An introduction to spatial risk assessment

Peter F. Craigmile

http://www.stat.osu.edu/~pfc/

Statistics for Environmental Evaluation
Quantifying Environmental Risk and Resilience Workshop
Glasgow, Scotland
5-8 January 2016

Introduction

• In traditionally risk assessments, we investigate the impact of risk factors upon some environmental variable of interest.

  – For example, we could use a regression model to understand the relationship between
  PM$_{10}$ (particulate matter less than 10 microns) (the response, $Y$)
  and
  meteorology and sources of pollution (the covariates, $x$).

• It gets more interesting if we observe the data over space.
(similarly time.)

Acknowledgments

• Thank you to NERC and The School of Mathematics and Statistics at the University of Glasgow.

• Some of my research is funded by the US National Science Foundation, NSF SES-1424481 and NSF DMS-1407604.

Application 1: Modelled $PM_{10}$ data over Glasgow
(with Katie Allison, Scottish Government and Marian Scott, U. Glasgow)

Modeled annual mean log $PM_{10}$ on a 1km × 1km grid over Glasgow obtained from the Scottish Air Quality website.

Incorporates background measures, traffic, and point sources of pollution.
Application 2: Arsenic in soils
(with Kate Calder, The Ohio State University and Jiang Zhang, Freddie Mac)

- The USSoils dataset contains measurements of over 46 compounds, analyzed at the USGS laboratories in Denver, CO [Boerngen and Shacklette, 1981].
  
  http://tin.er.usgs.gov/ussoils/

  - We considered Arsenic concentration in topsoils across the U.S. EPA Region 5 (Illinois, Indiana, Michigan, Minnesota, Ohio, Wisconsin).
  - Units are µg/m³, or parts per million (ppm).
  - We restrict our analysis to the surface layer, the so-called A horizon, corresponding to a depth of around 20cm.
  - No minimum detection limits for Arsenic (As).
  - The measurements range from 0.89–27.00 ppm, median is 4.8 ppm.

Application 3: Chesapeake Bay nitrogen levels
(With Noel Cressie, U. of Wollongong; Thomas Santer, Ohio State; Youlan Rao, former student)

- Interest in studying the nitrogen content of the Chesapeake Bay in the Mid-Atlantic region of the US. The watershed encompasses 6 states.

- Increased nitrogen levels has a wide ecological impact on the bay and surrounding areas. Cleanup of the bay is a non trivial exercise.

- Dataset: 49 measurements of the total surface nitrogen concentration (in mg/L) taken from the main bay area in May 1995.

- Question of interest: Describe the exceedance areas in the nitrogen levels for the purposes of characterization and/or remediation.
Exceedance regions – the scientific problem

- Suppose there is interest in picking out areas of a spatial domain for which a process exceeds some ‘value’.
  - We call these areas exceedance regions (sometimes called hotspots).
- This has application to quantifying risk and environmental remediation.
- A secondary question is “what should the exceedance value be?”

A naive approach

1. Build a spatial model for the data.
2. Given the model and data, produce predictions of the process at different locations in the area of interest.
3. Using these predictions, form an exceedance region by identifying those areas that exceed the given threshold value.

(We will discuss some issues with this approach towards the end of this lecture.)

Geostatistics, according to Diggle et al. [1998]

(making a change in their notation)

“Conventional geostatistical methodology solves the problem of predicting the realized value of a linear functional of a Gaussian spatial stochastic process $Y(s)$ based on observations $Z(s_i) = Y(s_i) + \epsilon(s_i)$ at sampling locations $s_i$, where the $\epsilon(s_i)$ are mutually independent, zero-mean Gaussian random variables.”

- Clearly there are other reasons why we want to carry out a geostatistical analysis.

