Bollettino di Geofisica Teorica ed Applicata Vol. 59, n. 1, pp. 71-94; March 2018
DOI 10.4430/bgta0217

Application of Machine Learning and Digital Music Technology to distinguish high from low gas-saturated reservoirs

P. Dell'Aversana, G. Carrasquero, G. Gabbriellini and A. Amendola

Eni S.p.A. Upstream and Technical Services, San Donato Milanese (MI), Italy

(Received: July 4, 2017; accepted: December 20, 2017)

ABSTRACT In this paper, we discuss a novel approach of pattern recognition, clustering and classification of seismic data based on commonly applied techniques in the domain of digital music and in musical genre classification. Our workflow starts with accurate conversion of seismic data from SEGY to Musical Instrument Digital Interface (MIDI) format. Then, we extract MIDI features from the converted data. These can be single-valued attributes related to instantaneous frequency and/or to the signal amplitude. Furthermore, we use multi-valued (or "high-level") MIDI attributes that have no equivalent in the seismic domain. For instance, we use MIDI features related to melodic, harmonic and rhythmic patterns in the data. We discuss an application to real data. We apply a Machine Learning approach to the MIDI-converted seismic data set with the purpose of accurate seismic facies classification. The final objective of the test is to distinguish between geological formations prevalently formed by clay, from two different gas-bearing sandy layers: one is a low gas-saturated reservoir and the other one is a high gas-saturated reservoir. In this paper, we present encouraging results. Considering the novelty of our approach, additional investigations are in progress on larger data sets, for a complete understanding of the physical meaning of the new "high-level" MIDI attributes.

Key words: pattern recognition, seismic classification, Machine Learning, digital music, MIDI attributes.

1. Introduction

Spectral decomposition methods have been applied in the hydrocarbon industry over the past few decades to extract the characteristic frequency components from seismic data and identify low frequency anomalies. These can be used as possible indicators of hydrocarbon accumulations (Klimentos *et al.*, 1995; Castagna and Sun, 2006; Wang, 2007). Some authors assume that the presence of low frequency anomalies in correspondence of thick gas reservoirs can be caused by variable attenuation of the seismic waves (Winkler and Nur, 1982; Kumar *et al.*, 2003). In fact, higher frequencies are expected to be attenuated more than the lower frequencies by a significant oil/gas filled reservoir. Other authors assume that the decrease in seismic velocity is the main factor causing low frequency anomalies associated with gas reservoir zones (Tai *et al.*, 2009). Castagna *et al.* (2003) apply a technique of Instantaneous Spectral Analysis (ISA) based on seismogram decomposition into constituent wavelets using wavelet transform methods for detecting low-frequency shadows associated with hydrocarbons. In their work, these authors discuss the benefits

and limitations of their approach compared with other time-frequency analysis techniques.

Despite the good results obtained by advanced spectral decomposition approaches, an important question remains open. This is the intrinsic difficulty to distinguish between scenarios of high or low gas saturation. In fact, it is well known that significant seismic amplitude anomalies, including low frequency anomalies, are often ascribed to low gas saturation in the reservoir (O'Brien, 2004). This misleading condition is often named "the fizz-water" effect. Batzle and Wang (1992) provide an excellent discussion on how gas distribution affects seismic attributes. The crucial question is that high seismic amplitudes and frequency variations can be clear evidences of the presence of gas in a reservoir, but they are poor quantitative indicators of the degree of gas saturation. Gas saturation of a few percent has an effect on P-wave velocity comparable to that of full gas saturation, making any distinction between the two scenarios difficult or even impossible. This dilemma remains unsolved even when applying advanced spectral decomposition analysis.

To tackle the above open question, in this paper, we introduce an approach based on new attributes derived through a different type of seismic data analysis. We start from applying advanced techniques of spectral decomposition to our data set. As in other approaches, but with a novel workflow, we try to identify amplitude/frequency anomalies and other "particular signatures" possibly associated with hydrocarbon accumulation. However, unlike previous researches in this field, we derived and applied the attributes in an unusual domain. This is the domain of digital music. We transform the seismic data into a musical format known as MIDI (Musical Instrument Digital Interface). This is a standard protocol used in digital music and represents a powerful symbolic format. As explained in this paper, this new "musical" representation of the seismic data offers a multitude of advantages. In the following sections, we discuss how the benefits of established geophysical techniques can be constructively combined with innovative methods recently developed in other domains. These fields include digital music, machine learning, Musical Information Retrieval (MIR), and musical genre classification.

2. From time series to S-spectrograms

In previous works, we introduced our approach based on sonification, for analysis and interpretation of geophysical data (Dell'Aversana, 2013, 2014; Dell'Aversana *et al.*, 2016a). We apply the Stockwell Transform (or S-Transform) in order to perform accurate time–frequency analysis. This transform applied to a generic time dependent signal $x(t)$, is given by (Stockwell *et al.*, 1996):

$$S(\tau, f) = \frac{|f|}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(t) e^{-i2\pi ft} e^{-\frac{(t-\tau)^2 f^2}{2}} dt \quad , \qquad (1)$$

where τ is the time where the S-Transform is calculated and f is the instantaneous frequency. The exponential function in the integral is frequency dependent. This allows the windowing function scrolling the time series as a function of the frequency. This transform is appropriate for analysing seismic signals, where instantaneous frequency information changes over time (non-stationary signals).

The left panel of Fig. 1 shows a real seismic trace and the relative Stockwell spectrogram.

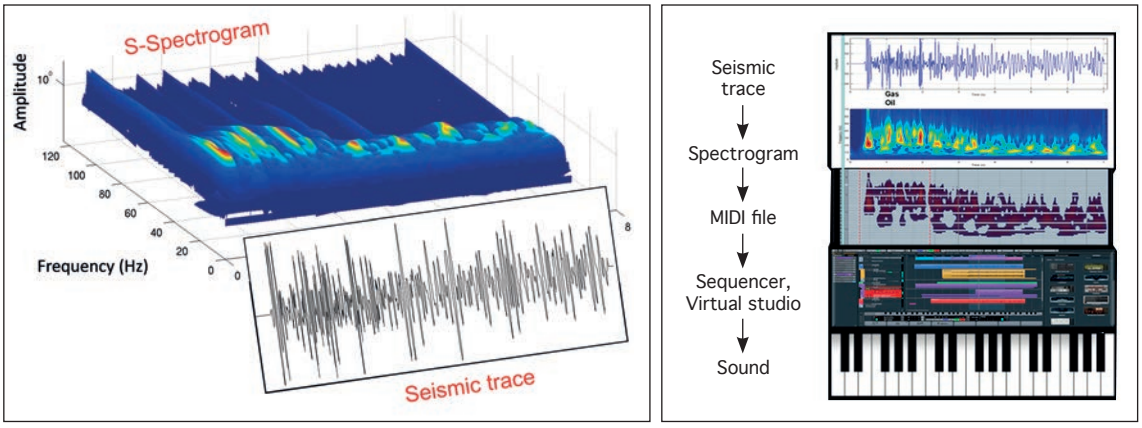


Fig. 1 - Left panel: example of a spectrogram of a seismic trace obtained by applying the Stockwell transform (after Dell'Aversana *et al.*, 2016a). Right panel: workflow through which seismic data are transformed into digital sounds.

3. From S-spectrograms to MIDI attributes

The next step in transforming geophysical data into musical files is to convert the physical quantities of a spectrogram (frequency, time, and amplitude) into "basic" MIDI attributes (such as pitch, sound intensity, and note length). The right panel of Fig. 1 shows how seismic data are progressively translated into audible sounds, moving from time series to spectrograms and from these to MIDI files. The MIDIs are then imported into digital music software ("sequencer", synthesizer, virtual mixer, etc.) and, after proper frequency transposition, are transformed into audible sounds.

We use the following mathematical relationship (Dell'Aversana *et al.*, 2016a) between the frequency f and the MIDI note number n:

$$f(n) = 440 \cdot 2^{(n-58)/12} \qquad (2)$$

In Eq. 2, the symbol n indicates the sequential number of MIDI notes. For instance, n = 108 corresponds to B8 (f = 7902 Hz).

As mentioned, the geophysical data transformed into MIDI files can easily be transposed into the audible frequency range (between 30 and 20000 Hz). These MIDI data can be played by using modern computer music tools, such as sequencers. Finally, audio analysis is performed simultaneously with interpretation of images, as a complementary tool (audio-video display).

Beside the basic MIDI attributes there are "high level MIDI attributes". These are, for instance, patterns of MIDI notes forming special melodic, harmonic and rhythmic trends. Taken together, these attributes form a sort of "musical texture" that can be related to specific musical genres. Such patterns and textures are very useful in the process of Musical Information Retrieval and in Musical Genre Classification because they allow distinguishing different musical categories and/or identifying individual songs in large databases.

4. Benefits of conversion from seismic to MIDI format

The conversion from seismic to MIDI formats offers a multitude of advantages. First, the MIDI standard represents a light symbolic representation of the seismic data. It allows creating discretized spectrograms of the original signal, where every single digital (musical) note represents a local spectral information. In fact, the tone of the note is derived from the instantaneous frequency of the signal, and the sound intensity is derived from the local amplitude of the original signal itself.

These MIDI files take up little memory (a few Kbyte for each seismic trace), as they do not contain the original waveform, but just the instructions to reproduce it using a "sequencer". Consequently, the conversion to MIDI brings meaningful information from a physical point of view and, at the same time, can be processed very quickly.

Many examples of these files can be found on the web (see for instance the YouTube Channel at the link https://www.youtube.com/channel/UCp4VG897AsFiSoBrvEnyRJA. A full e-lecture including examples of audio-video display can also be found here: https://www.youtube.com/watch?v=tGhICX2stTs.

The possibility to reproduce geophysical data through audio-video display offers the additional benefit of a dual perception of the same information. We have verified the effectiveness of this "multi-sensory" interpretation approach through individual and group tests, involving different types of geoscientists in the analysis/interpretation of audio-video display (Dell'Aversana *et al.*, 2016a). The feedback of the specialists was positive in about 90% of those interviewed in the tests, confirming that simultaneous audio-visual display of seismic data supports the interpretation process under many aspects (fault detection, anomaly detection, seismic facies analysis, and so forth).

