Introduction to Spatial Data and Models

Sudipto Banerjee\textsuperscript{1} and Andrew O. Finley\textsuperscript{2}

\textsuperscript{1}Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.
\textsuperscript{2}Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

June 22, 2009

Introduction to Spatial Data and Models

Researchers in diverse areas such as climatology, ecology, environmental health, and real estate marketing are increasingly faced with the task of analyzing data that are:

- highly multivariate, with many important predictors and response variables,
- geographically referenced, and often presented as maps, and
- temporally correlated, as in longitudinal or other time series structures.

⇒ motivates hierarchical modeling and data analysis for complex spatial (and spatiotemporal) data sets.

Type of spatial data

- **point-referenced data**, where $Y(s)$ is a random vector at a location $s \in \mathbb{R}^r$, where $s$ varies continuously over $D$, a fixed subset of $\mathbb{R}^r$ that contains an $r$-dimensional rectangle of positive volume;

- **areal data**, where $D$ is again a fixed subset (of regular or irregular shape), but now partitioned into a finite number of areal units with well-defined boundaries;

- **point pattern data**, where now $D$ is itself random; its index set gives the locations of random events that are the spatial point pattern. $Y(s)$ itself can simply equal 1 for all $s \in D$ (indicating occurrence of the event), or possibly give some additional covariate information (producing a marked point pattern process).

Exploration of spatial data

- **First Law of Geography**
  - **Mean**: first-order behavior
  - **Error**: second-order behavior (covariance function)
  - EDA tools examine both first and second order behavior
  - Preliminary displays: Simple locations to surface displays

First Law of Geography

Scallops Sites
Deterministic surface interpolation

- Spatial surface observed at finite set of locations
  \[ \mathcal{F} = \{ s_1, s_2, \ldots, s_n \} \]
- Tessellate the spatial domain (usually with data locations as vertices)
- Fit an interpolating polynomial:
  \[ f(s) = \sum_i w_i(\mathcal{F}; s) f(s_i) \]
- "Interpolate" by reading off \( f(s_0) \).
- Issues:
  - Sensitivity to tesselations
  - Choices of multivariate interpolators
  - Numerical error analysis

Drop-line scatter plot

Surface plot

Image contour plot

Locations form patterns
Point-level modelling refers to modelling of spatial data collected at locations referenced by coordinates (e.g., lat-long, Easting-Northing).

**Fundamental concept:** Data from a spatial process \{Y(s) : s ∈ D\}, where D is a fixed subset in Euclidean space.

**Example:** \(Y(s)\) is a pollutant level at site \(s\)

**Conceptually:** Pollutant level exists at all possible sites

**Practically:** Data will be a partial realization of a spatial process – observed at \{s_1, \ldots, s_n\}

**Statistical objectives:** Inference about the process \(Y(s)\); predict at new locations.

Suppose our spatial process has a mean, \(\mu(s) = E(Y(s))\), and that the variance of \(Y(s)\) exists for all \(s \in D\).

- **Strong stationarity:** If for any given set of sites, and any displacement \(h\), the distribution of \((Y(s_1), \ldots, Y(s_n))\) is the same as \((Y(s_1 + h), \ldots, Y(s_n + h))\).
- **Weak stationarity:** Constant mean \(\mu(s) = \mu\), and \(\text{Cov}(Y(s), Y(s + h)) = C(h)\): the covariance depends only upon the displacement (or separation) vector.
- **Strong stationarity implies weak stationarity**
- **The process is Gaussian if \(Y = (Y(s_1), \ldots, Y(s_n))\) has a multivariate normal distribution.**
- For Gaussian processes, strong and weak stationarity are equivalent.

**Variograms**

- Suppose we assume \(E[Y(s + h) - Y(s)] = 0\) and define
  \[ E[(Y(s + h) - Y(s))^2] = \text{Var}(Y(s + h) - Y(s)) = 2\gamma(h). \]

  This is sensible if the left hand side depends only upon \(h\). Then we say the process is **intrinsically stationary.**

- \(\gamma(h)\) is called the **semivariogram** and \(2\gamma(h)\) is called the **variogram.**

Note that intrinsic stationarity defines only the first and second moments of the differences \(Y(s + h) - Y(s)\). It says nothing about the joint distribution of a collection of variables \(Y(s_1), \ldots, Y(s_n)\), and thus provides no likelihood.

**Intrinsic Stationarity and Ergodicity**

- **Relationship between \(\gamma(h)\) and \(C(h)\):**
  \[
  2\gamma(h) = \text{Var}(Y(s + h)) + \text{Var}(Y(s)) - 2\text{Cov}(Y(s + h), Y(s)) = C(0) + C(0) - 2C(h) = 2[C(0) - C(h)].
  \]

- **Easy to recover \(\gamma\) from \(C\).** The converse needs the additional assumption of **ergodicity:** \(\lim_{|u| \to \infty} C(u) = 0\).
- So \(\lim_{|u| \to \infty} \gamma(u) = C(0)\), and we can recover \(C\) from \(\gamma\) as long as this limit exists.

  \[
  C(h) = \lim_{|u| \to \infty} \gamma(u) - \gamma(h).
  \]
Introduction to spatial data and models

Isotropy

When $\gamma(h)$ or $C(h)$ depends upon the separation vector only through the distance $|h|$, we say that the process is isotropic. In that case, we write $\gamma(|h|)$ or $C(|h|)$. Otherwise we say that the process is anisotropic.

If the process is intrinsically stationary and isotropic, it is also called homogeneous.

Isotropic processes are popular because of their simplicity, interpretability, and because a number of relatively simple parametric forms are available as candidates for $C$ (and $\gamma$). Denoting $|h|$ by $t$ for notational simplicity, the next two tables provide a few examples...

### Examples: Spherical Variogram

$$
\gamma(t) = \begin{cases}
  \tau^2 + \sigma^2 & \text{if } t \geq 1/\phi \\
  0 & \text{if } 0 < t \leq 1/\phi \\
  \tau^2 & \text{if } t = 0
\end{cases}
$$

- While $\gamma(0) = 0$ by definition, $\gamma(0^+) \equiv \lim_{t \to 0^+} \gamma(t) = \tau^2$; this quantity is the nugget.
- $\lim_{t \to \infty} \gamma(t) = \tau^2 + \sigma^2$; this asymptotic value of the semivariogram is called the sill. (The sill minus the nugget, $\sigma^2$ in this case, is called the partial sill.)
- Finally, the value $t = 1/\phi$ at which $\gamma(t)$ first reaches its ultimate level (the sill) is called the range, $R \equiv 1/\phi$.

### Some common isotropic variograms

<table>
<thead>
<tr>
<th>Model</th>
<th>Covariance function, $C(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$C(t)$ does not exist</td>
</tr>
</tbody>
</table>
| Spherical  | $C(t) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t \geq 1/\phi \\
|            | \sigma^2 \left[ 1 - \frac{3}{2} \phi t + \frac{3}{2} (\phi t)^2 \right] & \text{if } 0 < t \leq 1/\phi \\
|            | 0 & \text{otherwise} \end{cases}$                                                            |
| Exponential| $C(t) = \begin{cases} \sigma^2 \exp(-\phi t) & \text{if } t > 0 \\
|            | \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$                                            |
| Powered    | $C(t) = \begin{cases} \sigma^2 \exp(-|\phi t|) & \text{if } t > 0 \\
| exponential| \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$                                            |
| Matérn     | $C(t) = \begin{cases} \sigma^2 (1 + \phi t) \exp(-\phi t) & \text{if } t > 0 \\
| at $\nu = 3/2$ | \tau^2 + \sigma^2 & \text{otherwise} \end{cases}$                                            |

Notes on exponential model

$$
C(t) = \begin{cases} 
  \tau^2 + \sigma^2 & \text{if } t = 0 \\
  \sigma^2 \exp(-\phi t) & \text{if } t > 0
\end{cases}
$$

- We define the effective range, $t_0$, as the distance at which this correlation has dropped to only 0.05. Setting $\exp(-\phi t_0)$ equal to this value we obtain $t_0 \approx 3/\phi$, since $\log(0.05) \approx -3$.
- Finally, the form of $C(t)$ shows why the nugget $\tau^2$ is often viewed as a "nonspatial effect variance," and the partial sill ($\sigma^2$) is viewed as a "spatial effect variance."
The Matèrn Correlation Function

- Much of statistical modelling is carried out through correlation functions rather than variograms
- The Matèrn is a very versatile family:
  \[ C(t) = \begin{cases} \frac{\sigma^2}{2} \left( 2 \sqrt{\nu} t \phi \right)^{-\nu} K_\nu \left( 2 \sqrt{\nu} t \phi \right) & \text{if } t > 0 \\ \sigma^2 & \text{if } t = 0 \end{cases} \]

  \( K_\nu \) is the modified Bessel function of order \( \nu \) (computationally tractable)
- \( \nu \) is a smoothness parameter (a fractal) controlling process smoothness

How do we select a variogram? Can the data really distinguish between variograms?

