Spatial significance hotspot mapping using the Gi* statistic

Spencer Chainey

Director of Geographic Information Science

UCL Jill Dando Institute of Crime Science

Overview

- The value of testing for spatial significance
- Quick review of common hotspot mapping techniques
- LISA statistics
- Using the Gi* statistic to identify patterns of spatial significance

The value of significance testing

Statistical significance

- 95%, 99%, 99.9%
- E.g. 99%: 1 in 100 chance that the observation would have just occurred naturally

  i.e. what we are observing is extremely unusual

Example of spatial significance testing

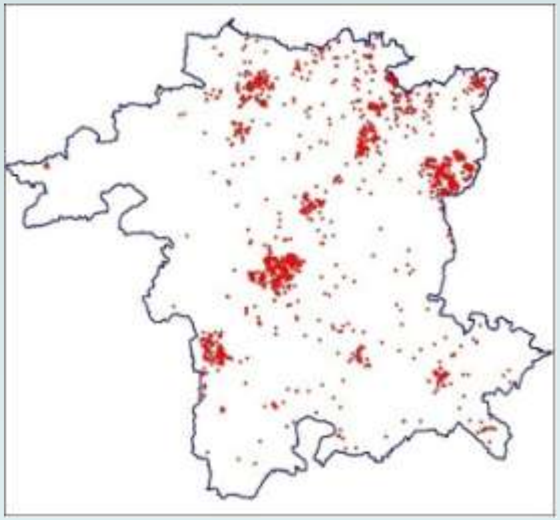
Nearest Neighbour Index (NNI)

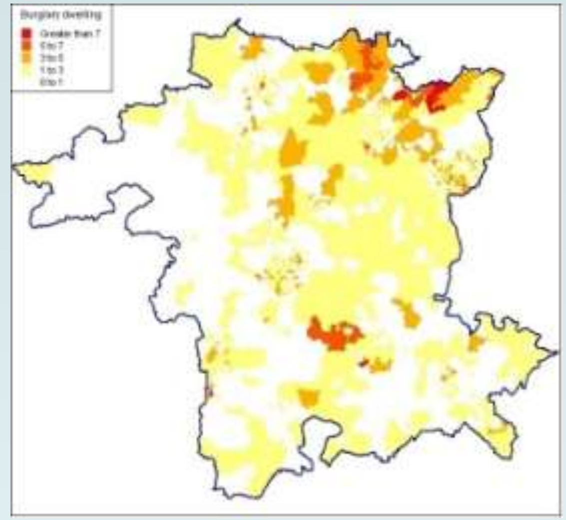
- Identifies if there is statistical evidence of clustering, and therefore hotspots, in point data
  - *How much data do I need before I can use a technique that maps where the hotspots are?*
- ***Interpretation of result:***
  - *If NNI = 1; point data is randomly distributed*
  - *If NNI < 1; point data shows evidence of clustering*
  - *If NNI > 1; point data is uniformly distributed*
- ***Statistical significance measure:*** *Test statistic (Z-score) and P value to indicate if result is statistically significant*
- Software
  - CrimeStat
  - ArcGIS Spatial Statistics Tools - **DEMO**

Review of common techniques

Hotspot mapping techniques

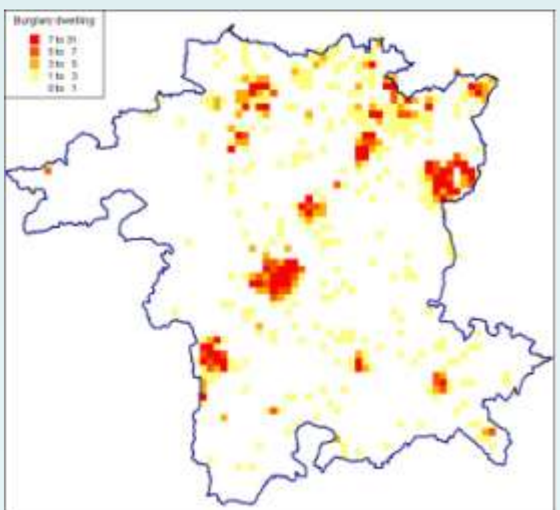
Point map

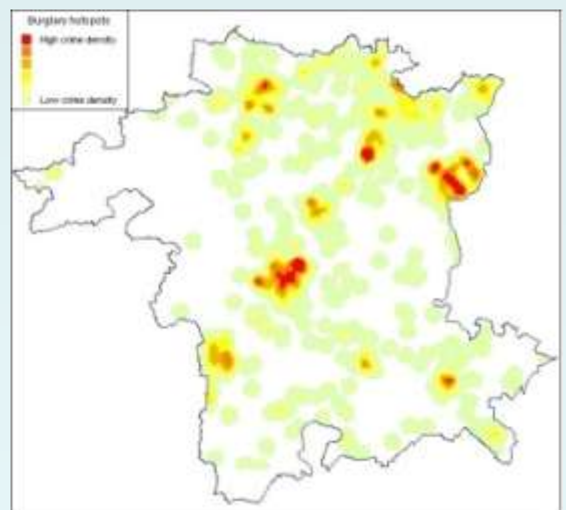




Thematic map of geographic administrative units

Grid thematic map





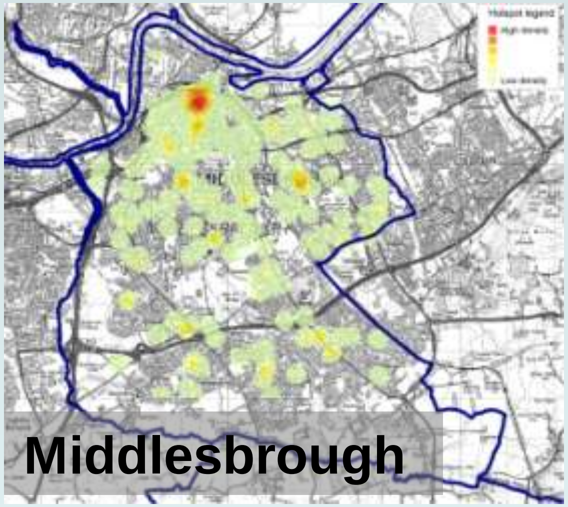
Kernel density estimation map

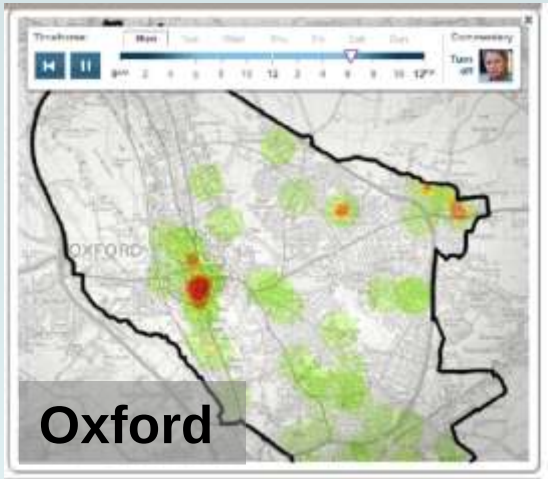
- *Best for location, size, shape and orientation of hotspot*
- *9 out of 10 intelligence professionals prefer it*

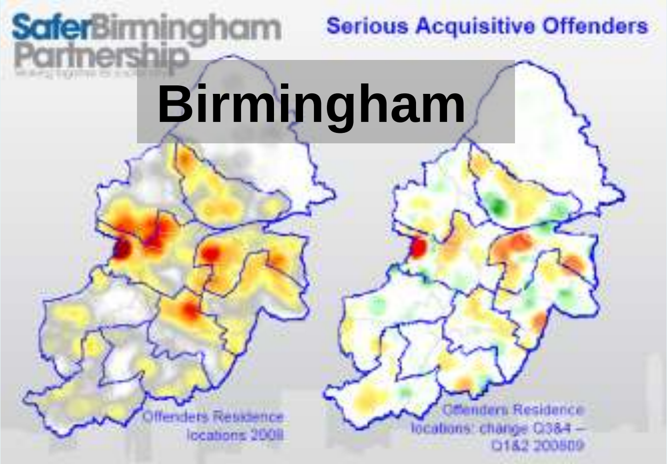
UCL Jill Dando Institute of Crime Science

