Spatial Statistics 18 (2016) 333–351

Contents lists available at ScienceDirect

Spatial Statistics

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

A hierarchical clustering method for multivariate geostatistical data

Francky Fouedjio

CSIRO Mineral Resources, 26 Dick Perry Avenue, Kensington, WA 6151, Australia

ARTICLE INFO

Article history:
Received 29 September 2015
Accepted 14 July 2016
Available online 1 August 2016

Keywords:
Clustering
Geostatistics
Non-parametric
Multivariate data
Spatial correlation
Spatial contiguity

ABSTRACT

Multivariate geostatistical data have become omnipresent in the geosciences and pose substantial analysis challenges. One of them is the grouping of data locations into spatially contiguous clusters so that data locations within the same cluster are more similar while clusters are different from each other. Spatially contiguous clusters can significantly improve the interpretation that turns the resulting clusters into meaningful geographical subregions. In this paper, we develop an agglomerative hierarchical clustering approach that takes into account the spatial dependency between observations. It relies on a dissimilarity matrix built from a non-parametric kernel estimator of the multivariate spatial dependence structure of data. It integrates existing methods to find the optimal number of clusters and to evaluate the contribution of variables to the clustering. The capability of the proposed approach to provide spatially compact, connected and meaningful clusters is assessed using multivariate synthetic and real datasets. The proposed clustering method gives satisfactory results compared to other similar geostatistical clustering methods.

1. Introduction

Multivariate data indexed by geographical coordinates have become increasingly frequent in the geosciences and pose real analysis challenges. A basic problem is the clustering of observations into spatially contiguous groups so that observations in the same group are similar to each other and

E-mail address: francky.fouedjiokameni@csiro.au.

http://dx.doi.org/10.1016/j.spasta.2016.07.003
2211-6753/

different from those in other groups. Some typical applications in the geosciences are (Schuenemeyer and Drew, 2011): (i) defining climate zones, (ii) determining zones of similar land use, (iii) identifying archaeological sites, (iv) delineating agricultural management areas, and (v) defining ore typologies.

In the non-spatial context, the problem of clustering observations is well-known and described in many textbooks from a descriptive to theoretical viewpoint. There are two principal clustering approaches namely, hierarchical and partitioning (Kaufman and Rousseeuw, 1990; Charu and Chandan, 2013). In the hierarchical approach, a tree-like structure is constructed using agglomerative or divisive procedures. In the partitioning approach, observations are divided into clusters once the number of clusters to be formed is specified. Very often, applied to geostatistical data, these non-spatial clustering algorithms have a tendency to produce spatially scattered clusters. However, this characteristic is undesirable for many geoscience applications (e.g., delineation of agricultural management zones).

In the geostatistical context, a more specific approach is needed. In fact, geostatistical data often show properties of spatial dependency and heterogeneity over the study region. Observations located close to one another in the geographical space might have similar characteristics. Furthermore, the mean, the variance, and the spatial dependence structure can be different from one subregion to another. Hence there is a need to obtain close related or contiguous clusters of data locations with similar attribute values. The clustering can be achieved in different ways, depending mainly on the measure used to quantify proximity among observations. It is important to point out that proximity in the attribute space does not ensure proximity in the geographical space. Thus, in addition to the proximity in the attribute space, proximity in the geographical space must be taken into account. Moreover, data locations belonging to the same cluster should usually be close to one another in the geographical space. To address these constraints, conventional non-spatial clustering approaches have been adapted. Existing approaches can be distinguished into four different categories: (i) non-spatial clustering with geographical coordinates as additional variables, (ii) non-spatial clustering based on a spatial dissimilarity measure, (iii) spatially constrained clustering, and (iv) model-based clustering.

The first category incorporates the spatial information by just considering geographical coordinates as additional variables. In other words, each observation is seen as a point in a dimensional space, including both the geographical space and the attribute space. Thereby non-spatial clustering methods like K-means clustering or agglomerative hierarchical clustering can be applied to this new space. In practice, the resulting clusters provided by this approach look too scattered spatially. Indeed, this approach does not distinguish between the geographical space and the attribute space.

The second category uses existing non-spatial clustering methods by modifying the dissimilarity measure between two observations to take explicitly into account the spatial dependence. Olivier and Webster (1989) were the first to propose an approach of this kind. In the univariate case, they suggested using the stationary variogram of data to weight the original dissimilarities between data locations; whereas in the multivariate case, the stationary variogram of the first principal component of data is employed. In the multivariate case, Bourgault et al. (1992) used the stationary multivariate variogram of data as the weighting function to decrease similarities of distant data locations. In the approaches proposed by Olivier and Webster (1989) and Bourgault et al. (1992), clustering proceeds in two main steps. The first step involves computing dissimilarities between all pairs of sampling locations from attribute values. These dissimilarities are modified by multiplying them by a function of geographical separation to form new dissimilarities. The second step finds latent roots and vectors of the resulting dissimilarity matrix and uses the leading vectors to apply the K-means clustering. Romary et al. (2015) pointed out that these methods have a tendency to produce a smooth dissimilarity matrix without however reinforcing the spatial contiguity between resulting clusters.

The third category is different from the second one in that it considers spatial contiguity constraints rather than spatial dissimilarities. Specifically, data locations are grouped together through a non-spatial clustering technique according to a set of spatial contiguity constraints. In the univariate case, Pawitan and Huang (2003) developed two spatially constrained clustering algorithms, hierarchical and non-hierarchical. The spatial connectivity of resulting clusters is imposed by a graph structuring data locations in the geographical space such as the Delaunay triangulation. However, the lengths of the edges of the graph are not accounted; this might produce spurious results. In the multi-