STATISTICAL RESEARCH REROT No. 4
ISBN Juni 2006
ISSN 0806 – 3842

Modelling Pollock egg counts from the western Gulf of Alaska by a zero-inflated Bayesian hierarchical space-time model

Ingunn Fride Tvete,
University of Oslo, Oslo, Norway,

Geir Storvik,
University of Oslo, Oslo, Norway

and Bent Natvig,
University of Oslo, Oslo, Norway

Summary. Data from egg sampling surveys often contain a large proportion of zeros. We examine the time and location of collected walleye pollock egg counts from the western Gulf of Alaska, from Kodiak Island to Unimah Pass, in the years 1978-2000. We take the excessive number of zeros in the data into account by taking a two stage modelling approach, resulting in a zero-inflated hierarchical space-time model. The binary (presence/absence of zeros) process is directly linked to the underlying intensity process. Hence, the latter process influences both the presence of zeros *and* the amount of eggs in non-zero observations. We fit our model for each year, and compare the underlying intensities over the years within regions of the sampled area.

Keywords: Dynamical systems; walleye pollock; egg counts; MCMC; zero-inflated; space-time modelling; Bayesian hierarchical modelling

Address for correspondence: Ingunn Fride Tvete, Department of Mathematics, University of Oslo, P.O. Box 1053 Blindern, N-0316 Oslo, Norway. E-mail: ift@math.uio.no

1. Introduction

Spatial structure and temporal changes in spawning habits are important aspects of understanding the population dynamics of many fish stocks. Large surveys performing egg counts during the spawning period have been performed for many species. In addition to variations in space and time, a typical pattern of such data is a large amount of zero observations. These zero observations can be informative about the spawning structure and neglecting them might lead to biased estimates and erroneous results.
In this paper we examine the time and location of collected walleye pollock egg counts from the western Gulf of Alaska, from Kodiak Island to Unimah Pass, in the years $1978-2000$ (with some years missing). According to Ciannelli et al. (2006) there are favorable geographical formations in the Shelikof strait, making it the main spawning region throughout the years. During the last decade there has been indications of less spawning within this area. Statistical analysis of changes in the spatial structure over the years might shed light on this. The counts are per m^2. In addition to a large amount of zeros, the data set contains many high non-zero counts. Using a traditional modelling approach without taking the zeros carefully into account might lead to biased estimates and erroneous results. The expected counts, given that something is observed, are large, and a natural approximation is the $\log$ transformed Gaussian distribution. Working within a Gaussian framework is preferable since a well-developed theory exists, making the computations and implementation nicer. The presence of zeros in the data set complicates the modelling. We aim to build a model that takes this into account. This can be done in several ways. One way of handling zero-inflated data, when the non-zero data follow a known distribution, is an approach in the spirit of Allcroft and Glasbey (2002 and 2003) and Glasbey and Nevison (1998). They think of the data as thresholded Gaussian variables, where the non-zeros form an upper part of the distribution. This model imposes an assumption of zero observations being an indication of small values of the underlying process, and is sensible in some applications, as in e.g. Natvig and Tvete (2006). In our situation the eggs will typically appear in clusters. A zero observation corresponds to missing a cluster. It is therefore natural for us to think of a zero observation as something occurring with low probability. We therefore choose to apply a method along the lines of Dobbie and Welsh (2001), but in a space-time setting and with continuous data. They describe a two stage modelling approach to zero-inflated data. They first model the presence/absence of zeros by a logistic model and then, conditioned on the presence of the non-zero counts, these are modelled by a truncated Poisson model. We also model the presence/absence of zeros by a logistic model, but then, conditioned on the presence of non-zero counts, these are modelled by a Gaussian space-time model. The egg counts are modelled according to an underlying egg intensity that both effects the zero and the non-zero observations. Such a connection was proposed by Lambert (1992), in a non-spatial setting, and commented on by Agarwal et al. (2002), in a spatial setting (but not carried through). Papers concering egg abundance where the excessive num-

bers of zeros are taken into consideration include Pennington (1983) and Fox et al. (2000). The problem of excessive zeros is also discussed in Brochers et al. (1997), but not modelled. Both Pennington (1983) and Fox et al. (2000) separate the zero from the non-zero observations. They model the non-zero observations as either lognomal or through a GAM approach, respectively. The analysis in Pennington (1983) is simple, not spatio-temporal and the probability of zero observations is not explicitly computed. The model in Fox et al. (2000) is spatio-temporal and the probability of zero observations is modelled similar to the expectation process for the non-zero observations, but no direct connection between them is made (i.e. the two processes just include the same covariates).
In our modelling we obtain a hierarchical model which we analyze in a Bayesian framework. Inference is performed through MCMC simulations. We choose to model each year separately and compare the fit over years.

A description of the data is given in Section 2. The model is presented in Section 3. Some comments on Bayesian inference is given in Section 4, while the results of fitting the model to separate years are presented in Section 5. A comparison over the years is done in Section 6, and some final remarks and a conclusion are given in Section 7.

2. Data

Our objective is to examine the time and location of walleye pollock spawning in the western Gulf of Alaska, from Kodiak Island to Unimah Pass, determining the area within which the majority of the pollock spawns. In the years $1978 - 2000$ (with 1980 missing) walleye pollock eggs were sampled within the area $52 - 61^\circ$ N, $139 - 168^\circ$ W. Sampling was also done in 1972, but with only a few samples from this year (43) and a fairly long time gap to the next year of sampling, it is omitted from the analysis. The samples are densities, eggs per m^2, taken at a given longitude, latitude, Julian day (hereafter only denoted day) and bottom depth, recorded together with other information. We only include samples taken between (and including) the 76th and 149th day, and bottom depths between (and including) 33 and 403 m below sea surface, which include most of the area and days within the year when eggs can be found. A similar restriction was made by Cianelli et al. (2006), who analyzed the non-zero data through GAM-modelling. The top plot in Figure 1 displays the sampled days within each year and the bottom plot the sampled days versus counts, all the years taken together, with counts above zero on the log scale. Traditionally the majority of walleye pollock spawn in the last week of March and the first week of April, with the Shelikof Strait as the preferred spawning grounds. From the bottom plot in Figure 1 we see a peak in count size around day 100. We can see from the top plot in Figure 1 that for some of the years the sampling went on for a few days, then stopped for a while, and then continued for a few days more. In 1979, 1983 and the years after 1994,

sampling was done only late in the season, many days after the peak days seen in the other years. These years were therefore neglected in this study.

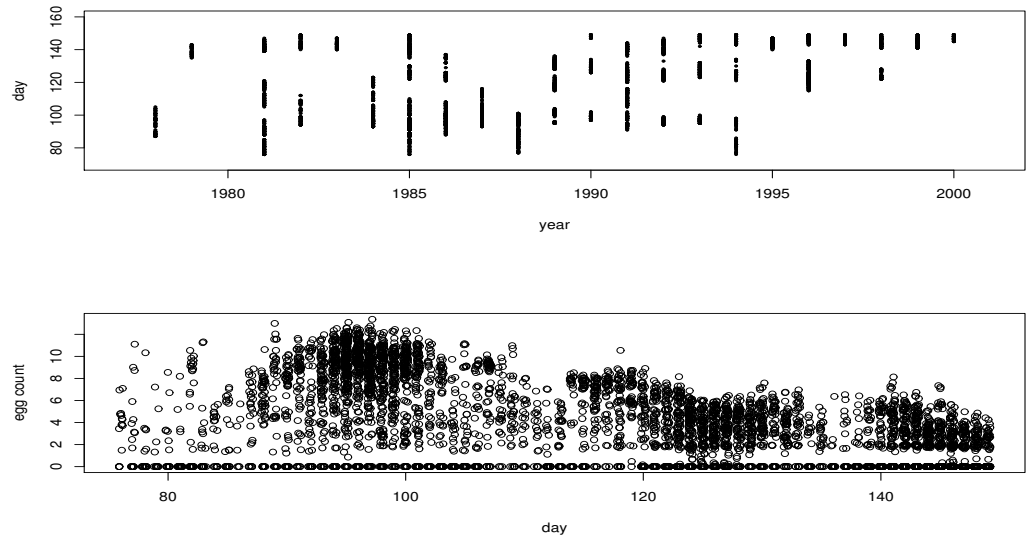

Figure 1: Top plot: range of the sampled days within each year. Bottom plot: the sampled days versus the counts, with counts above zero on the log scale.

For each year a different number of boats sampled at different locations in space and time. The amount of data varies from year to year. Table 2 in the Appendix displays some summary statistics for all the years. Over all the years, about 30% of the data consists of zero counts, for some of the years as much as 70%. Considering all the years, about 35 % of the counts above zero are greater than 200 and about 26 % are greater than 1000. These numbers vary over the years, but the tendency is the same for all years. Typically sampling was done within the same region several times a day, and often a few zero samples are collected on the same day within the same region as high counts. The plots in Figure 2 display two representative years regarding day and location of sampling, 1984 and 1993. The observations are plotted as circles, together with the borders of 5 regions within the total area. These regions will both be used in the modelling step (temporal structure depending on region) and to describe changes in spatial structure over the years.

The top plots in Figure 4 display the spatial distribution of the positive egg densities, on log scale, for 1984 and 1993. These plots have been generated in S-plus by standard kriging techniques (applying an exponential correlation structure). Figures 9 and 10 in the Appendix display the same for all the years analyzed. Considering first 3 sub-areas; the lower left part, the middle part (containing the Shelikof strait region) and the top right part, there are certain features over the years. For years where the sampling was done throughout the peak days, as in 1984, the zeros and the non-zeros in the lower

left part are often mixed. In the middle part, where most of the sampling was done and the highest values registered, there are considerably fewer zeros, but still some. In the top right part again zeros and non-zeros are mixed, for some years more than others.
In 1984, 64 out of a total of 198 observations are zero. The numbers for 1993 are 99 out of 300. For 1984 the maximal count is 98415.6 eggs per m^2, taken on day 95, and for 1993 the maximal count is 39643.85 eggs per m^2, taken on day 96. On day 95 in 1984 the 5 largest counts collected that year, all above 6000 eggs per m^2, were collected south west in the Shelikof strait, while on day 96 in 1993 the 3 largest counts collected that year, all above 2000, were collected somewhat more north-east in the Shelikof strait.
The eggs are not necessarily evenly distributed in the water, but could rather be aggregated in "hot spots", possibly depending on present eddies. When sampling, one might hit a "hot spot" and thereby achieve a high count, or miss a "hot spot" resulting in a zero count. The result is a dataset with some zeros and high counts close in space and space-time. Hence, we could have a very high local variation at certain areas and time points, compared to the overall variation. These eddies can occur and disappear, be small or large. An important question is whether this causes the process to be inhomogeneous. According to biologists the clustering effects of these eddies are random, in the sense that when having a "hot spot" at one place and a "none hot spot" at a nearby place at a given time point , the reverse is just as likely at a next time point. Another factor, not taken into consideration, is the diffusion and drift of eggs from spawning time to the time of sampling. According to biologists the Alaska Coastal current is a subsurface current and does not cause a serious displacement of the eggs, since spawning usually takes place at greater depths.

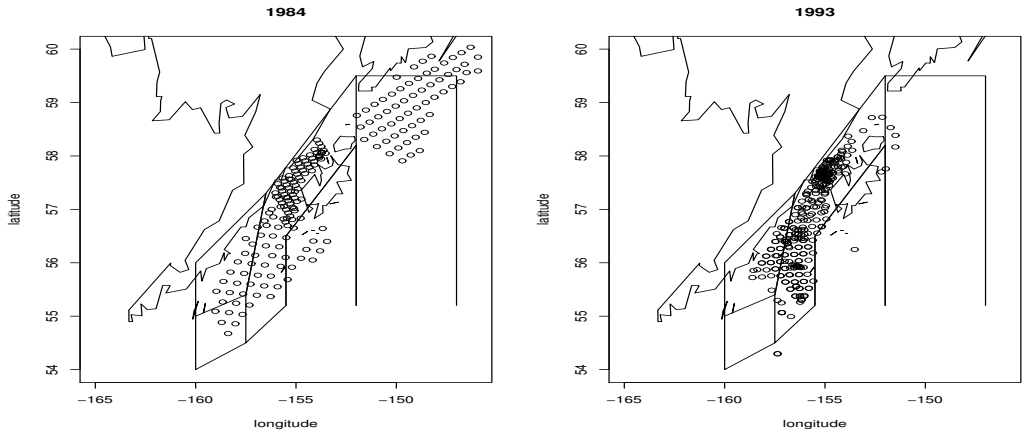

Figure 2: The spatial observations plotted as circles, together with the borders of 5 regions within the total area.

