Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geo-statistical Datasets

Abhirup Datta[1] Sudipto Banerjee[1] Andrew O. Finley[2]

Alan E. Gelfand[3]

[1]University of Minnesota, Minneapolis, MN, USA

[2]Michigan State University, East Lansing, MI, USA

[3]Duke University, Durham, NC, USA

PASI, 2014

Outline

U.S. Forest biomass data

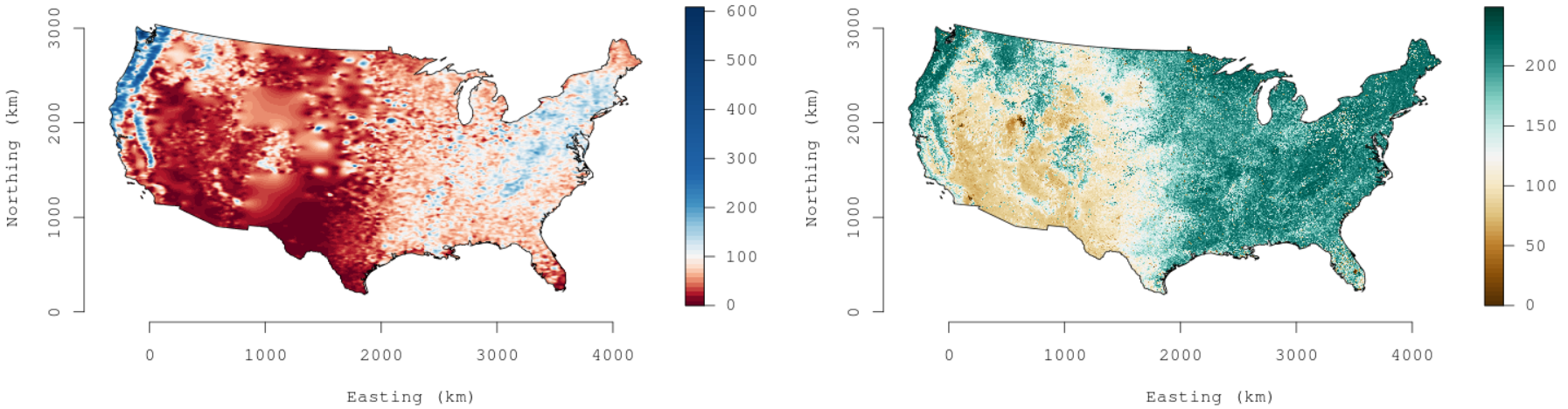


Figure: Observed biomass (left) and NDVI (right)

- Forest biomass data collected between 1999 and 2006 at 114,371 plots
- Normalized Difference Vegetation Index (NDVI) calculated in July 2006
- NDVI is a measure of greenness and is used as a covariate in Forest Biomass Regression Models

Non Spatial Model

Model

$$Biomass = \beta_0 + \beta_1 NDVI + error, \quad \hat{\beta}_0 = 1.043, \hat{\beta}_1 = 0.0093$$

