DOI 10.4430/bgta0325

An integrated multi-physics Machine Learning approach for exploration risk mitigation

P. Dell'Aversana

Eni S.p.A., Upstream and Technical services, San Donato Milanese, Milano, Italy

(Received: 17 December 2019; accepted: 10 March 2020)

ABSTRACT In this paper, we combine pre-stack depth migration of seismic data, cooperative modelling of Controlled Source Electromagnetic (CSEM) and gravity data, and constrained inversion of CSEM data, with Machine Learning (ML) classification approaches. Our objective is to obtain probability maps of hydrocarbon distribution aimed at mitigating the exploration risk and supporting the process of appraisal of hydrocarbon fields. We introduce a novel workflow divided into two linked branches: one consists of an iterative loop of modelling and inversion steps driving towards a multi-parametric Earth model; the other path of the workflow goes through the application of advanced statistical tools and takes the benefits of automatic learning and classification algorithms. These allow us combining the entire set of heterogeneous data/models into probabilistic maps of oil distribution at target depth. We applied our methodology to a complex data set in the Norway offshore, obtaining encouraging results.

Key words: geophysical data integration, Machine Learning, marine CSEM, gravity.

1. Introduction

The process of geophysical data integration involves several interrelated methods aimed at retrieving multi-parametric Earth models from multi-disciplinary measurements (Dell'Aversana, 2014). This process can be realised through cooperative modelling, single-domain, constrained, sequential and simultaneous joint inversion approaches. More frequently, the final Earth model is obtained through a combination of modelling and inversion approaches rather than using just one type of algorithm individually. This "systemic approach" is commonly implemented in a unique software platform where geophysicists and geologists can create interactively a complex integration workflow (Dell'Aversana *et al.*, 2016). In principle, such an integrated Earth model can significantly mitigate the exploration risk. However, especially in complex geological settings, interpreting simultaneously data and model parameters belonging to different geophysical domains can be difficult. For instance, it happens frequently that seismic, gravity and electromagnetic data are sampled with different acquisition density. Furthermore, they commonly show different intrinsic resolution. Consequently, the corresponding models (such as velocity, density and resistivity models) can be combined only marginally and with significant ambiguities and uncertainties. Finally, the Earth models obtained through modelling and/or inversion, are frequently non-deterministic; they often represent the solution of a stochastic inversion process with variable ranges of uncertainties in the model parameter distribution (Tarantola, 2005). In

© 2020 – OGS

summary, integration in geophysics generally represents a multi-physics, multi-scale statistical inverse problem.

Nowadays, such a difficult problem can be approached with the support of Machine Learning (ML) algorithms. Over the past few years, ML has radically changed many scientific sectors and even the daily routine of most of us. Self-driving cars, speech recognition, effective web search, and an improved understanding of the human genome, represent just few examples of the practical impact of ML on our life. This is "the subfield of computer science that gives computers the ability to learn without being explicitly programmed" (Samuel, 1959). Statistical (or mathematical) techniques are applied for retrieving a model from observed data, rather than codifying a specific set of instructions that define the model for that data (Bishop, 2006). There are different ML paradigms. In the case of supervised learning algorithms, ML techniques are used to train on model examples. In other words, previously unseen data can be classified using the rules generated during training on labelled data. Instead, unsupervised learning algorithms cluster the data based on similarities rather than model categories. There is an additional ML approach based on the concept of "reinforcement learning". This paradigm of learning is based on trial-and-error, and on some codified form of "rewards" or "punishments". The machine improves its performance using the feedback coming from the "external environment". Like humans, software agents learn for themselves to achieve successful strategies that lead to the greatest long-term rewards.

Nowadays, the various ML methods are massively applied in many sectors, such as medical, social and financial disciplines. The number of applications of ML has been growing over the past 10-15 years impressively in geosciences too, including geophysics (Aminzadeh and de Groot, 2006). Examples of applications are seismic facies recognition and classification, automatic interpretation of geophysical data, well log analysis, and so forth. Of course, ML can find useful applications in integration of multidisciplinary information, like seismic, electromagnetic, gravity, and magnetic data.

In this paper, we introduce an approach aimed at maximising the benefits of integrated geophysical models and ML methods. The idea is to combine big and heterogeneous data sets, in both data space and model space, using advanced techniques commonly applied in the domain of Data Science. The main objectives are mitigating the exploration risk and supporting the process of field appraisal.

2. Combining multi-physics and Machine Learning

2.1. Multi-domain measurements and multi-attributes

Characterising geophysical data in terms of attributes linked with relevant physical properties is a well-consolidated approach. For instance, in the seismic domain, there are many attributes based on properties like mean amplitude, average energy, frequency and AVO trends, coherence, dip, curvature, and so forth (Castagna and Backus, 1993). Geophysical attributes are used in electromagnetic applications too, such as in Marine Controlled Source Electromagnetic (Marine CSEM) methods and in Magnetotellurics (MT). In the first case, the attenuation of the electric and magnetic fields with offset represents important information depending on the 3D distribution of electrical resistivity (and on frequency). Consequently, the trend of electric and magnetic amplitude/phase can provide diagnostic information about the possible presence of hydrocarbon-filled reservoirs. In MT, simultaneous measurements of orthogonal components of the electric and magnetic fields are recorded (Cagniard, 1953). These data are used to calculate the impedance tensor. This is inverted for retrieving a model of electric resistivity. Furthermore, many MT attributes can be derived from the impedance tensor, providing useful insight about the Earth resistivity even before performing any inversion (Vozoff, 1990).

