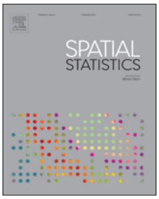
Spatial Statistics 16 (2016) 137–151

Contents lists available at ScienceDirect

Spatial Statistics

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



Application of optimal data-based binning method to spatial analysis of ecological datasets

Anna Tovo [a,*], Marco Formentin [a], Marco Favretti [a], Amos Maritan [b]

[a] *Department of Mathematics, University of Padova, Padova, Italy*

[b] *Department of Physics and Astronomy, University of Padova, INFN and CNISM, Padova, Italy*

ARTICLE INFO

Article history:
Received 9 December 2015
Accepted 18 February 2016
Available online 18 March 2016

Keywords:
Maximum a-posteriori Estimation
Spatial analysis
Cluster analysis

ABSTRACT

Investigation of highly structured datasets to unveil statistical regularities is of major importance in complex system research. The first step is to choose the scale at which to observe the process, the most informative scale being the one that includes the important features while disregarding noisy details in the data. In the investigation of spatial patterns, the optimal scale defines the optimal bin size of the histogram in which to visualize the empirical density of the pattern. In this paper we investigate a method proposed recently by K.H. Knuth to find the optimal bin size of an histogram as a tool for statistical analysis of spatial point processes. We test it through numerical simulations on various spatial processes which are of interest in ecology. We show that Knuth optimal bin size rule reducing noisy fluctuations performs better than standard kernel methods to infer the intensity of the underlying process. Moreover it can be used to highlight relevant spatial characteristics of the underlying distribution such as space anisotropy and clusterization. We apply these findings to analyse cluster-like structures in plants' arrangement of Barro Colorado Island rainforest.

© 2016 Elsevier B.V. All rights reserved.

[*] Corresponding author.
E-mail address: annatovo@math.unipd.it (A. Tovo).

http://dx.doi.org/10.1016/j.spasta.2016.02.006
2211-6753/© 2016 Elsevier B.V. All rights reserved.

0. Introduction

Nowadays, a huge quantity of data structured on different time and space scales are easily available. Analysis of these massive databases reveals that despite their diversity and complexity, natural phenomena are characterized by the emergence of regularities that are largely independent of biological and physiological details. One is the tendency, observed both in ecological communities and in human activities, to form spatial or temporal clusters (He et al., 1997; Condit et al., 2000; Plotkin et al., 2000; Adorisio et al., 2014). However, classifying a spatial point pattern as clustered rather than regular can be a challenging task because establishing the main features of its spatial density function strongly depends on the *scale* through which we look at it.

More generally, it is well known that the form of a data-based density function may depend on the algorithm (binning rule) used for the binning of the data (Etienne and Haegeman, 2010). In our view the main flaw of many binning rules (Sturges, 1926; Yule and Kendall, 1950; Doane, 1976; Freedman and Diaconis, 1981; Stone, 1984; Scott, 2015) is that they assume some knowledge on the data distribution. For example Sturges rule (Sturges, 1926) assumes that the data are normally distributed. This is a key point if you have in view applications to ecological datasets. In many cases it is not reasonable to assume such knowledge and the process generating the dataset must be considered unknown. Therefore any criteria based on some prior knowledge of the true density should not be applied as it often introduces a degree of arbitrariness that may produce biased conclusions.

In this paper we are concerned with the statistical analysis of spatial patterns describing the location of plants in a tropical forest study area. We intend to use a method proposed by K. H. Knuth (Knuth, 2013) to find the optimal bin size of a two-dimensional histogram. Knuth non parametric method selects the optimal scale from the data without any assumption on the underlying process that generated the data. We show that the Knuth method can be used to highlight relevant spatial characteristics on the underlying distribution such as space anisotropy and clusterization. We tested it against most currently used (Epanechnikov) kernel method for two-dimensional datasets and one-dimensional (Stone binning rule) and it results to be more efficient in detecting Complete Spatial Random processes and avoiding sample fluctuations. Therefore our analysis validates it as a reliable method for determining the intensity function of a pattern. Additionally, it is not subject to the virtual aggregation phenomenon (see Section 3.3). It correctly detects homogeneity or the presence of a gradient in the density function and the relative difference of the rectangular bin sides is a measure of the anisotropy of the pattern. It also allows to infer quantitative (cluster size) information on both first and second-order statistics. Thus, it is not only a rule to choose the bin size in which to organize the data. Indeed our analysis proves that Knuth bin size is a good indicator of how finely structured is the dataset and that it can be used as a trusted tool for the preliminary statistical analysis of a spatial dataset. We show which are the relevant information contained in the size and the shape of the optimal bin and how they are related to the spatial features of the process/dataset.

We test our findings on an ecological dataset consisting of the spatial coordinates of individuals belonging to 300 different species of plants located in a 50 ha rectangle of the Barro Colorado Island rainforest (BCI). In particular, we study cluster-like structures in plants' arrangement.

The present analysis should help inform future investigations of temporal or spatial features of different complex systems in ecology and human dynamics (Simini et al., 2012; Formentin et al., 2014; Sanli and Lambiotte, 2015).

1. On optimal binning rules

There exist diverse rules to determine the optimal number of bins of a histogram. Some of the most known (Scott, 1979; Freedman and Diaconis, 1981; Stone, 1984; Scott, 2015) rely on the minimization of the L^2 norm between the histogram and the true underlying density on which they assume some prior knowledge. Assuming such prior information is not reasonable for ecological datasets and the process generating the data must be considered unknown. Hence, any criteria determining the optimal bin size based on prior knowledge on the true density should not be applied. Moreover, some methods work well for unimodal densities while they are known to be suboptimal for multimodal ones. In particular, Freedman and Diaconis rule (Freedman and Diaconis, 1981) is not valid for uniform or

Download English Version:

https://daneshyari.com/en/article/1064510

Download Persian Version:

https://daneshyari.com/article/1064510

Daneshyari.com