AN INTRODUCTION TO SPATIAL POINT PROCESSES

Mohammad Mehdi Moradi

Geotec: Geospatial Technologies research group,
Institute of new imaging technology (INIT),
Jaume I University, Castellon, Spain.

GEO-C

POINT PATTERNS

A '*spatial point pattern*' is a dataset giving the observed spatial locations of things or events. For instance locations of:

· Trees in a forest,
· Stars in a star cluster,
· Road accidents,
· Earthquake epicentres,
· Mobile phone calls,
· Animal sightings,
· Nuclei in a microscopic section of tissue,
· Rare disease,
· Crime,
· Gold deposits mapped in a geological survey.

GEO·C

```
R> library(spatstat)
R> plot(cells,pch=20,cex=2,main="")
R> plot(redwood,pch=20,cex=2,main="")
```
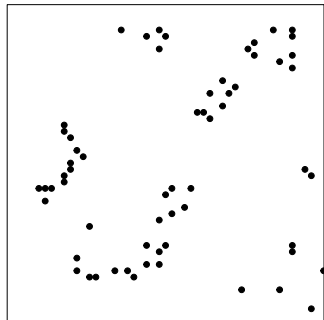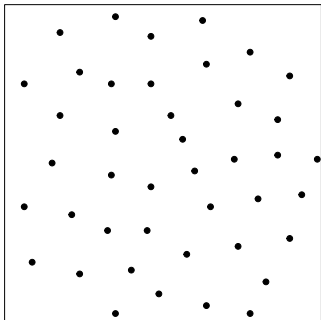

Figure: Left: Biological cells observed under optical microscopy in a histological section. Right: Locations of 62 seedlings and saplings of California redwood trees.

GEO·C

R> plot(ants,cex=2,main="",lwd=2)
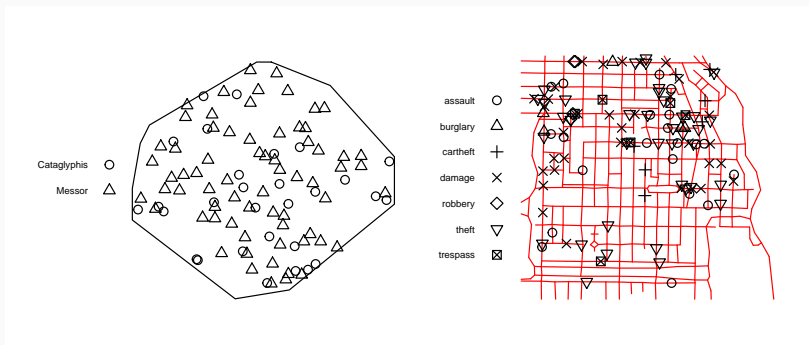R> plot(chicago,cex=2,main="",col=2,lwd=2)

Figure: Left: Spatial locations of nests of two species of ants. Right: Spatial locations of crimes reported in the period 25 April to 8 May 2002, in an area of Chicago (Illinois, USA)
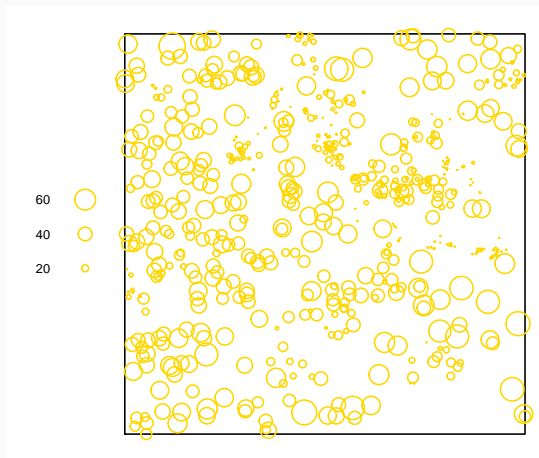
R> plot(longleaf,main="",cols="gold",lwd=2)

Figure: Locations and sizes of Longleaf pine trees. A marked point pattern

- 'Point data' cannot necessarily be treated as a 'point pattern'. For example, measurements of soil acidity at a series of sampling locations in a field would not normally be treated as a point pattern: the sampling locations are artificial, and irrelevant to the study of soil acidity.
- Statistical analysis of the spatial arrangement of points can reveal important features, such as a tendency for gold deposits to be found close to a major geological fault, or for cases of a disease to be more prevalent near a pollution source, or for bird nests to maintain a certain minimum separation from each other.
- A dataset may also include covariates, i.e. any data that we treat as explanatory, rather than as part of the 'response'. Type of covariates vary from a dataset to another one. For instance, in case of studying crime point patterns, distance from data points to some particular places might be considered as covariates.

