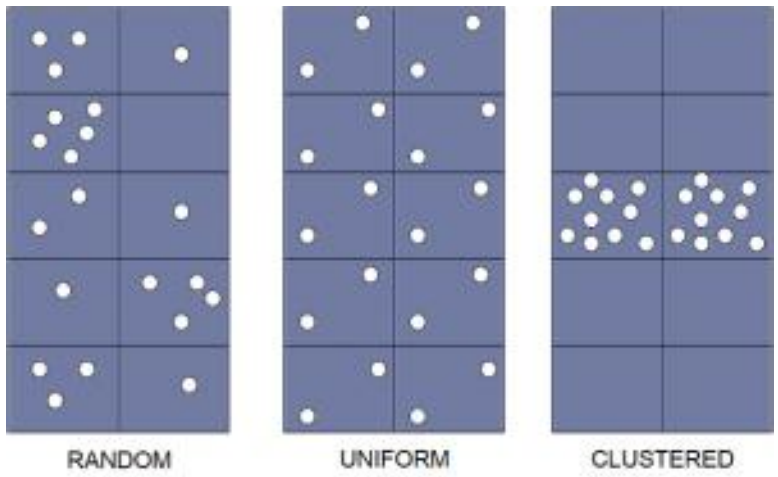
Cluster analysis for spatial data mining and visualisation

Spatial Data Mining: is a process to analyse large geographical databases and to extract implicit information from spatial data. Spatial data mining can be viewed as the search for interesting, useful and unexpected, but implicit spatial patterns.

How spatial data are distributed across a region?
Are they located randomly?
Are they clustered?
Can you see spatial dependency within data?



Acknowledgements:

Sections of this tutorial (including images) have been derived and or adapted from various publicly accessible resources and have been duly referenced.

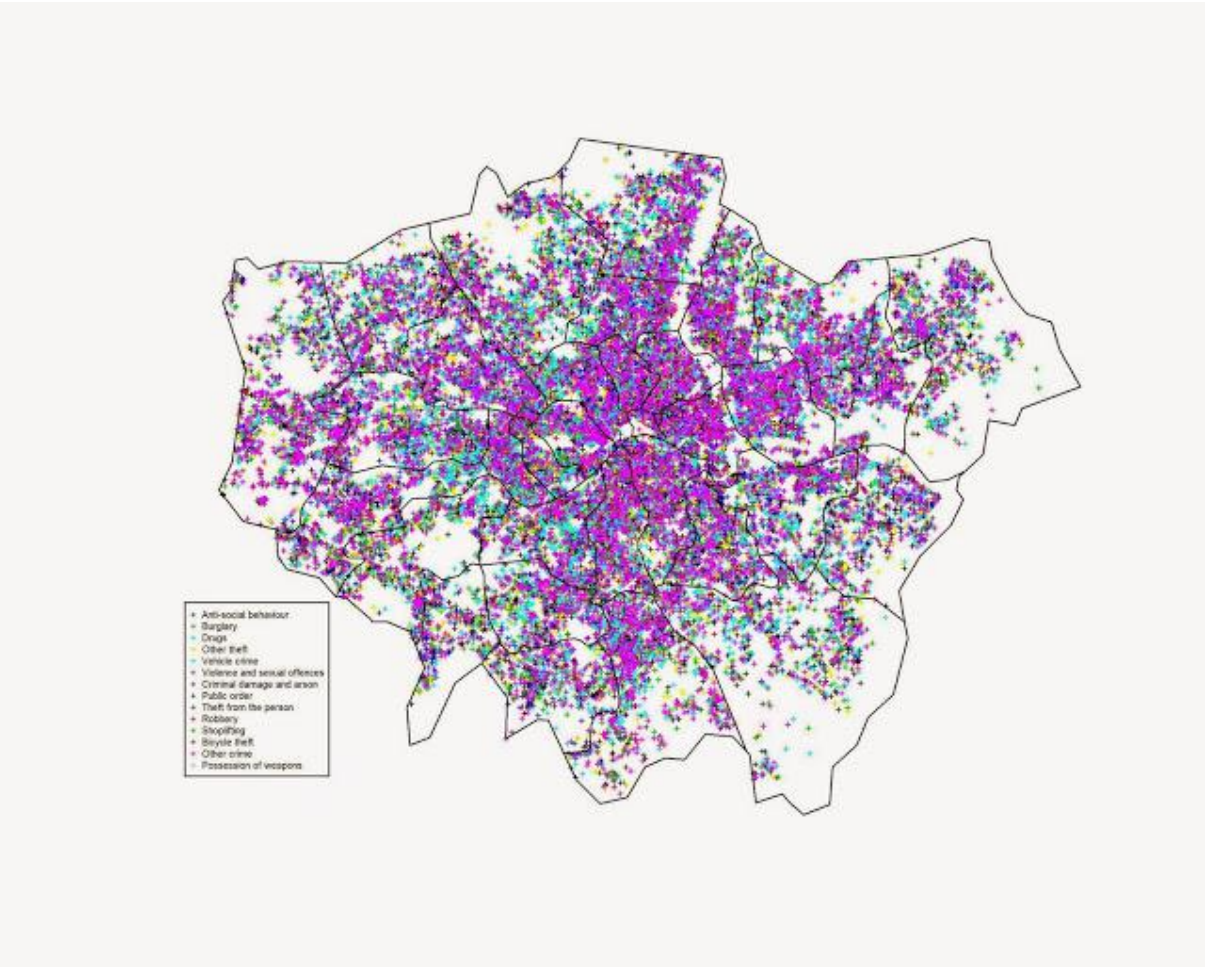
Please refer to the associated practical sheet for formal references and the reference list at the conclusion of this presentation for links to the original document as created by the corresponding author.

