TMA4250 Spatial Statistics Exercise 3, Spring 2019

Group members: Henrik Syversveen Lie, Øyvind Auestad

19.05.2019

Problem 1: Markov RF

This problem is based on observations of seismic data over a domain $\mathcal{D} \subset \mathbb{R}^2$. The objective is to identify the underlying {sand, shale} lithology distribution over $\mathcal{D}$, represented by $\{0,1\}$ respectively.

The observations are collected on a regular (75×75) grid L_D, and the seismic data are denoted $\{d(\mathbf{x}); \mathbf{x} \in L_D\}; d(\mathbf{x}) \in \mathbb{R}$, represented by the n-vector $\mathbf{d} \in \mathbb{R}^n$. We retrieve the observations from the R library MASS in the file *seismic.dat*.

Moreover, observations of the lithology distribution {sand, shale} in a geologically comparable domain $\mathcal{D}_c \subset \mathbb{R}^2$ is available. The lithology distribution is collected on a regular (66×66) grid L_{D_c}, which has the same spacing as L_D over $\mathcal{D}_c$. Here, we retrieve the observations from the same R library MASS in the file *complit.dat*.

We assume that the underlying lithology distribution can be represented by a Mosaic RF $\{l(\mathbf{x}); \mathbf{x} \in L_D\}; l(\mathbf{x}) \in \{0,1\} = \mathbb{L}$ represented by the n-vector $\mathbf{l}$. Then, we have $\mathbf{l} \in \mathbb{L}^n$, with $\mathbb{L}^n$ representing all possible binary n-vectors.

a)

The seismic data collection procedure defines the response likelihood model:

$$[d_i|\mathbf{l}] = \begin{cases} 0.02 + U_i, & \text{if } l_i = 0 - \text{ sand}, \\ 0.08 + U_i, & \text{if } l_i = 1 - \text{ shale} \end{cases} \quad ; i = 1, \ldots, n, \tag{1}$$

with $U_i; i = 1, \ldots, n$ i.i.d. $\text{Gauss}\{0, 0.06^2\}$.

Since the U_i's are all independent, d_i depends only on l_i. Hence we get the following likelihood model

$$[\mathbf{d}|\mathbf{l}] \sim p(\mathbf{d}|\mathbf{l}) = \prod_{i=1}^{n} p(d_i|\mathbf{l}) = \prod_{i=1}^{n} p(d_i|l_i) = \prod_{i=1}^{n} \left(\phi_1(d_i; 0.02, 0.06^2) I(l_i = 0) + \phi_1(d_i; 0.08, 0.06^2) I(l_i = 1) \right)$$

$$= \prod_{\substack{i=1 \\ l_i=0}}^{n} \phi_1(d_i; 0.02, 0.06^2) \prod_{\substack{i=1 \\ l_i=1}}^{n} \phi_1(d_i; 0.08, 0.06^2). \tag{2}$$

We then display the observations as a map in figure 1.

b)

We now give $\mathbf{l}$ a uniform constant prior, i.e. $p(\mathbf{l}) = \text{const}$. We obtain the following posterior model

$$[\mathbf{l}|\mathbf{d}] \sim p(\mathbf{l}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{l})}{\sum_{\mathbf{l}' \in \mathbb{L}^n} p(\mathbf{d}|\mathbf{l}')},$$
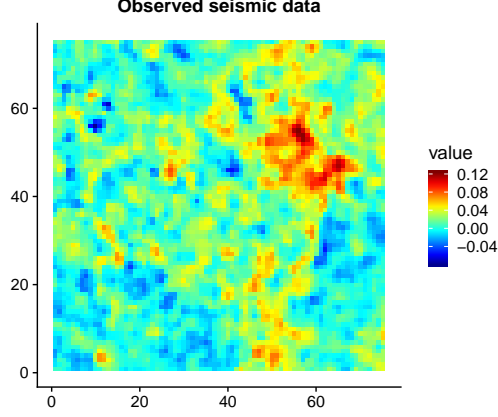
Figure 1: Observed values of **d** from the seismic data.

