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/

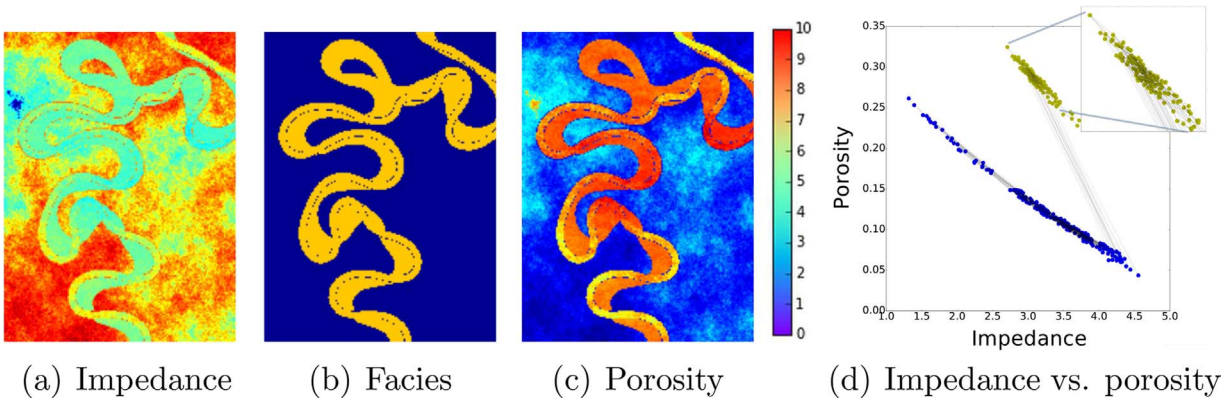


Fig. 1. Porosity estimation problem. The goal is to estimate (c) porosity (unknown at most of the locations) from (a) impedance (known) by using a linear relationship between them. However, this relationship depends on the (b) facies (unknown), and accurate facies estimation requires porosity measurements 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.

We show good performance of our proposed method on synthetic data and real data. In the synthetic data experiment, we compare the proposed method with baseline automatic estimation methods including the original TCRFR, and show preferable performance of our new method. In the real data experiment, we adopt the geostatistics result as a reference and show that our proposed method gives comparable result to the geostatistics one, while the original TCRFR fails. This result implies the possibility of our new method being a versatile (semi-)automatic alternative to the manual labored, time-consuming geostatistics approach.

Starting from a brief description of TCRFR in Section 2, we introduce our new approach in Section 3 and show experimental results in Section 4. We conclude the paper in Section 5.

2. Preliminaries

In this section, we briefly describe the background of the TCRFR model, which is the baseline of our novel proposed algorithm. Assume that, in a volume, we are given impedance observations $\boldsymbol{x} \in \mathbb{R}^D$ at all voxels (locations), and porosity observations $y \in \mathbb{R}$ at some voxels. We call the voxels with a porosity observation *labeled samples*, and the voxels without porosity observation *unlabeled samples*. We index the samples so that the labeled samples come first, namely, we are given a labeled sample set $\mathcal{S} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^D \times \mathbb{R}\}_{i=1}^{n}$ and an unlabeled sample set $\mathcal{U} = \{\boldsymbol{x}_i \in \mathbb{R}^D\}_{i=n+1}^{n+m}$. Typically, $n \ll m$ holds. Our goal is to infer porosities y for all unlabeled samples $\mathcal{U}$.

2.1. Ridge regression

A regression model can be trained on the labeled set and gives the porosity prediction for the unlabeled set. Assume that the porosity can be written as the sum of a parameterized function $f(\boldsymbol{x}; \boldsymbol{w})$ of impedance and an additive Gaussian observation noise:

$$y = f(\boldsymbol{x}; \boldsymbol{w}) + \epsilon, \quad \epsilon = y - f(\boldsymbol{x}; \boldsymbol{w}) \sim \mathcal{N}(0, \sigma^2),$$
$$p(y|\boldsymbol{x}, \boldsymbol{w}) \propto \exp\left(-\frac{1}{2\sigma^2}|y - f(\boldsymbol{x}; \boldsymbol{w})|^2\right),$$

where $f(\boldsymbol{x}; \boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ ($\langle \cdot, \cdot \rangle$ denotes the inner-product of vectors) is a linear regression function with an unknown parameter $\boldsymbol{w} \in \mathbb{R}^D$ that we estimate by using solely the labeled sample set $\mathcal{S}$. To avoid over fitting, we assume the Gaussian prior for $\boldsymbol{w}$: $p(\boldsymbol{w}) \propto \exp(-\frac{\lambda'}{2} \| \boldsymbol{w} \|^2)$. Then, the *maximum a posteriori (MAP)* estimator is obtained by maximizing the joint distribution of $\{y_i\}_{i=1}^n$ and $\boldsymbol{w}$:

$$\max_{\boldsymbol{w} \in \mathbb{R}^D} p(\{y_i\}_{i=1}^n | \{\boldsymbol{x}_i\}_{i=1}^n, \boldsymbol{w}) p(\boldsymbol{w}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \boldsymbol{w}) p(\boldsymbol{w}), \tag{1}$$

or, equivalently, minimizing the negative logarithm of the joint distribution $\min_{\boldsymbol{w} \in \mathbb{R}^D} \mathcal{L}_0(\boldsymbol{w})$, where

$$\mathcal{L}_0(\boldsymbol{w}) = \lambda' \| \boldsymbol{w} \|_2^2 + \sum_i \frac{|y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle|^2}{\sigma^2}. \tag{2}$$

The standard ridge regression, which is used in many inverse problems, minimizes Eq. (2). Once the optimal parameter vector $\boldsymbol{w}^*$ is obtained, the porosities for the unlabeled set $\mathcal{U}$ can be predicted by applying the regression function $y = f(\boldsymbol{x}; \boldsymbol{w}^*)$ for $\boldsymbol{x} \in \mathcal{U}$. In the context of regression, $\boldsymbol{x}$ is called the input, and y is called the output or regression target.

While ridge regression has been proven useful in dozens of applications, it still suffers from severe drawbacks that are likely to deteriorate the prediction accuracy in our setting. Namely, (a) ridge regression assumes that data is IID.[4] This allows us to compute the joint distribution over all samples by just multiplying the density functions at the samples, as in Eq. (1). Such samples are called IID. hence no spatial connections between data points are considered; and (b) the linear dependency assumption between input and output (e.g. impedances and porosities) only holds within the same facies, hence the resulting linear model by ridge regression will not be accurate in the presence of multiple facies.

[4] In the typical setting of regression, samples are assumed to be *independently and identically distributed (IID)*.

2.2. Transductive conditional random field regression

To overcome the restrictions of standard ridge regression, we need to take spatial relations between variables into account (i.e., nearby samples tend to belong to the same facies, and therefore the same impedance-porosity relation is applied), while inferring the type of facies π_i at each location i.

Conditional Random Fields (CRFs) (Lafferty et al., 2001) are probabilistic graph-based models used for labeling and segmenting structured data, such as sequences, trees, and lattices (Blei et al., 2004). CRFs are used to infer latent states π of given observations $\boldsymbol{x}$ with known dependency structure, i.e., $\pi = \mathrm{argmax}_{\hat{\pi}} \log p(\hat{\pi}|\boldsymbol{x}; \boldsymbol{v})$. Their excellent performance has been noted in many important applications, e.g., object classification in an image (He et al., 2004), natural language processing (Taskar et al.), and gene finding (Zeller et al., 2013). Here, $\boldsymbol{v}$ is a model parameter that is learned by either maximum likelihood or maximum a posteriori using multiple examples $\boldsymbol{x}$ with corresponding *known* latent states π. Since we assume latent states (i.e. facies) to be unknown in advance, direct application of CRFs to our problem setting is prohibitive.

For further information on CRFs, including parameter estimation and applications, we refer to a broad literature (see e.g. Blei et al., 2004; Lafferty et al., 2001; Sutton and Mccallum, 2010).

Fortunately, recently proposed Transductive Conditional Random Field Regression (TCRFR) combines the advantages of ridge regression and CRFs *without* the need of knowing facies in advance. It can be viewed as an ordinary ridge regression where the input-output (impedance-porosity) relations are coupled through latent variables (facies) π_i for each location i. Fig 2 illustrates the TCRFR model. Note that the gray-shaded nodes are observed, while the non-shaded and the orange-colored nodes are unknown and need to be estimated.

Let the regression model parameter be $\boldsymbol{u}$ and the CRF model parameter be $\boldsymbol{v}$. Then the TCRFR model is derived as (starting from the ridge regression model Eq. (1))

$$\begin{aligned}
&\max_{\boldsymbol{u}} p(\{y_i\}_{i=1}^n | \{\boldsymbol{x}_i\}_{i=1}^{n+m}, \boldsymbol{u}) p(\boldsymbol{u}) \\
&\geq \max_{\boldsymbol{u}, \boldsymbol{v}, \{\pi_i\}_{i=1}^{n+m}} p(\{y_i\}_{i=1}^n, \{\pi_i\}_{i=1}^{n+m}, \boldsymbol{v} | \{\boldsymbol{x}_i\}_{i=1}^{n+m}, \boldsymbol{u}) p(\boldsymbol{u}) \\
&= \max_{\boldsymbol{u}, \boldsymbol{v}, \{\pi_i\}_{i=1}^{n+m}} \prod_{i=1}^{n} p(y_i|\pi_i, \boldsymbol{x}_i, \boldsymbol{u}) p(\boldsymbol{u}) p(\{\pi_i\}_{i=1}^{n+m} | \{\boldsymbol{x}_i\}_{i=1}^{n+m}, \boldsymbol{v}) p(\boldsymbol{v}).
\end{aligned}$$

Choosing the Gaussian priors for $\boldsymbol{u}$ and $\boldsymbol{v}$ and taking the negative logarithm, we arrive at the final optimization problem:

$$\min_{\boldsymbol{u}, \boldsymbol{v}, \{\pi_i\}_{i=1}^{n+m}} \mathcal{L}(\boldsymbol{u}, \boldsymbol{v}, \{\pi_i\}_{i=1}^{n+m}), \tag{3}$$

where $\mathcal{L}(\boldsymbol{u}, \boldsymbol{v}, \{\pi_i\}_{i=1}^{n+m})$ is a convex combination of the objectives of the regression model and the conditional random field, i.e.,

$$\mathcal{L}(\boldsymbol{u}, \boldsymbol{v}, \{\pi_i\}_{i=1}^{n+m}) = \theta \mathcal{L}_{\mathrm{rr}}(\boldsymbol{u}, \{\pi_i\}_{i=1}^{n}) + (1 - \theta) \mathcal{L}_{\mathrm{crf}}(\boldsymbol{v}, \{\pi_i\}_{i=1}^{n+m}),$$

and

$$\begin{aligned}
\mathcal{L}_{\mathrm{rr}}(\boldsymbol{u}, \{\pi_i\}_{i=1}^{n}) &= \frac{\lambda}{2} \| \boldsymbol{u} \|_2^2 + \frac{1}{2} \sum_{i=1}^{n} |y_i - \langle \boldsymbol{u}, \Phi(\boldsymbol{x}_i, \pi_i) \rangle|^2, \\
\mathcal{L}_{\mathrm{crf}}(\boldsymbol{v}, \{\pi_i\}_{i=1}^{n+m}) &= \frac{1}{2} \| \boldsymbol{v} \Gamma^{\frac{1}{2}} \|_2^2 - \langle \boldsymbol{v}, \Psi(\{\boldsymbol{x}_i\}_{i=1}^{n+m}, \{\pi_i\}_{i=1}^{n+m}) \rangle \\
&\quad + \log Z(\{\boldsymbol{x}_i\}_{i=1}^{n+m}, \boldsymbol{v}).
\end{aligned}$$