Experiments: Analysing point patterns

- PPA: Study of the spatial arrangements of points in (usually 2-dimensional) space.
- The easiest way to visualize a 2-D point pattern is a map of the locations.

Data Preparation (using **spatstat** R package)**:** Crime dataset

1. Remove NAs from data
2. Remove all duplicates
3. Select Specific region


Anti-social behaviour
Burglary
Drugs
Other theft
Vehicle crime
Violence and sexual offences
Criminal damage and arson
Public order
Theft from the person
Robbery
Shoplifting
Bicycle theft
Other crime
Possession of weapons

Descriptive Statistics

PPA: Analyse the occurrence of points in a particular space.

- How many points are there?
- How many crimes are committed in a neighbourhood of a city?
- How does that compare to a differing neighbourhood?

Simple descriptive statistics: Count, Mean, Median, and Standard Deviation

- How dense a pattern is, where the center of a set of points is
- How dispersed these points are

1) Frequency *and Density*

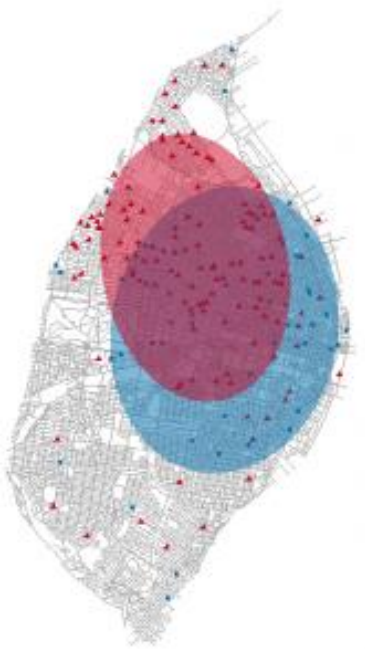
Mean centre: Mean in space

$$C = (\bar{x}, \bar{y}) = \left(\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n}\right)$$

