JOURNAL OF MULTIVARIATE ANALYSIS 43, 351–374 (1992)

Interpolation with Uncertain Spatial Covariances: A Bayesian Alternative to Kriging

NHU D. LE AND JAMES V. ZIDEK

University of British Columbia

Communicated by the Editors

In this paper a Bayesian alternative to Kriging is developed. The latter is an important tool in geostatistics. But aspects of environmetrics make it less suitable as a tool for interpolating spatial random fields which are observed successively over time. The theory presented here permits temporal (and spatial) modeling to be done in a convenient and flexible way. At the same time model misspecifications, if any, can be corrected by additional data if and when it becomes available, and past data may be used in a systematic way to fit model parameters. Finally, uncertainty about model parameters is represented in the (posterior) distributions, so unrealistically small credible regions for the interpolants are avoided. The theory is based on the multivariate normal and related distributions, but because of the hierarchical prior models adopted, the results would seem somewhat robust with respect to the choice of these distributions and associated hyperparameters.

1. INTRODUCTION

In this paper, a fully (hierarchical) Bayesian alternative to Kriging is derived. Kriging is a well-known method of spatial interpolation (cf. [9]). In the discrete space version of this problem, a real-valued spatial random field is supposed to be observed at g discrete, "gauged" sites to yield a $g \times 1$ data vector, $X^{(2)} = (X^{(21)}, ..., X^{(2g)})'$. The object of inferential interest is a $u \times 1$ vector, $X^{(1)} = (X^{(11)}, ..., X^{(1u)})'$, of unmeasured values at u "ungauged" sites. The spatial field is over $p = u + g$ discrete sites. In Kriging, $u = 1$ and the spatial covariance matrix of $X = (X^{(1)'}, X^{(2)'})$, Σ, is taken to be fully specified. The optimal linear unbiased interpolator of $X^{(1)}$ based on $X^{(2)}$ is then determined. However, this interpolator depends on Σ which, in practice, is unknown. Commonly the variogram, i.e., the expected squared difference between the values of the random field at any two sites, replaces

Received November 18, 1991; revised February 11, 1992.
AMS 1980 subject classifications: 62M20, 62H99.
Key words and phrases: posterior distributions, Kriging, spatial covariance estimators.

0047-259X/92 $5.00

the covariance of these values, although when Σ exists as is supposed here, the two are equivalent. In practice Σ (or the variogram) is estimated from preliminary data, $\{X_j^{(2)}, j=1, ..., n\}$ under the assumptions that (i) the field is spatially isotropic and (ii) the covariance of the field between any two sites is a parametric function of the distance between the sites. The preliminary data can then be used to estimate the parameters of the covariance. The estimate $\hat{\Sigma}$, replaces Σ in the best linear interpolator; the result is the Kriging interpolator. An estimate of the mean squared interpolation error is readily found in a similar way. In simple Kriging the mean of the random field is assumed to be constant over space, but this condition is relaxed somewhat in universal Kriging.

A deficiency of classical Kriging methodology is that it fails to incorporate uncertainty about Σ into its measure of interpolation error. This deficiency leads to unwarranted confidence in the interpolated values and, potentially, to seemingly valid decisions or regulatory actions which are, in fact, unjustified. Bayesian approaches to spatial interpolation avoid this deficiency, and such an approach is the subject of this paper.

Specific advantages of Bayesian interpolation over Kriging are that (i) uncertainty about Σ (and all other parameters in the analysis) is reflected in the assessed interpolation errors and (ii) preliminary covariance models which incorporate available knowledge may be introduced into the analysis without requiring their perpetual use. The second is a major advantage in applications such as environmental monitoring where repeated observations of the random field are made. Bayesian estimates of all uncertain quantities (including coefficients of the interpolators) change adaptively with the incoming data. However, to derive maximal benefit from this advantage, the support of all priors must be as broad as is feasible and realistic.

In this paper we seek to exploit these advantages. However, there is a growing body of related literature on Bayesian methods which we need to review first to put our work in perspective.

As noted by Pilz (undated manuscript), if we restrict ourselves to linear interpolators, the Bayesian approach like Kriging, depends on the prior distributions only through their first and second moments. Taking these moments as specified, Omre [10] characterizes the Bayesian interpolator. Omre and Halvorsen [11; hereafter, OH] extend the model of Omre [10] to include linear trend models with random coefficients. The resulting model seems identical to that proposed on non-Bayesian grounds by Fedorov and Mueller [5, 6] to bridge the gap between classical theory of optimal design in regression to that of designing spatial monitoring networks. OH find that the Bayesian approach yields a continuum of models between the extremes of simple and universal Kriging. The theme of OH is elaborated on by Omre, Halvorsen, and Berteig [12], who consider and illustrate two competing linear structures. Again empirical non-Bayesian estimates of the covariance structure are given.

In all three of the last cited papers of Omre, a final empirical Bayesian step is taken when the unknown components of the spatial covariance structure are estimated from the data. Obvious nonparametric (non-Bayesian) estimates are suggested under the assumption that these components are translation invariant. The resulting estimates are then incorporated into the analysis as if they had been fully specified *a priori*. As OH point out [11, p. 771] in this last step they are emulating what is done in Kriging; so their method is potentially susceptible to the same criticism as that which we leveled above at Kriging.

An empirical Bayesian step in a fully Bayesian analysis may be justified under certain circumstances. When the posterior distribution of the unobservables, conditional on the data and certain hyperparameters, is insensitive to the choice of these hyperparameters, it may well be practically expedient to adopt a diffuse prior for them. Or they may be estimated using the data distribution conditioned on the same hyperparameters. Either way, the final posterior will then be approximately equal to that which would be obtained by properly modeling out the small fraction of residual uncertainty left in the hyperparameters. However, Omre and his co-authors do not justify their empirical Bayes step in this way. In particular, the robustness of their procedures to the choice of their spatial covariance estimator is not clear.

The robustness of universal Kriging to the misspecification of the variogram has been the subject of several investigations surveyed by Pilz (*ibid.*). He considers the robustness of Bayesian interpolators with respect to the misspecification of the first and second moments of the joint (*a priori*) distribution of the regression parameters and data. He determines measures of robustness and determines a robust Bayesian interpolator when the spatial covariance is fully specified.

In this paper the spatial covariance is left completely unspecified in the first level of a hierarchical Bayesian model. Uncertainty about this parameter is then incorporated through a second level prior. Our approach differs from the partially hierarchical Bayesian approach of de Waal and Groenewald (undated manuscript), since the latter assumes an isotropic random field and a parametric spatial covariance structure. They put a prior distribution on some of the covariance parameters while fitting the others, albeit parametrically, in the spirit of the work of Omre and others cited above.

Although the paper of Loader and Switzer [8] is not fundamentally Bayesian, a Bayesian approach is used to motivate their interpolation procedure and spatial covariance estimator in the case considered here, where the random field is anisotropic. Their paper explicates aspects of the earlier paper of Switzer [16]. And it overlaps ours in that they are able to avoid parametrizing Σ by assuming as do Caselton, Kan, and Zidek [3, in a design context], and as we do here, that the covariance matrix of $X^{(2)}$, Σ_{22}, has the inverted Wishart distribution. However, unlike us they do not explicitly determine the joint posterior distribution of the unobservables, including the values of the random field at a multiplicity of ungauged sites. And while they assume $u=1$, we allow $u \geqslant 1$, which in conjunction with the explicit posterior distribution paves the way for the development of simultaneous credibility bounds, the subject of a future paper.

Our prior models and basic assumptions are given in Section 2 along with a statement of our main results. The derivations of these results is given in Section 3. We assume that the hyperparameters of our second-level priors are specified. So the implementation of our results entail their specification. A discussion of this problem and other aspects of our work constitutes Section 4.

