APPENDIX AVAILABLE ON THE HEI WEB SITE

Research Report 167

Assessment and Statistical Modeling of the Relationship Between Remotely Sensed Aerosol Optical Depth and $PM_{2.5}$ in the Eastern United States

Christopher J. Paciorek and Yang Liu

Appendix C. Statistical Details for Flexible Spatial Latent Variable Modeling

---

Correspondence may be addressed to Dr. Christopher J. Paciorek, Department of Statistics, 367 Evans Hall, University of California, Berkeley, CA 94720; e-mail: *paciorek@stat.berkeley.edu*.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83234701 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and therefore may not necessarily reflect the views of the Agency, and no official endorsement by it should be inferred. The contents of this document also have not been reviewed by private party institutions, including those that support the Health Effects Institute; therefore, it may not reflect the views or policies of these parties, and no endorsement by them should be inferred.

This document was reviewed by the HEI Health Review Committee
but did not undergo the HEI scientific editing and production process.

C. Statistical Details for Flexible Spatial Latent Variable Modeling

C.1. Derivation of the MRF weight matrix for the thin plate spline approximation

Rue and Held (2005, Sec. 3.4.2) give a basic overview of the second-order, two-dimensional intrinsic Gaussian MRF that approximates a thin plate spline, but do not provide the details regarding the boundary condition effects that allow one to construct the full weight matrix, Q. Here we provide the details, following the development in an unpublished manuscript by Y. Yue and P. Speckman, also given in Y. Yue's unpublished Ph.D. dissertation from the Department of Statistics at the University of Missouri, Columbia.

A thin plate spline minimizes a penalized likelihood where the penalty term for a function, $g(\cdot)$, defined for $(s_1, s_2) \in \Re^2$ is

$$J(g) = \iint_{\Re^2} \left[\left(\frac{\partial^2 g(s_1,s_2)}{\partial s_1^2} \right)^2 + 2\left(\frac{\partial^2 g(s_1,s_2)}{\partial s_1 \partial s_2} \right)^2 + \left(\frac{\partial^2 g(s_1,s_2)}{\partial s_2^2} \right)^2 \right] ds_1 \, ds_2. \tag{C1}$$

Taking the MRF as an approximation to $g(\cdot)$ on a regular grid of size $m_1 \times m_2$, a discretized approximation to $J(g)$ is

$$\sum_{i=3}^{m_1} \sum_{j=1}^{m_2} (\Delta_{s_1}^2 g_{i,j})^2 + 2 \sum_{i=2}^{m_1} \sum_{j=2}^{m_2} (\Delta_{s_1,s_2}^2 g_{i,j})^2 + \sum_{i=1}^{m_1} \sum_{j=3}^{m_2} (\Delta_{s_2}^2 g_{i,j})^2 \tag{C2}$$

where

$$\Delta_{s_1}^2 g_{i,j} = g_{i,j} - 2g_{i-1,j} + g_{i-2,j}$$

$$\Delta_{s_1,s_2}^2 g_{i,j} = g_{ij} - g_{i-1,j} - (g_{i,,j-1} - g_{i-1,j-1})$$

$$\Delta_{s_2}^2 g_{i,j} = g_{i,j} - 2g_{i,j-1} + g_{i,j-2}.$$

$\Delta_{s_1}^2 g_{i,j}$ and $\Delta_{s_2}^2 g_{i,j}$ are the second order backward difference operators in the horizontal and vertical directions. $\Delta_{s_1,s_2}^2 g_{i,j}$ is found as the first order backward difference operator in the vertical direction applied to the first order backward difference operators in the horizontal direction.

The difference calculations are done for every cell on the grid and summed to approximate the integral in (C1). Note that they cannot be calculated for cells on the lower and lefthand boundaries of the rectangular grid, nor can they be calculated for some of the terms for the cells in the second row from the bottom or second row from the left, hence some of the indices of summation in (C2) do not start at one. Then, to equate the elements of Q with the coefficients of $\{g_{i,j}\}$ in (C2), note that the discretized penalty can be expressed as

$g^T Qg = \sum_{k=1}^{m} g_k^2 Q_{kk} + 2\sum_{k=1}^{m} \sum_{l<k} g_k g_l Q_{kl}$ where $g = \text{Vec}(\{g_{i,j}\}) = \{g_k\}_{k=1,..,m}$ and $m = m_1 m_2$. Thus Q_{kk} is the coefficient multiplying g_k and Q_{kl} is one half the coefficient multiplying $g_k g_l$. Note that $g^T Qg = 0$ for the constant function and any linear functions of the two coordinates, so the rank of Q is $m-3$, and the prior puts infinite variance on the intercept and linear terms in the two coordinates.