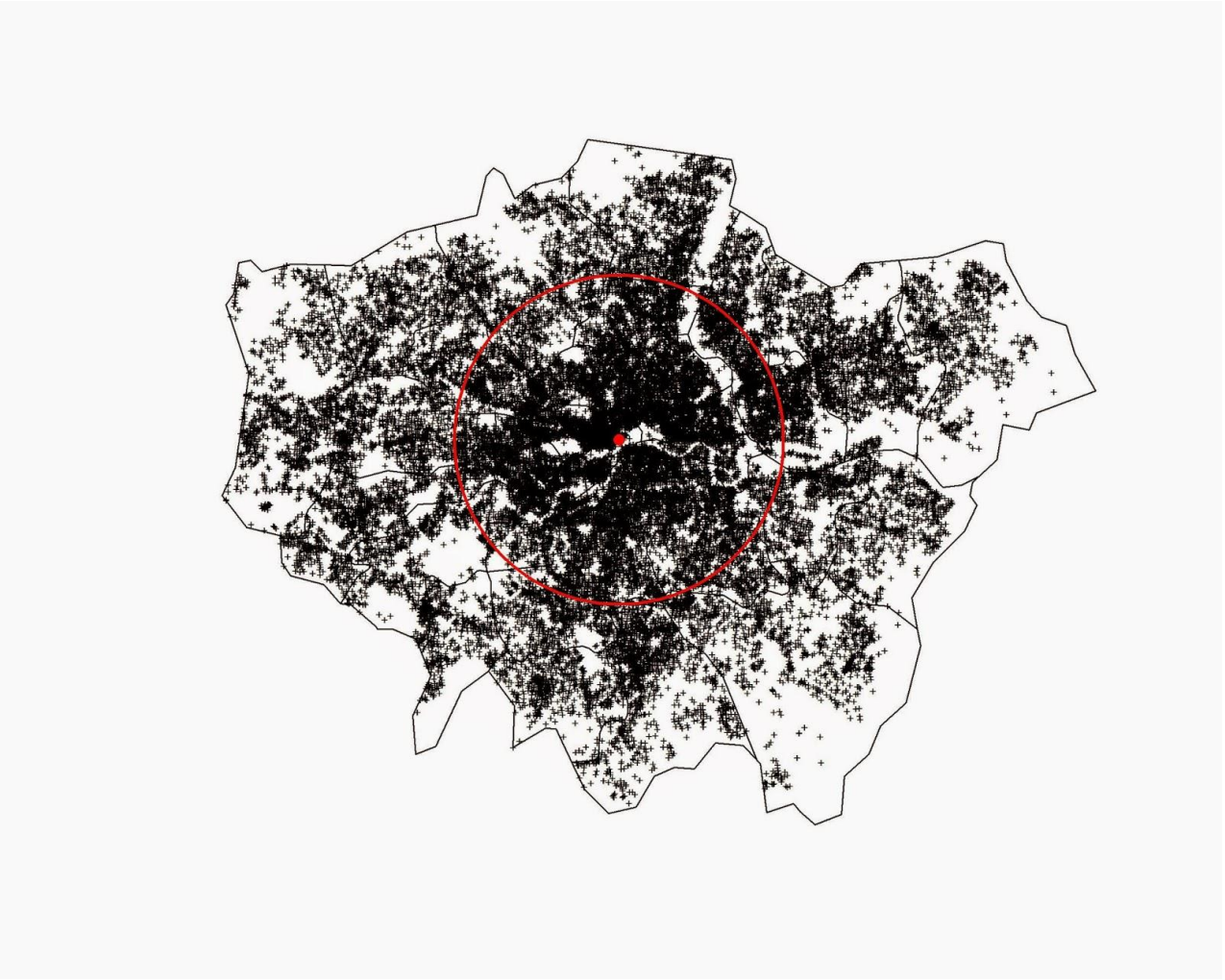
Standard distance: SD in space

Measure of spread in the 2D space

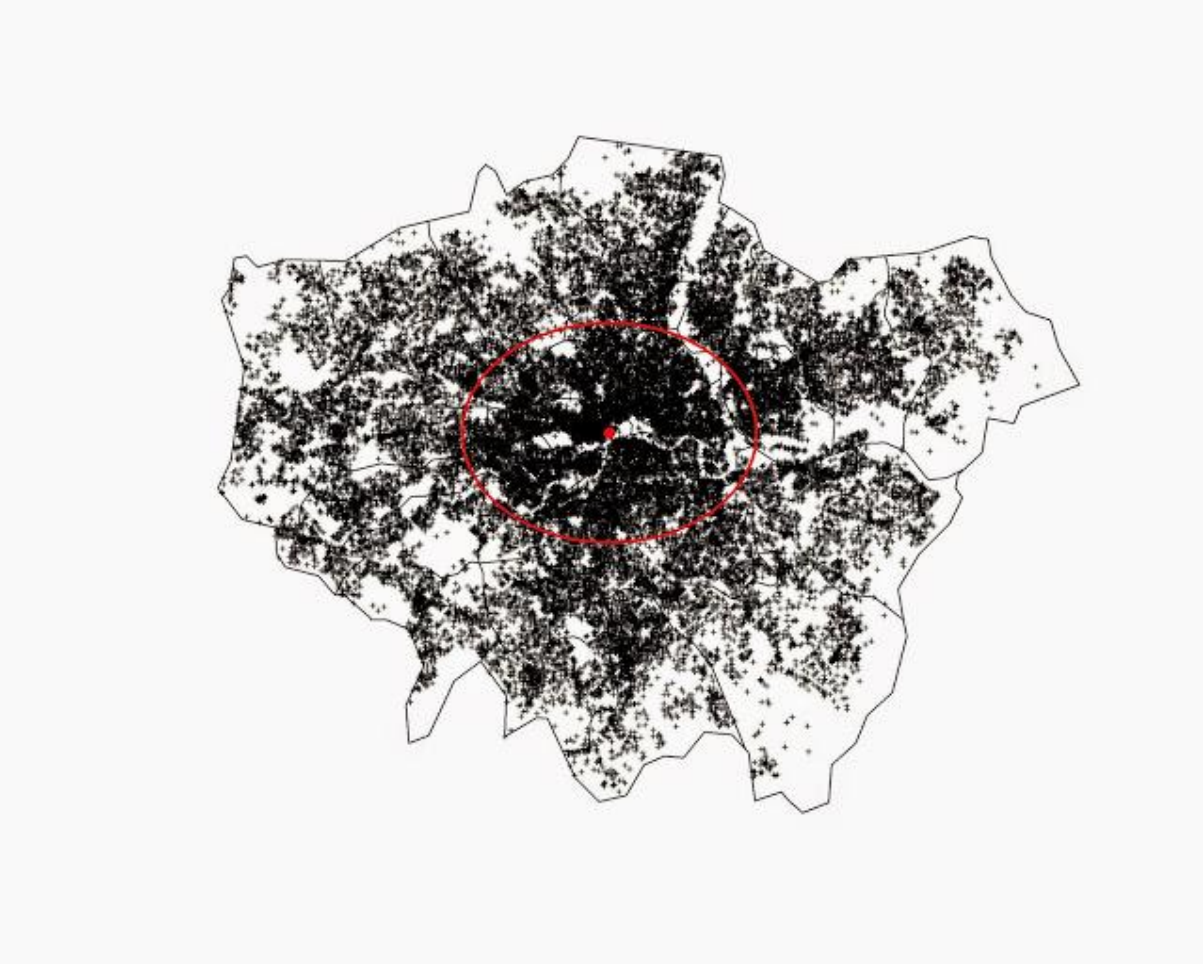
$$d = \sqrt{\frac{\sum_{i=1}^{n}[(x_i - \bar{x})^2 + (y_i - \bar{y})^2]}{n}}$$