Also in gravity and magnetic methods, the data are commonly interpreted using a similar attribute-based approach (Fairhead, 2015). Measurements of gravity provide information about densities of rocks. In fact, density variations within the Earth caused by geological variations result in gravity anomalies. These can be highlighted through various attributes, spatial derivatives and/or filters of the field measurements. For instance, the first vertical derivative of the Bouguer anomaly, represents a high pass filter of the original measurements that allows detecting density anomalies in the subsoil at relatively small spatial scale.

More in general, the density distribution in the Earth's interior can be determined with improved accuracy by measuring directly the spatial derivatives of the gravity vector (Condi and Talwani, 1999; Fairhead, 2015). This type of measurement is called gravity gradiometry which measures the variations in the acceleration due to gravity. Gravity gradiometry has predominantly been used to image subsurface geology as a valid support for hydrocarbon and mineral exploration. The most frequently used and intuitive component is the vertical gravity gradient, G_{zz}, which represents the rate of change of vertical gravity (g_z) with height (z). Full-tensor gradiometers measure the rate of change of the gravity vector in all three perpendicular directions (the gravity gradient tensor). The unit of gravity gradient is the Eotvos, which is equivalent to 10^{-9} s^{-2} (or 10^{-4} mGal/m). The derivatives of gravity represent the spectral power of the gravity gradient signals. In other words, they provide higher-frequency information about density distribution. For this reason, the gravity gradient anomalies are generally more localised to the source than the gravity anomaly.

2.2. The workflow

When large multidisciplinary geophysical data are available, combining many attributes generally requires high computational efforts, technical resources and professional skills. Our approach is aimed at supporting the integration process through application of algorithms and procedures commonly used in Advanced Analytics of "Big Data" and in the field of ML.

The workflow, summarised in Fig. 1, is divided into two complementary branches.

The left path of the workflow is performed through an optimised combination of iterative modelling, constrained, cooperative and joint inversion algorithms. This approach is applied using an integrated software platform that includes migration of seismic data and algorithms for modelling/inversion of seismic and non-seismic data. We previously called this platform "Quantitative Integration System", or briefly QUIS (Dell'Aversana, 2014; Dell'Aversana *et al.*, 2016). The objective of this part of the workflow is to derive, gradually, a multi-physics Earth model honouring complementary geophysical observations. When possible, we apply Simultaneous Joint Inversion. In general, this approach is very demanding from a computational point of view. Thus, it is generally limited to the reservoir layer(s), with the objective to retrieve relevant cross-properties such as porosity and fluid saturation (Dell'Aversana *et al.*, 2011). This part of the workflow produces a multi-physics layered Earth model, characterised by a set of properties like seismic velocity, resistivity, density and, eventually, porosity, fluid saturation, and so forth. The dashed arrow in Fig. 1 indicates that this part of the workflow can be characterised by an iterative circularity. This loop can be caused by possible feedback from one geophysical domain to another. In fact, the result of inversion/modelling obtained in one domain produces frequently an update into different geophysical domains.

The "right path" of the workflow consists mainly in the application of statistical, analytical, and classification tools. It starts from applying advanced analytics tools for statistical analysis of the various data sets and their mutual correlations. Then, the workflow continues by extracting geophysical "features" and organising them in the same format. These features can belong to seismic, gravity, magnetic, EM domains, in both data and model space. The red arrow in Fig. 1 indicates that the geophysical parameters, obtained through the left branch of the workflow, are treated as geophysical features in the ML process. All of them, together with other attributes belonging to the data space, concur to form a "global feature matrix". It includes many types of geophysical attributes, such as seismic amplitudes, gravity and electromagnetic attributes, spectral properties extracted from the data, seismic velocities, resistivity, and density. In many applications, it is better to normalise the features' values, in order to bring the different attributes into a comparable scale. The next step is defining a training data set consisting of labelled data. For instance, we can calibrate the multi-physics measurements in correspondence of wells, if available. These labelled data sets are used for training the "learner algorithms". In our workflow, we apply and compare many different types of classification methods, including Deep Neural Networks, Support Vector Machine, Random Forest, Bayesian Networks, and other algorithms. Finally, we select the approaches that produce the most reliable results, depending on the specific classification/prediction problem, on data quality, and on geological complexity. The effectiveness of the classification approach is properly tested using well-known methods, like cross-validation techniques, quantitative performance indexes, and confusion matrices. As shown in this paper,