3. The model

Working with space-time models, a challenge is to construct models that describe the covariance structure that is present in the given data. Correlation can be present within non-zero data, but the source of non-zero data can also influence the probability of zero data in nearby observations. Most of the variation from year to year is due to climatic variability. Since the latter has a low year-to-year dependence, each year is modelled separately. The model will be presented for one year, suppressing reference to year in the notation.

3.1. Modelling the observed egg densities

We model the egg densities through an underlying process $\{\lambda\}$, describing the intensity of eggs at spatio-temporal locations. The $\{\lambda\}$ process is interpreted as the underlying intensity for egg densities for a given small spatial area for a time span of a few hours. This process influences both the presence of zeros *and* the amount of eggs in non-zero observations. We let M_i be the egg count for observation i, at location longitude$_i$, latitude$_i$ and day$_i$, $i = 1, ..., n$, and assume

$$M_i = \begin{cases} 0 & \text{with probability } p_i \\ f(Z_i) & \text{with probability } 1 - p_i \end{cases}, \tag{1}$$

where both p_i and the expectation of Z_i depend on the intensity λ_i (with a slight abuse of notation we will use λ_i as the intensity at the spatio-temporal location for observation i and similarly for p_i and Z_i). f is an appropriate transformation that makes it reasonable to assume Z_i to be Gaussian distributed. Given $\{\lambda\}$ we further assume that the Z_i's are independent with expectation λ_i and common variance σ^2_λ. Hence, $Z_i \sim N(\lambda_i, \sigma^2_\lambda)$.

3.2. Modelling the presence/absence of zeros

The probability process p_i models the discrete presence/absence of zeros in the egg density process. The data indicates that regions with high density counts have fewer zeros and vice versa. We therefore choose to make p_i dependent upon the intensity process λ_i. In particular, we will assume $\text{logit}(p_i) = \xi - \kappa\lambda_i$. Local variations in λ_i is then reflected directly in p_i. The idea is that a high value of λ_i should impose a small value of p_i, and vice versa. Notice that we do not put any constraints on ξ and κ. Hence, they are in theory allowed to be negative, although we should expect κ to be positive.

3.3. Modelling the intensity process

The $\{\lambda\}$ process is specified as a combination of a regression term, a spatially varying term and an independent random variable term, given by

$$\lambda_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \mu_{s_i} + \epsilon_i, \tag{2}$$

where s_i is the spatial location corresponding to observation i.

$\mathbf{X}_i$ is a vector of covariates and $\boldsymbol{\beta}$ is a vector of corresponding regression parameters. μ_{s_i} explains spatial variation not captured in the regression term. The terms ϵ_i are independent random variables. They describe the local variation in the expectation process not accounted for in the term $\mathbf{X}_i^T\boldsymbol{\beta} + \mu_{s_i}$. We let $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, $i = 1, ..., n$.

$\{\mu_s\}$ is a stationary spatial process with

$$cov(\mu_{s+h}, \mu_s) = \sigma_\mu^2 e^{-\frac{|h|}{R_\mu}}. \tag{3}$$

3.4. Remarks on the model

Conditional on the spatio-temporal intensity process $\{\lambda\}$ and the spatio-temporal probability process $\{p\}$ the egg densities are either zero or follow a Gaussian distribution (on a transformed scale). Two counts are spatio-temporally dependent through $\{\lambda\}$ but conditionally independent given this process. To what degree the intensity process describes both the binary and the non-zero data is reflected in the variance σ_λ^2. A small value of σ_λ^2 indicates a high correlation between these data sets, and vice versa.

Although we have suppressed reference to year in the notation, we emphasize that all variables *and* parameters are assumed to change from year to year.

4. Bayesian inference

Inference will be made in a Bayesian setting with the computation carried out through Markov Chain Monte Carlo (MCMC) simulations. Let $\overrightarrow{\lambda} = (\lambda_1, ..., \lambda_n)^T$. Parallel to $\overrightarrow{\lambda}$ we define $\overrightarrow{p}$, $\overrightarrow{\epsilon}$ and $\overrightarrow{\mu}$. We write $\overrightarrow{\mu} \sim N(0, \Sigma_\mu)$, where the elements of Σ_μ are defined through (3).
With the hierarchical description of the model given in Section 3, the variables to be simulated are $\boldsymbol{\beta}, \overrightarrow{\mu}, \overrightarrow{\epsilon}, \sigma_\mu^2, R_\mu, \kappa, \xi, \sigma_\lambda^2$ and σ_ϵ^2. In order to obtain reasonable mixing properties of the MCMC algorithm, the updating is done in blocks. An option in these cases is to group some highly correlated parameters, and update all the parameters within each block simultaneously. For observations greater than zero, $\mathbf{X}_i^T\boldsymbol{\beta} + \mu_i + \epsilon_i \approx$

$f^{-1}(M_i)$, giving a negative correlation between $\boldsymbol{\beta}$, μ_i, ϵ_i, and a possible block is ($\boldsymbol{\beta}$, $\overrightarrow{\mu}$, $\overrightarrow{\epsilon}$). Since there is a one to one correspondence between ($\boldsymbol{\beta}$, $\overrightarrow{\mu}$, $\overrightarrow{\epsilon}$) and ($\boldsymbol{\beta}$, $\overrightarrow{\mu}$, $\overrightarrow{\lambda}$), and considering $\overrightarrow{\lambda}$ in this case leads to nicer equations, we define a block consisting of ($\boldsymbol{\beta}$, $\overrightarrow{\mu}$, $\overrightarrow{\lambda}$). Due to the zero observations the full conditional distribution of this block is a non-standard one, and we apply a Metropolis Hastings step. For details, see the Appendix. The other four blocks in our MCMC algorithm are (κ), (ξ), $(\sigma_\mu^2, \sigma_\lambda^2, \sigma_\epsilon^2)$ and (R_μ). Given $\boldsymbol{\beta}$, $\overrightarrow{\mu}$, $\overrightarrow{\lambda}$, κ and ξ, the variances $\sigma_\mu^2, \sigma_\lambda^2$ and σ_ϵ^2 can be sampled independently of each other. The details of the algorithm are again given in the Appendix.

A priori we let $\boldsymbol{\beta} \sim N(0, I_\beta \sigma_\beta^2)$, where I_β is an identity matrix. Prior assumption for R_μ is given by $R_\mu \sim U(a_{R_\mu}, b_{R_\mu})$. We assume a priori that $\kappa \sim N(0, \sigma_\kappa^2)$ and $\xi \sim N(0, \sigma_\xi^2)$. $\sigma_\beta^2, a_{R_\mu}, b_{R_\mu}$, σ_κ^2 and σ_ξ^2 are parameters to be specified. a_{R_μ} and b_{R_μ} are chosen to be $0.5 * min(|h|)$ and $0.5 * max(|h|)$, respectively, where h is given in equation (3).
For the simulations performed in the next section, we have specified the following. $\sigma_\beta^2 = 5$. $\sigma_\kappa^2 = 10$ and $\sigma_\xi^2 = 10$. The presicion $1/\sigma_\epsilon^2$ is Gamma distributed with expectation 1 and variance 50. The presicion $1/\sigma_\mu^2$ is Gamma distributed with expectation 1 and variance 100. Finally, the presicion $1/\sigma_\lambda^2$ is Gamma distributed with expectation 1 and variance 11.11.

5. Results

The years are analyzed separately, and results for 1984 and 1993 are discussed in detail. The MCMC chain was run with 100 000 iterations (where 60 000 iterations were disregarded as "burn-ins") to obtain posterior samples of the variables. In the analysis we have used longitude (lon) , latitude (lat), bottom depth ($depth$) (on log scale) and day (day) as covariates. Day includes both a linear term and a quadratic term, and is assumed to be region specific, using the 5 regions defined in Section 2, and displayed with numbering in Figure 6. In addition to the other three covariates and a constant term this results in a total of 14 regression coefficients. The covariates are transformed to be more uncorrelated to the constant term by subtracting the mean levels over all observations. This gives

$$\begin{aligned}\lambda_i = \beta_1 + \beta_2(lon_i - \bar{lon}) + \beta_3(lat_i - \bar{lat}) + \beta_4(depth_i - \bar{depth}) + \\ \beta_{5,l_i}(day_i - \bar{day}) + \beta_{6,l_i}(day_i - \bar{day})^2 + \mu_{s_i} + \epsilon_i, \end{aligned} \quad (4)$$

where l_i denotes the region corresponding to location of observation i, $l_i = 1, ..., 5$. For years where there are no observations in some of the regions the dimension of $\boldsymbol{X}_i$ is reduced. As noticed in Figure 2 some of the counts were sampled outside the 5 regions. They are taken to belong to their closest region. We have chosen to model the

intensity process on the log scale, that is, $f(Z_i) = \exp(Z_i)$.

Figure 3 displays the residuals $\log(M_i) - \hat{\lambda}$ for $M_i > 0$ plotted against day. There is an overweight of residuals above zero. This is natural as the residuals are only for egg counts above zero. The residuals are not too large and we see no apparent structure. Hence, we conclude that the model fit for the non-zero data is fairly good for 1984 and 1993.

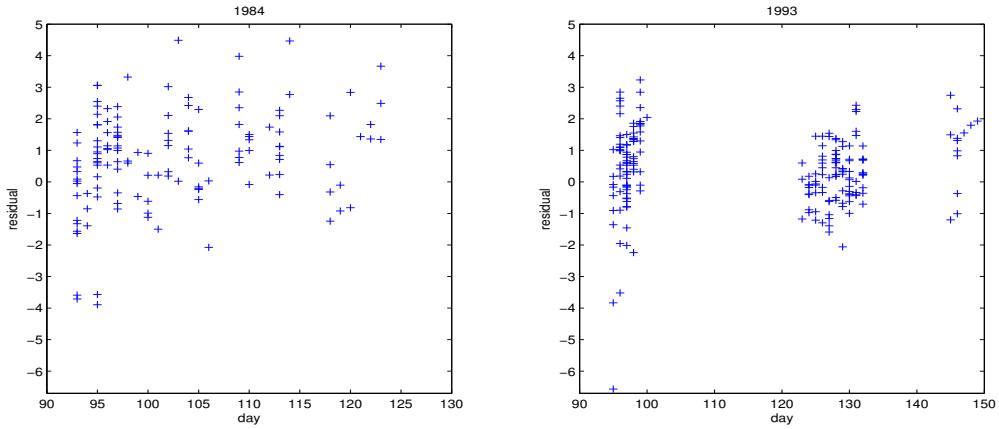

Figure 3: Residuals, $\log(M_i) - \hat{\lambda}_i$ for all $M_i > 0$, plotted against day.