Empirical Variogram:

\[ \gamma(t) = \frac{1}{2|N(t)|} \sum_{s_i, s_j \in N(t)} (Y(s_i) - Y(s_j))^2 \]

where \( N(t) \) is the number of points such that \( \|s_i - s_j\| = t \) and \( |N(t)| \) is the number of points in \( N(t) \).

Grid up the \( t \) space into intervals \( I_1 = (0, t_1), I_2 = (t_1, t_2), \) and so forth, up to \( I_K = (t_{K-1}, t_K) \). Representing \( t \) values in each interval by its midpoint, we define:

\[ N(t_k) = \{ (s_i, s_j) : \|s_i - s_j\| \in I_k \}, k = 1, \ldots, K. \]
Principles of Bayesian Inference

Sudipto Banerjee1 and Andrew O. Finley2

1 Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.
2 Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

June 22, 2009

Bayesian principles

- The key to Bayesian inference is “learning” or “updating” of prior beliefs. Thus, posterior information ≥ prior information.

- Is the classical approach wrong? That may be a controversial statement, but it certainly is fair to say that the classical approach is limited in scope.

- The Bayesian approach expands the class of models and easily handles:
  - repeated measures
  - unbalanced or missing data
  - nonhomogenous variances
  - multivariate data
  - and many other settings that are precluded (or much more complicated) in classical settings.

- Calculations (numerical and algebraic) are usually required only up to a proportionally constant. We, therefore, write the posterior as:
  \[ p(\theta | y, \lambda) \propto p(\theta | \lambda) \times f(y | \theta). \]

- If \( \lambda \) are known/fixed, then the above represents the desired posterior. If, however, \( \lambda \) are unknown, we assign a prior, \( p(\lambda) \), and seek:
  \[ p(\theta, \lambda | y) \propto p(\lambda)p(\theta | \lambda)f(y | \theta). \]
  The proportionality constant does not depend upon \( \theta \) or \( \lambda \):
  \[ \frac{1}{p(y)} = \int p(\lambda)p(\theta | \lambda)f(y | \theta) d\lambda d\theta \]

- The above represents a joint posterior from a hierarchical model. The marginal posterior distribution for \( \theta \) is:
  \[ p(\theta | y) = \int p(\lambda)p(\theta | \lambda)f(y | \theta) d\lambda. \]

Basics of Bayesian inference

We start with a model (likelihood) \( f(y | \theta) \) for the observed data \( y = (y_1, \ldots, y_n)' \) given unknown parameters \( \theta \) (perhaps a collection of several parameters).

Add a prior distribution \( p(\theta | \lambda) \), where \( \lambda \) is a vector of hyper-parameters.

The posterior distribution of \( \theta \) is given by:

\[
p(\theta | y, \lambda) = \frac{p(\theta | \lambda) \times f(y | \theta)}{p(y | \lambda)} = \frac{p(\theta | \lambda) \times f(y | \theta)}{\int f(y | \theta)p(\theta | \lambda)d\theta}
\]

We refer to this formula as Bayes Theorem.

Bayesian inference: point estimation

- Point estimation is easy: simply choose an appropriate distribution summary: posterior mean, median or mode.

  - **Mode** sometimes easy to compute (no integration, simply optimization), but often misrepresents the “middle” of the distribution – especially for one-tailed distributions.

  - **Mean**: easy to compute. It has the “opposite effect” of the mode – chases tails.

  - **Median**: probably the best compromise in being robust to tail behaviour although it may be awkward to compute as it needs to solve:

\[
\int_{-\infty}^{\theta_{median}} p(\theta | y) d\theta = \frac{1}{2}
\]
A simple example: Normal data and normal priors

Interpretation: Posterior mean is a weighted mean of prior mean and data.

\[ p(\theta | y) \propto N(\theta | \mu, \tau^2) \times N(y | \bar{y}, \sigma^2) \]

Posterior distribution of \( \theta \)

\[
\begin{align*}
\theta &\sim N(\mu, \tau^2) \\
\theta &\sim N(\mu, \tau^2) \times N(y | \bar{y}, \sigma^2) \\
\theta &\sim N(\theta | \mu, \tau^2) \times N(y | \bar{y}, \sigma^2) \\
\end{align*}
\]

Then clearly \( P(\theta \in (q_L, q_U) | y) = 1 - \alpha \).

This interval is relatively easy to compute and has a direct interpretation: The probability that \( \theta \) lies between \((q_L, q_U)\) is \( 1 - \alpha \). The frequentist interpretation is extremely convoluted.

Bayesian inference: interval estimation

The most popular method of inference in practical Bayesian modelling is interval estimation using credible sets. A 100(1 - \alpha)% credible set \( C \) for \( \theta \) is a set that satisfies:

\[
P(\theta \in C | y) = \int_C p(\theta | y) d\theta \geq 1 - \alpha.
\]

The most popular credible set is the simple equal-tail interval estimate \((q_L, q_U)\) such that:

\[
\int_{q_L}^{q_U} p(\theta | y) d\theta = \frac{\alpha}{2} = \int_{q_L}^{q_U} p(\theta | y) d\theta
\]

Another simple example: The Beta-Binomial model

Example: Let \( Y \) be the number of successes in \( n \) independent trials.

\[
P(Y = y | \theta) = f(y | \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}
\]

Prior: \( p(\theta) = Beta(a, b): \)

\[
p(\theta) \propto \theta^{a-1} (1 - \theta)^{b-1}.
\]

Prior mean: \( \mu = a/(a + b) \); Variance \( ab/((a + b)^2(a + b + 1)) \)

Posterior distribution of \( \theta \)

\[
p(\theta | y) = Beta(a + y, b + n - y)
\]

Sampling-based inference

- We will compute the posterior distribution \( p(\theta | y) \) by drawing samples from it. This replaces numerical integration (quadrature) by "Monte Carlo integration".

- One important advantage: we only need to know \( p(\theta | y) \) up to the proportionality constant.

- Suppose \( \theta = (\theta_1, \theta_2) \) and we know how to sample from the marginal posterior distribution \( p(\theta_2 | y) \) and the conditional distribution \( P(\theta_1 | \theta_2, y) \).

- How do we draw samples from the joint distribution: \( p(\theta_1, \theta_2 | y) \)?

We do this in two stages using composition sampling:

- First draw \( \theta_2^{(j)} \sim p(\theta_2 | y), j = 1, \ldots, M. \)

- Next draw \( \theta_1^{(j)} \sim p(\theta_1 | \theta_2^{(j)}, y) \).

This sampling scheme produces exact samples, \( \{\theta_1^{(1)}, \theta_2^{(1)}\} \) \( \sim p(\theta_1, \theta_2 | y) \) from the posterior distribution \( p(\theta_1, \theta_2 | y) \).

Gelfand and Smith (JASA, 1990) demonstrated automatic marginalization: \( \{\theta_1^{(1)}, \theta_2^{(1)}\} \sim p(\theta_1, \theta_2 | y) \) are samples from \( p(\theta_1 | y) \) and (of course) \( \{\theta_1^{(1)}, \theta_2^{(1)}\} \) \( \sim p(\theta_2 | y) \) are samples from \( p(\theta_2 | y) \).