2. Main Results

2.1. *Preliminaries and Notations*

Let $X_j = (X_j^{(1)'}, X_j^{(2)'})$ be a p-dimensional random vector where $X_j^{(1)}$ and $X_j^{(2)}$ are u- and g-dimensional vectors, respectively ($p = u + g$). Let z_j be a k-dimensional vector of covariates. Assume

$$X_j \mid z_j, B, \Sigma \overset{\text{independent}}{\sim} N_p(Bz_j, \Sigma), \tag{1}$$

where B denotes a $(p \times k)$ matrix of regression coefficients,

$$B = \begin{pmatrix} \beta_{11} & \cdots & \beta_{ik} \\ & \vdots & \\ \beta_{p1} & \cdots & \beta_{pk} \end{pmatrix} \equiv \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}.$$

The covariance matrix, Σ, is partitioned accordingly as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where, in general, for any two random vectors R and S,

$$\Sigma_{RS} \equiv \operatorname{Cov}(R, S') \equiv E[R - E(R)][S - E(S)]'.$$

It will be convenient to reparametrize Σ as $(\Sigma_{22}, \Sigma_{1|2}, \tau)$, where $\Sigma_{1|2}$ is a $(u \times u)$ matrix denoting the residual covariance of $X_j^{(1)}$-residuals after optimal linear prediction based on $X_j^{(2)}$; it is given by

$$\Sigma_{1|2} \equiv \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

The $(g \times g)$ matrix Σ_{22} is the covariance matrix of $X_j^{(2)}$ and τ is a $(u \times g)$ matrix representing the slope of the optimal linear predictor of $X_j^{(1)}$ based on $X_j^{(2)}$ given by

$$\tau \equiv \Sigma_{12}\Sigma_{22}^{-1}.$$

This transformation is achieved through the Bartlett [2] decomposition, $\Sigma = T\Delta T'$, where

$$\Delta = \begin{pmatrix} \Sigma_{1|2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} I & \tau \\ 0 & I \end{pmatrix}.$$

With this decomposition, Σ can be written as

$$\Sigma = \begin{pmatrix} \Sigma_{1|2} + \tau\Sigma_{22}\tau' & \tau\Sigma_{22} \\ \Sigma_{22}\tau' & \Sigma_{22} \end{pmatrix}. \tag{2}$$

We suppose the uncertainty about B and Σ can be represented by the conjugate prior distribution (cf. [14, p. 234]),

$$B \mid B^o, \Sigma, F \sim N_{kp}(B^o, \Sigma \otimes F^{-1}) \tag{3}$$

$$\Sigma \mid \Psi, m \sim W_p^{-1}(\Psi, m), \tag{4}$$

where $N_p(\mu, \Sigma)$ denotes the p-dimensional Gaussian distribution with mean μ and covariance matrix Σ. $W_p^{-1}(\Psi, m)$ denotes the p-dimensional inverted Wishart distribution with scale matrix Ψ and m degrees of freedom; it is proper if $p < m$. B^o and Ψ are the prior parameter matrices.

The conjugate prior distribution (4) for Σ can be equivalently presented in terms of the new parameters $(\Sigma_{22}, \Sigma_{1|2}, \tau)$ as

$$\begin{aligned} &\Sigma_{22} \mid \Psi, m \sim W_g^{-1}(\Psi_{22}, m-u) \\ &\Sigma_{1|2} \mid \Psi, m \sim W_u^{-1}(\Psi_{1|2}, m) \\ &\tau \mid \Sigma_{1|2}, \Psi \sim N_{ug}(\eta, \Sigma_{1|2} \otimes \Psi_{22}^{-1}). \end{aligned} \tag{5}$$

Here $(\Psi_{22}, \Psi_{1|2}, \eta)$ denotes the decomposition of the prior parameter matrix Ψ analogous to that of Σ; that is, $\eta = \Psi_{12}\Psi_{22}^{-1}$. Moreover, Σ_{22} is independent of $(\Sigma_{1|2}, \tau)$ when the prior distribution is proper. See Caselton *et al.* [3] for more details on this decomposition.

Let $D=((x_1^{(2)}, z_1), ..., (x_n^{(2)}, z_n))$ be the observed data, where, given the covariate z_j's, the $\{x_j^{(2)}\}$ are independent realizations of

$$X_j^{(2)} \mid B, \Sigma, z_j \sim N_g(B_2 z_j, \Sigma_{22}). \tag{6}$$

Note that this represents only the second component of the X_j in model (1). Hence, D represents the partially observed data; i.e., there are no observations for the first u-coordinates.

Define $\hat{B}_2$ and S as

$$\begin{aligned} \hat{B} &= CA^{-1} \\ S &= \sum_{j=1}^{n} (x_j^{(2)} - \hat{B}_2 z_j)(x_j^{(2)} - \hat{B}_2 z_j)' \end{aligned} \tag{7}$$

where

$$C = \sum_{j=1}^{n} x_j^{(2)} z_j' \tag{8}$$

and

$$A = \sum_{j=1}^{n} z_j z_j'.$$

Anderson [1, p. 291] shows that, conditional on B_2 as well as Σ_{22}, A, $\hat{B}_2$, and S are independent and

$$\hat{B}_2 \mid B_2, \Sigma_{22}, A \sim N_{gk}(B_2, \Sigma_{22} \otimes A^{-1}) \tag{9}$$

and

$$S \mid B_2, \Sigma_{22}, A \sim W_g(\Sigma_{22}, n-k).$$

Here $W_g(\Psi, m)$ denotes the g-dimensional Wishart distribution with scale matrix Ψ and m degrees of freedom.

2.2. *Results*

We now present the main results of this paper. Here (B_1^o, B_2^o) is a partition of B^o analogous to that of B. The proofs are presented in the next section.

THEOREM 2.1. *Let the prior distributions of B and $(\Sigma_{22}, \Sigma_{1|2}, \tau)$ be defined as in* (3) *and* (5). *Then*

$$B \mid D, B^o, \Sigma_{22}, \Sigma_{1|2}, \tau \sim N_{kp}(B^*, \Sigma^*), \tag{10}$$

where

$$B^* = B^o + \begin{pmatrix} \tau \\ I_g \end{pmatrix} (\hat{B}_2 - B_2^o)\, \hat{E}' \tag{11}$$

$$\Sigma^* = \Sigma \otimes F^{-1} - \left[\begin{pmatrix} \tau \\ I_g \end{pmatrix} (\Sigma_{22}\tau', \Sigma_{22}) \right] \otimes (\hat{E}F^{-1}) \tag{12}$$

$$\hat{E} = F^{-1}(A^{-1} + F^{-1})^{-1}$$

and A and $\hat{B}_2$ are defined as in (7)–(8). *The posterior distribution of* $(\Sigma_{22}, \Sigma_{1|2}, \tau)$ *is*

$$\begin{aligned} \Sigma_{22} \mid D, \Psi_{22}, m &\sim W_g^{-1}(\hat{\Psi}_{22}, m+n-u) \\ \Sigma_{1|2} \mid D, \Psi_{1|2}, m &\sim W_u^{-1}(\Psi_{1|2}, m) \\ \tau \mid D, \eta, \Sigma_{1|2} &\sim N_{ug}(\eta, \Sigma_{1|2} \otimes \Psi_{22}^{-1}), \end{aligned} \tag{13}$$

where

$$\hat{\Psi}_{22} = \Psi_{22} + S + (\hat{B}_2 - B_2^o)' (A^{-1} + F^{-1})^{-1} (\hat{B}_2 - B_2^o) \tag{14}$$

and S is given in (7).

Remark. The result (13) establishes the intuitive result that the posterior distributions of τ and $\Sigma_{1|2}$ are just their prior distributions; our knowledge of the covariance matrix Σ is updated by the data D only through Σ_{22}. This is reasonable, since Σ_{22} is independent of $(\Sigma_{1|2}, \tau)$, *a priori*. However, although the data cannot be used to improve the prior models of $\Sigma_{1|2}$, τ, the prior covariance matrix, Σ_{11}, for the ungauged sites and the cross-covariance matrix, Σ_{12}, between the gauged and ungauged sites can be improved with the data as shown in the following result.

