Venezuelan rainfall data analysed by using a Bayesian space–time model

Bruno Sansó and Lelys Guenni

Universidad Simón Bolívar, Caracas, Venezuela

[Received February 1996. Final revision August 1998]

Summary. We consider a set of data from 80 stations in the Venezuelan state of Guárico consisting of accumulated monthly rainfall in a time span of 16 years. The problem of modelling rainfall accumulated over fixed periods of time and recorded at meteorological stations at different sites is studied by using a model based on the assumption that the data follow a truncated and transformed multivariate normal distribution. The spatial correlation is modelled by using an exponentially decreasing correlation function and an interpolating surface for the means. Missing data and dry periods are handled within a Markov chain Monte Carlo framework using latent variables. We estimate the amount of rainfall as well as the probability of a dry period by using the predictive density of the data. We considered a model based on a full second-degree polynomial over the spatial co-ordinates as well as the first two Fourier harmonics to describe the variability during the year. Predictive inferences on the data show very realistic results, capturing the typical rainfall variability in time and space for that region. Important extensions of the model are also discussed.

Keywords: Bayesian estimation; Markov chain Monte Carlo method; Rainfall modelling; Space–time models; Truncated normal model

1. Introduction

We present an analysis of a set of data from the state of Guárico, a central region in Venezuela. The data consist of accumulated monthly rainfall and were collected during 16 years, from January 1968, at 80 different stations irregularly scattered in an area of roughly 250 × 300 km. The data can be obtained from

http://www.blackwellpublishers.co.uk/rss/

A schematic plot showing the locations of the 80 stations is seen in Fig. 1. The longitude and latitude are known for each of the stations, as well as the elevation in metres with respect to sea-level. The y-axis follows, roughly, a south–north direction; the Apure and Orinoco rivers correspond, roughly, to the x-axis. There are some hills in the northern part of the region, whereas the area becomes quite flat as one approaches the rivers. Three groups of seven stations each have been marked in Fig. 1 and the data corresponding to one station from each group are shown in Fig. 2. The data are recorded in millimetres of water collected; no mention is made of ‘traces’ of rain, as is customary in some sets of rainfall data.

As can be seen from Fig. 2 the data have a considerable number of missing values (up to 25% for some stations) and a marked seasonal pattern due to the dry–wet seasons cycle. Adding to the complexity of the data is the fact that, to account for their spatial behaviour, a model should consider all 80 stations jointly, implying the use of high dimensional distributions.
The approach presented in this study provides a mechanism to make predictions of rainfall data in space and time. This is particularly important in a region where the sparsity of data in space and the limited length of historical data become a limitation for many applications in hydrology, ecology and agriculture. At a monthly level, the amount of rainfall and the probability of a dry period are required quantities for many applications.

From Fig. 1 it is obvious that there is a much higher density of stations in the northern part of the region; this is because the population of the southern part is very small. Although most stations have an altitude of less than 300 m above sea-level, some, in the northern part of the region, are above 600 m. The fact that there can be a difference of several hundred metres in elevation between stations does not seem to be a relevant feature to include in a model for rainfall, as can be seen in Fig. 3, where the data corresponding to March, June and September are plotted against the elevation. The reason for this lack of dependence is that rainfall-generating mechanisms in this region are mainly produced by the convective activity of the intertropical convergence zone rather than the orographic effects.

The small dots in Fig. 4 correspond to the observed rainfall at stations 1 and 2 (other features in Fig. 4 will be explained in Section 4.1). It is clear that the variability in the data depends strongly on the month when they are collected. It is also evident that this is a very challenging set of data, since observations that could be considered as outliers are quite often present regardless of the month and location.

In Fig. 5 we present the median of the observed rainfall at each station corresponding to May and July as a function of $y$ (again we shall discuss Fig. 5 later in Section 4.1). We observe a very noisy behaviour reflecting high peaks of rain during certain years.

