Computationally-efficient Statistical Modeling for Spatial and Spatio-temporal Data

Chris Paciorek
Department of Statistics, University of California, Berkeley
and
Department of Biostatistics, Harvard School of Public Health

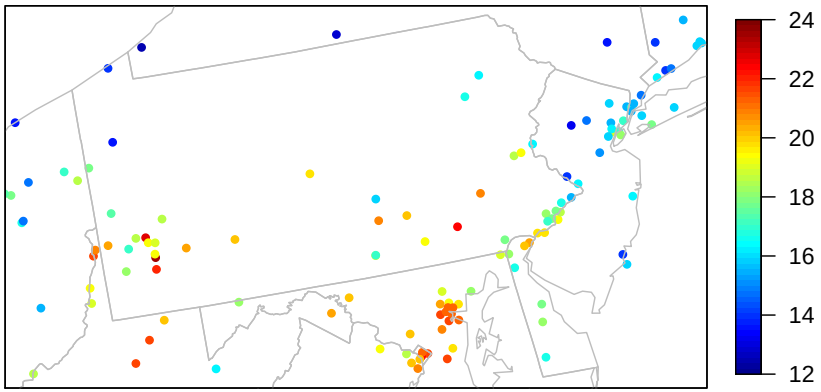
www.biostat.harvard.edu/~paciorek

Research supported by HEI 4746-RFA05-2/06-7 and NIH/NCI P01 CA134294-02

October 4, 2010

Spatial and spatio-temporal data

- Traditionally, spatial statistics has focused on relatively small observational datasets.



- To estimate (predict) the spatial surface at all locations, we estimate how the correlation of observations behaves as a function of the distance between the locations.
- In the last 10-15 years, there has been a lot of attention to larger datasets and computational impediments.
    - In particular, remote sensing and computer code output.

A basic hierarchical model

- Models for spatial data typically have a latent Gaussian process at their core:

$$\begin{aligned} \mathbf{Y} &\sim \mathcal{N}(\mathbf{g}, \sigma^2 \mathbf{I}) \\ \mathbf{g} &\sim \mathcal{N}(\mathbf{X}\beta, \Sigma) \end{aligned}$$

  Here $\mathbf{g}$ has a Gaussian process prior. The structure of Σ determines the behavior of the spatial field, $\mathbf{g}$.
- When one fits the models, the result is a tradeoff between fidelity to the data and constraints imposed by the process representation and its covariance (the prior).
- Uncertainty about our estimation of $\mathbf{g}$ reflects noise in the data and the strength of the prior.

Extending the model

- A simple model for counts of events:

$$\begin{aligned} \mathbf{Y} &\sim \text{Poi}(\exp(\mathbf{g})) \\ \mathbf{g} &\sim \mathcal{N}(\mathbf{X}\beta, \Sigma) \end{aligned}$$

- Depending on the data sources, much more complicated models build on this basic representation.

Covariance models

- The covariance matrix is based on a covariance function. At its simplest, one specifies a stationary covariance that is just a function of distance between the observations:
- Exponential:

$$\Sigma_{ij} = \tau^2 \exp\left(-\frac{d_{ij}}{\theta}\right)$$

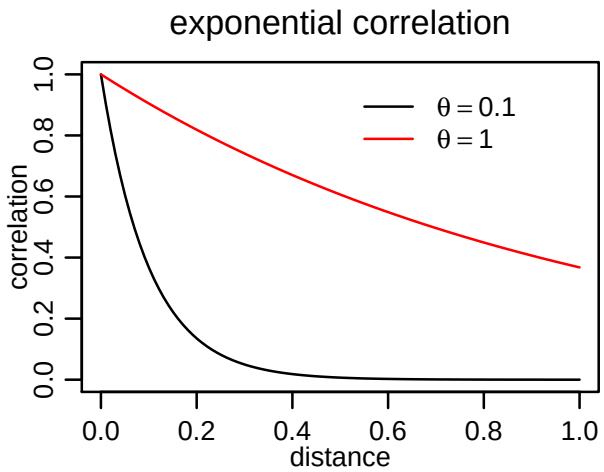
Processes are continuous but not differentiable (like Brownian motion)

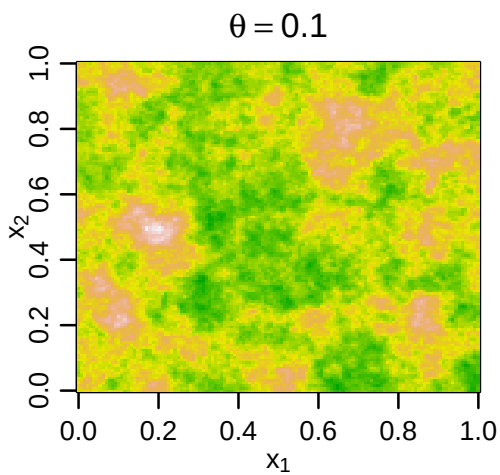
- Matern:

