Bayesian spatial modeling

Abhi Datta

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland

abhidatta.com @dattascience

Review of last lecture

- Basics of Bayesian inference – priors, posteriors, sampling, posterior (point and interval) estimates
- Example: Bayesian linear model
- Bayesian spatial GP model analysis using `rstan`

Bayesian inference for spatial linear model

- $y(s) = x(s)'\beta + w(s) + \epsilon(s)$, $w(s) \sim GP(0, C(\cdot, \cdot \mid \boldsymbol{\phi}))$, $\epsilon \overset{\text{iid}}{\sim} N(0, \tau^2)$
- Latent/unmarginalized model: For n locations, we have $y = N(X\beta + w, \tau^2 I)$, $w \sim N(0, C(\phi))$
- Assuming stationarity, $C(\boldsymbol{\phi}) = \sigma^2 R(\phi)$ where $R := R(\phi)$ is the correlation matrix
- Marginalized model: $y \sim N(X\beta, \sigma^2 R + \tau^2 I)$
- Letting $\theta = (\beta, \sigma^2, \tau^2, \phi)$ and $p(\theta)$ the prior, we have $p(\theta \mid y) \propto$

$$\frac{1}{\sqrt{|\sigma^2 R + \tau^2 I|}} \exp\left(-\frac{1}{2}(y - X\beta)'(\sigma^2 R + \tau^2 I)^{-1}(y - X\beta)\right) \times p(\theta).$$

- We will use `rstan` to sample multiple chains from this non-standard posterior

MCMC diagnostics: Trace plots

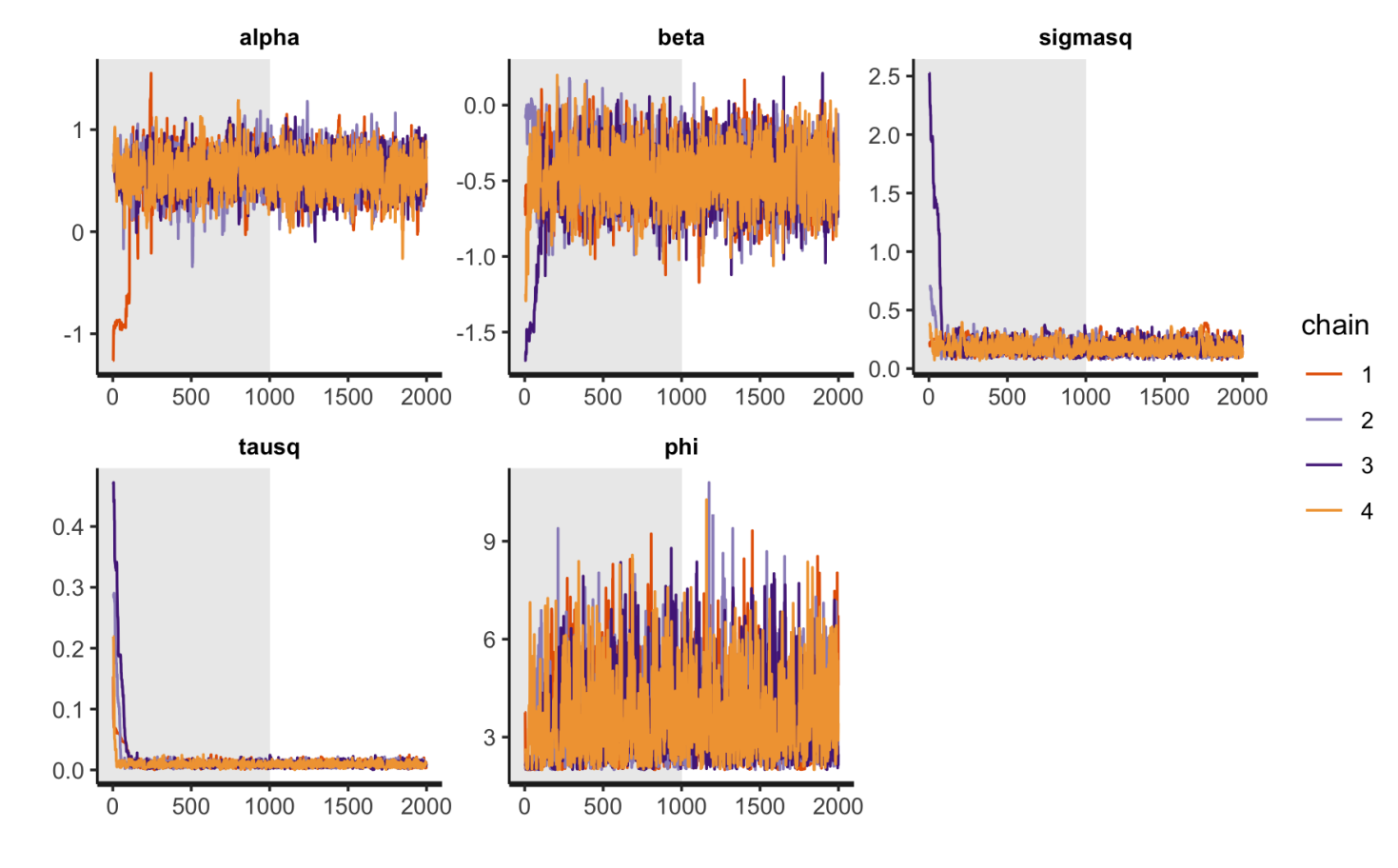


Figure: Trace plots for the marginalized model for Dataset 3

- Chains look well mixed
- Assessment of convergence from trace plots is subjective and may be unreliable