Parameter estimation of stochastic rainfall models has been a topic of intense research in the last 10 years. The estimation procedures are intrinsically linked to the nature of the
Fig. 2. Rainfall collected at three stations: (a) station 1; (b) station 2; (c) station 3
Fig. 3. Rainfall as a function of elevation from sea-level: (a) March; (b) June; (c) September

Fig. 4. Observed rainfall (•), mean estimated rainfall (●) and estimated probability of a dry month (△) as a function of the month: (a) station 1; (b) station 2
rainfall model itself and the timescale used to represent the process. There are models which
describe the rainfall process in continuous time and models describing the probabilistic
characteristics of rainfall accumulated on a given time period, say daily or monthly totals.
Different reviews of the available rainfall models have been presented: see for example
Continuous time models for a single site with parameters related to the underlying physical
rainfall process are particularly important for the analysis of data at shorter timescales, e.g.
hourly. Some of these models are described in Rodrı́guez-Iturbe et al. (1987, 1988) and
Waymire and Gupta (1981). Some of the features represented in these models are rain cells
and storms. Cowpertwait (1994) has provided a generalization of these models for different
 types of rain cell (e.g. convective and non-convective). Owing to difficulties in using max-
imum likelihood estimation in a closed form, the most widely used method for the parameter
estimation of such models is the method of moments. In this case, model parameters are
estimated from the properties of rainfall accumulated during a given time interval and
parameter estimates can vary greatly depending on the properties used in the fitting pro-
cedure. Recently Chandler (1995) has proposed a method by which maximum likelihood
estimation is possible by using a Fourier transform of the data rather than the data them-

The rainfall models for a single site is the spatial extension of these
models for multiple sites, to try to incorporate the intersite dependence but preserving
the marginal properties at each site. A more ambitious task is the modelling of rainfall con-
tinuously in time and space and original work on these types of model based on point process

\[\text{Bayesian Space–Time Model}\]
theory was presented by LeCam (1961) and further developed by Waymire et al. (1984) and Cox and Isham (1994). More recently Mellor (1996) has developed the modified turning bands model which reproduces some of the physical features of rainfall fields in space as rainbands, cluster potential regions and rain cells. Parameter estimation for continuous time–space models of rainfall is not trivial and validation of the rainfall characteristics represented by these models requires high resolution data in time and space.

When only data from a ground-based rainfall network are available and accumulated rainfall amounts for a particular time period are recorded, discrete time rainfall models are appealing. Discrete time models with a reduced number of parameters are a very useful tool in applications such as long-term prediction and planning, climate change studies and coarse timescales descriptions of hydrological processes. One of the difficulties of modelling rainfall is that it is a fairly skewed variable which has a point mass at 0. Smith (1994) proposed to break the problem into two parts: first analyse the pattern of wet and dry periods and then model the rainfall given that it is positive.

A simple model which has proved to be adequate for descriptions of rainfall at different timescales is the truncated normal model (Stidd, 1973; Hutchinson, 1995). Let \( z \) be the observed rainfall at a certain site and time interval; then suppose that

\[
  z = \begin{cases} 
    w^\beta & w > 0, \\
    0 & w \leq 0 
  \end{cases}
\]

where \( w \) follows a normal distribution with a given mean and variance.

Bardossy and Plate (1992) extended model (1) to a multisite framework with parameters conditioned on large scale atmospheric circulation patterns. Spatial dependence is modelled through the covariance structure of the multivariate normal process. More recently Glasbey and Nevison (1997) have used a truncated normal model with a different family of transformations.

In the traditional approach, the uncertainty in the estimates of the parameters of the probability distribution of the rainfall is not incorporated in the final predictive inference. Within a Bayesian approach, uncertainties in the parameter estimates are taken into account and expressed through the posterior distribution of them. This is clearly a key issue when the model is being used for prediction. Smith and Robinson (1997) have also used a Bayesian approach to fit a rainfall model which simulates rainfall occurrences and rainfall amounts separately.

We use Bayesian methodology to fit a model to rainfall data collected at different sites during repeated periods of time. The model is based on a transformed truncated multivariate normal representation of the data that uses model (1) as a base-line, where the mean of the multivariate normal distribution depends on both the time of the year and location and its covariance matrix depends on the distance between sites. On the basis of the characteristics of the data we shall consider a model which allows us to fit a seasonal pattern, a spatial correlation, the use of covariates, different variances for each month and different values of the transforming parameter allowing for a different tail behaviour at each month. This model is discussed in Section 2.