where the expression for the likelihood $p(\mathbf{d}|\mathbf{l})$ is given in equation 2. This gives the following posterior model

$$p(\mathbf{l}|\mathbf{d}) = \frac{\prod_{i=1}^{n}\left(\phi_1(d_i;0.02,0.06^2)I(l_i=0) + \phi_1(d_i;0.08,0.06^2)I(l_i=1)\right)}{\sum_{\mathbf{l}'\in\mathbb{L}^n}\left[\prod_{i=1}^{n}\left(\phi_1(d_i;0.02,0.06^2)I(l'_i=0) + \phi_1(d_i;0.08,0.06^2)I(l'_i=1)\right)\right]} \tag{3}$$

We want to simulate from the posterior Mosaic RF $\{l(\mathbf{x}); \mathbf{x} \in L_D|\mathbf{d}\}$. Because we have an independent prior, each point in the Mosaic RF will be independent, and we get the following posterior distribution

$$[l_i|d_i] \sim p(l_i|d_i) = \frac{p(d_i|l_i)}{\sum_{l'_i\in L} p(d_i|l'_i)} = \begin{cases} \frac{\phi_1(d_i;0.02,0.06^2)}{\phi_1(d_i;0.02,0.06^2)+\phi_1(d_i;0.08,0.06^2)} = 1-p_i, & \text{if } l_i=0, \\ \frac{\phi_1(d_i;0.08,0.06^2)}{\phi_1(d_i;0.02,0.06^2)+\phi_1(d_i;0.08,0.06^2)} = p_i, & \text{if } l_i=1 \end{cases}; i=1,\ldots,n.$$

This means that $[l_i|d_i]$ will be Bernoulli distributed with probability p_i as given above. We can simulate from the posterior distribution by simulating each point from a Bernoulli distribution with given p_i. In figures 2 and 3, we display in total 10 realizations of the posterior Mosaic RF.
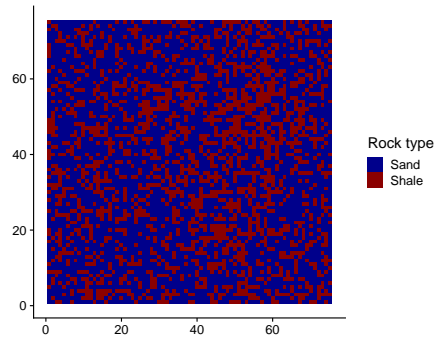

Figure 2: One realisation of the posterior RF with constant prior

We see that they are all very similar, and they look reasonable when compared to the observations in figure 1, taking into account that we have a constant prior. To expand on our analysis, we want to develop expressions for the expectation $E\{\mathbf{l}|\mathbf{d}\}$ and the variances in the diagonal terms of the matrix $\text{Var}\{\mathbf{l}|\mathbf{d}\}$.

We know that each point is independent, and each point conditioned on the observation in that point has a Bernoulli distribution with parameter p_i. From theory, we know that if $[l_i|d_i] \sim \text{Bernoulli}(p_i)$, then
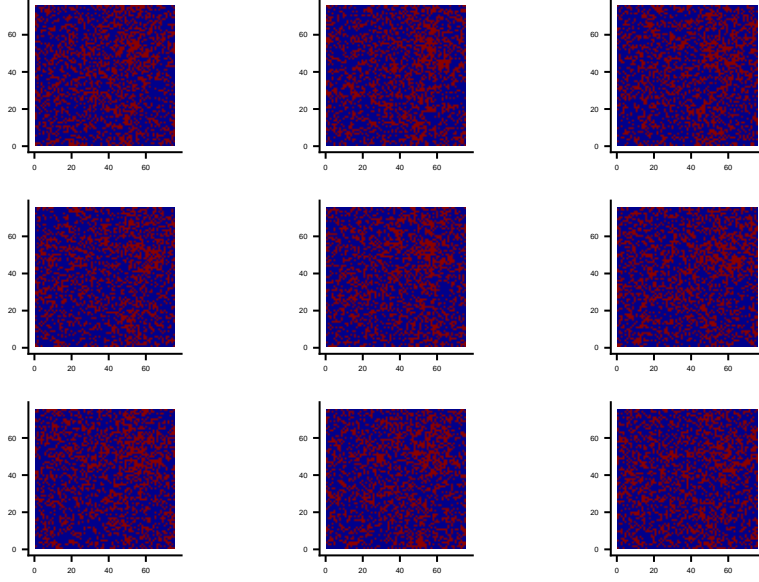
Figure 3: 9 realisations of the posterior RF with constant prior

