Available online at www.sciencedirect.com
ScienceDirect
Journal of the Korean Statistical Society 37 (2008) 3–10
JOURNAL OF THE KOREAN STATISTICAL SOCIETY
www.elsevier.com/locate/jkss

A modeling approach for large spatial datasets

Michael L. Stein

Department of Statistics, University of Chicago, Chicago, IL 60637, United States

Received 1 July 2007; accepted 1 September 2007
Available online 30 January 2008

Abstract

For Gaussian spatial processes observed at n irregularly sited locations, exact computation of the likelihood generally requires $O(n^3)$ operations and $O(n^2)$ memory. If we can write the covariance function of the process as a nugget effect plus a term of moderate rank, then both the number of computations and memory requirements can be greatly reduced. However, while such models can capture larger-scale structure of spatial processes, they have trouble describing the local behavior of spatial processes accurately. If the nugget effect is replaced by a covariance function with compact support, one can get a much better model for the local behavior of the process and still carry out exact likelihood calculations on quite large datasets. This approach is applied to compute the maximum likelihood estimate for 13,216 observations from a single orbit of TOMS (Total Ozone Mapping Spectrometer) measurements of total column ozone. It is also applied to obtain likelihood-based estimates using over one million observations from 83 orbits. Replacing the nugget effect by a compactly supported covariance function leads to huge increases in likelihood, but there is still clear evidence of model misfit.

1. Introduction

A major obstacle to using likelihood-based and Bayesian methods in spatial statistics is the computational burden of calculating the likelihood function when the number of observations, n, is large. Even for Gaussian processes, for which the likelihood has a convenient closed form expression, calculating the likelihood exactly requires $O(n^2)$ memory and $O(n^3)$ floating point operations, so that while sample sizes of $n = 1000$ are no longer a challenge, $n = 100{,}000$ remains out of reach for even large clusters of processors. Specifically, if $\mathbf{W}$ is the covariance matrix of the observations, one needs to be able to calculate a quadratic form in $\mathbf{W}^{-1}$ and the logarithm of the determinant of $\mathbf{W}$, both of which can be readily obtained from the Cholesky decomposition of $\mathbf{W}$. If the observations fall on a grid or have some other regular pattern, then it is sometimes possible to reduce the computations. However, many large spatial datasets do not have observations in a pattern that can be exploited in exact computations. In this case, one might resort to various approximations to the likelihood; see, for example, Caragea (2003), Fuentes (2007) and Stein, Chi, and Welty (2004). Another possibility is to use models that allow for efficient matrix calculations even for irregularly sited observations. For example, if the covariance function has compact support and its range is sufficiently small, then the sparseness of the resulting covariance matrix can be used to compute the Cholesky decomposition more efficiently, as noted by Furrer, Genton, and Nychka (2006). Indeed, Davis (2006) describes a set of freely available programs

E-mail address: stein@galton.uchicago.edu.

1226-3192/$ - see front matter
doi:10.1016/j.jkss.2007.09.001

for carrying out these and related computations. We might hope these compactly supported covariance functions are capable of accurately describing the local behavior of spatial processes, but, by construction, they cannot describe dependence over larger scales.

As noted by Cressie and Johanneson (2008), faster matrix computations are also possible when the covariance matrix has the form of a diagonal matrix plus a low rank matrix. Cressie and Johanneson (2008) showed how one can calculate kriging predictors and likelihood functions exactly with massive spatial datasets with this type of model. They then applied these results to calculate exactly kriging predictors under an estimated model using all of the TOMS data on a given day (around 170,000 observations), but chose not to use likelihood-based parameter estimates due to difficulties in maximizing likelihoods with large numbers of parameters. Stein (2007) also used covariance matrices that were a sum of a diagonal and a low rank matrix for TOMS data, but with a quite different form for the low rank part of the model. This work estimated parameters using both weighted least squares and maximum likelihood, where, following Cressie and Johanneson (2008), the specific structure of the covariance matrices was exploited to speed the likelihood calculations. While the model captures larger-scale behavior of the TOMS data reasonably well, it does a poor job of describing local dependence. The present work considers models whose covariance matrices are a sum of a sparse (but not diagonal) matrix plus a low rank matrix in an attempt to capture both small and large-scale spatial variation but still allow exact computation of the likelihood for large spatial datasets.

