A new class of nonstationary spatial models

Montserrat Fuentes\(^1\) and Richard L. Smith\(^2\)

ABSTRACT

Spatial processes are an important modeling tool for many problems of environmental monitoring. Classical geostatistics is based on processes which are stationary and isotropic, but it is widely recognized that real environmental processes are rarely stationary and isotropic. In this paper, a new class of nonstationary processes is proposed, based on a convolution of local stationary processes. This model has the advantage that the model is simultaneously defined everywhere, unlike "moving window" approaches, but it retains the attractive property that locally in small regions, it behaves like a stationary spatial processes. We discuss model fitting through exact and approximate likelihood maximization, and propose a hierarchical Bayes approach to allow predictive inference when the parameters of the model are unknown. Applications include obtaining the total loading of sulfur dioxide concentrations over different geo-political boundaries.

---

\(^1\)Montserrat Fuentes is an assistant professor at the Statistics Department, North Carolina State University (NCSU), Raleigh, NC 27695-8203, and a visiting scientist at the US Environmental Protection Agency (EPA). Tel.: (919) 515-1921, Fax: (919) 515-1169, E-mail: fuentes@stat.ncsu.edu. This research was sponsored by a National Science Foundation grant DMS 0002790, and by a US EPA award R-8287801.

\(^2\)Richard L. Smith is a professor at the Statistics Department, University of North Carolina, Chapel Hill, NC. Tel: (919) 962-2660, Fax: (919) 962-1279, ris@email.unc.edu.

Key words: air pollution, Bayesian statistics, change of support, kriging, Matérn covariance, nonstationary process, spatial-temporal statistics.
1 Introduction

A major focus of the Clean Air Act Amendments of 1990, has been the effect of atmospherically-transported pollutants on terrestrial and aquatic ecosystems. The 1990 amendments include new requirements that will appreciably reduce sulfur dioxide ($SO_2$) emissions. Monitoring data for $SO_2$ are analyzed as part of the process for assessing compliance with the Clean Air Act Amendments of 1990. The ground-level concentration of $SO_2$ will depend on the proximity to the source, the prevailing meteorology, and the nature and extent of atmospheric chemical reactions between the source and receptor. The interaction of these chemical and physical atmospheric processes and the source locations tend to produce data patterns that show large spatial variability. Because spatial patterns of $SO_2$ fluxes and concentrations are nonstationary processes, in the sense that the spatial structure changes with location, standard methods of spatial interpolation are inadequate. We present a new statistical methodology for prediction of nonstationary processes, with the objective of obtaining in an efficient way the total loading of $SO_2$ concentrations and fluxes over different geo-political boundaries.

In recent years, probably the most extensively studied method for nonstationary spatial processes is the deformation approach due to Sampson and Guttorp (1992); see also Guttorp and Sampson (1994), Guttorp, Meiring and Sampson (1994). Maximum likelihood versions of the method were developed by Mardia and Goodall (1993) and Smith (1996). In a series of papers best represented by Haas (1995), T. Haas has proposed an approach to nonstationary spatial kriging based on moving windows. Higdon, Swall and Kern (1999) give a model for accounting for heterogeneity in the spatial covariance function of a spatial process, using a moving average specification of a Gaussian process. Another approach has been developed by Nychka and Saltzman (1998) and Holland et al. (1999), that extends the “empirical orthogonal functions” (EOF) approach that is popular among atmospheric scientists.

In this paper we give a new methodology for spatial interpolation of nonstationary processes. More specifically, we represent the process locally as a stationary isotropic random field with some parameters that describe the local spatial structure. These parameters are allowed to vary across space and reflect the lack of stationarity of the process. We represent a nonstationary process $Z$ observed on a region $D$ as a
convolution of local stationary processes:

$$Z(x) = \int_D K(x - s)Z_{\theta(s)}(x)ds.$$  \hspace{1cm} (1)

where $K$ is a kernel function and $Z_{\theta(x)}$, $x \in D$ is a family of (independent) stationary Gaussian processes indexed by $\theta$.

If $K$ is a sharply peaked kernel function and $\theta(s)$ varies slowly with $s$, this has the property that for $x$ near $s$, the process “looks like” a stationary process with parameter $\theta(s)$. On the other hand, since $\theta(s)$ may vary substantially over the whole space, it also allows significant nonstationarity. The method has features in common with Haas’s approach, but in view of the representation (1), there is no problem about it being a well-defined process with a positive definite covariance function. We could even use a variant of Haas’s approach to estimate the model, for instance, by estimating $\theta(s)$ for a finite set of values of $s$ assuming stationarity within some window, and then smoothing the function $\theta(s)$ by kernels or splines. In the discussion to follow, however, we shall outline a number of more sophisticated estimation procedures.

This paper is organized as follows. Section 2 is a statement of the scientific problem discussed in this paper with a description of the environmental data. In section 3, we describe the proposed model for nonstationary processes, we present fitting algorithms to estimate the spatial structure and a geostatistical approach for prediction. Section 4 discusses prediction from a Bayesian point of view, to take into consideration in the prediction the uncertainty in the covariance parameters. In Section 5, we discuss the change-of-support problem that occurs when the supports of predictand and data are not same. Section 6 is an application of the methodology presented in this paper to the $SO_2$ data. Finally, in Section 7 we present some conclusions and final remarks.

2 Description of the data

We have two sources of data for fluxes and concentrations of $SO_2$:

1. The first sources of information are the regional scale air quality models. These models, e.g. Models-3, are run by EPA and the U.S. States and provide $SO_2$ areal concentrations and fluxes in regular grids in parts of the US (see Figure 1). The current resolution of Models-3 is $36 \times 36$ km.
2. EPA provides point measurements at 50 irregularly spaced sites in the eastern U.S. (see Figure 2). At each site, EPA measures concentrations and fluxes of different atmospheric pollutants.

