Numbers aren't nasty: a workbook of spatial concepts

David J. Unwin
Emeritus Professor in Geography
Birkbeck London
University of London

2010

Chapter 3 - Patterns of Point Objects

Acknowledgement

This Workbook collects together materials I have used over many years and in many different educational contexts. I am grateful to Dr. Nick Tate, University of Leicester for his suggestion that they could be gathered together under a heading of 'spatial literacy' and to the SPLINT initiative for the Fellowship that enabled me to do this. In preparation, I am grateful for additional discussion with Sarah Witham Bednarz, Chris Brunsdon, Jason Dykes, Nick Tate, and Harvey Miller. Students at Universities in Leicester, London, Hamilton (NZ), Christchurch, and Redlands together with those enrolled on my course for http://www.statistics.com have often unwittingly helped in the development of almost all the ideas and exercises that are presented. Some of the exercises aren't possible without use of software created by Luc Anselin, Ned Levine, the team that created 3Dfield, and Jo Wood. Alex Szumski 'road tested' the exercises and checked the manuscript for typos and the like. Any remaining errors are of course entirely my own responsibility.

I am grateful to Karl Grossner for permission to use Figure 1.1, to John Davis for permission to reference and use data (STRIAE and NOTREDAM) from his text *Statistics and Data Analysis in Geology* and to Wiley & Sons (New York) for permission to redraw and reuse my own figures 4.1, 4.3, 4.6, 5.4, 5.13, and 6.2 from O'Sullivan and Unwin (2003 & 2010) *Geographic Information Analysis*.

David Unwin
December 22nd 2010

Copyright and Intellectual Property

In the post World Wide Web, e-learning, world, attribution of Copyright to materials such as those in this Workbook isn't easy. There are ideas and possibly short sections of text that could be claimed by John Wiley & Sons, Birkbeck London, University of Leicester, University of Redlands (USA), Karen K Kemp (with whom I worked in Redlands) and myself. As sponsors of the project that has assembled them and provided their *raison d'etre* in a spatial literacy context, the SPLINT Centre for Excellence in Teaching and Learning might well also lodge a claim. Where there might be doubt, I have indicated this by use of an appropriate qualifying phrase such as 'source' or 'taken from'.

In this form and context these materials have been authored entirely by myself, so it is reasonable to claim them as © David J. Unwin (2010) and to assert the moral rights of authorship. They are made available under the understanding that they can be used in teaching and modified to suit local circumstances without and claim or charge provided that the source is recognized by the statement:

These materials have been developed from the Workbook 'Numbers aren't Nasty' assembled by David J. Unwin under the auspices of the (UK) Spatial Literacy in Teaching Initiative, (2010).

Any repurposing for commercial gain is subject to the usual 'all rights' © claim given above

Chapter 3: Patterns of Point Objects

3.1 Introduction

This chapter continues the examination and clarification of concepts relating to point objects, for which as argued in Chapter 1, appropriate, complex/ second order, concepts relate to words like 'distribution', 'dispersion', 'density', 'pattern' and 'scale' and, at higher level still, third order concepts relating to point process models, stationarity and isotropy/anisotropy. In this chapter we provide suggestions for exercises based mostly in the CRIMESTAT package that explore such second order concepts as dispersion, density and pattern in distributions of point located objects. The main exercise looks critically at familiar tests against the hypothesis of complete spatial randomness, otherwise perhaps known as 'no pattern'.

3.2 Exercise (8): Dotting the map

Aims and introduction

Dot maps show differences in the location and density of point located 'events'. Although numerous methods of point pattern analysis have been developed, it is only seldom that in practical studies we have suitable data for these analyses. The aim of this exercise is to show why this is so, and to force students critically to examine any dot/pin maps that they see.

Geometry, space and level

A set of located point objects when mapped in a metric space immediately presents complex/second order concepts referred to as distribution, dispersion, pattern, clustering, and density. This exercise uses visualization to address them.

Intended learning outcomes

After doing this exercise, students will

- Be able to recognize a simple dot or pin map;
- Understand the characteristics of a true point pattern suitable for statistical analysis as distinct from a simple dot density map;
- Appreciate the importance of the 'art and science' of cartography in determining the look of a map.

Resources needed

WWW browser with access to Google™.

Suggested student briefing

1. In order to understand some of the issues in 'dotting' a dot map watch the tutorial from Sara Fabrikant at http://www.csiss.org/streaming_video/csiss/fabrikant_dot_maps.htm. Pay particular attention to the distinction she makes between a 'one to one' and a 'many to one' mapping and to the importance of exact locations vs. data that are aggregated over areas.

It should be clear that for a 'one to one' mapping the basic data have to be suitable, with perhaps five basic conditions being necessary. These are:

- The pattern should be mapped/projected on the plane such that distance between the points are preserved;
- The study area should be determined objectively, with boundaries that are not arbitrary. In practice this is very hard to achieve;
- The pattern should be an enumeration of all the defined point objects in the study area;
- There should be a one-to-one correspondence between dots on the map and objects on the ground, one dot, one object;
- Locations should be proper, not for example arbitrary points within areas chosen to be in some sense representative.

2. Now go to *Google*™ (or similar search engine) to find a proper dot map that meets all five conditions. If you search for 'dot map', ask yourself several questions:

    - Is there a one to one between the dots and distinct 'events' such as the location of a crime, some facility or whatever? Often dotting is used as a cartographic symbol with a 'many to one' relationship to the phenomenon being mapped, for example '1 dot represents 2000 acres'. These are dot density maps of the 'choropleth' variety;

    - Are the locations 'proper'? Is each dot located at the correct place where the 'event' occurred or is to be found? Often dots are placed at the centroids of areas or in a stipple across an area, so the locations have no special meaning and can't be used in point pattern analysis;

    - If the two conditions above are met, is it a sample or a complete enumeration or census?