MCMC diagnostics: Density plots

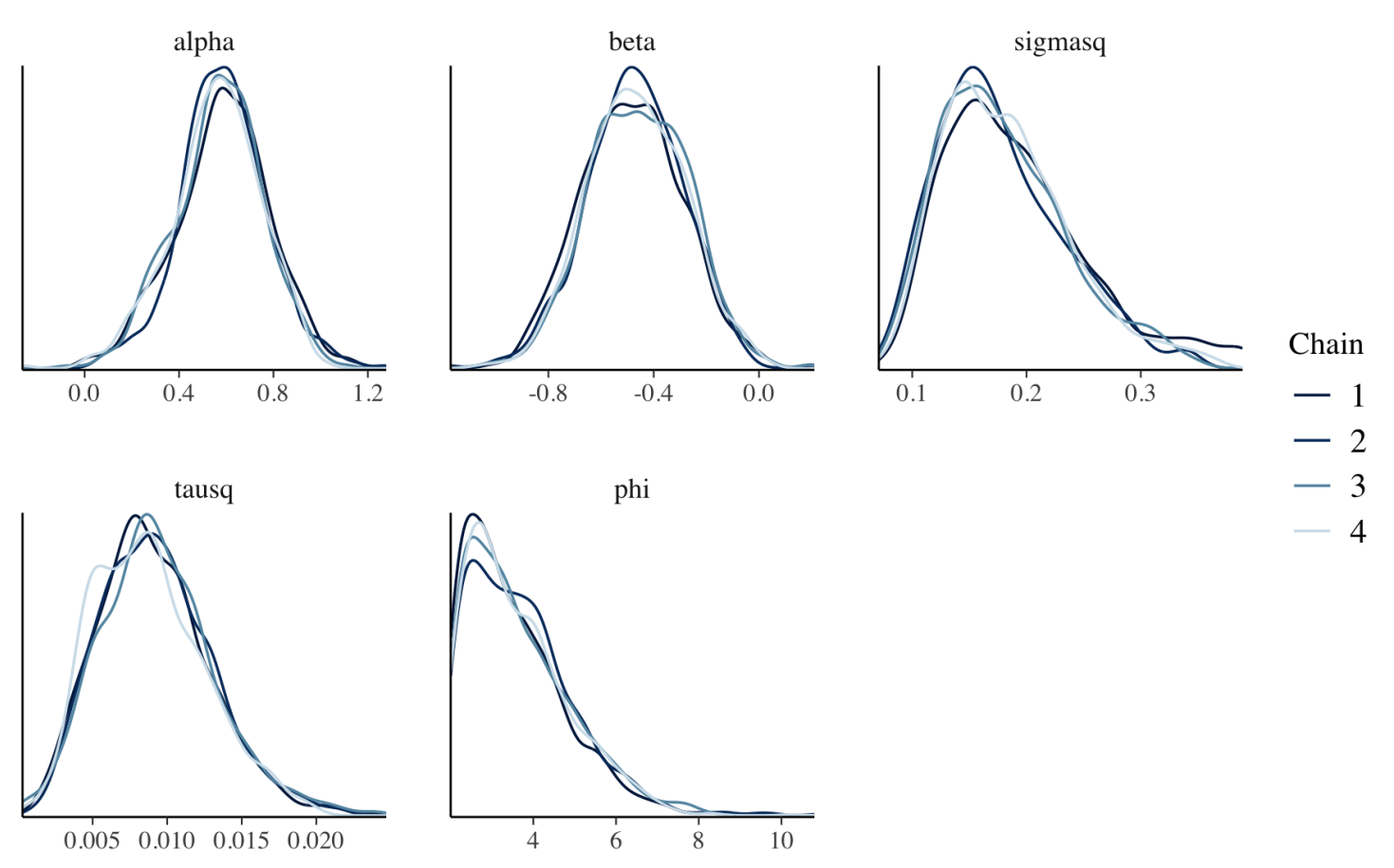


Figure: Density plots for the marginalized model for Dataset 3 using the `bayesplot` package

- Densities from different chains look very similar
- Assessment of convergence from density plots is subjective

MCMC diagnostics: Gelman-Rubin shrink factor

- Run chains of length N with overdispersed initial values
- Discard the first N_b draws of each chain as burn-in
- For each variable θ, calculate the within-chain variance
  $W = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{N-N_b-1} \sum_{i=N_b+1}^{N} (\theta^{(ij)} - \bar{\theta}_j)^2$
- For each variable θ, calculate the between-chain variance
  $B = \frac{N-N_b}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2$ where $\bar{\theta} = \frac{1}{m} \sum_{j=1}^{m} \bar{\theta}_j$
- Calculate the Gelman-Rubin shrink factor as
  $\widehat{R} = \sqrt{\frac{(1-\frac{1}{N-N_b})W + \frac{1}{N-N_b}B}{W}}$
- Thumb rule: $\widehat{R} > 1.1$ indicates lack of convergence

MCMC diagnostics: Other formal metrics

- Effective sample size (ESS): Sample size of the MCMC after adjusting the length of the MCMC run for autocorrelation among the samples. Higher ESS is better.
- Monte Carlo standard error: Standard error estimates for the parameters using ESS. Lower MCSE implies closer approximation by the Monte Carlo.
- Useful slides on MCMC diagnostics: `https://mc-stan.org/docs/2_18/reference-manual/effective-sample-size-section.html`

MCMC diagnostics

	Rhat	n_eff	se_mean
alpha	1.0052	1058.8104	0.0059
beta	1.0021	1689.7914	0.0046
sigmasq	1.0128	442.1063	0.0028
tausq	1.0024	1342.2045	0.0001
phi	1.0043	845.8044	0.0428

Figure: MCMC diagnostics for Dataset 3

- Stan has in-built functions to calculate $\widehat{R}$, ESS (n_{eff}), and MCSE of the posterior mean (se_{mean}) for each parameter
- Also, *coda* package in R calculates them for general MCMC outputs

Latent model

- So far we have used the marginalized model $y \sim N(X\beta, \sigma^2 R + \tau^2 I)$ to optimize (MLE) for $\theta = (\beta, \sigma^2, \tau^2, \phi)$ or sample (MCMC) from the posterior of θ
- Latent/unmarginalized model: We can also directly use the hierarchical model

$$y \sim N(X\beta + w, \tau^2 I),$$
$$w \sim N(0, \sigma^2 R(\phi))$$

  and sample from the joint posterior of θ and the latent random effects w
- Posterior $p(\theta, w|y)$ is proportional to the joint likelihood: $N(y \,|\, X\beta + w, \tau^2 I) \times N(w \,|\, 0, \sigma^2 R(\phi)) \times p(\theta)$
- MCMC is particularly convenient to sample from posteriors of such hierarchical models

Latent vs marginalized model

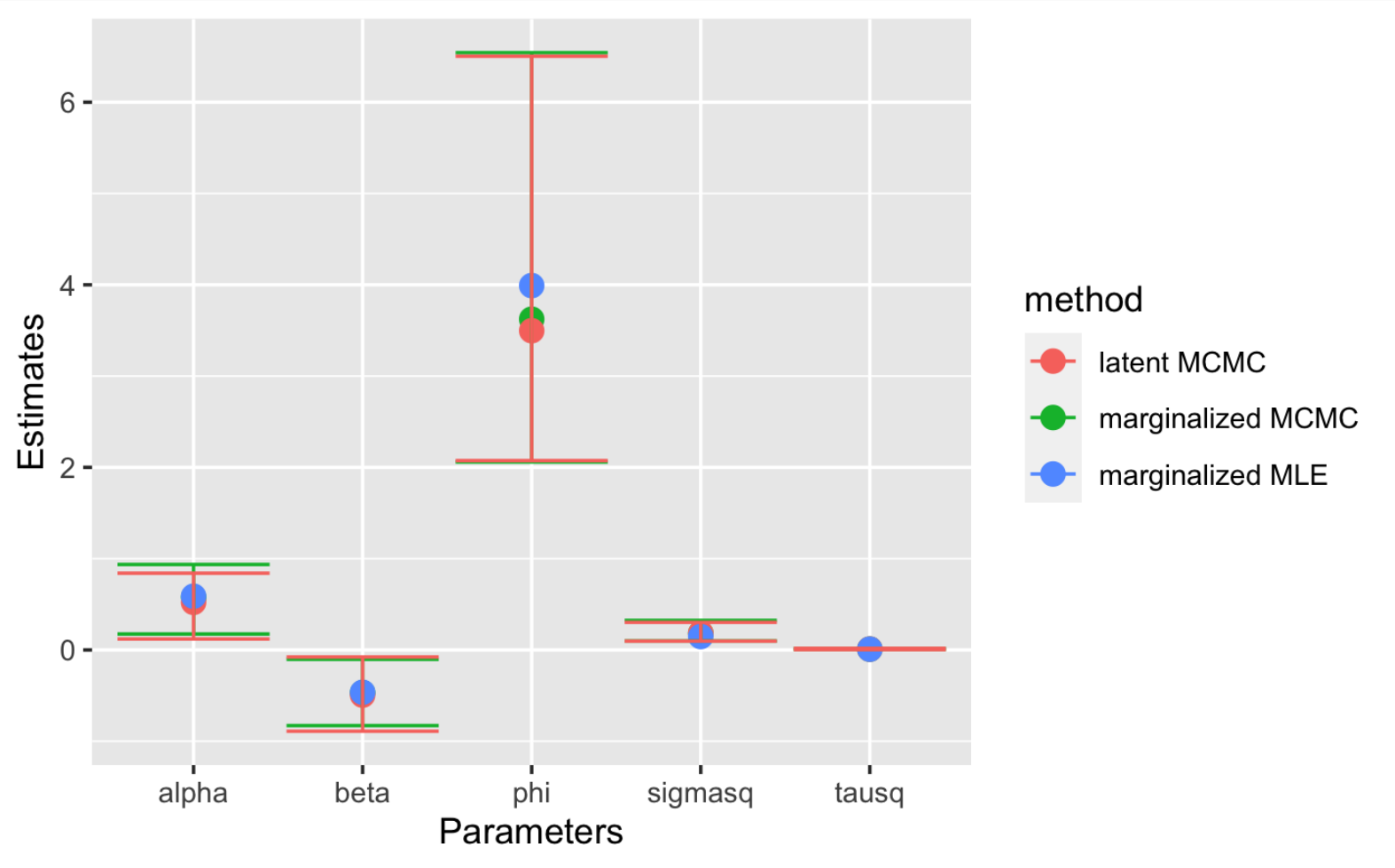


Figure: Posterior means and 2.5% and 97.5% quantiles from the latent and marginalized model, along with MLE for the marginalized model for Dataset 3.

