Extending Point-pattern analysis to polygons using vector representations

Peter Whigham

[1]Department of Information Science
University of Otago. Dunedin, New Zealand
Phone: +64 3 479-7391 Fax: +64 3 479-8311
Email: peter.whigham@otago.ac.nz

Presented at SIRC NZ 2013 – GIS and Remote Sensing Research Conference
University of Otago, Dunedin, New Zealand
August 29th – 30th 2013

Keywords and phrases**:** Ripley's K, point-pattern analysis, polygon pattern analysis, clustering

1.0 INTRODUCTION

Point pattern analysis is a fundamental approach in many disciplines. The concept of clustering, interaction between elements in space such as repulsion or attraction, and the overall structural description of a set of spatial objects often forms the basis for confirming hypotheses related to patterns driven by an underlying process. Pattern descriptions are generally based on the degree of clustering or dispersion based on a comparison with a random (typically Poisson) process. Simple methods for analysis include functions based on distance, such as the nearest neighbour distance and the empty space function. The development of 2nd order methods (Lieshout and Baddeley, 1996) for estimating these pattern types have led to Ripley's K and associated transforms (such as the L function), which are now a common approach in ecology and vegetation science.

Although point patterns have a wide range of application, the approximation of an object as a point can lead to a number of errors regarding the observed pattern statistics. In particular, a vector object represented as a centroid increases the distance between objects and may therefore mask patterns of clustering that may exist in the areal data. In addition vector objects may overlap, have holes or complex shape, all of which are eliminated by a reduction to a point representation.

The extension of point pattern analysis to a grid-based approach using O-ring statistics was first proposed by Wiegand and Moloney (Wiegand and Moloney, 2004), and extended to handle 2nd order statistics (Wiegand, Kissling et al., 2006). This approach used a categorical raster representation where the cell size (scale) was selected based on the smallest object to be represented. An object was represented by a grouping of adjacent cells. This approach has been successfully used in a number of ecological studies including habitat loss and fragmentation (Bruggeman, Wiegand and Fernandez, 2010), forest stand structure (Barbeito, Pardos et al., 2008) and the influence of grazing on species interactions and stress (Graff and Aguiar, 2011). Although allowing areal objects to be assessed in terms of Ripley's K and other measures, the grid-based representation meant that objects could not overlap. In addition, the simulation models required for comparing the observed patterns in space to a random configuration were restricted to rotating each object by 0, 90, 180 or 270 degrees and then randomly shifting the object within the grid.

This paper introduces a vector-based approach to realising distance-based and K-statistic measures which allows overlapping objects and arbitrary rotation during simulation. In addition, since the objects do not have to be mapped to a raster representation, the scale of the data is represented by the original detail used when the data was collected.

2.0 VECTOR-BASED PATTERN STATISTICS

The following pattern statistics for polygons are implemented in the open-source package R (R Development Core Team ,2011), and use the "sp", "spatstat" and "rgeos" libraries for point and vector representation, geometric operations and display. Issues regarding randomisation of patterns, edge correction and details for each algorithm will be given in the appropriate section. For the purposes of comparison to a point-based measure the operations will be compared to point patterns and random simple polygons generated with the point as a centroid.

2.1 Empty Space Distance

The empty space distance function represents the nearest distance to a polygon from an arbitrary location within the border window of the dataset. This is calculated for a set of point locations as a grid within the border window. Figure 1 shows a point-based empty space function as a map and a corresponding map for random rectangles with the point as the centroid.

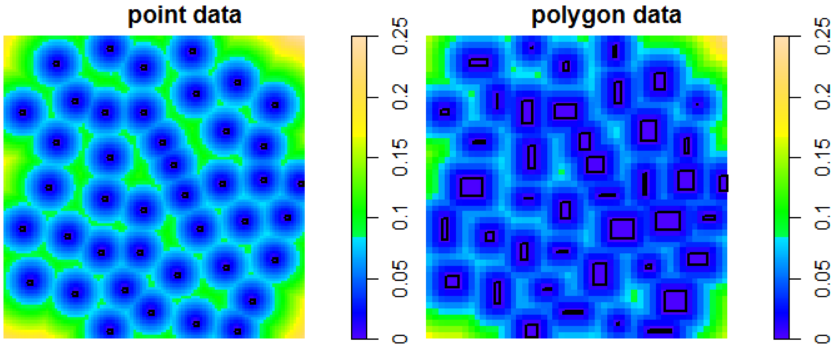

Figure 1. Empty Space Distance function for a set of points and random polygons with the same centroid

