Using statistical methods to analyse environmental extremes.

Emma Eastoe

Department of Mathematics and Statistics
Lancaster University

December 16, 2008

Focus of talk

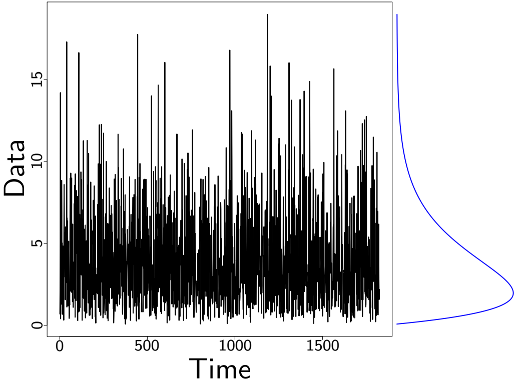
- Discuss statistical models used to predict the size and/or times of unusually large (extreme) events;
- Show how these models can be adapted to incorporate physical structure of data;
- Two examples - air pollution and river flow.

Unusually large events

- How high should a sea wall be built so that it is breached (on average) only once every 100 years? *Coastal engineer*.
- How high should a dam be built so that it floods (on average) only one year in every 10000? *Engineer, water company*.
- On how many days a year will ozone levels exceed safety levels? *Health worker, traffic planner*.
- Where should an oil rig be located to be best protected from the largest waves? *Oil company, engineer*.

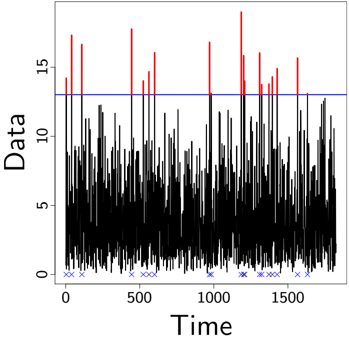
Background

- Data : time series of daily observations for the last n years;
- Assume data are independent and identically distributed (IID);
- *i.e.* data are an independent random sample from a probability distribution with constant parameters, *e.g.* Normal(μ, σ), Exponential(λ).



Background 2

- Interest is in predicting unusually large events
- Fit a statistical model based only on data from the upper tail of the underlying distribution;
- *i.e.* ignore small/medium-sized data;
- Peaks over threshold (POT): select a high constant threshold u and model rate and size of threshold exceedances.



Background 3

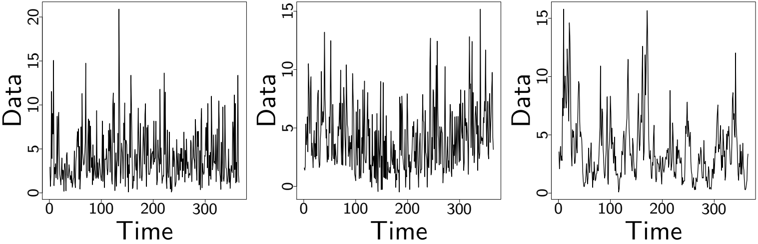
Mathematical results support the use of the following models for the rate and size of threshold exceedances:

- Rate : a Poisson process with parameter λ;
- Size : the two-parameter generalised Pareto distribution (GPD).

Use the fitted models to calculate the N-year return level, *i.e.* the level exceeded on average once every N years, for N much larger than the length of the data set n.

Model assumption that data are IID is usually too simplistic.

- Series displays trends;
- Series is correlated at short lags.



(a) IID (b) Trend (c) Correlated

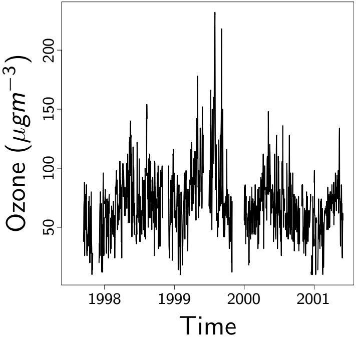
Part 1

Data with trend

- Mathematical theory which supports the threshold model only extends to a few special cases of trend;
- Modellers use the POT model but allow Poisson process and GPD parameters to vary with time and/or covariates;
- Parametric (GLM) or non-parametric (LOESS, GAM) models;
- Numerical difficulties with model fitting;
- Discuss an alternative approach - better motivated by theory *and* easier to fit.

Ozone 1

- Surface-level ozone data, centre of Reading;
- Summer peaks, winter troughs;
- Responds to changes in precursors (*e.g.* NO, NO_2) as well as sunshine, temperature and wind speed/direction.



