Brazilian Journal of Probability and Statistics
2014, Vol. 28, No. 2, 190–208
DOI: 10.1214/12-BJPS201

Modelling categorized levels of precipitation

Patrícia Lusié Velozo[a,b], Mariane B. Alves[a] and Alexandra M. Schmidt[a]

[a] *Universidade Federal do Rio de Janeiro*
[b] *Universidade Federal Fluminense*

Abstract. We propose a dynamic model to analyze polychotomous data subject to temporal variation. In particular, we propose to model categorized levels of rainfall across time. Our model assumes that the observed category is related to an underlying latent continuous variable, which is modelled according to a power transformation of a Gaussian latent process, centered on a predictor that assigns dynamic effects to observable covariates. The inference procedure is based on the Bayesian paradigm and makes use of Markov chain Monte Carlo methods. We analyze artificial sets of data and daily measurements of rainfall in Rio de Janeiro, Brazil. When compared to the fitting of the actual observed volume of rainfall, our categorized model seems to recover well the structure of the data.

1 Introduction

In different fields of science, such as atmospheric sciences, agriculture, and hydrology, understanding and forecasting levels of precipitation over a region, across time, is a key issue. Depending on the time scale, observed values of precipitation are either equal to 0 (dry period) or equal to a positive quantity. For this reason, it is important to have statistical models that account for this property of the data. There are in the literature different proposals to model levels of rainfall. Stid (1973) proposes a model which assumes that levels of precipitation are realizations from a normal distribution that has been truncated and transformed. Sansó and Guenni (1999a) propose a dynamic version of the model proposed by Stid (1973). More specifically, Sansó and Guenni (1999a) assume that levels of rainfall are a power transformation of a normal process, which, in turn, is centered on a dynamic linear predictor allowing covariates' effects to vary smoothly through time. Sansó and Guenni (1999b) extend the idea of the dynamic model to a spatio-temporal setting. De Oliveira (2004) proposes a model for rainfall fields that do not have continuous distributions, and possess a distinctive probabilistic structure that is not presented by standard random field models. His proposal is suitable for short to medium periods of time as it accounts for the zero inflation typically present in such rainfall data. Fernandes et al. (2009) pursue a different approach by assuming that observed rainfall is a realization from a mixture distribution between a variable with

Key words and phrases. Bayesian inference, cumulative link model, latent variable, ordinal data, probit model.

Received February 2011; accepted June 2012.

Bernoulli distribution, and another one assuming only positive values. They explore the exponential, gamma and lognormal distributions for the positive part of the model.

For some applications, the interest lies only in predicting if it will rain or not. In this case, one can propose models for precipitation occurrence by assuming, for example, a temporal logistic or probit regression. Alternatively, Hughes et al. (1999) propose a non-homogeneous hidden Markov model, relating precipitation occurrences to broad scale atmospheric circulation patterns.

Here we propose to consider that observed volumes of rainfall at each time t can be categorized into one of J categories. As pointed out by Fuentes et al. (2008), rain gauges are widely used to measure rainfall accumulation, but the information they provide is limited by their spatial and temporal resolution. Rainfall estimates are also obtained through remote sensing which provide information about rainfall at locations which do not have a ground monitor. We focus on situations in which the actual volume of rainfall for some time t at a particular location is unknown. However, it is known, through different sources of information (remote sense, physical model, etc.), in which range, e.g. dry, drizzle, rain, storm, the amount of rainfall at time t is.

The multinomial distribution is a natural choice to model polychotomous data. For ordinal responses, it is usual to model the cumulative distribution function, according to the so called cumulative link models, as seen in e.g., Agresti (1990) and Congdon (2005). The choice of a link function can be arbitrary or induced by data augmentation, which is a method frequently adopted to model categorical ordinal data. The idea is to assume that the categorical response is generated by an underlying, latent, continuous variable, supposed to be divided into intervals, each of which representing a category.

Albert and Chib (1993) develop exact Bayesian inference for polychotomous data by using data augmentation. The idea is to make use of an underlying normal regression structure on latent continuous data. On the other hand, Chen and Dey (2000) use scale mixture of multivariate normal link functions to model correlated ordinal response data.

On a pure spatial setting, De Oliveira (2000) proposes a model for binary random fields by clipping a Gaussian random field at a fixed level. Higgs and Hoeting (2010) extend the approach of De Oliveira (2000) to model ordinal, categorical spatial observations. Berret and Calder (2012) develop strategies to improve the inference of a Bayesian spatial probit regression model.

In the temporal context, Carlin and Polson (1992) assume that the categorical time series is a known function of an underlying continuous process which evolves according to a state-space model. Inference is performed under the Bayesian paradigm and they concentrate on the dichotomous case. Knorr-Held (1995) proposes a dynamic version of the cumulative probit model. In particular, a multivariate autoregressive structure is assumed for the regression coefficients and threshold parameters which define each of the categories. Cargnoni et al. (1997) discuss a

class of conditionally Gaussian dynamic models for non-normal, multivariate time series. They focus on multivariate time series of multinomial observations.

This paper is organized as follows. The next section proposes a model for temporal observations of categories of rainfall. Basically, we assume the latent approach of Albert and Chib (1993), but model the latent variable following Sansó and Guenni (1999a). Besides, we consider the bin boundaries that connect the latent variable with each of the J categories to be unknown. Therein we also discuss possible identifiability problems with the multinomial model. Then, in Section 3, we start by performing a simulation study to check if our proposed model is able to capture the true structure of the data when the truth is known. We provide an example with real data by analyzing observed temporal categories of rainfall in Rio de Janeiro, Brazil. As the actual volumes of rainfall are observed for this data set we also fit a model to the daily observed amount of rain and compare the predictions based on the categorized and continuous observations. Finally, Section 4 presents some concluding remarks and points to future avenues of research.

2 Proposed model

Let $Y_t = j$ be an ordinal categorical variable indicating that the response variable is in category j at time t. A possible way of modelling a categorized random variable Y_t is to consider that it has been generated from a continuous latent variable, Z_t, divided into intervals whose bin boundaries are unknown. The categorical variable is classified in category j if, and only if, the associated continuous variable falls within, say, λ_{j-1} and λ_j, that is

$$Y_t = j \quad \Longleftrightarrow \quad \lambda_{j-1} < Z_t \leq \lambda_j, \qquad j = 1, \ldots, J,$$

with $\lambda_0 = -\infty$, or the lowest value that Z_t can assume, and $\lambda_J = \infty$. Then one can model the cumulative probability that the response variable lies in category j, or below it, at time t as

$$\gamma_{tj} = \Pr(Y_t \leq j) = \Pr(Z_t \leq \lambda_j). \tag{2.1}$$

We propose to model categories of rainfall, treating the true volumes of rain Z_t, as a latent process. In particular, we follow Sansó and Guenni (1999a) to model this true process. Assume that Z_t is a transformation of a Gaussian latent variable ζ_t, given by:

$$\begin{aligned} Z_t &= \begin{cases} \zeta_t^{\alpha}, & \zeta_t > 0, \\ 0, & \zeta_t \leq 0. \end{cases} \\ \zeta_t &= \mathbf{F}_t^{'}\boldsymbol{\theta}_t + e_t, \qquad e_t \sim N(0, V_t), \\ \boldsymbol{\theta}_t &= \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim N_K(\mathbf{0}, \mathbf{W}_t), \end{aligned} \tag{2.2}$$

with $\alpha > 0$ and $\mathbf{F}_t$ being a vector of regressors, which may include trend and seasonal components, as well as other covariates at time t. The effects of the structural

components of $\mathbf{F}_t$ are described through $\boldsymbol{\theta}_t$, a vector with K coefficients, which may vary through time according to the stochastic dynamic structure described in the bottom line of (2.2). Note that current and past values of the state parameters $\boldsymbol{\theta}$ are related through a $K \times K$ evolution matrix $\mathbf{G}_t$. In the data analysis of Section 3.2 we explore models only with covariates, we assume $V_t = V$, and $\mathbf{G}_t$ is the identity matrix of dimension K, $\forall t$. The structure in (2.2) implies that Z_t is positive and zero inflated. As the underlying true process, Z_t, represents the volume of precipitation at time t, we fix $\lambda_0 = 0$.

