Statistics for extreme & sparse data

Aoibheann Brady

University of Bath

December 6, 2018

Plan

Overview of WTW problems & available data

The Problem

Climate Change = Bad!

4 key problems

- Volcanic eruptions/catastrophic event prediction.
- Windstorms — how to work out potential losses should one occur.
- Weather extreme indices.
- Parametric insurance applications.

Data

- **Windstorm data:** `http://www.europeanwindstorms.org/`
- **Volcanic eruption data:** `https://www.ngdc.noaa.gov/nndc/servlet/ShowDatasets?dataset=102557&search_look=50&display_look=50`
- **Climate data — station & gridded:** `https://www.metoffice.gov.uk/climate/uk/data`
- **Climate index data:** `https://www.climdex.org/datasets.html`

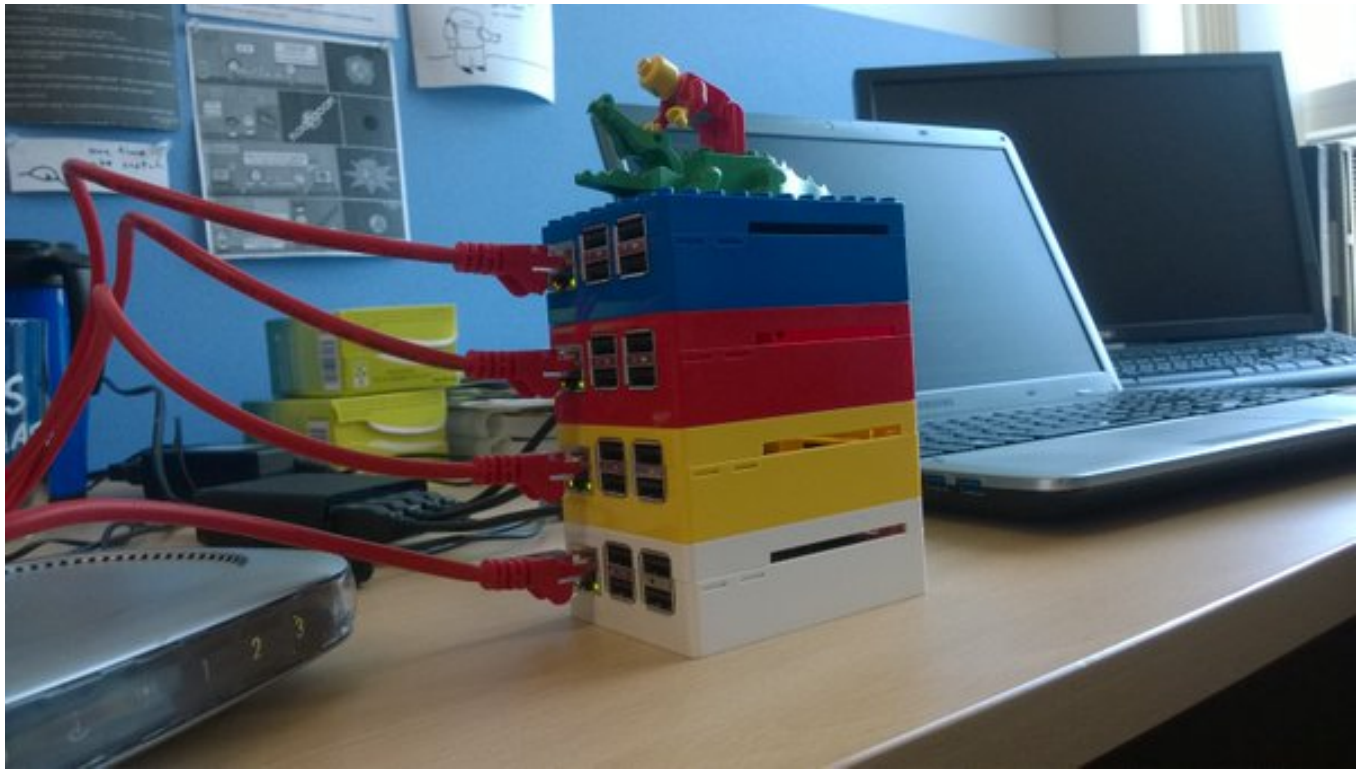
Key issues

- Dealing with **very sparse data** — both in space and in time.
- Ground monitoring stations are often very scattered and coverage of many locations is very poor.
- These events are **extreme in nature** so tend to be spread out in time.

