Points

Luc Anselin

THE CENTER FOR SPATIAL DATA SCIENCE
THE UNIVERSITY OF CHICAGO

http://spatial.uchicago.edu

classic point pattern analysis

spatial randomness

intensity

distance-based statistics

points on networks

Classic Point Pattern Analysis

- Classic Examples

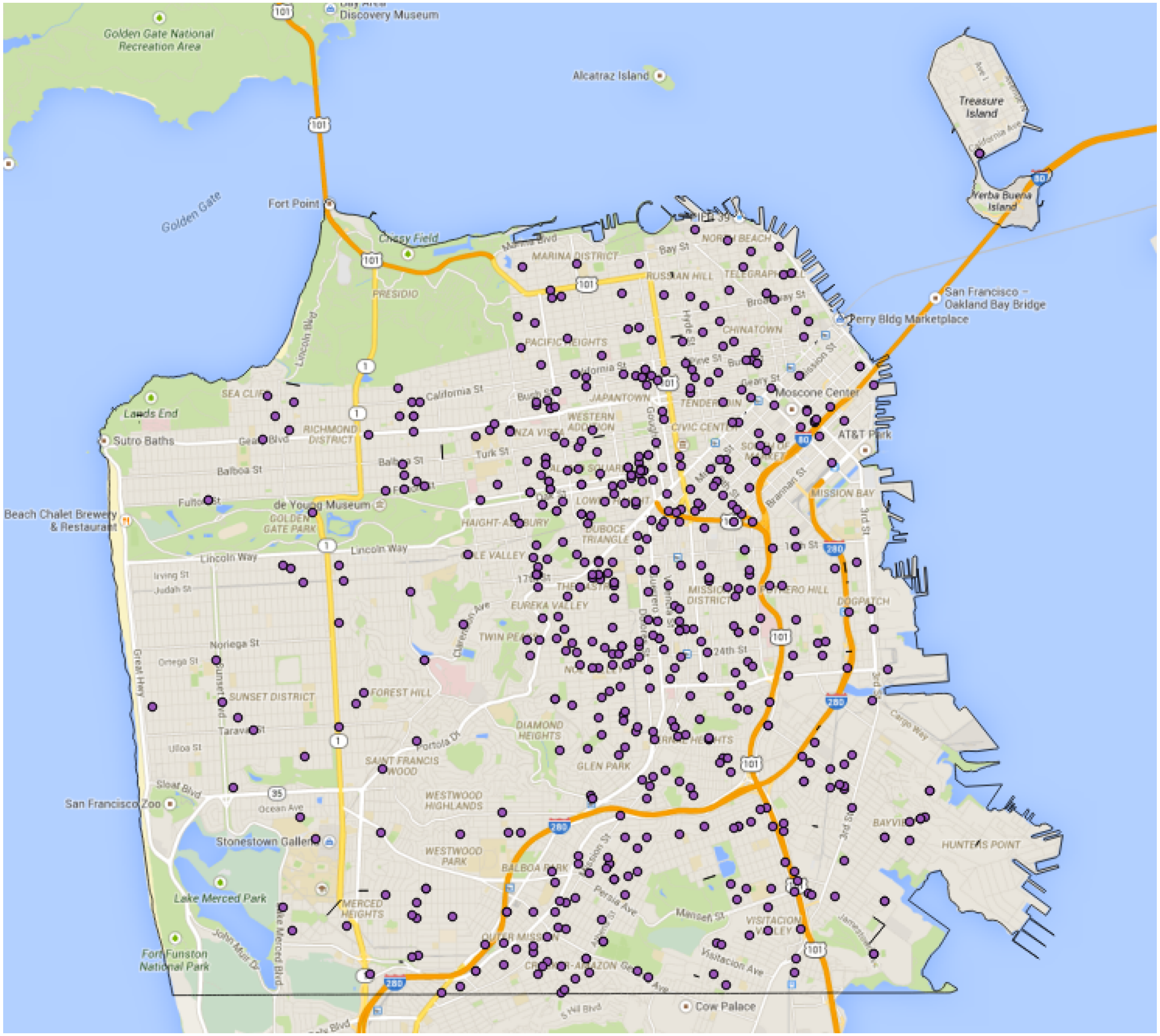
    forestry, plant species, astronomy

    locations of crimes, accidents

    locations of persons with a disease

    facility locations (economic geography)

    settlement patterns



SF car thefts, Aug 2012

- 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

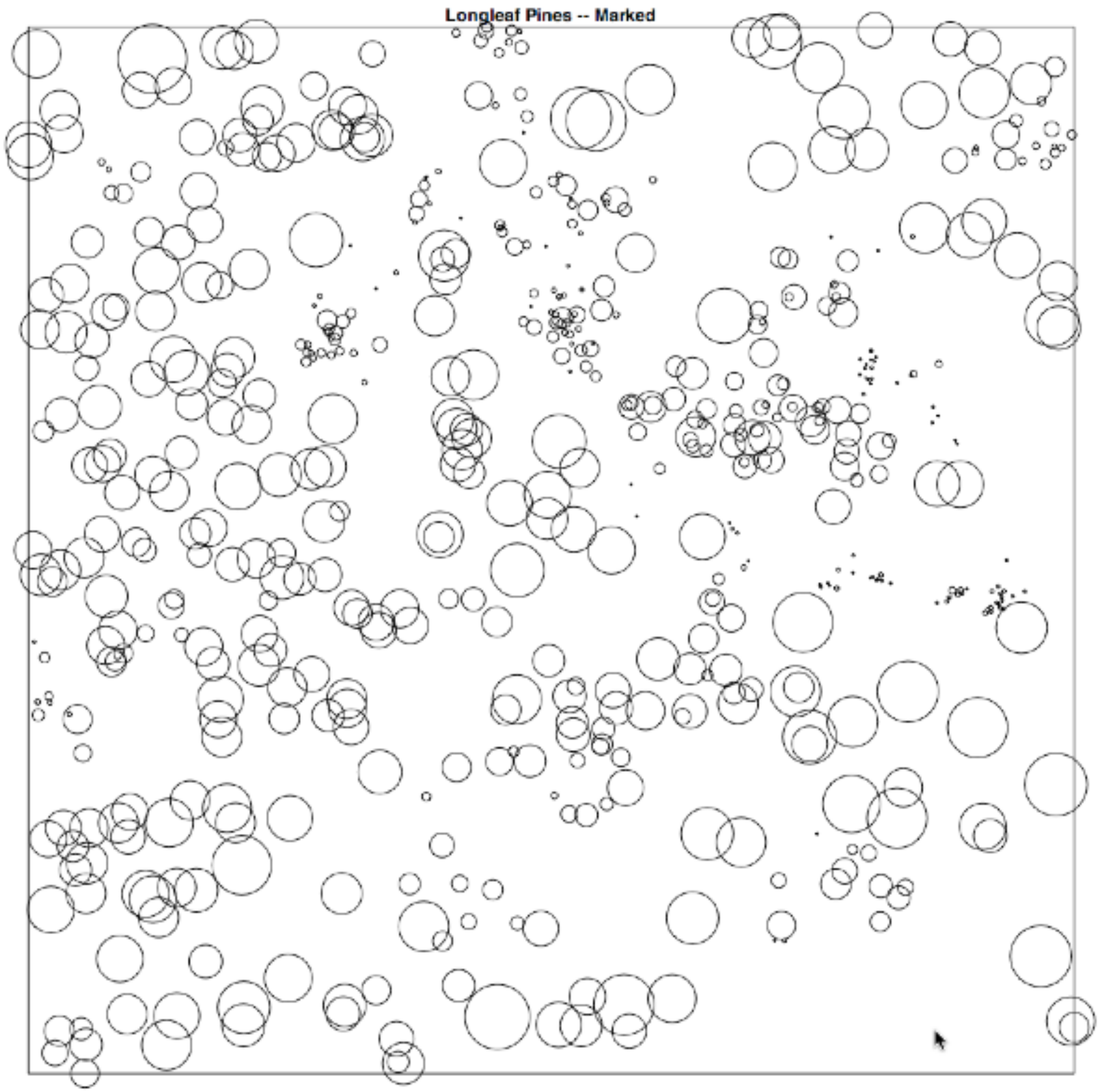
- Marked Point Pattern

    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

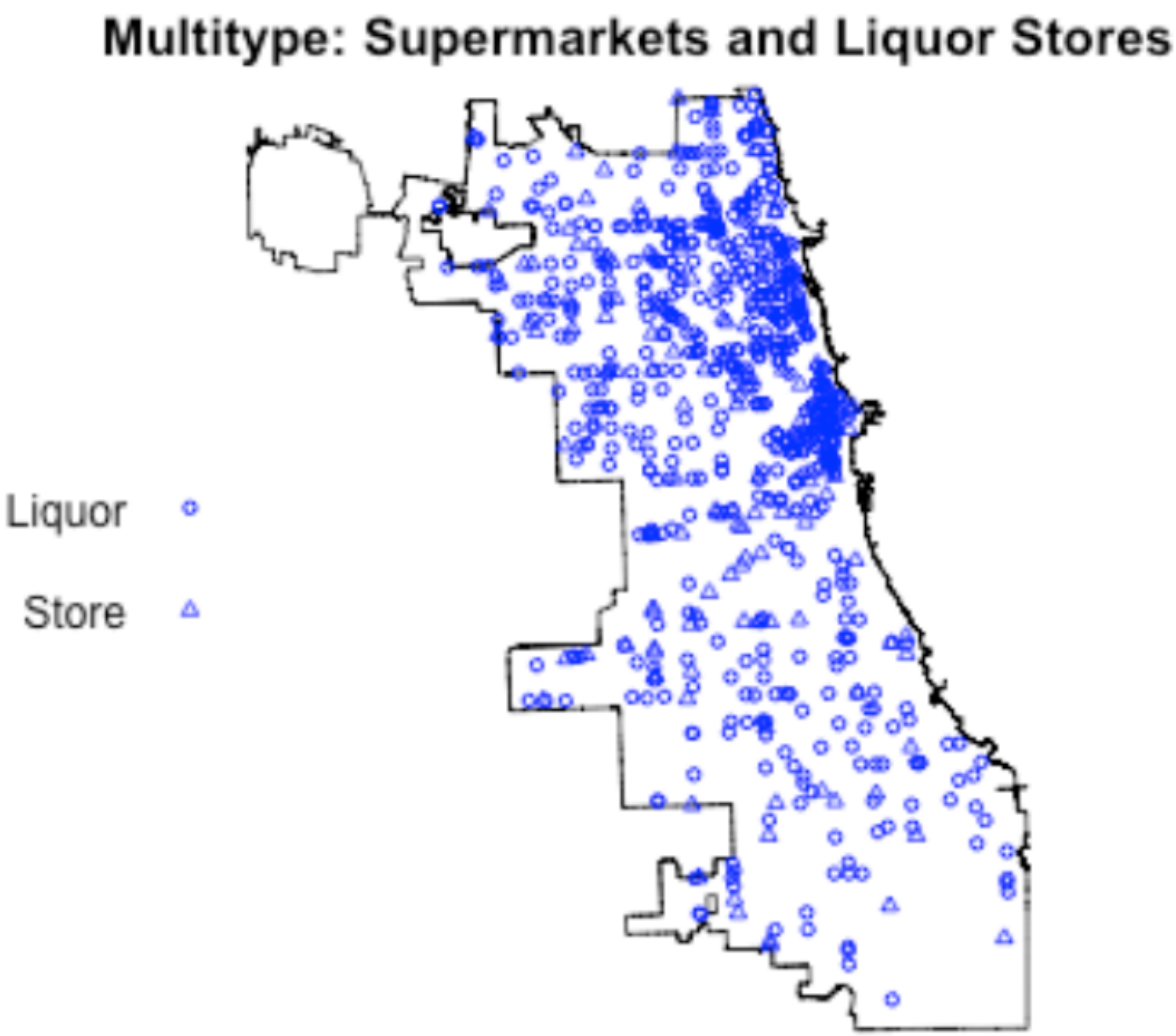
    multiple categories of events in one pattern

    research questions:

        patterning within a single type

        association between patterns in different types

        repulsion or attraction between types



Chicago multitype point pattern

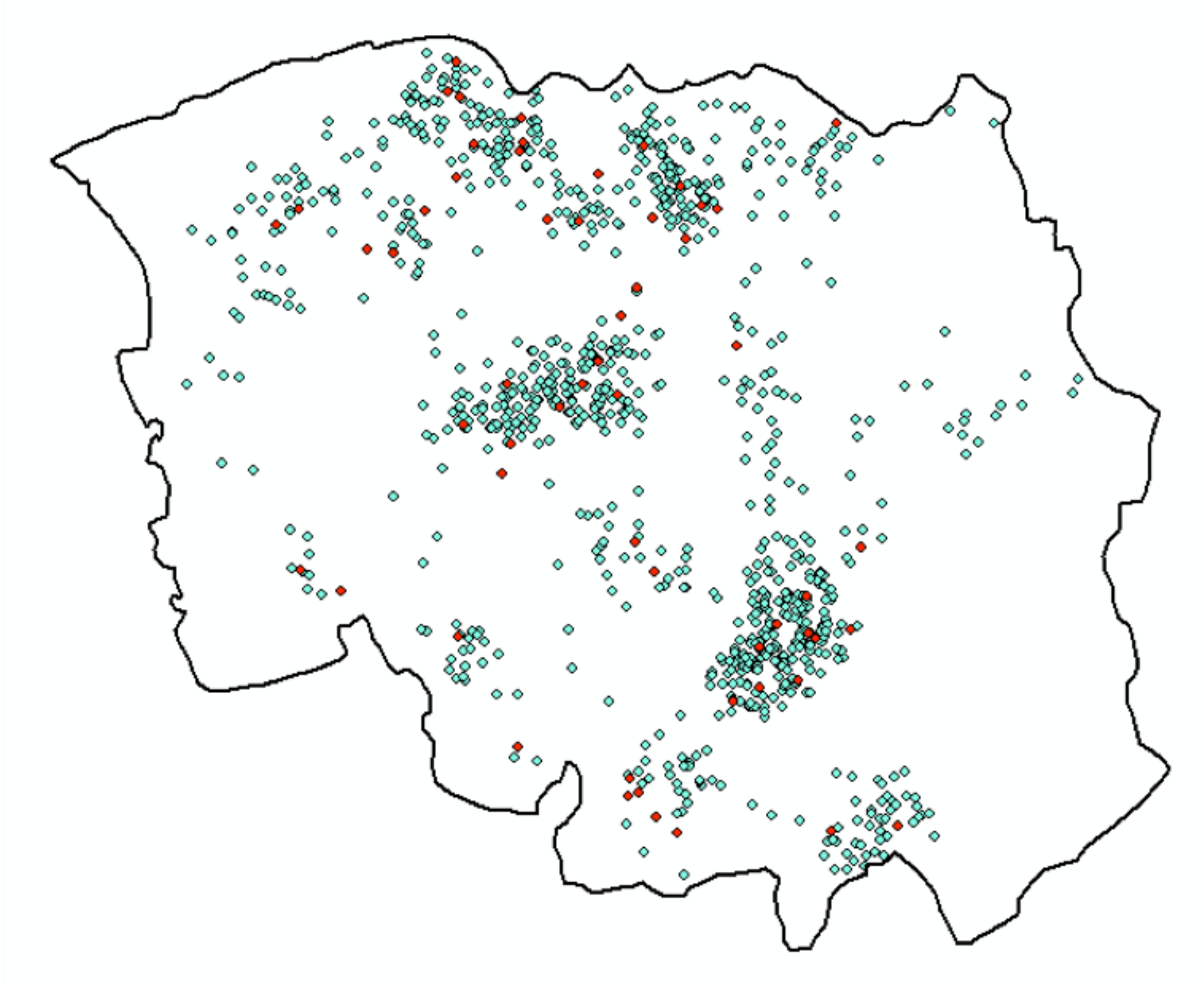
- Case-Control Design

  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

Spatial Randomness

- Complete Spatial Randomness
    - standard of reference
    - uniform distribution
        - each location has equal probability for an event
        - locations of events are independent
    - homogeneous planar Poisson process

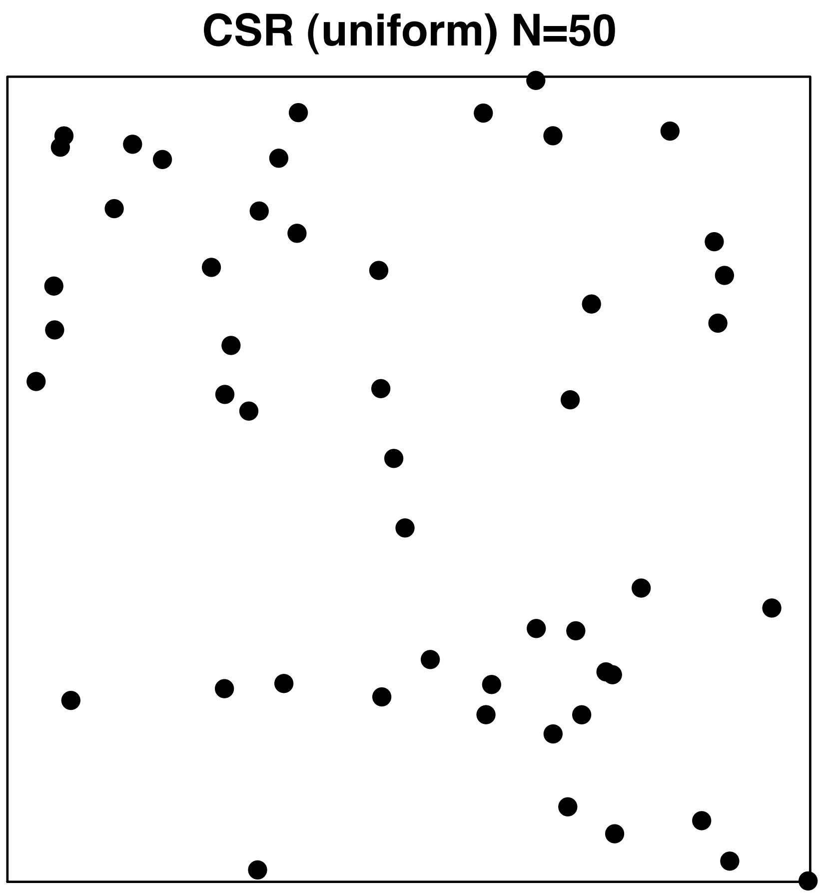
- Poisson Point Process

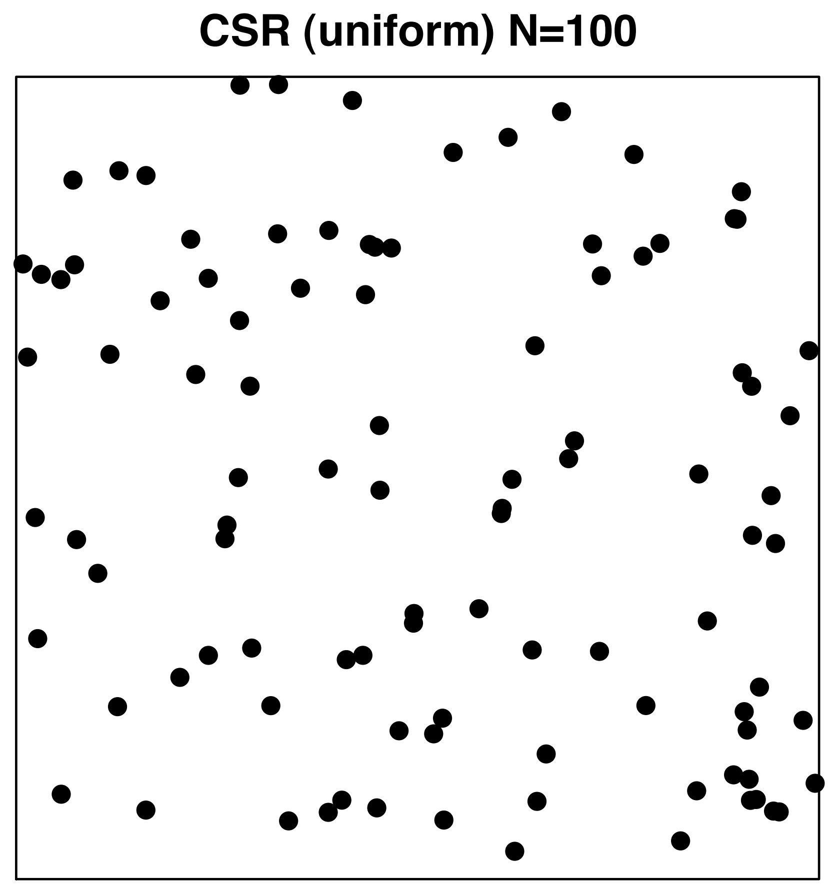
distribution for N points in area A, $N(A)$

intensity: $\lambda = N/|A|$ ($|A|$ is area of A)

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

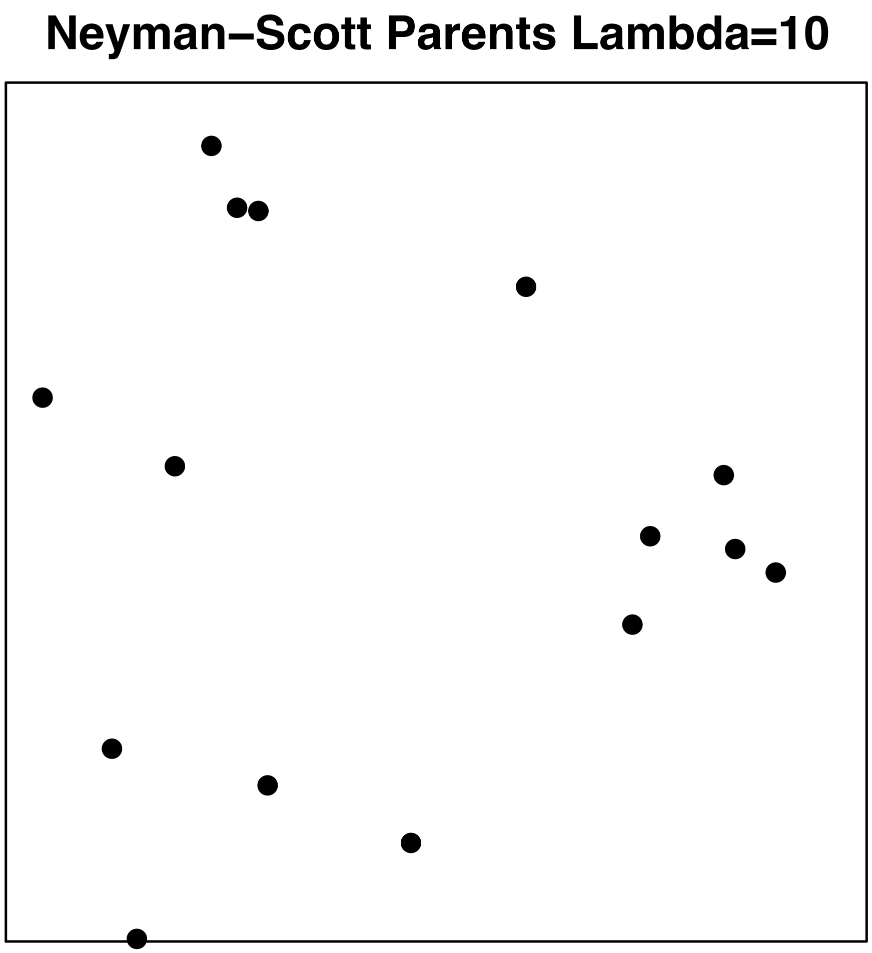
Poisson distribution: $N(A) \sim Poi(\lambda|A|)$

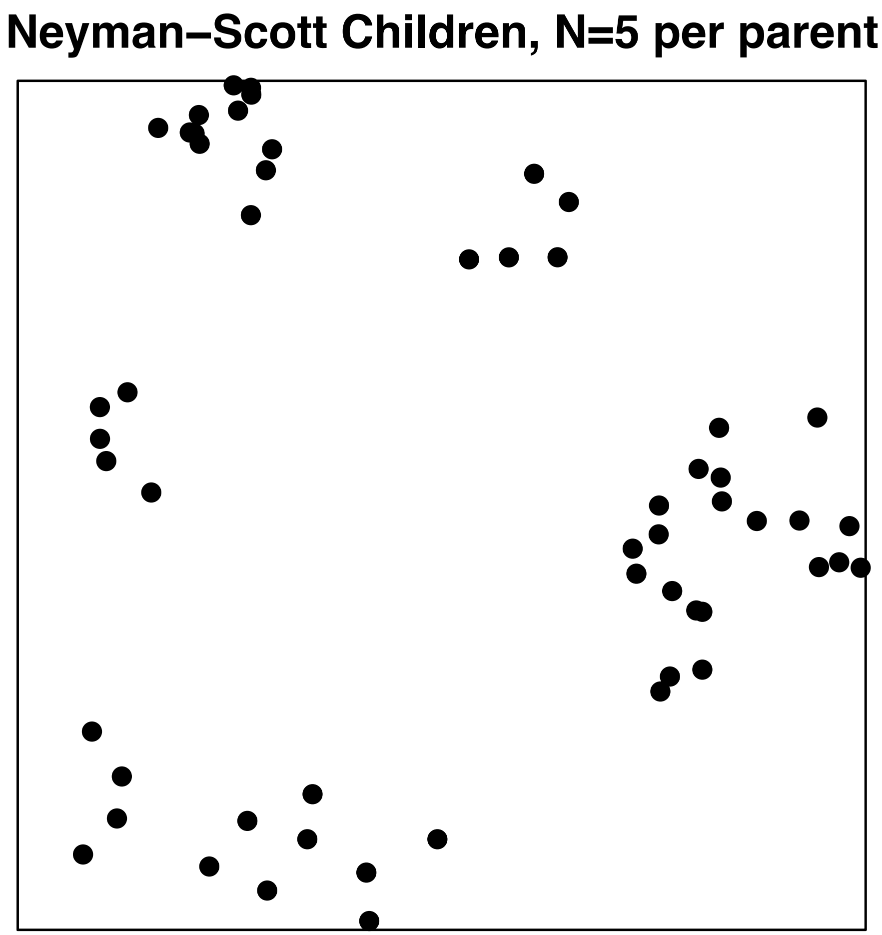




Simulated CSR - uniform with N fixed on unit square

- Contagious Point Distributions
  - two stages
    - distribution for “parents”
    - distribution for “offspring”
  - formal models
    - Poisson cluster process or Neyman-Scott process
    - Matern cluster process