This interactive audio-visual approach is intriguing and useful, but necessarily time consuming. Consequently, it can be applied to restricted areas of the geophysical data volume, for instance in selected portions of the data where imaging is affected by poor signal-to-noise ratio. A further benefit of having the data in MIDI format is that several processes of automatic analysis can be applied efficiently to large databases, as happens in the domain of digital music. For instance, automatic pattern recognition, data mining, clustering and classification, are all processes that can be made extremely efficient when applied to "light" volumes of MIDI files derived from geophysical data sets (Amendola *et al.*, 2017).

5. Machine Learning and Musical Information Retrieval for seismic data classification

In our previous works (Dell'Aversana, 2014; Dell'Aversana *et al.*, 2016b; Amendola *et al.*, 2017), we introduced the idea that standard seismic data converted into MIDI format can be analysed and classified using the same methods that are commonly applied for musical pattern recognition. We imported and applied some algorithms of Musical Information Retrieval (MIR) into the geophysical domain. The main goal of MIR is to recognize occurrences of a musical query pattern within a musical database, such as a collection of musical pieces available in Internet databases. Our approach addressed to geophysical data classification uses the same criteria, with the difference that the database consists of sounds obtained by

converting geophysical data into musical formats. In other words, we adapt fast and accurate MIR algorithms to geophysical purposes. In our previous applications, we demonstrated the validity of our approach at variable spatial scales. In fact, we showed that "high-level MIDI attributes" (such as those based on melodic, harmonic, and rhythmic patterns) can contribute to the correct classification of seismic facies. In the following paragraphs, we show how the same approach can be applied to the problem of distinguishing *high* from *low* gas-saturated sand reservoirs.

MIR commonly uses a wide range of classification approaches and algorithms (McKay, 2004). We applied a hybrid-supervised approach to our real data combining several types of learner algorithms, such as the k-Nearest Neighbours algorithm (or K-NN for short), Artificial Neural Networks (ANNs), Support Vector Machine (SVM), Random Forest and other methods.

In K-NN, an object is classified taking into account the properties of its neighbours (commonly forming the training set). These classifiers offer advantages in training speed; however, they are limited in modelling relationships between features. Thus, we used K-NN for a preliminary classification step based on basic MIDI features only (pitch and sound intensity distribution). The classification approach is based entirely on some type of distance that expresses the similitude of records estimated on one or more features. Instead, ANNs take into account the relationships between the features. ANNs consist of units (emulating human neurons) connected by links, each with an associated weight. Learning in ANNs takes place by modifying the values of the weights through an iterative optimization process. We used ANNs to refine the previous classification step, adding also "high level" MIDI attributes, such as melodic, harmonic and rhythmic patterns.

SVM is a machine learning technique that separates the attribute space with a hyperplane. It allows maximizing the margin between the instances of different classes or class values.

Besides K-NN, ANNs and SVM, other techniques used in this paper (such as Decision Tree, Naïve Bayes, and Random Forest) have already been applied and discussed in geophysics for many different purposes (Aminzadeh and de Groot, 2006). Consequently, we skip the detailed description of these methods. However, our approach to classify geophysical data transformed into MIDI files is new in geophysics. In the next section, we discuss a specific test to demonstrate the effectiveness of our method in distinguishing sandy/clay and shale formations, low gas-saturated and high gas-saturated sandy reservoirs.

6. A small-scale classification test

6.1. Problem description and methodological aspects

The methodological objective of this test is to verify if our classification approach based on MIDI attributes can distinguish, in principle, seismic facies related to geological formations with different rock-physical properties and fluid content. We selected an area where an exploration well encountered a stacked reservoir. This consists of unconsolidated gas bearing sandstones of Upper Pliocene, deposited in a turbidite channel system. The lateral continuity of the channels is interrupted by erosional effects and by variable clay content. The logs in the upper layer revealed low gas saturation, whereas the lower formation showed high gas saturation. Both targets are sealed by prevalent clay formations.

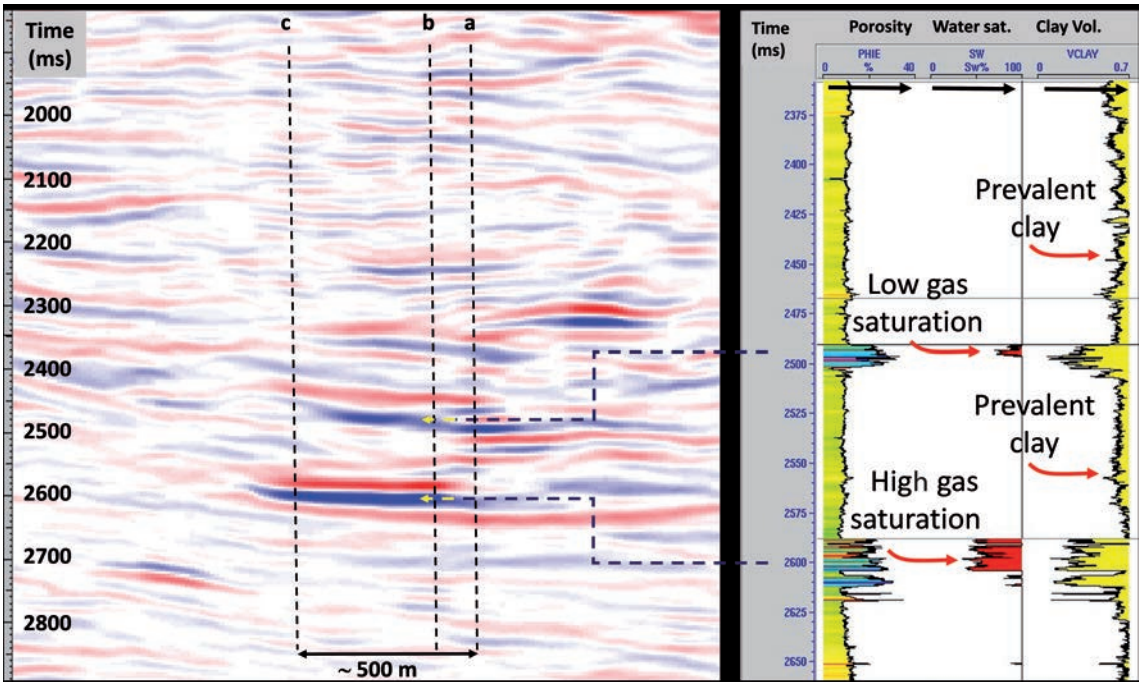


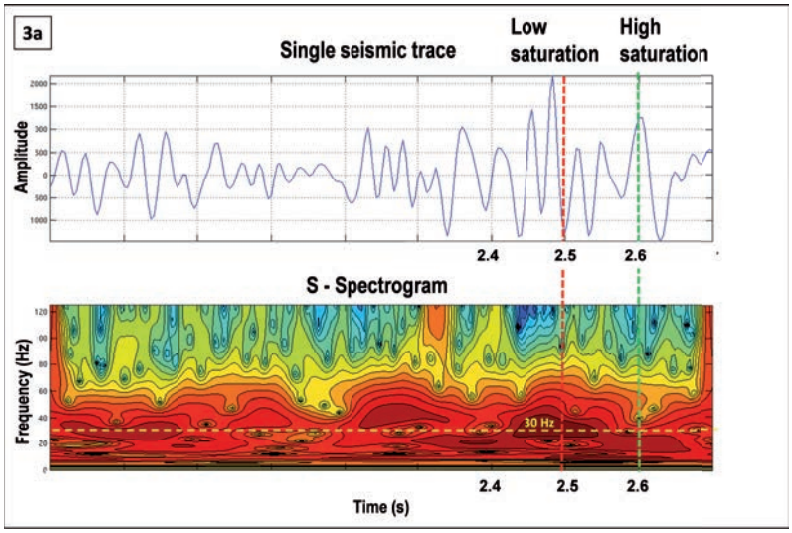
Fig. 2 - Left panel: seismic image of two gas-bearing layers with different degree of gas saturation. The exploration well is located roughly in the middle of the image. The letters "a", "b" and "c" on the top of the seismic section, indicate the location of the individual seismic traces shown in the next Figs. 3a, 3b, and 3c, respectively. Right panel: CPI logs showing porosity, water saturation and clay volume in the well crossing the seismic section. The vertical scale of the two panels are different, because they show data with different resolution. The arrows help to correlate the well logs with the seismic section.

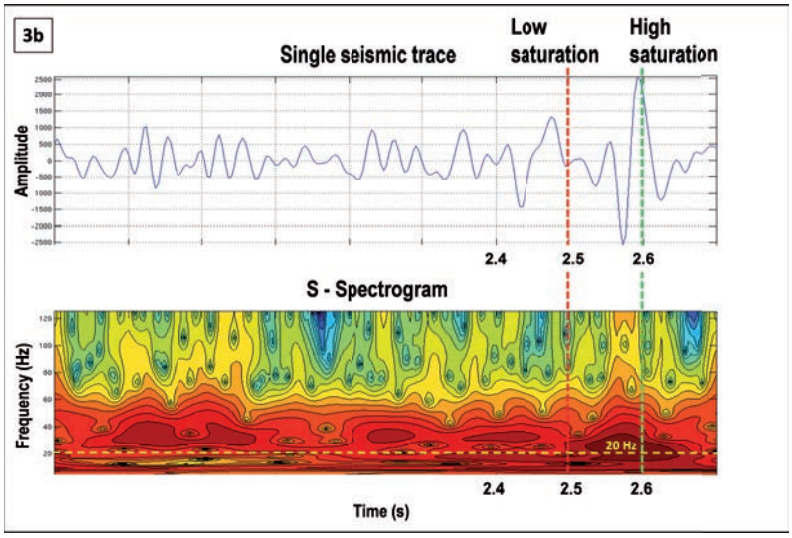
We start the discussion with a relatively small seismic data set. This enabled controlling the classification results vs. the wide range of possible parameters and attributes to select. In the final part of this paper, we discuss an extension of this limited small-scale test to a larger 2D data set. Furthermore, an additional application of our method to an industrial data set, but with different purposes, is discussed in a previous paper (Amendola *et al.*, 2017).