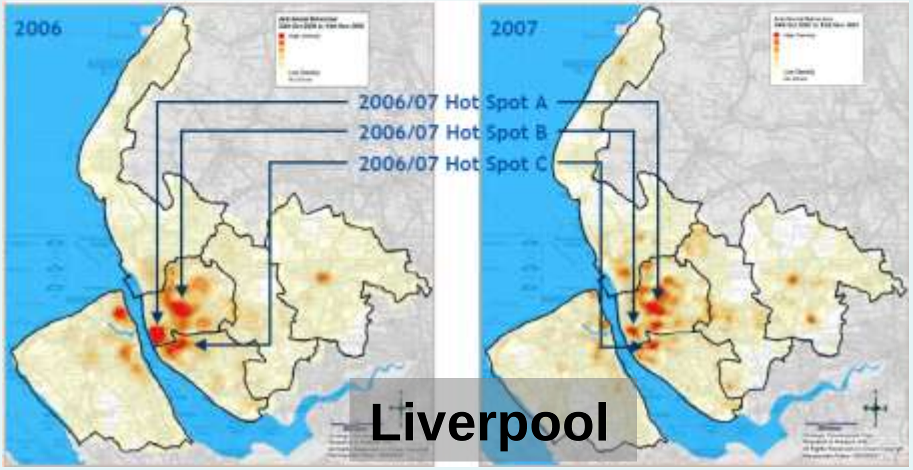
Kernel density estimation

Examples of use in presentations from the UK Crime Mapping Conference, 2009





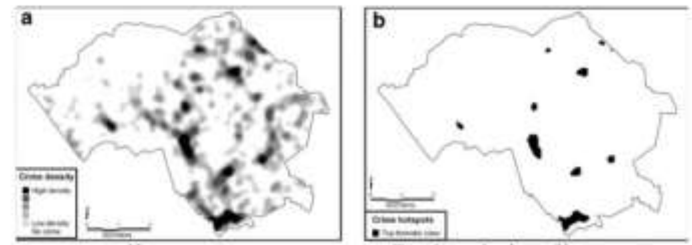




Comparing KDE to other methods

- Results from research
  - Prediction Accuracy Index

Chainey,S.P., Tompson,L., Uhlig,S. (2008). ***The utility of hotspot mapping for predicting spatial patterns of crime***. Security Journal



Hotspot map | Top thematic class of hotspot map

Figure 3 Hotspots were determined by selecting the uppermost thematic class calculated using the five classes and the default values generated from applying the quantile thematic range method in MapInfo.

Table 6 PAI values for different hotspot mapping techniques

Hotspot mapping technique	*Average PAI (01/01/2003)*	*Average PAI (13/03/2003)*
Spatial ellipses 250 m	1.74	2.25
Spatial ellipses 500 m	*1.24*	*1.52*
Spatial ellipses HSD	1.69	2.03
Thematic mapping of output areas	1.91	2.38
Thematic mapping of grids 250 m	2.00	2.34
Thematic mapping of grids HSD	2.06	2.63
Kernel density estimation	**2.90**	**3.41**

Values in bold indicate the highest values and values in italics indicate the lowest PAI values. Results are presented for each of the dates when hotspot maps were generated. These results show that KDE consistently produced the best hotspot maps for predicting future events.

Comparing KDE to other methods

Table 7 PAI values for different hotspot mapping techniques, by crime type

Hotspot mapping technique	*Residential burglary*	*Street crime*	*Theft from vehicle*	*Theft of vehicle*
(a) PAI values calculated from the 1 January 2003 measurement date				
Spatial ellipses 250 m	1.38	2.36	2.18	1.65
Spatial ellipses 500 m	1.34	*1.46*	1.54	*0.82*
Spatial ellipses HSD	1.43	2.45	2.12	1.29
Thematic mapping of output areas	*1.10*	4.20	*1.17*	1.18
Thematic mapping of grids 250 m	1.70	4.04	1.82	1.37
Thematic mapping of grids HSD	1.68	3.46	2.12	2.06
Kernel density estimation	**2.31**	**4.68**	**2.29**	**2.32**
(b) PAI values calculated from the 13 March 2003 measurement date				
Spatial ellipses 250 m	1.32	2.59	2.15	2.93
Spatial ellipses 500 m	1.31	*1.40*	*1.55*	1.82
Spatial ellipses HSD	1.29	2.63	2.63	*1.59*
Thematic mapping of output areas	*1.25*	3.32	2.93	2.01
Thematic mapping of grids 250 m	1.67	3.58	2.43	1.66
Thematic mapping of grids HSD	1.95	4.14	2.55	1.89
Kernel density estimation	**2.33**	**4.59**	**3.66**	**3.05**

Values in bold indicate the highest values and values in italics indicate the lowest PAI values. These results show that KDE consistently produced the best hotspot maps for predicting spatial patterns of crime for all crime types, and that in some cases STAC was not the worst performer. Instead, thematic mapping of output areas generated the lowest PAI values for residential burglary, and in one case for theft from vehicles.