Section 2 gives more details on the general statistical model and computational methods. Section 3 applies this approach to a dataset consisting of over one million observations of total column ozone as measured by TOMS over 83 orbits during May 1–6, 1990. The approach is first used to maximize the exact likelihood for a model including both local and large-scale terms based on the 13,216 from the first of these 83 orbits. This approach is extended to multiple orbits by ignoring the dependence across orbits and is used to obtain likelihood-based parameter estimates based on all 83 orbits available during this six-day period. Section 4 describes some diagnostics illustrating problems with the fitted models, which suggest that the modeling approach used here is not completely adequate. Section 5 briefly describes some possible extensions and other applications.

2. Model

Let us suppose that the covariance matrix of a Gaussian random vector can be written in the form

$$\mathbf{W}(\theta) = \theta_0 \mathbf{A} + \mathbf{P}\mathbf{K}(\theta_1)\mathbf{P}', \tag{1}$$

where $\mathbf{A}$ is a known $n \times n$ positive definite matrix, $\mathbf{P}$ a known $n \times p$ matrix, $\theta_0 \geq 0$ an unknown scalar, $\mathbf{K}(\theta_1)$ a $p \times p$ positive semidefinite matrix depending on the unknown (possibly vector-valued) parameter θ_1 and $\theta = (\theta_0, \theta_1)$. Such a model can naturally arise by considering a spatial process Z on some domain D that is the sum of independent Gaussian processes Z_1 and Z_2 with $\mathrm{cov}\{Z_1(x), Z_1(y)\} = \theta_0 K_0(x, y)$ for $x, y \in D$, $K_0(x, y) = 0$ for x and y sufficiently distant and

$$Z_2(x) = \sum_{j=1}^{p} U_j P_j(x), \tag{2}$$

where the P_j's are known functions and $(U_1, \ldots, U_p)'$ is multivariate normal. The form (1) is obtained if Z is observed at $x_1, \ldots, x_n$, $\mathbf{A}$ is the matrix with jkth element $K_0(x_j, x_k)$, $\mathbf{P}$ is the $n \times p$ matrix with jkth element $P_k(x_j)$ and $\mathbf{K}(\theta_1)$ is the covariance matrix for $(U_1, \ldots, U_p)'$.

If $\mathbf{A}$ is sufficiently sparse so that its Cholesky decomposition $\mathbf{C}'\mathbf{C}$ can be computed, then if p is not too large, the Cholesky decomposition of $\mathbf{W}(\theta)$ for any particular θ can be obtained with modest additional effort. Since $\mathbf{C}$ does not need to be recomputed as θ varies, this model can be used in practice even when computing this decomposition is fairly time-consuming. Indeed, for models of the form (1), all computations directly involving $\mathbf{C}$ can be done once and for all, so that the computations that need to be redone when θ varies are of order p^3, independent of n. Thus, this model is very convenient for likelihood or Bayesian calculations, for which one needs to consider many values for θ.

Let us spell out how to do the computations so that all of the order n computations can be done only once for all θ. Suppose, for simplicity, the mean of our Gaussian random observation vector $\mathbf{y}$ is $\mathbf{0}$. We can compute $(\mathbf{C}')^{-1}\mathbf{P}$ and $(\mathbf{C}')^{-1}\mathbf{y}$ once and for all, and from these, obtain the scalar $q = \mathbf{y}'\mathbf{A}^{-1}\mathbf{y}$, the p-vector $\mathbf{u} = \mathbf{P}'\mathbf{A}^{-1}\mathbf{y}$ and the $p \times p$ matrix $\mathbf{R} = \mathbf{P}'\mathbf{A}^{-1}\mathbf{P}$.