$E(l_i|d_i) = p_i$ and $Var(l_i|d_i) = p_i(1 - p_i)$. This gives

$$E\{\mathbf{l}|\mathbf{d}\} = \mathbf{p} = (p_1, \ldots, p_n), \quad \mathrm{Var}\{l_i|d_i\} = p_i(1 - p_i), i = 1, \ldots, n.$$
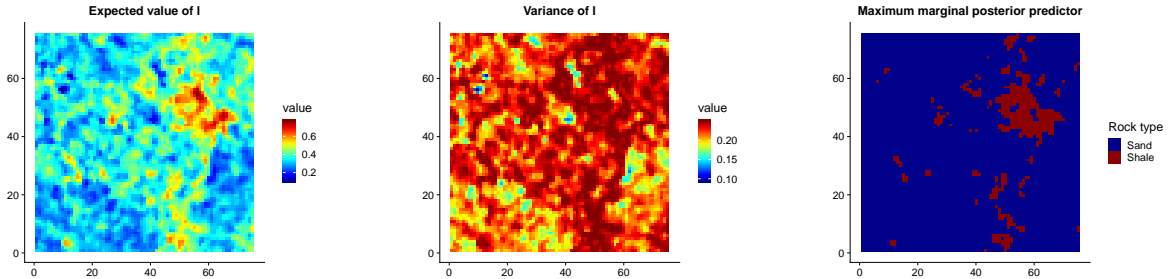
In addition, we want to develop an expression for the maximum marginal posterior predictor

$$\mathrm{MMAP}\{\mathbf{l}|\mathbf{d}\} = \hat{\mathbf{l}} = \operatorname*{argmax}_{\mathbf{l} \in \mathbb{L}^n}\{p(\mathbf{l}|\mathbf{d})\}.$$

Because each point is independent and $[l_i|d_i]$ is a Bernoulli distributed variable with parameter p_i, we get

$$\hat{l}_i = \begin{cases} 0, & \text{if } p_i < 0.5, \\ 1, & \text{if } p_i > 0.5. \end{cases}$$

Below, we have plotted the expecation, variance and MMAP.

We see that the expectation value $E(\mathbf{l}|\mathbf{d})$ is identical to the observed data in figure 1, only on a different scale. This is what we would expect, because we have a constant prior, so all the information about the lithology $\mathbf{l}$ comes from the observed data. The expectation value can also be interpreted as the probability that a point has a value of $l_i = 1$. Also, the variance is large over most of the domain, except for some minor areas where presumably the observed data either has a very large or small value.

c)

We then consider a Markov RF prior model for $\{l(\mathbf{x}); \mathbf{x} \in L_D\}$, represented by the n-vector $\mathbf{l}$, with a neighborhood system $\mathbf{n}_L$ consisting of the four closest neighbors of each grid node.

The Markov formulation of the RF is,

$$p(l_i|l_j; j \in \mathbf{n}_i) = const \times \exp\left(\beta \sum_{j \in \mathbf{n}_i} I(l_i = l_j)\right); i = 1, \ldots, n,$$

where $I(\cdot)$ is the indicator function.