It is worth noting that the inclusion of the Gaussian latent variable ζ_t implies that the link function that is implicitly assumed in the proposed formulation is a variation of a probit model, since:

$$\begin{aligned}\gamma_{tj} &= \Pr(Z_t \leq \lambda_j) = \Pr(\zeta_t \leq 0) + \Pr\big(0 < \zeta_t \leq \lambda_j^{1/\alpha}\big) \\ &= \Phi\bigg(\frac{\lambda_j^{1/\alpha} - \mathbf{F}_t'\boldsymbol{\theta}_t}{\sqrt{V}}\bigg).\end{aligned} \tag{2.3}$$

Hence, $\Phi^{-1}(\gamma_{tj}) = \rho_j - \mathbf{F}_t'\boldsymbol{\vartheta}_t$, with $\rho_j = \frac{\lambda_j^{1/\alpha}}{\sqrt{V}}$, $\boldsymbol{\vartheta}_t = \frac{\boldsymbol{\theta}_t}{\sqrt{V}}$ and $\Phi(\cdot)$ denoting the cumulative standard normal distribution.

2.1 Inference procedure

Let $\mathbf{y} = (y_1, \ldots, y_T)'$ denote a random sample from the categorical variable for T instants in time, and π_{tj} be the probability that the response variable lies in category j at time t, that is, $\pi_{tj} = \Pr(Y_t = j)$. From equation (2.1), follows that

$$\begin{aligned}\pi_{t1} &= \gamma_{t1}, \\ \pi_{tj} &= \gamma_{tj} - \gamma_{t,j-1}, \qquad j = 2, \ldots, J.\end{aligned} \tag{2.4}$$

Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{J-1})$ be the bin boundaries. Thus, the likelihood function is proportional to

$$\begin{aligned}&l(\mathbf{y}|\boldsymbol{\lambda}, \boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_T, V, \alpha) \\ &\quad\propto \prod_{t=1}^{T}\prod_{j=1}^{J} \pi_{tj}^{I(y_t=j)} \\ &\quad= \prod_{t=1}^{T}\big[\Phi(u_{t,1})\big]^{I(y_t=1)} \prod_{j=2}^{J}\big[\Phi(u_{t,j}) - \Phi(u_{t,j-1})\big]^{I(y_t=j)} \\ &\quad= \prod_{t=1}^{T}\Big\{I(y_t=1)\Phi(u_{t,1}) + \sum_{j=2}^{J} I(y_t=j)\big[\Phi(u_{t,j}) - \Phi(u_{t,j-1})\big]\Big\},\end{aligned} \tag{2.5}$$

with $I(A)$ denoting an indicator function, that is $I(A) = 1$ if A occurs, and 0 otherwise, and

$$u_{t,j} = \frac{\lambda_j^{1/\alpha} - \mathbf{F}_t' \boldsymbol{\theta}_t}{\sqrt{V}}, \qquad j = 1, \ldots, J; t = 1, \ldots, T.$$

Because of identifiability reasons Albert and Chib (1993) fix $V = 1$ and the first bin boundary at 0. De Oliveira (2000) and Higgs and Hoeting (2010) follow a similar approach and do not perform inference about the variance of the latent random variable. Indeed, examination of the likelihood function in equation (2.5) shows that there are identifiability issues with the intercept, the exponent, the variance and the bin boundaries. To understand this, first let $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{Kt})$; if the predictor contains an intercept ($\mathbf{F}_t' = (1 \;\; x_{2t} \;\; \cdots \;\; x_{Kt})$), the substitution of the parameters $\varphi = (\alpha, \boldsymbol{\lambda}, \boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_T, V)$ by $\varphi^* = (\alpha^*, \boldsymbol{\lambda}^*, \boldsymbol{\theta}_0^*, \ldots, \boldsymbol{\theta}_T^*, V^*)$, with $\alpha^* = k\alpha$, $\lambda_j^* = (a\lambda_j^{1/\alpha} - c)^{k\alpha}$, $j = 1, \ldots, J-1$, $\boldsymbol{\theta}_t^* = a\boldsymbol{\theta}_t - c\mathbf{e_1}$, $t = 1, \ldots, T$, $V^* = a^2 V$ for any $a, k > 0$, $c \in \mathbb{R}$ and K-dimensional vector $\mathbf{e_1} = (1, 0, \ldots, 0)$ implies that

$$\begin{aligned} u_{t,j}^* &= \frac{(\lambda_j^*)^{1/\alpha^*} - \mathbf{F}_t' \boldsymbol{\theta}_t^*}{\sqrt{V^*}} = \frac{(a\lambda_j^{1/\alpha} - c)^{k\alpha/(k\alpha)} - a\theta_{1t} + c - \sum_{k=2}^{K} a x_{kt}\theta_{kt}}{\sqrt{a^2 V}} \\ &= \frac{\lambda_j^{1/\alpha} - \mathbf{F}_t' \boldsymbol{\theta}_t}{\sqrt{V}} = u_{t,j}, \end{aligned}$$

impying that $l(\mathbf{y}|\varphi) = l(\mathbf{y}|\varphi^*)$. If the predictor does not have an intercept, it still follows that, for $c = 0$, $l(\mathbf{y}|\varphi) = l(\mathbf{y}|\varphi^*)$. Also, note that the likelihood function does not depend on the minimum value ($\lambda_0 = 0$) of the variable Z_t because the continuous variable ζ_t is defined in the real line. Then, the probability of the categorical variable falling within the first category is given by

$$\begin{aligned} \Pr(Y_t = 1) &= \Pr(\lambda_0 \le Z_t \le \lambda_1) = \Pr(Z_t = 0) + \Pr(0 < Z_t \le \lambda_1) \\ &= \Pr(\zeta_t \le 0) + \Pr\big(0 < \zeta_t \le \lambda_1^{1/\alpha}\big) = \Pr\big(\zeta_t \le \lambda_1^{1/\alpha}\big). \end{aligned}$$

Because of the reasons mentioned above we fix the parameters (α, V) in equation (2.2) at some reasonable values, and consider a predictor without intercept, and focus on the inference of the bin boundaries and the coefficients of the covariates.