For our test, we selected an ambiguous case drilled by an exploration well where seismic data show two significant reflection events and amplitude-frequency anomalies in correspondence of two reservoir layers (Fig. 2).

Only the lower channel showed high gas saturation, whereas the upper channel showed just low gas saturation (see right panel of Fig. 2). Unfortunately, the distinction between low and high gas saturation effect is ambiguous from the seismic data. These two reservoirs correspond to the reflection events at about 2500 and 2600 ms in the section of Fig. 2. These reflections both show high amplitude. If we analyse seismic data trace by trace, we see that distinguishing high from low gas saturation is difficult in both time and frequency domains. Unfortunately, this is a common problem in the entire considered exploration area, where sand-channels with low gas saturation show relatively high seismic response.

In this initial classification test, we used all the seismic traces sampling the reservoir channels and the clay seals, for a lateral extension of more than 500 m and a time interval of 300 ms. In this paragraph, we focus on just a few individual traces (see Fig. 2 for trace location). As shown in Figs. 3a and 3b, the spectrograms in correspondence of both high and low gas saturation show





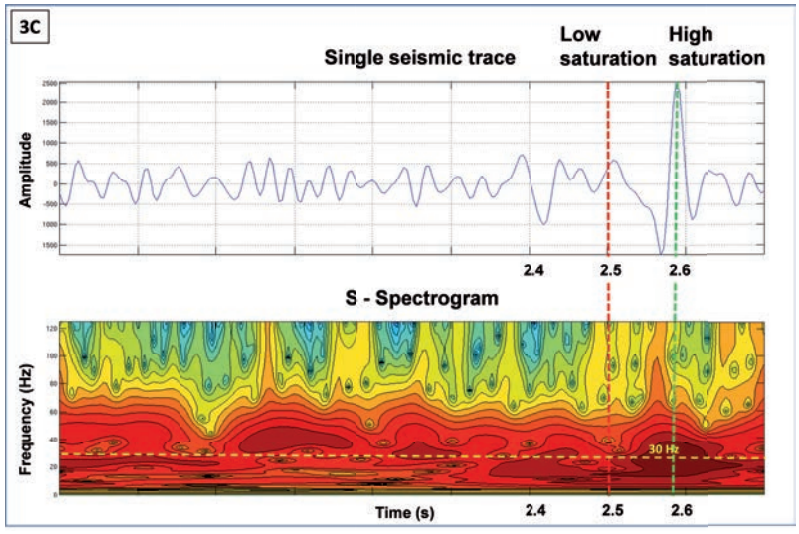


Fig. 3 - Examples of three (panels a, b, and c) seismic traces (upper part of the panel) and their corresponding spectrograms (lower part of the panel) in proximity of the same well showing the time-frequency responses for both low gas and high gas-saturation scenarios. Red and brown colours indicate high seismic amplitude. Vertical bars indicate approximately the core of the amplitude anomaly in correspondence of the two low gas and high gas-saturated layers (see Fig. 2 for the position of the three traces).

a significant low frequency anomaly. Fig. 3a shows that the main anomaly (centred around 30-35 Hz) appears in correspondence of the low gas-saturated layer. Instead, Fig. 3b shows another seismic trace not too far from the previous one (about 100 m), where the main anomaly (centred around 20-25 Hz) appears in the layer with high gas saturation.

The traces of Figs. 3a and 3b show significant differences in terms of relative amplitude in the two channels. In fact, as shown in the seismic section of Fig. 2, the seismic response can be discontinuous due to erosional effects affecting the channels. The clay volume can change at small spatial scale, affecting formational porosity and saturation.

Fig. 3c shows another seismic trace and its spectrogram located towards the left border of the reservoir (see Fig. 2). In this case, the main amplitude anomaly is centred at a prevalent frequency of about 30 Hz in correspondence of the high gas-saturated channel.

In summary, Figs. 3a, 3b and 3c show that the individual traces have variable "behaviours" in the two reservoir layers, even at relatively short lateral distance (in the range of hundreds metres or less). The key message is that, the amplitude of the anomalies and the frequency content might not be sufficiently diagnostic to distinguish high from low gas saturation cases in this complex geological situation.

Thus, we attempted to distinguish these different scenarios using a Machine Learning approach based on "high-level MIDI" attributes. These are mainly derived from patterns of sounds (in terms of MIDI musical notes), and not only from instantaneous spectral features.

To achieve our goal, we used a subset of the seismic data selected near the well as "labelled data" for training the automatic classifiers based on MIDI features. Of course, the reliability of the classifiers and their range of applicability increase if the training phase is expanded to larger data sets. Indeed, we increased the size of the training data set significantly when we expanded the test to the entire seismic section (as discussed in the final part of this paper).

We started defining an initial taxonomy consisting of three main classes: prevalent clay, low and high gas-saturation sands. We defined these three classes with the help of the CPI well logs shown in Fig. 2, based on the measured ranges of porosity, saturation and clay percentage. In our taxonomy, the class "prevalent clay" corresponds to 60-70% of clay percentage. "High saturation and low saturation" classes correspond to sandy channels characterized by the prevalence of sand and relatively small clay percentage (25-30%). In the channels, gas saturation ranges from about 70-75% to 5-10%, respectively. When gas saturation is near zero, the term "brine sands" should be more appropriate. However, we conventionally continue to use the class name "low gas saturation" so as not to complicate the classification taxonomy too much. Only in some figures (Figs. 20 and 21), we use the equivalent terminology "low gas-saturated or brine sands" for that class. We will see that the seismic response is, apparently, very similar in the various types of sands, independently of fluid saturation. This ambiguity makes interpreting the seismic facies and fluid distribution in this area extremely difficult.

Finally, as we can see from the CPI logs, porosity is rather variable, depending on clay percentage (see Fig. 2).

Then we performed the phase of feature extraction. We extracted MIDI features related to pitch, sound intensity, note duration, melodic, harmonic and rhythmic patterns. We started with classification tests based on basic features commonly used in "standard" seismic attribute analysis related to frequency and signal intensity (instantaneous spectral features). We progressively increased the number of the features, including also multi-valued features that have no equivalent "conventional" seismic attributes, like melodic, harmonic and rhythmic patterns.

6.2. MIDI features

Due to the novelty of our approach based on MIDI features, it is necessary to discuss some significant examples of these unusual attributes with additional details. Further descriptions can be found in Dell'Aversana *et al.* (2016b).

We start discussing how MIDI files can be displayed, in order to show their principal features. Fig. 4 shows an example of MIDI display (bottom panel) of the same seismic trace (top panel) shown in Fig. 3b. This lower display is known as "MIDI piano roll". Two among the most relevant MIDI features are shown. These are pitch and "MIDI velocity", related to instantaneous frequency and sound intensity, respectively. In the lower panel, the vertical axis represents the pitch through a virtual keyboard shown on the left boundary of the figure. It is commonly indicated in terms of musical notes. However, it is related to instantaneous frequency and is obtained from the S-spectrogram using Eq. 2. Note that the original frequency content of the seismic data has been transposed into the audible frequency range, so that we can listen to the derived MIDI file using a "sequencer software" package.

The different colours indicate sound intensity associated to the musical notes: red corresponds to high values and blue represents low values. This type of display is a sort of discrete spectrogram where each pixel represents a musical note with a certain sound intensity. The horizontal axis represents the "MIDI execution time" (settable by the user, so that the MIDI file can be listened at desirable execution rate). The direction from left to right corresponds to increasing travel time or, equivalently, increasing depth.

Fig. 5 is the same trace of the previous figure (top panel), but is now represented as a pitch histogram (bottom panel).

In this case, colours represent the variable pitch. The scale on the left shows that a different color is assigned to each musical note (pitch) in every octave. For instance, red is assigned to the musical note *C*, yellow is assigned to the musical note *F*, and so forth. Pitch histogram is a useful MIDI "multi-valued, or high-level, or polyphonic feature" that provides a discrete representation

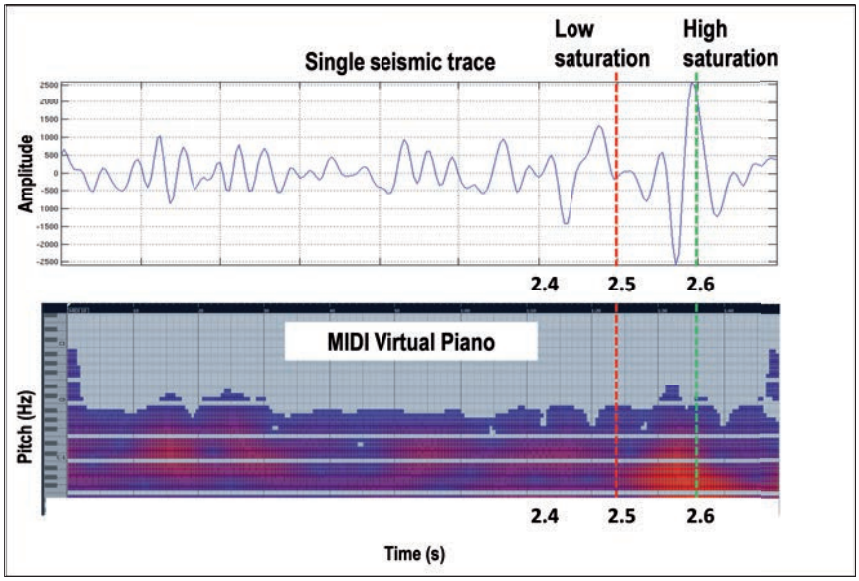


Fig. 4 - Same seismic trace of Fig. 3b (upper panel) and its MIDI piano roll display (lower panel).