Here, λ, θ, and Γ are hyper-parameters, while Z is the partition

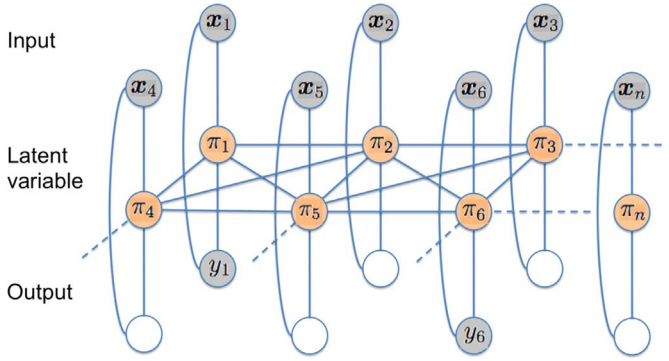


Fig. 2. The transductive conditional random field regression model.

function. Our model consists basically of two parts: a latent variable with spatial connections part and a regression part. Accordingly, we define two joint feature maps (cf. Tsochantaridis et al., 2005), Ψ and Φ. The feature map Ψ resembles a CRF corresponding to an undirected graph $G = (V, E)$ with binary edges E (i.e., spatial connections) and vertices V (i.e., impedance measurements), where each vertex represents a sample and the state space $S := \{1, \ldots, K\}$ depends on the number of facies K:

$$\Psi\left(\{\boldsymbol{x}_i\}_{i=1}^{n+m}, \{\pi_i\}_{i=1}^{n+m}\right) = \begin{pmatrix} \left(\sum_{(e1,e2)\in E} \mathbf{1}[\pi_{e_1} = s_1 \wedge \pi_{e_2} = s_2]\right)_{(s1,s2)\in S} \\ \left(\sum_{v\in V} \mathbf{1}[\pi_v = s]\phi(\boldsymbol{x}_v)\right)_{s\in S} \end{pmatrix}. \tag{4}$$

The joint feature map Φ for the regression part is

$$\Phi(\boldsymbol{x}, \pi) = \phi(\boldsymbol{x}) \otimes \Lambda(\pi), \tag{5}$$

where $\Lambda(\pi) \in \{0, 1\}^K$ with entries $(\Lambda(\pi))_k = 1$ if $\pi = k$ and 0 otherwise. $K \in \mathbb{N}^+$ is the number of facies we expect to encounter and ϕ is the feature function ϕ: $\mathbb{R}^D \to \mathcal{X}$ (i.e., impedances). Basically, the regression map is a K times replicated feature vector where all parts that do not correspond to the current active state π are set to zero. A simple illustrative example explaining the notation above is given in Appendix B. For further information and examples of joint feature maps, we refer to Tsochantaridis et al. (2005).

The training algorithm is given in Algorithm 1, which consists of 3 coordinate-wise optimization steps. After convergence, the regression target y_i (i.e., porosity) of each unlabeled sample is predicted by the regression model conditioned on the inferred parameter $\boldsymbol{u}^t$ and the latent variable π_i^t (i.e., facies). Hence, for each location i, we can report its corresponding porosity, as well as the type of facies π_i it belongs to.

Algorithm 1. Transductive Conditional Random Field Regression (TCRFR).

Put $t = 0$ and initialize $\boldsymbol{u}^t$ and $\boldsymbol{v}^t$ (e.g., randomly)
repeat
  t := t + 1
  Minimize Eq. (3) by splitting into 3 parts:
  (1) Update $\{\pi_i^t\}_{i=1}^{n+m}$ using the intermediate solutions $\boldsymbol{u}^{t-1}$ and $\boldsymbol{v}^{t-1}$
  (2) Update $\boldsymbol{u}^t$ with fixed $\{\pi_i^t\}_{i=1}^{n+m}$
  (3) Update $\boldsymbol{v}^t$ with fixed $\{\pi_i^t\}_{i=1}^{n+m}$
until $\forall\, i = 1, \ldots, n+m$: $\pi_i^t = \pi_i^{t-1}$
Predict unlabeled examples $\boldsymbol{x}_i = n+1, \ldots, n+m$ using the inferred states $\{\pi_i^t\}_{i=n+1}^{n+m}$ and regression parameter $\boldsymbol{u}^t$:
$y_i = \langle \boldsymbol{u}^t, \Phi(x_i, \pi_i^t) \rangle$

3. Proposed method

Although the original TCRFR method showed notable performance under a small proportion of labeled samples, it still requires ~ 5% of labeled samples (Görnitz et al., 2017), which is a much larger number than we can expect in practical geosciences field applications. To fill this gap, we propose an enhanced TCRFR variant equipped with two techniques inspired from image processing literature. The first idea is to apply a 3D graph-based segmentation, extended from Felzenszwalb and Huttenlocher (2004), in the seismic input volume. Based on the segmented volume, the neighborhood graph is constructed, through which TCRFR propagates the valuable label information. The second idea is to incorporate hand-labeled information on some voxels that geologists can provide with high confidence. Here, the hand-labeling is not to give a value for porosity, but for facies. The flow of our proposed method, called enhanced TCRFR (E-TCRFR), is shown in Fig. 3. These preprocessing steps and the mathematical framework of E-TCRFR are described in the following sections.

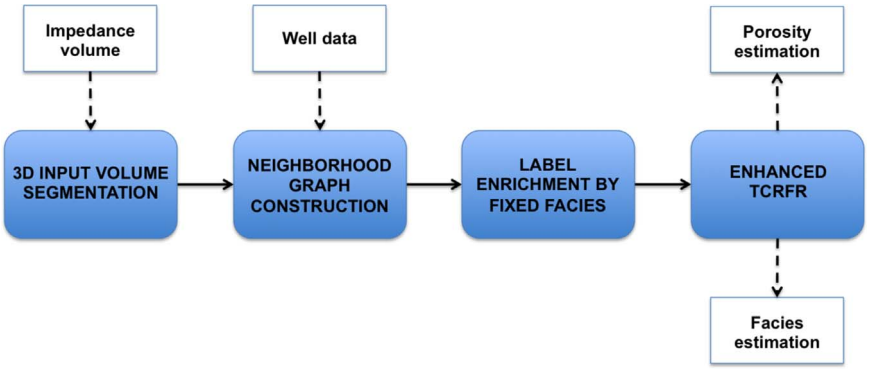


Fig. 3. Enhanced transductive conditional random field regression (E-TCRFR).

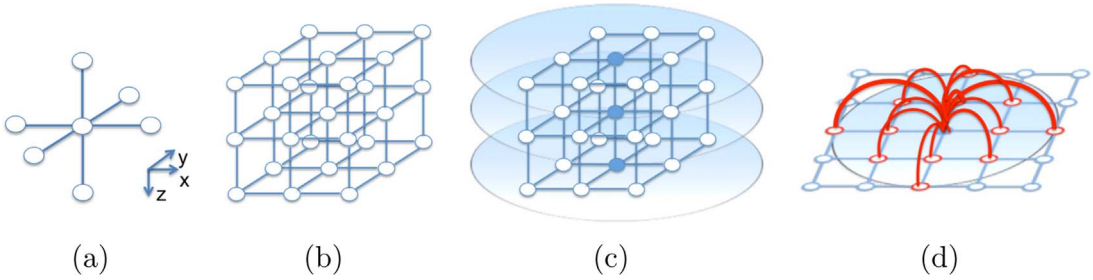


Fig. 4. Typical 3D graph connections in a volume. Blue voxels represent labeled samples in a well. White voxels represent unlabeled samples: (a) 6-tile voxel connections; (b) unlabeled connections in the volume; (c) radius for additional labeled (blue) to unlabeled (white) voxel connections in the horizontal slices; (d) example of connections from one labeled voxel to its neighbors, considering a radius of 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1. 3D input volume segmentation

In this step, we use graph-based image segmentation (Felzenszwalb and Huttenlocher, 2004) to define connected clusters (or *geobodies*, in this case) with similar features based solely on the impedance data. We extended the publicly available code for 2D images to 3D. The segmentation is applied to the whole volume. The impedance values are converted to RGB colors in a 256 color table.

The method adaptively adjusts the segmentation criterion based on the degree of variability in neighboring regions of the volume. The evidence for creating a boundary between two regions is given by comparing two quantities, one based on intensity differences across the boundary, and the other based on intensity differences between neighboring voxels within each region. Intuitively, the intensity differences across the boundary of two regions are perceptually important if they are large relative to the intensity differences inside at least one of the regions (Felzenszwalb and Huttenlocher, 2004). The details are described in Appendix A.

Note that we do not use the resulting segments as *super voxels*, in which all voxels are classified into a single facies category. Different facies can still occur inside some segments after this step. The graph construction step explained in the next subsection is essential for taking a good balance between the densely available impedance information and the sparsely available porosity information.

3.2. Neighborhood graph construction

By using the segmented 3D image obtained in the previous section, we create a neighborhood graph that describes the interaction between the latent variables (see Fig. 2). First, we connect in a 6-tile setting each voxel to the neighbors that belong to the same segment (see Fig. 4(a) and (b)). Then, we add edges from each labeled voxel (latent variable corresponding to the location where the porosity is observed) to its surrounding neighbors that are at the same depth in the volume (again, only if each pair of voxels belongs to the same segment). These additional connections consider the neighboring voxels that belong to a circle centered on the labeled voxel (Fig. 4(c)). A white voxel corresponds to an unlabeled sample, while a blue voxel represents a labeled voxel, i.e., the location where a well exists. The reason why we treat x, y directions and z direction differently is because the areal spatial continuity is usually greater than in the vertical direction (Deutsch, 2002). The circle radius is defined as the distance from this centered labeled voxel to its closest labeled neighbor. Fig. 4(d) shows an example of those additional connections (in red) from a labeled voxel, considering a radius equal to 2.

These additional edges help to propagate the reliable information contained in the labeled samples to the surrounding neighbors.

3.3. Incorporation of hand-labeling to facies

Although the graph construction explained in Section 3.2 improves the performance, in some real case scenarios it is still not sufficient to be used as a practical tool for the oil industry. To arrive at a better performance, we can incorporate hand-labeling information provided by the geologist.