For computational convenience we follow Albert and Chib (1993), and parameterize the likelihood in terms of the latent variables $\zeta_t, \ldots, \zeta_T$, that is

$$\begin{aligned} l(\mathbf{y}|\boldsymbol{\lambda}, \boldsymbol{\zeta}, \alpha) \propto \prod_{t=1}^{T} \Bigg[& I(y_t = 1) I\big(\zeta_t \le \lambda_1^{1/\alpha}\big) \\ & + \sum_{j=2}^{J} I(y_t = j) I\big(\lambda_{j-1}^{1/\alpha} < \zeta_t \le \lambda_j^{1/\alpha}\big) \Bigg], \end{aligned} \tag{2.6}$$

and hence the parameter vector to be estimated in the proposed model is $\boldsymbol{\psi} = (\boldsymbol{\lambda}, \boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_T, \boldsymbol{\zeta})$, as well as the evolution covariance matrices $\mathbf{W}_t$. In particular, we assume that $\mathbf{W}_t$ is a diagonal matrix, implying prior independence among the components of $\boldsymbol{\theta}_t$, $\forall t$. In order to specify $\mathbf{W}_t$ we make use of discount factors. The choice of such discounts reflects the rate of adaptation of $\boldsymbol{\theta}_t$ to new incoming data, that is, it implies a graduate decay on the information that observations previous to time t should bring to the estimation of $\boldsymbol{\theta}_t$. For details on the specification of discount factors and the relationship between such discounts and evolution errors' variances, see West and Harrison (1997, pp. 51, 193–202).

The prior specification for the components of the parametric vector $\boldsymbol{\psi}$ is as follows: for $\boldsymbol{\theta}_0$ we assign a multivariate normal distribution with mean vector $\mathbf{m}_0$ and covariance matrix $\mathbf{C}_0$; for $\boldsymbol{\lambda}$ we assign a joint prior distribution that can be factored in a product of conditionally truncated normal distributions, each with parameters m_{λ_j} and U_{λ_j}, defined in the interval (λ_{j-1}, ∞), for $j = 2, \ldots, J-1$ and defined in the interval $(0, \infty)$ for $j = 1$. We assume prior independence among the errors e_t such that, given $\boldsymbol{\theta}_t$ and V, $\zeta_1, \ldots, \zeta_T$ are conditionally independent, *a priori*. Therefore, the joint posterior distribution for the general model, conditional on α, V and the discount factors for $\mathbf{W}_t$, is proportional to

$$p(\boldsymbol{\psi}|y_1^T, \alpha, V, \mathbf{W}_t) \propto l(\mathbf{y}|\boldsymbol{\lambda}, \boldsymbol{\zeta}, \alpha)\, p(\boldsymbol{\theta}_0) \times \prod_{t=1}^{T} \left[p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathbf{W}_t)\, p(\zeta_t|\boldsymbol{\theta}_t, V)\right] \prod_{j=1}^{J-1} p(\lambda_j|\lambda_{j-1}).$$

As our continuous variable represents precipitation, the positive part of the distribution is typically skewed. The hydrological literature suggests the cubic root as a reasonable transformation to obtain normality. For this reason, in the examples in Section 3 we fix $\alpha = 3$. V is fixed at some reasonable value, we discuss this in more detail in Section 3.

The joint posterior distribution is analytically intractable and we resort to Markov chain Monte Carlo (MCMC) methods to obtain samples from the target distribution. In particular, we use a hybrid Gibbs algorithm (Geman and Geman (1984), Gelfand and Smith (1990)) with some Metropolis–Hastings steps (Metropolis et al. (1953), Hastings (1970)). The Appendix provides some details about the resultant full conditional posterior distributions and proposal distributions adopted in the MCMC sampling scheme.

2.2 Predictive inference

Let D_0 denote the set that summarizes all the information available to a forecaster at time $t = 0$. If the model is closed to external information, the available information at each time t is given by $D_t = \{D_{t-1}, y_t\}$. In most time series applications, one aims to predict future values Y_{T+h}, $h = 1, \ldots, H$, given the information available up to time T, D_T. Let $\mathbf{y}_f = (y_{T+1}, \ldots, y_{T+H})'$ be the future values at times

$T+1,\ldots,T+H$ and define $\boldsymbol{\psi}_f$ as the collection of parameters required for the distribution of $\mathbf{Y}_f$. Then the predictive distribution for $\mathbf{y}_f$, under model M, is given by:

$$\begin{aligned} p(\mathbf{y}_f|D_T,M) &= \int l(\mathbf{y}_f|\boldsymbol{\psi}_f,D_T,M)\,p(\boldsymbol{\psi}_f|D_T,M)\,d\boldsymbol{\psi}_f \\ &= \int l(\mathbf{y}_f|\boldsymbol{\psi}_f,M)\,p(\boldsymbol{\psi}_f|D_T,M)\,d\boldsymbol{\psi}_f \\ &= E_{\boldsymbol{\psi}_f|D_T,M}\big[l(\mathbf{y}_f|\boldsymbol{\psi}_f,M)\big], \end{aligned} \tag{2.7}$$

with $p(\boldsymbol{\psi}_f|D_T,M)$ obtained by updating $p(\boldsymbol{\psi}|D_T,M)$ through the evolution equation in the bottom line of equation (2.2) and $l(\mathbf{y}_f|\boldsymbol{\psi}_f,M)=\prod_{h=1}^{H} l(y_{T+h}|\boldsymbol{\psi}_{T+h},M)$. When looked at as a function of the model M, equation (2.7) gives the predictive likelihood for model M, which may be used as a criterion for model selection, see e.g. Alves et al. (2010).

Let $\boldsymbol{\psi}_m$ be the set of the parameters needed to describe the predictive likelihood of the model m and suppose that a Monte Carlo sample of size N of $p(\boldsymbol{\psi}_m|M=m,D_T)$ is available. Then the construction of a sample of $p(\boldsymbol{\psi}_m|M=m,D_T)$ follows directly and a Monte Carlo estimate for the predictive likelihood in (2.7) is given by

$$\begin{aligned} &\hat{E}_{\boldsymbol{\psi}_m|M=m,D_T}\big[l(\mathbf{y}_f|\boldsymbol{\psi}_m,M=m,D_T)\big] \\ &\quad=\frac{1}{N}\sum_{i=1}^{N}\prod_{h=1}^{H} l\big(y_{T+h}|\boldsymbol{\psi}_m^{(i)},M=m,D_T\big). \end{aligned} \tag{2.8}$$

3 Data analysis

In order to verify if the proposed inference procedure is able to recover the actual structure that generated the data, when that structure is known, artificial data sets were generated following equation (2.2). This simulation exercise is summarized in Section 3.1. Next, we fit our proposed model to a real data set, aiming at predicting categorized levels of precipitation. We also compare the performance of the prediction under the categorized formulation with a fitting to actual volumes of precipitation, which we call continuous formulation.