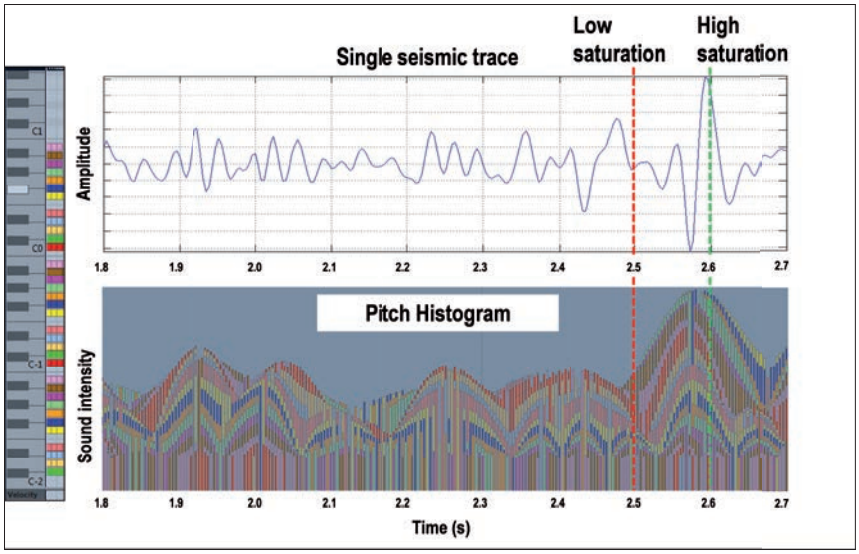


Fig. 5 - Same seismic trace of Fig. 3b (upper panel) and its MIDI pitch histogram (lower panel).

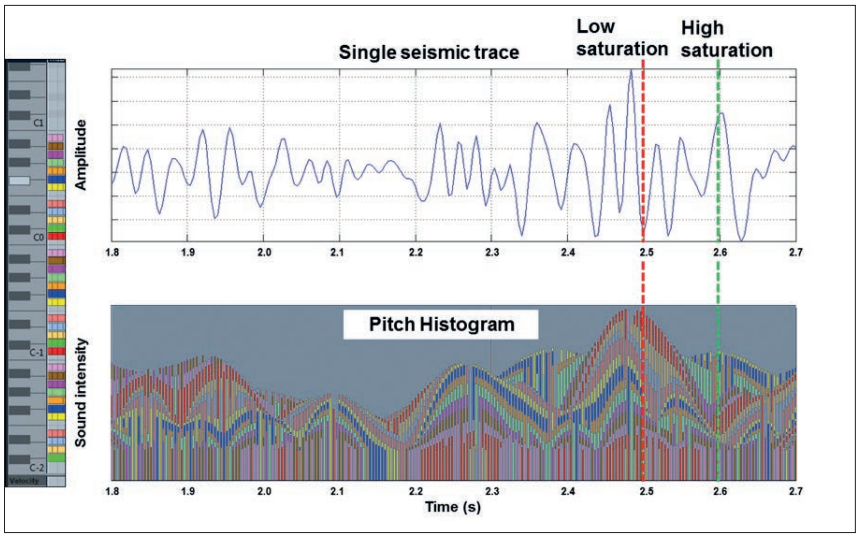


Fig. 6 - Same seismic trace of Fig. 3a (upper panel) and its MIDI pitch histogram (lower panel).

of the frequency content of seismic trace segments. It can be diagnostic for distinguishing different seismic facies. For instance, looking at Fig. 5, we can see that each hydrocarbon-bearing channel (high and low gas-saturated sands, respectively) shows its peculiar pitch histogram. Consequently, it is reasonable to expect that the pitch histogram can contribute to distinguish/classify different targets with different saturations.

For comparison, we report the pitch histogram of the trace of Fig. 3a. In particular, we can note the different "rhythmic" and "harmonic" trends of some pitches in the two distinct reservoir channels, in both Figs. 5 and 6. These variable pitch trends suggest that MIDI features based on ensembles of MIDI notes (often called "high order MIDI attributes") can be useful for automatic classification of different seismic facies. This point will be further exploited and verified in the next sections.

MIDI spectrograms and pitch histograms can be analysed, processed, filtered, combined, mixed, stacked and quickly integrated with other information, using the advanced technology commonly applied in the industry of digital music. Fig. 7 shows a low pass filter of the pitch histogram of the same seismic trace of Fig. 3b, highlighting the low-frequency anomaly in the layer with high gas saturation.

6.3. Classification workflow

After clarifying the basics of MIDI display tools (MIDI piano roll and pitch histogram), we can now describe some details of our Machine Learning classification approach. This is based on the same criteria and workflow commonly used in Musical Genre Classification. We simply adapted the workflow to MIDI files extracted from our converted seismic data. For clarity, we summarize the entire procedure with a synoptic block diagram (Fig. 8).

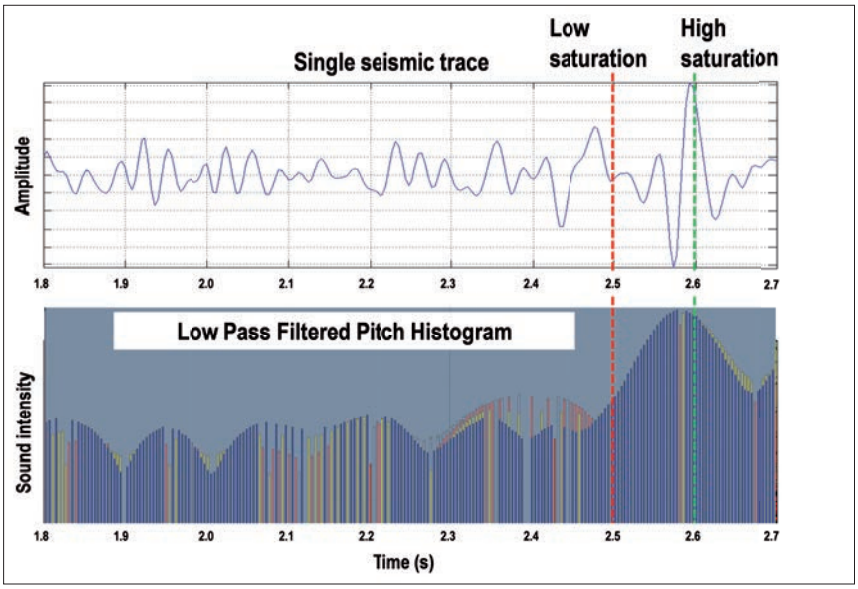


Fig. 7 - Same seismic trace of Fig. 3b (upper panel) and its filtered pitch histogram (lower panel).

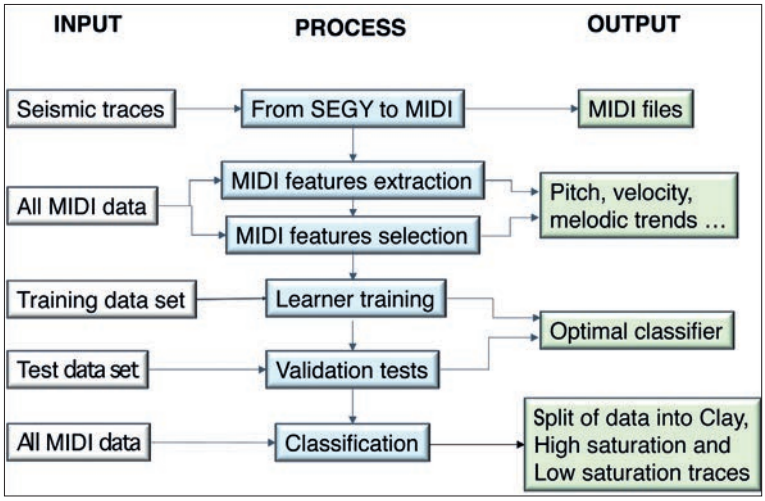


Fig. 8 - Block diagram of the entire classification workflow.

First, we transformed all the original SEGY seismic traces into MIDI files, following the procedure explained above (using the S-Transform). Then we split each MIDI trace into MIDI segments corresponding to 100 ms of the original SEGY file. This size of the time window is adequate for full sampling the seismic events of interest. These effectively happen in time ranges in the order of 50-100 ms (see the marked reflection events in the seismic traces and in their correspondent MIDI displays shown in Figs. 3 to 7). Furthermore, this segment length allows detecting patterns of MIDI notes and extracting high-level MIDI features based on melodic, harmonic, and rhythmic properties from the data. Using smaller time-segments could prevent to extract these attributes with a significant statistical meaning.

We imported all the MIDI files into a software platform of digital music for visualizing, processing and listening to our data (a commercial sequencer). We, then, uploaded all the MIDI files into another platform for MIR and Musical Genre Classification. We extracted the MIDI attributes from our MIDI files and started the phase of training, using a subset of labelled data. We performed the training phase on a fixed number of traces near the wells. A subset of the data was used as "test data set" for estimating the effectiveness of the training (cross-validation tests). At this initial stage, the training data was relatively small: it consisted of about 40 MIDI traces corresponding to seismic segments of 100 ms representative of the different classes ("high gas-saturated sands", "low gas-saturated sands", "prevalent clay"). These samples were extracted from seismic traces around one exploration well in a distance range of about 150 m. However, as explained in the following, we included many additional samples (many hundreds of MIDI files only for the training) when we expanded our approach to the entire seismic section. We discuss the different performance of our classifiers on the training data in the next paragraph, using confusion matrices and various classification indexes.

In our training, we started from the simplest attributes (related to pitch and sound intensity) moving progressively to multi-valued features (melodic, harmonic, and rhythmic patterns). At each trial, we evaluated the relative performance of the classifier(s) for each ensemble of attributes. Our classification process was aimed at classifying the data into three main classes: "low gas-saturation layer", "high gas-saturation layer" and "prevalent clay layer" associated to different seismic segments for each seismic trace. We tested and used a combination of classifiers including K-NN, ANN, SVM, Random Forest, Decision Tree and other methods.

6.4. Statistical analysis of the MIDI attributes

In order to show the different diagnostic power of the various MIDI features, we discuss briefly the results of a statistical analysis of some of the most relevant MIDI attributes extracted from the seismic traces. The physical meaning of the main MIDI attributes will appear progressively clear during the discussion.

First, we verified the statistical distributions of the MIDI features in the two reservoir channels characterized by different gas saturation and in the formation with prevalent clay sealing both reservoirs.

In the following graphs, we plot the normalized values of MIDI features vs. their probability density curves (Botev *et al.*, 2010) observed in the training data set (labelled subset of our data). Plotting the normalized values centred on the zero allows better comparison between different attributes having different physical meaning.