We require the geologist to choose one or more 2D slices of the input volume, and manually label some pixels to one of the facies categories. Note that the geologist annotate only pixels where the corresponding facies can be safely considered with high confidence. This is possible, for instance, in the synthetic dataset for the shale facies (blue colored in Fig. 1(b)) that are distant from the sand facies (yellow colored in Fig. 1(b)). It is also possible to annotate some points as the sand facies. An example annotation is shown in Fig. 5(a), where the red points correspond to the well locations, the black dots correspond to the manually annotated “sand” pixels, and the white stripes correspond to manually annotated “shale” pixels. The connections from the hand-labeled facies voxels have a 6-tile setting (Fig. 4(a)).

Geologists are used to make several assumptions about the geological model, mainly during the geologic evaluation of a reservoir in the exploration phase, where the available labeled data (i.e., porosity) is really scarce. The hand-labeling step is not necessarily a requirement for the E-TCRFR method to work, but it can substantially improve the porosity prediction results, if the geologist detains sufficient expert knowledge to assign some hand-labeling facies. It is important to keep in mind that even one pixel in one slice in the whole volume can be already of great help for the method, as this valuable information is propagated throughout the whole 3D segment that pixel (voxel) belongs to in the volume, due to the graph structure.

3.4. Enhanced TCRFR

Here we give a mathematical description of our proposed method. The enhanced TCRFR (E-TCRFR) solves the following problem:

$$\min_{\boldsymbol{u}\in\mathcal{H}_1,\boldsymbol{v}\in\mathcal{H}_2,\{\pi_i\in\mathcal{Z}\}_{i=1}^{n+m}} \mathcal{L}(\boldsymbol{u},\boldsymbol{v},\{\pi_i\}_{i=1}^{n+m}) \quad \text{subject to} \quad \{\pi_k\}_{k\in\mathcal{M}} = \Omega, \tag{6}$$

where $\mathcal{M}$ is the index set of the hand-labeled voxels and $\Omega \in \mathcal{Z}^{|\mathcal{M}|}$ the corresponding set of facies categories. Note that we treat the hand-labels as hard constraints, assuming that the geologists give labels only when they are confident. $\mathcal{L}(\boldsymbol{u},\boldsymbol{v},\{\pi_i\}_{i=1}^{n+m})$ is a convex combination of the objectives of the regression model and the conditional random field:

$$\mathcal{L}(\boldsymbol{u},\boldsymbol{v},\{\pi_i\}_{i=1}^{n+m}) = \theta\mathcal{L}_{\text{rr}}(\boldsymbol{u},\{\pi_i\}_{i=1}^{n}) + (1-\theta)\mathcal{L}_{\text{crf}}(\boldsymbol{v},\{\pi_i\}_{i=1}^{n+m}), \tag{7}$$

where

$$\mathcal{L}_{\mathrm{rr}}(\boldsymbol{u}, \{\pi_i\}_{i=1}^{n}) = \frac{\lambda}{2} \| \boldsymbol{u} \|_2^2 + \frac{1}{2} \sum_{i=1}^{n} |y_i - \langle \boldsymbol{u}, \Phi(\boldsymbol{x}_i, \pi_i) \rangle|^2, \tag{8}$$

$$\mathcal{L}_{\mathrm{crf}}(\boldsymbol{v}, \{\pi_i\}_{i=1}^{n+m}) = \frac{1}{2} \| \boldsymbol{v}\Gamma^{\frac{1}{2}} \|_2^2 - \sum_{c=1}^{C} \langle \boldsymbol{v}, \Psi(\{\boldsymbol{x}_i\}_{i\in\mathcal{I}_c}, \{\pi_i\}_{i\in\mathcal{I}_c}) \rangle + \log Z(\{\boldsymbol{x}_i\}_{i=1}^{n+m}, \boldsymbol{v}). \tag{9}$$

In Eq. (9), we incorporate the result by graph-based segmentation. C denotes the number of segments, and the disjoint set $\mathcal{I}_c$ for $c = 1,...,C$ consists of the voxel indices within each of the segments. The joint feature map $\Psi(\{\boldsymbol{x}_i\}_{i\in\mathcal{I}_c}, \{\pi_i\}_{i\in\mathcal{I}_c})$ is constructed according to the neighborhood graph construction explained in Section 3.2.

For drastically saving computation time, we skip the MAP-inference for some segments in the following way. Each segment $c = 1,\ldots, C$ satisfies either of the following:

(a) The segment contains no labeled sample nor hand-labeled voxels;
(b) The segment contains only a single labeled samples or voxels hand-annotated to a signal facies category;
(c) The segment contains multiple labeled samples and/or voxels hand-annotated to multiple facies categories.

For the segments satisfying (a), our E-TCRFR cannot do much, because the voxels in the segments are completely unlabeled. For those segments, we assign the same facies category to the voxels in each segment by majority voting based on the impedance. Also for the segments satisfying (b), we assign the same facies category to the voxels in each segment, however, in this case, the category is the one estimated for the labeled voxel from the porosity or the one given to the hand-annotated voxels in the segment. The full MAP-inference is applied only to the segments satisfying (c). Although (a) and (b) are exceptional cases where the full MAP inference is not necessary, those cases apply to many segments under very sparsely labeled scenario. Since the computation for those cases is done with *constant* time complexity, this strategy provides a huge boost in runtime performance.

For case (c), the optimization is performed in a similar fashion to the original TCRFR. The pseudo-code is given in Algorithm 2.

Algorithm 2. Enhanced TCRFR algorithm.

Put $t = 0$ and initialize $\boldsymbol{u}^t$ and $\boldsymbol{v}^t$ (e.g., randomly)
repeat
  t := t + 1
  $c = 0$
  repeat
    Update $\{\pi_i^t\}_{i\in\mathcal{I}_c}$ according to setting (a), (b), or (c) for the current partition using the intermediate solutions $\boldsymbol{u}^{t-1}$ and $\boldsymbol{v}^{t-1}$
    $c := c + 1$
  until $c = C$
  (2) Update $\boldsymbol{u}^t$ with fixed $\{\pi_i^t\}_{i=1}^{n+m}$
  (3) Update $\boldsymbol{v}^t$ with fixed $\{\pi_i^t\}_{i=1}^{n+m}$
until $\forall\, i = 1,\ldots,N$: $\pi_i^t = \pi_i^{t-1}$
Predict unlabeled examples $\boldsymbol{x}_i = n + 1,\ldots,n + m$ using the inferred states $\{\pi_i^t\}_{i=n+1}^{n+m}$ and regression parameter $\boldsymbol{u}^t$: $y_i = \langle \boldsymbol{u}^t, \Phi(x_i, \pi_i^t) \rangle$

Fig. 5 illustrates how the additional techniques work in an example situation. TCRFR assumes that the impedance for the whole volume and the porosity observation at the wells (red dots in Fig. 5(a)) are given. Additionally, E-TCRFR assumes hand-labeled voxels (black dots and white stripes in Fig. 5(a)), and it performs graph-based segmentation on the impedance data (Fig. 5(b)). Note that Fig. 5(b) shows a slice of the segmented volume and each disconnected segment in this slice with the same color actually belongs to a single 3D segment (for example the cyan regions in the upper part).

The two figures on the right show the facies estimation results by E-TCRFR without (Fig. 5(c)) and with (Fig. 5(d)) the hand-labeled annotations. We can see that the hand-labels are helpful in the regions where there are no wells.

4. Experimental results

In this section, we conduct experiments on synthetic and real reservoir datasets. On the synthetic data, we compare the performance results with the provided ground truth for different criteria: for prediction, we show the median absolute error (MDAE) and the coefficient of determination (R^2) score; for clustering (latent variable estimation) accuracy, we show the adjusted rand score (ARS). On the real dataset, we visually compare the results obtained with our proposed method with the ones provided by the classical geostatistics approach (Deutsch and Journel, 1998).

4.1. Empirical evaluation on synthetic seismic data

We use the second layer of the Stanford VI synthetic 3D reservoir

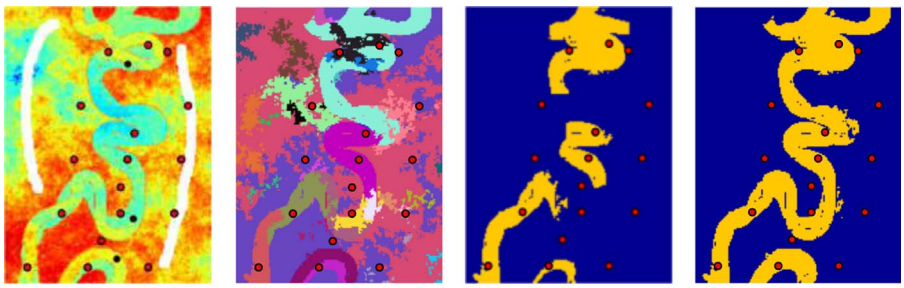


Fig. 5. Illustration of the effect of E-TCRFR. (a) Impedance input slice with well locations (red dots) and fixed facies given by annotators. The white stripes represent "shale" facies and the black dots represent "sand" facies; (b) the 3D graph-based segmentation image; (c) the estimated facies by E-TCRFR *without hand-labeling*; (d) the estimated facies by (full) E-TCRFR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

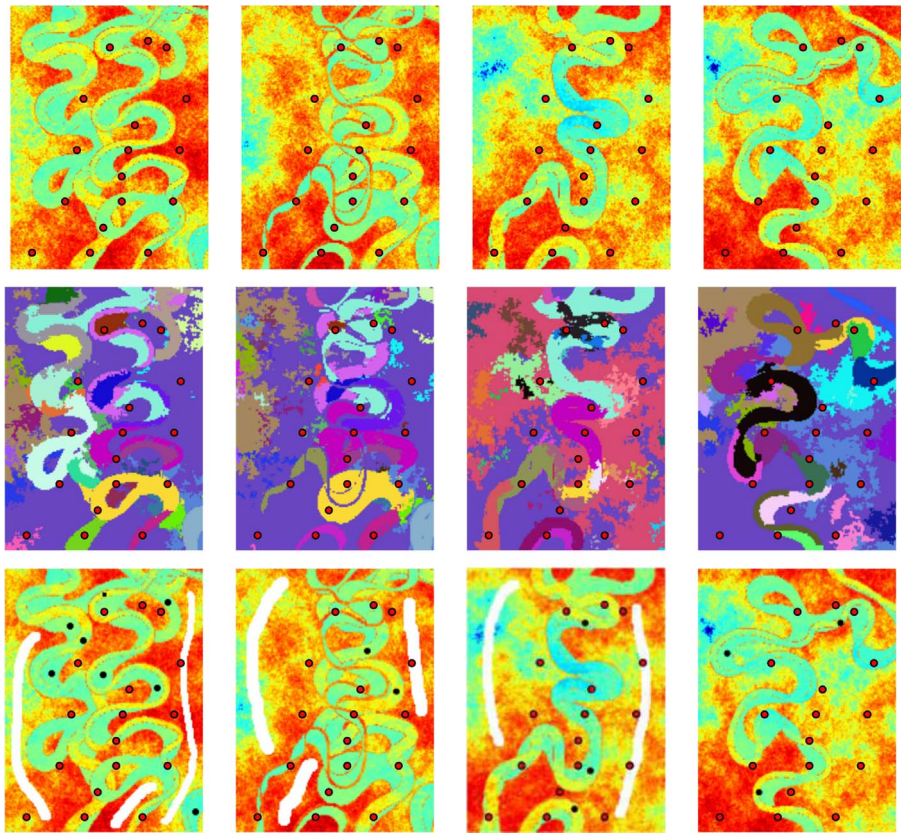


Fig. 6. Four distinct regions used for the synthetic data experiment. The red dots correspond to the well locations. Top row: the shear impedance input; Middle row: (a slice) of the graph-based segmented volume; Bottom row: manual annotations given by a geologist, where the white stripe voxels are annotated as the "shale" facies, and the black dot voxels are annotated as "sand" facies. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(Castro et al., 2005) (150 × 200 × 40 voxels), consisting of meandering sand channels. For reservoir exploration purposes, it is enough to segment the meandering depositional system from the shale in this example Miall, 2014, so we simplify the data model for our experiments by merging the point bar and channel sands in one facies (sand), and the floodplain and boundary in another one (shale).