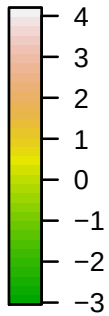
$$\Sigma_{ij} = \tau^2 \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\sqrt{\nu} d_{ij}}{\theta}\right) \mathcal{K}_\nu \left(\frac{2\sqrt{\nu} d_{ij}}{\theta}\right)$$

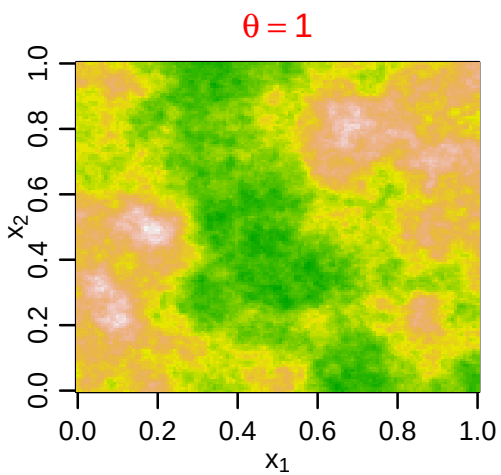
Processes are continuous and the number of derivatives increases with ν.

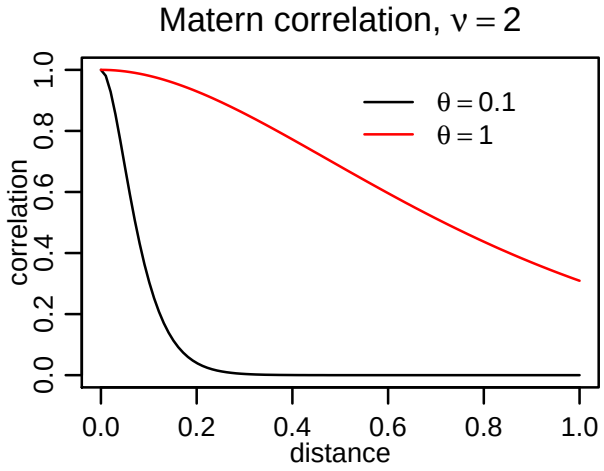
Covariance models illustrated

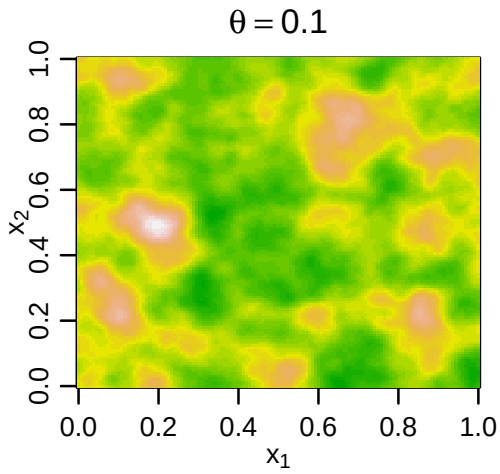


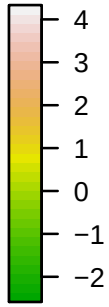


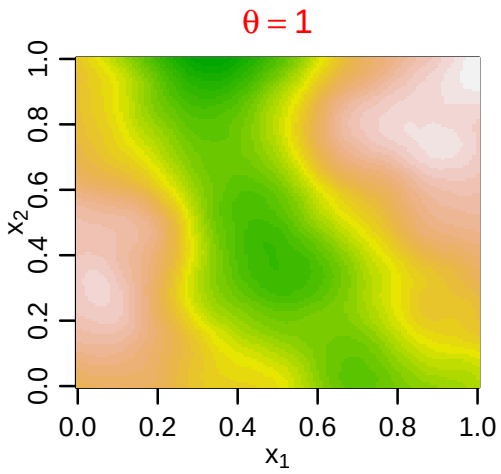












Discrepancy

- Suppose we have a computer code (e.g., climate model, hydrologic model, etc.). We might embed the code in a statistical representation to do model validation or model parameter estimation (calibration).
- The difference between the code and the truth is likely correlated spatially. So we might consider

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{Pm} + \mathbf{D}, \sigma^2 \mathbf{I})$$

  where $\mathbf{D}$ is a discrepancy term, spatially-correlated.
- Our spatial statistical approach (the Gaussian process) could then be employed here in terms of D:

$$\mathbf{D} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

Scientific goals

- Prediction (i.e., mapping, interpolation)
- Regression/association analysis that accounts for the spatial correlation
- Borrowing strength spatially in the face of large pixel-specific uncertainty
- Decomposing signal and noise and detection of patterns/anomalous areas