When the number of parameters in the model is large and there is no conjugacy, a Markov chain Monte Carlo (MCMC) method can be used to explore the joint posterior distribution of the parameters as well as the predictive distribution (see, for example, Gelfand and Smith (1990) and Tanner (1993)). In the case of a model for rainfall MCMC methodology provides a flexible framework allowing for the modelling of dry periods and missing values as censored
data that can be handled as latent variables. In Section 3 we present an MCMC method for
drawing samples of the posterior distribution of the parameters and producing predictive
inference.
In Section 4 we describe the results of the analysis of the data. We start by fitting a model
based on a second-order polynomial on the co-ordinates of each station to capture the spatial
variability and a function of the six Fourier harmonics to capture the seasonality. This model
is criticized by using posterior samples from the MCMC run and a second fit is attempted
with a reduced number of parameters. The conclusions, presented in the last section, high-
light the fact that the model provides a very realistic predictive inference, since it captures the
typical behaviour of the rainfall in the area. Nevertheless additional improvements can be
made to the model at the expense of a significant computational cost.

2. A space–time model
Let \(z_{it}\) be the observed rainfall at site \(i\) and time \(t\), measured in months, where \(i = 1, \ldots, k\)
and \(t = 1, \ldots, n\). In our case \(n = 192\) and \(k = 80\). Suppose that the data have a natural
periodicity of length \(p\). Since we think of annual periodicity, \(p = 12\) for our data. Using
model (1) we can model \(z_{it}\) as

\[
\begin{align*}
  z_{it} = \begin{cases} 
    w_{it} & \text{if } w_{it} > 0, \\
    0 & \text{if } w_{it} \leq 0
  \end{cases}
\end{align*}
\]

where \(\beta_i > 0\) and \(w_{it}\) is a normal random variable. Letting \(w_t = (w_{1t}, \ldots, w_{kt})'\) then \(w_t \sim N_k(\mu_t, \sigma_t^2 \Sigma)\), i.e. we assume that to obtain the vector of observed rainfall at time \(t\) a multi-
variate normal variate is truncated and transformed. Note that the transformation is assumed
to vary over time but remains constant over all sites, as considered in Bardossy and Plate

Writing the likelihood when the rainfall observations follow model (2) is a cumbersome
task for large \(k\). Nevertheless it is a simple exercise to write it for the (unobserved) \(w_t\); this
fact will be used in the next section to obtain an MCMC run to explore the joint posterior
distribution of the parameters.

Let us now consider some structure for \(\mu_t\) and \(\Sigma\). Assume that \(\mu_t = \mu_0 + \mu_{1t}\) where
\(\mu_0 = X_0 b_0\) describes the spatial structure of the mean, \(X_0\) is a \(k \times r\) matrix and \(b_0\) an \(r\)-
dimensional vector, so \(r\) covariates are being considered in the model. These will include some
functions of the spatial location of each site. The serial structure of the mean will be expressed
in terms of Fourier harmonics, so

\[
\mu_{1t} = 1_k \sum_{s=1}^{q} \{ c_s \cos(\alpha st) + d_s \sin(\alpha st) \}
\]

where \(p = 2q, q = 6, \alpha = 2\pi/p, d_q = 0\) and \(1_k\) is a \(k\)-dimensional vector of 1s. We shall also
suppose that \(\sigma_i^2 = \sigma_j^2\) and \(\beta_i = \beta_j\) whenever \(t = l\) modulo \(p\); this is also true for \(\mu_t\). Letting
\(b = (b'_0, c_1, d_1, \ldots, c_q)'\) we can then write \(\mu_t = X_t b\) for the appropriate matrices \(X_t\). Given \(b,\)
\(w_t\) are serially uncorrelated.

We shall assume that \(\Sigma_{ij} = \exp(-\lambda d_{ij})\) for some \(\lambda > 0\) and \(d_{ij}\) the distance between sites \(i\)
and \(j\), after projection of the locations of the sites on a horizontal plane. This is a common
function in hydrology applications for the correlation function of an isotropic random field
which guarantees that \(\Sigma_{ij}\) is symmetric and positive definite. A wider choice of functions can
be considered without altering the structure of the model. A more general approach would be
to consider a completely unknown covariance matrix; this will increase enormously the number of parameters in the model, possibly producing overparameterization, a problem that can be tackled by considering a proper prior distribution for the covariance matrix. The usual approach is to consider a normal–Wishart prior to preserve conjugacy. For a clear illustration of this approach applied to spatial interpolation problems see Brown et al. (1995).

