Latent Gaussian models: Approximate Bayesian inference (INLA)

Jo Eidsvik

January 30, 2018

Schedule

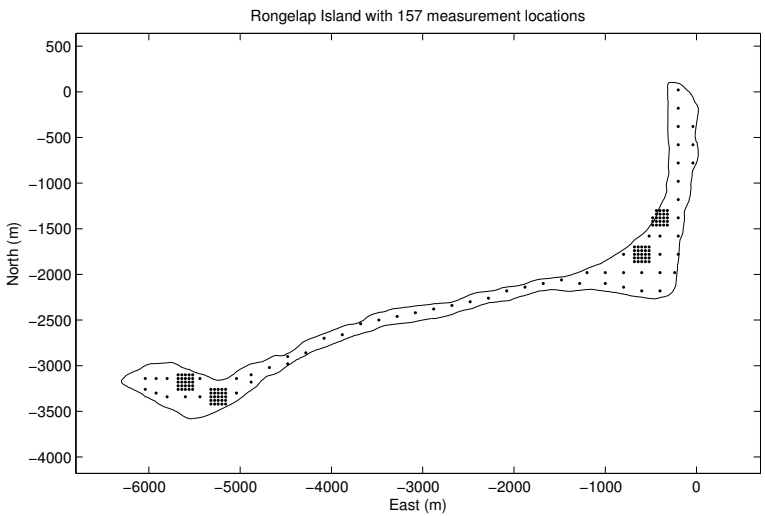
- 16 Jan: Gaussian processes (Jo Eidsvik)
- 23 Jan: Hands-on project on Gaussian processes (Team effort, work in groups)
- **30 Jan: Latent Gaussian models and INLA (Jo Eidsvik)**
- 6 Feb: Hands-on project on INLA (Team effort, work in groups)
- 12-13 Feb: Template model builder. (Guest lecturer Hans J Skaug)
- To be decided...

Plan for today

- Latent Gaussian models.
- Prior for parameters of Gaussian process. (We will be Bayesian today.)
- Laplace approximation and numerics for inference, INLA, (Rue et al., 2009)
- INLA shown for geostatistical applications. (Eidsvik et al., 2009)

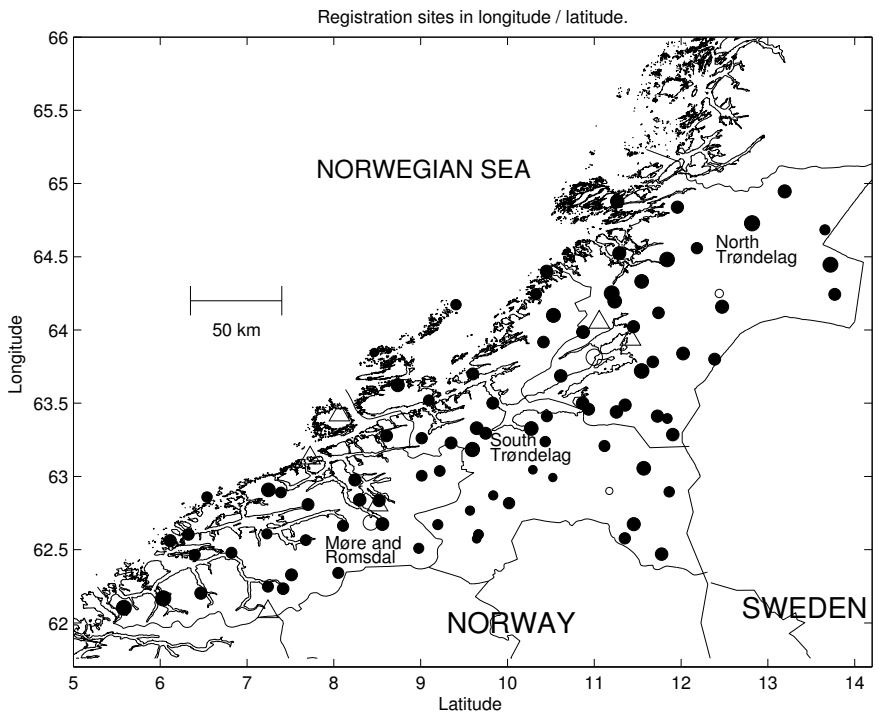
Examples of spatial latent Gaussian models

Radioactivity counts: Poisson



Example of spatial latent Gaussian models

Number of days with rain for $k = 92$ sites in September-October 2006.



Objective

Main goals:

- Fit model parameters of statistical covariance model.
- Predict latent intensity or risk at all spatial sites.

- Outlier detection.
- Spatial design.

Statistical model

Consider the following hierarchical model

1. Observed data $\boldsymbol{y} = (y_1, \ldots, y_k)$ where

$$\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\eta}) = \pi(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{k} \pi(y_i \mid x_{s_i})$$

   Often exponential family: Normal, Poisson, binomial, etc.
   $\log \pi(y_i|x_{s_i}) = \frac{y_i x_{s_i} - b(x_{s_i})}{a(\phi)} + c(\phi, y_i)$. $b(x)$ canonical link.

2. Latent Gaussian field $\boldsymbol{x} = (x_1, \ldots, x_k)$

$$\pi(\boldsymbol{x} \mid \boldsymbol{\eta}) = N[\mu \mathbf{1}_k, \ \boldsymbol{\Sigma}(\boldsymbol{\eta})]$$

3. Prior for hyperparameters $\pi(\boldsymbol{\eta})$

NOTE : Last point means we are Bayesian today!

Mixed models - Normal linear case

Common model

- $y_i = \boldsymbol{H}_i\boldsymbol{\beta} + x_i + \epsilon_i$
- y_i is observation. (Could be y_{ij}, individual or group i, replicate j.)
- $\boldsymbol{\beta}$ fixed effect. Prior $\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$.
- x_i Gaussian random effect having a structured covariance model with parameter $\boldsymbol{\eta}$. (Could be $\boldsymbol{U}_{ij}\boldsymbol{x}$ for group or individual i.)
- ϵ_i is random (unstructured) measurement noise. $\epsilon_i \sim N(0, \tau^2)$.