3. If searching using 'dot map' doesn't reveal anything, try instead a search using the post-GIS term for the same type of map which seems to be 'pin map'.

Comment/answers

Almost all of the images returned using a 'dot map' search will actually be dot density maps that do not meet the five conditions. Searches using 'pin maps' seem to do better. It is probable that students will find genuine examples in some crime maps and/or maps in epidemiology. It is worth emphasizing exactly

what dot density maps show, which is area aggregated data and thus make the point that just because dots are used in the representation it does not mean that the data themselves relate to point objects. The simple conclusion is that we seldom have 'pure' point data at precise locations on the plane of the sort required by almost all the standard methods of point pattern analysis. This is a very important lesson!

Personally, I've never been sure that they are all that useful unless they show *rates of occurrence*. In crime pattern analysis, the dots might useful because they tell the police where to deploy their resources, but in epidemiology and criminology surely it is the rate that matters, relative to some underlying factors?

In addition, a majority of the maps returned will be cartographically awful, hardly worth drawing in the first place.

Suggestions for modification

Discussion of the conditions for point pattern analysis to be sensible can be extended further. For example, a case can be made that some methods of analysis do allow use of sampled data (most obviously using randomly distributed quadrats and/sampling nearest neighbour distances in ecology)

3.3 Exercise (9): Drawing your own pin map

Aims and introduction

There are two exercises here, the first of which simply uses *Google*™ to produce maps, whilst the second uses Microsoft *Excel* ™ to produce maps of three supplied point data sets. The overall aim is to introduce the idea of patterns in point data that are revealed by the maps.

Geometry, space and level

As in Exercise (8), a set of located point objects when mapped in a metric space immediately presents complex/second order concepts referred to as distribution, dispersion, pattern, clustering, and density. This exercise uses visualization to address them.

Intended learning outcomes

After doing this exercise, students will be able to:

- Produce dot/pin maps of any facilities recorded in the Google™ Maps data base and/or
- Map any point located data provide as (x, y) co-ordinate pairs in a Cartesian system;
- Criticize the cartography they employ;
- Suggest possible descriptions for the point patterns revealed.

Resources needed

Approach (a) requires a web browser and (b) needs Microsoft *Excel*™ or, should you prefer it, whatever standard GIS you use. Approach (b) also requires access to the three supplied data sets called BOOK, BANK and SNOW.

Suggested student briefing

a) The lazy way

1. In your web browser got to http://www.google.com and select the Map option;
2. If you live in, or know of, any reasonably large city, enter text in the search box as 'coffee shops in xyz', where xyz is the name of your city. If coffee shops don't appeal then try some other suitable facility;

3. The result will be returned as a pin map of the type discussed in Exercise (8);
4. In 3.1 above, five criteria were suggested to use in testing whether or not such a map is of a genuine point pattern that might be analyzed using methods to be introduced in the next Section. Evaluate your result in the light of these five criteria;
5. Finally, in your own words how would you describe the patterns revealed? Are the point locations 'clustered', 'random' or 'regular'?

b) Doing it yourself using Microsoft *Excel* ™

Three text files of (*x, y*) co-ordinates of some point 'events' are provided:

Book: These are the 12 sample data taken from Table 5.2 page 131 of O'Sullivan and Unwin (2010)

Bank: This is a famous data set that has been analyzed many times, notably by the statistician Brian Ripley. These data were taken from the website associated with the text by Davis (2002). It gives 47 (x, y) pairs giving the location (in a projected Euclidean co-ordinate system), of dark magnetite crystals in a polished cross-section of a rock called anorthosite. The interest is in whether or not this distribution is random within the section. Co-ordinates are on a 100x100 grid, with its origin at the bottom left, but no units are given (assume cm?) and the rock in question forms part of the doorway to one of the banks in the city of Cambridge, England. The origin of these data is simply to remind you that not all 'spatial analysis', even in a GIS, must be 'geo-spatial'.

Snow: This is probably the most famous point data set ever to be analyzed. It consists of the locations of 578 deaths from cholera recorded by Dr. John Snow in the Soho area of London during an outbreak of cholera in 1854. Snow mapped these data as a dot map and was able to show that they clustered around a single water pump (no piped water in those days!) in Broad (now Broadwick) Street. Acting on his advice, the authorities removed the handle from the pump and the epidemic ended soon after, although it may well have already been past its peak. The events are celebrated by a facsimile of the pump and in the naming of a nearby pub the 'John Snow'. All epidemiologists and all spatial analysts should at some time make a pilgrimage to the street and have a drink in the pub that now bears John Snow's name.

Snow's work is widely regarded as the birth of scientific epidemiology, and his demonstration that the vector for cholera was water-borne led to massive investment in UK during the second half of the nineteenth century to provide safe public water supplies. Of course, the story isn't as simple as it is sometimes

suggested. For a recent scientific account, see Brody, H. *et al*., (2000). For a 'popular' account that by Steven Johnson (2006) is highly recommended.

These data were digitized at the request of Professor Waldo Tobler (UCSB) by Rusty Dodson of the US National Center for Geographic Information Analysis from a reprint of Snow's book *On Cholera* (Oxford University Press, London).

Although the origin is at (0, 0) these data have arbitrary co-ordinates that range on X from 8.280715 to 17.938930 and on Y from 6.090047 to 16.972760. In the real world, one full unit (e.g. 1.0000) represents about 54m on the ground, so the minimum enclosing rectangle has an area of about 0.3km^2. Note that the coordinate system provides for a lot of unused 'white space' around these points, which you might judge should not be included in any analysis.