Now we are prepared to carry out the computations needed for a specific θ. Let $\mathbf{Q}(\theta_1)'\mathbf{Q}(\theta_1)$ be the Cholesky decomposition of $\mathbf{K}(\theta_1)$, which, since $\mathbf{K}(\theta_1)$ may be nearly or exactly singular, may require pivoting to compute. Next, let $\mathbf{J}(\theta)'\mathbf{J}(\theta)$ be the Cholesky decomposition of $\mathbf{I} + \frac{1}{\theta_0}\mathbf{Q}(\theta_1)\mathbf{R}\mathbf{Q}(\theta_1)'$, where $\mathbf{I}$ is the identity matrix. Then it is straightforward to show that

$$\det\{\mathbf{W}(\theta)\} = \theta_0^n \det(\mathbf{A})[\det\{\mathbf{J}(\theta)\}]^2$$

and, writing $\|\cdot\|$ for Euclidean norm,

$$\mathbf{y}'\mathbf{W}(\theta)^{-1}\mathbf{y} = \frac{q}{\theta_0} - \frac{1}{\theta_0^2}\|\{\mathbf{J}(\theta)'\}^{-1}\{\mathbf{Q}(\theta_1)\mathbf{u}\}\|^2,$$

from which one immediately obtains the likelihood at θ, given by $-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det\{\mathbf{W}(\theta)\} - \frac{1}{2}\mathbf{y}'\mathbf{W}(\theta)^{-1}\mathbf{y}$.

In practice, the sparse matrix part of the covariance may be of the more general form $\mathbf{A}(\theta_0)$, with θ_0 potentially a vector. In this case, it would make sense to decompose θ_0 into components ϕ_0 (a scalar) and α_0 (possibly a vector) so that $\mathbf{A}(\theta_0)$ can be written in the form $\phi_0\mathbf{A}(\alpha_0)$ and then organize any computations so that α_0 is changed as infrequently as possible.

3. Application

This section applies the method just described to observations of total column ozone as measured by TOMS on May 1–6, 1990. This TOMS instrument was on a sun-synchronous polar-orbiting satellite making around 13.8 orbits each day. Krueger, Bhartia, McPeters, Herman, Wellemeyer, and Jaross (1998) provide detailed documentation on the TOMS data and Stein (2007) describes the preprocessing used to obtain the actual dataset studied here. Each orbit spans all latitudes (except near the South Pole, since TOMS uses reflected sunlight) and a longitude band that is about 30° wide near the equator and wider as one heads poleward. As in Stein (2007), the data are analyzed on a logarithmic scale and a regression approach is used to remove persistent spatial variation over the period May 1–6, so the analyses here are based on the residuals from this regression. We only model the spatial variation here and consider analyzing all 13,216 observations from the first orbit on May 1 using exact likelihoods and then analyzing the more than one million observations from the six-day period using the likelihood obtained by ignoring dependence across orbits.

Stein (2007) shows that the ozone residuals are clearly not isotropic on the sphere but argues that, to a decent approximation, possess what Jones (1963) calls axial symmetry. Specifically, consider a random field Z on the sphere with coordinates designated by latitude L and longitude ℓ. Then Z is (weakly) axially symmetric if its mean depends only on latitude and there exists a function K on $[-\frac{1}{2}\pi, \frac{1}{2}\pi]^2 \times (-\pi, \pi]$ such that, for all (L, ℓ) and (L', ℓ'),

$$\operatorname{cov}\{Z(L, \ell), Z(L', \ell')\} = K(L, L', \ell - \ell'). \tag{3}$$

We will call the function K an axially symmetric covariance function.