Working out the algebra, for a given cell, k, we can write the set of neighbors and the elements of that cell's row in Q in a spatial representation where the center element is Q_{kk} and the non-zero Q_{kl} values are represented in terms of the relative positions of the cells of g_k and g_l on the grid. We see that the neighbor structure and elements of Q corresponding to cells in the interior and various categories of cells near the boundary, are:

(a)

	2	−8	2	
1	−8	20	−8	1
	2	−8	2	
		1		

(b)

	2	−8	2	
1	−8	19	−8	1
	2	−6	2	

(c)

	2	−8	2	
	−6	18	−8	1
	2	−6	2	

(d)

	2	−6	2	
	−4	10	−6	1

(e)

	2	−6	2	
1	−6	11	−6	1

(f)

		−4	2	
		4	−4	1

where (a) is for an interior cell, at least two cells away from any boundary, (b) is for a cell two cells from one boundary and one cell from another boundary, (c) is for a cell one cell from each boundary, (d) is for a cell on one boundary and one cell from another other boundary, (e) is for a cell on one boundary and two or more cells from another boundary, and (f) is for a cell in one of the four corners.

Note that Yue and Speckman (unpub.) provide a computational shortcut for computing Q as

$$Q = I_{m_2} \otimes Q_{m_1}^{(2)} + 2Q_{m_1}^{(1)} \otimes Q_{m_2}^{(1)} + Q_{m_2}^{(2)} \otimes I_{m_1}$$

where $Q_n^{(1)}$ and $Q_n^{(2)}$ are the one-dimensional first and second order $n \times n$ structure (weight) matrices provided in Rue and Held (2005, pp. 95, 110):

$$Q_n^{(1)} = \begin{pmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix}, Q_n^{(2)} = \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}.$$

C.2. Spatial model structure

Our basic model is a model with two likelihoods and additive mean terms, in particular

$$\begin{aligned} Y &\sim N_{n_Y}(Z_y b_y + Z_L b_L + P_\delta \delta, V_Y) \\ A &\sim N_{n_A}(P_\varphi \varphi + Z_a b_a + \beta_1 P_A Z_L b_L, V_A) \\ \delta &\sim N(0, \sigma_h^2 I) \\ \varphi &\sim N_{m-3}(0, (\kappa Q)^-) \\ b = \{b_y, b_L, b_a\} &\sim N(0, \Lambda) \end{aligned}$$

where $Z_y b_y$ is the matrix representation of $\sum_p \beta_{y,p}(x_{y,p})$ and $Z_L b_L$ is the matrix representation of $\sum_p \beta_{L,p}(x_{L,p}) + g(s)$ with b_L the collection of combined coefficients for the regression smooths and the spatial term, as well as including an intercept for Y. Similarly, $Z_a b_a$ represents the influence of explanatory variables for the proxy unrelated to latent $PM_{2.5}$ (cloud cover in the case of the AOD model). δ are site-specific effects that account for correlation between monitors placed at the same site. We denote the variance of φ using the generalized inverse to indicate that the prior is proper in an $m-3$ dimensional space, fixing the mean and coefficients for linear terms of the spatial coordinates to zero. In our sampling, the three parameters are identified by the likelihood for A, so we sample these parameters implicitly as part of φ and therefore omit a separate intercept for A. Given the limited number of observation locations, we use five knots for each regression smooth and 55 knots for the spatial residual. Knots were placed either uniformly over the range of covariate values or at equally-spaced quantiles to achieve reasonable spread over the covariate spaces, but given the use of penalized splines, results should be robust to the exact placement of knots. Λ is the prior covariance matrix of b (non-informative for the fixed effect components and with exchangeable priors amongst the coefficients for a given regression smooth term, following Ruppert et al. (2003)).

We integrate over the conditional normal distribution of φ with mean $M_\varphi = V_\varphi(P_A^T V_A^{-1}(A - Z_a b_a - P_A Z_L b_L))$ and variance $V_\varphi = (P_\varphi^T V_A^{-1} P_\varphi + \kappa Q)^{-1}$ to obtain

$$A \sim N_{n_A-3}(Z_a b_a + \beta_1 P_A Z_L b_L, \Sigma_A).$$

Here $\Sigma_A^{-1} = V_A^{-1} - V_A^{-1} P_\varphi V_\varphi P_\varphi^T V_A^{-1}$ and $|\Sigma_A|^{-\frac{1}{2}}$ can be expressed as

$$\frac{1}{|\Sigma_A|^{\frac{1}{2}}} = \frac{\kappa^{\frac{m-3}{2}} |Q|^{\frac{1}{2}}}{|V_A|^{\frac{1}{2}} |V_\varphi^{-1}|^{\frac{1}{2}}}.$$