Comparing KDE to other methods

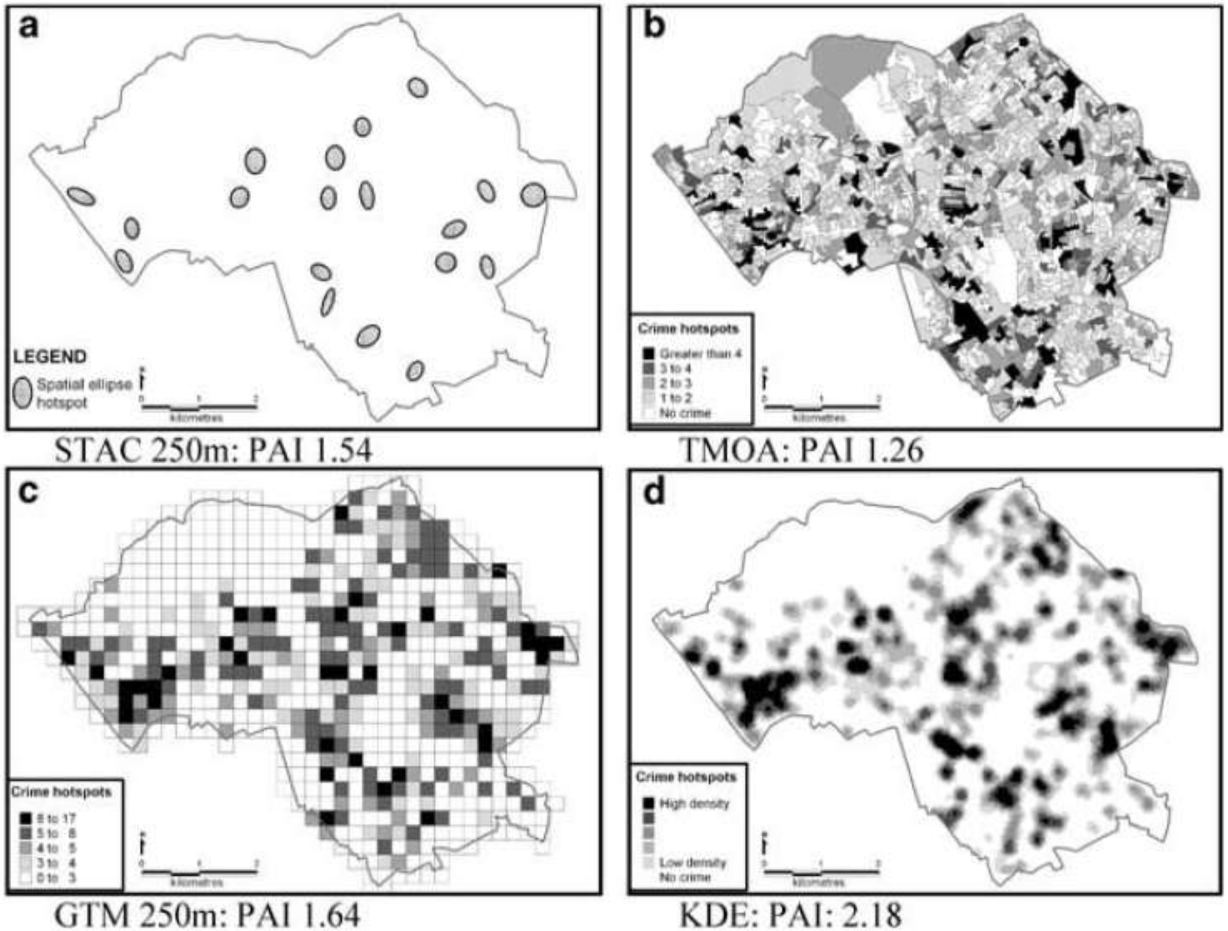
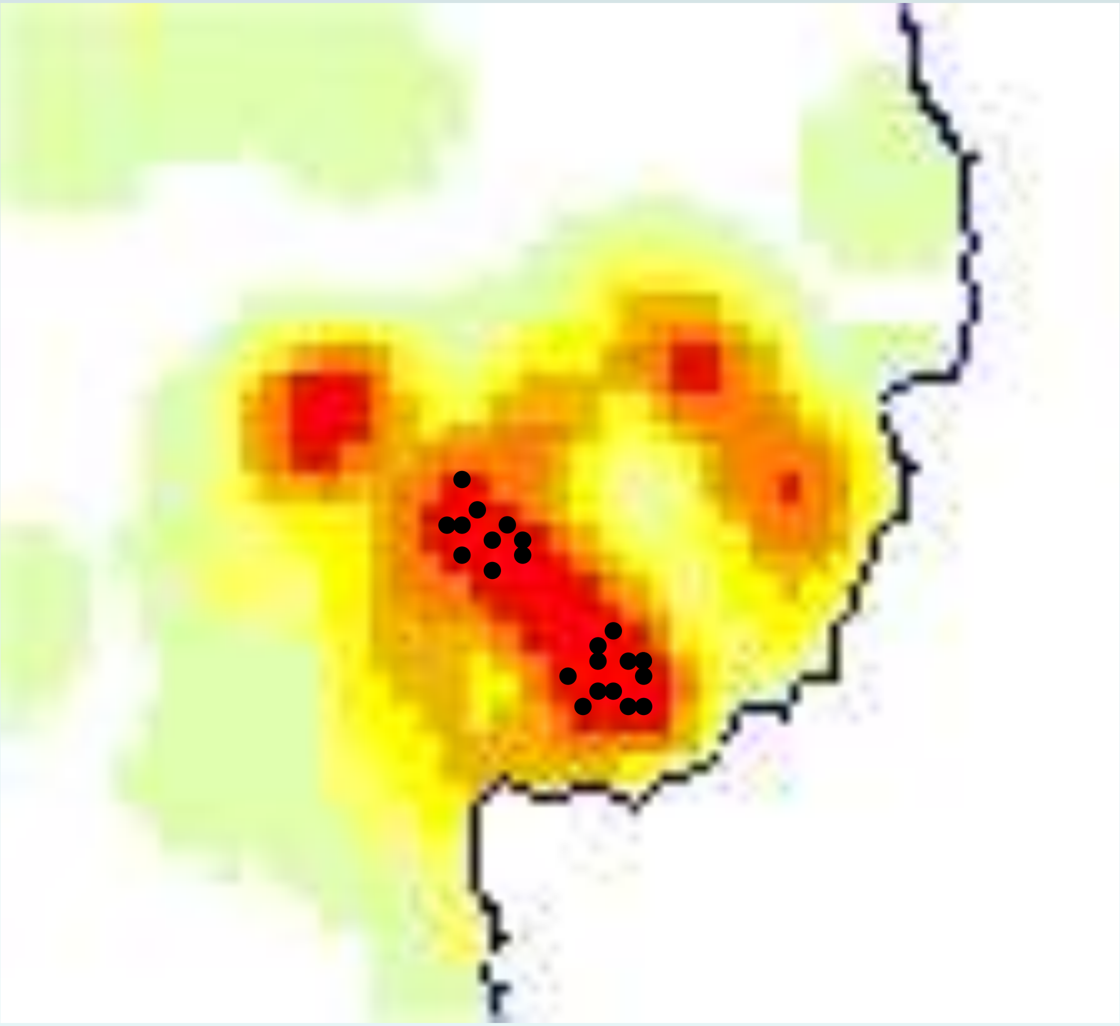
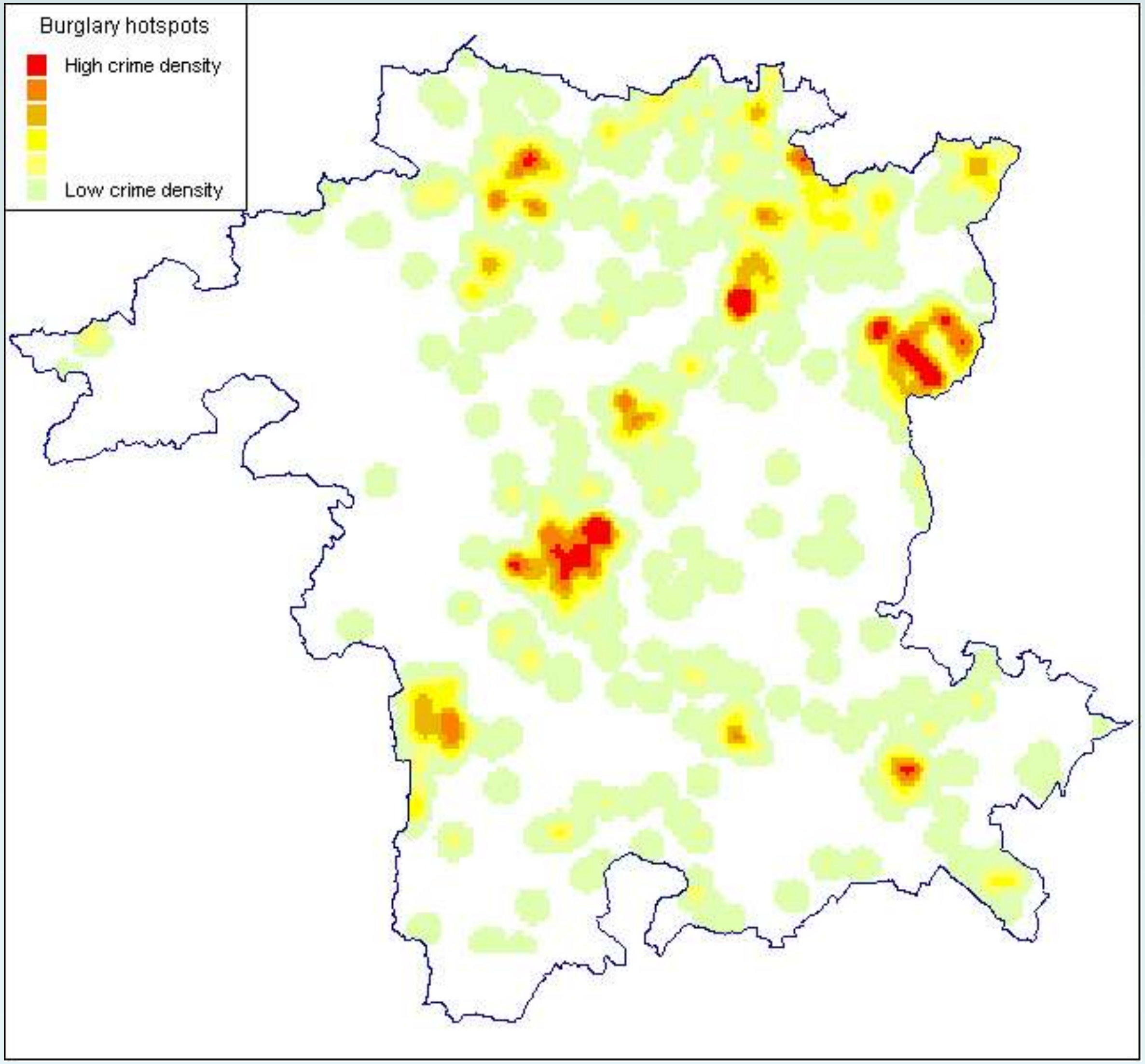


Figure 4. Hotspot maps generated from 3 months of residential burglary input data (measurement date of the 1 January 2003) using (a) STAC, (b) thematic mapping of output areas, (c) grid thematic mapping and (d) KDE. Each map is shown with its PAI value, based on 1 month of measurement data.

KDE weaknesses: smoothes between areas

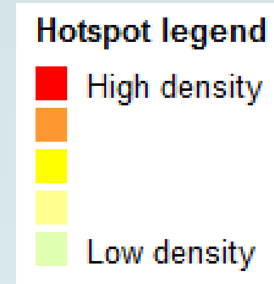


KDE weaknesses: attention drawn to the big blobs

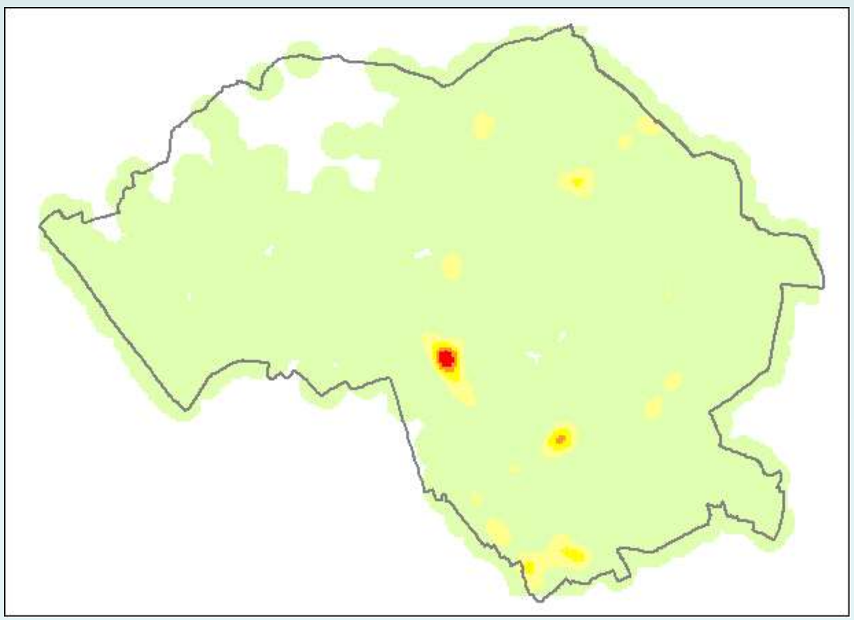


KDE weaknesses: how many hotspots?!