In effect, composition sampling has performed the following "integration":

\[
p(\theta_1 | y) = \int p(\theta_1 | \theta_2, y) p(\theta_2 | y) d\theta_2.
\]
Bayesian predictions

1. Suppose we want to predict new observations, say \( \tilde{y} \), based upon the observed data \( y \). We will specify a joint probability model \( p(\tilde{y}, y | \theta) \), which defines the conditional predictive distribution:

\[
p(\tilde{y} | y, \theta) = \frac{p(\tilde{y}, y | \theta)}{p(y | \theta)}
\]

2. Bayesian predictions follow from the posterior predictive distribution that averages out the \( \theta \) from the conditional predictive distribution with respect to the posterior:

\[
p(\tilde{y} | y) = \int p(\tilde{y} | y, \theta) p(\theta | y) \, d\theta.
\]

This can be evaluated using composition sampling:

- First obtain: \( \theta^{(j)} \sim p(\theta | y) \), \( j = 1, \ldots, M \)
- For \( j = 1, \ldots, M \) sample \( \tilde{y}^{(j)}(\tilde{y} | y, \theta^{(j)}) \)
- The \( \{\tilde{y}^{(j)}\}_{j=1}^{M} \) are samples from the posterior predictive distribution \( p(\tilde{y} | y) \).

Some remarks on sampling-based inference

- Direct Monte Carlo: Some algorithms (e.g., composition sampling) can generate independent samples exactly from the posterior distribution. In these situations there are NO convergence problems or issues. Sampling is called exact.

- Markov Chain Monte Carlo (MCMC): In general, exact sampling may not be possible/feasible. MCMC is a far more versatile set of algorithms that can be invoked to fit more general models. Note: anywhere where direct Monte Carlo applies, MCMC will provide excellent results too.

- Convergence issues: There is no free lunch! The power of MCMC comes at a cost. The initial samples do not necessarily come from the desired posterior distribution. Rather, they need to converge to the true posterior distribution. Therefore, one needs to assess convergence, discard output before the convergence and retain only post-convergence samples. The time of convergence is called burn-in.

- Diagnosing convergence: Usually a few parallel chains are run from rather different starting points. The sample values are plotted (called trace-plots) for each of the chains. The time for the chains to “mix” together is taken as the time for convergence.

- Good news! All this is automated in WinBUGS. So, as users, we need to only configure how to specify good Bayesian models and implement them in WinBUGS.
Marginal and conditional distributions

With a flat prior on $\beta$ we obtain, after some algebra, the conditional posterior distribution:

$$p(\beta | \sigma^2, y) \propto N(\beta | (X^T X)^{-1} X^T y, \sigma^2 (X^T X)^{-1}).$$

The conditional posterior distribution of $\beta$ would have been the desired posterior distribution had $\sigma^2$ been known.

Since that is not the case, we need to obtain the marginal posterior distribution by integrating out $\sigma^2$ as:

$$p(\beta | y) = \int p(\beta | \sigma^2, y) p(\sigma^2 | y) d\sigma^2$$

Can we solve this integration using composition sampling? YES: if we can generate samples from $p(\sigma^2 | y)$!

Ingredients of a linear model include an $n \times 1$ response vector $y = (y_1, \ldots, y_n)^T$ and an $n \times p$ design matrix (e.g., including regressors) $X = [x_1, \ldots, x_p]$, assumed to have been observed without error. The linear model:

$$y = X\beta + \epsilon; \quad \epsilon \sim N(0, \sigma^2 I)$$

The linear model is the most fundamental of all serious statistical models encompassing:

- ANOVA: $y$ is continuous, $x_1$'s are categorical
- REGRESSION: $y$ is continuous, $x_1$'s are continuous
- ANCOVA: $y$ is continuous, some $x_1$'s are continuous, some categorical.

Unknown parameters include the regression parameters $\beta$ and the variance $\sigma^2$. We assume $X$ is observed without error and all inference is conditional on $X$.

For Bayesian analysis, we will need to specify priors for the unknown regression parameters $\beta$ and the variance $\sigma^2$.

Consider independent flat priors on $\beta$ and $\log \sigma^2$:

$$p(\beta) \propto 1; \quad p(\log(\sigma^2)) \propto 1$$

or equivalently

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

None of the above two “distributions” are valid probabilities (they do not integrate to any finite number). So why is it that we are even discussing them?

It turns out that even if the priors are improper (that’s what we call them), as long as the resulting posterior distributions are valid we can still conduct legitimate statistical inference on them.

So, we need to find the marginal posterior distribution of $\sigma^2$. With the choice of the flat prior we obtain:

$$p(\sigma^2 | y) \propto \frac{1}{\sigma^2(n-p)/2 + 1} \exp\left(\frac{(n-p)\sigma^2}{2\sigma^2}\right) = IG\left(\sigma^2 \mid \frac{n-p}{2}, \frac{(n-p)\sigma^2}{2}\right),$$

where $\sigma^2 = \hat{\sigma}^2 = \frac{1}{n-p} y^T (I - P_X) y$.

This is known as an inverted Gamma distribution (also called a scaled chi-square distribution)

$$IG(\sigma^2 | (n-p)/2, (n-p)\hat{\sigma}^2/2).$$

In other words: $[(n-p)\hat{\sigma}^2/\sigma^2 | y] \sim \chi^2_{n-p}$ (with $n-p$ degrees of freedom). A striking similarity with the classical result: The distribution of $\hat{\sigma}^2$ is also characterized as $(n-p)\hat{\sigma}^2/\sigma^2$ following a chi-square distribution.
More generally, if $\theta = (\theta_1, \ldots, \theta_M)$ are the parameters in our model, we provide a set of initial values $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_M^{(0)})$ and then perform the $j$-th iteration, for $j = 1, \ldots, M$, by updating successively from the full conditional distributions:

\[
\begin{align*}
\theta_1^{(j)} &\sim p(\theta_1^{(j)} | \theta_2^{(j-1)}, \ldots, \theta_M^{(j-1)}, y) \\
\theta_2^{(j)} &\sim p(\theta_2 | \theta_1^{(j)}, \theta_3^{(j)}, \ldots, \theta_M^{(j-1)}, y) \\
&\quad \ldots \\
\theta_M^{(j)} &\sim p(\theta_M | \theta_1^{(j)}, \ldots, \theta_{M-1}^{(j)}, y)
\end{align*}
\]

(the generic $j^{th}$ element)

\[
\begin{align*}
\theta_1^{(j)} &\sim p(\theta_1 | \theta_2^{(j)}, \ldots, \theta_M^{(j-1)}, y) \\
\theta_2^{(j)} &\sim p(\theta_2 | \theta_1^{(j)}, \theta_3^{(j)}, \ldots, \theta_M^{(j-1)}, y) \\
&\quad \ldots \\
\theta_M^{(j)} &\sim p(\theta_M | \theta_1^{(j)}, \ldots, \theta_{M-1}^{(j)}, y)
\end{align*}
\]

The Gibbs sampler

- The marginal distribution of each individual regression parameter $\beta_j$ is a non-central univariate $t_{n-p}$ distribution. In fact,

\[
\frac{\beta_j - \hat{\beta}_j}{s / \sqrt{X^T X_j}} \sim t_{n-p}.
\]

The 95% credible intervals for each $\beta_j$ are constructed from the quantiles of the $t$-distribution. The credible intervals exactly coincide with the 95% classical confidence intervals, but the interpretation is direct: the probability of $\beta_j$ falling in that interval, given the observed data, is 0.95.

- Note: an intercept only linear model reduces to the simple univariate $N(y | \mu, \sigma^2/n)$ likelihood, for which the marginal posterior of $\mu$ is:

\[
\frac{\mu - \bar{y}}{s / \sqrt{n}} \sim t_{n-1}.
\]
The Metropolis-Hastings algorithm

- The Metropolis-Hastings algorithm: Start with an initial value for \( \theta = \theta^{(0)} \).
- Select a candidate or proposal distribution from which to propose a value of \( \theta \) at the \( j \)-th iteration: \( \theta^{(j)} \sim q(\theta^{(j-1)}, \nu) \). For example, \( q(\theta^{(j-1)}, \nu) = N(\theta^{(j-1)}, \nu) \) with \( \nu \) fixed.
- Compute
  \[
  r = \frac{p(\theta^*) | \theta^{(j-1)} \cdot q(\theta^{(j-1)} | \theta^*) \nu)}{p(\theta^{(j-1)} | \theta^{(j-1)} \cdot q(\theta^{(j-1)} | \theta^*) \nu)}
  \]
  - If \( r \geq 1 \) then set \( \theta^{(j)} = \theta^* \).
  - If \( r \leq 1 \) then draw \( U \sim (0,1) \). If \( U \leq r \) then \( \theta^{(j)} = \theta^* \). Otherwise, \( \theta^{(j)} = \theta^{(j-1)} \).
- Repeat for \( j = 1, \ldots M \). This yields \( \theta^{(1)}, \ldots, \theta^{(M)} \), which, after a burn-in period, will be samples from the true posterior distribution. It is important to monitor the acceptance ratio \( r \) of the sampler through the iterations. Rough recommendations: for vector updates \( r \approx 20\% \), for scalar updates \( r \approx 40\% \). This can be controlled by "tuning" \( \nu \).
- Popular approach: Embed Metropolis steps within Gibbs to draw from full conditionals that are not accessible to directly generate from.