Latent vs marginalized model

- Unmarginalized model has n additional parameters w
- May lead to slow MCMC convergence

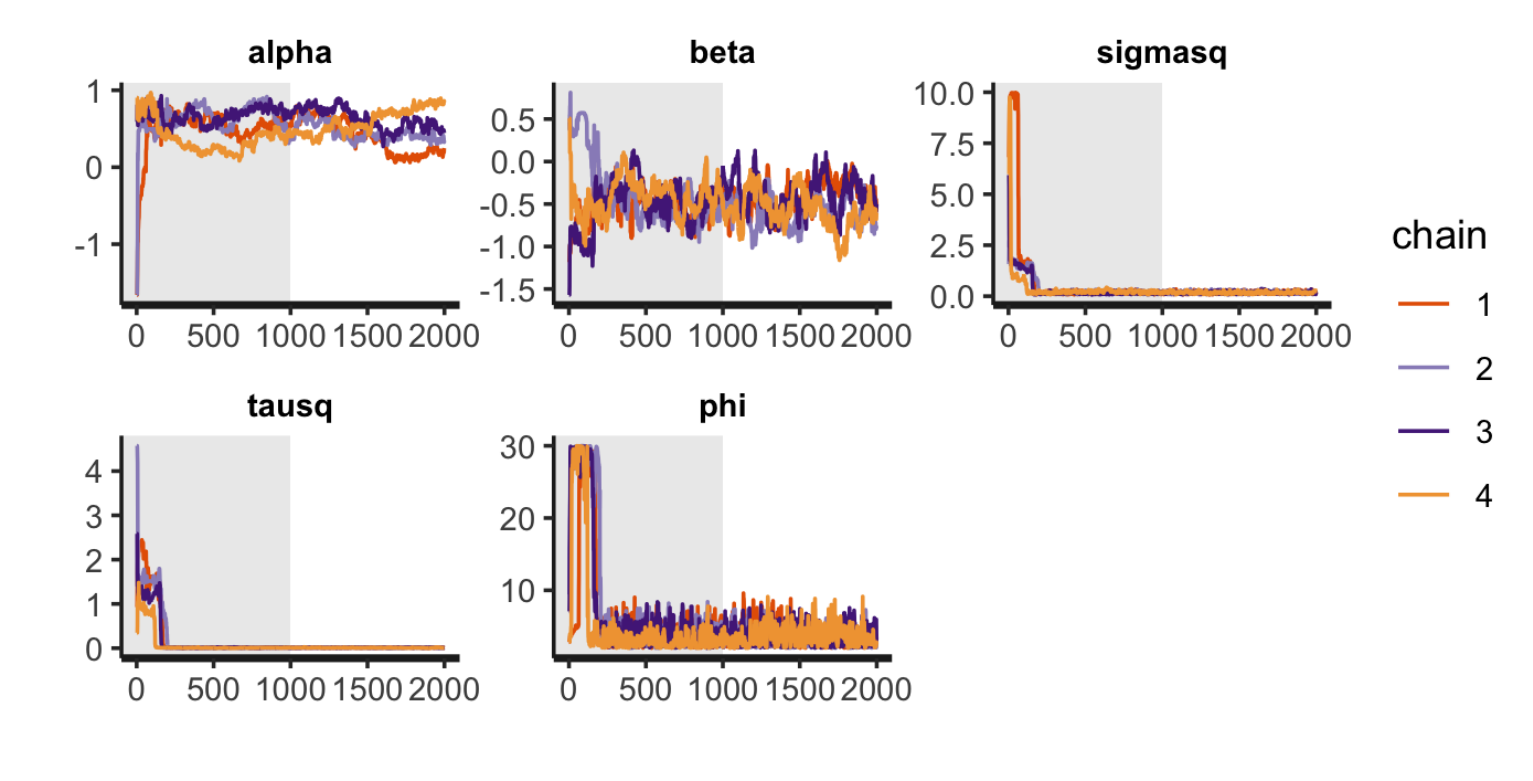


Figure: Trace plots for the latent model for Dataset 3.

Latent vs marginalized model

- Unmarginalized model has n additional parameters w
- May lead to slow MCMC convergence

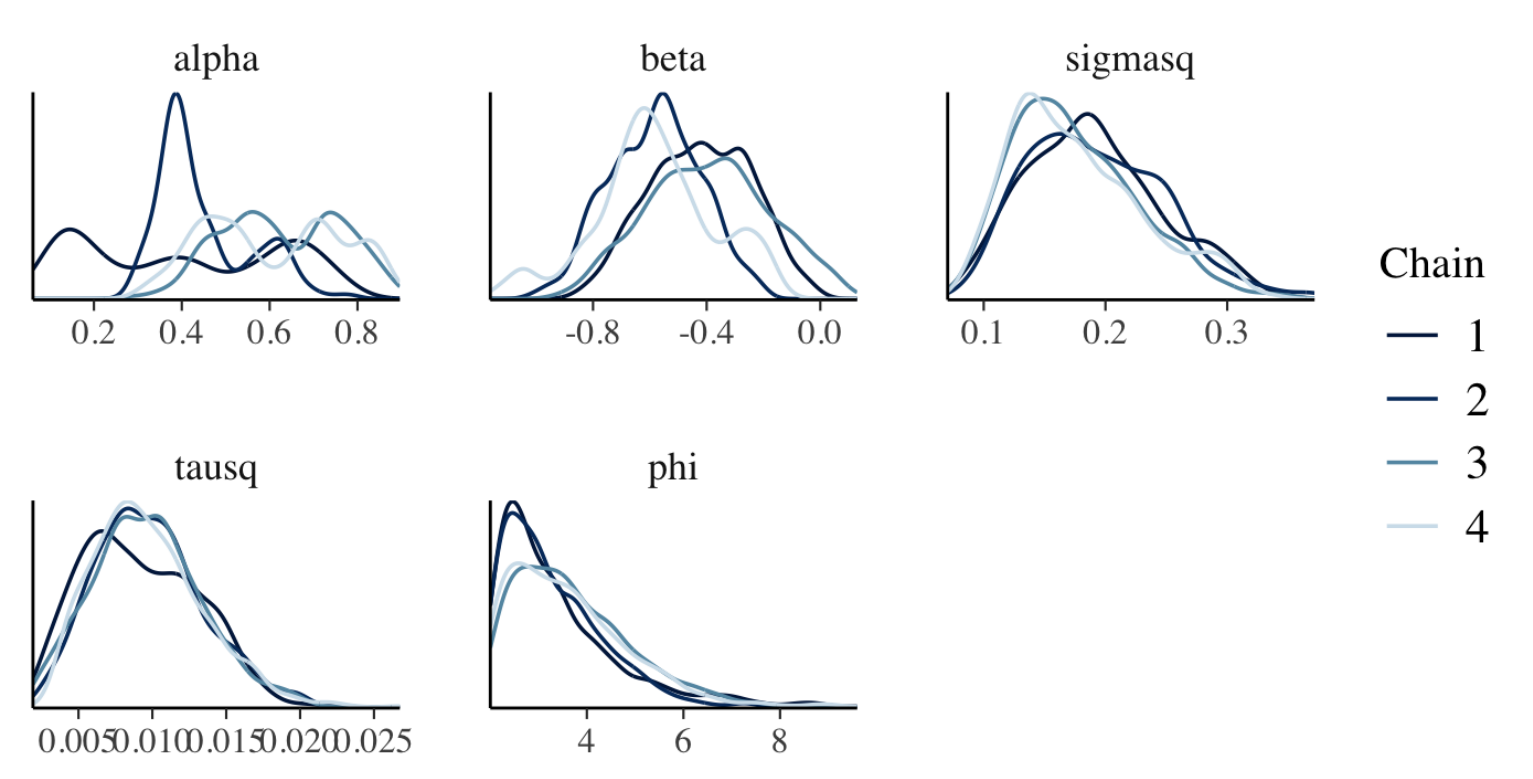


Figure: Density plots for the latent model for Dataset 3.