realized N=15

overall λ=10x5

realized N=55

Simulated Neyman-Scott process

- Heterogeneous Poisson Process

  spatially varying intensity $\lambda(s)$

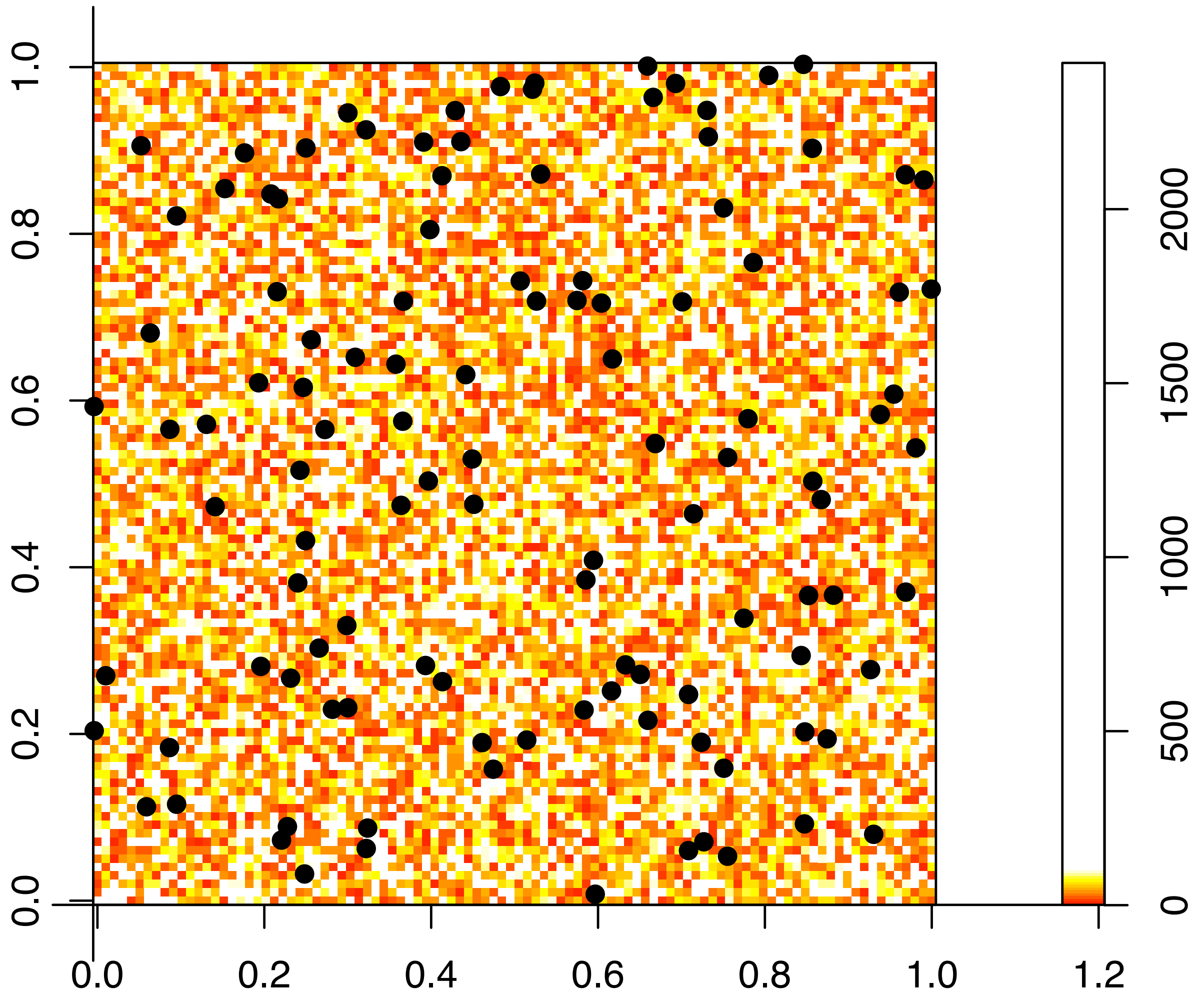
    mean intensity is integral of the location-specific intensities over the region

  source of variability

    function for $\lambda(s) = f(z)$ with covariates

    doubly stochastic process with $\lambda(s) \sim \Lambda(s)$

Log Gaussian Point Process



$\ln Z \sim N(4.1,1)$ $E[\lambda] \approx 100$ average $\lambda = 113$

Intensity

- Average Intensity

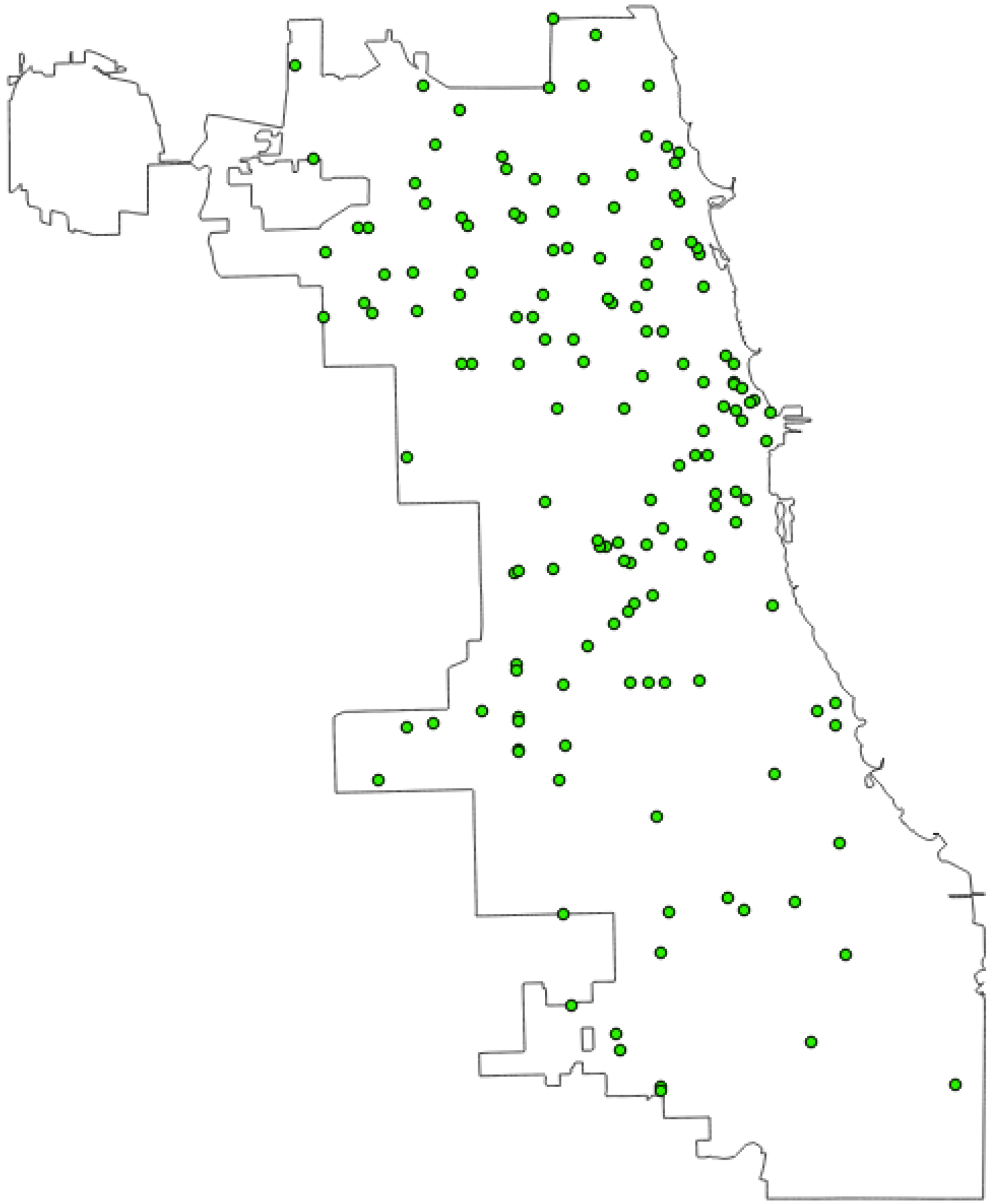
  first moment of a point pattern distribution

  number of points per unit area

  intensity: $\lambda = N/|A|$

  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



Chicago supermarkets - City boundary

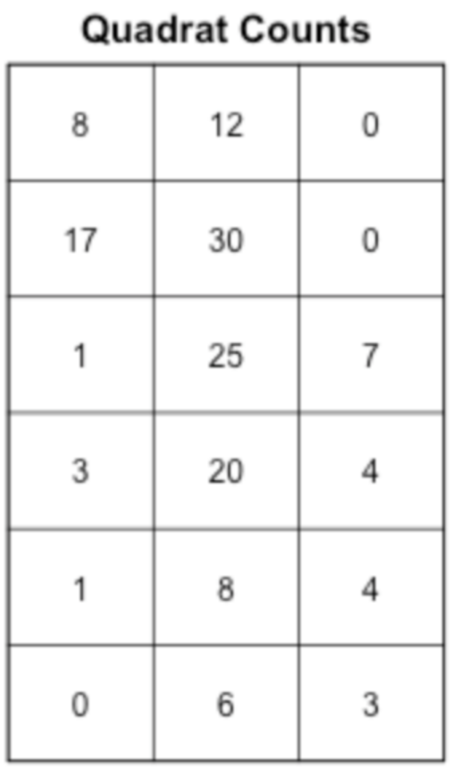
- Quadrat Counts

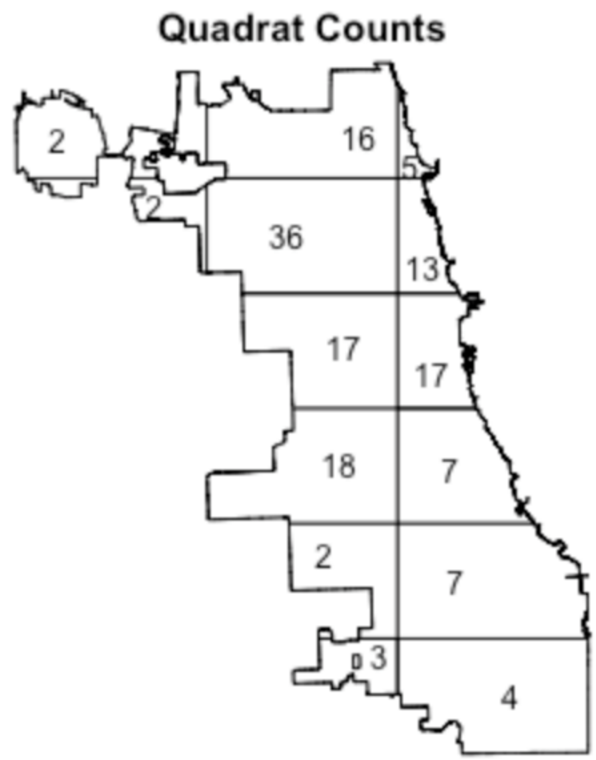
  assess the extent to which intensity is constant across space