As an alternative to the Markov formulation, we can formulate the Markov RF prior model with the Gibbs formulation, which is a n-dimensional joint distribution for all the grid nodes. To do this, we first specify the clique system. The clique system $\mathbf{c}_L : \{\mathbf{c}_1, \ldots, \mathbf{c}_{n_c}\}$ will be a two-node clique system where each clique consists of the two nearest nodes in horizontal or vertical direction. Having defined the cliques, we specify the Gibbs formulation of the Markov RF prior model as,

$$\mathbf{l} \sim p(\mathbf{l}) = const \times \prod_{i=1}^{n} v_{0l}(l_i) \times \prod_{\mathbf{c} \in \mathbf{c}_L} v_{1\mathbf{c}}(l_j; j \in \mathbf{c})$$

$$= const \times \prod_{<i,j>} \exp\left(\beta I(l_i = l_j)\right)$$

$$= const \times \exp\left(\beta \sum_{<i,j>} I(l_i = l_j)\right),$$

where $< i, j >$ represents the cliques of all two nearest neighbors in the grid L_D.

We have a response likelihood model, and a Markov RF prior model, which is a conjugate prior model to the response likelihood model. We want to develop expressions for the posterior model on Gibbs and Markov formulation, so we begin by developing the joint distribution of the spatial variable of interest and the observations,

$$[\mathbf{l}, \mathbf{d}] \sim p(\mathbf{l}, \mathbf{d}) = p(\mathbf{d}|\mathbf{l})p(\mathbf{l})$$

$$= const \times \prod_{i=1}^{n} p(d_i|l_i) \times \prod_{i=1}^{n} v_{0l}(l_i) \prod_{\mathbf{c} \in \mathbf{c}_L} v_{1\mathbf{c}}(l_j; j \in \mathbf{c})$$

$$= const \times \prod_{i=1}^{n} \left(\phi_1(d_i; 0.02, 0.06^2)I(l_i = 0) + \phi_1(d_i; 0.08, 0.06^2)I(l_i = 1)\right) \times \exp\left(\beta \sum_{<i,j>} I(l_i = l_j)\right).$$

To find the Gibbs formulation of the posterior model, we condition on $\mathbf{d}$ by dividing by $p(\mathbf{d})$. Because this expression is constant in $\mathbf{l}$ and we have not specified the normalizing constant, the resulting distribution will be the same as above, although with a different normalizing constant,

$$[\mathbf{l}|\mathbf{d}] \sim p(\mathbf{l}|\mathbf{d}) = [p(\mathbf{d})]^{-1} p(\mathbf{l}, \mathbf{d})$$

$$= const \times \prod_{i=1}^{n} \left(\phi_1(d_i; 0.02, 0.06^2)I(l_i = 0) + \phi_1(d_i; 0.08, 0.06^2)I(l_i = 1)\right) \times \exp\left(\beta \sum_{<i,j>} I(l_i = l_j)\right).$$

To obtain the Markov formulation, we take the joint distribution and for each l_i condition on the other nodes $\mathbf{l}_{-i}$, to get,

$$\begin{bmatrix} l_i \\ \mathbf{d} \end{bmatrix}\Big| \mathbf{l}_{-i} \sim p(l_i, \mathbf{d}|\mathbf{l}_{-i}) = p(\mathbf{d}|\mathbf{l})p(l_i|\mathbf{l}_{-i})$$

$$= const \times \prod_{i=1}^{n} \Big(\phi_1(d_i; 0.02, 0.06^2)I(l_i = 0) + \phi_1(d_i; 0.08, 0.06^2)I(l_i = 1) \Big) \times \exp\Big(\beta \sum_{j \in \mathbf{n}_i} I(l_i = l_j) \Big),$$

$$i = 1, 2, \ldots, n.$$

The posterior model for the Markov model formulation is then obtained by first dividing by $p(\mathbf{d}_{-i}|\mathbf{l}_{-i})$ and further computing the normalizing constant as $[p(d_i|\mathbf{l}_{-i})]^{-1}$,