Latent vs marginalized model

- Unmarginalized model has n additional parameters w
- May lead to slow MCMC convergence

	Rhat	n_eff	se_mean
alpha	2.1252	8.1519	0.0660
beta	1.1916	17.0248	0.0514
sigmasq	1.0305	251.5815	0.0035
tausq	1.0092	183.8859	0.0003
phi	1.0487	172.2260	0.0908

Figure: MCMC diagnostics for the latent model for Dataset 3.

Latent vs marginalized model

- Unmarginalized model has n additional parameters w
- May lead to slow MCMC convergence
- Marginalized model only has $p + 3$ parameters
- The latent hierarchical model generalizes more easily to other settings (non-Gaussian response, spatial misalignment)
- The latent model has advantages in certain computational approximations for the likelihood for large data

Bayesian predictions

- To predict new observations $\tilde{y}$, based upon the observed data y, we specify a joint probability model $p(\tilde{y}, y \,|\, , \theta)$, which defines the conditional predictive distribution:

$$p(\tilde{y} \,|\, y, \theta) = \frac{p(\tilde{y}, y \,|\, , \theta)}{p(y \,|\, \theta)}.$$

- Posterior predictive distribution is $p(\tilde{y} \,|\, y) = \int p(\tilde{y} \,|\, y, \theta) p(\theta \,|\, y) d\theta$.
- 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 $\tilde{y}^{(j)} \sim p(\tilde{y} \,|\, y, \theta^{(j)})$
- The $\{\tilde{y}^{(j)}\}_{j=1}^{M}$ are samples from the posterior predictive distribution $p(\tilde{y} \,|\, y)$.

Bayesian predictions from the (non-spatial) linear model

- Suppose we have observed the new predictors $\tilde{X}$, and we wish to predict the outcome $\tilde{y}$. We specify $p(\tilde{y}, y \mid \theta)$ to be a normal distribution:

$$\begin{pmatrix} y \\ \tilde{y} \end{pmatrix} \sim N\left(\begin{bmatrix} X \\ \tilde{X} \end{bmatrix} \beta, \sigma^2 I\right)$$

- Note $p(\tilde{y} \mid y, \beta, \sigma^2) = p(\tilde{y} \mid \beta, \sigma^2) = N(\tilde{y} \mid \tilde{X}\beta, \sigma^2 I)$.

- The posterior predictive distribution:

$$\begin{aligned} p(\tilde{y} \mid y) &= \int p(\tilde{y} \mid y, \beta, \sigma^2) p(\beta, \sigma^2 \mid y) d\beta d\sigma^2 \\ &= \int p(\tilde{y} \mid \beta, \sigma^2) p(\beta, \sigma^2 \mid y) d\beta d\sigma^2. \end{aligned}$$

- By now we are comfortable evaluating such integrals:
  - First obtain: $(\beta^{(j)}, \sigma^{2(j)}) \sim p(\beta, \sigma^2 \mid y)$, $j = 1, \ldots, M$
  - Next draw: $\tilde{y}^{(j)} \sim N(\tilde{X}\beta^{(j)}, \sigma^{2(j)} I)$.

Bayesian predictions latent spatial model

- For the latent model, total parameters $\theta^* = (\theta, \{w(s)\})$
- To predict at a new location s_0, we have $p(\tilde{y}(s_0)|y, \theta^*) = p(\tilde{y}(s_0)|\theta^*) = N(\tilde{y}(s_0)|x(s_0)'\beta + w(s_0), \tau^2)$.
- Posterior samples for $w(s)$ are already generated for all s in the training data locations S
- Posterior predictive distributions $\tilde{y}(s_0)$ can be obtained using composition sampling:
  - For $s_0 \notin S$, generate samples from $w(s_0) \,|\, w, \theta$ using kriging

$$w(s_0)^{(j)} \sim N\left(r_{s_0}^{(j')}(R^{(j)})^{-1}w, \sigma^{2(j)}(1 - r_{s_0}^{(j')}(R^{(j)})^{-1}r_{s_0}^{(j')})\right)$$

  - $r_{s_0}^{(j')} = cor(w(s_0), w|\phi^{(j)})$ and $R^{(j')} = cor(w|\phi^{(j)})$
  - For any s, generate $\tilde{y}(s)^{(j)} = N(x(s)'\beta^{(j)} + w(s)^{(j)}, \tau^{2(j)})$

Bayesian predictions for the marginalized model

- Two ways to predict $\tilde{y}(s) \,|\, y$
- Via recovering w
  - The marginalized model integrates out the w's
  - With $\Sigma = \sigma^2 R + \tau^2 I$, the joint distribution of w and y is

$$\begin{pmatrix} w \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ X\beta \end{pmatrix}, \begin{pmatrix} \sigma^2 R & \sigma^2 R \\ \sigma^2 R & \Sigma \end{pmatrix} \right)$$

  - We can recover them after the MCMC using

$$w \,|\, y, \theta \sim N\left(\sigma^2 R \Sigma^{-1}(y - X\beta),\ \sigma^2 R - \sigma^4 R \Sigma^{-1} R\right)$$

  - Generate $w(s_0) \,|\, w, \theta$ and then $\tilde{y}(s_0) \,|\, w(s_0), \theta$ similar to the latent model
- Direct approach (not requiring samples of w):
  - Use kriging for the response process $y(s)$ to generate

$$\tilde{y}(s_0) \,|\, y, \theta \sim N\left(x(s_0)'\beta + \sigma^2 r_{s_0}' \Sigma^{-1}(y - X\beta),\ \tau^2 + \sigma^2 - \sigma^4 r_{s_0}' \Sigma^{-1} r_{s_0}\right)$$