1. For BOOK, BANK and SNOW produce simple dot maps. This can be done in Microsoft *Excel* ™, provided care is taken to scale the X, Y axes appropriately as follows. Go FILE>OPEN>FIND and navigate to where you have saved BANK.TXT. Then chose 'delimited' and set this to 'space' with the data type set as 'general'. This should incorporate the two columns into Microsoft *Excel™* ;
2. In the chart wizard, it's a simple matter to chose the (X,Y) scatter plot. What I seem unable to do is to stretch the axes on these plot such that they have the same scale (you can set the range), and have usually done this before printing by clicking on the two horizontal axes of the display and pulling them out until I get the desired equal scale;
3. If you have access to *ArcGIS*™ or similar, you should be able easily to draw 'proper' maps of these three distributions;
4. In Exercise (8), five criteria were suggested to use in testing whether or not such a map is of a genuine point pattern that might be analyzed using methods to be introduced in the next section. Evaluate your result in the light of these five criteria;
5. In your own words how would you describe the patterns revealed? Are the point locations 'clustered', 'random' or 'regular'? How do they differ?
6. How do you think the choices made for the 'frame' will affect the descriptions you have given?

Comment/answers

Book

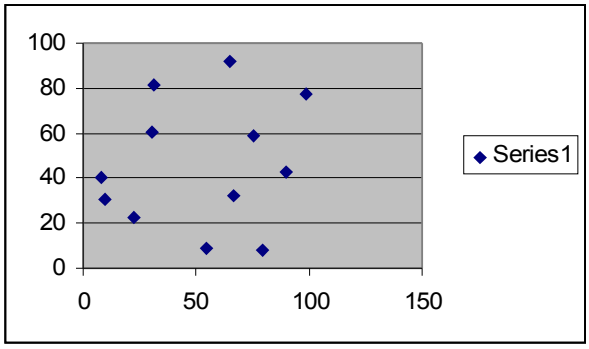

Figure 3.1 Dot map for Book

Figure 3.1 shows the distribution of events in BOOK. I have used Microsoft *Excel™*, copied into *WORD*. The box could be re-sized such that the scales on X and Y are the same, noting that the range of value on X is greater. Visually I would say that the pattern looks fairly random, but with only 12 events how can one tell?

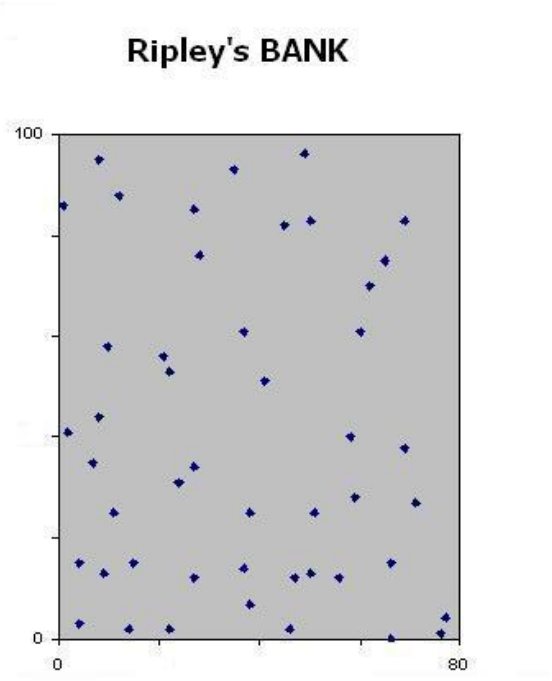

Figure 3.2 Dot map of BANK

Figure 3.2 shows the distribution of events in BANK. This is also from Microsoft *Excel*™, but scaled and transferred into *PAINT* where it has been edited a bit more. It ought to have a scale, as we are now well on the way towards a proper 'map'. Visually I'd describe it as 'more regular than random'.

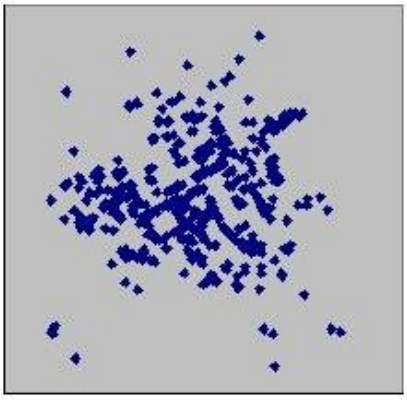

Figure 3.3 Dot map of SNOW

Figure 3.3 shows the distribution of events in SNOW. The usual interpretation is that the cases 'cluster', that is they are more aggregated than random, with the clustering around a specific point, the Broad Street water pump. Note that in this hypothesis, we only have one cluster, so what is the value of the standard CSR model in this case?

One obvious point that the SNOW data show is the dependence of what we see on the 'edges of space' that we chose to use. We can make this look even more clustered by simply extending the frame. Zooming in to a subset of these same data might well make them look random or even dispersed. In the case of BANK, zooming out would gradually make hem look more aggregated than random. In other words, the choice of frame is critical in the visualizations and what should be done is basically to proceed carefully unless there is a 'natural' frame.

Suggestions for modification

An obvious extension is to ask students to run a kernel density estimate over these data, with three possible reasons for interest:

a) As a means of locating 'hotspots' in the patterns;

b) To initiate a discussion of band width, kernel function and even the appropriateness of the underlying 'geography' (For example, in the Snow case should we use street walking distances and not straight lines?);

c) To show the value of a transformation from a pattern of discrete objects (the events) into a spatially continuous field of density estimates.

3.4 Exercise (10): Proportionate symbol maps

Aims and introduction

Proportional symbol maps show differences in the location and magnitude of point located 'events' and are appropriate for visualizing what statisticians call a 'marked' point pattern. What we now have is a pattern of discrete point objects/events, but in each case we have an additional 'weight' attached to each event. Almost all the basic methods of point pattern analysis can be modified by use of such weights.