- Points may be observed in different spaces. They might happen in $\mathbb{R}^2$ such as *"cells"* and *"redwood"*. Some may happen in one-dimensional space such as *"Chicago crime data"*. Another type could be those recorded in space and time such as earthquake epicentre locations and times.
- Working with point patterns, we assume the location of points are already projected. If not, please project them before any analysis.
- When working with multi-type point patterns such as *"ants"*, make sure the location of all types are projected in a same ref system.
- In case of point patterns on network such as *"Chicago crime data"*, make sure all data point and network itself are projected using a same ref system.
- Mathematically speaking, different spaces may demand their own special attention.
- Other types of point patterns include **replicated patterns** and **point patterns on sphere**.

STATISTICAL METHODOLOGY FOR POINT PATTERNS

Aiming at analysing a point pattern, one might be interested in:

· Spatially varying distribution of the points. Is there any hot-spot? How events use space?

· The type of interaction between points.

· Is there any evidence of clustering?

· How is the spatially variation of marks?

· How covariates affect the spatially varying distribution of points?

· Are different types of points affected by each other?

A *point process* is a random mechanism whose outcome is a point pattern. If the experiment could be repeated under identical conditions, the observed point pattern would be different each time.

· **Important note**: If the researcher concludes that "the points are completely random" or "the points are uniformly spread", this cannot be a literal description of the point pattern dataset. Data are numbers that were observed and recorded: **they are fixed, not random**. A point in a point pattern is a discrete object, and cannot be 'spread uniformly' like butter. Instead, the conclusion is that the data points were generated by some mechanism which was completely random, or which spread the points uniformly, etc.

We assume $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ as a point pattern generated by a point process X. The number n of points in the pattern is not fixed in advance, and may be any finite nonnegative number including zero.

A 'completely random' point process is characterised by two key properties:

Homogeneity: the points have no preference for any spatial location;

Independence: information about the outcome in one region of space has no influence on the outcome in other regions.

It is a realistic model of some physical phenomena, such as radioactivity, rare events, and extreme events. It serves as a benchmark or standard reference model of completely random patterns, against which other patterns can be compared.

A point process X is usually labeled by an intensity function λ which governs the spatially varying distribution of points so that,

$$\mathbb{E}[N(X \cap A)] = \int_A \lambda(u)\mathrm{d}u, \tag{1}$$

where N is a count function.

Poisson process: X is a Poisson process with intensity function $\lambda > 0$ on $\mathbb{R}^2$ if

- For any $B \subset \mathbb{R}^2$, the number of points of X in B is Poisson distributed with parameter $\int_B \lambda(u)\mathrm{d}u$.
- The number of points in $B_1, B_2, \ldots, B_m \subset \mathbb{R}^2$ are independent random variables.

In order to understand the spatially varying distribution of the points, the key point is to estimate the intensity function λ.

Kernel smoothing:

Simple explanation: place one square of chocolate on each data point. Using a hair dryer we apply heat to the chocolate so that it melts slightly. The result is an undulating surface of chocolate; the height of the surface represents the estimated intensity function of the point process. The total mass of chocolate is unchanged.

Mathematically speaking:

Consider a point pattern **x** observed in window $W \subset \mathbb{R}^2$. The usual kernel estimators of the intensity function are, with uniform correction,

$$\widehat{\lambda}^U(u) = \sum_{i=1}^n \frac{\kappa_\sigma(u - x_i)}{C_{W,\sigma}(u)}, \qquad u \in W, \tag{2}$$

and with Jones-Diggle correction,

$$\widehat{\lambda}^{JD}(u) = \sum_{i=1}^n \frac{\kappa_\sigma(u - x_i)}{C_{W,\sigma}(x_i)}, \qquad u \in W, \tag{3}$$

where σ is the smoothing bandwidth parameter, κ is a kernel density function and

$$C_{W,\sigma}(u) = \int_W \kappa_\sigma(u - v) \mathrm{d}v, \qquad v \in W. \tag{4}$$

Properties:

· Intensity estimate (2) is unbiased if the true intensity is constant, i.e.
  $\mathbb{E}[\widehat{\lambda}^U(u)] = \lambda$.