We divide the volume layer in four zones (in the z direction) that present distinct geometry shapes (Fig. 6) and apply our method on

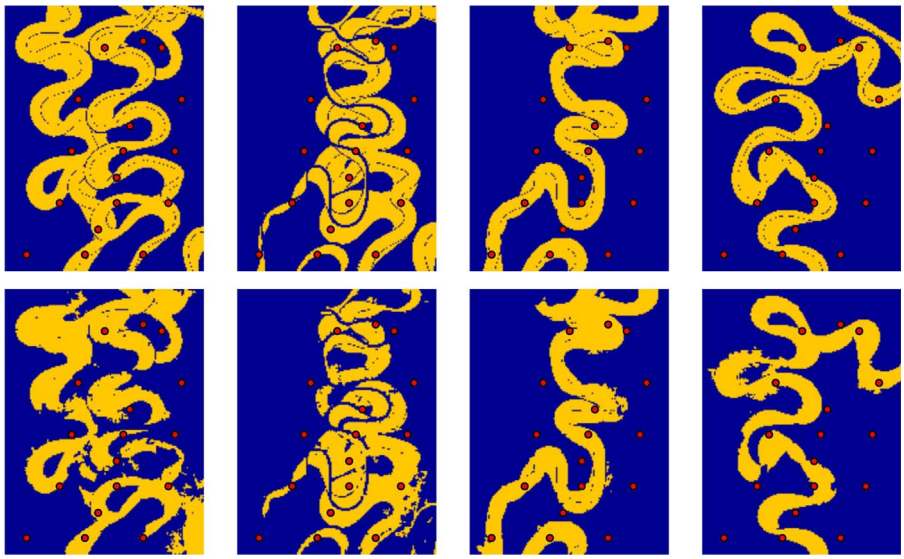


Fig. 7. Estimated facies by E-TCRFR for 4 different regions in the synthetic data. The red dots correspond to the well locations. Top row: the ground truth facies; Bottom row: estimated facies by E-TCRFR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

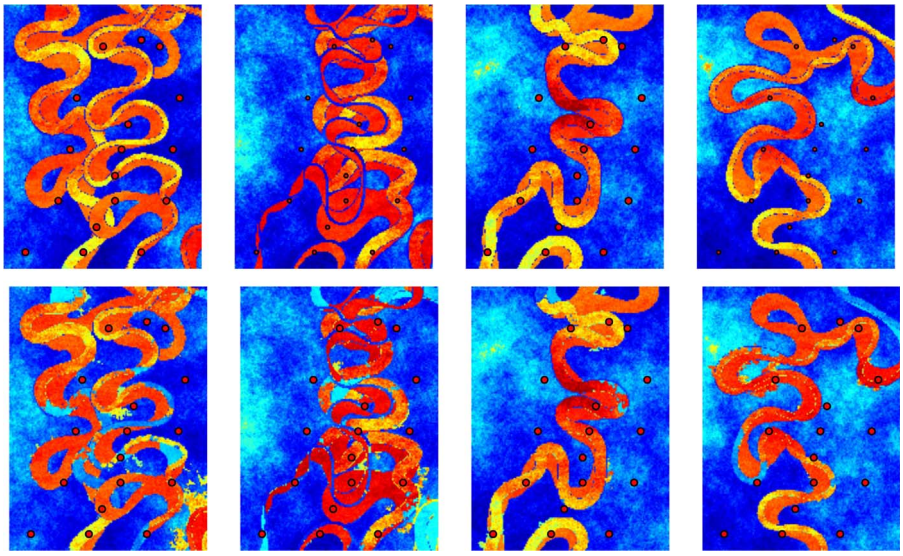


Fig. 8. Estimated porosity by E-TCRFR for 4 different regions in the synthetic data. The red dots correspond to the well locations. Top row: the ground truth porosity; Bottom row: estimated porosity by E-TCRFR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Performance on synthetic seismic data.

Zone	# Slices	MDAE	R2	ARS
1	12	0.17122	0.82054	0.68739
2	2	0.17492	0.80693	0.69699
3	10	0.16267	0.90446	0.86022
4	6	0.16527	0.91091	0.86165

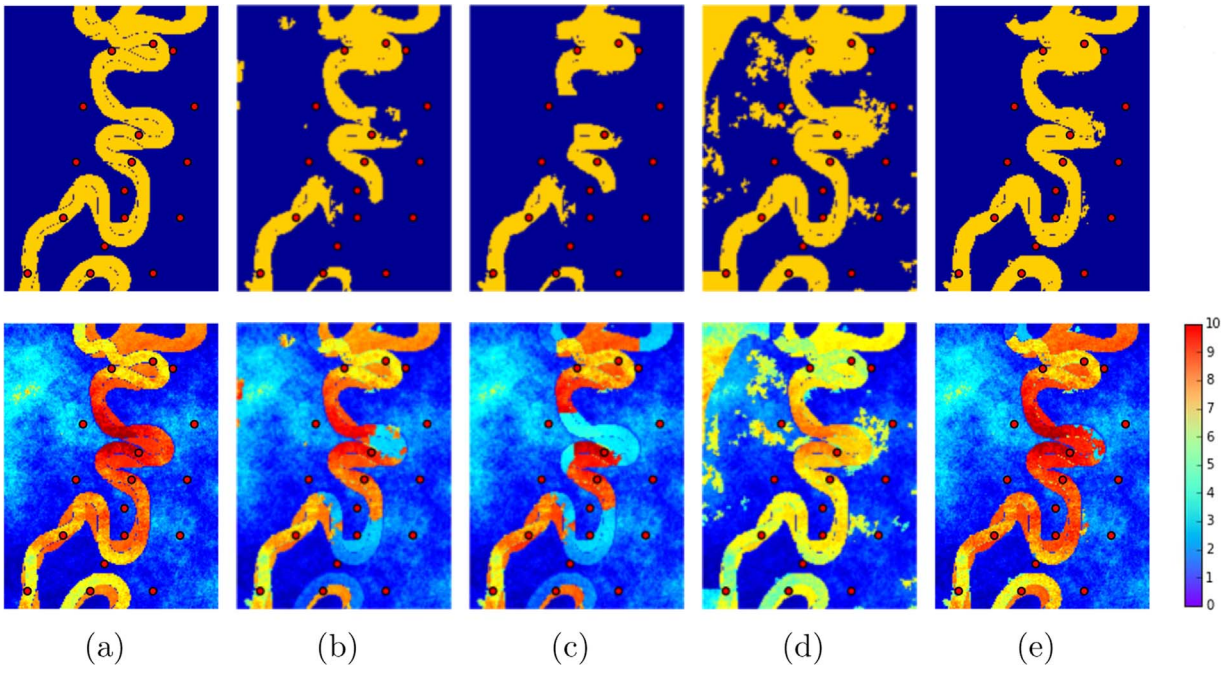


Fig. 9. Facies (top) and porosities (bottom) results for different TCRFR methods. (a) ground truth; (b) original TCRFR; (c) TCRFR with graph construction based on the segmented volume; (d) TCRFR with manual annotations; (e) (full) E-TCRFR.

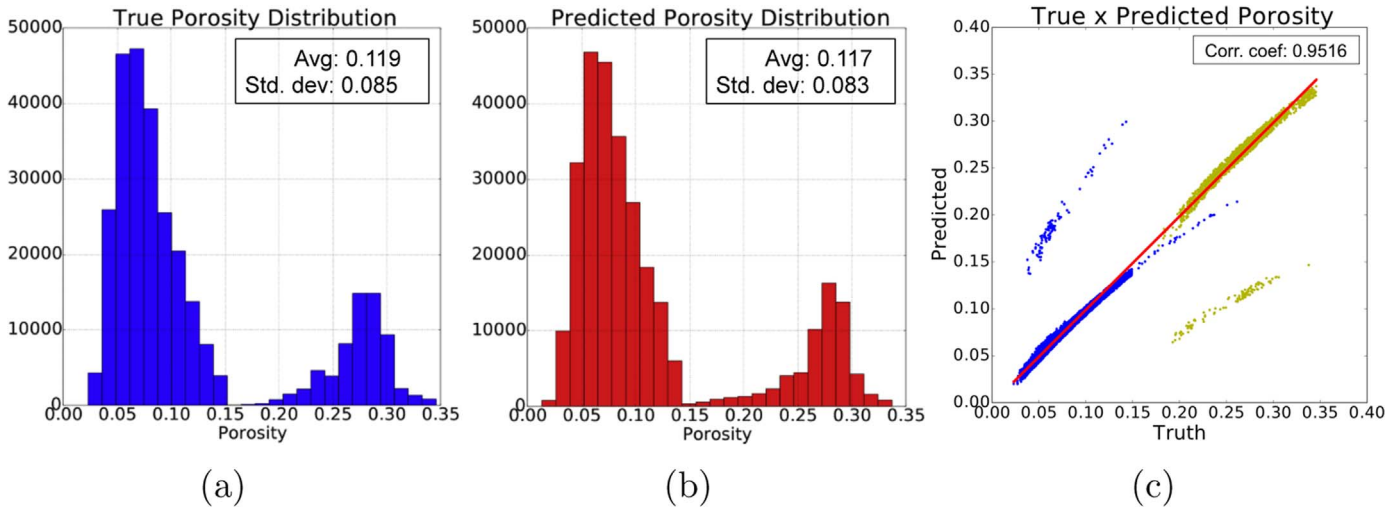


Fig. 10. Porosity statistics: (a) true porosity distribution; (b) estimated porosity distribution by E-TCRFR; (c) true vs. estimated porosity cross plot. Sand facies samples in yellow and shale facies samples in blue.

each of them separately. For the input data, we use the shear impedance volume.

Porosity observation is given at 17 production wells available in the reservoir (P1–P6 and P21–P31 in Castro et al. (2005)). For each of the four zones described above, we use all the porosity observations at the 17 wells.