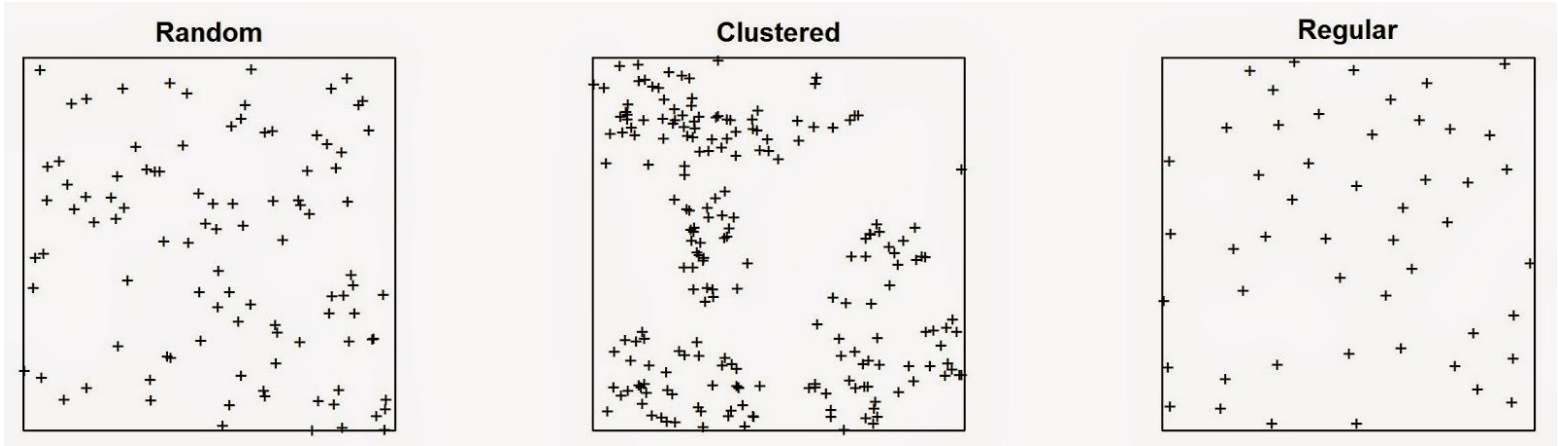
Visual feeling of the spread of data around mean centre



Standard Deviational Ellipse is a modified version of standard distance that captures the shape of this distribution by showing any directional bias in the pattern.



