Computers & Geosciences 103 (2017) 173–182

Contents lists available at ScienceDirect

Computers & Geosciences

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

Research paper

A machine learning approach to the potential-field method for implicit modeling of geological structures

Ítalo Gomes Gonçalves*,[1], Sissa Kumaira, Felipe Guadagnin

Campus Caçapava do Sul, Universidade Federal do Pampa, Brazil

ARTICLE INFO

Keywords:
Implicit modeling
Machine learning
Potential field
Kriging
3D geological modeling
Compositional data analysis

ABSTRACT

Implicit modeling has experienced a rise in popularity over the last decade due to its advantages in terms of speed and reproducibility in comparison with manual digitization of geological structures. The potential-field method consists in interpolating a scalar function that indicates to which side of a geological boundary a given point belongs to, based on cokriging of point data and structural orientations. This work proposes a vector potential-field solution from a machine learning perspective, recasting the problem as multi-class classification, which alleviates some of the original method's assumptions. The potentials related to each geological class are interpreted in a compositional data framework. Variogram modeling is avoided through the use of maximum likelihood to train the model, and an uncertainty measure is introduced. The methodology was applied to the modeling of a sample dataset provided with the software Move™. The calculations were implemented in the R language and 3D visualizations were prepared with the `rgl` package.

1. Introduction

Implicit methods for 3D geological modeling have risen in popularity over the last decade (Calcagno et al., 2008; Maxelon et al., 2009; Caumon et al., 2013; Hillier et al., 2013; Jessell et al., 2014; Vollgger et al., 2015; Wu et al., 2015). Their advantages over explicit models include reproducibility, automation, easy model update with newly acquired information, minimal user-induced bias, and the straightforward incorporation of multi source information (Vollgger et al., 2015). As stated by McLennan and Deutsch (2006), a good methodology for implicit modeling should be simple, realistic, and provide some measure of uncertainty in its results.

Geostatistics is the technique of choice for modeling spatial variations of properties in geoscientific related problems (Chilès and Delfiner, 1999; Goovaerts, 1997; Isaaks and Srivastava, 1989). In geological and/or structural modeling, different forms of kriging have been used to model geological surfaces (Carr et al., 2001; Cowan et al., 2003, 2004; Vollgger et al., 2015), orientation data (Gumiaux et al., 2003), and geological surfaces constrained by orientation data (Aug, 2004; Calcagno et al., 2008; Chilès et al., 2004; Lajaunie et al., 1997). The latter was termed as "potential-field method". It is possible to find implementations of these methods in commercial software.

Machine learning comprises a set of statistical techniques that have a wide range of applications, such as spam detection, handwritten text and speech recognition, and recommender systems. As defined by Flach (2012), machine learning is "*the study of algorithms and systems that improve their knowledge or performance with experience*", with the "experience" coming from the available data. In other words, a machine learning algorithm effectively "learns" how to perform a particular task. Recently there have been works that apply machine learning methods in the geosciences, such as decision trees to aid in mineral prospection (Rodriguez-Galiano et al., 2015) and geological mapping (Cracknell and Reading, 2014), support vector machines for geological modeling (Smirnoff et al., 2008; Wang et al., 2014), clustering methods to aid standard geostatistics (Kapageridis, 2014) and identify homogeneous domains in wide areas (Romary et al., 2014), and attempts to bridge the gap between machine learning and geostatistics (Hristopulos, 2015). Furthermore, kriging itself is a machine learning technique widely used for classification and regression tasks (Rasmussen and Williams, 2006) and function optimization (Chan, 2010) in any number of dimensions, although it is referred to as Gaussian Process in the machine learning literature (Rasmussen and Williams, 2006).