· Intensity estimate (3) provides mass conservation, i.e. $\int_W \widehat{\lambda}^{ID}(u) = n$

Note: Smoothing parameter σ plays an important role in estimating intensity function using kernel smoothing.

Note: Estimators (2) and (3) are implemented in function 'density.ppp' in R package *'spatstat'*.

· Likelihood cross-validation: Maximising

$$\sum_{i=1}^{n} log(\widehat{\lambda}^{-i}(x_i)) - \int_W \widehat{\lambda}(u)du, \tag{5}$$

where $\widehat{\lambda}^{-i}$ is the corresponding "leave-one-out" at x_i, defined by omitting the contribution from x_i from the sum in (2) or (3) accordingly. This is implemented in *'spatstat'* by function 'bw.ppl'.

· Scott's rule of thumb for 2 dimension:

$$\sigma_x = (4n)^{-1/6}s_x, \qquad \sigma_y = (4n)^{-1/6}s_y, \tag{6}$$

where 's' is the sample standard deviation of the Cartesian coordinate values for the data locations. This is implemented in *'spatstat'* by function 'bw.scott'.

```
# redwood
R> d <- density.ppp(redwood,sigma = bw.ppl(redwood))
R> plot(d,main="",ribwid=0.04,ribsep=0.02)
R> plot(redwood,pch=20,cex=2,add=T)
# ants
R> d1 <- density.ppp(ants,sigma = bw.scott(ants))
R> plot(d1,main="",ribwid=0.04,ribsep=0.02)
R> plot(ants,cex=2,add=T,lwd=2)
# chicago
R> d2 <- density.lpp(chicago,sigma = 100)
R> plot(d2,main="",ribwid=0.04,ribsep=0.02)
R> plot(chicago,pch=20,add=T,cex=1.4)
R> plot(d2,style = "w",main="",adjust = 1.3)
```
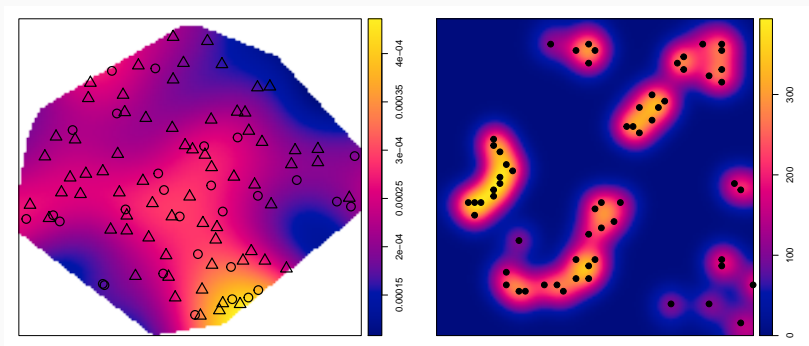
Figure: Left: Estimated intensity for *ants* data. Right: Estimated intensity for *Redwood* data.

Geo·C

Note: Intensity estimation for point patterns on networks are different than estimates (2) and (3), however it has the same interpretation adapted to the network. We do not discuss that here.
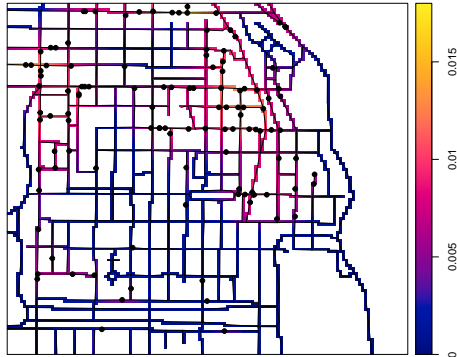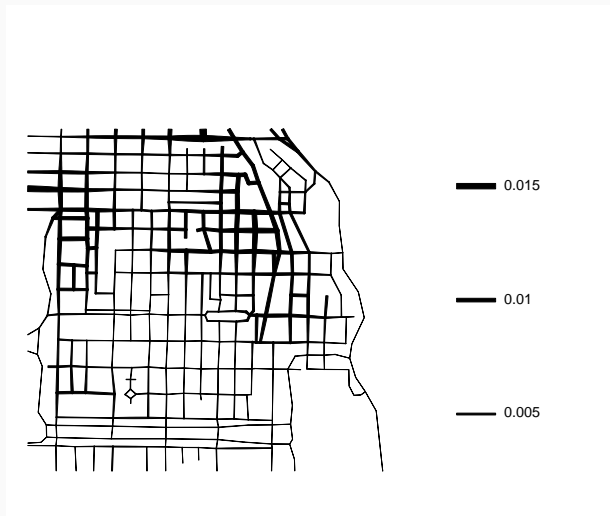

