Statistical Methods for Combining Measurements and Models, with Application to Predicting Particulate Matter

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

February 11, 2011

Modern Regression

Modern statistical regression involves a vast array of extensions to the usual

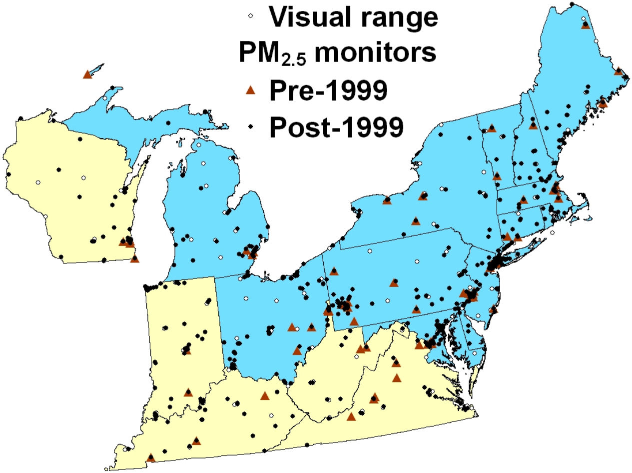
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

model.

In particular, $\mathbf{X}\beta$ is often replaced by

- Complicated interactions of the X variables (e.g., regression trees),
- Known functional forms of the X variable(s), $f(x, \theta)$ (nonlinear regression),
- Unknown functions of the X variable(s) (nonparametric regression),
- Terms to capture spatial, temporal and spatio-temporal structure in the outcome, and
- Hierarchical structure that cluster observations into groups of similar observations.

PM Example



Spatio-temporal Statistical Modeling

- A spatio-temporal statistical model (Yanosky et al. 2009, Environmental Health Perspectives; Paciorek et al. 2009, Annals of Applied Statistics):
  - First stage for monthly spatial variation:

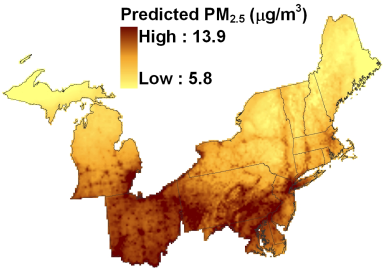
$$\log \mathrm{PM}_{it} = \mu_i + \mathbf{X}_{it}^{st}\boldsymbol{\beta}^{st} + g_t(s_i) + \epsilon_{it}$$

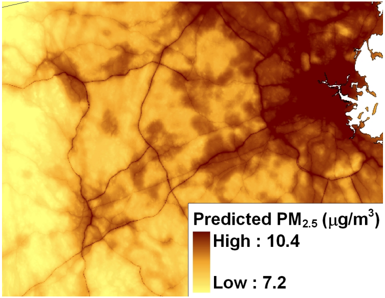
  - Second stage model for spatial-only effects:

$$\mu_i = \mathbf{X}_i^{s}\boldsymbol{\beta}^{s} + g_\mu(s_i) + \delta_i$$

- X^{st}'s are spatio-temporally-varying predictors (e.g., meteorology), while X^{s}'s vary only spatially (e.g., population density, road effects).
- Spatio-temporal ($g_t(s)$) and spatial ($g_\mu(s)$) terms account for additional spatio-temporal structure.

PM Predictions (Ambient Exposure Estimates)





$PM_{2.5}$ predictions: northeast US (left) and greater Boston (right)

Predictions from the model at individual residences can then be used as estimated exposures in health analyses.

Advantages and Disadvantages of Empiricism

Advantages:

- Predictions are empirically driven (calibrated and 'validated')
- Adding additional explanatory variables and spatio-temporal structure to the model is easily done and readily evaluated in terms of whether predictions are improved.
- Prediction error is readily quantifiable.

Disadvantages:

- Observations are expensive
- Data-driven estimation of effects of important variables is difficult because of data requirements and complicated relationships.
  - Ex. interaction of wind direction, wind speed, source location, receptor location

An 'Easy' Use of a Computer Model

The CALINE model represents air pollution from line sources based on Gaussian diffusion.

One might simply make predictions from the computer model and include as an explanatory variable in a statistical model.