In examining how well the model is able to capture the spatial structure in the data we plot the spatial term $spat_i = \hat{\beta}_1 + \hat{\beta}_2(lon_i - \bar{lon}) + \hat{\beta}_3(lat_i - \bar{lat}) + \hat{\beta}_4(depth_i - \bar{depth}) + \hat{\mu}_{s_i}$ in the bottom row in Figure 4. Comparing these plots with the spatial plots of the data in the top row, keeping in mind the different scales, we see that the model captures the main features of the spatial structure in the data. The high and low intensity regions are roughly identified.
For considering the temporal fit we define $temp_i = \hat{\beta}_{5,l_i}(day_i - \bar{day}) + \hat{\beta}_{6,l_i}(day_i - \bar{day})^2$. If the model fit is good the term $\log(M_i) - spat_i$ for all $M_i > 0$ should be close to the temporal term (i.e. σ_ϵ^2 is small) (see (2)). Hence, we plot in Figure 5 both these two terms against day of sampling. If the spatial fit is good (displayed in Figure 4) and the $temp_i$ is close to $\log(M_i) - spat_i$ for all $M_i > 0$ over the sampled days, we conclude that the temporal fit is satisfactory. $\log(M_i) - spat_i$, for all $M_i > 0$, are plotted as points with region coding (blue). $temp$ are also plotted as points with region coding, but connected with lines (red). From the plots in Figure 5 we see that the temporal effect on egg density for several regions are not necessarily linear. We especially notice how the temporal effect on egg density varies for different regions. As an example, consider region 1, containing the Shelikof strait. In 1984, where most of the sampling in this region was done early, the temporal effect on egg density in this

region is slightly convex. In 1993, where the sampling in this region was done from the peak days throughout the season, it is piecewise linear with large negative slopes.

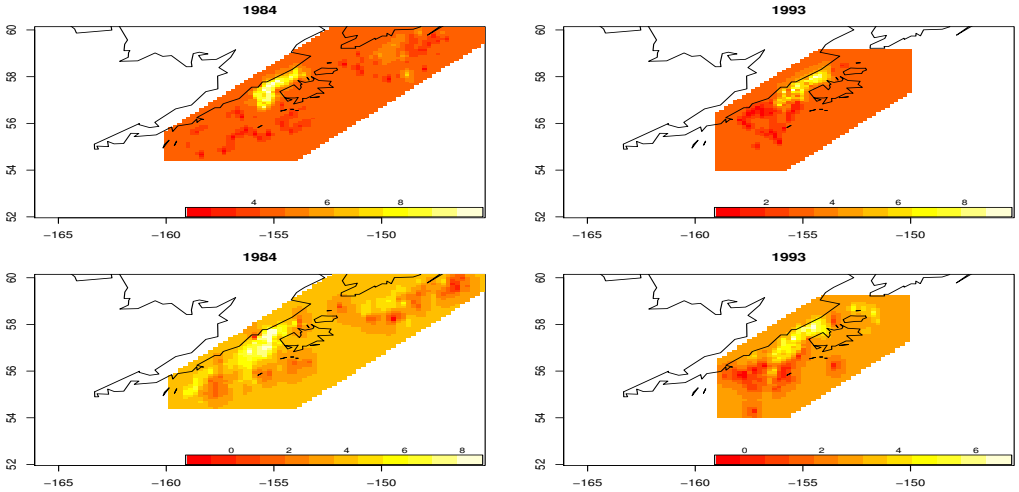

Figure 4: Top row: spatial distribution of the collected egg densities for 1984 and 1993 based on smoothing the non-zero data. Bottom row: the spatial fit, $\hat{\beta}_1 + \hat{\beta}_2(lon_i - \bar{lon}) + \hat{\beta}_3(lat_i - \bar{lat}) + \hat{\beta}_4(depth_i - \bar{depth}) + \hat{\mu}_{s_i}$, $i = 1, ..., n$, for the same years.

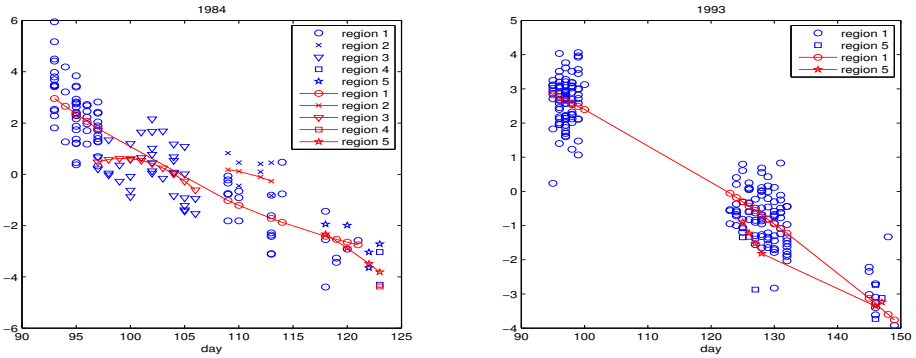

Figure 5: The temporal fit: $\log(M_i) - spat_i$, for all $M_i > 0$, $i = 1, ..., n$, are plotted as points with region coding (blue). $temp_i$, $i = 1, ..., n$, are also plotted as points with region coding, but connected with lines (red).

Table 1 displays the MCMC posterior means with standard deviations for the space-time independent variables entering the model for the years 1984 and 1993. The estimated coefficients for longitude and latitude, β_2 and β_3, indicate a positive north east direction for the spatial influence of the egg densities for all the three years. The estimated coefficient for bottom depth, β_4, indicates an increasing egg density with greater

depths. Many of the temporal coefficients, β_{5,l_i} and β_{6,l_i}, $r_i = 1, ..., 5$, have negative signs for all the three years, resulting in the temporal fit on egg density as displayed in Figure 5. σ^2_λ is estimated to be small for 1984 and 1993. That is, the intensity process describes both the binary data and the non-zero data well for these two years.

Parameter	*1984*	*1993*
β_1	4.2116(0.1420)	3.2065(0.0883)
β_2	0.7880(0.2245)	1.2608(0.1218)
β_3	-0.5320(0.0941)	-0.1922(0.1258)
β_4	2.3610(0.2776)	1.2989(0.0810)
β_{5,l_1}	-22.591(1.119)	-12.273(0.220)
β_{5,l_2}	11.738(4.163)	-24.970(11.067)
β_{5,l_3}	-26.157(5.864)	-31.088(11.031)
β_{5,l_4}	-21.455(11.881)	
β_{5,l_5}	-7.232(9.719)	-39.164(06.425)
β_{6,l_1}	3.755(2.268)	-0.737(0.248)
β_{6,l_2}	-16.510(4.245)	-24.509(11.752)
β_{6,l_3}	-27.390(3.112)	-13.395(11.884)
β_{6,l_4}	-0.875(6.070)	
β_{6,l_5}	-6.796(5.594)	10.598(2.323)
σ^2_μ	2.0585(0.3451)	0.6364(0.2204)
R_μ	0.2621(0.0654)	0.2558(0.2243)
ξ	2.1161(0.3240)	0.5964(0.2035)
κ	0.8464(0.0744)	1.0363(0.1360)
σ^2_ϵ	0.4235(0.0484)	0.4869(0.0462)
σ^2_λ	0.4069(0.0535)	0.1050(0.0109)

Table 1: Posterior means and standard deviations, based upon the MCMC simulations of the parameters.

Results from fitting all the years are displayed in the Appendix. Figures 11 and 12 display the spatial fit for all the years, as described in the beginning of this section, and shown in Figure 4 for 1984 and 1993. Figures 13 and 14 display the temporal fit, also as described in the beginning of this section, and as shown in Figure 5 for 1984 and 1993. Comparing the spatial fit for all the years in Figure 11 and 12 with the plots in Figure 9 and 10 in the Appendix, displaying the distributions of egg densities, and keeping in mind the different scales, we see that the spatial fit is captured by the model for most years. For some years there are a few extreme values in the Shelikof strait, and the model does not seem to be able to capture these in the spatial fit, as is seen for e.g. 1986. There might be a too abrupt jump from small to extreme densities. High values in the spatial fit is typically found on the Shelf of the Alaska Peninsula, along the Alaska Peninsula and in particular in the Shelikof strait. 1991 seems to be

an exception, with low values along the Alaska Peninsula and high values in the far north-east. Comparing the plots in Figures 11 and 12 with those in Figures 9 and 10 for 1978 the spatial fit is not as good as for the other years. In this year as much as 67.12 % of the observations are zero. In this year the estimated σ_μ^2, σ_ϵ^2 and σ_λ^2 were 2.96, 2.14 and 2.10, respectively. σ_λ^2 is for a few of the other years estimated to be large as well, but for those years σ_μ^2 and σ_ϵ^2 are estimated to be small.

Looking at the temporal fit for the years in Figures 13 and 14 we see that the temporal effect on egg density clearly varies for the different regions, according to day of sampling over the years. That is, there clearly is an interaction in space and time. For many of the years the temporal effect on egg density is linear or close to linear for many regions, but far from all. In sampling prior to around day 100 there tends to be an increasing egg density with day, while sampling posterior to this day the relationship is opposite. This is natural, and in accordance with the idea of a spawning peek around day 100.

6. Comparing the intensity level over years

As stated in Section 5, our ultimate goal is to compare the fitted underlying intensity process λ over the years. We wish to make the comparison on day 100, the approximate peak day. That is, we are interested in comparing the fitted $\lambda_i^{spat} = \beta_1 + \beta_2(lon_i - \bar{lon}) + \beta_3(lat_i - \bar{lat}) + \beta_4(depth_i - \bar{depth}) + \beta_{5,l_i}(100 - \bar{day}) + \beta_{6,l_i}(100 - \bar{day})^2 + \mu_{s_i}$ over the years. Let us consider possible changes in the 5 regions presented in section 2. To compare the fitted λ_i^{spat} over years we choose a center point and a few surrounding reference points within the 5 regions, being the same for each year. The center and reference points are displayed in Figure 6. In each of the reference points, r, we calculate λ_r^{spat}. Letting $\overrightarrow{\mu}_r$ be the vector of μ values for the reference points, we note that

$$\begin{pmatrix} \overrightarrow{\mu} \\ \overrightarrow{\mu}_r \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\mu & \Sigma_{\mu r} \\ \Sigma_{\mu r}^T & \Sigma_r \end{pmatrix} \right),$$

where Σ_μ is the covariance matrix of $\overrightarrow{\mu}$, given by (3), $\Sigma_{\mu r}$ is the covariance matrix of $\overrightarrow{\mu}$ and $\overrightarrow{\mu}_r$, and Σ_r is the covariance matrix of $\overrightarrow{\mu}_r$, also found by (3).
Since $[\overrightarrow{\mu}_r | \overrightarrow{\mu}] \sim N(\Sigma_{\mu r}^T \Sigma_\mu^{-1} \overrightarrow{\mu}, \Sigma_r - \Sigma_{\mu r}^T \Sigma_\mu^{-1} \Sigma_{\mu r})$ we can sample $\overrightarrow{\mu}_r$. Having $\overrightarrow{\mu}_r$ we can compute λ_r^{spat}. In order to obtain one value representing the egg density at the center point of each region, we compute, for each region, $\sum_{r=1}^R w_r \lambda_r^{spat}$, where w_r is the weight for reference point r and R the total number of reference points in the given region. We choose w_r to be the Euclidean distance between the center point and the reference point, scaled to sum to 1. Letting l denote region, $l = 1, ..., 5$, we write $\lambda_l^{sc} = \sum_{r=1}^R w_r \lambda_{l,r}^{spat}$ for each region. This is done for every 100th iteration in the MCMC sampling, giving the posterior distribution for λ_l^{sc} in every region l. We omit 1978

since this year is not fitted well.