COROLLARY 2.2. *Let Σ_o be the mean of the prior distribution of Σ given by* (4); *that is,*

$$\Sigma_o = \frac{1}{m-p-1} \Psi.$$

The posterior mean, $\tilde{\Sigma}$, of Σ is then given by

$$\tilde{\Sigma} = \alpha \Sigma_o + (1-\alpha)\, \hat{\Sigma}, \tag{15}$$

where $\alpha = (m-p-1)/(m+n-p-1)$ and

$$\hat{\Sigma} = \begin{pmatrix} \left[\dfrac{1 + \operatorname{tr}(\tilde{S}\Psi_{22}^{-1})}{m-u-1} \right] \Psi_{1|2} + \eta \tilde{S} \eta' & \eta \tilde{S} \\ \tilde{S}\eta' & \tilde{S} \end{pmatrix}$$

and

$$\tilde{S} = \frac{1}{n}(\hat{\Psi}_{22} - \Psi_{22}).$$

An anonymous referee has raised an interesting issue of whether our methodology would allow for a smoothness constraint on the posterior mean of Σ, given that the prior specification is smooth. From Eq. (15) the posterior mean is a weighted average of the prior and the estimate based on the observed data. So it is smooth if the estimate is; however, in our methodology, there is no mechanism to ensure that the estimate is smooth.

We would also remark that the Bartlett's decomposition shows that any symmetric matrix, Σ, is nonnegative definite if $\Sigma_{1|2}$ and Σ_{22} are. It follows that $\tilde{\Sigma}$ in Corollary 2.2 is nonnegative definite. Likewise from Eq. (13), we may conclude that the posterior distribution of Σ gives probability one to nonnegative definite matrices.

Using the available data D and the posterior distributions of B and Σ, we obtain the predictive distribution of the random vector X_f with sampling distribution described by (1), given the new covariate vector z_f. To present the result we let, in general, $t_r(\mu, H, v)$ denote the r-variate t-distribution with density function given by

$$f(x) \propto |H|^{-1/2} [v + (x-\mu)' H^{-1}(x-\mu)]^{-(v+r)/2}.$$

THEOREM 2.3. *Let the posterior distributions of B and Σ be defined as in Theorem 2.1. The predictive distribution of $X_f = (X_f^{(1)'}, X_f^{(2)'})$, given a covariate vector z_f and the prior parameters B^o and $(\Psi_{22}, \Psi_{1|2}, \eta)$, is*

$$X_f^{(2)} \mid D \sim t_g\left(a_2 + b, \frac{c-d}{l}\hat{\Psi}_{22}, l\right) \tag{16}$$

$$\begin{gathered} X_f^{(1)} \mid X_f^{(2)} = x_f^{(2)}, \\ D \sim t_u\left(a_1 + \eta(x_f^{(2)} - a_2), \frac{c + (a_2 - x_f^{(2)})' \Psi_{22}^{-1}(a_2 - x_f^{(2)})}{q} \Psi_{1|2}, q\right), \end{gathered} \tag{17}$$

where $\hat{\Psi}_{22}$ is given by (14), $l = m + n - u - g + 1$, $q = m - u + 1$, *and*

$$\hat{E} = F^{-1}(A^{-1} + F^{-1})^{-1}$$

$$a \equiv \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = B^o z_f$$

$$b = (\hat{B}_2 - B_2^o)\hat{E} z_f$$

$$c = 1 + z_f' F^{-1} z_f$$

$$d = z_f' \hat{E} F^{-1} z_f.$$

COROLLARY 2.4. *Given the predictive distribution in Theorem 2.3, the predictive means of $X_f^{(1)}$ and $X_f^{(2)}$, given z_f are*

$$\begin{aligned} \mu_1 &\equiv EX_f^{(1)} = a_1 + \eta b \\ \mu_2 &\equiv EX_f^{(2)} = a_2 + b. \end{aligned} \tag{18}$$

The value of the hierarchical approach we have adopted lies in the very general covariance structure introduced at the first level. We are then free to model the covariance structure at the second level through the various hyperparametric objects in Eqs. (3) and (4). Where possible the past and future data will update these prior models to achieve increasingly more realistic versions of their level-one counterparts, B and Σ. But whether or not such updating is possible, our uncertainty about the level-one parameters is reflected in the marginal likelihood of the hyperparameters specified by Theorem 2.3. Consequently (credible) intervals for the interpolates of Theorem 2.3 will realistically account for that uncertainty through the heavy-tailed Student's t-distributions described there.

At level two of the hierarchy, we may specify (hyper) parametric models for objects like Ψ, for example, to accommodate our remaining prior knowledge about the random field under consideration. The paper of Loader and Switzer [8] exemplifies that approach. This will leave a small number of additional (third stage) unspecified parameters. In a strictly Bayesian approach, we would add another distribution to our hierarchy to accomodate our uncertainty about these parameters. However, as noted in the Introduction, the posterior distribution of quantities of inferential interest will not be especially sensitive to the choices made of these parameters and this suggests alternatives to a fully Bayesian approach. One possibility would be a robust Bayesian approach like that of Pilz [13]. Another would be to introduce a diffuse prior on these remaining unspecified quantities. But the simplest option is that of estimating them from the marginal distribution of the data conditional on these quantities, the empirical Bayes solution.

At any event, Eq. (17) yields a Bayesian alternative to Kriging which permits a lot of prior modeling. Errors of interpolation are expressed by the covariance in Eq. (17). The latter depends on the covariates but not on the data unless they are used to estimate the third-stage hyperparameters. The Student's t-distribution reflects our uncertainty about level-one parameters.

The simplest special case of our theory perhaps, is that in which the station records are "exchangeable." To be precise, referring to displays (1) and (3), $k = 1$, $EB = B^0 = \mathbf{1}\mu^0$, $z_j = 1$ for all j, and $F = k^0$, where $\mathbf{1}$ denotes the $p \times 1$ vector all of whose elements are 1, and μ^0 as well as k^0 are specified constants. Here μ^0 is the overall level of the β_j's in $B = (\beta_1, ..., \beta_p)'$. For example, μ^0 might be between 4.5 and 5.5 when X's coordinates represent the pH of acidic precipitation, depending on the geographical area; 5.5 might be used for the so-called "pristine" areas. The value chosen for k^0 would depend on how well μ^0 represents the population of the β_j's in Eq. (1). This would depend partly on the size of the geographical area under consideration.

For the model in the last paragraph it seems reasonable that $\Psi = \sigma^2 \times \xi$, where $\xi_{ij} = 1$ or ρ, according as $i = j$ or $i \neq j$, $i, j = 1, ..., p$. It is easily shown that

$$E[X_f^{(1)} \mid X_f^{(2)} = x_f^{(2)}, D] = [(1 - \alpha_1)\, \mu^0 + \alpha_1 \bar{x}_f^{(2)}]\, \mathbf{1}_u$$

and

$$E[X_f^{(2)} \mid D] = (1 - \alpha_2)\, \mu^0 \mathbf{1}_g + \alpha_2 \bar{x}^{(2)},$$

where $\alpha_1 = g\rho/[(1 - \rho) + g\rho]$, $\alpha_2 = n/[k^0 + n]$, $\bar{x}_f^{(2)} = (1/g) \sum_m x_f^{(2m)}$ (averaging over g gauged stations), and $\bar{x}^{(2)} = (1/n) \sum_m x_m^{(2)}$ (averaging over n times points). Thus when n is large, $E[X_f^{(2)} \mid D] \simeq \bar{x}^{(2)}$. In contrast, $E[X_f^{(1)} \mid X_f^{(2)} = x_f^{(2)}, D]$, remains unchanged as n changes. However, if μ^0 were not specified, it would be natural to estimate it by $\hat{\mu}^0 = (1/g) \sum_{m=1}^{g} \bar{x}^{(2m)}$ (i.e., an overall average of all observations) and this estimate could then be used in the expression for $E[X_f^{(1)} \mid X_f^{(2)} = x_f^{(2)}, D]$.

A simple extension of the model of the last paragraph for the hypothetical application under consideration would cluster, on statistical or some other grounds, the stations into homogeneous subgroups, $l = 1, ..., L$. We would then have $B = (B_1', ..., B_L')$, where $B_l^0 = 1_{s_l} \mu^{0l}$ with s_l denoting the size of the lth cluster and μ^{0l} playing for cluster l, the role played in the last paragraph by the scalar μ^0, itself. And μ^{0l} might be estimated from the data for that cluster, if the posterior were not deemed to be unduly sensitive to this choice.

