

spBayes: An R Package for Univariate and Multivariate Hierarchical Point-referenced Spatial Models

Andrew O. Finley¹, Sudipto Banerjee²,
and Bradley P. Carlin²

¹Michigan State University, Departments of Forestry and Geography

²University of Minnesota, Division of Biostatistics

July 29, 2007

Outline:

- Need for new software tools
- **spBayes** overview
- Spatial models and Bayesian inference
- Illustration
- Future direction
- Summary

Spatial data is used in an array of disciplines – economics, genetics, natural sciences, sociology, . . .

Spatial data is used in an array of disciplines – economics, genetics, natural sciences, sociology, . . .

Trends in data availability and analytical needs:

- Increasing volume of spatial data
- Addressing interrelated variables of interest
- Summarizing uncertainty at many levels (Hierarchical modeling)
- Delivering spatial analysis products to end users

Scope of interest:

- *Point-referenced* or geostatistical data vs. *areally referenced* data
- Point locations defined on \mathbb{R}^2 (e.g., latitude-longitude or Easting-Northing)
- Multivariate generalized linear model
- Simple spatial dependence structures – stationary and isotropic spatial processes

Current software options:

- **BUGS** (Bayesian Inference Using Gibbs Sampling)
 - Flexible coding environment
 - Too slow (e.g., matrix operations)

Current software options:

- **BUGS** (Bayesian Inference Using Gibbs Sampling)
 - Flexible coding environment
 - Too slow (e.g., matrix operations)

- **geoR** and **geoRglm**
 - Univariate only
 - Limited prior specifications
 - Too canned and cumbersome

Current software options:

- **BUGS** (Bayesian Inference Using Gibbs Sampling)
 - Flexible coding environment
 - Too slow (e.g., matrix operations)
- **geoR** and **geoRglm**
 - Univariate only
 - Limited prior specifications
 - Too canned and cumbersome
- **Model specific coding**
 - Efficient code – C, C++, FORTRAN
 - Optimized libs – **BLAS** (Basic Linear Algebra Subprogs.), **LAPACK** (Linear Algebra Package)
 - Too time consuming

spBayes features

Models:

- Univariate and multivariate spatial regression
- Choice of spatial and non-spatial covariance matrices
- Fully Bayesian specification

Diagnostics:

- Model choice – DIC (Deviance Information Criterion)
- **CODA** (Convergence Diagnosis and Output Analysis) ready output

Interface and underlying code:

- R package
- Written in C++ and **BLAS**, **LAPACK**, and **SPARSKIT**
- R's foreign language interface permits OS portability and processor optimization

spBayes features

Models:

- Univariate and multivariate spatial regression
- Choice of spatial and non-spatial covariance matrices
- Fully Bayesian specification

Diagnostics:

- Model choice – DIC (Deviance Information Criterion)
- **CODA** (Convergence Diagnosis and Output Analysis) ready output

Interface and underlying code:

- R package
- Written in C++ and **BLAS**, **LAPACK**, and **SPARSKIT**
- R's foreign language interface permits OS portability and processor optimization

spBayes features

Models:

- Univariate and multivariate spatial regression
- Choice of spatial and non-spatial covariance matrices
- Fully Bayesian specification

Diagnostics:

- Model choice – DIC (Deviance Information Criterion)
- **CODA** (Convergence Diagnosis and Output Analysis)
ready output

Interface and underlying code:

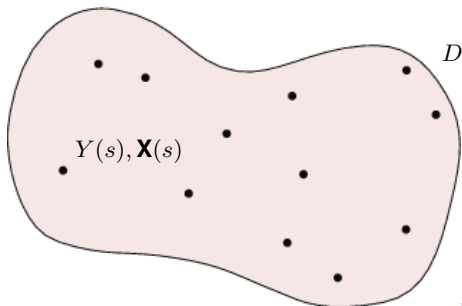
- R package
- Written in C++ and **BLAS**, **LAPACK**, and **SPARSKIT**
- R's foreign language interface permits OS portability and processor optimization