Note that the impropriety in the prior for φ carries over into this marginal likelihood for A, resulting in $m-3$ rather than m in the exponent of κ and in Σ_A^{-1} being singular, with three zero eigenvalues, but our subsequent calculations all involve Σ_A^{-1} so no inversion is needed. Equivalently, we do not have a legitimate data-generating model for A because of the prior impropriety, with information in three of the linear combinations in the quadratic form in the exponent of the marginal likelihood contributing zero to the marginal likelihood because the variance for those combinations is infinite. We can avoid calculating the non-existent determinant of Q because this is a constant with respect to the model parameters. Note that the impropriety is analogous to that in the marginal likelihood obtained from integrating over the mean in a simple normal mean problem with an improper prior for the mean.

We can then integrate over the joint distribution for $b = \{b_y, b_L, b_a\}$, where we construct Z_Y and Z_A such that $Z_Y b = Z_y b_y + Z_L b_L$ and $Z_A b = Z_a b_a + P_A Z_L b_L$ by adding columns with all zeroes as necessary. The conditional distribution for b is normal with mean, $M_b = V_b (Z_Y^T V_Y (Y - P_\delta \delta) + Z_A^T \Sigma_A^{-1} A)$ and $V_b = (Z_Y^T V_Y^{-1} Z_Y + Z_A^T \Sigma_A^{-1} Z_A + \Lambda^{-1})^{-1}$. We collect the remaining parameters as $\theta = \{\beta_1, \sigma^2_{\text{sub}}, \sigma^2_\delta, \sigma^2_\varepsilon, \sigma^2_A, \sigma^2_\alpha, \sigma^2_{b,L}, \sigma^2_{b,a}, \sigma^2_{b,y}, \sigma^2_{b,p}, \kappa\}$ where the variance components, $\sigma^2_{b,y}$, $\sigma^2_{b,L}$, and $\sigma^2_{b,a}$, are vectors with one variance component for each smooth regression term in a given sum of regression smooths and σ^2_{sub}, σ^2_δ, σ^2_ε, σ^2_A, and σ^2_a are parameters used to construct V_Y and V_A, described below. The marginal posterior for the remaining parameters and δ is

$$P(\theta, \delta \mid A, Y) \propto |\Lambda|^{-\frac{1}{2}} |V_Y|^{-\frac{1}{2}} |\Sigma_A|^{-\frac{1}{2}} |V_b|^{\frac{1}{2}} P(\delta) P(\theta) \cdot$$
$$\exp\left(-\frac{1}{2} \left((Y - P_\delta \delta)^T V_Y^{-1} (Y - P_\delta \delta) + A^T \Sigma_A^{-1} A - M_b^T V_b^{-1} M_b \right) \right), \quad \text{(C3)}$$

which we use to sample θ via blocked Metropolis. Depending on the model, in some cases we use a single block and in other cases subblocks. We use adaptive MCMC to tune the proposal covariance matrix throughout the chain (Andrieu and Thoms 2008)

The key computational impediments involve the determinant of V_φ^{-1}, which can be calculated based on sparse matrix operations since both of its components are sparse; note that in our models P_φ is a simple mapping matrix assigning elements of φ to the proxy values, but might be used more elegantly to realign between different grids, in which case it would still be sparse but with non-zero weights reflecting the overlap of cells between the different grids. Next V_b is a dense matrix whose size corresponds to the number of basis coefficients, which can be

computationally burdensome when we use a large number of knots for g or the total number of knots used for all the regression smooth terms is large. Finally, we must compute $\Sigma_A^{-1}A$ and $\Sigma_A^{-1}Z_A$, the latter being more burdensome because Z_A is a matrix with as many non-zero columns as there are coefficients in $\{b_a, b_L\}$. Considering the representation of Σ_A^{-1}, note that we can easily compute $V_A^{-1}Z_A$ and then $P_\varphi^T V_A^{-1}Z_A$ because V_A^{-1} is diagonal and P_φ^T is sparse. Next we use sparse matrix operations to solve the system of equations $V_\varphi P_A^T V_A^{-1}Z_A$ (recall that we calculate V_φ^{-1} quickly as the sparse matrix sum of two sparse matrices). We use the spam package in R for sparse matrix calculations.

