mathbf{s}_0),\hat{S}_{\mathrm{ok}}(\mathbf{s}_0)])\}.$$

Such a bias correction is not generally available for other nonlinear functionals of $S(\cdot)$. Thus, using the OK predictor $\check{X}_{\mathrm{ok}}$ instead of $\hat{X}_{\mathrm{ok}}$ is being generous to the methodology. (No bias correction was made in the simulation experiment in Section 5.)

Define the CK predictor of $X(\mathbf{s}_0)$ as

$$\hat{X}_{\mathrm{ck}}(\mathbf{s}_0) \equiv \exp(\hat{S}_{\mathrm{ck}}(\mathbf{s}_0)), \quad \mathbf{s}_0 \in D,$$

where $\hat{S}_{\mathrm{ck}}(\mathbf{s}_0)$ is the CK predictor of $S(\mathbf{s}_0)$. Observe that since $\exp(\cdot)$ is a smooth function, the CK predictor of $X(\mathbf{s}_0)$ is at least approximately unbiased, whatever the distribution of $S(\cdot)$. However, under the model that $S(\cdot)$ is Gaussian, the CK predictor $\hat{X}_{\mathrm{ck}}(\mathbf{s}_0)$ is exactly unbiased. It is not difficult to show that its MSPE is given by

$$\mathrm{MSPE}(\hat{X}_{\mathrm{ck}}(\mathbf{s}_0)) \equiv 2\{\exp(2\mu+\sigma^2)\}\{\exp(\sigma^2)-\exp(\mathrm{cov}[S(\mathbf{s}_0),\hat{S}_{\mathrm{ck}}(\mathbf{s}_0)])\}.$$

Define the "partitioned" CM predictor of $\mathbf{X} \equiv (\exp(S(\mathbf{s}_1)),\ldots,\exp(S(\mathbf{s}_{279})))'$ as

$$\hat{\mathbf{X}}_{\mathrm{pc}} \equiv (\exp(\hat{S}_{\mathrm{pc}}(\mathbf{s}_1)),\ldots,\exp(\hat{S}_{\mathrm{pc}}(\mathbf{s}_{279})))',$$

where $\hat{S}_{\mathrm{pc}}(\mathbf{s}_i)$ denotes the ith element of $\hat{\mathbf{S}}_{\mathrm{pc}}$; $i=1,\ldots,279$, and $\hat{\mathbf{S}}_{\mathrm{pc}}$ is the "partitioned" CM predictor of $\mathbf{S} \equiv (S(\mathbf{s}_1),\ldots,S(\mathbf{s}_{279}))'$, partitioned into 93 subvectors each of length 3; see (4.7). (The set of configurations of triangular locations within which covariances were matched are $\{(6.76,-20.80),(7.76,-19.06),(8.76,-20.80)\}$, $\{(9.76,-19.06),(10.76,-20.80),(11.76,-19.06)\}$, and so on.) Denote the ith element of $\hat{\mathbf{X}}_{\mathrm{pc}}$ as $\hat{X}_{\mathrm{pc}}(\mathbf{s}_i)$. Then, it is not difficult to show that $\hat{X}_{\mathrm{pc}}(\cdot)$ is exactly unbiased for $X(\cdot)$, and

$$\mathrm{MSPE}(\hat{X}_{\mathrm{pc}}(\mathbf{s}_i)) \equiv 2\{\exp(2\mu+\sigma^2)\}\{\exp(\sigma^2)-\exp(\mathrm{cov}[\hat{S}_{\mathrm{pc}}(\mathbf{s}_i),S(\mathbf{s}_i)])\}; \quad i=1,\ldots,279,$$

where $\mathrm{cov}[\hat{S}_{\mathrm{pc}}(\mathbf{s}_i),S(\mathbf{s}_i)]$ is the ith diagonal element of $\mathrm{cov}[\hat{\mathbf{S}}_{\mathrm{pc}},\mathbf{S}]$. From (4.6), we can show that

$$\mathrm{cov}[\hat{\mathbf{S}}_{\mathrm{pc}},\mathbf{S}] = \mathbf{1}_{279}\{1/(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})\}\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{C} + \mathbf{P}'_{p1}\mathbf{Q}_{p1},$$

where $\mathbf{1}_{279}$ is a vector of 279 ones, $\mathbf{1}$ is the usual vector of $n(=30)$ ones, and $\mathbf{P}_{p1}$ and $\mathbf{Q}_{p1}$ are symmetric matrices obtained by the spectral decomposition method as given by (5.2).

Define the CS predictor of $X(\mathbf{s}_0)$ as

$$\hat{X}_{\mathrm{cs}}(\mathbf{s}_0) \equiv \frac{1}{L}\sum_{l=1}^{L}\exp(S_{C,l}(\mathbf{s}_0)), \quad \mathbf{s}_0 \in D,$$

where $S_{C,l}(\mathbf{s}_0)$ is defined in Eq. (3.9). Observe that since all first and second moments of each realization of the conditional simulation match those of $S(\mathbf{s}_0)$ and $S(\cdot)$ is Gaussian, then $\hat{X}_{\mathrm{cs}}(\mathbf{s}_0)$ is unbiased. Now, for L large, we expect $\hat{X}_{\mathrm{cs}}(\cdot)$ to be like $\check{X}_{\mathrm{ok}}(\cdot)$, modulo the effect of estimating the unknown mean μ.

