Microsoft

Machine Learning, Analytics, & Data Science Conference

June 2–3, 2016 • MSCC

An Introduction to the Statistics of Spatial Data

Soren Johansen
GBB TSP AA WE

Microsoft

Session Goals

- Session Goals
  - To give an overview of the key ideas/concepts of Spatial Statistics
  - To give real life examples of spatial statistics data
  - To show how the theory works with R
  - The literature on Spatial Statistics

Agenda

Introduction

Spatial Kernel Density Estimation

Distance to Nearest Event

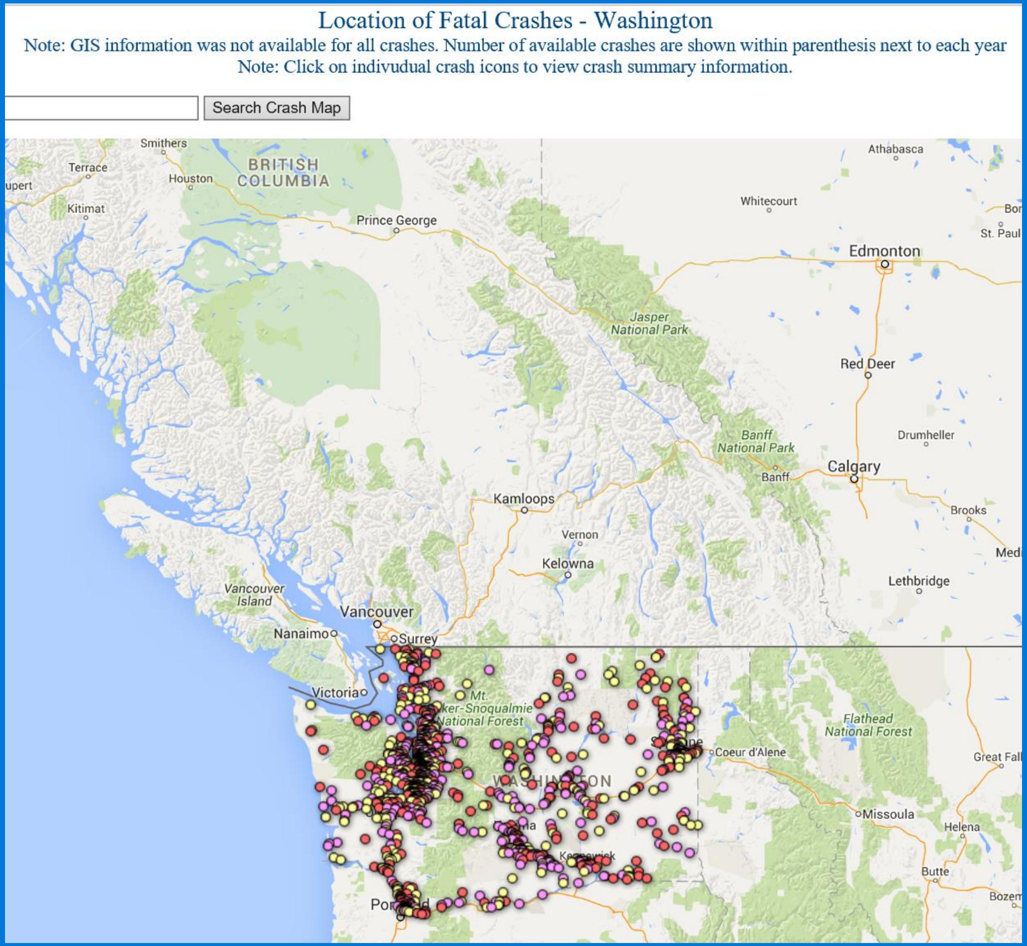