In the situation described in the last paragraph, X_i would be partitioned to reflect the clustering. The between cluster covariances might be approximated by 0, while that within clusters might be given the intraclass correlation structure with just two parameters that might also be estimated from the corresponding within cluster data sets. Alternatively, one might adopt a components-of-variance model, with one component for overall variation and independent cluster specific components to represent the individual clusters. This too would lead to a substantial reduction in the number of parameters to be fitted.

An extension of the exchangeable means models discussed above, would be the exchangeable regressions model. In Eq. (3) we would now have

$B^0 = \mathbf{1} \otimes \beta^0$, where $\beta^0 = (\beta^{01}, ..., \beta^{0k})$ is a vector of regression coefficients. These exchangeable regression models, like their counterpart for means, have the well-known advantage that they allow "borrowing from strength." When this is justified and the amount of data is small, this advantage can be substantial. Variations on this theme emerge when clustering is permitted. For brevity, we will not provide any further details here.

Analogous to the components-of-variance model proposed above, is a fixed effects linear model in Eq. (1). Now the "covariates" in (1) would be the appropriate indicator variables to pick out the fixed effects appropriate to the individual coordinates of the X_j vectors. Theorem 2.3 could then be used to obtain estimates of these fixed effects if they were of inferential interest.

It should be emphasized that the regression model in Eq. (1) allows time varying covariates. This is to enable trend and seasonality to be incorporated directly into the Bayesian analysis; there is no need to remove them in an ad hoc preliminary analysis. In this case, z_j would contain a coordinate, t_j, to represent the time while the ith row of B would have slope and intercept coefficients for station i. As well, z_j would contain $\cos \alpha t_j$ and $\sin \alpha t_j$ to account for seasonality. Here α determines the period and might well depend on the station, i, in which case there would have to be different cos and sin coordinates in z_j for each station. A technically awkward feature of the resulting model is that it is nonlinear in α.

We could also include in z_j, spatial modeling coordinates, $f_l(x_i)$, $i = 1, ..., p$, $l = 1, ..., L$, where the f_l's are specified functions and x_i represents the location vector of station i. This is done in the Bayesian extensions to universal Kriging and the models of Fedorov and Mueller [5, 6] described in the Introduction. Such spatial models may well be important in applications such as those of geostatistics. However, they are of little relevance in the context of environmental analysis where location has little explanatory value (for a discussion, see [17]).

3. Proofs

We now prove the results stated in the previous section. For expository convenience, we suppress the prior parameters B^o, Ψ, and m in all the conditional distributions in this section. Moreover, to make the proofs easier to follow, we will use β to represent a strung-out version of B, where each row of B is transposed and stacked up to obtain vector β. This convention is also used for the matrix B with subscripts and superscripts when necessary. The conditional density of X and Y given D is denoted by $f(X, Y \mid D)$.

Proof of Theorem 2.1. Ignoring irrelevant normalizing constants, we may rewrite the posterior density as

$$\begin{aligned} &f(B \mid D, \Sigma_{22}, \Sigma_{1\mid 2}, \tau) \\ &\quad \propto f(D \mid B, \Sigma_{22}, \Sigma_{1\mid 2}, \tau)\, f(B \mid \Sigma_{22}, \Sigma_{1\mid 2}, \tau) \\ &\quad \propto f(\hat{B}_2 \mid B, \Sigma_{22}, \Sigma_{1\mid 2}, \tau)\, f(S \mid B, \Sigma_{22}, \Sigma_{1\mid 2}, \tau)\, f(B \mid \Sigma_{22}, \Sigma_{1\mid 2}, \tau), \end{aligned}$$

since $\hat{B}_2$ and S are sufficient statistics for B_2 and Σ_{22} and they are also independent of each other. Thus by (9)

$$\begin{aligned} f(B \mid D, \Sigma_{22}, \Sigma_{1\mid 2}, \tau) &\propto f(\hat{B}_2 \mid B, \Sigma_{22}, \Sigma_{1\mid 2}, \tau)\, f(B \mid \Sigma_{22}, \Sigma_{1\mid 2}, \tau) \\ &\propto f(B \mid \hat{B}_2, \Sigma_{22}, \Sigma_{1\mid 2}, \tau). \end{aligned}$$

Hence, the distribution of $B \mid D$ is the same as that of $B \mid \hat{B}_2$, given the covariance parameters $(\Sigma_{22}, \Sigma_{1\mid 2}, \tau)$.

Combining this last result with (3) and (9) yields

$$B \mid D, \Sigma_{22}, \Sigma_{1\mid 2}, \tau \sim N_{kp}(\mu^*, \Sigma^*), \tag{19}$$

where μ^* and Σ^* can be obtained in the manner now described.

The conditional distribution of $\hat{B}_2$ in (9) can equivalently be represented as

$$\hat{\beta}_2 \mid \beta, \Sigma \sim N_{kg}(e\beta, e(\Sigma \otimes A^{-1})\, e'), \tag{20}$$

where

$$e = (0_{kg \times ku}, I_{kg})$$

with $0_{kg \times ku}$ denoting a $(kg \times ku)$ matrix of zeroes and I_{kg} denoting a $(kg \times kg)$ identity matrix.

Now from general theory,

$$E(\hat{\beta}_2 \mid \beta) = E\hat{\beta}_2 + \Sigma_{\hat{\beta}_2 \beta} \Sigma_{\beta\beta}^{-1}(\beta - E\beta),$$

where for any random vectors R and S,

$$\Sigma_{RS} \equiv \operatorname{cov}(R, S').$$

Thus, conditional on Σ,

$$E(\hat{\beta}_2 \mid \beta) = e\beta^o + \Sigma_{\hat{\beta}_2 \beta} \Sigma_{\beta\beta}^{-1}(\beta - \beta^o).$$

But according to Eq. (20),

$$E(\hat{\beta}_2 \mid \beta) = \varepsilon\beta.$$

It follows that conditional on Σ,

$$\begin{aligned}\Sigma_{\hat{\beta}_2\beta} &= e\Sigma_{\beta\beta}\\ &= e(\Sigma\otimes F^{-1})\\ &= (\Sigma_{22}\tau', \Sigma_{22})\otimes F^{-1},\end{aligned}\tag{21}$$

the second and third of these last equations deriving from (4) and (2), respectively. Moreover, combining (3) and (20) gives, again conditional on Σ,

$$\begin{aligned}\Sigma_{\hat{\beta}_2\hat{\beta}_2} &= e(\Sigma\otimes A^{-1})\,e' + e(\Sigma\otimes F^{-1})\,e'\\ &= e[\Sigma\otimes(A^{-1}+F^{-1})]\,e'\\ &= \Sigma_{22}\otimes(A^{-1}+F^{-1}).\end{aligned}\tag{22}$$

Combining (21) and (22) yields

$$\begin{aligned}\Sigma_{\beta\hat{\beta}_2}\Sigma^{-1}_{\hat{\beta}_2\hat{\beta}_2} &= ((\Sigma_{22}\tau', \Sigma_{22})\otimes F^{-1})'\,(\Sigma_{22}\otimes(A^{-1}+F^{-1}))^{-1}\\ &= \begin{pmatrix}\tau\\ I_g\end{pmatrix}\otimes\hat{E},\end{aligned}\tag{23}$$

where

$$\hat{E} = F^{-1}(A^{-1}+F^{-1})^{-1}.$$

Hence, it follows from general theory that conditional on Σ,

$$\begin{aligned}\mu^* \equiv E\beta\,|\,\hat{\beta}_2 &= E\beta + \Sigma_{\beta\hat{\beta}_2}\Sigma^{-1}_{\hat{\beta}_2\hat{\beta}_2}(\hat{\beta}_2 - E\hat{\beta}_2)\\ &= \beta^o + \begin{pmatrix}\tau\\ I_g\end{pmatrix}\otimes\hat{E}(\hat{\beta}_2 - e\beta^o),\end{aligned}$$

the last equation being a consequence of Eq. (23). Equivalently,

$$\mu^* = B^o + \begin{pmatrix}\tau\\ I_g\end{pmatrix}(\hat{B}_2 - B^o_2)\,\hat{E}'.\tag{24}$$