Point-referenced spatial models and Bayesian inference (in a nutshell)

Simple linear model + random spatial effects

$$Y(s) = \mu(s) + w(s) + \epsilon(s),$$

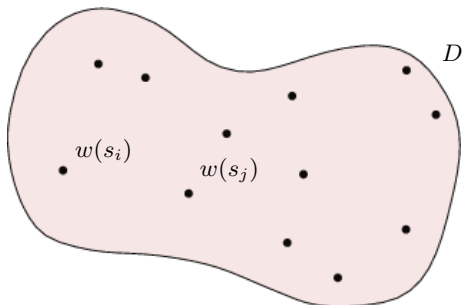
- Response: $Y(s)$ at some site
- Mean: $\mu = \mathbf{X}(s)^T \boldsymbol{\beta}$
- **Spatial random effects:** $w(s) \sim GP(\mathbf{0}, \sigma^2 \rho(\theta; \|s - s'\|))$
- Non-spatial variance: $\epsilon(s) \stackrel{iid}{\sim} N(\mathbf{0}, \tau^2)$



Spatial Gaussian process (GP):

- Say $w(s) \sim GP(\mathbf{0}, \sigma^2 \rho(\cdot))$ and

$$\text{Cov}(w(s), w(s')) = \sigma^2 \rho(\theta; \|s - s'\|)$$



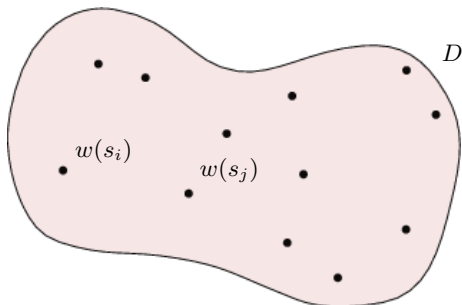
Spatial Gaussian process (GP):

- Say $w(s) \sim GP(\mathbf{0}, \sigma^2 \rho(\cdot))$ and

$$\text{Cov}(w(s), w(s')) = \sigma^2 \rho(\theta; \|s - s'\|)$$

- Let $\mathbf{w} = [w(s_i)]_{i=1}^n$, then

$$\mathbf{w} \sim MVN(\mathbf{0}, \sigma^2 H(\theta)), \text{ where } H(\theta) = [\rho(\theta; \|s_i - s_j\|)]_{i,j=1}^n$$

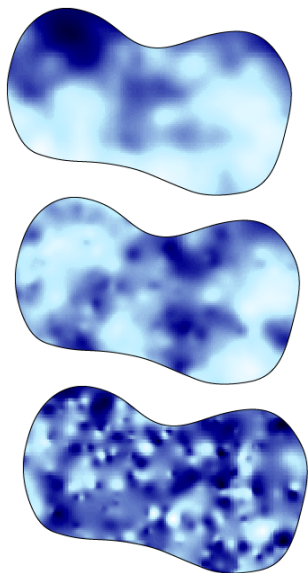


Realization of a Gaussian process:

- Changing θ and holding $\sigma^2 = 1$:

$\mathbf{w} \sim MVN(\mathbf{0}, \sigma^2 H(\theta))$, where

$$H(\theta) = [\rho(\theta; \|s_i - s_j\|)]_{i,j=1}^n$$



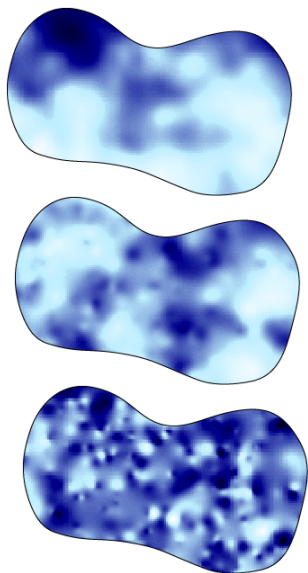
Realization of a Gaussian process:

- Changing θ and holding $\sigma^2 = 1$:

$$\mathbf{w} \sim MVN(\mathbf{0}, \sigma^2 H(\theta)), \text{ where}$$
$$H(\theta) = [\rho(\theta; \|s_i - s_j\|)]_{i,j=1}^n$$

- Correlation model for $H(\theta)$:
e.g., exponential decay

$$\rho(\theta; t) = \exp(-\theta t) \text{ if } t > 0.$$



Realization of a Gaussian process:

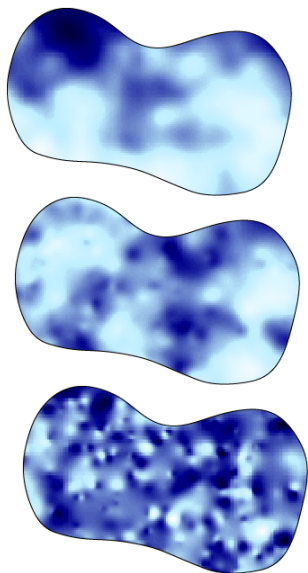
- Changing θ and holding $\sigma^2 = 1$:

$$\mathbf{w} \sim MVN(\mathbf{0}, \sigma^2 H(\theta)), \text{ where}$$
$$H(\theta) = [\rho(\theta; \|s_i - s_j\|)]_{i,j=1}^n$$

- Correlation model for $H(\theta)$:
e.g., exponential decay

$$\rho(\theta; t) = \exp(-\theta t) \text{ if } t > 0.$$

- Other **valid** models e.g., Gaussian, Spherical, Matérn, *etc.*



Multivariate spatial model

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

- Response: $\mathbf{Y}(s) = [Y_i(s)]_{i=1}^q$
- Mean: $\boldsymbol{\mu} = \mathbf{X}^T(s)\boldsymbol{\beta}$, where $\mathbf{X}^T(s) = [\mathbf{x}_i^T(s)]_{i=1}^q$
- Spatial random effects: $\mathbf{w}(s) \sim MVGP(\mathbf{0}, K(s, s'; \boldsymbol{\theta}))$
- Non-spatial variance: $\boldsymbol{\epsilon}(s) \sim MVN(\mathbf{0}, \Psi)$

Multivariate spatial model

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

- Response: $\mathbf{Y}(s) = [Y_i(s)]_{i=1}^q$
- Mean: $\boldsymbol{\mu} = \mathbf{X}^T(s)\boldsymbol{\beta}$, where $\mathbf{X}^T(s) = [\mathbf{x}_i^T(s)]_{i=1}^q$
- Spatial random effects: $\mathbf{w}(s) \sim MVGP(\mathbf{0}, K(s, s'; \boldsymbol{\theta}))$
- Non-spatial variance: $\boldsymbol{\epsilon}(s) \sim MVN(\mathbf{0}, \Psi)$

Care is needed in choosing $K(s, s'; \boldsymbol{\theta})$ – insure symmetric and positive definite $qn \times qn$ covariance matrix.

Multivariate spatial model

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

- Response: $\mathbf{Y}(s) = [Y_i(s)]_{i=1}^q$
- Mean: $\boldsymbol{\mu} = \mathbf{X}^T(s)\boldsymbol{\beta}$, where $\mathbf{X}^T(s) = [\mathbf{x}_i^T(s)]_{i=1}^q$
- Spatial random effects: $\mathbf{w}(s) \sim MVGP(\mathbf{0}, K(s, s'; \boldsymbol{\theta}))$
- Non-spatial variance: $\boldsymbol{\epsilon}(s) \sim MVN(\mathbf{0}, \Psi)$

Care is needed in choosing $K(s, s'; \boldsymbol{\theta})$ – insure symmetric and positive definite $qn \times qn$ covariance matrix.