Interpolation of Point Pattern

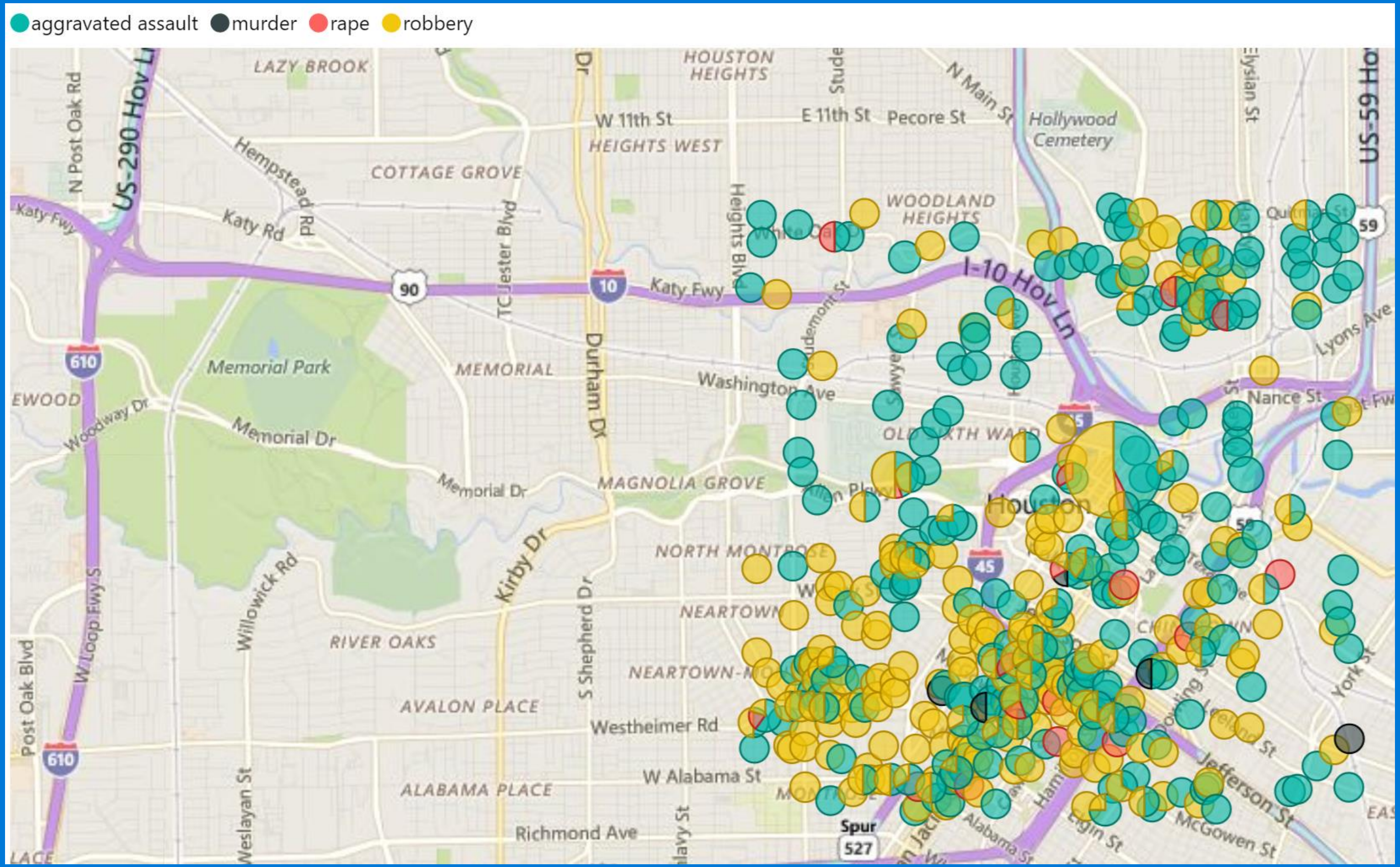
Spatial Prediction Models

Q&A

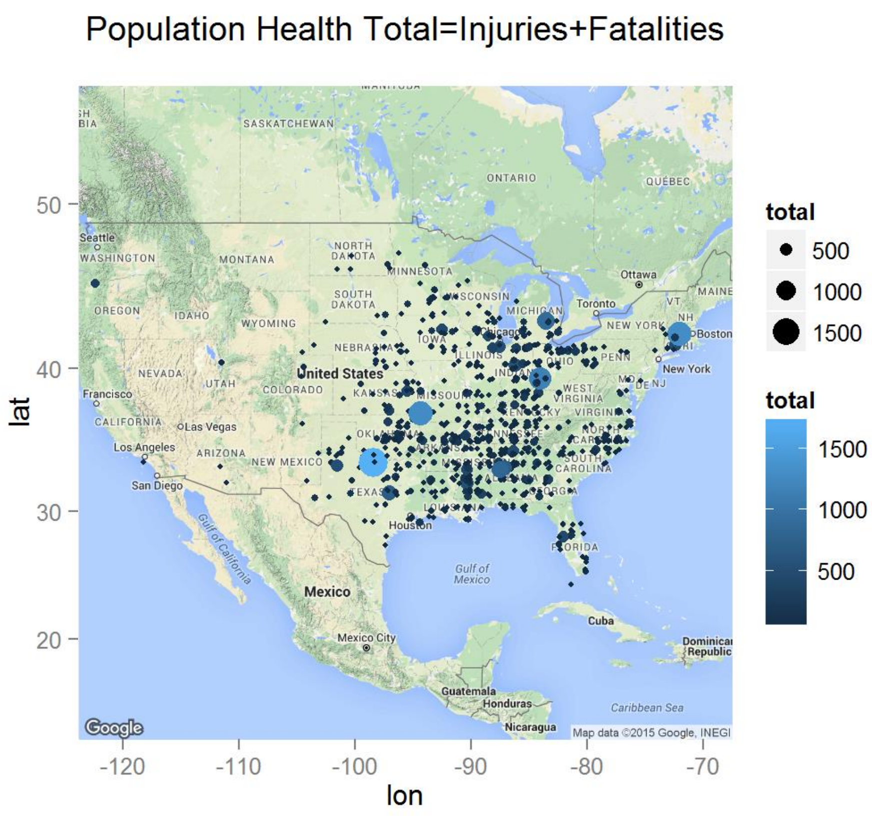
Introduction

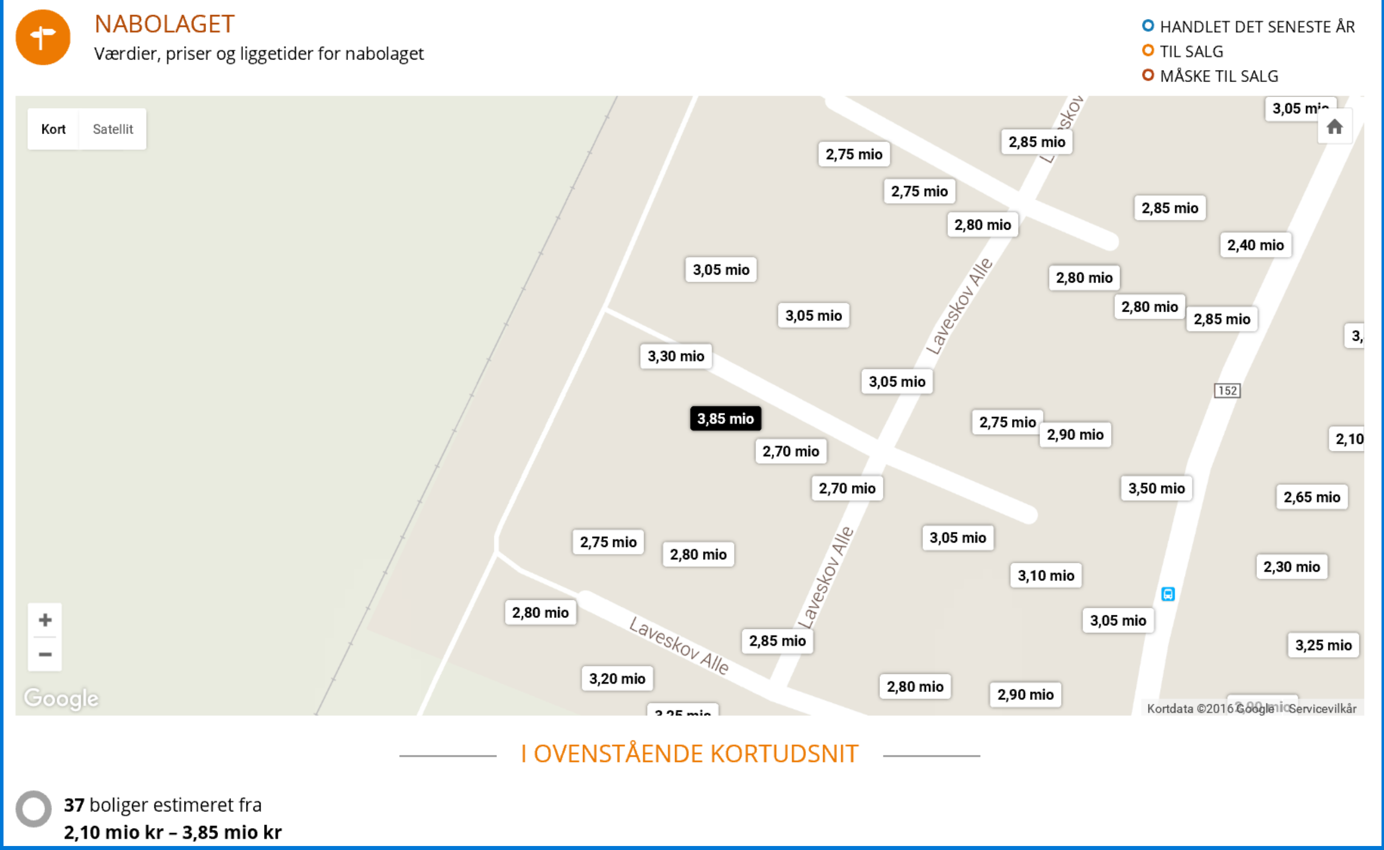
Examples of Spatial Statistics Data





Examples of Spatial Statistics Data





[Revolution blog]

Spatial Statistics

- Point processes (locations and spatial patterns of individuals/events)
- Maps of a continuous response variable(kriging)
- Spatially explicitly responses affected by the identity, size and proximity of neighbours