It is not difficult to show that its MSPE is given by

$$\mathrm{MSPE}[\hat{X}_{\mathrm{cs}}(\mathbf{s}_0)] \equiv (1+1/L)\{\exp(2\mu+\sigma^2)\}\{\exp(\sigma^2)-\exp(\mathbf{c}(\mathbf{s}_0)'\boldsymbol{\Sigma}^{-1}\mathbf{c}(\mathbf{s}_0))\},$$

where $\mathbf{c}(\mathbf{s}_0) = \mathrm{cov}[\mathbf{Z}, S(\mathbf{s}_0)]$. For L large, this MSPE is slightly more than half the size of the MSPE that would result from a predictor based on only one realization of the conditional simulation.

Rather than choose a large value of L (e.g., $L = 1600$ was chosen in the simulation of Section 5), we shall calibrate CS to CM as follows. Define the spatial variance of any surface $X^*(\cdot)$ as

$$\mathrm{SV}(X^*) \equiv \frac{1}{|B|}\int_B (X^*(\mathbf{u}) - \bar{X}^*)^2\,\mathrm{d}\mathbf{u},$$

where $\bar{X}^* \equiv \frac{1}{|B|}\int_B X^*(\mathbf{u})\,\mathrm{d}\mathbf{u}$. We solve for L in:

$$\mathrm{SV}(\hat{X}_{\mathrm{cs}}) = \mathrm{SV}(\hat{X}_{\mathrm{cm}}), \tag{6.2}$$

and obtain L=10. (Note that L satisfying (6.2) may vary from one set of CS to another.) Choice of this value of L allows us to compare qualitatively the prediction surfaces of CS and CM. Any difference will be due to the inherently different approaches of the two methods. (We already know that a large value of L will yield a surface like that of OK.)

Contour maps of the bias-corrected OK predictions and root MSPEs, the CK predictions and root MSPEs, the CM predictions and root MSPEs, and the CS predictions and root MSPEs (with $L = 10$) of PM_{10} over D are presented in Figs. 5–8, respectively. The sampling sites and the prediction sites are shown in each of the figures. Notice that the map of the OK predictions appears to be a little smoother than that of CK which, in turn, is somewhat smoother than that of CS. But the remarkable difference is between these rather smooth maps and that of CM, which shows features of local peaks and valleys. Indeed, the CM-prediction contour map in Fig. 7 is beginning to take some semblance of a single realization from a conditional simulation! Yet, we have constructed it to have MSPE-optimality properties. We appear to have the best of both worlds: a more realistic-looking map, but with MSPEs that have been minimized (subject to covariance-matching constraints). Also notice that the root MSPEs of the four predictors are fairly similar: those of OK are the smallest and those of CM are the largest. We observe that the results presented in Figs. 5–8 are consistent with the theory established in Section 4.

Inspection of Figs. 5–8 shows that the region D is well within PM_{10} compliance on day 226, 1996, since the largest predicted value is well below the 24-hour standard of 150 $\mu\mathrm{g/m}^3$. However, for illustrative purposes, suppose that we wish to know the proportion of D for which PM_{10} exceeds the cutoff value of 28.75 $\mu\mathrm{g/m}^3$ (third quartile of the data). The predictand describing this quantity is defined as

$$F_X(28.75) \equiv \frac{1}{|D|}\sum_{\mathbf{s}\in D} I(\exp(S(\mathbf{s})) \leqslant 28.75). \tag{6.3}$$