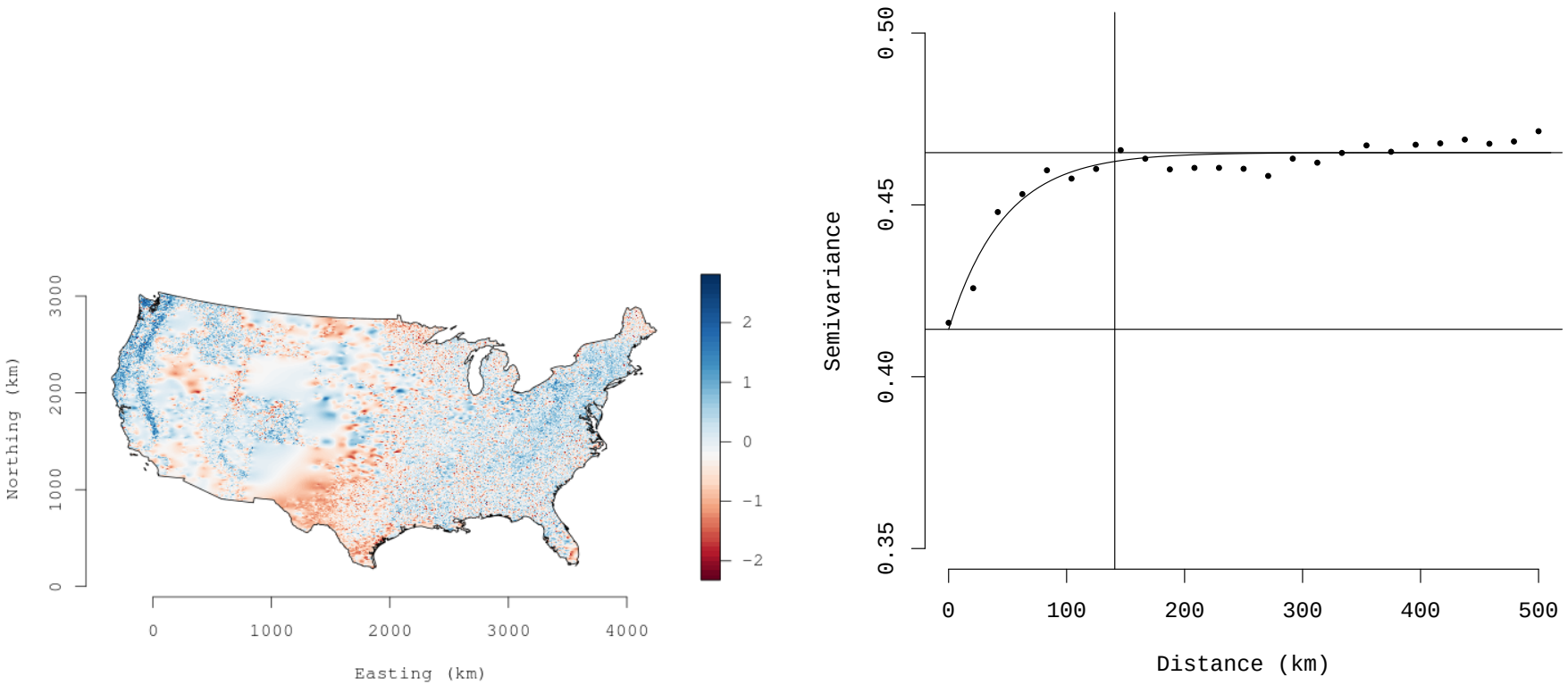


Figure: Heat map (left) and variogram (right) of residuals reflecting spatial correlation

Gaussian process models

Full rank model

- $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ denote locations where data is observed
- $y(s_i)$ denote the response at the i^{th} location
- $y = (y(s_1), y(s_2), \ldots, y(s_n))'$
- $y = X\beta + w + \epsilon$, $\epsilon \sim N(0, \tau^2 I)$
- $w = (w(s_1), w(s_2), \ldots, w(s_n))'$ are spatial random effects
- $w \sim GP(0, C(\theta))$, $C(\theta)$ is a valid spatial covariance matrix

Computation issues

- Needs to store the n^2 pairwise distances to compute $C(\theta)$
- $C(\theta)$ is dense, computing $C(\theta)^{-1}$ uses $O(n^3)$ flops
- Computationally infeasible for such massive datasets

Computation issues

- Needs to store the n^2 pairwise distances to compute $C(\theta)$
- $C(\theta)$ is dense, computing $C(\theta)^{-1}$ uses $O(n^3)$ flops
- Computationally infeasible for such massive datasets

Low rank models

- Regresses w on $\tilde{w}$ realized at r locations (knots) where $r \ll n$
- $O(nr^2)$ flops but requires large r for massive datasets
- Has known performance issues (Stein 2013 [2])

Composite likelihoods (Vecchia 1988 [3], Stein 2004 [1])

- y_{i_m} denote the realizations of the GP at m nearest neighbors of s_i among $s_1, s_2, \ldots, s_{i-1}$
- $p_{CL}(y) = p(y(s_1)) \prod_{i=2}^{n} p(y(s_i) \mid y_{i_m})$ where m is very small (~ 10)
- Computationally efficient parameter estimation
- Various other choices for neighbor sets
- Confidence intervals and model evaluation based on inappropriate asymptotics
- Not model based: estimation and prediction based on different likelihoods
- Not always possible to recover the residual spatial surface i.e. w

