AALBORG UNIVERSITY

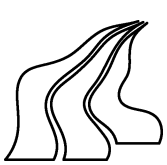
Analysis of Spatial Count Data using Kalman Smoothing

by

Claus Dethlefsen

April 2004 R-2004-12

Department of Mathematical Sciences
Aalborg University
Fredrik Bajers Vej 7 G ▪ DK - 9220 Aalborg Øst ▪ Denmark
Phone: +45 96 35 80 80 ▪ Telefax: +45 98 15 81 29
URL: www.math.aau.dk/research/reports/reports.htm



ISSN 1399–2503 ▪ On-line version ISSN 1601–7811

Analysis of Spatial Count Data using Kalman Smoothing

Claus Dethlefsen
Dept. of Mathematical Sciences
Aalborg University
Fr. Bajers Vej 7G
9220 Aalborg, Denmark

Abstract

This paper considers spatial count data from an agricultural field experiment. Counts of weed plants in a field have been recorded in a project on precision farming. Interest is in mapping the weed intensity so that the dose of herbicide applied at any location can be adjusted to the amount of weed present at the location. We elaborate on a link between state space models and Markov random fields. The observations are modelled as independent Poisson counts conditional on a Gaussian Markov random field. We employ the fact that the model may be written as a state space model which may be analysed by combining approximate Kalman filter techniques with importance sampling.

1 Introduction

We analyse a data set kindly put to our disposal by Danish Institute of Agricultural Sciences. The data were collected in connection with a project in precision farming at the Danish Institute of Agricultural Sciences. Counts of weed plants on a field were recorded in 1993, 1994 and 1995. Interest is in mapping the weed intensity so that the dose of herbicide applied at any location can be adjusted to the amount of weed present at the location.

Along with the weed counts, 11 explanatory variables were also measured. Among these variables, Christensen et al. [2000] found that the intensity of weed was related to the percentage of organic matter in the soil and that there is a north-south decreasing trend in the data. Here, we model the relation between counts of the species *Viola arvensis* in year 1994 and the two explanatory variables.

The weed counts are displayed in Figure 1, using the actual values. Missing values are shown with a star, and zero counts are not shown. The horizontal axis in Figure 1 corresponds to the ploughing direction, and the counts are observed within $0.25m^2$ circular frames with spacing $20m$. The five contiguous missing values in the rows 4–6 from the top correspond to a peat bog.

As in Christensen and Waagepetersen [2002], we have transformed the two explanatory variables. The explanatory variable measuring the percentage of or-

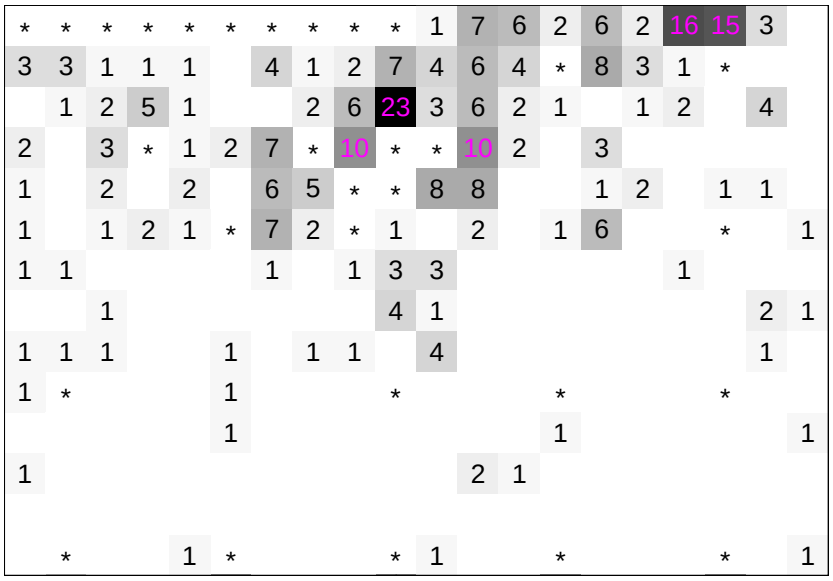


Figure 1: Counts of weed plants at locations with spacing $20m$. Unobserved points are marked by "*". Zero counts are not printed. Darker regions mean higher concentration of weed. The ploughing direction is left-right.

ganic matter in the soil was first transformed using a logit transformation, then the average was subtracted and, finally, the values were divided by the maximum value. The second explanatory variable is the second coordinate of the sites, corresponding to the north-south trend. The variable was transformed by subtracting the average and dividing by the largest value.