Given the posterior for the remaining parameters (C3), we can derive the closed form normal conditional distribution for δ, which has mean and variance,

$$M_\delta = V_\delta(P_\delta^T V_Y^{-1}Y - P_\delta^T V_Y^{-1}Z_Y V_b(Z_Y^T V_Y^{-1}Y + Z_A^T\Sigma_A^{-1}A))$$
$$V_\delta^{-1} = \sigma_\delta^2 I + P_\delta^T V_Y^{-1}P_\delta - P_\delta^T V_Y^{-1}Z_Y V_b Z_Y^T V_Y^{-1}P_\delta.$$

Sampling from this distribution efficiently involves sparse matrix calculations similar to those just described.

Posterior samples of φ and b can be drawn off-line from the conditional distributions indicated above; we choose to draw them every 10 MCMC iterations.

V_Y is modeled using a diagonal heteroscedastic variance, $(V_Y)_{ii} = \sigma_\varepsilon^2 / n_i + k(n_i)\sigma_{\text{sub}}^2$ where the first term is the variance of the sum of independent daily instrument errors. The second reflects subsampling variability in the average of n_i instrument-error-free daily values relative to the average of the error-free daily values over all the days in the month, $\text{Var}(\sum_{d\in\text{subsample}} L_d / n_i - \sum_d L_d / n_{\text{month}})$, under the simplifying assumption of independence between true daily pollution values, $L_d \overset{iid}{\sim} N(L, \sigma_{\text{sub}}^2)$, which gives $k(n_i) = \frac{1}{n_i} - \frac{1}{n_{\text{month}}}$, where n_{month} is the number of days in the month. Note that for simplicity we fixed σ_ε^2 at $\sigma_\varepsilon^2 = 1.5$, estimated in advance from co-located monitors, to enhance identifiability and because it has a small contribution to the overall error variance. For monitors not co-located with another monitor, we integrated over the prior for the δ values for those sites, which added a term, σ_h^2, to $(V_Y)_{ii}$, but we sampled the values of δ for sites with co-located monitors within our primary MCMC to avoid introducing off-diagonal elements into V_Y as this would have obviated some of the computational efficiencies in the calculations outlined above. For models involving CMAQ, which is available for all days, as the proxy, we take $V_A = \sigma_A^2 I$. For models involving AOD, we use a diagonal heteroscedastic variance analogous to V_Y that reflects the number of daily values in each monthly average, plus a homoscedastic term reflecting the fundamental discrepancy between AOD and true $PM_{2.5}$: $(V_A)_{ii} = \sigma_A^2 + k(n_i)\sigma_\alpha^2$.

We use several covariates calculated at the grid level for individual cells: elevation at the cell

centroids, population density, and total length of roads in three road classes. Area $PM_{2.5}$ emissions from the 2002 EPA National Emissions Inventory (NEI) are calculated as density of emissions per county and the value for the county of the grid cell centroid is assigned to the grid cell. Population density, road density, and area emissions are log-transformed to reduce sparsity and pull in extremely large values in the right tail, and we truncated the values of some outlying covariates to reduce extrapolation problems. We used the NEI point source emissions strength and location data in the flexible buffer modeling described in Appendix D, creating a basis matrix that contributes columns to Z_L. For the observation likelihood, we calculate the source strength-weighted sum of distance-weighted contributions from $PM_{2.5}$ primary source point emissions within a maximum distance (100 km) for each monitor, omitting sources emitting less than five tons in 2002. For the proxy likelihood and for prediction on the grid, we take a subgrid of 16 points within each grid box and calculate the average sum of contributions from the point emissions within 100 km of the points in the subgrid, as a simple approximation to the true integral of the point emission effect over the grid cell.

Some CMAQ pixels overlap four km cells both on land and those in the ocean or Great Lakes with the cells primarily over water having undefined covariate values for some covariates. We treat the CMAQ value in a CMAQ pixel as reflecting the weighted average of $L(\cdot)$ from only the land-based four km grid cells, with weights in P_A summing to one for each CMAQ pixel. Exploratory analysis indicated that CMAQ-estimated $PM_{2.5}$ in pixels on the land-water boundary was often high, such that not normalizing the weights to sum to one distorted our model fitting. The problem is that not normalizing reduces the contribution from L to the mean of the CMAQ proxy, increasing the discrepancy between the proxy values and L in cases (such as the New York City area) where the CMAQ proxy is much larger than the estimated grid cell values that are driven by monitors with lower values. We do not include CMAQ values in the likelihood for pixels with 60% or more overlap with four km cells that do not intersect land in the U.S.