To simplify the notation, let \( \hat{\beta}_1, \ldots, \hat{\beta}_p, \sigma^2, \theta = (\lambda, \beta, \sigma^2, b) \) and \( z \) be the vector of all the observations. To complete the model we need to specify a prior distribution for the parameters. We use the non-informative prior

\[
\pi(\sigma^2, \beta, \lambda, b) \propto \frac{1}{\lambda} \prod_{i} \frac{1}{\beta_i \sigma_i^2},
\]

which assumes independence and local uniformity on \( b \) and \( \log(\lambda), \log(\beta_i) \) and \( \log(\sigma_i^2) \).

### 3. A Markov chain Monte Carlo method

To obtain approximate samples from the joint posterior distribution of the parameters we run a Gibbs sampler as proposed in Gelfand and Smith (1990). For a clear explanation of the algorithm see Casella and George (1992). We need the full conditionals of the parameters without having to write explicitly the joint density of the observations. We use the data augmentation approach proposed in Smith and Roberts (1993) for models with censored data. We can consider that the variables \( w_{it} \) have been censored and then transformed. Let \( v_{it} \leq 0 \) be a latent variable corresponding to the (censored) negative \( w_{it} \); then, letting \( v \) be the vector of all \( v_{it} \), \( \pi(\theta | z, v) \) is the joint posterior for the parameters that would be obtained if there had been no censoring. This posterior is the product of the joint density of \( z \) and \( v \), which is the product of multivariate normals with some of the components transformed, multiplied by the prior for \( \theta \).

Hydrological data are usually collected over long time periods and frequently there are a number of missing data; the latent variable approach can be used to handle these. Let \( u_{it} \) be the latent variable that corresponds to a missing \( w_{it} \), and let \( u \) be the vector of such variables; then \( u \) plays the same role as \( v \) in the MCMC sampling, only that there is no constraint on its components. The augmented data vector can be written as \((z, v, u)\); more explicitly, \( w_{it} \) can be redefined as

\[
w_{it} = \begin{cases} u_{it} & \text{if } z_{it} \text{ is missing}, \\ \frac{1}{\beta_i} v_{it} \quad & \text{if } z_{it} > 0, \\ v_{it} & \text{if } z_{it} = 0. \end{cases}
\]

Letting \( J_l = \{ j: 1 \leq j \leq n, j = l \mod p \} \) for all \( l = 1, \ldots, p \), and \( m \) the number of elements in \( J_l \), in our case is 16, we note that the transformation in equations (2) introduces a Jacobian \( \frac{1}{\beta_l} / \beta_l \) into the likelihood; thus we have that the joint posterior density is given by

\[
\pi(\lambda, \beta, \sigma^2, b, v, u | z) \propto \frac{\left| \Sigma(\lambda) \right|^{-n/2}}{\lambda} \prod_{l=1}^p \left[ \frac{1}{\beta_l} \left( \frac{1}{\sigma_l^2} \right)^{(mk+2)/2} \prod_{j \in J_l, z_{ij} > 0} \frac{1}{\beta_l} \right] \\
\times \exp \left\{ -\frac{1}{2\sigma_l^2} (w_j - X_j b)' \Sigma^{-1} (w_j - X_j b) \right\}.
\]
We can now write the full conditionals for the parameters as well as the latent variables \( v \) and \( u \). Starting with \( \sigma^2 \), we have that

\[
\pi(\sigma_i^2|b, \beta, \lambda, z, v, u) \propto \left( \frac{1}{\sigma_i^2} \right)^{(mk+2)/2} \exp \left\{ - \frac{1}{2\sigma_i^2} \sum_{j \in J_i} (w_j - \mu_i)^\top \Sigma^{-1} (w_j - \mu_i) \right\}
\]

for all \( l = 1, \ldots, p \). This corresponds to an inverse gamma density.