The data were first presented and analysed by Walter et al. [1997]. In Christensen and Waagepetersen [2002], the data set was analysed using Markov chain Monte Carlo (MCMC) methods. They used Langevin-Hastings updates to simulate from the posterior distribution of a generalised linear mixed model. The random effects were modelled by a spatial stationary Gaussian field and, conditional on this field, the weed counts were assumed to be independent Poisson observations.

Our model is based on a Poisson observation model conditional on a Gaussian Markov random field and assumes non-stationary spatial random effects. The methodology described here is thus an alternative to the approach in Christensen and Waagepetersen [2002].

Lavine [1999] showed that a Gaussian Markov random field model may be written in the form of a state space model. We elaborate on this and express the Poisson-Gaussian model as a non-Gaussian state space model. Then, we may use Kalman filter techniques for making inference. For basic references on state space methodology, see Harvey [1989], West and Harrison [1997] and Durbin and Koopman [2001]. Our approach is not based on MCMC methods, but on iterated extended Kalman smoothing, which may be combined with importance sampling for exact simulation, see Durbin and Koopman [2000]. Using this method, we avoid the MCMC problems of ensuring that the Markov chain is

mixing well and assessing whether the chain has converged or not.

The programmes used in the analysis are available from `www.math.aau.dk/~dethlef/PhD` and have been written using **R** (see R Development Core Team [2003]).

2 Model for Spatial Count Data

The data are arranged on an $I \times J$ grid, where $I = 14$ and $J = 20$. Let y_{ij} be the weed count in site ij corresponding to the location $(20 \cdot (15 - i), 20 \cdot j)$, where $i = 1, \ldots, I$ is the row index and $j = 1, \ldots, J$ is the column index. The indexing begins in the upper left corner in Figure 1. Let θ_{ij} be the unobserved random effect at site ij. The weed counts for row i are collected in the vector $\mathbf{y}_i$ with corresponding unobserved vector $\boldsymbol{\theta}_i$.

As prior model for the vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i=1,\ldots,I}$, we use a Markov random field given by Besag [1974],

$$p(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{P} \boldsymbol{\theta}\right), \tag{1}$$

where $\mathbf{P}$ is the $IJ \times IJ$ precision matrix

$$\mathbf{P} = \tau^{-2}\mathbf{T}_I \otimes \mathbf{I}_J + \tau^{-2}\mathbf{I}_I \otimes \mathbf{T}_J,$$

with

$$\mathbf{T}_l = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 2 & -1 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}_{l \times l}.$$

The parameter τ^2 measures the smoothness of the Markov random field, and is estimated by maximum likelihood estimation. Note that the distribution in (1) is improper, since the precision matrix is singular. Our interest is, however, in the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$, which is proper in our case.

Let $\mathbf{z}_{ij}^\top$ be the 2-dimensional row vector containing the covariates measured at location ij: The second coordinate and the percentage of organic matter. Let $\boldsymbol{\beta} = (\beta_1, \beta_2)$ be the corresponding vector of coefficients. We aim at assessing the posterior distribution of $\boldsymbol{\beta}$ and assign the prior distribution

$$p(\boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{0}, 100 \cdot \mathbf{I}). \tag{2}$$

We assume that the observations are independent Poisson observations conditional on $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$,

$$y_{ij}|(\boldsymbol{\theta}, \boldsymbol{\beta}) \sim \mathrm{Po}(\overbrace{\exp(\underbrace{\mathbf{z}_{ij}^\top \boldsymbol{\beta} + \theta_{ij}}_{\lambda_{ij}})}^{\mu_{ij}}). \tag{3}$$

Here, μ_{ij} denotes the weed intensity.