Checking spatial randomness

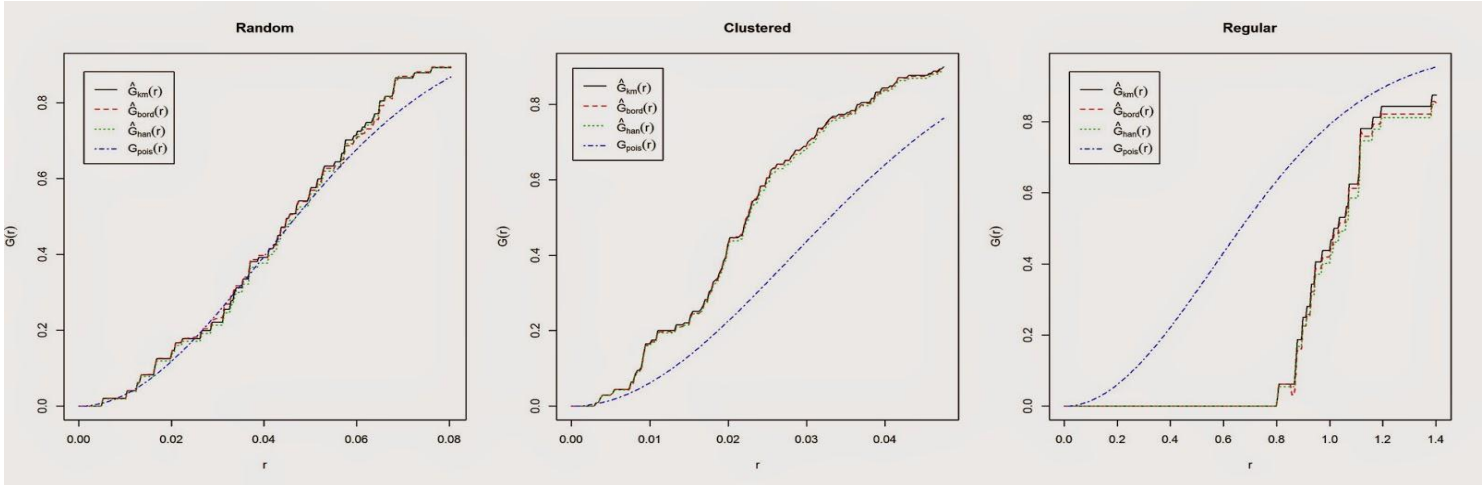


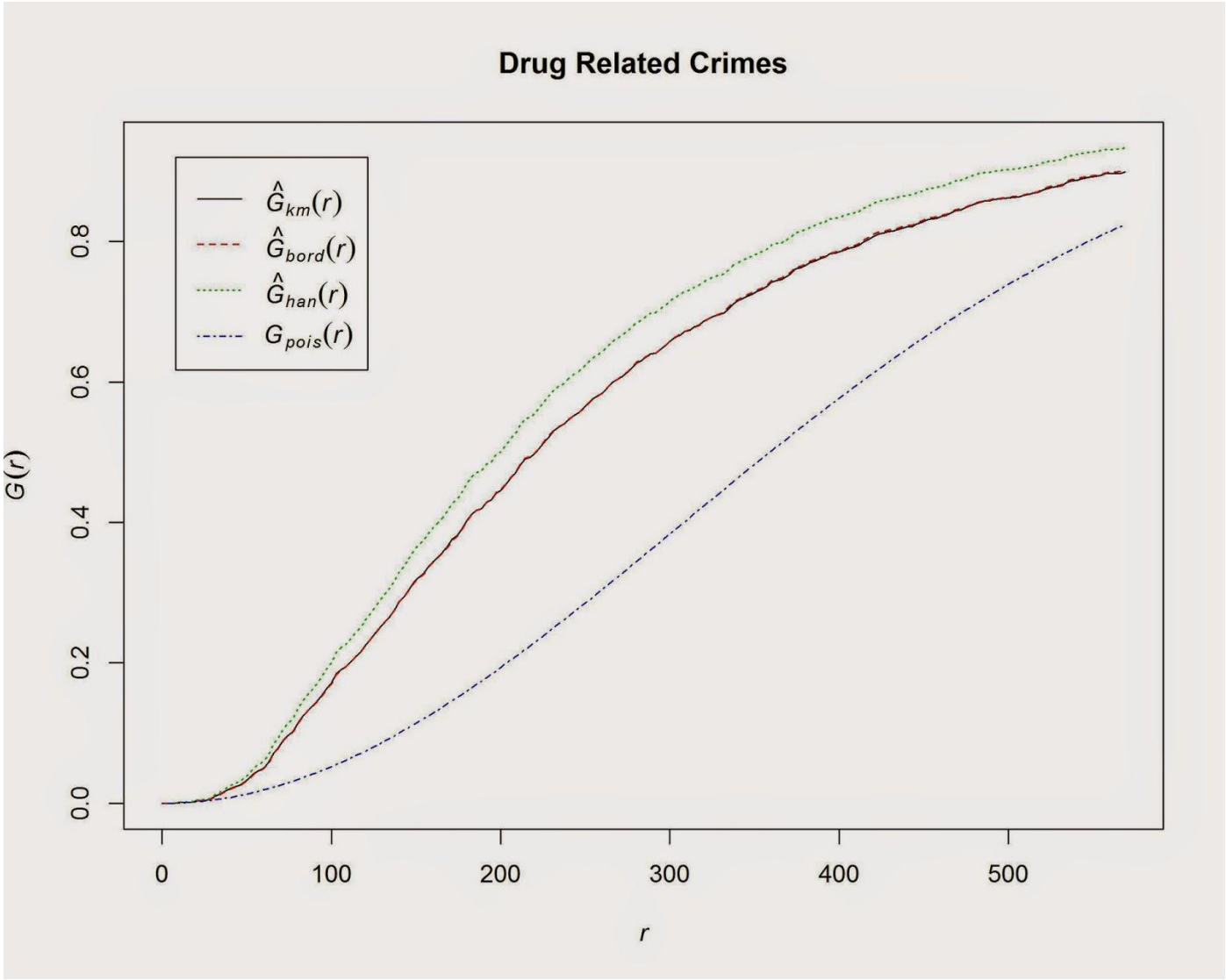
Each point is independent from each other