Mixed models - Inference

Can integrate out $\boldsymbol{\beta}$.

$$\pi(\boldsymbol{x}|\boldsymbol{\eta}) = \int \pi(\boldsymbol{x}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})d\boldsymbol{\beta} = N[\boldsymbol{H}\boldsymbol{\mu}_\beta, \boldsymbol{H}\boldsymbol{\Sigma}_\beta\boldsymbol{H}' + \boldsymbol{\Sigma}(\boldsymbol{\eta})]$$

Still a challenge to do inference on $\boldsymbol{\eta}$.

Mixed models - Inference

Common situation that has been hard to infer effectively:

- Frequentist, $\hat{\boldsymbol{\eta}}$: Laplace approximations or estimating equations.
- Bayesian $\pi(\boldsymbol{\eta}|\boldsymbol{y})$: Markov chain Monte Carlo.
- Inference not enough, wish to do model criticism, outlier detection, design, etc. Such goals require fast tools!

Mixed models - GLM

Likelihood is Poisson, binomial, or similar.

- Frequentist: Breslow and Clayton (1993).
- Bayes: Diggle, Tawn and Moyeed (1998), Christensen, Roberts and Sköld (2003), Diggle and Ribeiro (2007).

MCMC - Markov chain Monte Carlo

Around 2000 MCMC was very popular for inference and prediction in latent Gaussian models.
Today MCMC is still very popular, but not for latent Gaussian models.
These models are solved by Laplace approximations (Frequentist), or INLA (Bayes).

Typical MCMC algorithm

Initiate $\boldsymbol{\eta}^1$, $\boldsymbol{x}^1$.
Iterate for $i = 1, \ldots, B$

- Propose $\boldsymbol{\eta}^* | \boldsymbol{x}, \boldsymbol{y}$.
- Accept (Set $\boldsymbol{\eta}^{i+1} = \boldsymbol{\eta}^*$) or reject (Set $\boldsymbol{\eta}^{i+1} = \boldsymbol{\eta}^i$).
- For all j. Propose $x_j^* | \boldsymbol{x}_{1:j-1}^{i+1}, \boldsymbol{x}_{j+1:k}^{i}, \boldsymbol{\eta}^{i+1}, \boldsymbol{y}$. Accept ($x_j^{i+1} = x_j^*$) or reject ($x_j^{i+1} = x_j^i$).

Converges to sampling from the joint distribution. All properties of distribution can be extracted from MCMC samples.
Mixing of Markov chain can be very slow. Blocking helps, but not always.
Gibbs sampler requires conjugate priors. Fast updates, but mixing not better.

Inference

Posterior

$$\pi(\boldsymbol{x}, \boldsymbol{\eta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\eta})\, \pi(\boldsymbol{x} \mid \boldsymbol{\eta})\, \pi(\boldsymbol{y} \mid \boldsymbol{x})$$

In most cases the main tasks are:

- *PREDICTION:* Posterior marginals for x_j, $j = 1, \ldots, n$

$$\pi(x_j \mid \boldsymbol{y})$$

- *PARAMETER ESTIMATION:* Posterior marginals for η_j

$$\pi(\eta_j \mid \boldsymbol{y})$$

Inference

Split the joint density

$$\pi(\boldsymbol{x}, \boldsymbol{\eta}, \boldsymbol{y}) = \pi(\boldsymbol{\eta})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x}) = \pi(\boldsymbol{y})\pi(\boldsymbol{\eta} \mid \boldsymbol{y})\pi(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})$$

Clearly:

$$\pi(\boldsymbol{\eta} \mid \boldsymbol{y}) = \frac{\pi(\boldsymbol{\eta})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x})}{\pi(\boldsymbol{y})\pi(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})} \propto \frac{\pi(\boldsymbol{\eta})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x})}{\pi(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})}$$

Marginalization:

$$\pi(\boldsymbol{x}_j \mid \boldsymbol{y}) = \int_{\boldsymbol{\eta}} \pi(\boldsymbol{\eta} \mid \boldsymbol{y})\pi(\boldsymbol{x}_j \mid \boldsymbol{\eta}, \boldsymbol{y}) d\boldsymbol{\eta}$$

Inference

Laplace approximation

$$\hat{\pi}(\boldsymbol{\eta} \mid \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{\eta})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x})}{\hat{\pi}(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\hat{\boldsymbol{m}}(\boldsymbol{\eta},\boldsymbol{y})}$$

Use a *Gaussian* approximation $\hat{\pi}(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})$.
$\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}}(\boldsymbol{\eta}, \boldsymbol{y}) = argmax_{\boldsymbol{x}}[\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x})]$.

Approximate conjugacy

The Laplace approximation relies on approximate conjugacy. If the full conditional for $\boldsymbol{x}$ is Gaussian, the formula is exact. When we insert a Gaussian approximation at the mode, the approximation depends on the non-Gaussian likelihood. This cannot be bimodal.

$$\hat{\pi}(\boldsymbol{\eta} \mid \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{\eta})\pi(\boldsymbol{x}|\boldsymbol{\eta})\pi(\boldsymbol{y} \mid \boldsymbol{x})}{\hat{\pi}(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\hat{\boldsymbol{m}}(\boldsymbol{\eta},\boldsymbol{y})}$$

The error of the Laplace approximation (under weak regularity conditions) is *relative* and $\mathcal{O}(k^{-1})$ (Tierney and Kadane, 1986).