Fig. 9 shows the distribution of the normalized "Most common pitch" of the MIDI notes sampled in the seismic segments related to the two reservoir intervals and in the prevalent clay

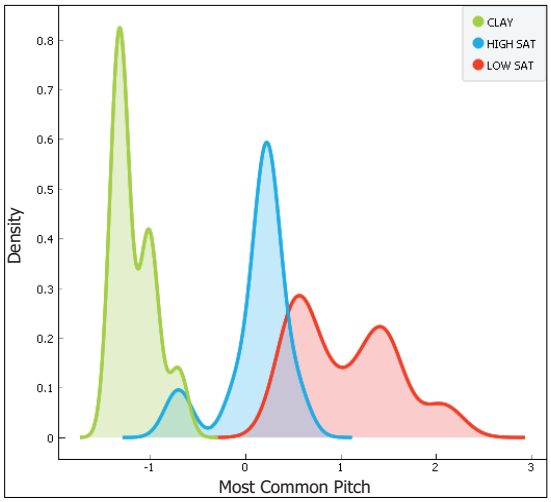


Fig. 9 - Probability density distribution of the normalized "Most common pitch". In this and in the following figures, blue indicates the "high gas-saturated layer" class, red indicates the "low gas-saturated layer" class, green indicates the "prevalent clay layer" class.

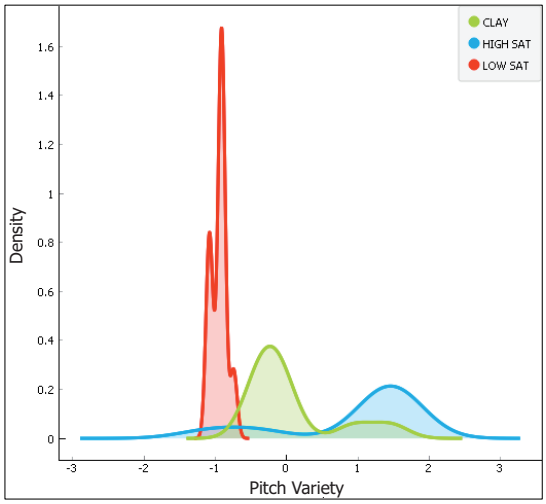


Fig. 10 - Probability density distribution of the normalized "Pitch variety".

formation. We have already clarified that the MIDI pitch is (not linearly) related to the instantaneous frequency of the seismic data through Eq. 2. The "Most common pitch" is calculated as the MIDI pitch value of the most frequent pitch divided by the number of possible pitches in the selected time interval (100 ms in our case).

Despite its relative simplicity, the "Most common pitch" is an important MIDI feature that can contribute, partially, to classify the data. However, we see that there is some ambiguity in separating prevalent low-gas from high gas-saturated layers. Thus, using this feature without any other attribute could lead to an ambiguous data classification. Fortunately, in the practice of Musical Genre Classification, a great variety of independent MIDI features is available. As explained earlier, we can use ensembles of features combining (using simultaneously) many of these attributes for our classification purposes.

Pitch variety is another MIDI feature that takes into account the degree of heterogeneity of pitches used in a fixed time interval. It is a "second-order" feature (derived from other basic features). It shows an interesting diagnostic power for discriminating the different classes, at least in this specific case (Fig. 10).

The MIDI feature called "Variation of Dynamics" shows an additional interesting statistical distribution in our training data (Fig. 11). It is the standard deviation of loudness levels of all MIDI notes in a fixed time interval and is related to variability of seismic signal intensity. From Fig. 11, we can see that this feature allows a good class separation, enabling a clear distinction of low gas saturation class from the other two classes.

Another interesting MIDI feature is "Average note duration" (standard deviation of note durations in a fixed time interval). This MIDI attribute is linked to both the melodic and rhythmic properties of the data. Its statistical distribution (Fig. 12) shows, although there is some overlap between the clay and low saturation classes, that this MIDI attribute can be optimal for clear identification of the high saturation class.

Fig. 13 shows that the MIDI feature called "Rhythmic variability" is complementary to "Average note duration". It is reasonable to expect proper classification of our data set in all of the three classes using these two features together. We will verify this assumption in the next paragraph where we discuss the classification results.

We can extract many rhythmic attributes from MIDI data. Almost all of them can be useful for classification purposes because they enable detecting important patterns in the MIDI data.

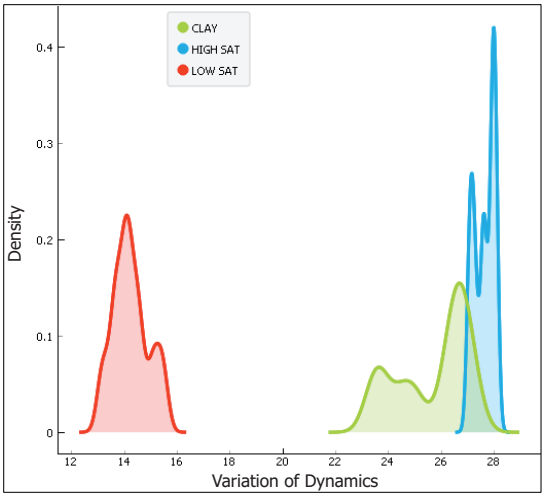


Fig. 11 - Probability density distribution of the normalized "Variation of Dynamics".

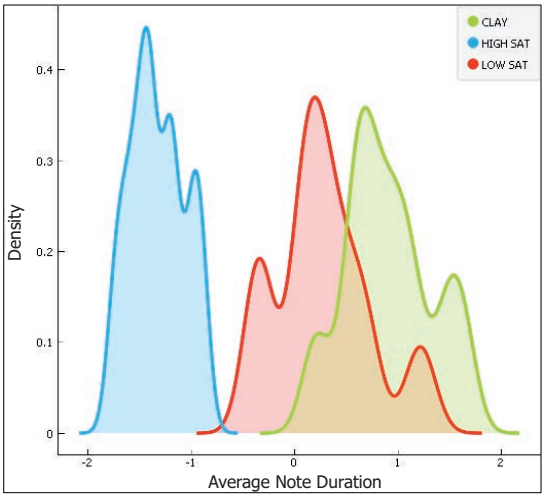


Fig. 12 - Probability density distribution of the normalized "Average Note Duration".

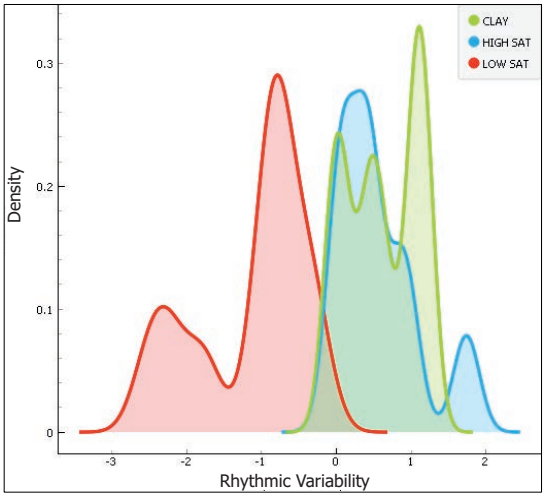


Fig. 13 - Probability density distribution of the normalized "Rhythmic Variability".

They are related to autocorrelation properties of the signal. It is well known that autocorrelation allows comparing a signal with versions of itself delayed by successive intervals (lags). In other words, it allows finding repeating patterns characterized by different periodicities within a signal. In terms of MIDI musical data, the relative strength of different rhythmic pulses can be found by autocorrelation. In our specific case, we used the following autocorrelation function (McKay, 2004):

$$A(lag) = 1/N \cdot \Sigma_{n=0}^{N-1} x(n) \cdot x(n\text{-}lag). \qquad (3)$$

In Eq. 3, n represents the input sample index (in MIDI Tick, that is the MIDI sampling time unit), N is the total number of MIDI Ticks, x is the sequence of MIDI notes and lag is the delay in MIDI ticks ($0 \leq lag < N$).

We could continue with this type of histogram adding similar figures for many other MIDI features, but the message should be clear. Some features associated to pitch, rhythmic, dynamics, harmonic, and melodic properties of the data can cooperate for data classification.

6.5. Training

The analysis of MIDI attributes on labelled traces (training data set) allowed us to optimize the workflow for the classification of unlabelled traces. In fact, using the most sensitive MIDI features, we tested different classification methods and evaluated their effectiveness on a test data set. This process is commonly (but not exclusively) performed using a technique called "Cross-validation test". This is a model validation technique aimed at assessing how the results of a statistical analysis will generalize to an independent data set. It works by partitioning a labelled sample of data into complementary subsets; we perform the classification on one labelled subset (called the training set), and then validate the analysis on the other labelled subset (called the validation set or testing set). In our specific test, we applied several learning algorithms and classification methods including ANNs, Decision Tree, Random Forest, Naïve Bayes, KNN, and SVM. We used all the MIDI features described above (plus other MIDI features not explicitly discussed here), considering their mutual complementarity. However, we verified that some features work more effectively than others.

The effectiveness of the different classification methods can be estimated through the "Confusion matrix" that gives the number/proportion of instances between the predicted and actual class. Fig. 14 shows, for example, the Confusion matrices for four different learner algorithms: Decision Tree, ANN, Random Forest, and SVM. All these matrices provide a quantitative measure of the recognition rate. The principal diagonal of each matrix shows the match (in %) between predicted and actual class.

We can see that the predictions obtained by Decision Tree and ANN produced some mismatch between predicted and actual values. We verified through many tests (not discussed here) that classification results improve if we set (by trial and error) the parameters properly in the ANN algorithm applied for the training (such as the momentum and the learning rate).

However, at least for this specific test, Random Forest and SVM seem to work more efficiently than the other learners do. Their performance depends on the MIDI features used.

In order to provide a complete assessment of the quality of the classification results, Table 1 shows two quantitative indexes of the classification performance and reconstruction rate for each individual method (including other methods used in our test and not mentioned in Fig. 14). "Classification accuracy" (CA) is the proportion of correctly classified examples. "Precision" is the proportion of true positives among instances classified as positive.