The hyperparameter θ, λ, Γ are set in the following way: Let $\Gamma = \nu I$ with $\nu \in \mathbb{R}^+$ and I being the identity matrix with ones on the diagonal and zeros elsewhere. We then adjusted ν such that the terms in Eqs. (8) and (9) are leveled. This simple heuristic proved to work reasonably well on toy data. For the choice of θ and λ we applied leave-one-out cross-validation for each of the 17 wells, obtaining $\theta = 0.99$ and $\lambda = 1$.

Figs. 7 and 8 show the quality of facies and porosity estimation, respectively, by E-TCRFR. We see that E-TCRFR (bottom row in each figure) accurately estimates the ground-truth (top row). Table 1 shows quantitative results with the performance criteria.

Fig. 9 compares the performance of E-TCRFR and the original TCRFR.[5] Comparing with the ground truth (Fig. 9(a): facies (top) and porosity (bottom)), we clearly see that E-TCRFR (Fig. 9(e)) outperforms the original TCRFR (Fig. 9(b)). Fig. 9(c) and (d) show the results with other variants of TCRFR, where just one of the new techniques, i.e., the new graph-construction and the incorporation of manual annotation, is applied. We see that in this case both techniques are essential for good performance of our proposed method. In particular, in Fig. 9(c) the sand (yellow) channel is disconnected because of the lack of label information which is compensated by hand-labeling in Fig. 9(e), while in Fig. 9(d) the facies of the main sand channel is accurately estimated, but the method incorrectly classifies shale (blue) regions (mainly on the top left corner of the slice) as sand. As a result, the regression model for the sand facies is inaccurately trained, which results in a poor porosity prediction over the sand channel regions.

Fig. 10 shows some statistics of the true and the estimated porosity distributions by E-TCRFR. Fig. 10(a) and (b) show that the distribution of the estimated porosity is quite similar to the distribution of the true one.

Fig. 10(c) shows that there are two small clusters with misclassified samples: The blue one on the top shows shale (low) porosity samples that were incorrectly classified into the sand facies; the yellow one on the bottom shows sand (high) porosity samples incorrectly classified in the shale facies.

We performed sensitivity analysis on a slice of the synthetic data, progressively adding Gaussian noise over the impedance input with impedance values varying from 0% to 100%. Fig. 11 presents the results. The top row of Fig. 11(a)–(e) shows the impedance input data. The second row presents the corresponding estimated facies and the third row shows the estimated porosity. The coefficient of determination (R2) and median absolute error (MDAE) results are presented in Fig. 11(f). It can be observed that even with 20% Gaussian noise the R2 performance is still close to 85%, while the MDAE increases linearly with the noise.

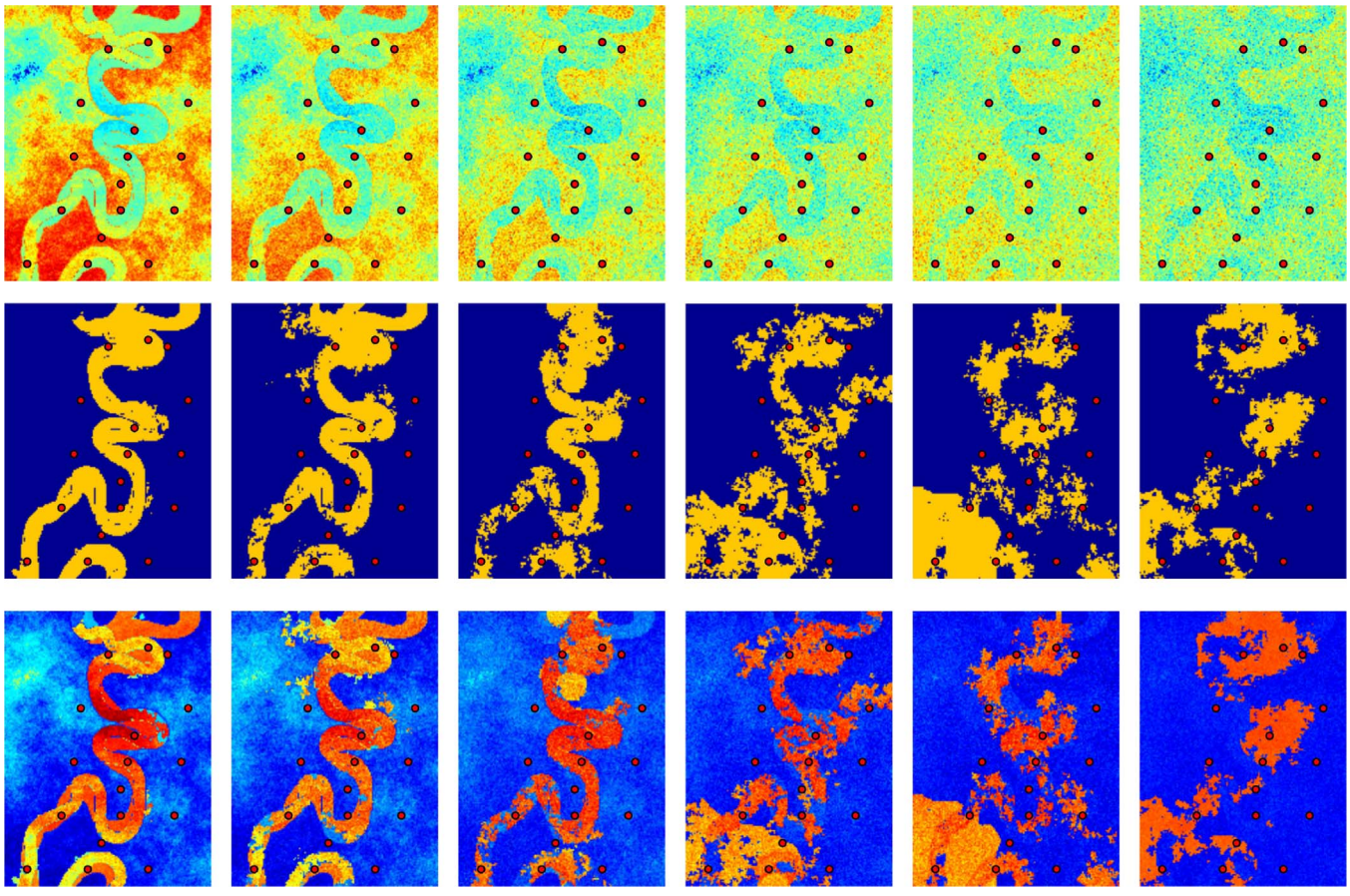
Fig. 12 presents sensitivity analysis considering the hand-labeled facies. In Fig. 12(a) some portions of the sand channel are not correctly identified by the method. In Fig. 12(b) we added a black ("sand") point on the upper half of the slice and, as a result, a good portion of the channel is now detected by E-TCRFR. Adding a second black point to the bottom half of the slice in Fig. 12(c) makes it possible for the method to connect the whole sand channel. Fig. 12(d) and (e) illustrate that adding more black points to the slice do not necessarily further improve the overall result, showing that E-TCRFR just requires a minimum number of hand-labeled points to provide a good performance. The MDAE and R2 plots shown in Fig. 12(f) and (g) present the corresponding increase in the method's performance as a result the added hand-labeled facies.

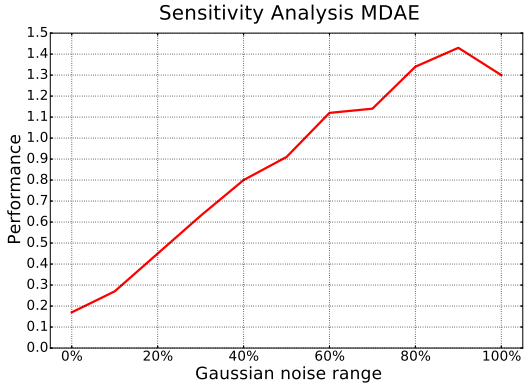
E-TCRFR execution time is approximately linear in the number of samples as shown in Fig. 13, where the method is executed varying the number of contiguous slices in the volume from one (30,000 samples) to 10 (300,000 samples).

4.2. Porosity prediction on real seismic data

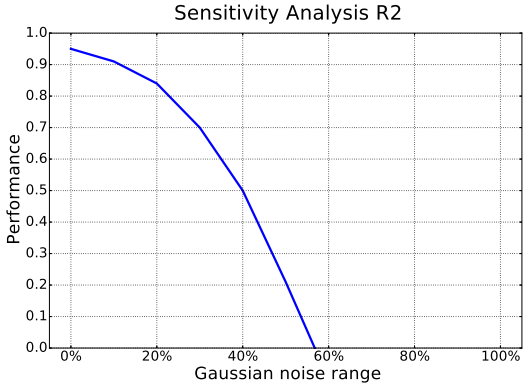
We now apply our proposed method to a real carbonate reservoir, located in the offshore coast of Brazil (cf. Fig. 14). It covers an area of approximately 100 square kilometers, with 460 m in depth. The reservoir is part of the sedimentary rock formation whose depositional model is presented in Fig. 15. It comprises a carbonate platform with progressive shallowing cycles strongly related to subsidence, salt tectonics and sea level oscillations. The reservoir is composed of oolitic/oncolytic calcarenites developed in high energy environments (oolitic shoals). These shoals were developed in the highest parts of the

[5] We omit comparison with the previous methods other than TCRFR, since they were shown to be outperformed by the original TCRFR (Görnitz et al., 2017).





(g)



(h)

Fig. 11. E-TCRFR sensitivity analysis on a slice of the synthetic data: increasing Gaussian noise applied to the impedance input, from 0% to 100% standard deviation over the original values. Top row: the impedance input; Second row: the estimated facies; Third row: the estimated porosity; Bottom row: sensitivity analysis plots for median absolute error (MDAE) and coefficient of determination (R2) with increasing Gaussian noise over the original input impedance.