The Metropolis-Hastings algorithm: Start with an initial value for \( \theta = \theta^{(0)} \).
- Select a candidate or proposal distribution from which to propose a value of \( \theta \) at the \( j \)-th iteration: \( \theta^{(j)} \sim q(\theta^{(j-1)}, \nu) \). For example, \( q(\theta^{(j-1)}, \nu) = N(\theta^{(j-1)}, \nu) \) with \( \nu \) fixed.
- Compute
  \[
  r = \frac{p(\theta^* | \theta^{(j-1)}) q(\theta^{(j-1)} | \theta^*) \nu)}{p(\theta^{(j-1)} | \theta^{(j-1)}) q(\theta^{(j-1)} | \theta^*) \nu)}
  \]
  - If \( r \geq 1 \) then set \( \theta^{(j)} = \theta^* \).
  - If \( r \leq 1 \) then draw \( U \sim (0,1) \). If \( U \leq r \) then \( \theta^{(j)} = \theta^* \). Otherwise, \( \theta^{(j)} = \theta^{(j-1)} \).
- Repeat for \( j = 1, \ldots M \). This yields \( \theta^{(1)}, \ldots, \theta^{(M)} \), which, after a burn-in period, will be samples from the true posterior distribution. It is important to monitor the acceptance ratio \( r \) of the sampler through the iterations. Rough recommendations: for vector updates \( r \approx 20\% \), for scalar updates \( r \approx 40\% \). This can be controlled by "tuning" \( \nu \).
- Popular approach: Embed Metropolis steps within Gibbs to draw from full conditionals that are not accessible to directly generate from.

Example: For the linear model, our parameters are \((\beta, \sigma^2)\). We write \( \theta = (\beta, \log(\sigma^2)) \) and, at the \( j \)-th iteration, propose \( \theta^* \sim N(\theta^{(j-1)}, \Sigma) \). The log transformation on \( \sigma^2 \) ensures that all components of \( \theta \) have support on the entire real line and can have meaningful proposed values from the multivariate normal. But we need to transform our prior to \( p(\beta, \log(\sigma^2)) \).

Let \( z = \log(\sigma^2) \) and assume \( p(\beta, z) = p(\beta|z)p(z) \). Let us derive \( p(z) \). REMEMBER: we need to adjust for the Jacobian. Then \( p(z) = p(z) | d \log p(z) | = p(z) e^{1/2} \). The position here is \( z = e^{z} \).

Let \( p(\beta) = 1 \) and \( p(z^2) = f_G(z^2, u, h) \). Then log posterior is:
\[
\log p(\beta, z) = -\frac{1}{2}(u + 1)z + \frac{1}{2}z^2 (Y - X\beta)^2 (Y - X\beta).
\]

A symmetric proposal distribution, say \( q(\theta^* | \theta^{(j-1)}, \Sigma) = N(\theta^{(j-1)} | \Sigma) \), cancels out in \( r \). In practice it is better to compute \( \log r(\theta^*) = \log(\theta^* | y) - \log(\theta^{(j-1)} | y) \). For the proposal, \( N(\theta^* | \Sigma) \), it is \( d = \text{var}(\theta^*) = p + 1 \).

If \( \log r(\theta^*) \geq 0 \) then set \( \theta^{(j)} = \theta^* \). If \( \log r(\theta^*) \leq 0 \) then draw \( U \sim (0,1) \). If \( U \leq r \) then \( \theta^{(j)} = \theta^* \). Otherwise, \( \theta^{(j)} = \theta^{(j-1)} \).

Repeat the above procedure for \( j = 1, \ldots, M \) to obtain samples \( \theta^{(1)}, \ldots, \theta^{(M)} \).

In principle, the Gibbs sampler will work for extremely complex hierarchical models. The only issue is sampling from the full conditionals. They may not be amenable to easy sampling – when these are not in closed form. A more general and extremely powerful – and often easier to code - algorithm is the Metropolis-Hastings (MH) algorithm.

This algorithm also constructs a Markov Chain, but does not necessarily care about full conditionals.
Hierarchical Modelling for Univariate Spatial Data

Sudipto Banerjee$^1$ and Andrew O. Finley$^2$

1 Department of Forestry & Department of Geography, Michigan State University, Lansing, Michigan, U.S.A.
2 Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

June 22, 2009

Algorithmic Modelling

- Spatial surface observed at finite set of locations $\mathcal{S} = \{s_1, s_2, ..., s_n\}$
- Tessellate the spatial domain (usually with data locations as vertices)
- Fit an interpolating polynomial:
  
  $f(s) = \sum_i w_i(\mathcal{S}; s)f(s_i)$

  "Interpolate" by reading off $f(s_0)$.

- Issues:
  - Sensitivity to tessellations
  - Choices of multivariate interpolators
  - Numerical error analysis

Simple linear model

$Y(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s})$,

- Response: $Y(\mathbf{s})$ at location $\mathbf{s}$
- Mean: $\mu = \mathbf{x}^T(\mathbf{s})\beta$
- Error: $\epsilon(\mathbf{s}) \iid N(0, \tau^2)$

Simple linear model

$Y(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s})$,

Assumptions regarding $\epsilon(\mathbf{s})$:

- $\epsilon(\mathbf{s}) \iid N(0, \tau^2)$
- $\epsilon(\mathbf{s}_i)$ and $\epsilon(\mathbf{s}_j)$ are uncorrelated for all $i \neq j$
Spatial Gaussian processes (GP):
- Say \(w(\mathbf{s}) \sim GP(0, \sigma^2 \rho(\cdot))\) and
  \[
  \text{Cov}(w(\mathbf{s}_1), w(\mathbf{s}_2)) = \sigma^2 \rho(\|\mathbf{s}_1 - \mathbf{s}_2\|)
  \]
- Let \(w = [w(\mathbf{s}_i)]_{i=1}^n\), then
  \[
  w \sim N(0, \sigma^2 R(\phi)), \text{ where } R(\phi) = [\rho(\phi; \|\mathbf{s}_i - \mathbf{s}_j\|)]_{i,j=1}^n
  \]

Realization of a Gaussian process:
- Changing \(\phi\) and holding \(\sigma^2 = 1\):
  \[
  w \sim N(0, \sigma^2 R(\phi)), \text{ where } R(\phi) = [\rho(\phi; \|\mathbf{s}_i - \mathbf{s}_j\|)]_{i,j=1}^n
  \]
- Correlation model for \(R(\phi)\):
  e.g., exponential decay
  \[
  \rho(\phi; t) = \exp(-\phi t) \text{ if } t > 0.
  \]
- Other valid models e.g., Gaussian, Spherical, Matérn.
- Effective range,
  \[
  t_0 = \ln(0.05)/\phi \approx 3/\phi
  \]

Univariate spatial regression
Simple linear model + random spatial effects
\[
Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s}),
\]
- Response: \(Y(\mathbf{s})\) at some site
- Mean: \(\mu = \mathbf{x}'(\mathbf{s})\beta\)
- Spatial random effects: \(w(\mathbf{s}) \sim GP(0, \sigma^2 \rho(\phi; \|\mathbf{s}_1 - \mathbf{s}_2\|))\)
- Non-spatial variance: \(\epsilon(\mathbf{s}) \sim N(0, \tau^2)\)