3.1 Artificial data

Based on equation (2.2), we generated $L=25$ samples, each of length $T=169$, with $J=4$ categories and used $K=2$ covariates, such that $\mathbf{F}_t'=(x_{1t},x_{2t})$, where x_1 and x_2 are the same covariates used in the analysis of the real data in Section 3.2. After fixing $\boldsymbol{\theta}_0=(3.0,3.5)$, $W=0.0001$, $V=0.1$, and $\alpha=3$, we generated the true values for θ_{1t}, θ_{2t}, and ζ_t. The true bin boundaries were fixed at $\lambda_1=0.5$,

Table 1 *Prior mean* (m_λ) *and variance* (U_λ) *of the associated normal distributions for the bin boundaries,* λ_1, λ_2, *and* λ_3, *for the simulation study*

Prior	m_{λ_1}	m_{λ_2}	m_{λ_3}	$U_\lambda = U_{\lambda_j}\ \forall j = 1, 2, 3$
I	0.5	7.5	15.0	10.0
II	1.0	7.0	13.0	10.0
III	1.0	7.0	13.0	50.0
IV	1.0	7.0	13.0	100.0
V	1.0	10.0	20.0	100.0

$\lambda_2 = 7.5$, $\lambda_3 = 15$. Once these values were established, we obtained the observed values y_t as follows: if $z_t \in [0.0, 0.5]$ then $y_t = 1$, else if $z_t \in (0.5, 7.5]$ then $y_t = 2$, else if $z_t \in (7.5, 15.0]$ then $y_t = 3$, else if $z_t \in (15.0, \infty)$ then $y_t = 4$.

For each of the $L = 25$ samples, we fitted the same model used to generate the data, and assigned the following prior distributions: for θ_{01} and θ_{02}, independent, zero mean normal distributions, each with variance 10. The variances of the evolution equation of the parameters of the covariates, $\mathbf{W} = \text{diag}(W_1, W_2)$, were estimated using discounting factors, and these were fixed at 0.98. The exponent α and the variance V were fixed at their true values.

We explored five different prior specifications for the bin boundaries λ_js, $j = 1, 2, 3$. All of them assumed prior normal distributions for λ_j, truncated on λ_{j-1}, and the mean and variance of these normals are shown in Table 1. Prior distribution I assumes the prior mean (m_{λ_j}) equal to the respective true values, and the prior variance is assumed reasonably big. The other priors consider m_{λ_j} with values different from the ones used to generate the data. Priors II, III and IV assume the same prior mean but the prior variance are fixed, respectively, at 10, 50, and 100. Finally, prior V assigns values for m_{λ_j} greater than the ones used to generate the data, and assume a large prior variance. This is to investigate the effect of the prior uncertainty on the posterior distribution of the bin boundaries.

For each sample, and prior specification, we let the MCMC run for 100,000 iterations, considered the first 10,000 as burn in, and stored every 90th iteration to avoid autocorrelation among the sampled values. Convergence of the chains was checked through trace plots.

Panels in columns 2 to 4 of Figure 1 compare the 95% posterior credible intervals (solid lines) of the bin boundaries λ_1, λ_2, and λ_3 (rows) with their respective 95% prior credible interval (dashed lines), for each of the $L = 25$ samples, under priors I, III, and V (columns). Priors II and IV provided very similar results, and are not shown here. The first column of this figure shows how many observations, among the 169, fell in each of the categories. As expected, there are fewer observations in category 3. For all priors and most of the samples, the 95% posterior credible intervals contain the respective values used to generate the data. As the

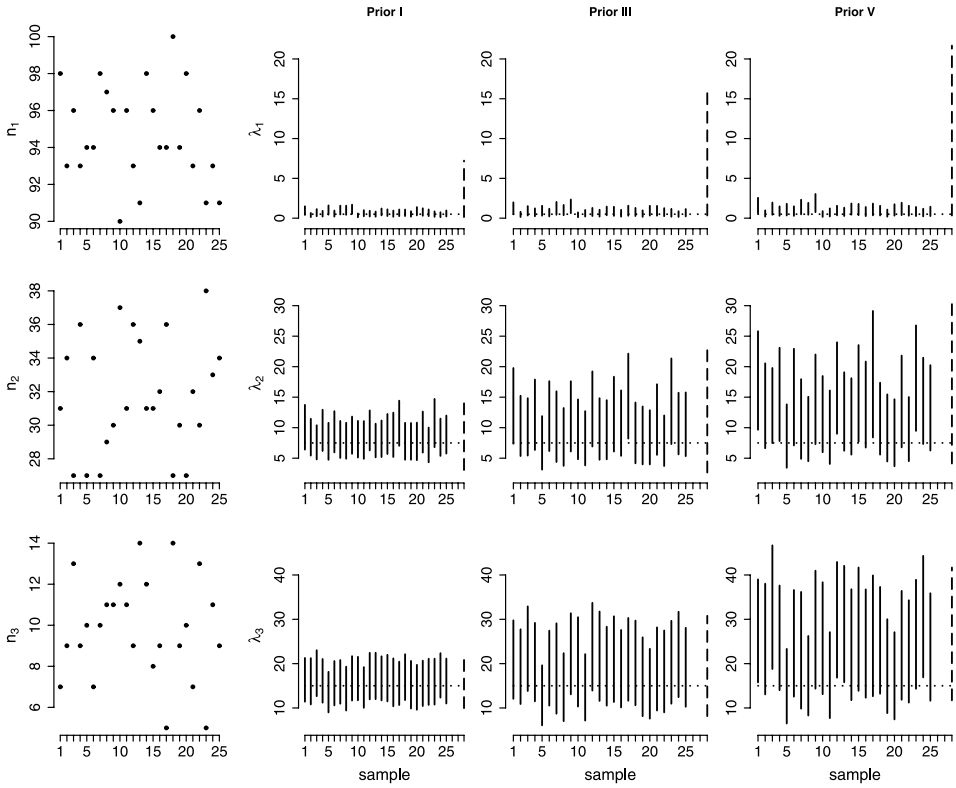

Figure 1 *Panels in the first column show the number of observations that fell in each of the* $J = 3$ *categories,* n_1, n_2, *and* n_3 (*rows*), *for each of the* $L = 25$ *samples. Panels in columns* 2 *to* 4 *show the* 95% *posterior credible interval for* λ_1, λ_2, *and* λ_3 (*rows*), *under prior specifications I, III, and V* (*columns*). *Dashed lines in panels of columns* 2 *to* 4 *represent the* 95% prior *credible interval of the respective* λ_j. *And the horizontal dotted line in each panel is the respective true value of* λ_j.

last category has the smallest number of observations it results in posterior distributions which are very similar to their respective prior distributions (last row of Figure 1). As there is a lot of information about category 1 in the likelihood, the posterior distribution for λ_1 is very concentrated when compared to the respective prior, and this is independent of the magnitude of the prior variance (see first row of Figure 1). Clearly, the posterior distributions of λ_2 and λ_3 seem to be sensitive to the variance of the prior distribution. We believe this is related to the number of observations that fell within each category.

3.2 Analyzing daily categories of precipitation in the city of Rio de Janeiro

In this subsection, two approaches are compared, both aiming at modelling precipitation data. In the first approach we assumed that the available information is on categorized precipitation occurrence and that the actual amount of rain is un-

known, being treated as a latent process, as described in Section 2. In the second approach, we follow Sansó and Guenni (1999a), and model volumes of rainfall, then we compare the resultant categorical predictions under both approaches.

The volumes of rainfall were obtained from the National Institute of Meteorology—INMET, Brazil, and comprise daily ground observations on volumes of rainfall in the city of Rio de Janeiro. We have also available daily records on average wind speed, average humidity and average temperature, but preliminary analysis showed no significant effect of average wind speed on rainfall. The sampling period ranges from September 22nd, 2005 to March 19th, 2006, comprising 179 observations. We held out the last $H = 10$ observations for model selection and predictive purposes, and kept $T = 169$ observations for the inference procedure.

