The importance of managing co-variate and distance biases in point-pattern clustering

P.A. Whigham*[1], B. de Graaf[2], R. Srivastava[3], and P. Glue[4]

[1]Information Science Department, University of Otago, Dunedin, New Zealand

[2]Department of Preventive and Social Medicine, University of Otago, Dunedin, New Zealand

[3]Department of Psychological Medicine, University of Otago, Dunedin, New Zealand

[4]Department of Psychological Medicine, University of Otago, Dunedin, New Zealand

*Email: peter.whigham@otago.ac.nz

Abstract

Assessing social conditions for individuals often involves the evaluation of point-based patterns over some geographic region. This paper demonstrates the influence of co-variate information on observed clustering patterns and a method for assessing the role of distance metrics on clustering. The approaches are applied to assess social contagion of deliberate self-harm for a set of individuals observed over a two year period in a small community in New Zealand.

Keywords: point pattern, Ripley's K, clustering, Manhattan distance.

1 Introduction

Point-pattern analysis is commonly used to assess clustering or dispersion of a set of observations in a bounded geographic region using methods such as the empty space function, pairwise and nearest neighbour distance measures (Bailey and Gatrell (1995); Anselin and Rey (2010); O'Sullivan and Unwin (2010)). Here we modify the approach of Ripley's K pairwise measure (Ripley (1977); Diggle and Chetwynd (1991)) to account for co-variate effects of observed clustering patterns. In addition, since the observed point patterns are a subset of the possibly observed locations a stochastic resampling method is required to address the unobserved data. In addition, since the distance metric is fundamental for measuring clustering we examine the influence of metric choice by using a rotated Manhattan distance to assess bias in spatial structure (Whigham et al. (2016)). Note that we only consider homogeneous point patterns which assume a constant intensity across the study area. Although this assumption may not be always justified, for our situation the introduction of a static set of candidate points (where observations have not been made) means that we are only interested in observed deviations for a constant background process.

Previous clustering methods for self-harm behaviour used area-based counts for index events which aligned with other area-based covariates (such as social deprivation). This allowed regression approaches to be constructed to account for covariates and spatial lag (Anselin et al. (1996)). Morans

*

I or other count-based methods could then be used to assess clustering (Fortune and Hawton (2005); Rehkopf and Buka (2006)). Here we have point-based data of residential addresses for people who have presented to the Emergency Department or Emergency Psychiatric Service Team between January 2011 and December 2012 within the Invercargill region, New Zealand. The initial data ($n = 254$) was reduced to those that intersected residential parcels in the urban region of Invercargill, resulting in $n = 136$. For a point-based approach to clustering we will require handling possible co-variate properties of space, handle the unobserved (possible) locations of people, and assess the effect of distance measures on the significance of observed pattern. Details regarding the data collection and a detailed analysis of the methods may be found in Whigham et al. (2016).

2 Methods

The second order moment (Ripleys K) for an unlabelled, homogeneous, isotropic point process observed as a set of points $x_i \in \Re^2$ is defined as [5]:

$$K(r) = \lambda_{-1} E[points \leq r \in x_i] \tag{1}$$

where λ is the intensity of the point process per unit area. For an isotropic process comparisons with $K(r)$ are normally based on the homogeneous Poisson process $Kpois(r) = \pi r^2$ (Diggle and Chetwynd (1991)). For a homogeneous process the assumption is that the process λ can be approximated by the number of points divided by the observed region area. For our derivation of K(r) observations are constrained to a finite set of possible locations. Hence λ is set to the number of points divided by the maximum distance between any two points in x_i.

Our problem describes a set of observed locations where people live who have been diagnosed with deliberate self-harm. Since our question is to consider whether there is social contagion involved with the geographic pattern of these people we need to address the issue that they could have lived in any location within the city, but also that where they live may be related to socio-economic factors. This is done by creating a set of marked point patterns representing the location of individuals with self-harm, with the remaining geographic locations (points) created using the residential parcel footprints and placing a point at the centre of each residential block. We refer to the complete set of points as W.

This results in a dataset where each point x_i has an associated mark from a finite set of marks M (which in our case is the deprivation index Salmond and Crampton (2012)), defining a marked point pattern:

$$y = (x_1, m_1), \ldots, (x_n, m_n), x_i \in W, m_i \in M \tag{2}$$

We observe a set of q marked points $q \in y$, where $q \ll n$ and want to determine if the set q deviates from complete spatial randomness. In addition, since the marked pattern may be spatially correlated to the process generating the point pattern, the distribution of the observed marks of q must be taken into account when simulating a random sample from y. Initially (since q is fixed), we construct the discrete cumulative distribution function $F(X)$ for the q marks that defines the probability of observing a point given a certain level of deprivation. This will be used to select points from W as possible locations of deliberate self-harm based on the observed distribution of q. This will allow the clustering due to deprivation to be adjusted for social structure. Initially we use the Euclidean distance metric (Minkowski Distance L_2 for clustering.

```
Result: $K(r_i) = \lambda_{-1} * P$ for distance $r_i$
$P = \{\}$;
while $q$ points not selected do
    $\rho$ = uniform random number in [0,1);
    Select mark $m_k$ from distribution $F(X)$;
    Select points $t = \{(x_i, m_k) \in y$;
    Randomly select point $s \in t$;
    $P = P \cup s$;
end
```