Directed acyclic graphs

- $w(s) \sim p(\cdot)$ be any stochastic process over a domain $\mathcal{D}$
- $\mathcal{S}^* = (s_1^*, s_2^*, \ldots, s_k^*)$ is a finite set of locations in $\mathcal{D}$
- $G_{\mathcal{S}^*}$ denotes a directed graph on $\mathcal{S}^*$
- $N(s_i^*)$ denotes the set of directed neighbors for s_i^* in the graph
- $p(w_{\mathcal{S}^*})$ is the probability density of realizations of $w(s)$ over $\mathcal{S}^*$

Theorem

If $G_{\mathcal{S}^*}$ is closed and acyclic, then the composite likelihood given by

$$\tilde{p}(w_{\mathcal{S}^*}) = \prod_{i=1}^{k} p(w(s_i^*) \mid w_{N(s_i^*)}) \tag{1}$$

is a valid multivariate density over $\mathcal{S}^*$

Extension to process

- For any location s outside $\mathcal{S}^*$, $N(s)$ denote a set of directed neighbors of s in $\mathcal{S}^*$
- For any $\mathcal{S} = (s_1, s_2, \ldots, s_n)$ outside $\mathcal{S}$ define,

$$\tilde{p}(w_{\mathcal{S}} \mid w_{\mathcal{S}^*}) = \prod_{i=1}^{n} p(w(s_i) \mid w_{N(s_i)}) \tag{2}$$

Theorem

If $\mathcal{S}^*$ is fixed and $G_{\mathcal{S}^*}$ is closed and acyclic, then the finite dimensional densities given in equations (1) and (2) define a stochastic process over the entire domain $\mathcal{D}$ that satisfies Kolmogorov's consistency criteria.

Nearest Neighbor Gaussian Process

Theorem

If $w(s) \sim GP(0, C(\theta))$ be a stationary Gaussian process over $\mathcal{D}$ and every location s in $\mathcal{D}$ has at most m directed neighbors, then

1. $\tilde{p}(w_{\mathcal{S}^*})$ is joint density from the model $w_{\mathcal{S}^*} \sim N(0, \tilde{C}_{\mathcal{S}^*})$ where $\tilde{C}^{-1}_{\mathcal{S}^*}$ is sparse with at most km^2 non-zero entries
2. $\tilde{p}(w_{\mathcal{S}} \mid w_{\mathcal{S}^*})$ is the density for $w_{\mathcal{S}} \mid w_{\mathcal{S}^*} \sim N(B_{\mathcal{S}} w_{\mathcal{S}^*}, F_{\mathcal{S}})$ where $B_{\mathcal{S}}$ is sparse with atmost nm non-zero entries. $F_{\mathcal{S}}$ is diagonal
3. $\tilde{p}(\cdot)$ defines a new Gaussian Process derived from the parent Gaussian Process

