Spatial Models: A *Quick* Overview
Astrostatistics Summer School, 2018

Murali Haran

Department of Statistics
Penn State University

May 21, 2018

What this tutorial will cover

- I will explain why spatial models may be useful to scientists in many disciplines.
- I will outline types of spatial data and some basic concepts.
- The idea is to give you enough information so you know when you might have a spatial data problem and where you could look to find help.

What are spatial data?

- Data that have locations associated with them.
- Assumption: their locations are important in how we interpret and analyze the data. The locations themselves may also be central to the scientific questions of interest.
- Dependence is modeled as a function of distance between points, often dependence (or correlation) between data decreases with distance. Can also model process as being attractive (or repulsive) so presence of a data point increases (or reduces) the probability of another data point appearing nearby.

Some reasons to use spatial models

- Fitting an inappropriate model for the data, say by ignoring dependence, may lead to incorrect conclusions. e.g. underestimated variances
- Can lead to superior estimators (e.g. lower mean squared error).
- Sometimes learning about spatial dependence is central to the scientific questions. e.g. when finding spatial clusters, regions of influence/dependence.

Types of Spatial Data

There are three main categories of spatial data (though it is not always obvious how to classify data into these categories):

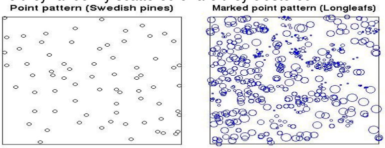
- **Spatial point processes**: When a spatial process is observed at a set of locations and the locations themselves are of interest. e.g. galaxies in space
- **Geostatistical data**: When a spatial process that varies continuously is observed only at a few points e.g. mineral concentrations at various drilling locations
- **Lattice data**: When a spatial process is observed on a regular or irregular grid. Often this arises due to aggregation of some sort, e.g. averages over a pixel in an image

Spatial Point Process Data examples

Locations of pine saplings in a Swedish forest.

Location, diameter of longleaf pines (*marked* point process).

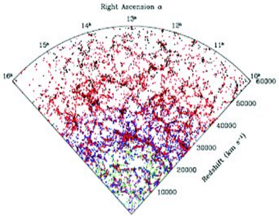
Are they randomly scattered or are they clustered?



(from Baddeley and Turner `R` package, 2006)

The galaxy distribution: 3D spatial point process

2d location in sky, 1d from redshift as a surrogate for distance.

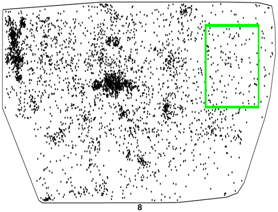


(Tegmark et al. 2004, The three-dimensional power spectrum of galaxies from the Sloan Digital Sky Survey,

Astrophys. J. 606, 702-740)

Shapley concentration: 2D spatial point process

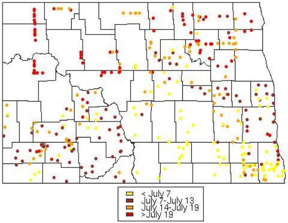
Nearby rich supercluster of galaxies. Several thousand galaxy redshift measurements. Galaxies show statistically significant clustering on small scales (Baddeley, 2008)



Geostatistical (point-referenced) data examples

Spatial analogue to continuous-time time series data.

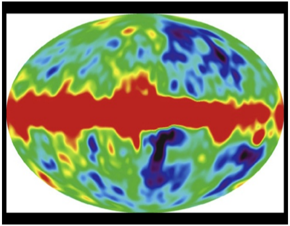
Wheat flowering dates by location (below):



Courtesy Plant Pathology, PSU and North Dakota State.

CMB

Cosmic microwave background



(ESA and the Planck Collaboration: tiny temperature fluctuations that correspond to regions of slightly different densities, representing the seeds of all future structure: the stars and galaxies of today)

Lattice data example

Spatial analogue to discrete-time data, e.g. images

http://www.scantips.com

Types of Spatial Data

- **Spatial point processes**
- Geostatistical data
- Lattice data

Spatial Point Processes: Introduction

- **Spatial point process**: The *locations* where the process is observed are random variables, process itself may not be defined; if defined, it is a **marked spatial point process**.
- Some stochastic mechanism generates the locations/point pattern. Based on the observed locations, we want to learn about the underlying mechanism
- **Observation window**: the area where points of the pattern can be observed. Important since absence of points in a region within observation window is informative.