  quadrat = polygon

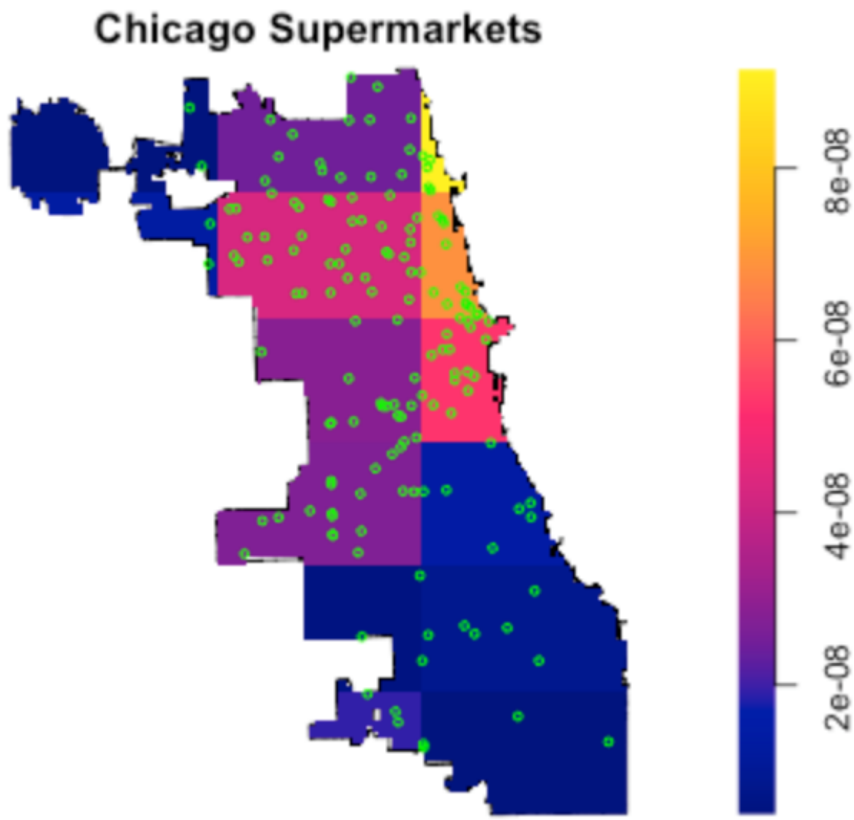
  count the points in the quadrant

  visualize counts, intensity map





Quadrat counts - alternative configurations



Quadrat count intensity graph

intensity = count / area

- Intensity Function

    spatial heterogeneity

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

    estimating $\lambda(s)$

        non-parametric kernel function

- Kernel Density Estimation

  non-parametric approach

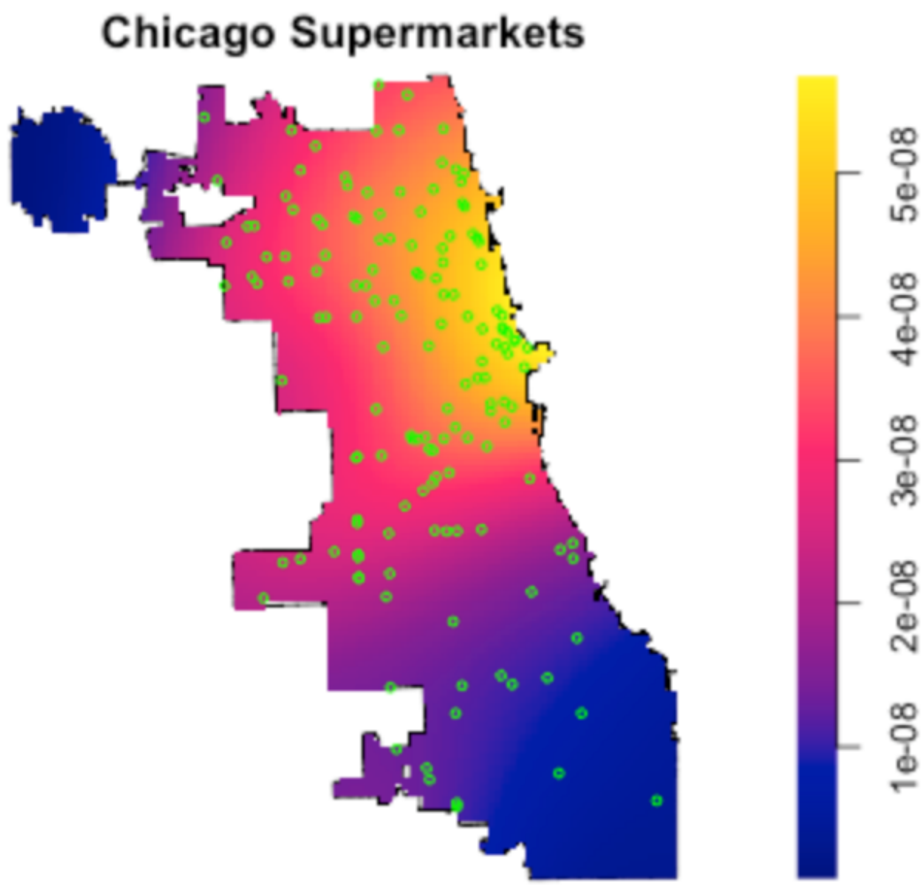
  weighted moving average of the data

- $f(u) = (1/N_b) \sum_i K[(u - u_i)/b]$

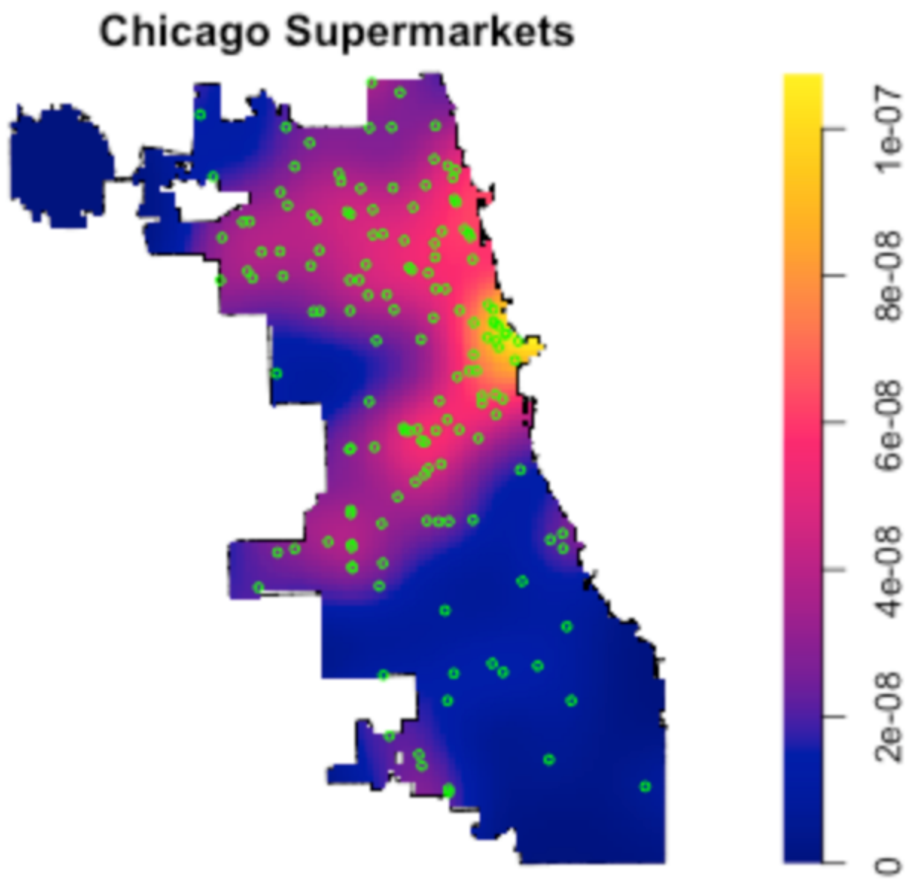
  u is any location

  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



Chicago supermarket locations
Gaussian kernel
bw = 14259



Chicago supermarket locations
Gaussian kernel
bw = 6071

Distance-Based Statistics

Nearest Neighbor Functions

- Terminology

    events and points

        event: observed location of an event

        point: reference point (e.g., point on a grid)

    distances

        event-to-event distance

        point-to-event distance

Nearest Neighbor Statistic

principle

under CSR the nearest neighbor distance between points has known mathematical properties

testing strategy = detect deviations from these properties

- Nearest Neighbor Statistic (2)
  - implementation
    - event to nearest event
    - point to nearest event
    - characterize this distribution relative to CSR
  - many nearest neighbor statistics
    - G function (event to event)
    - F function (point to event)
    - J function (combination)

- G Function - Event-to-Event Distribution

  cumulative distribution of nearest neighbor distances

  - $G(r) = n^{-1} \#(r_i \leq r)$

    proportion of nearest neighbor distances that are less than r

  plot estimated G(r) against r

  implementation: many types of edge corrections

- G under CSR

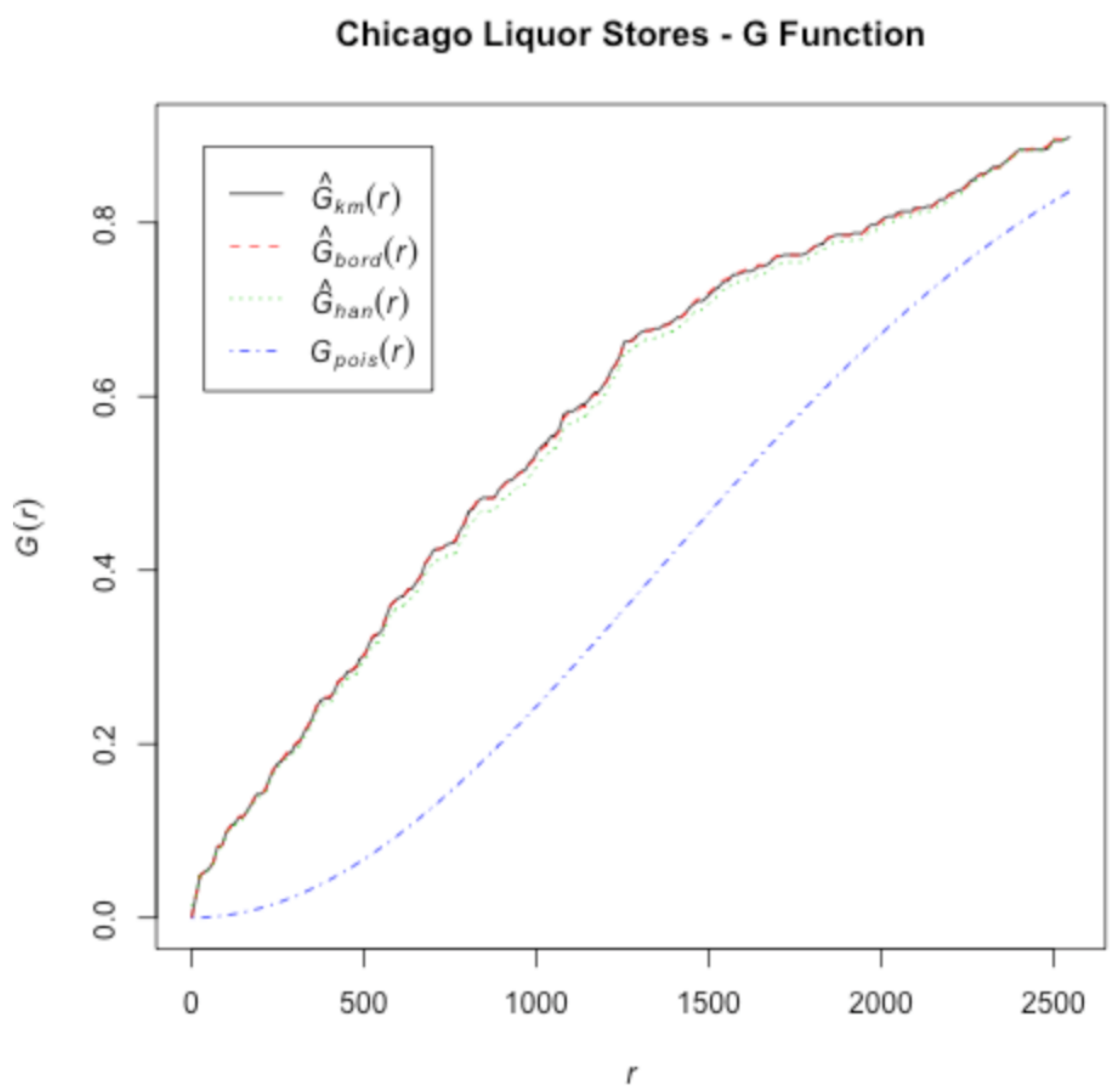
  nearest neighbor at distance r implies that no other points are within a circle with radius r

