Computers & Geosciences 33 (2007) 952–965

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

Delineation of support domain of feature in the presence of noise

Tao Pei[a,*], A-Xing Zhu[a,b], Chenghu Zhou[a], Baolin Li[a], Chengzhi Qin[a]

[a]*State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, CAS, 11A, Datun Road Anwai, Beijing 100101, China*

[b]*Department of geography, University of Wisconsin Madison, 550N, Park Street, Madison, WI 53706-1491, USA*

Received 15 May 2006; received in revised form 15 November 2006; accepted 17 November 2006

Abstract

Clustered events are usually deemed as feature when several spatial point processes are overlaid in a region. They can be perceived either as a precursor that may induce a major event to come or as offspring triggered by a major event. Hence, the detection of clustered events from point processes may help to predict a forthcoming major event or to study the process caused by a major event. Nevertheless, the locations of existing clustered events alone are not sufficient to identify the area susceptible to a potential major future event or to predict the potential locations of similar future events, so it is desirable to know the shape and the size of the region (the "territory" of feature events) that the feature process occupies. In this paper, the support domain of feature (SDF), the region over which any feature event has the equivalent likelihood to occur, is employed to approximate the "territory" of feature events. A method is developed to delineate the SDF from a region containing spatial point processes. The method consists of three major steps. The first is to construct a discrimination function for separating feature points from noise points. The second is to divide the entire area into a regular mesh of points and then compute a fuzzy membership value for each grid point belonging to the SDF. The final step is to trace the boundary of the SDF. The algorithm was applied to two seismic cases for evaluation, one is the Lingwu earthquake and the other is the Longling earthquakes. Results show that the main earthquakes in both areas as well as most aftershocks triggered by them fell into the estimated SDFs. The case study of Longling shows that the algorithm can deal with a region containing more than two processes.

Keywords: Nearest-neighbor; Cluster; Spatial point process; Poisson process; Spatial data mining; Earthquake prediction; EM algorithm

1. Introduction

Many natural phenomena manifest themselves as spatial point processes which produce numerous events in space, such as earthquakes, landslides and craters. For clarification, we define an occurrence of a phenomenon located at a single point as an event in contrast to a simple geometric point. An event set is often not completely randomly distributed. Some events assemble in a restricted region while other events are dispersed over the remaining area. The former, distributed with higher intensity, is viewed as a cluster or "hotspot" (Brimicombe, 2003) and probably reveals some meaningful pattern (Jemal et al., 2002; Steenberghen et al., 2004), while the latter, distributed with lower intensity, is considered

*Corresponding author. Tel.: +86 10 64888960; fax: +86 10 64889630.

E-mail addresses: peit@lreis.ac.cn (T. Pei), axing@lreis.ac.cn, axing@geography.wisc.edu (A.-X. Zhu), zhouch@lreis.ac.cn (C. Zhou), libl@lreis.ac.cn (B. Li), qincz@lreis.ac.cn (C. Qin).

0098-3004/$ - see front matter
doi:10.1016/j.cageo.2006.11.010

as noise or background (Fraley and Raftery, 2002). In many natural processes, the clustered events can be deemed either as a precursor that may induce a huge event to come or as offspring triggered by a huge event. Hence, the detection of clustered events from point processes may help us to predict a forthcoming major event or allow us to study the process caused by a major event (Wu et al., 1990; Ogata, 2001; Pei et al., 2006). The issue of how to detect a feature/cluster from a data set in the presence of noise has been extensively discussed (Ester et al., 1996; Byers and Raftery, 1998; Pei et al., 2006; Wang et al., 2006). Nevertheless, the locations of clustered events may not reveal the region in which a major event or successive events may take place, and the region may be very important in the case of constructing a natural disaster prevention plan, such as a quakeproof plan. It is thereby desirable to know the region (the "territory" of feature events) where future similar events might occur not only for identifying the area susceptible to a potential major future event but also for predicting the potential locations of future events. However, the locations of existing feature events alone are not sufficient to reveal the shape and size of the "territory" of clustered events. Therefore, it is necessary to construct a method that can precisely delineate the support domain of feature (SDF), which is referred to as the "territory" occupied by feature (clustered) events, in the presence of noise (background).

The problem of how to estimate the boundary of a point process, which is referred to as the convex hull of a point process, has been widely discussed (Ripley and Rasson, 1977; Davis et al., 1988; Hall et al., 2002; Chiu and Molchanov, 2003). However, their methods may overestimate the area of the support domain of a point process when it is distributed over a region with a concave shape, and also are restricted to only one process. If two Poisson processes with different intensities are superimposed, the support domain of one is included in that of the other. The methods for estimating the convex hull, because they are designed for approximating the boundary of one point process, will not be applicable due to the interference caused by the noise.

There are a number of studies addressing the issue on contouring clustered events in the presence of background noise. Banfield and Raftery (1993) proposed a model-based clustering method in which the clustered events are presumed to be mixtures of Gaussian density. The clustered events can be contoured at a specific value after the parameters of each Gaussian model have been derived. Fraley and Raftery (1998) constructed a multivariate normal mixture model for the sake of accommodating the different components originating from noise and feature. In their method, the Bayesian Information Criterion is utilized to determine the number of components. Jin et al. (2005) established a scalable model-based clustering framework. This algorithm is superior in analyzing large sets of complicated data and also significantly reduces the runtime as opposed to traditional algorithms. Although these methods may contour the clustered events in light of the difference in densities between the feature and noise, all of these ideas are built on the predefined model and therefore suffer from two limitations: a point process is rarely a Gaussian mixture and results from these methods are sensitive to departure from the model.

Allard and Fraley (1997) constituted a maximum likelihood estimator for a mixture of uniform point processes using the Voronoï tessellation defined by the data themselves. Although the method based on the Voronoï tessellation can determine the support domain of feature events by connecting the Voronoï polygons containing the feature events, it is limited in that the support domain of feature must be restricted to a single connected component without holes, and the boundary of features is of a specific geometry. Huo and Lu (2004) presented a digraph-based algorithm to estimate the boundary of the higher concentration regions (HCRs) from point processes. Although the method can adapt to HCRs with arbitrary shapes, users have to predefine the geometric constraint to the boundary of the HCR and specify the center(s) of the underlying region in advance.

In this paper, the support domain of feature is regarded as the "territory" occupied by feature events. In fact, the SDF is defined as the region in which a given feature event is equally likely to occur anywhere. An algorithm is developed in this paper to delineate the SDF from a region containing spatial point processes. The algorithm assumes that the entire region contains two point processes with different intensities. Any point in the region can be classified as feature or noise based on the distance between itself and its kth nearest event. Then, the entire region is divided into a mesh grid and each node is endowed with a fuzzy membership value of belonging to the SDF. Finally, the boundary of the