$$\begin{bmatrix} l_i \Big| \begin{matrix} \mathbf{d} \\ \mathbf{l}_{-i} \end{matrix} \end{bmatrix} \sim p(l_i|\mathbf{d}, \mathbf{l}_{-i}) = p(l_i|d_i, l_j, j \in \mathbf{n}_i^l)$$

$$= \left[\sum_{l_i' \in \mathbb{L}} p(d_i|l_i')v_{0l}(l_i')w_l(l_i'|l_j; j \in \mathbf{n}_i^l) \right]^{-1} \times p(d_i|l_i)v_{0l}(l_i)w_l(l_i|l_j; j \in \mathbf{n}_i^l)$$

$$= \frac{p(d_i|l_i) \times \exp\Big(\beta \sum_{j \in \mathbf{n}_i} I(l_i = l_j) \Big)}{\phi_1(d_i; 0.02, 0.06^2)\exp\Big(\beta \sum_{j \in \mathbf{n}_i} I(l_j = 0) \Big) + \phi_1(d_i; 0.08, 0.06^2)\exp\Big(\beta \sum_{j \in \mathbf{n}_i} I(l_j = 1) \Big)}$$

$$i = 1, 2, \ldots, n,$$

with $p(d_i|l_i)$ as in equation 1.

To simulate realizations from $\{[l(\mathbf{x}|\mathbf{d}]; \mathbf{x} \in L\}$, we need to use an iterative algorithm. The most frequently used algorithm is a McMC/Gibbs algorithm with a single-site proposal pdf based on the Markov formulation of the Markov RF. We implement the following algorithm:

$$
\begin{aligned}
&\text{Define } g(\mathbf{l}'|\mathbf{l}): \\
&i \sim Uni[1, 2, \ldots, n] \\
&l_i' \sim p(l_i|d_i, l_j; j \in \mathbf{n}_i) \\
&\mathbf{l}' = (l_1, \ldots, l_{i-1}, l_i', l_{i+1}, \ldots, l_n) \\
&\text{Initiate:} \\
&\mathbf{l}^0 \text{ such that } p(\mathbf{l}^0|\mathbf{d}) > 0 \\
&\text{do for } i = 1, 2, \ldots \\
&\text{Generate: } \mathbf{l}^i \sim g(\mathbf{l}|\mathbf{l}^{i-1}) \\
&\text{end do}
\end{aligned}
\tag{4}
$$

Asymptotically as $i \to \infty$ then $\mathbf{l}^i \to \mathbf{l}^s \sim p(\mathbf{l}|\mathbf{d})$.

A prediction of the spatial variable, represented by the n-vector $\hat{\mathbf{l}}$ is usually defined based on a marginal maximum posteriori criterion:

$$\hat{\mathbf{l}} = \underset{\mathbf{l} \in \mathbb{L}^n}{\mathrm{argmax}}\{p(\mathbf{l}|\mathbf{d})\}.$$

The associated uncertainty quantification is provided by probability spatial variables, $\{p(l(\mathbf{x}) = l|\mathbf{d}); \mathbf{x} \in L_D\}$, represented by the n-vector $\mathbf{p}_l$ for each $l \in \mathbb{L}$,

$$\mathbf{p}_l = [p_{li} = p(l_i = l|\mathbf{d})]_{i=1,\ldots,n}.$$

We then note that the expected value $E(\mathbf{l}|\mathbf{d}) = \mathbf{p}_l$ and the diagonal terms of $\text{Var}(\mathbf{l}|\mathbf{d})$ will be $p_{li}(1 - p_{li}), i = 1, \ldots, n$.

Both the prediction and probability spatial variables must be assessed by sampling based inference, i.e. we generate a set of realizations $\mathbf{l}^s; s = 1, 2, \ldots, n_s$ and assess them by counting estimators.