Some questions

- What kind of attraction/repulsion exists in the process?
- Is there regular spacing between locations or do locations show a tendency to cluster?
- Does the probability of observing the event vary according to some factors? (Need to relate predictors to observations in a regression type setting.)
- Pattern arose through spread mechanism? e.g. clustering of 'offspring' near 'parents'?
- Can we estimate the overall count from only partial observations?
- Interest in measurements associated with points ("marked patterns")? e.g. diameter of trees, magnitude of galaxies

Some definitions for spatial point processes

- A spatial point process is a stochastic process, a realization of which consists of a countable set of points $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ in a bounded region $S \in \mathbb{R}^2$
- The points $\mathbf{s}_i$ are called **events**
- For a region $A \in S$, $N(A) = \#(\mathbf{s}_i \in A)$, $N(A)$ is random
- The **intensity measure** $\Lambda(A) = E(N(A))$ for any $A \in S$.
- If $\Lambda(A)$ can be written as

$$\Lambda(A) = \int_A \lambda(\mathbf{s}) d\mathbf{s} \quad \text{for all A} \in S,$$

then $\lambda(\mathbf{s})$ is called the **intensity function**.

Stationarity and isotropy

- The process is **stationary** if its distribution is invariant to translation in space
- The process is **isotropic** if its distribution is invariant to rotation in space
- Hard to assess based on having only a single realization of the process (which is typically the case): stationary process can look non-stationary (or vice-versa) within bounded window

Intensity of Poisson point process

Let $d\mathbf{s}$ denote a small region containing location $\mathbf{s}$.

- First-order intensity function of a spatial point process:

$$\lambda(\mathbf{s}) = \lim_{d\mathbf{s}\to 0} \frac{E(N(d\mathbf{s}))}{|d\mathbf{s}|}.$$

- Second-order intensity function of a spatial point process:

$$\lambda^{(2)}(\mathbf{s}_1, \mathbf{s}_2) = \lim_{d\mathbf{s}_1\to 0} \lim_{d\mathbf{s}_2\to 0} \frac{E\{N(d\mathbf{s}_1)N(d\mathbf{s}_2)\}}{|d\mathbf{s}_1||d\mathbf{s}_2|}.$$

- Covariance density of a spatial point process

$$\gamma(\mathbf{s}_1, \mathbf{s}_2) = \lambda^{(2)}(\mathbf{s}_1, \mathbf{s}_2) - \lambda(\mathbf{s}_1)\lambda(\mathbf{s}_2).$$

Spatial point process modeling

Spatial point process models can be specified by :

- A deterministic intensity function
- A random intensity function
- Major classes of models:
  - Poisson Processes: models for no interaction patterns
  - Cox processes: models for aggregated patterns
  - Inhibition processes: models for interacting patterns
  - Markov processes: models for attraction and/or repulsion
- Poisson process: basis for exploratory tools and constructing more advanced point process models.

Poisson Process

Poisson process on $\mathbf{X}$ defined on S with intensity measure Λ and intensity function λ, satisfies for any bounded region $B \in S$ with $\Lambda(B) > 0$:

1. $N(B) \sim$Poisson($\Lambda(B)$).
2. Conditional on $N(B)$, event locations in B are *independent* and distributed according to pdf proportional to $\lambda(\mathbf{s})$.

- **Homogeneous Poisson process**: The intensity function, $\lambda(\mathbf{s})$ is constant for all $\mathbf{s} \in S$.
- **Non-homogeneous Poisson process**: $\lambda(\mathbf{s})$ deterministic, varies with $\mathbf{s}$. Example: model $\lambda(\mathbf{s})$ as a function of spatially varying covariates

Cox Process

Also called *doubly stochastic Poisson process* (Cox, 1955)

- Natural extension of a Poisson process: Consider the intensity function of the Poisson process as a realization of a random field. We assume $\Lambda(A) = \int_A \lambda(\mathbf{s})d\mathbf{s}$.
  - Stage 1: $N(A)|\Lambda \sim$Poisson($\Lambda(A)$).
  - Stage 2: $\lambda(\mathbf{s})|\Theta \sim f(\cdot;\Theta)$ so that λ is stochastic, a nonnegative random field parametrized by Θ.

Markov Point Processes

- Point patterns may require a flexible description that allows for the points to interact. Simple: inhibition processes
- Markov point processes are models for point processes with interacting points (attractive or repulsive behavior can be modeled).
- 'Markovian' in that intensity of an event at some location **s**, given the realization of the process in the remainder of the region, depends only on information about events within some distance of **s**.
- Origins in statistical physics, used for modeling large interacting particle systems.