3 State Space Formulation in Gaussian Case

Our aim is to write the model given by (1)-(3) as a non-Gaussian state space model. However, first we treat the case with the Gaussian observation model,

$$\mathbf{y}_i|(\boldsymbol{\theta}_i, \boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{Z}_i^\top \boldsymbol{\beta} + \boldsymbol{\theta}_i, \boldsymbol{\Sigma}_i),$$

where the design matrix $\mathbf{Z}_i^\top$ has $\mathbf{z}_{ij}^\top$ in the jth row. Lavine [1999] showed that in this Gaussian case, the model can be expressed as a Gaussian state space model, evolving following the rows.

$$\begin{pmatrix} \mathbf{y}_i \\ \mathbf{x}_i \end{pmatrix} \Bigg| \ (\boldsymbol{\theta}_i, \boldsymbol{\beta}) \sim \mathcal{N}\left[\begin{pmatrix} \mathbf{Z}_i^\top \boldsymbol{\beta} + \boldsymbol{\theta}_i \\ \mathbf{H}\boldsymbol{\theta}_i \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_i & \mathbf{0} \\ \mathbf{0} & \tau^2 \mathbf{I}_{J-1} \end{bmatrix} \right]$$

$$\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1} \sim \mathcal{N}(\boldsymbol{\theta}_{i-1}, \tau^2 \mathbf{I}_J)$$

$$p(\boldsymbol{\theta}_1) \propto 1,$$

where

$$\mathbf{H} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix}.$$

Thus $\mathbf{y}_i$ are the observed rows, $\boldsymbol{\theta}_i$ are the corresponding latent variables and $\mathbf{x}_i$ are so-called pseudo observations. The analysis of the model is carried out conditional on the pseudo observations being observed to zero as this ensures the equivalence of the state space model with the Markov random field model. In other words, $p(\boldsymbol{\theta}|\mathbf{x} = \mathbf{0})$ is the Markov random field prior (1) and $p(\boldsymbol{\theta}|\mathbf{x} = \mathbf{0}, \mathbf{y})$ is the posterior.

If we introduce a more compact notation, letting $\mathbf{Y}_i = (\mathbf{y}_i, \mathbf{x}_i)$, $\boldsymbol{\Theta}_i = (\boldsymbol{\theta}_i, \boldsymbol{\beta})$, we may write the model as

$$\begin{aligned} \mathbf{Y}_i|\boldsymbol{\Theta}_i &\sim \mathcal{N}(\mathbf{F}_i^\top \boldsymbol{\Theta}_i, \mathbf{V}_i) \\ \boldsymbol{\Theta}_i|\boldsymbol{\Theta}_{i-1} &\sim \mathcal{N}(\boldsymbol{\Theta}_{i-1}, \mathbf{W}_i) \\ \boldsymbol{\Theta}_0 &\sim \mathcal{N}(\mathbf{m}_0, \mathbf{C}_0), \end{aligned}$$

where

$$\mathbf{F}_i^\top = \begin{bmatrix} \mathbf{I} & \mathbf{Z}_i^\top \\ \mathbf{H} & \mathbf{0} \end{bmatrix} \qquad \mathbf{V}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i & \mathbf{0} \\ \mathbf{0} & \tau^2 \mathbf{I} \end{bmatrix}$$
$$\mathbf{W}_i = \begin{bmatrix} \tau^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