Gaussian process computations

- Σ is a dense matrix, often of size $n \times n$, as is $\Sigma^* = \Sigma + \sigma^2 \mathbf{I}$.
- Standard computations involved in maximizing the likelihood, doing kriging, and doing Bayesian estimation involve $\mathbf{\Sigma}^{*-1}\mathbf{z}$ and $\mathbf{\Sigma}^{*-1}\mathbf{Z}$ as well as $|\mathbf{\Sigma}^*|$
  - Maximimation typically involves a small number of iterations.
  - MCMC for Bayesian estimation can involve more than 10,000 iterations.
- The usual computational approach is to (in each iteration) compute the Cholesky decomposition, $\mathbf{L}^\top \mathbf{L} = \mathbf{\Sigma}^*$ and solve the necessary systems of equations.
- The Cholesky is $O(n^3)$, and this often is the rate-limiting step.

Computational strategies

- **Do the covariance calculations faster**
  - Threaded BLAS/Lapack (GotoBLAS, MKL, ACML)
  - Distributed calculation of the Cholesky decomposition (Scalapack)
- Approximate the covariance calculations
- Low-rank approximation of the spatial process
- **Sparse representation of the covariance (Markov random fields)**

Fast Cholesky decomposition

- Threaded linear algebra packages can greatly speed up the Cholesky in a multicore environment (shared memory): GotoBLAS, MKL, ACML
- Scalapack does the Cholesky block-wise in a distributed environment but entails a lot of communication overhead.
- For some of our statistical calculations (e.g., $\boldsymbol{\Sigma}^{*-1}\mathbf{z}$), we can avoid a lot of communication cost by doing multiple steps within the same call to each slave node.
- Current implementation:
  - Using R and Rmpi, divide up the calculation into blocks, a la Scalapack, and for each block do multi-threaded computations (in R) based on threaded MKL.
  - Joint work with Tina Zhuo (Georgia Tech), Prabhat (LBL), Cari Kaufman (UC Berkeley)

Markov random field (MRF) models

- MRF models extend the idea of a Markov chain to two dimensions.
- Value at one location are conditionally independent of the other values, given a small number of 'neighbors'.
  - Note that MRF models are used for data associated with areas rather than points.
- For a Gaussian model, this gives us

$$g_i | g_{j \neq i} \sim \mathcal{N}\left(\sum_{j \in N(i)} w_j g_j, \frac{1}{\kappa \cdot \#N(i)} \right)$$

MRF models and sparse matrices

- The Hammersley-Clifford theorem tells us that this conditional specification also gives a legitimate joint distribution,

$$\mathbf{g} \sim \mathcal{N}(\mathbf{0}, (\kappa \mathbf{Q})^{-})$$

  where $\mathbf{Q}$ is a sparse precision, i.e., inverse covariance, matrix.
- There are non-negative values on the diagonal, representing the number of neighbors for that row's location.
- There are non-zero values on the off-diagonals for i, j pairs that are neighbors.

Possible MRF models

Standard CAR

$$\begin{array}{ccc} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{array}$$

Thin plate spline MRF approximation

$$\begin{array}{ccccc} & & 1 & & \\ & 2 & -8 & 2 & \\ 1 & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{array}$$

- The standard CAR model puts a weight of one on the cardinal neighbors (left).
  - The limit of a standard CAR model as the grid becomes finer is the de Wijs process: 2D Brownian motion, which is not differentiable (Besag and Mondal 2005)
- The neighborhood structure on the right gives us a MRF that approximates a thin plate spline (Rue and Held 2005).
  - This MRF model gives smoother processes (albeit on a grid), in concordance with the thin plate spline being differentiable.