Before we proceed to generate realizations, we need to estimate the β to be used. We use data from a geologically comparable domain to estimate β by a maximum pseudo-likelihood process. First, we display the observations from the domain D_c as a map in figure 4.
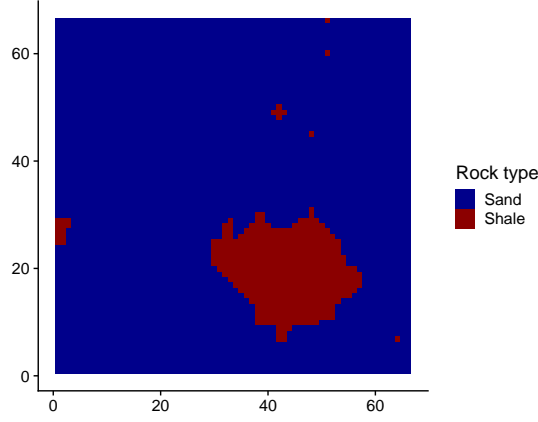

Figure 4:   Observations of lithology distribution in a geologically comparable domain to the seismic data.

Then, we define the distribution of the vector $\mathbf{l}$, given the model parameter β, as a Gibbs formulation of a Markov RF,

$$\mathbf{l} \sim p(\mathbf{l}; \beta) = const \times \exp\left(\beta \sum_{<i,j>} I(l_i = l_j)\right).$$

Ideally, we would like to use this distribution to estimate β, but the normalizing constant is difficult to assess because we have to sum over 2^n terms, which is not computanionally feasible. As an alternative, we define the distribution as a Markov formulation,

$$[l_i|l_j, \beta; j \in \mathbf{n}_i] \sim p(l_i|l_j, \beta; j \in \mathbf{n}_i) = \left[\sum_{l_i' \in \mathbb{L}} \exp\left(\beta \sum_{j \in \mathbf{n}_i} I(l_i' = l_j)\right)\right]^{-1} \times \exp\left(\beta \sum_{j \in \mathbf{n}_i} I(l_i = l_j)\right).$$

Then, we can define the pseudo-likelihood of observing the data, given the parameter β, as the product of these univariate distributions,

$$\hat{p}(\mathbf{d}; \beta) \approx const \times \prod_{i=1}^{n} \sum_{[l_i'; l_j' \in \mathbf{n}_i] \in \mathbb{L}} \prod_{j=i; j \in \mathbf{n}_i} p(d_j|l_j') \times p(l_i'|l_j'; j \in \mathbf{n}_i).$$

This expression does not contain any cumbersome normalizing constant. Because we assume exact observations of the underlying lithology in figure 4, we have the following likelihood,

$$[d_i|l_i] \sim p(d_i|l_i) = \delta_{d_i}(l_i),$$

which is inserted into the pseudo-likelihood, together with the distribution of $\mathbf{l}$ to achieve

$$\hat{p}(\mathbf{d}; \beta) \propto \prod_{i=1}^{n} \left[\sum_{l_i' \in \mathbb{L}} \exp\left(\beta \sum_{j \in \mathbf{n}_i} I(l_i' = l_j)\right)\right]^{-1} \times \exp\left(\beta \sum_{j \in \mathbf{n}_i} I(l_i = l_j)\right).$$

To find the optimal β, we maximize the logarithm of this pseudo-likelihood, giving

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\{\log(\hat{p}(\mathbf{d};\beta)\} = \underset{\beta}{\operatorname{argmax}}\left\{ \sum_{i=1}^{n}\left(\beta\Big[\sum_{j\in\mathbf{n}_i}I(l_i = l_j)\Big] - \log\Big[\sum_{l'_i\in\mathbb{L}}\exp\Big(\beta\sum_{j\in\mathbf{n}_i}I(l_i = l_i)\Big)\Big]\right)\right\}.$$

We use wrapping boundary conditions on the data displayed in figure 4 and estimate β by the procedure above using the R function `optim()` to find the maximum pseudo-loglikelihood. This gives $\hat{\beta} = 1.67$.