- Advantages:
  - Simplicity
  - Black box representation of complicated relationships
- Disadvantages:
  - Computational cost of running the model for 10,000s of prediction locations, potentially for many time steps.
  - Input data requirements, including emissions information
  - Model errors; in particular error from model 'extrapolation'

Hierarchical (Latent Variable) Modeling

- A basic hierarchical spatial model

$$
\begin{aligned}
Y_i &\sim \mathcal{N}(g_i, \sigma^2) && \text{measurement model} \\
\mathbf{g} = (g_1, \ldots, g_n) &\sim \mathcal{N}(\mu, \mathbf{\Sigma}(\rho)) && \text{process model} \\
\{\sigma^2, \mu, \rho\} &\sim P && \text{parameter model (Bayesian)}
\end{aligned}
$$

  Here $\mathbf{g}$ is the unknown latent spatial field (the state of the system). The structure of $\mathbf{\Sigma}$ determines the behavior of the spatial field, $\mathbf{g}$.
- When one fits the model, estimating $\mathbf{g}$, the result is a tradeoff between fidelity to the data and constraints imposed by the process representation and its covariance ($\mathbf{\Sigma}$).
- When statistical distributions are used to characterize all the unknowns in the model, including ρ, these are Bayesian hierarchical models.

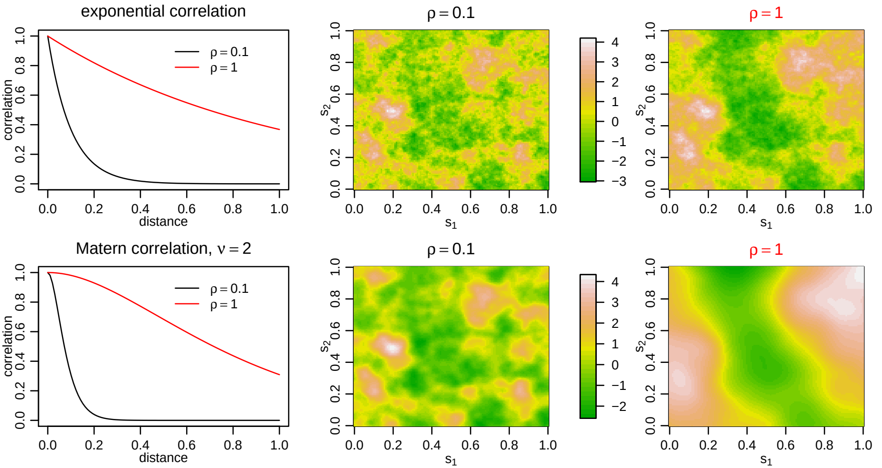
Two Statistical Models for Spatial Fields

$$\mathbf{g} \sim \mathcal{N}(\mu, \mathbf{\Sigma}(\rho)) \quad \text{process}$$

1. Gaussian process models for g.
   1. The elements of $\mathbf{\Sigma}(\rho)$ are the pairwise covariances between g_i and g_j (i.e., for any pair of locations).
   2. The covariance is assumed to be a known function of (a) the distance between a pair of locations and (b) θ.
   3. θ controls how 'quickly' the field varies spatial (i.e., frequency or 'wiggliness')
2. Markov random field models for g
   1. g is evaluated only for areas and not individual locations
   2. $\mathbf{\Sigma}(\rho)$ is actually represented as $\kappa Q = \Sigma^{-1}$
   3. Q represents what pairs of locations are 'neighbors' and how to weight neighbors to make a prediction at a location of interest

Similar ideas are used for spatio-temporal fields.

Gaussian process models illustrated



Misalignment

Data may be collected at differing spatial and temporal scales.

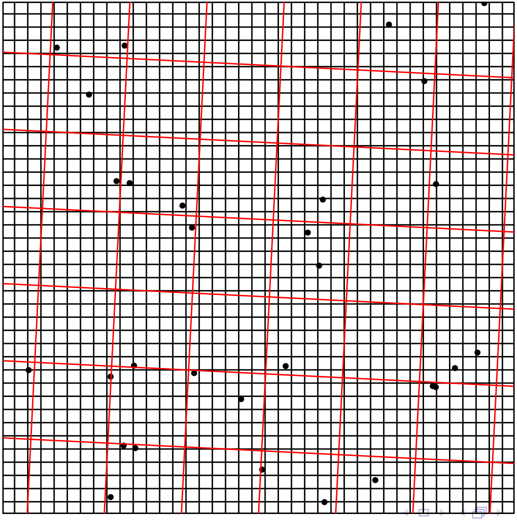
- Proxy information (computer model output, remote sensing output) is often aggregated spatially.
  - Conceptually:

$$A_i \sim \mathcal{N}\left(\int_{R_i} g(s)ds, \sigma^2\right)$$

  - Numerically: suppose g is represented on a very fine spatial grid

$$A_i \sim \mathcal{N}(\mathbf{K}_i^\top \mathbf{g}, \sigma^2)$$

Misalignment visualized



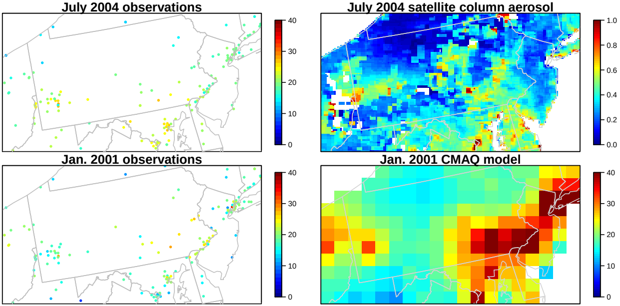
Treating Proxy Information as Data

- Let's consider both observations, Y, and proxy information, A:

$$Y_i \sim \mathcal{N}(\mathbf{K}_{Y,i}^{\top}\mathbf{g}, \sigma^2)$$
$$A_j \sim \mathcal{N}(\mathbf{K}_{A,j}^{\top}\mathbf{g}, \tau^2)$$

- Why treat the proxy as data rather than explanatory variable?
  - Remote sensing information might be considered 'measurements'.
  - It allows us to more readily deal with missingness.
  - We want to relate the information to the latent field and account for discrepancy, with a potential model 'validation' analysis.

Discrepancy



A modified model:

$$Y_i \sim \mathrm{N}(\mathbf{K}_{Y,i}^\top \mathbf{g}, \sigma^2)$$
$$A_j \sim \mathrm{N}(\mathbf{K}_{A,j}^\top \mathbf{D} + \mathbf{K}_{A,j}^\top \mathbf{g}, \tau^2)$$

Flexible Spatial Discrepancy Modeling

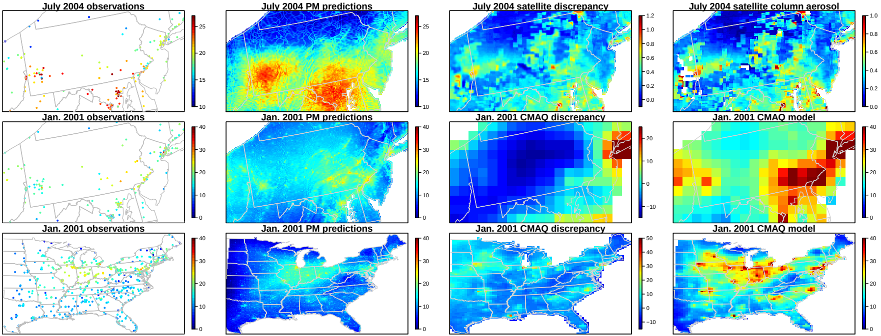
- Consider additive bias as a spatial discrepancy process, $D(\cdot)$:

$$\begin{aligned} Y_i &\sim \mathcal{N}(\mathbf{K}_{y,i}^\top \mathbf{g}, \sigma_y^2) \\ A_j &\sim \mathcal{N}(\mathbf{K}_{A,j}^\top \mathbf{D} + \beta_1 \mathbf{K}_{A,j}^\top \mathbf{g}, \sigma_a^2) \\ \mathbf{g} &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_g(\rho_g)) \\ \mathbf{D} &\sim \mathcal{N}(\mathbf{Z}\boldsymbol{\alpha}, \boldsymbol{\Sigma}_D(\rho_D)) \end{aligned}$$

- X and Z are predictor variables for the pollution process and the discrepancy term, respectively.
- We can explore the relationship of the proxy and gold standard through analysis of the spatial scales of $D(\cdot)$.

Predicted PM

Y | $\mathbf{PM = g}$ | **D** | **A**



Results

- Satellite AOD:
    - The model fitting suggests there is little common spatial pattern to PM and AOD observations.
        - The discrepancy term, D, varies at both small and large scales.
    - As a result the model discounts AOD in predicting PM.
- Atmospheric Chemistry Model (CMAQ):
    - Stronger 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.
- Paciorek and Liu (2011) and Paciorek (submitted).

Difficulties

- Heavy data requirements.
- Discrepancy is highly-structured, complex, and may be hard to represent stochastically.
  - Misspecification of the proxy component of the model may cause us to downweight the proxy relative to the observations.
- At the same time, it is unknown, so a statistical treatment is natural.