Thin plate spline

- A thin plate spline is the function that results from minimizing

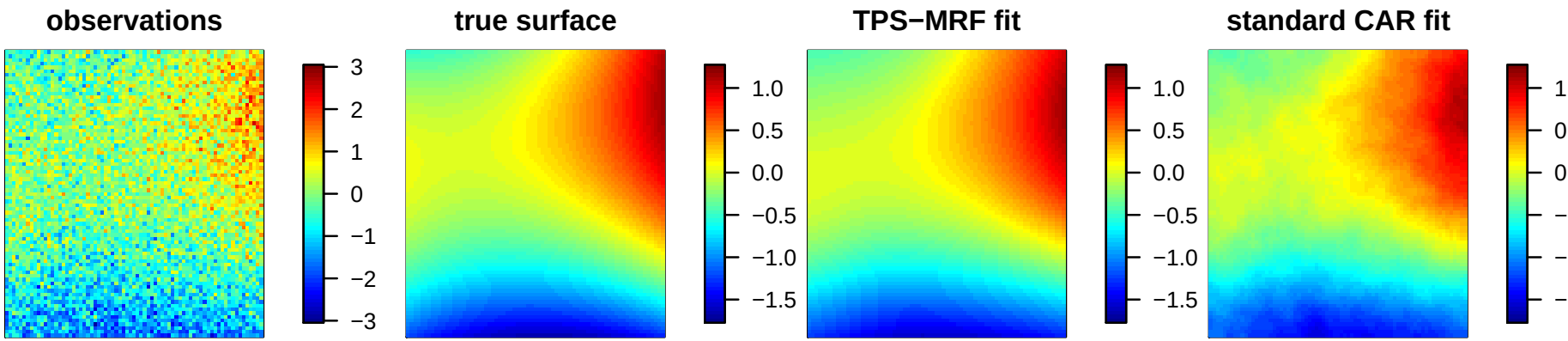
$$\sum_{i=1}^{n}(y_i - g(s_i))^2 + \lambda J(g)$$

where $J(g)$ penalizes 'wiggliness' in the surface.

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

- The MRF approximates a thin plate spline by deriving weights based on a discrete approximation to this penalty.
  - The result is higher order neighbors and oscillating weights.

