Spatial Data Analysis

Week 10: Point Pattern Data I

Meredith Franklin

Department of Statistical Sciences and School of the Environment

November 7, 2025

Point Pattern Data: A Historical Perspective

In 1946, R.D. Clarke wrote a report about heavily bombed region of South London.

AN APPLICATION OF THE POISSON DISTRIBUTION

BY R. D. CLARKE, F.I.A.
of the Prudential Assurance Company, Ltd.

READERS of Lidstone's *Notes on the Poisson frequency distribution* (*J.I.A.* Vol. LXXI, p. 284) may be interested in an application of this distribution which I recently had occasion to make in the course of a practical investigation.

During the flying-bomb attack on London, frequent assertions were made that the points of impact of the bombs tended to be grouped in clusters. It was accordingly decided to apply a statistical test to discover whether any support could be found for this allegation.

An area was selected comprising 144 square kilometres of south London over which the basic probability function of the distribution was very nearly constant, i.e. the theoretical mean density was not subject to material variation anywhere within the area examined. The selected area was divided into 576 squares of $\frac{1}{4}$ square kilometre each, and a count was made of the numbers of squares containing 0, 1, 2, 3, ..., etc. flying bombs. Over the period considered the total number of bombs within the area involved was 537. The expected numbers of squares corresponding to the actual numbers yielded by the count were then calculated from the Poisson formula:

$$Ne^{-m}(1+m+m^2/2!+m^3/3!+\ldots),$$

where $$N=576 \quad \text{and} \quad m=537/576.$$

The result provided a very neat example of conformity to the Poisson law and might afford material to future writers of statistical text-books.

The actual results were as follows:

No. of flying bombs per square	Expected no. of squares (Poisson)	Actual no. of squares
0	226·74	229
1	211·39	211
2	98·54	93
3	30·62	35
4	7·14	7
5 and over	1·57	1
	576·00	576

The occurrence of clustering would have been reflected in the above table by an excess number of squares containing either a high number of flying bombs or none at all, with a deficiency in the intermediate classes. The closeness of fit which in fact appears lends no support to the clustering hypothesis.

Applying the χ^2 test to the comparison of actual with expected figures, we obtain $\chi^2 = 1{\cdot}17$. There are 4 degrees of freedom, and the probability of obtaining this or a higher value of χ^2 is ·88.

Point Pattern Data: A Historical Perspective

- During WWII, Germany launched 1,358 V-2 Rockets at London.
- The V-2 had speed and a trajectory that made it invulnerable to interception, but its guidance systems were primitive, so it was thought that it couldn't hit specific targets.
- After strikes began in 1944, bomb damage maps were interpreted as showing that impact sites were clustered.
- If the V-2 strikes were clustered, then the guidance systems were more sophisticated than thought.
- R.D. Clarke set out to analyze these data to determine if the data were clustered or not.

Point Pattern Data: A Historical Perspective

- Clarke took a 12 km x 12 km region and sliced it up in to a grid of 576 squares, (144 km^2, so each grid square is 1/4 km^2).
- For each square, Clark recorded the total number of observed bomb hits. There were 537 total in the study area.
- He then recorded the number of squares with $k = 1, 2, 3, \ldots$ hits.
- The expected number of squares with k hits was derived from the Poisson distribution $\sum_{k=1}^{n} \frac{e^{-\lambda}\lambda^k}{k!}$ where $\lambda = \frac{537}{576}$ and $n = 576$.

No. of flying bombs per square	Expected no. of squares (Poisson)	Actual no. of squares
0	226·74	229
1	211·39	211
2	98·54	93
3	30·62	35
4	7·14	7
5 and over	1·57	1
	576·00	576

Point Pattern Data: A Historical Perspective

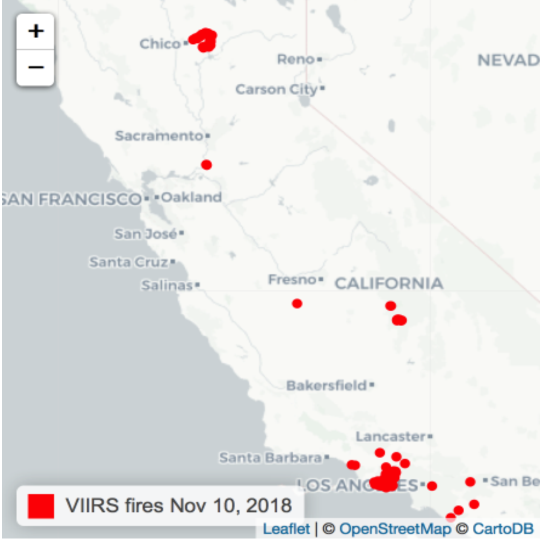
- Doing the cross tabulation of observed vs expected, he found $\chi^2 = 1.17$ which with 4 degrees of freedom (n-1 groups, the 0 group was excluded) has p-value=0.88.
- The occurrence of clustering would have been reflected in an excess of squares with a high number of bombs or none at all.
- The insignificant p-value and the closeness of fit of the data to the Poisson distribution indicates that the V-2 impact sites were random rather than clustered.

Point Pattern Data

- The goal of point pattern analysis is to assess whether the occurrences of an event show spatial structure.
- Distinguish between sampling locations and event locations.
- In geostatistics, "points" are sampling locations where a variable is measured; these form a set of spatial random variables whose dependence is studied via the covariance function.
- Point patterns consist of event locations; we focus on the presence/absence and configuration of events rather than measured values.
- Core question: do the observed events in the domain arise from a completely random spatial process, or do they exhibit clustering or inhibition (regularity)?
- In R, we use the `spatstat` ecosystem for point pattern analysis.

