Computers & Geosciences 106 (2017) 33–48

Contents lists available at ScienceDirect

Computers & Geosciences

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

Research paper

Porosity estimation by semi-supervised learning with sparsely available labeled samples

Luiz Alberto Lima[a,b,*,1], Nico Görnitz[c,**,1], Luiz Eduardo Varella[a], Marley Vellasco[b], Klaus-Robert Müller[c,d,e,***], Shinichi Nakajima[c,d]

[a] Petrobras, 20031-170 Rio de Janeiro, Brazil
[b] Pontifícia Univ. Católica do Rio de Janeiro, 22543-900 Rio de Janeiro, Brazil
[c] Berlin Institute of Technology, Machine Learning Group, 10587 Berlin, Germany
[d] Berlin Big Data Center, 10587 Berlin, Germany
[e] Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 136-713, Republic of Korea

ARTICLE INFO

Keywords:
Porosity estimation
Facies classification
Latent variable
Conditional random fields
Ridge regression

ABSTRACT

This paper addresses the porosity estimation problem from seismic impedance volumes and porosity samples located in a small group of exploratory wells. Regression methods, trained on the impedance as inputs and the porosity as output labels, generally suffer from extremely expensive (and hence sparsely available) porosity samples. To optimally make use of the valuable porosity data, a semi-supervised machine learning method was proposed, *Transductive Conditional Random Field Regression* (TCRFR), showing good performance (Görnitz et al., 2017). TCRFR, however, still requires more labeled data than those usually available, which creates a gap when applying the method to the porosity estimation problem in realistic situations. In this paper, we aim to fill this gap by introducing two graph-based preprocessing techniques, which adapt the original TCRFR for extremely weakly supervised scenarios. Our new method outperforms the previous automatic estimation methods on synthetic data and provides a comparable result to the manual labored, time-consuming geostatistics approach on real data, proving its potential as a practical industrial tool.

1. Introduction

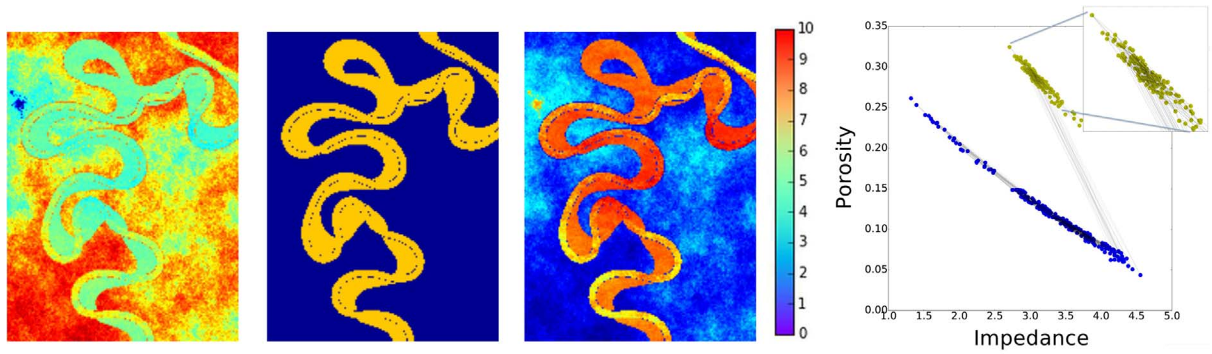
Porosity, the fraction of void space over the total rock volume, is a key indicator for existence of a petroleum reservoir—void space can store hydrocarbons (Schlumberger, 2015).

Porosity can be directly measured at wells once they are drilled but, because of drilling costs, it is typically estimated from indirect sources like seismic impedances obtained from reflections of sonic waves. Fig. 1 illustrates the porosity estimation problem, adapted from Castro et al. (2005). The following three facts make accurate porosity estimation a hard task:

1. **Hidden structure governs the regression relationship:** porosity estimation typically relies on the inverse correlation between seismic impedance and porosity. However, the correlation coefficients and offsets heavily depend on the sedimentary discontinuities provided by distinct geological *facies*. It is known that porosity usually averages linearly and has low variability within each facies (Deutsch, 2002). Therefore, once the facies structure is known, porosities can be estimated from impedances by simple linear regression methods. Nevertheless, facies estimation is an intricate task, due to the many complex geometric shapes that can co-exist in the reservoir.
2. **Seismic impedance alone is not informative for facies estimation:** one might hope that facies can be estimated from the seismic impedance alone. The marginal distribution of the impedance, however, does not give sufficient information for estimating facies. This is illustrated in Fig. 1(d). Each point indicates the impedance and the porosity at a location, and the color indicates the facies (the lines connect neighboring locations). If we have no information on the porosity, we have to estimate the facies only from the impedance (x-axis), which is not very accurate due to the overlapping marginal distribution of the impedance between two facies.