Bayesian predictions example

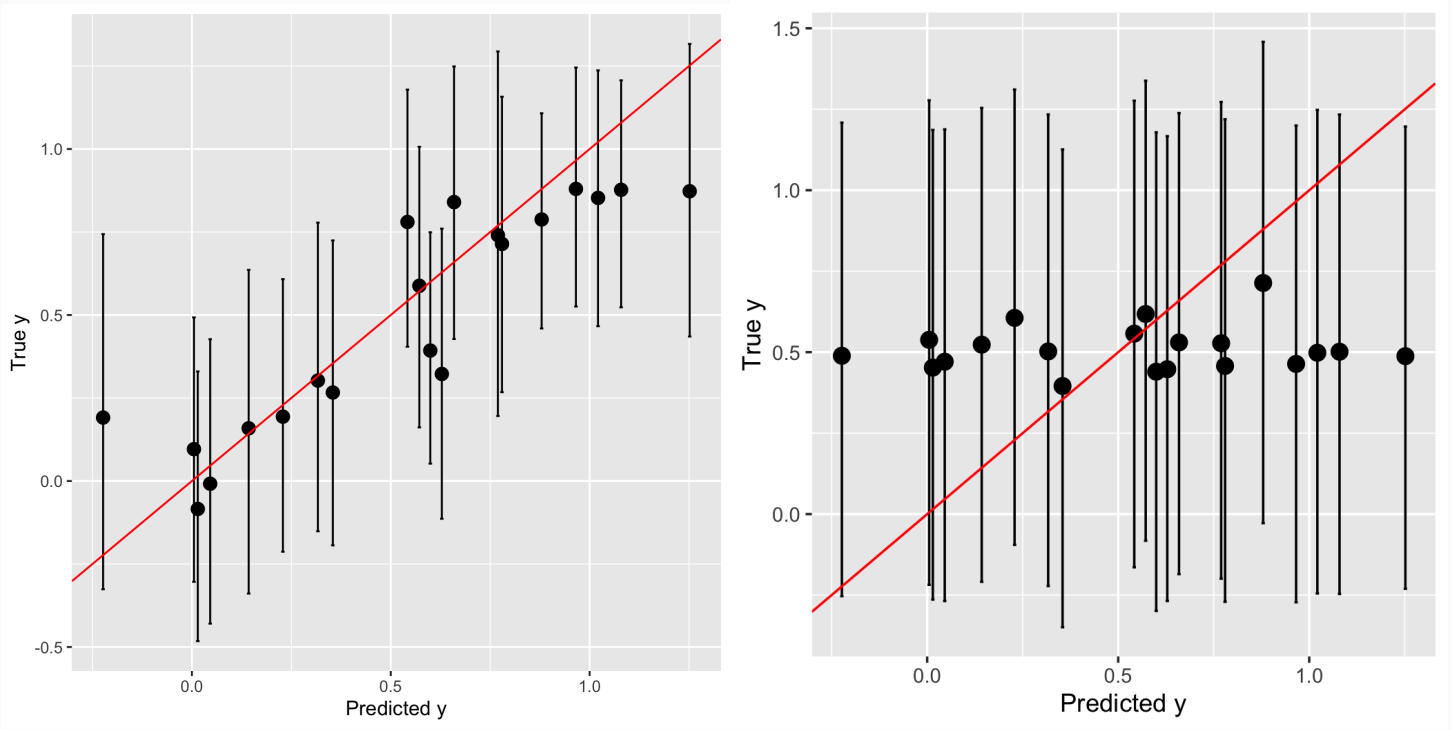


Figure: Out-of-sample predictions (posterior means and 95% prediction intervals from the latent spatial model (left) and non-spatial linear model (right) and for Dataset 3.

Model comparison using Bayesian output

- For Bayesian models using MCMC sampling, we can use Deviance Information Criterion (DIC) (Spiegelhalter, 2002 [1])
- DIC for the likelihood $p(y|\theta) = \prod_i p_i(y_i|\theta)$ is based on the deviance $D(y, \theta) = -2 \log p(y|\theta) = -2 \sum_{i=1}^{n} \log p_i(y_i|\theta)$
- $DIC = D(y, E_{\theta|y}(\theta)) + 2p_{DIC} = -2 \log p(y|E_{\theta|y}(\theta)) + 2p_{DIC}$
- $p_{DIC} = E_{\theta|y}\left(D(y, \theta)\right) - D(y, E_{\theta|y}(\theta))$ can be interpreted as effective number of parameters in the model and hence penalizes more complex models
- Similar to AIC for MLE: $AIC = -2 \log(y|\hat{\theta}_{MLE}) + 2k$, k is the number of parameters

[1] https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353

Model comparison using Bayesian output

- For Bayesian models using MCMC sampling, we can use Deviance Information Criterion (DIC) (Spiegelhalter, 2002 [1])
- DIC for the likelihood $p(y|\theta) = \prod_i p_i(y_i|\theta)$ is based on the deviance $D(y, \theta) = -2 \log p(y|\theta) = -2 \sum_{i=1}^{n} \log p_i(y_i|\theta)$
- $DIC = D(y, E_{\theta|y}(\theta)) + 2p_{DIC} = -2 \log p(y|E_{\theta|y}(\theta)) + 2p_{DIC}$
- $p_{DIC} = E_{\theta|y}\left(D(y, \theta)\right) - D(y, E_{\theta|y}(\theta))$ can be interpreted as effective number of parameters in the model and hence penalizes more complex models
- Similar to AIC for MLE: $AIC = -2 \log(y|\hat{\theta}_{MLE}) + 2k$, k is the number of parameters
- DIC can be calculated from the posterior samples as: $D(y, \frac{1}{M} \sum_{j=1}^{M} \theta^{(j)}) + 2(\frac{1}{M} \sum_{j=1}^{M} D(y, \theta^{(j)}) - D(y, \frac{1}{M} \sum_{j=1}^{M} \theta^{(j)}))$
- Many Bayesian and spatial R-packages offer DIC (`rjags`, `R2OpenBUGS`, `spBayes`)

[1] https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353

Model comparison using Bayesian output

- An alternative to DIC is the Widely applicable (or Watanabe-Akaike) Information Criterion (WAIC) (Watanabe, 2010) [2]
- $WAIC = -2 \log \prod_i E_{\theta|y} p_i(y_i|\theta) + 2p_{WAIC}$
- $p_{WAIC} = E_{\theta|y}\left(D(y,\theta)\right) + 2 \log \prod_i E_{\theta|y} p_i(y_i|\theta)$
- WAIC can be calculated from the posterior samples as:
  $-2 \sum_i \log \left(\frac{1}{M} \sum_{j=1}^{M} \log p_i(y_i|\theta^{(j)}) \right) +$
  $2 \left(\frac{1}{M} \sum_{j=1}^{M} D(y,\theta^{(j)}) + 2 \sum_i \log \left(\frac{1}{M} \sum_{j=1}^{M} \log p_i(y_i|\theta^{(j)}) \right) \right)$
- The `loo` package computes WAIC from a Stan output
- More details about WAIC in Gelman et al. (2014) [3]

[2] https://www.jmlr.org/papers/volume11/watanabe10a/watanabe10a.pdf

[3] http://www.stat.columbia.edu/~gelman/research/published/waic_understand3.pdf

Model comparison using Bayesian output

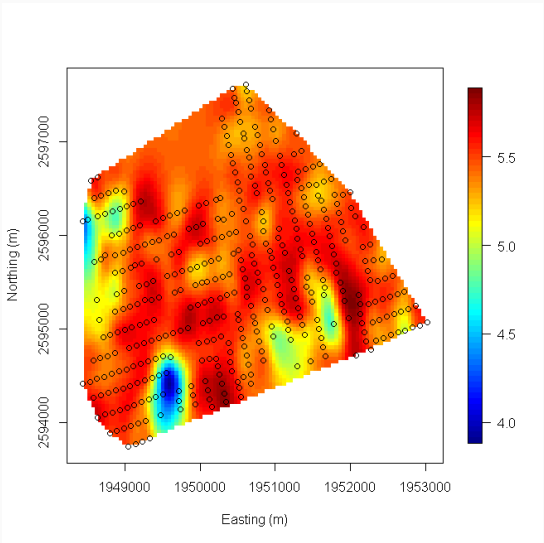
Table: WAIC for spatial and non-spatial linear models for Dataset 3.