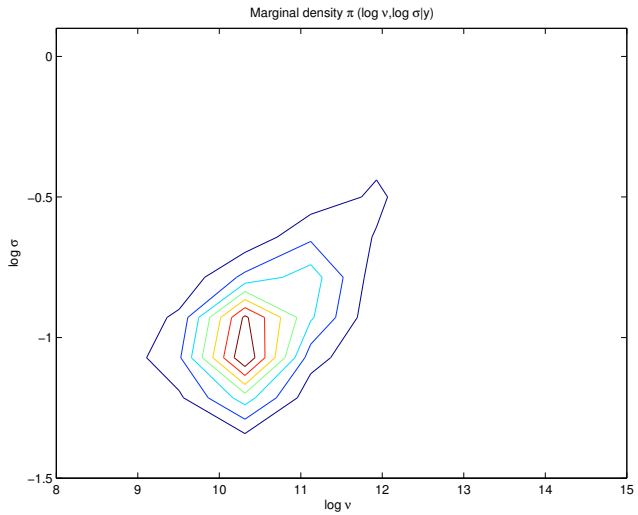
Gaussian approximation of full posterior

$$\pi(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \mu\mathbf{1}_n) + \sum_{i=1}^{k} \log \pi(y_i|x_{s_i})\right)$$

- $\log \pi(y_i|x_{s_i}) = \frac{y_i x_{s_i} - b(x_{s_i})}{a(\phi)} + c(\phi, y_i)$. $b(x)$ is canonical link.
- Expand GLM part $\log \pi(y_i|x_{s_i})$ to second order.
- Iterative solution to posterior mode $\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}}(\boldsymbol{\eta}, \boldsymbol{y})$. ('Scoring').
- $\hat{\boldsymbol{m}} = \mu\mathbf{1}_n - \boldsymbol{\Sigma}\boldsymbol{A}'[\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}' + \boldsymbol{P}]^{-1}(\boldsymbol{z}(\boldsymbol{y}, \hat{\boldsymbol{m}}) - \mu\boldsymbol{A}\mathbf{1}_n)$.
- Fit Gaussian approximation from Hessian at posterior mode: $\hat{\pi}(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y}) = N(\hat{\boldsymbol{m}}, \hat{\boldsymbol{V}})$.
- $\boldsymbol{P} = \boldsymbol{P}(\hat{\boldsymbol{m}})$. Size $k \times k$ matrix inversion required.

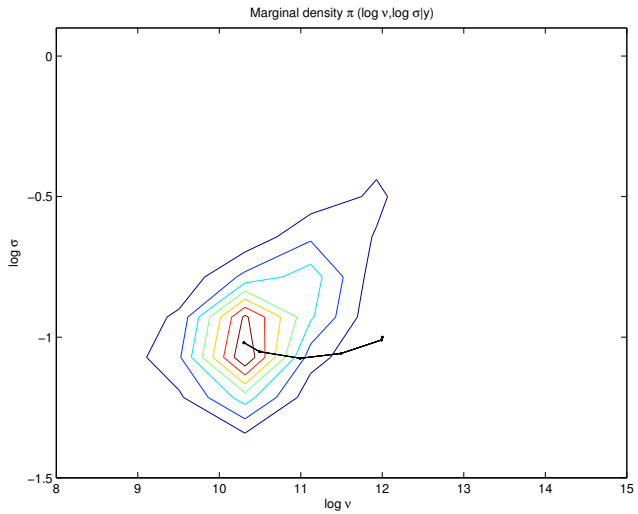
Practical implementation

Numerical approximation of $\tilde{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$



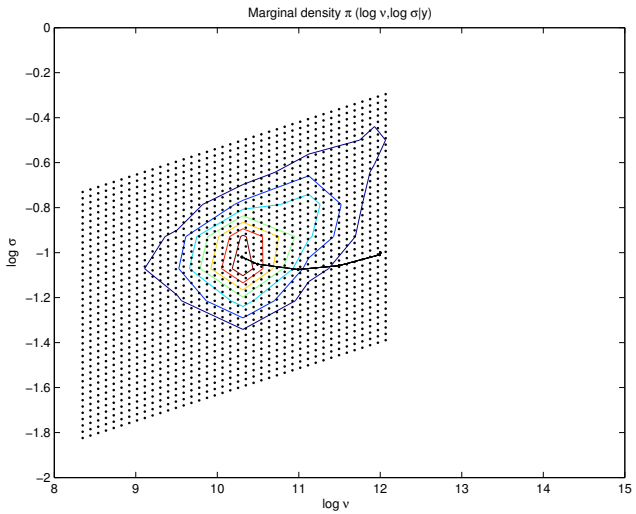
Practical implementation

Numerical approximation of $\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$, Step 1: Find mode



Each step requires $\boldsymbol{m}(\boldsymbol{\eta}, \boldsymbol{y})$, $\hat{\pi}(\boldsymbol{x} \mid \boldsymbol{\eta}, \boldsymbol{y})$ and Laplace.

Numerical approximation of $\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$, Step 2: Use Hessian at mode to set grid



Direct approximation of $\pi(x_j|\boldsymbol{y})$

Direct mixture approach for marginal prediction:

$$\hat{\pi}(x_j|\boldsymbol{y}) = \sum_l \hat{\pi}(x_j|\boldsymbol{\eta}_l, \boldsymbol{y})\hat{\pi}(\boldsymbol{\eta}_l|\boldsymbol{y})$$

$\hat{\pi}(x_j|\boldsymbol{\eta}_l, \boldsymbol{y}) = N(\hat{m}_j, \hat{V}_{j,j})$.
$\hat{m}_j = \hat{m}_j(\boldsymbol{\eta}_l, \boldsymbol{y})$, $\hat{V}_{j,j} = \hat{V}_{j,j}(\boldsymbol{\eta}_l, \boldsymbol{y})$.
Element j of posterior mode and j, j of full posterior covariance.

A frequentist solution would just plug in $\hat{\eta}$, the approximate MLE.