We denote it by Nearest Neighbor Gaussian Process – $NNGP(0, \tilde{C}(\theta))$

Nearest Neighbor Gaussian Process

- Sparse precision matrices

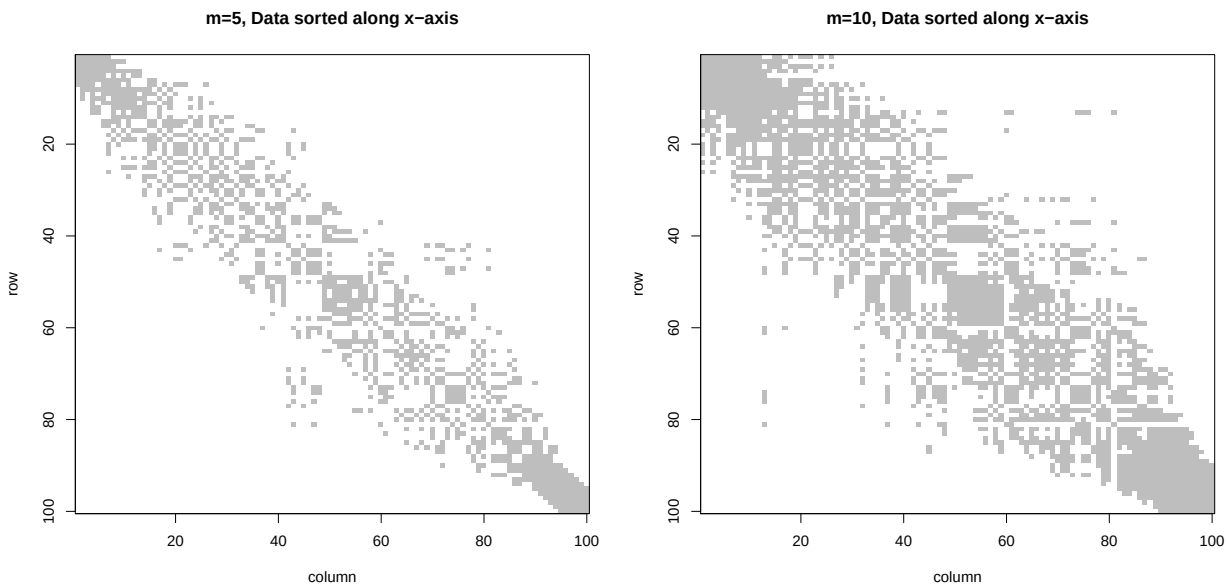


Figure: Sparse precision matrices for NNGP

- Any choice of neighbor sets used by Vecchia and Stein can be used
- Easily embeds into a hierarchical setup

Hierarchical NNGP model

- $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ denote locations where data is observed
- $\mathcal{S}_1 = \mathcal{S} \setminus \mathcal{S}^* = \{s_1, s_2, \ldots, s_r\}$

Model

$$y(s) = X(s)'\beta + w(s) + \epsilon(s) \qquad w(s) \sim NNGP(0, \tilde{C}(\theta))$$
$$\epsilon \sim N(0, \tau^2 I) \qquad \tau^2 \sim IG(a_\tau . b_\tau)$$
$$\beta \sim N(\mu_\beta, V_\beta) \qquad \theta \sim \pi(\theta)$$

Likelihood

$$N(y \,|\, X\beta + w_{\mathcal{S}}, \tau^2 I) \times N(w_{\mathcal{S}_1} \,|\, B_{\mathcal{S}_1} w_{\mathcal{S}^*}, F_{\mathcal{S}_1}) \times N(w_{\mathcal{S}^*} \,|\, 0, \tilde{C}_{\mathcal{S}^*})$$
$$\times \quad N(\beta \,|\, \mu_\beta, V_\beta) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times \pi(\theta)$$

Gibbs' sampler

- Conjugate full conditionals for β, τ^2
- Sequential updates or sparse Cholesky block update for full conditional of $w(s_i^*)$'s
- Block update for full conditionals of $w(s_i)$'s, $i = 1, 2, \ldots, r$
- Metropolis Hastings step for updating θ

Gibbs' sampler

- Conjugate full conditionals for β, τ^2
- Sequential updates or sparse Cholesky block update for full conditional of $w(s_i^*)$'s
- Block update for full conditionals of $w(s_i)$'s, $i = 1, 2, \ldots, r$
- Metropolis Hastings step for updating θ

Storage and computation

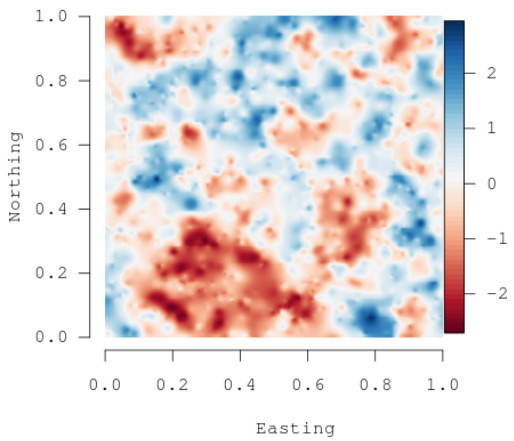
- Never needs to store $n \times n$ distance matrix. Stores $n + k$ small $m \times m$ matrices
- Total flop count per iteration of Gibbs' sampler is linear in $n + k$