Figure: Estimated intensity for *chicago* crime data.

Figure: Estimated intensity for *chicago* crime data.

Uniform correction,

$$\widehat{\lambda}^U(u) = \sum_{i=1}^{n} w_i \frac{\kappa_\sigma(u - x_i)}{C_{W,\sigma}(u)}, \qquad u \in W, \tag{7}$$

and with Jones-Diggle correction,

$$\widehat{\lambda}^{JD}(u) = \sum_{i=1}^{n} w_i \frac{\kappa_\sigma(u - x_i)}{C_{W,\sigma}(x_i)}, \qquad u \in W. \tag{8}$$

Applying (7) to 'longleaf' data.

```
R> d4 <- density(longleaf,sigma = bw.ppl(longleaf),weights =marks(longleaf))
R> plot(d4,main = "",ribwid=0.04,ribsep=0.02)
R> plot(longleaf,add=T)
```

The diameter-weighted intensity is the average total diameter of trees per unit area of forest.
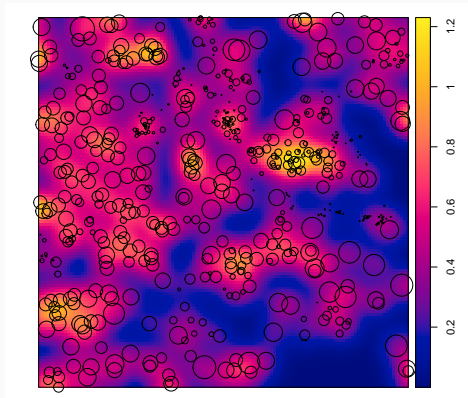

Figure: Weighted estimated intensity for *longleaf* data. Weights are diameter of trees.

- There are cases where each data point represents one or more events that occurred at the same spatial location, such as multiple disease cases at the same residence. It is then appropriate to weight each residential location by the number of cases, so that the intensity would be the average total number of cases per unit area, not the number of affected residences per unit area.

- If weight is the number of vehicles involved in the traffic accident that occurred at locations, then the weighted intensity estimate is the spatially-varying number of vehicles involved in accidents per unit length.

In spatial relative risk estimation, data points are classified into different types, and we seek to estimate the spatially-varying relative frequency of each type of point.

For simplicity, assume there are only two types of points, and we observe two point patterns x; y containing the points of the first and second types respectively, on the same window *W*. The goal is to estimate the logarithmic relative risk

$$\rho(u) = log\frac{\lambda_X(u)}{\lambda_Y(u)}, \qquad u \in W, \tag{9}$$

where λ_X and λ_Y are the intensity functions of the underlying point processes *X*; *Y* of points of each type.