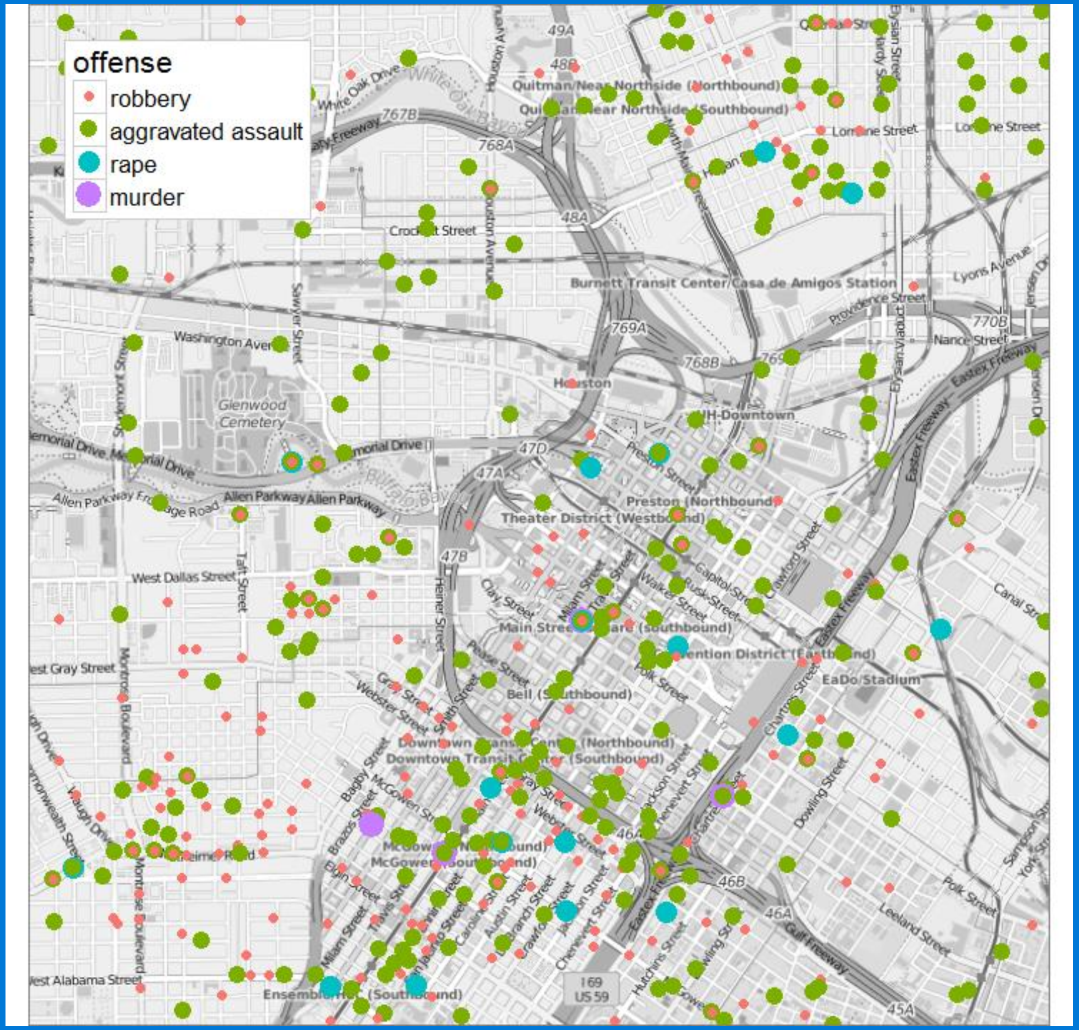
Spatial Kernel Density Estimation

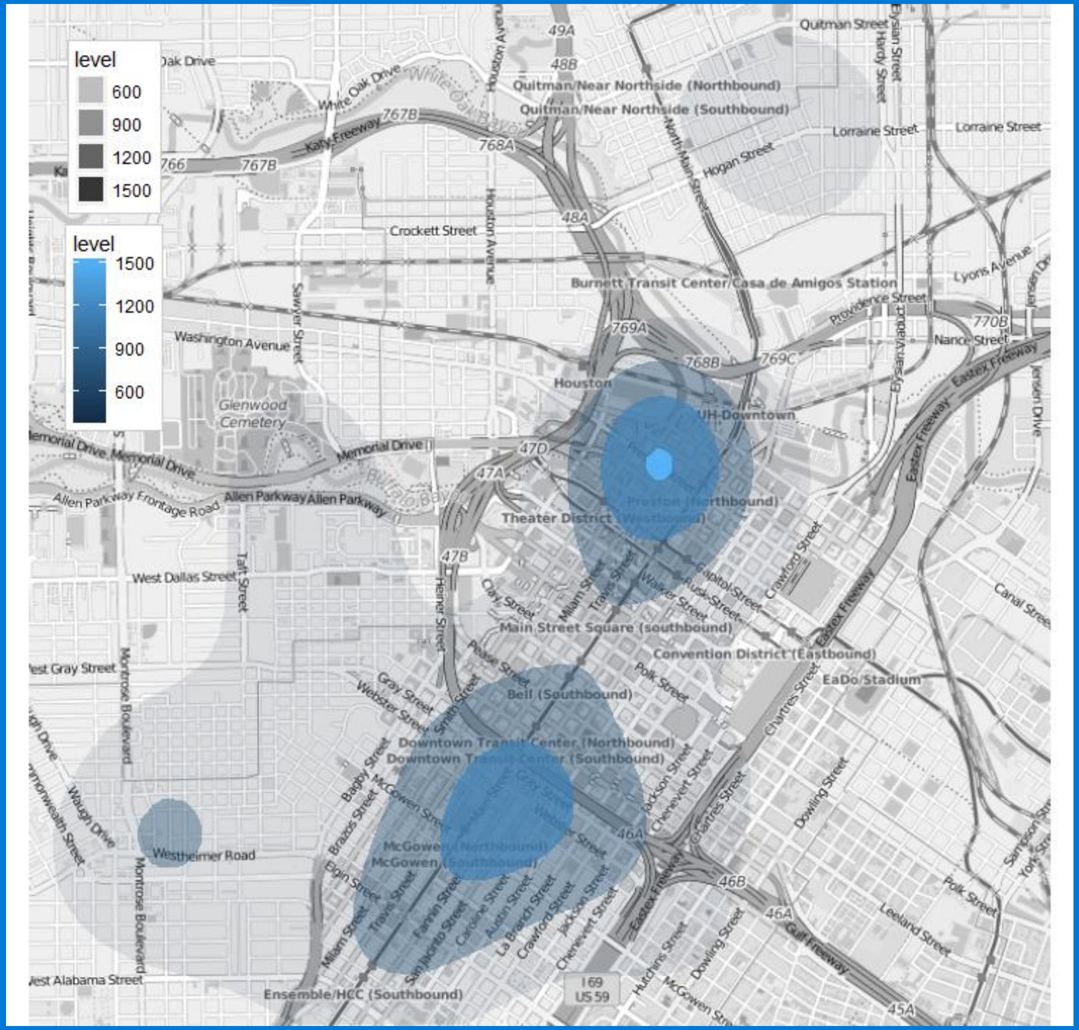
Spatial Kernel Density Estimation

- A kernel is Function maps a location onto a probability density
- $f(x, y) = \frac{1}{nh_x h_y} \sum_i k\left(\frac{x-x_i}{h_x}, \frac{y-y_i}{h_y}\right)$
- h_x & h_y are called bandwidth functions in the x and y directions
- Kernels are used to estimate intensities and densities
- The R packages spatstat and GIStools can do the job
- Mainly used for exploratory spatial analysis

Spatial Kernel Density Estimation

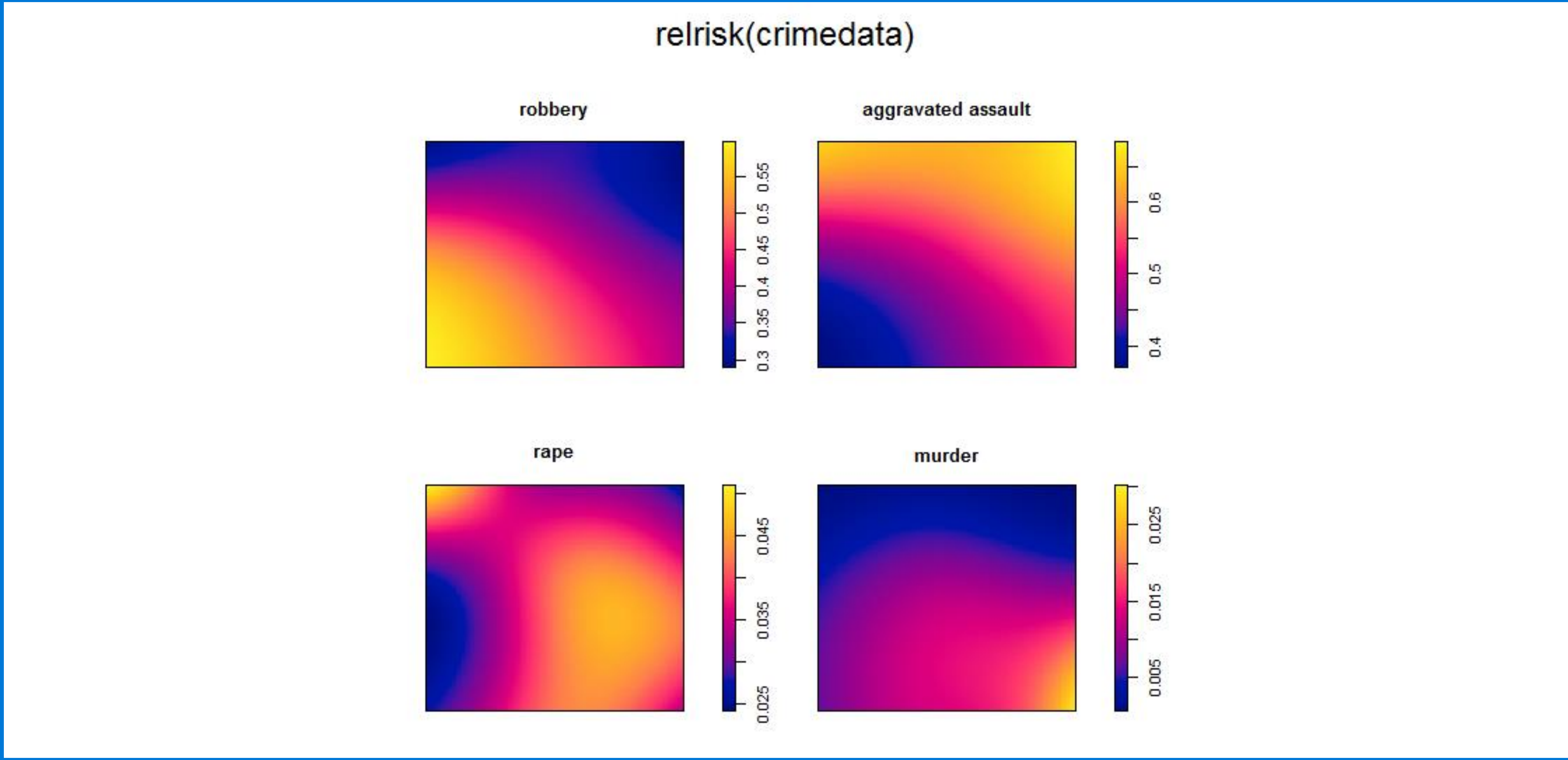
- The package ggmap is used for visualization