$$K(s, s'; \boldsymbol{\theta}) = \mathbf{A}[\oplus_{k=1}^q \rho_k(s, s'; \boldsymbol{\theta}_k)]\mathbf{A}^T$$

Model parameterization

1 Unmarginalized likelihood:

- First stage: $\mathbf{Y} | \boldsymbol{\beta}, \mathbf{w}, \Psi \sim MVN(\mathbf{X}^T \boldsymbol{\beta} + \mathbf{w}, I_n \otimes \Psi)$

Model parameterization

1 Unmarginalized likelihood:

- First stage: $\mathbf{Y}|\boldsymbol{\beta}, \mathbf{w}, \Psi \sim MVN(\mathbf{X}^T \boldsymbol{\beta} + \mathbf{w}, I_n \otimes \Psi)$
- Second stage: $\mathbf{w}|\mathbf{A}, \boldsymbol{\theta} \sim MVN(\mathbf{0}, \Sigma_w)$,
where $\Sigma_w = [K(s_i, s_j; \boldsymbol{\theta})]_{i,j=1}^n$

Model parameterization

1 Unmarginalized likelihood:

- First stage: $\mathbf{Y}|\boldsymbol{\beta}, \mathbf{w}, \Psi \sim MVN(\mathbf{X}^T\boldsymbol{\beta} + \mathbf{w}, I_n \otimes \Psi)$
- Second stage: $\mathbf{w}|\mathbf{A}, \boldsymbol{\theta} \sim MVN(\mathbf{0}, \Sigma_w)$,
where $\Sigma_w = [K(s_i, s_j; \boldsymbol{\theta})]_{i,j=1}^n$

2 Marginalized likelihood (used in **spBayes**):

$$\mathbf{Y}|\Omega \sim MVN(\mathbf{X}^T\boldsymbol{\beta}, \Sigma_w + I_n \otimes \Psi), \text{ where } \Omega = \{\boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\theta}, \Psi\}.$$

Sampling and quantities of interest

spBayes MCMC sampling scheme for Ω :

- β – Gibbs (default) or Metropolis-Hastings (MH)
- \mathbf{A} , Ψ , θ – MH but soon slice sampling

Recover \mathbf{w} given Ω samples:

$$P(\mathbf{w} | Data) \propto \int P(\mathbf{w} | \Omega, Data) P(\Omega | Data) d\Omega.$$

Prediction given Ω samples and **new** $\mathbf{X}(s_0) \dots \mathbf{X}(s_m)$:

$$P(\mathbf{w}^* | Data) \propto \int P(\mathbf{w}^* | \mathbf{w}, \Omega, Data) P(\mathbf{w} | \Omega, Data) P(\Omega | Data) d\Omega d\mathbf{w},$$

$$P(\mathbf{Y}^* | Data) \propto \int P(\mathbf{Y}^* | \Omega, Data) P(\Omega | Data) d\Omega.$$

Sampling and quantities of interest

spBayes MCMC sampling scheme for Ω :

- β – Gibbs (default) or Metropolis-Hastings (MH)
- \mathbf{A} , Ψ , θ – MH but soon slice sampling

Recover \mathbf{w} given Ω samples:

$$P(\mathbf{w} | Data) \propto \int P(\mathbf{w} | \Omega, Data) P(\Omega | Data) d\Omega.$$

Prediction given Ω samples and **new** $\mathbf{X}(s_0) \dots \mathbf{X}(s_m)$:

$$P(\mathbf{w}^* | Data) \propto \int P(\mathbf{w}^* | \mathbf{w}, \Omega, Data) P(\mathbf{w} | \Omega, Data) P(\Omega | Data) d\Omega d\mathbf{w},$$

$$P(\mathbf{Y}^* | Data) \propto \int P(\mathbf{Y}^* | \Omega, Data) P(\Omega | Data) d\Omega.$$