One of the main objectives of the work presented here is to evaluate Models-3, we need then measures of how well Models-3 output and real data agree. We present a methodology that uses Models-3 output to come up with probabilistic predictions of $SO_2$ concentrations at the point measurement sites. In our approach we take into account the change of support problem between Models-3 (areal measurements) and CASTNet (point measurements). This application requires, as a first step, fitting some random field model to Models-3, but it is unrealistic to assume that a stationary isotropic model applies over such a large geographical area. Therefore, we need nonstationary spatial models.

3 A new model for nonstationary spatial processes

Consider a Gaussian spatial process $Z(\mathbf{x})$, where $\mathbf{x}$ varies over a domain $D$ contained in a $d$-dimensional Euclidean space $\mathbb{R}^d$ for some $d > 1$. Typically, $d = 2$. A spatial process $Z$ is generally modeled as follows:

$$Z(\mathbf{x}) = \mu(\mathbf{x}) + \epsilon(\mathbf{x})$$

where $\mu(\mathbf{x})$ represents the spatial trend (large scale structure), and the process $\epsilon$ represents some spatially correlated zero-mean noise. The covariance of the process $\epsilon$ is defined as:

$$\text{cov}(\epsilon(\mathbf{s}_1), \epsilon(\mathbf{s}_2)) = C(\mathbf{s}_1, \mathbf{s}_2; \theta)$$

where $\theta$ is the parameter associated with the covariance function. In our model, the parameter function $\theta$ will be allowed to vary across space to reflect the lack of stationarity of the process.

We represent $Z$ as a convolution of local stationary processes:

$$Z(\mathbf{x}) = \int_D K(\mathbf{x} - \mathbf{s}) Z_{\theta(\mathbf{s})}(\mathbf{x}) d\mathbf{s}. \quad (2)$$

where $K$ is a kernel function and $Z_{\theta(\mathbf{s})}(\mathbf{x}), \mathbf{x} \in D$ is a family of (independent) stationary Gaussian processes indexed by $\theta$.

The covariance of $Z_{\theta(\mathbf{s})}$ is stationary with parameter $\theta(\mathbf{s})$,

$$\text{cov} \{ Z_{\theta(\mathbf{s}_1)}(\mathbf{s}_1), Z_{\theta(\mathbf{s})}(\mathbf{s}_2) \} = C_{\theta(\mathbf{s})}(\mathbf{s}_1 - \mathbf{s}_2).$$
The covariance \( C(s_1, s_2; \theta) \) of \( Z \) is a convolution of the local covariances \( C_{\theta(s)}(s_1 - s_2) \),

\[
C(s_1, s_2; \theta) = \int_D K(s_1 - s)K(s_2 - s)C_{\theta(s)}(s_1 - s_2)ds.
\]

We assume that \( \theta(s) \) is a continuous function on \( s \). The process \( Z_{\theta(s)} \) could have a Matérn stationary covariance:

\[
C_{\theta(s)}(x) = \frac{\sigma_s^2}{2^{\nu_s - 1} \Gamma(\nu_s)\alpha_s^{\nu_s}} (2\nu_s^{1/2}|x|/\rho_s)^{\nu_s} K_{\nu_s}(2\nu_s^{1/2}|x|/\rho_s),
\]

where \( K_{\nu_s} \) is a modified Bessel function and \( \theta(s) = (\nu_s, \sigma_s, \rho_s) \), with the parameter \( \rho_s \) measuring how the correlation decays with distance, generally this parameter is called the range; \( \sigma_s \) is the variance of the random field, i.e. \( \sigma_s = \text{var}(Z_{\theta(s)}(x)) \), the covariance parameter \( \sigma_s \) is usually referred to as the sill; and the parameter \( \nu_s \) measures the degree of smoothness of the process \( Z_{\theta(s)} \), the higher the value of \( \nu_s \) the smoother \( Z_{\theta(s)} \) would be, e.g. when \( \nu_s = \frac{1}{2} \), we get the exponential covariance function. If we consider the limit as \( \nu_s \to \infty \) we get the Gaussian covariance

\[
C_{\theta(s)}(x) = \sigma_s e^{-|x|^2/\rho_s^2}.
\]

Each stationary process \( Z_{\theta(s)}(x) \) has a mean function \( \mu_s \) that is constant, i.e. \( \mu_s \) does not depend on \( x \). However, the mean of the nonstationary process \( Z \) could change with location. The mean of \( Z \) is a convolution of the means for the stationary processes \( Z_{\theta(s)} \). We have

\[
E\{Z(x)\} = \mu(x) = \int_D K(x - s)\mu_s ds.
\]

As an alternative approach, we propose a parametric model for the mean of \( Z \),

\[
E\{Z(x)\} = \mu(x; \beta),
\]

where \( \mu \) could be a polynomial function on \( x \) with coefficients \( \beta \).

In (3) every entry requires an integration. Since each such integration is actually an expectation with respect to a uniform distribution, then we propose Monte Carlo integration. We propose to draw an independent set of locations \( s_m, m = 1, 2, ..., M \) over \( D \). Hence, we replace \( C(s_1, s_2; \theta) \) with

\[
\hat{C}(s_1, s_2; \theta) = M^{-1} \sum_{m=1}^M K(s_1 - s_m)K(s_2 - s_m)C_{\theta(s_m)}(s_1 - s_2)
\]

(5)
In this notation, the “hat” denotes a Monte Carlo integration which can be made arbitrarily accurate and has nothing to do with the data $Z$.

It is useful to note that, if we define a process $\hat{Z}$ for any $s \in D$ as follows:

$$\hat{Z}(s) = M^{-1} \sum_{m=1}^{M} K(s - s_m) Z_{\theta(s_m)}(s)$$  \hspace{1cm} (6)

then $\hat{Z}(s)$ is a Monte Carlo integration for $Z(s)$ as given in (2). As we increase $M$ and the number of observations $s_1, \ldots, s_M$ becomes more dense on the domain $D$, then the covariance function of the process $\hat{Z}$, $\hat{C}$, converges to $C$, which the covariance function of the process $Z$. Thus, $\hat{Z}(s)$ as defined in (6) converges in distribution to $Z(s)$. This type of asymptotic model, where as we increase $M$ the locations become more dense over the fixed domain, is called fixed-domain asymptotics or infill asymptotics.

