Computers and Geosciences 120 (2018) 82–96

Contents lists available at ScienceDirect

Computers and Geosciences

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

Research paper

Towards justifying unsupervised stationary decisions for geostatistical modeling: Ensemble spatial and multivariate clustering with geomodeling specific clustering metrics

Ryan Martin*[,1], Jeff Boisvert[2]

6-241 Donadeo Innovation Centre for Engineering, 9211-116 Street, University of Alberta, Edmonton, Alberta, T6G 1H9, Canada

ARTICLE INFO

Keywords:
Estimation domains
Spatial clustering
Ensemble clustering
Stationarity

ABSTRACT

The subdivision of samples into stationary sets is one of the first decisions in a resource modeling workflow where geologically and statistically related samples are grouped for further geostatistical modeling. Unsupervised learning algorithms (clustering) with modifications to consider spatial correlation have several benefits in the decision of stationarity, they are: automatic and repeatable; assess uncertainty; and provide a framework for checking existing groupings. However, subjective parameterization remains a critical limitation for the application of spatial clustering in geostatistical workflows. In this work, two main extensions to the current state of research are proposed: 1) a combined spatial-multivariate metric that describes cluster quality in both multivariate and Cartesian space by measuring multivariate compactness and spatial contiguity; and 2) a novel random-path spatial-multivariate ensemble clustering algorithm to reduce the reliance on subjective clustering parameterization. The metrics developed in this work quantify clustering quality based multivariate and spatial properties. Moreover, the proposed clustering algorithm allows the user to control the spatial contiguity and multivariate compactness of the final clusters by modifying a single parameter. The clustering algorithm is introduced and demonstrated on a synthetic test dataset, then further application to more complex datasets are explored to demonstrate the clustering properties and simplicity of tuning the algorithm to prefer spatially contiguous clusters or multivariate compactness. The proposed algorithm outperforms other clustering algorithms for the datasets tested based on the metrics developed.

1. Introduction

Rock samples collected to characterize natural resources provide an incomplete view of the subsurface since the properties measured from the samples vary spatially and at all scales. A goal of geostatistics is to estimate the distribution of uncertainty of a property of interest (e.g., grade, porosity, value, etc) at an unsampled location given the nearby data. Two important assumptions are often made: 1) the nearby data are related to one another and represent relevant geological processes; and 2) the unsampled locations are also part of these geological processes.

Geostatistical modeling makes various assumptions of stationarity. In a geostatistical context, a stationary RF implies invariance of the first and second order statistics ($\mu(\mathbf{u})$, $\sigma(\mathbf{u})$, $Cov(\mathbf{u})$) for all $\mathbf{u} \in \mathbf{A}$. In the strict sense, geological domains are rarely stationary; a trend model or local variograms are required to satisfy stationarity in the presence of elemental zoning or complex geological structures (Qu and Deutsch, 2017; Boisvert and Deutsch, 2011). However, many non-stationary features can be accounted for by considering sub-setting the geological dataset so that each sub-population is stationary. Typically sub-sets are generated by considering the geological attributes recorded in core logs. Two typical scenarios include: 1) either there are too many distinct geological units defined in the logs and decisions about merging units must be made; or 2) the available logging information does not inform clearly defined stationary populations. In the first case the geostatistician must adopt a merging workflow that simultaneously considers the spatial, multivariate and geological properties to define a smaller number of modeling categories (Rossi and Deutsch, 2014). In the second case, decisions about how to subset the population into distinct spatial-multivariate sets must be made. In the univariate case, grade domains are often considered where subsets of the dataset are

* Corresponding author.
E-mail address: rdm1@ualberta.ca (R. Martin).
[1] Authorship: development of main algorithm implementation, development of dual space-metrics, manuscript writing, case studies.
[2] Authorship: conceptualization of the work, preliminary implementation of algorithm, supervisory guidance, manuscript writing.

https://doi.org/10.1016/j.cageo.2018.08.005
Received 31 December 2017; Received in revised form 7 August 2018; Accepted 21 August 2018
Available online 25 August 2018
0098-3004/

generated based on project specific cutoffs (Leuangthong and Srivastava, 2012) and each subset is modeled independently. Typically the grade variables are lognormally distributed and the heteroscedastic nature of the variable can be used to justify this style of subdivision (Manchuk et al., 2009).

Cluster analysis is a popular technique used to learn the dominant groups present in a given dataset and includes algorithms such as K-means or hierarchical clustering. Consider a geostatistical dataset consisting of surface samples, drill cores or wells. At each location, M continuous and/or discreet properties are recorded, such as grain size, rock type, alteration type and intensity, geophysical properties, or measurements of the values of interest with whole rock lithogeochemical analyses. Clustering algorithms partition the M dimensional attribute space, i.e. hyper-space, so that samples within a partition are highly related and different from samples in other partitions. Traditional clustering methods are poorly suited to partitioning geostatistical domains because they do not consider the spatially correlated nature of the samples.

Methods to consider the spatial correlation between samples in clustering algorithms have been developed (Oliver and Webster, 1989; Ambroise and Govaert, 1998; Scrucca, 2005; Fouedjio,; Chautru et al., 2016; Romary et al., 2015). Although several methods exist to treat the spatial correlation of the samples, all algorithms utilize a single parameterization that is subjectively chosen by the practitioner to generate classes that are subjectively assessed as reasonable. All truly unsupervised classification algorithms require some domain knowledge to justify the results and ensure the resulting classes are reasonable (Strehl and Ghosh, 2002). Romary et al. (2015) note that expert domain knowledge should guide the parameter inference of their classification algorithm so that the resulting classes are reasonable and better suited to the geostatistical workflows that follow. However the results of two differently parameterized clustering runs may be significantly different and, to our knowledge, objective measures of clustering configurations for geostatistical applications have not been developed.