A hierarchical (geo)statistical model for point referenced data

- Let the domain, $D$, be a bounded subset of $\mathbb{R}^p$.
- Let $\{Y(s) : s \in D\}$ be a geostatistical stochastic process observed in $D$.
- Suppose we observe data $Z(s_i)$ at locations $s_i \ (i = 1, \ldots, n)$
- Traditional hierarchical approach:
  1. Write down a model describing the distribution of $Z(s)$ given $Y(s)$.
     (The data model)
  2. Write down the distribution of $Y(s)$.
     (The process model)
Possible data models

1. \( Z(s_i) \) are independent \( N(Y(s_i), \sigma^2) \).

   Generalizations: Letting the variance vary over space to capture known non-stationarities in the measurement collection process (e.g., in remote sensing).

2. More generally, \( Z(s_i) \) given \( Y(s_i) \) is drawn from an exponential family where \( g(E(Z(s_i))) = Y(s_i) \) for some link function \( g(\cdot) \).

   Examples of exponential families include binomial, Poisson, exponential, gamma.

[See e.g., Diggle et al., 1998, for more details].

Assuming stationarity

- The process \( Y(s) \) is (weakly) **stationary** if
  \[ \mu_Y(s) = \mu_Y, \]
  for all locations \( s \in D \) and
  \[ C_Y(s, s') = C_Y(s - s'), s, s' \in D. \]

- Further, our spatial process is **isotropic** if
  \[ C_Y(s, s') = C_Y(||s - s'||), \]
  for some norm \( ||\cdot|| \). A stationary process is **anisotropic** if it is not isotropic.

- Examples of stationary covariance functions: exponential, Gaussian, Matérn.

The latent geostatistical process \( Y \)

- Consider a Gaussian process model for \( \{Y(s) : s \in D\} \), with finite variance.

  (Remember a Gaussian process is a process which is jointly normal at any set of locations.)

- Our process is characterized by the **mean function**
  \[ \{\mu_Y(s) = E(Y(s)) : s \in D\} \]
  and the **covariance function**
  \[ C_Y(s, s') = \text{cov}(Y(s), Y(s')), \]
  a symmetric, non-negative definite function.

- We can also use the **variogram**, \( 2\gamma_Y(s, s') = \text{var}(Y(s) - Y(s')) \), as a measure of spatial dependence.

The allure of stationarity

- Advantages:

- Disadvantages:
The simplest nonstationarity – nonstationarity in the mean

- Simplest model: For covariates \( \{x_k(s) : s \in D, k = 1, \ldots, K\} \) let
  \[
  \mu_Y(s) = \beta_0 + \sum_{k=1}^{K} \beta_k x_k(s).
  \]
- The covariates could just be functions of the spatial locations.
- More generally they could be basis functions (e.g., wavelets, thin-plate splines).
- See, e.g., Kent and Mardia [1994] and Nychka [2000] for a discussion of the link between thin-plate splines and spatial prediction ("kriging").
- Simplest changepoint models for the mean involve covariates with spatial step functions (e.g., the Haar wavelet).

Nonstationary covariance structures

- Currently an area of very active interest in spatial statistics, especially in the context of massive datasets.
- Key issue: The spatial covariance function, \( C_Y(\cdot, \cdot) \), must be symmetric and non-negative definite.
- Can be tricky to guarantee non-negative definiteness, unless a constructive approach is taken.
- See http://u.osu.edu/sses/autumn-2014/ for a nice review of nonstationary covariance functions.

Chesapeake Bay exploratory analysis

- Trimodal distribution of log concentrations.
- Increasing trend from south to north.
- Evidence of anisotropy to the data.
  \( \rightarrow \) Resolved if we scale the easting direction by 2.5.
- Oscillatory pattern to the spatial correlations.
  \( \rightarrow \) Possible due to the mixing of three different processes (north, middle and south) in Chesapeake Bay.
  \( \rightarrow \) As a first cut, we model the relationship close to the origin with an exponential variogram function.
The data and process of interest

- Suppose \( \{Y(s) : s \in D\} \) is a random field model for the true total nitrogen concentrations. We assume \( Y(s) > 0 \) and \( D \subset \mathbb{R}^2 \).
- We observe these concentrations with measurement error. Let the observed process be \( \{Z(s) : s \in D\} \).
- Let \( Z = (Z(s_1), \ldots, Z(s_n))^T \) denote the vector of observed concentrations at the sites \( s_1, \ldots, s_n \) (here \( n = 49 \)).
- Let \( Y = (Y(s_1), \ldots, Y(s_n))^T \) be the true nitrogen concentration at those sites.

