A latent structure model for high river flows

T. Economou [1], R. Vitolo[1], T. C. Bailey [1], E. Waterhouse[2] and Z. Kapelan[1]

[1] School of Engineering, Computer Science and Mathematics, University of Exeter, Harrison Building, North Park Road, Exeter, EX4 4QF, UK
[2] Department of Geography, Durham University, UK

Abstract: River discharge in the UK exhibit significant clustering of high-flow events on multidecadal timescales. A hidden semi-Markov model is constructed for the study of such multidecadal variability. The model includes time dependent covariates of climatological nature as well as a random effect driven by the hidden Markov states to account for possible non-explicit low-frequency climatic processes. The model is applied to an illustrative data set for river Severn in the UK.

Keywords: Hidden semi-Markov models; floods; ultralow-frequency variability.

1 Introduction

Floods are natural catastrophes which may have a disastrous effect in economic terms. For example, UK floods in summer 2007 resulted in the largest flood-related aggregate insured loss in the UK (Lane, 2007). These events have generated considerable concern in the insurance industry and, therefore, interest in better understanding the associated statistical properties. The problem of determining return periods for high-flow river discharge is particularly delicate if the underlying physical process is intrinsically non-stationary, for example due to climate change or to anthropogenic causes (e.g. change in land usage). A recent study of a number of catchments in UK has revealed remarkable hydrological volatility in the past: major floods appear to be characterised by significant spatio-temporal clustering (Robson, 2002). The pronounced temporal variability of flood occurrences has been described in terms of 'flood rich' and 'flood poor' periods which may extend for multiple decades (Robson, 2002; Lane, 2007).
Low-frequency behaviour at decadal timescales is a possible explanation for the clustering behaviour described above, somewhat contradicting the climate change hypothesis. Recent increasing trends in the frequency of flooding in certain catchments may be explained as irregularly recurring patterns of variability, occurring on multidecadal timescales. Climatic variability at such low frequencies has indeed been observed and described in

a fairly large number of studies and it is crucial to understand which climatic trends can be attributed to natural variability of the climate system, rather than to anthropogenic forcing. In this paper we consider a hidden semi-Markov model in an effort to try and capture features of the variability in flooding that may be driven by unobserved processes.

2 Model Specification

Consider a data set which is a time series $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$ of yearly counts of single days for which a flood event was recorded assuming that the river was observed for N years. Here we assume a non-stationary Poisson model for y_t $(t = 1, \ldots, N)$, where the mean $\Lambda(x_t)$ may depend on possibly time dependent covariates x_t. Furthermore we assume that the mean depends on a hidden state S_t of a semi-Markov chain at time t where $S_t \in \{1, 2, \ldots, S\}$ is the state space of the process. The mean is characterised as follows:

$$\Lambda\left(\boldsymbol{x}_t; S_t\right) = \exp\left\{\theta_{S_t} + \boldsymbol{\beta}\boldsymbol{x}_t\right\}$$

so that given the state S_t, y_t is Poisson distributed with a mean that depends on time through covariates $\boldsymbol{x}_t = (x_{1t}, x_{2t}, \ldots, x_{pt})$ which have associated parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. According to the state of a hidden semi-Markov chain, $\Lambda\left(\boldsymbol{x}_t; S_t\right)$ will be different for each state through the state dependent intercept θ_{S_t}. Here we assume that the resulting hidden semi-Markov Poisson (HSMP) model occurs in discrete time mainly due to the nature of flood data but also because it reduces the complexity of the model in a way. Note that through the hidden chain, some correlation structure is introduced in the counts y_t.
To derive the likelihood of the HSMP model consider first the likelihood of the Poisson model over a period τ given the state s of the chain during that period:

$$\ell(y_1, y_2, \ldots, y_\tau | s) = \prod_{i=1}^{\tau} \frac{e^{-\Lambda(\boldsymbol{x}_i; s)} \Lambda\left(\boldsymbol{x}_i; s\right)^{y_i}}{y_i!} \tag{1}$$