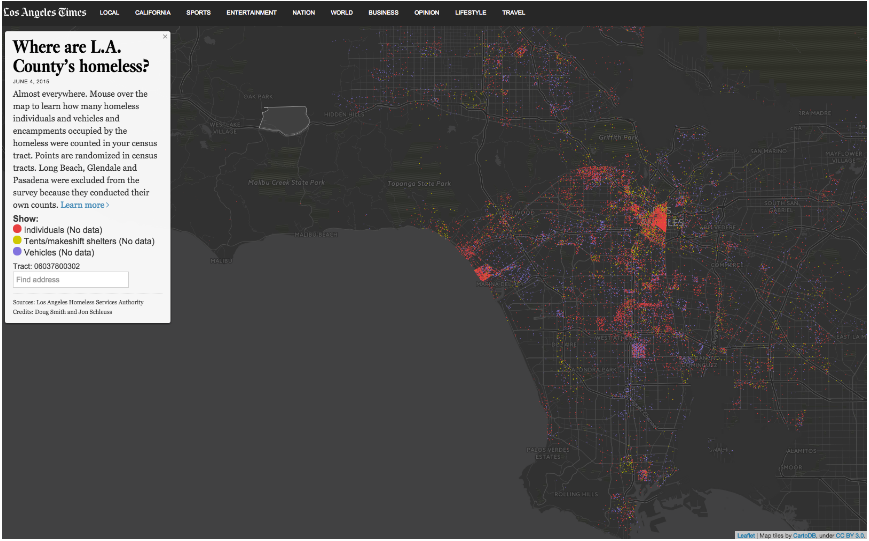
Point Pattern Data: Example

Locations of satellite detected fires in California



Point Pattern Data: Example

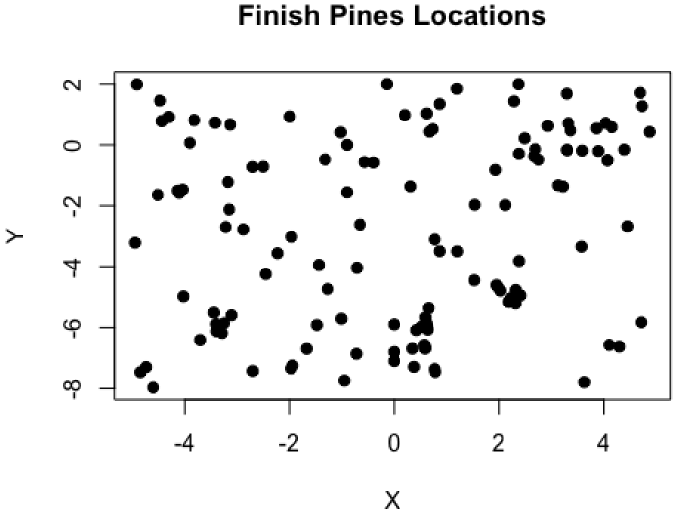
LA Homeless count: point locations of homeless



http://graphics.latimes.com/homeless-los-angeles-2015/

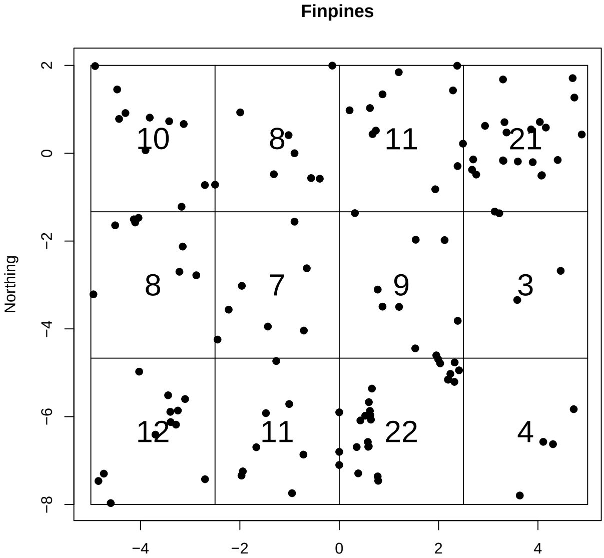
Point Pattern Data: A Classic Example

Locations of Finish pine trees in Finland forest



Quadrant Count

A simple quadrant count of the Finnpines points.

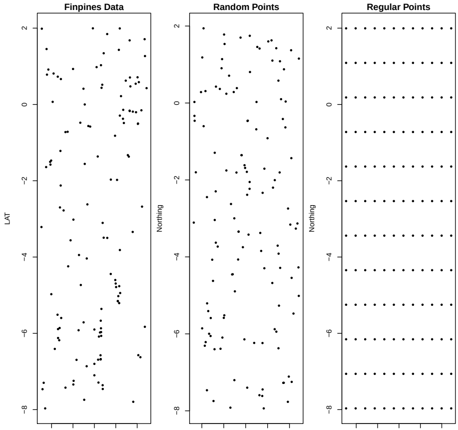


Point Pattern Data

Questions of interest:

- Are points closer together than expected by chance (clustering)?
- Are points more regularly spaced than expected by chance (inhibition)?
- What model could reproduce the observed pattern?

Comparing the Finpines data to a spatially random process and to a regular pattern:



Point Pattern Data: Objectives

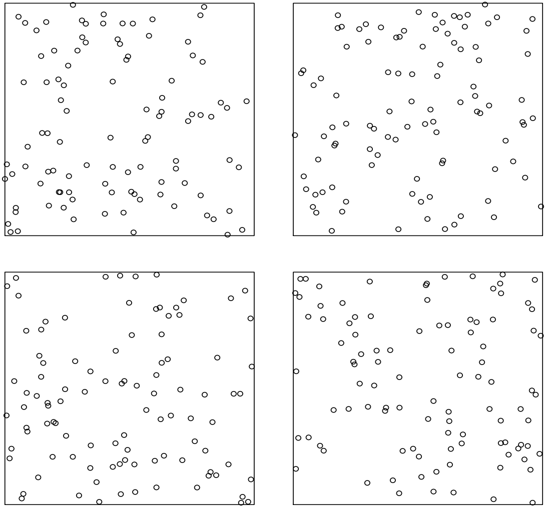
- Let $s = (x, y)$ denote a spatial location and let D be the study domain.
- Represent the pattern by event locations $\{s_i\} \subset D$ and the counting process $N(A) =$ number of events in region $A \subseteq D$.
- Null hypothesis: no spatial pattern beyond randomness (complete spatial randomness, CSR).
- Find statistics to test for clustering or regularity.
- Develop models to generate spatial patterns (Homogeneous Poisson Process, Inhomogeneous Poisson Process, Cluster/Cox processes, Simple inhibition processes).

Complete Spatial Randomness

- Terms you may see: spatial randomness, random pattern, at random, by chance.
- Under CSR, events are uniformly distributed over D and occur independently.
- Equivalently: any location/region within D is equally likely to contain a point, and points do not influence one another.
- Testing CSR is the most basic analysis for point pattern data.