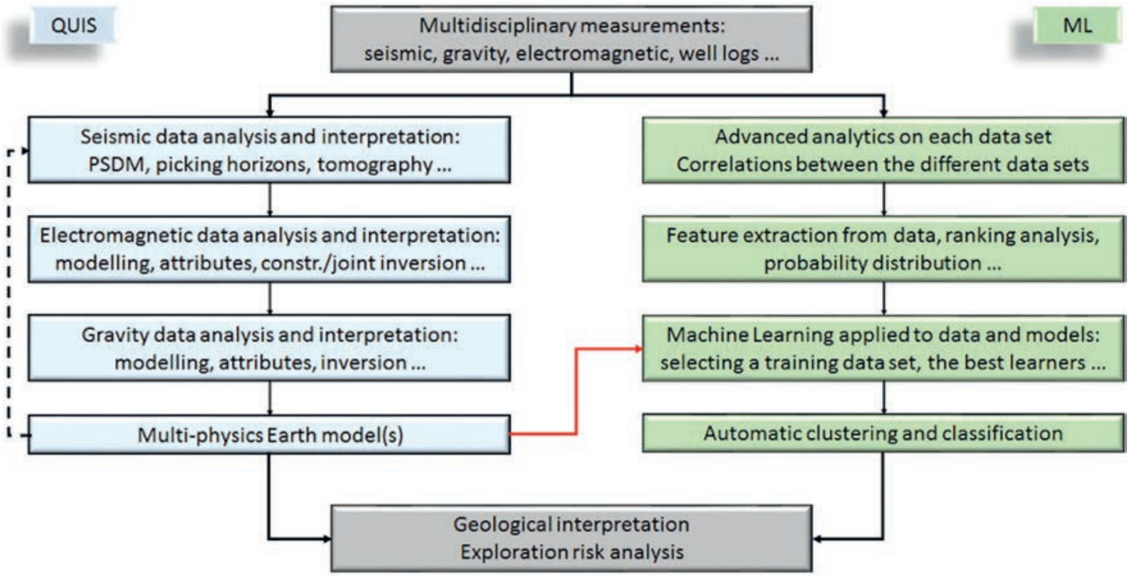

QUIS
Multidisciplinary measurements: seismic, gravity, electromagnetic, well logs ...
ML
Seismic data analysis and interpretation: PSDM, picking horizons, tomography...
Electromagnetic data analysis and interpretation: modelling, attributes, constr./joint inversion ...
Gravity data analysis and interpretation: modelling, attributes, inversion ...
Multi-physics Earth model(s)
Advanced analytics on each data set Correlations between the different data sets
Feature extraction from data, ranking analysis, probability distribution ...
Machine Learning applied to data and models: selecting a training data set, the best learners ...
Automatic clustering and classification
Geological interpretation Exploration risk analysis

Fig. 1 - Workflow combining QUIS and ML.

these matrices allow comparing the reliability of different methods by plotting predicted vs. actual values (referred to the labelled data set) for each selected attribute and for each classification algorithm. After selecting the most effective learner(s), we perform the classification of unlabelled data sets. Finally, the results are mapped in terms of probabilistic spatial distribution of classes.

As schematically shown in Fig. 1, the left and the right paths converge towards the same target: geological interpretation and risk assessment. The statistical analysis of many types of attributes (seismic, EM, gravity), provides probability maps of potential hydrocarbon distributions or of other scenarios of interest. Finally, the results produced through the ML workflow are evaluated and interpreted under the light of the geophysical/geological models obtained through the left branch of the workflow.

3. A case history

In order to explain with additional details the approach introduced in the previous section, we discuss a case history in a difficult exploration area where we applied the entire workflow of Fig. 1. We have already discussed (partially) the same data set in previous works (Dell'Aversana *et al.*, 2012, 2016). In this paper, we use the same data set for illustrating the benefits of ML methods when these are used in cooperation with modelling, inversion and geological interpretation. Thus, this new article represents the right complement to our previous published researches.

3.1. Overview

The area of the test is located in the Barents Sea, offshore the coast of Finnmark. The structure is a large roll-over cut by numerous faults, defining a complex geological setting characterised by separate fault blocks. The field is explored by wells penetrating hydrocarbon-bearing rocks of Upper, Middle, and Lower Triassic ages. Permian carbonates have been drilled in the south fault compartment. Seabed is relatively smooth with a decrease in depth towards the SW. Water depth ranges from 260 to 440 m in the studied area. Several wells have confirmed the presence of hydrocarbons in stacked sandy reservoirs in separate fault blocks on the field. Two independent Marine CSEM surveys have been performed in the area, with the main goal to reduce the uncertainties related to how extensive the accumulations are. These uncertainties depend mainly on the sealing properties of the bounding faults, and the faults inside the roll-over structure. The CSEM data were recorded using a set of acquisition parameters optimised through pre-survey sensitivity analysis. We acquired 11 CSEM profiles for a total of 350 km of towing lines. We used two central frequencies for each towing line: 0.15 and 0.50 Hz. In fact, the pre-survey 2D/3D modelling indicated that a central frequency of 0.50 Hz would be optimal for detecting the shallowest target. Furthermore, CSEM data recorded with that frequency should be affected only marginally by the deep Carbonate platform underlying the stacked reservoir. On the other side, the lowest frequency can be useful for detecting the resistivity trend of the lowest reservoir and, partially, the carbonates below.