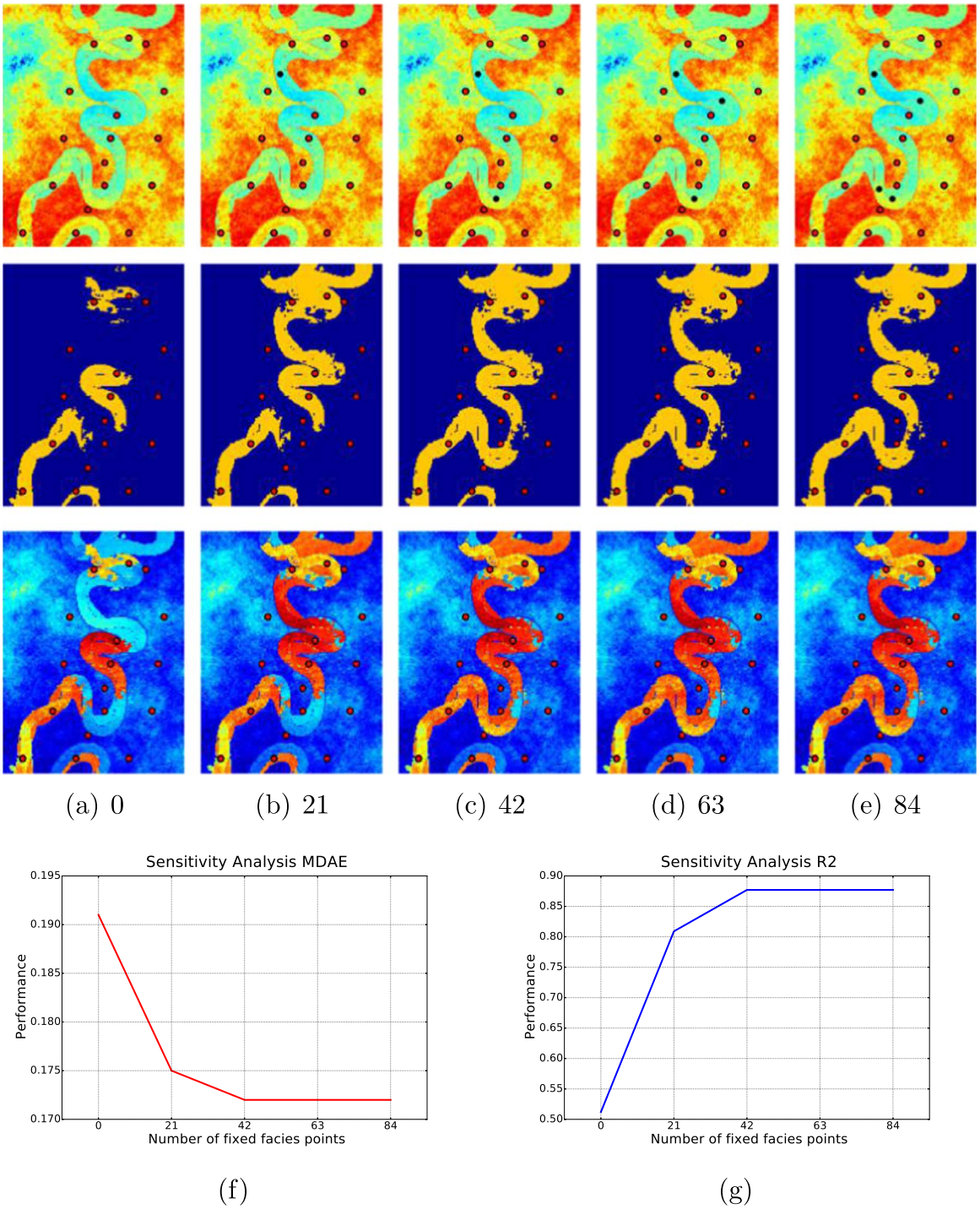


Fig. 12. E-TCRFR sensitivity analysis on a slice of the synthetic data for increasing number of hand-labeled facies defined by the geologist. Top row: the impedance input; Second row: the estimated facies; Third row: the estimated porosity; Bottom row: sensitivity analysis plots for median absolute error (MDAE) and coefficient of determination (R2) with increasing number of hand-labeled facies.

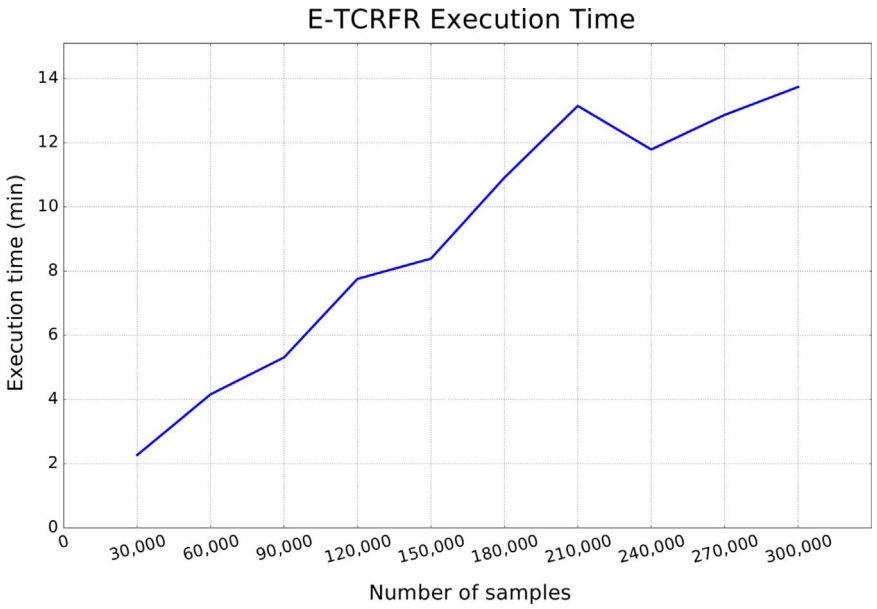


Fig. 13. E-TCRFR execution time (in minutes) from one to ten contiguous slices. Each slice contains 30,000 samples.

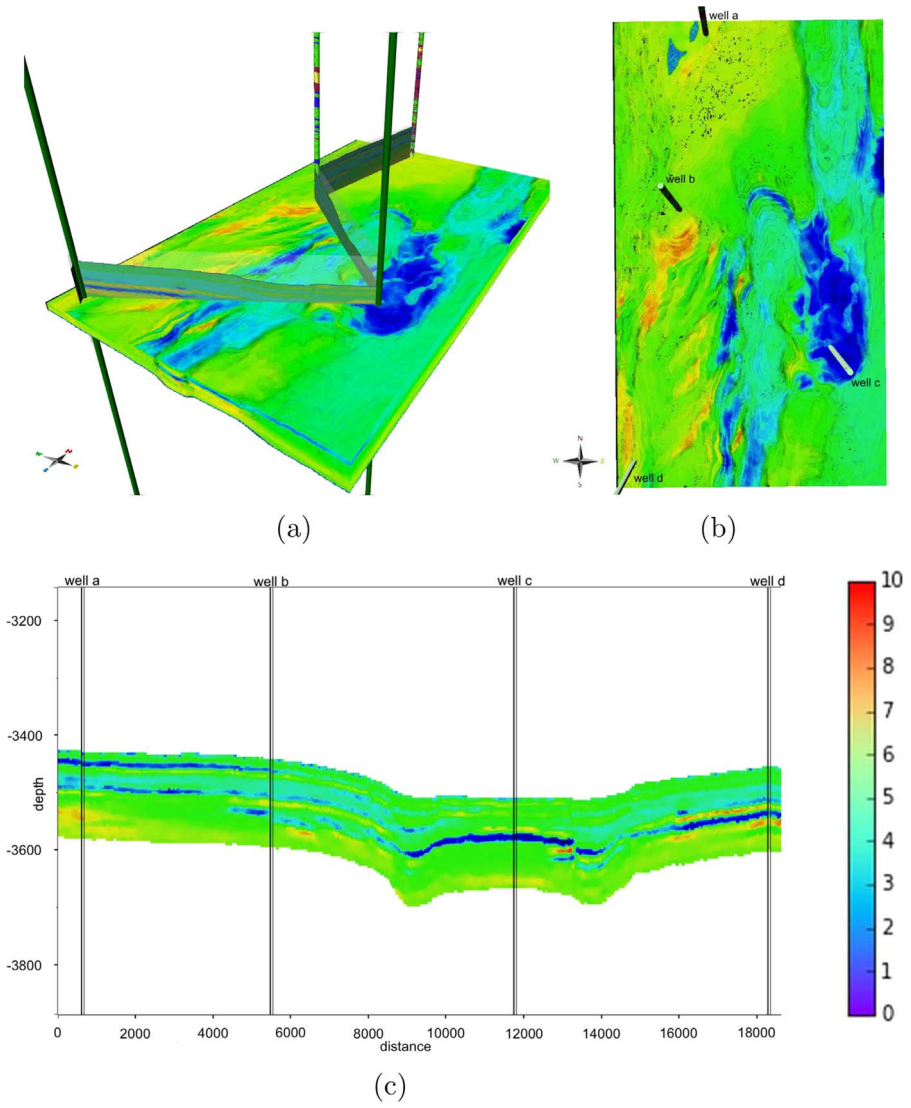


Fig. 14. The real data reservoir: (a) 3D view of an acoustic impedance subvolume in the reservoir with a cut section passing along the four wells; (b) map view; (c) section view passing along the four wells.

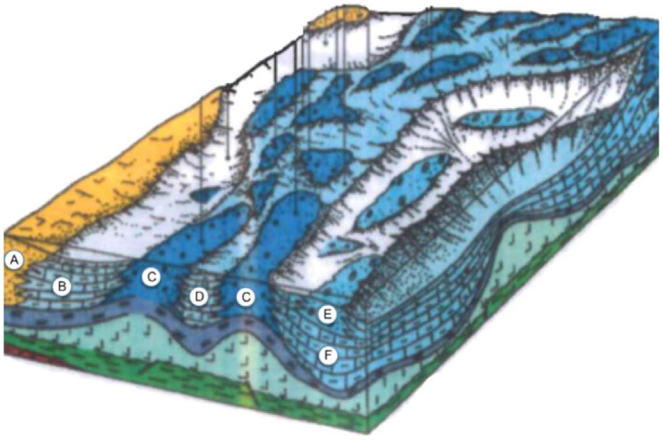


Fig. 15. Depositional model: (A) terrigenous, tidal plain; (B) wackstones/packstones; (C) oolitic grainstones; (D) peloidal packstones; (E) oncolytic packstones; (F) wackstones/mudstones (open sea).

structures generated by the movement of the salt that accumulated in that area during middle Albian. The variations in the tectonic regime and/or fluctuations of the sea level promoted the cyclicity in the depositional system characterized by the intercalation of sediments with high and low energy. The extensive deposits of low energy, formed during high-sea level correspond to seals to distinct reservoir units. Three facies groups occur in this region: grainstone at the bar crests with high energy sediments; oolitic/oncolytic packstones with moderate to low energy sediments at the flanks of the shoals; and peloidal packstones and wackestones in depressions located around the bars.

Table 2
Comparison between TCRFR and E-TCRFR on the real dataset. Errors are evaluated by using the geostatistics estimation as reference.

Method	RMSE	MDAE
TCRFR	0.89065	0.48472
E-TCRFR	0.44430	0.17930

The volume data in the reservoir region comprises $313 \times 549 \times 74$ voxels of acoustic impedance samples and four exploratory wells with a total of 121 (*impedance*, *porosity*) pair samples in seismic resolution. The impedance volume was previously obtained using constrained sparke-spike inversion in the Jason™ Workbench. We chose a subvolume with 6 contiguous time slices and used all the porosity values provided by the four wells, so the number of labeled samples correspond to approximately 0.01% of the total number of samples in this subvolume. We also estimated the porosity with the traditional geostatistics approach. The algorithm used was 3D Kriging with Locally Varying Mean (LVM).