Complete Spatial Randomness

Four realizations of a uniform (CSR) point process in a 1×1 box:



Properties of Complete Spatial Randomness

A CSR point process has:

- **Uniformity:** Any location has equal probability of containing a point.
- **Independence:** Point locations are independent; one point does not change the likelihood of another elsewhere.
- **Poisson counts:** CSR is modeled by a homogeneous Poisson process. For any region A, $N(A) \sim \text{Poisson}(\lambda|A|)$.
- **Intensity:** λ (points per unit area) is constant over D.
- **Expectation:** $E[N(D)] = \lambda|D|$, where $|D|$ is the area of D.

Complete Spatial Randomness

A point process with CSR is a stationary **homogeneous Poisson point process (HPP)**.

- The HPP is the simplest stochastic mechanism for spatial point patterns and serves as an idealized CSR benchmark; even if not strictly true in practice, it can be a useful approximation.

Homogeneous Poisson Process

Defining the homogeneous Poisson process (HPP)

- **Homogeneity:** intensity λ is constant across the study area (analogous to a constant mean in geostatistics).
- $\lambda = \frac{E[N(D)]}{|D|}$ is the expected number of events per unit area.
- $N(D)$ denotes the number of events in region D.
- The distribution of $N(D)$ is

$$\mathbb{P}\big(N(D) = k\big) = \frac{e^{-\lambda|D|}(\lambda|D|)^k}{k!}, \quad k = 0, 1, 2, \ldots$$

- $E[N(D)] = \lambda|D|, \quad \mathrm{Var}[N(D)] = \lambda|D|.$

Homogeneous Poisson Process: Postulates

HPP can be characterized by:

- **Independent increments:** For disjoint regions A and B, $N(A)$ and $N(B)$ are independent.
- **Small-area probabilities (orderliness):**

$$\lim_{|A|\to 0} \frac{\mathbb{P}[N(A) \geq 2]}{|A|} = 0,$$

  so the probability of two or more events in an infinitesimal area is negligible.
- **Homogeneity (constant rate):**

$$\lim_{|A|\to 0} \frac{\mathbb{P}[N(A) = 1]}{|A|} = \lambda > 0,$$

  i.e., the chance of a single event depends only on area, not location within D.

Homogeneous Poisson Process: Properties

Some properties of HPPs:

- The total number of events varies across realizations, with mean density λ.
- Location independence: the presence of one point does not affect probabilities of others nearby.
- For any disjoint A, B, $N(A)$ and $N(B)$ are independent (independent increments).
- Independence implies that knowing the count in one region gives no information about counts in a disjoint region.

Homogeneous Poisson Process: Properties

Conditional properties

- **Conditional uniformity:** Given $N(A) = n$, the n points are i.i.d. uniformly distributed over A (no clustering or inhibition).
- **Thinning:** Keep each point independently with probability p; the retained points form an HPP with intensity $p\lambda$.
- **Superposition:** The union of two independent HPPs on the same region with intensities λ_1 and λ_2 is an HPP with intensity $\lambda_1 + \lambda_2$.

Homogeneous Poisson Process: Properties

Why independence under thinning and superposition matters

- Enables construction of *cluster processes*.
- Defines **marked Poisson processes**: either (i) drop points from an overall HPP and independently assign marks, or (ii) superpose independent HPPs for each mark type. These constructions are equivalent by thinning/superposition.

Homogeneous Poisson Process: Conditionality

Given a fixed number of points, locations are i.i.d. uniform

- In 1D or 2D, conditioning on $N(A) = n$ removes λ: the n point locations are i.i.d. $\mathrm{Unif}(A)$.
- Equivalently: given N points of a Poisson process in A, their unordered locations are i.i.d. uniform on A.

Homogeneous Poisson Process: Conditionality (proof idea)

Uniform distribution given one point on a line

- Consider $[0,x]$ and an interval $[a,b] \subset [0,x]$ of length $\ell = b - a$.
- Let X be the point location, and $E = \{N([0,x]) = 1\}$.

$$\mathbb{P}(X \in [a,b] \mid E) = \frac{\mathbb{P}(N([a,b]) = 1)\,\mathbb{P}(N([0,a] \cup [b,x]) = 0)}{\mathbb{P}(N([0,x]) = 1)} = \frac{(\lambda \ell e^{-\lambda \ell})\, e^{-\lambda(x-\ell)}}{\lambda x e^{-\lambda x}} = \frac{\ell}{x},$$

which is the uniform distribution on $[0,x]$.

Homogeneous Poisson Process: Conditionality

Proof: Uniform distribution given a fixed number of points

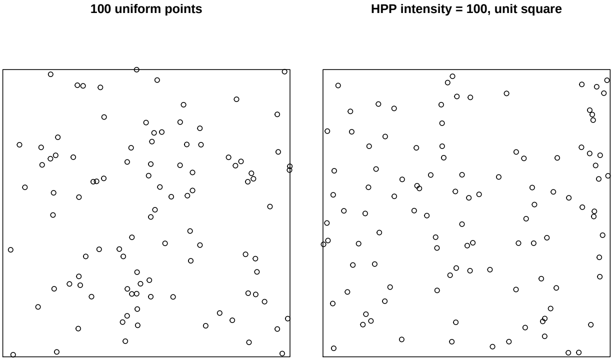
- $\mathbb{P}(X \in [a,b], E)$ equals the probability of exactly one point in $[a,b]$ (length $\ell = b-a$) times the probability of zero points in the complement $[0,a] \cup [b,x]$ (combined length $x-\ell$), by independent increments.
- Plugging into the Poisson probabilities:

$$\mathbb{P}(X \in [a,b] \mid E) = \frac{\mathbb{P}(N([a,b]) = 1)\,\mathbb{P}(N([0,a] \cup [b,x]) = 0)}{\mathbb{P}(N([0,x]) = 1)}$$
$$= \frac{(\lambda\ell)e^{-\lambda\ell}\, e^{-\lambda(x-\ell)}}{(\lambda x)e^{-\lambda x}} = \frac{\ell}{x},$$

which is the uniform probability on $[0,x]$.