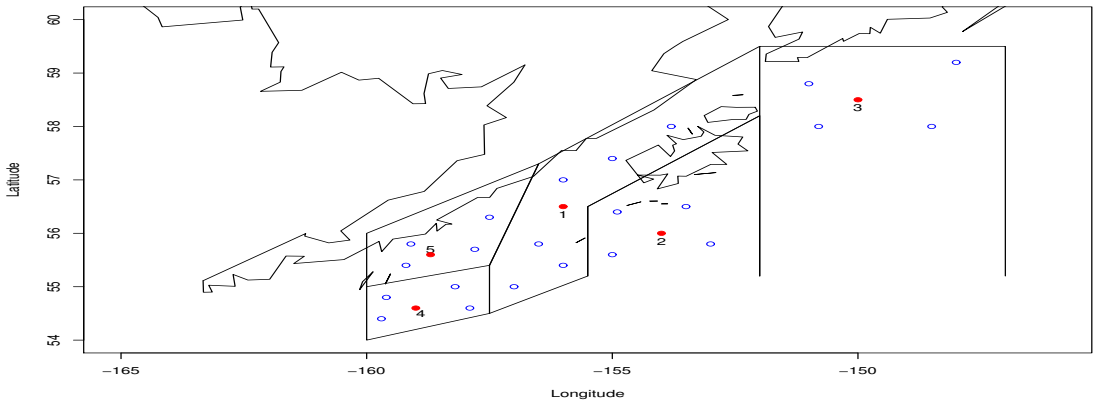

Figure 6: Solid, red circles display the center points. Open, blue circles display reference points.

The estimated means and 95 % credibility bands based on the posterior distributions of λ^{sc} for the 5 regions are displayed in Figure 7, along with the mean and 95 % credibility bands for the sum over the 5 regions. For region 1 the estimated level is slightly increasing from 1981 through 1984, thereafter slightly decreasing through 1988. From 1988 to 1989 there is an abrupt upward shift in the estimated level, reaching a peak in 1990, and thereafter it decreases. In region 2 the estimated level is fairly stable, with small estimated credibility bands, through 1991. In 1992 there is a large abrupt downward shift in the estimated level, which increases over the next years. In region 3 the level is more or less increasing through 1988. In 1989 the estimated level abruptly shifts downwards, where it more or less remains through 1993, when it shifts somewhat down again. In region 4 the estimated level fluctuate over the years through 1991. In 1992 it is also here an abrupt downward shift in the estimated level, but it shifts up in 1993, and back down again in 1994. In region 5 the estimated level varies somewhat over the years through 1992, then making an abrupt upward shift in 1993 to an estimated level remaining more or less the same in 1994. In addition, we notice the wide credibility bands for the last three years for all the regions. This could be due to the fact that, although sampling covered day 100, the bulk of the sampling occurred later in the years. In addition the majority of the late sampling took place rather to the south-west of the Shelikof strait, that is the south-west part of region 1, the south-east part of region 5 and the north-east part of region 4. This is opposed to the early sampling these years, which took place in the Shelikof strait. This results in some uncertainty about the level early in the season. When considering the bottom right plot, for the sum over the five regions, we again see the abrupt shift downwards from 1991 to 1992. Then, in 1993 it shifts back again to the old level. We also here notice the large credibility bands for the last three years.

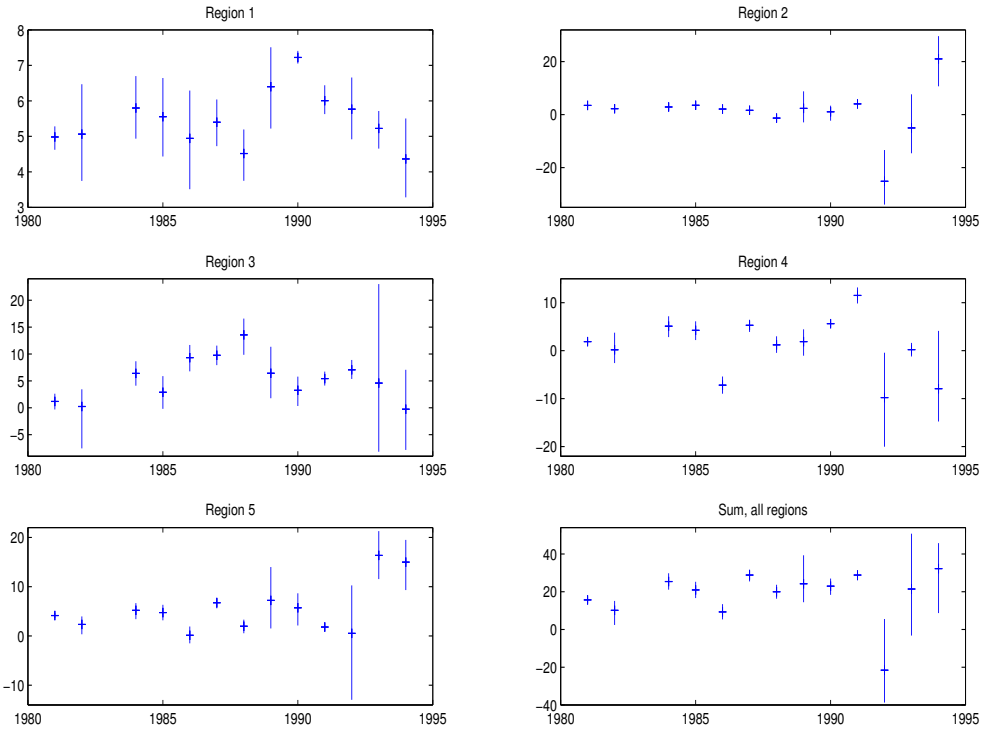

Figure 7: Mean and 95 % credibility bands based on the posterior distribution of λ^{sc} over the years for the 5 regions and their sum.

It is also of interest to examine how the estimated λ^{spat} in each region varies over the years relative to the other regions. Let $\lambda_l^{sc*} = \lambda_l^{sc} - \sum_{k=1}^{5} \lambda_k^{sc}/5$, and computing this for every 100th iteration we obtain the posterior distribution of λ_l^{sc*}. The resulting means and 95 % credibility bands based on the posterior distributions of λ^{sc*} for the 5 regions are displayed in Figure 8.

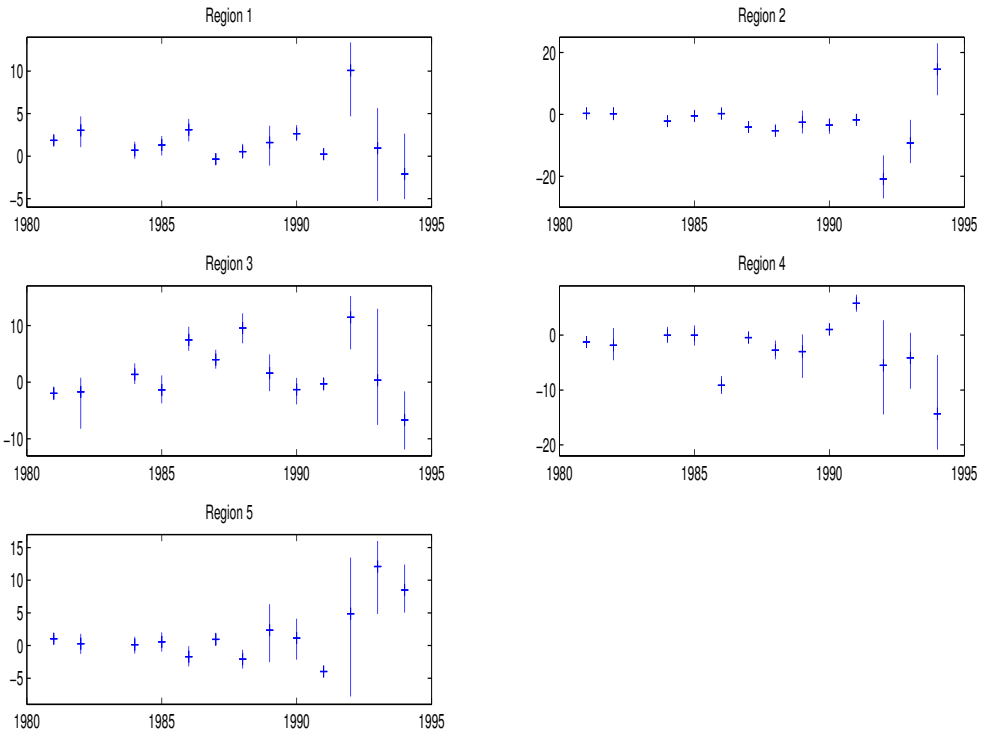

Figure 8: Mean and 95 % credibility bands based on the posterior distribution of λ^{sc*} over the years for the 5 regions.

Looking at Figure 8 we see an abrupt change from 1991 to 1992 in all the regions. The relative estimated level in region 1 compared to the other years is mostly positive, with the opposite being the case for region 2 and 4. We also notice that the relative estimated level is fairly stable in regions 1, 2 and 5 before 1992, compared to region 3 and 4, where it fluctuates quite a lot. We also notice that in the years 1986 and 1988 the relative estimated level shifts somewhat upwards in region 3, with the opposite being the case in region 4. As in Figure 7 we also here notice the wide credibility bands for the last three years.

7. Final remarks and conclusions

In settings where it is appropriate to think of a zero observation as something occuring with low probability a popular approach is the two stage model applied here. By applying such a model we are able to capture the space-time process in the egg densities

well for most years. That is, the space-time regression term $\mathbf{X}_i^T\boldsymbol{\beta}$ and the spatial $\overrightarrow{\mu}$ term describe most of the variation in the data, as is seen by the low estimated σ_λ^2 for most years.

In our model we have included an extra random effect in the modelling of the positive data, that is $\text{var}(Z_i) = \sigma_\lambda^2$. As an alternative we could instead have modelled this random effect in the binary process, that is, suppressing space-time notation, $\text{var}(\text{logit}(p)|\xi, \kappa, \lambda) = \sigma_\lambda^2$. This is an alternative approach worth keeping in mind for this type of modelling in the future.

The local variation in the expectation process not accounted for in $\mathbf{X}_i^T\boldsymbol{\beta} + \mu_{s_i}$, ϵ_i, has an estimated small variance for all the years except 1978, and we conclude that the term $\mathbf{X}_i^T\boldsymbol{\beta} + \mu_{s_i}$ is able to capture the major space-time variation in the underlying intensity process for the egg densities. The temporal dependency on the underlying intensity varies with region. This dependency is often linear or close to linear, but not always. Generally the temporal effect increases with day before the peak day of about 100. Afterward the effect is usually the opposite. How strong this dependency is, varies between the different regions. This space-time interaction is an important feature of the model, and ignoring it could lead to erroneous results.

Analyzing temporal changes over years could give valuable information regarding possible major shifts in spawning habits, and a contribution to explaining environmental changes in the area. Instead of fitting the model for each year separately we could have constructed one model for all the years. When obtaining data for a new year we would then have to run the model for all the years (with data from the previous years and the new year). When looking at the data we find it plausible to allow for both the binary and the underlying intensity process to be different over the years. If one should choose to incorporate the year-to-year dynamics the model would need the flexibility to allow many different changes, both smooth and abrupt. This is a complex modelling task. Hence, we model each year separately.

We are particularly interested in possible changes in the spatial structure of the egg densities over the years. We are looking for evidence of walleye pollock changing preferences regarding spawning area, and if we find it, whether the spatial spawning pattern is changing in a particular direction as the years pass by. An examination of the fitted zero-inflated spatio-temporal Gaussian distribution model over the years sheds light on this. The objective of this study has been to examine possible changes in the spatial influence on the spawning pattern. By examining the underlying intensity process λ over the years that had a sampling interval covering day 100 we can from Figure 7 see some main features. In region 1 and 3 there is a shift in the estimated level in 1989. In region 1 it shifts upwards and thereafter decreases, while in region

3 it shifts downwards to a fairly stable level until 1994. In region 2 and 4 there is a shift downwards in 1992, while in region 5 there is a shift upwards in 1993. When comparing the regions relative to each other in Figure 8 we see a shift in 1992, upwards in region 1, 3 and 5, downwards in region 2 and 4. Hence, regions 1, 3 and 5 became more important relative to the other regions in 1992, while regions 2 and 4 became less important. In 1993 and 1994 regions 1, 3 and 4 loose some of their relative importance, while in regions 2 and 5 it increases.