The size of the sample, $M$, is selected using the following iterative algorithm. We first start with a systematic sample of size $M$, where $M$ is small, and we increase $M$ by adding a new sample point at a time. At each step of the iterative approach we draw a new sample point in between two neighboring points in the current sample sequence. Thus, in each iteration we decrease by half the distance between two neighboring draws. We iterate this process till a Akaike information criterion (AIC) (Akaike, 1974) suggests no significant improvement in the estimation the estimation of the nonstationary covariance of $Z$ by increasing $M$ (i.e. decreasing the distance between draws in the sample sequence).

The optimal predictor of $Z(s_0)$, based on observations of the process $Z$ at locations $s_1, \ldots, s_N$ over the fixed-domain $D$, is $E(Z(s_0)|Z)$ where $Z = (Z(s_1), \ldots, Z(s_N))$. If we assume that $Z$ is Gaussian and the parameters are known functions then this optimal predictor is the simple kriging predictor. If we assume unknown mean, then the optimal predictor would be superior to the universal-kriging predictor.

The conditional distribution of $Z(s_0)$ given $Z$ is

$$f(Z(s_0)|Z, \beta, \theta) = N(\mu_0, V_\theta)$$  \hspace{1cm} (7)

where $\mu_0$ is the conditional mean (kriging predictor) and $V_\theta$ the covariance matrix of $Z(s_0)$ given $Z$. We replace (7) with

$$\hat{f}(Z(x_0)|Z, \beta, \theta) = N(\mu_0, \hat{V}_\theta).$$

Thus, the solution $\hat{E}(Z(s_0)|Z)$ is actually $E(\hat{Z}(s_0)|\hat{Z})$. As we increase $M$, $\hat{E}(Z(s_0)|Z)$ converges to the
optimal predictor of $Z(s_0)$ given $Z$. Assuming all parameters known, $E(\tilde{Z}(s_0)|\bar{Z})$ can be easily calculated using the simple kriging equations for the process $\tilde{Z}$.

3.1 Choice of weight function and bandwidth

The weight function $K(s-s_m)$ could be simply the square inverse distance between $s$ and $s_m$. In this section we discuss another approach, the representation of the weight as a kernel function with compact support and bandwidth $h$. As in the related problem of density function estimation (Bartlett, 1963) we may expect the behavior of the spatial predictor to depend more critically on the bandwidth $h$ than the mathematical form of the kernel function $K$. The results given by Priestly and Chao (1972) imply that the asymptotic mean-square-error of a kernel regression estimator is almost independent on the choice of the kernel function, although there is some marginal advantage in using $K_4$, the quadratic weight-function given by

$$K_4(u) = \frac{3}{4}(1 - u^2).$$

Similar conclusions have been reached by Epanechnikov (1969) and Rosenblatt (1971). This weight function was also recommended by Clark (1977) and it is the kernel function used in this paper. The convolution-based estimator used in this paper was first introduced by Clark (1977) in the following regression setting. Supposed we are given observations $\{(x_i, y_i), i = 1, \ldots, n\}$ satisfying the model

$$y_i = F(x_i) + \varepsilon_i$$

where the $\varepsilon_i$ are independent random variables with zero mean and constant variance $\sigma^2$, and the abscissae $\{x_i\}$ are assumed to be known without error. The problem is to estimate the unknown regression function $F$. Clark (1977) proposed an estimator $f$ of $F$ defined by a kernel convolution;

$$f(x) = \int K(x-t)f_t dt$$

where $K$ is a kernel function with certain bandwidth $h$, and $f_t$ is the value at the abscissa $t$ of a first-order interpolating spline for $\{(x_i, y_i), i = 1, \ldots, n\}$. Some simulations studies ran by Clark (1977) showed that the behavior of the convolution-based regression estimator using $K_4$ is very similar to a cubic spline, but the convolution-based estimator offers important computational advantages over spline functions. Clark (1977) also showed that the convolution-based regression estimator is again almost independent of the choice of $K$.  

7
In practice, the choice of $h$ is crucial, since this parameter determines the degree of smoothing. To preserve the general "shape" of the data the bandwidth should be small (Clark, 1977).

In a regression setting, reducing the size of the bandwidth reduces the bias but increases the variance. In our spatial setting, since the variance might change with location, we actually do not gain anything by increasing $h$. The shape of the process is represented by the parameter $\theta$, which accounts for the lack of stationarity of $Z$. Thus, we need to choose $h$ as small as possible to preserve this general shape. Ideally we would like $h$ to be zero. However, we need to ensure that each point of prediction $x$ has at least one weight $K(x - s_i)$, for $i = 1, \ldots, M$, greater than zero, where the $\{s_i\}$ are the $M$ draws from the systematic sample. Thus, we choose the smallest value of $h$ such that for all $x \in D$ there is at least one $s_i$, with $K(x - s_i) > 0$. For instance, if we use $K_4$ and have $M$ draws separated by a distance $l$ the optimal value of $h$ is $l/\sqrt{2}$.

When the distance between neighboring points of the sample sequence $s_1, \ldots, s_M$ varies, we could also allow the bandwidth to change with location. Thus, we define $K(u - s_i) = \frac{1}{h_i} K \left( \frac{u - s_i}{h_i} \right)$, for $i = 1, \ldots, M$. In this setting, we should again choose the minimum possible value for each bandwidth, ensuring that for each $u \in D$ if $s_i$ is the closest sample point to $u$, then the bandwidth $h_i$ should satisfy $K(u - s_i) > 0$ where $K(u - s_i) = \frac{1}{h_i} K \left( \frac{u - s_i}{h_i} \right)$.