There are areas where the number of events is higher than average

Each subarea has the same number of events

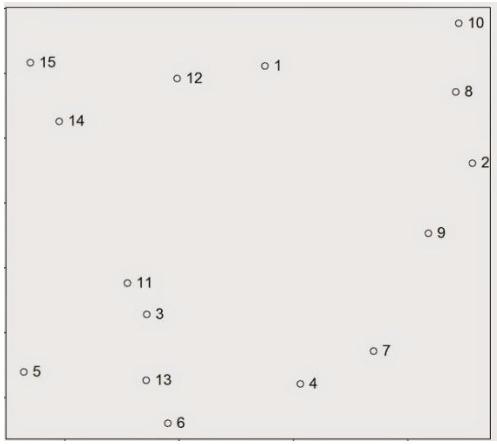
G function

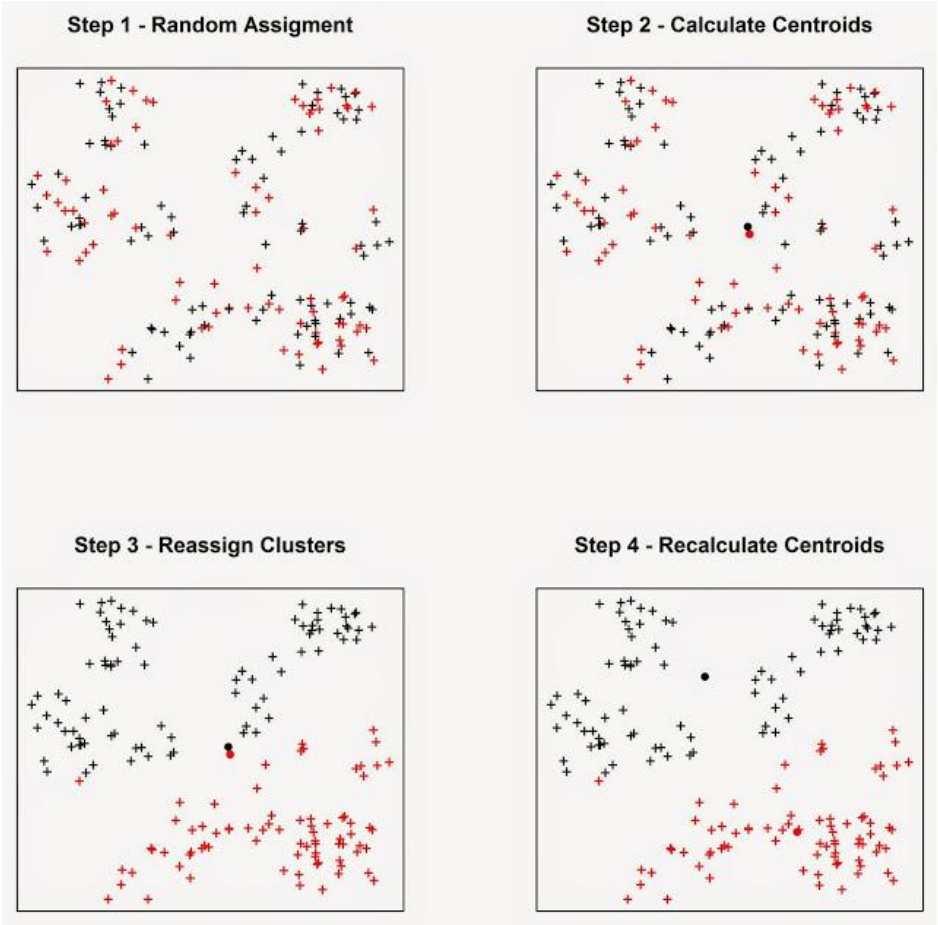



Drug Related Crimes
$\hat{G}_{km}(r)$
$\hat{G}_{bord}(r)$