The full conditional of \( b \), \( \pi(b|\sigma^2, \beta, \lambda, z, v, u) \), is proportional to

\[
\exp \left\{ \frac{1}{2} (\hat{b} - b)^\top X^\top BD (\sigma_i^{-2} \Sigma^{-1}) X (\hat{b} - b) \right\}
\]

where BD denotes a block diagonal matrix,

\[
X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}
\]

and

\[
\hat{b} = \left( \sum_{i=1}^p \sigma_i^{-2} X_i^\top \Sigma^{-1} X_i \right)^{-1} \left( \sum_{i=1}^p \sigma_i^{-2} X_i^\top \Sigma^{-1} \tilde{v}_i \right)
\]

where the average is taken over all \( j \in J \). The former shows that the full conditional for \( b \) is a multivariate normal distribution with mean \( \hat{b} \) and covariance matrix \( (\Sigma \sigma_i^{-2} X_i^\top \Sigma^{-1} X_i)^{-1} \).

The full conditional for \( \beta_i \), \( \pi(\beta_i|b, \lambda, \sigma^2, z, u, v) \), \( l = 1, \ldots, p \), is proportional to

\[
\frac{1}{\beta_i} \left( \prod_{z_i > 0} \frac{z_i^{1/\beta_i-1}}{\beta_i} \right) \exp \left\{ - \frac{1}{2\sigma_i^2} \sum_{j \in J_i} (w_j - \mu_i)^\top \Sigma^{-1} (w_j - \mu_i) \right\};
\]

since there is no obvious family from which to generate a sample of this density, a Metropolis step will be used, following Müller (1991). This proceeds as follows: in the previous formula transform to the log-scale and obtain the density \( \pi(\delta_i|b, \lambda, \sigma^2, z, u, v) \) where \( \delta_i = \log(\beta_i) \); then, given the current value \( \delta_i^0 \), generate a candidate value \( \delta_i^\ast \) from a given distribution \( \pi(\delta_i) \), which is called the jumping distribution. Accept the new sample with probability

\[
\min \left\{ \frac{\pi(\delta_i^\ast|b, \lambda, \sigma^2, z, u, v)}{\pi(\delta_i^0|b, \lambda, \sigma^2, z, u, v)}, 1 \right\}.
\]

We have chosen a uniform \( (\delta_i^0 - a, \delta_i^0 + a) \) distribution as the jumping distribution, the parameter \( a \) having been set to achieve an acceptance rate higher than 30%. A clear account of the Metropolis–Hastings algorithm can be read in Chib and Greenberg (1995).

A similar step will be taken to draw a sample of \( \lambda \), since the full conditional of \( \lambda \), \( \pi(\lambda|\sigma^2, \beta, b, v, u, z) \), is proportional to

\[
\frac{1}{\lambda} |\Sigma(\lambda)|^{-n/2} \exp \left\{ - \frac{1}{2\sigma_i^2} \sum_{j \in J_i} (w_j - \mu_i)^\top \Sigma(\lambda)^{-1} (w_j - \mu_i) \right\}.
\]

We must now consider the problem of drawing samples from the latent variables \( u \). To simplify the notation suppose that at time \( t \) there are \( k_t \) missing values and that they correspond to the first \( k_t \) components of \( w_t \), so that the partition \( (u_{1t}, w_{1t}) \) is obtained. Consider the corresponding partitions.
\[ \Sigma^{-1} = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \]

and \( \mu = (\mu_1, \mu_2) \); then, following Johnson and Kotz (1972), the full conditional of \( u \) is a \( k \)-variate normal distribution with mean \( \mu_1 - (w_{12} - \mu_2) / L_{21} L_{11}^{-1} \) and covariance matrix \( \sigma^2 \Sigma^{-1} \).

The full conditional of \( v \) can be obtained in a similar way, only that a truncated multivariate normal distribution of, say, dimension \( k \), mean \( \mu \) and covariance matrix \( L \) can be achieved by use of the following algorithm.

**Step 1:** use a Cholesky decomposition (see for example Golub and Van Loan (1983)) to obtain \( L \), \( K \), lower triangular.

**Step 2:** generate a multivariate standard normal vector \( z_{1, \ldots, p} \) that satisfies the constraints \( z_1 < -\mu_1 / K_{11} \) and \( z_i < -(\mu_i + \sum_{j=1}^{i-1} K_{ij} z_j) / K_{ii} \) for \( i = 1, \ldots, k \).