Second, consider a simple semi-Markov chain which is often defined by an initial state distribution $\boldsymbol{\pi} = (\pi(1), \pi(2), \ldots, \pi(S))$, a transition probability matrix $\boldsymbol{P} = \{p_{ij}\}$ where $p_{ii} = 0$, $\sum_j p_{ij} = 1$ and a vector of holding time distributions $\boldsymbol{h}(\tau) = \{h_i(\tau)\}$. So the chain starts at a state s_1 say, according to $\pi(s_1)$ and holds that state for a time interval τ_{s_1} according to distribution $h_{s_1}(\tau_{s_1})$, it then enters a new state s_2 according p_{s_1,s_2} and the process repeats itself analogously. The likelihood of a realisation $(\tau_{s_1}, \tau_{s_2}, \ldots, \tau_{s_n})$ of this chain involving n state changes is

$$\pi(s_1) h_{s_1}(\tau_{s_1}) \prod_{j=2}^{n} p_{s_{j-1},s_j} h_{s_j}(\tau_{s_j}) \tag{2}$$

Note that $h_i(\tau)$ can be any discrete distribution and if it is geometric then the chain is Markov and not semi-Markov. The reason for considering a semi-Markov chain here is because it increases model flexibility by not imposing a specific structure on $h_i(\tau)$.
Now suppose that both the time series $\boldsymbol{y}$ and the semi-Markov chain have been observed. Then the joint likelihood $L(y_1, \ldots, y_N; \tau_{s_1}, \ldots, \tau_{s_n})$ is obtained by combining (1) and (2):

$$\pi(s_1)h_{s_1}(\tau_{s_1})\ell(y_1, \ldots, y_{\tau_{s_1}}|s_1)p_{s_1,s_2}h_{s_2}(\tau_{s_2})\ell(y_{\tau_{s_1}+1}, \ldots, y_{\tau_{s_1}+\tau_{s_2}}|s_2) \times \text{etc.}$$

But since the chain is not observed, we sum over all possible states $s_i \in \{1, 2, \ldots, S\}$ and and all possible holding times $\tau_i \in \{1, 2, \ldots, \infty\}$ to obtain the marginal likelihood $L(y_1, \ldots, y_N)$ of the HSMP model as:

$$\sum_{\tau_{s_1}=1}^{\infty} \cdots \sum_{\tau_{s_n}=1}^{\infty} \sum_{s_1=1}^{S} \cdots \sum_{s_n=1}^{S} L(y_1, \ldots, y_N; \tau_{s_1}, \ldots, \tau_{s_n}) \tag{3}$$

In general, the HSMP likelihood is a function of the parameters in $\Lambda\left(\boldsymbol{x}(t); S_t\right)$, the unknown initial distribution $\boldsymbol{\pi}$ and transition matrix $\boldsymbol{P}$ and also the parameters $\boldsymbol{\phi}$ of the specified holding distributions $h_i(\tau_i)$. Given the complexity of the model we adopt an MCMC approach to model fitting considering that the computational cost in evaluating (3) is great given any reasonable observation interval and number of proposed states. By our assumption, the data $\boldsymbol{y}$ for the HSMP model are expressed in discrete time steps meaning that recursive algorithms used in the Hidden Markov models literature (MacDonald and Zucchini, 1997) can be analogously modified for efficient calculation of the HSMP likelihood. The idea in such recursion is to consider a variable $\alpha_t(j)$ sequentially at each discrete time step $t \in \{1, \ldots, N\}$, where:

$$\alpha_t(j) = \Pr(y_1, \ldots, y_t \text{ and chain exits state } j \text{ at time } t)$$

One can then compute $\alpha_t(j)$ recursively:

$$\begin{aligned}
\alpha_1(j) &= \pi(j)h_j(1)\ell(y_1|j) \\
\alpha_2(j) &= \pi(j)h_j(2)\ell(y_1, y_2|j) + \sum_{i \neq j} \alpha_1(i)p_{ij}h_j(1)\ell(y_2|j) \\
\alpha_3(j) &= \text{etc...}
\end{aligned}$$

Then $\sum_{j=1}^{S} \alpha_N(j)$ is equivalent to (3). Note that in the recursive expression for $\alpha_N(j)$, we replace $h_j()$ with its upper tail to account for right censoring in the final state duration. Once the likelihood is efficiently evaluated, it can be used in conjunction with Metropolis-Hastings to provide a computationally feasible estimation procedure for the parameters of the HSMP model. Here we use a combination of the random walk and the independence Metropolis samplers (Gilks et al., 1996) but do not provide any details.