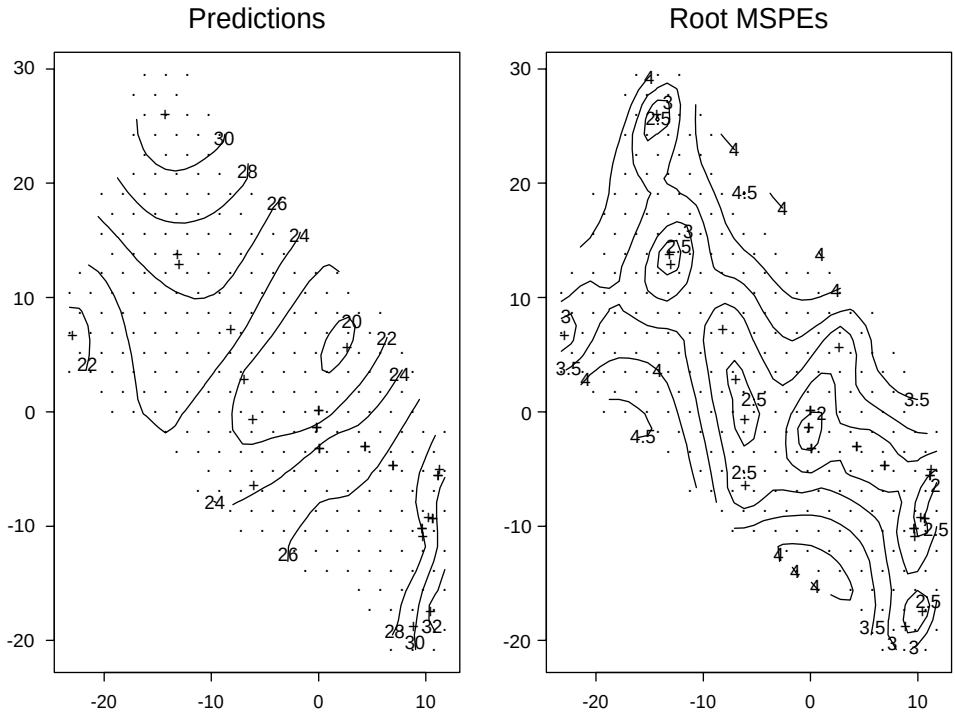

Fig. 5. Contour maps of bias-corrected ordinary kriging predictions and root MSPEs of PM_{10}. There are 279 prediction locations marked ".", and 22 sampling locations marked "+".

Further, suppose we wish to predict the average amount of PM_{10} that exceeds the cutoff value. The predictand describing this quantity can be written as

$$Q_X(28.75) \equiv \frac{1}{|D|} \sum_{\mathbf{s} \in D} \exp(S(\mathbf{s}))I(\exp(S(\mathbf{s})) > 28.75)/(1 - F_X(28.75)). \tag{6.4}$$

Ordinary kriging (OK), constrained kriging (CK), covariance-matching constrained kriging (CM), and conditional simulation for a given L (CS(L)) were used to predict (6.3) and (6.4). In predicting $F_X(28.75)$, OK, CK, CM, CS(10), and CS(1600) yielded the values 0.853, 0.810, 0.860, 0.824, and 0.828, respectively. In predicting $Q_X(28.75)$, OK, CK, CM, CS(10), and CS(1600) yielded the values 30.68, 31.77, 31.56, 31.13, and 30.95, respectively.

Which prediction do we choose in each case? We believe that $S(\cdot)$ is a Gaussian spatial process, that $\hat{\tau}^2 = 0.007$ is small compared with $\hat{\sigma}^2 = 0.052$, and that $\hat{a} = 22.7$ suggests a fairly strong spatial dependence structure in the signal. Armed with this knowledge and based on the results of the simulation experiment in Section 5, we should rule out OK as a candidate in both cases. Indeed, by checking the cell in Table 2 (where $F_S(q(0.75))$ is the predictand) and the cell in Table 3 (where $Q_S(q(0.75))$ is the predictand) for which the random field is Gaussian (i.e., $S(\cdot) = S_1(\cdot)$), $\tau^2 = 0.1$, and

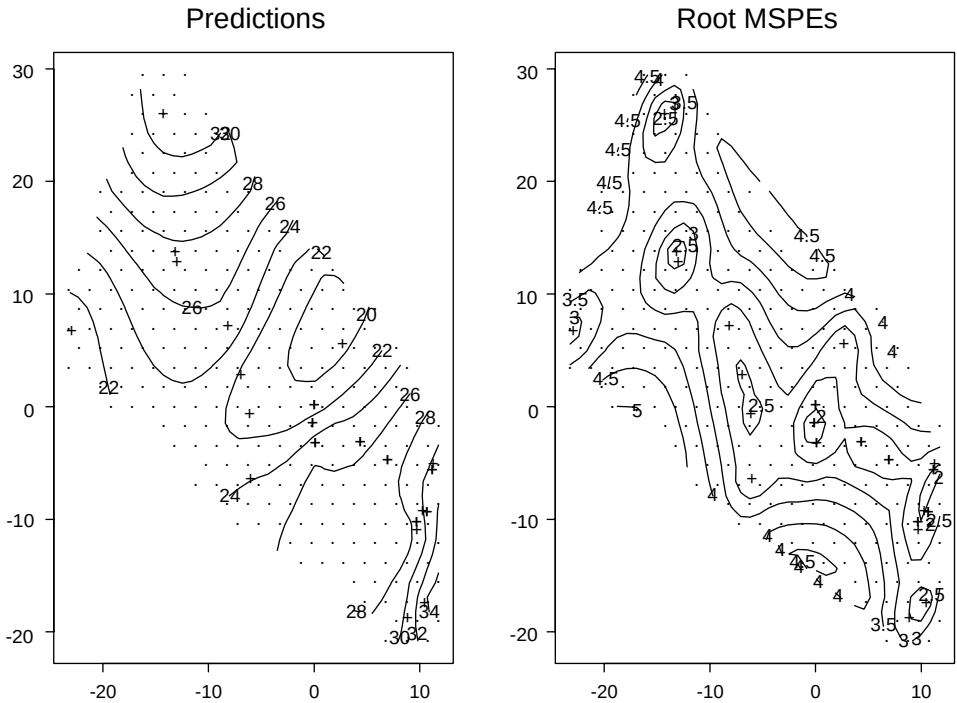

Fig. 6. Contour maps of constrained kriging predictions and root MSPEs of PM_{10}. There are 279 prediction locations marked ".", and 22 sampling locations marked "+".