The interpretation of the whole data set was based on the systemic approach showed in Fig. 1, aimed at combining CSEM data with pre-existing seismic and gravity data. We used an integrated interpretation platform including algorithms of forward and inverse modelling of multidisciplinary data. In the following paragraphs, we are going to discuss just few examples of the interactive modelling and inversion workflow applied to each CSEM line. Additional details of the whole interpretation workflow can be found in Dell'Aversana *et al*. (2012, 2016).

After showing some illustrative examples of multi-physics modelling/inversion for one selected line, we will focus the discussion on the new aspects of the workflow. We will show how complementary ML algorithms allow extracting probabilities maps of hydrocarbon distribution from the multi-parametric models obtained from seismic and non-seismic data.

3.2. Data

The data set consisted of a seismic 3D volume migrated through 3D Kirchhoff pre-stack depth migration, 2D CSEM data, and regional gravity data.

A total of 172 CSEM receivers were deployed on the sea floor in two different surveys covering the whole area of interest. Two fundamental source frequencies, 0.15 and 0.50 Hz, were used along over 350 km of CSEM towing lines. As mentioned above, this electromagnetic acquisition was aimed at illuminating two different targets (two separate sandy reservoirs) located at different depth. For that reason, we decided to acquire each individual CSEM line twice using two central frequencies (0.15 and 0.50 Hz) and their corresponding harmonics.

Fig. 2 shows, in the left panel, part of the CSEM layout focused on the target area, co-rendered with a depth map of the top of the upper reservoir (in colours) and of the carbonates (contours), as interpreted from seismic data. In the right panel, colours indicate the sum of the normalised magnitude of electric and magnetic fields. It is calculated just by normalising to 1 the sum of the normalised electric and magnetic magnitudes. The normalisation of each measured value of the electric and magnetic fields is done with respect to a conductive CSEM response, simulated in a uniform resistivity half space. The large yellow area about in the middle of the map represents the high EM response caused by the proved presence of the stacked hydrocarbon reservoirs.

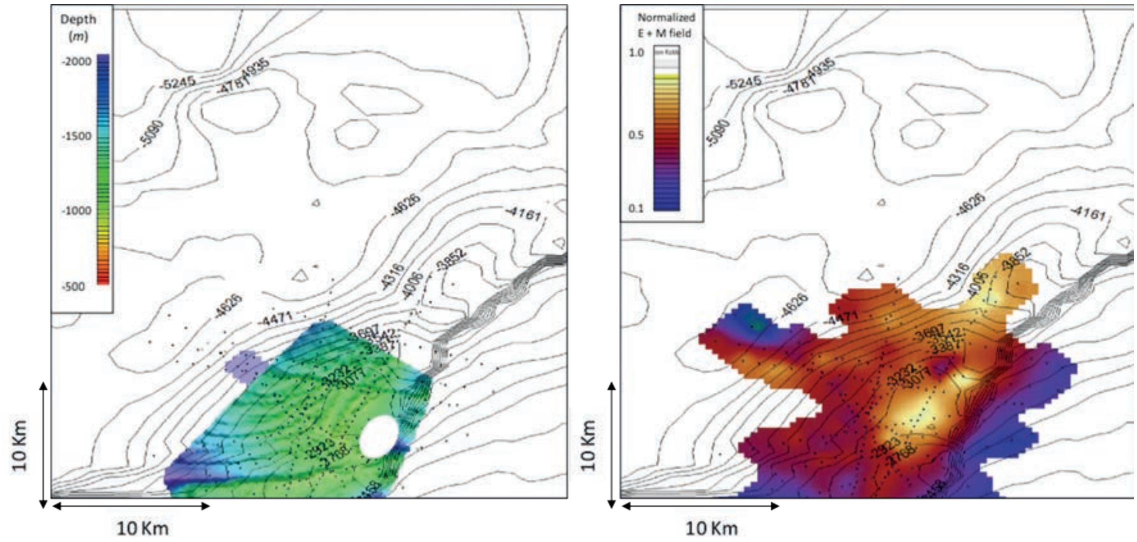


Fig. 2 - Left panel: depth contour lines (in m) of the carbonate platform co-rendered with the map (in colours) of the upper reservoir. The black dots are part of the CSEM layout. Right panel: CSEM normalised magnitude (normalised electric + normalised magnetic magnitude). Note the large anomaly about in the middle of the CSEM layout. Other anomalies appear at north and NW.

Fig. 3 shows a depth map of the Permian carbonate platform (contour lines), co-rendered with the first vertical derivative of the Bouguer anomaly. The area showed in the figure includes the area of the hydrocarbon field. The yellow symbols represent some of the CSEM receivers, indicating the area of the electromagnetic survey. The two lines indicate the position of two of the 11 CSEM towing lines. We can see the general correlation between the carbonates and the gravity response. The line highlighted in yellow (Line 02) corresponds to the seismic section showed in Fig. 4. In that section, we can see the interpreted tops of the two main reservoirs and of the carbonate platform. In this section and in the following, we are going to discuss the key steps of the modelling and inversion work performed on Line 02.

Line 01 indicated in the Fig. 3, is another important reference line, because it crosses the entire reservoir from SW to NE, almost orthogonally to Line 02. We discussed extensively the modelling and inversion work performed along Line 01, showing the details of our integrated approach in a previous paper (Dell'Aversana *et al.*, 2012). We do not repeat that discussion in order to avoid excessive redundancy.