Our convolution model could be also considered a generalization to spatial problems of the delta sequences used for probability density estimation (eg. Walter and Blum (1979)). With delta sequences, small values of $h$ are ideal to preserve the “shape”.

### 3.2 Geostatistical approach

In a geostatistical approach, the covariance structure is estimated first and then the estimated covariance is used for interpolation. Thus, we could estimate the values of the parameters $\beta$, and $\theta(s)$ for all $s \in D$ using a likelihood approach. The likelihood here is well defined

$$f(Z|\beta, \theta) = N(\mu_\beta, C_\theta)$$

where $\mu_\beta = [\mu(s_1; \beta), \ldots, \mu(s_N; \beta)]^T$, and $C_\theta$ is the covariance matrix of $Z$ obtained in (3). We replace $f(Z|\beta, \theta)$ with

$$\hat{f}(Z|\beta, \theta) = N(\hat{\mu}_\beta, \hat{C}_\theta)$$
which is $f(\hat{Z}|\beta, \theta)$, the likelihood for the process $\hat{Z}$, $\hat{C}_\theta$ is the covariance of $Z$ defined in (5). We have seen
that $\hat{Z}(s)$ converges in distribution to $Z(s)$. Thus, we could use the likelihood for $\hat{Z}$ to estimate the values of the parameters
$\beta, \theta(s_1), \ldots, \theta(s_M)$. A pseudo-kriging prediction for $Z$ could be obtained using the form of the covariance for $\hat{Z}$ and the maximum likelihood estimated parameters from $\hat{f}$. This pseudo-kriging
predictor asymptotically has the same properties as the kriging predictor for $Z$. However, in practice this
approach is much easier to implement than estimating $\theta(s)$ for all $s \in D$, and then use the covariance of $Z$
to get the kriging predictor.

In principle, if $Z$ is Gaussian is straightforward to write down the exact likelihood function and hence
to maximize it numerically with respect to the unknown parameters. Kitanidis (1983) and Mardia and
Marshall (1984) were the first to advocate estimating spatial processes in this way. The evaluation of
the likelihood function requires computing the inverse and determinant of the model covariance matrix. In
general, environmental datasets are very large and calculating the determinants that we have in the likelihood
function can be often infeasible. Spectral methods could be used to approximate the likelihood and obtain
the maximum likelihood estimate of the covariance parameters, avoiding the calculation of determinants
and inverse of matrixes. Whittle (1954) proposed the following approximation to the Gaussian negative log
likelihood of a stationary process $Z_i$ observed on a grid $n_1 \times n_2$ with $N = n_1 n_2$, here we write $Z_i$ to denote
$Z_{\theta(s_i)}$:

$$
\sum_{\mathbf{j} \in J_N} \left\{ \log f_i(2\pi \mathbf{j}/\mathbf{n}) + I_i,N(2\pi \mathbf{j}/\mathbf{n}) f_i(2\pi \mathbf{j}/\mathbf{n})^{-1} \right\}, \quad (8)
$$

the sum is considered at the Fourier frequencies $(2\pi \mathbf{j}/\mathbf{n})$, for $\mathbf{f} \in J_N$, with

$$
J_N = \left\{ [-n_1/2], \ldots, n_1 - n_1/2 \right\} \times \left\{ [-n_2/2], \ldots, n_2 - n_2/2 \right\} \quad (9)
$$

and $\lfloor u \rfloor$ denotes the largest integer less or equal than $u$. The sums are often substituted by integrals, but
we prefer to sum over Fourier frequencies and use the fast Fourier transform (FFT) to efficiently calculate
(8). The function $I_i,N(\omega)$ in (8) is the periodogram at a frequency $\omega$, defined as

$$
I_i,N(\omega) = (2\pi)^{-2} (n_1 n_2)^{-1} \sum_{s_1=1}^{n_1} \sum_{s_2=1}^{n_2} Z_i(s) \exp\{-is^T \omega\}^2, \quad (10)
$$

where $s = (s_1, s_2)$. The function $f_i$ is the spectral density of the process $Z_i$, which is the Fourier transform
of the covariance function of $Z_i$. 

9
In our setting, to use Whittle’s approximation to the likelihood we need to assume that the weight function $K(s - s_i)$ has compact support, i.e. it is only positive in a subregion $S_i$ centered at $s_i$, so data from different subregions can be used to identify which stationary process $Z_i$ is being used, and to calculate then the corresponding periodogram.

In this section we ignored the effect of estimating the parameters $(\beta, \theta)$ on the subsequent predictions. In the next section, we present a Bayesian approach to take into account in the prediction the uncertainty in the trend and covariance parameters.

4 Bayesian prediction

The main objective of the previous section was to develop the tools that provide a satisfactory technique for interpolation of nonstationary fields when the covariance structure is at least partially unknown. The covariance structure is estimated first, and then the estimated covariance is used for interpolation. The properties of the interpolants based on an estimated covariance structure are not well understood, and it is common practice to ignore the effect of the uncertainty in the covariance structure on subsequent predictions.

A Bayesian approach to interpolation of spatial processes will provide a general methodology for taking into account the uncertainty about parameters on subsequent predictions.

Suppose that an appropriate model for $Z(x)$ is

$$Z(x) = \mu(x; \beta) + \epsilon(x)$$

with

$$\mu(x; \beta) = \sum_{i=1}^{q} f_i(x) \beta_i$$

where $\beta_i, i = 1, \ldots, q$ are unknown parameters, $\beta = \{\beta_1, \ldots, \beta_q\}$, $f(x) = \{f_1(x), \ldots, f_q(x)\}$, and $\epsilon$ is a zero-mean spatially correlated process, with a nonstationary covariance. In geostatistics, the dimension space $q$ is generally considered fixed. However, we could treat $q$ as a hyperparameter and use a reversible jump Markov Chain Monte Carlo (MCMC) (Green, 1995) for model selection. Denison, Mallick and Smith (1998) showed
a similar approach using splines and wavelets as basis functions for univariate non parametric regression, this algorithm could be extended to a multivariate case and applied to the spatial model presented in this paper.