Jones (1963) shows how all continuous axially symmetric covariance functions can be written in terms of series expansions. Let P_n^m be the Legendre polynomial of degree n and order m and $\overline{P}_n^m$ its normalized version (normalized so its squared integral on $[-1, 1]$ is 1). Consider functions of the form

$$K(L, L', \ell) = \sum_{m=-N}^{N} \sum_{j,k=|m|}^{N} \mathrm{e}^{im\ell}\overline{P}_j^m(\sin L)\overline{P}_k^m(\sin L')g_m(j, k) \tag{4}$$

with the complex $g_m(j, k)$'s satisfying $g_{-m}(j, k) = g_m(j, k)^*$, where * indicates complex conjugate, which ensures that the resulting process is real. Note that this constraint forces $g_0(j, k)$ to be real. Define $\mathbf{G}_m(N)$ to be the $(N-m+1)\times(N-m+1)$ matrix whose jkth entry is $g_m(m+j-1, m+k-1)$. Then K in (4) is an axially symmetric covariance function and, furthermore, every continuous axially symmetric covariance function is a pointwise limit of functions of this form. The number of columns of the matrix $\mathbf{P}$ in (1), which we denoted by p, equals $(N+1)^2$. As noted in Stein (2007), this model yields $\frac{1}{3}N^3 + N^2 + \frac{5}{3}N + 1$ real parameters and this cubic growth in N makes it difficult to take N even moderately large if one needs to, for example, maximize a likelihood over these parameters. As in Stein (2007), we will parameterize $\mathbf{G}_m(N)$ using its (complex unless $m = 0$) Cholesky decomposition as a way to guarantee that $\mathbf{G}_m(N)$ is positive semidefinite.

Stein (2007) fitted this model with $N = 7$ together with a nugget effect (a white noise term) to describe the within orbit variation of the TOMS data on May 1–6, 1990. A weighted least squares (wls) fit was able to capture many of the large-scale features in the spatial dependence, providing a decent approximation to the rather dramatic changes with latitude in these dependencies. However, the model does a poor job of describing the small-scale (i.e., up to a few degrees angular distance) dependencies. Plots in Stein (2007) suggest that dependencies on these scales do not change all that much with latitude and, furthermore, exhibit approximate isotropy. Thus, one might hope to gain a large improvement in fit by adding a compactly supported isotropic covariance function to the model (4). Here, we will use what is called the pentaspherical model, an isotropic model in which the covariance between observations a distance d apart is

$$K(d) = \theta_0 \left\{ 8 - 15\frac{d}{\alpha} + 10\left(\frac{d}{\alpha}\right)^3 - 3\left(\frac{d}{\alpha}\right)^5 \right\} 1\{d \le \alpha\}, \tag{5}$$

where $1\{\cdot\}$ is an indicator function so that $K(d)$ is identically 0 for $d > \alpha$. The model gets its name from the fact that it can be obtained by taking a moving average over balls of a five-dimensional white noise process. The distance d is straight line distance, but one can trivially rewrite the model in terms of great circle distances. The reason for using the pentaspherical model rather than the more familiar spherical model is that the pentaspherical model is twice differentiable in d for $d > 0$, which has both statistical and computational advantages (see Stein (1999, p. 53)) over the only once differentiable spherical model.

By adding a term of the form (5) to the covariance structure in (4) together with a nugget effect, we might hope to gain a good fit to the data on both large and small scales. We could try directly maximizing the likelihood with respect to the 179 parameters, in which case, we would want to write the sum of the nugget and pentaspherical terms as

$$\theta_0[\alpha_0 1\{d = 0\} + K(d/\phi_0)], \tag{6}$$

updating the values of α_0 and ϕ_0 only rarely in the iterations and putting some upper limit on ϕ_0 so that sparse matrix methods remain effective. Rather than carry out a full maximization, I first fit this 179-parameter model using the wls procedure described in Stein (2007) and obtained an estimated range $\widehat{\phi_0}$ of just over 4°. I then fixed this parameter and fit, via maximum likelihood, the other 178 parameters, taking care to update α_0 as rarely as possible. Maximization of the likelihood for fixed α_0 was carried out using the nlm routine in *R*. Stein (2007) found the maximum likelihood estimates for the model including just a nugget and spherical harmonic terms. For completeness, I also computed the maximum likelihood estimates of α_0 and θ_0 without any spherical harmonic terms and again setting $\widehat{\phi_0}$ to its wls estimate. Somewhat surprisingly, in this last case the mle of α_0 turned out to be 0. I wrote my own *R* routines to do the computations and made use of specific patterns of the observation locations to exploit the sparseness of $\mathbf{A}(\alpha_0)$ so that all computations could be reduced to Cholesky decompositions of unstructured matrices of size no greater than 1800. Davis (2006) describes a set of methods and software that can be used with arbitrary sparse matrices and do not require any user preprocessing. Once these calculations are done for any particular α_0, all further matrix operations are on matrices of size 64×64, since $(N + 1)^2 = 64$ for $N = 7$.