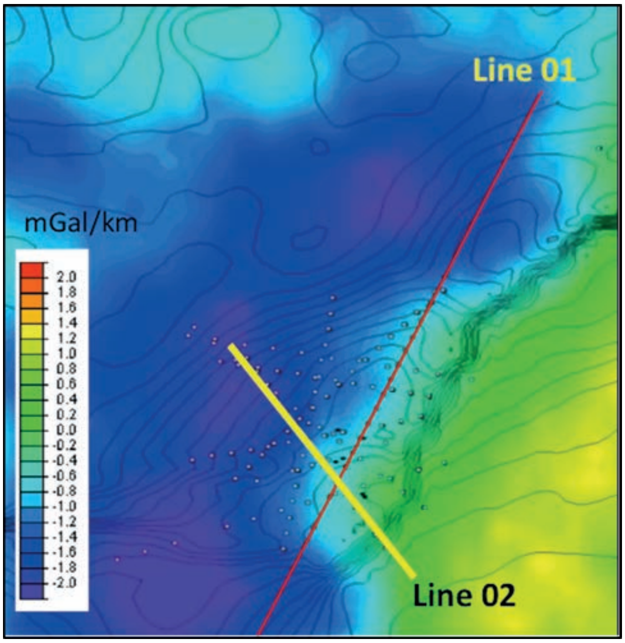


Fig. 3 - Depth map of the Permian carbonate platform (contour lines), co-rendered with the first vertical derivative of the Bouguer anomaly. Two CSEM towing lines are displayed. The yellow line is the one discussed in this section.

3.3. Iterative modelling and inversion

In this paragraph, we summarise the main modelling and inversion steps performed on Line 02, following the scheme of the left branch of Fig. 1. Using the constraints derived from seismic interpretation of the 3D Pre-Stack Depth Migrated (PSDM) volume, we started with 3D CSEM modelling. Fig. 5 is a scheme of the iterative modelling workflow. It shows conceptually how the resistivity model was progressively built through a trial-and-error approach. We started from a uniform conductive half space embedding resistivity anomalies consistently with the main seismic horizons. Then, we progressively updated the resistivity model in order to fit the CSEM data. As a first guess, we filled the layers with resistivity values derived from the well logs and extrapolating

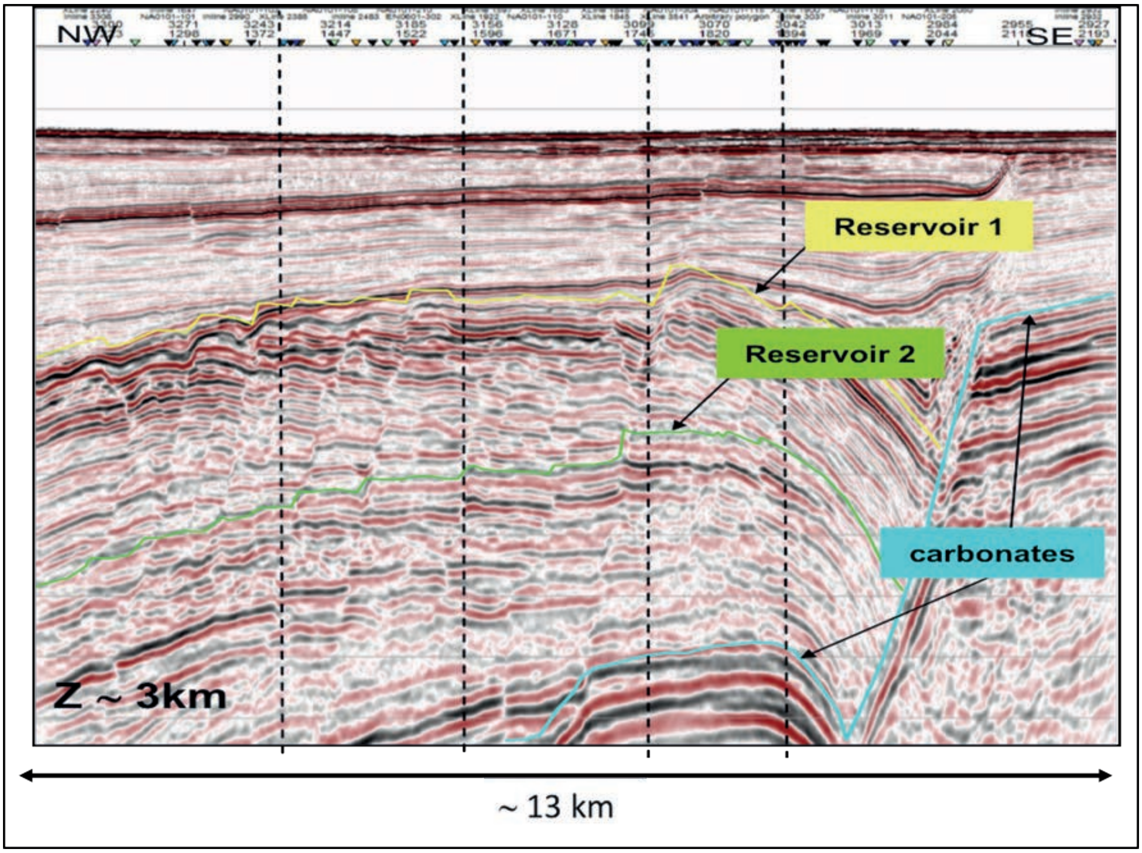
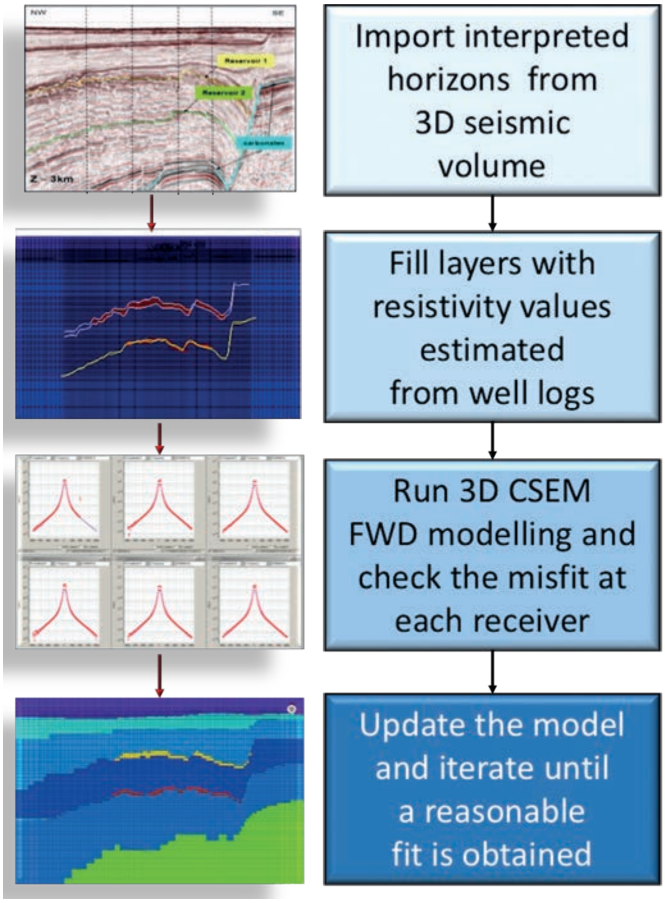