The Kalman filter recursively yields $p(\boldsymbol{\Theta}_i|D_i)$, the conditional distribution of $\boldsymbol{\Theta}_i$ given all information available, D_i, at current row i,

$$\boldsymbol{\Theta}_i|D_{i-1} \sim \mathcal{N}(\overbrace{\mathbf{m}_{i-1}}^{\mathbf{a}_i}, \overbrace{\mathbf{C}_{i-1} + \mathbf{W}_i}^{\mathbf{R}_i})$$

$$\mathbf{Y}_i|D_{i-1} \sim \mathcal{N}(\overbrace{\mathbf{F}_i^\top \mathbf{a}_i}^{\mathbf{f}_i}, \overbrace{\mathbf{F}_i^\top \mathbf{R}_i \mathbf{F}_i + \mathbf{V}_i}^{\mathbf{Q}_i})$$

$$\boldsymbol{\Theta}_i|D_i \sim \mathcal{N}(\underbrace{\mathbf{a}_i + \overbrace{\mathbf{R}_i \mathbf{F}_i \mathbf{Q}_i^{-1}}^{\mathbf{A}_i}(\mathbf{Y}_i - \mathbf{f}_i)}_{\mathbf{m}_i}, \underbrace{\mathbf{R}_i - \mathbf{A}_i \mathbf{Q}_i \mathbf{A}_i^\top}_{\mathbf{C}_i}).$$

Assessment of the state vector, $\boldsymbol{\Theta}_i$, using all available information, D_I, is called Kalman smoothing and we write $(\boldsymbol{\Theta}_i|D_I) \sim \mathcal{N}(\widetilde{\mathbf{m}}_i, \widetilde{\mathbf{C}}_i)$. Starting with $\widetilde{\mathbf{m}}_I = \mathbf{m}_I$ and $\widetilde{\mathbf{C}}_I = \mathbf{C}_I$, the Kalman smoother is a backwards recursion, $i = I-1, \ldots, 1$, with

$$\widetilde{\mathbf{m}}_i = \mathbf{m}_i + \mathbf{B}_i(\widetilde{\mathbf{m}}_{i+1} - \mathbf{a}_{i+1})$$

and

$$\widetilde{\mathbf{C}}_i = \mathbf{C}_i + \mathbf{B}_i(\widetilde{\mathbf{C}}_{i+1} - \mathbf{R}_{i+1})\mathbf{B}_i^\top,$$

where $\mathbf{B}_i = \mathbf{C}_i \mathbf{R}_{i+1}^{-1}$. It is often computationally faster to use the mathematically equivalent disturbance smoother, see Koopman [1993].

The log likelihood function for a vector of hyperparameters $\boldsymbol{\psi}$, *e.g.* τ^2 and components of $\boldsymbol{\Sigma}_i$, is given by

$$\begin{aligned} l(\boldsymbol{\psi}) &= \sum_{i=1}^{n} \log p(\mathbf{y}_i|\mathbf{y}_1, \ldots, \mathbf{y}_{i-1}, \boldsymbol{\psi}) \\ &= c - \frac{1}{2}\sum_{i=1}^{n} \left\{ \log|\mathbf{Q}_i^*| + \|\mathbf{y}_i - \mathbf{f}_i^*\|^2_{\mathbf{Q}_i^{*-1}} \right\}, \end{aligned} \tag{4}$$

where $\mathbf{f}_i^*$ and $\mathbf{Q}_i^*$ are the first J components of $\mathbf{f}_i$ and $\mathbf{Q}_i$ that corresponds to $\mathbf{y}_i$, and $\|\mathbf{x}\|^2_{\boldsymbol{\Sigma}} = \mathbf{x}^\top \boldsymbol{\Sigma} \mathbf{x}$ and c is a constant. The log likelihood for a given value of $\boldsymbol{\psi}$ can thus be obtained directly from the Kalman filter. The expression (4) can then be maximised numerically yielding the maximum likelihood estimate.

4 State Space Formulation of the Poisson-Gaussian Model

In the weed count application, the observation model is not Gaussian and the result of Lavine [1999] cannot be applied directly. However, we may combine the result with the framework of Durbin and Koopman [2001]. Amongst other models, they treated state space models with observations from the exponential family, including the Poisson case. Following their approach, we linearise the observation model and use the Kalman smoother iteratively to obtain an approximating Gaussian state space model. The procedure is called iterated extended Kalman smoothing. The approximating Gaussian state space model has the same posterior mode and curvature at the mode as the non-Gaussian model.