Model	WAIC
Non-spatial	84
Spatial (latent)	-90

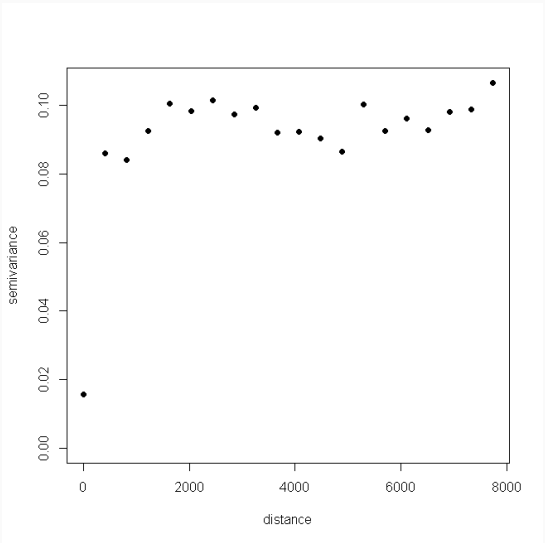
Of course, with holdout data we can also use the usual prediction based comparison metrics like RMSPE (using posterior means or medians), out-of-sample CP (coverage probability) and CIW (confidence interval width) based on posterior quantiles

BEF data analysis in spBayes

- Dataset available in spBayes on long-term research studies on the Bartlett Experimental Forest, Bartlett, NH
- Forest inventory data for 437 locations
- Variables include species specific basal area and biomass; inventory plot coordinates; slope; elevation; and tasseled cap brightness (TC1), greenness (TC2), and wetness (TC3) components from spring, summer, and fall 2002 Landsat images



log biomass



Variogram

MCMC free Bayesian inference

- The marginalized model can be reparametrized as: $N(y, X\beta, \sigma^2(R(\phi) + \alpha I))$ where $\alpha = \tau^2/\sigma^2$
- If ϕ and α is fixed, we can do exact conjugate sampling with Normal prior $\beta \sim (\mu, \sigma^2 V)$ for $\beta|\sigma^2$ and Inverse-Gamma prior for σ^2. *bayesGeostatExact* does that
- Fixed values of ϕ and α can be chosen from the variogram

	2.5%	25%	50%	75%	97.5%
(Intercept)	-0.624	0.267	0.728	1.182	2.079
Elevation	0.000	0.001	0.001	0.001	0.001
Slope	-0.017	-0.013	-0.011	-0.008	-0.004
Brightness	-0.001	0.006	0.010	0.013	0.021
Greenness	0.000	0.004	0.007	0.009	0.014
Wetness	0.015	0.021	0.024	0.028	0.034
σ^2	0.072	0.079	0.083	0.087	0.094
τ^2	0.014	0.016	0.016	0.017	0.019

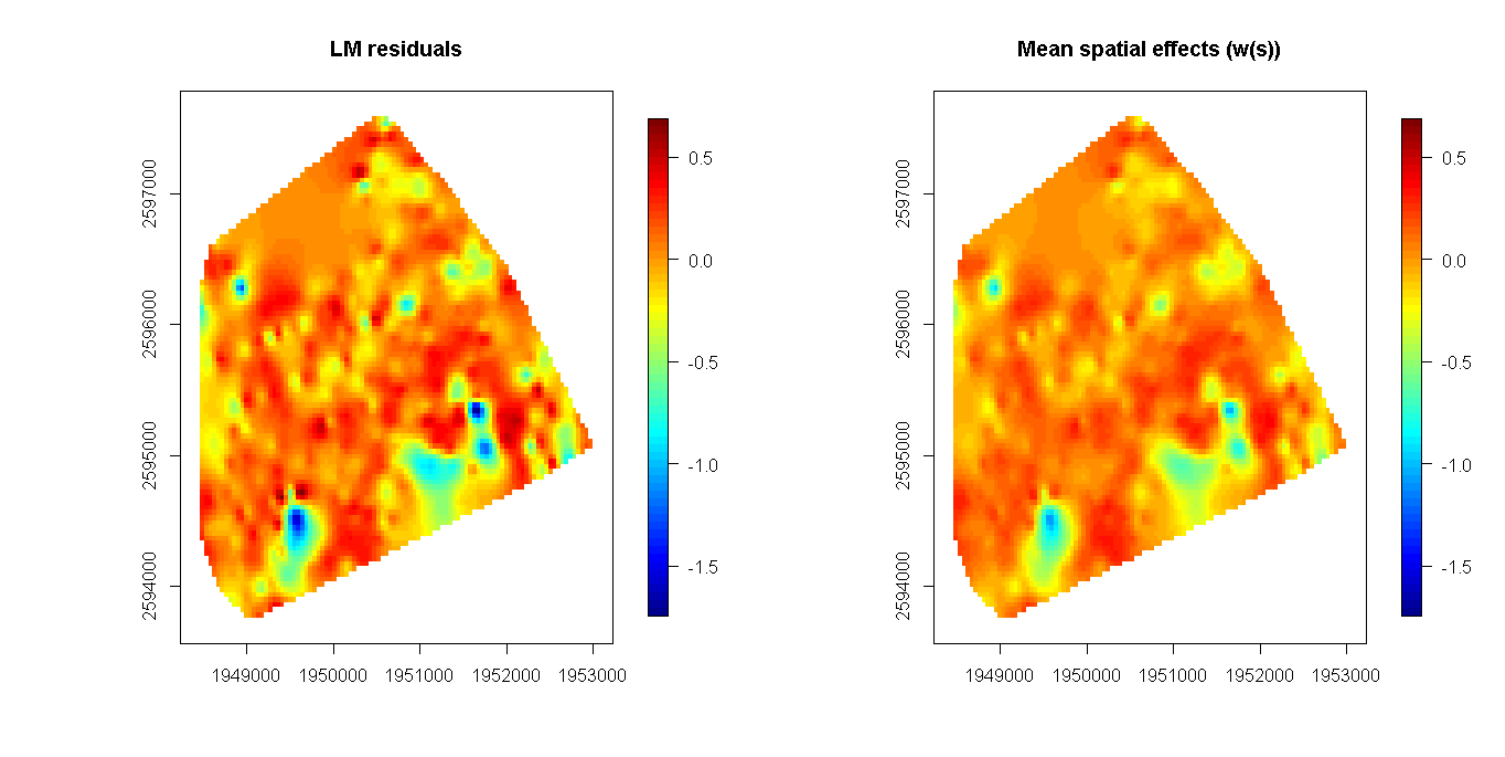
Full Bayesian inference

- *spLM* function
- Even marginalizes out β to make the chain only 3 dimensional

	2.5%	25%	50%	75%	97.5%
(Intercept)	-0.253	0.937	1.586	2.069	3.189
Elevation	0.000	0.000	0.000	0.001	0.001
Slope	-0.017	-0.011	-0.008	-0.005	0.002
Brightness	-0.005	0.006	0.010	0.015	0.025
Greenness	-0.005	0.003	0.005	0.008	0.014
Wetness	0.007	0.015	0.019	0.023	0.032
σ^2	0.042	0.074	0.086	0.095	0.108
τ^2	0.005	0.010	0.015	0.030	0.063
ϕ	0.004	0.008	0.010	0.012	0.016