Choice of S, neighbors, m

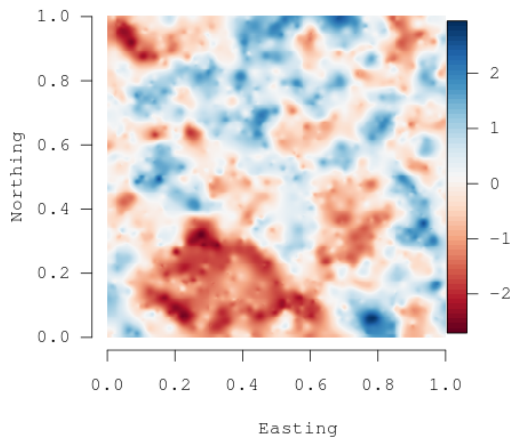
- Fully model based estimation and prediction
- Posterior distribution of the parameters and the residual surface w
- Both storage and computation is linear in $n + k$. Scalabale to massive datasets.
- k can be as large as n. Not a low rank process
- $\mathcal{S}^*$ is usually a large grid on the domain with k close to n.
- $\mathcal{S}^*$ can be chosen to be $\mathcal{S}$ if $\mathcal{S}$ is large and uniformly spread over $\mathcal{D}$
- All competing choices of $\mathcal{S}^*$, $N(s)$ and m can be compared using standard model comparison metrics

Simulation experiments

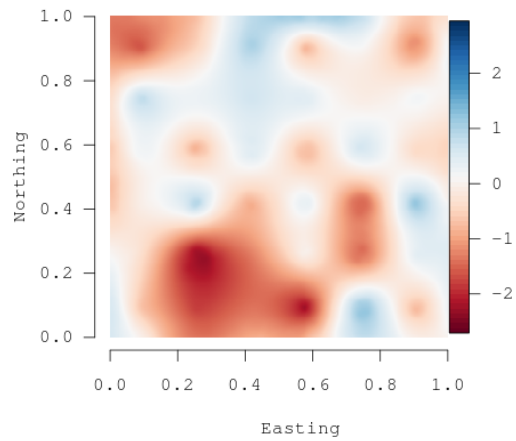
- 2500 locations on a unit square
- $y(s_i) = \beta_0 + \beta_1 X(s_i) + w(s_i) + \epsilon(s_i)$
- Single covariate generated from $N(0, 1)$
- $R(\nu, \phi)$ denotes Matern correlation function with smoothness ν and decay ϕ
- Spatial effects generated from $GP(0, \sigma^2 R(\nu, \phi))$
- Candidate models: Full GP, Low rank GP (PPGP) with 64 knots, NNGP $\{\mathcal{S}^* = \mathcal{S}\}$ and NNGP $\{\mathcal{S}^* \neq \mathcal{S}\}$