Hierarchical modelling
- First stage:
  \[
  y(\beta, w, \tau^2) \sim \prod_{i=1}^n N(Y(\mathbf{s}_i) | \mathbf{x}'(\mathbf{s}_i)\beta + w(\mathbf{s}_i), \tau^2)
  \]
- Second stage:
  \[
  w(\sigma^2, \phi) \sim N(0, \sigma^2 R(\phi))
  \]
- Third stage: Priors on \(\Omega = (\beta, \tau^2, \sigma^2, \phi)\)
- Marginalized likelihood:
  \[
  y | \Omega \sim N(X(\beta), \sigma^2 R(\phi) + \tau^2 I)
  \]
- Note: Spatial process parametrizes \(\Sigma\):
  \[
  y = X\beta + \epsilon, \epsilon \sim N(0, \Sigma), \Sigma = \sigma^2 R(\phi) + \tau^2 I
  \]

Bayesian Computations
- Choice: Fit \([y | \Omega] \times [\Omega]\) or \([y | \beta, w, \tau^2] \times [w | \sigma^2, \phi] \times [\Omega]\).
- Conditional model:
  - conjugate full conditionals for \(\sigma^2, \tau^2\) and \(w\) – easier to program.
- Marginalized model:
  - need Metropolis or Slice sampling for \(\sigma^2, \tau^2\) and \(\phi\). Harder to program.
  - But, reduced parameter space \(\Rightarrow\) faster convergence
  - \(\sigma^2 R(\phi) + \tau^2 I\) is more stable than \(\sigma^2 R(\phi)\).
- But what about \(R^{-1}(\phi)\) ?? EXPENSIVE!
Univariate spatial models

Where are the \( w \)'s?

- Interest often lies in the spatial surface \( w|y \).
- They are recovered from

\[
[w|y, X] = \int [w|\Omega, y, X] \times [\Omega|y, X] d\Omega
\]

using posterior samples:

- Obtain \( \Omega^{(1)}, \ldots, \Omega^{(G)} \sim [\Omega|y, X] \)
- For each \( \Omega^{(g)} \), draw \( w^{(g)} \sim [w|\Omega^{(g)}, y, X] \)

**NOTE:** With Gaussian likelihoods \( [w|\Omega, y, X] \) is also Gaussian. With other likelihoods this may not be easy and often the conditional updating scheme is preferred.

- Often we need to predict \( Y(s) \) at a new set of locations \( \{s_0, \ldots, s_m\} \) with associated predictor matrix \( X \).

- Sample from predictive distribution:

\[
[y|y, X, \tilde{X}] = \int [\tilde{y}|y, \Omega, X, \tilde{X}] d\Omega
\]

\[
= \int [\tilde{y}|y, \Omega, X, \tilde{X}] \times [\Omega|y, X] d\Omega,
\]

\( [\tilde{y}|y, \Omega, X, \tilde{X}] \) is multivariate normal. Sampling scheme:

- Obtain \( \Omega^{(1)}, \ldots, \Omega^{(G)} \sim [\Omega|y, X] \)
- For each \( \Omega^{(g)} \), draw \( \tilde{y}^{(g)} \sim [\tilde{y}|y, \Omega^{(g)}, X, \tilde{X}] \).

Prediction: Summary of \( [Y(s)|y] \)
Univariate spatial models

Modelling temperature: 507 locations in Colorado.

Simple spatial regression model:

\[ Y(s) = X^T(s)\beta + w(s) + \epsilon(s) \]

- \( w(s) \sim GP(0, \sigma^2(\cdot; \phi, \nu)) \)
- \( \epsilon(s) \sim N(0, \tau^2) \)

### Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.131, 3.866</td>
</tr>
<tr>
<td>Elevation</td>
<td>-0.527, -0.333</td>
</tr>
<tr>
<td>Precipitation</td>
<td>0.002, 0.072</td>
</tr>
<tr>
<td>( \phi )</td>
<td>0.051, 1.245</td>
</tr>
<tr>
<td>Range</td>
<td>38.8, 476.3</td>
</tr>
<tr>
<td>( \tau^2 )</td>
<td>0.022, 0.092</td>
</tr>
</tbody>
</table>

Temperature residual map

Residual map with elev. as covariate

Elevation map
Hierarchical Modelling for non-Gaussian Spatial Data

Sudipto Banerjee¹ and Andrew O. Finley²

¹ Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.
² Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

June 22, 2009

Spatial Generalized Linear Models

- First stage: \( Y(s_i) \) are conditionally independent given \( \beta \) and \( w(s_i) \), so \( f(y(s_i)|\beta, w(s_i), \gamma) \) equals
  \[
  h(y(s_i), \gamma) \exp\left( \gamma \left[ g(y(s_i), \eta(s_i)) - \psi(\eta(s_i)) \right] \right)
  \]
  where \( g(E(Y(s_i))) = \eta(s_i) = x(s_i) \beta + w(s_i) \) (canonical link function) and \( \gamma \) is a dispersion parameter.

- Second stage: Model \( w(s) \) as a Gaussian process:
  \[
  w \sim N(0, \sigma^2 R(\phi))
  \]

- Third stage: Priors and hyperpriors.
  - No process for \( Y(s) \), only a valid joint distribution
  - Not sensible to add a pure error term \( \epsilon(s) \)

- Often data sets preclude Gaussian modelling: \( Y(s) \) may not even be continuous
  - Example: \( Y(s) \) is a binary or count variable
    - species presence or absence at location \( s \)
    - species abundance from count at location \( s \)
    - continuous forest variable is high or low at location \( s \)

- Replace Gaussian likelihood by exponential family member
  - Diggle Tawn and Moyeed (1998)

Illustration

Binary spatial regression: forest/non-forest
We illustrate a non-Gaussian model for point-referenced spatial data:
  - Objective is to make pixel-level prediction of forest/non-forest across the domain.
  - Data: Observations are from 500 georeferenced USDA Forest Service Forest Inventory and Analysis (FIA) inventory plots within a 32 km radius circle in MN, USA.
  - The response \( Y(s) \) is a binary variable, with
    \[
    Y(s) = \begin{cases} 
    1 & \text{if inventory plot is forested} \\
    0 & \text{if inventory plot is not forested}
    \end{cases}
    \]
  - Observed covariates include the coinciding pixel values for 3 dates of 30 × 30 m resolution Landsat imagery.

Illustration from:

We fit a generalized linear model where
\[ Y(s_i) \sim \text{Bernoulli}(p(s_i)), \quad \logit(p(s_i)) = x^T(s_i)\beta + w(s_i). \]

Assume vague flat for \( \beta \), a Uniform \((3/32, 3/0.5)\) prior for \( \phi \), and an inverse-Gamma \((2, \cdot)\) prior for \( \sigma^2 \).

Parameters updated with Metropolis algorithm using target log density:
\[
\ln(p(\Omega | Y)) \propto -\sigma + 1 + n \ln(\sigma^2) - \frac{\sigma_n}{\sigma^2} - \frac{1}{2} \ln(|R(\phi)|) - \frac{1}{2} \sigma^2 w^T R(\phi)^{-1} w
\]
\[
+ \sum_{i=1}^n Y(s_i) \left( x^T(s_i)\beta + w(s_i) \right) - \frac{n}{2} \ln \left( 1 + \exp(x^T(s_i)\beta + w(s_i)) \right)
\]
\[
+ \ln(\sigma^2) + \ln(\phi - \phi_a) + \ln(\phi_b - \phi).
\]

Covariates and proximity to observed FIA plot will contribute to increase precision of prediction.

