Descriptive vs Statitical Point Pattern Analysis

Descriptive analysis:

- set of quantitative (and graphical) tools for characterizing spatial point patterns
- different tools are appropriate for investigating first- or second-order effects (e.g., kernel density estimation versus sample G function)
- can shed light onto whether points are clustered or evenly distributed in space

Limitation:

- no assessment of *how* clustered or *how* evenly-spaced is an observed point pattern
- no yardstick against which to compare observed values (or graph) of results

Statistical analysis:

- assessment of whether an observed point pattern can be regarded as one (out of many) realizations from a particular spatial process
- measures of confidence with which the above assessment can be made (how likely is that the observed pattern is a realization of a particular spatial process)

Some General Terminology

Null hypothesis:

- spatial process that is hypothesized to be the generating mechanism of spatial patterns
- in this class, we'll focus on the null hypothesis of complete spatial randomness (CSR)

Samples:

- realizations from the above process
- in this lecture, the set of alternative point patterns that could result from the null process of CSR

Sampling distribution of a statistic:

- *sample statistic* = a summary measure, e.g., mean or entire CDF, characterizing (and thus computed from) a sample
- *sampling distribution of a statistic* = distribution, e.g., histogram, of such a summary measure computed from many alternative samples generated from the null process

A Particular Example I

Null hypothesis:

- complete spatial randomness (CSR) as a mechanism for generating point patterns

Statistic:

- mean event-to-nearest-event (E2NE) distance;
  here the variable is the minimum distance between events (E2NE), and the selected summary statistic is the mean of those distances:

$$\bar{d}_{min} = \frac{1}{n} \sum_{i=1}^{n} d_{min}(\mathbf{s}_i)$$

$d_{min}(\mathbf{s}_i)$ = distance between i-th event and its nearest neighbor event

Sampling distribution of mean E2NE:

1. generate (simulate) one realization of a point pattern under CSR
2. compute $\bar{d}_{min}$ from that realization
3. repeat steps (1) and (2) many, say 1000, times
4. histogram of 1000 $\bar{d}_{min}$ values is the sampling distribution of mean E2NE distance *under the null hypothesis of CSR*

sampling distribution under CSR can often be analytically derived, without resorting to simulation

A Particular Example II

Two realizations under CSR of point patterns with $n = 50$ events:

$\bar{d}_{min} = 7.13$ $\bar{d}_{min} = 8.13$

Sampling distribution or histogram of $\bar{d}_{min}$ values from 500 simulated (under CSR) point patterns with $n = 50$ events

A Particular Example III

Two realizations under CSR of point patterns with $n = 100$ events:

$\bar{d}_{min} = 4.88$ $\bar{d}_{min} = 5.67$

Sampling distribution or histogram of $\bar{d}_{min}$ values from 500 simulated (under CSR) point patterns with $n = 100$ events

Looking at Observed Point Patterns I

Two observed point patterns with $n = 100$ events:

$\bar{d}_{min} = 7.86$ $\bar{d}_{min} = 1.17$

Question: Could these two patterns be realizations under CSR?

Obviously no, and this can be said with great confidence; pattern on left has much larger mean E2NE distance than expected under CSR, and vise versa for pattern on right

Looking at Observed Point Patterns II

Observed point pattern with $n = 100$ events:

$$\bar{d}_{min} = 5.18$$

Question: Is this pattern more clustered than a CSR-generated one?

Most probably no, since observed $\bar{d}_{min} = 5.18$ (black bar) lies at the center of the sampling distribution under CSR

Looking at Observed Point Patterns III

Observed point pattern with $n = 100$ events, and sampling distribution of $\bar{d}_{min}$ under CSR:

$$\bar{d}_{min} = 4.65$$

Question: Is this pattern more clustered than a CSR-generated one?

Equivalent question: Since small $\bar{d}_{min}$ values indicate clustering, is observed $\bar{d}_{min}$ on left side of sampling distribution under CSR?