To give a reference base, I will compare the maximized loglikelihoods under these various models to the maximized loglikelihood under a white noise model. As reported in Stein (2007), the series expansion plus nugget model has a loglikelihood of 7372 units larger than the pure nugget model. The pentaspherical model plus nugget (which, as just noted, has an estimated nugget of 0) has a loglikelihood of 17,805 units greater than the white noise model. Thus, this very simple isotropic model with all covariances greater than 4° degrees equal to 0 has a loglikelihood more than 10,000 greater than the 177-parameter series expansion plus the nugget model. This result is not so unexpected, since, as far as the likelihood is concerned, most of the information in the data is about dependence on small scales, so a model that better describes this dependence will generally have larger likelihood even if it provides a poor description of dependence on large scales. Finally, consider combining a nugget effect, the pentaspherical model with range fixed at the wls estimate and the series expansion. This calculation yields a loglikelihood 18,809 greater than the white noise model. Thus, adding the 176 parameters from the series expansion increases the loglikelihood by 1004 over the nugget plus pentaspherical model, which may seem like a modest amount considering the large number of parameters and the large number of observations. However, data from a single orbit provides very little information about the 176 parameters in $\mathbf{K}(\theta_1)$. Indeed, we have even less information about θ_1 than if we directly observed $(U_1, \ldots, U_{64})'$

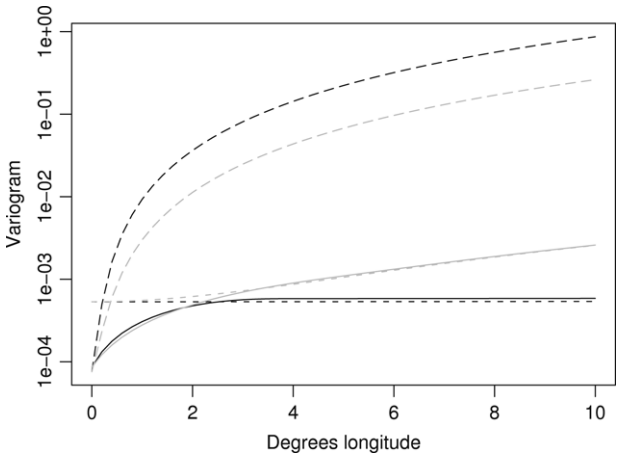

Fig. 1. Variograms along the east–west direction at 0° (black curves) and 40°N (gray curves). Long-dashed curves give likelihood-based estimate, solid curves give wls fit and short-dashed curves give wls fit omitting pentaspherical term.

in (2), which themselves are obviously insufficient to estimate 176 covariance parameters sensibly. Clearly, we need either more data or a more restricted model or both.

To obtain somewhat more stable estimates, I will use the data from all 83 orbits during this six-day period. I will ignore the dependence across orbits and maximize the sum of the loglikelihoods across the orbits as a way of obtaining parameter estimates. Removing the spatially persistent pattern from the data as we have done here (see Stein (2007)) does serve to decrease the statistical dependence across orbits. Nevertheless, there clearly is dependence across orbits even in the residuals we analyze here, so we cannot take the likelihoods that emerge from this calculation seriously, but the point estimates obtained should still be sensible. In particular, the score equations we obtain by setting the gradient of this "likelihood" to 0 give unbiased estimating equations for the parameters (Stein et al., 2004).