Applications to data

- We now have a framework for thinking about spatial point processes.
- Starting point/exploratory data analysis
  - test for complete spatial randomness (Poisson model) say via distance methods, quadrant count method etc. (established literature on point processes)
  - Estimate the intensity function, starting with assuming constant intensity (easy: total count/area of observation)
  - Estimate second order properties: K function, paired correlation, two-point correlation
- Based on results above, possibly fit a model

Exploring homogeneity/clustering

Assume process is stationarity (same intensity everywhere) and isotropic (direction/rotations do not affect the process)

$K(d) = \frac{1}{\lambda}E(\text{number of events within distance d of an arbitrary event})$.

- If process is clustered: Each event is likely to be surrounded by more events from the same cluster. $K(d)$ will therefore be *relatively large* for small values of d.
- If process is randomly distributed in space: Each event is likely to be surrounded by empty space. For small values of d, $K(d)$ will be *relatively small*.

Can obtain an intuitive estimator for $K(d)$ for a given data set.

Ripley's *K* Function

Let λ be the intensity of the process.

- Effective method for seeing whether the processs is completely random in space.

  $K(d) = \frac{\text{Mean number of events within distance d of an event}}{\lambda}$
- This can be estimated by

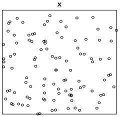
  $$\hat{K}(d) = \frac{\sum_{i \neq j} w_{ij} I(d_{ij} \leq d)}{\hat{\lambda}}$$

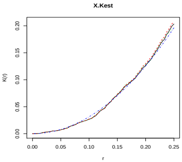
  where $\hat{\lambda} = N/|A|$ with $|A|$ as the total area of the observation window and N is the observed count.
- Note: K can also be viewed as an integral of the two point correlation function as used by astronomers (cf. Martinez and Saar, 2002).

Ripley's K for homogeneous Poisson Process

Process was simulated with intensity function $\lambda(x, y) = 100$.

homogeneous Poisson Process   Ripley's K





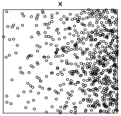
blue=K function under complete spatial randomness

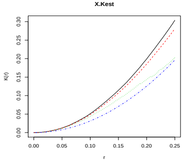
black (and red and green) are various versions of estimates of the K function

Ripley's K for inhomogeneous Poisson Process (Eg.1)

Process was simulated with intensity function $\lambda(x, y) = 100\exp(3x)$.

Inhomogeneous Poisson Process   Ripley's K





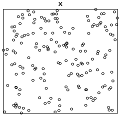
blue=K function under complete spatial randomness

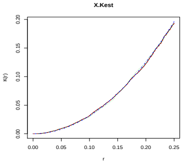
black (and red and green) are various versions of estimates of the K function

Ripley's K for inhomogeneous Poisson Process (Eg.2)

Process was simulated with intensity function $\lambda(x, y) = 100\exp(y)$.

Inhomogeneous Poisson Process   Ripley's K





blue=K function under complete spatial randomness

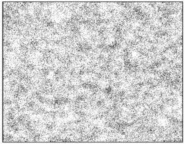
black (and red and green) are various versions of estimates of the K function

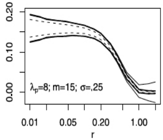
Exploring non-homogeneity

- A common way to study spatial point processes is to compare the realization of the process (observations) to a homogeneous Poisson process. This kind of exploratory data analysis or hypothesis test-based approach can be a useful first step.
- Estimating errors on the 2-point correlation or K function has been derived for a random Poisson process (Ripley 1988; Landy & Szalay 1993) or, if a model for the underlying process is known, from a parametric bootstrap (Einsenstein et al. 2005).
- But Poisson errors may be too small for spatially correlated samples.

Exploring non-homogeneity: recent developments

Loh (2008) recommends a ‘marked point boostrap’ resampling procedure. Figures below show a simulated clustered process, and the resulting 2-point correlation function with Poisson (dashed) and bootstrap (solid) 95% confidence error bands respectively.

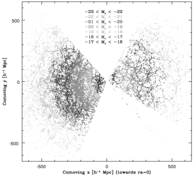




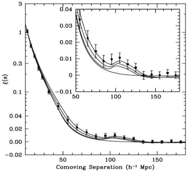
Loh 2008, A valid and fast spatial bootstrap for correlation functions, Astrophys. J. 681, 726-734

Example: Galaxy clustering (Sloan Digital Sky Survey)

Galaxy distribution



Two-point correlation function



Distribution of 67,676 galaxies in two slices of the sky showing strong anisotropic clustering (Tegmark et al. 2004).
Bottom: Two-point correlation function showing the faint feature around 100 megaparsec scales revealing cosmological Baryonic Acoustic Oscillations (Eisenstein et al. 2005).