a=30, we see that OK has the largest absolute bias in both cases. The cell in question in Table 2 suggests that in predicting $F_S(q(0.75))$, CM has the smallest absolute bias but CS(1600) has the smallest MSPE (and a relatively large absolute bias). By also noting that $F_X(28.75)$ has an extra degree of nonlinearity (i.e., $X(\cdot) = \exp(S(\cdot))$), we might prefer CM in this case. The corresponding cell in Table 3 suggests that in predicting $Q_S(q(0.75))$, CS(1600) performs best in terms of unbiasedness and the MSPE, with CM taking second place. Also noting the extra degree of nonlinearity in $Q_X(28.75)$, it may be difficult to choose between CM and CS(1600) in this case.

7. Conclusions

Several conclusions can be drawn from this paper. The first is that if a measurement-error component is explicitly added to a spatial model, then this may have a profound effect on the nature of spatial prediction. Indeed, popular nonlinear kriging predictors such as indicator (co)kriging and disjunctive kriging, are biased in the presence of measurement error. On the other hand, linear kriging predictors can be adjusted easily to filter out measurement error. In addition, the spatial model with additive measurement error can handle replicated data (i.e., multiple observations at a single sampling

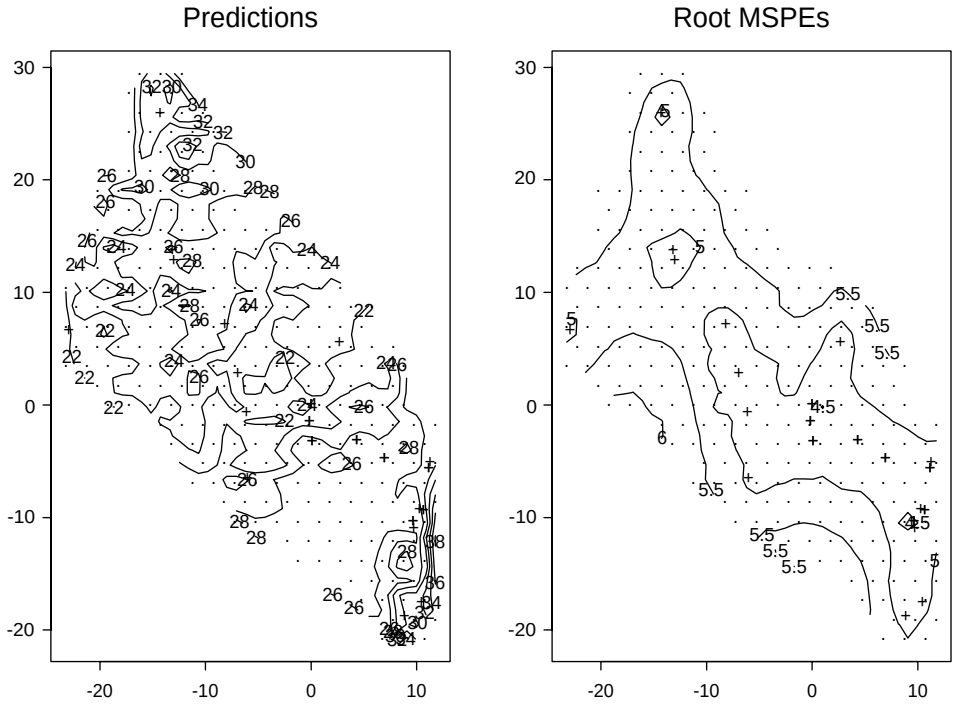

Fig. 7. Contour maps of covariance-matching constrained kriging predictions and root MSPEs of PM_{10}. There are 279 prediction locations marked ".", and 22 sampling locations marked "+".

location) and data where the same underlying spatial phenomenon is observed via two or more distinct types of measuring instrument.

The second conclusion is that for the prediction of nonlinear spatial functionals, covariance-matching constrained kriging has good unbiasedness properties in almost all cases, with some exceptions as noted in Section 5. Its MSPE properties are sometimes good or best, but almost never worst among the four predictors, and they usually beat those of constrained kriging. In terms of bias and MSPE, the performance of conditional simulation is sometimes very good and sometimes very poor, depending on the random field and the degree of nonlinearity of the predictand. While conditional simulation matches *all* means, variances, and covariances, it requires a lot of computing time and storage of simulations. Covariance-matching constrained kriging presents a compromise between conditional simulation and ordinary kriging, where some covariances are matched but only one prediction surface is computed. (Olea and Pawlowsky, 1996, take an empirical, data-driven approach to compensating for the smoothing inherent in kriging.) Ordinary kriging usually performed very poorly for nonlinear predictands, but in a few cases did better than expected. To sum up, covariance-matching constrained kriging fills a useful niche in predicting nonlinear spatial functionals, especially when the predictand has a high degree of nonlinearity and when the distribution of the underlying signal is unknown.