Sampling and quantities of interest

spBayes MCMC sampling scheme for Ω :

- β – Gibbs (default) or Metropolis-Hastings (MH)
- \mathbf{A} , Ψ , θ – MH but soon slice sampling

Recover \mathbf{w} given Ω samples:

$$P(\mathbf{w} | Data) \propto \int P(\mathbf{w} | \Omega, Data) P(\Omega | Data) d\Omega.$$

Prediction given Ω samples and **new** $\mathbf{X}(s_0) \dots \mathbf{X}(s_m)$:

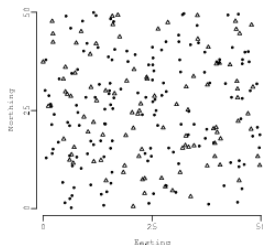
$$P(\mathbf{w}^* | Data) \propto \int P(\mathbf{w}^* | \mathbf{w}, \Omega, Data) P(\mathbf{w} | \Omega, Data) P(\Omega | Data) d\Omega d\mathbf{w},$$

$$P(\mathbf{Y}^* | Data) \propto \int P(\mathbf{Y}^* | \Omega, Data) P(\Omega | Data) d\Omega.$$

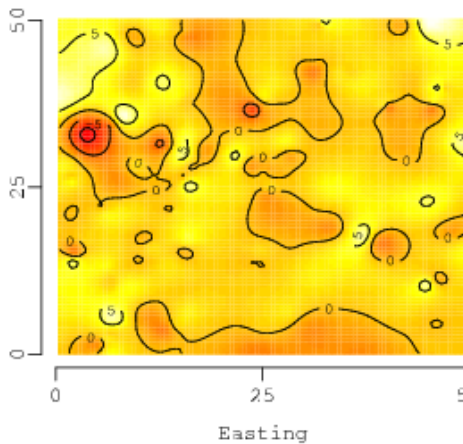
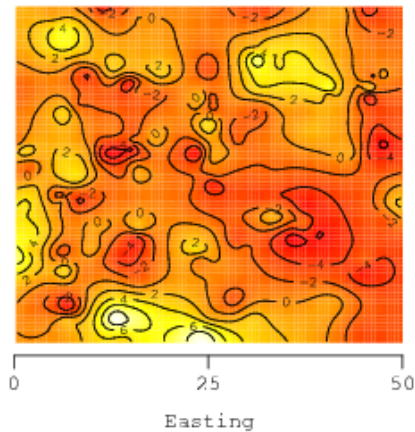
Synthetic data

Given two response variables (i.e., $q = 2$), exponential spatial correlation function, and

$$\beta = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{K}(\mathbf{0}; \theta) = \begin{pmatrix} 1 & -2 \\ -2 & 8 \end{pmatrix}, \Psi = \begin{pmatrix} 9 & 0 \\ 0 & 2 \end{pmatrix}, \theta = \begin{pmatrix} 0.6 \\ 0.1 \end{pmatrix}.$$



Synthetic data (cont'd)

 Y_1  Y_2 

Potential candidate models fit with `ggT.sp` in **spBayes**

Model 1: $\mathbf{w} = \mathbf{0}$ (non-spatial)

Model 2: $\mathbf{A} = \sigma I_q$ and $\Psi = \mathbf{0}$

Model 3: $\mathbf{A} = \sigma I_q$ and $\Psi = \tau^2 I_q$

Model 4: $\mathbf{A} = \text{diag}[\sigma_i]_{i=1}^q$ and $\Psi = \text{diag}[\tau_i^2]_{i=1}^q$

Model 5: \mathbf{A} and $\Psi = \text{diag}[\tau_i^2]_{i=1}^q$

Model 6: $\mathbf{A}, \Psi = \text{diag}[\tau_i^2]_{i=1}^q$, and $\boldsymbol{\theta} = \{\phi_k\}_{k=1}^q$