Proof: Uniform distribution given a fixed number of points

- The uniform pdf on $[0,x]$ is $f(u) = 1/x$, so the result ℓ/x is exactly the probability that a uniform point falls in an interval of length ℓ.
- The same argument extends to any fixed number of points and to 2D: given $N(A) = n$, locations are i.i.d. $\mathrm{Unif}(A)$.



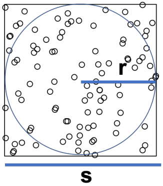
Homogeneous Poisson Process: Area proportionality

Property: Expected number of points is proportional to area

- For any subregion $A \subset D$, the expected count is $E[N(A)] = \lambda |A|$.
- To build intuition, we can use this to simulate/estimate the constant π.

Application of uniform intensity: estimating π

- Generate an HPP in a square of side s.
- Inscribe a circle of radius r in the square.



- Let n be the total number of points and m the number inside the circle.
- By area proportionality, $\frac{m}{n} \approx \frac{\pi r^2}{s^2}$, so

$$\hat{\pi} = \frac{ms^2}{nr^2}.$$

- If the circle is inscribed, $r = s/2$, yielding the familiar $\hat{\pi} \approx 4m/n$.

Tests of Complete Spatial Randomness

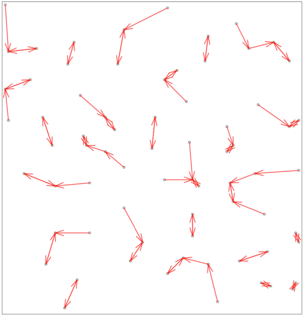
Most point pattern analyses begin with a test of CSR:

- If CSR is not rejected, further modeling may be unnecessary.
- Tests are exploratory tools that help formulate hypotheses about clustering or inhibition.
- CSR is a dividing hypothesis between **clustered** and **regular** patterns.

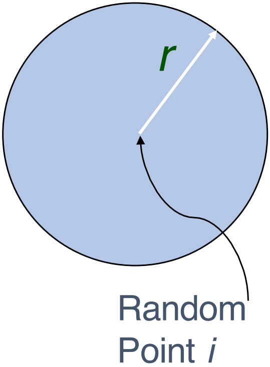
Tests of CSR: Nearest-neighbor G

Exploring CSR: Nearest neighbors

- For the i^{th} point, let D_i be the nearest-neighbor distance.
- Define $G(r) = \mathbb{P}(D_i \leq r)$, the CDF of NN distances.



Nearest Neighbor dists



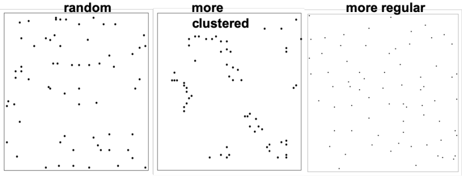
- Under homogeneous CSR (ignoring edge effects),

$$G(r) = \mathbb{P}(D_i \leq r) = 1 - \exp(-\lambda \pi r^2).$$

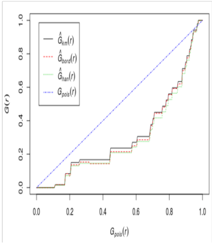
Tests of CSR: Nearest-neighbor G

Evaluating CSR via nearest neighbors

- Let $\hat{G}(r)$ be the empirical CDF of observed NN distances.
- Compare $\hat{G}(r)$ to $G(r) = 1 - e^{-\lambda \pi r^2}$.
- **Interpretation:** $\hat{G}(r) \gg G(r)$ suggests clustering; $\hat{G}(r) \ll G(r)$ suggests inhibition (regularity).
  Note: apply edge corrections near boundaries.



NN Distance Distribution for "more regular" pattern



Tests of CSR: Nearest-neighbor G

- Let $G(h)$ be the CDF of nearest-neighbor distances; estimate by

$$\hat{G}(h) = \frac{\#\{D_i \leq h\}}{n}.$$

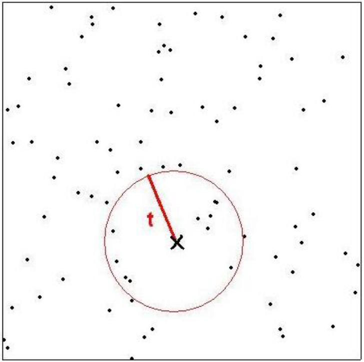
- Finite-window approximation: for CSR in area $|D|$,

$$G(h) \approx 1 - \left(1 - \pi h^2/|D|\right)^{n-1}$$

and, for large n (or in the plane), $G(h) \approx 1 - e^{-\lambda \pi h^2}$ with $\lambda = n/|D|$.

Tests of CSR: Ripley's K

- $K(r) = \lambda^{-1}\,\mathbb{E}[N_0(r)]$, where $N_0(r)$ is the number of *other* events within distance r of an arbitrary event.
- $K(r)$ summarizes second-order interaction; under CSR, $K(r) = \pi r^2$.



Tests of CSR: Ripley's K

- Regularity $\Rightarrow$ fewer neighbors within r than CSR.
- Clustering $\Rightarrow$ more neighbors within r than CSR.
- A simple estimator (no edge correction) is

$$\hat{K}(r) = \hat{\lambda}^{-1} \frac{1}{n} \sum_{i} \sum_{j \neq i} \mathbf{1}\{d(i,j) \leq r\}, \quad \hat{\lambda} = \frac{n}{|D|}.$$

Tests of CSR: Ripley's K (edge correction)

- To mitigate boundary bias, use edge-corrected estimators:

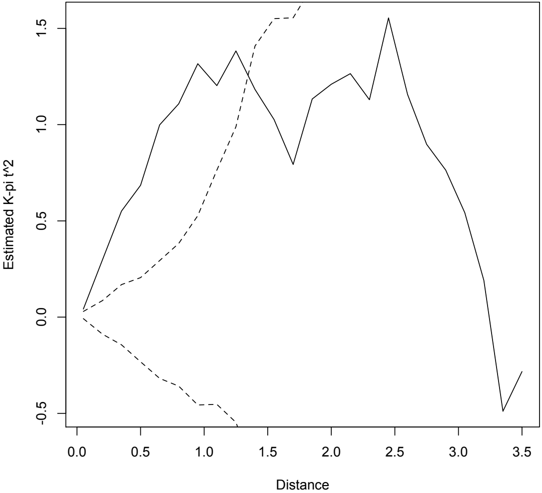
$$\hat{K}_{\mathrm{ec}}(r) = \hat{\lambda}^{-1}\frac{1}{n}\sum_i\sum_{j\neq i} w_{ij}(r)\,\mathbf{1}\{d(i,j) \le r\}.$$

- **Border (guard) correction:** $w_{ij}(r) = \mathbf{1}\{d(i,\partial D) > r\}$ (depends on i only).
- Other options: isotropic or translation corrections (as in `spatstat`).