Again doing a crude search over α_0 values to minimize computations, I maximized this approximate likelihood over the 178 parameters (with $\widehat{\phi}_0$ fixed to its wls value), although convergence is slow and it is not certain whether I have found even a local maximum of the approximate likelihood. Fig. 1 plots some fitted variograms for pairs of points at the same latitude under this and the wls fits, as well as the wls fit from Stein (2007), which omits the pentaspherical term. At the shortest lags, the two fits based on the model including the pentaspherical term agree reasonably well and are quite a bit lower than the wls fit without the pentaspherical term, which Stein (2007) noted overestimates the variogram at these shortest lags. However, the likelihood fit is much larger (by several orders of magnitude) than the two wls fits at longer lags, which, by construction, track the observed variogram values at these lags rather well and are nearly indistinguishable beyond lags of 3° longitude. Note further that the likelihood fit gives larger variogram values at the equator than at 40°N, even though the wls fits (and the empirical variograms, see Stein (2007)) show just the opposite, at least at longer lags. This kind of radical disagreement between the empirical variogram and the likelihood-based estimate is most likely an evidence of model misspecification. The next section explores this issue further.

4. A useful diagnostic

To better understand why the likelihood-based estimate so badly overestimates the variogram at long lags, it helps to look at variances of contrasts of more than two observations. When observations are on a regular grid, one can compare empirical to fitted generalized variograms (Chilès & Delfiner, 1999), basically, variances of higher order differences. The TOMS Level 2 data are not on a grid, but there are near repetitions of observation patterns rotated by some angle about the Earth's axis. For example, Fig. 2 shows observation locations from three scans of the instrument, one for each of the first three orbits on May 1, 1990, the scan selected being the first on that orbit with an observation north of 40°N. One orbit out of the 83 available over May 1–6, 1990 has missing observations around this latitude, leaving 82 near "replicates" of the same pattern of 35 observations. The 82 sets of observations follow (nearly) the same distribution under axial symmetry, but they are not independent. However, it is still informative to compare fitted and empirical variances for different linear combinations of the observations. Specifically, for $j = 1, \ldots, 82$ and $k = 1, \ldots, 35$,

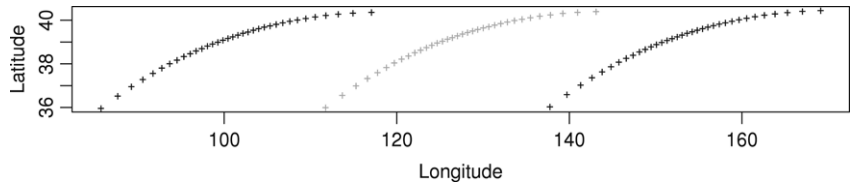

Fig. 2. Locations of observations for scans from first three orbits on May 1, 1990 with first observation just north of 40°N.

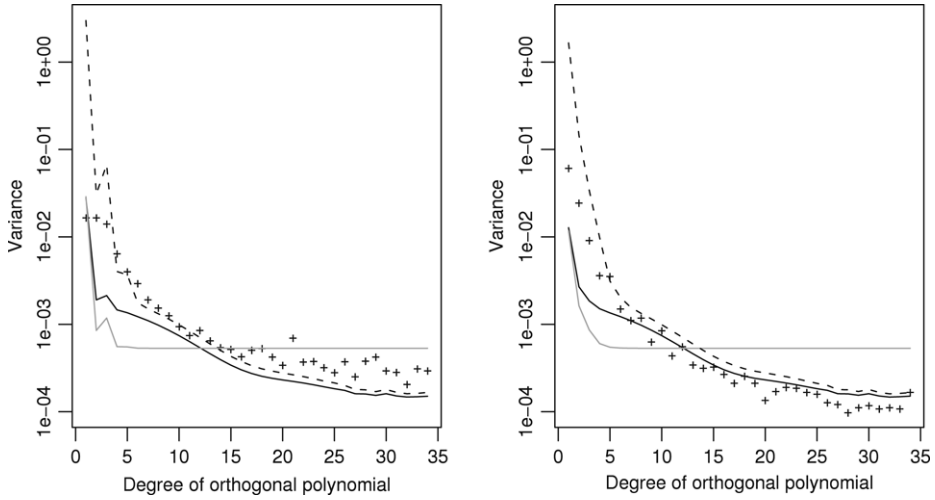

Fig. 3. Empirical variances ("+") and fitted variances (dashed for likelihood-based estimate, solid for wls and gray for wls without pentaspherical term) for orthogonal polynomial contrasts. Left plot is for scans beginning at equator; right plot for scans beginning at 40°N.