Model 7: \mathbf{A}, Ψ , and $\boldsymbol{\theta} = \{\phi_k\}_{k=1}^q$

Specifying a model using `ggT.sp`

Recall Model 6: $\mathbf{A}, \Psi = \text{diag}[\tau_i^2]_{i=1}^q$, and $\theta = \{\phi_k\}_{k=1}^q$

Step 1: Define parameters – priors, hyperpriors

```
K.prior <- prior(dist="IWISH", df=2, S=diag(c(3,6)))  
Psi.prior.1 <- prior(dist="IG", shape=2, scale=7)  
Psi.prior.2 <- prior(dist="IG", shape=2, scale=5)  
phi.prior <- prior(dist="UNIF", a=0.06, b=3)
```

Step 2: Define parameters' sampling info – starting value, MH tuning, etc.

```
var.update.control <-  
list(  
  "K"=list(sample.order=0, starting=diag(1, 2),  
           tuning=diag(c(0.1, 0.5, 0.1)), prior=K.prior),  
  
  "Psi"=list(sample.order=1, starting=1,  
            tuning=0.3, prior=list(Psi.prior.1, Psi.prior.2))  
  
  "phi"=list(sample.order=2, starting=0.5,  
            tuning=0.5, prior=list(phi.prior, phi.prior))  
)  
  
beta.control <-  
list(update="GIBBS", prior=prior(dist="FLAT"))
```

Step 3: Run control – number of samples, etc.

```
run.control <-  
  list("n.samples"=5000, "sp.effects"=TRUE)
```

Step 4: Call ggt.sp

```
ggt.Model.6 <- ggt.sp(  
  formula=list(Y.1~1, Y.2~1),  
  run.control=run.control,  
  coords=coords,  
  var.update.control=var.update.control,  
  beta.update.control=beta.control,  
  cov.model="exponential")
```

Step 5: Prediction if desired

```
sp.pred <- sp.predict(ggt.Model.6,  
  pred.coords=coords, pred.covars=covars)
```

Step 3: Run control – number of samples, etc.

```
run.control <-  
  list("n.samples"=5000, "sp.effects"=TRUE)
```

Step 4: Call ggt.sp

```
ggt.Model.6 <- ggt.sp(  
  formula=list(Y.1~1, Y.2~1),  
  run.control=run.control,  
  coords=coords,  
  var.update.control=var.update.control,  
  beta.update.control=beta.control,  
  cov.model="exponential")
```

Step 5: Prediction if desired

```
sp.pred <- sp.predict(ggt.Model.6,  
  pred.coords=coords, pred.covars=covars)
```

Step 3: Run control – number of samples, etc.

```
run.control <-  
  list("n.samples"=5000, "sp.effects"=TRUE)
```

Step 4: Call ggt.sp

```
ggt.Model.6 <- ggt.sp(  
  formula=list(Y.1~1, Y.2~1),  
  run.control=run.control,  
  coords=coords,  
  var.update.control=var.update.control,  
  beta.update.control=beta.control,  
  cov.model="exponential")
```

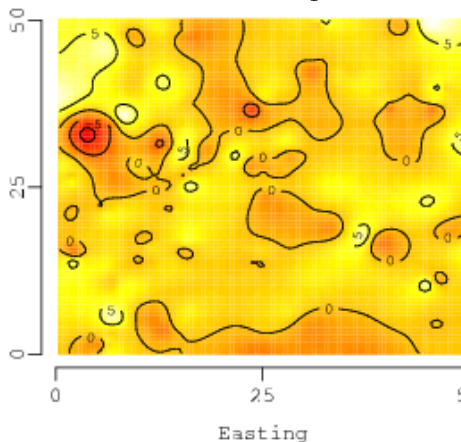
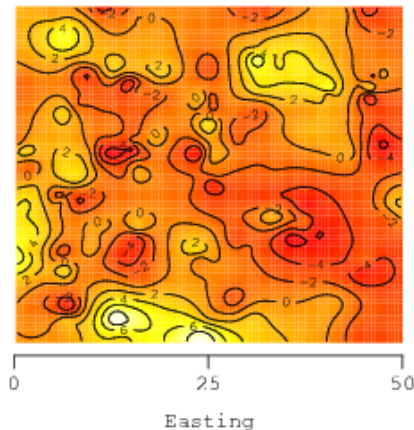
Step 5: Prediction if desired

```
sp.pred <- sp.predict(ggt.Model.6,  
  pred.coords=coords, pred.covars=covars)
```