Fig. 14 and Table 1 show very good performance, especially for Random Forest and SVM methods. These results might appear rather suspicious, due to overfitting/overtraining problems. The *k*-fold validation method we used helped limit this type of problem. As confirmed by various authors (McKay, 2004), generally a value of $k = 10$ (where k is the number of folds in which we split the data) is what is suggested to minimize the over-fit and minimize the bias in training the population selection. However, in our cross-validation tests, we tested several values for k (from 2 to 10), using different percentages of data for training (from 20 to 66%).

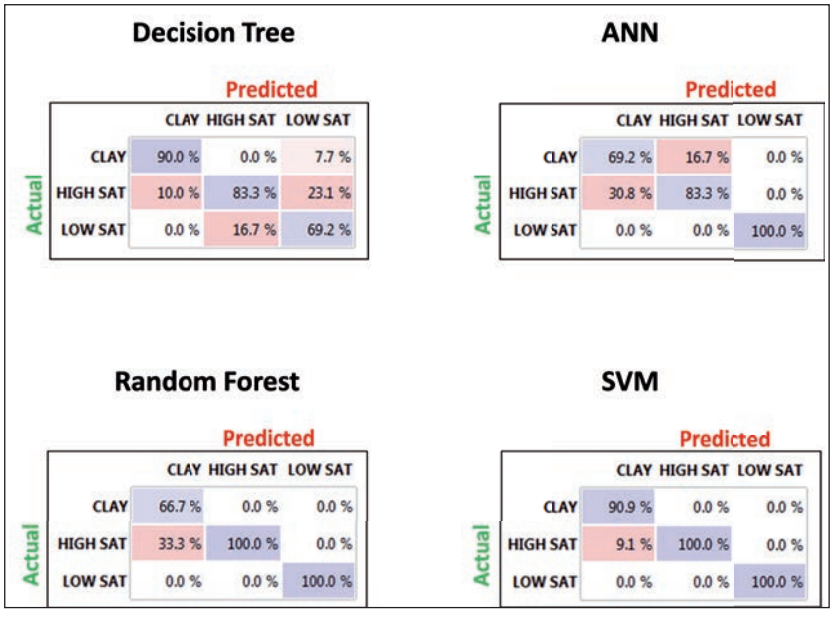


Fig. 14 - Confusion matrix showing the recognition rate for four different learner algorithms: Decision Tree, ANN, Random Forest, and SVM.

Table 1 - Indexes of the classification performance for the various methods.

Method	CA	Precision
Decision Tree	0.793	0.808
ANN	0.828	0.842
AdaBoost	0.862	0.901
Naive Bayes	0.931	0.931
Random Forest	0.966	0.969
SVM	0.966	0.969

6.6. Classification

After identifying the most promising learners and the most sensitive MIDI features on the training data set, we moved to the final part of the workflow, the classification of unlabelled data.

Naturally, we focused our efforts on the learners showing the best performance on the labelled data (such as SVM and Random Forest). However, we did not neglect analysing the results of the remaining learners too (such as Decision Tree and Naïve Bayes). As an example of our classification results, we show (in Fig. 15) the map of the different classes plotted in a two-dimensional feature space applying the SVM learner and the MIDI features of normalized "Average note duration", "Rhythmic variability" and "Average melodic interval".

The classification maps in different two-dimensional feature spaces shown in Fig. 15 are encouraging. They tell us that the selected MIDI features are able to drive the SVM learner towards a proper classification of our unlabelled data. However, there are a few points that seem to fall in the wrong domains. For instance, in both panels we can see red samples classified as "low saturation" falling in the green domain of the "clay" class.

The classification result improves if we project the data into a new 2D feature space obtained through Principal Component Analysis (Fig. 16). In this new transformed algebraic space, the three classes are better separated, although a few samples appear isolated in the top of the image. It is possible that they belong to a fourth category that is different from the three classes assumed in our initial taxonomy. That observation will motivate us to extend the test and include an additional class in our taxonomy. We discuss this test extension in the final part of the paper.

Finally, Figs. 17, 18, and 19 show some examples of normalized MIDI attributes for the classified MIDI files. They are plotted along the reservoir distance in each one of the three layers (low gas-saturated reservoir at the top, prevalent clay in the centre, and high gas-saturated reservoir at the bottom). These figures help understand the relative values of the MIDI features and how they are complementary for distinguishing the three classes. We notice some apparent outliers, like for instance the large yellow circle at the right end of the clay layer in Fig. 17. These points could be explained as noisy or misclassified data. In fact, we assumed only three classes in our taxonomy. However, each layer could show internal geological variations. Consequently, the number of classes explaining the data could be higher than three (see next paragraph).

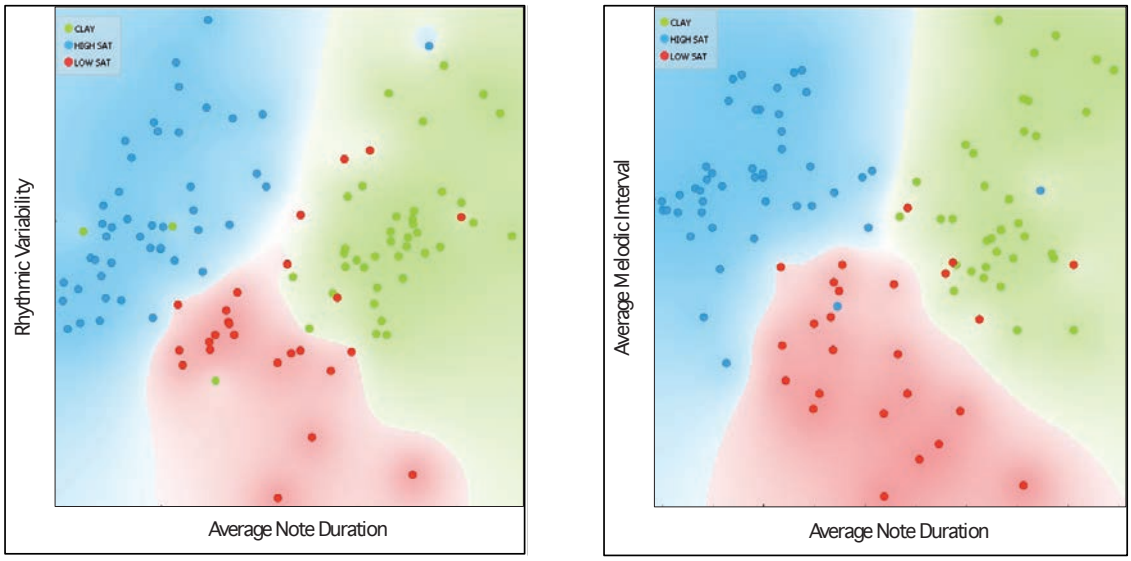


Fig. 15 - Map of classification results using the SVM learner and two different couples of MIDI features: "Average Note Duration" and "Rhythmic Variability" (left), "Average Note Duration" and "Average Melodic Interval" (right). Each unlabelled MIDI time-segment (coloured circles) is plotted in the two-dimensional space at a location defined by the values of its MIDI features. We assigned a different colour to each class.

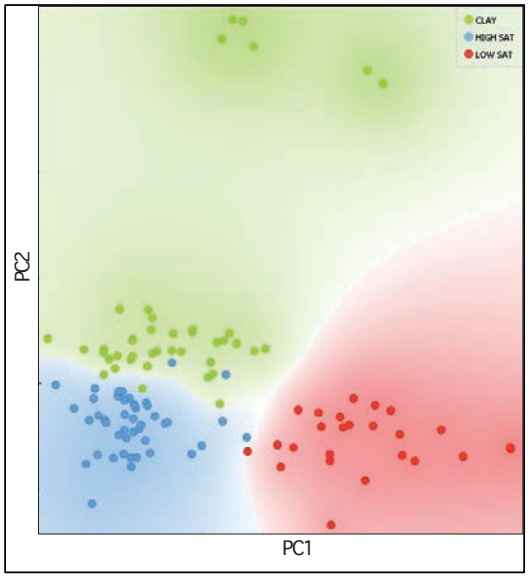


Fig. 16 - Principal Component Analysis map. The first two principal components (PC1 and PC2) allow a clear separation of the traces in three distinct classes. Each MIDI time-segment is plotted in the new two-dimensional space (coloured circles) at a location defined by the two principal components PC1 and PC2.

7. Extending the test to the entire section

After testing our approach within a limited portion of the seismic section (showed in Fig. 2), we extended our classification approach to the entire section. We increased significantly both the training and the unlabelled data sets (in the order of several thousands of MIDI files for the last one). Consequently, although this data set is not as big as a typical "industrial data set", it has a significant statistical value for validating our approach.

We applied the same workflow described in the previous sections, but this time we extended

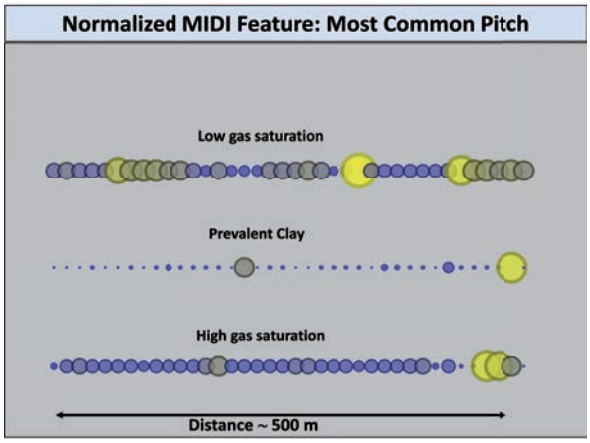


Fig. 17 - Normalized Most Common Pitch for the classified MIDI files along the reservoir distance. The size of the symbols is proportional to the normalized value of the attribute. The colour scale is set in such a way that the normalized values increase from blue to yellow.