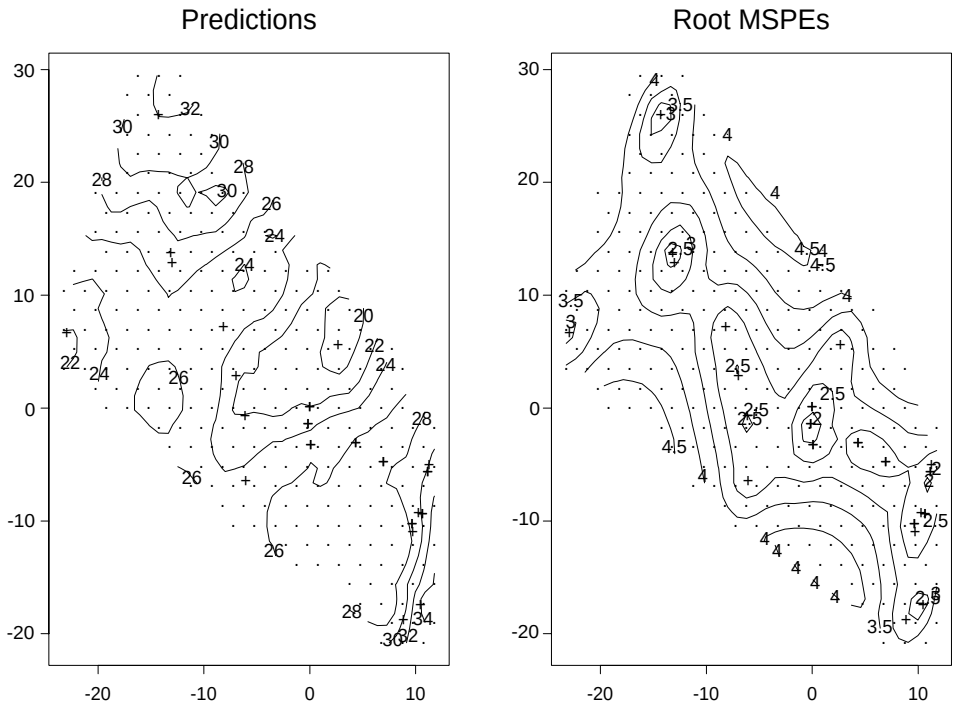

Fig. 8. Contour maps of conditional simulation predictions (with L calibrated to be 10) and root MSPEs of PM_{10}. There are 279 prediction locations marked ".", and 22 sampling locations marked "+".

Acknowledgements

The research presented in this article was partially supported by the US Environmental Protection Agency under Co-operative Agreement CR822919-01-0 and Assistance Agreement R827257-01-0; the article has not been subjected to review by the EPA and thus does not reflect the view of the agency and no official endorsement should be inferred. The research was also partially supported by the Office of Naval Research under grant N00014-99-0214.

Appendix

Before we prove Propositions 2–4, we need to establish the following lemma.

Lemma. *Assume that* $\mathbf{P}$, $\mathbf{Q}$, *and* $\boldsymbol{\Sigma}$ *are p.d., where* $\mathbf{P}$ *and* $\mathbf{Q}$ *are defined in the statement of Proposition* 1, *along with* $\mathbf{P}_1 \equiv \mathbf{P}^{1/2}$, $\mathbf{Q}_1 \equiv \mathbf{Q}^{1/2}$, *and* $\mathbf{K} \equiv \mathbf{Q}_1^{-1}\mathbf{P}_1$. *Then for any* $m \times m$ *matrix* $\mathbf{A}$ *satisfying* (C1) *and* (C2),

$$(\mathbf{A}_0 - \mathbf{A})'\mathbf{C} = (\boldsymbol{\Sigma}_m - \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}_0)\mathbf{K}^{-1},$$

where $\mathbf{A}_0$ *is given by* (4.3).