To assess the posterior model $p(\mathbf{l}|\mathbf{d})$ of the seismic data, we use algorithm 4 to simulate from the posterior distribution. Because this is an MCMC algorithm, we need some way to justify that our algorithm has converged. To do this, we run four simulations, one starting with $\mathbf{l}^0 = \mathbf{0}$, one starting with $\mathbf{l}^0 = \mathbf{1}$, one starting from the MMAP in section b) and one looking like a chess board. We then display the proportion of nodes with $l_i = 1$ for each iteration in the algorithm.
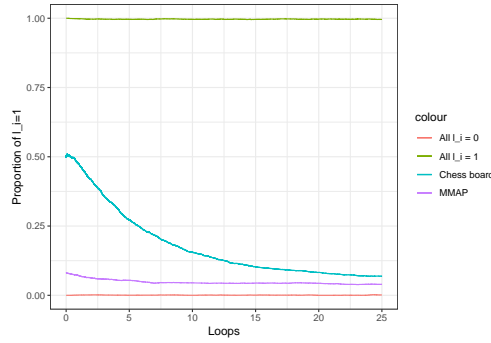

Figure 5: Proportion of nodes with $l_i = 1$ for four different MCMC chains with $\beta = 1.67$. Each loop consists of $75 \cdot 75$ iterations.

From figure 5, we see that the MCMCs starting with all $l_i = 0$ and all $l_i = 1$ get stuck and do not converge. We suspect that this is a result of β being too large. If we compare the training data set in figure 4 with the seismic data in figure 1 and the predicitons from section b), we see that the training data maybe has too little spread and a too "compact" structure. To account for this, we reduce the the estimated parameter somewhat, and let $\beta = 1.4$, which should in theory produce results with more spread. We then run the same algorithm and display four MCMCs with the same initial values as above.
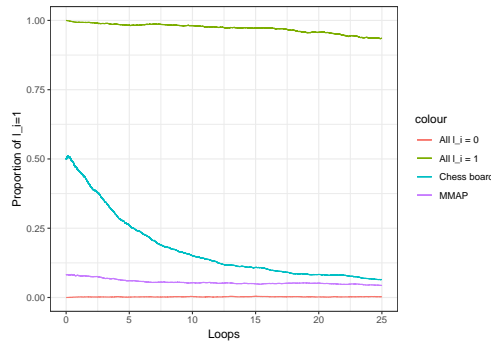

Figure 6: Proportion of nodes with $l_i = 1$ for four different MCMC chains with $\beta = 1.4$. Each loop consists of $75 \cdot 75$ iterations.

From figure 6 we see that the chain still gets stuck in $\mathbf{l}^0 = \mathbf{0}$ and converges slowly, although somewhat faster than for $\beta = 1.67$. We therefore procede with $\beta = 1.4$. We use 15 loops and use as initial $\mathbf{l}^0$ the MMAP from section b). This MMAP is our best best estimate so far, and will therefore be our best hope of the MCMC converging in feasible time. We then display 9 independent realizations as maps in figure 7.
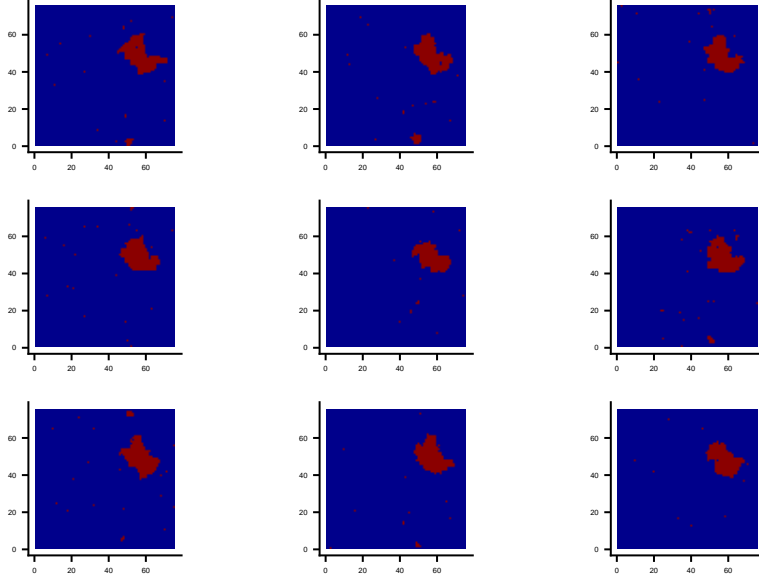
Figure 7: 9 realisations of the posterior RF with Markov RF prior model