Parameter estimates (posterior medians and upper and lower 2.5 percentiles):

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Estimates: 50% (2.5%, 97.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (( \theta_0 ))</td>
<td>82.39 (49.56, 120.46)</td>
</tr>
<tr>
<td>AprilTC1 (( \theta_1 ))</td>
<td>-0.27 (-0.45, -0.11)</td>
</tr>
<tr>
<td>AprilTC2 (( \theta_2 ))</td>
<td>0.17 (0.07, 0.29)</td>
</tr>
<tr>
<td>AprilTC3 (( \theta_3 ))</td>
<td>-0.24 (-0.43, -0.08)</td>
</tr>
<tr>
<td>JulyTC1 (( \theta_4 ))</td>
<td>-0.04 (-0.25, 0.17)</td>
</tr>
<tr>
<td>JulyTC2 (( \theta_5 ))</td>
<td>0.09 (0.01, 0.19)</td>
</tr>
<tr>
<td>JulyTC3 (( \theta_6 ))</td>
<td>0.01 (0.15, 0.16)</td>
</tr>
<tr>
<td>OctTC1 (( \theta_7 ))</td>
<td>-0.43 (-0.68, -0.22)</td>
</tr>
<tr>
<td>OctTC2 (( \theta_8 ))</td>
<td>-0.03 (0.19, 0.14)</td>
</tr>
<tr>
<td>OctTC3 (( \theta_9 ))</td>
<td>-0.26 (-0.46, -0.07)</td>
</tr>
</tbody>
</table>

\( \sigma^2 = 1.358 (0.39, 2.42) \)

\( \phi = 0.00182 (0.00065, 0.0032) \)

\( \log(0.05)/\phi \) (meters) = 1644.19 (932.33, 4606.7)

Classification of 15 20 \times 20 pixel areas (based on visual inspection of imagery) into non-forest (•), moderately forest (◦), and forest (no marker).
Hierarchical Modelling for Multivariate Spatial Data

Sudipto Banerjee and Andrew O. Finley

1 Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.
2 Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

June 22, 2009

Each location contains \( m \) spatial regressions

\[
Y_k(s) = \mu_k(s) + w_k(s) + \epsilon_k(s), \quad k = 1, \ldots, m.
\]

- **Mean:** \( \mu(s) = [\mu_k(s)]_{k=1}^m = [X_k^T(s)]_{k=1}^m \)
- **Cov:** \( \text{Cov}(w(s)) = \text{MVGPF}(0, \Gamma_w(\cdot, \cdot)) \)
  \[
  \Gamma_w(s, s') = [\text{Cov}(w(s), w_k(s'))]_{k, k'=1}^m = \Sigma \伽马 \Psi
  \]
- **Error:** \( \epsilon(s) = [\epsilon_k(s)]_{k=1}^m \sim \text{MVN}(0, \Psi) \)
- \( \Psi \) is an \( m \times m \) p.d. matrix, e.g. usually \( \text{Diag}(\tau_k^2)_{k=1}^m \).

### Properties:
- \( \text{Cov}(w(s), s') = \Gamma_w(s, s') \)
- \( \Gamma_w(s, s') \) is p.d. and \( \Gamma_w(s, s) = \text{Var}(w(s)) \).
- For sites in any finite collection \( \mathbb{S} = \{s_1, \ldots, s_n\} \):
  \[
  \sum_{i=1}^n \sum_{j=1}^n u_i^T \Gamma_w(s_i, s_j) u_j \geq 0 \quad \text{for all} \quad u_i, u_j \in \mathbb{R}^m.
  \]
- Any valid \( \Gamma_w \) must satisfy the above conditions.
- The last property implies that \( \Sigma_w \) is p.d.
- In complete generality:
  - \( \Gamma_w(s, s') \) need not be symmetric.
  - \( \Gamma_w(s, s') \) need not be p.d. for \( s \neq s' \).

### Moving average or kernel convolution of a process:
- Let \( Z(s) \sim \text{GP}(0, r(s)) \). Use kernels to form:
  \[
  w_j(s) = \int \kappa_j(u) Z(s + u) du = \int \kappa_j(s - s') Z(s') ds'
  \]
- \( \Gamma_w(s - s') \) has \( (i, j) \)-th element:
  \[
  [\Gamma_w(s - s')]_{i,j} = \int \kappa_i(s - s' + u) \kappa_j(u') r(u - u') du du'
  \]
- **Convolution of Covariance Functions:** 
  \( \rho_1, \rho_2, \ldots, \rho_m \) are valid covariance functions. Form:
  \[
  [\Gamma_w(s - s')]_{i,j} = \int \rho_i(s - s' - t) \rho_j(t) dt dt.
  \]

### Point-referenced spatial data often come as multivariate measurements at each location.

**Examples:**
- **Environmental monitoring:** stations yield measurements on ozone, NO, CO, and PM\(_{2.5}\).
- **Community ecology:** assemblages of plant species due to water availability, temperature, and light requirements.
- **Forestry:** measurements of stand characteristics age, total biomass, and average tree diameter.
- **Atmospheric modeling:** at a given site we observe surface temperature, precipitation and wind speed.

- We anticipate dependence between measurements:
  - at a particular location
  - across locations

### Multivariate Spatial Modelling

Each location contains \( m \) spatial regressions:

\[
Y_k(s) = \mu_k(s) + w_k(s) + \epsilon_k(s), \quad k = 1, \ldots, m.
\]

**Mean:** \( \mu(s) = [\mu_k(s)]_{k=1}^m = [X_k^T(s)]_{k=1}^m \)

**Cov:** \( \text{Cov}(w(s)) = \text{MVGPF}(0, \Gamma_w(\cdot, \cdot)) \)

\[
\Gamma_w(s, s') = [\text{Cov}(w(s), w_k(s'))]_{k, k'=1}^m = \Sigma \伽马 \Psi
\]

**Error:** \( \epsilon(s) = [\epsilon_k(s)]_{k=1}^m \sim \text{MVN}(0, \Psi) \)

**\( \Psi \)** is an \( m \times m \) p.d. matrix, e.g. usually \( \text{Diag}(\tau_k^2)_{k=1}^m \).

### Convolution of Covariance Functions:

\( \rho_1, \rho_2, \ldots, \rho_m \) are valid covariance functions. Form:

\[
[\Gamma_w(s - s')]_{i,j} = \int \rho_i(s - s' - t) \rho_j(t) dt dt.
\]
Constructive approach

- Let \( v_k(s) \sim GP(0, \rho_k(s, s')) \), for \( k = 1, \ldots, m \) be \( m \) independent GP's with unit variance.
- Form the simple multivariate process \( v(s) = [v_k(s)]_{k=1}^m \):
  \[
  v(s) \sim MVGP(0, \Gamma_v(\cdot, \cdot))
  \]
  with \( \Gamma_v(s, s') = \text{Diag}(\rho_k(s, s')) \) for \( k = 1, \ldots, m \).
- Assume \( w(s) = A(s)v(s) \) arises as a space-varying linear transformation of \( v(s) \). Then:
  \[
  \Gamma_w(s, s') = A(s) \Gamma_v(s, s') A^T(s')
  \]
  is a valid cross-covariance function.

Constructive approach, contd.

- If \( A(s) = A \):
  - \( w(s) \) is stationary when \( v(s) \) is.
  - \( \Gamma_w(s, s') \) is symmetric.
  - \( \Gamma_v(s, s') = \rho(s, s') I_m \Rightarrow \Gamma_w = \rho(s, s') AA^T \)

- Last specification is called intrinsic and leads to separable models:
  \[
  \Sigma_w = H(\phi) \otimes \Lambda; \; \Lambda = AA^T
  \]

Multivariate spatial modelling

- Choice: Fit as \( [y|\Omega] \times [\Omega] \) or as \( [y|\beta, w, \Psi] \times [w|\Phi] \times [\Omega] \).
- Conditional model:
  - Conjugate distributions are available for \( \Psi \) and other variance parameters. Easy to program.
- Marginalized model:
  - need Metropolis or Slice sampling for most variance-covariance parameters. Harder to program.
  - But reduced parameter space (no \( w \)'s) results in faster convergence
  - \( \Sigma_w(\Phi) + I \otimes \Psi \) is more stable than \( \Sigma_w(\Phi) \).

- But what about \( \Sigma_w^{-1}(\Phi) \)?? Matrix inversion is EXPENSIVE \( O(n^3) \).