Proof. Since (C1) implies that $\mathbf{A}'\mathbf{X} = \mathbf{X}'_m$, we have

$$\begin{aligned}\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}_0\mathbf{K}^{-1} &= \mathbf{A}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1})\mathbf{C} + \mathbf{A}'\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m\mathbf{K}^{-1} \\ &= \mathbf{A}'\mathbf{C} + \mathbf{X}'_m(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}_m\mathbf{K}^{-1} - \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{C})\end{aligned}$$

and hence,

$$\mathbf{A}'\mathbf{C} = \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}_0\mathbf{K}^{-1} - \mathbf{X}'_m(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}_m\mathbf{K}^{-1} - \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{C}).$$

Now,

$$\mathbf{A}'_0\mathbf{C} = \mathbf{K}'\mathbf{C}'(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\Sigma}^{-1}\mathbf{C} + \mathbf{X}'_m(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{C}.$$

Therefore,

$$\begin{aligned}(\mathbf{A}_0 - \mathbf{A})'\mathbf{C} &= \mathbf{K}'\mathbf{C}'(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\Sigma}^{-1}\mathbf{C} \\ &\quad - (\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}_0 - \mathbf{X}'_m(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m)\mathbf{K}^{-1} \\ &= \mathbf{P}'_1(\mathbf{Q}_1^{-1})'\mathbf{Q} - (\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}_0 - \boldsymbol{\Sigma}_m + \mathbf{P})\mathbf{K}^{-1},\end{aligned}$$

since $\mathbf{P} \equiv \boldsymbol{\Sigma}_m - \mathbf{X}'_m(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m$ and $\mathbf{Q} \equiv \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C} - \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$. Therefore,

$$\begin{aligned}(\mathbf{A}_0 - \mathbf{A})'\mathbf{C} &= \mathbf{P}'_1\mathbf{Q}_1 + (\boldsymbol{\Sigma}_m - \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}_0)\mathbf{K}^{-1} - \mathbf{P}\mathbf{P}_1^{-1}\mathbf{Q}_1 \\ &= (\boldsymbol{\Sigma}_m - \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}_0)\mathbf{K}^{-1}. \qquad \square\end{aligned}$$

Proof of Proposition 2. (i) Differentiate $f(\mathbf{A})$ with respect to $\mathbf{A}$ and equate with zero to obtain

$$f'(\mathbf{A}) \equiv \left(\frac{\partial f(\mathbf{A})}{\partial a_{ij}}\right) = 2\mathbf{C}\mathbf{y}\mathbf{y}' + 2\mathbf{X}\boldsymbol{\Lambda}_1 - 2\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Lambda}_2 = \mathbf{0},$$

where $\boldsymbol{\Lambda}_1 \equiv (\boldsymbol{\lambda}_{11}, \ldots, \boldsymbol{\lambda}_{1m})$ is a $(p+1) \times m$ matrix and $\boldsymbol{\Lambda}_2 \equiv (\lambda_{2ij})$ is an $m \times m$ symmetric matrix. The expression for $f'(\mathbf{A})$ follows from

$$(\partial/\partial\mathbf{a}_\ell)\left(\sum_{i=1}^{m}\mathbf{a}'_i\mathbf{X}\boldsymbol{\lambda}_{1i}\right) = \mathbf{X}\boldsymbol{\lambda}_{1\ell}$$

and

$$(\partial/\partial\mathbf{a}_\ell)\left(\sum_{i=1}^{m}\sum_{j=i}^{m}\lambda_{2ij}(\mathbf{a}'_i\boldsymbol{\Sigma}\mathbf{a}_j - C(B_i, B_j))\right) = 2\sum_{k=1}^{m}\lambda_{2i\ell}\boldsymbol{\Sigma}\mathbf{a}_i.$$

Thus,

$$\mathbf{A}\boldsymbol{\Lambda}_2 = \boldsymbol{\Sigma}^{-1}(\mathbf{C}\mathbf{y}\mathbf{y}' + \mathbf{X}\boldsymbol{\Lambda}_1),$$

since $\boldsymbol{\Sigma}$ is p.d. Consequently,

$$\boldsymbol{\Lambda}_1 = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}[\mathbf{X}_m\boldsymbol{\Lambda}_2 - \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{C}\mathbf{y}\mathbf{y}'],$$

since (C1) implies that $\mathbf{X}'\mathbf{A} = \mathbf{X}_m$. Hence,

$$\mathbf{A}\boldsymbol{\Lambda}_2 = \boldsymbol{\Sigma}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1})\mathbf{C}\mathbf{y}\mathbf{y}' + \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m\boldsymbol{\Lambda}_2. \tag{A.1}$$

Also, (C2) implies that

$$\begin{aligned}\Lambda_2\boldsymbol{\Sigma}_m\Lambda_2 &\equiv \Lambda_2\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}\Lambda_2\\ &= \mathbf{y}\mathbf{y}'\mathbf{C}'\boldsymbol{\Sigma}^{-1}\{\boldsymbol{\Sigma} - 2\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\\ &\quad + \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\}\boldsymbol{\Sigma}^{-1}\mathbf{C}\mathbf{y}\mathbf{y}'\\ &\quad + \Lambda_2\mathbf{X}_m'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m\Lambda_2,\end{aligned}$$

since $\mathbf{y}\mathbf{y}'\mathbf{C}'\{\mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\}\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} = \mathbf{0}$. Therefore,

$$\begin{aligned}\Lambda_2\boldsymbol{\Sigma}_m\Lambda_2 &= \mathbf{y}\mathbf{y}'(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C} - \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{C})\mathbf{y}\mathbf{y}'\\ &\quad + \Lambda_2\mathbf{X}_m'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m\Lambda_2.\end{aligned}$$

