Point Pattern Analysis

Concepts

Luc Anselin

http://spatial.uchicago.edu
some examples

terminology

intensity
Some Examples
• Classic Examples

- forestry, plant species, astronomy
- locations of crimes, accidents
- locations of persons with a disease
- facility locations (economic geography)
- settlement patterns
SF car thefts as points in space, Aug 2012
Chicago supermarket locations as points in space, 2014
Manufacturing locations, Duranton and Overman (2005)

(a) Basic Pharmaceuticals (SIC2441)

(b) Pharmaceutical Preparations (SIC2442)
Terminology
• Events

points are the location of an event of interest

all points are known

= mapped pattern

selection bias

events are mapped, but non-events are not
Research Questions

is the pattern random or structured in some fashion

clustered: closer than random

dispersed/regular: farther than random

what is the process that might have generated the pattern
Classic Point Pattern Analysis

points located on an isotropic plane

no directional effect

distance as straight line distance
Marked Point Pattern

both location and value

e.g., location and employment of manufacturing plants, trunk size of trees

patterns in the location of the points and in the values association with the locations

= spatial autocorrelation
Classic data set: longleaf pines
- Multi-Type Pattern

multiple categories of events in one pattern

research questions:

patterning within a single type

association between patterns in different types

repulsion or attraction between types
Classic data set: Lansing Woods data on multiple tree types
Multitype: Supermarkets and Liquor Stores

Chicago multitype point pattern
• Points on a Network

realistic locations

events located on actual network, not floating in space

network distance

replaces straight line distance

shortest path on the network
Los Angeles riot locations
Baghdad IED locations
• Case-Control Design

take into account background heterogeneity

non-uniform “population at risk”

pattern for event of interest = case

pattern for background population = control
Classic case-control data set: Lancashire cancers
Classic case-control data set: Lancashire cancers
Intensity
• Average Intensity

first moment of a point pattern distribution

number of points per unit area

intensity: $\lambda = \frac{N}{|A|}$

area depends on bounding polygon
• Bounding Polygon

  classic unit square

  unrealistic but used in classic example data sets

  actual regional boundary (GIS)

  bounding box

  convex hull
• Bounding Box

rectangle

upper right corner = min x, max y

lower left corner = max x, min y
Chicago supermarkets - bounding box
Convex Hull

tightest fit around points

several algorithms
Chicago supermarkets - convex hull
• Adjustments

rescaled bounding box and convex hull

Ripley-Rasson transformation

add small area outside traditional bounding box or convex hull
Sensitivity of Intensity Estimate

Chicago supermarkets $n = 149$
(units of area are square feet $\times 10^8$)

- Boundary: 2.31
- Bounding Box: 1.55
- Convex Hull: 2.53
- R-R Bounding Box: 1.51
- R-R Convex Hull: 2.34
complete spatial randomness (CSR)
clustered patterns
regular/dispersed patterns
advanced models
Complete Spatial Randomness (CSR)
• **Definition of CSR**

  standard of reference

  uniform distribution

  each location has equal probability for an event

  locations of events are independent

  homogeneous planar Poisson process
• Poisson Point Process

distribution for \( N \) points in area \( A \), \( N(A) \)

intensity: \( \lambda = N/|A| \) \((|A| \text{ is area of } A)\)

therefore \( N = \lambda |A| \) points randomly scattered in a region with area \(|A|\)

Poisson distribution: \( N(A) \sim Poi(\lambda |A|) \)
• Recap - Poisson Distribution

single parameter: mean = $\lambda |A|$

mean = variance

• $P[N(A)=y] = \exp(-\lambda |A|)(\lambda |A|)^y / y!$

$P[N(A)=0] = \exp(-\lambda |A|)$
Probability Density for Poisson with Mean=2

Counts

0.00 0.05 0.10 0.15 0.20 0.25
Probability Density for Poisson with Mean=5
• Example: CSR with intensity 5 points/km²

with region as circle with radius r, area = \( \pi r^2 \)

e.g., with \( r = 0.01 \), area \( \approx 0.03 \)

probability of zero points in the circle

\[ P[N(A)=0] \approx e^{-5\times0.03} \approx e^{-0.16} \approx 0.85 \]
Simulating CSR

- Total number of points fixed
- Coordinates x, y random uniform
- Total number of points follow Poisson distribution
  - \( \lambda \) and |A| given
  - Number of points in area A from Poisson(\( \lambda |A| \))
Simulated CSR - uniform with N fixed on unit square
Simulated CSR using Poisson distribution for $N$ on unit square

CSR (Poisson) Lambda = 50

N=53

CSR (Poisson) Lambda = 100

N=91
• Limitations of CSR

  not a natural process

  very few actual processes are CSR

  not interesting, purely random

  used as a reference

  null hypothesis of CSR
Clustered Patterns
Properties of a Clustered Pattern

- events more grouped than CSR
- some higher densities
- many point pairs at short distances
- overdispersion
  - variance > mean
- greater variation in densities than CSR
• Source of Clustering

contagion

presence of an event at one location affects the probability of an event at another location

correlated point process

heterogeneity

varying intensity ($\lambda$) with location

heterogeneous Poisson point process
• Contagious Point Distributions

  two stages

  distribution for “parents”

  distribution for “offspring”

formal models

  Poisson cluster process or Neyman-Scott process

  Matern cluster process
Neyman–Scott Parents Lambda=10

realized $N=15$

overall $\lambda=10 \times 5$

Neyman–Scott Children, N=5 per parent

realized $N=55$

Simulated Neyman-Scott process
• Heterogeneous Poisson Process

- spatially varying intensity $\lambda(s)$
  - mean intensity is integral of the location-specific intensities over the region