Algorithm 1: Ripley's $K(r_i)$ for one sample of points from W

Algorithm 1 creates point locations that follow the co-variate distribution of deprivation and can be used as the comparison against the observed distribution of identified individuals. Hence any deviation from randomness in the spatial pattern can be inferred as resulting from some other process such as social contagion. Although this approach handles co-variate issues in space there is still the issue of the appropriateness of Euclidean distance when considering neighbourhood. Since Invercargill largely follows a grid-based road layout, distance is probably more meaningful as a Manhattan distance (Minkowski L_1). This also allows a test of the significance of the layout of roads versus the distance measure since L_1 is sensitive to rotation. To address the effect of distance bias due to the orientation of the streets a set of simulations can be done where the spatial configuration is rotated and L_1 used as the distance metric for each rotation when calculating $K(r)$. Evidence of clustering for all rotations would give confidence in the observed pattern being independent of the arbitrary spatial structure of the road network.

3 Results

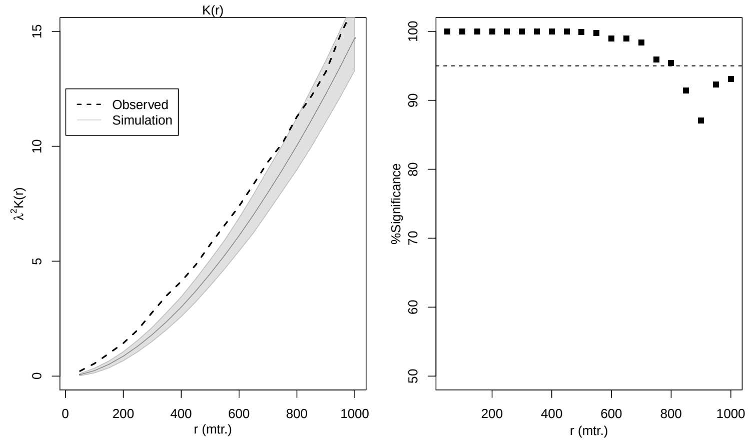

Figure 1: $K(r)$ using Euclidean distance without covariates. Note significant clustering observed up to $\approx 800m$.

Figure 1 shows the result of 1000 independent runs for Ripley's K estimated using the original points within the study region of residential housing. Since $K(r)$ is sensitive to the defined study area (used to calculate λ), and the observed self-harm addresses were correlated to deprivation, the significant clustering up to $\approx 800m$ has to be interpreted with some caution.

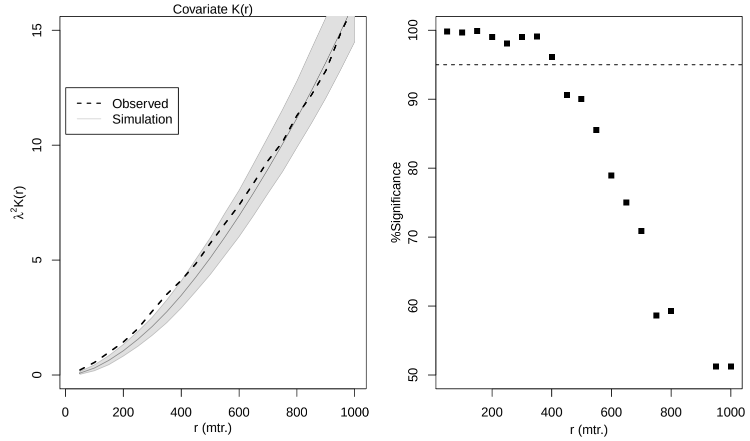

Figure 2: $K(r)$ using Euclidean distance with points simulated based on the co-variate deprivation. Note the reduced significant clustering observed to $\approx 400m$ compared with Figure 1.

Figure 2 shows the clustering estimate for $K(r)$ when the co-variate of deprivation is handled as part of the assessment. The significant reduction in clustering distance is therefore a result of the inherent clustering of the deprivation index observed in the Census data. However, there still appears to be some clustering that may indicate social contagion for deliberate self-harm. Since Euclidean distance has been used, but social distance is not likely to follow a straight line relationship, it is necessary to test the dependence on this metric.

Figure 3 shows the one-sided test for significance using the L_1 (Manhattan) distance metric for various rotations of the study area whilst handling co-variate deprivation. Evidence for clustering for all of these rotations would be a strong indicator that there is some additional social behaviour influencing the resulting clustering. Since Figure 3 shows that for distances up to $\approx 400m$ there is still evidence of clustering suggests some other factor, such as a social relationship between people, may be an influence on behaviour.