MRF models: surface predictions



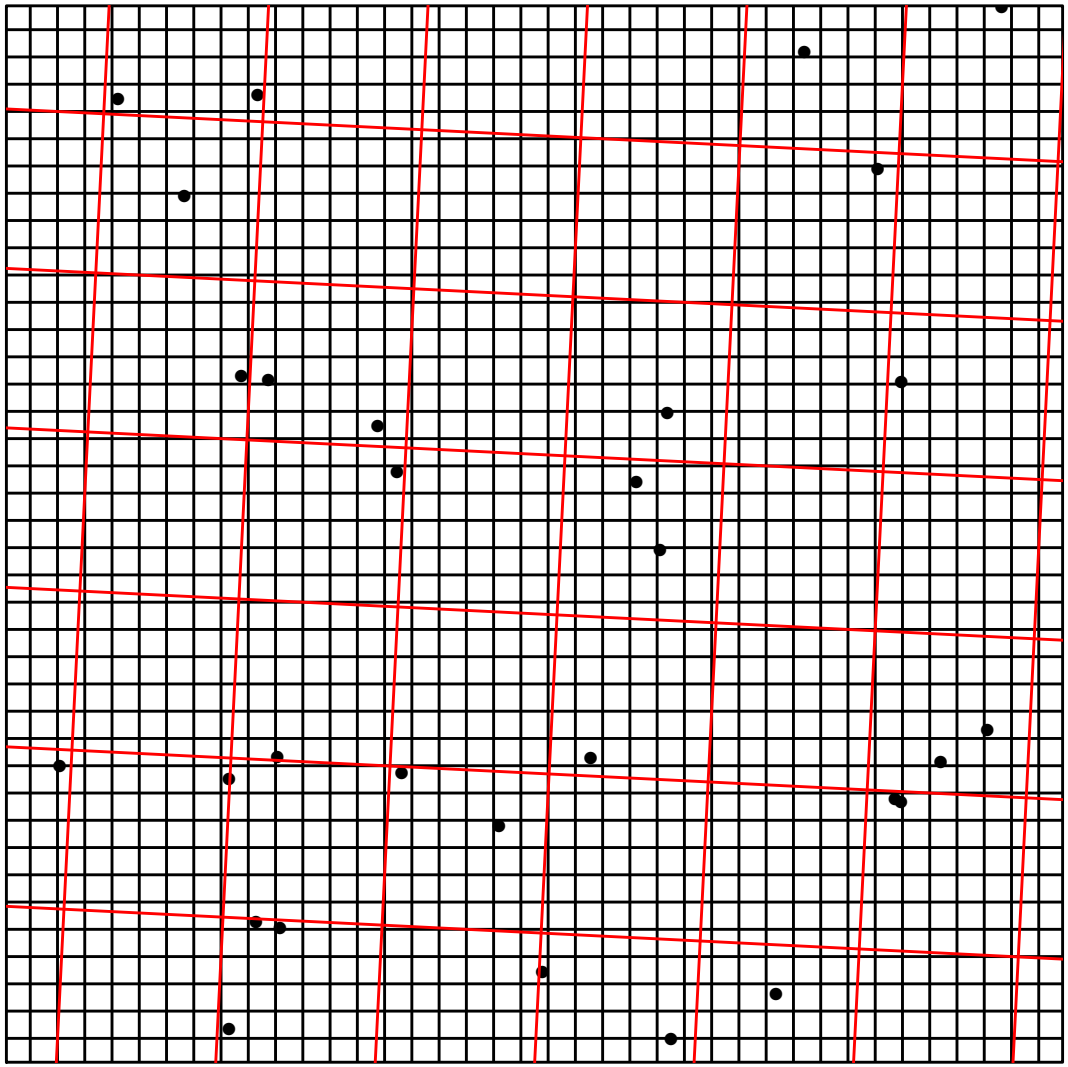
Using MRFs

- In many environmental applications, we will have some point-level data, and we will have areally-aggregated data (e.g., if the 'data' are computer code output or remote sensing).
- The core approach in my recent work is to represent the latent spatial process on a fine regular grid,

$$\{\mathbf{Y}, \mathbf{A}\} \sim \mathcal{N}(\mathbf{Pg}, \sigma^2\mathbf{I})$$
$$\mathbf{g} \sim \mathcal{N}(\mathbf{0}, (\kappa\mathbf{Q})^{-})$$

- Each point observations can be taken to relate to the g_i for the grid cell the point falls in, $\mathbf{P}_i^\top \mathbf{g}$ where $\mathbf{P}_i$ picks out the correct cell.
- Areal data can be taken to relate to $\mathbf{P}_i^\top \mathbf{g}$, a weighted average of the overlapped grid cells, approximating the integral based on the piecewise constant representation, $\mathbf{g}$.
- The approach handles spatially-misaligned datasets in a coherent fashion, tackling the regridding problem.

MRFs and spatial misalignment



Spatio-temporal extension

We can extend the spatial model by assuming an autoregressive structure in time:

$$\mathbf{g}_t \sim \mathcal{N}(\mathbf{g}_\mu + \rho(\mathbf{g}_{t-1} - \mathbf{g}_\mu), \kappa\mathbf{Q})$$

Some analytic manipulations give us the distribution of $\mathbf{g} = \{\mathbf{g}_1, \ldots, \mathbf{g}_T\}$:

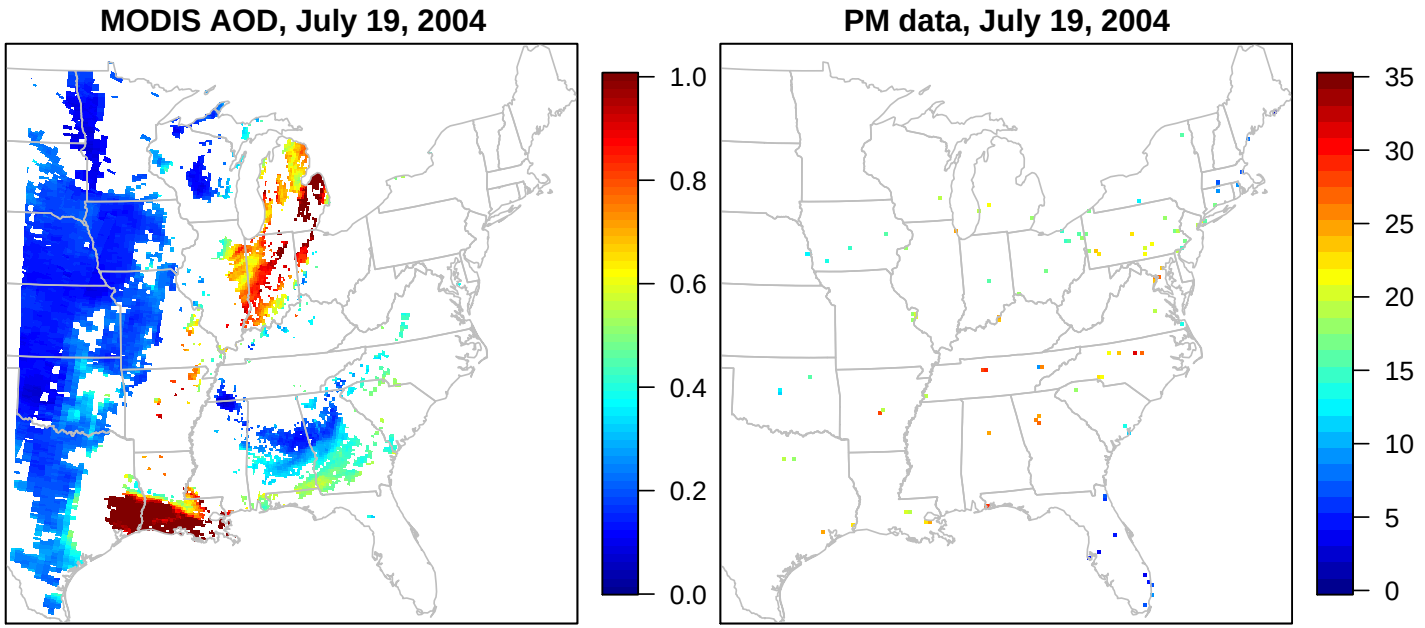
$$\mathbf{g} \sim \mathcal{N}(\mathbf{0}, (\mathbf{H}(\kappa) \otimes \mathbf{Q})^-)$$

where $\mathbf{H}$ is T by T.

