Advanced hotspot analysis: spatial significance mapping using Gi*
Overview

• Quick review of common hotspot mapping techniques
• The value of significance testing
• Advanced hotspot mapping
  – Using the Gi* statistic to identify patterns of spatial significance
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
Kernel density estimation

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

- Leeds
- Middlesbrough
- Oxford
- Birmingham
- Liverpool
Comparing KDE to other methods

• Results from research
  – Prediction Accuracy Index


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

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

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

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: smooths 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  

Which hotspots are statistically significant? i.e. is the spatial concentration of these crime incidents really unusual?
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
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. count of crime per 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:

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

**Parameters:**

- **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
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

**Input data for Excel**

- Coordinates for my grid cell centroids
- Value for each grid cell

![Input data table and grid example](image-url)
Gi* statistic

An example

Video clip on my YouTube channel showing use of Rooks Case to calculate Gi* statistic in Excel
http://youtu.be/s1cNp-YI3yo
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

- 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: $\geq 1.645$
  - 95% significant: $\geq 1.960$
  - 99% significant: $\geq 2.576$
  - 99.9% significant: $\geq 3.291$ (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
Operational steps

1. Create a grid in my GIS
2. Calculate a count of crime per grid cell
3. Export data in to Excel format (e.g. xls, csv)
   - 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)
Gi* statistic and ArcGIS

Operational steps

1. Create a grid
2. Calculate a count of crime per grid cell
3. Run Gi*: ‘Hotspot Analysis – Getis and Ord Gi*’
4. Display and interpret the results (using the Z score statistical significance threshold values)
Step 1: Create 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 (70m cells)

Video clip on my YouTube channel showing how you can use Hawths Tools in ArcGIS 9.3 or lower to create a user defined grid:
http://youtu.be/sNH42FnYlk0
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 clip on my YouTube channel showing how to generate a count of the number of crimes for each grid cell in ArcGIS: http://youtu.be/XVqswGaBAh8
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 to 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 clip on my YouTube channel showing how to enter settings into ArcGIS for running Gi*:
http://youtu.be/pA_FXNFj0cM
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: $\geq 1.645$
- 95% significant: $\geq 1.960$
- 99% significant: $\geq 2.576$
- 99.9% significant: $\geq 3.291$
Step 4: Displaying and interpreting the results

Video clip on my YouTube channel showing the ‘whims and fancies’ of legend thematic threshold selection for KDE, how to enter thematic thresholds for Gi* using Z score statistical significance values and display the results in ArcGIS:
http://youtu.be/ypRJhB19Tj0
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)
  - Recent study (Manchester), Gi* areas predicted 52% of burglary over next 3 months
Gi* negative features

- Not as visually alluring as KDE (but watch this space ...)
- Not available in all the most popular GIS
  - But Rookcase (University of Ottawa)
- Identifies too many areas as 99.9%
  - Currently exploring use of corrected Bonferonni approach rather than Gaussian for determining Z-score thresholds as originally proposed by Getis and Ord
- Identifies some cells that have low crime count
  - Need to remember it is considering a neighbourhood rather than just that cell in its calculation
Summary

- 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
- Some negative features, but none that are show stoppers
- 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
• My YouTube channel!
  www.youtube.com/user/SpencerPC

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
Follow me on Twitter @SpencerChaine