Inference

- So far exploratory data analysis. Powerful but full inference with a model may provide richer set of tools and scientific conclusions.
- Recently developed algorithms and software (`spatstat in R`) make it easier to fit at least relatively simple or standard point process models. More advanced models need more specialized code

Spatial point processes: computing

- R command: `spatstat` function `ppm` fits models that include spatial trend, interpoint interaction, and dependence on covariates, generally using MPL.
- Maximum pseudolikelihood (MPL) often works well (Baddeley, 2005) but can work very poorly when there is strong dependence
- More rigorous but more complicated: Markov chain Maximum Likelihood (cf. C.J.Geyer’s chapter in “MCMC in Practice”, 1996)
- Bayesian models are becoming more common but not much software available

Types of Spatial Data

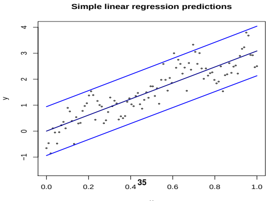
- Spatial point processes
- **Geostatistical data**
- Lattice data

Continuous-domain/geostatistical data

- We will now talk briefly about models for continous-domain spatial data: useful for interpolation and for regression with dependent data.
- Important non-spatial use in the context of astronomy: approximating complex computer models (that may take a long time to run) by probabilistic interpolation across computer model runs at a few parameter settings. Given how the computer model behaves at a few sets of inputs (parameters), approximate how the model will behave at other input settings: "Gaussian process emulation".

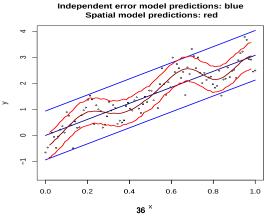
The importance of dependence (contd.)

Toy example: simple linear regression with the correct mean but assuming iid error structure. $Z(s_i) = \beta s_i + \epsilon_i$, where ϵ_is are iid. Does not capture the data/data generating process well even though trend (β) is estimated correctly.

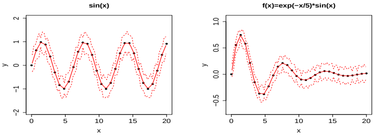


The importance of dependence (contd.)

Model: linear regression with correct mean, now assuming dependent error structure. This picks up the 'wiggles'. Independent error model: blue. Dependent error model: red.



Fitting complicated mean structures



Functions: $f(x) = \sin(x)$ and $f(x) = \exp(-x/5)\sin(x)$.
Same model used both times: $f(x) = \alpha + \epsilon(x)$, where $\{\epsilon(x),\ x \in (0, 20)\}$ is a Gaussian process, α is a constant.
Note:the dependence is being introduced to indirectly capture the non-linear structure, not to model dependence per se.

Spatial (linear) model for geostatistics and lattice data

- Spatial process at location $\mathbf{s}$ is $Z(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s})$ where:
  - $\mu(\mathbf{s})$ is the mean. Often $\mu(\mathbf{s}) = X(\mathbf{s})\beta$, $X(\mathbf{s})$ are covariates at $\mathbf{s}$ and β is a vector of coefficients.
- Model dependence among spatial random variables by imposing it on the errors (the $w(\mathbf{s})$'s).
- For n locations, $\mathbf{s}_1, \ldots, \mathbf{s}_n$, $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^T$ can be jointly modeled via a zero mean Gaussian process (GP), for geostatistics

Gaussian Processes

- Gaussian Process (GP): Let Θ be the parameters for covariance matrix $\Sigma(\Theta)$. Then:
$$\mathbf{w}|\Theta \sim N(0, \Sigma(\Theta)).$$
  This implies:
$$\mathbf{Z}|\Theta, \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \Sigma(\Theta))$$
- We have used the simplest multivariate distribution (the multivariate normal). We will specify $\Sigma(\Theta)$ so it reflects spatial dependence.
- Need to ensure that $\Sigma(\Theta)$ is positive definite for this distribution to be valid, so we assume some valid parametric forms for specifying the covariance.

Gaussian Processes: Example

- Consider the popular **exponential** covariance function.
- Let $\Sigma(\Theta) = \kappa I + \psi H(\phi)$ where I is the $N \times N$ identify matrix. Note that $\Theta = (\kappa, \psi, \phi)$ and $\kappa, \psi, \phi > 0$.
- The i, jth element of the matrix H, $H(\|\mathbf{s}_i - \mathbf{s}_j\|; \phi)_{ij} = \exp(-\phi\|\mathbf{s}_i - \mathbf{s}_j\|)$.
- Note: covariance between i, jth random variables depends only on distance between $\mathbf{s}_i$ and $\mathbf{s}_j$, and does not depend on the locations themselves (implying *stationarity*) and only depends on the magnitude of the distance, not on direction (implying *isotropy*).
- Extremely flexible models, relaxing these conditions, can be easily obtained though fitting them can be more difficult.