Geometry, space and level

As Exercise (8) and (9), a set of located point objects when mapped in a metric space immediately presents complex/second order concepts referred to as distribution, dispersion, pattern, clustering, and density, but in this case we have two sources of variation in geographic space and but with a primitive/first order notion of magnitude added.

Intended learning outcomes

After doing this exercise, students will

- Be able to recognize a true proportionate symbol map;
- Understand the difficulty of simultaneously associating both variation in magnitude and variation in geographic space;
- Be able to distinguish such maps and the data on which they are based from maps that use similar symbolism but to display area aggregated data;
- Appreciate the importance of the 'art and science' of cartography in determining the look of a map.

Resources needed

WWW browser with access to *Google*™.

Suggested student briefing

1. Go to *Google*™ (or similar search engine) to find a proper proportionate symbol map that meets all the five conditions noted in Exercise (8). Ask yourself:

2. Is there a one to one between the dots and distinct 'events' such as the location of a crime, some facility or whatever?

3. What is the numerical variable that is attached to each event in the pattern?

4. Are the locations 'proper'? Is each dot located at the correct place where the 'event' occurred or is to be found?

5. If the conditions above are met, is it a sample or a complete enumeration or census?

6. Is the way the symbol used is related to the magnitude of the variable being displayed appropriate?

7. Can you make 'sense' of the distribution?

Comment/answers

The results are likely to be much more satisfactory, cartographically speaking, than for dot/pin maps, but there is a real difference between maps in which the symbol refers to an exact spatial location (such as, for example, size-graduated circles to show the output from a series of point located factories) and those that refer to data that are an aggregate for a specified area and are usually located at some central point within the area (such as a population map of the Counties in a State). In fact, examples of the former will be hard to find. Almost always the maps found will actually have symbols (circles are favourite, but beware some of the bizarre symbols that were found) used as a form of area symbolism and their locations were at some arbitrary point (usually the centroids) within the areas to which the aggregate data refer. Exercise (22) on choropleth mapping makes some further points about this sort of data, especially the folly of mapping absolute totals when using area aggregated data. It is worth using the results to point out the problem of isolating effects related to the geography of the locations themselves at the same time as their magnitudes.

Suggestions for modification

The main interest in this exercise is likely to be the weird and wonderful shapes used by some web cartographers to visualize the located quantities. It is well-known that use of even the simple circle with its area proportional to the value of the located datum can mislead. Human beings simply do not 'see' circle area in this way. The classic study and the suggested 'law' that corrects for it is by Flannery (1971).

Students may well also find maps that have as their symbols graduated pie charts showing the proportions of some constituent of the total. These can

display an enormous amount of data, but whether these visualizations are effective is moot, and might form the basis of an in-class discussion about the balance between map clarity and data volume/character.

3.5 Exercise (11): Centrography

Aims and introduction

The best way to learn something about point pattern analysis is to do it. This exercise uses public domain software and three supplied data sets to go through typical analyses, but at the same time highlighting the practical implications of some of the difficulties. It will probably take students around 4-5 hours to complete all the tasks.

- To demonstrate computation of simple basic point pattern measures with different types of patterned data;
- To illustrate some of the problems and issues that might emerge in such use, notably the influence of the area used and the need to understand edge effects.

Geometry, space and level

A set of located point objects when mapped in a metric space immediately presents complex/second order concepts referred to as distribution, dispersion, pattern, clustering, and density. This exercise uses simple arithmetic to address them.

Intended learning outcomes

After doing this exercise, students will

- Realize that the boundaries of the space we choose greatly affect these types of measure;
- Understand that centrography does not explicitly capture the notion of pattern in a distribution of point events;
- Critically assess the situations in which these measures might be used to compare different distributions in the same area and/or the change in a distribution over a sequence of time slices.

Resources needed

CrimeStat III, produced by Ned Levine Associates for use by police forces interested in the spatial distribution of crime, is available as a free download and can be used to analyze almost any point pattern, not just the distribution of crimes. Although it will compute many of the measures we have discussed, it doesn't compute a quadrat analysis or some of the more esoteric measures, such as G(*d*) and F(*d*), nor does it have serious production graphical capabilities.

Likewise, if you have an *Apple*™ machine you'll need to run it in Windows emulation mode. The graphic deficiencies can be overcome, either by use of *ArcGIS*™ 'shape' (.shp) files as export and import, or by using ASCII text files (.txt) files imported into Microsoft *Excel*™.

Suggested student briefing

1. Visit the website at http://www.icpsr.umich.edu/CRIMESTAT/ and download the *Crimestat III* software. It is best to download all the associated files at the same time. Follow the instructions to install the program;

   Almost all the problems you might have when using *Crimestat III* will be associated with errors made at the data description and entry stages, so it pays to take care.

   In Data:

   - Set file characteristics at ASCII
   - Select file, say BOOK, and navigate to it
   - Check the SPACE SEPARATOR and ensure that there are 0 header lines and 2 columns.

   On the Data Set Up screen, take care to ensure that you:

   - Set X as column 1
   - Set Y as column 2
   - The remaining fields should be either <none> or <blank>
   - Set the type of co-ordinate system to 'Projected', units to 'm' (they are actually arbitrary).

   Several of the *Crimestat* routines require either a 'reference' file, a 'measurement' file, or both. Experience suggests that although you can get some results without creating and saving these, it's often better to create do this right at the start. Note, too, that saving the parameters, available under the 'options' menu, saves both reference and measurement files. If you do this, you need to think hard about what area to input and about the density of estimates you want in, for example, the K(*d*) function and/or kernel density estimation routines, so you may need to revisit this step!