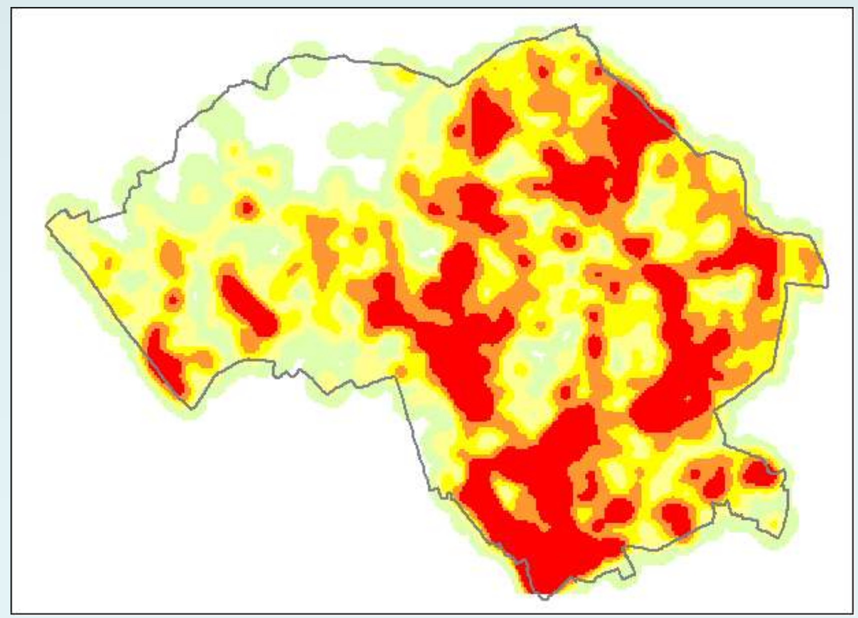
- Thematic thresholds to apply?
- Left to the whims and fancies of the map producer
- Trial and error, experimentation, experience, whatever suits your circumstance



One main hotspot



Lots of hotspots!



Local Indicators of Spatial Association (LISA statistics)

UCL Jill Dando Institute of Crime Science

Global statistics of spatial association

- Spatial autocorrelation
  - Moran’s I and Geary’s C
  - In practice are of marginal value for crime data
  - Global statistics may help inform the nature of the general distribution of crime
  - But may only summarise an enormous number of possible disparate spatial relationships in our data

UCL

UCL Jill Dando Institute of Crime Science

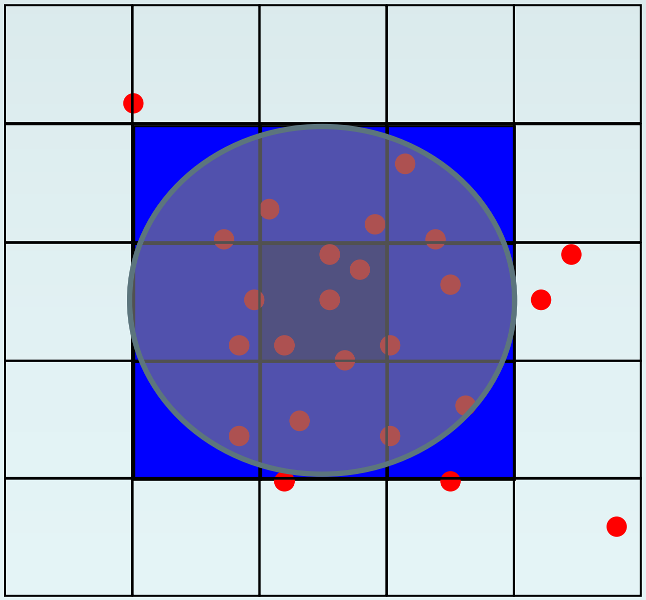
LISA statistics

- Identify the local association between an observation and its neighbours, up to a specified distance from the observation

- LISA statistics help inform the nature of the local distribution of crime

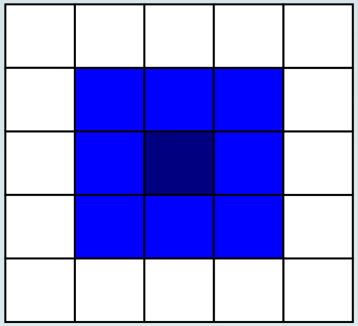
LISA statistics

- Requires data to be aggregated to some form of geographic unit (e.g. Census block, grid cell)
  - Adjacency/contiguity (i.e. which neighbours to consider)
    - Units within a specified radius



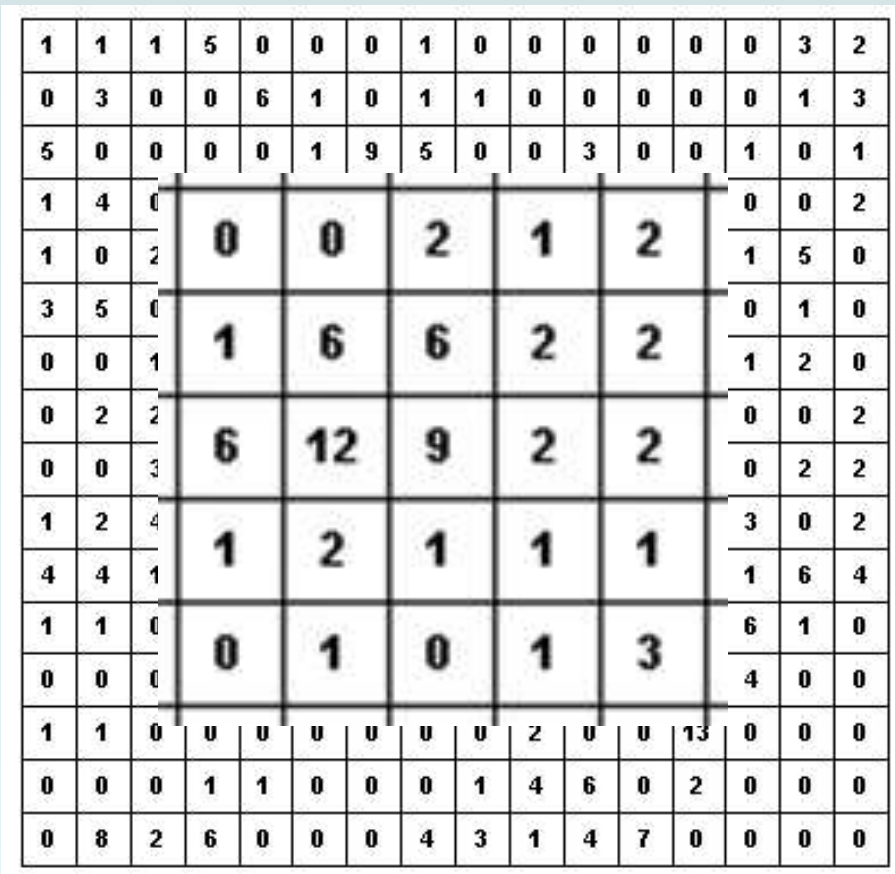
LISA statistics

- Local Moran’s I and Local Geary’s C
  - Compare if the value for each observation is similar to those that neighbour it
  - Effectively produce Moran’s I or Geary’s C for each cell



- Gi and Gi*
  - Compare local averages to global averages
- Application of a spatial significance test
  - *Where are the really unusual patterns of spatial association?*
  - *What’s hot and what’s not hot?*
  - Identifies if local pattern of crime is (statistically) significantly different to what is generally observed across the whole study area
- Gi and Gi* have become the most popular amongst crime analysts

Gi and Gi* statistics



- Each cell is a georeferenced object with a value associated with it
- Eighth row, eighth column = 9
- Null hypothesis: there is no association between the values of crime counts at site *i* and its neighbours, which we will call the *j*s, up to a distance of d, measured from *i* in all directions
  - The sum of values at all the *j* sites within a radius *d* of *i* is not more (or less) than one would expect by chance given all the values in the entire study area (both within and beyond the distance *d*).