  $P[y=0]$ is $\exp(-\lambda\pi r^2)$ under Poisson distribution

  the probability of finding a nearest neighbor is then the complement of this

- $P[r_i < r] = 1 - \exp(-\lambda\pi r^2)$

  reference function, plot $1 - \exp(-\lambda\pi r^2)$ against r



G function with reference curve for CSR

- Inference

    analytical results intractable or only under unrealistic assumptions

    mimic CSR by random simulation

    - random pattern for same n

    - compute G(r) for each random pattern

    - create a simulation envelope

G function with randomization envelope using min and max for each r

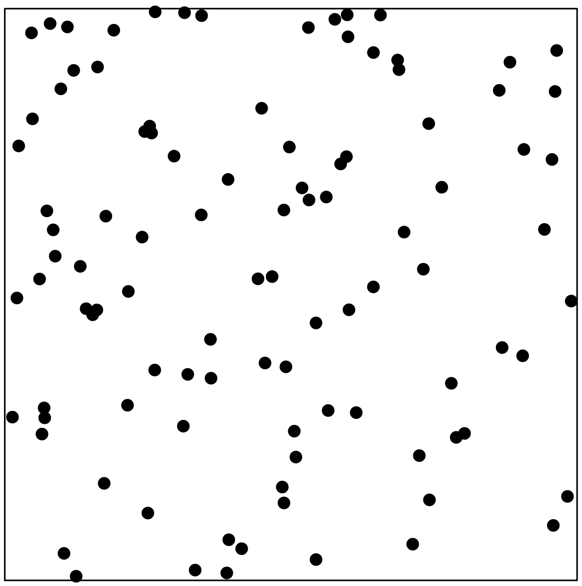
- Interpretation

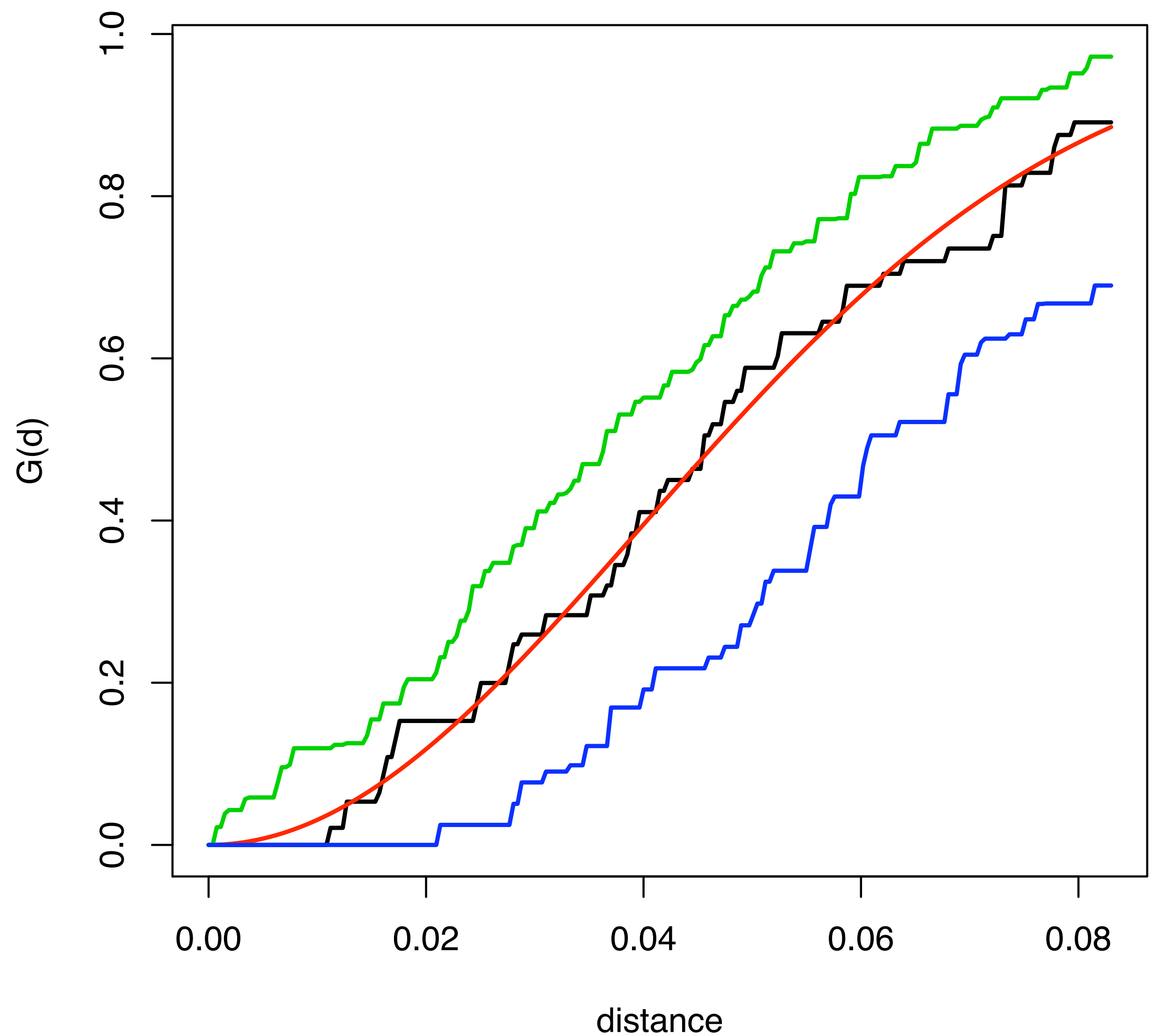
  clustering

    G(r) function above randomization envelope

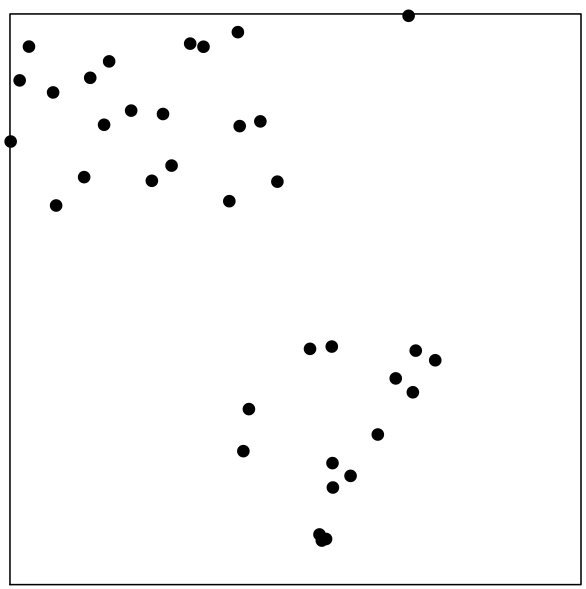
  inhibition

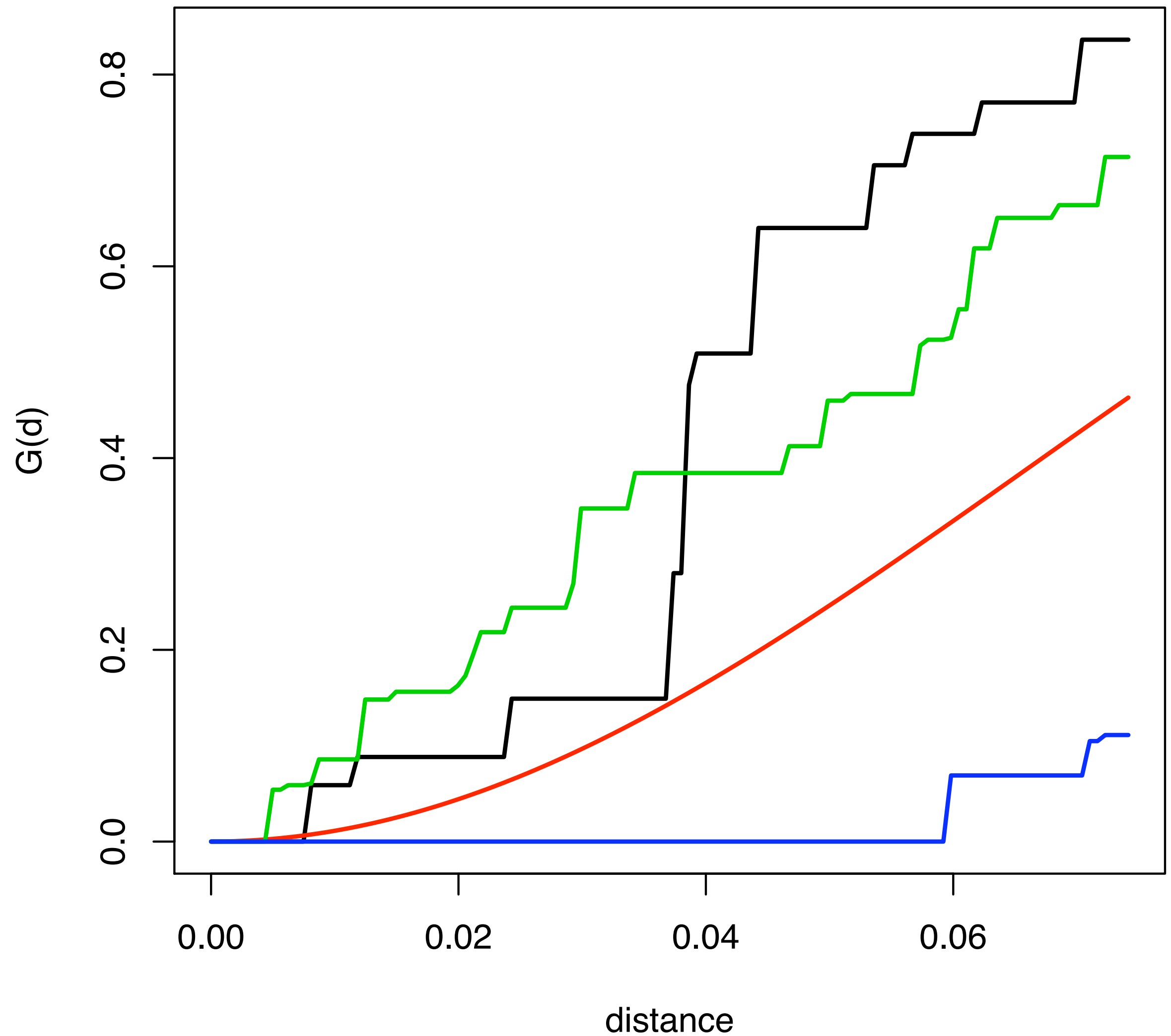
    G(r) function below randomization envelope



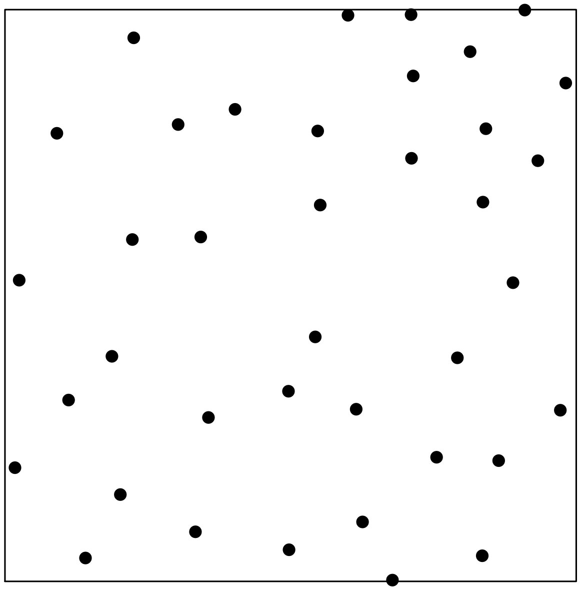


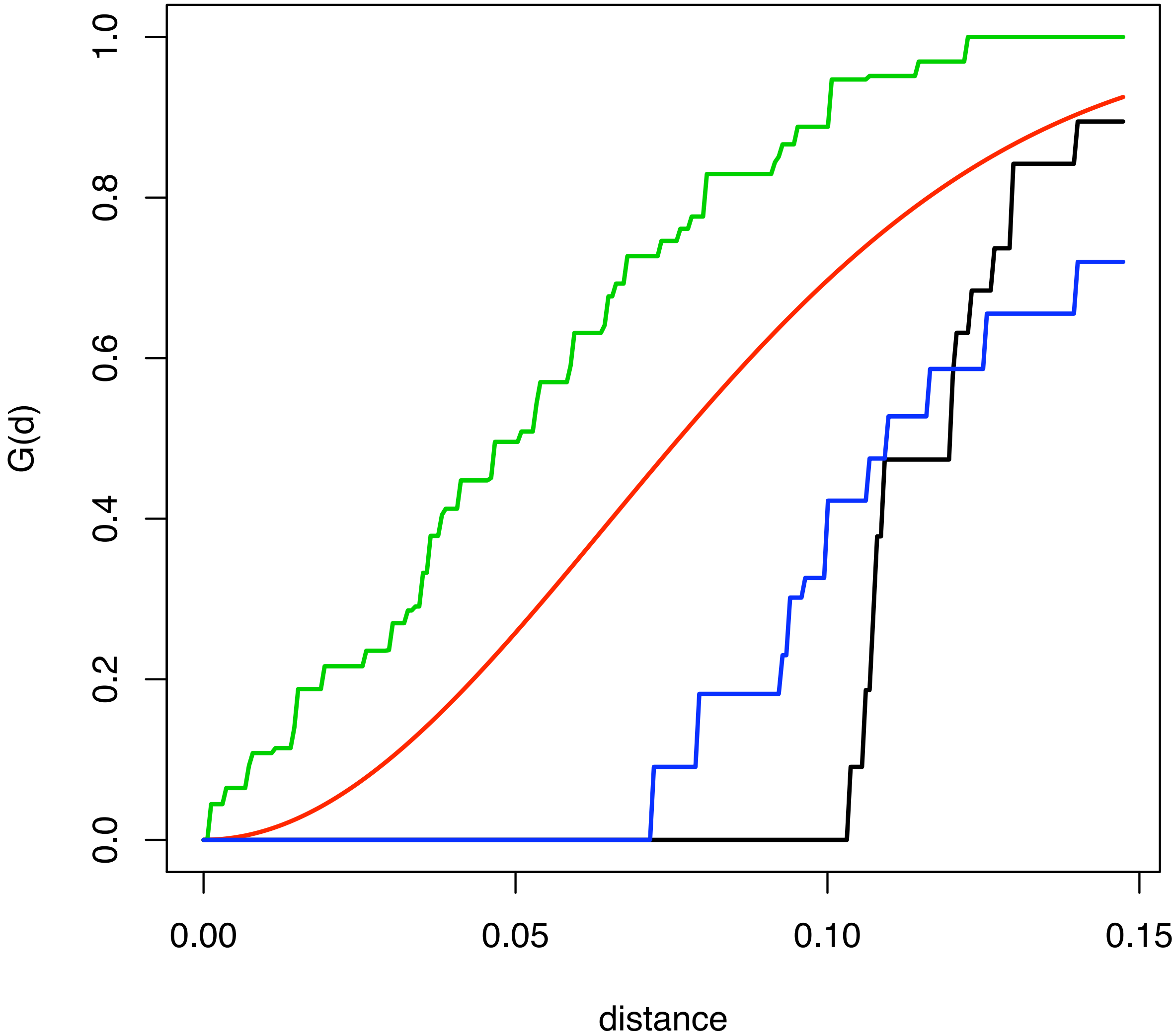
G for CSR





G for Poisson Clustered Process





G for Matern II Inhibition Process

Second Order Statistics

- Beyond Nearest Neighbor Statistics

  nearest neighbor distances do not fully capture the complexity of point processes

  instead, take into account all the pair-wise distances

  as a density function or as a cumulative density function

- Second Order Statistics

    second order statistics exploit the notion of covariance

    based on the number of other points within a given radius of a point

    pair correlation function, or g-function

    Ripley's K and Besag's L function

- Ripley's K Function

best known second order statistic

so-called reduced second order moment

$\lambda K(r) = E[N_0(r)]$

$E[N_0(r)]$ is the expected number of events within a distance r from an arbitrary event