At coordinate level, we write the non-Gaussian state space model as

$$y_{ij}|(\boldsymbol{\theta}_i, \boldsymbol{\beta}) \sim \mathrm{Po}(\overbrace{\exp(\underbrace{\mathbf{z}_{ij}^\top \boldsymbol{\beta} + \theta_{ij}}_{\lambda_{ij}})}^{\mu_{ij}}) \quad (5)$$

$$x_{ij}|(\theta_{ij}, \theta_{i,j+1}) \sim \mathcal{N}(\theta_{ij} - \theta_{i,j+1}, \tau^2) \quad (6)$$

$$\theta_{ij}|\theta_{i-1,j} \sim \mathcal{N}(\theta_{i-1,j}, \tau^2). \quad (7)$$

As initialization, we have used

$$p(\boldsymbol{\theta}_0) \sim \mathcal{N}(\mathbf{0}, 100 \cdot \mathbf{I})$$

Theoretically, the prior for $\boldsymbol{\theta}_0$ should be improper, but we have not implemented this feature and expect, that the above prior will behave similarly.

To linearise the observation model, we use initial values $\widetilde{\boldsymbol{\beta}}^{(0)}$ and $\widetilde{\boldsymbol{\theta}}_i^{(0)}$. In the kth iteration of the iterated extended Kalman smoother, we assume that $\widetilde{\boldsymbol{\beta}}^{(k-1)}$ and $\widetilde{\boldsymbol{\theta}}_i^{(k-1)}$ are given. Let

$$\widetilde{\boldsymbol{\lambda}}_i^{(k-1)} = \mathbf{Z}_i^\top \widetilde{\boldsymbol{\beta}}^{(k-1)} + \widetilde{\boldsymbol{\theta}}_i^{(k-1)}$$

and let $\widetilde{\mathbf{V}}_i^{(k)}$ be a diagonal matrix with

$$\widetilde{V}_{ij}^{(k)} = \exp(-\widetilde{\lambda}_{ij}^{(k-1)})$$

as jth diagonal element. Finally, let $\widetilde{\mathbf{y}}_i^{(k)}$ be a vector with jth element $\widetilde{y}_{ij}^{(k)} = \widetilde{\lambda}_{ij}^{(k-1)} + \widetilde{V}_{ij}^{(k)} y_{ij} - 1$.

We now have the matrix form of the approximating state space model

$$\begin{pmatrix} \widetilde{\mathbf{y}}_i^{(k)} \\ \mathbf{x}_i \end{pmatrix} \Bigg| \begin{pmatrix} \boldsymbol{\theta}_i \\ \boldsymbol{\beta} \end{pmatrix} \sim \mathcal{N}\left[\begin{bmatrix} \mathbf{I} & \mathbf{Z}_i^\top \\ \mathbf{H} & \mathbf{0} \end{bmatrix} \begin{pmatrix} \boldsymbol{\theta}_i \\ \boldsymbol{\beta} \end{pmatrix}, \begin{bmatrix} \widetilde{\mathbf{V}}_i^{(k)} & \mathbf{0} \\ \mathbf{0} & \tau^2 \mathbf{I} \end{bmatrix} \right]$$

$$\begin{pmatrix} \boldsymbol{\theta}_i \\ \boldsymbol{\beta} \end{pmatrix} \Bigg| \begin{pmatrix} \boldsymbol{\theta}_{i-1} \\ \boldsymbol{\beta} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \boldsymbol{\theta}_{i-1} \\ \boldsymbol{\beta} \end{pmatrix}, \begin{bmatrix} \tau^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right],$$

which is analysed using the Kalman smoother conditional on $\mathbf{x}_i = \mathbf{0}$. When the Kalman smoother is iterated until convergence, we obtain an approximating state space model with the same posterior mode and curvature at the mode as the non-Gaussian model (5)– (7), see Durbin and Koopman [2001].

We have not made any action towards eliminating edge effects arising from choosing a Markov random field as a prior. One way of eliminating the effect would be to put extra frames of missing values around the measured area as suggested by Besag and Higdon [1999].

5 Results

The parameter τ^2 was estimated using maximum likelihood estimation. As an approximation to the likelihood, we have used the log likelihood (4) from the approximating state space model.