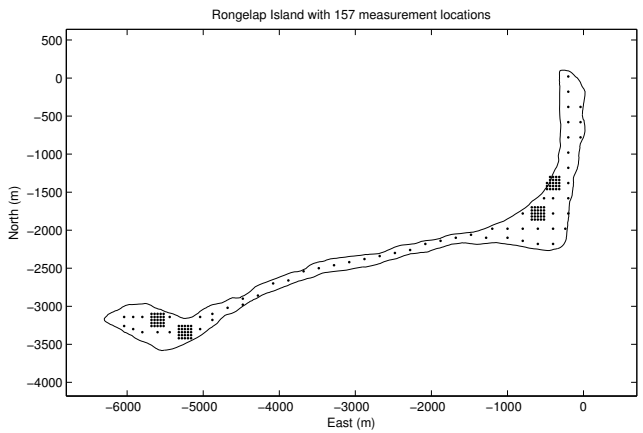
Nested approximation of $\pi(x_j|\boldsymbol{y})$

$$\pi(x_j|\boldsymbol{y},\boldsymbol{\theta}) \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})}{\pi(\boldsymbol{x}_{-j}|x_j,\boldsymbol{y},\boldsymbol{\theta})},$$

Using the Laplace approximation again, for fixed x_j.
$\hat{\pi}(\boldsymbol{x}_{-j}|x_j,\boldsymbol{y},\boldsymbol{\theta})$ approximated by a Gaussian (for each x_j on a grid or design points).

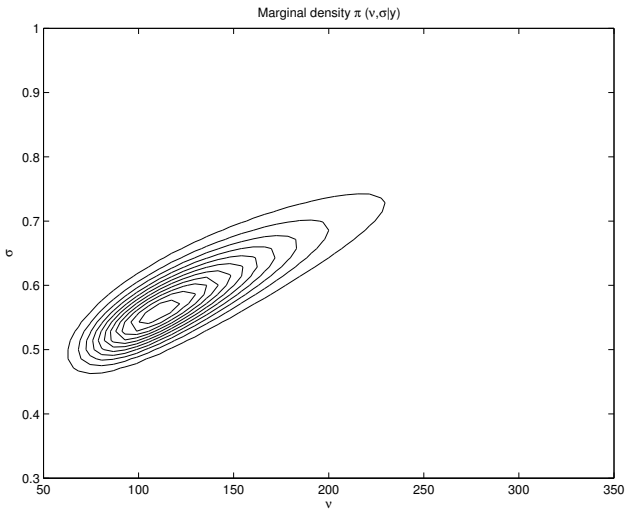
Example: Rongelap

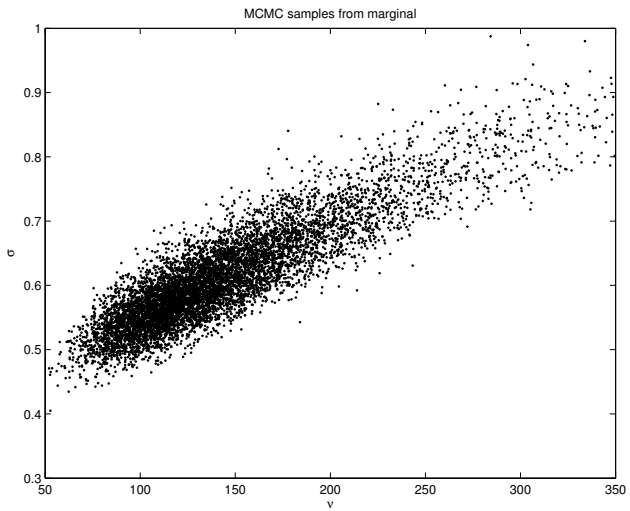
- Radioactivity counts at 157 registration sites. Poisson counts.
- $\boldsymbol{\Sigma}(\boldsymbol{\eta})$ defined from exponential covariance function.
  $\boldsymbol{\eta} = (\nu, \sigma)$, range and standard deviation.



Marginals for $\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$

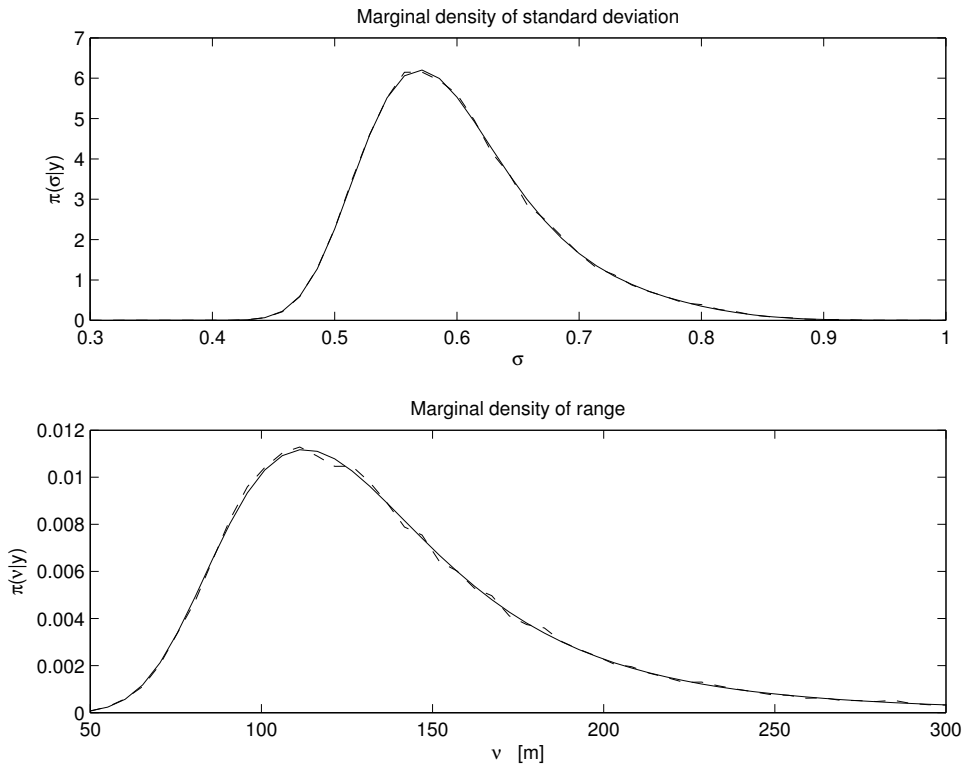
Laplace approximation+numerics (left) and solutions with MCMC (right). Left) Seconds. Right) Minutes.