Spatial Kernel Density Estimation

- Spatstat relrisk function estimates the densities for each mark

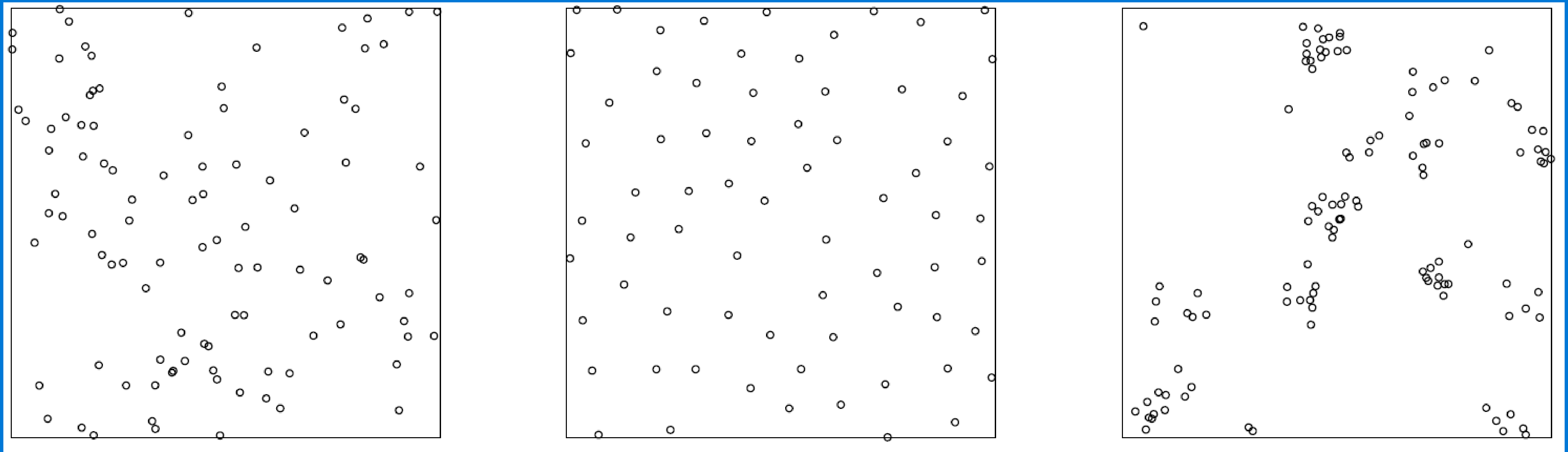


Distance to Nearest Event

Distance to Nearest Event

- In many cases the motivation for analyzing point pattern data is to determine whether the points appear to have been placed independently of each other or whether they exhibit some kind of interpoint dependence.

- Which of the following pictures are independent?



Distance to Nearest Event

Complete Spatial Randomness (CSR):

The events are distributed independent at random and uniform over the study area. This implies that there are no regions where the events are more (or less) likely to occur and that the presence of a given event does not modify the probability of other events appearing nearby.

When the events are independent and the marginal densities are uniform = poisson

Distance to Nearest Event

- *The K*-function

$$K(d) = \lambda^{-1} E(N_d)$$

- Where N_d is the number of events within a distance *d* of a randomly chosen event from all recorded events. λ is the intensity of the process.

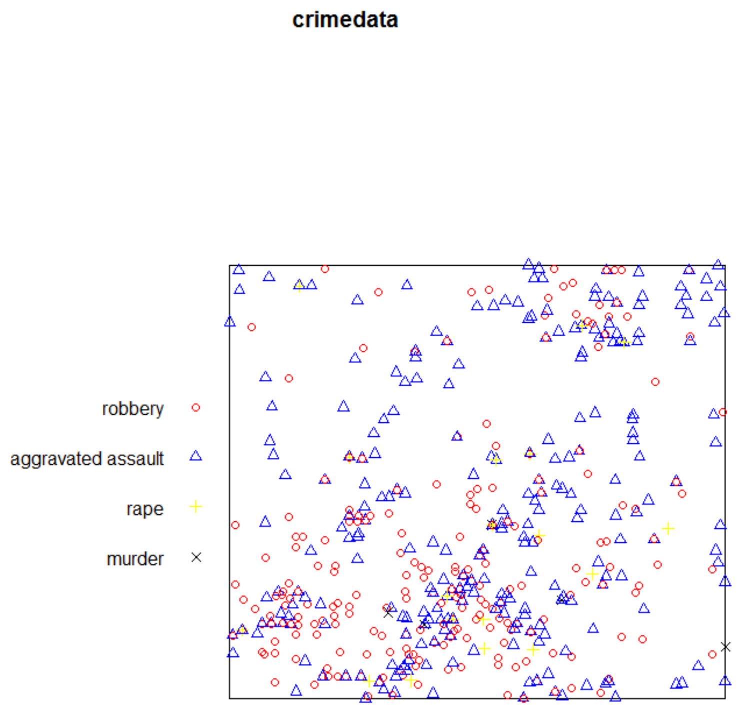
- Under CSR the value of the *K*-function is: $K(d) = \pi d^2$

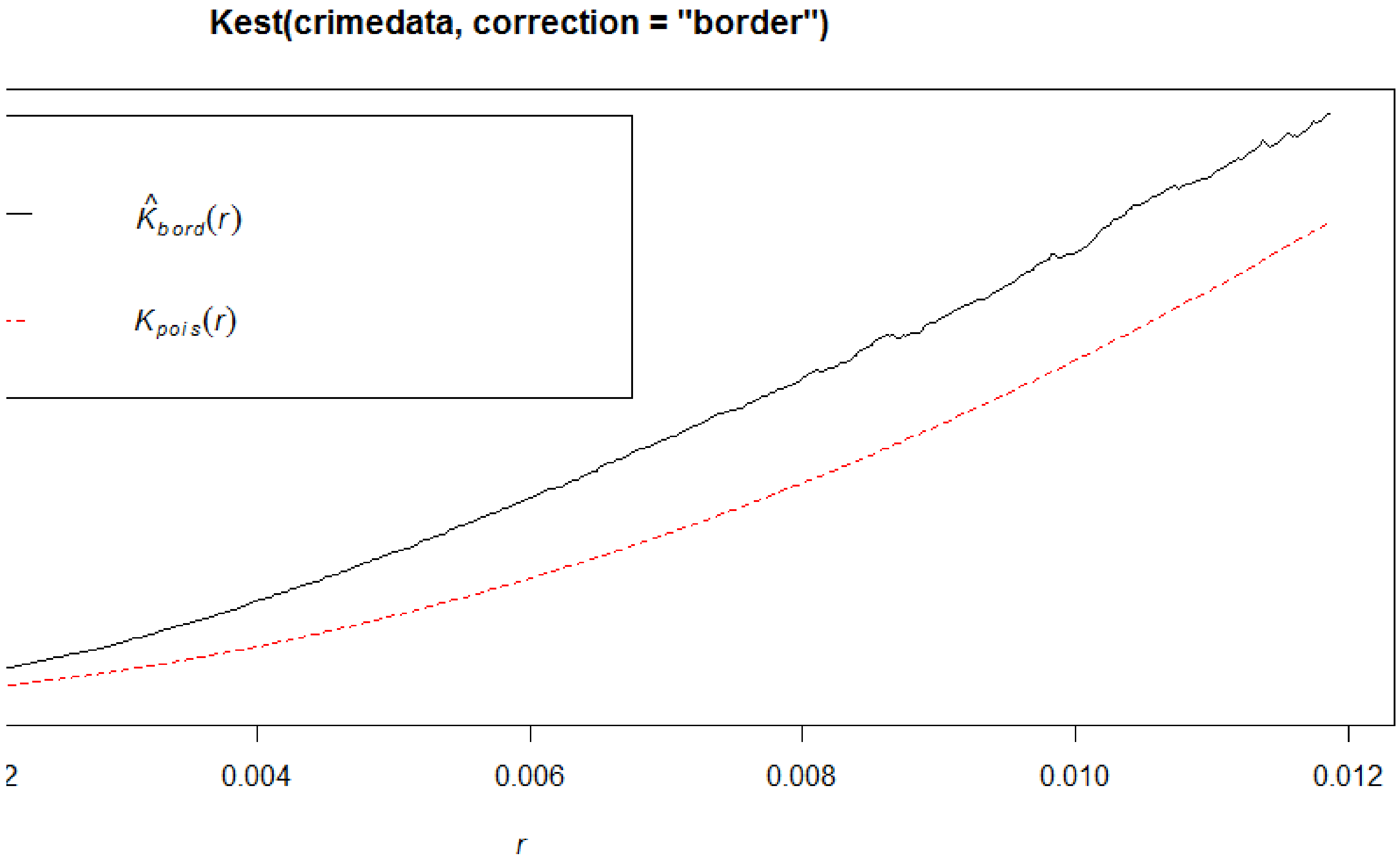
- Ripley's *K*-function (with correction)