Ozone 2

First remove trends from full data set $\{Y_t\}$ using covariates $\{\mathbf{x}_t\}$, then model extremes which should be IID.

- Model trends in mean μ and scale σ by supposing that transformed data $\{Y_t^\lambda\}$ are a sample from a Normal$(\mu(\mathbf{x}_t), \sigma(\mathbf{x}_t))$ distribution;
- Model mean and variance as linear functions of covariates, *e.g.*

$$\mu(\mathbf{x}_t) = \boldsymbol{\mu}'\mathbf{x}_t = \mu_0 + \mu_1 \mathbf{x}_{1,t} + \ldots + \mu_p \mathbf{x}_{p,t}$$

  where $\boldsymbol{\mu}$ is a vector of regression coefficients.
- Standardise the data by

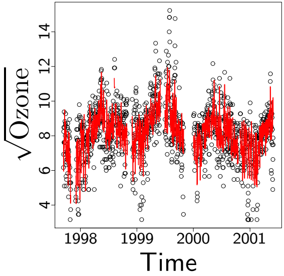
$$Z_t = \frac{Y_t^\lambda - \mu(\mathbf{x_t})}{\sigma(\mathbf{x_t})}.$$

To model the extremes:

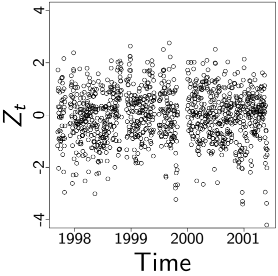
- Select a high constant threshold u_z, *e.g.* 99% quantile of the standardised series $\{Z_t\}$;
- Model the rate and size of threshold exceedances of u_Z using Poisson process and GPD models with constant parameters;
- Could include covariates in POT model parameters, especially if extremes might respond in a different way to covariates;
- Effective threshold now varies in time:

$$u(\mathbf{x_t}) = \left[u_Z \sigma(\mathbf{x_t}) + \mu(\mathbf{x_t})\right]^{1/\lambda}.$$

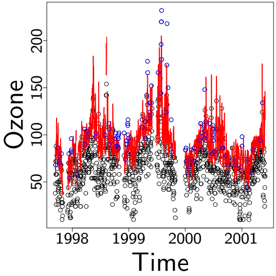
Modelled mean $\mu(\mathbf{x_t})$ and variance $\sigma^2(\mathbf{x_t})$ of square root of ozone.



(a) Mean



(b) Standardised ozone



(c) Effective threshold

Return levels

From this model, some estimated return levels (with 95% confidence intervals) are

Return period	5-yr	10-yr	100-yr
Return level	218	234	289
($\mu g m^{-3}$)	(197,247)	(208,271)	(239,387)

These seem realistic

- The 5-year return level is exceeded on only three days over the four years of observed data (at the end of July/beginning of August);
- Neither the 10- nor the 100-year return levels is exceeded at all.

Part 2

River flow 1

Time-series of daily flows at Kingston on the River Thames (1883-2006):

What is the best statistical model for the maximum annual flow at this site?

Could extract annual maxima and model these, but leads to loss of information ...

Instead use models for the number of events in a year and the size of the peak flow in an event ...

River flow 1

Time-series of daily flows at Kingston on the River Thames (1883-2006):

What is the best statistical model for the maximum annual flow at this site?

Could extract annual maxima and model these, but leads to loss of information ...

Instead use models for the number of events in a year and the size of the peak flow in an event ...

- What is the most appropriate model for the annual number of events?

River flow 1

Time-series of daily flows at Kingston on the River Thames (1883-2006):

What is the best statistical model for the maximum annual flow at this site?

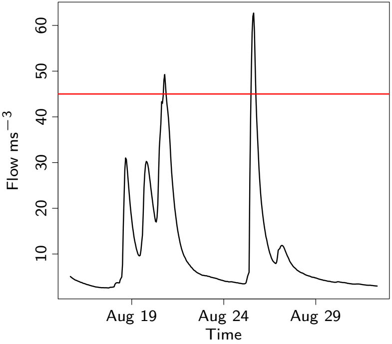
Could extract annual maxima and model these, but leads to loss of information ...

Instead use models for the number of events in a year and the size of the peak flow in an event ...

- What is the most appropriate model for the annual number of events?
- How does the model used for this affect the implied annual maxima distribution?

River flow 2

- Data show strong correlation at short time lags;
- Select a high threshold u to identify associated flow events;
- Independent events begin with a threshold exceedance and end after m consecutive non-exceedances.