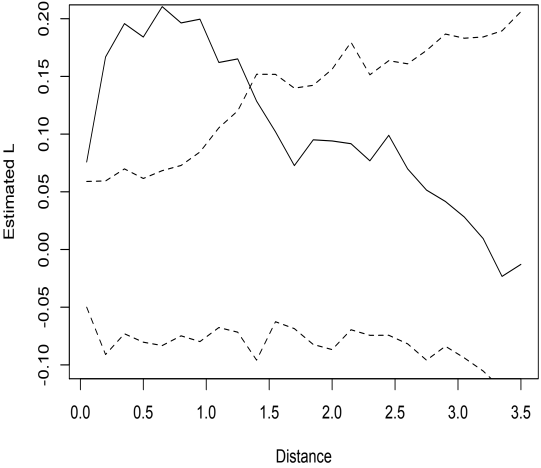
Tests of CSR: K/L plotting

- Under CSR, $K(r) = \pi r^2$; define $L(r) = \sqrt{K(r)/\pi}$.
- Plot $L(r) - r$ vs. r: under CSR, $L(r) - r = 0$ (flat line).
- Deviations above 0 indicate clustering; below 0 indicate inhibition.

Tests of CSR: Ripley's K (example)



Tests of CSR: Ripley's L (example)



Tests of CSR: Monte Carlo

- Choose a test statistic (e.g., G, K, L).
- Simulate N_{sim} CSR patterns in the same window and compute the statistic for each.
- Compare the observed statistic to the simulated distribution; use rank/envelope tests.
- Account for edge effects (e.g., border/isotropic corrections).

Tests of CSR: Monte Carlo envelopes for K/L

- Simulate m CSR patterns (same window, same n or same λ).
- Compute $\hat{K}^{(b)}(r)$ (or $\hat{L}^{(b)}(r)$) for $b = 1, \dots, m$.
- Pointwise envelopes: $U(r) = \max_b \hat{L}^{(b)}(r)$, $L(r) = \min_b \hat{L}^{(b)}(r)$.
- Plot $\hat{L}_{\text{obs}}(r) - r$ with envelopes; rank-based/global envelopes give formal tests.
- With m simulations, a one-sided rank test at size $k/(m+1)$ rejects if the observed curve ranks among the k most extreme.

Tests of CSR: Rank idea for K

Test statistic (rank idea):

- Let $\hat{K}^{(0)}(r)$ be observed, and $\hat{K}^{(1)}(r), \ldots, \hat{K}^{(m)}(r)$ be from simulations.
- For each r, order $\hat{K}^{(b)}(r)$ and record the rank of $\hat{K}^{(0)}(r)$ among $m+1$ curves.
- A one-sided test of size $k/(m+1)$ rejects if the observed ranks are among the k most extreme over r (with appropriate global envelope procedure).

Tests of CSR: Global envelope test (rank-based)

Goal: Test the whole curve (all r in an interval I) against CSR while controlling for multiple r's.

Null (CSR): The observed point pattern is a realization of a homogeneous Poisson process on the study window D (complete spatial randomness), i.e., for all r in the test range I, $K(r) = \pi r^2$ (equivalently L(r)=r). (Specify whether the null is fixed-N CSR or Poisson CSR with intensity $\lambda = n/|D|$)

Under CSR in window D, use

$$T(r) = K(r) - \pi r^2 \quad \text{or} \quad T(r) = L(r) - r, \qquad r \in I.$$

Steps:

- Simulate m CSR patterns on D; compute $T^{(b)}(r)$, $b = 1, \ldots, m$, and the observed $T^{(0)}(r)$.
- For each $r \in I$, rank $T^{(0)}(r)$ among $\{T^{(b)}(r)\}_{b=0}^{m}$ (two-sided extremeness).
- Build the *global* $(1-\alpha)$ envelope from the ranks across all $r \in I$.
- Reject CSR at level α if the observed curve exits the global envelope at any $r \in I$.

Tests of CSR: Global envelope test (rank-based)

- Monte Carlo p-value:

$$p = \frac{1 + \#\{\text{simulated curves more extreme than observed (globally)}\}}{m+1}.$$

Interpretation: $T(r) > 0$ above the envelope indicates *clustering* at scale r; $T(r) < 0$ indicates regular pattern (inhibition).

- Rank envelopes are robust to monotone transforms: $K \leftrightarrow L$.)

Tests of CSR: Alternatives

Curve-level test statistics (no envelope plot):

- **DCLF** (Diggle–Cressie–Loosmore–Ford): integrated, variance-weighted squared deviation over I.
- **MAD** (Maximum Absolute Deviation): largest $|T(r)|$ over I.
- p-values via Monte Carlo: rank the observed statistic among m simulations.

Key choices that affect power and validity:

- *Null model*: CSR with fixed N (conditional) vs Poisson CSR (random N).
- *Distance range* I: typically $\sim$40–60% of the shortest window side to limit edge bias.
- *Edge correction*: "border" or "translation" for K/L; apply consistently to data and simulations.
- *Simulation size*: larger m (e.g., 199/499/999) gives tighter envelopes and more stable p.

Reporting: Show $T(r)$ with the global envelope and give the global p; state I, edge correction, null (fixed-N vs Poisson), and m.