The kriging point predictor $\hat{Z}(x_0)$ is the best linear unbiased predictor

$$
\hat{Z}(x_0) = \beta_\theta C^{-1}_\theta Z + b_\theta \hat{\beta}
$$

(11)

where $\hat{\beta}$ is the generalized least squares estimate of $\beta$, $C_\theta = \{\text{cov} (Z(x_i), Z(x_j))\}_{n\times n}$ is the covariance matrix defined in (3) with parameter $\theta$ that varies over $D$, $F = \{f_i(x_i)\}_{n\times q}$, $c_\theta = \{C_\theta(x_0, x_i)\}_{n\times 1}$, and $b_\theta = f(x_0) - F' C^{-1}_\theta c_\theta$. The kriging prediction error variance is

$$
\hat{V}_\theta = C_\theta(x_0, x_0) - \beta_\theta C^{-1}_\theta c_\theta + b_\theta (F' C^{-1}_\theta F)^{-1} b_\theta.
$$

(12)

If $\theta$ is known and the process is Gaussian, then the predictive density may be taken a normal with mean given by (11) and variance (12). Although traditional kriging is based on stationary processes, the above equations do not depend on stationarity and are general so long as the covariance components are known.

We might take the (improper) prior density of $\beta$ to be constant and model the covariance parameter $\theta$ as a spatial function $\theta(s)$ for $s \in D$ with vague spatial prior $\pi(\theta)$, then we get

$$
\pi(\beta, \theta) \propto \pi(\theta).
$$

This leads to a marginal posterior distribution of $\theta$, given $Z$, of the form

$$
p(\theta|Z) \propto \pi(\theta) C_\theta^{-1/2} F' C^{-1}_\theta F^{-1/2} \exp(-S/2)
$$

(13)

where $S = (Z - F' \hat{\beta}) C^{-1}_\theta (Z - F' \hat{\beta})$.

For spatial prediction, the quantity of interest is the predictive distribution for $Z(x_0)$

$$
p(Z(x_0)|Z) \propto \int p(Z(x_0)|\theta, Z) p(\theta|Z) d\theta
$$

We present now a method to efficiently obtain the posterior distribution of $\beta$ and $\theta$, and the predictive distribution to sample $Z(s_0)$. We define $\hat{Z}$ as in (6), then $\hat{Z}$ is a Monte Carlo integration for $Z$. The likelihood for the process $Z$ is well defined

$$
f(Z|\beta, \theta) = N(\mu_\beta, C_\theta)
$$
where \( \mu_\beta = [\mu(s_1; \beta), \ldots, \mu(s_N; \beta)]^T \), and \( C_\beta \) is the covariance matrix of \( Z \). Hence, given a prior on \( \beta \) and \( \theta \) a Bayesian model for the posteriors would be completely specified. We replace \( f(Z|\beta, \theta) \) with \( \tilde{f}(Z|\beta, \theta) = N(\mu_\beta, \tilde{C}_\beta) \) which is \( f(\tilde{Z}|\beta, \theta) \), the likelihood for the process \( \tilde{Z} \).

The conditional distribution of \( Z(s_0) \) given \( Z \) is defined as follows

\[
    f(Z(s_0)|Z, \beta, \theta) = N(z_0, V_\theta)
\]

where \( z_0 \) is the conditional mean (kriging predictor) and \( V_\theta \) is the covariance matrix of \( Z(s_0) \) given \( Z \). Therefore, given a prior on \( \beta \) and \( \theta \) the Bayesian model for spatial prediction would be completely specified. We simply replace (14) with

\[
    \tilde{f}(Z(x_0)|Z, \beta, \theta) = N(\tilde{z}_0, \tilde{V}_\theta).
\]

Thus, the solution \( \tilde{f}(Z(s_0)|Z) \) is actually \( f(\tilde{Z}(s_0)|\tilde{Z}) \). In the Bayesian approach if we use (5) as an approximation of (3), the Bayesian predictive distribution we obtain to sample \( Z(x_0) \) is actually \( \tilde{f}(Z(x_0)|Z) \). Note that \( \tilde{f}(Z(x_0)|Z) = f(\tilde{Z}(x_0)|\tilde{Z}) \). In practice, we will work with \( \tilde{f} \) converting to \( f \), to sample \( Z(s_0) \) rather than sampling \( Z(s_0) \) through the \( \tilde{Z} \). Hence, we need \( \tilde{Z}(s) \xrightarrow{D} Z(s) \). If the covariance for each local stationary process \( Z_{\theta(s)} \) is a continuous function on \( s \in D \), then the distribution of \( \tilde{Z} \) converges to the distribution of the process \( Z \).

5 Hierarchical Bayesian model

The parameter function \( \theta(s) \) for \( s \in D \), measures the lack of stationarity of the process \( Z \). If would be natural to treat \( \theta(s) \) as a stochastic process, with correlated errors. Thus, we could consider a hierarchical Bayesian approach to model and take into the account the spatial structure of the parameter \( \theta(s) \) in the prediction of \( Z \). The essence of the hierarchy that we suggest is the specification of \( Z \) as a parameterized process model:

\[
    [\text{process, parameters}] = [\text{process} \mid \text{parameters}] \ [\text{parameters}] .
\]

Bayes’ Theorem provides \( [\text{process, parameters} \mid \text{data}] \) by combining with an observational data model \( [\text{data} \mid \text{process, parameters}] \).