Combining multiple sources of data

Source 1: Data from Monitoring Stations

Source 2: Numerical Model Output



Combining data sources?

Why?: Need to accurately assess exposure (e.g. in volcanic eruption) to estimate financial or health risks.

Types of data:

- Directly measured data (e.g. pollutant concentrations) from monitoring networks
- Numerical model output, which provides estimates of the average level in grid cells.

Directly measured data provides a **finer level** on a spatial scale but **considerable missingness**.

Modelling involves **numerically solving complex systems** of differential equations capturing various physical processes. This output spans large spatial domains with no missingness.

Combining data sources?

Why?: Need to accurately assess exposure (e.g. in volcanic eruption) to estimate financial or health risks.

Types of data:

- Directly measured data (e.g. pollutant concentrations) from monitoring networks
- Numerical model output, which provides estimates of the average level in grid cells.

Directly measured data provides a **finer level** on a spatial scale but **considerable missingness**.

Modelling involves **numerically solving complex systems** of differential equations capturing various physical processes. This output spans large spatial domains with no missingness.

Combining data sources?

Why?: Need to accurately assess exposure (e.g. in volcanic eruption) to estimate financial or health risks.

Types of data:

- Directly measured data (e.g. pollutant concentrations) from monitoring networks
- Numerical model output, which provides estimates of the average level in grid cells.

Directly measured data provides a **finer level** on a spatial scale but **considerable missingness**.

Modelling involves **numerically solving complex systems** of differential equations capturing various physical processes. This output spans large spatial domains with no missingness.

Combining data sources?

Why?: Need to accurately assess exposure (e.g. in volcanic eruption) to estimate financial or health risks.

Types of data:

- Directly measured data (e.g. pollutant concentrations) from monitoring networks
- Numerical model output, which provides estimates of the average level in grid cells.

Directly measured data provides a **finer level** on a spatial scale but **considerable missingness**.

Modelling involves **numerically solving complex systems** of differential equations capturing various physical processes. This output spans large spatial domains with no missingness.

What Are the Benefits of Data Fusion?

1. **Advantages**
   - Fusion can improve assessment of exposure at high resolution.
   - Allows us to combine observational data on the current state of the atmosphere with a short-range numerical forecast to obtain initial conditions for a numerical atmospheric model.
2. **Disadvantages**
   - Many methods proposed in atmospheric data assimilation are ad-hoc, algorithmic and don't address the "change of support" problem.

Change of Support Problem

Spatial data can be collected at point locations or associated with areal units.

The **spatial support** is the volume/shape/size/orientation associated with each spatial measurement.

The **change of support problem** is concerned with inference about the values of a variable at a level different from that which it has been observed. It involves studying statistical properties of a spatial process as we change the spatial support (Banerjee, Carlin and Gelfand, 2004).

Change of Support Problem

Spatial data can be collected at point locations or associated with areal units.

The **spatial support** is the volume/shape/size/orientation associated with each spatial measurement.

The **change of support problem** is concerned with inference about the values of a variable at a level different from that which it has been observed. It involves studying statistical properties of a spatial process as we change the spatial support (Banerjee, Carlin and Gelfand, 2004).

Change of Support Problem

Spatial data can be collected at point locations or associated with areal units.

The **spatial support** is the volume/shape/size/orientation associated with each spatial measurement.

