Spatial Statistics 14 (2015) 400–411

Contents lists available at ScienceDirect

Spatial Statistics

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

Spatial cluster detection using nearest neighbor distance

Avner Bar-Hen [a,*], Mathieu Emily [b], Nicolas Picard [c]

[a] *Université Paris Descartes, Laboratoire MAP5 (CNRS-UMR 8145), 45, rue des Saints-Pères, 75270 Paris, France*

[b] *Agrocampus Ouest - IRMAR UMR CNRS 6625, 65, rue de Saint Brieuc, 35042 Rennes Cedex, France*

[c] *CIRAD, BP 4035, Libreville, Gabon*

ARTICLE INFO

Article history:
Received 6 March 2015
Accepted 15 July 2015
Available online 31 July 2015

Keywords:
Classification
Cluster detection
Point process
Nearest-neighbor distance

ABSTRACT

Motivated by the analysis of the impact of ecological processes on spatial distribution of tree species, we introduce in this paper a novel approach to detect spatial cluster of points. Our procedure is based on an iterative transformation of the distance between points into a measure of closeness. Our measure has the advantage of being independent of an arbitrary cluster shape and allowing adjustment for covariates. The comparison of the observed measure of closeness to a reference point process leads to a hierarchical clustering of spatial points. The selection of the optimal number of clusters is performed using the Gap statistic. Our procedure is illustrated on a spatial distribution of the *Dicorynia guianensis* species in the French Guiana *terra firme* rainforest.

1. Introduction

The spatial pattern of a given population of a tree species can be viewed as the result of interactions between the biology of the population and other ecological processes on the abiotic and biotic environments of the population. Describing the spatial distribution of a tree species, as a product of these processes, is an important tool for understanding its dynamic. Points process modeling is a useful tool for modeling the spatial determinants of tree species distribution while considering that

* Corresponding author.
E-mail address: avner.bar-hen@mi.parisdescartes.fr (A. Bar-Hen).

http://dx.doi.org/10.1016/j.spasta.2015.07.006

the exact locations of trees is random (Cressie, 1993; Diggle, 1983; Ripley, 1988; Stoyan et al., 1995; Stoyan and Stoyan, 1994; Daley and Vere-Jones, 1988).

One main purpose of point process modeling is the identification of spatial clusters that characterize spatial areas exhibiting a high concentration of events or points. Since detecting spatial pattern of events is essential in many fields (medicine, cosmology with spatial clustering of galaxies, social sciences and criminology, agronomy and more), a substantial literature has been dedicated to the issue of spatial clustering (Murray et al., 2014). The two most popular cluster detection approaches are the spatial scan statistics (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997; Patil and Taillie, 2004; Duczmal and Assuncao, 2004; Tango and Takahashi, 2005; Demattei et al., 2007) and spatial autocorrelation (Ord and Getis, 1995; F Dormann et al., 2007; Stojanova et al., 2013). On one hand, spatial scan statistics aim at scanning the studied area using windows of an imposed shape (circles, ellipses or squares): based on a likelihood ratio test, spatial clusters are defined by the windows that group together an abnormally high number of cases. On the other hand, spatial autocorrelation methods rely on the significance of local indicators calculated according to a weighted neighborhood matrix between observed points. Both classes of methods are based on a pre-existing spatial structure: the geometric shape of the scanning window for spatial scan statistics and the spatial weights set in the neighborhood matrix for autocorrelation indicators. The use of arbitrary spatial structures is a clear limitation regarding cluster detection since actual cluster structures are not restricted to regular shapes nor to known weighted neighborhood. Furthermore, the use of pre-defined spatial structures is likely to mask the underlying cause of clusters, such as the relationship between clusters and their ecological environment for instance. Therefore, existing methods are likely to fail at distinguishing true clusters from covariates dependence.

In this paper, we introduce a novel statistical procedure to detect spatial cluster based on a transformation of the 2-dimensional observed point process into a collection of 1-dimensional ordered trajectories. The two main advantages of our approach are the following. First, our transformation of the data does not depend on an imposed spatial structure. Next, transformed data are compared to a reference Poisson point process that can be either homogeneous or inhomogeneous. The use of an inhomogeneous Poisson process as a reference allows to account for the effect of covariates.

The paper is organized as follows. In Section 2, details of our method of detection are given with respect to stationary Poisson process in $\mathbb{R}^2$. The main idea is to consider that distances within clusters should be smaller than distances between clusters. Based on the nearest neighbor distance distribution of a Poisson process, we propose to transform the distance between neighbor points into a measure of closeness. Our measure of closeness is then combined with classification theory to obtain a dissimilarity matrix of the observed points. Clusters are finally defined by thresholding the classification tree obtained from the dissimilarity matrix with the Gap statistic.

In Section 3, we performed simulation studies, with respect to predefined spatial cluster patterns. Two main features of our approach are illustrated. Our first goal is to evaluate the benefit of using the Gap statistic for choosing an optimal cutting threshold with respect to nine other well-known methods. Our second goal is to compare the accuracy of our procedure to scan statistics based methods.

Finally, Section 4 is devoted to the study of the spatial distribution of a tree species: *Dicorynia guianensis*. *Dicorynia guianensis* is a tropical tree species of the French Guiana characterized by an aggregated spatial pattern (Loubry, 1993; Collinet, 1997). The application of our methodology leads to the definition of clusters that are consistent with the ecological knowledge of the plot. Furthermore, we focus on the influence of the slope on the aggregation and show that the clustering is slightly modified. Such modification can be explained by the interaction that exists between clustering ecological processes, such as the range of dispersion, and topological properties.

2. Cluster detection

This section is devoted to the description of the methodology to detect cluster(s). Let Φ be a point process in $\mathbb{R}^2$ and $\varphi = \{x_1, \ldots, x_n\}$ a realization of Φ in a bounded set $A \subset \mathbb{R}^2$. As illustrated in our motivated example, x_i typically characterizes the spatial coordinates of a tree. Let further introduce d as a given distance in $\mathbb{R}^2$, such as the Euclidean distance for example.