Tests of CSR: Inter-event distances

- Let $H(h)$ be the CDF of inter-event distances in D; estimate by

$$\hat{H}(h) = \frac{\#\{d_{ij} \leq h : \ i < j\}}{\binom{n}{2}}.$$

- For specific windows (square/circle), theoretical $H(h)$ is known (e.g., Bartlett, 1964 for unit circle):

$$H(h) = 1 + \pi^{-1}\left[2(h^2-1)\cos^{-1}(h/2) - h\left(1 + \frac{h^2}{2}\right)\sqrt{1 - \frac{h^2}{4}}\right], \quad 0 \leq h \leq 2.$$

Tests of CSR: Inter-event distances

- As with Ripley's K we need the distribution of our statistic $\hat{H}(h)$ under CSR
- Visual test: plot $\hat{H}(h)$ vs. $H(h)$ with Monte Carlo envelopes under CSR.
- Simulate m CSR patterns, compute $\hat{H}^{(b)}(h)$, and form pointwise envelopes

$$\hat{H}_U(h) = \max_b \hat{H}^{(b)}(h), \quad \hat{H}_L(h) = \min_b \hat{H}^{(b)}(h).$$

CSR and Basic Tests (recap)

Complete Spatial Randomness (CSR): A homogeneous Poisson point process on window D with constant intensity λ; counts in any region A satisfy $N(A) \sim \text{Poisson}(\lambda|A|)$ and locations are independent.

Quadrat count test:

- Partition D into cells of area a; under CSR, cell counts are $\text{Poisson}(\lambda a)$.
- Goodness–of–fit via χ^2 or dispersion index.
- *Caveats:* choice of cell size/location affects power; ensure expected counts are not too small; edge effects.

Ripley's K and L:

- $K(r) = \lambda^{-1}\,\mathbb{E}[N_0(r)]$; under CSR, $K(r) = \pi r^2$.
- Often plot $L(r) - r$ with $L(r) = \sqrt{K(r)/\pi}$; under CSR, $L(r) - r \equiv 0$ (easier to read).
- *Testing:* compare the observed curve to *Monte Carlo envelopes* under CSR (prefer global/rank envelopes).
- *Interpretation:* above $\Rightarrow$ clustering; below $\Rightarrow$ regular (inhibition).

CSR and Basic Tests (recap, con't)

Nearest-neighbour (G):

- $G(r)$**:** CDF of event→nearest-event distance. Approx. CSR reference in the plane: $G(r) \approx 1 - e^{-\lambda \pi r^2}$.
- *Interpretation:* above CSR at small $r \Rightarrow$ clustering; below $\Rightarrow$ regular (inhibition).

Inter-event distances (H):

- $H(r)$**:** CDF of pairwise (inter-event) distances. CSR reference depends on window geometry (closed forms exist for simple shapes).
- *Interpretation:* above CSR $\Rightarrow$ more close pairs (clustering); below $\Rightarrow$ regular (inhibition).

Extras (in lab): $F(r)$ empty-space CDF (random location→nearest event, CSR $\approx 1 - e^{-\lambda \pi r^2}$); $J(r) = (1-G)/(1-F)$ with CSR baseline 1.

Workflow for testing CSR with G, K/L, and H

Monte Carlo testing is recommended

- Choose summary $S(r) \in \{K, L, G, H\}$ and distance range I.
- Simulate m CSR patterns on D and compute $S^{(b)}(r)$.
- Build **global rank envelopes** and compare the observed curve, look at global p-value.

Considerations:

- *Edge correction:* use the same for data and sims (e.g., `border`/`translation` for K/L; `km`/`rs` for G; `rs`/`km` for H).
- *Distance range:* restrict to ∼40–60% of the shortest window side to limit edge bias.
- *Transform:* plot $K(r) - \pi r^2$ or $L(r) - r$; L is easier to read and often stabilizes variance.
- *Computation:* H is heavier; use a fine r grid (small step), possibly fewer sims at first.

Alternatives test For a single p-value without envelopes, use **DCLF** (integrated squared deviation) or **MAD** (max deviation) on $K/L/G/H$.

Inhomogeneous Poisson Process

- A generalization of the HPP is the Inhomogeneous (heterogeneous) Poisson Process (IPP). The IPP occurs when the intensity λ is not constant over the region.
- Many cases homogeneity in intensity is not realistic, for example the locations of trees in a forest may be irregular due to geographic features such as soil, rock, slope or other terrain irregularities.
- In the case of IPP, the intensity is a function that varies spatially, $\lambda(s)$.
- In the IPP, the intensity is a function of location, allowing for more flexibility to model patterns with varying densities.

Inhomogeneous Poisson Process (IPP)

- **Intensity function:** $\lambda(s) \geq 0$ on window $D \subset \mathbb{R}^2$ (units: points per unit area).
- **Local (limit) definition:** $\lambda(s)$ is the *local expected density of points* at s:

$$\lambda(s) \; = \; \lim_{|ds| \to 0} \frac{\mathbb{E}[N(ds)]}{|ds|},$$

- **Counts over regions:** For any $A \subset D$,

$$N(A) \sim \text{Poisson}\big(\Lambda(A)\big), \qquad \Lambda(A) = \int_A \lambda(s)\, ds, \qquad \mathbb{E}[N(A)] = \Lambda(A).$$

- **Independence:** If $A_1, \ldots, A_k$ are disjoint, then $N(A_1), \ldots, N(A_k)$ are independent.
- **Special case (HPP):** If $\lambda(s) \equiv \lambda$ is constant, then $N(A) \sim \text{Poisson}(\lambda |A|)$.

IPP: Intensity Function

- For an inhomogeneous Poisson process, $\lambda(s)$ is closely related to the density of events over the domain.
- The intensity function $\lambda(s)$ can depend on spatial covariates (such as elevation, population density, or environmental factors) or follow a general trend.
- Density estimators provide an estimate of intensity (e.g. kernel density).

IPP: Kernel Intensity Estimator (2D)

Goal: nonparametric estimate of the intensity $\lambda(s)$ from points $s_1, \dots, s_n$ in window D.

$$\boxed{\hat{\lambda}_h(s) = \frac{1}{h^2} \frac{\sum_{i=1}^{n} \kappa\left(\frac{\|s-s_i\|}{h}\right)}{e(s)}}$$

- $h > 0$: bandwidth (smoothing scale).
- $\kappa(\cdot)$: isotropic kernel with $\int_{\mathbb{R}^2} \kappa(\|u\|)\, du = 1$ (e.g., Gaussian, biweight).
- $e(s)$: edge correction (fraction of kernel mass inside D at location s); compensates boundary bias.
- Interpretation: local average of nearby points weighted by κ, properly scaled and edge-corrected.

In d dimensions use $1/h^d$. With anistropic bandwidth matrix H: $\hat{\lambda}(s) = |H|^{-1/2} \sum_i \kappa\big(\|H^{-1/2}(s-s_i)\|\big)/e_H(s)$.