$K(r) = \lambda^{-1} E[N_0(r)]$ is the K function

Estimating the K Function

expected events within distance r

$$E[N_0(r)] = n^{-1} \sum_i \sum_{j \neq i} I_h(r_{ij} < r)$$

for each event, sum over all other events within the given distance band, for increasing distances

cumulative function

edge corrections

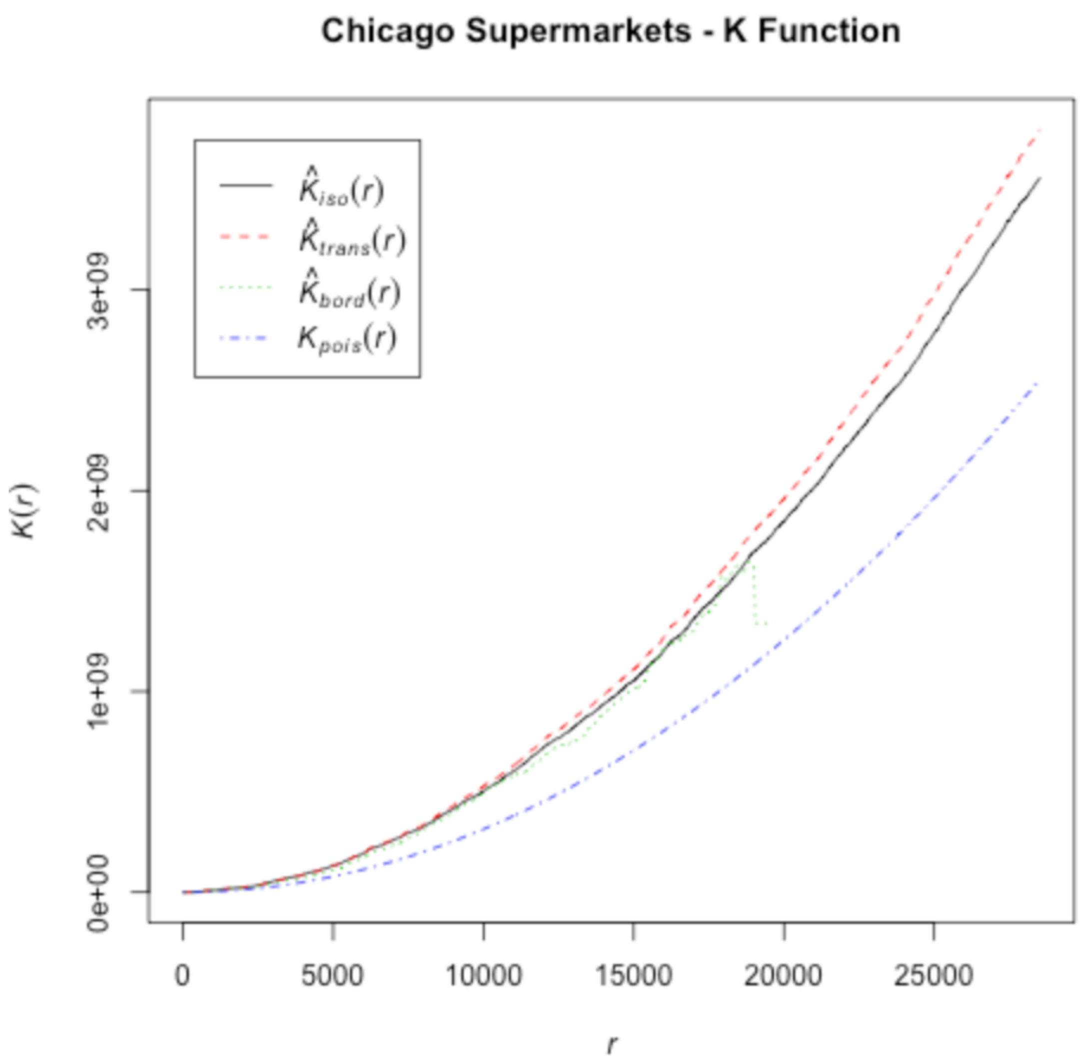
- Inference and Interpretation

  for CSR, $K(r) = \pi r^2$

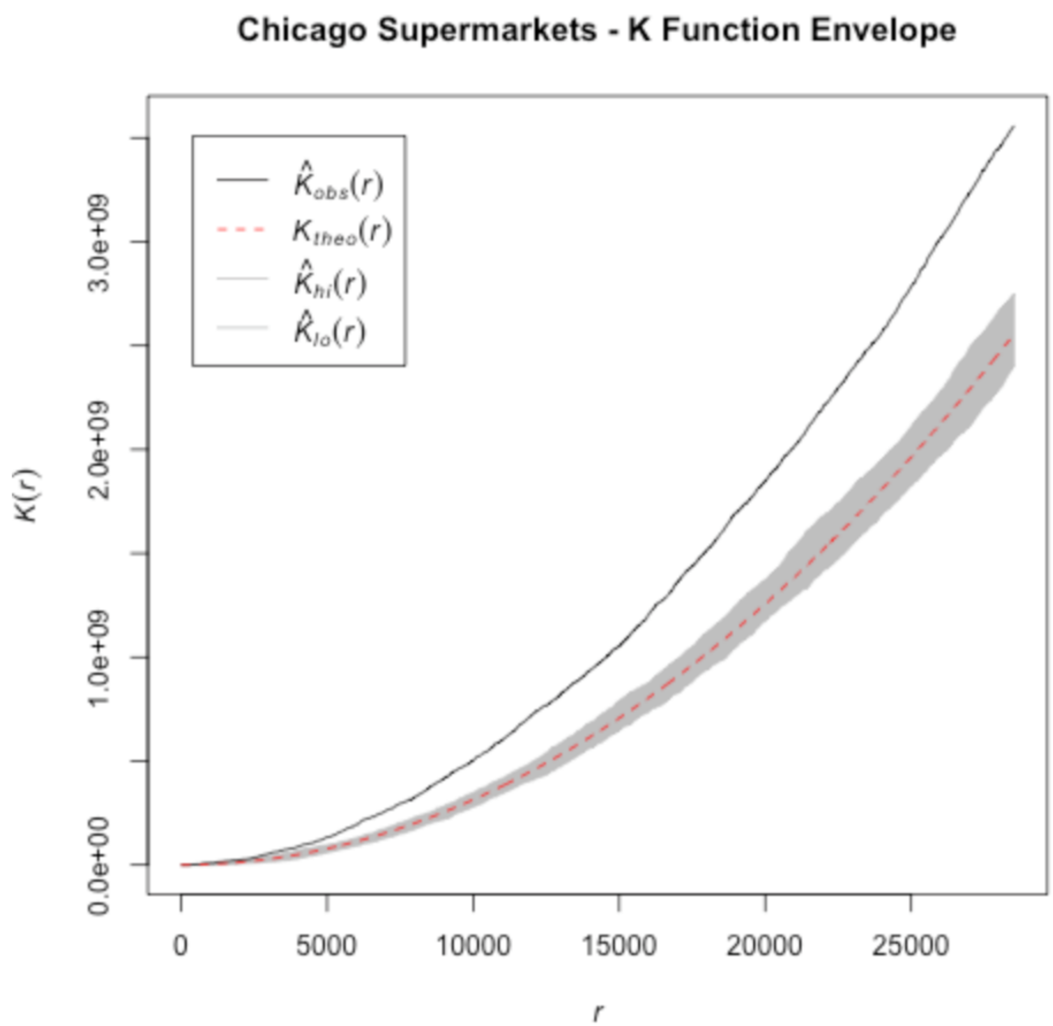
  $K(r) > \pi r^2$ implies clustering

  $K(r) < \pi r^2$ implies inhibition (regular process)

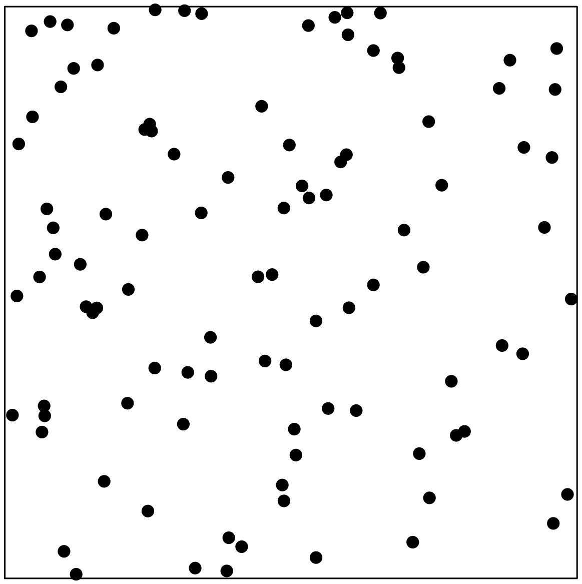
  use randomization envelope for inference