In order to further promote the application of machine learning in the geosciences, the present work approaches the original potential-field method by Lajaunie et al. (1997) from another perspective, recasting it as a multi-class classification problem. Classification is done within the compositional data framework from Tolosana-Delgado

* Correspondence to: Av. Pedro Anunciação, 111, Caçapava do Sul, RS, Brazil.
E-mail address: italogoncalves@unipampa.edu.br (Í.G. Gonçalves).
[1] Source code is available on https://github.com/italo-goncalves/geomod3D.git

http://dx.doi.org/10.1016/j.cageo.2017.03.015
Received 3 October 2016; Accepted 21 March 2017
Available online 22 March 2017

et al. (2008) with the probabilistic treatment by Rasmussen and Williams (2006). The covariance function parameters are inferred through maximization of the log-likelihood (Mardia and Marshall, 1984), eliminating the need for complicated manual variography of orientation data (Chilès et al., 2004; Calcagno et al., 2008). The model also does not depend on structural data to work and does not make assumptions on the structures' polarity (*i.e.* the younging direction). The algorithms were implemented in the R language (Core Team, 2017) and tested on a dataset contained within the software Move™ (Midland Valley Exploration, 2014). The article is structured in the following manner: Sections 2 and 3 lay out the theory and methodology; Section 4 presents a case study; in Section 5 the results, strengths and limitations of the method are discussed; and Section 6 presents the conclusions and suggestions for future work.

2. Problem statement

Consider a point in space with coordinates $\mathbf{x} = (x, y, z)^T$ and geological class label given by $L(\mathbf{x}) = c$, $c \in 1, 2, \ldots, C$, where C is the number of geological classes involved in the problem (geological formations, domains, lithologies, *etc.*). Associated with each class there is an unknown probability $\pi_c(\mathbf{x}) = \Pr(L(\mathbf{x}) = c)$, as well as an extra probability $\pi_{C+1}(\mathbf{x})$ related to an "unknown" class (its role in the assessment of model uncertainty shall be explained later). Compositional data theory (Pawlowsky-Glahn and Buccianti, 2011; Buccianti et al., 2006) defines the simplex, a C-dimensional subspace embedded in a $(C + 1)$-dimensional space in which the probabilities are contained. This is so due to the natural constraint that $\sum_{c=1}^{C+1} \pi_c(\mathbf{x}) = 1$ and $\pi_c > 0 \ \forall \ c$. The theory states that the best way to work with data constrained in this way is through log-odds. One way of doing so is the central log-ratio transformation (clr):

$$\phi_c(\mathbf{x}) = \ln \pi_c(\mathbf{x}) - \frac{1}{C+1} \sum_{i=1}^{C+1} \ln \pi_i(\mathbf{x}) \tag{1}$$

where $\phi_c(\mathbf{x})$ is the log-transformed value and the second term in the right side corresponds to the logarithm of the geometric mean of the probabilities. This transformation has the property that $\sum_{c=1}^{C+1} \phi_c(\mathbf{x}) = 0$. The probabilities can be recovered with a back-transform, also known in the machine learning community as the softmax function (Bishop, 2006; Flach, 2012):

$$\pi_c(\mathbf{x}) = \frac{\exp\phi_c(\mathbf{x})}{\sum_{c=1}^{C+1} \exp\phi_c(\mathbf{x})} \tag{2}$$

The log-transformed probabilities are usually called coordinates, due to the vector space structure of the simplex. Here they are denominated *potential components* in order to establish the link with the potential field from Lajaunie et al. (1997). In that work the authors define the potential as a scalar field, with the geological surfaces of interest being modeled as different isovalues in that field. Here the potential field is defined as a vector field $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_C(\mathbf{x}))^T$, and the potential components $\phi_c(\mathbf{x})$ are all linked due to their compositional origin. Implicit modeling amounts to estimating $\boldsymbol{\phi}$ over the region of interest, the details of which are given in the next section.