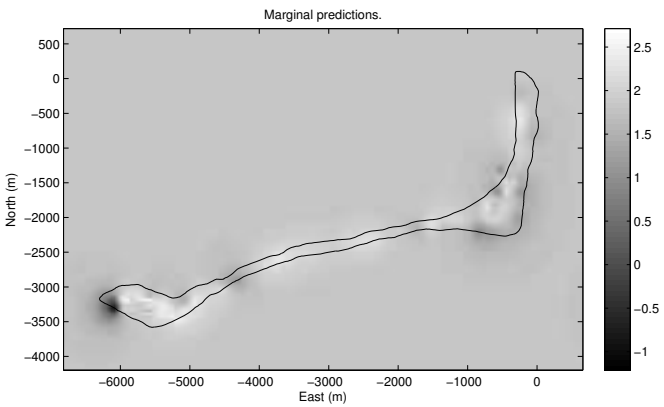


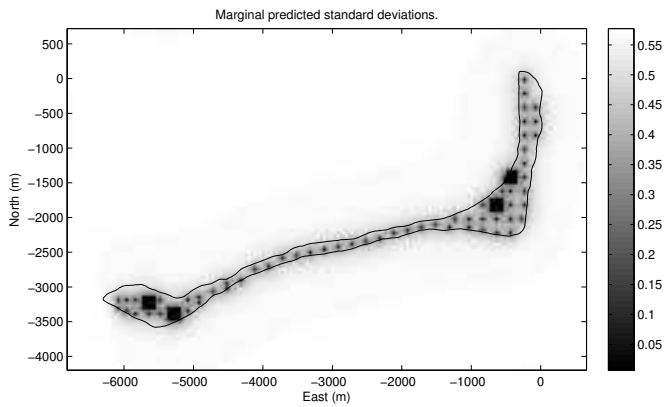
Marginals $\hat{\pi}(\boldsymbol{\eta}|\boldsymbol{y})$

Laplace approximation (solid) and MCMC (dashed).



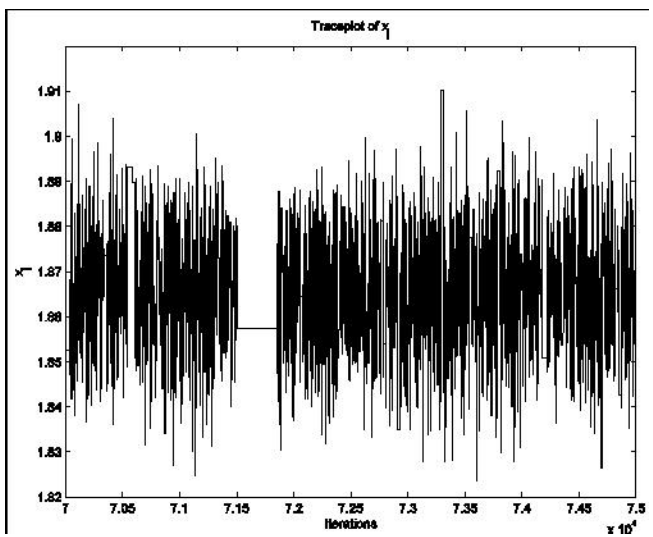
Prediction $\hat{E}(x_j|\boldsymbol{y})$ and $\hat{V}(x_j|\boldsymbol{y})$

