University of Colorado, Denver
MATH 6384 - Spatial Data Analysis
Fall 2025

Midterm Study Guide
Brady Lamson
06.11.2025

Midterm Study Guide

Contents

1. Summary Table

Test	Null Hypothesis	Test Statistic	Reference Hypothesis	Detects	Comments
K-function	CSR	$(K(h) - \pi h^2)$	CSR	Global clustering	Uses edge correction
L-function	CSR	$(L(h) - h)$	CSR	Global clustering	Simplifies $K(h)$
Log RR	RLH	$(r(s) = 0)$	RLH	Local clusters	Monte Carlo envelopes
Diff K	RLH	$(K_c - K_{\text{ctrl}})$	RLH	Global clustering	No location info
CEPP	CRH	Max cases in n^* pop	CRH	Regional cluster	Fixed population
Besag–Newell	CRH	Min pop for c^* cases	CRH	Regional cluster	Fixed cases

2. 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.

NOTE: A **cluster** is different from **clustering**.

- A **cluster** is a collection of cases inconsistent w/ the null.
- **Clustering** is when the data overall is clustered together w/ respect to the spatial domain of interest.

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. Provides local insight into where patterns differ ("There is a cluster here").

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. This order provides insight into global aspects of the data. Patterns of clustering or regularity but not specifically where ("This data has clustering in it").

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.

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 effectively averaging the number of events within distance h of each event in the data set. Basically, on average, how many neighbors does each event have?

Note that this undercounts near the boundary. This is where edge corrections come in. We use some weighting on the points near the boundary. This weight is proportional to the amount of the circle outside the study area. Makes sure stuff near the edge is able to provide equivalent information.

We transform over the **L function** for ease of comparisons to CSR. Basically we do this:

$$\hat{L}(h) = \sqrt{\hat{K}_{ec} \frac{h}{\pi}}$$

Why? Recall the value of $K(h)$ under CSR. This removes the square and the π. Makes stuff real freackin simple cause

$$\hat{L}(h) - h = 0$$

So we have a baseline to compare against!

If the L-function of our observed data deviates a ton from this we get evidence of clustering or regularity.

Okay so uh, $L(h) - h = 0$ but $L(h)$ is random. So how do we know how far from 0 we can be without a departure from CSR? Simulations woo woo woo. Basically we simulate a ton of datasets, collect the $L(h) - h$ for each of those datasets and create envelopes for each value of h. This gives us an idea of how much variance we should expect at each distance value.

We can create min/max envelopes, quartile envelopes, whatever.

Other Concepts

Stationarity and Isotropy

We'll get back to these later

3. Chapter 6 - Point Data for Cases and Controls

Case Control Point Data: What it says on the tin get real. Cases vs. non-cases for point data.

One important note that we assume the non-cases to be an independent random sample from subjects free of the disease of interest.

So that's why we compare against the control. How does the behavior of the cases differ from the controls?

control can mean a lot of different stuff depending on the context. It can be a different disease for instance. How does covid compare against the flu for example.

Relative Risk

We don't know actual risk of disease typically. We estimate it using the rates, or proportions, of those with the disease. When we're talking relative risk, we're basically look at the fraction of two risk rates and looking at relative increases or decreases.

We use rate ratios to estimate relative risk.

$$\frac{\frac{\text{\# of cases at s}}{\text{\# at risk at s}}}{\frac{\text{total \# of cases}}{\text{total \# at risk}}}$$

From here we can see what areas have way higher rate ratios beyond what we would expect by random chance.

The two approaches for tackling this are:

- **constant risk hypothesis** for regional count data
- **random labeling hypothesis** for point case-control data

Random Labeling Hypothesis

Pretty straightforward honestly. Is the behavior we're seeing really that different from if we just randomly assigned the case/non-case labels across points?

Assumes a constant probability of case-control assignments at all locations.

Constant Risk Hypothesis

Assumes cases reflect a random sample of the at-risk population where the probability of selection is the same everywhere.

Assumes a known background risk.

Really it's a similar thing to random labeling hypothesis. Does the proportion of cases/non-cases in a given region deviate that much from the overall behavior of the study area?

Different Scenarios to Consider

- S1: The random labeling hypothesis addresses the question: "Are the N_1 case locations observed consistent with a random assignment of N_1 events among all of the event locations?"
- S2: The constant risk hypothesis, with a **fixed** number of cases, addresses the same question as S1.
- S3: The constant risk hypothesis, with a **random** number of cases, addresses the question, "Are the case locations observed consistent with each of the N locations observed having probability $\frac{N_1}{N}$ of being a case?"

To be more precise:

- S2: Is there evidence of clustering of 592 leukemia cases among $1,057,673$ persons at risk?
- S3: Is there evidence of clustering of leukemia cases among $1,057,673$ persons at risk where each person has a probability of $\frac{592}{1057673}$ of contracting leukemia in the time interval under study?

In S2, we KNOW how many cases there are. We can randomly re-distribute them and see behavior. In S3, that info just gives us the probability of assigning the label. So we're treating the number of cases as random and comparing against that.

Log Relative Risk

With this we're comparing the density of cases at a location relative to all of the cases to the density of the controls relative to all of the controls.

So like

$$\frac{\text{density of cases}}{\text{density of controls}}$$

We take the log of this for computational purposes. Referring to the relative risk section, we want to see how much the relative risk changes at various locations. If this ratio is constant across locations, $r(s) = 0$. What we're interested in is the locations where $r(s) > 0$ and larger departures from 0 provide more and more evidence of a cluster at s.