$\hat{G}_{han}(r)$
$G_{pois}(r)$
G(r)
0.0
0.2
0.4
0.6
0.8
0
100
200
300
400
500
r

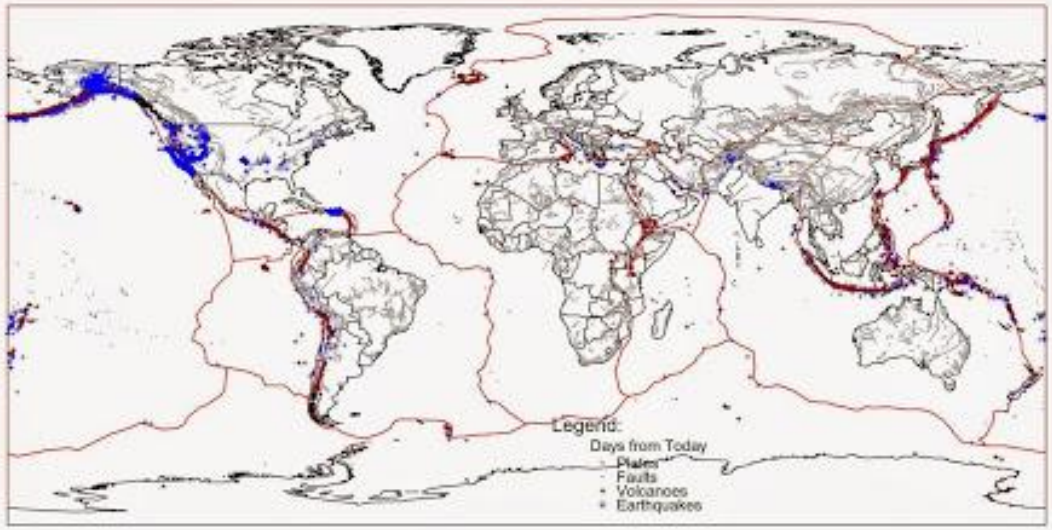
Cluster Analysis: Theoretical Background

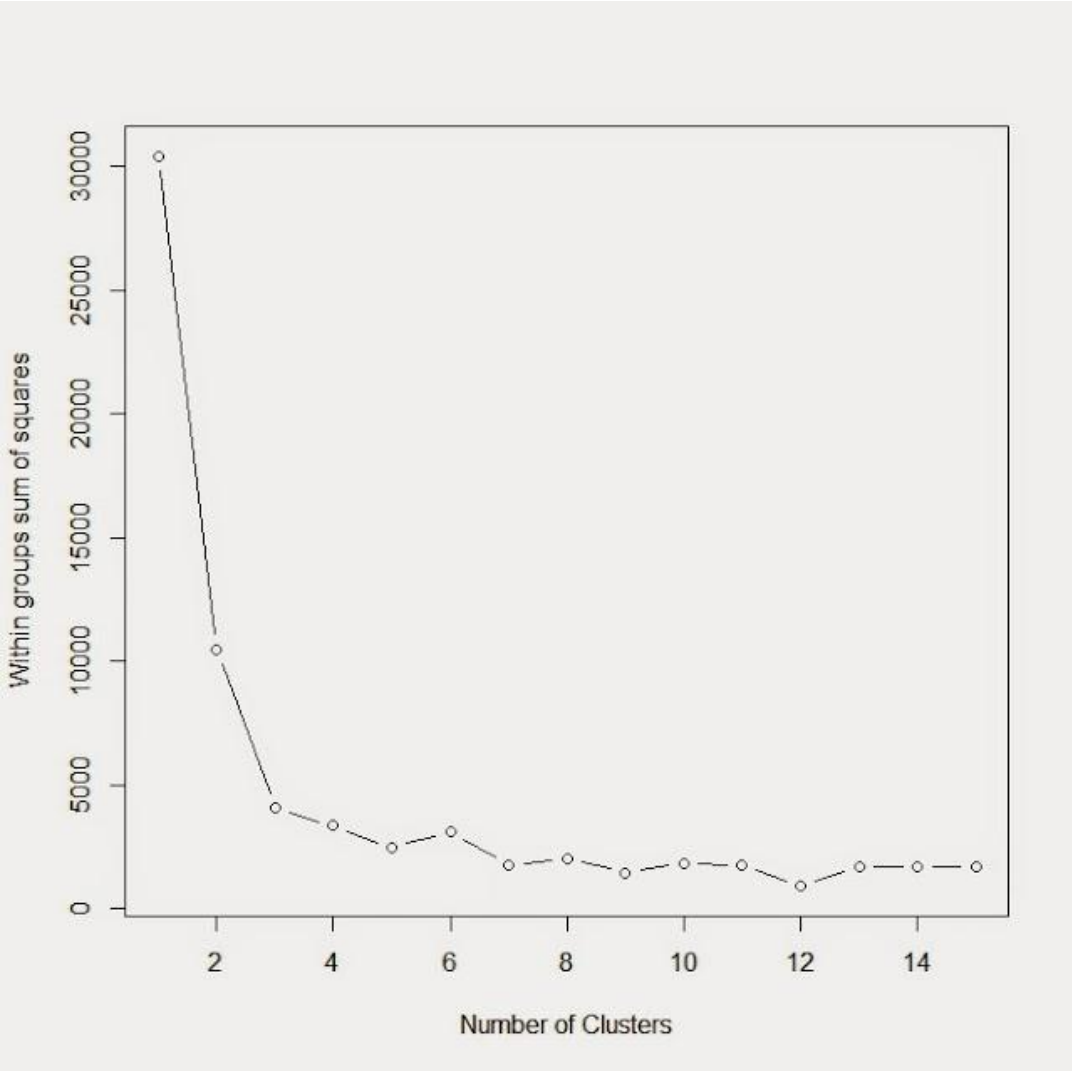
$$W_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2$$