A data and process model for concentrations

- Assume the data model is
  \[
  \log Z | \log Y, \sigma^2 \sim N_n(\log Y, \sigma^2 I_n).
  \]
  \( (N_n(m, V) \) denotes a \( n \)-variate normal distribution with mean \( m \) and covariance \( V \)).
- Then the process model is
  \[
  \log Y | \beta, \theta \sim N_n(X\beta, C_\theta),
  \]
  where \( X \) is some design matrix and \( C_\theta \) is a matrix which models the spatial covariances between sites. From the exploratory analysis
  - the design matrix represents a linear slope from south to north.
  - \( C_\theta \) is associated with an exponential covariance kernel.

The likelihood

- The likelihood for observed log concentrations is
  \[
  f(\log z | \sigma^2, \beta, \theta) = \left(2\pi\right)^{-n/2} \det(\Sigma|)^{-1/2} \exp\left\{-\frac{1}{2} (\log z - X\beta)^T \Sigma^{-1} (\log z - X\beta)\right\},
  \]
  where \( \Sigma = C_\theta + \sigma^2 I_n \).
- Straightforward to obtain the maximum likelihood (ML) estimates (could also use restricted maximum likelihood (REML) here).

Game plan

1. We model the true nitrogen concentrations using the available data.
2. We predict the true concentrations at a set of sites of scientific interest.
3. We determine those locations which exceed the spatial process of true nitrogen concentrations by a certain amount.
Predicting the log nitrogen concentrations

- Based on the maximum likelihood estimates we can use universal kriging [Cressie, 1993, p.361] to predict the log nitrogen concentrations at a given set of sites.
  - The practicum contains the R code.
  - We will discuss the mathematics in the next lecture.

- Problem: the predictions do not account for the variability in the parameter values!
  (We could take a Bayesian approach.)

Defining an exceedance set, and naively estimating it

- Let $B$ be some fixed region which is a subset of the region of interest, $D$.
- Let $C$ be some pre-specified cutoff.
- Then the exceedance set associated with this cutoff is
  \[ e_{B,C} = \{ s \in B : I(Y(s) \geq C) = 1 \} . \]
- The problem is that we do not observe the actual $Y$ process.
- A naive estimate is to plug-in the kriging estimate, $\hat{Y}(s)$:
  \[ \hat{e}_{B,C} = \{ s \in B : I(\hat{Y}(s) \geq C) = 1 \} . \]
What is wrong with the naive estimate?

The spatial cumulative distribution function (CDF)

- How do we set the cutoff probabilistically?
- Again let $B$ be some fixed region which is a subset of $D$.
- For some $y \in \mathbb{R}$ the spatial CDF is defined by
  \[
  F_B(y) = \frac{\int_B I(Y(s) \leq y) \, ds}{|B|},
  \]
  where $I(\cdot)$ denotes the indicator function and $|B|$ denotes the area of $B$.
- We can then define the inverse spatial CDF as
  \[
  F_B^{-1}(p) = \inf\{u \in \mathbb{R} : F_B(u) \geq p\}.
  \]
- For some probability $p$ let the cutoff $C = F_B^{-1}(p)$.

Conclusions

- Hierarchical statistical models are useful for modeling complicated processes
- we will talk more about this in later lectures.
- Evaluation of exceedance sets can depend on the model, and can be defined in a probabilistic way.
- For a more formal technique for estimating the exceedance set and threshold jointly, see Zhang et al. [2008].
- See French and Sain [2013] for a spatio-temporal extension of this work.
References