$$K(d) = \left(n(n-1)\right)^{-1} \lambda^{-1} \sum_{i=1} \sum_{j \neq i} \frac{2I(d_{ij}<d)}{w_{ij}}$$

Distance to Nearest Event

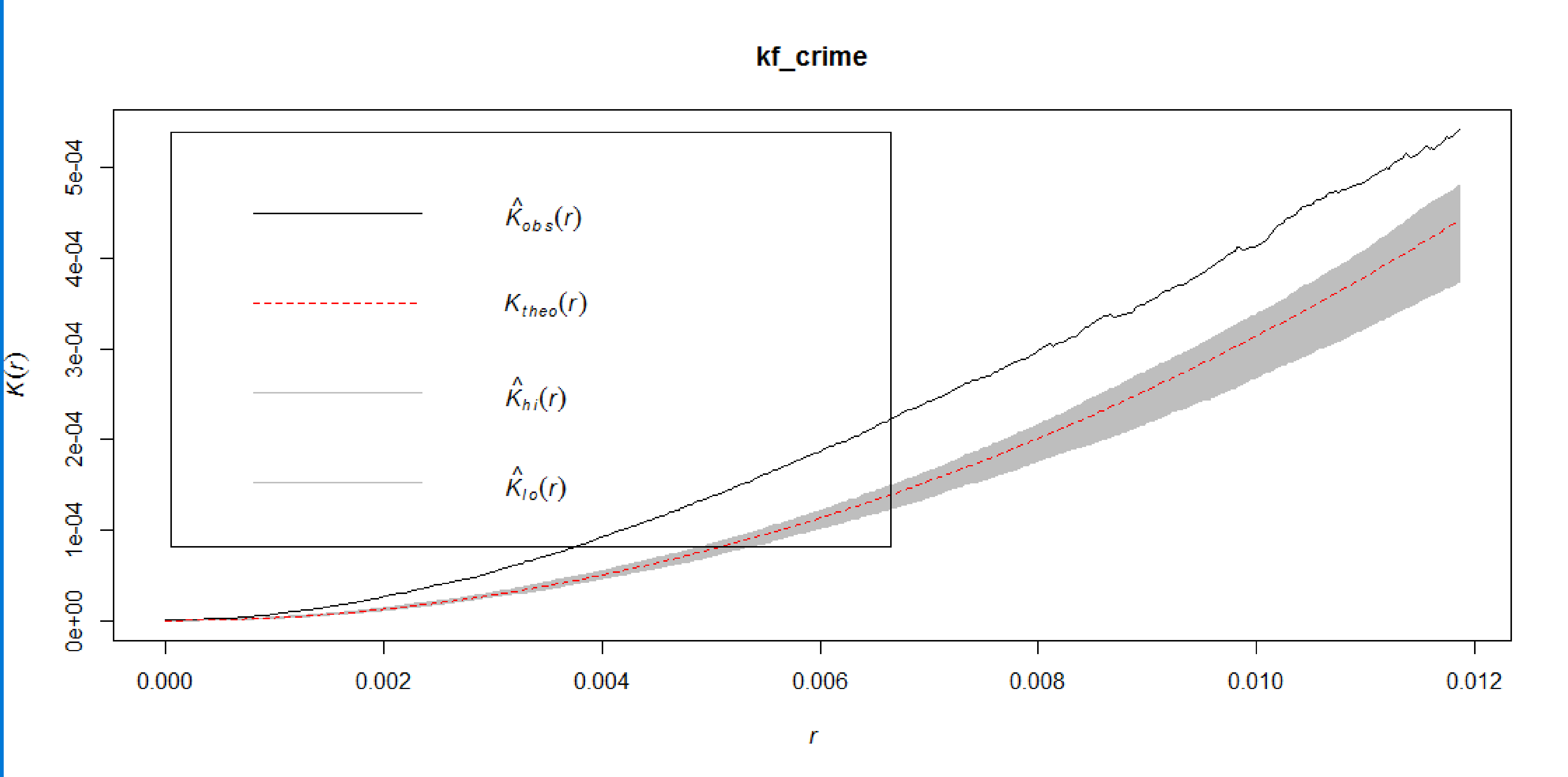
- Case: Crime in Houston.





Distance to Nearest Event

- Simulation analysis(envelope) – Confidence intervals



Distance to Nearest Event

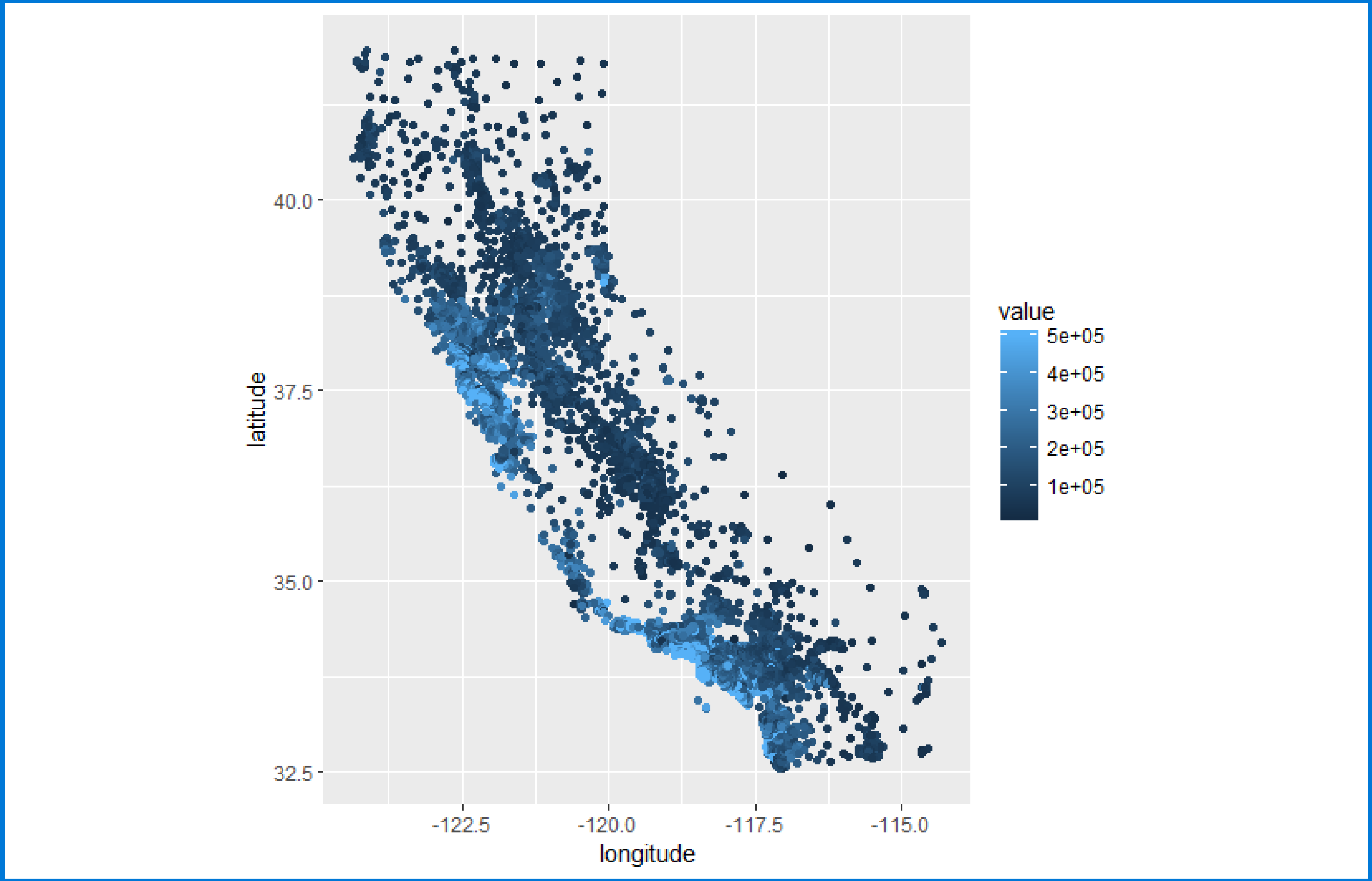
- Inference – Hypothesis test (mad and dclf test)

```
 mad.test(crimedata,Kest)
+
Generating 99 simulations of CSR  ...
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30,
31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45,
46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60,
61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75,
76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90,
91, 92, 93, 94, 95, 96, 97, 98,  99.

Done.

Maximum absolute deviation test of CSR
Monte Carlo test based on 99 simulations
Summary function: K(r)
Reference function: theoretical
Alternative: two.sided
Interval of distance values: [0, 0.0118630749999999]
Test statistic: Maximum absolute deviation
Deviation = observed minus theoretical

data:  crimedata
mad = 0.00015184, rank = 1, p-value = 0.01
```