Fig. 16 shows (a) a time slice of the seismic impedance input with manual annotation; (b) the graph-based segmentation result; (c) the estimated facies by the original TCRFR; (d) the estimated facies by the

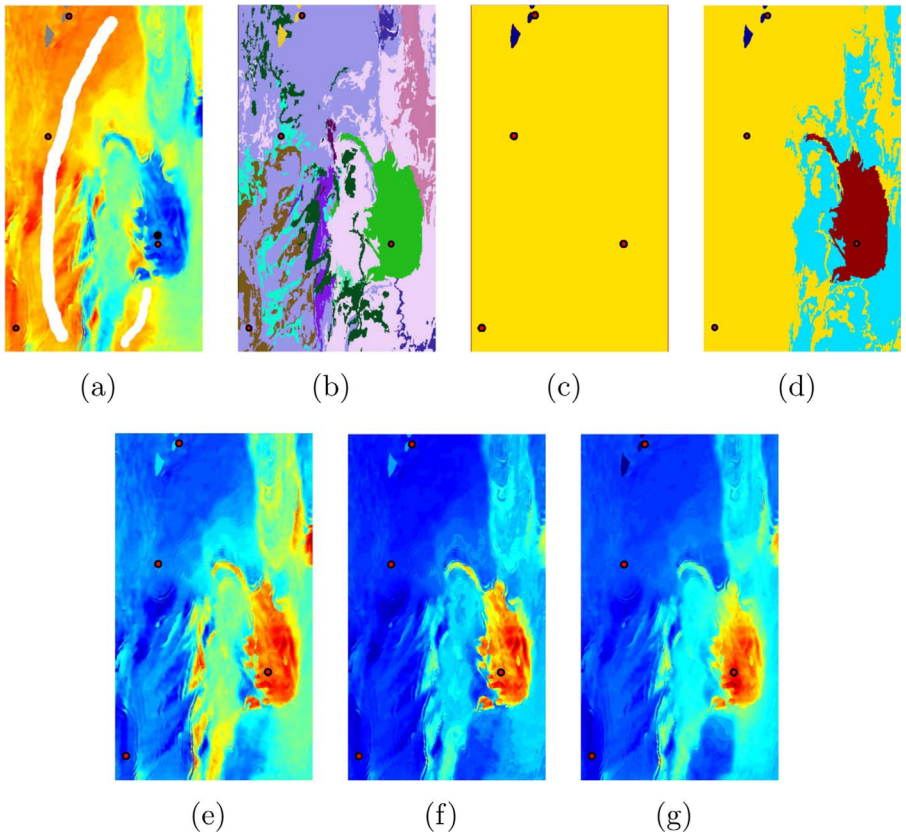


Fig. 16. Estimated facies and predicted porosity for one slice in the real data: (a) impedance input with manual annotations; (b) graph-based segmentation; (c) facies estimated by original TCRFR; (d) facies estimated by E-TCRFR; (e) porosity estimated by original TCRFR; (f) Porosity estimated by E-TCRFR; (g) porosity estimated by geostatistics.

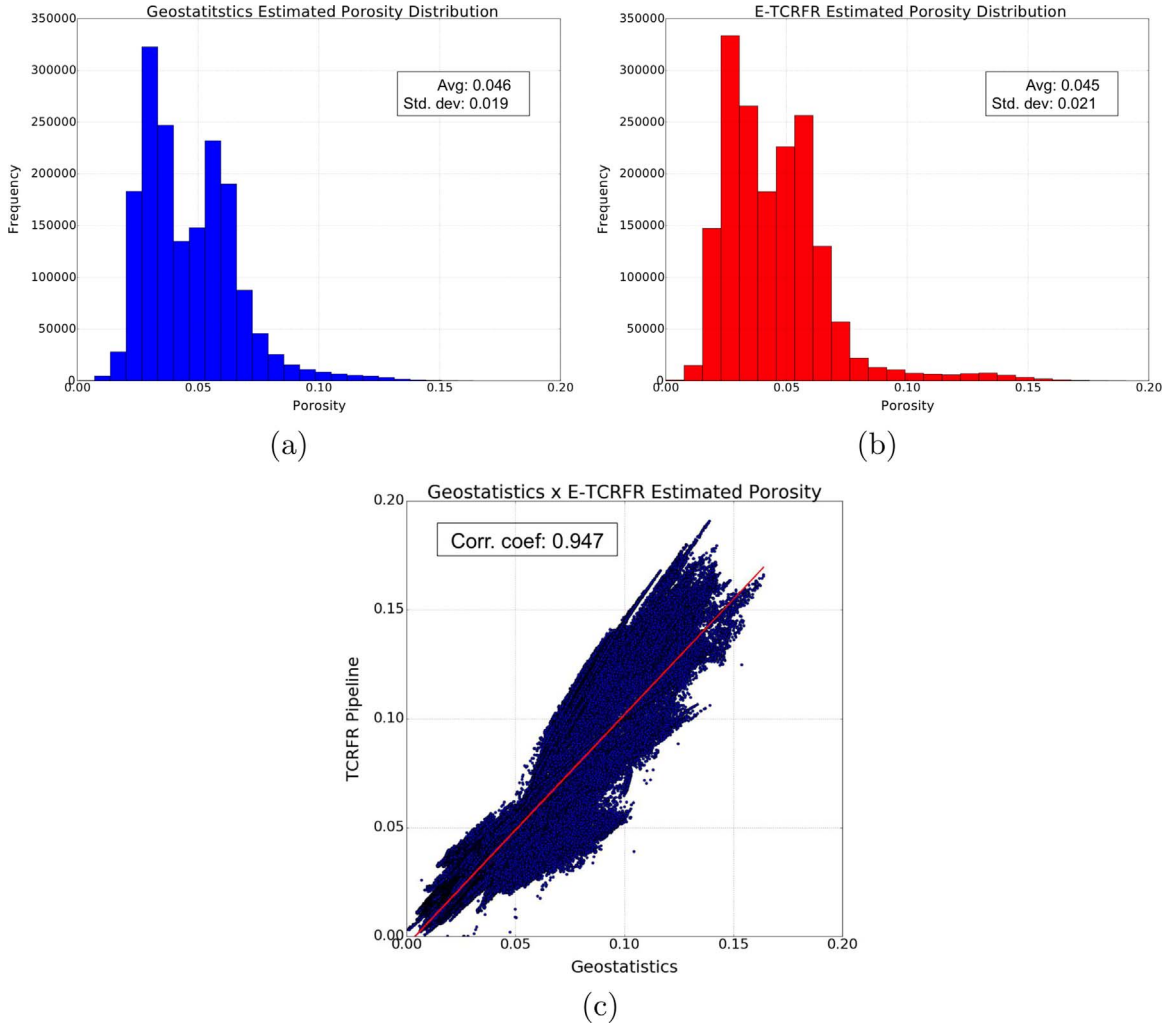


Fig. 17. Porosity statistics: (a) geostatistics porosity distribution; (b) E-TCRFR porosity distribution; (c) geostatistics vs. E-TCRFR porosity cross plot.

improved TCRFR, (e) the estimated porosity by the original TCRFR; (f) the estimated porosity by E-TCRFR; and (g) the estimated porosity by geostatistics. Here we used the same hyperparameters estimated for the synthetic case.

We can observe a significant gain from the original TCRFR—our improved TCRFR (f) gives a more similar result to the geostatistics estimation (g) than the original TCRFR (e). As we see in Fig. 16(c), the original TCRFR is not able to correctly estimate the facies, as the number of labeled samples (121 in this case) is much smaller than the unlabeled samples (more than a million). Only one facies was found, leading to just one regression model. With E-TCRFR, three facies were estimated.

Table 2 shows the RMSE and MDAE errors by original TCRFR and E-TCRFR from the geostatistics estimation and Fig. 17 shows the estimated porosity histograms and a cross plot comparing the geostatistics and E-TCRFR estimation results. Again, one can observe that the results from the improved TCRFR are similar to the ones obtained with the geostatistical approach.

It is worth noting that our E-TCRFR gives sharper contours than the geostatistics estimation. Although we cannot argue with only the current results that this is an advantage of E-TCRFR over the geostatistics estimation, it might imply that our semi-automatic method has a potential to improve even geostatistics estimation.

In general, even with partial facies overlap, E-TCRFR is able to estimate the different facies present in a reservoir as long as the corresponding regression models are distinct, i.e., if the slope and/or intercept for each facies linear regressor is different from all the others.

5. Conclusion

Machine learning techniques have been applied to many fields in recent decades. However, in many cases there still remains a gap between the state-of-the-art machine learning methods and automatic industrial tools. The oil industry is not an exception. Our work in this paper to improve transductive conditional random field regression (TCRFR) is exactly an effort to fill this gap, and make a machine learning method a practically useful tool.

Equipped with two image processing techniques, our enhanced TCRFR has taken a significant step forward in this direction, getting closer to become a practical option for geologists in real applications. Future work will aim to further enhance our method, minimizing human interactions. Outside the application field of the oil industry, where we mainly encounter spatial statistics, we consider TCRFR of

promise for time series analysis, where data may even have complex spatio-temporal dependencies.

Acknowledgments

LAL and LEV acknowledge the support of Petrobras. NG was supported by BMBF ALICE II Grant 01IB15001B. He also acknowledges the support by the German Research Foundation through the Grant DFG MU 987/6-1 and RA 1894/1-1. KRM thanks for partial funding by the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology in the BK21 Program. KRM and SN were supported by the German Ministry for Education and Research as Berlin Big Data Center BBDC, funding mark 01IS14013A. KRM is corresponding author.

Appendix A. Detail of graph-based segmentation (Felzenszwalb and Huttenlocher, 2004)

The volume is represented as a graph $G = (V; E)$, where each node $v_i \in V$ corresponds to a voxel and the edges in E connect pairs of neighboring voxels $(v_i, v_j) \in E$ in a 6-connected tile, as shown in Fig. 4(a). A weight $w((v_i, v_j))$ is associated with each edge. This weight is a non-negative measure of the dissimilarity between neighboring elements v_i and v_j based on the RGB color intensity difference between the voxels that it connects: $w((v_i, v_j)) = |I(p_i) - I(p_j)|$. $I(p_i)$ represents the color intensity of voxel p_i.

The method is executed once for each of the red, green, and blue color components. Two neighboring voxels are set in the same cluster only if they independently belong to each of the same red, green, and blue clusters.

The segmentation S is a partition of V into clusters such that each cluster $C_i \in S$ corresponds to a connected component in a graph $G' = (V; E')$, where $E' \subseteq E$. The algorithm compares inter-cluster differences to intra-cluster differences for every pair of adjacent clusters to create a boundary. The idea is that edges between two vertices in the same cluster should have relatively low weights, and edges between vertices in different cluster should have higher weights.

Appendix B. A simple example of joint feature maps

Our model consists basically of two parts: a latent variable with spatial connections part and a regression part. The defined two joint feature maps for the CRF are Ψ, and, for the ridge regression part Φ. The feature map Ψ resembles a CRF corresponding to an undirected graph $G = (V, E)$ with binary edges E (i.e., spatial connections) and vertices V (i.e., impedance measurements), where each vertex represents a sample and the state space $S := \{0, \ldots, K-1\}$ depends on the number of facies K.