From this analysis we see that changes occurred in the Western Gulf of Alaska over the years examined. In 1989 there is a sudden increase in the estimated level in the major spawning region within our area; the Shelikof strait and the area to the south-west of it, while at the same time there is a shift in the north-east region, from an increasing estimated level to a lower, more stable level for the next years. Even though the estimated level in the region containing the Shelikof strait and the area to the south-west of it decreases over the last years it becomes relative important as spawning ground in 1992, at the expense of the area south of Kodiak island and the area to the north-east. In the next years the Shelikof strait and the area to the south-west of it, the north-east region and the south-west region became relative less important, but the region along the Alaska peninsula and the region south of the Kodiak island became more important. Keeping in mind that few samples were taken in the region in the north-east and to the south of Kodiak island we sum up at the end of 1994: the Shelikof strait and the area to the south-west of it lost some of it's relative importance as a major spawning region at the expense of the area along the Alaska peninsula.

Acknowledgments

This analysis is based on walleye pollock egg densities (number of eggs per m^2) collected during the ichthyoplankton surveys of the Alaska Fisheries Science Center (AFSC, Seattle) in the gulf of Alaska, extracted from the Ichthyoplankton Cruise Database (IchBase), conducted by AFSC and partner institutions in the gulf of Alaska. For further details, see Cianelli et al. (2006). We are grateful to Lorenzo Cianelli and Nils Christian Stenseth at the Centre for Ecological and Evolutionary Synthesis (CEES), University of Oslo, for their biological expertise and valuable comments and suggestions.

Appendix

Plots and statistics

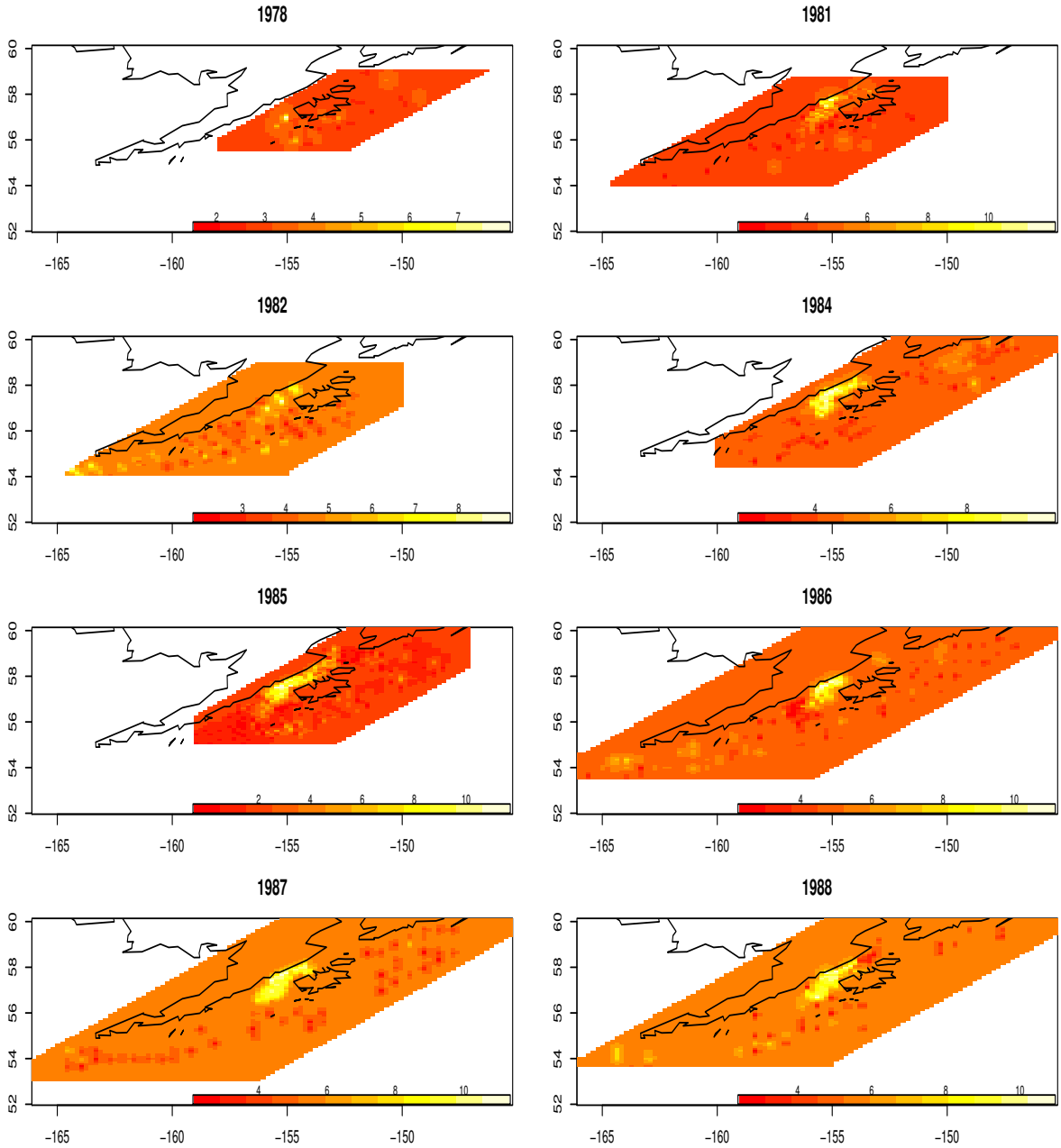

Figure 9: The spatial distribution of the positive egg densities, on log scale, collected in the given area for the years 1978, 1981, 1982, 1984, ..., 1988.

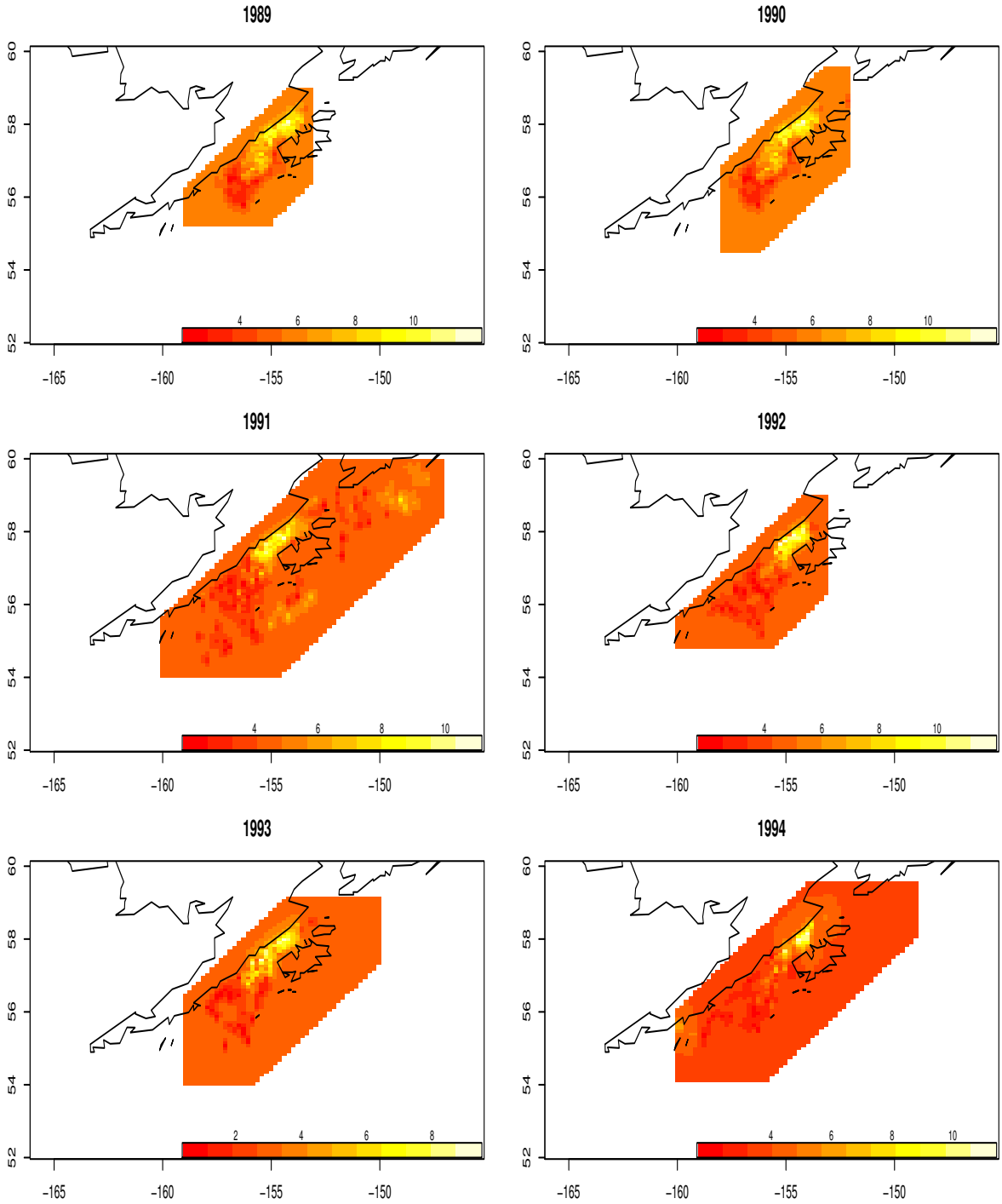

Figure 10: The spatial distribution of the positive egg densities, on log scale, collected in the given area for the years 1989, ..., 1994.

Table 2 displays a statistical summary for all the years. Longitude, latitude, bottom depth and Julian day are all on the original scale. The third column displays the percentage of zero observations for each year. The last four columns contain the minimum and maximum values. Comparing this table with the corresponding one in Ciannelli et al. (2006) we notice that for 1981 and 1999 they differ in that in Ciannelli et al. (2006) the number of observations greater than zero is one less. This is due to the inclusion of bottom depth 33 meters in our analysis, while Ciannelli et al. (2006) considered bottom depths greater than 33 meters.

year	*no. obs* (> 0)	$\%$ 0	*lon*	*lat*	*depth*	*day*
1978	73(24)	67.12	55.9483, 59.0450	-155.4250, -149.5100	37, 367	87, 105
1979	44(14)	68.18	54.2750, 59.0083	-161.6667, -148.8700	45, 250	135, 143
1981	424(298)	29.72	54.0500, 58.3300	-163.3833, -151.5000	33, 375	76, 147
1982	120(93)	22.50	54.0333, 57.8833	-164.4000, -151.7000	35, 370	94, 149
1983	53(16)	69.81	55.2383, 57.2567	-158.7100, -154.6333	41, 297	140, 147
1984	198(134)	32.32	54.6750, 60.0350	-159.0167, -145.9333	34, 299	93, 123
1985	445(324)	27.19	55.2133, 59.5433	-158.2050, -148.5000	33, 390	76, 149
1986	293(245)	16.38	53.6633, 60.2117	-165.5833, -139.3500	55, 382	88, 137
1987	239(171)	28.45	53.6550, 59.6667	-165.0833, -147.7500	40, 380	93, 116
1988	305(258)	15.41	54.0000, 59.6667	-164.3333, -147.0000	37, 363	77, 101
1989	334(302)	9.58	55.6483, 59.0617	-157.4767, -151.7050	57, 355	95, 136
1990	214(185)	13.55	55.2500, 59.0367	-157.4650, -151.5167	49, 320	97, 149
1991	426(328)	23.01	54.4817, 59.7817	-158.8317, -147.8433	39, 329	91, 144
1992	308(214)	30.52	55.1100, 58.3700	-158.6083, -153.6433	36, 336	94, 147
1993	300(201)	33.00	54.2950, 58.7250	-158.6150, -151.4883	40, 375	95, 149
1994	248(161)	35.08	54.8467, 59.0883	-159.9750, -150.1117	41, 328	76, 149
1995	94(53)	43.62	54.2505, 57.7177	-163.5147, -154.7777	63, 314	140, 147
1996	425(285)	32.94	54.0445, 59.9032	-164.7272, -147.9328	37, 304	115, 149
1997	100(38)	62.00	55.0938, 58.3390	-158.6032, -153.5037	34, 311	143, 149
1998	193(102)	47.15	55.0933, 58.7058	-158.3785, -151.3877	40, 350	122, 149
1999	136(66)	51.47	54.1667, 57.7300	-164.7217, -154.5990	33, 305	141, 149
2000	76(26)	65.79	54.1735, 56.8183	-164.7155, -155.8172	34, 293	145, 149

Table 2: Statistical summary for the years.

