

BioSS Journal Club

Adam Butler, May 2005

Discussion of: **Smith, R. L. (1989)**

Extreme Value Analysis of Environmental Time Series:
An Application to Trend Detection in Ground-Level Ozone.

Statistical Science, 4(4), 367-393.

Structure of talk

- 1 : Ground-level ozone & human health
- 2 : Statistical strategy
- 3 : Extreme value methods: some background
- 4 : The point-process model
- 5 : Results of the data analysis
- 6 : Issues from the printed discussion
- 7 : Subsequent papers

1 : Ground-level ozone & human health

- **Health risk:**

levels of ground-level ozone are an indicator of air quality

- **Regulatory framework:**

≤ 3 exceedances of a 12pphm threshold over a 3y period

- **Process understanding:**

ozone levels are related to both emissions & meteorology

- **Monitoring data:**

95433 hourly records at Houston, Texas, during 1973-1986

2 : Statistical strategy

- **Parameters of interest:**

Crossing rates of high levels &/or n -year return values

- **Extreme value approach:**

Use a parametric model for data above a *high threshold* u

Approach justified by asymptotic theory as $u \rightarrow \infty$

Smith adopts a threshold of $u = 8\text{pphm}$

Incorporate trend and seasonality into the parameters

- **Declustering:**

Data exhibit strong short-range dependence

Concentrate on modelling the *cluster maxima*

Requires specification of a cluster interval: Smith uses 72h

3 : Extreme value methods - some background

- The extreme value condition

= Let X_1, X_2, \dots be an iid sequence with d.f. F

= Let $M_n = \max(X_1, \dots, X_n)$

= We look for sequences $a_n > 0$ and b_n such that

$$\mathbb{P} \left\{ \left(\frac{M_n - b_n}{a_n} \right) \leq x \right\} \rightarrow H(x),$$

or, equivalently,

$$n\{1 - F(a_n x + b_n)\} \rightarrow -\log H(x),$$

where $H(x)$ is nondegenerate.

- **The GEV distribution**

If (3.2) holds then (Fisher & Tippett, 1928) H has a *Generalised Extreme Value distribution*

$$H(x; \mu, \sigma, \xi) = \exp[-\{1 - k(x - \mu)/\sigma\}^{1/k}]$$

with parameters $\mu, \sigma > 0$ and ξ .

- **The GPD distribution**

If (3.2) holds then as $u \rightarrow \infty$ the conditional distribution

$$(X > u + y | X > u)$$

tends to a *Generalised Pareto distribution*

$$1 - (1 - ky/\sigma)^{1/k}$$

- **Dependence sequences**

Under a mild mixing condition, we have that

$$\mathbb{P}(M_n \leq x) \approx [F(x)]^{n\theta},$$

where $0 \leq \theta \leq 1$ is the *extremal index*

- **Statistical application**

- = Assume that block maxima follow a GEV distribution, for sufficiently long blocks
- = Assume that exceedances of u follow a GPD distribution, for a sufficiently high threshold u
- = Estimate the parameters of the GEV or GPD distribution, using block maxima (GEV) or exceedances of u (GPD)

4 : The point-process model

- **Theoretical result:**

= Assume condition (3.2) holds for sequences a_n & b_n

= Let X_1, \dots, X_n denote a random sample from F

= Let P_n denote the PP on \mathbb{R}^2 with points at

$$\left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right)$$

= P_n converges (as $n \rightarrow \infty$) to a process P with intensity $(t_2 - t_1)[1 - k(x - \mu)/\sigma]^{1/k}$, for $0 \leq t_1 \leq t_2 \leq 1$.

- **Statistical application:**

= View exceedances of a high threshold u as realisations of the *limiting* point process P

= Estimate parameters μ , σ and k by maximum likelihood

= Build a regression model around the parameters:

$$\mu_{ij} = \alpha_j + i\beta_j, \sigma_{ij} = \sigma_j, k_{ij} = k_j$$

(for period j of year i)

= Derive:

$$\tau(y) = (M/365) \sum_j \{1 - k_j(y - \mu_{ij})/\sigma_j\}^{1/k_j}$$

5 : Results of the data analysis

- **Shape parameter:**

Values of k imply a distribution with finite upper endpoint;
Some evidence that k is not common across months.

- **Temporal trend:**

Incorporate by splitting data or by linear trend in location;
Weak evidence for a decreasing trend in the extremes;
Strength of evidence depends on which outcome variable.

- **Sensitivity & diagnostics:**

Relatively insensitive to threshold, cluster interval & period;
Weak downward trend is also visible from the raw data.

6 : Issues from the printed discussion

- **Strength of evidence for a downward trend:**

Parametric form of the temporal trend model;

Method of inference and model selection.

- **Declustering:**

Removing short-range dependence, and the diurnal signal.

- **Long-range dependence:**

Do correlations persist at long (temporal) lags ?

- **Measurement error & calibration:**

Ozone measurements are manual, & open to human error;

Recalibration of instrumentation, quality assessment.

- **Spatial aspects:**

Trends may be due to highly localised effects;

Regulations refer to *all* locations, not just to sampled ones;

Design issues, relating to the choice of sampling locations.

- **Objectives of the analysis:**

Statistical vs scientific significance of a trend in ozone;

End use: inform policy, or contribute directly to control ?

“Legislators, industry, environmentalists and the public in general will be clamoring for information on whether and how quickly the air is improving... It may be a statistician’s duty to inform them that such a determination is impossible to make.”

7 : Subsequent work

- **Regression modelling:**

Parametric (Coles, 2001; Katz et. al., 2002)

Nonparametric (e.g. Chavez-Demoulin & Davison, 2005)

R: `evd`, `ismev`, `evir`, `evdbayes`, `extRemes`, `fExtremes`

- **Bayesian methods:**

Coles & Powell (1996), Stephenson & Tawn (accepted)

- **Spatial extremes:**

Geostatistical: Yun & Smith (2003)

Smoothing/linkage: Dixon & Tawn (1998)

- **Multivariate extremes:** to be continued next week...