The **change of support problem** is concerned with inference about the values of a variable at a level different from that which it has been observed. It involves studying statistical properties of a spatial process as we change the spatial support (Banerjee, Carlin and Gelfand, 2004).

Methods of Combining Sources

Kriging, a method of interpolation for which the interpolated values are modeled by a Gaussian process, is often used to interpolate between sites. *Not ideal when data is sparse!!!*.

If one is interested at a grid level, can use **upscaling** to make area-level inferences based on point data.

Conversely, can use **downscaling** when observing at an area level (e.g. gridded data) and want to make point-level inferences. We'll focus primarily on this.

Methods of Combining Sources

Kriging, a method of interpolation for which the interpolated values are modeled by a Gaussian process, is often used to interpolate between sites. *Not ideal when data is sparse!!!*.

If one is interested at a grid level, can use **upscaling** to make area-level inferences based on point data.

Conversely, can use **downscaling** when observing at an area level (e.g. gridded data) and want to make point-level inferences. We'll focus primarily on this.

Methods of Combining Sources

Kriging, a method of interpolation for which the interpolated values are modeled by a Gaussian process, is often used to interpolate between sites. *Not ideal when data is sparse!!!*.

If one is interested at a grid level, can use **upscaling** to make area-level inferences based on point data.

Conversely, can use **downscaling** when observing at an area level (e.g. gridded data) and want to make point-level inferences. We'll focus primarily on this.

Downscaling

Two types of downscaling:

- **Dynamical downscaling:** Output from Global Climate Models used to drive regional numerical model in higher spatial resolution. Physical/numerical approach!
- **Statistical downscaling:** A statistical relationship is established from observations between large scale variables (atmospheric surface pressure) and a local variable (a site's wind speed). The relationship is used on the GCM data to obtain the local variables from the GCM output.

Downscaling in practice

A tutorial on **working with spatial data & downscaling** can be seen at:
`http://www.wmo.int/pages/prog/wcp/agm/meetings/korea2016/documents/Tutorial_Statistical_Downscaling.pdf`

Another good source is this paper by Shaddick, Thomas et al (2018) which explains how to fit it within a **Bayesian framework**.
`https://pubs.acs.org/doi/10.1021/acs.est.8b02864`

Downscaling in practice

A tutorial on **working with spatial data & downscaling** can be seen at:
`http://www.wmo.int/pages/prog/wcp/agm/meetings/korea2016/documents/Tutorial_Statistical_Downscaling.pdf`

Another good source is this paper by Shaddick, Thomas et al (2018) which explains how to fit it within a **Bayesian framework**.
`https://pubs.acs.org/doi/10.1021/acs.est.8b02864`

Working with extreme-valued data

Potential options

Two heavy-tailed models that can allow large observations arise out of Extreme Value Theory:

1. **Generalised Extreme Value distribution (GEV):** This is the limit distribution for the sample maxima, the largest observation in each period of observation.
2. **Generalised Pareto Distribution (GPD):** This is the limit distribution of losses which exceed a high enough threshold.

Given count data for the events exceeding a given threshold, one can also use **Poisson regression** to investigate whether the frequency of these events is changing.

GEV distribution for block maxima

May wish to look at a series of **block maxima**

$$M_n = \max\{X_1, ..., X_n\},$$

a sequence of *iid* RVs with some common distribution function. The M_n represents the maximum (or minimum!) of this process over n units of time.
Can use the **generalized extreme value distribution** as a model for these block maxima.

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{1}$$

The model has a location parameter, μ, a scale parameter, σ, and a shape parameter, ξ.

GPD: Is the magnitude of events changing?

Given peak data X, for a large threshold u, the distribution of $(X - u)$ conditioned on $X > u$ may be approximated by:

$$H(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{1/\xi}.$$

This function is defined on $\{y : y > 0 \ \& \ \xi y/\bar{\sigma} > 0\}$, and $\bar{\sigma} = \sigma + \xi(u - \mu)$.