In general, our prior distributions for hyperparameters were non-informative, with normal priors with large variances (and also lower and upper bounds to prevent the MCMC from wandering in flat parts of the posterior) for location parameters and uniform scale parameters on the standard deviation scale (Gelman 2006), with large upper bounds. In all cases, the posterior distributions were much more peaked than the prior distributions and away from the bounds, except for some of the variance components for the coefficients of the regression smooths, which we restricted to avoid overly wiggly smooth terms. Furthere exploration of why these smooths tend toward less smooth functions than expected scientifically and on whether simple linear relationships would suffice and might even improve out-of-sample prediction would be worthwhile.

We ran the MCMC for 10,000 iterations during the burn-in and 25,000 subsequently, retaining every 10th iteration to reduce storage costs. We found reasonable convergence and mixing based on effective sample size calculations and trace plots. We did not run multiple chains for a given month and validation set because of our use of multiple months and validation sets, noting that predictive performance also helps to justify the adequacy of our fitting.

C.3. Spatio-temporal model structure

The spatio-temporal model structure builds on the spatial model structure but with autoregressive structure for the basis coefficients, $b_{g,t}, t = 1,...,T = 12$, of the monthly spatial residual surfaces, $g_t(\cdot)$, and an exchangeable structure for monthly discrepancy terms, $\varphi_t, t = 1,...,12$, as well as month-specific $\beta_{1,t}$ with independent non-informative priors. Calculations based on sparse matrix routines follow those described for the spatial model but with $D \otimes Q$ in place of κQ. Because of the increase in dimensionality, it is difficult to work with grids as large as in the spatial model. Our spatio-temporal model works with the 19×11 CMAQ grid, giving us a $19 \times 11 \times 12 = 2508$ dimensional φ. A major cause of slowdown is that our b vector is now much higher dimensional, as it includes $55 \times 12 = 660$ basis coefficients for the 12 residual spatial surfaces. Given the increase in sample size (albeit not locations), we use 10 rather than 5 knots for the regression smooth terms, allowing for the possibility of estimating additional nonlinearity. We ran the MCMC for 10,000 iterations during the burn-in and 50,000 subsequently, retaining every tenth iteration to reduce storage costs, again finding reasonable convergence and mixing.

C.4. Subnational model structure

Spatio-temporal modeling of small-scale spatial variation for the eastern U.S. is computationally challenging, so we fit separate spatial models for each month. We represent φ and g as TPS-MRFs on the $73 \times 77 = 5621$ dimensional CMAQ grid over the eastern U.S., each with its own precision parameter. Covariate effects are represented on the original four km base grid (now of dimension $669 \times 677 = 452{,}913$). Pre-computation of Z_A in advance of the MCMC involves the large matrix multiplication of a basis matrix and an averaging matrix that represents the weighted average of four km cells within each CMAQ pixel based on the amount of overlap. Z_Y also represents the product of a mapping matrix and the original basis matrices for the covariates. Note that we rely on the covariates to represent small-scale variation in residual spatial variability, such that representing g on the 36-km CMAQ grid is sufficient. Posterior assessment indicates that g is quite smooth, supporting this approach. Given the relatively large sample size, for this model we again use 10 rather than 5 knots for each regression smooth term.

Our integration over φ now involves integration over $\varphi^* = \{g, \varphi\}$ followed by integration over b, which no longer includes b_g. The first integration is done by representing the joint likelihood as

$$\begin{pmatrix} Y \\ A \end{pmatrix} \sim N_{n_Y + n_A}\left(\begin{pmatrix} Z_Y \\ Z_A \end{pmatrix} b + \begin{pmatrix} P_Y & 0 \\ \beta_1 P_A & P_A \end{pmatrix} \begin{pmatrix} g \\ \varphi \end{pmatrix} + \begin{pmatrix} P_\delta \\ 0 \end{pmatrix} \delta, \begin{pmatrix} V_Y & 0 \\ 0 & V_A \end{pmatrix} \right),$$

followed by analogous calculations as in the original spatial model to integrate over b and determine the marginal posterior (up to the normalizing constant) for the remaining parameters and the conditional normal posterior for δ given the remaining parameters and the data. Note

that P_Y and P_A simply map from the CMAQ grid cells to the observations and CMAQ values.

For our point source emissions covariate, computational demands required that we consider only point sources emitting more than 10 tons per year within 50 km of a given location, with our integral approximation using a subgrid with four, rather than 16, points within each four km cell.

We ran the MCMC for 10,000 iterations during the burn-in and 20,000 subsequently, retaining every 10th iteration to reduce storage costs, again finding reasonable convergence and mixing.

References

Andrieu C, Thoms J. 2008. A tutorial on adaptive MCMC. Statistics and Computing 18: 343–373.

Gelman A. 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis 1: 515–534.

Rue H, Held L. 2005. Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall, Boca Raton.