Gi and Gi* statistics

- What's the difference between Gi and Gi*?
  - Gi* statistic includes the value of the point in its calculation
  - Gi excludes this value and only considers the value of its nearest neighbours (within *d*) against the global average (which also does not include the value at site *i*)
- Gi* is the more popular of the two statistics because it considers all values within *d*
- Equation:

0	0	2	1	2
1	6	6	2	2
6	12	9	2	2
1	2	1	1	1
0	1	0	1	3

$$G_i^*(d) = \frac{\sum_j w_{ij}(d)x_j - W_i^* \bar{x}^*}{s^*\{[(nS_{1i}^*) - W_i^{*2}]/(n-1)\}^{1/2}}, \quad \text{for all } j,\ x_j \neq 0$$

Gi* statistic

- Does local spatial association exist?
  - Lots of high counts of crime close together
    - Gi* values will be positive for each cell
  - Lots of low counts of crime close together
    - Gi* values will be negative for each cell

- Software
  - Rook's Case Excel Add-in (University of Ottawa)
  - ArcGIS 9.2 and above (Spatial Statistics Toolkit)

Gi* statistic

An example

- Calculating the Gi* statistics for our 16x16 matrix dataset
- **Lag distance** – distance at which we wish to explore local spatial association
  - Cell size for this example is 125m
  - Set lag distance to 177m - all immediate surrounding cells for each cell will be considered

    i.e. the distance to cells in a diagonal direction from each cell of interest is 177m (by Pythagoras theorem)
- **Lags** – if we calculate our statistics against a lag of 1 then we only consider nearest neighbours within one lag distance of each point
  - A lag of 4 for our 16x16 matrix will calculate Gi* values within a distance d of 177, 354, 531, 708 i.e. multiples of 177

1	1	1	5	0	0	0	1	0	0	0	0	0	0	3	2
0	3	0	0	6	1	0	1	1	0	0	0	0	0	1	3
5	0	0	0	0	1	9	5	0	0	3	0	0	1	0	1
1	4	0	2	0	5	0	0	0	1	1	0	0	0	0	2
1	0	2	3	0	3	6	0	1	2	0	0	0	1	5	0
3	5	0	4	0	0	0	2	1	2	1	1	0	0	1	0
0	0	1	1	8	1	6	6	2	2	0	1	0	1	2	0
0	2	2	2	4	6	12	9	2	2	3	6	2	0	0	2
0	0	3	8	5	1	2	1	1	1	5	0	0	0	2	2
1	2	4	2	1	0	1	0	1	3	0	0	2	3	0	2
4	4	1	0	0	1	1	1	0	2	1	4	2	1	6	4
1	1	0	0	0	0	0	0	1	4	5	2	2	6	1	0
0	0	0	2	0	0	1	0	2	6	1	3	0	4	0	0
1	1	0	0	0	0	0	0	0	2	0	0	13	0	0	0
0	0	0	1	1	0	0	0	1	4	6	0	2	0	0	0
0	8	2	6	0	0	0	4	3	1	4	7	0	0	0	0

Gi* statistic

An example

- Run Rook’s Case
- Excel spreadsheet is populated with Gi* Z scores statistics for each point, and for each lag
  - The Gi* statistic is listed under the ‘z-Gi*(d)’
  - Gi is ‘z-Gi(d)’
- Cell 120
  - This is the point with the value of 9 in the eighth column of the eighth row

Gi* value =

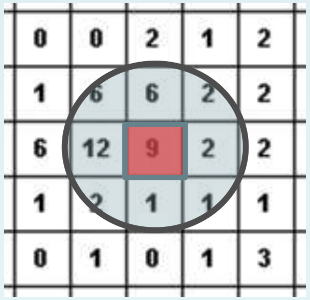
Gi* statistic

An example – VIDEO SHOWN DURING PRESENTATION

Gi* statistic

An example

- Run Rook's Case
- Excel spreadsheet is populated with Gi* statistics for each point, and for each lag
  - The Gi* statistic is listed under the 'z-Gi*(d)'
  - Gi is 'z-Gi(d)'
- Cell 120
  - This is the point with the value of 9 in the eighth column of the eighth row



Gi* value = 4.1785

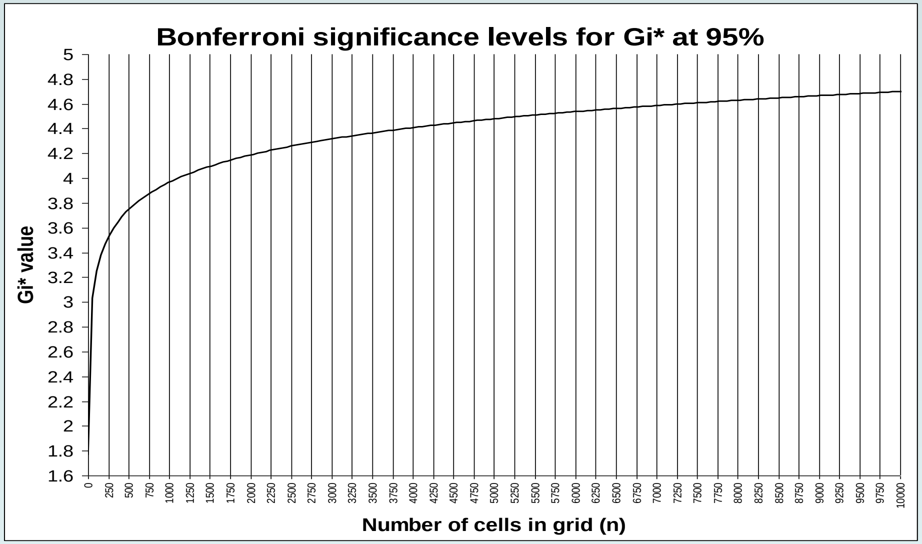
– *Gi* value is positive*

– *In relative terms (to the pattern across the whole study area), lots of cells with high counts of crime close together*

Gi* statistic

Statistical significance

- Ord and Getis suggest Bonforonni test (a statistical procedure that performs multiple tests to determine levels of significance in a data sample)



  - A common significance level to use is 95%
  - 95% significance level for our sample of 256 records is approximately 3.55
  - But difficult to apply: no software to calculate this!?
  - And current process (finger along a graph) is inadequate

Gi* statistic

- Gi* results are Z scores
  - Z scores indicate the place of a particular value in a dataset relative to the mean, standardized with respect to the standard deviation
  - Z = 0 is equivalent to the sample/data mean
  - Z < 0 is a value less than the mean
  - Z > 0 is a value greater than the mean
- Recall: Gi* compares local averages to global averages
  - Identifies if local pattern of crime is different to what is generally observed across the whole study area
- Z score is used extensively in determining confidence thresholds and in assessing statistical significance

Gi* statistic

Statistical significance

- Z score values for levels of statistical significance:
  - 90% significant: >= 1.645
  - 95% significant: >= 1.960
  - 99% significant: >= 2.576
  - 99.9% significant: >= 3.291 (if a cell has this value, then something exceptionally unusual has happened at this location in terms of the spatial concentration of crime)

  Universal Z score values: the same values apply, regardless of crime type, the location of your study area, the size of your study area ...
- Cell 120 - point with the value of 9 in the eighth column of the eighth row
  - Gi* value = 4.1785
  - Greater than 99.9% significant