4 Conclusion

This extended abstract has focused on demonstrating some of the issues associated with using point-pattern analyses for understanding social behaviour. In particular, socio-economic factors have been shown to have a significant effect on the measured properties of the data. In addition, we have demonstrated a method for assessing the bias due to the arbitrary orientation of spatial data and the relationship to a distance metric. This also highlights issues associated with problems where the placement of points is constrained - a situation that is common when data describes positions related to housing, residential behaviour and fixed structures.

There are a number of limitations to the approach presented here. The limited data of two years of observations (people with emergency admission for self-harm) may introduce a bias, however this has also meant that non-stationarity in time could be ignored. The small sample size may have biased the result, and certainly reduces the ability of the study to infer generalisations to other communities. From a methods perspective the assumption of a homogeneous intensity (i.e.

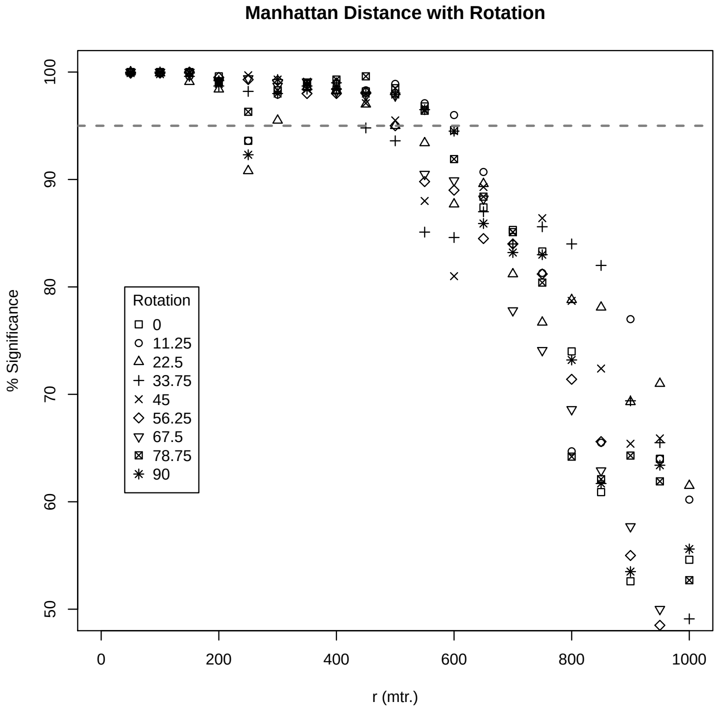

Figure 3: Rotated Manhattan distance and adjusting for co-variate deprivation. Note the significant clustering observed to $\approx 400m$

λ is constant) simplified the analyses, however this may not be true. The difficult with handling inhomogeneous point pattern intensities is that incorporating and managing co-variate information becomes more difficult. In addition we have only used a single co-variate (deprivation) whereas other measures of social or economic structure, or the clustered spatial pattern of other residential features (such as alcohol outlets) may influence behaviour and therefore observed clustering. Finally, there are extensions to Ripley's K that involve marked point patterns, such as the cross K function (Diggle and Chetwynd (1991)), that could be used as part of this research. However, the same issues with co-variates and biased distance metrics would have to be addressed.

5 References

Anselin, L., A. K. Bera, R. Florax, and M. J. Yoon
1996. Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, 26(1):77–104.

Anselin, L. and S. Rey
2010. *Perspectives on Spatial Data Analysis: Advances in Spatial Science.* Springer-Verlag Berlin Heidelberg.

Bailey, T. C. and A. C. Gatrell
1995. *Interactive Spatial Data Analysis.* Harlow Essex, England; New York, NY: Longman Scientific & Technical ; J. Wiley. OCLC: 32468517.

Diggle, P. J. and A. G. Chetwynd
1991. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, 47(3):1155–1163.

Fortune, S. A. and K. Hawton
2005. Deliberate self-harm in children and adolescents: A research update. *Current Opinion in Psychiatry*, 18(4):401–406.

O'Sullivan, D. and D. Unwin
2010. *Geographic Information Analysis*. Wiley.

Rehkopf, D. H. and S. L. Buka
2006. The association between suicide and the socio-economic characteristics of geographical areas: A systematic review. *Psychological Medicine*, 36(2):145–157.

Ripley, B. D.
1977. Modelling Spatial Patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–192.

Salmond, C. E. and P. Crampton
2012. Development of New Zealand's deprivation index (NZDep) and its uptake as a national policy tool. *Canadian Journal of Public Health = Revue Canadienne De Sante Publique*, 103(8 Suppl 2):S7–11.

Whigham, P. A., B. de Graaf, R. Srivastava, and P. Glue
2016. Managing distance and covariate information with point-based clustering. *BMC medical research methodology*, 16(1):115.