Similarly,

$$\begin{aligned}\Sigma^* \equiv \Sigma_{\beta\,|\,\hat{\beta}_2} &= \Sigma_{\beta\beta} - \Sigma_{\beta\hat{\beta}_2}\Sigma^{-1}_{\hat{\beta}_2\hat{\beta}_2}\Sigma_{\hat{\beta}_2\beta}\\ &= \Sigma\otimes F^{-1} - \left[\begin{pmatrix}\tau\\ I_g\end{pmatrix}\otimes\hat{E}\right][(\Sigma_{22}\tau', \Sigma_{22})\otimes F^{-1}] \qquad (25)\\ &= \Sigma\otimes F^{-1} - \left[\begin{pmatrix}\tau\\ I_g\end{pmatrix}\Sigma_{22}(\tau', I_g)\right]\otimes(\hat{E}F^{-1}), \qquad (26)\end{aligned}$$

the second equation deriving from (21) and (23). Thus, (19), (24), and (26) yield (10)–(12).

The posterior distributions of $(\Sigma_{22}, \Sigma_{1|2}, \tau)$ can now be obtained. For notational convenience, we use $\Gamma \equiv (\Sigma_{22}, \Sigma_{1|2}, \tau)$ and rewrite the posterior density as

$$\begin{aligned}
f(\Gamma \mid D) &\propto f(\Gamma)\, f(D \mid \Gamma) \\
&\propto f(\Gamma) \int f(D \mid B, \Gamma)\, f(B \mid \Gamma)\, dB \\
&\propto f(\Gamma) \int f(\hat{B}_2 \mid B, \Gamma)\, f(S \mid B, \Gamma)\, f(B \mid \Gamma)\, dB \\
&\propto f(\Gamma)\, f(S \mid \Gamma) \int f(\hat{B}_2 \mid B, \Gamma)\, f(B \mid \Gamma)\, dB \\
&\propto f(\hat{B}_2 \mid \Gamma)\, f(S \mid \Gamma)\, f(\Gamma).
\end{aligned}$$

More explicitly,

$$\begin{aligned}
&f(\Sigma_{22}, \Sigma_{1|2}, \tau \mid D) \propto f(\hat{B}_2 \mid \Sigma_{22}, \Sigma_{1|2}, \tau)\, f(S \mid \Sigma_{22}, \Sigma_{1|2}, \tau)\, f(\Sigma_{22}, \Sigma_{1|2}, \tau) \\
&\propto f(\hat{B}_2 \mid \Sigma_{22})\, f(S \mid \Sigma_{22})\, f(\Sigma_{22})\, f(\tau \mid \Sigma_{1|2})\, f(\Sigma_{1|2}).
\end{aligned}$$

This decomposition shows that the posterior distributions of τ and $\Sigma_{1|2}$ remain the same as their priors and, hence, proves the last two relations in (13). Moreover, it shows that the posterior density of Σ_{22} can be written as

$$f(\Sigma_{22} \mid D) \propto f(\hat{B}_2 \mid \Sigma_{22})\, f(S \mid \Sigma_{22})\, f(\Sigma_{22}). \tag{27}$$

But (3), (9), and (22) imply

$$\hat{B}_2 \mid \Sigma_{22} \sim N_{kg}(B_2^o, \Sigma_{22} \otimes (A^{-1} + F^{-1})^{-1}). \tag{28}$$

Substituting the densities defined by (5), (9), and (28) into $f(\Sigma_{22} \mid D)$ in (27) gives

$$\begin{aligned}
f(\Sigma_{22} \mid D) \propto\ & |\Sigma_{22}|^{-k/2} \exp\{-\tfrac{1}{2}[\mathrm{tr}[\Sigma_{22}^{-1}(\hat{B}_2 - B_2^o) \\
& \times (A^{-1} + F^{-1})^{-1} (\hat{B}_2 - B_2^o)']\} \\
& \times |\Sigma_{22}|^{-(n-k)/2} \exp\{-\tfrac{1}{2}[\mathrm{tr}(\Sigma_{22}^{-1} S)]\} \\
& \times |\Sigma_{22}|^{-(m-u+g+1)/2} \exp\{-\tfrac{1}{2}[\mathrm{tr}(\Sigma_{22}^{-1} \Psi_{22})]\} \\
\propto\ & |\Sigma_{22}|^{-(m+n-u+g+1)/2} \exp\{-\tfrac{1}{2}\mathrm{tr}[\Sigma_{22}^{-1}(\hat{\Psi}_{22})]\}
\end{aligned}$$

where $\hat{\Psi}_{22}$ is given in (14). Thus,

$$\Sigma_{22} \mid D \sim W_g^{-1}(\hat{\Psi}_{22}, m+n-u).$$

This proves the first relation in (13) and hence completes the proof. ∎

Proof of Corollary 2.2. To find the posterior mean of Σ, we need the expectations of all relevant terms in the decomposition (2), namely, Σ_{22}, $\tau\Sigma_{22}$, $\Sigma_{1|2}$, and $\tau\Sigma_{22}\tau'$. It is obvious by the results expressed by Eq. (13) that

$$E\Sigma_{22} = \frac{1}{m+n-u-g-1} \hat{\Psi}_{22}$$

$$E\tau\Sigma_{22} = \frac{1}{m+n-u-g-1} (\eta\Psi_{22})$$

$$E\Sigma_{1|2} = \frac{1}{m-u-1} \Psi_{1|2},$$

where $\hat{\Psi}_{22}$ is in (14).
Moreover,

$$\begin{aligned} E\tau\Sigma_{22}\tau' &= E[E(\tau\Sigma_{22}\tau') \mid \tau] = E[\tau(E(\Sigma_{22})\tau'] \\ &= \frac{1}{m+n-u-g-1} E(\tau\hat{\Psi}_{22}\tau') \\ &= \frac{1}{m+n-p-1} E[E(\tau\hat{\Psi}_{22}\tau') \mid \Sigma_{1|2}] \\ &= \frac{1}{m+n-p-1} E[\eta\hat{\Psi}_{22}\eta' + \mathrm{tr}(\hat{\Psi}_{22}\Psi_{22}^{-1})\,\Sigma_{1|2}] \\ &= \frac{1}{m+n-p-1} (\eta\hat{\Psi}_{22}\eta') \\ &\quad + \frac{\mathrm{tr}(\hat{\Psi}_{22}\Psi_{22}^{-1})}{(m+n-p-1)(m-u-1)} \Psi_{1|2}. \end{aligned}$$

Substituting these in the posterior expectation of (2) gives the posterior mean $\tilde{\Sigma}$,

$$\tilde{\Sigma} = \frac{1}{m+n-p-1}$$
$$\times \begin{pmatrix} \left(\dfrac{m+n-u-g-1+\operatorname{tr}(\hat{\Psi}_{22}\Psi_{22}^{-1})}{m-u-1}\right)\Psi_{1|2}+\eta\hat{\Psi}_{22}\eta' & \eta\hat{\Psi}_{22} \\ \hat{\Psi}_{22}\eta' & \hat{\Psi}_{22} \end{pmatrix}$$
$$= \frac{1}{m+n-p-1}$$
$$\times \begin{pmatrix} \left(1+\dfrac{n-g+\operatorname{tr}(\hat{\Psi}_{22}\Psi_{22}^{-1})}{m-u-1}\right)\Psi_{1|2}+\eta\hat{\Psi}_{22}\eta' & \eta\hat{\Psi}_{22} \\ \hat{\Psi}_{22}\eta' & \hat{\Psi}_{22} \end{pmatrix}$$
$$= \alpha\Sigma_o + (1-\alpha)\hat{\Sigma}.$$