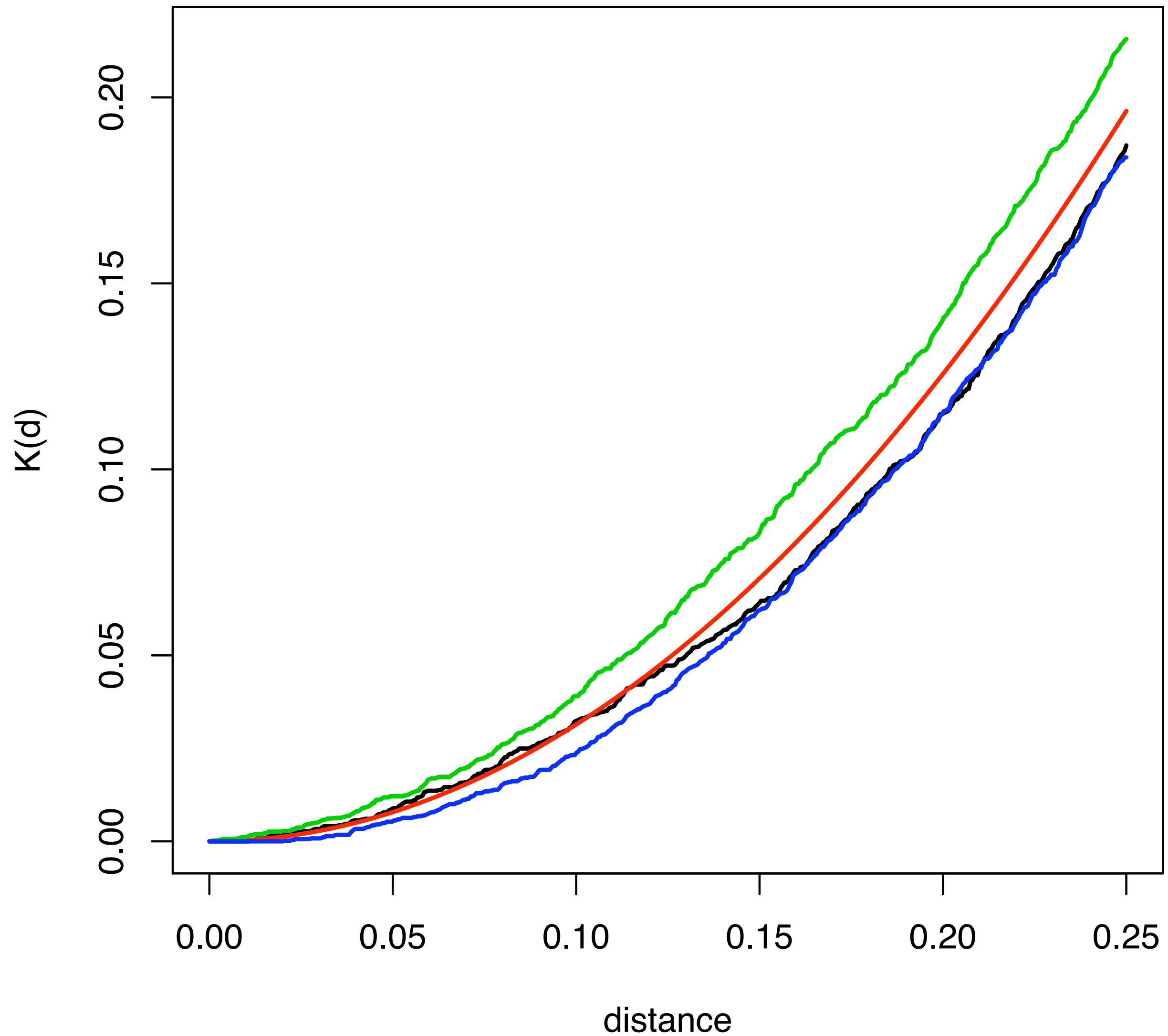


K function with reference line for CSR

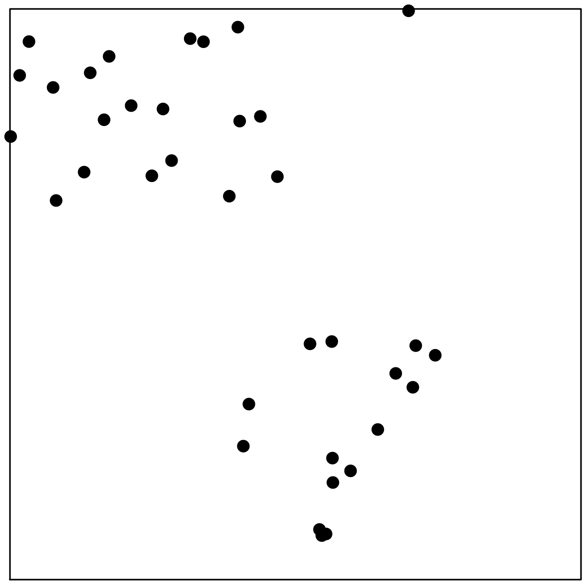


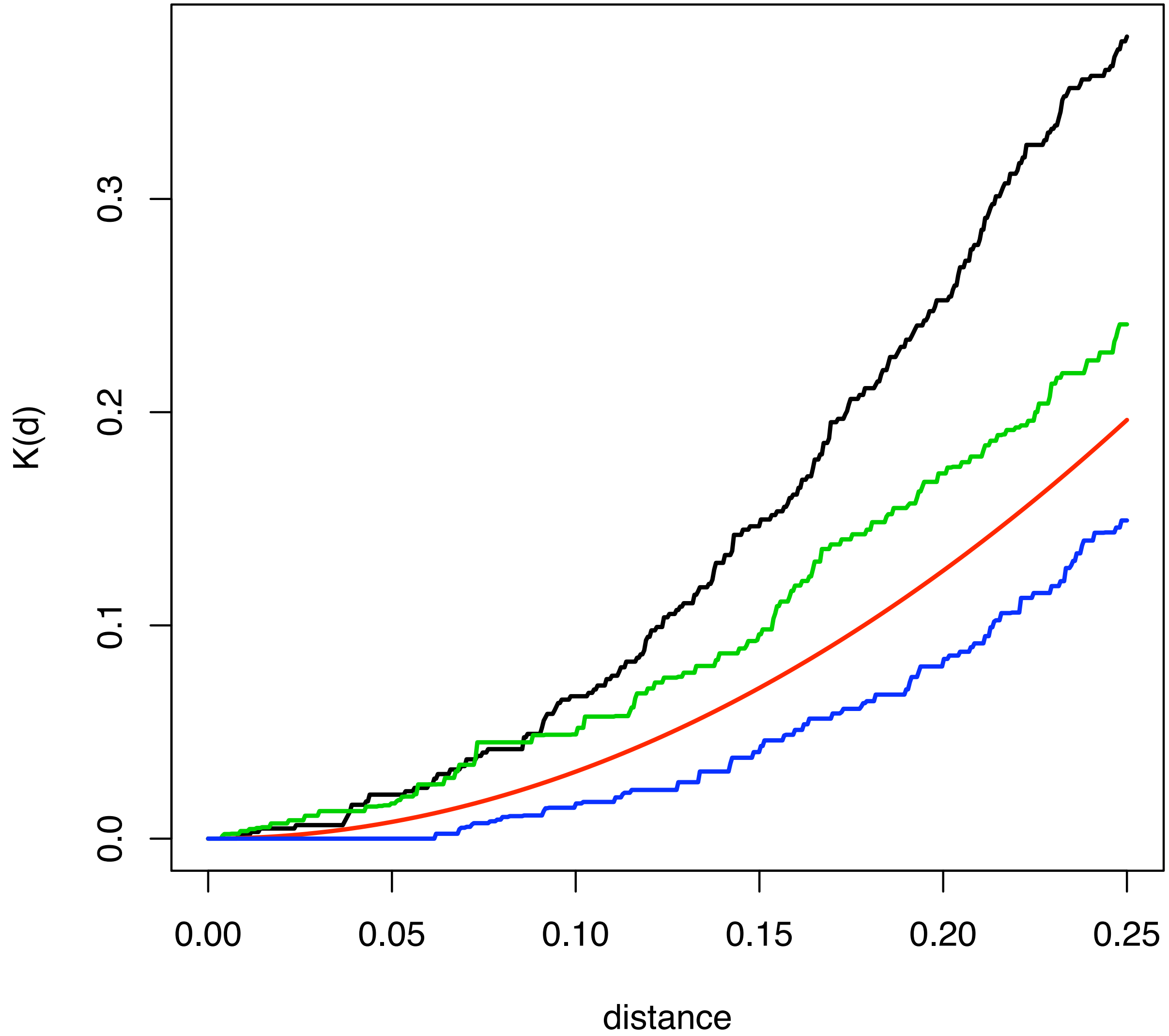
K function with randomization envelope using min and max for each r