This family of distributions is known as the **generalised Pareto family of distributions**. The *size* of threshold exceedances may be approximated by a member of this family.

Poisson regression: Is the frequency of these events changing?

We can fit a **generalised linear model**. We have count data for the numbers of peaks over threshold for each year, so can assume a **Poisson distribution**.

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = x_i^T \beta$$

where $y_i \sim \text{Poisson}(\mu_i)$ and we use the natural log link $g(\mu) = \log(\mu)$.

How about both? A point process representation.

We want something that looks both at the size and number of exceedances using the GPD style approach.

The **point process model** describes both the magnitude of threshold exceedances and the rate at which the threshold u is exceeded.

It is parameterised by three parameters – location, scale and shape.

Calculating extreme weather indices

What are extreme weather indices?

These are usually values derived from daily temperature & precipitation data.

There is a focus in **changes in extremes**, rather than the usual change in means.

Examples include number of days during which temperature exceeds its long-term 90th percentile, or annual count of days when rain exceeds 10mm.

What are extreme weather indices?

These are usually values derived from daily temperature & precipitation data.

There is a focus in **changes in extremes**, rather than the usual change in means.

Examples include number of days during which temperature exceeds its long-term 90th percentile, or annual count of days when rain exceeds 10mm.

What are extreme weather indices?

These are usually values derived from daily temperature & precipitation data.

There is a focus in **changes in extremes**, rather than the usual change in means.

Examples include number of days during which temperature exceeds its long-term 90th percentile, or annual count of days when rain exceeds 10mm.

How can we calculate them?

Two ways:

- Station-level index to gridded index
  - Calculate the index at a station level (e.g. by calculating **threshold exceedances**, possibly by GPD or contour sets).
  - Then grid these indices using techniques such as cubic spline interpolation.
- Grid to index
  - Simulate from posterior distribution having fitted a spatial model.
  - Calculate indices as before.

Problem! Stations not evenly distributed across the global land area $\Longrightarrow$ difficult to accurately compute global averages.

A simple average of data from all available stations would result in a representation biased toward areas of higher station density.

How can we calculate them?

Two ways:

- Station-level index to gridded index
  - Calculate the index at a station level (e.g. by calculating **threshold exceedances**, possibly by GPD or contour sets).
  - Then grid these indices using techniques such as cubic spline interpolation.
- Grid to index
  - Simulate from posterior distribution having fitted a spatial model.
  - Calculate indices as before.

Problem! Stations not evenly distributed across the global land area $\Longrightarrow$ difficult to accurately compute global averages.

A simple average of data from all available stations would result in a representation biased toward areas of higher station density.

How can we calculate them?

Two ways:

- Station-level index to gridded index
  - Calculate the index at a station level (e.g. by calculating **threshold exceedances**, possibly by GPD or contour sets).
  - Then grid these indices using techniques such as cubic spline interpolation.
- Grid to index
  - Simulate from posterior distribution having fitted a spatial model.
  - Calculate indices as before.

Problem! Stations not evenly distributed across the global land area $\implies$ difficult to accurately compute global averages.

A simple average of data from all available stations would result in a representation biased toward areas of higher station density.

Prediction of catastrophic events

Prediction of catastrophic events

Multiple potential outputs of interest to insurers & general population.

- Can be considered as rare event prediction — see extreme value distributions covered earlier.
- Predicting when an event is likely to occur based on behaviour prior to the event. When to raise the alarm? How to be aware of risk to life?
- Very sparse data — e.g. volcanic eruption where some are very well studied (such as Vesuvius) and many others aren't. Can we apply knowledge across?

Changepoint Analysis Model

We often wish to identify times when the probability distribution of a stochastic process or times series changes. Generally, we will want to look at whether a change (or changes) has occurred, and identifying the times of such a change.