Although Figure 1 is useful for visualisation, it does not assess the degree of clustering or dispersion with the data. For this to occur we define an empty space function F which defines the cumulative distribution function of the empty space distance. See Baddeley (2008) for details of this definition for point-based processes. One issue that arises with all functions estimated from a sample of points within a study region is the correction for bias due to edge effects. With point data this is often done based on the empirical cumulative distribution function. For polygons a similar approach is used, however any operation that increases the size of a polygon beyond the study window is clipped. The empty space function F(r) defines the observed number of points found from an arbitrary point with increasing distance (r). A comparison of this count versus a Poisson process is used to compare the observed distribution to a random point pattern with the same intensity. Hence if $F(r) > F_{pois}(r)$ this indicates that the points are regularly space, while $F(r) < F_{pois}(r)$ implies the point pattern is clustered. For points this can be formally defined, however with polygons a simulation is required that distributes the polygons randomly within the study region and calculates the F statistic for each randomisation. For both point and polygon patterns a confidence interval can be created that indicates a significance level based on the number of simulations. A significance level of 0.05 is obtained when the number of simulations is 39 (Baddeley, 2008).

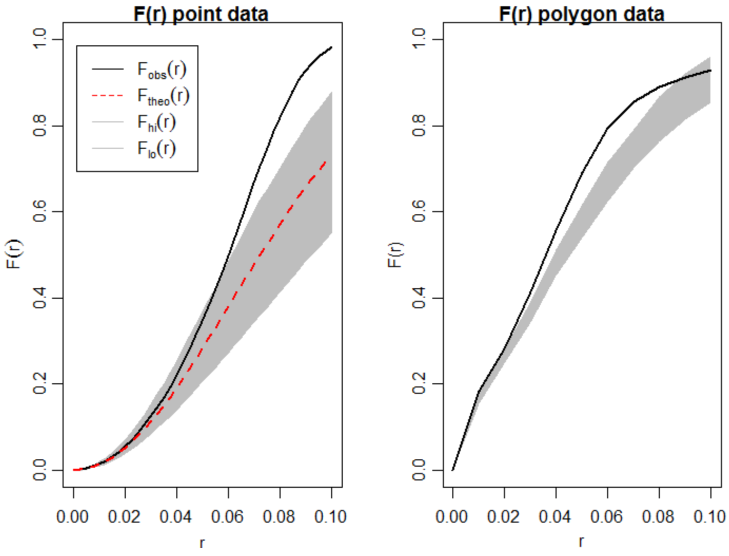

Figure 2. Empty space function F(r) for the point and polygon data

Fig. 2 shows the empty space function F(r) for the point and polygon data with a 95% confidence interval for a simulated poisson process. The main feature to note is that the polygon data shows regularity in the distribution of the polygons for smaller distance values (r). In addition there is a suggestion of random patterning for large r. Hence the use of points to approximate the polygon data underestimates the amount of regularity in the data, even though the randomised polygons for each point are quite small (see Fig. 1).

2.2 Ripley's K Estimate

The K function was first introduced by Ripley (Ripley, 1977) based on the distance between observed points for stationary patterns. Hence, given an intensity of points λ, Ripley defined λK(r) as the expected number of points that would be found within a distance r of an arbitrary observed point. A theoretical Poisson model can then be compared against the observed K(r) to indicate clustering or dispersion (regularity) within the pattern.

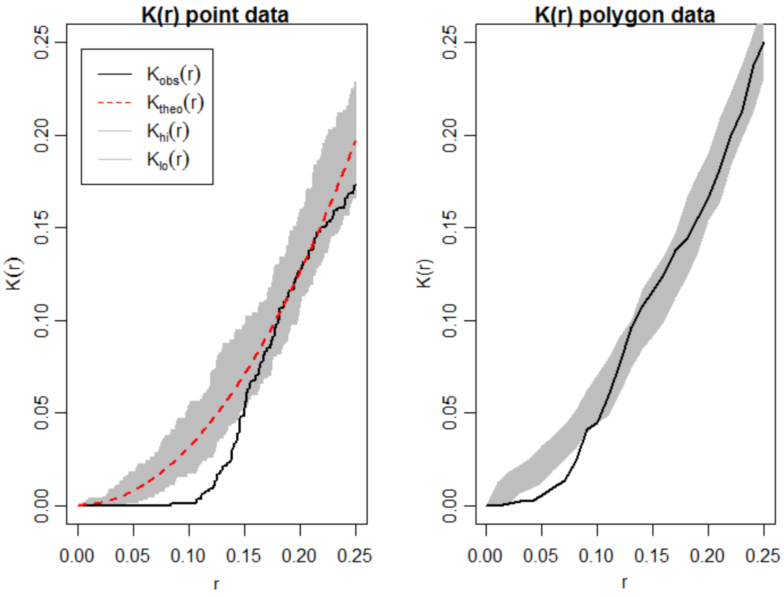

Figure 3. Ripley's K(r) estimate for point and polygon data

Figure 3 shows the effect of polygon objects approximated with point centroids for Ripley's K estimate. For the point data the indication is that the points are dispersed for distances up to r ≈ 0.15, whereas the polygon data suggests a random pattern once r > 0.8. There is also some suggestion that for the polygon data there is a clustering at a distance scale of approximately 0.13, whereas the point estimate for K(r) does not suggest clustering at any scale. Clearly patterns may change significantly if objects are represented as points rather than their correct areal representation. Figure 4 shows an example using real polygon data.