In our little example, let G be a simple chain of length n with hidden variables π and 2-dimensional observations $\boldsymbol{x}$ (x^0 and x^1) much like a hidden Markov model. For further simplicity, we assume the feature mapping function ϕ to be the identity. The latent variables (facies) can take 2 distinct states 0 and 1 (i.e. $S = \{0, 1\}$ and a realization of $\pi \in S$).

$$
\begin{aligned}
\Psi(\{\boldsymbol{x}_i\}_{i=1}^{n+m}, \{\pi_i\}_{i=1}^{n+m}) &= \begin{pmatrix} \left(\sum_{(e1,e2)\in E} \mathbf{1}[\pi_{e_1} = s_1 \wedge \pi_{e_2} = s_2] \right)_{(s1,s2)\in\{0,1\}} , \\ \left(\sum_{v\in V} \mathbf{1}[\pi_v = s]\boldsymbol{x}_v \right)_{s\in\{0,1\}} \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^{n-1} \mathbf{1}[\pi_i = 0 \wedge \pi_{i+1} = 0] \\ \sum_{i=1}^{n-1} \mathbf{1}[\pi_i = 0 \wedge \pi_{i+1} = 1] \\ \sum_{i=1}^{n-1} \mathbf{1}[\pi_i = 1 \wedge \pi_{i+1} = 0] \\ \sum_{i=1}^{n-1} \mathbf{1}[\pi_i = 1 \wedge \pi_{i+1} = 1] \\ \sum_{i=1}^{n} \mathbf{1}[\pi_1 = 0] x_i^0 \\ \sum_{i=1}^{n} \mathbf{1}[\pi_1 = 0] x_i^1 \\ \sum_{i=1}^{n} \mathbf{1}[\pi_1 = 1] x_i^0 \\ \sum_{i=1}^{n} \mathbf{1}[\pi_1 = 1] x_i^1 \end{pmatrix} \in \mathbb{R}^8. \qquad \text{(B.1)}
\end{aligned}
$$

Hence, our joint feature map consists basically of two parts: a part that counts the number of state transitions and another emission part that counts observations in a specific state. An illustrative example hereof is given in Fig. B.18.

The joint feature map Φ for the regression part is much simpler (again, we assuming ϕ is the identity)

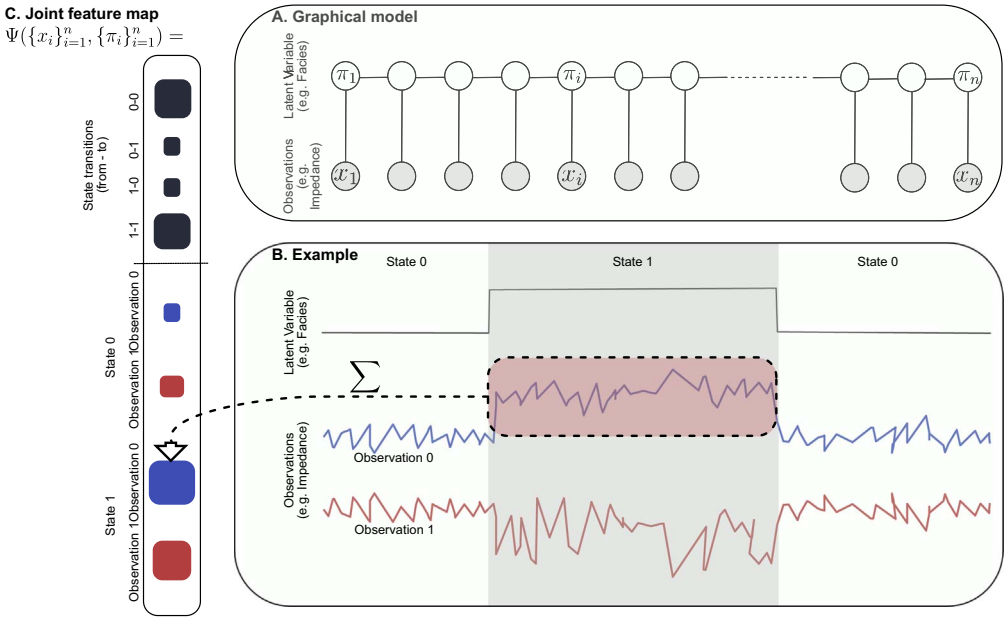


Fig. B.18. An illustrative example of joint feature maps. (A) depicts the graphical model: a Markov random field with binary connections between latent variables (π, 2 states: State 0 and State 1) and between latent variables and observations (2 dimensional, continuous) $\boldsymbol{x}$. (B) gives a concrete example with two observations (blue and red respectively) and latent variables (black) while (C) gives an intuition how this examples translates into a joint feature map. Here, the red box sums the measurements from Observation 0 and State 1 and sets them accordingly in the joint feature map.

$$\Phi(\boldsymbol{x}, \pi) = \phi(\boldsymbol{x}) \otimes \Lambda(\pi) = \boldsymbol{x} \otimes \Lambda(\pi), \tag{B.2}$$

where $\Lambda(\pi) \in \{0, 1\}^K$ with entries $(\Lambda(\pi))_k = 1$ if $\pi = k$ and 0 otherwise. Hence, when $K = 2$ the feature map can be re-written as

$$\Phi(\boldsymbol{x}, \pi) = \phi(\boldsymbol{x}) \otimes \Lambda(\pi) = \begin{pmatrix} x^0 & \text{if } \pi = 0 \text{ else } 0 \\ x^1 & \text{if } \pi = 0 \text{ else } 0 \\ x^0 & \text{if } \pi = 1 \text{ else } 0 \\ x^1 & \text{if } \pi = 1 \text{ else } 0 \end{pmatrix} \in \mathbb{R}^4. \tag{B.3}$$

References

Avseth, P., Mukerji, T., Mavko, G., 2010. Quantitative Seismic Interpretation. Cambridge University Press.

Blei, D.M. , Ng, A.Y., Jordan, M.I., Wallach, H.M., Hinton, G.E., Osindero, S., Teh, Y.-W., 2004. Conditional random fields: an introduction (Technical Report MS-CIS-04-21). Department of Computer and Information Science, University of Pennsylvania.

Boykov, Y., Jolly, M.-P. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: Proceedings of the Eighth IEEE International Conference on Computer Vision. ICCV 2001, vol. 1, pp. 105–112.

Caers, J., 2005. Petroleum Geostatistics. Society of Petroleum Engineers.

Castro, S., Caers, J., Mukerji, T., 2005. The stanford VI reservoir. 18th Annual Report. Stanford Center for Reservoir Forecasting (SCRF), pp. 1–73.

Connolly, P., Hughes, M., 2016. Stochastic inversion by matching to large numbers of pseudo-wells. Geophysics 82, M7–M22.

Deutsch, C.V., Journel, A.G., 1998. GSLIB – Geostatistical Software Library and User's Guide 2nd ed.. Oxford University Press.

Deutsch, C.V., 2002. Geostatistical Reservoir Modeling. Oxford University Press.

Doyen, P.M., 2007. Seismic Reservoir Characterization: An Earth Modeling Perspective. EAGE Publications.

Dubrule, O., 2003. Geostatistics for Seismic Data Integration in Earth Models. SEG.

Eidsvik, J., Omre, H., Mukerji, T., Mavko, G., Avseth, P., 2002. Seismic reservoir prediction using Bayesian integration of rock physics and Markov random fields; a North Sea example. Lead. Edge 21, 290–294.

Eidsvik, J., Avseth, P., Omre, H., Mukerji, T., Mavko, G., 2004. Stochastic reservoir characterization using prestack seismic data. Geophysics 69, 978–993.

Felzenszwalb, P.F., Huttenlocher, D.P., 2004. Efficient graph-based image segmentation. Int. J. Comput. Vis. 59 (2), 167–181.

Görnitz, N., Lima, L.A., Varella, L.E., Müller, K.-R., Nakajima, S., 2017. Transductive regression for data with latent dependency structure. IEEE Trans. Neural Netw. Learn. Syst.. http://dx.doi.org/10.1109/TNNLS.2017.2700429.

Grana, D., Rossa, E.D., 2010. Probabilistic petrophysical-properties estimation integrating statistical rock physics with seismic inversion. Geophysics 75, O21–O37.

Gunning, J., Glinsky, M., 2004. Delivery: an open-source model-based Bayesian seismic inversion program. Comput. Geosci. 30, 619–636.

He, X., Zemel, R., Carreira-Perpinan, M., 2004. Multiscale conditional random fields for image labeling. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 695–702.

Lafferty, J., McCallum, A., Pereira, F., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the International Conference on Machine Learning (ICML). Citeseer, p. 282–289.

Larsen, A., Ulvmoen, M., Omre, H., Buland, A., 2006. Bayesian lithology/fluid prediction and simulation on the basis of a Markov-chain prior model. Geophysics 71, R69–R78.

Mavko, G., Mukerji, T., Dvorkin, J., 2009. The Rock Physics Handbook 2nd ed.. Cambridge University Press.

Miall, A.D., 2014. Fluvial Depositional Systems. Springer Geology.

Mukerji, T., Jørstad, A., Avseth, P., Mavko, G., Granli, J.R., 2001a. Mapping lithofacies and pore-fluid probabilities in a north sea reservoir: seismic inversions and statistical rock physics. Geophysics 66, 988–1001.

Mukerji, T., Avseth, P., Mavko, G., Takahashi, I., 2001b. Statistical rock physics: combining rock physics, information theory, and geostatistics to reduce uncertainty in seismic reservoir characterization. Lead. Edge 20, 313–319.

Sams, M., Atkins, D., Said, P., Parwito, E., van Riel, P., Stochastic inversion for high resolution reservoir characterisation in the Central Sumatra Basin, In: Proceedings of the SPE Asia Pacific Improved Oil Recovery Conference.

Schlumberger, 2015. The Oilfield Glossary. URL ⟨http://www.glossary.oilfield.slb.com/Terms/p/porosity.aspx⟩.

Shimodaira, H., 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. J. Stat. Plan. Inference 90, 227–244.

Spikes, K., Mukerji, T., Dvorkin, J., Mavko, G., 2007. Probabilistic seismic inversion based on rock-physics models. Geophysics 72, R87–R97.

Sugiyama, M., Krauledat, M., Mueller, K.-R., 2007. Covariate shift adaptation by importance weighted cross validation. J. Mach. Learn. Res. 8, 985–1005.

Sutton, C., Mccallum, A., 2010. An Introduction to Conditional Random Fields for Relational Learning. Arxiv 7, 93.

Taskar, B., Klein, D., Collins, M., Koller, D., Manning, C., Max-margin parsing. In: Empirical Methods in Natural Language Processing.

Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y., 2005. Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. 6, 1453–1484.

Zeller, G., Görnitz, N., Kahles, A., Behr, J., Mudrakarta, P., Sonnenburg, S., Raetsch, G., 2013. mTim: Rapid and Accurate Transcript Reconstruction from RNA-Seq Data, arXiv: 1309.5211.