Interpolation of Point Pattern

Interpolation of Point Patterns

- The problem: Interpolation of measurements $(z_1, , , , , z_n)$ at locations $(x_1, , , , , , , x_n)$ – the goal is to estimate the value of *z* at some new point *x*.



Interpolation of Point Patterns

- Three different techniques

1. Nearest neighbour interpolation
    - Find i such that $|x_i - x|$ is minimized
    - The estimate of z is z_i
2. Inverse distance weighting
3. Kriging

Interpolation of Point Patterns

Inverse distance weighting (IDW)

- Estimate the value of *z* at location *x* by weighted mean of nearby observations.
- Observations of *z* at points closer to *x* should be give more importance in the interpolation.

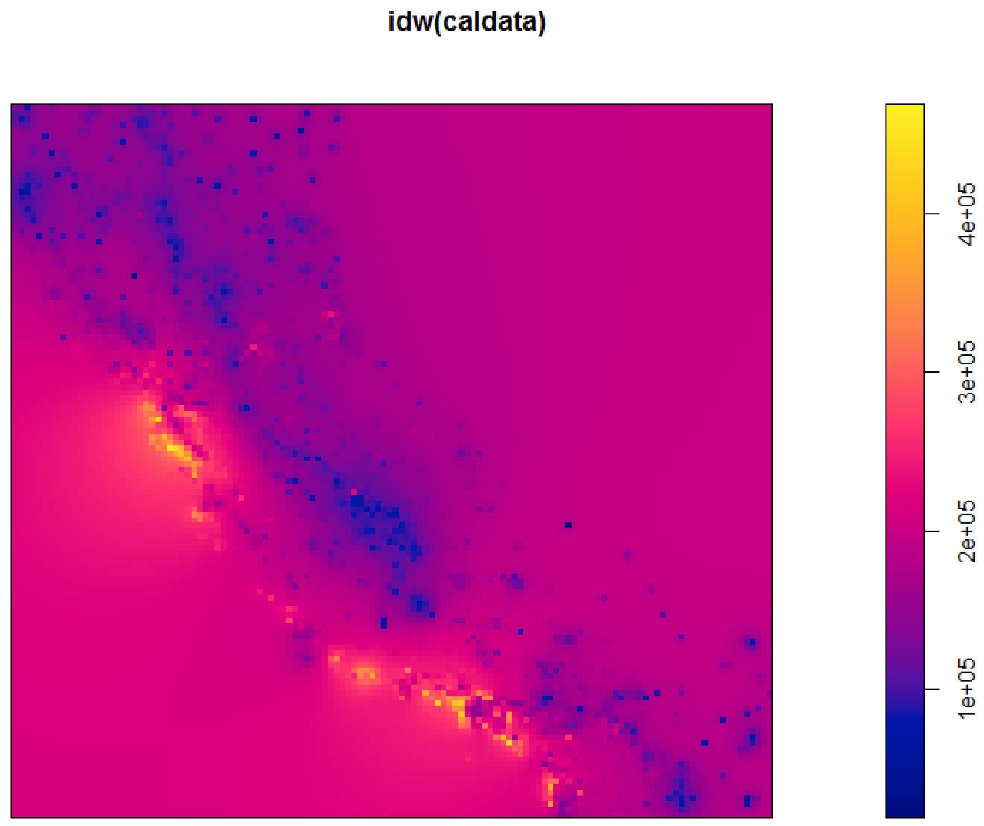
- $\hat{z}(x) = \frac{\sum_i w_i z_i}{\sum_i w_i}$

- Where

  $w_i = |x - x_i|^{-\alpha}$

Interpolation of Point Patterns

Example IDW of California house data



Interpolation of Point Patterns

Kriging

- The observed values z_i is modelled to be the outcome of random process

Ordinary kriging: $f(x) = \mu$

$$z_i = f(x_i) + v(x_i) + \varepsilon_i$$

Nugget effect

iid $N(0, \sigma^2)$

Deterministic trend function

Random function

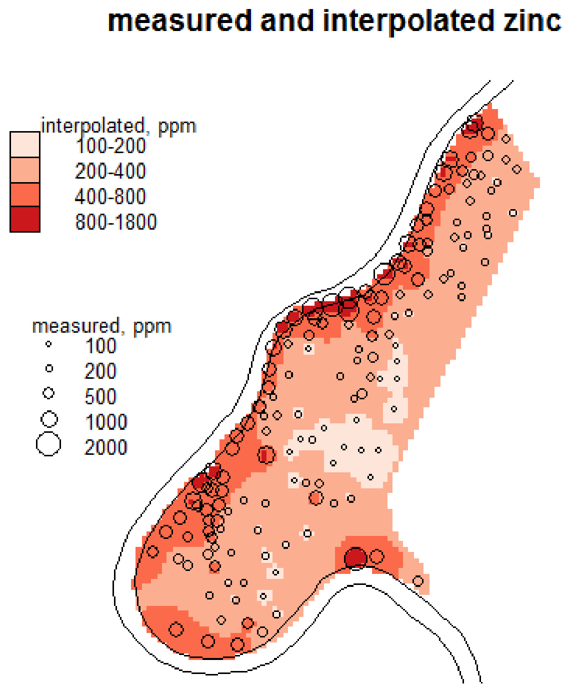
Interpolation of Point Patterns

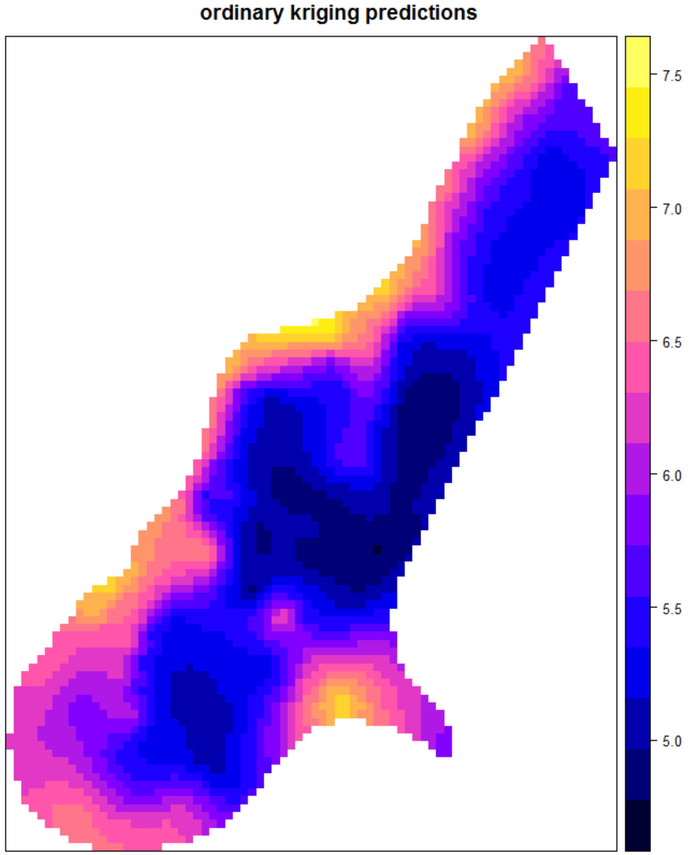
Kriging

- Assumptions:
- The function *v* is not specified directly but deduced by “working backwards” from $\gamma(d)$ and the observed data
- The idea is to specify a correlation/covariance structure for *v*
- Stationarity: The correlation function depends only on the distance between two vectors
- Typically the relationship is defined in terms of the variogram $\gamma(d)$
- If the process is stationary : $\gamma(d) = \frac{1}{2}E[(v(x_1) - v(x_2))^2]$
- The degree of correlation between $v(x_1)$ and $v(x_2)$ is assumed to decrease as distance increase

