

An improved detection of clusters in complex ecological systems by using the the Ripley's K -function

Una migliore individuazione della presenza di gruppi in sistemi ecologici complessi mediante la funzione K di Ripley

Giancarlo Ferrara, Mariangela Sciandra, Giada Adelfio¹

Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli",

Università degli Studi di Palermo,

e-mail: ferrara@dssm.unipa.it, sciandra@dssm.unipa.it, adelfio@dssm.unipa.it

Riassunto: L'obiettivo del lavoro consiste nella definizione di un nuovo approccio per l'individuazione di una eventuale struttura di raggruppamento presente nelle unità in assenza di una classificazione iniziale delle stesse. La procedura proposta per la determinazione di tale struttura si avvale dell'utilizzo della funzione K di Ripley, utilizzata in letteratura nell'ambito dell'analisi di dati spaziali. In particolare, viene qui generalizzato l'utilizzo della funzione K , sostituendo lo spazio fisico delle coordinate geografiche con quello definito dalle variabili osservate. Nel lavoro si applica tale metodologia in un sistema ecologico complesso come quello costituito dalle praterie di *Posidonia oceanica*.

Keywords: Cluster analysis, K -function, complex ecological systems

1. Introduction

The goal of clustering methods is to detect homogeneous subsets of units. These methods generally require the definition of a distance function, to evaluate the "similarity" among units, the choice both of a particular partitioning method and of the variables helpful for detecting clusters (Everitt, 2001). The redundancy of variables is sometimes dealt with the use of transformed observed variables by principal components, based on the linearity assumption of the transformations. As an alternative, we suggest to check the presence of a cluster-structure on the space of the observed variables through the Ripley's K -function; then, cluster analysis will be carried out by considering those variables for which the test of homogeneity is rejected. Therefore, our proposal can be considered as a preliminary tool for detecting those variables that will be used in subsequent clustering methods. In the next section an application in a complex ecological system is introduced in order to reduce the wide number of explanatory variables that characterizes such complex systems.

2. The proposed method with application

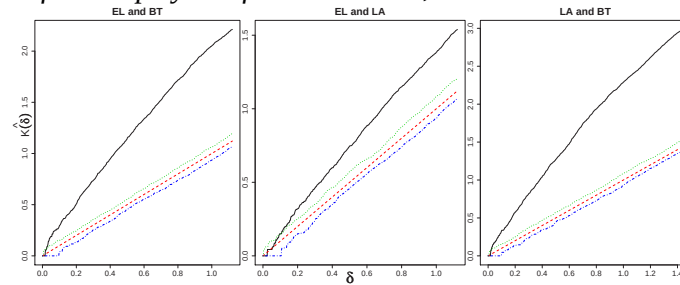
Given a point process on $\mathbb{R}^2$, the Ripley's K -function is defined as the expected number of further events within distance δ of any given point divided by the overall rate λ :

$$K(\delta) = \lambda^{-1} E[\text{n. of extra events within distance } \delta \text{ of an arbitrary event}], \quad \delta \geq 0. \quad (1)$$

¹Address of correspondence: Viale delle Scienze, Edificio 13, 90128 - Palermo (Italy)

The K function, being a measure of the distribution of the inter-point distances (Ripley, 1976), captures the spatial dependence between different regions of a point process. Its estimation is based on replacing the expectation operator in (1) with the sample mean. Let N be a point process defined on $A \subseteq \mathbb{R}^2$ and $\{s_1, \dots, s_n\}$ be n realizations of the process on A . A simple estimator of $K(\delta)$ is: $\hat{K}(\delta) = \hat{\lambda}^{-1} \sum_i \sum_{i \neq j} I(|s_i - s_j| \leq \delta)/n$, where $\hat{\lambda}$ is the estimator of the overall intensity given by $n/|A|$, i.e. the observed number of events per unit of the area $|A|$. Other versions of the previous estimator were proposed (Ripley, 1988), dealing with corrections for boundary effects. Estimates of $K(\delta)$ are usually provided to identify and quantify the apparent deviation from the Poisson randomness ($K(\delta) = \pi\delta^2$), towards clustering, for $K(\delta)$ greater than the corresponding Poisson value, or inhibition, otherwise. To illustrate the proposed procedure, we consider a dataset concerning *Posidonia oceanica* (*P.o.*) meadows (Micheli *et al.*, 2005); $n = 95$ transects of the Sicilian coast were sampled and some morphological variables, such as eroded leaves (%) (EL), brown tissue (%) (BT) and leaf area (LA), were observed for the plants belonging to each transept. The goal of the analysis is to classify each transept with respect to the sea-pollution level. Since *P.o.* can be considered as a sea pollution bio-indicator, it can be used to provide a classification of the strips compared on the basis of the observed variables. A first attempt has been to use some hierarchical clustering methods on all the observed variables. Since this approach did not allow to identify a well defined cluster-structure, we used the proposed clustering method. In particular, the Ripley's K -function is used on the variables space in order to identify those variables for which a cluster-structure can be detected. This procedure have allowed to reduce the dimension of the variables space finding out the confounding variables causing previous unclear classification. For example, Figure 1 displays the fitted K -function on the two-dimensional spaces defined by BT, EL and LA; the biggest slopes observed for BT are a proof of the importance of the same to identify clusters. The use of the K -function is here restricted to two dimensional spaces, although future search will be aimed to its generalization to more dimensional spaces.

Figure 1: The fitted Ripley's K function on EL, BT and LA.



References

- Everitt B. (2001) *Cluster analysis*, Arnold.
- Micheli C., Paganin P., Peirano A., Caye G., Meinesz A. and Bianchi C.N. (2005) Genetic variability of *posidonia oceanica* (L.) delile in relation to local factors and biogeographic patterns, *Aquatic botany*, 82, 3, 210–221.
- Ripley B. (1988) *Statistical inference for spatial processes*, Cambridge University Press.
- Ripley B.D. (1976) The second-order analysis of stationary point processes, *Journal of Applied Probability*, 13, 2, 255–266.