Midterm Study Guide

Contents

1. Chapter 5 - Analysis of Point Patterns

Chapter 5 sets up the fundamentals of spacial data analysis. We look at concepts such as **clustering** and **regularity** and some methods of identifying them. In this chapter we primarily think of these with respect to some behavior that is, for lack of better wording, totally random, to serve as our baseline.

With this baseline we can look at the behavior of observed points in space and see how that behavior compares to **complete spatial randomness**.

Basic Definitions

Definition 0.1 — **Point**

Any location where an event could occur

Definition 0.2 — **Event**

A location where an event did occur.

Definition 0.3 — **Point Pattern Data**

Consistents of a collection of observed event locations and a spatial domain of interest.

NOTE: The spatial domain of interest is super important. All results are with respect to it, different spatial domains can reach different conclusions.

CSR and Stochastic Processes

Definition 0.4 — **Stochastic Process**

A Collection of random variables

Definition 0.5 — **Spatial Point Process**

A stochastic process where each RV is the location of an event.

Definition 0.6 — **SPP Realization**

A collection of locations generated under the spatial point process model.

- This is regardless of whether we know what that model is.

Definition 0.7 — **Complete Spatial Randomness**

CSR is a situation where, given a spatial domain of interest, all points are equally likely to produce an event.

Events are independent and uniformly distributed.

Regularity and Clustering

Definition 0.8 **Regularity**

Points are consistently spaced apart from one another, they aren't right on top of eachother. Think houses in a neighborhood.

Definition 0.9 **Clustered Data**

Events are **clustered** when they occur more frequently near one another than one would expect under certain assumptions.

Simulating CSR Data

2-stage approach for simulating a realization of CSR in a study area D.

1. Generate a total number of points, $N(D)$, from a Poisson distribution with mean $\lambda\ |D|$
2.
   - if D is rectangular, we may generate u and v coordinates using uniform random number generators on the intervals corresponding to the width and height of D, respectively.
   - if D is NOT rectangular, things get a bit funkier. One option is to embed D within a larger rectangle R, and generate event locations uniformly in R until $N(D)$ events occur within D.

So basically, use the poisson distribution to generate the number of points. Use uniform distribution to generate the point coordinates.

Monte Carlo Testing

General gist here is that we can't do something as simple as compare directly against a distribution for these hypothesis tests. We instead use simulation as CSR is easy to approximate and we can then, in turn, simulate the test statistic distribution.

So we have to generate fake datasets to build up enough of a collection of simulated test statistics to approximate its distribution.

Procedure:

1. Calcuate T for the observed data. T_{obs}.
2. Generate a ton of simulated data sets assuming CSR. N_{sim}.
3. Calc the test statistic for each simulated dataset.
4. Count the number of all test statistics (including observed) that are large or larger than the observed test statistic. Denote this l.
5. The estimated p-value is:

$$\hat{Pr}[T \geq T_{\text{obs}} \mid H_0 \text{ is true}] = \frac{l}{N_{\text{sim}} + 1}$$

When CSR is unrealistic

Sometimes we don't expect stuff to follow CSR. Say like, human populations across a state. Those are clustered in cities and such.

In many situations the **constant risk hypothesis** is more appropriate.

Definition 0.10 **Constant Risk Hypothesis**

Also used to assess "no clustering". Under CRH, every person has the same risk of disease during the observation period, regardless of location.

- Clusters of cases in high populations violate the CSR but not the CRH because we expect more cases in high population areas.

Instead of a homogeneous poisson process, we can think of this as a **heterogeneous poisson process** where the intensity depends on the population of an area.

First and 2nd order Properties

1st order: Mean or average. Example: Intensity function $\lambda(s)$. Kinda like 1st moment being the mean I would assume.

2nd order: pertains to variance. So just like moments then.

What does 2nd order mean in the context of spacial data? Shows how often events occur within a given distance of other events. It's well, the spread of data in space.

Spatial Density and Intensity

Interpretation

If we're talking interpretation are we talking like plots?

Estimating using Kernel Smoothing

Kernel functions much like it does in CNNs. Various methods of smoothing out nearby points to get an overall idea of the density.

These use weighted averages of nearby points to smooth things out.

Bandwidth: The variance of the kernel. A larger bandwidth smooths things out more, removing local variation in the data. A larger bandwidth is KINDA like using less bins in a histogram, everything gets lumped into the same bins.

Density Function: Defines the probability of observing an event at location s.

Intensity Function: Defines the expected number of events per unit area at location s.

These two functions only differ by a constant. They give the same info.

$$f(s) = \frac{\lambda(s)}{D}$$

Choosing the right kernel

Most give similar results, though kernels with finite support are a lot less of a pain computationally.

2. Case Control Point Data

3. Random Labeling Hypothesis

4. Log Relative Risk

5. K Functions

Most common form of 2nd order analysis

$$K(h) = \frac{E[\#\text{ of events within } h \text{ of a randomly chosen event}]}{\lambda}$$

Note:

$$K(h) = \text{var}(N(A))$$

These things are equivalent. How nice.

Under CSR,

$$K(h) = \frac{\lambda \pi h^2}{\lambda} = \pi h^2$$

We approximate $K(h)$ by

Definition, how to estimate

- Edge correction?

- How to interpret.
- Transformation to L functions
- How to construct pointwise envelopes for inference
- How to interpret pointwise envelopes for inference

Difference in K Functions

Other Concepts

Stationarity and Isotropy

We'll get back to these later

6. Geographical Analysis Machine

7. Spatial Scan Method

This ones like the circular scan method yeah? Take a point, expand a circle out from it. Each time we hit a new observation we create a new window. Track cases/non-cases. The behavior of these specific windows are how we determine if they're a possible cluster or not.

8. Q Nearest Neighbors

9. Regional Count Data

10. Constant Risk Hypothesis

11. Misc Stuff

What is the ecological fallacy? Modifiable areal unit problem? Scales of clustering that can be detected?

12. CEPP, Besag-Newell, Spatial Scan methods

- What are the test statistics
- What are the null and alternative hypotheses for each test?
- What conclusions can be drawn from the tests?