Results

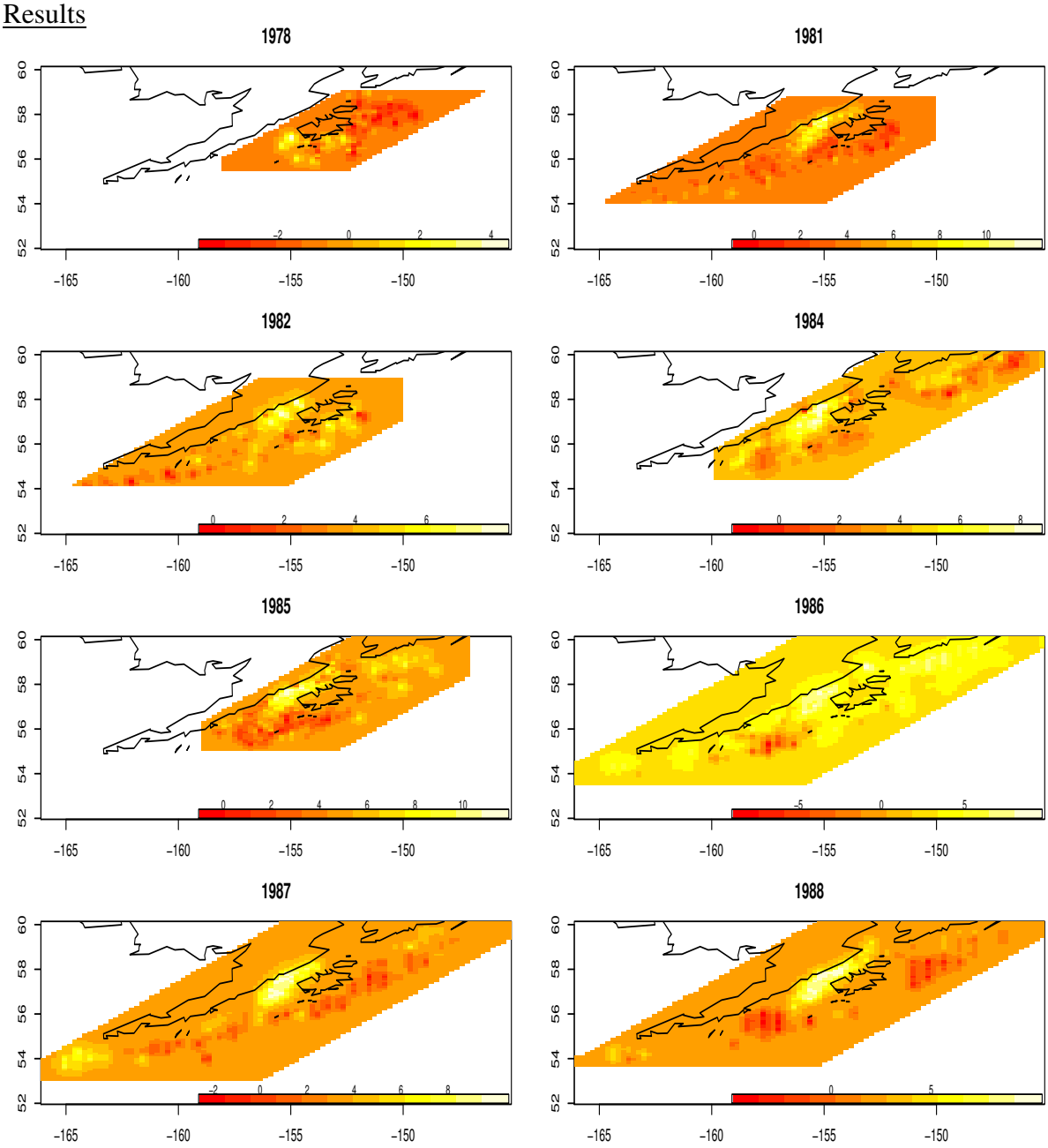

Figure 11: The spatial fit: $\hat{\beta}_1 + \hat{\beta}_2(lon_i - \bar{lon}) + \hat{\beta}_3(lat_i - \bar{lat}) + \hat{\beta}_4(depth_i - \bar{depth})$ $+\hat{\mu}_{s_i}, i = 1, ..., n.$ 1978, 1981, 1982, 1984, ..., 1988.

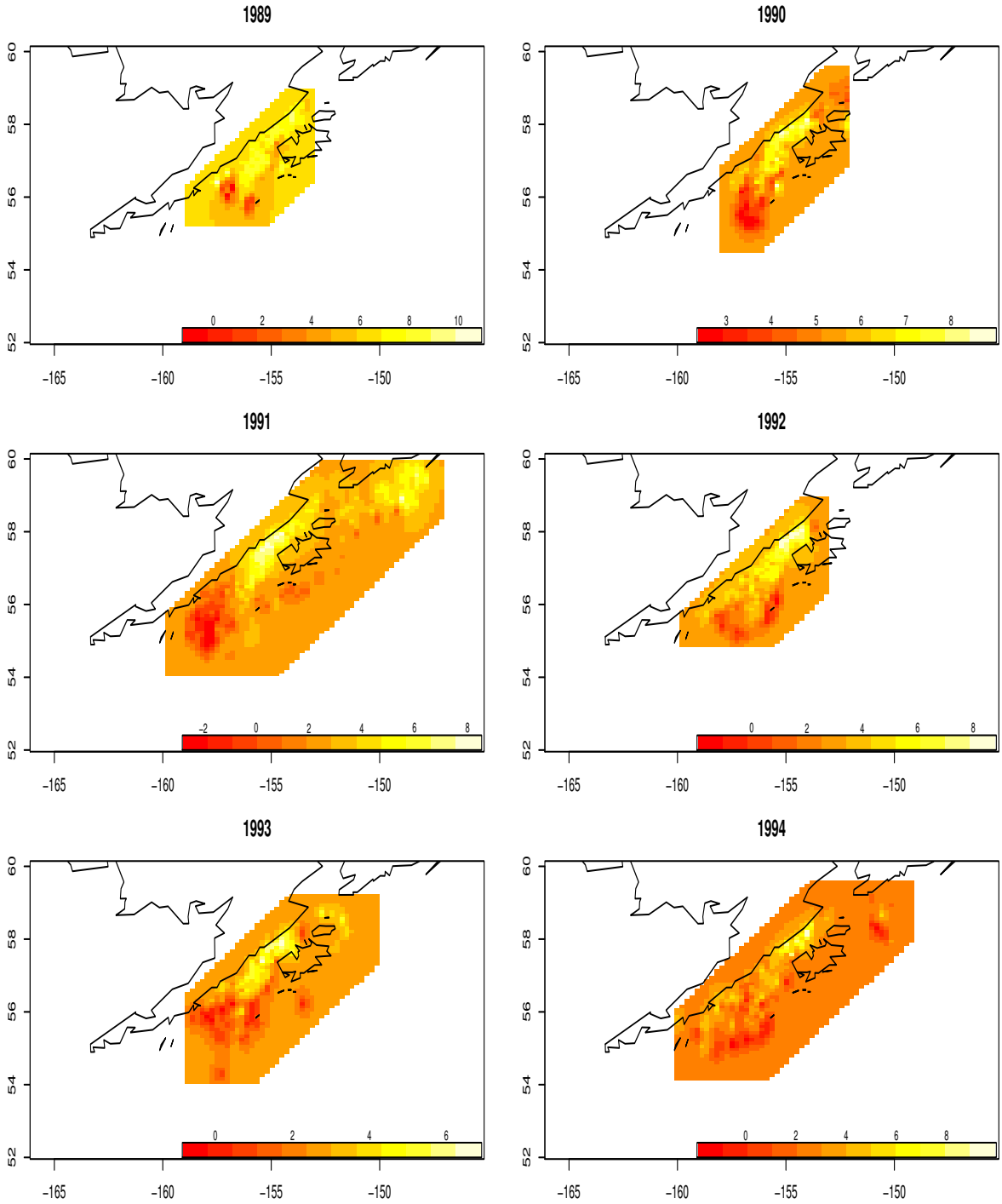

Figure 12: The spatial fit: $\hat{\beta}_1 + \hat{\beta}_2(lon_i - \bar{lon}) + \hat{\beta}_3(lat_i - \bar{lat}) + \hat{\beta}_4(depth_i - \bar{depth})$ $+\hat{\mu}_{s_i}$, $i = 1, ..., n$. 1989, ..., 1994.

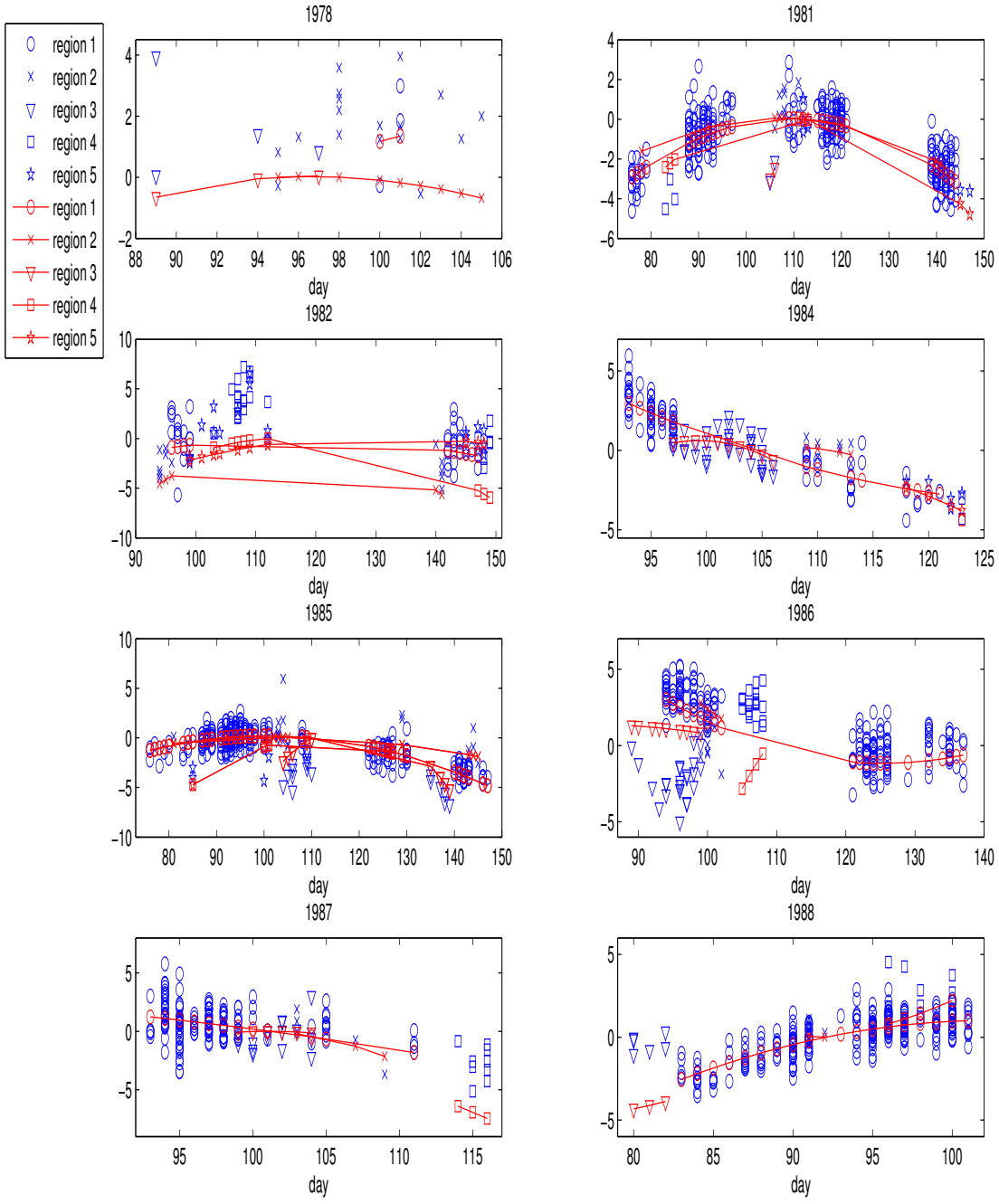

Figure 13: The temporal fit: $\log(M_i) - spat_i$, for all $M_i > 0$, $i = 1, ..., n$, are plotted as points with region coding (blue). $temp_i$, $i = 1, ..., n$, are also plotted as points with region coding, but connected with lines (red). $1978, 1981, 1982, 1984, ..., 1988$.