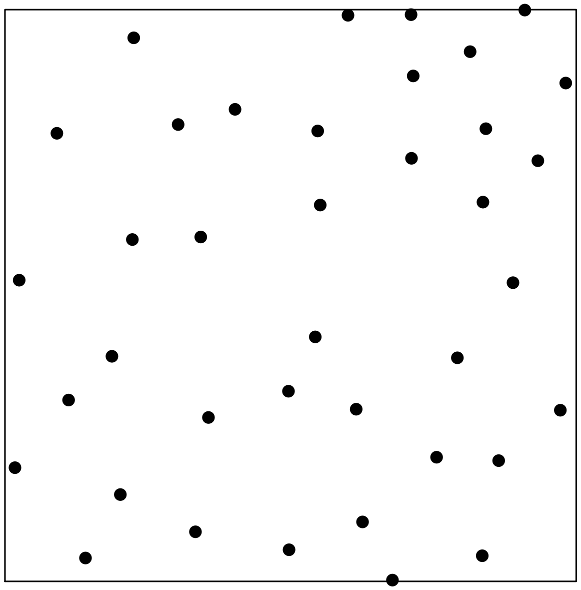


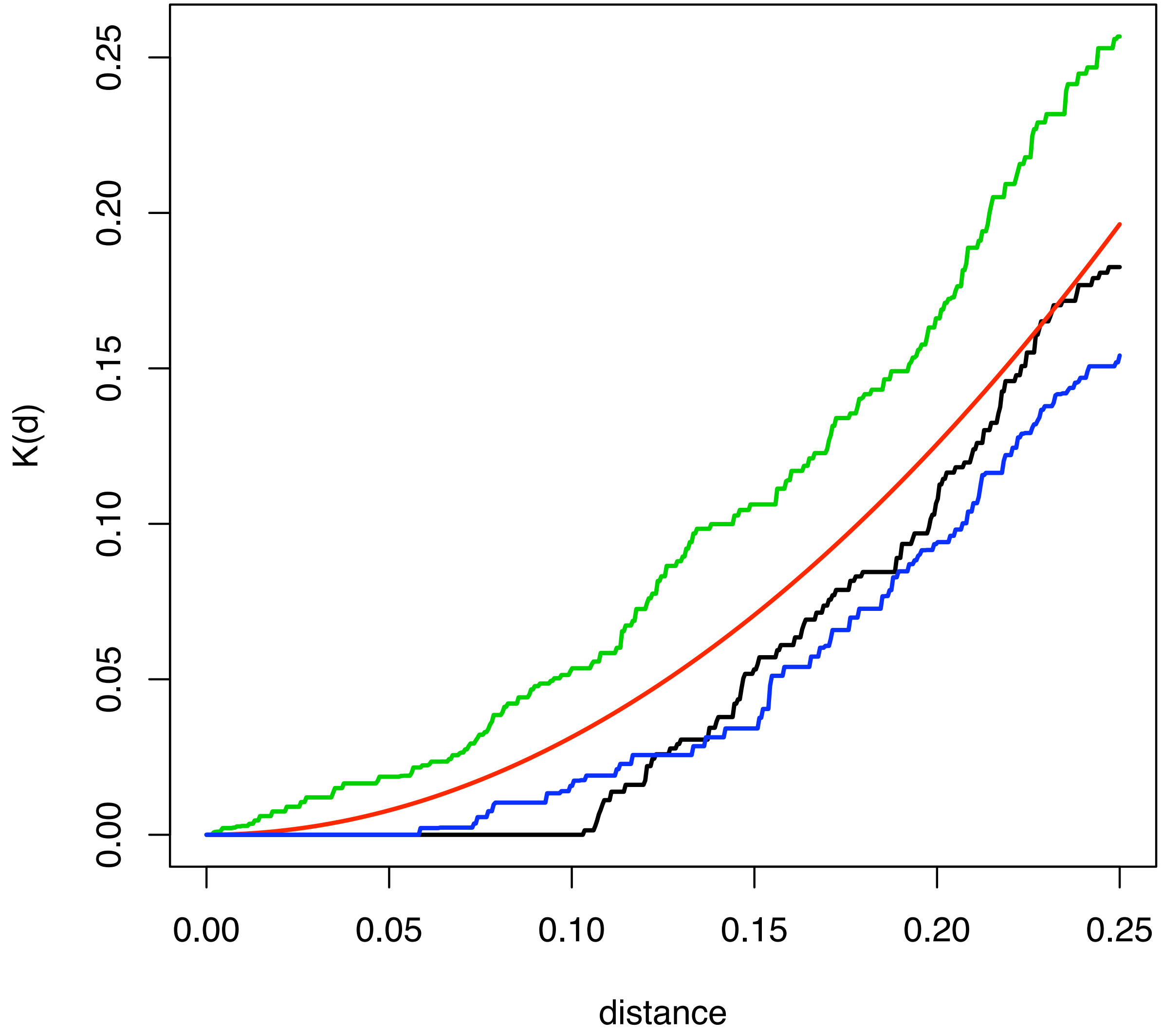
K for CSR





K for Poisson Cluster Process





K for Matern II Inhibition Process

Points on Networks

- Points on a Network

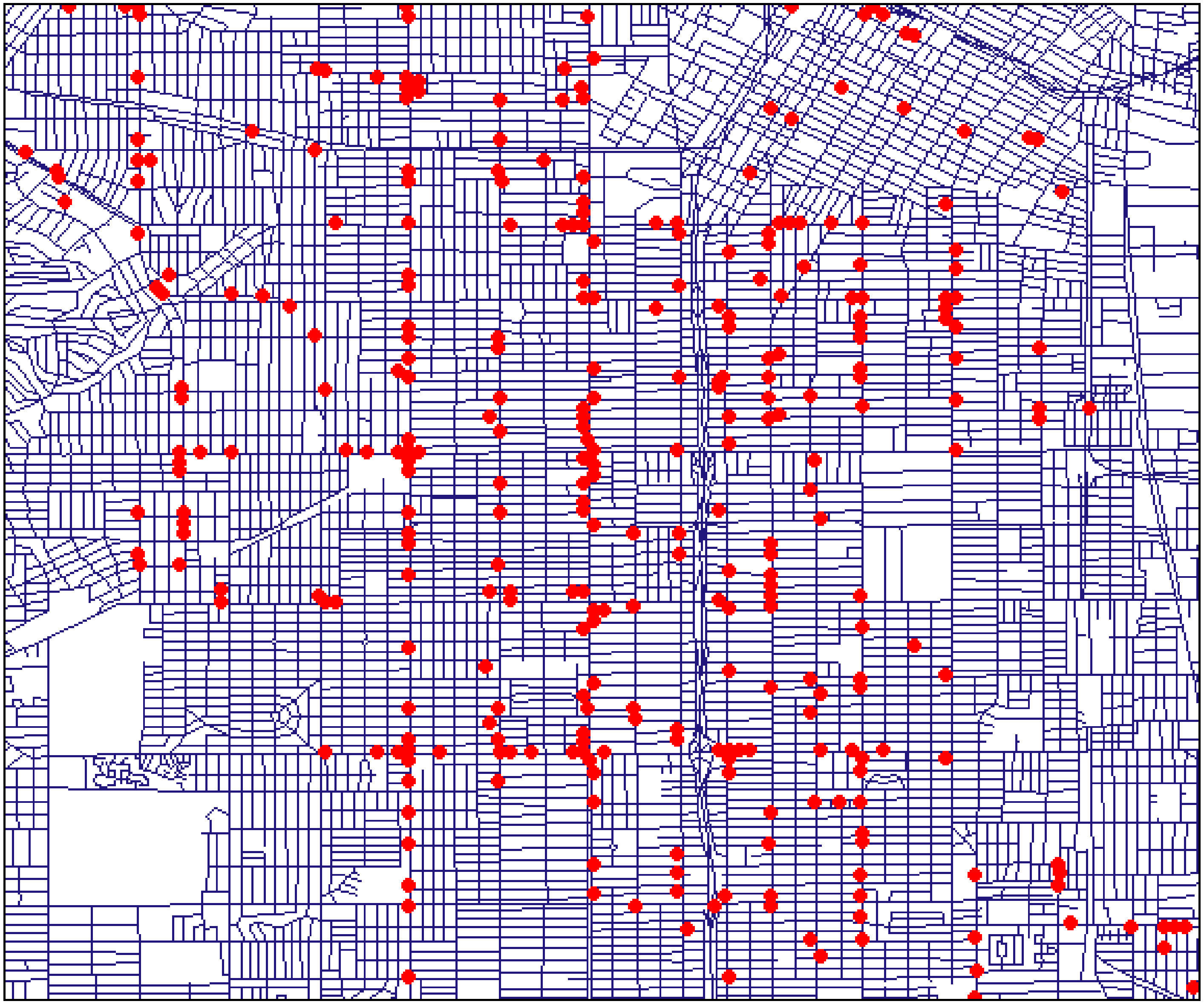
  - realistic locations

    - events located on actual network, not floating in space

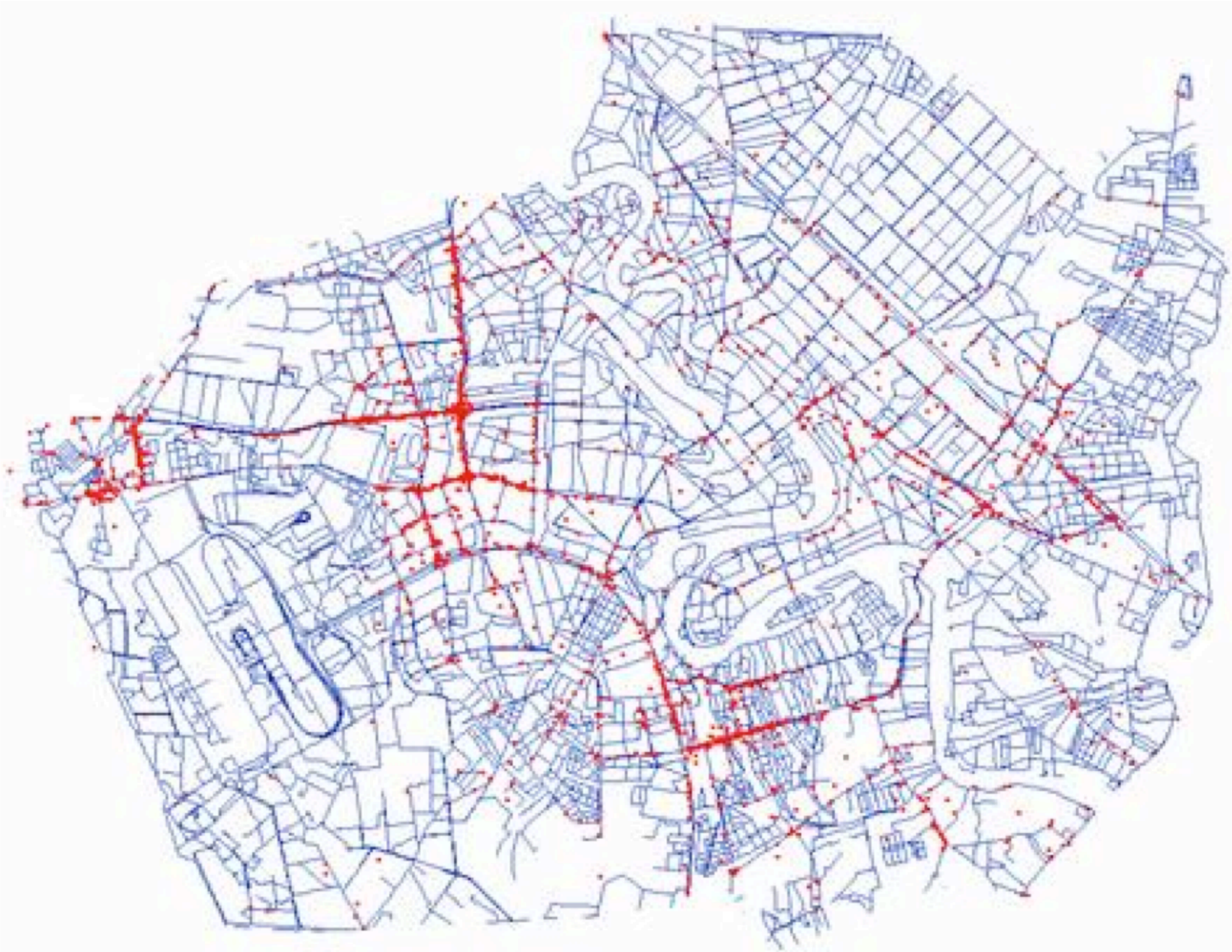
  - network distance

    - replaces straight line distance

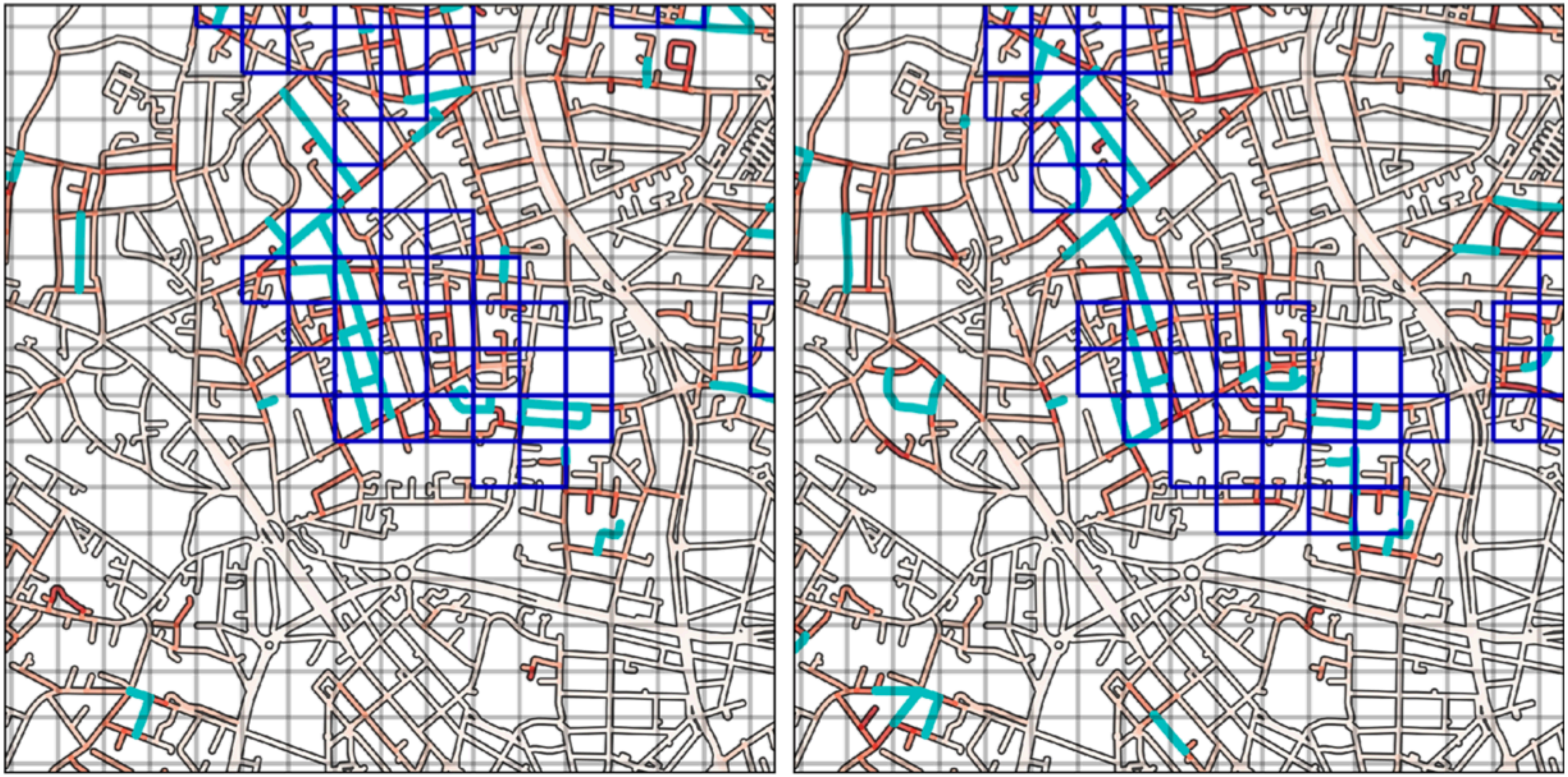
    - shortest path on the network



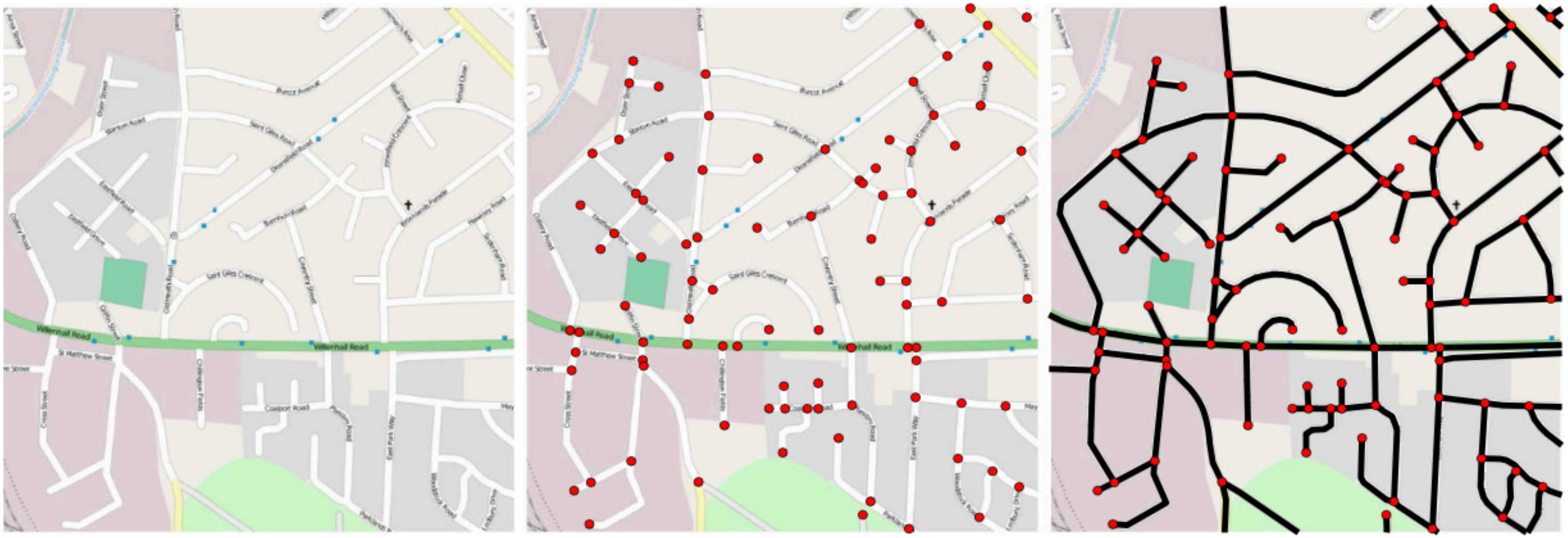
Los Angeles riot locations



Baghdad IED locations



network heat maps (kernel density)
Source: Rosser et al (2017)



from events to points on network segments

Source: Rosser et al (2017)

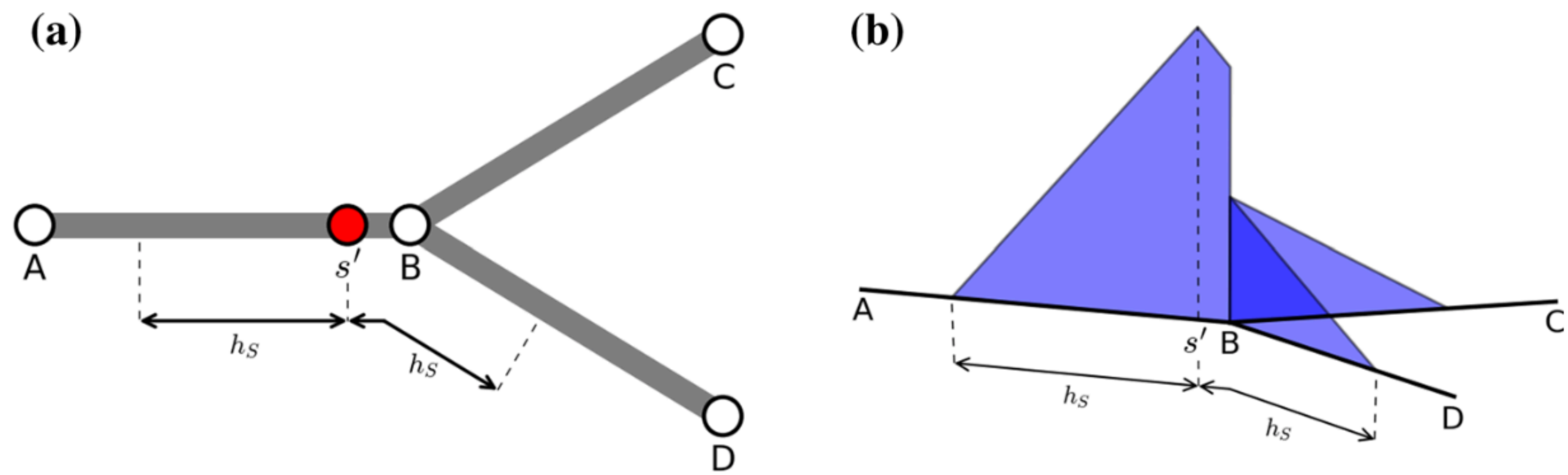


Fig. 3 Kernel calculation on networks: **a** for a kernel centred at s', a one-dimensional kernel function must be adapted to apply to each of the branches BC and BD; **b** the 'equal-split' approach, in which the remaining density at a junction is divided equally between the ongoing branches (this shows a linear kernel as used in our work)

kernel function on a network

Source: Rosser et al (2017)

Tool 1: Construction of a node-adjacency data set.
Tool 2: Assignment of a point to the nearest point on a network.
Tool 3: Aggregation of attribute values.
Tool 4: Network Voronoi diagram.
Tool 5: Random point generation.
Tool 6: Network cross *K*-function method.
Tool 7: Network *K*-function method.
Tool 8: Partition of a Polyline.
Tool 9: Assignment of polygon attributes to the nearest line segment.
Tool 10: Nearest-neighbor distance method.
Tool 11: Conditional nearest-neighbor distance method.
Tool 12: Polygon centroids generation.
Tool 13: Network Huff model.

SANET functionality

Source: Okabe et al (2016)

- Network Segments

    aggregate data by street segment

        e.g., accidents per traffic intensity

    street segments spatial weights

        define contiguity

        use shortest path distance

    network LISA