Note: It is recommended to use a common bandwidth to estimate numerator and denominator in (9). See R package *'sparr'*.

```
R> library(spatstat)
R> library(sparr)
R> antstypes <- split(ants)
R> cmbw <- LSCV.risk(antstypes$Cataglyphis,antstypes$Messor,method =
"hazelton")
R> dCataglyphis <- density.ppp(antstypes$Cataglyphis,sigma = cmbw)
R> dMessor <- density.ppp(antstypes$Messor,sigma = cmbw)
R> plot(dMessor/dCataglyphis,main = "",ribwid=0.04,ribsep=0.02)
R> npoints(antstypes$Messor)/npoints(antstypes$Cataglyphis)
R> [1] 2.344828
```
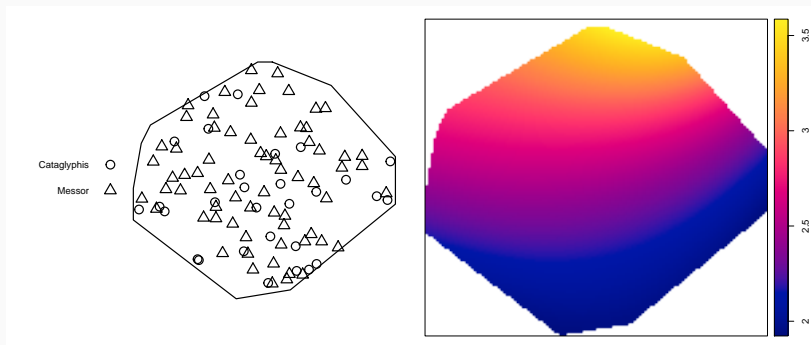
Figure: Left: Ants point pattern. Right: Estimated relative risk for the two different types of ants using the ratio of two estimated intensities.

A very popular technique for analysing spatial correlation in point patterns.

The inhomogeneous *K*-function for point processes in $\mathbb{R}^2$ is defined as

$$K_{inhom}(r) = \mathbb{E}\left[\sum_{x_i \in X} \frac{1\{0 < \|u - x_i\| \leq r\}}{\lambda(x_i)}\middle| u \in X\right]. \tag{10}$$

For Poisson processes in $\mathbb{R}^2$, $K_{inhom}(r) = \pi r^2$ and for Poisson processes on networks $K_{inhom}(r) = r$.

Note: $K > \pi r^2 (K > r)$ is an indication of clustering behaviour while $K < \pi r^2 (K < r)$ denote inhibition between points.

Note: *K*-function for multi-type patters, replicated patterns, patterns on networks, etc is different. However, it somehow keeps the same interpretation after adaption to the new space/condition. We do not discuss that here!

Note: There are different estimators for *K*-function. We do not discuss them here.

```
R> d3 <- density.lpp(chicago,sigma = 100)
R> kchig <- linearKinhom(chicago,lambda = d3,normpower = 2)
R> plot(kchig,main="")

R> kcells <- Kinhom(cells,normpower=2)
R> plot(kcells,main = "")

R> kants <- envelope(ants,fun = Kinhom,nsim=200,normpower=2)
R> plot(kants,main = "")

R> klong <- envelope(longleaf,Kinhom,normpower=2,nsim=200)
R> plot(klong,main="")
```
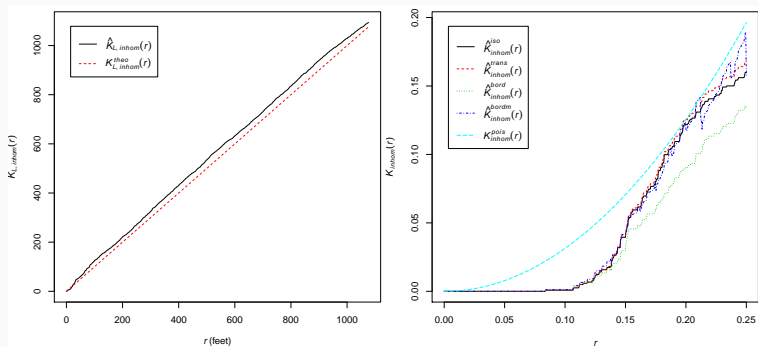
Figure: Estimated *K*-function for; Left: *chicago* data. Right: *cell* data.
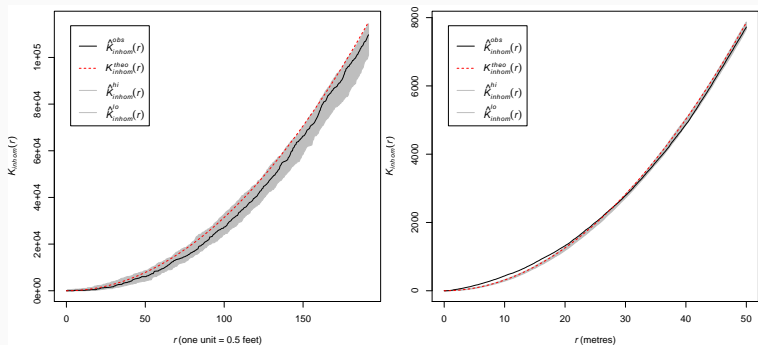
Figure: Estimated *K*-function together with simulation envelopes; Left: *ants* data. Right: *longleaf* data.