**Step 3:** compute \( K z + \mu \); this will have the desired distribution.

When the components of \( \mu \) are positive and large, the probability of sampling a negative variate is very small; an efficient way to generate from the tails of a univariate normal distribution is then needed. One possibility is to use a rejection algorithm based on the fact that the tails of a normal distribution behave approximately as an exponential distribution.

### 3.1. Predictive inference

The results of fitting a model by using the ideas discussed in Section 2 are very naturally presented in terms of the predicted rainfall on a certain site at a given time. The predictive distribution of, say, \( z_t^* \), a new vector of observations at time \( t \), given the observed rainfall \( z \), can be obtained from the samples of the distribution of \( \theta \) given \( z \) that are provided by the MCMC method. We have that

\[ \pi(z_t^* | z) = \int \pi(z_t^* | \theta) \pi(\theta | z) \, d\theta \approx \frac{1}{N} \sum_{i=1}^{N} \pi(z_t^* | \theta^{(i)}) \tag{4} \]

where the approximation is obtained by using the law of large numbers, \( \theta^{(i)} \) is the \( i \)th sample from the posterior distribution of \( \theta \) and \( N \) denotes the total number of such samples.

Formula (4) shows that the joint predictive distribution of the rainfall can be approximated by a mixture of truncated transformed multivariate normal distributions. This is also true for points that are not at the original sites, provided that \( \pi(z_t^* | \theta) \) is assumed as the conditional density, so equation (4) can be used for interpolation.

### 4. Fitting the model

We used the methodology presented in the previous two sections to fit a model with a full second-order polynomial, \( b_1 + b_2 x + b_3 y + b_4 x^2 + b_5 y^2 + b_6 xy \), where \( x \) and \( y \) are respectively the east–west and north–south planar co-ordinates of the stations. To improve the numerical behaviour of the MCMC algorithm, the actual fit used an orthogonalized version of these covariates obtained by applying a \( QR \)-decomposition (see, for example, Golub and Van Loan (1983)). The seasonal pattern was modelled using all six harmonics.

The MCMC algorithm was run for 5000 iterations after discarding the first 2000 iterations. Preliminary runs of the algorithm showed a very strong correlation between the \( \beta_i \) and the \( \sigma^2_i \).
To avoid convergence problems related to this fact, the transformations $\sigma_t^2 = \eta_t/\beta_t$ were used; the MCMC algorithm was run using $\eta_t$ instead of $\sigma_t^2$ and the results were transformed back. Several independent chains were run using different starting values to assess convergence.

Fig. 6 shows the histograms of the samples of the coefficients of the polynomial. All coefficients are seen to be significant, except perhaps $b_6$. We note that the coefficients corresponding to $x$ and $x^2$ have smaller absolute values than those corresponding to $y$ and $y^2$ respectively.

Fig. 7 shows the seasonal effects over $\mu_t$, for each $t$, together with 95% probability intervals for harmonics 1 and 2. The first harmonic is clearly dominant with a maximum in August and a minimum in February, which are among the wettest and the driest months respectively. The second harmonic has its most significant effects in January and July. The other four harmonics produced probability intervals which, in most months, included 0 and so were judged not to be significant.

Table 1 shows the means of the 35 parameters of interest, whereas Table 2 shows the standard deviations. We can see that large values of $\beta_t$ correspond to small values of $\sigma_t^2$ and these occur during the dry months. The estimation of $\beta_4$, corresponding to April, presented problems. In fact during the MCMC run the imputed values of $\beta_4$ were very unstable. This was because few data are available in April, since out of 192 observations 91 are 0 or missing,
the highest ratio of any month. We tackle the problem by imposing the restriction $1.5 \leq \beta_4 \leq 5$.

The mean value of $\lambda$ implies that the latent variable $w$ has a correlation of roughly 85% between the two nearest stations, which are 1.5 km apart, whereas the correlations are below 5% for stations more than 30 km apart, something that happens roughly in 75% of the cases.