Recovery of $w(s)$

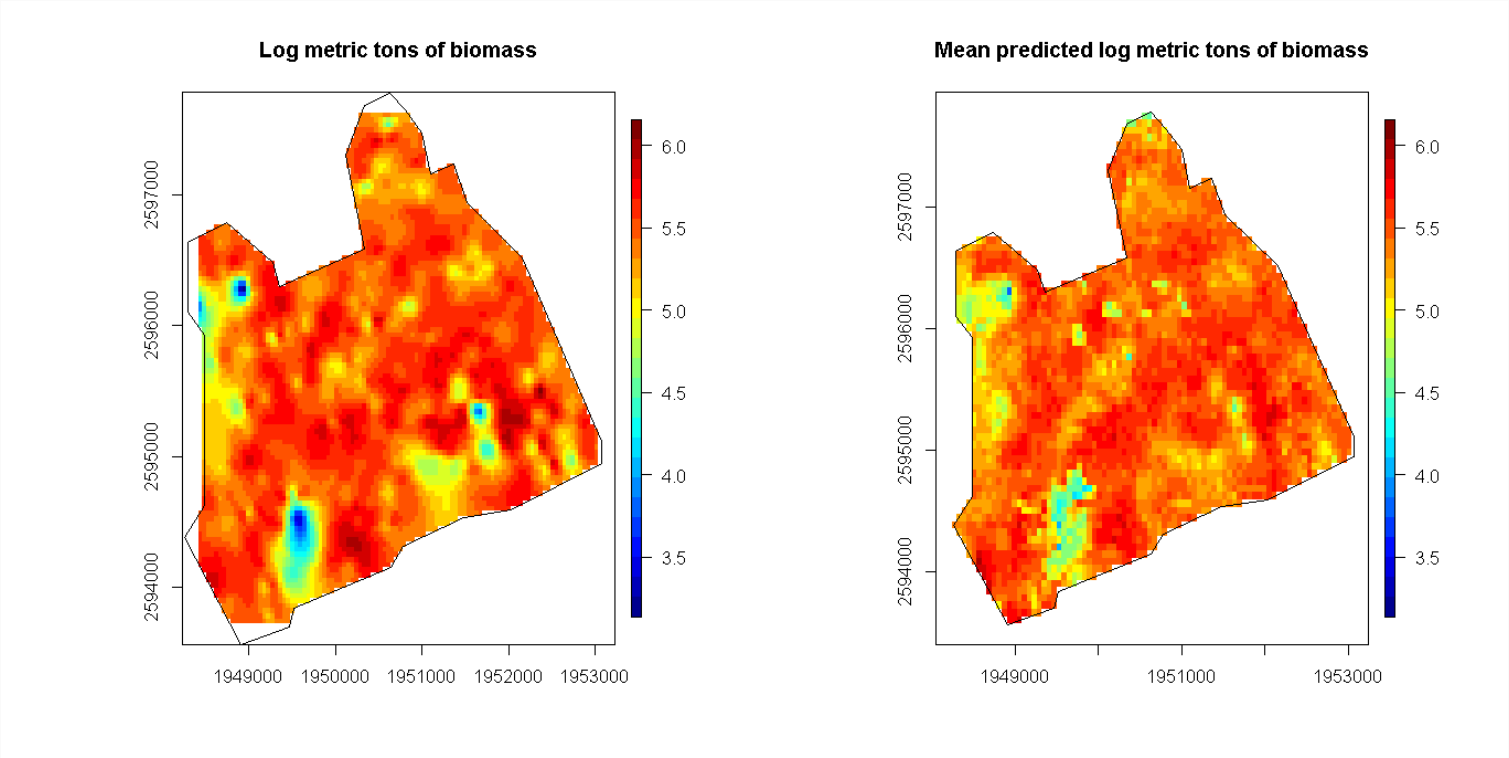
- *spRecover* function recovers both β and w



- *spDiag* calculates the DIC after recovery of w

Kriging

- *spPredict* function

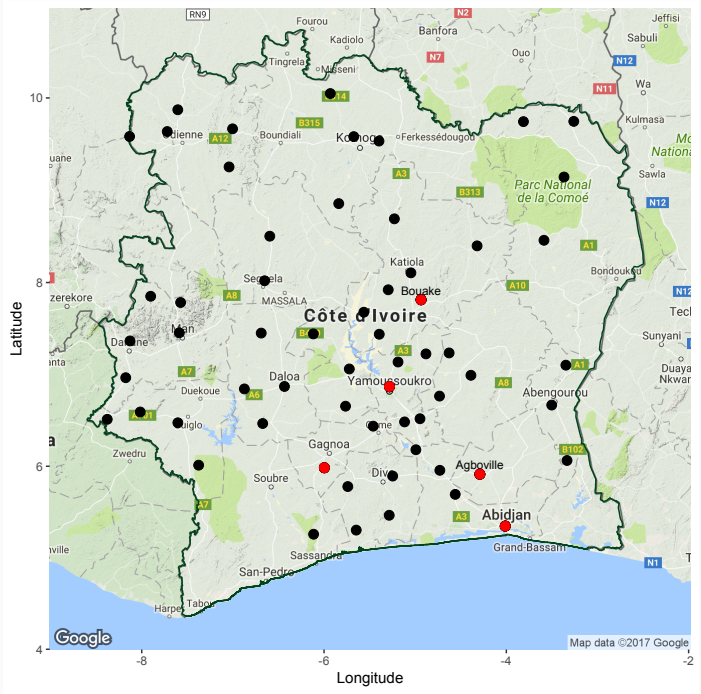


spBayes vs STAN

- STAN is a general package for running MCMC
- spBayes specializes in spatial GP regression and should be the preferred choice if the spatial (generalized) linear model is the suitable choice for the data analysis
- However, STAN is a great tool for more complex Bayesian models involving spatial data

A hierarchical modeling example

- Goal: Prediction of MSM population size at 61 regions of Côte d'Ivoire
- Data on MSM population (as proportion of the relevant male population) available for only 5 cities (red dots)
- Need to predict MSM population size at the remaining 56 regions (black dots) using a linear regression model



A hierarchical modeling example

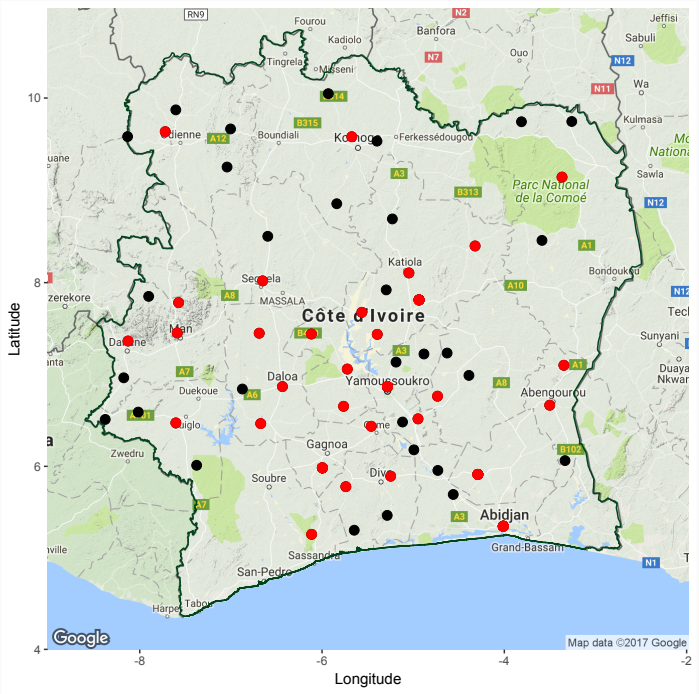
- Goal: Prediction of MSM population size at 61 regions of Côte d'Ivoire
- Data on MSM population (as proportion of the relevant male population) available for only 5 cities (red dots)
- Need to predict MSM population size at the remaining 56 regions (black dots) using a linear regression model
- Multiple datapoints available for each of the 5 cities (from multiple surveys)
- Preliminary analysis suggests important covariates are log of male population and HIV prevalence
- Details at *Datta et al., "Bayesian estimation of MSM population in Côte d'Ivoire", Statistics and Public Policy, (2019).*

Predicting MSM in Côte d'Ivoire

- This seems to be a straight-forward regression model, $\log(\text{ MSM \%}) = \beta_0 + \beta_1 \log(\text{male popn.}) + \beta_2 \text{ HIV prev.} + \text{error}$

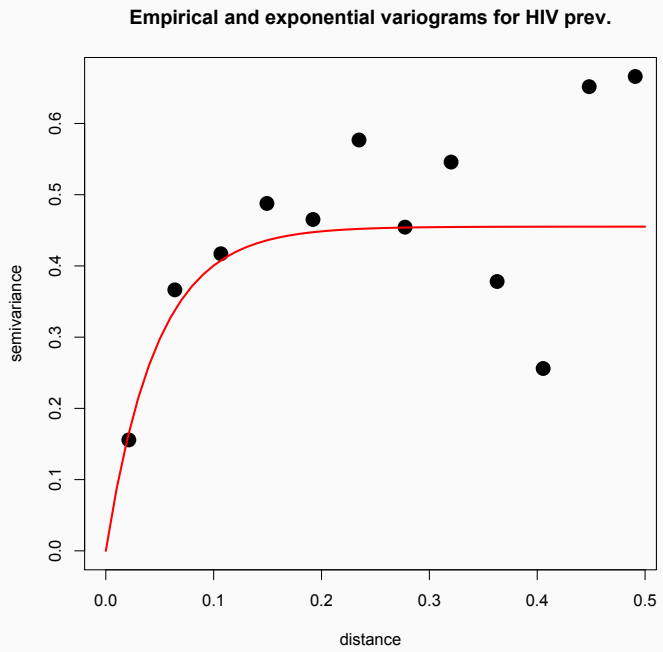
Predicting MSM in Côte d'Ivoire

- This seems to be a straight-forward regression model, $\log(\text{ MSM \%}) = \beta_0 + \beta_1 \log(\text{male popn.}) + \beta_2$ HIV prev. $+$ error
- HIV prevalence is missing in nearly 50% of the regions where we want to predict