Stage 1:
The process $Z$ is as a convolution of local stationary processes:

$$Z(x) = \int_D K(x - s)Z_{\theta(s)}(x)ds.$$  

(15)

where $K$ is a kernel function and $Z_{\theta}(x)$, $\mathbf{x} \in D$ is a family of (independent) stationary Gaussian processes indexed by $\theta$. Thus, the distribution of $Z$ given $\theta$ is Gaussian:

$$[Z|\theta]$$

is Gaussian

Stage 2:

We model the parameter function $\theta$ as a spatial process with correlated errors,

$$\theta(s) = \mu_\theta(s) + \epsilon_\theta(s)$$  

(16)

$\mu_\theta(s)$ represents the spatial trend (large scale structure), and the process $\epsilon_\theta(s)$ represents some spatially correlated zero-mean noise. We assume here $\mu_\theta(s)$ is a polynomial function in $s$ with coefficients, $\beta_0$, and the process $\epsilon_\theta(s)$ is Gaussian with mean zero and a Matérn stationary covariance, $\text{cov}(\epsilon_\theta(x + y), \epsilon_\theta(x)) = C_{\tau_0}(y)$, with parameters $\tau_0$. A Bayesian Information Criterion (BIC) in terms of the likelihood of $Z$, could be implemented as part of this hierarchical model to choose the value of $k$ (dimensionality of $\beta_0$).

Thus, we have in stage 2:

$$[\theta|\beta_0, \tau_0]$$

(17)

The parameter function $\theta$ could be also modeled using a Bayesian version of the median polish approach (e.g. Cressie, 1993). Assume the process $Z$ is observed in a grid $n_1 \times n_2$, then we propose the following model for the parameter function $\theta$

$$\theta(s) = a + r_i + c_j + \epsilon_\theta(s)$$

(18)

where $s = (s_i, s_j)$, for $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. The hyperparameters $a, r_i$ for $i = 1, \ldots, n_1$, and $c_j$ for $j = 1, \ldots, n_2$, are unknown and correspond to the $\beta_0$ parameter function in (17).

The kernel function $K$ in the representation of $Z$ is a kernel smoother with bandwidth $h$. The function $K$ is not representing the lack of stationarity. The crucial parameter here to explain the lack of stationarity of $Z$ is $\theta$. However, the bandwidth parameter $h$ instead of considering it fixed, could be also incorporated
in this hierarchy. Thus, in stage 1, we could have \( Z(\theta, h) \), and the Bayesian model would be completely specified once we give a prior distribution to \( h \) and \( \theta \).

In practice, we might not observe the true underlying process \( Z \), instead we observe a process \( Z_O \), which is \( Z \) plus some random noise, measurement error,

\[
Z_O(s) = Z(s) + \epsilon(s),
\]

where \( \epsilon(s) \) is a Gaussian process with zero-mean and uncorrelated errors with variance \( \sigma^2 \), and the process \( Z \) is the convolution of stationary processes as in (15). The parameter \( \sigma^2 \) is a hyperparameter with a vague prior. Thus, in this case the three stages of the hierarchy would be:

**Stage 1:**

\[
[Z_O|\theta, h, \sigma^2] \text{ is Gaussian}
\]

**Stage 2:**

\[
[\theta|\beta_0, \tau_0] \text{ is Gaussian}
\]

**Stage 3:** We specify the priors for the hyperparameters \( \beta_0, \tau_0 \), where \( \beta_0 \), could be modeled using the Bayesian median polish model (18).

If the goal is to predict \( Z_O \) at a location \( x_0 \), the Bayesian solution is the predictive distribution of \( Z_O(x_0) \) given the observations \( Z = (Z_O(s_1), \ldots, Z_O(s_N)) \),

\[
p(Z_O(x_0)|Z) \propto \int p(Z_O(x_0)|Z, \theta(s)) p(\theta(s)) ds \int dh \sigma^2 d\theta,
\]

(19)
a MCMC approach could be used to simulate \( m \) values from the posterior of the parameters \( \theta, h, \) and \( \sigma^2 \), and the predictive distribution could be approximated by the Rao-Blackwellized estimator:

\[
p(Z_O(x_0)|Z) \approx \frac{1}{m} \sum_{i=1}^{m} p(Z_O(x_0)|Z, \theta(s)^{(i)}) \quad \text{for } s \in D, h^{(i)}, \sigma^2^{(i)}
\]

where \( \theta(s)^{(i)} \) for \( s \in D, h^{(i)}, \sigma^2^{(i)} \) constitute the \( i \)-th draw from the posterior distribution. Notice that to sample from the posterior of \( \theta \) we need to specify first the priors of the hyperparamters \( \beta_0 \) and \( \tau_0 \):

\[
p(\theta(s)|Z) \propto \int p(Z|\theta(s)) p(\theta(s)|D)p(\theta(s)|\beta_0, \tau_0)d\tau_0 \, d\beta_0.
\]
Thus, once we determine the priors for the parameters $h$ and $\sigma$, and for the hyperparameters $\beta_0$ and $\tau_0$, the Bayesian model is completely specified. The joint posterior for $\theta, h$, and $\sigma^2$ is defined as follows:

$$p(\theta(s) \text{ for } s \in D, h, \sigma^2 | Z) \propto p(Z | \theta(s) \text{ for } s \in D, h, \sigma^2) p(h) p(\sigma^2) \int p(\theta(s) \text{ for } s \in D | \beta_0, \tau_0) p(\beta_0) p(\tau_0) d\tau_0 d\beta_0.$$ 

where $p(h), p(\sigma^2), p(\beta_0)$, and $p(\tau_0)$ are the prior distributions for $h$, $\sigma$, $\beta_0$ and $\tau_0$.

6 Change of Support

The change-of-support problem occurs when the supports of predictand and data are not same. In this case we are interested in making predictions about the random process at a point location $Z(x_0)$ from data on block averages, $Z(B_1), \ldots, Z(B_N)$.

Here, we observe $Z$, the output of Models-3, averaged over regions, $B_1, \ldots, B_N$ of dimensions 36km x 36km and we want to predict $Z$ at a location of interest $x_0$, e.g. at the location where we have a CASTNet site.