Following the suggestion of a reviewer, we classified the observed volumes of rainfall onto $J = 4$ categories, based on the observed quartiles of the time series. These $J = 4$ categories might be interpreted as 1: "dry period," 2: "drizzle," 3: "rainy day," and 4: "storm." As we have daily observations, there are many observed values equal to 0. Thus, in order to compute the observed quartiles, we removed the zeros from the dataset and the resulting series was used to categorize the rain volumes. The values used as the bin boundaries to define the categorized time series were 0.7, 4.2 and 12.7.

The parameters of the prior distributions for $\boldsymbol{\theta}_0$ were: $\mathbf{m}_0 = 0$ and $\mathbf{C}_0 = 10$. The proposed model assumes fixed values for α, V, and discount factors were adopted for the diagonal elements of $\mathbf{W}_t$ (see equation (2.2)). We fixed $\alpha = 3$ as it is suggested in the hydrological literature that a cubic root transformation of precipitation approximates well the normal distribution. The values of the discount factor δ and the value of V were fixed after some preliminary model fitting. This preliminary analysis was performed by assigning the prior mean of the bin boundaries equal to the observed quartiles, with prior variance $U_\lambda = 5$, and different values of δ and V were explored. Then we computed the predictive likelihood for each fitted model based on the $H = 10$ observations held out from the inference procedure. We also compared the fitted number of observations in each category with the observed ones. After analyzing these preliminary fits, we decided for fixing $V = 2$ and $\delta = 0.99$.

Once the values of α, V and δ were fixed, we fitted 5 different models. They differ in terms of the prior information about the bin boundaries. A grid of values was specified for the parameters of the truncated normal prior distributions for λ_1, $\lambda_2|\lambda_1$ and $\lambda_3|\lambda_2$, according to Table 2. The different prior specifications are based on two different sets of prior information. Priors 1, 2 and 3 are centered on the observed quartiles adopted as the bin boundaries in the construction of the categorized time series. These priors differ just in terms of the magnitude of their variances. The other priors are based on private communication of the authors with Professor Pedro Dias and Dr. America Espinoza, experts on rainfall modelling from IAG-USP. They suggested that for the time of the year we are considering in this example (spring–summer), they believe that reasonable values

Table 2 *Parameters of the prior distribution specifications for the bin boundaries* λ_j, $j = 1, 2, 3$ *for the rainfall dataset*

Prior specification	μ_{λ_1}	μ_{λ_2}	μ_{λ_3}	$U_\lambda = U_{\lambda_j}\ \forall j$
1	0.7	4.2	12.7	5.0
2	0.7	4.2	12.7	10.0
3	0.7	4.2	12.7	50.0
4	0.25	4.0	11.25	10.0
5	0.25	4.0	11.25	50.0

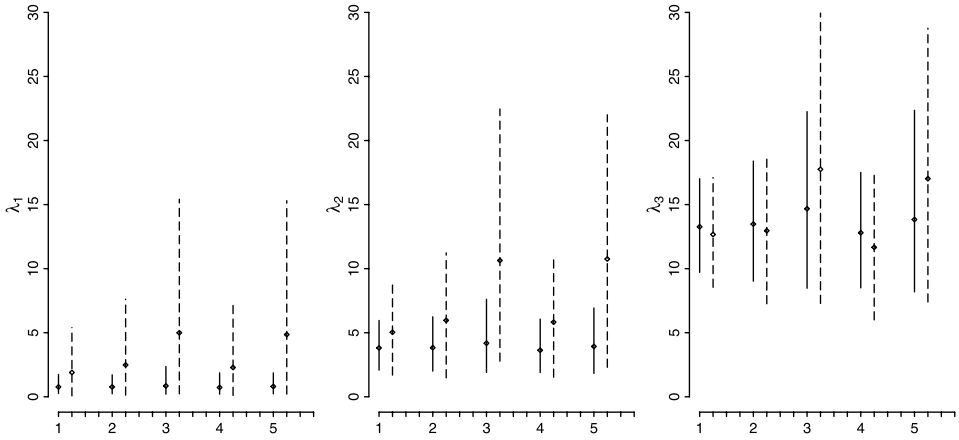

Figure 2 *Median* (*solid circles*) *and* 95% *credible intervals* (*lines*) *of the marginal prior* (*dashed lines*), *and respective posterior distributions* (*solid lines*) *of the bin boundaries under the five different prior specifications for* λ_1, λ_2, *and* λ_3 (*columns*).

for the prior mean of the bin boundaries are 0.25, 4.0, and 11.25. Therefore, we explore two other priors, 4 and 5, whose mean specification is as suggested by them, and prior 5 has variance greater than prior 4. For each fitted model, we let the MCMC algorithm run for 100,000 iterations. We considered the first 10,000 as burn in, and stored every 90th iteration to avoid autocorrelation among the sampled values. Convergence of the chains was checked through the traces of multiple chains starting from very different values.

The marginal posterior distribution of the bin boundaries λ_1 and λ_2, seem not to be sensitive to the prior distribution. Also, there seems to be a clear gain of information when we compare the marginal prior distribution of λ_1 and λ_2 with their respective marginal posterior distributions. On the other hand, for most of the 5 prior distributions considered for λ_3, the posterior distribution does not differ much from their respective priors. See Figure 2 for details.

Table 3 shows the values of the predictive likelihood based on the last $H = 10$ observations held out from the inference procedure. Jeffreys (1961) suggests a cri-

Table 3 *Predictive likelihoods based on the* $H = 10$ *observations held out from the inference procedure, for each of the* 5 *different prior specifications for the bin boundaries. All numbers must be multiplied by* 10^{-4}

	Prior 1	Prior 2	Prior 3	Prior 4	Prior 5
Predictive likelihood ($\times 10^{-4}$)	7.0	6.3	5.6	6.2	5.4

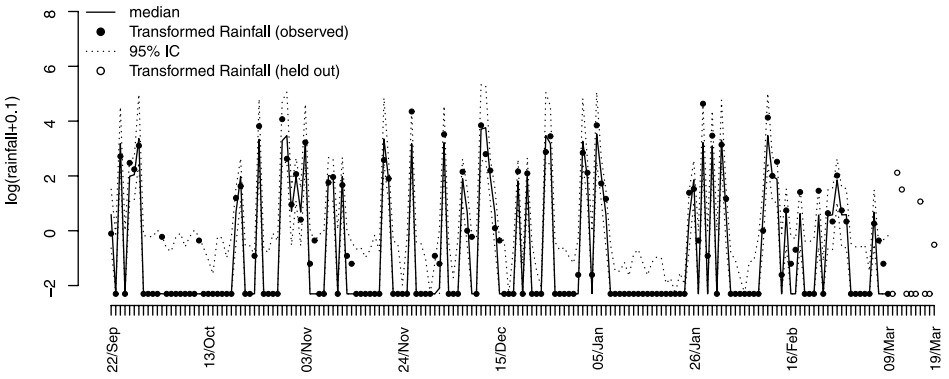

Figure 3 *Observed time series of precipitation* (*solid circles*)*—in the log scale—together with the summary of the posterior distribution of* $\log(Z_t + 0.1)$. *The solid line is the posterior median of* $\log(Z_t + 0.1)$ *and the dotted lines represent the limits of the* 95% *posterior credible interval. The hollow circles represent the observations held out for prediction purposes.*

teria for model comparison based on the magnitude of Bayes factors, according to which a factor lying on the interval $[1, 3.2)$ shows weak evidence of one hypothesis against another. Comparing the first four prior specifications against the fifth one and following Jeffreys' reasoning, we conclude that the different adopted prior distributions do not seem to interfere much with temporal prediction. In the remaining of this section, unless otherwise stated, the posterior and predictive results associated to the categorized formulation refer to the model fitted under the fourth prior specification in Table 2, which resulted in the best performance between the models that were built based on experts' knowledge.