Here $\alpha = (m-p-1)/(m+n-p-1)$, Σ_o denotes the prior mean given by

$$\Sigma_o = \frac{1}{m-p-1}\Psi = \frac{1}{m-p-1}\begin{pmatrix} \Psi_{1|2}+\eta\Psi_{22}\eta' & \eta\Psi_{22} \\ \Psi_{22}\eta' & \Psi_{22} \end{pmatrix},$$
$$\hat{\Sigma} = \begin{pmatrix} \left[\dfrac{1+\operatorname{tr}(\tilde{S}\Psi_{22}^{-1})}{m-u-1}\right]\Psi_{1|2}+(\eta\tilde{S}\eta') & \eta\tilde{S} \\ \tilde{S}\eta' & \tilde{S} \end{pmatrix}$$

and

$$\tilde{S} = \frac{1}{n}\left[S + (\hat{B}_2 - B_2^o)'(A^{-1}+F^{-1})^{-1}(\hat{B}_2 - B_2^o)\right].$$

Hence (15) is proved. ∎

Proof of Theorem 2.3. Write the predictive density of X_f given the covariate z_f, suppressing the prior parameters $(\Psi_{22}, \Psi_{1|2}, \eta)$ for simplicity, as

$$f(X_f \mid D) \propto \int f(X_f \mid D, \Sigma_{22}, \Sigma_{1|2}, \tau)\, f(\Sigma_{22}, \Sigma_{1|2}, \tau \mid D)\, \partial\Sigma_{22}\, \partial\Sigma_{1|2}\, \partial\tau$$
$$\propto \int f(X_f \mid D, \Sigma_{22}, \Sigma_{1|2}, \tau)\, f(\Sigma_{22} \mid D)$$
$$\times f(\tau \mid \Sigma_{1|2}, D)\, f(\Sigma_{1|2} \mid D)\, \partial\Sigma_{22}\, \partial\Sigma_{1|2}\, \partial\tau. \tag{29}$$

The distribution of $(X_f \mid \Sigma_{22}, \Sigma_{1|2}, \tau, D)$ is obtained next.

Recall Eqs. (1) and (10),

$$X_f \mid B, \Sigma_{22}, \Sigma_{1|2}, \tau, D \sim N_{kp}(Bz_f, \Sigma)$$
$$B \mid \Sigma_{22}, \Sigma_{1|2}, \tau, D \sim N(B^*, \Sigma^*),$$

where Σ, B^*, and Σ^* are given by (2), (11), and (12), respectively. Thus,

$$X_f \mid \Sigma_{22}, \Sigma_{1|2}, \tau, D \sim N_p(B^* z_f, \Sigma^{**}). \tag{30}$$

Here, from Eq. (11),

$$B^* z_f = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} \tau \\ I_g \end{pmatrix} b \tag{31}$$

with

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = B^o z_f$$
$$b = (\hat{B}_2 - B_2^o)\, \hat{E} z_f$$
$$\hat{E} = F^{-1}(A^{-1} + F^{-1})^{-1}.$$

And, from Eq. (12),

$$\begin{aligned} \Sigma^{**} &= \Sigma + \Sigma \otimes (z_f' F^{-1} z_f) \\ &\quad - \left[\begin{pmatrix} \tau \\ I_g \end{pmatrix} (\Sigma_{22}\tau', \Sigma_{22}) \right] \otimes (z_f' \hat{E} F^{-1} z_f) \\ &= \Sigma + \Sigma (z_f' F^{-1} z_f) \\ &\quad - \left[\begin{pmatrix} \tau \\ I_g \end{pmatrix} (\Sigma_{22}\tau', \Sigma_{22}) \right] (z_f' \hat{E} F^{-1} z_f), \end{aligned}$$

since $(z_f' F^{-1} z_f)$ and $(z_f' \hat{E} F^{-1} z_f)$ are just scalars. So

$$\Sigma^{**} = c\Sigma - d \begin{pmatrix} \tau \\ I_g \end{pmatrix} (\Sigma_{22}\tau', \Sigma_{22})$$

with $c = 1 + z_f' F^{-1} z_f$ and $d = z_f' \hat{E} F^{-1} z_f$ or

$$\begin{aligned} \Sigma^{**} &= \left[cI_p - d \begin{pmatrix} 0 & \tau \\ 0 & I_g \end{pmatrix} \right] \Sigma \\ &= \begin{bmatrix} cI_u & -\tau d \\ 0 & (c-d)\, I_g \end{bmatrix} \Sigma. \end{aligned}$$

Hence,

$$|\Sigma^{**}| = |c|^u\, |c-d|^g\, |\Sigma_{22}|\, |\Sigma_{1|2}| \tag{32}$$

and

$$(\Sigma^{**})^{-1} = \Sigma^{-1}\begin{pmatrix} cI_u & -\tau d \\ 0 & (c-d)\,I_g \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} \Sigma_{1|2}^{-1} & -\Sigma_{1|2}^{-1}\tau \\ -\tau'\Sigma_{1|2}^{-1} & \Sigma_{22}^{-1}+\tau'\Sigma_{1|2}^{-1}\tau \end{pmatrix} \begin{pmatrix} \frac{1}{c}I_u & \frac{d}{c(c-d)}\tau \\ 0 & \frac{1}{(c-d)}I_g \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{c}\Sigma_{1|2}^{-1} & -\frac{1}{c}\Sigma_{1|2}^{-1}\tau \\ -\frac{1}{c}\tau'\Sigma_{1|2}^{-1} & \frac{1}{c-d}\Sigma_{22}^{-1}+\frac{1}{c}\tau'\Sigma_{1|2}^{-1}\tau \end{pmatrix}$$

$$= \frac{1}{c-d}\begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} + \frac{1}{c}\begin{pmatrix} \Sigma_{1|2}^{-1} & -\Sigma_{1|2}^{-1}\tau \\ -\tau'\Sigma_{1|2}^{-1} & \tau'\Sigma_{1|2}^{-1}\tau \end{pmatrix}. \tag{33}$$

Using (30)–(32), the density

$$\begin{aligned} f(X_f \mid \Sigma_{22}, \Sigma_{1|2}, \tau, D) &\propto |\Sigma_{22}|^{-1/2}\,|\Sigma_{1|2}|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}\begin{pmatrix} X_f^{(1)}-a_1-\tau b \\ X_f^{(2)}-a_2-b \end{pmatrix}'(\Sigma^{**})^{-1}\begin{pmatrix} X_f^{(1)}-a_1-\tau b \\ X_f^{(2)}-a_2-b \end{pmatrix}\right\} \\ &\propto |\Sigma_{22}|^{-1/2}\,|\Sigma_{1|2}|^{-1/2}\exp\{-\tfrac{1}{2}(\mathrm{I}+\mathrm{II})\}. \end{aligned} \tag{34}$$

The quantities I, II in Eq. (34) derive from the decomposition of $(\Sigma^{**})^{-1}$ given in Eq. (33). More explicitly,

$$\begin{aligned} I &= \frac{1}{c-d}\begin{pmatrix} X_f^{(1)}-a_1-\tau b \\ X_f^{(2)}-a_2-b \end{pmatrix}'\begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix}\begin{pmatrix} X_f^{(1)}-a_1-\tau b \\ X_f^{(2)}-a_2-b \end{pmatrix} \\ &= \frac{1}{c-d}(X_f^{(2)}-a_2-b)'\,\Sigma_{22}^{-1}(X_f^{(2)}-a_2-b) \end{aligned} \tag{35}$$

and

$$\begin{aligned} \mathrm{II} &= \frac{1}{c}\begin{pmatrix} X_f^{(1)}-a_1-\tau b \\ X_f^{(2)}-a_2-b \end{pmatrix}'\begin{pmatrix} \Sigma_{1|2}^{-1} & -\Sigma_{1|2}^{-1}\tau \\ -\tau'\Sigma_{1|2}^{-1} & \tau'\Sigma_{1|2}^{-1}\tau \end{pmatrix}\begin{pmatrix} X_f^{(1)}-a_1-\tau b \\ X_f^{(2)}-a_2-b \end{pmatrix} \\ &= \frac{1}{c}\{[X_f^{(1)}-a_1-\tau b)'\,\Sigma_{1|2}^{-1}-(X_f^{(2)}-a_2-b)'\,\tau'\Sigma_{1|2}^{-1}] \\ &\quad \times (X_f^{(1)}-a_1-\tau b)+[-(X_f^{(1)}-a_1-\tau b)'\,\Sigma_{1|2}^{-1}\tau \\ &\quad +(X_f^{(2)}-a_2-b)'\,\tau'\Sigma_{1|2}^{-1}\tau](X_f^{(2)}-a_2-b)\} \end{aligned}$$