Fig. 4 - Pre-stack depth migrated seismic section extracted from the 3D seismic cube along the direction highlighted by the yellow line in Fig. 3 (Line 02 in Fig. 3).

them laterally. Initially, the fit between modelled and observed electromagnetic responses was relatively poor (probably because of the different physical meaning of the resistivity obtained through induction logs and the CSEM method). In order to reduce the initial misfit, we applied an iterative forward modelling, updating manually the resistivity of each layer. This method is very time consuming. However, it allowed us building a reasonable starting model for the following optimised inversion. At the same time, it allowed us exploring interactively the resistivity model space constrained by the interpreted seismic horizons.

The next step was inversion of CSEM data. We applied a complex workflow consisting of a sequence of different inversion approaches. We used both commercial as well as a proprietary software packages (Chiappa *et al.*, 2017) and, then, we compared the results. We started with a 2.5D Bayesian inversion for each individual line of both electric and magnetic components, constrained by the main seismic horizons (top and bottom of upper reservoir and lower reservoir). This initial inversion step was useful because the output resistivity sections are directly comparable with the seismic sections and with the gravity modelling results. However, we run also 3D anisotropic inversion (both constrained and unconstrained) of the entire data set, including both frequencies (0.15 and 0.50 Hz), plus their first 2 harmonics.



Import interpreted horizons from 3D seismic volume

Fill layers with resistivity values estimated from well logs

Run 3D CSEM FWD modelling and check the misfit at each receiver

Update the model and iterate until a reasonable fit is obtained

Fig. 5 - Conceptual scheme of the iterative/interactive 3D CSEM modelling workflow.

The bottom panel of Fig. 6 shows a 2D resistivity section obtained by constrained inversion of electric and magnetic CSEM data. The upper panel shows again the seismic section, for comparison purposes. The areas marked by the "gaps" indicate possible lateral edges of the reservoirs. The presence of these gaps was suggested by the analysis of the "CSEM attributes of symmetry" (Dell'Aversana and Zanoletti, 2010). These attributes help to identify lateral resistivity discontinuities with high accuracy. They are based on the comparison of the out-towing and in-towing branches of the CSEM response for each receiver.

Fig. 7 shows the results of the iterative gravity modelling. In this case, the modelling is aimed at fitting the first vertical derivative of the Bouguer anomaly. We started with an initial guess model using the main interfaces as interpreted on the seismic section. The initial density values for each layer were derived from the available well logs. Both densities and layer geometry were updated through iterative forward modelling, constrained by the wells, and taking into account the known geology of the area.