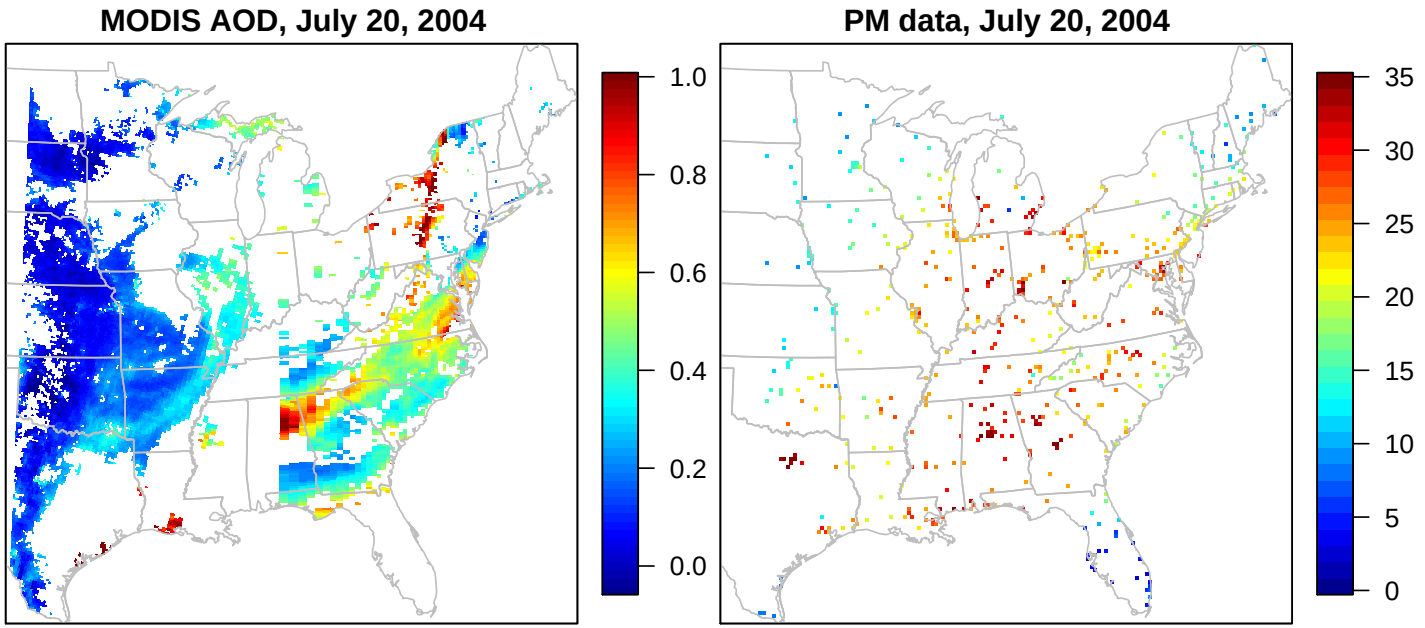
Particulate matter (PM)

- Particulate matter (soot) is produced from emissions from vehicles, power plants, and other sources.
- It is one of the key air pollutants regulated by US EPA, with a huge amount of scientific work focused on estimating the short-term and long-term health effects of PM.
- Typically health analysis has relied on a network of about 1000 monitors throughout the US.
- Ideally, we could get additional information from remote sensing (satellite aerosol optical depth – AOD) or atmospheric chemistry modeling (CMAQ).

Combining information



Combining information



Challenges of proxy information

- Systematic spatial (and temporal) discrepancy between proxy and truth
  - White noise error structure often implausible
  - This impacts predictions, prediction uncertainty, and assessment of proxy usefulness
  - Ignoring the discrepancy leads to overinterpreting patterns in the proxy
  - Proxy may not directly quantify the process of interest, hence 'discrepancy' rather than 'error' or 'bias'; e.g. AOD is a vertical column metric.
- Spatial misalignment of gridded proxy information and point-level observations
- Proxy datasets are usually very large
  - Working with standard Gaussian processes is infeasible