2.3 Randomization of polygons within a window

The use of simulations for estimating the poisson model for polygons has one significant difference from a point pattern – polygons may overlap. If the original polygon dataset had overlaps (or the objects of interest could overlap) then randomisation of the polygons is simple. However, many polygon patterns are constrained to not overlap, and therefore the randomisation of these polygons must also satisfy this property. Unfortunately the random placement of polygons within a window without overlap is a difficult problem to solve efficiently. The current approach is that if overlapping is not allowed a set number of trials are conducted for each polygon, placing it randomly within the window until no overlap occurs. If the polygon cannot be placed it is ignored and the process continues until all polygons have been tried. This may result in fewer polygons being used for the Poisson estimate, however it avoids issues with overlapping randomisation creating artefacts that are especially noticeable when r ≈ 0.

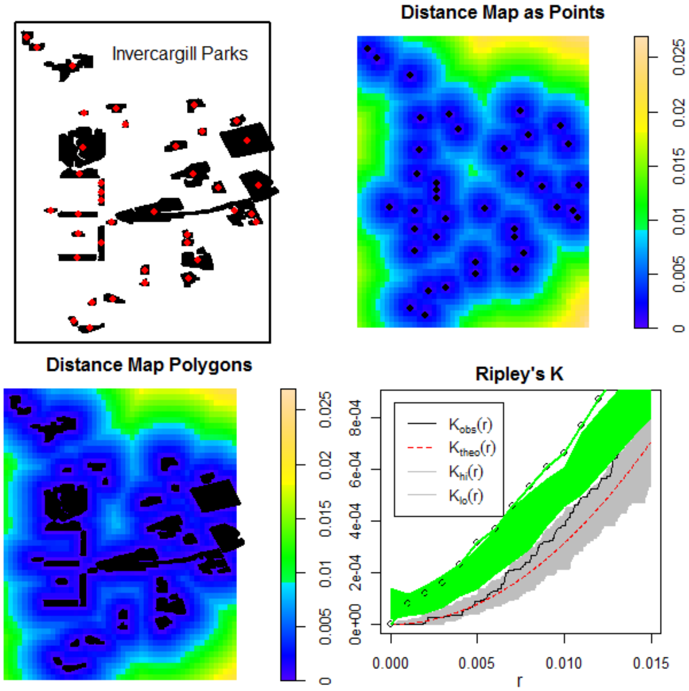

Figure 4. Ripley's K(r) and Distance maps for Invercargill Parks. The polygon Ripley's K is in green

3.0 CONCLUSIONS

This paper has presented the extension of two standard point-based functions to estimate whether a polygon dataset exhibits random, clustered or dispersed (regular) patterns in space. The reduction of polygons to points is likely to bias the estimate of the distribution properties of these data and therefore these types of operations are required if a true measure of the patterning in space is to be estimated. Other standard point-based functions have also been extended to polygon representations and will be described in detail in later publications.

REFERENCES

Baddeley, A. (2008) Analysing spatial point patterns in R. Workshop Notes, Version 3. CSIRO. URL http://www.csiro.au/files/files/piph.pdf

Barbeito, I, Pardos, M., Calama, R. and I. Cannellas (2008) Effect of stand structure on Stone pine (*Pinus pinea* L.) regeneration dynamics, Forestry, Vol. 81, No. 5, 2008. doi:10.1093/forestry/cpn037

Bruggeman, D.J., Wiegand, T and Fernandez, N (2010) The relative effects of habitat loss and fragmentation on population genetic variation in the red-cockaded woodpecker (Picoides borealis), Molecular Ecology19, 3679–3691.

Graff, P. and Aguiar, M. R. (2011) Testing the role of biotic stress in the stress gradient hypothesis. Processes and patterns in arid rangelands. Oikos 120: 1023–1030, 2011

Leishout, M. and Baddeley, A. (1996) A nonparametric measure of spatial interaction in point patterns. Statistica Netherlands, Vol 50(3) 344-361.

R Development Core Team (2011). R: A language and environment for Statistical Computing, Vienna, Australia. URL http://www.R-project.org/.

Ripley, B.D. (1977) Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, series B*, 39:172–212.