That is,

$$\Lambda_2\mathbf{P}\Lambda_2 = \mathbf{y}\mathbf{y}'\mathbf{Q}\mathbf{y}\mathbf{y}'. \tag{A.2}$$

Now (A.2) can be equivalently written as

$$(\boldsymbol{\Lambda}_2^*)'\boldsymbol{\Lambda}_2^* = (\mathbf{y}^*\mathbf{y}^{*'})'(\mathbf{y}^*\mathbf{y}^{*'}),$$

where $\boldsymbol{\Lambda}_2^* \equiv \mathbf{P}_1\boldsymbol{\Lambda}_2\mathbf{Q}_1$ and $\mathbf{y}^* \equiv \mathbf{Q}_1\mathbf{y}$. Its solution is $\boldsymbol{\Lambda}_2^* = \mathbf{y}^*\mathbf{y}^{*'}$, or equivalently,

$$\boldsymbol{\Lambda}_2 = \mathbf{P}_1^{-1}\mathbf{Q}_1\mathbf{y}\mathbf{y}',$$

provided the right-hand side is symmetric. This is true if $\mathbf{y} = \mathbf{V}_j$ in Proposition 1, since then

$$\mathbf{P}_1^{-1}\mathbf{Q}_1\mathbf{y}\mathbf{y}' = \sum_{i=1}^{m} v_i^{-2}\mathbf{V}_i\mathbf{W}_i'\mathbf{V}_j\mathbf{V}_j' = v_j^{-2}\mathbf{V}_j\mathbf{V}_j'; \quad j = 1,\ldots,m.$$

Then it is a simple matter to demonstrate that $\mathbf{A} = \mathbf{A}_0$, $\mathbf{y} = \mathbf{V}_j$, and $\boldsymbol{\Lambda}_2 = \mathbf{P}_1^{-1}\mathbf{Q}_1\mathbf{y}\mathbf{y}'$ satisfy (A.1) with $\boldsymbol{\Lambda}_2$ symmetric. Thus $\mathbf{A} = \mathbf{A}_0$ is a stationary point of $f(\mathbf{A})$.

(ii) It only remains to prove that $\mathbf{y}'\mathbf{M}_A\mathbf{y}$, subject to $\mathbf{A}$ satisfying (C1) and (C2), is in fact minimized by $\mathbf{A} = \mathbf{A}_0$. Now $\mathbf{M}_A = 2\boldsymbol{\Sigma}_m - \mathbf{A}'\mathbf{C} - \mathbf{C}'\mathbf{A}$, due to (C2). Hence,

$$\begin{aligned}\mathbf{M}_A - \mathbf{M}_{A_0} &= (\mathbf{A}_0 - \mathbf{A})'\mathbf{C} + \mathbf{C}'(\mathbf{A}_0 - \mathbf{A})\\ &= (\boldsymbol{\Sigma}_m - \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}_0)\mathbf{K}^{-1} + (\mathbf{K}^{-1})'(\boldsymbol{\Sigma}_m - \mathbf{A}_0'\boldsymbol{\Sigma}\mathbf{A}),\end{aligned}$$

due to Lemma. Hence,

$$\mathbf{y}'(\mathbf{M}_A - \mathbf{M}_{A_0})\mathbf{y} = \mathbf{y}'(\boldsymbol{\Sigma}_m - \mathbf{A}'\boldsymbol{\Sigma}\mathbf{A}_0)\mathbf{K}^{-1}\mathbf{y} + (\mathbf{K}^{-1}\mathbf{y})'(\boldsymbol{\Sigma}_m - \mathbf{A}_0'\boldsymbol{\Sigma}\mathbf{A})\mathbf{y}$$
$$= v_j^{-2}\mathbf{V}_j'\{(\mathbf{A}_0 - \mathbf{A})'\boldsymbol{\Sigma}(\mathbf{A}_0 - \mathbf{A})\}\mathbf{V}_j \geqslant 0,$$

where $\mathbf{y} = \mathbf{V}_j$; $j = 1, \ldots, m$. □

Proof of Proposition 3. Constraint (C1) is satisfied, since

$$\mathbf{A}_0'\mathbf{X} = \mathbf{K}'\mathbf{C}'(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\Sigma}^{-1}\mathbf{X} + \mathbf{X}_m'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$$
$$= \mathbf{X}_m'.$$