From figure 7 we see that all the realisations have a compact body of shale in the topright corner, with some minor spreads of shale elsewhere. To further investigate our new model, we look at $E(\mathbf{l}|\mathbf{d})$, the diagonal terms of $\mathrm{Var}(\mathbf{l}|\mathbf{d})$ and the prediction $\hat{\mathbf{l}} = \mathrm{MMAP}(\mathbf{l}|\mathbf{d})$ through simulating several independent Markov RFs and then assessing these by counting estimators.
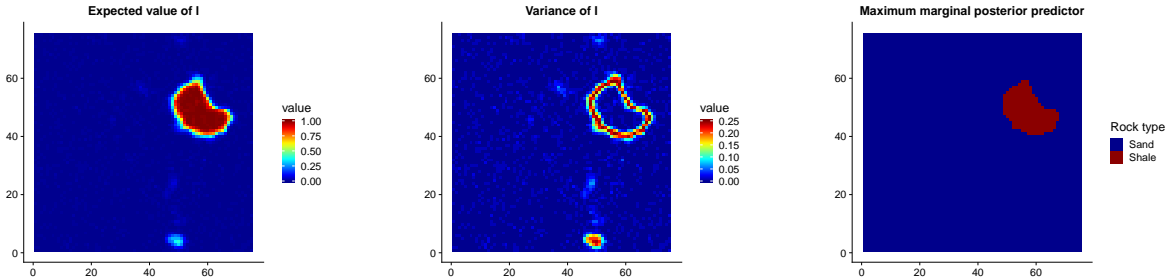
For the expectation and variance we get,

$$E(l_i|d_i) = p_i = \frac{1}{n_s}\sum_{s=1}^{n_s} l_i^s, \quad \mathrm{Var}(l_i|d_i) = p_i(1-p_i).$$

For the MMAP$(\mathbf{l}|\mathbf{d})$, this will be for each node l_i the mode of all the n_s samples, or

$$\hat{l}_i = \mathrm{Mode}\big(\{l_i^s\}_{s=1,\ldots,n_s}\big) = \underset{l_i' \in \mathbb{L}}{\mathrm{argmax}}\left[\sum_{s=1}^{n_s} I(l_i' = l_i^s)\right].$$

We simulate 100 independent realizations with 15 loops in each realization and then display the above characteristics as maps.

As opposed to the model with constant prior, the expectation of this model is not equal to the observed data. The expectation has a solid body with value close to 1 in the top right part of the domain. The variance is approximately zero everywhere, except on the edge of the body and in a small area in the bottom of the domain. The MMAP is zero everywhere, except for the solid body in the top right.

d)

Comparing the two models, we see that the constant prior model has much higher variance in its prediction and a much larger spread in the MMAP and expectation. From the model specifications, it is clear that the Markov RF prior gives a larger probability of two neighbors having equal value. Hence, we get less spread in the prediction and only one large body of shale, as opposed to several smaller areas of shale. The parameter β in the Markov RF prior was estimated by a dataset containing one large body of shale, and only a few other small areas of shale. It is then logical for the posterior to only have one large body of shale, even though we reduced β somewhat. By reducing the value of β even more, the Markov RF prior model would resemble the constant prior more and get a larger spread.