   You are now ready to analyze some data!

2. Using BOOK, and to gain confidence, show that the mean center of the 12 events is at (52.575, 46.175). In addition, record the average density, and standard distance;

3. Use BANK and then SNOW to do the same things;

4. Do these numbers tell you very much? Do they help differentiate the patterns?

5. On your plot of the SNOW data, locate the mean center and confirm that it does indicate something useful.

Comment/answers

Assuming projected data with co-ordinates in m Table 3.1 shows the results, but the exact numerical values aren't important:

File	n	Mean X	Mean Y	Density (m assumed)	Standard distance
BOOK	12	52.6	46.2	0.001571	43.98
BANK	47	36.6	40.1	0.006442	39.24
SNOW	578	13.0	4.7	5.500000	2.56

Table 3.1 Centrographic measures for BOOK, BANK and SNOW

Obviously one needs to convert the apparent density units into those appropriate for the particular data set. These centrographic measures tell us very little, at least in these applications. Answers will all lie close to the centre not simply of the data co-ordinates but of the frame chosen, around (50, 50). They are useful to compare patterns of 'events' when these events are of different kinds in the same geographic area, for example the locations of stores of differing types across a city area. They are also sometimes useful in tracking the evolution of a pattern over time. One minor use I can see is for the so-called 'standard deviational ellipse' (not circle) which can indicate a pattern of events that has some directional bias. Maybe, just maybe, the SNOW analysis shows a third useful thing they can do?

Suggestions for modification

1. Read the PDF files from the manual Chapter 1 and then Chapter 2, sections I (*Data Setup*), II (Spatial Description) and III (*Spatial Modeling*);
2. There is no need to go further than Chapter 2, but you might also read appropriate bits of Chapter 4.1 – 4.17 (*Centrographic Statistics*,

Chapter 5.1, 5.7, and 5.40 and Chapter 8.1 – 8.14 on kernel density estimation);

3. You might also like to follow in its entirety the example given at the end of Chapter 3 (page 3.32 et seq.) using a supplied .dbf file and the 'general sample data' found in a ZIP file. If you do this take care to name the columns correctly.

3.6 Exercise (12): Nearest neighbor statistics

Aims and introduction

In elementary texts such as Unwin (1981), the most often used measure of spatial pattern is the classic nearest neighbour statistic devised originally by Clarke and Evans (1954). The logic behind this statistic is described by O'Sullivan and Unwin (2010, pages 130-132 and 143-145). Basically the so-called *R-index* is the ratio of the mean of the observed distances from each event in the pattern to its nearest neighbour to the expected mean distance under the hypothesis that the pattern is random. A statistical significance test can be developed, since both the expected mean distance and its variance are readily obtained from simple mathematics. The approach has a number of possible traps for the unwary, not least of which are the choice of the 'frame' in which the events are considered to be present and the possible impact of unwelcome effects at the edges of the distribution when the number of events is low. This exercise uses *Crimestat III* and the same data as in Exercise (11) to illustrate these issues.

Geometry, space and level

A set of located point objects when mapped in a metric space immediately presents complex/second order concepts referred to as distribution, dispersion, pattern, clustering, and density. This exercise introduces the analytical/third order concept of a spatial process.

Intended learning outcomes

After doing this exercise, students will:

- Be able to conduct a nearest neighbour analysis using *Crimestat III*;
- Understand that, if the R-index is the ratio of the observed to expected mean distances to nearest neighbour, the 'expected is relative to some hypothesis most obviously that of complete spatial randomness;
- Be able to assess the statistical significance of the departure of the computed *R*-index from 1.0 using Student's *t*;
- Discover how importance the choice of frame is to the results obtained, preferably by showing how this is taken up into the calculations by way of the study region area used to find the expected mean distance under the null hypothesis that of compete spatial randomness;
- Note that a general test such as this might not be what is wanted in a specific case study such as John Snow's problem;
- Finally, understand that scale effects can be addressed using the same approach to examine distance to second, third, etc, nearest neighbours.

Resources needed

Crimestat III together with the point pattern data sets used in Exercise (11).

Suggested student briefing

For these exercises, unless specifically requested, DO NOT set the area in the measurement parameters section of the file set up.

1. Using BOOK use *>spatial description>distance analysis I>nearest neighbour analysis* to confirm that the mean distance to nearest neighbour for these 12 points is 21.62 as given on page ;

2. Using BANK confirm that the Clark and Evans R with no edge correction and using the *Crimestat III* default way of finding the area from the so-called minimum enclosing rectangle as (77-1)*(96-0) = 7296 cm^2 is:

$$R = \frac{d_{obs}}{d_{\exp}} = \frac{7.81}{6.23} = 1.2539$$

   With a t-value of 3.33, this is significantly different from random at p=0.001. Given that we have observed mean distance to nearest neighbor greater than expected, we infer that the distribution of points is 'more dispersed than random'.

3. However, there are two well-documented issues with this statistic, which make it not very 'GISable' and the consequences of both can be illustrated using these same data. First, there is a critical dependence on the area used in the calculation of the expected mean distance. In computing the index, *Crimestat III* defaults to use the area given by the range of the co-ordinates on the X and Y-axes which gives an area of 7296 cm^2. To illustrate this, repeat the analysis, but in this case in 'measurement parameters' set the area to be that of the entire frame, which is 100 units x 100 units = 10,000 units-squared. Confirm that we now get:

$$R = \frac{d_{obs}}{d_{\exp}} = \frac{7.81}{7.29} = 1.0710$$

   With a t-value of 0.9317, this isn't statistically different from random. Note that there is no 'natural' boundary for these data: in fact I suspect

that the size was determined by Professor Ripley's unwillingness to measure any more crystals!