Definition

Given a dataset $X_{1:n} = \{X_1, X_2, ..., X_n\}$ of real-valued observations we wish to find an ideal set of K non-overlapping intervals where the observations are homogeneous within each. For K segments, the change-point model expresses the distribution of X given some segmentation $S_{1:n} = \{S_1, S_2, ..., S_n\} \in M_k$ as:

$$P_\theta(X_{1:n}|S_{1:n}) = \prod_{i=1}^{n} \beta_{S_i}(X_i) = \prod_{s=1}^{K} \prod_{i, S_i = s} \beta_s(X_i)$$

Changepoint Analysis Model

We often wish to identify times when the probability distribution of a stochastic process or times series changes. Generally, we will want to look at whether a change (or changes) has occurred, and identifying the times of such a change.

Definition

Given a dataset $X_{1:n} = \{X_1, X_2, ..., X_n\}$ of real-valued observations we wish to find an ideal set of K non-overlapping intervals where the observations are homogeneous within each. For K segments, the change-point model expresses the distribution of X given some segmentation $S_{1:n} = \{S_1, S_2, ..., S_n\} \in M_k$ as:

$$P_\theta(X_{1:n}|S_{1:n}) = \prod_{i=1}^{n} \beta_{S_i}(X_i) = \prod_{s=1}^{K} \prod_{i, S_i = s} \beta_s(X_i)$$

Hidden Markov Models

Definition

A *Hidden Markov Model (HMM)* is a statistical Markov model in which the system being modelled is assumed to be a Markov process with some unobserved, or hidden, states. It is a particular kind of Bayesian network (dynamic!). A typical HMM is given by the joint probability distribution:

$$P(X_{1:n}|S_{1:n}) = P(S_1)P(X_1|S_1)\prod_{i=2}^{n} P(S_i|S_{i-1})P(X_i|S_i)$$

HMM Example

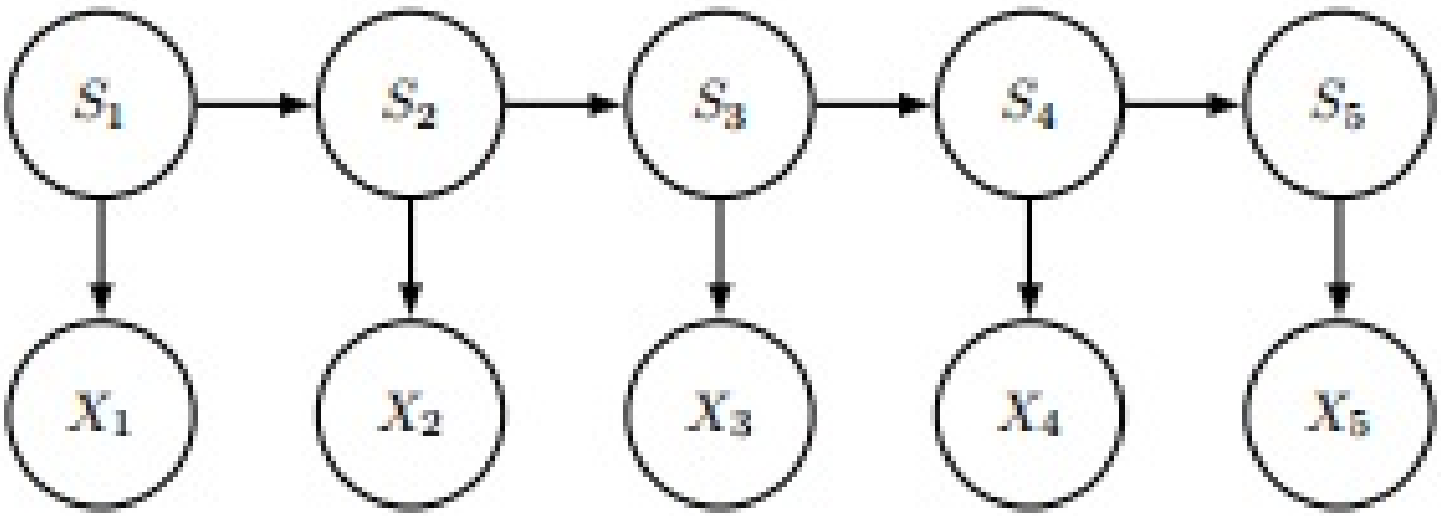


Figure: HMM topology with n=5. S_i are the hidden states and X_i the observed states, $i = 1, ..., 5$