IPP: Kernel Density

- There are various kernel functions, including Gaussian:

$$\kappa(s) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{s^2}{2h^2}\right)$$

  where s is the distance from the point where the density is being estimated, h is the bandwidth that controls the degree of smoothing.
- Estimating the density function is done by

$$\hat{f}(s) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{s-s_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{(s-s_i)^2}{2h^2}\right)$$

- Each data point s_i contributes a Gaussian-shaped bump to the density estimate, centered at s_i and with spread controlled by h . The estimated density $\hat{f}(s)$ at s is the average of these contributions.

IPP: Kernel Density

- Another kernel is the Epanechnikov kernel:

$$\kappa(s) = \frac{3}{4h}\left(1 - \frac{s^2}{h^2}\right) \quad \text{for } s \leq h$$

- The Epanechnikov kernel is parabolic, assigning weights that decrease quadratically with distance and become zero at distance h (the bandwidth).
- the biweight kernel is also a popular parabolic kernel:

$$\kappa(s) = \frac{15}{16h}\left(1 - \frac{s^2}{h^2}\right)^2 \quad \text{for } s \leq h$$

- In practice, the choice of h (bandwidth, or smoothing parameter) has a greater impact on the results than the specific kernel shape.

IPP: Kernel Density

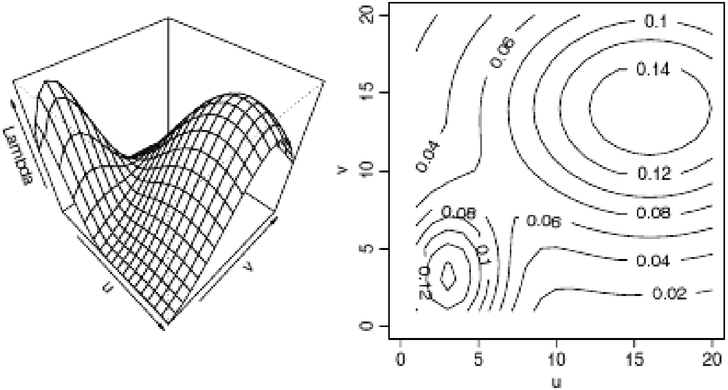


FIG. 5.5 Example intensity function, $\lambda(s)$, for a heterogeneous Poisson point process defined for $s = (u, v)$ and $u, v \in (0, 20)$.

IPP: Kernel Density

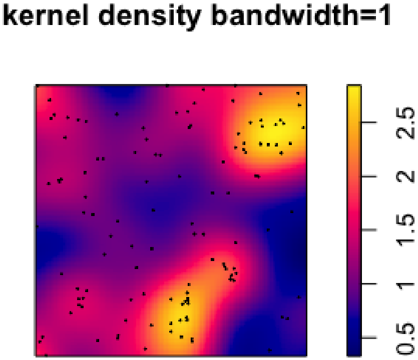
- The inhomogeneous Poisson process shows lack of events between the modes
- More events around the mode (16,14) and a narrower peaked area around (3,3)
- Collections of events suggest areas of higher intensity
- In order to do kernel estimation we need to choose:
  - the kernel type (e.g. Gaussian)
  - the bandwidth (e.g. radius of the area where smoothing is applied). Smaller bandwidths produce more localized densities.

IPP: Intensity Estimates with GAM

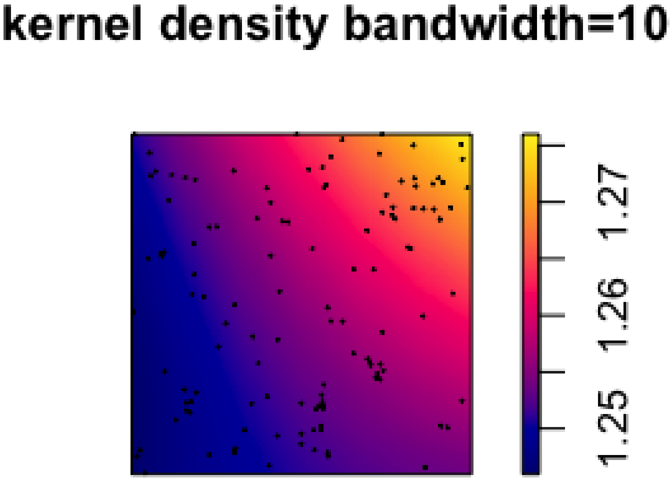
- Smoothing splines can be used to estimate a smooth surface for $\lambda(s)$ over the study region. Thin-plate splines are a common choice for spatial data, as we saw in geostatistics.
- Splines provide a smooth, flexible estimate of $\lambda(s)$ without assuming a specific parametric form, but they may not incorporate covariates directly.
- Can use $\hat{\lambda}(s) = \beta_0 + f(x,y)$ where $f(x,y)$ is a smooth function estimated by the thin-plate spline.

IPP: Kernel Density in R

Example: finpines data
plot(density(finpines,1))
points(finpines,pch=19,cex=0.1)



IPP: Kernel Density in R



IPP: Parametric Intensity

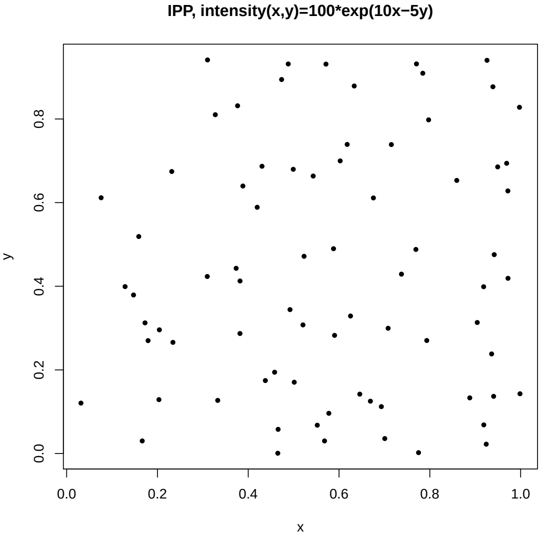
- Example of varying intensity function $\lambda(s)$ could be that intensity varies with location due to environmental heterogeneity
- Example if D i a square unit and N(D)=100
- $\lambda(x,y) = 100 * \exp(10x - 5y)$
- $\lambda(x,y) = 100 * \exp(-10x + 5y)$
- The intensity function $\lambda(s)$ can also be modeled as a function of spatial covariates (e.g., elevation, population density).
- Example: $\lambda(s) = \exp(\beta_0 + \beta_1 x_1(s) + \beta_2 x_2(s) + \cdots + \beta_p x_p(s))$ where $x_1(s), x_2(s), \ldots, x_p(s)$ are covariates at location s , and $\beta_0, \beta_1, \ldots, \beta_p$ are parameters estimated from the data.