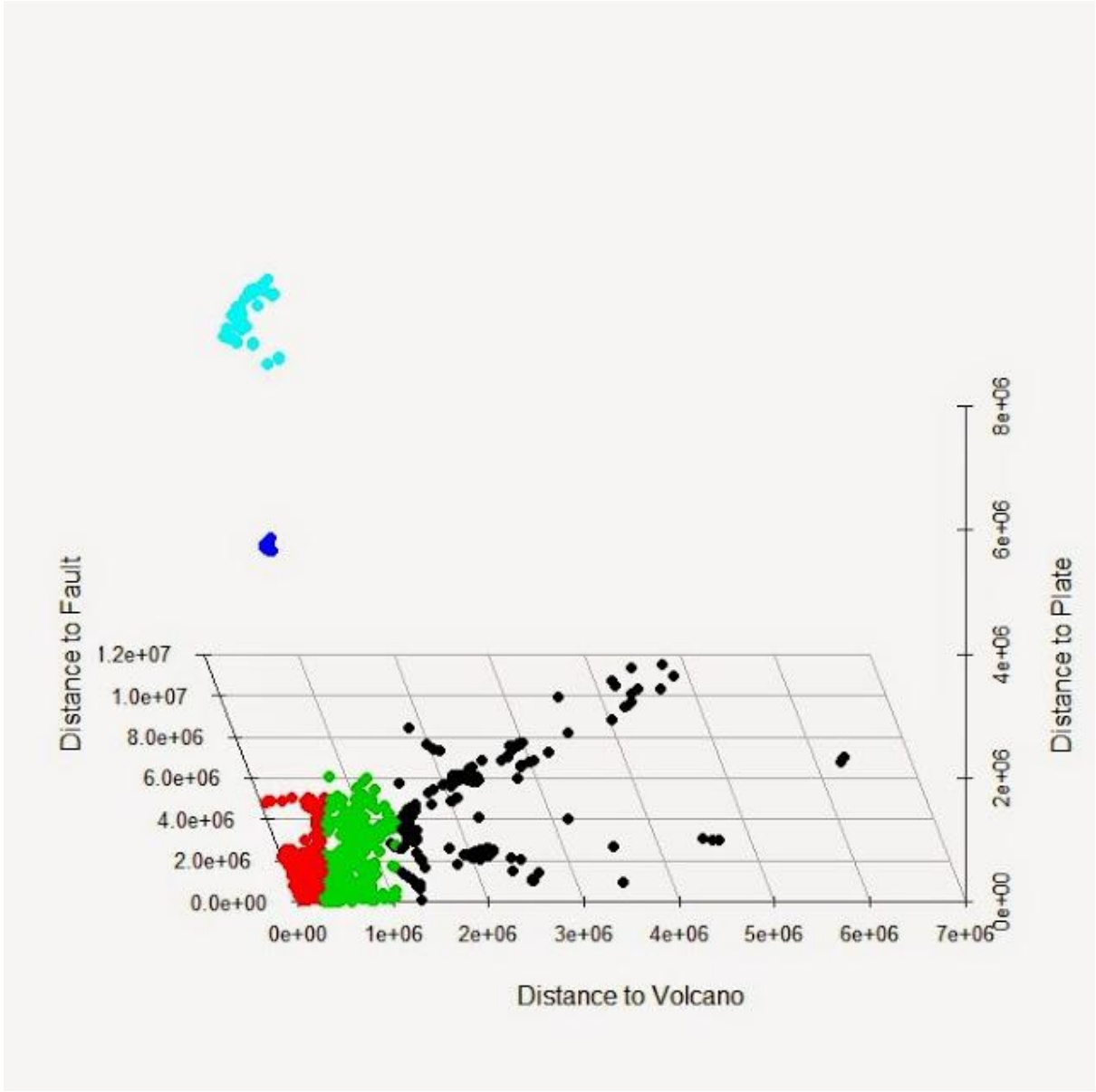




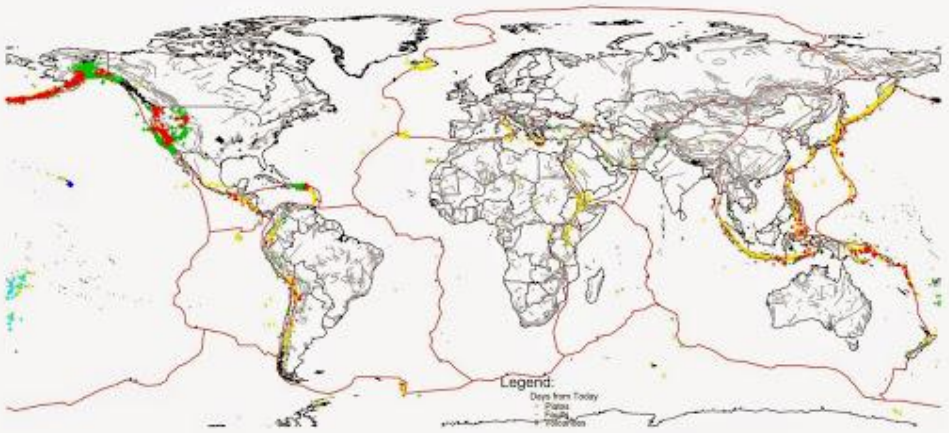
Earthquakes in the last 30 days

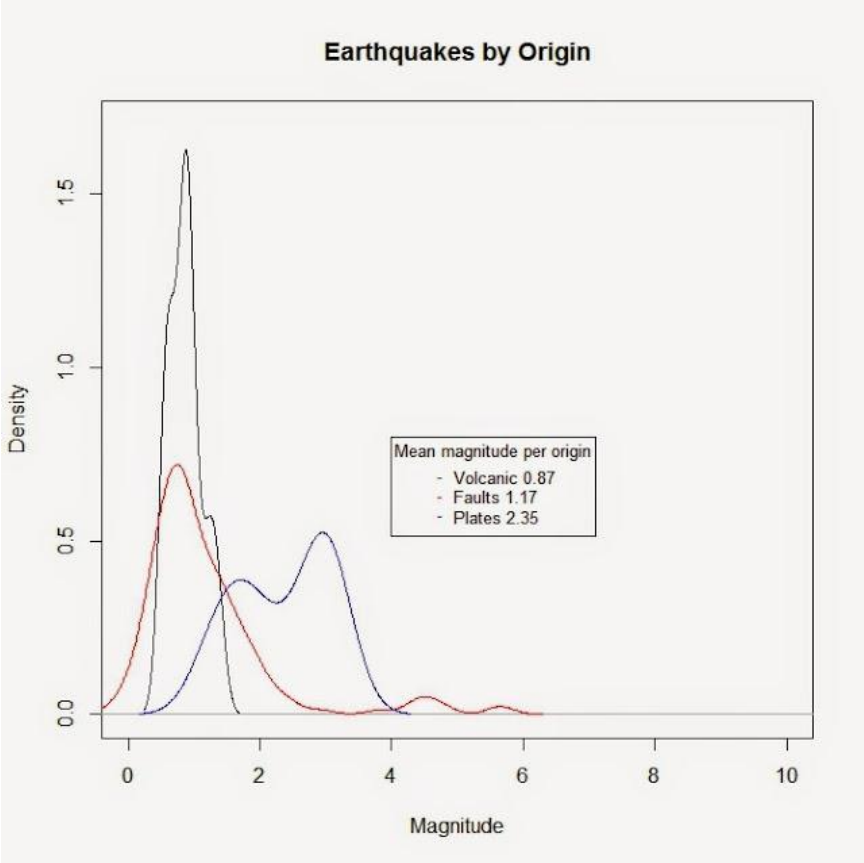



Within groups sum of squares
0
5000
10000
15000
20000
25000
30000
2
4
6
8
10
12
14
Number of Clusters


Distance to Fault
1.2e+07
1.0e+07
8.0e+06
6.0e+06
4.0e+06
2.0e+06
0.0e+00
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
7e+06
Distance to Volcano
Distance to Plate
0e+00
2e+06
4e+06
6e+06
8e+06

Earthquakes in the last 30 days




Earthquakes by Origin
Density
0.0
0.5
1.0
1.5
0
2
4
6
8
10
Magnitude
Mean magnitude per origin
- Volcanic 0.87
- Faults 1.17
- Plates 2.35

These slides have been compiled from various important resources

References:

Baddley, A. (2010) Analysing spatial point patterns in R. Workshop Notes Version 4.1, CSIRO. Online, viewed 31 August 2016, http://research.csiro.au/software/wp-content/uploads/sites/6/2015/02/Rspatialcourse_CMIS_PDF-Standard.pdf

Han, J., Kamber, M. and Tung, K. H. (2001) Spatial Clustering Methods in Data Mining: A Survey. In Harvey J. Miller and Jiawei Han (eds.), *Geographic Data Mining and Knowledge Discovery*, CRC Press. Boca Raton, FL, USA.

Ervin, D. (2016) Point pattern analysis. Advanced Spatial Analysis, University of California Santa Barbara. Online, viewed 31 August 2016, http://gispopsci.org/point-pattern-analysis/

Veronesi, F. (2015a) Introductory point pattern analysis of open crime data in London. R Tutorial for Spatial Statistics. Online, viewed 31 August 2016, http://r-video-tutorial.blogspot.com.au/2015/05/introductory-point-pattern-analysis-of.html

Veronesi, F. (2015b) Cluster analysis on earthquake data from USGS. R-bloggers. Online

Briggs, R. (2007) GISC 6382 (Spring)