The first vertical derivative works as a high pass filters. It allows modelling some important details of the geometry of the main layers, eventually updating the previous interpretation steps. Indeed, the gravity modelling allowed to update, partially, the layers geometries in the resistivity

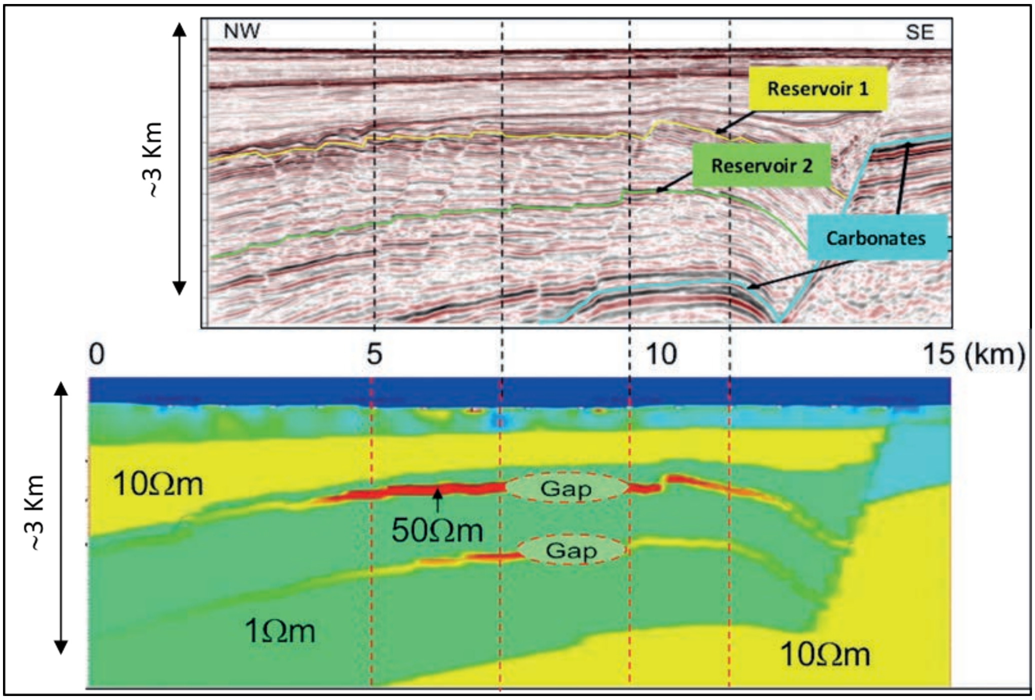


Fig. 6 - Comparison between the seismic section showed in Fig. 4 and the CSEM inverted resistivity model along the same section (Line 02 in Fig. 3).

model; furthermore it supported and improved the interpretation of the seismic section (see Fig. 8).

Fig. 8 shows a synoptic view of the various modelling/inversion results in the three different domains. It shows a comparison between the PSDM seismic section, the CSEM inverted section and the density modelling. Both the gravity and the CSEM sections are consistent with the main features of the seismic section. We notice a general correspondence between the spatial distribution of density and resistivity, even though there are some differences. Indeed, the value added by the gravity modelling was to introduce some "large-scale" (low spatial frequency) geometrical updates to the CSEM inversion results. On the other side, the CSEM symmetry attributes allowed detecting two possible resistive gaps in both reservoirs (as discussed in Dell'Aversana *et al*., 2012). The gravity modelling cannot sense these small-scale gaps. Thus, this example shows how CSEM and gravity data, both supported by seismic, bring complementary information.

4. The Machine Learning approach

We applied the same integration workflow of Line 02 to each one of the remaining ten CSEM towing lines. We skip the detailed discussion for each line, in order to avoid redundant descriptions of the same methodological concepts. Additional details can be found in our previous papers (Dell'Aversana *et al*., 2012, 2016). Instead, in this section, we discuss the part of the workflow based on the ML approach (right branch of Fig. 1).