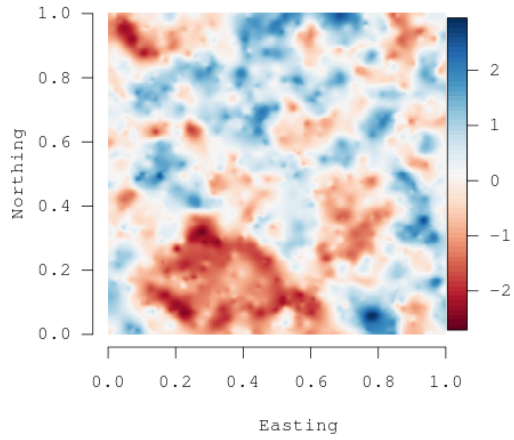
(a) True w



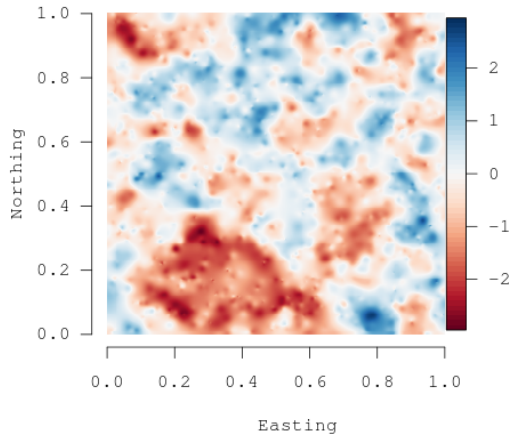
(b) Full GP



(c) PPGP 64 knots



(d) NNGP $m = 10$ $\mathcal{S} = \mathcal{S}^*$



(e) NNGP $m = 10$ $\mathcal{S} \neq \mathcal{S}^*$

Table: Univariate synthetic data analysis

	True	NNGP ($\mathcal{S}^* \neq \mathcal{S}$) $m = 10, k = 2000$	NNGP ($\mathcal{S}^* = \mathcal{S}$) $m = 10$	Predictive Process 64 knots	Full Gaussian Process
β_0	1	0.99 (0.71, 1.48)	1.00 (0.62, 1.31)	1.30 (0.54, 2.03)	1.03 (0.69, 1.34)
β_1	5	5.00 (4.98, 5.03)	5.01 (4.99, 5.03)	5.03 (4.99, 5.06)	5.01 (4.99, 5.03)
σ^2	1	1.09 (0.89, 1.49)	0.96 (0.78, 1.23)	1.29 (0.96, 2.00)	0.94 (0.76, 1.23)
τ^2	0.1	0.07 (0.04, 0.10)	0.10 (0.08, 0.13)	0.08 (0.04, 0.13)	0.10 (0.08, 0.12)
ϕ	12	11.81 (8.18, 15.02)	12.93 (9.70, 16.77)	5.61 (3.48, 8.09)	13.52 (9.92, 17.50)
p_D	–	1491.08	1243.32	1258.27	1260.68
DIC	–	1856.85	2390.65	13677.97	2364.80
G	–	33.67	77.84	1075.63	74.80
P	–	253.03	340.40	200.39	333.27
D	–	286.70	418.24	1276.03	408.08
RMSPE	–	1.22	1.2	1.68	1.2
Run time (Minutes)	–	**20.29**	**14.40**	43.36	560.31

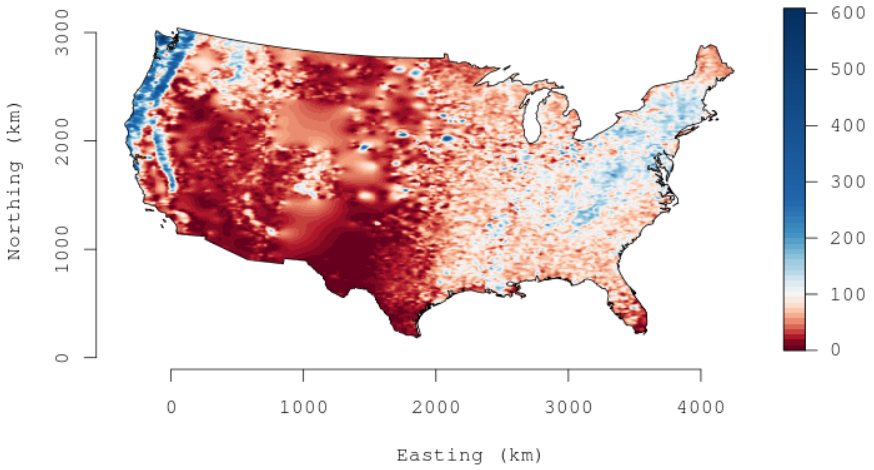
- Parameter estimates for all models are similar
- NNGP performs at par with Full GP, PPGP performs worse
- NNGP yields huge computational gains

Back to the Forest biomass dataset

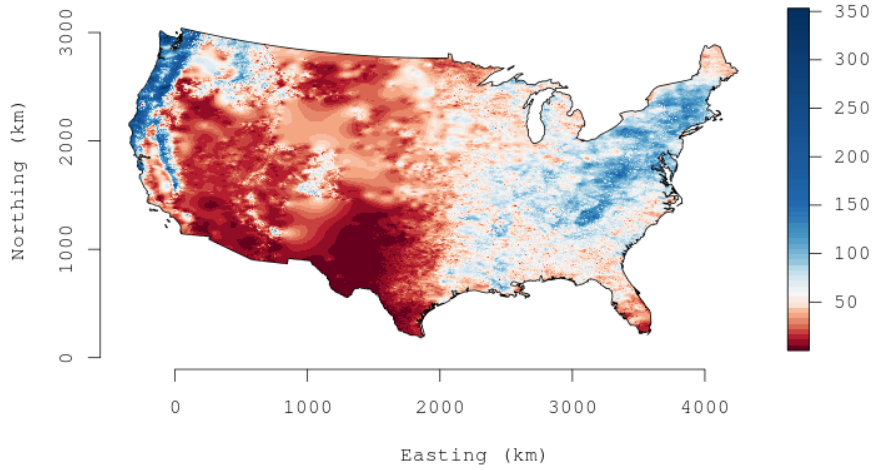
- $n = 114,371$
- Full GP and PPGP storage requirements $\gg$ 38 gibabytes available
- We use a hierarchical spatially varying coefficients NNGP model