In our proposed model, Z_t is a latent random variable. As we have the actual observations of precipitation, we can compare the posterior distribution of Z_t with the actual observations of rainfall, which is shown in Figure 3. Clearly, the posterior medians of the latent variables Z_t recover the structure of the volumes of precipitation quite well, with all observations falling within the 95% posterior credible intervals for Z_t.

A comparison with an analysis using the actual observed volumes of rainfall. We now compare the results obtained under the proposed categorized formulation with one that assumes that volumes—and not just categories—of rainfall are observed.

The latter is modeled based on the proposal of Sansó and Guenni (1999a). Our aim is to evaluate how well the categorized approach compares to the continuous one, in terms of estimation of model parameters and prediction of future values. Basically, following equation (2.2), we write down a likelihood function for the daily volumes of precipitation, in this case Z_t denotes the actual observed amount of rain. Thus it is only necessary to estimate the variance V of the latent variable $\boldsymbol{\zeta}$, the regression coefficients $\boldsymbol{\theta}$, the exponent α and the latent variable during the dry periods (when it rains and the value of the exponent is known, the latent variable is deterministic according to equation (2.2)). In order to make a fair comparison with the categorical model, we also fix here $V = 2$ and $\alpha = 3$.

Figure 4 shows the summary of the posterior distribution of the regression coefficients, which are positive for humidity and negative for temperature, under both approaches. The estimated coefficients exhibit quite similar temporal trajectories, regardless of the adopted approach. However, the ranges of the 95% posterior credible intervals are slightly wider under the categorical likelihood.

We now compare the categorized and continuous formulations through their respective posterior predictive distributions for times $T + h$, $h = 1, 2, \ldots, 10$. That is, for each formulation we compare the posterior distribution of the following probability $\Pr(Y_{T+h} = j|\boldsymbol{\psi}_f, D_T)$, which provides the posterior probability that

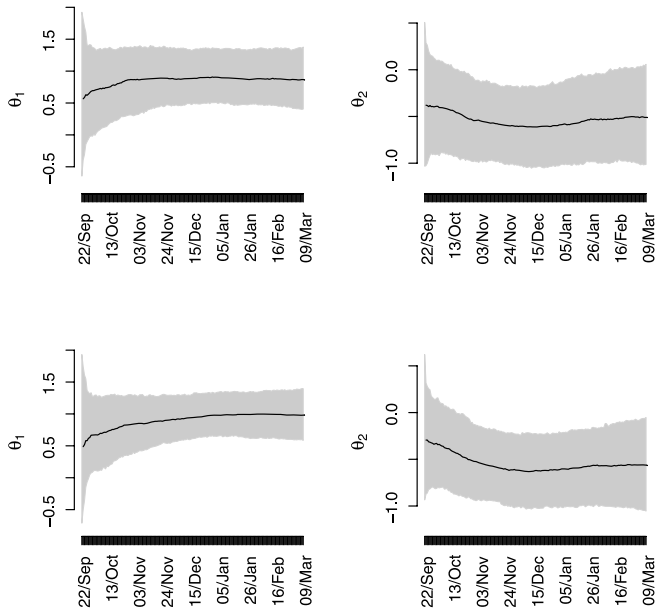

Figure 4 *Posterior summary of the evolution of the regression coefficients for humidity* (θ_1) *and temperature* (θ_2), *under the categorized model* (*first row*) *and the continuous formulation* (*second row*). *Solid lines represent the posterior median and the shaded areas represent the* 95% *posterior credible intervals*.

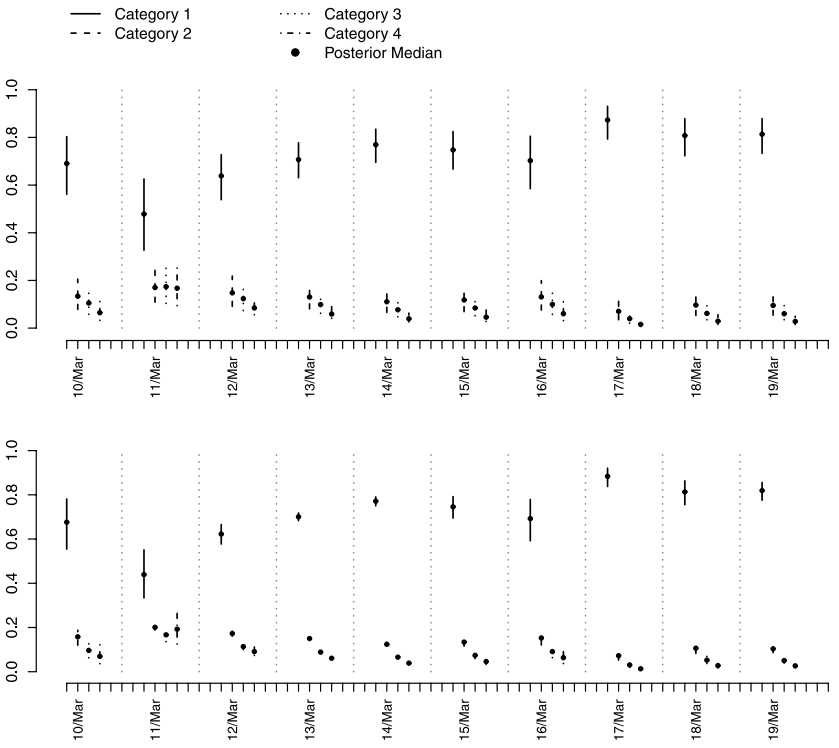

Figure 5 *Summary of the posterior distribution of* $\Pr(Y_{T+h} = j|\boldsymbol{\psi}_f, D_T)$, $j = 1, 2, 3$, $h = 1, \ldots, 10$, *for the last* 10 *observations, held out from the inference procedure under the categorized* (*top panel*) *and the continuous* (*bottom panel*) *formulations. The solid circle is the posterior median and the solid line is the* 95% *posterior credible interval. The observations for times* $T + 1, \ldots, T + 10$ *were respectively* $\mathbf{y}_f = (1, 3, 3, 1, 1, 1, 2, 1, 1, 1)$.

the observation at time $T + h$ falls within category j. Again, aiming at performing a fair comparison between the two approaches, we fix the bin boundaries for both models at the posterior medians obtained under the categorized formulation and, for the other parameters, we use the posterior samples obtained under each approach. More specifically: given an estimate of $\mathbf{W}_T$ and a posterior sample of $\boldsymbol{\theta}_T$, it is possible to obtain a sample of $\boldsymbol{\theta}_{T+h}$, for each $h = 1, \ldots, 10$, following the evolution equation presented in the bottom line of (2.2). Therefore, given the samples of $\boldsymbol{\theta}_{T+h}$, as well as the posterior samples of $\lambda_1, \ldots, \lambda_J$ and conditional on $\mathbf{F}_{T+h}$, the evaluation of $\Pr(Y_{T+h} = j|\boldsymbol{\psi}_f, D_T)$ is straightforward, using equations (2.3) and (2.4). Figure 5 compares the predictive posterior distribution, $\Pr(Y_{T+h} = j|\boldsymbol{\psi}_f, D_T)$, $j = 1, 2, 3, 4$, $h = 1, \ldots, 10$, for the last 10 observations, held out from the inference procedure under both approaches. The real values of those observations are $\mathbf{y}_f = (1, 3, 3, 1, 1, 1, 2, 1, 1, 1)$. As may be seen, the approach based on categorized observations provides equivalent results to those obtained under the formulation based on the observed volumes of rain.