Flexible spatial discrepancy modeling

- A hierarchical Bayesian model:

$$\begin{aligned}
\mathbf{Y} &\sim \mathcal{N}(\mathbf{X}_y\boldsymbol{\beta}_y + \mathbf{P}_y\mathbf{g}, \sigma_y^2\mathbf{I}) \\
\mathbf{A} &\sim \mathcal{N}(\mathbf{P}_A\mathbf{D} + \beta_1\mathbf{P}_A\mathbf{g}, \sigma_a^2\mathbf{I}) \\
\mathbf{g} &\sim \text{MRF}(\mathbf{X}_g\boldsymbol{\beta}_g, \kappa_g\mathbf{Q}) \\
\mathbf{D} &\sim \text{MRF}(\mathbf{X}_D\boldsymbol{\beta}_D, \kappa_D\mathbf{Q})
\end{aligned}$$

- Latent processes, $g(\cdot)$ and $D(\cdot)$, are represented on a fine grid.
- We can explore the relationship of the proxy and gold standard through analysis of spatial variation in $\mathbf{D}$.
- $\mathbf{X}_y\boldsymbol{\beta}_y$ involves regression terms that explain sub-grid scale variation in the point measurements, while $\mathbf{X}_g\boldsymbol{\beta}_g$ and $\mathbf{X}_D\boldsymbol{\beta}_D$ are regression effects on the grid-scale process and the discrepancy term, respectively.

Markov chain Monte Carlo

- MCMC involves setting up a Markov chain on $\boldsymbol{\theta}$ whose stationary (long-run) distribution is the posterior distribution,

$$P(\boldsymbol{\theta}|\mathbf{Y} = \mathbf{y}, \mathbf{A} = \mathbf{a})$$

  where $\boldsymbol{\theta} = \{\mathbf{g}, \mathbf{D}, \ldots\}$ are the unknowns, including quantities you want to predict, and $\mathbf{y}$ are the data values you have.
- We need to run the MCMC for many iterations (often 10000s, 100000s)
- Running the MCMC involves calculating various quantities related to the posterior distribution above, which involves computations with the covariance or inverse covariance matrix of the sort we have discussed.

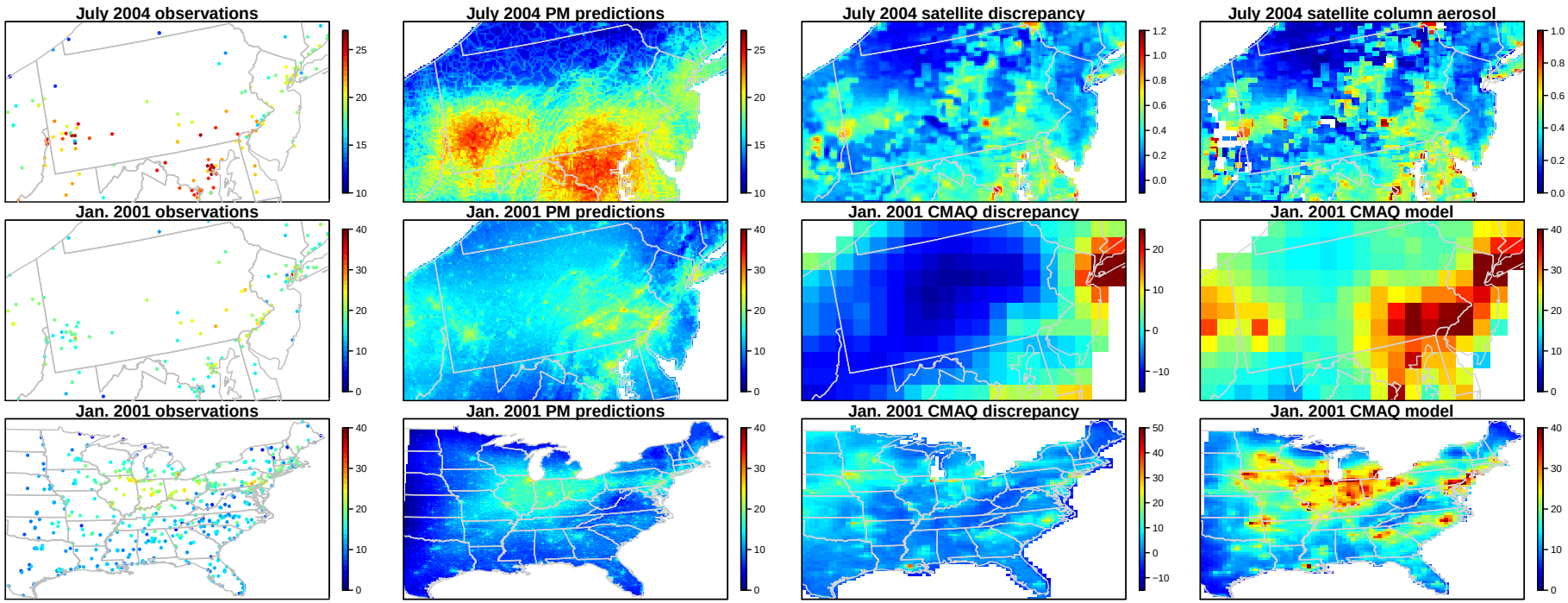
Exploiting sparsity

- If possible, we integrate over the those components of θ that we can do analytically.
- Then in marginal posterior computations, exploit the sparse structure appropriately.