The covariance for the block averages is defined as follows

$$\text{cov}(Z(B_i), Z(B_j)) = \int_{B_i} \int_{B_j} C(u, v) \, du \, dv / |B_i||B_j|,$$  

where

$$C(u, v) = \text{cov}(Z(u), Z(v))$$

$C$ is a nonstationary spatial function. In practice, for each pixel $B_i$ we draw an independent set of locations $u_{ir}, r = 1, 2, \ldots, L_i$ uniformly over $B_i$, and we approximate the integral in (20) with the following expression (Gelfand et al, 2000)

$$\hat{\text{cov}}(Z(B_i), Z(B_j)) = L_i^{-1} L_j^{-1} \sum_r \sum_{r'} C(u_{ir}, u_{jr'}).$$  

(21)

Therefore, we approximate the point prediction $\hat{E}(Z(x_0)|Z)$, where $Z = (Z(B_1), \ldots, Z(B_N))$ with $\hat{E}(Z(x_0)|Z)$, which is obtained using (21) as an approximation of (20).

It is useful to note that, if we define $\tilde{Z}(B_i) = L_i^{-1} \sum_r Z(u_{ir})$ for $i = 1, \ldots, N$, then the solution $\tilde{E}(Z(s_0)|Z)$ is actually $E(Z(s_0)|Z)$. In the Bayesian approach if we use (21) as an approximation of (20), the Bayesian predictive distribution we obtain to sample $Z(x_0)$ is actually $\tilde{f}(Z(x_0)|Z)$. Note that $\tilde{f}(Z(x_0)|Z) = f(Z(x_0)|\tilde{Z})$. 

15
Hence, we need $\hat{Z}(B_i) \xrightarrow{p} Z(B_i)$. If $\text{cov}(Z(u),Z(v))$ is a continuous function on $u$ and $v$, then $Z$ is mean square continuous (see e.g. Stein, 1999b), which implies $\hat{Z}(B_i) \xrightarrow{p} Z(B_i)$. By the definition of the nonstationary covariance $\text{cov}(Z(u),Z(v))$ in (3), if the local stationary covariances are Matérn, then $\text{cov}(Z(u),Z(v))$ is continuous on $u$ and $v$.

7 Application

Our first goal is to understand and quantify the spatial structure of air pollutants using the output of the regional scale air quality models (Models-3). Models-3 estimates hourly concentrations and fluxes of different pollutants. As an example we examine sulfur dioxide. The spatial domain, $D$, is a regular grid $(81 \times 87)$, the dimensions of each pixel on the grid are $36 \text{km} \times 36 \text{km}$. Models-3 provides the estimated average concentration for each pixel. Figure 1 shows the weekly average concentrations of $SO_2$ for the week starting July 11, 1995. For the purpose of illustrating the need of the technique presented in this paper, Figures 4 and 5 show the posterior distributions of covariance parameters at the selected sites plotted in Figure 3. We used vague gamma priors for all the covariance parameters, except for the sill parameter that we used

$$p(\sigma) \propto \sigma^{-1},$$

this is a uniform prior for $\log(\sigma)$. The sill parameter is changing with location as illustrated by the variation in the distributions in Figure 5. Thus, this indicates lack of stationarity. We can clearly appreciate this fact in Figure 6. Figure 6 shows a map of the modes for the posterior distributions of the sill parameter of the Matérn covariance for the Models-3 $SO_2$ concentrations, for the week starting July 11, 1995. The value of $M$ (number of sampling locations) in this application is 81. We used a BIC criterion combined with a systematic sampling approach to calculate $M$. Figure 6 indicates clearly deviation from stationarity. Figure 7 shows the sample variance at each pixel, obtained using the hourly observations of Models-3 $SO_2$ concentrations, for the week starting July 11, 1995. The different spatial patterns observed in Figures 6 and 7, indicate that if we standardize the response (i.e. we divide the response by the sample standard deviation at each location) it would not be enough to get stationarity. Because, the sill parameter (Figure 6) is not only representing the local variance (Figure 7), but is also capturing the local spatial correlation structure.
We implemented here the nonstationary model (6) with a kernel function with compact support, the kernel we used was the quadratic weight-function $K_4$ presented in Section 3.1. We applied the algorithm discussed in Section 3.1 to choose the bandwidth, trying to preserve the general “shape” of the data.

The evaluation of the regional air quality models is crucial. We use the ground truth from the Clean Air Status and Trends Network (CASTNet) measurements (see Figure 2), to evaluate weakly average concentrations of $SO_2$ estimated by Models-3. The coordinates for Models-3 do not match the location of the CASTNet sites. Thus, we spatially interpolate the output of the models at the location of the CASTNet sites. We should do the interpolation by taking into consideration the spatial structure of the pollutant concentrations. Furthermore, since Models-3 concentrations are block averages over the $36km \times 36km$. We use here a Bayesian approach to the interpolation technique for nonstationary fields taking also into account the change-of-support problem (we calculated the block averages covariances drawing a set of 4 locations in each pixel). We use posterior predictive checks (PPC) as suggested by Rubin (1984) for validation of the air quality numerical models. Thus, we compare the posterior predictive distributions of the physical models at different locations to the observed data, and we judge if the numerical models generate data that are similar to the CASTNet data (Figure 3). Figure 8 shows the predictive values of the Models-3 $SO_2$ weekly average concentrations at the CASTNet selected sites. As expected, we get very high variability at the Indiana site, this site is very close to several coal power plants, and therefore the $SO_2$ levels can be very high or very low depending on wind speed, wind direction, and on the atmospheric stability. The sites in Maine and Florida have the lowest $SO_2$ levels and variability. The agricultural site in Illinois and the site in North Carolina have similar behavior regarding $SO_2$ levels. The site in NC is not far from the Tennessee power plants, and the site in Illinois is also relatively close to some Midwestern power plants. The site in Michigan, which is very close to the lake Michigan and relatively far from power plants has also low $SO_2$ levels.