CODA ready posterior samples held in `ggt.sp` object.
 For example, summary of `ggt.Model.6$p.samples`:

Parameter	Estimates: 50% (2.5%, 97.5%)
$\beta_{1,0}$	1.086 (0.555, 1.628)
$\beta_{2,0}$	-0.273 (-1.619, 1.157)
K _{1,1}	1.801 (0.542, 6.949)
K _{2,1}	-1.784 (-3.604, -0.357)
K _{2,2}	8.253 (4.645, 13.221)
$\Psi_{1,1}$	7.478 (3.020, 10.276)
$\Psi_{2,2}$	2.276 (0.832, 5.063)
ϕ_1	1.024 (0.243, 2.805)
ϕ_2	0.193 (0.073, 0.437)

Finally, given the model object and **new** $\mathbf{X}(s_0) \dots \mathbf{X}(s_m)$

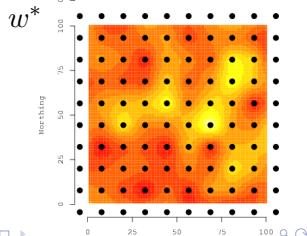
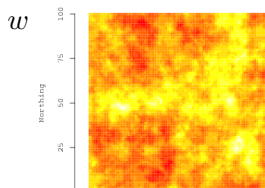
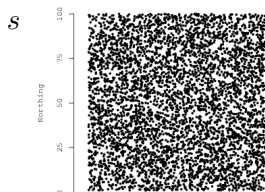
 Y_1^*  Y_2^* 

Future of **spBayes**

- Extend to other GLMs
- More priors and sampling routines (e.g., slice sampling)
- Include general model specification (e.g., **MCMC** package's `metrop`)
- Address “BIG-N” challenges

New function `sp.lm`

- Fully Bayesian `lm` with spatial effects
- Minimal arguments for fitting and prediction
- “BIG-N” through *Predictive Process*
 - Knot-based dimension reduction
 - Fit models with $n > 10,000$ with common desktop computer
 - See Banerjee et al. (2007)



Summary

spBayes begins to meet a statistical computing need:

- Flexible model specification
- Efficient MCMC computation
- Useful parameter and predictive inference
- Portable and scalable code base

This work was supported by:

NASA Headquarters under the Earth System Science Fellowship Grant NGT5

NSF Grant 0706870

USDA Forest Service Forest Inventory and Analysis program

University of Minnesota, School of Statistics

References:

Finley, A.O., Banerjee, S., and Carlin. B.P. (2007). spBayes: An R package for Univariate and Multivariate Hierarchical Point-referenced Spatial Models *Journal of Statistical Software*.

<http://www.jstatsoft.org> 19(4).

Finley, A.O., Banerjee, S., Ek, A.R. and McRoberts, R.E. (In press). Bayesian multivariate process modeling for prediction of forest attributes. *Journal of Agricultural, Biological, and Environmental Statistics*.

Banerjee, S., and Finley, A.O. (In press). Bayesian multiresolution modeling of spatially replicated data. *Journal of Statistical Planning and Inference*.

Finley, A.O., Banerjee, S., and McRoberts, R.E. (In press). A Bayesian approach to quantifying uncertainty in multi-source forest area estimates. *Environmental and Ecological Statistics*.



Questions