Model

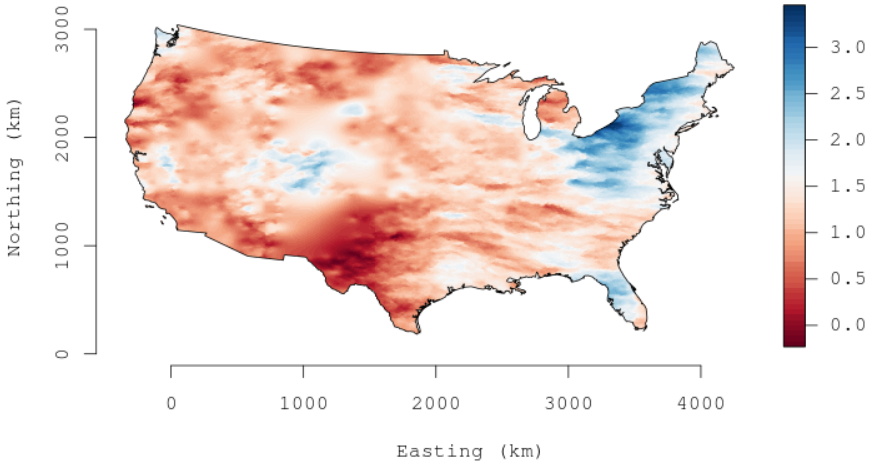
- $Biomass(s) = \beta_0(s) + \beta_1(s)NDVI(s) + \epsilon(s)$
- $w(s) = (\beta_0(s), \beta_1(s))' \sim$ Bi-variate $NNGP(0, \tilde{C}(\theta))$
- $\mathcal{S}^* = \mathcal{S}, m = 5$
- Computation time 46 hrs



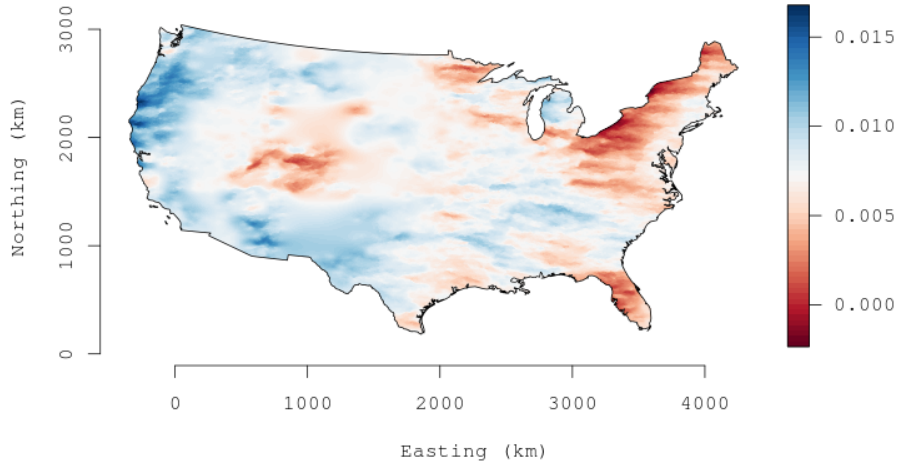
(a) Observed biomass



(b) Fitted biomass



(c) $\beta_0(s)$



(d) $\beta_{NDVI}(s)$

Conclusions

- Model based framework for a large class of neighbor based composite likelihood techniques
- Unified platform for estimation, prediction and model comparison
- Easily extends to multivariate spatial processes
- Seamlessly adapts into a hierarchical setup
- Posterior predictions, recovery of spatial surface
- Superior performance, massive computation and storage gains over existing models
- Possible extension to spatio-temporal models, spatial GLMs

References

[1] Stein, M. L., Chi, Z., and Welty, L. J. (2004), *Approximating Likelihoods for Large Spatial Data Sets*, Journal of the Royal Statistical Society. Series B (Methodological), 66, 275-296.

[2] Stein M.L. (2013), *Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data*, Spatial Statistics.

[3] Vecchia, A. V. (1988), *Estimation and Model Identification for Continuous Spatial Processes*, Journal of the Royal Statistical Society. Series B (Methodological), 50, 297-312.