ARTICLE IN PRESS

Spatial Statistics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

Spatially significant cluster detection

Alan T. Murray [a,*], Tony H. Grubesic [b], Ran Wei [a]

[a] *GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287, United States*

[b] *Geospatial Intelligence and Planning, College of Earth, Ocean and Atmospheric Sciences, Oregon State University, Corvallis, OR 97331, United States*

A R T I C L E I N F O

Article history:
Received 12 August 2013
Accepted 7 March 2014
Available online xxxx

Keywords:
Clustering
Hot spots
Spatial scan
Optimization
ESDA

A B S T R A C T

Cluster discovery techniques are a fundamental group of exploratory methods designed to identify areas exhibiting elevated levels of disease, risk, danger, etc. Given the intent of cluster detection, spatial structure plays an important role and must be taken into account appropriately if meaningful clusters are to be found. This paper discusses contiguity and the ways in which it is central to local clusters that may be of interest for planners, managers and policy makers. While spatial contiguity is widely considered an important condition of a cluster, most detection approaches employ *a priori* artificial structure, leading to disingenuous significance and unintended spatial biases that hinders meaningful discovery and interpretation. The basis for significance is reviewed, and methods for maximizing likelihood are detailed. An approach is presented for addressing spatial contiguity explicitly in cluster detection without the use of arbitrarily shaped scan windows. A case study using crime events within a major urban region is presented, with empirical results used to illustrate capabilities for identifying significant and meaningful clusters.

1. Introduction

There are many situations where analysts are interested in detecting significant spatial clusters, including law enforcement and criminology (Sherman and Weisburd, 1995; Harries, 1999; Eck et al.,

* Corresponding author.
E-mail address: atmurray@asu.edu (A.T. Murray).

http://dx.doi.org/10.1016/j.spasta.2014.03.001
2211-6753/

2005), ecology (Stohlgren et al., 1999; Myers et al., 2000), epidemiology (Kulldorff et al., 1997; Heymann and Rodier, 2001), astrophysics (Gladders and Yee, 2000; Kim et al., 2002) and geography (Ord and Getis, 1995; Anselin, 1995; Murray and Estivill-Castro, 1998; Grubesic, 2006; Rogerson and Yamada, 2009), among others. Often referred to as ''hot spots'', these geographic areas exhibit a higher concentration of events or objects than their surrounding areas. For example, in crime analysis hot spots often correspond to a single address, a block, a street, a neighborhood or some other frame of spatial reference where there is an elevated or unusually high occurrence of criminal activity (Bowers and Johnson, 2005).

Although approaches for hot spot detection and spatial cluster discovery are varied, including visual techniques (Julesz, 1962) and exploratory spatial data analysis (ESDA) (Dubes and Jain, 1980; Murray and Estivill-Castro, 1998), many are structure imposing. Consider, for example, hierarchical and non-hierarchical cluster analysis (Everitt, 1993; Kaufman and Rousseeuw, 2005). Both techniques require that the number of cluster groups, k, be specified a priori. This basic constraint ensures that a specific structure, a cluster count, is imposed upon any and all results (Murray and Grubesic, 2013). Many spatial scan statistics (Naus, 1965; Loader, 1991; Kulldorff, 1997) are also structure imposing, using a pre-defined, geometric window in the assessment of potential clusters. For example, the Geographical Analysis Machine (Openshaw et al., 1987) uses a grid of points with circles of various radii over the grid. If the number of incidents within a given circle exceeds the expected number of incidents for an underlying population (and is statistically significant), the geographic extent of the circle is considered a hot spot. Similarly, both the circular and elliptical spatial scan statistics (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997; Kulldorff et al., 2006) use pre-defined geometric shapes for identifying and testing for potential clusters.

There are two issues associated with *a priori* structured approaches. First, the use of pre-defined geometric shapes can mask the actual spatial morphology of hot spots. Simply put, actual spatial clusters do not necessarily correspond to generic geometries like circles, ellipses, squares or any other shapes used to tessellate space. Second, imposed structures can impede statistical inference, masking the underlying causes of clusters and/or the relationship between clusters and their social, economic and ecological environment. To address these and other issues, an evolving area of research is focused on finding irregularly shaped clusters (Duczmal et al., 2006, 2007, 2008; Pei et al., 2011; Costa et al., 2012), though the theoretical underpinnings are only beginning to evolve.

Given recognized limitations, the purpose of this paper is to review the implications of assumed spatial structure in cluster detection approaches. Further, this paper develops a likelihood maximization approach without an assumed spatial window. In the next section, we provide an overview of clustering approaches with assumed spatial structure, discussing their strengths and weaknesses. This includes the role of spatial contiguity and how it can influence hot spot detection. Section 3 presents criteria for likelihood maximization and uses this in an optimization-based approach for detecting spatial clusters. This is followed by a case study that uses crime events within a major urban region to illustrate the utility of the developed approaches. The paper concludes with a discussion of the results and implications for future research.

2. Cluster detection

Approaches to support cluster analysis are generally designed to impose structure on observations/events in order to account for similarity of some sort. Similarities are quantified in different ways, including visual appearance, event frequency, geographic location, time, attribute, or a multivariate combination of space, time, etc. From a statistical perspective, cluster analysis typically seeks groupings that minimize within-group variance (Fisher, 1958; Everitt, 1993; Kaufman and Rousseeuw, 2005). This often reflects the desires of substantive contexts. For example, law enforcement agencies, public health practitioners, ecologists and the like are interested in unusual concentrations of activity (or non-activity) because of social, economic and policy implications associated with elevated rates of crime, disease, environmental degradation, etc., respectively, within or between regions. While identifying significant heterogeneities across the landscape is important for intervention efforts and/or policy development, the process of identifying anomalies requires the development of robust