Multivariate spatial modelling

- Recovering the \( w \)'s?
  - Interest often lies in the spatial surface \( w(y) \).
  - They are recovered from
    \[
    [w(y, X)] = \int [w(\Omega, y, X)] [\Omega|y, X] d\Omega
    \]
    using posterior samples:
    - Obtain \( \Omega(1), \ldots, \Omega(n) \sim [\Omega|y, X] \)
    - For each \( \Omega(n) \), draw \( w(\Omega(n), y, X) \)

- NOTE: With Gaussian likelihoods \( [w(\Omega, y, X)] \) is also Gaussian. With other likelihoods this may not be easy and often the conditional updating scheme is preferred.
Multivariate spatial modelling

- Often we need to predict \( Y(s) \) at a new set of locations \( \{s_0, \ldots, s_m\} \) with associated predictor matrix \( X \).
- Sample from predictive distribution:
  \[
  \begin{align*}
  \hat{y}(y, X, \tilde{X}) &= \int \hat{y}(y, \Omega, X, \tilde{X})d\Omega \\
  &= \int [\hat{y}(y, \Omega, X, \tilde{X}) | \Omega] \times [\hat{y}(y, \Omega, X, \tilde{X}) | \Omega] \Omega \).
  \end{align*}
\]

\( \hat{y}(y, \Omega, X, \tilde{X}) \) is multivariate normal. Sampling scheme:
- Obtain \( \Omega^{(1)}, \ldots, \Omega^{(G)} \sim \{\hat{y}(y, X) \}
- For each \( \Omega^{(g)} \), draw \( \tilde{y}^{(g)} \sim \{\hat{y}(y, \Omega^{(g)}, X, \tilde{X}) \}

Illustration from:


Slight digression – why we fit a model:
- Association between response and covariates, \( \beta \), (e.g., ecological interpretation)
- Residual spatial and/or non-spatial associations and patterns (i.e., given covariates)
- Subsequent prediction

Study objectives:
- Evaluate methods for multi-source forest attribute mapping
- Find the “best” model, given the data
- Produce maps of biomass and uncertainty, by tree species

Study area:
- USDA FS Bartlett Experimental Forest (BEF), NH
- 1,053 ha heavily forested
- Major tree species: American beech (BE), eastern hemlock (EH), red maple (RM), sugar maple (SM), and yellow birch (YB)

Response variables:
- Metric tons of total tree biomass per ha
- Measured on 437 \( \frac{1}{10} \) ha plots
- Models fit using random subset of 218 plots
- Prediction at remaining 219 plots
Lower DIC is better.

Candidate models

Each model includes 55 covariates and 5 intercepts, therefore, \( X^T \) is 1090 \( \times \) 60.

Different specifications of variance structures:

1. Non-spatial multivariate \( \text{Diag}(\psi) = \tau^2 \)
2. \( \text{Diag}(K) \), same \( \phi \), \( \text{Diag}(\psi) \)
3. \( K \), same \( \phi \), \( \text{Diag}(\psi) \)
4. \( \text{Diag}(K) \), different \( \phi \), \( \text{Diag}(\psi) \)
5. \( K \), different \( \phi \), \( \text{Diag}(\psi) \)
6. \( K \), different \( \phi \), \( \psi \)

Selected model

- Model 5: \( K \), different \( \phi \), \( \text{Diag}(\psi) \)
- Parameters: \( K = 15 \), \( \phi = 5 \), \( \text{Diag}(\psi) = 5 \)

Focus on spatial cross-covariance matrix \( K \) (for brevity).

Posterior inference of \( \text{cor}(K) \), e.g., 50 (2.5, 97.5) percentiles:

<table>
<thead>
<tr>
<th></th>
<th>BE</th>
<th>EH</th>
</tr>
</thead>
<tbody>
<tr>
<td>RM</td>
<td>-0.20 (-0.23, -0.15)</td>
<td>0.45 (0.26, 0.66)</td>
</tr>
<tr>
<td>SM</td>
<td>-0.20 (-0.22, -0.17)</td>
<td>-0.12 (-0.16, -0.09)</td>
</tr>
<tr>
<td>YB</td>
<td>0.07 (0.04, 0.08)</td>
<td>0.22 (0.20, 0.25)</td>
</tr>
</tbody>
</table>

These relationships expressed in mapped random spatial effects, \( w \).
Summary

Proposed Bayesian hierarchical spatial methodology:
- Partition sources of uncertainty
  - Provides hypothesis testing
  - Reveal spatial patterns and missing covariates
- Allow flexible inference
  - Access parameters’ posterior distribution
  - Access posterior predictive distribution
- Provide consistent prediction of multiple variables
  - Maintains spatial and non-spatial association

Extendable model template:
- Cluster plot sample design – multiresolution models
- Non-continuous response – general linear models
- Obs. over time and space – spatiotemporal models
Spatio-temporal Models

Specification:
- Again point-referenced vs. areal unit data
- Continuous time vs. discretized time

⇒ association in space, association in time
For point-referenced data, \( t \) continuous, Gaussian
\[
Y(t, s) = \mu(t, s) + \epsilon(t, s)
\]
where
\[
\epsilon(t, s) \sim N(0, \sigma^2)
\]
non-Gaussian data, \( g(E(Y(t, s)) = \mu(t, s) + \epsilon(t, s) \)

Don’t treat time as a third coordinate \((s, t)\)

\[
\text{Cov}(Y(s, t), Y(s', t')) = C(s - s', t - t')
\]

Dynamic spatiotemporal models

Measurement Equation
\[
Y(t, s) = \mu(s, t) + \epsilon(s, t); \quad \epsilon(s, t) \sim N(0, \sigma^2)
\]
\[
\mu(s, t) = X(s, t)\beta(t)
\]
\[
\tilde{\beta}(t) = \beta(t) + \epsilon(s, t)
\]

Transition Equation
\[
\beta_t = \beta_{t-1} + \eta_t; \quad \eta_t \sim N_p(0, \Sigma_{\eta})
\]
\[
\beta(s, t) = \beta(s, t-1) + W(s, t)
\]
Hierarchical Modelling for Large Spatial Datasets

Sudipto Banerjee\textsuperscript{1} and Andrew O. Finley\textsuperscript{2}

\textsuperscript{1} Department of Forestry \& Department of Geography, Michigan State University, Lansing Michigan, U.S.A.
\textsuperscript{2} Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

June 22, 2009

The Big $n$ issue

\textbf{Univariate spatial regression}

\[ Y = X\beta + w + \epsilon, \]

- Estimation involves $(\sigma^2 I(\phi) + \tau^2 I)^{-1}$, which is $n \times n$.
- Matrix computations occur in each MCMC iteration.
- Known as the “Big-N problem” in geostatistics.
- Approach: Use a model $Y = X\beta + Zw^* + \epsilon$. But what $Z$?

Knots: A “Knotty” problem??

- Knot selection: Regular grid? More knots near locations we have sampled more?
- Formal spatial design paradigm: maximize information metrics (Zhu and Stein, 2006; Diggle and Lophaven, 2006)
- Geometric considerations: space-filling designs (Royle & Nychka, 1998); various clustering algorithms
- Compare performance of estimation of range and smoothness by varying knot size.
- Stein (2007, 2008): method may not work for fine-scale spatial data
- Still a popular choice – seamlessly adapts to multivariate and spatiotemporal settings.

A rectified predictive process is defined as

\[ \hat{w} \sim \tilde{w} + \tilde{e}, \text{ where} \]

\[ \hat{w}(s) \overset{ind}{\sim} N(0, \sigma_w^2 (1 - R(s, \phi)) R^{-1} \otimes (\phi) R(s, \phi)). \]

Maximum likelihood estimates of $\tau^2$:

<table>
<thead>
<tr>
<th># of Knots</th>
<th>Predictive Process $\hat{\tau}^2$</th>
<th>Rectified Predictive Process $\hat{\tau}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1.56941</td>
<td>1.00786</td>
</tr>
<tr>
<td>36</td>
<td>1.65688</td>
<td>1.15386</td>
</tr>
<tr>
<td>64</td>
<td>1.45169</td>
<td>1.08358</td>
</tr>
<tr>
<td>100</td>
<td>1.37916</td>
<td>1.09657</td>
</tr>
<tr>
<td>225</td>
<td>1.27391</td>
<td>1.08358</td>
</tr>
<tr>
<td>400</td>
<td>1.22429</td>
<td>1.09489</td>
</tr>
<tr>
<td>625</td>
<td>1.21127</td>
<td>1.09998</td>
</tr>
<tr>
<td>exact</td>
<td>1.14414</td>
<td>1.14414</td>
</tr>
</tbody>
</table>
Illustration Univariate random effects models

Genetic effects model:

\[ Y_i = x_{i1}^T \beta + a_i + d_i + \epsilon_i, \]

- Common feature is systematic heterogeneity among observational units (i.e., violation of \( \epsilon \sim N(0, \tau^2 I_n) \))
- Spatial heterogeneity arises from:
  - soil characteristics
  - micro-climates
  - light availability
- Residual correlation among units as a function of distance and/or direction = erroneous parameter estimates (e.g., biased \( h^2 \))

Univariate random effects models

Modeling genetic variation in Scots pine (Pinus sylvestris L.), long-term progeny study in northern Sweden.