Gaussian Processes: Inference

- The model completely specifies the likelihood, $\mathcal{L}(\mathbf{Z}|\Theta, \boldsymbol{\beta})$.
- This means we can do likelihood-based inference:
  - Estimation using maximum likelihood (MLE) or Bayes (if we place priors on $\Theta, \boldsymbol{\beta}$)
  - Prediction using plug-in MLE or posterior predictive (Bayes).

Gaussian Processes: Computing

- For likelihood based inference: `R`'s `geoR` package by Ribeiro and Diggle.
- For Bayesian inference:
  - `R`'s `spBayes` package by Finley, Banerjee and Carlin.
  - `WINBUGS` software by Spiegelhalter, Thomas and Best.
- Very flexible packages: can fit many versions of the linear Gaussian spatial model. Also reasonably well documented.
- Warning: With large datasets ($>$1000 data points), matrix operations (of order $O(N^3)$) become very slow. Either need to be clever with coding or modeling. Above software (except spBayes in some cases) will not work.

Models for lattice data

- We have not discussed lattice data models here.
- Worth noting that lattice models like Gaussian Markov random fields may have computationally advantages (due to sparse matrices) and hence may be useful for continous-domain data
- `GeoDa` package at `https://www.geoda.uiuc.edu/`(free) by Luc Anselin
- `R`'s `spdep` package by Roger Bivand et al.
- Bayesian inference: `WINBUGS` includes `GeoBUGS` which is useful for fitting such models.
- `INLA` by H. Rue and co-authors

Useful ideas for non-spatial data

Some spatial modeling techniques may be useful in non-spatial scenarios:

- Gaussian processes: Useful for modeling complex relationships of various kinds. Examples: flexible nonparametric regression, classification. see Rasmussen and Williams (2005) online book
- Fast approximations for complex computer models.
- Ideas for modeling time series, particularly multivariate time series.

Summary: spatial data types and associated models

General spatial process: $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$, D is set of locations.

- **Spatial point process**: $D = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ is a random collection of points on the plane. Ordinarily $Z(\mathbf{s})$ does not exist. For marked point process, $Z(\mathbf{s})$ is a random variable. Usual (basic) models: Poisson process, Cox process.
- **Geostatistics**: D is a fixed subset of $\mathbb{R}^2$ (or $\mathbb{R}^3$ in 3D case). $Z(\mathbf{s})$ is a random variable at each location $\mathbf{s} \in D$. Usual (basic) model: Gaussian process.
- **Lattice data**: $D = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ is a fixed regular or irregular lattice, on $\mathbb{R}^2$ (or $\mathbb{R}^3$). $Z(\mathbf{s})$ is a random variable at each location $\mathbf{s} \in D$. Usual (basic) model: Gaussian Markov random field.

References: Geostatistics and Lattice Processes

Geostatistics and Lattice Data:

- Schabenberger and Gotway (2005) "Statistical Methods for Spatial Data Analysis". A fairly comprehensive easy to read book on spatial models for (in order of emphasis): geostatistics, lattice data and point processes.
- Cressie (1994) "Statistics for Spatial Data". This is a comprehensive guide to classical spatial statistics, but it is considerably more technical than the other two references listed here.
- S. Banerjee, B.P. Carlin and A.E. Gelfand (2004) “Hierarchical Modeling and Analysis for Spatial Data”. This is a textbook on Bayesian models for spatial data.

References: Spatial Point Processes

- Møller and Waagepetersen review article “Modern spatial point processes modelling and inference (with discussion)” (Scandinavian Journal of Statistics, 2007) or online chapter: http://people.math.aau.dk/~jm/spatialhandbook.pdf
- Baddeley and Turner’s `R spatstat` package.
- Baddeley et al. “Case Studies in Spatial Point Process Modeling” (2005).
- P.J.Diggle’s online lecture notes: http://www.maths.lancs.ac.uk/~diggle/spatialepi/notes.ps

References: Spatial Point Processes

- P.J.Diggle “Stat Analysis of Spatial Point Patterns” (2003)
- “Modern statistics for spatial point processes” by J.Møller and R.P.Waagepeterson (2004).
- V.J.Martinez and E.Sarr “Statistics of the Galaxy Distribution.”

Acknowledgments:

- Much of the material and examples in this tutorial were drawn from several of the listed references, Ji-Meng Loh’s notes for the Penn State astrostatistics tutorial in 2013, and from Eric Feigelson.