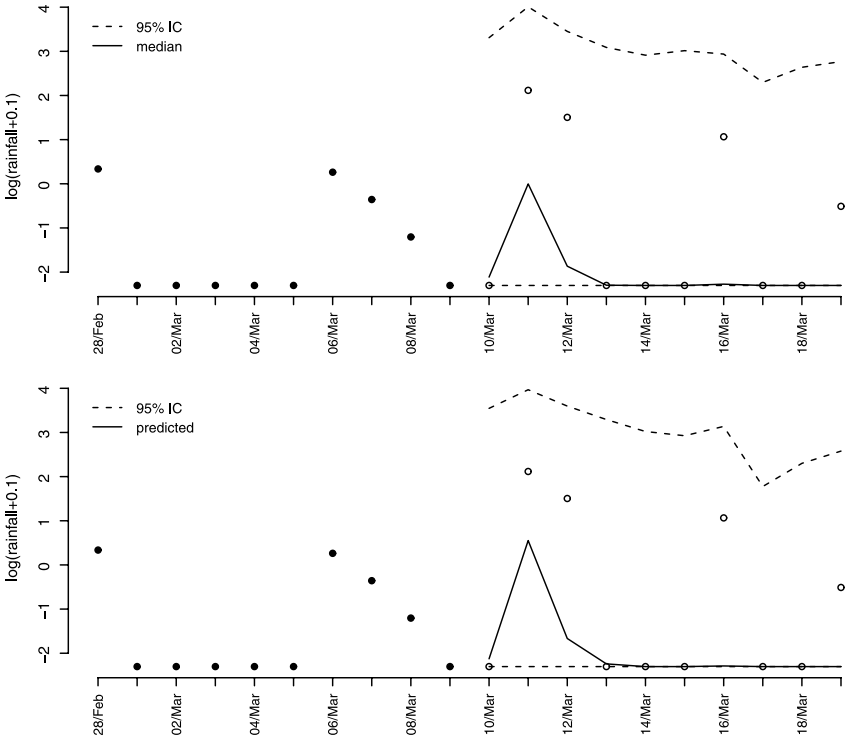

Figure 6 *Summary of the posterior predictive distributions median* (*solid line*) *and limits of the* 95% *posterior credible intervals* (*dashed lines*), *under the categorized* (*top panel*) *and the continuous* (*bottom panel*) *formulations*, *for the volumes of rainfall* (*in the log scale*). *The solid circles represent the last* 10 *observations which were used in the inference procedure and the hollow circle is the observed volume held out for prediction.*

Figure 6 shows the summary of the posterior predictive distribution for the volumes of rainfall, Z_t, transformed to the log scale, referring to the last ten observations held out from the inference procedure under both approaches. Clearly, the credible intervals under both approaches contain the observed volumes of rainfall and the categorized approach provides predictive estimates of the latent volumes Z_{T+h} which are equivalent to the predictions based on observed volumes Z_t, $t = 1, \ldots, T$.

4 Discussion

We proposed a model for polycothomous data that vary across time. More specifically, we concentrated on the problem of modelling observed categories of rainfall. We extended the work of Albert and Chib (1993) assuming that the underlying continuous variable follows the model proposed by Sansó and Guenni (1999a). This extension imposes a temporal structure on the probit link function.

In Section 2, we showed that we must impose some restrictions to the proposed model in order to be able to obtain estimates of the parameters of interest. Analysis of artificial data suggest that the posterior distribution of the bin boundaries (λ_i's) are sensitive to their prior distributions depending on the number of observations present in each category. The analysis of daily measurements of rainfall in Rio de Janeiro suggests that the categorized approach is able to recover reasonably well the underlying continuous process, when compared to the model that makes use of the actual observed volumes of rainfall (Section 3.2).

Following the suggestion of Dias and Espinosa (Private Communication, 2008), we assumed the bin boundaries fixed across time because we had only Spring/Summer observations. If a longer time series, covering different seasons of the year is investigated, we suggest to change the prior distribution of the λs accordingly. In this case, the MCMC described in the appendix has to be adapted, since different bin boundaries should be used for different instants in time.

Although the modelling of the volumes of rainfall tend to be more flexible, the proposed categorical model might be used as the top layer of a hierarchical model which accounts for different sources of information on rainfall, e.g., ground-based measurements, remote sense, physical models, among others. Combining the information from these different sources is challenging and is a current subject of research.

Appendix: Full conditional posterior distributions

In what follows, the full conditional posterior distributions based on the likelihood function in (2.6), which makes use of the latent variables ζ, are described. Let $\boldsymbol{\psi} = (\boldsymbol{\lambda}, \boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_T, \boldsymbol{\zeta})$ and $\boldsymbol{\psi}_{-\beta}$ be the vector $\boldsymbol{\psi}$, except for a component β. We assume V, α and $\mathbf{W}$ known.

Full conditional distribution of the bin boundaries $\boldsymbol{\lambda_1, \ldots, \lambda_{J-1}}$

The full conditional posterior distribution of $\boldsymbol{\lambda}$ is given by

$$
\begin{aligned}
p(\boldsymbol{\lambda}|\boldsymbol{\psi}_{-\boldsymbol{\lambda}}, y_1^T) \propto \prod_{j=1}^{J-1} \Bigg\{ & \exp\Bigg\{ -\frac{1}{2} \sum_{j=1}^{J-1} \Bigg(\frac{\lambda_j - m_{\lambda_j}}{\sqrt{U_{\lambda_j}}} \Bigg)^2 \Bigg\} \\
& \times \Bigg[1 - \Phi\Bigg(\frac{\lambda_{j-1} - m_{\lambda_j}}{\sqrt{U_{\lambda_j}}} \Bigg) \Bigg]^{-1} \Bigg\} \\
& \times \prod_{j=1}^{J-1} \big[1\big(\max\big\{ \max\{Z_t : y_t = j\}, \lambda_{j-1} \big\} < \lambda_j \\
& < \min\big\{ \min\{Z_t : y_t = j+1\}, \lambda_{j+1} \big\} \big) \big].
\end{aligned}
$$

This distribution is analytically intractable, hence we use Metropolis–Hastings steps to obtain samples from it. A product of truncated normal distributions, each one centered on the current value of each cut point, is adopted as proposal density for this step, so that $q(\boldsymbol{\lambda}^p|\boldsymbol{\lambda}^c) = q_1(\lambda_1^p|\boldsymbol{\lambda}^c)\prod_{j=2}^{J-1} q_j(\lambda_j^p|\lambda_{j-1}^p, \boldsymbol{\lambda}^c)$, with

$$q_j(\lambda_j^p|\lambda_{j-1}^p, \boldsymbol{\lambda}^c) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\left(\frac{\lambda_j^p - \lambda_j^c}{\sqrt{\sigma^2}}\right)^2\right\} \Bigg/ \left(\Phi\left(\frac{\min\{Z_t : y_t = j+1\} - \lambda_j^c}{\sqrt{\sigma^2}}\right) - \Phi\left(\frac{\max\{\max\{Z_t : y_t = j\}, \lambda_{j-1}^p\} - \lambda_j^c}{\sqrt{\sigma^2}}\right)\right).$$

The supports of the densities $q_1, \ldots, q_{J-1}$ are given by:

$$\max\{\max\{Z_t : y_t = 1\}, 0\} < \lambda_1^p < \min\{Z_t : y_t = 2\},$$
$$\max\{\max\{Z_t : y_t = j\}, \lambda_{j-1}^p\} < \lambda_j^p < \min\{Z_t : y_t = j+1\}, \qquad j = 2, \ldots, J-1$$

and σ^2 is tuned to provide reasonable acceptance rates.