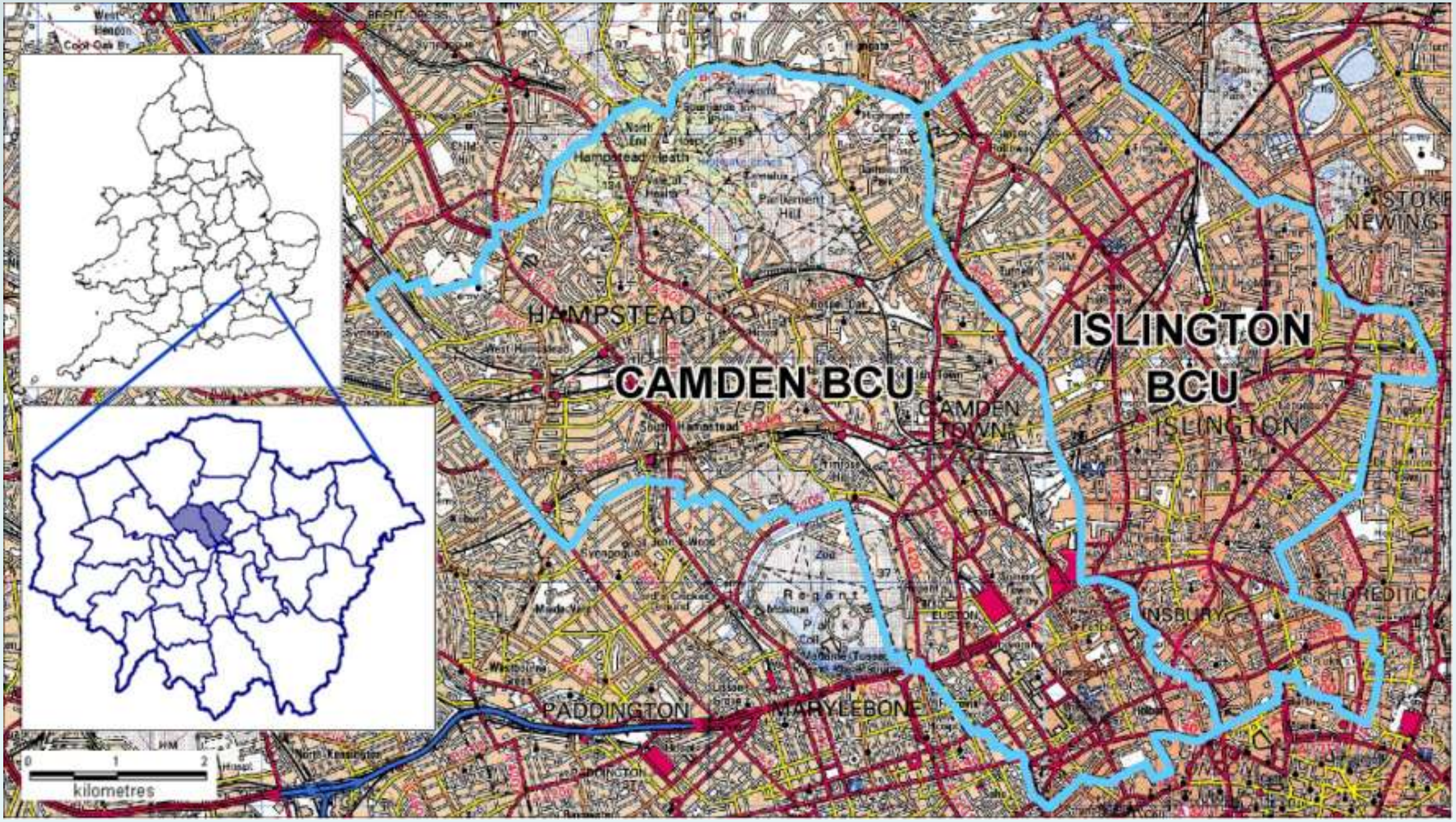
Gi* statistic and Rook's Case

1. Generate a grid in my GIS
2. Calculate a count of crime per grid cell
3. Export my data
   - X, Y, count
   - Open in Excel
4. Run Rook's Case
5. Import results to my GIS
6. Join my results to my grid
7. Thematically map the results (using the Z score statistical significance threshold values)

Another example - study area

London Metropolitan Police: Camden and Islington BCUs

Hotspots of robbery from a person (street robbery/mugging)



Step 1: Input data – creating a grid

ArcGIS

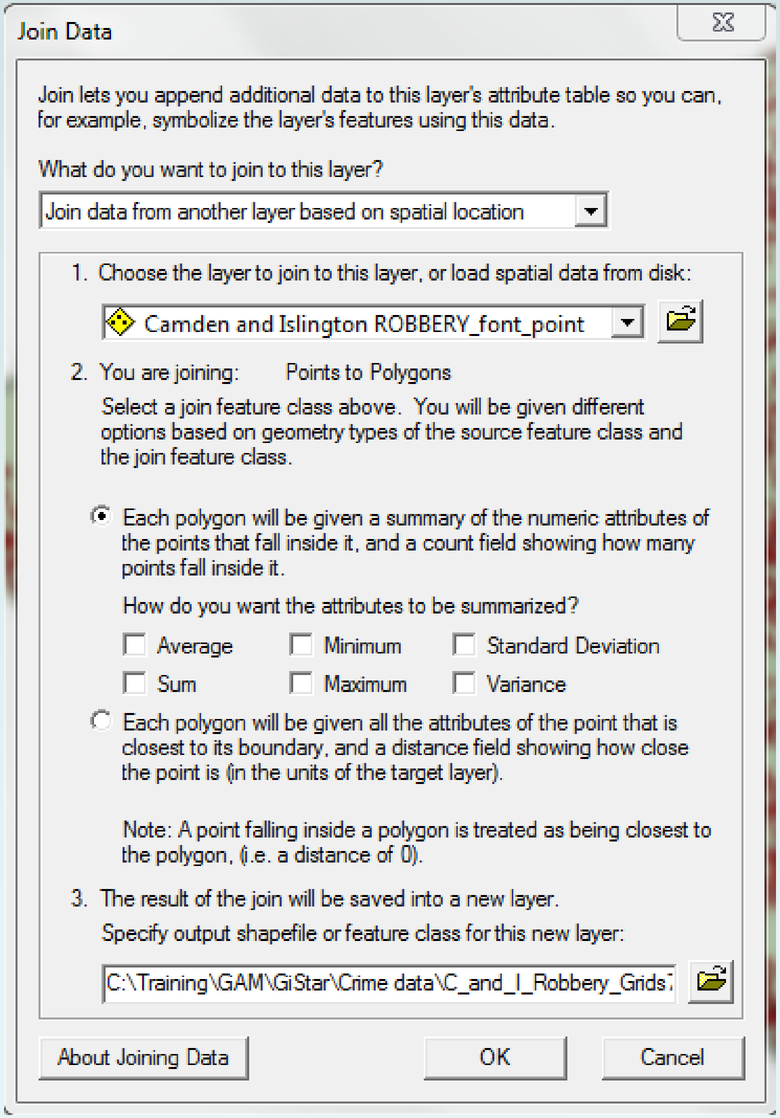
- v9.3 or lower: use Hawth's Tools (free) or some other grid creating tool
- v10: 'fishnet' tool built in
- Grid cell size?
  - Good starting point: divide shorter side of MBR by 100

Very important: we need to cookie cut our grid cell lattice to our study area

Step 1: Input data – creating a grid – VIDEO SHOWN DURING PRESENTATION

Step 2: Input data – count of crime

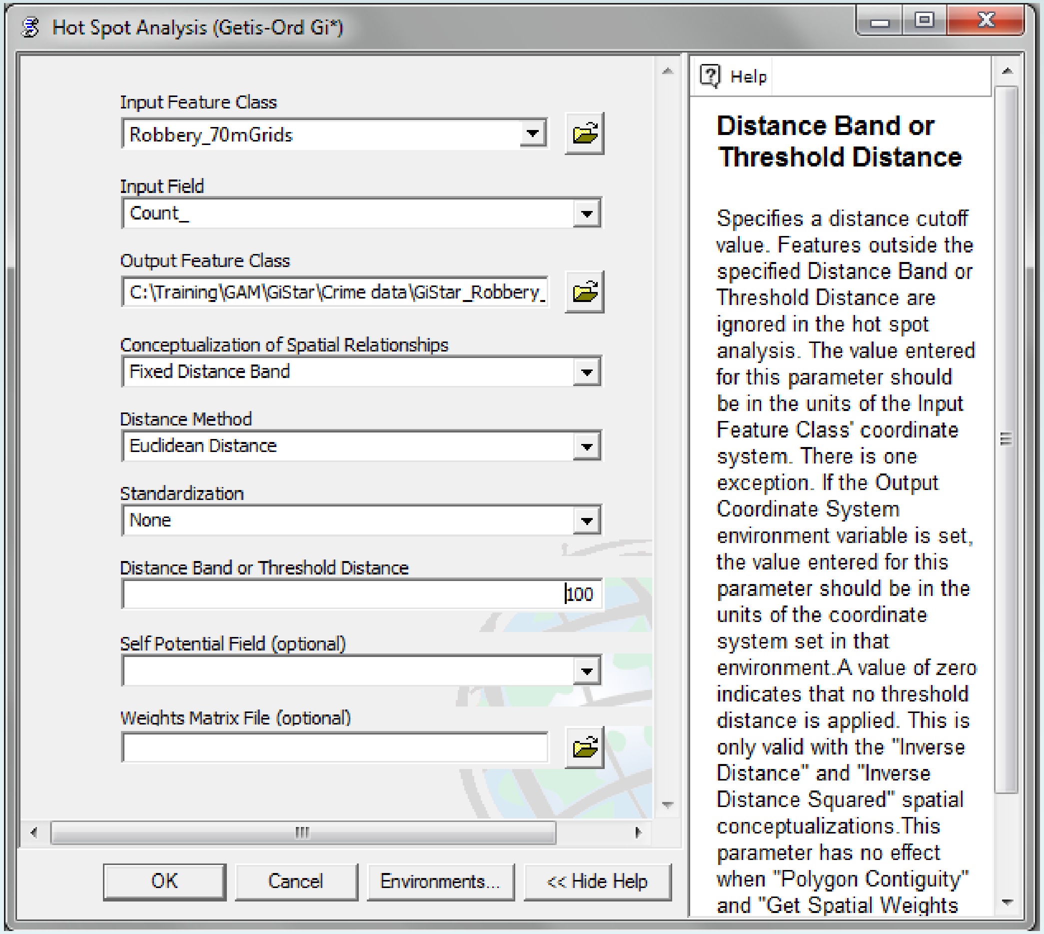
- ArcGIS
  - Geographically referenced grid lattice (geodatabase file or shape file)
  - Count of crime in each grid cell
    - Do this by performing a ***Join*** against the grid cells data



Step 2: Input data – count of crime – VIDEO SHOWN DURING PRESENTATION

Step 3: Running Gi*

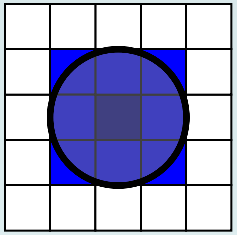
- ArcGIS
  - Spatial Statistics Toolbox>Mapping Clusters
  - Hot Spot Analysis (Getis – Ord Gi*)
  - Lag distance (known in ArcGIS as Distance Band or Threshold Distance)
    - Why 100m?