Parameter estimation

- Consider observations, $Y_1, \ldots, Y_n$ corresponding to evaluation of the computer model at input vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
- If $f(\mathbf{x}; \boldsymbol{\theta})$ is the output of the computer model, optimize a cost function with respect to model parameters, $\boldsymbol{\theta}$, e.g.,

$$\text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i; \boldsymbol{\theta}))^2$$

- This is equivalent to assuming that the observation error is $\mathcal{N}(0, \sigma^2)$.
- In some sense this is now an empirical nonlinear regression model.
  - How are we to interpret the parameter values?

Discrepancy

Of course the difference between an observation and the model output is more than just noisiness in the measurements.

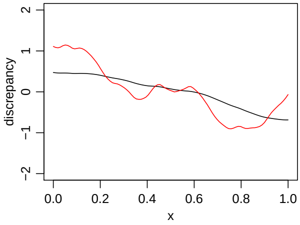
1. Observations and model output may be at different aggregations (i.e., misaligned)
2. There will be systematic discrepancy between the model output and the truth.
   - This discrepancy might be thought to relate to certain factors. Perhaps build in a regression relationship.
   - 'Systematic' implies some 'correlation' in some dimension of the input space.

Gaussian processes in variable space

- A basic discrepancy model

$$Y_i \sim \mathcal{N}(f(x_i;\theta) + D(x_i;\rho), \sigma^2)$$

  - We might assume $D(x_i)$ and $D(x_j)$ are similar for x_i similar to x_j.
- We can represent this assumption as another Gaussian process model, $\mathbf{D} \sim \mathcal{N}(0, \mathbf{\Sigma}_D(\rho))$.



Emulators

- Suppose the code is computationally demanding and can only be run m times at m values of the input vector, $\mathbf{x}_1, \ldots, \mathbf{x}_m$.
- Gaussian processes are also commonly used in these scenarios as a low-dimensional approximation to the computer model.
- Let $\mathbf{f} = f(\mathbf{x}_1; \theta), \ldots, f(\mathbf{x}_m; \theta)$.
- If we represent f as a Gaussian process in the space of $\mathbf{x}$,

$$\mathbf{f} \sim \mathrm{N}(\mu, \boldsymbol{\Sigma}_f(\rho_f)),$$

  then we can predict (interpolate) the output of the code at any new input vector, $\mathbf{x}^*$.
    - Here ρ controls how 'quickly' the model output changes as one changes the input values.

The State-of-the-statistical Art in Computer Models

- Goals: calibration, uncertainty quantification, and prediction with computationally-demanding computer models.
- Approach: build a single statistical model

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + D(\mathbf{x}_i, \rho_D) + \epsilon$$

  that

  1. Estimates (tunes/calibrates) unknown parameters in the computer model.
  2. Includes a discrepancy term for which we estimate parameters that help quantify model inadequacy.
  3. Includes an emulation component to limit the number of code runs needed.

- The approach is still somewhat unsatisfying in that characterizing discrepancy is data-needy. Also, it doesn't help much with uncertainty quantification in cases where data are sparse or unavailable (e.g., climate model projections).

Data Assimilation (DA)

- Computer models often represent systems evolving over time.
- The goal of DA to use available data reflecting the system state to tweak the model toward reality, while relying on the model to capture the core dynamics in a way that sparse data could not.
- At the core, we have another hierarchical statistical model, a 'state-space' model

$$\mathbf{Y}_t \sim \mathcal{N}(\mathbf{K}_t \mathbf{g}_t, \sigma^2 \mathbf{I})$$
$$\mathbf{g}_t \sim \mathcal{N}(f(\mathbf{g}_{t-1}, \mathbf{x}_t), \boldsymbol{\Sigma}_g)$$

Kalman Filter (KF)

$$\mathbf{Y}_t \sim \mathcal{N}(\mathbf{K}_t \mathbf{g}_t, \sigma^2 \mathbf{I})$$
$$\mathbf{g}_t \sim \mathcal{N}(f(\mathbf{g}_{t-1}, \mathbf{x}_t), \mathbf{\Sigma}_g)$$

- The algorithm to update the estimate of the state vector to get $\hat{\mathbf{g}}_t$ (i.e., $E(\mathbf{g}_t|\mathbf{Y}_t)$ and $\text{Var}(\mathbf{g}_t|\mathbf{Y}_t)$ is the Kalman Filter (KF), but is really just calculation of the posterior distribution in a Bayesian hierarchical model.
- Note that $\mathbf{\Sigma}_g$ here represents model error/uncertainty in the state vector, treated stochastically.
- DA is often done using the ensemble KF, propagating a sample, $\{\mathbf{g}_t^i\}$ based on the model (f), and using the empirical covariance of $\{\mathbf{g}_t^i\}$ to estimate $\mathbf{\Sigma}_g$.