We used the maximizer `optimize` in **R** and chose $[-10, 10]$ as the interval for $\log \tau$. Using broader intervals, we ran into numerical difficulties. The resulting estimate was $\hat{\tau}^2 = 2.32 \cdot 10^{-9}$ and an approximate log likelihood of -460. The maximum was obtained on the edge of the interval given, but we have chosen to retain this estimate.

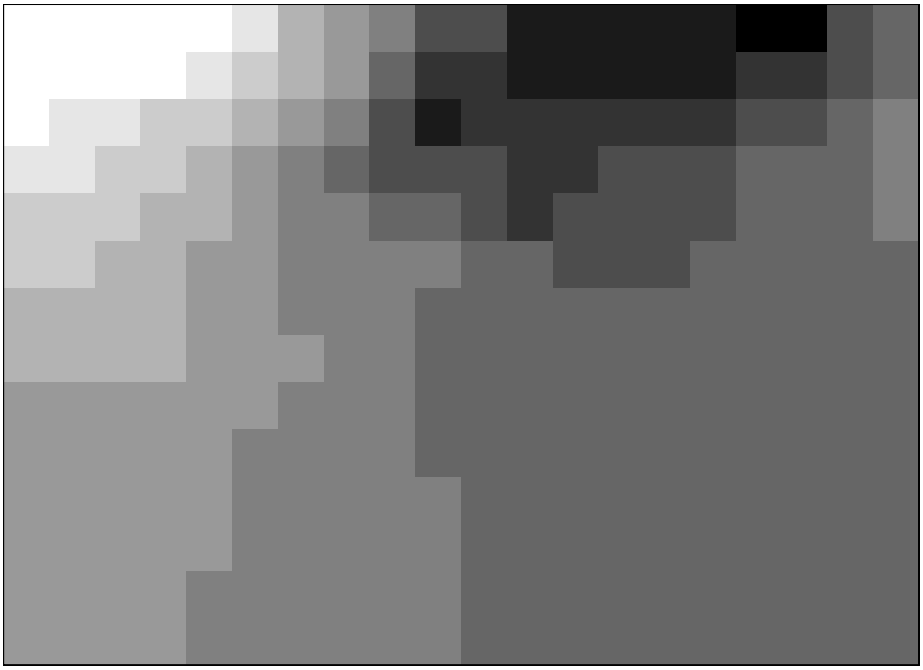


Figure 2: Posterior mean of random effects.

Using the iterated extended Kalman smoother with the maximum likelihood estimate for τ^2, we get

$$\begin{aligned} \hat{\beta}_1 &= 2.13 \\ \hat{\beta}_2 &= 2.42. \end{aligned}$$

These estimates are within the reported 95% credible intervals from Christensen and Waagepetersen [2002], respectively $]1.84, 3.08[$ and $]0.82, 2.94[$.

The average of the posterior mean of the random effects, θ_{ij}, we call the intercept, and we estimate this to be -0.543, which is just beyond the estimated 95% credible interval $]-0.41, 0.67[$ from Christensen and Waagepetersen [2002].

The estimated random effects are shown in Figure 2. The values are very small and the effect is negligible compared with the effect of the covariates. The random effects were not reported in Christensen and Waagepetersen [2002].

The posterior mean of the weed intensity is shown in Figure 3. This map can be used by the farmer to adjust the dose of herbicide applied at the different locations on the field.