2.1 Incorporating structural data

The modeled geological bodies are expected to conform to the structural measurements obtained in the field. In order for the model to exhibit this behavior, the iso-potential surfaces must pass tangentially to the measured orientation lines and planes. Given a point s at position $\mathbf{x}_s$, $s \in 1, 2, \ldots, S$, with an associated structural direction $\widehat{\mathbf{d}}_s$ (a unit vector), this tangent constraint implies that

$$\frac{\partial\phi_c(\mathbf{x}_s)}{\partial\widehat{\mathbf{d}}_s} = \langle\nabla\phi_c(\mathbf{x}_s), \widehat{\mathbf{d}}_s\rangle = 0, \quad c \in 1, 2, \ldots, C \tag{3}$$

where $\frac{\partial\phi_c(\mathbf{x}_s)}{\partial\widehat{\mathbf{d}}_s}$ is the directional derivative of ϕ_c at $\mathbf{x}_s$ along $\widehat{\mathbf{d}}_s$, $\nabla\phi_c(\mathbf{x}_s)$ is its gradient, and $\langle\cdot, \cdot\rangle$ is the scalar product. In other words, the gradient of every potential component must be orthogonal to the measured directions. If a point belongs to a structural plane there are two structural directions associated with it, one along dip and the other along strike. Note that, unlike the original formulation from Lajaunie et al. (1997), for this method the structural measurements are completely optional, although they represent an important contribution to the final model. Furthermore, due to the way that the potential field is defined here, there is no need to assume an arbitrary modulus for the gradient or to state its structural polarity.

3. Problem formulation

3.1. Kriging estimation of the potential field

Rasmussen and Williams (2006) define the Gaussian Process (GP) as "*a collection of random variables, any finite number of which have a joint Gaussian distribution*". It can be seen as a distribution over functions, specified by a mean function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{y})$, with $\mathbf{x}$ and $\mathbf{y}$ being points in space. It is assumed that the potential components defined above are distributed as GPs:

$$\phi_c(\mathbf{x}) \sim \mathcal{GP}(m_c(\mathbf{x}), k_c(\mathbf{x}, \mathbf{y})), \quad c \in 1, 2, \ldots, C \tag{4}$$

In order to apply this model to the geological data, one must deal with the matter of how to assign potential vectors to the data points, as no meaningful probability can be derived for them. All one knows is that a point is either inside a geological class or in a boundary between two classes, which only indicates the dominant class (*i.e.* the one with highest potential component and probability) but gives no hint as to its absolute value. Note that simply assigning 0/1 values for the probabilities is not possible, as this would result in an infinite potential. Tolosana-Delgado et al. (2008) deal with this by assigning an arbitrary probability $1 - b$ to a point's true class and $\frac{b}{C}$ to the others, with $0 < b < 0.5$. It is shown that this results in a constant scale factor β in log-transformed space. In order to stay aligned with the definition of the GP given above, the approach proposed here is to assume the potential components are independent, normally distributed random variables with variance ${\sigma_0}^2$. The mean potential component $\phi_c(\mathbf{x}_p)$ at a data point p, $p \in 1, 2, \ldots, P$, is then given by

$$\phi_c(\mathbf{x}_p) = \begin{cases} 1, & \text{if } \mathbf{x}_p \text{ belongs to class } c \\ -\frac{1}{C}, & \text{if } \mathbf{x}_p \text{does not belong to class } c \\ \frac{1-\frac{1}{C}}{2} = \frac{C-1}{2C}, & \text{if } \mathbf{x}_p \text{lies at the boundary between } c \text{ and another class} \end{cases} \tag{5}$$

The potential component of the unknown class is calculated as $\phi_{C+1}(\mathbf{x}) = -\sum_{c=1}^{C} \phi_c(\mathbf{x})$ in order to respect the zero-sum property of clr-transformed variables. Due to the uncertainty over the potential components, back-transforming through (2) is no longer correct. Instead, one should estimate the proportion of the multivariate normal probability density that favors each class, which can be done through simulation (Rasmussen and Williams, 2006). However, the quantities of interest here are the positions of the geological boundaries, so the calculation of probabilities is a secondary issue.

Once P data points and S structural points (hereafter called training points) are observed and Q test points (where the potential is to be estimated) are defined, conditioning the joint distribution of the potential components $\boldsymbol{\phi}_{cP}$ and $\boldsymbol{\phi}_{cQ}$ and directional derivatives of a given class on the known data yields (Rasmussen and Williams, 2006;