Step 3: Running Gi*

Lag distance (ArcGIS: Distance Band or Threshold distance)

- Want too include all immediate neighbours in calculation
- Calculated in relation to cell size
- SQRT((70*70)+(70*70)) = 98.99
  - 70 is the cell size we chose
- We'll round it up to 100 to ensure we capture all immediate neighbours: no more, no less



Step 3: Running Gi* – VIDEO SHOWN DURING PRESENTATION

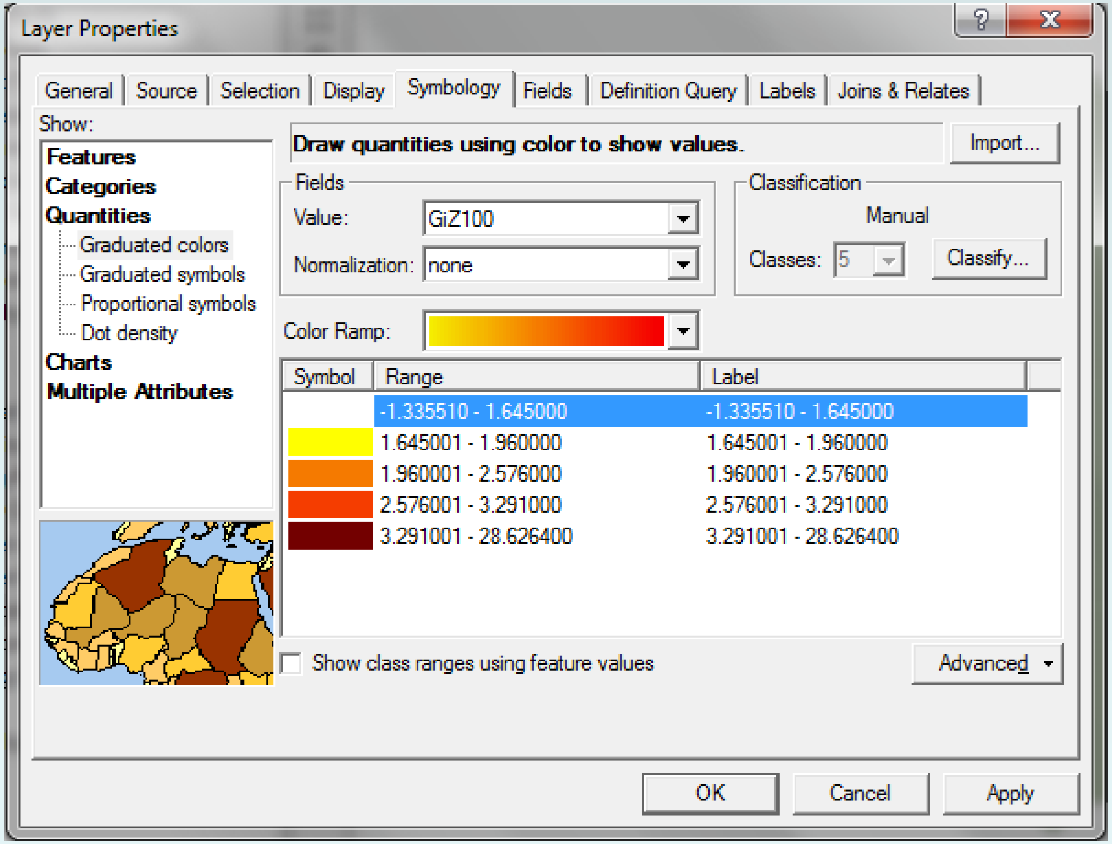
Step 4: Displaying and interpreting the results

- Gi* results are Z score values
- Use these to determine thematic class values
  - 90% significant: >= 1.645
  - 95% significant: >= 1.960
  - 99% significant: >= 2.576
  - 99.9% significant: >= 3.291

Step 4: Displaying and interpreting the results

Thematic class values:

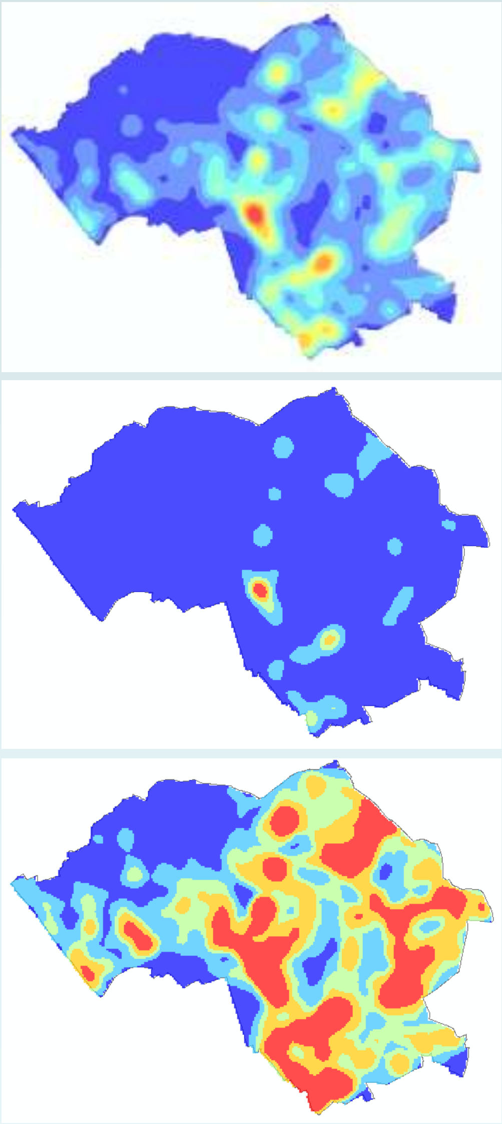
- 90% significant: >= 1.645
- 95% significant: >= 1.960
- 99% significant: >= 2.576
- 99.9% significant: >= 3.291