- source of variability
  - function for $\lambda(s) = f(z)$ with covariates
  - doubly stochastic process with $\lambda(s) \sim \Lambda(s)$
\[ \lambda(x,y) = 100 \times \exp(-3x) \]

Average Intensity = 32
\[ \ln Z \sim N(4.1, 1) \quad \mathbb{E}[\lambda] \approx 100 \quad \text{average } \lambda = 113 \]
True Contagion and Apparent Contagion

inverse problem

identify process from pattern

impossible to identify contagion from heterogeneity
Regular/Dispersed Patterns
• Properties of a Regular Pattern
  
  events less grouped than CSR

  few higher densities, empty space

  many point pairs at larger distances

  underdispersion

  variance < mean

  less variation in densities than CSR
- **Source of Dispersion**
  - repulsion, competition
  - inhibition processes
    - minimum permissible distance (no points closer)
  - Matern processes
Simulated Inhibition Point Processes (Matern)
Advanced Models
• Interaction Point Processes

Markov point processes

Strauss process

pairwise interaction point processes

area interaction point processes

many formal models, with parameters
Point Pattern Analysis

Descriptive Statistics

Luc Anselin

http://spatial.uchicago.edu
centrography
quadrat counts
intensity functions
Centrography
• **Descriptive Statistics of the Point Locations**

  statistics of the x,y coordinates

  central tendency

  mean, median, minimum distance

  dispersion

  standard deviation box, standard distance circle, standard deviation ellipse
• Central Tendency

mean center = average of x and average of y

median center = median of x and median of y

central feature = point of minimum cumulative distance
Central Points: mean (+), median (x) and min distance
• **Standard Deviation Box**

  centered on the mean center

  sides of rectangle is proportional to the standard deviation in x and y dimension

  actually two standard deviations
Standard deviation box
• Standard Distance Circle

centered on the mean center

radius is hypothenuse of triangle with standard deviation in x and y direction

radius is square root of sum of variance in x and variance in y direction
Standard distance circle
• Standard Deviation Ellipse

centered on mean center

major elliptical axis follows angle of greatest dispersion

use standard deviation along each orthogonal axis

reflects both location and direction
Standard deviation ellipse
Applications of Centrography

visual summary of pattern

changes/differences in central tendency

compactness and directionality

changes/differences over time
Centrographic comparison of supermarkets and liquor stores
Chicago - 2014
Quadrat Counts
• Quadrat Counts

assess the extent to which intensity is constant across space

quadrat = polygon

count the points in the quadrant

visualize counts, intensity map
Quadrat counts - alternative configurations
Quadrat count intensity graph

intensity = count / area
Quadrat Count Test

quadrat counts as a goodness of fit test

null hypothesis: expected count

\[ n_m = \text{total points} / \text{total subregions} = n / m \]

\[ X^2 = \sum_i (n_i - n_m)^2 / n_m \sim \chi^2(m-1) \]

chi-square goodness-of-fit test

also Monte Carlo inference option
Quadrat count test = 167.9 with 17 d.f., p < 0.0000
• **Drawbacks**

  arbitrariness of quadrant selection

  scale issue, MAUP

  shape: square, rectangle, circle, irregular polygon

  fixed or random quadrants

  ignores spatial autocorrelation

  inference may be spurious
Intensity Functions
• Intensity Function

spatial heterogeneity

intensity $\lambda(s)$ varies with location $s$

estimating $\lambda(s)$

non-parametric kernel function
• **Intensity vs. Density**

  intensity function

  expected number of events per unit area at location

  integrates to overall mean number of events

  density function

  probability of an event at a given location

  integrates to one

  differ by a constant of proportionality
• **Kernel Density Estimation**

  non-parametric approach

  weighted moving average of the data

• \( f(u) = \left( \frac{1}{N_b} \right) \sum_{i} K[(u - u_i)/b] \)

  \( u \) is any location

  \( K \) is the kernel function (a function of distance)

  \( b \) is the bandwidth, i.e., how far the moving average is computed with \( N_b \) as the number of observations within the bandwidth
Normal kernel (Source: Crimestat Manual)
Effect of kernel bandwidth (Source: CrimeStat Manual)
• Example Kernel Functions

Gaussian:

\[ K(z) = \left(\frac{1}{\sqrt{2\pi b}}\right) \exp\left[-\frac{1}{2b^2}z^2\right] \]

most commonly used in point pattern analysis

bandwidth corresponds to standard deviation

Triangle:

\[ K(z) = \left[\frac{1}{b(2-b)}\right] (1 - z) \mathbb{I}(z \leq 1) \]

Quartic:

\[ K(z) = \left(\frac{15}{16b}\right) [1 - z^2]^2 \mathbb{I}(z \leq 1) \]
Chicago supermarket locations
Gaussian kernel
bw = 14259
Chicago supermarket locations
Gaussian kernel
bw = 6071