Interpolation of Point Patterns

Kriging





Spatial Prediction Models

Spatial Prediction Models

- Moran's I index for spatial correlation

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2}$$

- Where w_{ij} is a weights matrix W, specifying the degree of dependency between polygons *i* and *j*.
- The lagged mean $\bar{z}$:

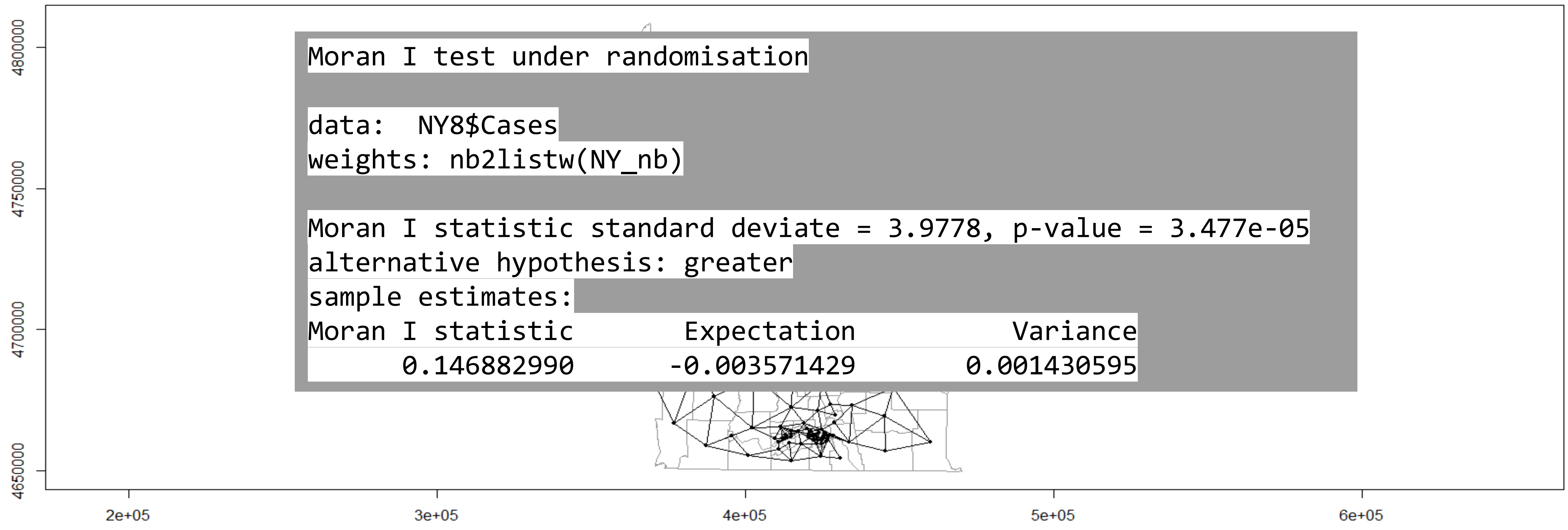
$$\bar{z}_i = \sum_{j \in \delta_i} \frac{1}{|\delta_i|} z_i$$

δ_i is the number of neighbours to *i*

$\frac{1}{|\delta_i|}$ is the weight w_{ij}

Spatial Prediction Models

- Example Remember to speak about the data!




```
Moran I test under randomisation

data:  NY8$Cases
weights: nb2listw(NY_nb)

Moran I statistic standard deviate = 3.9778, p-value = 3.477e-05
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.146882990      -0.003571429       0.001430595
```

Spatial Prediction Models

- Spatial Autoregression

SAR – Simultaneous autoregression

$$z_i = \mu + \sum_{j=1} b_{ij}(z_j - \mu) + \varepsilon_i \longrightarrow \text{iid } N(0, \sigma^2)$$

- $b_{ii} = 0$ and $b_{ij} = 0$ if polygon i is not adjacent to polygon j
- $b_{ij} = \lambda\, w_{ij}$ here λ is a parameter specifying the degree of spatial dependence
- $\lambda = 0\,, > 0, < 0$

CAR- Conditional autoregression

$$z_i | \{z_i : j \neq i\} \sim N(\mu + \sum_{j=1} c_{ij}(z_j - \mu), \tau_i^2)$$

- $c_{ii} = 0$ and $c_{ij} = 0$ if polygon i is not adjacent to polygon j
- $c_{ij} = \lambda\, w_{ij}$ here λ is a parameter specifying

  the degree of spatial dependence
- $\lambda = 0\,, > 0, < 0$
- τ_i^2 is the conditional variance of z_i given $\{z_j | j \neq i\}$