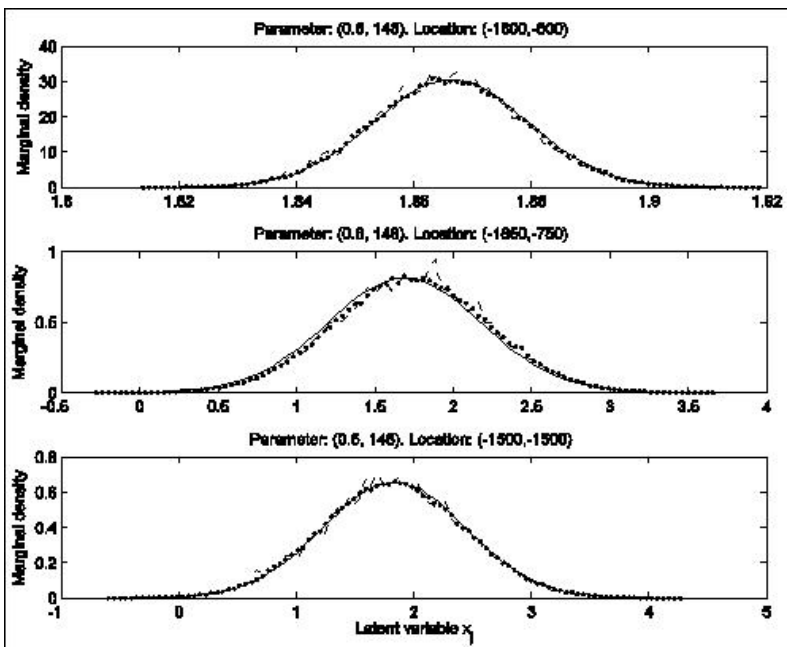




Marginals $\hat{\pi}(x_j|\eta, y)$

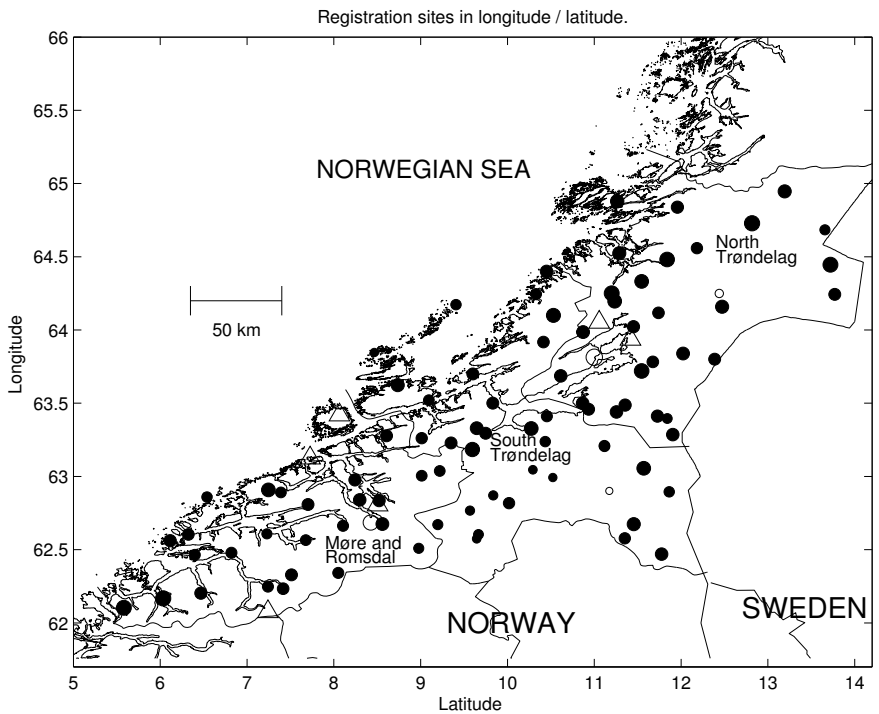
Conditional prediction at one spatial site MCMC (dashed), Importance sampling (dotted) and direct Gaussian approximation (solid).





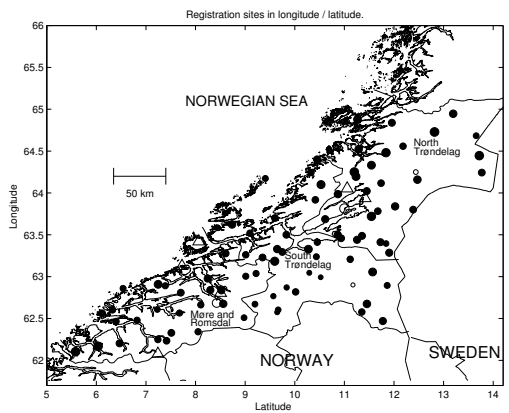
Example: Precipitation in Middle Norway

Number of days with rain for $k = 92$ sites in September-October 2006.



Example: Precipitation in Middle Norway

Binomial data $y_i = \text{Binomial}[\frac{e^{x s_i}}{1+e^{x s_i}}, 61]$.
Standard GLM gives no significance to East, North, Altitude.
Include only spatial trend.



- Outlier detection
- Spatial design

Outlier detection

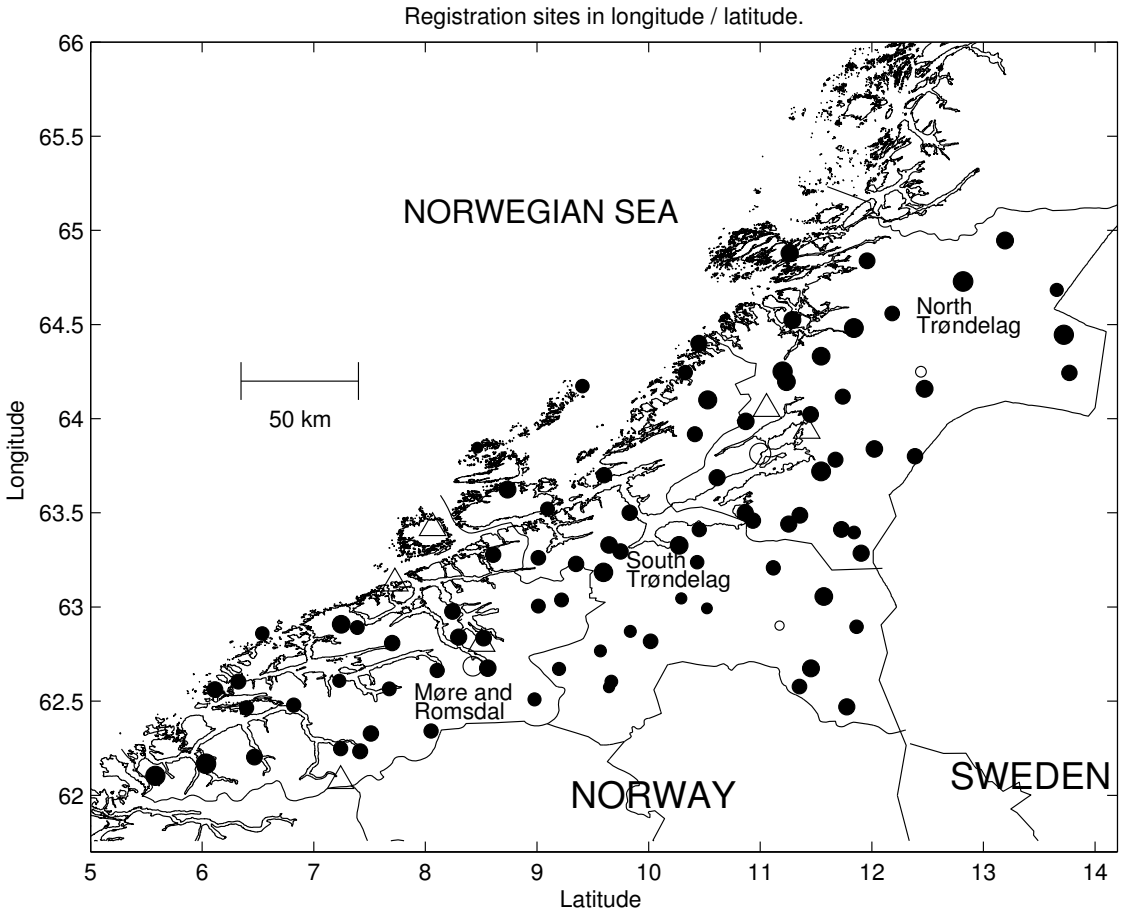
Use crossvalidation $\pi(y_i \mid \boldsymbol{y}_{-i})$.