Quantitative genetics: studies the inheritance of polygenic traits, focusing upon estimation of additive genetic variance, \( \sigma^2_A \), and the heritability \( h^2 = \sigma^2_A / \sigma^2_T \), where the \( \sigma^2_T \) represents the total genetic and unexplained variation.

A high heritability, \( h^2 \), should result in a larger selection response (i.e., a higher probability for genetic gain in future generations).

Parameter credible intervals, 50% (2.5%, 97.5%) for the non-spatial models Scots pine trial.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta )</td>
<td>72.53 (69.66, 75.08)</td>
<td>72.27 (70.04, 74.57)</td>
<td></td>
</tr>
<tr>
<td>( \sigma^2_A )</td>
<td>31.94 (18.30, 49.85)</td>
<td>25.23 (14.12, 43.96)</td>
<td></td>
</tr>
<tr>
<td>( \sigma^2_D )</td>
<td>~</td>
<td>22.37 (11.24, 40.11)</td>
<td></td>
</tr>
<tr>
<td>( \tau^2 )</td>
<td>133.60 (121.18, 144.70)</td>
<td>116.14 (100.51, 127.76)</td>
<td></td>
</tr>
<tr>
<td>( h^2 )</td>
<td>0.19 (0.12, 0.28)</td>
<td>0.15 (0.09, 0.26)</td>
<td></td>
</tr>
</tbody>
</table>

Illustration from:

Data overview:
- established in 1971 (by Skogforsk)
- partial diallel design of 52 parent trees
- 8,160 planted randomly on 2.2m squares
- 1997 reinventory of 4,970 surviving trees, height, DBH, branch angle, etc.
Previous approaches to accommodating residual spatial dependence:
- Manipulate the mean function
- constructing covariates using residuals from neighboring units (see e.g., Wilkinson et al., 1983; Besag and Kempton, 1986; Williams, 1986)
- Geostatistical
  - spatial process formed $AR(1)_{col} \otimes AR(1)_{row}$ (Martin, 1990; Cullis et al., 1998)
  - classical geostatistical method (Zimmerman and Harville, 1991)

All are computationally feasible, but ad hoc and/or restrictive from a modeling perspective.

**Trick to sample genetic effects:**

Gibbs draw for random effects, e.g.,
$$a|\cdot \sim MVN(\mu_a, \Sigma_a),$$

where calculating $\Sigma_a = \frac{1}{\sigma_a^2} A^{-1} + \frac{1}{\tau^2} I$ is computationally expensive!

However $A$ and $D$ are known, so use initial spectral decomposition i.e., $A^{-1} = P^T \Lambda^{-1} P$.

Thus, $\Sigma_a = P^T \left( \frac{1}{\sigma_a^2} \Lambda^{-1} + \frac{1}{\tau^2} I \right)^{-1} P$ to achieve computational benefits.

**Trick** does not work for $w$. Rather, we proposed the knot-based **predictive process**.

**Corresponding predictive process model:**

$$w(s_i) \sim C(s_i - s_j)^T \Sigma^{-1} (s_i - s_j), \Sigma = G(\psi, \Lambda).$$

$w$ can accommodate complex spatial dependence structures, e.g., anisotropic Matérn correlation function:
$$\rho(s_i, s_j; \theta) = \left( 1 + \left( \frac{s_i - s_j}{\kappa} \right)^2 \right)^{-\nu/2},$$
where $d_{ij} = (s_i - s_j)^T \Sigma^{-1} (s_i - s_j), \Sigma = G(\psi, \Lambda).$

So, $\epsilon \sim N(0, \tau^2 I_n)$. Consider a spatial model.
Genetic + spatial effects models

- Candidate spatial models (i.e., specifications of $C^*(\theta)$):
  - AR(1)$_{col}$ $\otimes$ AR(1)$_{row}$
  - isotropic Matérn
  - anisotropic Matérn
- Each model evaluated using 64, 144, and 256 knot grids.
- Model choice using Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002)

### Table: Model comparisons using the DIC criterion for the Scots pine dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>$P_D$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-spatial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Add.</td>
<td>306.40</td>
<td>15,618.09</td>
</tr>
<tr>
<td>Add. Dom.</td>
<td>555.92</td>
<td>15,547.85</td>
</tr>
<tr>
<td>Spatial Isotropic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64 Knots</td>
<td>639.77</td>
<td>14,877.51</td>
</tr>
<tr>
<td>144 Knots</td>
<td>739.61</td>
<td>14,814.89</td>
</tr>
<tr>
<td>256 Knots</td>
<td>802.29</td>
<td>14,771.64</td>
</tr>
<tr>
<td>Spatial Anisotropic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64 Knots</td>
<td>678.82</td>
<td>14,884.13</td>
</tr>
<tr>
<td>144 Knots</td>
<td>748.89</td>
<td>14,823.90</td>
</tr>
<tr>
<td>256 Knots</td>
<td>806.46</td>
<td>14,781.53</td>
</tr>
</tbody>
</table>

### Parameter credible intervals, 50% (2.5%, 97.5%) for the isotropic Matérn and 64 and 256 knots Scots pine trial:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>64 Knots</th>
<th>256 Knots</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>72.53 (69.00, 76.05)</td>
<td>74.21 (69.66, 79.66)</td>
</tr>
<tr>
<td>$\sigma^2_a$</td>
<td>26.87 (17.14, 41.82)</td>
<td>33.03 (18.19, 53.69)</td>
</tr>
<tr>
<td>$\sigma^2_d$</td>
<td>11.69 (6.00, 34.27)</td>
<td>13.96 (7.65, 27.05)</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>41.84 (23.71, 73.34)</td>
<td>50.36 (30.24, 88.10)</td>
</tr>
<tr>
<td>$\nu$</td>
<td>89.55 (72.11, 99.65)</td>
<td>80.75 (67.90, 96.16)</td>
</tr>
<tr>
<td>$h^2$</td>
<td>0.83 (0.31, 1.46)</td>
<td>0.47 (0.26, 1.28)</td>
</tr>
<tr>
<td>$\phi^2$</td>
<td>0.05 (0.02, 0.09)</td>
<td>0.04 (0.02, 0.09)</td>
</tr>
<tr>
<td>Eff. Range</td>
<td>71.00 (44.66, 127.93)</td>
<td>74.59 (45.22, 129.83)</td>
</tr>
</tbody>
</table>

Decrease in $\tau^2$ due to removal of spatial variation, results in increase in $h^2$ (i.e., ~ 0.25 vs. ~ 0.15 with confounding).

**Summary**

Challenge - to meet modeling needs:

- ensure computationally feasible
  - reduce algorithmic complexity = cheap tricks (e.g., spectral decomp. of $A$ prior to MCMC)
  - reduce dimensionality = predictive process
- maintain richness and flexibility
  - focus on the model not how to estimate the parameters = embrace new tools (MCMC) for estimating highly flexible hierarchical models
- truly acknowledge sources of uncertainty
  - propagate uncertainty through hierarchical structures (e.g., recognize uncertainty in $C(\theta)$)

**Predictive process – balance model richness with computational feasibility (e.g., $4,970 \times 4,970$ vs. $64 \times 64$).**
Some computing notes

June 22, 2009

Computing environment

Many core and contributed packages (including spBayes) call Basic Linear Algebra Subprograms (BLAS) and LAPACK (Linear Algebra PACKage) Fortran libraries.

Substantial computing gains:
- processor specific threaded BLAS/LAPACK implementation (e.g., Intel's Math Kernel Library or AMD's Core Math Library (ACML))
- processor specific compilers (e.g., Intel's icc/ifort)

Time needed to collect 100 MCMC samples using spLM and threaded vs. non-threaded BLAS/LAPACK on a Intel Core 2 Quad processor and Ubuntu 8.10 Linux OS. R compiled with GNU gcc and gfortran.