Further,

$$\mathbf{A}_0'\boldsymbol{\Sigma}\mathbf{A}_0 = \mathbf{K}'\mathbf{C}'\boldsymbol{\Sigma}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1})\mathbf{C}\mathbf{K} + \mathbf{X}_m'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m,$$

since $(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\Sigma}^{-1}\mathbf{X}' = \mathbf{0}$. And, since $\mathbf{P} = \boldsymbol{\Sigma}_m - \mathbf{X}_m'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}_m$ and $\mathbf{Q} = \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C} - \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$, we have

$$\mathbf{A}_0'\boldsymbol{\Sigma}\mathbf{A}_0 = \mathbf{K}'\mathbf{Q}\mathbf{K} + \boldsymbol{\Sigma}_m - \mathbf{P} = \boldsymbol{\Sigma}_m,$$

which is (C2). □

Proof of Proposition 4. Since $\boldsymbol{\Sigma}$ is p.d., we can write $\boldsymbol{\Sigma}^{-1} = \mathbf{F}'\mathbf{F}$ for a nonsingular $\mathbf{F}$. Then,

$$\mathbf{Q} = \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C} - \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$$
$$= \mathbf{G}'\mathbf{G},$$

where $\mathbf{G} = \mathbf{J}\mathbf{F}\mathbf{C}$, and $\mathbf{J} \equiv (\mathbf{I} - \mathbf{F}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{F}')$ is symmetric and idempotent. Hence $\mathbf{Q}$ is n.n.d. □

References

Aldworth, J., 1998. Spatial prediction, spatial sampling, and measurement error. Ph.D. Thesis, Iowa State University, unpublished.

Aldworth, J., Cressie, N., 1999. Sampling designs and prediction methods for Gaussian spatial processes. In: Ghosh, S. (Ed.), Multivariate Analysis, Design of Experiments and Survey Sampling. Marcel Decker, New York, pp. 1–54.

Besag, J.E., Green, P.J., 1993. Spatial statistics and Bayesian computation. J. Roy. Statist. Soc. B 55, 25–37.

Cressie, N., 1988. Spatial prediction and ordinary kriging. Math. Geol. 20, 405–421 [Erratum: Math. Geol. 21, 493–494].

Cressie, N., 1993a. Statistics for Spatial Data, Revised Edition. Wiley, New York.

Cressie, N., 1993b. Aggregation in geostatistical problems. In: Soares, A. (Ed.), Geostatistics Tróia 1992, Vol. 1. Kluwer Academic Publishers, Dordrecht, pp. 25–36.

Cressie, N., Johannesson, G., 2001. Kriging for cut-offs and other difficult problems. In: Monestiez, P. et al. (Eds.), geoENV-III—Geostatistics for Environmental Applications. Kluwer, Dordrecht, pp. 299–310.

Cressie, N., Kaiser, M.S., Daniels, M.J., Aldworth, J., Lee, J., Lahiri, S.N., Cox, L.H., 1999. Spatial analysis of particulate matter in an urban environment. In: Gomez-Hernandez, J. et al. (Ed.), geoENV2—Geostatistics for Environmental Applications. Kluwer Academic Publishers, Dordrecht, pp. 41–52.

Diggle, P.J., Tawn, J.A., Moyeed, R.A., 1998. Model-based geostatistics. Appl. Statist. 47, 299–326.

Goldberger, A.S., 1962. Best linear unbiased prediction in the generalized linear regression model. J. Amer. Statist. Assoc. 57, 369–375.

Gotway, C.A., 1994. The use of conditional simulation in nuclear waste site performance assessment. Technometrics 36, 129–161.

Gotway, C.A., Cressie, N., 1993. Improved multivariate prediction under a general linear model. J. Multivariate Anal. 45, 56–72.

Journel, A.G., 1974. Geostatistics for conditional simulation of ore bodies. Econom. Geol. 69, 673–687.

Journel, A.G., 1983. Nonparametric estimation of spatial distributions. J. Internat. Assoc. Math. Geol. 15, 445–468.

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

Lajaunie, C., 1990. Comparing some approximate methods for building local confidence intervals for predicting regionalized variables. Math. Geol. 22, 123–144.

Matheron, G., 1963. Principles of geostatistics. Econom. Geol. 58, 1246–1266.

Matheron, G., 1976. A simple substitute for conditional expectation: the disjunctive kriging. In: Guarascio, M., David, M., Huijbregts, C. (Eds.), Advanced Geostatistics in the Mining Industry. Reidel, Dordrecht, pp. 221–236.

Myers, D.E., 1982. Matrix formulation of co-kriging. J. Internat. Assoc. Math. Geol. 14, 249–257.

Olea, R.A., Pawlowsky, V., 1996. Compensating for estimation smoothing in kriging. Math. Geol. 28, 407–417.

Rivoirard, J., 1994. Introduction to Disjunctive Kriging and Non-linear Geostatistics. Oxford University Press, Oxford.

Schott, J.R., 1997. Matrix Analysis for Statistics. Wiley, New York.

Serra, J., 1982. Image Analysis and Mathematical Morphology. Academic Press, London.

US Environmental Protection Agency, 1996. Review of national assessment of scientific and technical information. OAQPS Staff Paper, EPA Report no. 452/R-96-013, Office of Air Quality Planning and Standards, US EPA, Research Triangle Park, NC.

Ver Hoef, J.M., Cressie, N., 1993. Multivariable spatial prediction. Math. Geol. 25, 219–240. [Errata: Math. Geol. 26, 273–275].