3 Model Application

We examine daily discharge data for the river Severn at Bewdley (UK) in the 85 year period between 1922 and 2006. A flood is recorded when the discharge passes a certain threshold and the response $\boldsymbol{y}$ is defined as the number of days a flood has occurred in a year. Covariates $x_1(t)$ and $x_2(t)$ are used corresponding to the yearly averages of Atlantic multidecadal oscillation (AMO) and North Atlantic oscillation (NAO) indexes between 1922 and 2006 respectively. The possible presence of other, not explicit low-frequency processes is accounted for by the hidden semi-Markov chain. Specifically we assume two hidden states S_t in the chain where each has a Poisson holding time with a different mean. The model is then

$$\begin{aligned} y_t &\sim \text{Pois}(\Lambda(x_{1t}, x_{2t}; S_t)) \qquad t = 1, \ldots, 85 \\ \Lambda(x_{1t}, x_{2t}; S_t) &= \exp\{\theta_{S_t} + \beta_1 x_{1t} + \beta_2 x_{2t}\} \qquad S_t \in \{1, 2\} \end{aligned}$$

10000 samples were collected from the posterior distribution of each parameter and from the posterior predictive distributions of the fitted values. In figure (1) the black line represents the actual values $\boldsymbol{y}$, the dashed line shows the fitted values calculated as the means of the posteriors and bold lines show the 95% credible intervals calculated as the 95% quantiles of the posteriors (note that the lower interval is 0 for all years)

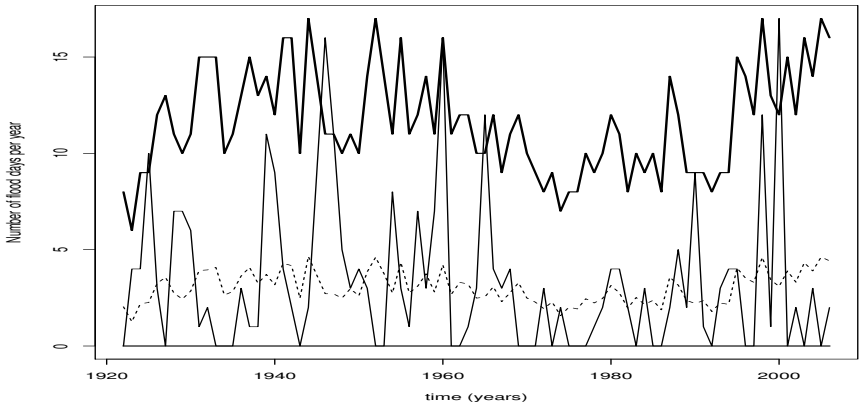

FIGURE 1. Fitted and actual values of the response.

3.1 Conclusion and Discussion

Figure 1 shows that the model is able to capture the increased variance in prevalence between 1930 and 1960 and in the last part of the record: these

are two 'flood-rich' periods for the Severn, separated by a long 'flood-poor' period between 1960 and the late 1990s. Although this behaviour is mainly explained by the covariate AMO and not the hidden chain, the model did identify two hidden states, one with a very small prevalence in time but with a higher value of θ_{S_t} in the Poisson mean which is what the data is suggesting looking at the 'spikes' of large values in the observed counts in Figure (1). This is reflected in the sufficiently high credible intervals.
Possible extensions to the model include the introduction of a (spatial) random effect to facilitate a multi-catchment scenario. Also, one of the main point of interest will be to try and characterise the hidden states of the Markov model in terms of physical processes which would be possibly useful in setting up improved seasonal or interannual forecasts of high-flows.

Acknowledgments: The authors are indebted to Prof S. Lane for insightful discussions.

References

Gilks,W.R., Richardson S. and Spiegelhalter D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall

Lane, S.N. (2007). The 2007 UK summer floods: a scientific perspective *www.willisresearchnetwork.com/Lists/Publications/DispForm.aspx?ID=6*.

MacDonald, I. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*. Chapman and Hall.

Robson, A.J. (2002). Evidence for trends in UK flooding. *Philosophical transactions of the Royal Society A*, **360**, 1327.