An alternative tool is the pair correlation function $g(r)$ which contains contributions only from inter-point distances equal to r.

$$g_{inhom}(r) = \frac{K'_{inhom}(r)}{\pi r^2} \tag{11}$$

where $K'(r)$ is the derivative of the K-function with respect to r.

Simply speaking: $g(r)$ is the probability of observing a pair of points of the process separated by a distance r, divided by the corresponding probability for a Poisson process.

Note: $g_{inhom}(r) > 1$ indicates clustering behaviour whereas $g_{inhom}(r) < 1$ shows inhibition.

```
R> gchig <- linearpcfinhom(chicago,lambda = d3,normpower = 2)
R> plot(gchig,main="")

R> gcells <- pcfinhom(cells,normpower=2)
R> plot(gcells,main="")

R> gants <- envelope(ants,pcfinhom,nsim=200,normpower=2)
R> plot(gants,main="")

R> glong <- envelope(longleaf,pcfinhom,nsim=200,normpower=2)
R> plot(glong,main="")
```
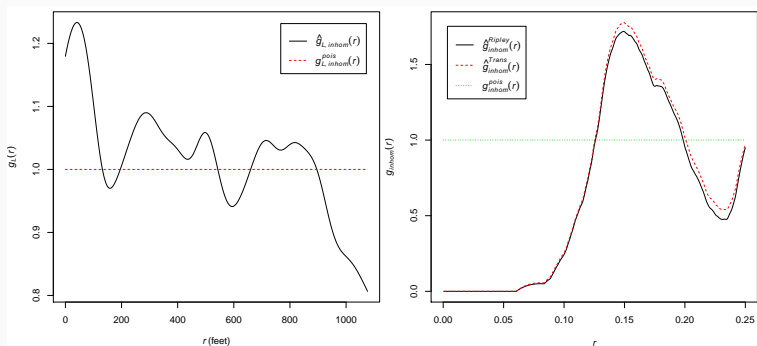
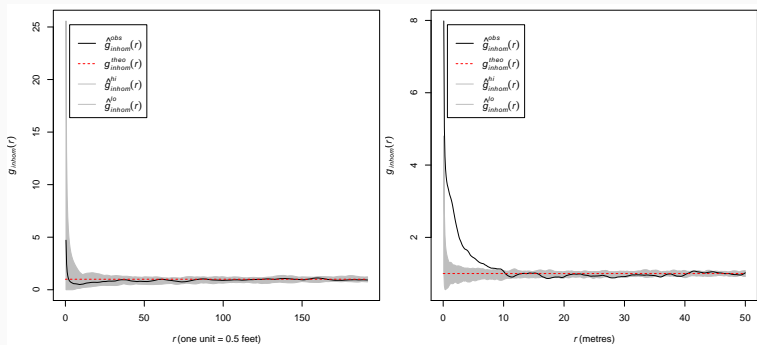Figure: Estimated *g*-function; Left: *chicago* data. Right: *cells* data.

Figure: Estimated *g*-function together with simulation envelopes; Left: *ants* data.
Right: *longleaf* data.

MORE TO KNOW

- Estimating intensity function parametrically.
- Adaptive estimators.
- Adaptive smoothing.
- Other bandwidth selectors.
- Space-time point processes.
- Marked space-time point processes.
- Higher-order summary statistics.
- Details of point processes on linear networks.
- Point processes on sphere.
- Replicated point patterns.
- And many other things ;)

Note: For more details, see Diggle (2013) and Baddeley, et al (2015).

REFERENCES

- Baddeley, A., Rubak, E., & Turner, R. (2015). Spatial point patterns: methodology and applications with R. CRC Press.
- Baddeley, A., & Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. Journal of statistical software, 12(6), 1-42.
- Davies, T. M., Hazelton, M. L., & Marshall, J. C. (2011). Sparr: analyzing spatial relative risk using fixed and adaptive kernel density estimation in R. Journal of Statistical Software, 39(i01).
- Diggle, P. J. (2013). Statistical analysis of spatial and spatio-temporal point patterns. CRC Press.

Geo·C