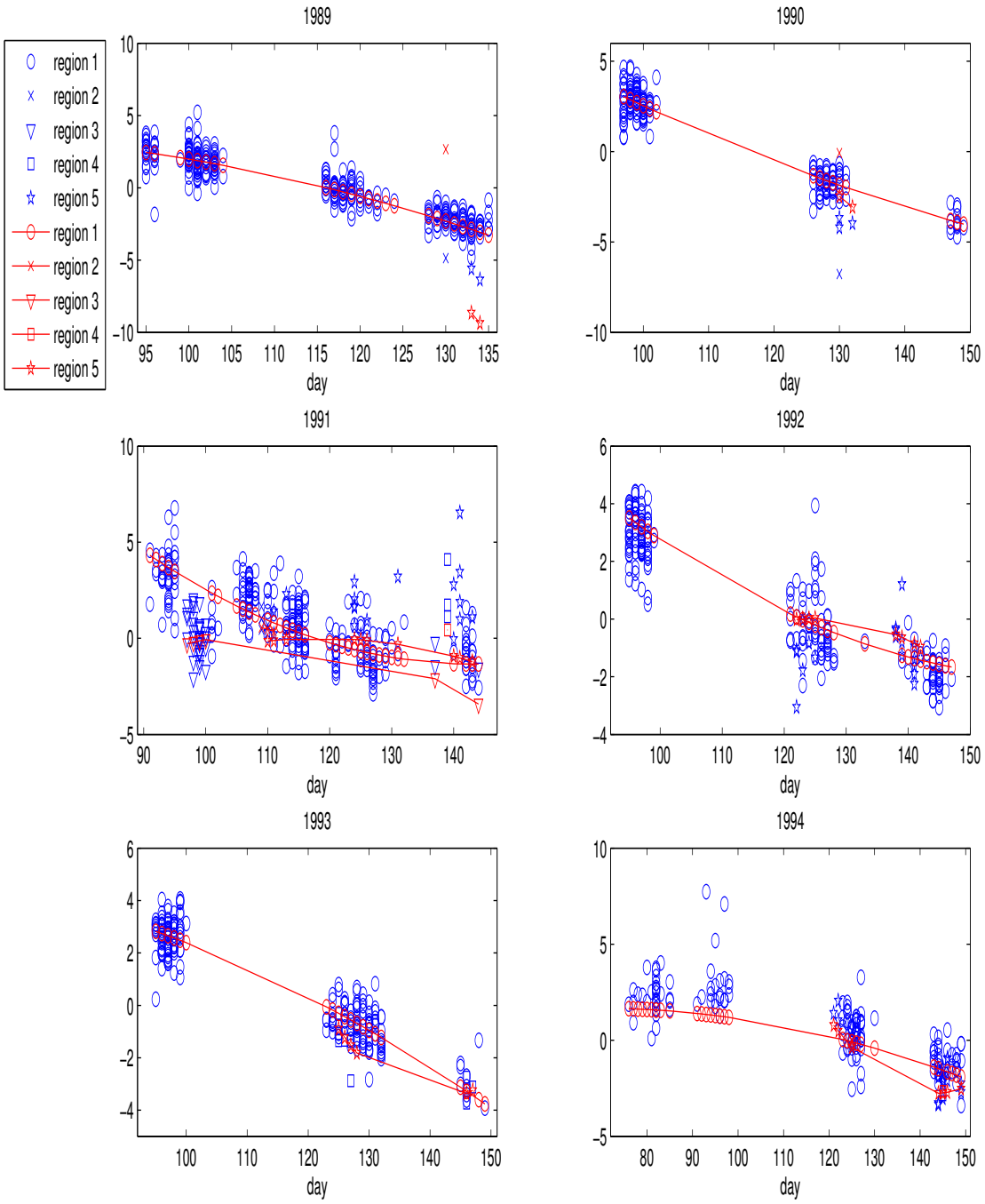

Figure 14: The temporal fit: $\log(M_i) - spat_i$, for all $M_i > 0$, $i = 1, ..., n$, are plotted as points with region coding (blue). $temp_i$, $i = 1, ..., n$, are also plotted as points with region coding, but connected with lines (red). $1989, ..., 1994$.

The MCMC algorithm

We here describe the details in the MCMC algorithm applied. The sampling is performed sequentially through the five blocks defined in Section 4. We describe sampling of each of these blocks below.
The vector of regression parameters $\boldsymbol{\beta}$ is of length n_β. $\boldsymbol{X}$ is the matrix of covariates of dimension $n_\beta \times n$. We define some matrices and vectors to ease the notation.

- 0_n is a vector of dimension n of zeros.
- 1_n is a vector of dimension n of ones.
- I_n is an $n \times n$ identity matrix.
- E_M is an $n \times n$ matrix with entries on the main diagonal equal to one if $M_i > 0$, zero elsewhere.

1. Sampling $\overrightarrow{\theta} = (\boldsymbol{\beta}, \overrightarrow{\mu}, \overrightarrow{\lambda})^T$

   In sampling $\overrightarrow{\theta}$ we take the approach of Rue and Held (2005, p. 135 - 142). We let $\pi[\overrightarrow{\theta}\,|\cdot]$ denote the full conditional distribution of $\overrightarrow{\theta}$. We denote the other conditional distributions in a similar way. We further use the notation $\doteq$ to mean equality up to an additive constant.

$$
\begin{aligned}
\log\, \pi[\overrightarrow{\theta}\,|\cdot] &= \log\, \pi[\boldsymbol{\beta}, \overrightarrow{\mu}, \overrightarrow{\lambda}\,|\cdot] \\
&\doteq log\{\pi[\boldsymbol{\beta}|\sigma_\beta^2]\pi[\overrightarrow{\mu}\,|\Sigma_\mu]\pi[\overrightarrow{\lambda}\,|\boldsymbol{\beta}, \overrightarrow{\mu}, \sigma_\epsilon^2]\pi[\overrightarrow{M}|\,\overrightarrow{\theta}, \cdot]\} \\
&\doteq -\tfrac{1}{2}\{\tfrac{1}{\sigma_\beta^2}\boldsymbol{\beta}^T\boldsymbol{\beta} + \overrightarrow{\mu}^T\Sigma_\mu^{-1}\overrightarrow{\mu} + \tfrac{1}{\sigma_\epsilon^2}(\overrightarrow{\lambda} - (X^T\boldsymbol{\beta} + \overrightarrow{\mu}))^T(\overrightarrow{\lambda} - (X^T\boldsymbol{\beta} + \overrightarrow{\mu}))\} \\
&\quad -\tfrac{1}{2}\tfrac{1}{\sigma_\lambda^2}(\overrightarrow{Z} - \overrightarrow{\lambda})^T E_M(\overrightarrow{Z} - \overrightarrow{\lambda}) \\
&\quad + \log\{\Pi_{i=1}^n(\tfrac{1}{1+\exp(\xi-\kappa\lambda_i)})^{I(M_i=1)} \times \Pi_{i=1}^n(\tfrac{\exp(\xi-\kappa\lambda_i)}{1+\exp(\xi-\kappa\lambda_i)})^{I(M_i=0)}\} \\
&\doteq -\tfrac{1}{2}\{\boldsymbol{\beta}^T(\tfrac{1}{\sigma_\beta^2}I_\beta + \tfrac{1}{\sigma_\epsilon^2}XX^T)\boldsymbol{\beta} + \overrightarrow{\mu}^T(\Sigma_\mu^{-1} + \tfrac{1}{\sigma_\epsilon^2}I_n)\overrightarrow{\mu} + \overrightarrow{\lambda}^T(\tfrac{1}{\sigma_\epsilon^2}I_n + \tfrac{1}{\sigma_\lambda^2}E_M)\overrightarrow{\lambda} \\
&\quad + \tfrac{2}{\sigma_\epsilon^2}\boldsymbol{\beta}^T X\overrightarrow{\mu} - \tfrac{2}{\sigma_\epsilon^2}\boldsymbol{\beta}^T X\overrightarrow{\lambda} - \tfrac{2}{\sigma_\epsilon^2}\overrightarrow{\mu}^T\overrightarrow{\lambda} - 2\tfrac{1}{\sigma_\lambda^2}\overrightarrow{\lambda}^T E_M\overrightarrow{Z}\} + \textstyle\sum_{i=1}^n g(\lambda_i),
\end{aligned}
$$

   where $g(\lambda_i) = (\xi - \kappa\lambda_i)I(M_i = 0) - \log(1 + \exp(\xi - \kappa\lambda_i))$.

   Define $\overrightarrow{\eta}_\theta = (0_{n_\beta}^T, 0_n^T, \tfrac{1}{\sigma_\lambda^2}(E_M\overrightarrow{Z})^T)^T$ and Σ_θ^{-1} by

$$\Sigma_\theta^{-1} = \begin{bmatrix} \frac{1}{\sigma_\beta^2} I_\beta + \frac{1}{\sigma_\epsilon^2} X X^T & | & \frac{1}{\sigma_\epsilon^2} X & | & -\frac{1}{\sigma_\epsilon^2} X \\ - & - & - & - & - \\ \frac{1}{\sigma_\epsilon^2} X^T & | & \Sigma_\mu^{-1} + \frac{1}{\sigma_\epsilon^2} I_n & | & -\frac{1}{\sigma_\epsilon^2} I_n \\ - & - & - & - & - \\ -\frac{1}{\sigma_\epsilon^2} X^T & | & -\frac{1}{\sigma_\epsilon^2} I_n & | & \frac{1}{\sigma_\epsilon^2} I_n + \frac{1}{\sigma_\lambda^2} E_M \end{bmatrix},$$

This gives
$\log\ \pi[\overrightarrow{\theta}\,|\cdot] \doteq -\frac{1}{2}(\overrightarrow{\theta} - \Sigma_\theta \overrightarrow{\eta}_\theta)^T \Sigma_\theta^{-1} (\overrightarrow{\theta} - \Sigma_\theta \overrightarrow{\eta}_\theta) + \sum_{i=1}^n g(\lambda(i))$.

$g(\lambda_i)$ contains non-quadratic terms of λ_i, resulting in a non-Gaussian distribution for $\pi[\overrightarrow{\theta}\,|\cdot]$, and we apply a Metropolis Hastings step. This requires a proposal distribution. We consider an approach where a Gaussian proposal is derived through a Taylor series expansion of g around $\lambda_{0,i}$. In particular, we approximate $g(\lambda_i) \approx a_i + b_i\lambda_i - \frac{1}{2}c_i\lambda_i^2$, where $b_i = g'(\lambda_{0,i}) - g''(\lambda_{0,i})\lambda_{0,i}$ and $c_i = -g''(\lambda_{0,i})$ (we do not need to specify a_i). We define $\overrightarrow{b}$ and $\overrightarrow{c}$ parallel to $\overrightarrow{\lambda}$.