What does this tell you about the idea of 'randomness' in point patterns?

4. Second, there is also an effect at the edges of the distribution. For small numbers of events the effects can be quite dramatic, shifting the null value from 1 upwards into the 'more dispersed than random' range. This arises because points near the edges of the distribution are forced to find neighbors within the space, when in reality it is probable that their true nearest neighbors would be at some shorter distance outside the frame. This clearly biases the mean upwards. I know of four ways of handling this issue. The first is that used by *Crimestat III* in which such points are handled by taking the distance from a border event to the frame edge (assuming either a rectangular or circular frame) if this distance is less than any measured distance to the nearest event within the frame. Section 5.11 of the *Crimestat III* manual explains this in more detail. To see what happens, re-set the 'measurement parameters' area to zero and then run the program again, but this time ticking the 'rectangle' edge correction box. This will correct things as:

$$R = \frac{d_{obs}}{d_{\exp}} = \frac{6.40}{6.23} = 1.0274$$

The second places a 'guard area' around the frame and proceeds as usual but allows points near the edge of the frame to find neighbors within the guard region. Of course, the nearest neighbour distances of these guard points are not themselves included in the analysis. The third approach uses a series of edge corrections, established by a combination of mathematics and experiment. Possibly the most elegant approach is the fourth, which wraps round both edges of the frame, to meet their opposite side and then proceeds in the usual way. Notice that we now have three possible values for the nearest neighbor statistic for these same data, dependent on what we assume about the area of the region and how any edge effects are handled by the software. If nothing else this should alert you to the need to take extreme care when using this approach, the more so if you don't actually know precisely how any GIS you use does its calculations.

5. Load SNOW and repeat the above analyses. It is instructive to experiment with different values for the region area. Does the nearest neighbor statistic help in any meaningful way in testing Snow's hypothesis?

6. A third problem with the distance to nearest neighbour is that, by taking only the nearest neighbour distances it only indicates the nature of any global patterning at this 'scale'. This deficiency can be circumvented by repeating the analysis (the mathematics is essentially the same) for successive 'orders' of neighbors and so examining the patterning at successively longer distance scales. *Crimestat III* lets you do this on the basic *distance analysis 1* screen by asking for as many neighbors as desired to be considered. Do this for both BANK and SNOW but when you have the results, use the GRAPH option to get simple plots of the change in *R* with order of neighbour.

Comment/answers

See test above. The sensitivity of the test to the definition of the study area and with small *n* to edge effects comes as a surprise. It is useful to point out that Snow didn't really need to do any statistical analysis to get his point across. Steven Johnson's (2006) book should be referenced for the complete story.

Suggestions for modification

In Snow's second map, the usual display of a point pattern that we have seen so far was supplemented by a line enclosing all the houses that from his local knowledge Snow knew to be closer to the infected Broad Street pump than they were to any other. In essence this was what future spatial analysts would call a Voronoi diagram or Thiessen network and it showed with astonishing clarity that not only did the cases cluster, they clustered around the suspect pump with only a few exceptions of deaths to people for whom the pump was not the nearest source of water. The pattern of streets in 1854 wasn't the same as it now is, but a rough approximation to the Snow's border can be obtained by computing and displaying the Voronoi/Thiessen network on top of a dot/pin map of the deaths.

If students have access to a GIS capable of computing the Voronoi tessellation a display of these point data with the Voronoi diagram for all 13 pumps in the area is a very convincing demonstration of the power of simple, almost geometric 'spatial analysis'. *3DField*, used in Exercise (30) will compute this as well.

3.7 Exercise (13): Ripley's K statistic

Aims and introduction

In view of the problems with the single number approach, it is hardly surprising that spatial statisticians have tried to characterize pattern using distance functions such as those discussed in the text. Of these, Ripley's K(*d*) is the most satisfactory. This exercise is a simple introduction to the approach. The theory behind the approach is introduced in O'Sullivan and Unwin (2010, pages 135-137 and 146-148.

Geometry, space and level

A set of located point objects when mapped in a metric space immediately present complex/second order concepts referred to as distribution, dispersion, pattern, clustering, and density. This exercise uses both the analytical/third order concept of a spatial process and visualization to address them.

Resources needed

Crimestat III together with the point pattern data sets used in Exercise (11)

Intended learning outcomes

After doing this exercise, students will:

- Be able to conduct a point pattern analysis using Ripley's *K*(*d*) function approach in Crimestat *III*;
- Understand the process by which a mean value of *K*(*d*) at some distance *d* is estimated;
- Be able to interpret a graph of computed values of *K*(*d*) against distance, *d*;
- Understand how theoretical values for a random distribution can be derived and used to convert *K*(*d*) into the *L*(*d*) function in which the theoretical expectation for a random pattern is zero at all distances;
- Understand that this approach enables investigators to examine the scale at which a pattern of point events can be said to 'cluster';
- Be able to assess the statistical significance of the departure of the computed K*(d)* using the randomization approach.