$$=\frac{1}{c}\{[(X_f^{(1)}-a_1)'\,\Sigma_{1|2}^{-1}-(X_f^{(2)}-a_2)'\,\tau'\Sigma_{1|2}^{-1}]\,(X_f^{(1)}-a_1-\tau b)$$

$$+[-(X_f^{(1)}-a_1)'\,\Sigma_{1|2}^{-1}\tau+(X_f^{(2)}-a_2)'\,\tau'\Sigma_{1|2}^{-1}\tau]\,(X_f^{(2)}-a_2-b)\}$$

$$=\frac{1}{c}\{(X_f^{(1)}-a_1)'\,\Sigma_{1|2}^{-1}(X_f^{(1)}-a_1)-(X_f^{(1)}-a_1)'\,\Sigma_{1|2}^{-1}\tau b$$

$$-(X_f^{(2)}-a_2)'\,\tau'\Sigma_{1|2}^{-1}(X_f^{(1)}-a_1)+(X_f^{(2)}-a_2)'\,\tau'\Sigma_{1|2}^{-1}\tau b$$

$$-(X_f^{(1)}-a_1)'\,\Sigma_{1|2}^{-1}\tau(X_f^{(2)}-a_2)+(X_f^{(1)}-a_1)'\,\Sigma_{1|2}^{-1}\tau b$$

$$+(X_f^{(2)}-a_2)'\,\tau'\Sigma_{1|2}^{-1}\tau(X_f^{(2)}-a_2)-(X_f^{(2)}-a_2)'\,\tau'\Sigma_{1|2}^{-1}\tau b\}$$

$$=\frac{1}{c}\{(X_f^{(1)}-a_1)'\,\Sigma_{1|2}^{-1}(X_f^{(1)}-a_1)-2(X_f^{(2)}-a_2)'$$

$$\times\tau'\Sigma_{1|2}^{-1}(X_f^{(1)}-a_1)+(X_f^{(2)}-a_2)'\,\tau'\Sigma_{1|2}^{-1}\tau(X_f^{(2)}-a_2)\}$$

$$=\frac{1}{c}\{[\tau(X_f^{(2)}-a_2)-(X_f^{(1)}-a_1)]'$$

$$\times\Sigma_{1|2}^{-1}[\tau(X_f^{(2)}-a_2)-(X_f^{(1)}-a_1)]\}. \tag{36}$$

Substituting (34) to (29) yields

$$f(X_f|D)\propto\int_{\Sigma_{22}}|\Sigma_{22}|^{-1/2}\exp\left\{-\frac{1}{2}\mathrm{I}\right\}f(\Sigma_{22}|D)\,\partial\Sigma_{22}$$

$$\times\int_{\Sigma_{1|2}}\int_{\tau}|\Sigma_{1|2}|^{-1/2}\exp\left\{-\frac{1}{2}\mathrm{II}\right\}$$

$$\times f(\tau|\Sigma_{1|2},D)\,\partial\tau f(\Sigma_{1|2}|D)\,\partial\Sigma_{1|2}. \tag{37}$$

But, by Eqs. (13) and (35),

$$\int_{\Sigma_{22}}|\Sigma_{22}|^{-1/2}\exp\left\{-\frac{1}{2}\mathrm{I}\right\}f(\Sigma_{22}|D)\,\partial\Sigma_{22}$$

$$\propto\int|\Sigma_{22}|^{-1/2}\,|\Sigma_{22}|^{-(m+n-u+g+1)/2}$$

$$\times\exp\left\{-\frac{1}{2(c-d)}(X_f^{(2)}-a_2-b)'\right.$$

$$\left.\times\Sigma_{22}^{-1}(X_f^{(2)}-a_2-b)-\frac{1}{2}\operatorname{tr}\hat{\Psi}_{22}\Sigma_{22}^{-1}\right\}\partial\Sigma_{22}$$

$$\propto \int |\Sigma_{22}|^{-(m+n-u+g+2)/3}$$

$$\times \exp\left\{-\frac{1}{2}\operatorname{tr}\Sigma_{22}^{-1}\left[\hat{\Psi}_{22}+\frac{1}{(c-d)}(X_f^{(2)}-a_2-b)\right.\right.$$

$$\left.\left.\times(X_f^{(2)}-a_2-b)'\right]\right\}\partial\Sigma_{22}$$

$$\propto \left|\hat{\Psi}_{22}+\frac{1}{(c-d)}(X_f^{(2)}-a_2-b)(X_f^{(2)}-a_2-b)'\right|^{-(m+n-u+1)/2} \tag{38}$$

Similarly, by Eq. (36),

$$\int_\tau |\Sigma_{1|2}|^{-1/2}\exp\left\{-\frac{1}{2}\,\mathrm{II}\right\} f(\tau\,|\,\Sigma_{1|2}, D)\,\partial\tau$$

$$\propto \int |\Sigma_{1|2}|^{-1/2}\exp\left\{-\frac{1}{2c}\left[\tau(X_f^{(2)}-a_2)-(X_f^{(1)}-a_1)\right]'\right.$$

$$\left.\times\Sigma_{1|2}^{-1}\left[\tau(X_f^{(2)}-a_2)-(X_f^{(1)}-a_1)\right)\right\}$$

$$\times|\Sigma_{1|2}\otimes\Psi_{22}^{-1}|^{-1/2}\exp\left\{-\frac{1}{2}(\tau-\eta)'(\Sigma_{1|2}\otimes\Psi_{22}^{-1})^{-1}(\tau-\eta)\right\}\partial\tau.$$

This integral is equivalent to the marginal density of $(X_f^{(1)}-a_1)$ when

$$(X_f^{(1)}-a_1)\,|\,\tau\sim N_u(\tau(X_f^{(2)}-a_2), c\Sigma_{1|2})$$

and

$$\tau\sim N_{ug}(\eta, \Sigma_{1|2}\otimes\Psi_{22}^{-1}).$$

So from the general theory of the multivariate normal distribution,

$$(X_f^{(1)}-a_1)\sim N_u(\eta(X_f^{(2)}-a_2), \Sigma_{1|2}(c+(X_f^{(2)}-a_2)'\,\Psi_{22}^{-1}(X_f^{(2)}-a_2)))$$

$$\sim N_u(\eta(X_f^{(2)}-a_2), w\Sigma_{1|2}),$$

where $w=(c+(X_f^{(2)}-a_2)'\,\Psi_{22}^{-1}(X_f^{(2)}-a_2))$.