Then $\pi[\overrightarrow{\theta}\,|\cdot]$ is approximated by $\widetilde{\pi}[\overrightarrow{\theta}\,|\cdot]$, where

$$\begin{aligned}\log \widetilde{\pi}[\overrightarrow{\theta}\,|\cdot] &\doteq -\tfrac{1}{2}(\overrightarrow{\theta} - \Sigma_\theta \overrightarrow{\eta}_\theta)^T \Sigma_\theta^{-1} (\overrightarrow{\theta} - \Sigma_\theta \overrightarrow{\eta}_\theta) + \textstyle\sum_{i=1}^n (a_i + b_i\lambda(i) - \frac{1}{2}c_i\lambda(i)^2) \\ &\doteq -\tfrac{1}{2}\overrightarrow{\theta}^T(\Sigma_\theta^{-1} + diag(\overrightarrow{c}_\theta))\overrightarrow{\theta} + (\overrightarrow{\eta}_\theta + \overrightarrow{b}_\theta)^T\overrightarrow{\theta},\end{aligned}$$

where $\overrightarrow{b}_\theta = (0_{n_\beta}^T, 0_n^T, \overrightarrow{b}^T)^T$ and $\overrightarrow{c}_\theta = (0_\beta^T, 0_n^T, \overrightarrow{c}^T)^T$.
Using that
$g'(\lambda_i) = -\kappa I(M_i = 0) + \kappa\frac{\exp(\xi-\kappa\lambda_i)}{1+\exp(\xi-\kappa\lambda_i)}$ and $g''(\lambda_i) = -\kappa^2\frac{\exp(\xi-\kappa\lambda_i)}{(1+\exp(\xi-\kappa\lambda_i))^2}$,
we get
$b_i = -\kappa I(M_i = 0) + \kappa\frac{\exp(\xi-\kappa\lambda_{0,i})}{1+\exp(\xi-\kappa\lambda_{0,i})} + \kappa^2\frac{\exp(\xi-\kappa\lambda_0(i))}{(1+\exp(\xi-\kappa\lambda_{0,i}))^2}\lambda_{0,i}$ and
$c_i = \kappa^2\frac{\exp(\xi-\kappa\lambda_{0,i})}{(1+\exp(\xi-\kappa\lambda_{0,i}))^2}$.

The approximated distribution for $[\overrightarrow{\theta}\,|\cdot]$ is given by

$$\widetilde{\pi}[\overrightarrow{\theta}\,|\cdot] \sim N((\Sigma_\theta^{-1} + diag(\overrightarrow{c_\theta}))^{-1}(\overrightarrow{\eta_\theta} + \overrightarrow{b_\theta}), (\Sigma_\theta^{-1} + diag(\overrightarrow{c_\theta}))^{-1}).$$

This distribution depends on $\overrightarrow{\lambda_0}$ since both $\overrightarrow{b}$ and $\overrightarrow{c}$ depend on $\overrightarrow{\lambda_0}$. This approximation is used as the proposal distribution in a Metropolis Hastings step. We have choosen $\overrightarrow{\lambda_0}$ to be the current $\overrightarrow{\lambda}$ value.

We constrain the μ_is so that $\sum_{i=1}^n \mu_i = 0$, in order to avoid possible identifiability problems. We therefore sample a proposal $\overrightarrow{\theta}_p$ from $\widetilde{\pi}[\overrightarrow{\theta}\,|\cdot]$ conditioned on $\sum_{i=1}^n \mu_i = 0$. We do this by applying a "trick" described in Rue and Held (2005, p. 39 - 40). Denote $A = (0_{n_\beta}^T, 1_n^T, 0_n^T)$. We want to sample from $\widetilde{\pi}[\overrightarrow{\theta}\,|A\overrightarrow{\theta} = 0]$. Denoting $E_{\widetilde{\pi}}(\overrightarrow{\theta}\,|\cdot) = \nu$ and $Cov_{\widetilde{\pi}}(\overrightarrow{\theta}\,|\cdot) = Q^{-1}$ we can compute

the expectation $E_{\widetilde{\pi}}(\overrightarrow{\theta}\,|A\overrightarrow{\theta}=0)=\nu-AQ^{-1}(AQ^{-1}A^T)^{-1}A\nu$ and covariance $Cov_{\widetilde{\pi}}(\overrightarrow{\theta}\,|A\overrightarrow{\theta}=0)=Q^{-1}-Q^{-1}A^T(AQ^{-1}A^T)AQ^{-1}$.
Hence, $E_{\widetilde{\pi}}(\overrightarrow{\theta}\,|A\overrightarrow{\theta}=0)$ is just a transformation of $E_{\widetilde{\pi}}(\overrightarrow{\theta}\,|\cdot)=\nu$.

2. Sampling σ_μ^2

   We let $\bar{\Sigma}_\mu^{-1}$ denote Σ_μ^{-1} without the factor$\frac{1}{\sigma_\mu^2}$, i.e. $\Sigma_\mu^{-1}=\frac{1}{\sigma_\mu^2}\bar{\Sigma}_\mu^{-1}$.

   $$\begin{aligned}\pi[\sigma_\mu^2|\cdot] &\propto \pi[\sigma_\mu^2|q_\mu,r_\mu]\pi[\overrightarrow{\mu}\,|\Sigma_\mu]\\ &\propto \tfrac{1}{\Gamma(q_\mu)}r_\mu^{-q_\mu}\tfrac{1}{(\sigma_\mu^2)^{q_\mu+1}}\exp\{-\tfrac{1}{r_\mu\sigma_\mu^2}\}\times\tfrac{1}{(\sigma_\mu^2)^{n/2}}\exp\{-\tfrac{1}{2\sigma_\mu^2}\overrightarrow{\mu}^T\bar{\Sigma}_\mu^{-1}\overrightarrow{\mu}\}\end{aligned}$$

   This gives: $\pi[\sigma_\mu^2|\cdot]\sim IG(\frac{n}{2}+q_\mu,[\frac{1}{r_\mu}+\frac{1}{2}\overrightarrow{\mu}^T\bar{\Sigma}_\mu^{-1}\overrightarrow{\mu}]^{-1})$.

3. Sampling σ_ϵ^2

   As for σ_μ^2 we get:
   $\pi[\sigma_\epsilon^2|\cdot]\sim IG(\frac{n}{2}+q_\epsilon,[\frac{1}{r_\epsilon}+\frac{1}{2}(\overrightarrow{\lambda}-(X^T\beta+\overrightarrow{\mu}))^T(\overrightarrow{\lambda}-(X^T\beta+\overrightarrow{\mu}))]^{-1})$.

4. Sampling σ_λ^2

   We have: $\pi[\sigma_\lambda^2|\cdot]\sim IG(\frac{n}{2}+q_\lambda,[\frac{1}{r_\lambda}+\frac{1}{2}(\overrightarrow{Z}-E_M\overrightarrow{\lambda})^T(\overrightarrow{Z}-E_M\overrightarrow{\lambda})]^{-1})$.

5. Sampling R_μ

   $$\pi[R_\mu|\cdot]\propto\pi[\overrightarrow{\mu}\,|\Sigma_\mu]\pi[R_\mu|a_{R_\mu},b_{R_\mu}]\propto|\Sigma_\mu|^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}\overrightarrow{\mu}^T\Sigma_\mu^{-1}\overrightarrow{\mu}\}\times\tfrac{1}{b_{R_\mu}-a_{R_\mu}}$$

   R_μ enters the equation above on the right hand side through Σ_μ, and the full conditional distribution for R_μ is a non-standard one. We employ a Metropolis Hastings step. We take the approach of Rue and Held (2005, p. 143-144) by letting the proposal R_μ^* be given by $R_\mu^*=fR_\mu$. R_μ is the present value and f a scaling factor having the density $\pi(f)\propto 1+\frac{1}{f}$ for $f\in[\frac{1}{F},F]$, and zero otherwise. F is a tuning parameter. This is a convenient approach, as the proposal distribution is symmetric, i.e. $\frac{\widetilde{\pi}(R_\mu^*|R_\mu)}{\widetilde{\pi}(R_\mu|R_\mu^*)}=1$.

6. Sampling ξ and κ

ξ and κ enters the model both through p and are sampled in a similar way. We therefore only present the sampling of ξ.

$$\pi[\xi|\cdot] \propto \pi[\xi|\sigma_\xi^2]\pi[\overrightarrow{p}|\xi,\kappa,\overrightarrow{\lambda}] \propto \exp(-\tfrac{1}{2}\tfrac{1}{\sigma_\xi^2}\xi^2) \times \Pi_{i=1}^n (\tfrac{\exp(\xi-\kappa\lambda(i))}{1+\exp(\xi-\kappa\lambda(i))})^{I(M(i)=0)}$$

This is a non-standard distribution and we apply a Metropolis Hastings step with proposal distribution $\widetilde{\pi}(\xi)$. $\widetilde{\pi}(\xi)$ is a Gaussian distribution with expectation given by the current value of ξ and constant variance. Since $[\xi|\cdot]$ can be written as $\exp(-\frac{1}{2}\frac{1}{\sigma_\xi^2}\xi^2 + \sum_{i=1}^n g(\xi,\kappa,\lambda(i)))$ we could employ the method of Held and Rue (2005) as done for the $\overrightarrow{\theta}$ block. This was tried out, but rejected as a possible approach since the proposal was practically never accepted. Probably the resulting proposal distribution from a second order Taylor expansion around $\sum_{i=1}^n g(\lambda_i)$ is just not a good one.

References

Agarwal, D.K., Gelfand, A.E., Citron-Pousty S. (2002) Zero-inflated models with application to spatial count data, *Environmental and ecological statistics*, 9(4), 341-355.

Allcroft D.J. and Glasbey C.A. (2002) Spatial disaggregation of rainfall using a latent Gaussian Markov random field. In *Statistical Modeling in Society: Proceedings of the 17th International Workshop on Statistical Modeling*, (eds M. Stasinopoulos and G. Touloumi), 85-93.

Allcroft, D.J. and Glasbey, C.A. (2003) A latent Gaussian Markov random field model for spatio-temporal rainfall disaggregation, *Applied Statistics*, 52, 487-498.

Brochers, D.L., Buckland, S.T., Priede, I.G. and Ahmadi, S. (1997) Improving the precision of the daily egg production method using generalized additive models, *Can. J. Fish. Aquat. Sci.*, 54, 2727-2742.

Ciannelli, L., Bailey, K., Chan, K.S. and Stenseth, N. C. (2006) Phenological and geographical patterns of Walley Pollock spawning in the Western Gulf of Alaska, *submitted*.

Dobbie, M. J., Welsh, A.H. (2001) Modelling Correlated Zero-inflated Count Data, *Aust, N.Z.J.Stat.*, 43(4), 431 − 444.

Fox, C.J., O'Brian, C.M., Dickey-Collas, M. and Nash, R.D.M. (2000) Patterns in the spawning of cod (*Gadus morhua L.*), sole (*Solea solea L.*) and plaice (*Pleuronectes platessa L.*) in the Irish Sea as determined by generalized additive modelling, *Fisheries Oceanography*, 9, 33-49.

Glasbey, C.A. and Nevison, I.M. (1998) Rainfall modelling using a latent Gaussian variable. In *Modelling Longitudinal and Spatially Correlated Data, Methods, Applications and Future Directions*, (eds T.G. Gregoire), Springer-Verlag , New York , 122,

233-242.
Lambert, D. (1992) Zero-inflated Poisson regression, with application to defects in manufacturing, *Technometrics*, 34, 1-14 .
Natvig, B. and Tvete, I.F. (2006) Bayesian modeling of earthquake data, *submitted*.
Pennington, M. (1983) Efficient Estimators of Abundance, for Fish and Plankton Surveys, *Biometrics*, 39, 1, 281-286.
Rue, H. and Held, L. (2005) In *Gaussian Markov Random Fields, Theory and Applications*, CRC Press/Chapman and Hall, 39-40, 135-143.