Full conditional distribution of the regression coefficients $\boldsymbol{\theta}$

As the elements of $\boldsymbol{\theta}$, θ_t, evolve smoothly through time, we make use of the Forward Filtering and Backward Sampling algorithm (FFBS) (Frühwirth-Schnatter (1994) and Carter and Kohn (1994)) to obtain samples from the posterior full conditional of $\boldsymbol{\theta}$.

As $\zeta_t = \mathbf{F}_t'\boldsymbol{\theta}_t + e_t$, $e_t \sim N(0, V)$ and $\boldsymbol{\theta}_t \sim N(\boldsymbol{\theta}_{t-1}, \mathbf{W})$, $t = 1, \ldots, T$, $\boldsymbol{\theta}_0 \sim N(\mathbf{m}_0, \mathbf{C}_0)$, then the full conditional posterior distribution of the states $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T)$ is proportional to $p(\boldsymbol{\theta}|\boldsymbol{\psi}_{-\boldsymbol{\theta}}, \mathbf{W}, Y_1^T) \propto \prod_{t=1}^{T}\{p(\zeta_t|\boldsymbol{\theta}_t, V)\}p(\boldsymbol{\theta}|\mathbf{W})$. The variance $\mathbf{W}$ is obtained through the specification of discount factors.

Full conditional distribution of the latent variable $\boldsymbol{\zeta_t}$

The full conditional posterior distribution of ζ_t, $t = 1, \ldots, T$, is given by $p(\zeta_t|\boldsymbol{\psi}_{-\zeta_t}, y_1^T) \propto l(y_1^T|\boldsymbol{\psi})p(\zeta_t|\boldsymbol{\theta}_t, V)$. Therefore, $p(\zeta_t|\boldsymbol{\psi}_{-\zeta_t}, \mathbf{W}, y_t = j) \propto N(\mathbf{F}_t'\boldsymbol{\theta}_t, V)[I(\zeta_t \leq \lambda_1^{1/\alpha})I(y_{t1} = 1) + \sum_{j=2}^{J} I(\lambda_{j-1}^{1/\alpha} < \zeta_t \leq \lambda_j^{1/\alpha})I(y_{tj} = 1)]$, which follows a truncated normal distribution and is easy to sample from.

Acknowledgments

Most of this work was developed while P. L. Velozo was a M.Sc. student at IM-UFRJ. Velozo is grateful to CAPES for the financial support during her M.Sc. studies. Schmidt was partially supported by CNPq and FAPERJ. The authors are

grateful to Pedro L. S. Dias (IAG-USP/LNCC) and America M. Espinosa (IAG-USP) for fruitful discussions about rainfall modelling, and also to an anonymous reviewer whose suggestions greatly improved the presentation of the paper.

References

Agresti, A. (1990). *Categorical Data Analysis*. *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR1044993

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679. MR1224394

Alves, M. B., Gamerman, D. and Ferreira, M. A. R. (2010). Transfer functions in dynamic generalized linear models. *Statistical Modelling* **10**, 3–40. MR2758054

Berret, C. and Calder, C. A. (2012). Data augmentation strategies for the bayesian spatial probit regression model. *Computational Statistics & Data Analysis* **56**, 478–490.

Cargnoni, C., Müller, P. and West, M. (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *Journal of the American Statistical Association* **92**, 640–647.

Carlin, B. P. and Polson, N. G. (1992). Monte Carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and Smith, A. F. M., eds.) 577–586. Oxford, UK: Oxford Univ. Press.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–553. MR1311096

Chen, M.-H. and Dey, D. K. (2000). A unified Bayesian approach for analysing correlated ordinal response data. *Brazilian Journal of Probability and Statistics* **14**, 87–111. MR1838453

Congdon, P. (2005). *Bayesian Models for Categorical Data*. *Wiley Series in Probability and Statistics*. Chichester: Wiley. MR2191351

De Oliveira, V. (2000). Bayesian prediction of clipped Gaussian random fields. *Computational Statistics & Data Analysis* **34**, 299–314.

De Oliveira, V. (2004). A simple model for spatial rainfall fields. *Stochastic Environmental Research and Risk Assessment* **18**, 131–140.

Fernandes, M. V. M., Schmidt, A. M. and Migon, H. S. (2009). Modelling zero-inflated spatio-temporal processes. *Statistical Modelling* **9**, 3–25. MR2750828

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis* **15**, 183–202. MR1263889

Fuentes, M., Reich, B. J. and Lee, G. (2008). Spatial-temporal mesoscale modelling of rainfall intensity using gage and radar data. *The Annals of Applied Statistics* **4**, 1148–1169. MR2655653

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409. MR1141740

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Higgs, M. D. and Hoeting, J. A. (2010). A clipped latent variable model for spatially correlated ordered categorical data. *Computational Statistiscs & Data Analysis* **54**, 1999–2011. MR2640303

Hughes, J. P., Guttorp, P. and Charles, S. P. (1999). A non-homogeneous hidden Markov model for precipitation ocurrence. *Applied Statistics* **48**, 15–30.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford, UK: Oxford Univ. Press. MR0187257

Knorr-Held, L. (1995). Dynamic cumulative probit models for ordinal panel-data; a Bayesian analysis by Gibbs sampling. Technical Report 386, Ludwig-Maximilians-Universität, Munich, Germany.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092.

Sansó, B. and Guenni, L. (1999a). A stochastic model for tropical rainfall at a single location. *Journal of Hydrology* **214**, 64–73.

Sansó, B. and Guenni, L. (1999b). Venezuelan rainfall data analysed by using a Bayesian space-time model. *Applied Statistics* **48**, 345–362.

Stid, C. K. (1973). Estimating the precipitation climate. *Water Resources Research* **9**, 1235–2141.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR1482232

P. L. Velozo
Depto. de Estatística
Instituto de Matemática e Estatística
Universidade Federal Fluminense
24020-140 Niterói, RJ
Brazil
E-mail: patricia@dme.ufrj.br

M. B. Alves
A. M. Schmidt
Depto. de Métodos Estatísticos
Instituto de Matemática
Universidade Federal do Rio de Janeiro
Caixa Postal 68530
21945-970 Rio de Janeiro, RJ
Brazil
E-mail: mariane@im.ufrj.br
alex@im.ufrj.br