* Corresponding author at: Petrobras, 20031-170 Rio de Janeiro, Brazil.
** Corresponding author.
*** Corresponding author at: Berlin Institute of Technology, Machine Learning Group, 10587 Berlin, Germany.
E-mail addresses: lual@petrobras.com.br (L.A. Lima), nico.goernitz@tu-berlin.de (N. Görnitz), klaus-robert.mueller@tu-berlin.de (K.-R. Müller).
[1] Authors contributed equally.

http://dx.doi.org/10.1016/j.cageo.2017.05.004
Received 2 September 2016; Received in revised form 3 May 2017; Accepted 6 May 2017
Available online 27 May 2017
0098-3004/



(a) Impedance (b) Facies (c) Porosity (d) Impedance vs. porosity

Fig. 1. Porosity estimation problem. The goal is to estimate (c) porosity (unknown at most of the locations) from (a) impedance (known) by using a linear relationship between them. However, this relationship depends on the (b) facies (unknown), and accurate facies estimation requires porosity measurements because of the overlapped marginal distribution of the impedance (d). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

3. **Lack of labeled samples:** the measurements of the porosity at wells in the reservoir are used as labeled data, with which the regression model is trained. If those labeled data are available densely enough to capture the hidden facies structure, we can still estimate porosity by using *local* regression models. However, drilling a well is extremely costly and typically conducted only at the locations where a petroleum reservoir is highly likely to exist.[2] Thus, labeled samples are typically available only for a small number of locations.

As a result, standard geostatistics approaches (Deutsch and Journel, 1998; Dubrule, 2003; Caers, 2005; Larsen et al., 2006) are manual labor, time-consuming processes, demanding considerable expert knowledge during design parameterization.

Fig. 1(d), on the other hand, also implies some hope to achieve accurate porosity estimation. First, there is a clear separation between the two facies in the joint space of impedance and porosity, i.e., the joint distribution is not overlapping.[3] Second, the edges between neighboring locations are sparse between the two facies, while dense in each facies category, i.e., facies tend to be the same in neighboring locations, as we can also observe in Fig. 1(b). These facts imply that we could perform porosity estimation by optimally using the sparsely available porosity information and propagating this information based on the neighborhood spatial structure.

Motivated by this observation, recently a semi-supervised structural learning technique, called Transductive Conditional Random Field Regression (TCRFR) (Görnitz et al., 2017), was proposed. TCRFR is an extension of Conditional Random Field (CRF), a popular graph-based machine learning techniques where known (or observed) and unknown variables are expressed as nodes, and their probabilistic dependencies are expressed as edges (a short introduction of CRF is given in Appendix B). TCRFR can be used to estimate porosity from impedance on seismic volumes conditioned on the porosity values from the available wells in the reservoir. The method is able to infer the hidden or *latent* states of geological facies by combining the local, labeled and accurate porosity information in those wells with the plentiful but imprecise impedance information available everywhere in the reservoir volume. That accurate information is propagated in the reservoir based on conditional random field probabilistic graphical models. The original TCRFR, applied to 2D time slices, presented a good performance with 5% $(impedance, porosity)$ pairs of labeled data. Although accurate estimation from only 5% labeled data is a notable achievement in machine learning, it is still a large number in a real porosity estimation setting scenario, where only a few wells are typically available in the reservoir. In this paper, we tackle the problem of porosity estimation under realistic scenarios by refining and specializing the original TCRFR method. More specifically, we introduce two additional techniques, mainly inspired from the image processing literature, to enhance the performance of TCRFR. The first one is an extension of the original graph-based image segmentation method proposed in Felzenszwalb and Huttenlocher (2004), using its result to determine the neighboring graph structure. In other words, we use the impedance spatial structure to determine how the label information should propagate through the graph.

The second technique relies upon manual annotation of facies categories. This procedure is based on a common assumption in image segmentation, i.e., there are pixels that can be easily labeled by hand for annotators (Boykov and Jolly, 2001). For example, annotating pixels for the shale facies (blue colored in Fig. 1(b)) which are far from the sand facies (yellow colored in Fig. 1(b)) is relatively easy for geologists, and from this process we can establish a practical semi-automatic porosity estimation. Additionally, we extend the original TCRFR method to allow it to work with the 3D segmented data and manually fixed facies.

Note that prediction of porosity and other reservoir variables has also been addressed in several geophysics applications that, e.g., combine rock physics models with seismic inversion. Rock physics fundamentals are described in Mukerji et al. (2001a, 2001b), Doyen (2007), Mavko et al. (2009), Avseth et al. (2010). Petrophysical seismic inversion formulations are depicted in Mukerji et al. (2001a, 2001b), Gunning and Glinsky (2004), Eidsvik et al. (2004), Spikes et al. (2007), Connolly and Hughes (2016). Gaussian mixture models for estimation of reservoir variables from seismic inversion and rock physics is presented in Grana and Rossa (2010). Lithology and fluid prediction classification based on Markov chain models are described in Eidsvik et al. (2002), Larsen et al. (2006). Also, joint inversion approaches for lithology and elastic properties have been proposed by Sams et al., Doyen (2007), among others. In this paper, we focus on porosity estimation from already inverted seismic impedance volumes and sparse porosity samples located in a few exploratory wells, a typical problem faced by geologists during the evaluation of a reservoir in the exploration phase. Compared to the previous approaches, the proposed method automates porosity prediction and facies classification, learning the model directly from the available data.

[2] This tendency of well locations can induce a bias (Deutsch, 2002)—the labeled data are usually available only in high porosity regions, which results in biased statistics of observed rock properties. However, the bias is not extreme if porosity samples are available at regular intervals along the wells, which typically goes through low porosity areas. Further improvement by adapting for this issue, called in statistics *covariate shift adaptation* (Shimodaira, 2000; Sugiyama et al., 2007), is left as future work.

[3] In real data, such clear separation is not always observed, but, in general, separation is much easier in the joint space.