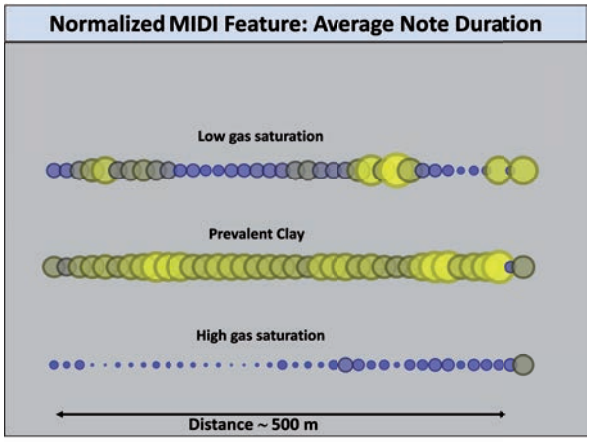


Fig. 18 - Normalized Average Note Duration for the classified MIDI files along the reservoir distance. The size of the symbols is proportional to the normalized value of the attribute. The colour scale is set in such a way that the normalized values increase from blue to yellow.

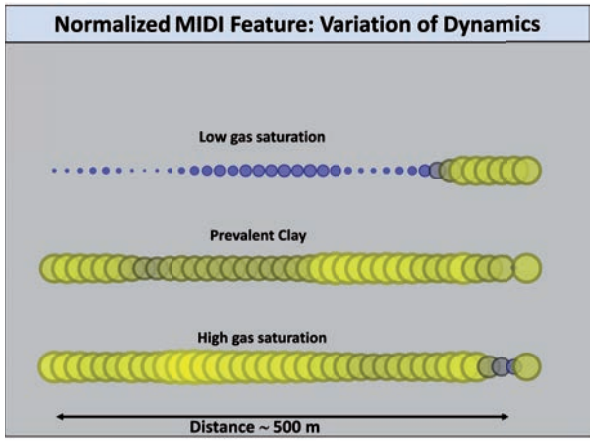


Fig. 19 - Normalized Variation of Dynamics for the classified MIDI files along the reservoir distance. The size of the symbols is proportional to the normalized value of the attribute. The colour scale is set in such a way that the normalized values increase from blue to yellow.

our analysis far away from the well location, in a distance range of about 5 km. Fig. 20 shows the results of our analysis in a time interval of 600 ms centred at target depth (see the rectangular shadowed area in Fig. 20). Fig. 21 shows a zoom of the same results qualitatively co-rendered with the same portion of the seismic section of Fig. 2.

In this test, we extended our analysis to include an additional class called “Shale rock” in our classification taxonomy. The choice was driven by the analysis of the CPI in other wells in an adjacent area. Furthermore, as mentioned, the results of the small-scale test discussed in previous

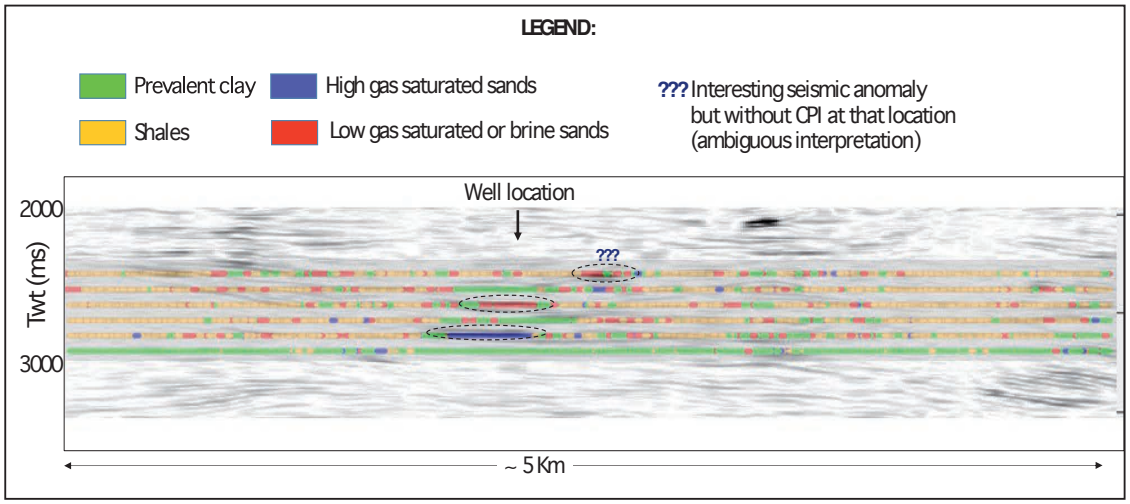


Fig. 20 - Classification results using Random Forest method co-rendered with the entire original seismic section (this is displayed in black and white, in the background, to better highlight the classification results). In the shadowed area, colours mark different classes identified through our MIDI based approach. We recall that we segmented the seismic traces into time windows of 100 ms. For each trace, the coloured symbols are positioned in the middle of the correspondent time segment, in order to allow a qualitative co-rendering of the classification results with the seismic section.

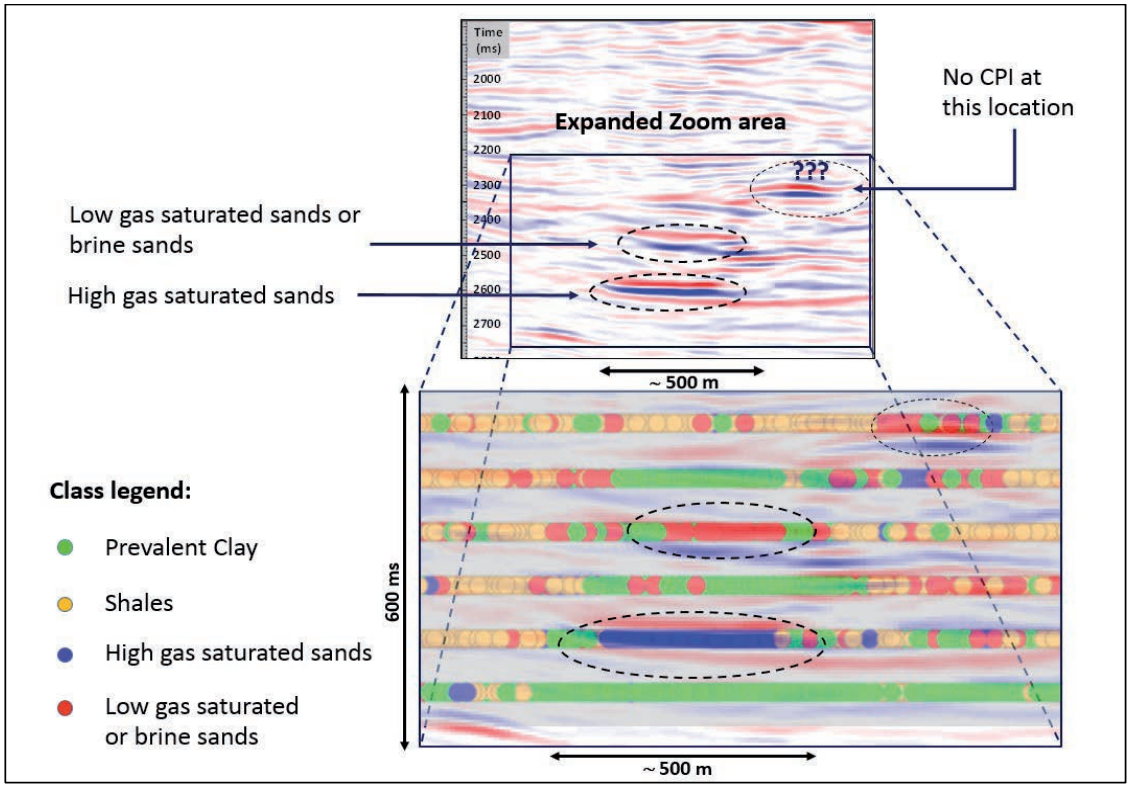


Fig. 21 - Zoom of the same classification results of Fig. 20, qualitatively co-rendered with the same portion of the seismic section of Fig. 2 (zoom area). Similar to Fig. 20, the continuous blue and red sequences of symbols inside the ellipses mark the two high and low gas-saturated sand channels, respectively, drilled by the exploration well. The uppermost ellipse highlights a further seismic event classified as "low gas-saturated or brine sand", through our MIDI-based approach and without the support of CPI.

sections, suggested the possibility of a fourth class for clustering properly the data. This new class is referred to a lithological type characterized almost exclusively by clay (percentage of clay greater than 80-85%). Together with the other class called "Prevalent clay", it takes into account the variable content of clayey rocks forming the geological background.

Both Figs. 20 and 21 show just one example of classification using the Random Forest method, co-rendered with the seismic section. However, it is worth noting that in this application (expanded to the entire section) we also used the other classification methods, like ANN, SVM, Decision Tree, and so forth. Every segment of 100 ms of each seismic trace is classified into one of the four classes; then it is marked with a specific colour and superimposed on the seismic section.

Looking at Figs. 20 and 21, the first ellipse from the bottom includes the zone that corresponds to the high gas-saturated sandy channel drilled at about 2600 ms. The ellipse immediately above, marks a "continuous red area" corresponding to the low gas-saturated sandy channel drilled by the same well. Both channels gradually change (laterally) into prevalent clay/shale formation (where symbols change from red to green/orange). These two channels are the same as those shown in the section of Fig. 2 (see Fig. 21 for comparison).

Our classification results are geologically reasonable and are in good agreement with the well logs. In fact, looking at the colours explained in the legend of Figs. 20 and 21, we can recognize the sedimentary series encountered by the well in the time interval 2200-2800 ms (consisting of the two sandy reservoirs with different gas saturation and their respective clayey seals, as shown in the CPI of Fig. 2). Observing the complete section, we see that there are many areas marked as "prevalent clay" (green), "shale" (orange) and "low gas-saturated sands or brine sands" (red). Other wells drilled in the adjacent area effectively demonstrate that there is a large diffusion of shale and sandy formations characterized by variable percentage of clay, and showing variable gas saturation. In many cases, gas saturation is near zero, and it would be more appropriate to speak of "brine sands". However, we did not introduce that further classification term as a separate class, so as not to overly complicate the taxonomy.