Homogeneous Poisson model

- Let N_i be the number of events in year i;
- Could assume N_i's are an independent random sample from a Binomial$(365, p)$ distribution;
- Number of observations large and probability of an event p small so use Poisson approximation;
- Assume that N_i's are an independent random sample from a Poisson(λ) distribution;
- Interpretation: λ is the mean number of events per year.

Over-dispersion

- A consequence of this model choice is that the between-year variance in the number of events is equal to the mean of the number of events per year;
- However, previous studies have shown that the between-year variance is larger than the mean for most rivers in the UK;
- Reason - unobserved (‘missing’) covariates, *e.g.* precipitation, antecedent soil conditions?
- This extra variation between years cannot be captured by homogeneous Poisson model.

Latent variables (or random effects)

Build a model for the 'missing' covariates (latent variables).

1. Denote by γ_i the latent variable for year i and assume these are an independent random sample from a Gamma$(1/\alpha, 1/\alpha)$ distribution;
2. Assume N_i are independent in time and follow a Poisson distribution with mean varying from year to year,

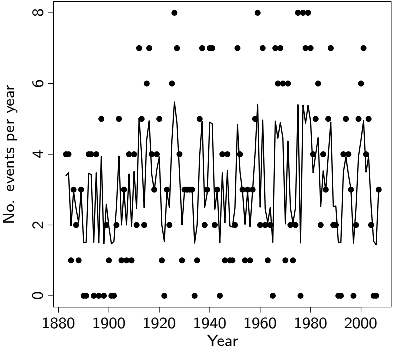
$$\lambda_i = \lambda\gamma_i, \quad \lambda > 0$$

Under this model

- The mean number of events per year is λ;
- Between-year variance has increased from λ to $\lambda(1 + \lambda\alpha)$.

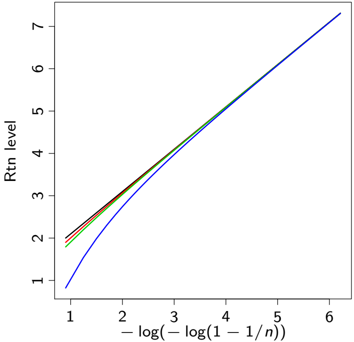
R. Thames, Kingston

- Average of 3.3 events per year and between-year variance of 5.7;
- Fitting the latent variable model gives parameter estimates of $\hat{\lambda} = 3.3\ (2.9, 3.8)$ and $\hat{\alpha} = 0.26\ (0.14, 0.43)$;
- Observed (dots) and estimated mean (line) number of events



Annual maximum

- 3 to 500 year return levels;
- Homogeneous Poisson process (black)
- Latent variables with $\alpha = 0.5$ (red), $\alpha = 1$ (green) and $\alpha = 5$ (blue)



- At higher return levels we have 'averaged over' all values of the latent variables;
- Bigger α implies more years with zero events, hence lower short period return levels.

Extension - within year variability

- Could model the point process δ_{ij} where

$$\delta_{ij} = \begin{cases} 1 & \text{if event peak on day } j \text{ of year } i; \\ 0 & \text{if no event peak on day } j \text{ of year } i. \end{cases}$$

- Inhomogeneous Poisson process model with rate parameter

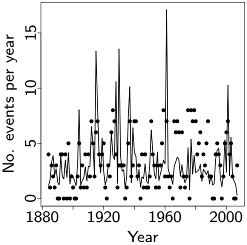
$$\lambda_{ij} = \gamma_i \exp\{\boldsymbol{\beta}' \mathbf{x}_{ij}\},$$

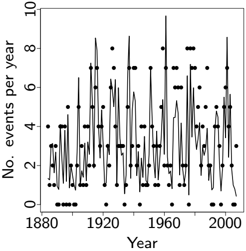
  $\boldsymbol{\beta}$ are regression coefficients, $\mathbf{x}_{ij}$ are covariates and γ_i are as earlier.
- Use latent variables as a diagnostic for covariate selection;
- Distribution of annual maxima by simulation.

R. Nairn at Firhall

Covariates - baseflow and 3-month aggregated rainfall.

- Left - covariates only;
- Right - covariates and latent variables;





- Over-estimation of mean number of events in covariates only model.

Conclusions

- Statistical models for extreme values provide a useful way to predict unusual events;
- These models can be adapted to a variety of applications;
- And can be constructed to incorporate known physical structure or relationships;
- Any questions, comments, possible applications are welcome ...

Thank-you!