Two main contributions of this work are (1) assess the results of spatial clustering and (2) reduce reliance on subjective parameterizations for spatial clustering of geostatistical domains. First, a novel random-path spatial clustering algorithm with dual-space search is developed. The algorithm implements consensus clustering where many individual clusterings of the data are combined to generate the most likely clustering (Strehl and Ghosh, 2002; Topchy et al., 2005). In this work, random-path and iterative dual-space merging generate individual clusterings, and the results are processed using an ensemble post-processing workflow.

Secondly, validation of spatial clustering techniques is an outstanding issue. Typically a clustering result is validated either with internal or external measures: e.g., measuring the compactness of the configuration or the error in predicting a set of known classes (Halkidi et al., 2001; Strehl and Ghosh, 2002; Tibshirani and Walther, 2005). In a geostatistical context the true labels are not known. Instead, a set of metrics that quantify the desirable properties of a set of spatial-multivariate classes for the geostatistical workflow are developed. These metrics permit objective comparisons between different spatial-multivariate clusterings of the same dataset.

The rest of the paper is organized as follows: Section 2 reviews existing methods for spatial clustering. Section 3 discusses methods for validation and develops novel spatial-multivariate metrics. Section 4 outlines the details of the proposed algorithm for finding dominant spatial classes in geostatistical datasets, and demonstrates the algorithm and parameter choices on a simple synthetic dataset. Finally, the results of considering this style of spatial clustering are demonstrated on a 2D and 3D dataset in Section 5.

2. Background

Stationarity is rarely a property of geological domains, rather it entails a series of decisions made by the geostatistician to subdivide samples into groups and/or choose how to model the non-stationary features present in the modeling domains. Numerous techniques are available to model geostatistical realizations of a non-stationary RF, such as modeling with a trend or locally varying anisotropy (Boisvert and Deutsch, 2011; Fouedjio, 2016b; Rossi and Deutsch, 2014; Qu and Deutsch, 2017). Better decisions of stationarity lead to better resource estimates since the statistical homogeneity of the domains is improved and the parameters inferred for modeling in each domain are more representative of each sub-population (Rossi and Deutsch, 2014).

2.1. Clustering and spatial clustering

Clustering algorithms partition a set of N samples of M variables into mutually exclusive groups based on the similarity of samples. Individual clustering algorithms typically differ in the form of the clusters found, the types of (dis)similarity metrics used, and the discreet or fuzzy partitioning of the samples (Ester et al., 1996; Jain, 2010; Sander and Ester, 1998). Clustering algorithms are well suited to partitioning M dimensional attribute space, however applications to geostatistical datasets are limited because the spatial correlation and geological properties of the samples are not considered (Rossi and Deutsch, 2014).

Spatial clustering describes methods developed to address spatial correlation between samples taken at a set of points in Cartesian space (Oliver and Webster, 1989; Ambroise and Govaert, 1998; Scrucca, 2005). The goal is to generate classes that are spatially contiguous and have distinct multivariate properties. Two general strategies have been applied: 1) some form of neighborhood constraint to modify relatedness of distant and uncorrelated samples (Oliver and Webster, 1989; Ambroise and Govaert, 1998; Fouedjio,; Romary et al., 2015); or 2) generating a secondary dataset calculated from the original data with local autocorrelation statistics (Scrucca, 2005). Oliver and Webster (1989) justified the variogram model as a method to increase the relatedness of nearby points in clustering. In their work the spatial compactness of clusters was tuned by modifying the range and shape of the variogram model. Fouedjio (Fouedjio,) extends the kernel-based modifier to the multivariate case by incorporating the direct and cross-variogram measures to modify the dissimilarity matrix. The methodologies from Ambroise and Govaert (1998) and [Romary et al. (2012), (Romary et al., 2015)] introduce spatial constraints to conventional clustering to ensure only samples in related neighborhoods are paired.

The other method for generating spatially connected multivariate clusters uses local autocorrelation statistics (Scrucca, 2005; Ord and Getis, 1995). The first step is to generate a new dataset by calculating local autocorrelation measures for each variable given the nearby data in the cartesian space [e.g., Morans, Getis local autocorrelation; 8]. The new variable at each location captures the magnitude and sign of the original value as well as the relatedness to nearby points. The new dataset is then used in conventional clustering algorithms, such as K-means, to determine the classes that have both spatial and multivariate significance.

However, common to all available clustering and spatial clustering algorithms is the issue of parameterization and validation; the number of clusters, the method of integrating spatial correlation, how the related spatial neighborhood is determined, the maximum distance of ‘relatedness’, the form of the clusters, etc, must be chosen and justified with expert knowledge or somehow compared against known classes in the dataset. Although statistical techniques can be used to determine values for some parameters [e.g., the number of clusters; (Tibshirani and Walther, 2005), (Tibshirani et al., 2001)], others decisions are more subjective (e.g., which method to integrate spatial information). These are not problems with spatial clustering alone, instead these are general issues for implementing clustering as the algorithms cannot be applied in a truly unsupervised sense; some domain knowledge is