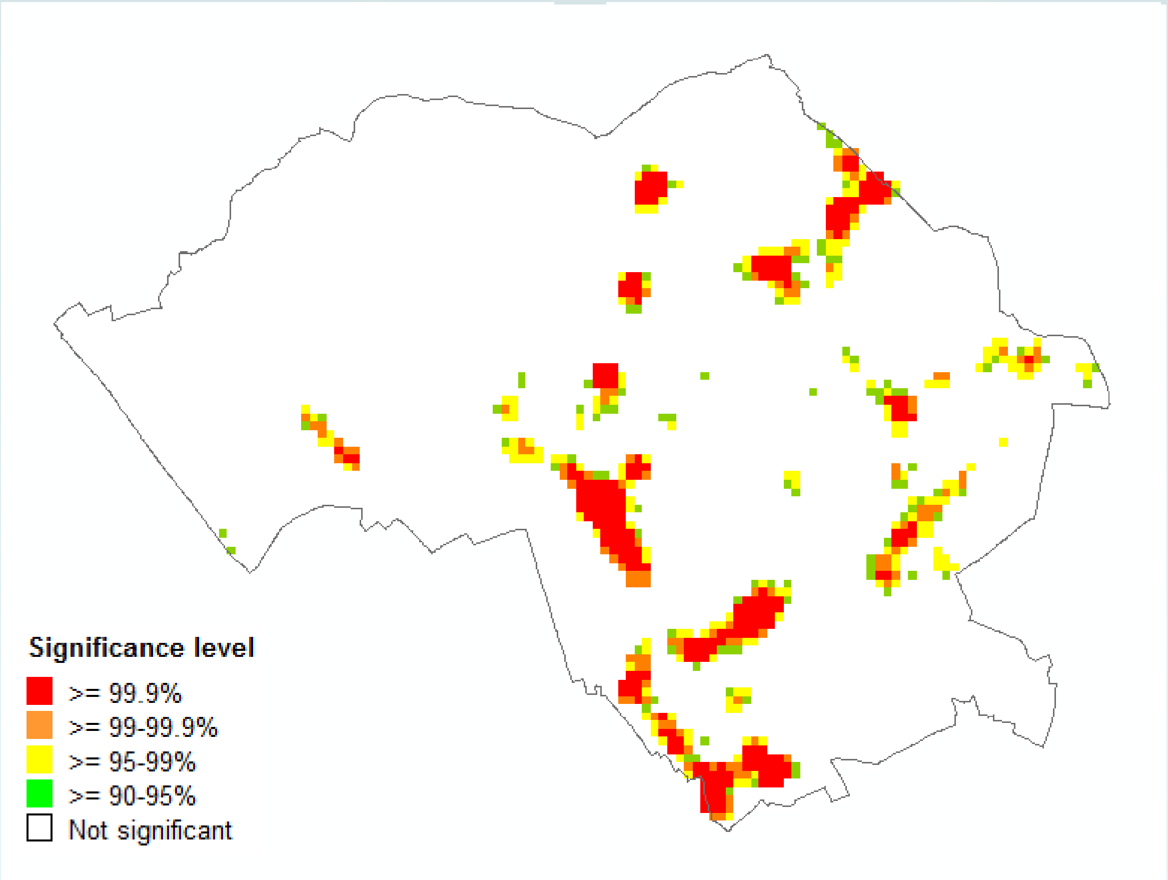


Step 4: Displaying and interpreting the results

– VIDEO SHOWN DURING PRESENTATION

Kernel density estimation and Gi*

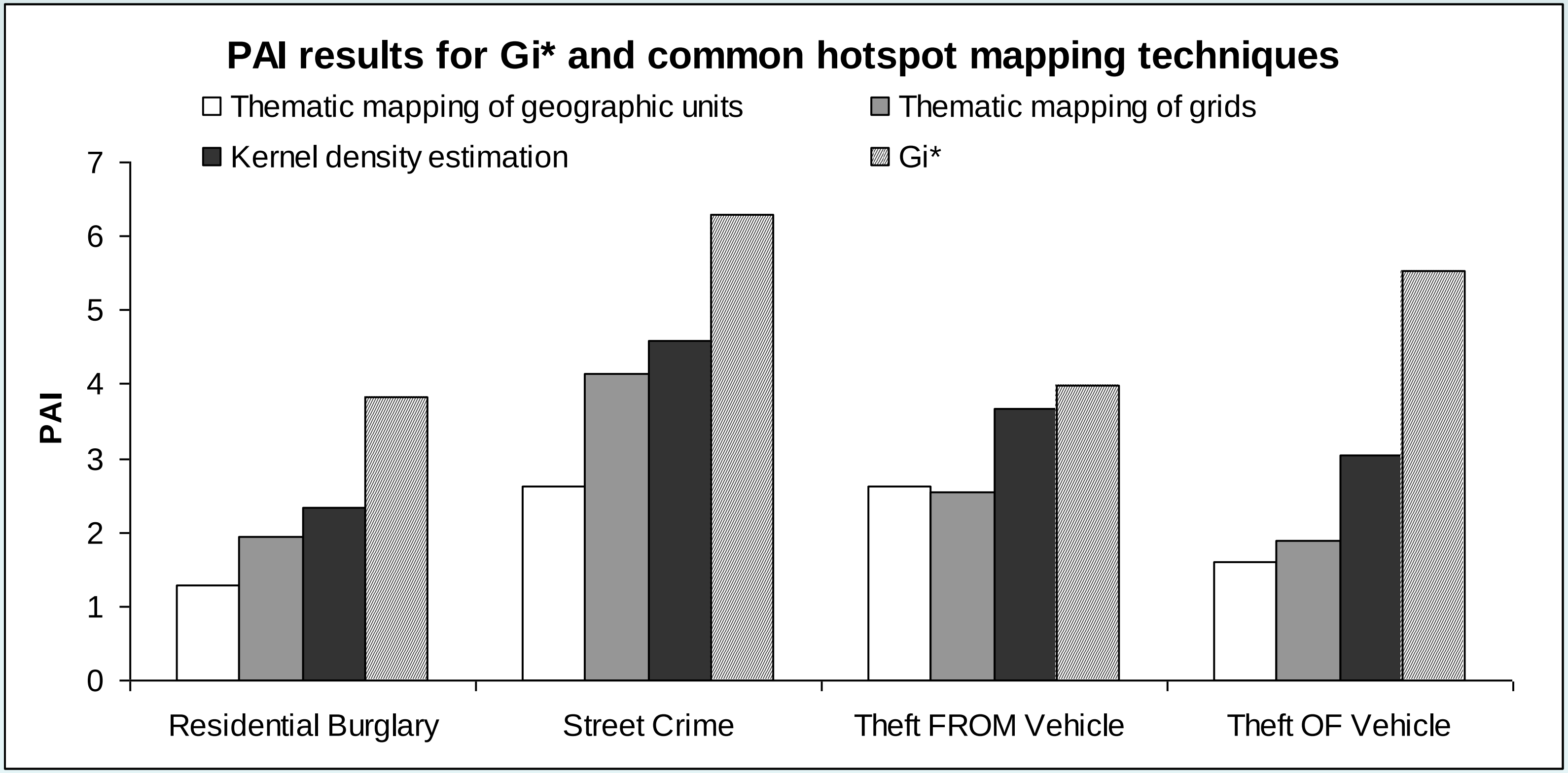




90% significant: Gi* z score > 1.645; 95% significant: Gi* z score > 1.960; 99% significant: Gi* z score > 2.576; 99.9% significant: Gi* z score > 3.291

Predictive accuracy of Gi* and common hotspot mapping techniques

- **Results from research** - higher Prediction Accuracy Index (PAI), better it is at predicting where crime will happen
  - Gi* gives best results (shown for 95% significance level)



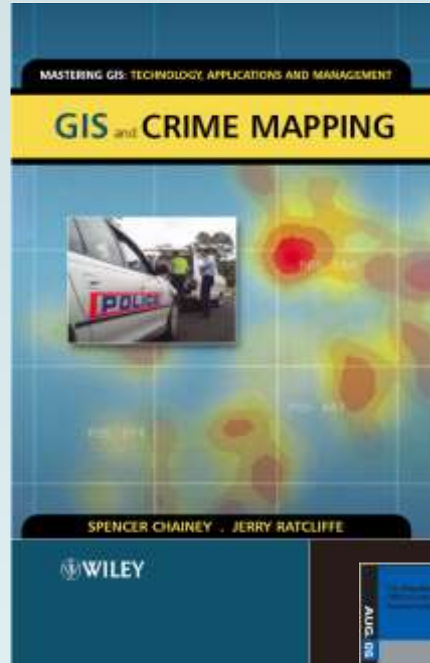
Summary: advantages of using Gi*

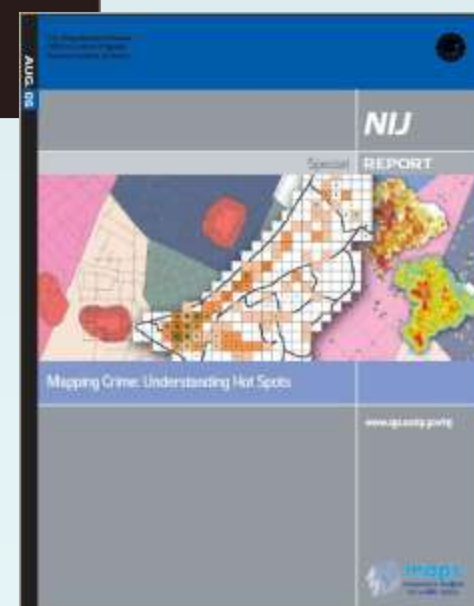
- Adds statistical significance to hotspot analysis
  - Which are the hotspots that are significant?
  - Where is there something really unusual going on?
- Better at predicting where crime will occur
  - In comparison to KDE and other common techniques
- Compensates for the over-smoothing created from KDE and *whims and fancies* of thematic threshold settings
- Negative features
  - Not as visually alluring as KDE
  - Not available in all the most popular GIS
    - But Rookcase (University of Ottawa)
- Does it replace KDE?
  - No, complements it

Thankyou

More information

- A couple of decent books!
- Research journal articles by Getis and/or Ord
- Rook's Case Help
- ESRI ArcGIS Help





Spencer Chainey
The Jill Dando Institute of Security and Crime Science
University College London
E: s.chainey@ucl.ac.uk
W: www.ucl.ac.uk/jdi