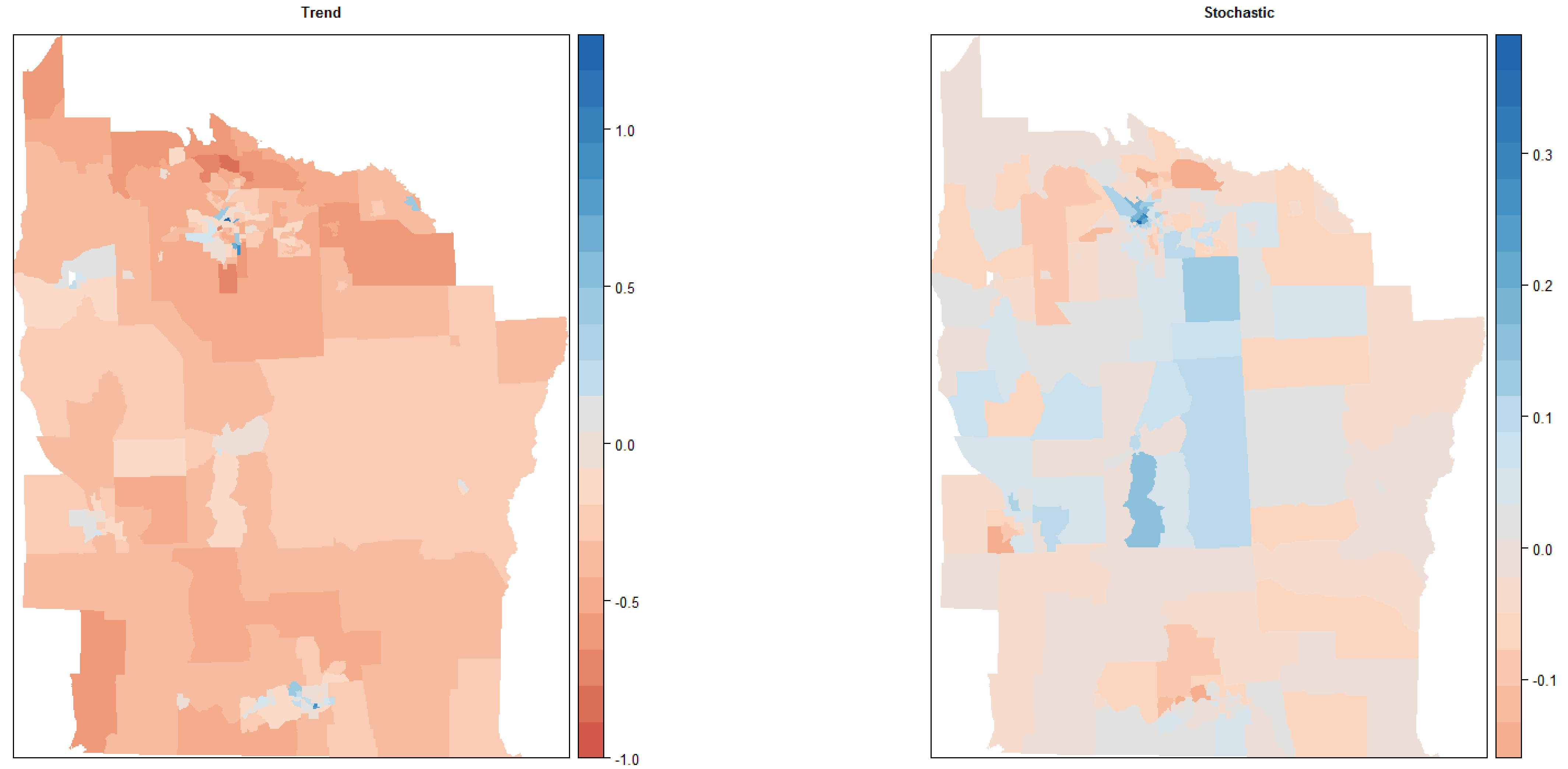
Spatial Prediction Models

- Example SAR – Simultaneous autoregression

```
Call: spautolm(formula = Z ~ PEXPOSURE + PCTAGE65P +
PCTOWNHOME, data = NY8,
    listw = NYlistw)

Residuals:
     Min       1Q   Median       3Q      Max
-1.56754 -0.38239 -0.02643  0.33109  4.01219

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -0.618193   0.176784 -3.4969 0.0004707
PEXPOSURE    0.071014   0.042051  1.6888 0.0912635
PCTAGE65P    3.754200   0.624722  6.0094 1.862e-09
PCTOWNHOME  -0.419890   0.191329 -2.1946 0.0281930

Lambda: 0.040487 LR test value: 5.2438 p-value: 0.022026
Numerical Hessian standard error of lambda: 0.017199

Log likelihood: -276.1069
ML residual variance (sigma squared): 0.41388, (sigma:
0.64333)
Number of observations: 281
Number of parameters estimated: 6
AIC: 564.21
```

Spatial Prediction Models

- Example SAR – Simultaneous autoregression



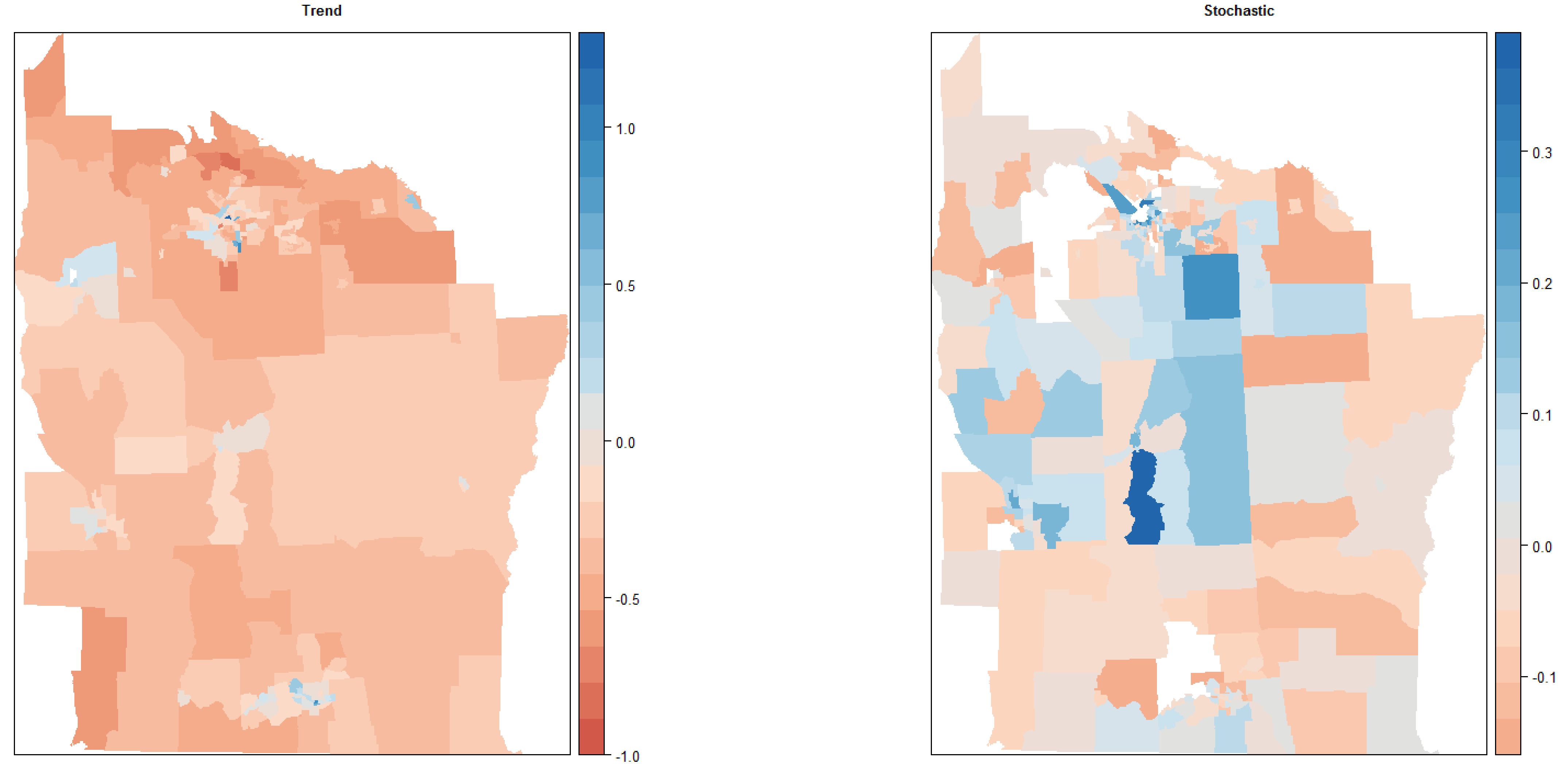
Spatial Prediction Models

- Example CAR – Conditional autoregression

```
Call: spautolm(formula = Z ~ PEXPOSURE + PCTAGE65P +
PCTOWNHOME, data = NY8,
    listw = NYlistw, family = "CAR")

Residuals:
      Min        1Q    Median        3Q       Max
-1.539732 -0.384311 -0.030646  0.335126  3.808848

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -0.648362   0.181129 -3.5796 0.0003442
PEXPOSURE    0.077899   0.043692  1.7829 0.0745986
PCTAGE65P    3.703830   0.627185  5.9055 3.516e-09
PCTOWNHOME  -0.382789   0.195564 -1.9574 0.0503053

Lambda: 0.084123 LR test value: 5.8009 p-value: 0.016018
Numerical Hessian standard error of lambda: 0.030868

Log likelihood: -275.8283
ML residual variance (sigma squared): 0.40758, (sigma:
0.63842)
Number of observations: 281
Number of parameters estimated: 6
AIC: 563.66
```

Spatial Prediction Models

- Example CAR – Conditional autoregression



Appendix

R packages

Spatstat

- Creation, manipulation and plotting of point patterns,
- Exploratory data analysis
- Simulation of point process models,
- Parametric model fitting,
- Hypothesis tests and diagnosis

Spdep

- Compute basis spatial statistics such as Moran’s I
- Create neigbour objects of class nb
- Create list objects of class lw,
- Work out neighbour relations from polygons
- Colour mapped regions on the basis of dervied statistics

References

Books:

- Bivand, R.S., Pebesma, E.J. And Gómez-Rubio, V.G. (2008) Applied Spatial Data Analysis with R. New York: Springer.
- Brunsdon, C. And Comber, L. (2015) An Introduction to R for Spatial Analysis & Mapping. London: Sage.
- Baddeley, A. & Rubak, E and Turner, R. (2015) Spatial Point Patterns: Methodology and Applications with R. Chapman and Hall/CRC Press, 2015.

- Papers:

- http://spatstat.github.io/resources/spatstatJSSpaper.pdf

- R Code

- https://github.com/sojohan/MLADS

Q&A

Rate this session!

Go to: http://aka.ms/mlads-surveys
to complete a 1-question poll about this session

You will also receive a post-event survey where you can provide detailed feedback.

Network

To ensure the strongest connectivity please use the dedicated MLADS network.

SSID: MLADS 2016
Password: MLADS2016!

Microsoft