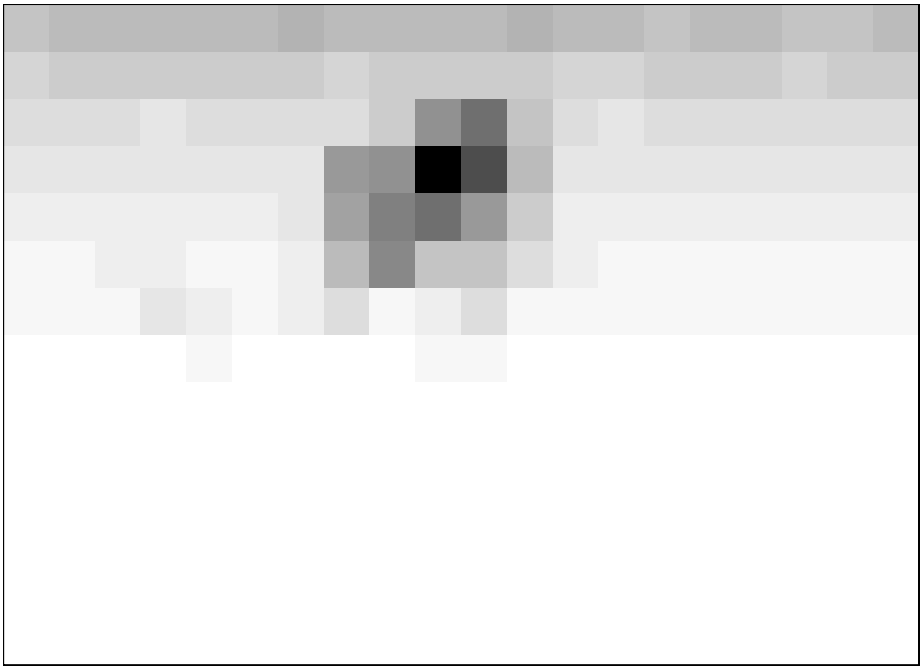


Figure 3: Posterior mean of weed intensity.

6 Discussion

In this paper we have formulated a model for the analysis of spatial count data. Our analysis has been based on the approximative analysis by the iterated extended Kalman smoother. The approximating state space model can be used as importance density to provide exact sampling of quantities of interest, but we have not implemented this.

Writing Markov random field models as state space models following Lavine [1999], makes it possible to use Kalman filter techniques to extend and analyse more complex Markov random field models. In Dethlefsen [2002] it is shown how the methodology may be adopted for restoring digital images with focus on finding edges in the image. However, the new class of models also have applications within agricultural experiments, see *e.g.* Besag and Higdon [1999] and within disease mapping, see *e.g.* Knorr-Held and Rue [2002].

Our results in the weed count analysis are very close to the results reported by Christensen and Waagepetersen [2002], but we conclude that the effect of the two explanatory variables overshadows the random effects. However, we had problems finding the maximum likelihood estimate and must be cautious in our conclusions.

Acknowledgements

I am indebted to my Ph.D. supervisor Søren Lundbye-Christensen for inspiring discussions. Also, we thank the Danish Institute of Agricultural Sciences for providing the weed count data.

References

J. Bernardo, J. Berger, A.P. Dawid, and A.F.M. Smith, editors. *Bayesian Statistics 6*, 1999. Oxford University Press.

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36(2):192–236, 1974.

J. Besag and D. Higdon. Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society, Series B*, 61(4): 691–746, 1999.

O.F. Christensen, J. Møller, and R. Waagepetersen. Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte Carlo. Technical report, R-00-2009, Aalborg University, 2000.

O.F. Christensen and R. Waagepetersen. Bayesian prediction of spatial count data using generalised linear mixed models. *Biometrics*, 58:280–286, 2002.

C. Dethlefsen. *Space time problems and applications*. PhD thesis, Aalborg University, 2002.

J. Durbin and S.J. Koopman. Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with discussion). *Journal of the Royal Statistical Society, Series B*, 62(1): 3–56, 2000.

J. Durbin and S.J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.

A.C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, 1989.

L. Knorr-Held and H. Rue. On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 2002.

S.J. Koopman. Disturbance smoother for state space models. *Biometrika*, 80 (1):117–126, 1993.

M. Lavine. Another look at conditionally Gaussian Markov random fields. In Bernardo et al. [1999].

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2003. URL `http://www.R-project.org`. ISBN 3-900051-00-3.

A.M. Walter, T. Heisel, and S. Christensen. Shortcuts in weed mapping. In J.V. Stafford, editor, *Precision Agriculture 1997*, pages 777–784. BIOS Scientific Publishers Ltd., 1997.

M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, 2nd edition, 1997.