The irregular spatial distribution of these formations properly reflects the true geological/sedimentary complexity of the area and the consequent irregular seismic response. We remark that in such a complex scenario, our approach has a predictive value. For instance, there is a clear seismic event included in the uppermost ellipse and highlighted by the question marks in both Figs. 20 and 21 (it is much clearer in the last figure). It has the same seismic appearance as the other reflection event included in the first ellipse from the bottom. Consequently, its interpretation could be "high gas-saturated sands". Unfortunately, we do not have any CPI information to calibrate this further seismic reflection. Thus, the interpretation would remain ambiguous, unless we do not use our MIDI-based approach. In fact, as shown in Figs. 20 and 21, our method allows classifying that event as a probable "low gas-saturated or brine sand", using the Random Forest method (and/or other classification methods). This is an example of how the MIDI attributes can point the interpretation of seismic facies and of fluid distribution towards probabilistic solutions even in very difficult cases.

8. Open questions and work-hypotheses

The statistical analysis and the classification workflow discussed in this paper is based on a completely new category of seismic attributes derived from SEGY data transformed into MIDI

files. This is an unusual approach for clustering, classifying and interpreting geophysical data. Consequently, several open questions and doubts can emerge.

These questions may be divided in two main categories, depending if they are focused on methodological or physical aspects.

The first methodological perplexity can derive from the fact that information of a data set does not change under domain transformation. Hence, quantitative interpretation of the post-stack data in time-domain should not be different from an interpretation in the "Stockwell domain" or in the "MIDI domain".

A second question is why high-level MIDI features, such as rhythm and melody, should improve the performance of the classification with respect to a learner trained only with attributes based on instantaneous properties of the spectrogram.

A third doubt about our methodology is related to the effectiveness of MIDI pitch-histogram. This is nothing other than a coded representation of the S-spectrogram. Why should the first type of representation work better than the second one?

The above questions are correlated in some way. First, it is worth noting that transforming seismic data into MIDI notes does not create any new information. Indeed, this is not our intention. On the other hand, MIDI files have the advantage of a different representation of the same data. Such a new representation can be useful for many reasons. First because MIDI files transform the spectrograms into discrete entities (generally called MIDI events). These can be easily codified into new types of "second order attributes". Consequently, data can be represented in the form of patterns (ensembles of MIDI notes and chords). In such a way, signal patterns, and not only local signals, can be highlighted and used for automatic classification. Rhythmic, melodic, and harmonic properties implicitly contained in the seismic data can appear explicitly and can be used as discriminant features, as shown in our examples and in many figures of this paper. All the statistical distributions of Figs. 9 to 13 show the complementarity of the many available MIDI features. These figures explain why high-level MIDI features, such as rhythm and melody, should improve the performance of the classification with respect to a learner trained, for instance, exclusively on the instantaneous frequency. Our ranking analysis performed on many MIDI features confirms that these "ensemble" properties of the data (associated to patterns of notes), can classify data better than features based exclusively on "local" spectral properties. This is one reason why many experts of Musical Genre Classification and MIR prefer using ensembles of "high-level" MIDI features together with "low-level" attributes extracted directly from the spectrograms (McKay, 2004).

Of course, MIDI is not the only type of discrete representation of time series. We use MIDI because it is a consolidated standard protocol in the domain of digital music. Consequently, after transforming the geophysical data into MIDI files, we can use an extremely large library of algorithms, tools, software platforms, codes, and methods successfully applied in music (such as in Musical Genre Classification and in MIR).

Finally, MIDI files can be listened to, adding a new perception in the domain of geophysical data analysis. This is not a secondary aspect of our approach, as is widely discussed in previous papers (Dell'Aversana *et al.*, 2016a).

An additional open question, focused on physical/geological aspects, is why high and low saturated intervals should have different rhythmic variability, sound intensity or pitch. The question can be reformulated more explicitly: "there are many rock-physical models explaining

why gas-saturated intervals produce low frequency anomaly. On the other hand, it is unclear why high and low saturated intervals should have different rhythmic variability, sound intensity or pitch".

We suggest the hypothesis that different rhythmic (and even melodic/harmonic) properties in the seismic response can reflect the variability of different sedimentary sequences (such as a sequence of sand and clay). These, in turn, influence porosity and fluid distribution in the formation. Consequently, it is reasonable to expect some type of correlation between sedimentary features, percentage of clay and sand, gas saturation and high-level MIDI attributes. For sure, this hypothesis requires further investigation through modelling, lab and field experiments. However, the fact is that the above-discussed MIDI attributes are very well correlated with the degree of gas saturation in the specific data set here analyzed.

9. Conclusions and future steps

In this paper, we introduced a novel methodology addressed to automatic classification of seismic facies and to distinguish between low and high gas-saturated sand channels. It is based on combination of several classifiers such as ANNs, Decision Tree, Random Forest, SVM, and other methods. The particularity of our approach is that the classifiers are applied to MIDI features extracted from seismic data converted into musical formats. This methodology allowed us to use, for our classification, a set of new features that have no equivalent in the seismic attribute domain. We tested our approach on a real data set consisting of seismic traces related to clay and shale formations, to low gas-saturated and high gas-saturated sandy reservoirs. We verified that the MIDI attributes (pitch, melodic, harmonic, and rhythmic features) help distinguishing the data in separate seismic facies, consistently with the well data.

In conclusion, the classification of seismic data based on Machine Learning applied to a "light" symbolic format like MIDI, seems to work effectively for discriminating seismic facies, even in terms of fluid distribution. However, we should point out that the final classification results depend on a multitude of parameters. Particularly important is the training phase on the labelled data set. Wrong training commonly leads to wrong classification performance, independently of the effectiveness of the MIDI features in clustering the data. Of course, we have the possibility to use unsupervised classification approaches; however, in this first experimental phase of our method, we preferred to classify our data with the help of labelled training data sets. Indeed, we had labelled data to use in our test, due to the availability of well data in this area for calibrating the seismic response.

Finally, almost 100% of the traces have been correctly divided into their respective classes (as expected from well data). Although our test is confined to a relatively small seismic data set, it is focused on a relevant seismic interval including key vertical and lateral changes. For that reason, we think that our test produced statistically significant and encouraging results about the effectiveness of our approach. Considering the novelty here introduced, additional investigations are of course necessary for a complete understanding of the physical meaning of the new "high-level" MIDI attributes. We are working on industrial data sets and applying our MIDI-based classification method to 2D and 3D data with encouraging results.

Acknowledgments. The authors acknowledge the technical and scientific contribution of Alfonso Iunio Marini, Massimo Fervari and Andrea Gennari, for the many interesting and fruitful discussions about seismic attributes and geophysical data interpretation.

REFERENCES

Amendola A., Gabbriellini G., Dell'Aversana P. and Marini A.J.; 2017: *Seismic facies analysis through musical attributes*. Geophys. Prospect., doi:10.1111/1365-2478.12504.

Aminzadeh F. and de Groot P.; 2006: *Neural networks and other soft computing techniques with applications in the oil industry*. EAGE Publications 129, 164 pp.

Batzle M. and Wang Z.; 1992: *Seismic properties of pore fluids*. Geophys., **57**, 1396-1408.

Botev Z.I., Grotowski J.F. and Kroese D.P.; 2010: *Kernel density estimation via diffusion*. Ann. Stat., **38**, 2916-2957, doi:10.1214/10-AOS799.

Castagna J.P. and Sun S.; 2006: *Comparison of spectral decomposition methods*. First Break, **24**, 75-79.

Castagna J.P., Sun S. and Siegfried R.W.; 2003: *Instantaneous spectral analysis: detection of low-frequency shadows associated with hydrocarbons*. The Leading Edge, **22**, 120-127.

Dell'Aversana P.; 2013: *Listening to geophysics: audio processing tools for geophysical data analysis and interpretation.* The Leading Edge, **32**, 980-987, doi:10.1190/tle32080980.1.

Dell'Aversana P.; 2014: *A bridge between geophysics and digital music. Applications to hydrocarbon exploration*. First Break, **32**, 51-56.

Dell'Aversana P., Gabbriellini G. and Amendola A.; 2016a: *Sonification of geophysical data through time-frequency transforms*. Geophys. Prospect., **65**, 146-157.

Dell'Aversana P., Gabbriellini G., Marini A.I. and Amendola A.; 2016b: *Application of Musical Information Retrieval (MIR) techniques to seismic facies classification. Examples in hydrocarbon exploration*. AIMS Geosci., **2**, 413-425, doi:10.3934/geosci.2016.4.413.

Klimentos T.; 1995: *Attenuation of P- and S-waves as a method of distinguishing gas and condensate from oil and water*. Geophys., **60**, 447-458.

Kumar G., Batzle M. and Hofmann R.; 2003: *Effect of fluids on attenuation of elastic waves*. In: Expanded Abstracts 73th Ann. Int. Mtg., Soc. Explor. Geophys., pp. 1592-1595.

McKay C.; 2004: *Automatic Genre classification of MIDI recordings.* Thesis submitted June, Music Tech. Area Dept. of Theory, Faculty of Music, McGill Univ., Montreal, Canada, 150 pp.

O' Brien J.; 2004*: Seismic amplitudes from low gas saturation sands.* The Leading Edge, **23**, 1236-1243.

Stockwell R.G., Mansinha L. and Lowe R.P.; 1996: *Localization of the complex spectrum: the Stransform*. IEEE Trans. on Signal Process., **44**, 998-1001, doi:10.1109/78.492555.

Tai S., Puryear C. and Castagna J.P.; 2009: *Local frequency as a direct hydrocarbon indicator*. In: Expanded Abstracts 79th Ann. Int. Mtg, Soc. Explor. Geophys., pp. 2160-2164, doi:10.1190/1.3255284.

Wang Y.; 2007: *Seismic time-frequency spectral decomposition by matching pursuit*. Geophys., **72**, V13-V20.

Winkler K. and Nur A.; 1982: *Seismic attenuation: effects of pore fluids and frictional sliding*. Geophys., **47**, 1-15.

Corresponding author: Paolo Dell'Aversana
Eni S.p.A. Upstream and Technical Services
Via Emilia 1, 20097 San Donato Milanese (MI), Italy
Phone: +39 02 52063217; e-mail: Paolo.Dell'Aversana@eni.com