Spatial model for HIV prevalence

- Empirical variogram of HIV prevalence suggests spatial dependence



- We can use kriging to impute missing HIV prevalence
- Leave-one-out cross validation suggests kriging offers 20% improved predictive accuracy than simple mean imputation for HIV prevalence

Hierarchical Bayesian model

- $y_j(s_i)$ is the estimate of population proportion for the i^{th} region based on the j^{th} survey
- $N(s_i)$ is male population, $H(s_i)$ is HIV prevalence

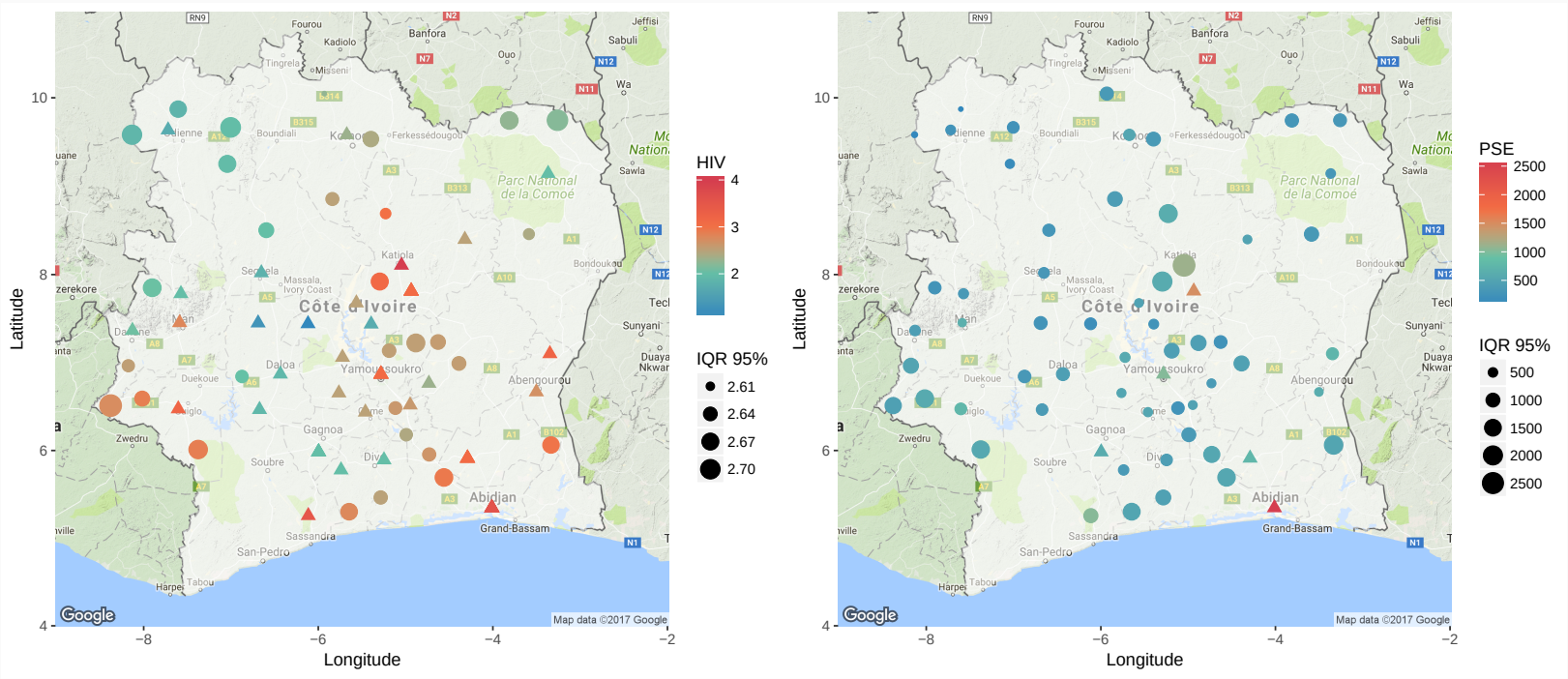
$$\prod_{i=1}^{5} \prod_j N(y_j(s_i) \mid \beta_0 + \beta_1 \log\{N(s_i)\} + \beta_2 H(s_i), \tau^2 w_{ij}) \times \quad \Leftarrow \text{Regression model}$$

$$N(H(S) \mid \mu 1, \Sigma(\sigma^2, \phi)) \times \quad \Leftarrow \text{Spatial model for HIV prevalence}$$

$$N(\beta \mid 0, 10^6 I) \times N(\mu \mid 0, 10^6) \times \text{Unif}(\phi \mid 0, 10) \times$$
$$\text{Gamma}(1/\tau^2 \mid 0.01, 0.01) \times \text{Gamma}(1/\sigma^2 \mid 2, 1) \quad \Leftarrow \text{Priors}$$

- We use a mean constant mean parameter in the second-stage GP model for HIV prevalence
- spBayes cannot implement hierarchical models like this
- We can use STAN to run the MCMC

Predictions and uncertainty estimates



HIV

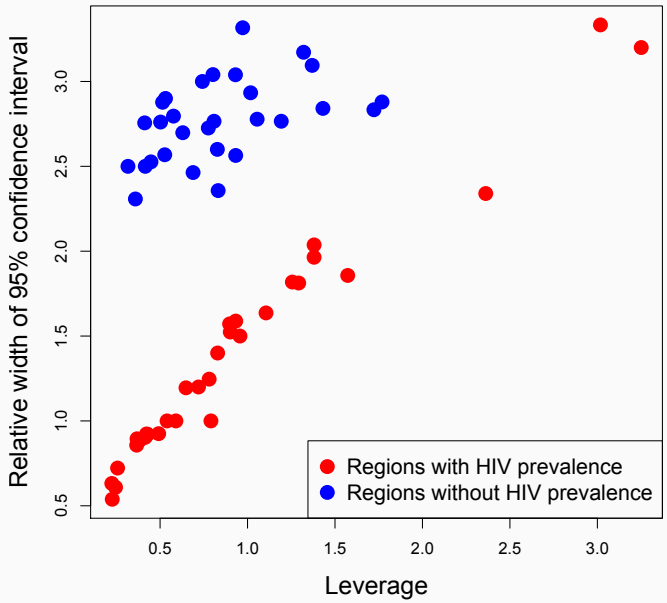
Population size

Value of a fully Bayesian model

- We can simply use a two step model where we first use kriging to predict HIV and use the predicted HIV in the regression predictions.
- Why do we need the fully Bayesian model that does everything together?

Value of a fully Bayesian model

- We can simply use a two step model where we first use kriging to predict HIV and use the predicted HIV in the regression predictions.
- Why do we need the fully Bayesian model that does everything together?



- Proper uncertainty quantification: Using the Bayesian model, regions with predicted HIV has higher uncertainty

Summary

- Using STAN package to run the MCMC for marginalizedd and latent models
- MCMC diagnostics
- Spatial predictions from Bayesian output
- Model comparison using Bayesian output (DIC and WAIC)
- MCMC-free and fully Bayesian spatial analysis using spBayes package