**Table 1.** Posterior means of the parameters†

<table>
<thead>
<tr>
<th>( \sigma_t^2 )</th>
<th>( \beta_t )</th>
<th>( \lambda )</th>
<th>( b_0 )</th>
<th>( c_1 )</th>
<th>( d_1 )</th>
<th>( c_2 )</th>
<th>( d_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.63</td>
<td>3.66</td>
<td>8.94</td>
<td>43.73</td>
<td>-3.35</td>
<td>-2.30</td>
<td>1.11</td>
<td>0.24</td>
</tr>
<tr>
<td>0.97</td>
<td>3.68</td>
<td>8.94</td>
<td>-1.75</td>
<td>-3.35</td>
<td>-2.30</td>
<td>1.11</td>
<td>0.24</td>
</tr>
<tr>
<td>2.49</td>
<td>3.37</td>
<td>8.94</td>
<td>-6.73</td>
<td>-3.35</td>
<td>-2.30</td>
<td>1.11</td>
<td>0.24</td>
</tr>
<tr>
<td>1.85</td>
<td>4.94</td>
<td>8.94</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>1.91</td>
<td>4.71</td>
<td>8.94</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td>4.54</td>
<td>2.85</td>
<td>8.94</td>
<td>-0.044</td>
<td>-0.044</td>
<td>-0.044</td>
<td>-0.044</td>
<td>-0.044</td>
</tr>
<tr>
<td>3.67</td>
<td>2.60</td>
<td>8.94</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>5.73</td>
<td>2.39</td>
<td>8.94</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td>5.55</td>
<td>2.32</td>
<td>8.94</td>
<td>-0.044</td>
<td>-0.044</td>
<td>-0.044</td>
<td>-0.044</td>
<td>-0.044</td>
</tr>
<tr>
<td>2.38</td>
<td>2.59</td>
<td>8.94</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>1.04</td>
<td>3.01</td>
<td>8.94</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td>0.69</td>
<td>3.47</td>
<td>8.94</td>
<td>-0.044</td>
<td>-0.044</td>
<td>-0.044</td>
<td>-0.044</td>
<td>-0.044</td>
</tr>
</tbody>
</table>

†The values of \( \sigma_t^2 \) and \( \beta_t \) are from left to right for from January to December.

**Table 2.** Posterior standard deviations of the parameters†

<table>
<thead>
<tr>
<th>( \sigma_t^2 )</th>
<th>( \beta_t )</th>
<th>( \lambda )</th>
<th>( b_0 )</th>
<th>( c_1 )</th>
<th>( d_1 )</th>
<th>( c_2 )</th>
<th>( d_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.248</td>
<td>0.146</td>
<td>0.227</td>
<td>6.23</td>
<td>0.129</td>
<td>0.084</td>
<td>0.076</td>
<td>0.088</td>
</tr>
<tr>
<td>0.118</td>
<td>0.141</td>
<td>0.161</td>
<td>0.130</td>
<td>0.130</td>
<td>0.084</td>
<td>0.084</td>
<td>0.084</td>
</tr>
<tr>
<td>0.153</td>
<td>0.161</td>
<td>0.057</td>
<td>0.084</td>
<td>0.084</td>
<td>0.084</td>
<td>0.084</td>
<td>0.084</td>
</tr>
<tr>
<td>0.623</td>
<td>0.570</td>
<td>0.548</td>
<td>0.078</td>
<td>0.078</td>
<td>0.078</td>
<td>0.078</td>
<td>0.078</td>
</tr>
<tr>
<td>0.449</td>
<td>0.548</td>
<td>0.548</td>
<td>0.040</td>
<td>0.040</td>
<td>0.040</td>
<td>0.040</td>
<td>0.040</td>
</tr>
<tr>
<td>0.485</td>
<td>0.485</td>
<td>0.485</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
</tr>
<tr>
<td>0.214</td>
<td>0.214</td>
<td>0.214</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
</tr>
<tr>
<td>0.107</td>
<td>0.107</td>
<td>0.107</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
</tr>
<tr>
<td>0.062</td>
<td>0.062</td>
<td>0.062</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
</tr>
<tr>
<td>0.099</td>
<td>0.099</td>
<td>0.099</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
<td>0.038</td>
</tr>
</tbody>
</table>

†The values of \( \sigma_t^2 \) and \( \beta_t \) are from left to right for from January to December.
Fig. 8. Observed and predicted densities of rainfall for grouped observations in May: (a) group 1; (b) group 2; (c) group 3
Fig. 9. Observed and predicted densities of rainfall for grouped observations in July: (a) group 1; (b) group 2; (c) group 3
4.1. Predictions

Recalling the discussion in Section 3.1 we obtained samples of the predicted rainfall by sampling from multivariate normal distributions, truncating and then transforming. We used in each case the values of the parameters that correspond to the $i$th iteration. We present our results in terms of the median of those samples; these are compared with the medians of the observed rainfall, at each station, during May and July in Fig. 5. A tendency to underestimate the rainfall for stations with a small $y$-co-ordinate is present.