define Z_{jk} to be the kth observation from the jth orbit. Next, for $\ell = 0, \ldots, 34$ and $k = 1, \ldots, 35$, let $\lambda_{\ell+1,k}$ be the $(\ell+1, k)$th element of the Gram–Schmidt orthogonalization of the 35×35 matrix with $(\ell+1, k)$th element ℓ^k. The use of orthogonal polynomials to define the contrasts is somewhat arbitrary and not completely satisfactory since the observation locations are not equally spaced on a line, but calculations that take into account this uneven spacing give very similar results. Fig. 3 plots the empirical variances $\frac{1}{82}\sum_{j=1}^{82}\left(\sum_{k=1}^{35}\lambda_{\ell k}Z_{jk}\right)^2$ for $\ell = 1, \ldots, 34$, as well as similar results for scans in which the first observation is just north of the equator. The empirical variances tend to go down as the degree of the polynomial increases as we might expect, since there is more variation over larger spatial scales and the higher order terms correspond to higher spatial frequencies and hence smaller spatial scales. If the observations within an orbit were independent and identically distributed, we would expect to see no pattern in the empirical variances as the degree of the orthogonal polynomial increases. Note that for $\ell > 2$, there is more variation at the equator than at 40°N, an important feature of the data that cannot be seen by looking at the empirical variograms in Stein (2007).

Fig. 3 also plots estimated values for $\mathrm{var}\left(\sum_{k=1}^{35}\lambda_{\ell k}Z_{1k}\right)$ for the three different fitted models shown in Fig. 1. The wls fit from Stein (2007) performs quite badly for most degrees ℓ at both latitudes, indicating that this model provides a poor description of the process, despite its decent agreement with the empirical variograms in Stein (2007). Note in particular that the estimated variances under this fit hardly change for $\ell > 5$, which reflects the fact that the smooth spherical harmonic terms hardly affect the variances when ℓ is not small and the estimated variances are almost entirely due to the nugget effect. Adding the pentaspherical term to the model greatly improves the fit of the wls procedure in this figure. The great improvement here is in stark contrast to the improvement in the wls criterion function, which was lowered by only about 2% by adding the pentaspherical term. Overall, the likelihood-based estimate appears to do noticeably better than the wls fit at the equator, especially for moderate k, and somewhat worse at 40°N. At both latitudes, especially at the equator, the likelihood-based estimate is far too large for degree $\ell = 1$, which we should expect from Fig. 1. For larger ℓ, the data show more variation at the equator than at 40°N, and it is presumably this feature of the data that the likelihood is trying to capture, which explains why, in Fig. 1, the likelihood-based fitted

variogram at the equator is so large. Specifically, in trying to use the very smooth spherical harmonics to capture the greater high frequency variation near the equator, the likelihood-based estimate is forced to overestimate the low frequency variation particularly badly in that same latitude band.

The implications of these results for statistical modeling and analysis of spatial processes on a global scale are not completely clear to me. However, since Bayesian approaches using MCMC calculations are coming into regular use for modeling large spatial and spatial-temporal datasets (see, for example, Le and Zidek (2006)) and that Bayesian methods may work poorly when there is nothing in the model space that approximates reality reasonably well, the need for caution is apparent. This is particularly important if one has data with high spatial resolution but one is at least in part interested in the behavior of the process at scales much larger than the typical distance between neighboring observations.

5. Discussion

The ability to capture both local and large-scale behavior in a spatial model for which exact likelihood computations are feasible even for very large datasets makes the modeling approach described here attractive in principle, despite its shortcomings in the present application. Further extensions that could expand its applicability are possible. For example, we only need that $\mathbf{A}$ be sparse; it does not have to be generated from a stationary and isotropic model. Thus, we could allow geometric anisotropies or even nonstationarities as long as the resulting $\mathbf{A}$ is sparse. Allowing the scale of variation in the pentaspherical model to vary with latitude would undoubtedly help with fitting the TOMS data considered here.