$$P(\theta_{\text{reduced}}|\mathbf{A},\mathbf{Y}) \propto |\mathbf{\Lambda}|^{-\frac{1}{2}}|\mathbf{V}_Y|^{-\frac{1}{2}}|\mathbf{\Sigma}_A|^{-\frac{1}{2}}|\mathbf{V}_b|^{\frac{1}{2}}P(\theta)\cdot$$
$$\exp(-\tfrac{1}{2}(\mathbf{Y}^T\mathbf{V}_Y^{-1}\mathbf{Y}+\mathbf{A}^T\mathbf{\Sigma}_A^{-1}\mathbf{A}-\mathbf{M}_b^T\mathbf{V}_b^{-1}\mathbf{M}_b))$$
$$\mathbf{V}_b = (\mathbf{Z}_Y^T\mathbf{V}_Y^{-1}\mathbf{Z}_Y+\mathbf{Z}_A^T\mathbf{\Sigma}_A^{-1}\mathbf{Z}_A+\mathbf{\Lambda}^{-1})^{-1}$$
$$\mathbf{\Sigma}_A^{-1} = \mathbf{V}_A^{-1}-\mathbf{V}_A^{-1}\mathbf{P}\mathbf{V}_D\mathbf{P}^T\mathbf{V}_A^{-1}$$
$$\mathbf{V}_D = (\mathbf{P}^T\mathbf{V}_A^{-1}\mathbf{P}+\kappa\mathbf{Q})^{-1}$$

Predicted PM

Y **PM = g** **D** **A**



Results

- Satellite AOD:
  - The model fitting suggests there is little common spatial pattern to PM and AOD observations.
    - The discrepancy term, $D(\cdot)$, varies at both small and large scales.
  - As a result the model discounts AOD in predicting PM.
- Atmospheric Chemistry Model (CMAQ):
  - More apparent relationship between CMAQ output and PM.
    - The discrepancy term also varies at small and large scales, but more of the variation in the proxy appears to be signal than for AOD.
  - Statistical model still heavily discounts the proxy.

Cross-validation predictive ability, R^2 (RMSPE)

Time scale	Proxy?	mid-Atlantic, 2004, MODIS AOD	mid-Atlantic, 2001, CMAQ, space-time	mid-Atlantic, 2001, CMAQ, spatial models	eastern U.S., 2001, CMAQ
Monthly[1]	w/ proxy	0.806 (1.80)	0.640 (2.60)	0.755 (2.14)	0.827 (1.71)
	no proxy	0.808 (1.79)	0.686 (2.42)	0.777 (2.04)	0.826 (1.72)
	as regr.				0.849 (1.60)
Yearly[2]	w/ proxy	0.668 (1.00)[3]	<0[4] (1.97)[3]	0.503 (1.32)[3]	0.800 (1.21)
	no proxy	0.650 (1.03)[3]	0.169 (1.70)[3]	0.584 (1.20)[3]	0.835 (1.09)
	as regr.				0.849 (1.05)

[1] Including monthly averages based on at least five daily observations.
[2]Including yearly averages (averages of monthly values) based on at least nine months with at least five daily observations.
[3] Excludes one site outside Pittsburgh just downwind of a major industrial facility.
[4]Squared correlation of held-out data and predictions is 0.473, but observations vs. predictions are not centered on the one to one line, so error sum of squares exceeds total sum of squares.

Conclusions: computation

- Certain statistical representations of spatial and spatio-temporal processes improve computational efficiency.
  - Markov random field models lead to sparse matrix calculations.

Conclusions: discrepancy

- In many problems involving computer codes, characterizing the discrepancy between the code and the truth is important.
  - White noise error, while convenient, is generally not appropriate.
  - Distinguishing correlated noise from correlated signal is difficult and likely sensitive to modeling assumptions.
    - Case study: Is there useful information in the proxies that the current model structure is not exploiting?
  - When we don't have gold standard data, such as in uncertainty quantification for climate model projections, prior assumptions about the correlation structure of the discrepancy will be critical.
    - What can be said about uncertainty in climate projections at regional scales?