To illustrate the behaviour of the predictive densities we pooled the observations of the stations in each of the groups identified in Fig. 1. This was done since the number of observations per station is too small to produce illustrative histograms.

In Figs 8(a) and 9(a) we present the histograms of the observed rainfall for each of the groups with the estimated predicted density superimposed. In Figs 8(b) and 9(b) we present QQ-plots that correspond to the observations versus the predicted values. We note that, even when the observations are remarkably noisy, the predictive densities have a reasonable behaviour, except for the rather long tails in most of the cases. Fig. 8 corresponds to May and Fig. 9 to July.

To illustrate the spatial behaviour of the predicted rainfall, we show the dimensions and contour plots of the surfaces that correspond to the predicted medians. These can be seen in Fig. 10 for May and July. The peaks that we observe in the surfaces are due to the variability in the predictions; 300 samples were used at each point to produce the plot.

Fig. 11 shows the predicted probabilities of having no rainfall during March and January. These were calculated using the proportions of negative or null samples obtained. There are
quite high probabilities of not receiving any rainfall during those months; this is expected, since these two months are part of the dry season.

The behaviour of the predictions in time is illustrated in Fig. 4 where the predicted rainfall in stations 1 and 2 is compared with the observed rainfall for each month. Large observations such as those observed in June and November do not affect the estimation in a severe way. It is seen that from June to December the probability of a dry month, estimated by the model, is negligible.

5. Conclusions

We have proposed a truncated normal model to analyse rainfall collected at different sites during several periods of time. The model was fitted by making intensive use of an MCMC method that provides great flexibility to consider a large number of parameters and to model dry periods as well as missing values by using latent variables. The method has been developed for a challenging application involving a large number of sites and a relevant number of missing values. The model is seen to perform quite well for this application and produces sensible predictions for the amount of monthly rainfall and the probability of a dry month; these seem fairly robust with respect to unusually large observations.

The results are likely to be extended to the modelling of tropical rainfall in areas of little geographical diversity. In this particular example two important factors, the sparsity of data of the rainfall network and the limited length of historical records, make the application even
more valuable. This is likely to be the case in any region with limited data sources. One difficulty that was encountered with the model is that it tends to produce posterior predictive distributions with rather long tails. This can be due to the considerable variability observed in the data from year to year which produces large estimates of the transformation parameters $\beta$, for some months. Other transformations could be considered, the Box and Cox family being an obvious candidate. An interesting extension will be to consider correlation structures other than the structure proposed (see, for example, Cressie (1993)).

The model proposed is sufficiently flexible to allow for extensions and generalizations in many directions. It is worth noting, though, that the computational burden of sampling the posterior distribution of the parameters is already quite large, so adding extra complexity to the model must be carefully considered.

We can see two main directions for extensions of the model: one is to take into account a more sophisticated description of the spatial behaviour of the data, i.e. to consider other covariates, such as urban development or a more specific description of the geography; this could modify the design matrix $X$ as well as the parameterization of $\Sigma$. In the example that was presented no information was available about covariates. The other direction, which we think is important when considering rainfall accumulated over shorter periods of time, is to model a more complex serial dependence of the data. We are currently considering a modification of the model that would allow for short-term forecasts by dynamically varying the parameters of the linear structure of the model (Sansó and Guenni, 1998).

Acknowledgements

The first author wants to acknowledge the productive environment provided by the people and the resources of the Institute of Statistics and Decision Science at Duke University where this work was partially done while he was a visiting scholar; particular mention is due to the enlightening discussions that he had with Peter Müller and Mike West. We also acknowledge the useful comments of the Joint Editor and two referees that led to substantial improvements in the paper. This work has been partially supported by grant I-06 Banco Interamericano de Desarrollo–Consejo Nacional de Investigaciones Científicas y Tecnológicas, Venezuela.

References