Using this last expression and Eq. (38) in (37) gives

$$f(X_f \mid D) \propto \left| \hat{\Psi}_{22} + \frac{1}{(c-d)} (X_f^{(2)} - a_2 - b)(X_f^{(2)} - a_2 - b)' \right|^{-(m+n-u+1)/2}$$

$$\times \int |w\Sigma_{1|2}|^{-1/2} \exp\left\{ -\frac{1}{2w} v'\Sigma_{1|2}^{-1} v \right\} f(\Sigma_{1|2} \mid D)\, \partial\Sigma_{1|2}$$

$$\propto \left| \hat{\Psi}_{22} + \frac{1}{(c-d)} (X_f^{(2)} - a_2 - b)(X_f^{(2)} - a_2 - b)' \right|^{-(m+n-u+1)/2}$$

$$\times \int |w\Sigma_{1|2}|^{-1/2} |\Sigma_{1|2}|^{-(m+u+1)/2}$$

$$\times \exp - \frac{1}{2w} v'\Sigma_{1|2}^{-1} v - \frac{1}{2} \operatorname{tr} \Sigma_{1|2} \Psi_{1|2} \Big\} \partial\Sigma_{1|2},$$

where $v = (X_f^{(1)} - a_1) - \eta(X_f^{(2)} - a_2)$

$$\propto \left| \hat{\Psi}_{22} + \frac{1}{(c-d)} (X_f^{(2)} - a_2 - b)(X_f^{(2)} - a_2 - b)' \right|^{-(m+n-u+1)/2}$$

$$\times |w|^{-u/2} \int |\Sigma_{1|2}|^{-(m+u+2)/2}$$

$$\times \exp\left\{ -\frac{1}{2} \operatorname{tr} \Sigma_{1|2}^{-1} \left[\frac{1}{w} vv' + \Psi_{1|2} \right] \right\} \partial\Sigma_{1|2}$$

$$\propto \left| \hat{\Psi}_{22} + \frac{1}{(c-d)} (X_f^{(2)} - a_2 - b)(X_f^{(2)} - a_2 - b)' \right|^{-(m+n-u+1)/2}$$

$$\times |w|^{-u/2} \left| \frac{1}{w} vv' + \Psi_{1|2} \right|^{-(m+1)/2}$$

$$\propto |\hat{\Psi}_{22}|^{-1/2} (1 + (X_f^{(2)} - a_2 - b)((c-d)\, \hat{\Psi}_{22})^{-1}$$
$$\times (X_f^{(2)} - a_2 - b)')^{-(m+n-u+1)/2}$$
$$\times |w\Psi_{1|2}|^{-1/2} (1 + v'(w\Psi_{1|2})^{-1} v)^{-(m+1)/2} \tag{39}$$

Relation (39) shows that

$$X_f^{(1)} \mid X_f^{(2)} = x_f^{(2)}, D$$
$$\sim t_g\left(a_1 + \eta, (x_f^{(2)} - a_2), \frac{c + (a_2 - x_f^{(2)})' \Psi_{22}^{-1} (a_2 - x_f^{(2)})}{q} \Psi_{1|2}, q \right)$$

$$X_f^{(2)} \mid D \sim t_u\left(a_2 + b, \frac{c-d}{l} \hat{\Psi}_{22}, l \right),$$

where $q = m - u + 1$ and $l = m + n - u - g + 1$. This completes the proof. ∎

Proof of Corollary 2.4. The second equation in (18) is obtained by just taking the expectation of a multivariate-t distribution. The first one is obtained as

$$\begin{aligned} EX_f^{(1)} &= E[EX^{(1)} \mid X_f^{(2)}] \\ &= a_1 + \eta(EX_f^{(2)} - a_2) \\ &= a_1 + \eta(a_2 + b - a_2) \\ &= a_1 + \eta b. \quad \blacksquare \end{aligned}$$

4. Concluding Remarks

In this paper we have developed an alternative to Kriging for spatial interpolation. In the Introduction we indicate how it differs from Kriging and how it relates to other Bayes methods in the same context.

The method proposed here is very general and such things as trends and seasonality are readily incorporated in the spatial–temporal models while allowing model misspecifications to be corrected as data becomes available. At the same time, uncertainty about the model is honestly reflected in the heavier tails of the (posterior) distributions presented in Section 2. This makes the posterior distribution somewhat robust against model misspecification. And it ensures that (credible) regions for the interpolants are not misleadingly small, as such regions can be in Kriging when the variability of the spatial covariance matrix estimator is ignored while using it as if it were known, to compute the optimal linear interpolator.

Kriging has been adopted in *environmetrics*, although its roots are in *geostatistics*. And there are important differences between these two contexts which makes Kriging less successful in the former than in the latter. In environmetrics unlike geostatistics, one must commonly deal with time series of spatial observations. Hence, there is a need for methods which allow one to incorporate temporal models. Moreover, the data series can be used to update these and other models and there is a need for a theory which does this in some systematic way. By contrast, spatial modeling can be much less important and meaningful in environmetrics than in geostatistics (see the discussion at the end of Section 2). Thus a model like that of Fedorov and Mueller [5, 6], which represents monitoring station means in terms of functions of its location, is not so compelling in environmetrical applications as in those of geostatistics.

Currently we are working on the implementation of our theory in particular, the specification of the hyperparameters. Of special interest is the prior covariance structure, Ψ. One possibility would be an isotropic parametric prior model like that of Loader and Switzer [8]. One could then add an additional level in the prior distribution's hierarchy as discussed in Section 2, or more simply estimate the parameters involved, an approach akin to adopting a diffuse prior on the remaining unspecified hyperparameters. This last stage of hierarchical modeling has the important effect of tying the "gauged" and "ungauged" stations together so that now the available data does get drawn in to the estimation of covariances and cross-covariances involving the ungauges sites. In short this additional step allows us to "borrow from strength."

In general, the assumption of isotropy seems excessively strong. To us a more appealing idea, the one we are currently adopting in a design context, is that of estimating, Ψ, using the nonparametric approach (see [15, 7]).

An anonymous referee has correctly pointed out that in some situations it may not be feasible to identify a sufficiently accurate structural model for the underlying time series, as to validate our assumption of no autocorrelation. In this case, an extension of our methodology would be required.

Acknowledgments

The authors thank a referee whose comments have improved the presentation. Their research was partially by a grant from Natural Sciences and Engineering Research Council of Canada.

References

[1] Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

[2] Bartlett, M. S. (1933). On the theory of statistical regression. *Proc. Roy. Soc. Edinburgh* **53** 260–283.

[3] Caselton, W. F., Kan, L., and Zidek, J. V. (1992). Quality data networks that minimize entropy. In *Statistics in the Environmental and Earth Sciences* (P. Guttorp and A. Walden, Eds.). Griffin, London.

[4] de Waal, D. J., and Groenevald, P. C. N. (Undated). "Spatial Bayesian estimation." Unpublished manuscript.

[5] Federov, V., and Muller, W. (1988). Two approaches in optimization of observing networks. In *Optimal Design and Analysis of Experiments* (Y. Dodge, V. V. Fedorov, and H. P. Wynn, Eds.). Elsevier Science, Amsterdam/New York.

[6] Federov, V., and Muller, W. (1989). Comparison of two approaches in the optimal design of an observation network. *Statistics* **3** 339–351.

[7] Guttorp, P., Sampson, P.D., and Newman, K. (1991). Nonparametric estimation of spatial covariance and application monitoring network evaluation. In *Statistics in the Environmental and Earth Sciences* (P. Guttorp and A. Walden, Eds.). Griffin, London.

[8] Loader, C., and Switzer, P. (1989). "Spatial Covariance Estimation for Monitoring Data." SIMS Technical Report 133. Stanford University.

[9] MATHERON, G. (1971). "The Theory of Regionalized Variables and Its Applications." Centre de Geostatistique, Ecole des Mines and Paris.

[10] OMRE, H. (1987). Bayesian Kriging-merging observations and qualified guesses in Kriging, *Math. Geol.* **19** 25–39.

[11] OMRE, H., AND HALVORSEN, K. B. (1989). The Bayesian bridge between simple and universal Kriging, *Math. Geol.* **21** 767–786.

[12] OMRE, H., HALVORSEN, K. B., AND BERTIG, V. (1989). A Bayesian approach to Kriging. In *Geostatistics*, Vol. 1 (M. Armstrong, Ed.). Kluwer Academic, Norwell, MA.

[13] PILZ, J. (Undated). Robust Bayes linear prediction of regionalized variables. Unpublished manuscript.

[14] PRESS, J. S. (1972). *Applied Multivariate Analysis.* Holt, Rinehart, & Winston, New York.

[15] SAMPSON, P., AND GUTTORP, P. (1992). Spatial covariance estimation by scaled-meeting scaling and biorthogonal grids. *J. Amer. Statist. Assoc.*, to appear.

[16] SWITZER, P. (1989). Non-stationary spatial covariances estimated from monitoring data. In *Geostatistics*, Vol. 1 (M. Armstrong, Ed.). Kluwer Academic, Norwell, MA.

[17] WU, S., AND ZIDEK, J. V. (1992). An entropy based review of selected nadp/ntn network sites for 1983–86. *Atmosph. Environ.* **26A** 2089–2103.