Suggested student briefing

Much modern work in point pattern analysis uses the K(*d*) function approach developed by Ripley (1976) which is based on all the distances between events in the pattern. Computation of Ripley's K(*d*) is easy to explain but very tedious to do except by computer. All we do is to place circles, of each radius *d* , centered on each of the events in the pattern and find the number of events that fall into that circle. Doing this with the same radius d centered on each and every event allows calculation of a mean number for this distance. All we then do is to repeat this for a series of distances. Each mean count is divided by the overall study area event density to give *K*(*d*). Formally this is:

$$K(d) = \frac{\sum_{i=1}^{n} \#[S \in C(\mathbf{s}_i, d)]}{n\lambda} = \frac{a}{n} \cdot \frac{1}{n} \sum_{i=1}^{n} \#[S \in C(\mathbf{s}_i, d)]$$

Remember that $C(\mathbf{s}_i, d)$ is a circle of radius *d* centered at $\mathbf{s}_i$ and the operation specified by the numerator is the 'number of' (#) 'events', *S*, 'included in' ($\in$) that circle. Because *all* distances between events are used, over a range of distances, this function is much more informative about the patterning than any single number such as the *R*-index ever could be, but what values would we expect if the pattern is random? In fact this is easy to calculate, at least if there are no problems with edge effects and the definition of the area of interest. Since πd^2 is the area of each circle, and λ is the mean density of events per unit area, the expected value of *K*(*d*) is simply

$$E(K(d)) = \frac{\lambda \pi d^2}{\lambda} = \pi d^2$$

Because the expected function depends distance *squared*, both the expected and observed *K*(*d*), this function can become very large as *d* increases and it is difficult to see small differences between expected and observed values when they are plotted on appropriately scaled axes. The usual way round this problem is to convert the expected value of *K*(*d*) to zero, by dividing it by π, taking the square root, and then subtracting *d*. as

$$L(d) = \sqrt{\frac{K(d)}{\pi}} - d$$

The result is another function of distance, this time called the *L* function. If the pattern is random performing the same operations on the observed values of *K*(*d*), we should get values near zero. Where *L*(*d*) is above zero, there are more events at the corresponding spacing than would be expected under IRP/CSR; where it is below zero, there are fewer events than expected.

Crimestat III will provide an estimate of K(*d*), but to do so it has to have a defined reference file of a grid of locations. It also produces all the data needed to plot L(*d*), together with a simulate envelope around this for use in evaluating the significance of departures from the random expectation for L(*d*), which is zero.

1. Start *Crimestat III*
2. Using the *distance analysis 1* screen, set up and compute the *K*(*d*) function for both BANK and SNOW. The output is the L(*d*) (called *t*' on screen)

One problem with these functions is that at large *d* edge effects enter into consideration, where a substantial proportion of each circle is outside the study area. In these cases by definition there are no events outside the study region, so the number of events in the circles is lower than would be expected based on an assumption of uniform density. This is a problem in almost all work in spatial statistical analysis: we almost have to either to break some assumption made in the derivation of the theoretical values or to attempt some corrections that take them into account. Nowadays, plentiful computer power enables us to use a simulation approach to this problem. No matter which statistic we are interested in, the procedure is always very simple: use a computer to generate a large number of patterns according to some hypothesis we have about the process. In this case we'd simply use the computer's random number generator to give randomly located point 'events'. Next, for each pattern we measure the statistic to give an expected distribution of values against which the observed values of the same statistic can be compared. This approach lacks mathematical elegance, but it enables us to allow for things like edge effects, simply by using the same study region in the simulations as in our observed data. Such a simulation approach is known as a *Monte Carlo procedure*, and is widely used in modern statistics, but it is computationally very intensive especially when the number of events in the observed and hence also the simulated patterns is large. There is also controversy about how many simulated patterns should be used. Some purists recommend use of a very large number, say 999, whereas those willing to take a bigger risk in their assessment might only use 99, but it really rather depends on the extent to which the statistical; assessment is important.

3. For both BANK and SNOW use *Crimestat III* to compute and plot the *L*(*d*) function. In doing this, use the simulation routine with, say 99 runs to get and plot an estimate of a confidence envelope that can be used to assess the significance of the observed values. Note that use of the GRAPH button will plot the so-called simulation envelope and that it is then easy to see at what distances the observed *L*(*d*) is outside and thus

indicative of a distance scale at which the pattern is more/less regular than expected;

4. In doing this note that for SNOW with *n* = 578 the simulation will take a perceptible length of time even on a fairly quick machine!
5. Describe how the two patterns differ and the extent to which these results confirm your previous analyses with the same patterns.

Comment/answers

Figure 3.4 shows the results for BANK (left) and SNOW (right) using *Crimestat III* with no edge corrections and with simulation envelopes based on 99 runs.

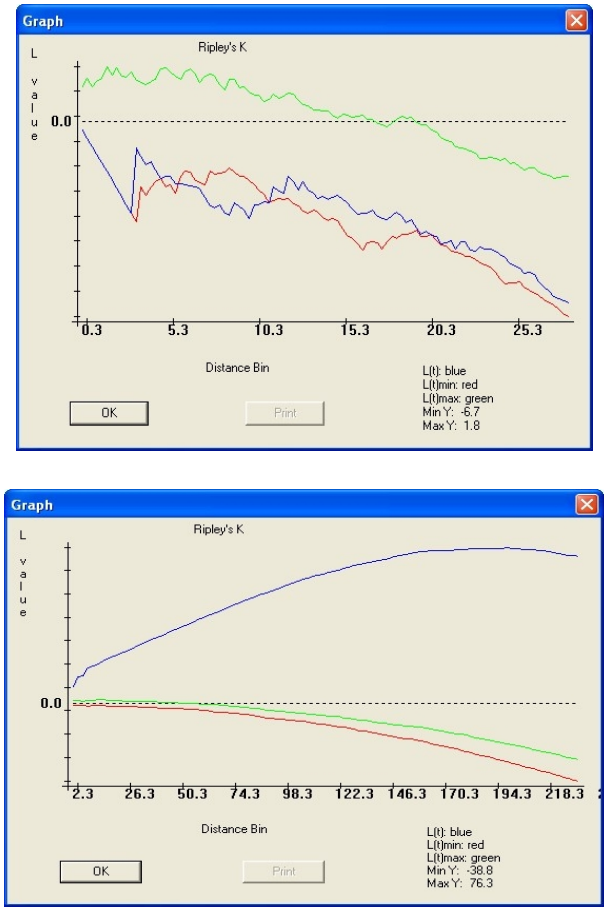

Figure 3.4 Results for Ripley's K(d) function for BANK and SNOW

On this display the observed L function is in blue and the green and red lines show the extreme values above and below in the simulations. It can be seen that the observed function for BANK is often within the simulation envelope but is always less than zero, indicating that at all distance scales there a fewer events at the specified distance than expected. This should confirm the idea that these data are more regular than random. For SNOW at every distance scale the L

function is well above zero, indicative of a pattern of events that is more aggregated (clustered) than random.