IPP: Parametric Intensity

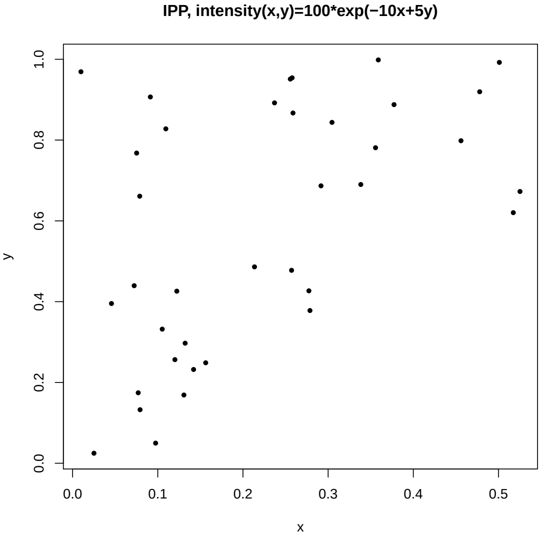
Intensity functions are often modeled as exponential functions.

- Ensures positive values ($\lambda(s)$ must be positive).
- Linear relationship with covariates on the log scale.
- Aligns with GLM, specifically Poisson regression.

IPP: Parametric Intensity



IPP: Parametric Intensity



IPP: Parametric Intensity

- We might see that cases of respiratory disease differ with respect to distance from a point source of environmental pollutions_0

$$\lambda(s) = \lambda_0(s) f(||s - s_0||, \theta)$$

- Where $\lambda_0(s)$ models the variation in population density
- $f(\cdot, \theta)$ models how the impact of the source varies with distance $\cdot(s - s_0)$ and angle θ.
- Alternatively you could write this as an exponential decay with distance: $\lambda(s) = \exp(\beta_0 + \beta_1 \text{Distance}(s) + \beta_2 X(s))$ where β_0 is the baseline intensity or intercept, distance is the distance between the pollution source and each point, β_1 would be the distance effect, and β_2 is the estimate for other important covariates such as population density.
- Other variations such as $\lambda(s) = \exp\left(\beta_0 + \frac{\beta_1}{\text{Distance}(s)}\right)$ would work too.

IPP Recap

Model: Points occur independently with spatially varying intensity $\lambda(s)$ on window D.

$$N(A) \sim \text{Poisson}\left(\int_A \lambda(s)\,ds\right), \quad \text{independent for disjoint } A.$$

First vs. second order:

- *First-order* structure $= \lambda(s)$ (trend / density variation).
- *Second-order* interaction $=$ dependence between points (clustering/regular pattern -inhibition, next week).

IPP assumption: After accounting for $\lambda(s)$, points are *independent* (no residual interaction).
Examples: Tree locations varying with topography, disease cases varying with population density, proximity to sources, etc.

IPP Recap: Estimating and Modelling $\lambda(s)$

Nonparametric (kernel) estimation.

$$\hat{\lambda}_h(s) = \frac{1}{h^2} \sum_{i=1}^{n} \kappa\left(\frac{\|s - s_i\|}{h}\right) \frac{1}{e(s)}$$

Edge factor $e(s)$ corrects for boundary truncation; choose bandwidth h.

Parametric (log-linear) IPP.

$\log \lambda(s) = \beta_0 + \beta^\top z(s)$ (covariates: elevation, slope, distance-to-road, population, ...)

Offsets/exposure. If risk varies by known surface $\rho(s)$, use $\log \lambda(s) = \log \rho(s) + \beta_0 + \beta^\top z(s)$.

In R (spatstat).

- Kernel: `density.ppp(X, sigma = ..., edge = TRUE)`; helpers: `bw.diggle`, `bw.scott`.
- Parametric IPP: `ppm(X ~ covariates)`; predict intensity: `predict(fit, type="trend")`.
- Covariate effect (nonparametric): `rhohat(X, Z)`.

IPP: Diagnostics & Tests

Goodness-of-fit for $\lambda(s)$:

- Quadrat GOF relative to fitted $\hat{\lambda}(s)$: `quadrat.test(fit, nx, ny)`.
- Residual fields (raw/Pearson) & smoothing: `residuals(fit, type="pearson")`, then `Smooth(...)`.
- Lurking variable plots vs. covariates or coordinates: `diagnose(fit)`.

Residual check (after trend removed):

- Inhomogeneous summaries: $K_{\text{inhom}}(r), L_{\text{inhom}}(r), g_{\text{inhom}}(r)$ computed with $\hat{\lambda}(s)$.
- Under a correct IPP (no interaction): $K_{\text{inhom}}(r) \approx \pi r^2$, $L_{\text{inhom}}(r) - r \approx 0$.
- Monte Carlo test: simulate from fitted IPP, build **global rank envelopes**.
- Use the same edge correction and distance range for data and simulations.
- If intensity varies strongly, prefer inhomogeneous K/L (and/or g) over homogeneous versions.

IPP: R Functions

- **Bandwidth** matters for kernel $\hat{\lambda}$: try `bw.diggle` and sensitivity checks.
- **Offsets/exposure**: supply known at-risk surface as `offset(log(exposure))` in `ppm()`.
- **Pair correlation**: $g_{\text{inhom}}(r)$ with `pcfinhom()` gives scale-resolved residual interaction.
- **G/F/J under inhomogeneity**: inhom variants exist (`Ginhom`, `Finhom`, `Jinhom`) but are usually secondary—prefer K_{inhom}, L_{inhom}, or g_{inhom} for testing independence after trend.
- **Simulation**: simulate IPP via `rpoispp(lambda = lam.im)` or `rpoispp(function(x,y){...}, win=...)`.