$$\hat{\pi}(y_i \mid \boldsymbol{y}_{-i}) = \int_{x_{s_i}} \sum_l \hat{\pi}(\boldsymbol{\eta}_l \mid \boldsymbol{y}_{-i}) \hat{\pi}(x_{s_i} \mid \boldsymbol{\eta}_l, \boldsymbol{y}_{-i}) \pi(y_i \mid x_{s_i}) dx_{s_i}$$

Inference separately for each y_i. I.e. k times. Approximate predictive percentiles
$\sum_{y_i=0}^{y_{lower}} \hat{\pi}(y_i \mid \boldsymbol{y}_{-i}) = \alpha/2$, $\sum_{y_i=0}^{y_{upper}} \hat{\pi}(y_i \mid \boldsymbol{y}_{-i}) = 1 - \alpha/2$.
Compare (y_{lower}, y_{upper}) with observed y_i.

Results : Outlier detection

Results $\alpha/2 = 0.01$: detect 4 outliers (open circles).



Spatial design

Prospective view: $\boldsymbol{y} \rightarrow (\boldsymbol{y}, \boldsymbol{y}_a)$.
$\boldsymbol{y}_a$ extra data at 'new' spatial registration sites.
'Imagine' these observations - do not acquire them.
Design criterion is: Integrated prediction variance.

$$\hat{I} = \sum_{\boldsymbol{y}_a} \sum_{j} \hat{V}(x_j \mid \boldsymbol{y}, \boldsymbol{y}_a) \hat{\pi}(\boldsymbol{y}_a | \boldsymbol{y})$$

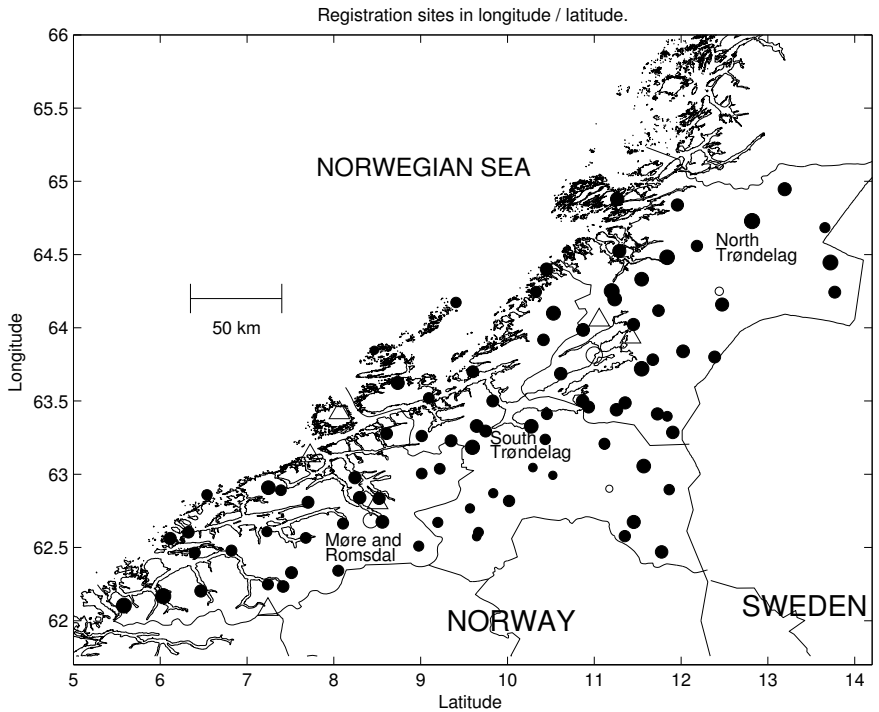
Results: Spatial design

Results of *three* design.

0: Existing design with 88 points (outliers excluded).

A: Currently installed stations, 88 plus 10 known sites (4 outlier sites and 6 sites out of service).

B: 88 plus $10 = 2 \cdot 5$ new random sites around two existing sites (50km radius).