An obvious way to obtain a more general model than the one used here is just to increase p, the number of columns of $\mathbf{P}$. In our application, p was only 64 and one could certainly increase this number considerably and still be able to calculate the exact likelihood repeatedly. Perhaps the more meaningful constraint is on the number of parameters in the model. If we do not put any constraint on $\mathbf{K}$ in (1) other than it is positive semidefinite, then it has $p(p+1)/2$ independent parameters. In the application here, the axial symmetry forces $\mathbf{K}$ to be block diagonal, which reduces the number of parameters considerably. However, if we took $N = 19$ in (4) so that p would be a still manageable value of 400, the resulting number of parameters in the axially symmetric model is 2680, which would make for a challenging function to maximize or to sample effectively in an MCMC scheme and, more importantly, is just too many parameters for a spatial covariance function even if one had many replicates of the process. Clearly, it would be helpful if one could place further constraints on the model to reduce the number of parameters, but still leave enough flexibility to capture large-scale nonstationarity.

The computational ideas presented here can be applied to the prediction problems for spatial data. For example, the commonly used spatial interpolation method of kriging corresponds to calculating optimal linear predictors acting as if the estimated covariance function is the truth. Both the point predictions and the covariance matrix of the prediction errors can be obtained from the Cholesky decomposition of the estimated covariance matrix, although, since the determinant is not needed for this problem, one could also consider using iterative methods for sparse matrices (Saad, 2003). Conditional simulations (Chilès & Delfiner, 1999) can also be conveniently carried out under a model with a compactly supported term plus a term with moderate rank since conditional simulation of Gaussian processes just requires a kriging step and an unconditional simulation of the process at the prediction location. Unconditional simulation is easy under this model, since the low rank terms and the contributions from the compactly supported part of the model can be simulated independently.

Disclaimer

Although the research described in this article has been funded wholly or in part by the United States Environmental Protection Agency through STAR cooperative agreement R-82940201-0 to the University of Chicago, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

References

Caragea, P. (2003). Approximate likelihoods for spatial processes. *Ph.D. Dissertation at UNC*. http://www.stat.unc.edu/postscript/rs/caragea.pdf.
Chilès, J., & Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty*. New York: Wiley.

Cressie, N., & Johanneson, G. (2008). Fixed rank kriging for large spatial datasets. *Journal of the Royal Statistical Society B*, *70*, 209–226.

Davis, T. A. (2006). *Direct methods for sparse linear systems*. Philadelphia: Society for Industrial and Applied Mathematics.

Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, *102*, 321–331.

Furrer, R., Genton, M. G., & Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, *15*, 502–523.

Jones, A. H. (1963). Stochastic processes on a sphere. *Annals of Mathematical Statistics*, *34*, 213–217.

Krueger, A. J., Bhartia, P. K., McPeters, R. D., Herman, J. R., Wellemeyer, C. G., Jaross, G., et al. (1998). *ADEOS total ozone mapping spectrometer (TOMS) data products user's guide*. Greenbelt, MD: National Aeronautics and Space Administration, Available at http://www.toms.gsfc.nasa.gov/datainfo/adeos_userguide.pdf.

Le, N. D., & Zidek, J. V. (2006). *Statistical analysis of environmental space-time processes*. New York: Springer.

Saad, Y. (2003). *Iterative methods for sparse linear systems* (2nd ed.). Philadelphia: Society for Industrial and Applied Mathematics.

Stein, M. L. (1999). *Interpolation of spatial data: Some theory for Kriging*. New York: Springer.

Stein, M. L. (2007). Spatial variation of total column ozone on a global scale. *Annals of Applied Statistics*, *1*, 191–210.

Stein, M. L., Chi, Z., & Welty, L. J. (2004). Approximating likelihoods for large spatial datasets. *Journal of the Royal Statistical Society B*, *66*, 275–296.