Suggestions for modification

A variation on Ripley's *K*(*d*) function called the *O-ring statistic* (Wiegand and Moloney, 2004) or the neighborhood density function (Perry *et al*., 2006), has been used. This is easily computed by noting that the original *K*(d) function is cumulative with the proportion of events from 0.0 to 1.0 at each circle radius plotted as a function of the radius, *d.* These more recent functions plot the actual proportion in a series of annuli centered on each event.

3.8 Further Exercises: Other things you can do with *Crimestat III*

This doesn't exhaust the options for work on point patterns based on *Crimestat III*. There are perhaps three additional options that you might like to explore using these same data at some future time:

1. Computing a kernel density estimation and then exporting the results for visualization into a GIS or other mapping program such as *3Dfield* (see Chapter 6);
2. So-called 'hot spot' analysis, which uses a modification of standard cluster analysis to find areas of above average local spatial point density, or 'hot spots'. In both criminology and epidemiology such concentrations have obvious practical significance;
3. Computing and exporting selected distance matrices for analysis in other software.

There is a useful 'how to do it' guide by Luc Anselin *An Introduction to Point Pattern Analysis using CrimeStat* that shows how to export from that package in *ArcView*™ or *ArcGIS*™ available at http://geodacenter.asu.edu/system/files/points.pdf.

3.9 References

Brody, H. *et al.*, (2000) Map-making and myth-making in Broad Street: the London cholera epidemic 1854. *The Lancet*, 356 (9223):64-68.

Davis, J.C. (2002) *Statistics and Data Analysis in Geology*, (3rd Edition, John Wiley & Sons)

Flannery, J.L. (1971) The relative effectiveness of some common graduated point symbols in the presentation of quantitative data' *Canadian Geographer*, 8: 96-109. See also Unwin, D.J. (1981) *Introductory Spatial Analysis* (London: Methuen), pages 33-36

Johnson, S. (2006) *The Ghost Map* (London: Penguin)

Levine, N. and Associates (2007) *Crimestat III: a spatial statistics program for the analysis of crime locations*, at http://www.icpsr.umich.edu/CRIMESTAT/

O'Sullivan, D. and D. Unwin (2010) *Geographic Information Analysis* (2nd Edition, Hoboken, NY: John Wiley and Sons)

Perry, G.L.W., Miller, B.P. and N.J. Enright (2006) A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecology* 187: 59-82

Ripley, B. D. (1976), The second-order analysis of stationary point processes, *Journal of Applied Probability, 13,* 255–266.

Wiegand, T and K.A, Moloney (2004) Rings, circles and null models for point patterns analysis in ecology. *Oikos*, 1104: 209-229

Acknowledgement

This Workbook collects together materials I have used over many years and in many different educational contexts. I am grateful to Dr. Nick Tate, University of Leicester for his suggestion that they could be gathered together under a heading of 'spatial literacy' and to the SPLINT initiative for the Fellowship that enabled me to do this. In preparation, I am grateful for additional discussion with Sarah Witham Bednarz, Chris Brunsdon, Jason Dykes, Nick Tate, and Harvey Miller. Students at Universities in Leicester, London, Hamilton (NZ), Christchurch, and Redlands together with those enrolled on my course for http://www.statistics.com have often unwittingly helped in the development of almost all the ideas and exercises that are presented. Some of the exercises aren't possible without use of software created by Luc Anselin, Ned Levine, the team that created 3Dfield, and Jo Wood. Alex Szumski 'road tested' the exercises and checked the manuscript for typos and the like. Any remaining errors are of course entirely my own responsibility.

I am grateful to Karl Grossner for permission to use Figure 1.1, to John Davis for permission to reference and use data (STRIAE and NOTREDAM) from his text *Statistics and Data Analysis in Geology* and to Wiley & Sons (New York) for permission to redraw and reuse my own figures 4.1, 4.3, 4.6, 5.4, 5.13, and 6.2 from O'Sullivan and Unwin (2003 & 2010) *Geographic Information Analysis*.

David Unwin
December 22nd 2010

Copyright and Intellectual Property

In the post World Wide Web, e-learning, world, attribution of Copyright to materials such as those in this Workbook isn't easy. There are ideas and possibly short sections of text that could be claimed by John Wiley & Sons, Birkbeck London, University of Leicester, University of Redlands (USA), Karen K Kemp (with whom I worked in Redlands) and myself. As sponsors of the project that has assembled them and provided their *raison d'etre* in a spatial literacy context, the SPLINT Centre for Excellence in Teaching and Learning might well also lodge a claim. Where there might be doubt, I have indicated this by use of an appropriate qualifying phrase such as 'source' or 'taken from'.

In this form and context these materials have been authored entirely by myself, so it is reasonable to claim them as © David J. Unwin (2010) and to assert the moral rights of authorship. They are made available under the understanding that they can be used in teaching and modified to suit local circumstances without and claim or charge provided that the source is recognized by the statement:

These materials have been developed from the Workbook 'Numbers aren't Nasty' assembled by David J. Unwin under the auspices of the (UK) Spatial Literacy in Teaching Initiative, (2010).

Any repurposing for commercial gain is subject to the usual 'all rights' © claim given above