Local Test - Checking for Clusters

So this **can** be a tool for identifying **clusters**.

We can check for these using monte carlo simulations and the random labeling hypothesis.

A significant p-value here indicates that, at location s, there is significant evidence of a local cluster of cases relative to controls.

We can do monte carlo envelopes here too and check each grid point. If $r(s)$ is larger than the $1 - \frac{\alpha}{2}$ quantile then the cases are clustered relative to controls at s. If its smaller than $\frac{\alpha}{2}$ we have clustering of controls relative to cases.

- Note, these conclusions are w/ respect to what we expect under the random labeling hypothesis.

Global Test - Checking for Clustering

Kelsall and Diggle have a global test.

$$H_0 : r(s) = 0 \forall s \in D$$
$$H_a : r(s) \neq 0 \text{ for at least one } s \in D$$
$$T = \int_D (r(u))^2 du$$

We do the same monte carlo stuff as always for this test. Simulate a ton of datasets, gather a ton of simulated test statistics. See how many are at least as extreme as observed.

Difference in K Functions

2nd order test stuff here. We compare K-functions for cases and controls. Gives us insight into spatial scales of clustering, but not exact locations of clustering. The test itself just tells us if there is any evidence of clustering at any of the given spatial scales.

$$KD(h) = K_{\text{case}}(h) - K_{\text{control}}(h)$$

$$H_0 : KD(h) = 0 \forall h \in [0, h^*]$$
$$H_a : KD(h) > 0 \text{ for at least one } h \in [0, h^*]$$

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.

$$H_0 : \text{There are no clusters of cases in the study area}$$
$$H_a : \text{There is at least one cluster of cases in the study area}$$

Pendantic version,

H_0: "There are no windows where the most likely cluster is more unusual than what is expected under the random labeling hypothesis"

H_a: "There is at least one window where the most likely cluster is more unusual than what is expected under the random labeling hypothesis"

There is ALWAYS a most likely cluster even if it isn't a cluster.

With this we can also identify secondary and so on likely clusters. It's just the 2nd, 3rd, 4th, etc highest test statistic for windows that dont overlap.

Q Nearest Neighbors

Checks patterns of cases near other cases. A cluster here is "wow thats more cases near eachother than what we would expect under the RLH".

This is the one with the large table showing stuff like $T_{q_2} - T_{q_1}$. The key thing here is "how much of the significance in T_{q_2} is explained by T_{q_1}".

We check which q scales are significant then check if those contrasts are also significant.

4. Chapter 7 - Regional Count Data

Basics

This kinda data is more anonymized. It's just data aggregated into some more general space like counties, states, etc. So you'd get like counts in a state or whatever.

We can do a lot of the same stuff here, comparing proportions in specific regions vs what we would expect there.

What is regional count data?

- What do you need to perform spatial analysis?

Ecological Fallacy: This occurs when associates between outcomes and potential risk factors observed in groups are extrapolated to individuals.

- Example: Breast cancer mortality rates being higher in countries w/ high fat consumption when compared w/ low fat consumption. The issue here is that with this type of data we don't know that the women who died from breast cancer also had high fat consumption.

Modified Areal Unit Problem: Occurs when association changes based on the way the data are grouped. Think clusters spread between multiple regions. You don't have as strong of evidence for a specific region here. You have strong evidence for region 1+2 but not. region 1 OR 2.

Spatial Scale Problems: How much is the data aggregated? This is about balancing information between really large and really small regions. How do we decide how big a region should be? If regions are way too small then any case may be immediately significant. If they're too big we lose a lot of local information. State vs county vs town and so on.

- Clustering at small scales doesn't imply clustering at larger scales. Same thing the other direction.

What does regional data need?

- A set of counts observed for each region
- Enough info to determine the counts expected for each region. Population counts and stuff.
- Region borders

Geographical Analysis Machine (GAM)

Can be used for both case-control point data and regional data. More for exploratory analysis than actual statistical inference.

Basic process:

1. Construct circles of various distances
2. Count cases and at risk people within the circles.
3. Calculate local incidence proportion
4. Display circles exceeding some proportion threshold
   - No reason this can't be flipped to show extremely low proportions as well.

The goal here is to help identify areas with unusually low or high rates.

Cluster Evaluation Permutation Procedure - CEPP

This is the method that constructs windows at the centroid of every region. We have a fixed number of persons at risk to count up to: n^*. If a region doesn't contain enough people, we then take people from the nearest region. We tend to end up taking fractions of persons-at-risk from these nearby regions.

Test statistic here is the maximum number of cases across ALL windows.

This is under the constant risk hypothesis.

H_0: There is no window with n^* persons-at-risk that has significantly more cases than what is expected under the constant risk hypothesis.

H_a: There is at least one winmdow with n^* persons-at-risk that has significantly more cases than what is expected under the constant risk hypothesis.

Same kinda thing here where we have most likely clusters, second most likely clusters, etc.

Besag and Newell Approach

Instead of fixed population per window, we have fixed cases per window.

Same thing different flavor. Looking for most compact window containing that number of cases.

H_0: The most compact window (in terms of population) with at least c^* cases is not significantly more compact than we would expect under the constant risk hypothesis.

H_a: The most compact window (in terms of population) with at least c^* cases is significantly more compact than what we would expect under the constant risk hypothesis.