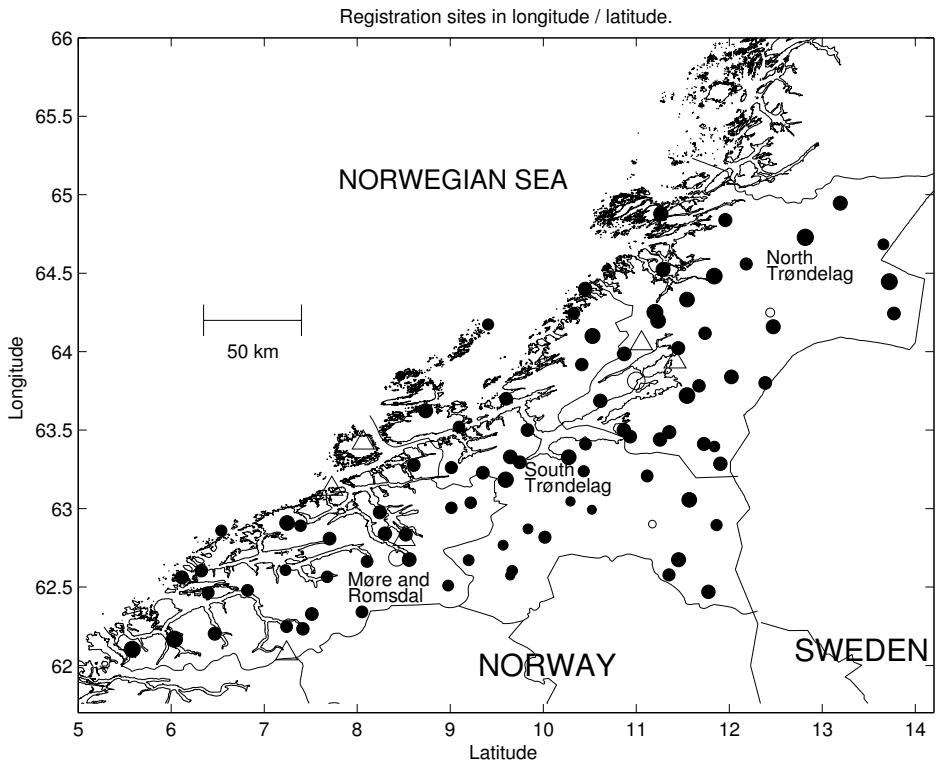
Results: Spatial design

Results of *three* designs.

0: Existing design: $\hat{I}_0 = 18.68$.

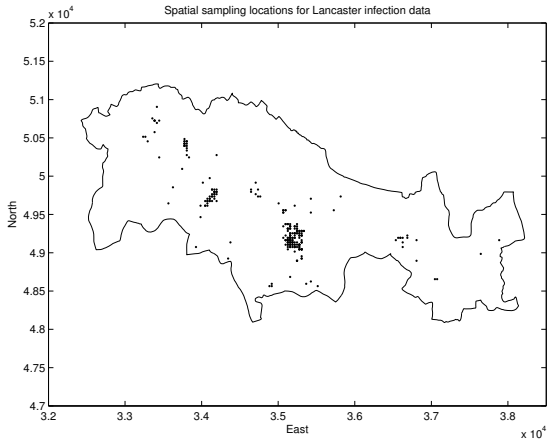
A: Currently installed stations: $\hat{I}_A = 17.94$.

B: Random around two existing sites: $\hat{I}_B = 17.85$



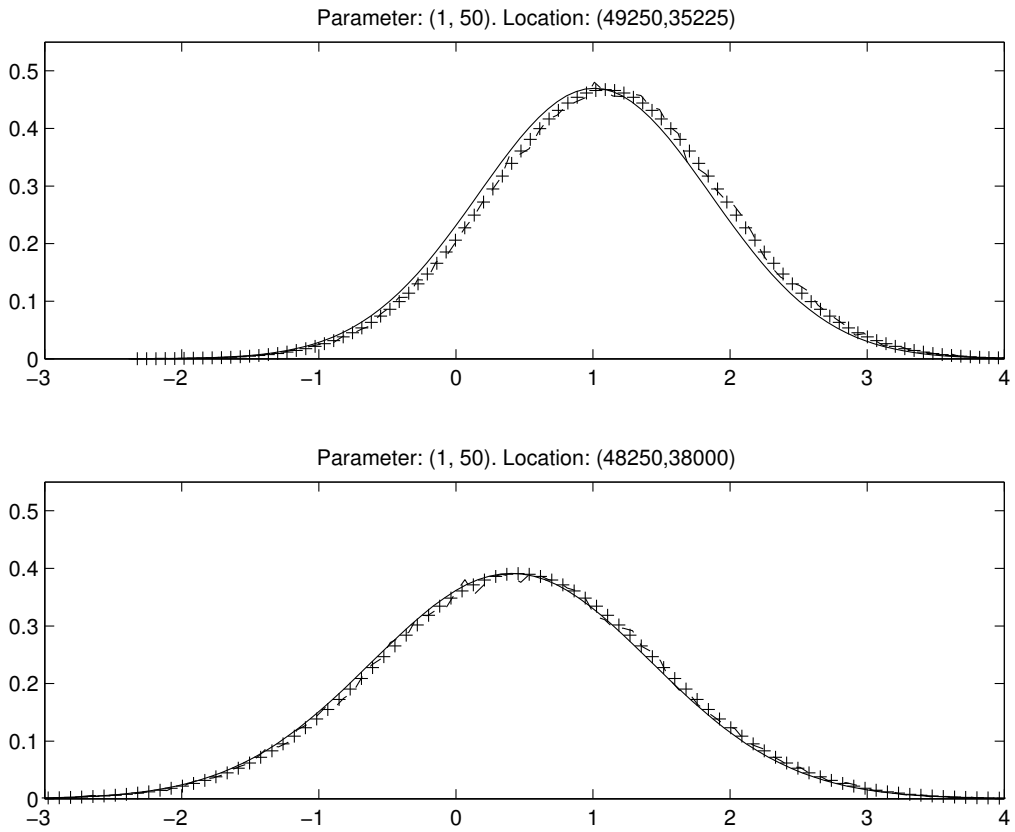
Example: Lancaster disease map

- Number of infections in different regions.
- Binomial data (with small counts).



Example: Lancaster disease map

LA, INLA and MCMC prediction at one site, for two parameter sets.



INLA contribution

- Mixed GLMs with latent Gaussian models cover wide range of applications
- The approximations work well for latent Gaussian models
- Generic routines. Software-friendly. Deterministic results (no Monte Carlo error)
- Enlarge scope of models

Conditions for INLA

- $\dim(\boldsymbol{\eta})$ is not too high
- No. of reg. sites $k < 10000$
- Marginals only. Bi-trivariate possible
- Likelihood must be well-behaved, not multimodal.

INLA vs MCMC

MCMC is very general. It explores all aspects of the joint posterior.
Approximate inference (INLA) is much faster. It is tailored to special tasks, such as marginals.
Applicable to much more than spatial data.

INLA software

INLA software: http://www.r-inla.org
Easy to call:
¿inla(y x + f(nu,model="iid"), family = c("poisson"), data = data, control.predictor=list(link=1))
Rue et al. (2009)
Routine runs on Gaussian Markov random fields.