HMMs for changepoint analysis

We can look at changepoint analysis as a HMM where:

- Data are the observations
- Unknown segmentations are the hidden states

So HMM adaptations **can identify changepoints by observations where a switch in hidden states is most likely to occur**. HMM approaches are also useful in inferential procedures - posterior estimations, etc.

Alternative: Treat it as an optimal stopping problem!

Suppose we have a continuously observable process behaving as a standard Brownian motion up to time τ_1, then as a B.M. with known drift afterwards.

At stopping time τ_2 some observable event occurs — e.g. a volcanic eruption or other.

Alternative: Treat it as an optimal stopping problem!

Suppose we have a continuously observable process behaving as a standard Brownian motion up to time τ_1, then as a B.M. with known drift afterwards.

At stopping time τ_2 some observable event occurs — e.g. a volcanic eruption or other.

Alternative: Treat it as an optimal stopping problem!

Suppose we have a continuously observable process behaving as a standard Brownian motion up to time τ_1, then as a B.M. with known drift afterwards.

At stopping time τ_2 some observable event occurs — e.g. a volcanic eruption or other.

Alternative: Treat it as an optimal stopping problem!

Løkka (2007) investigates the problem of **detecting changes in a system's behaviour prior to the event occurring**, taking into account information given by non-occurrence of the observable event.

Also looks at where it is favourable to "raise the alarm" before this event.

Løkka (2007) shows that this problem can be reduced to a **1D optimal stopping problem**, & derive an explicit solution.

Alternative: Treat it as an optimal stopping problem!

Løkka (2007) investigates the problem of **detecting changes in a system's behaviour prior to the event occurring**, taking into account information given by non-occurrence of the observable event.

Also looks at where it is favourable to "raise the alarm" before this event.

Løkka (2007) shows that this problem can be reduced to a **1D optimal stopping problem**, & derive an explicit solution.

Alternative: Treat it as an optimal stopping problem!

Løkka (2007) investigates the problem of **detecting changes in a system's behaviour prior to the event occurring**, taking into account information given by non-occurrence of the observable event.

Also looks at where it is favourable to "raise the alarm" before this event.

Løkka (2007) shows that this problem can be reduced to a **1D optimal stopping problem**, & derive an explicit solution.

Historical data

How to include historical data

All will be revealed at PSS on Thursday!

References

V. J. Berrocal, A. E. Gelfand & D. M. Holland (2012)

Space-Time Data Fusion under Error in Computer Model Output: An Application to Modeling Air Quality

Biometrics 68, 837-848

S. Banerjee, B. P. Carlin & A. E. Gelfand (2004)

Hierarchical Modelling and Analysis for Spatial Data.

Boca Ration, FL: Chapman & Hall/CRC.

References

T. M. Luong, V. Perduca & G. Nuel (2012)
Hidden Markov Model Applications in Change-Point Analysis
arXiv:1212.1778

R. Killick and I. A. Eckley (2014)
changepoint: An R Package for Changepoint Analysis
Journal of Statistical Software 58(3) 1-19

A. Lokka (2007)
Detection of Disorder Before an Observable Event
Stochastics: An International Journal of Probability and Stochastic Processes, Volume 79, Issue 3-4, 2007

References

Alexander, L. V., et al. (2006)

Global observed changes in daily climate extremes of temperature and precipitation.

Journal of Geophysical Research: Atmospheres 111.D5

https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2005JD006290