Answer: The area under the curve of the sampling distribution to the left of observed $\bar{d}_{min} = 4.65$ is an indication of how unlikely is the observed pattern to be generated by CSR: *the smaller that area, the more unlike is the pattern to be a realization under CSR*

NOTE: if we were asking whether the observed pattern was more even (less clustered) than a CSR-generated one, we would be looking at the area under the curve to the right of observed $\bar{d}_{min} = 4.65$, since we would be interested in larger (than CSR-related) such distance values

P-Value of An Observed Sample Statistic

P-value = Area under the curve of the sampling distribution to the direction of the alternative hypothesis from the observed statistic
$\longrightarrow$ probability of observing a $\bar{d}_{min}$ value ≤ 4.65 in this case

NOTE: direction dependence in defining the P-value comes into play for one-sided tests; when we are just interested in whether the null hypothesis holds or not, no matter the direction of the alternative hypothesis (two-sided test), the final P-value is defined as twice the above P-value

Interpretation: The P-value is an indication of how unlikely is the observed pattern to be generated by the null hypothesis: *the smaller the P-value, the more unlike is the pattern to be a realization under the null hypothesis, here CSR*

Note: Any P-value is associated with a null hypothesis, since a P-value is computed from a sampling distribution which in turn is generated under a null hypothesis

Sampling Distribution of G Function Under CSR

Sampling distribution: of $\hat{G}(d)$ under CSR computed from 500 simulated point patterns within a square region of area $|A| = 100 \times 100$:

Interpretation: Plots provide envelope of simulated minimum and maximum $G(d)$ values for assessing whether an observed point pattern (not available here) can be regarded a realization from a CSR null process;
this is done by: (i) comparing the observed $\hat{G}(d)$ value (not available here) with the expected (mean) curve, and (ii) assessing its relative position within the envelope

The larger n is (more events in the domain), the tighter the envelop

Assessing Observed Ghat Plots I

Two observed point patterns with $n = 100$ events:

Question: Could these two patterns be realizations under CSR?

Most probably no, since the observed $\hat{G}(d)$ curve lies outside the simulation envelope

Assessing Observed Ghat Plots II

Observed point patterns with $n = 100$ events:

Question: Could this pattern be a realization under CSR?

Most probably yes, since the observed $\hat{G}(d)$ curve lies very close to the mean simulated plot and is well within the simulation envelope

Analytically-Derived Sampling Distributions

Analytical derivations:

- for simple domains, e.g., rectangles, there are mathematical formulae that provide the expected values of sample statistics under CSR
- in other words, people have already calculated what is the mean of a very large number of simulated $\bar{d}_{min}$ or $\hat{G}(d)$ values under CSR, without ever touching a computer
- these formulae have been derived before the advent of powerful computers, and have been used for a long time in point pattern analysis
- since, no simulation runs are involved, such analytically-derived formulae can be easily used without the need to resort to computer-intensive simulation procedures

Limitations:

- analytically-derived formulae need to account for the fact that events near the boundary of the study region do not have the same number of neighbors as events in the middle of that region
- such edge effects can be taken care of when the study region has simple geometry, e.g., for rectangles

In general, if you have access to computer software that can perform simulation, do not use analytically-derived formulae...

CSR-Expected Mean Nearest Neighbor Distance

- **Definition:** average of all $d_{min}(\mathbf{s}_i)$ values:

$$\bar{d}_{min} = \frac{1}{n}\sum_{\alpha=1}^{n} d_{min}(\mathbf{s}_i)$$

- single number does not suffice to describe point pattern

Checking for CSR:

1. compute expected value of mean nearest-neighbor distance, under CSR:

$$E\{\bar{d}_{min}\} = \frac{1}{2\sqrt{\lambda}}$$

λ = overall intensity of point pattern =
(# of points within study region) / (area of region)

2. form ratio R: $\boxed{R = \frac{\bar{d}_{min}}{1/(2\sqrt{\lambda})} = 2\bar{d}_{min}\sqrt{\lambda}}$

3. interpretation: $R < 1 \Rightarrow$ observed nearest neighbor distances shorter than expected $\Rightarrow$ tendency towards clustering
$R > 1 \Rightarrow$ tendency towards evenly spaced events

Result depends heavily upon study area definition (used to compute λ)

CSR-Expected G and F Functions

G function definition:

- proportion of event-to-nearest-neighbor distances $d_{min}(\mathbf{s}_\alpha)$ no greater than given distance cutoff d
- cumulative distribution function (CDF) of all n event-to-nearest-event distances:

$$\hat{G}(d) = \frac{\#[d_{min}(\mathbf{s}_i) \leq d]}{n}$$

F function definition:

- proportion of point-to-nearest-neighbor distances $\tilde{d}_{min}(\tilde{\mathbf{s}}_p)$ no greater than given distance cutoff d
- cumulative distribution function (CDF) of all m point-to-nearest-event distances:

$$\hat{F}(d) = \frac{\#[\tilde{d}_{min}(\tilde{\mathbf{s}}_p) \leq d]}{m}$$

Expected G and F function under CSR:

$$E\{G(d)\} = E\{F(d)\} = 1 - e^{-\lambda\pi d^2}$$

Checking for CSR:

compare empirical functions $\hat{G}(d)$ and $\hat{F}(d)$ with their theoretical counterparts $E\{G(d)\}$ and $E\{F(d)\}$ under CSR

Examples of G Functions

solid lines indicate expected value of $G(d)$ under CSR:

$$E\{G(d)\} = 1 - e^{-0.01\pi d^2}$$

- for *clustered events*, $\hat{G}(d)$ rises sharply at short distances, and levels off at large d-values

Examples of F Functions

solid lines indicate expected value of $F(d)$ under CSR:

$$E\{F(d)\} = 1 - e^{-0.01\pi d^2}$$

- for *clustered events*, $\hat{F}(d)$ rises slowly at short distances, and more rapidly at longer distances

Example with Evenly Spaced Points

solid lines indicate expected value of $G(d)$ and $F(d)$ under CSR:
$1 - e^{-0.01\pi d^2}$

- for *evenly-spaced events*, $\hat{G}(d)$ rises slowly at short distances, and then increases rapidly

The K Function

Looking *beyond* nearest neighbors

Concept:

1. construct set of concentric circles (of increasing radius d) around each event
2. count number of events in each distance "band"
3. cumulative number of events up to radius d around *all* events becomes the sample K function $\hat{K}(d)$

Formal definition:

$$\begin{aligned} K(d) &= \frac{E\{ \text{ \# of events within distance } d \text{ of any arbitrary event } \}}{E\{ \text{ \# of events within study area } \}} \\ &\simeq \frac{1}{\lambda}\frac{1}{n}\#\{d_{ij} \leq d, i = 1, \ldots, n, j = 1, \ldots, n\} = \hat{K}(d) \end{aligned}$$

CSR-Expected K Function

Under CSR:

- $E\{K(d)\} = \frac{\lambda \pi d^2}{\lambda} = \pi d^2$
- this can become a very large number (due to d^2), and consequently small differences between $\hat{K}(d)$ and $E\{K(d)\}$ cannot be easily resolved
- use L function instead:

$$\hat{L}(d) = \sqrt{\frac{\hat{K}(d)}{\pi}} - d$$

with $E\{L(d)\} = 0$

Interpreting the L function:

- for $\hat{L}(d) > 0 \Rightarrow$ more events separated by distance d than expected under CSR $\Rightarrow$ *clustered events*
- watch out for edge effects ...

Examples of L Functions

Expected appearance:

- for $\hat{L}(d) > 0 \Rightarrow$ more events separated by distance d than expected under CSR $\Rightarrow$ *clustered events*
- watch out for edge effects . . .

Recap

Statistical analysis of spatial point patterns:

- allows to quantify departure of results obtained via exploratory tools, e.g., $\bar{d}_{min}$ or $\hat{G}(d)$, from expected such results derived under specific null hypotheses, here CSR hypothesis
- can be used to assess to what extent observed point patterns can be regarded as realizations from a particular spatial process (here CSR)

Sampling distribution of a test statistic:

- lies at the heart of any statistical hypothesis testing procedure, and is tied to a particular null hypothesis
- simulation and analytical derivations are two alternative ways of computing such sampling distributions (the latter being increasingly replaced by the former)

Watch out for *edge effects*...