The graph on the left in Figure 9 presents a naive approach for evaluation of Models-3. This graph shows Models-3 versus CASTNet, without doing any spatial interpolation of Models-3. In this graph we simply have the values of Models-3 for the pixels that are the closest to each CASTNet site, without considering the change of support. On the other hand, the graph on the right in Figure 9 shows the modes of the predictive Bayesian distributions for Models-3 at the CASTNet locations versus the ground truth obtained.
from CASTNet. Thus, in this graph we are taking into account the change of support and the uncertainty in the covariance function. The uncertainty in the estimated Models-3 values in this figure depends on location and is represented by the variance of the distributions in Figure 8. This Bayesian approach gives more reliable prediction errors, by taking into account the uncertainty in the covariance parameters, and the PPC are a powerful tool for validation of physical models. This figure is just an illustration of the powerful application of the technique presented in this paper, though it does arises some important questions that are currently being discussed with the Models-3 group.

8 Conclusions and final remarks

We introduce in this paper a new statistical methodology for nonstationary models, the spatial field is represented locally as a stationary isotropic random field, but the parameters of the stationary random field are allowed to vary continuously across space. Kernel functions are used to ensure that the field is well-defined but also continuous. New fitting algorithms are developed. The methods are extended to prediction/interpolation questions using Bayesian approaches to account for parameter uncertainty. In this paper we also take into account the change of support problem that occurs when data and predictant have different spatial resolution.

The spatial model proposed in this paper could be also applied to spatial temporal data. Consider $Z$ a spatio-temporal process,

$$Z(x, t) = \mu(x, t; \beta) + \epsilon(x, t)$$

where

$$E(Z(x, t)) = \mu(x, t; \beta)$$

and the process $\epsilon$ represents some spatial-temporally correlated zero-mean noise. We could represent $Z(x)$ using the model in Section 3 where $x$ now varies over a domain $D$ contained in a 3-dimensional Euclidean space $\mathbb{R}^3$. However, if we assume that the local stationary processes have a separable covariance, then we model $Z$ as follows:

$$Z(x, t) = \int_D K(x - s)Z_{\Phi|s} (x, t) ds.$$  \hspace{1cm} (22)
where $\phi(s) = (\theta(s), \rho(s))$ and the local processes $Z_{\phi(s)}$ have a separable spatial-temporal covariance,

$$\text{cov} \left( Z_{\phi(s)}(s_1, t_1), Z_{\phi(s)}(s_2, t_2) \right) = C_{\theta(s)}^{(1)}(s_1 - s_2) \cdot C_{\rho(s)}^{(2)}(|t_1 - t_2|)$$  \hspace{1cm} (23)

where $C_{\theta(s)}^{(1)}$ is a stationary spatial covariance with parameter $\theta(s)$, for instance a Matérn (4), and $C_{\rho(s)}^{(2)}$ is a stationary temporal covariance with parameter $\rho(s)$. For instance we could model $C_{\rho(s)}^{(2)}$ as follows:

$$C_{\rho(s)}^{(2)}(|t_1 - t_2|) = \frac{\sigma_s^2 \phi_s^{|t_1 - t_2|}}{1 - \phi_s^2}$$

where $\rho(s) = (\sigma_s^2, \phi_s)$. This is the covariance of an autoregressive AR(1) temporal model. Forms such as (23) have a history in spatial-temporal modeling; see e.g. Mardia and Goodall (1993) and references therein.

Then, the covariance of $Z$ can be written as follows:

$$\text{cov} \left( Z(s_1, t_1), Z(s_2, t_2) \right) = \int_D K(s_1 - s) K(s_2 - s) C_{\theta(s)}^{(1)}(s_1 - s_2) \cdot C_{\rho(s)}^{(2)}(|t_1 - t_2|) \, ds$$ \hspace{1cm} (24)

this is a nonstationary spatial-temporal covariance. Note that the covariances for the processes $Z_{\phi(s)}$ are separable but the covariance of $Z$ is not separable. For each local process $Z_{\phi(s)}$ spatial association at a fixed time point is captured through $C^{(1)}$; decay in such association over time is captured by $C^{(2)}$.

9 Acknowledgments

The authors would like to acknowledge the helpful insight and assistance provided by the US EPA Office of Research Development, RTP, NC, in particular by Dr. Peter Finkelstein and Dr. Robin Dennis.

References


Figure 1: Output of Models-3, weekly average of $SO_2$ concentrations (ppb), for the week of July 11, 1995. The resolution is 36 km$^2$. 
Figure 2: Weekly average of $SO_2$ concentrations (ppb) at the Clean Air Status and Trends Network (CASTNet) sites, for the week of July 11, 1995.
SO2 concentrations (CASTNet)

Figure 3: Weekly average of $SO_2$ concentrations (ppb) at some selected CASTNet sites, for the week of July 11, 1995.
Air pollutant: SO2

Figure 4: Posterior distributions for the range parameter (km) of the Matérn covariance for Model-3 SO₂ concentrations, for the week starting July 11, 1995. At the 6 selected locations showed in Figure 3.
Figure 5: Posterior distributions for the sill parameter of the Matérn covariance for Models-3 $SO_2$ concentrations, for the week starting July 11, 1995. At the 6 selected locations showed in Figure 3.
Figure 6: Map of the modes of the posterior distributions for the sill parameter of the Matérn covariance for Models-3 $SO_2$ concentrations, for the week starting July 11, 1995.
Figure 7: Sample variance for each pixel, obtained using hourly observations of Models-3 \( SO_2 \) concentrations, for the week starting July 11, 1995.
Figure 8: Predictive distributions for the Models-3 SO$_2$ concentrations, at the 6 selected locations showed in Figure 3, for the week starting July 11, 1995.
Figure 9: The graph on the left shows CASTNet measurements for the week starting July 11, 1995, versus the values of Models-3 for the pixels that are the closest to each CASTNet site, without considering the change of support. The graph on the right shows the CASTNet measurements versus the modes of the predictive Bayesian distributions for Models-3 at the CASTNet locations.