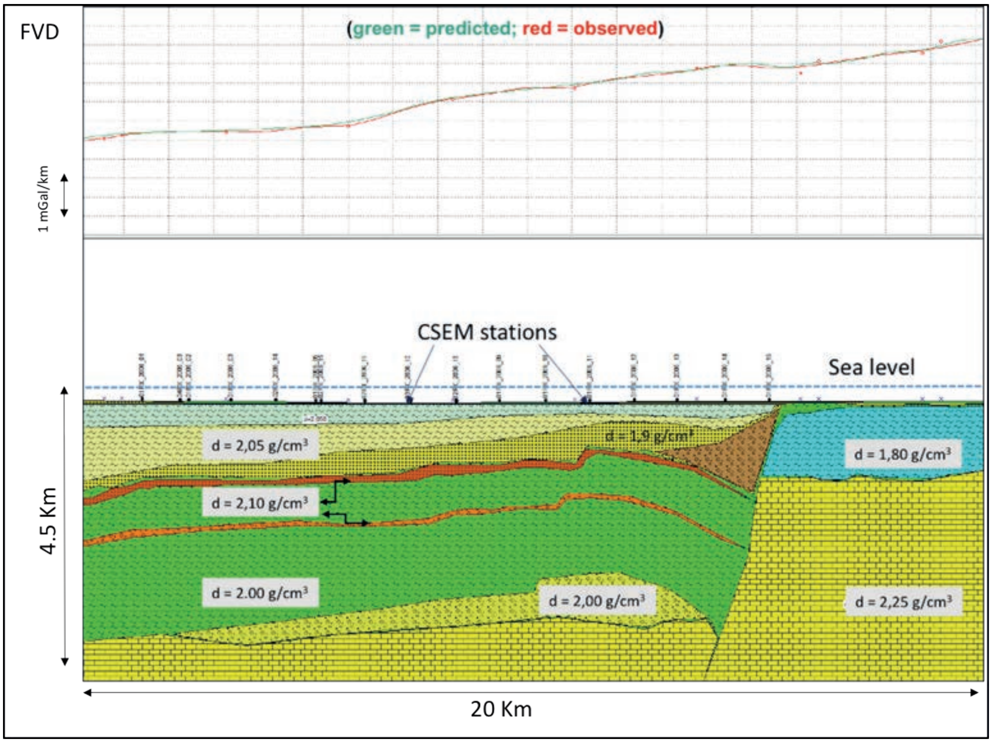


Fig. 7 - Modelling of first vertical derivative of the Bouguer anomaly along the same direction (but laterally extended) of the seismic and CSEM sections discussed earlier (Line 02 in Fig. 3). The green curve is the predicted response; the red curve is the interpolated curve of the observed first vertical derivative. Additional values concur in shaping the interpolated curve. They are the values observed laterally to the selected profile of Line 02, and are not showed in the graph of the interpolated anomaly.

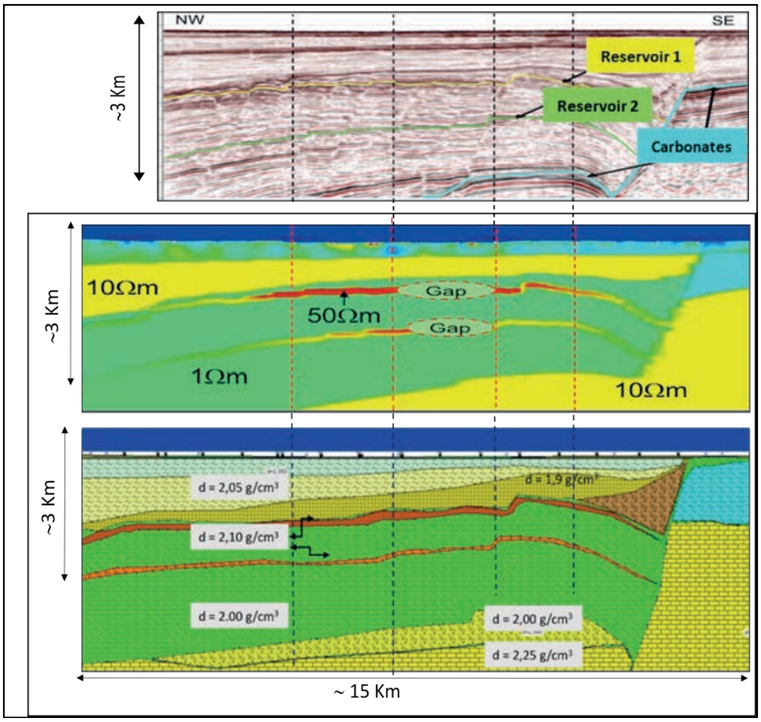


Fig. 8 - Synoptic view of the various migration/inversion/modelling results in the three different domains (seismic, CSEM, and gravity, respectively) along Line 02. The vertical dashed lines are markers helpful for correlating the mean features of the three sections.

The basic idea is to use, simultaneously, the entire available information in both data space and model space, in order to perform a multi-attribute automatic classification. The final purpose is to distinguish areas with variable probability to have oil or brine in the reservoirs, in order to predict the potential lateral extension of the productive layers. A key concept is that information is huge and heterogeneous at the same time in both data and model spaces. The presence of several exploration wells in the area allowed us calibrating locally our attribute maps. It means that the geophysical data measured in the drilled area were used as labelled information for training the learning algorithms on known data, and for classifying the data far away from the wells. For instance, a high CSEM response in correspondence of a proven oil discovery, is labelled as "Oil category" for the corresponding CSEM attribute at that location. Consequently, in this case history, we were able to apply a supervised ML approach. The CSEM information is available for many offsets and for many frequencies, providing a big data set for training the learning algorithms. Indeed, as specified earlier, we acquired CSEM data using two different fundamental frequencies. Furthermore, there are many CSEM data related to the various harmonics. The same abundance of information arises for gravity data, especially if we use various filters and spatial derivatives of the Bouguer anomaly. Of course, seismic information is well distributed in the whole exploration area and can be calibrated near the wells for creating a robust training data set. In summary, every piece of information can bring its own contribution to train the learning algorithms and to classify the data and the models into one scenario or into another.

In this specific case, we are interested mainly in two possible scenarios (binary classification): oil filled or brine filled reservoir. In the following analysis, we will see that no one of the available geophysical data/models can allow, individually, distinguishing between the two classes. Instead, when the entire information is integrated with the help of ML algorithms, a probabilistic result can be obtained. Such result can be finally used for producing hydrocarbon probability maps at reservoir depth.

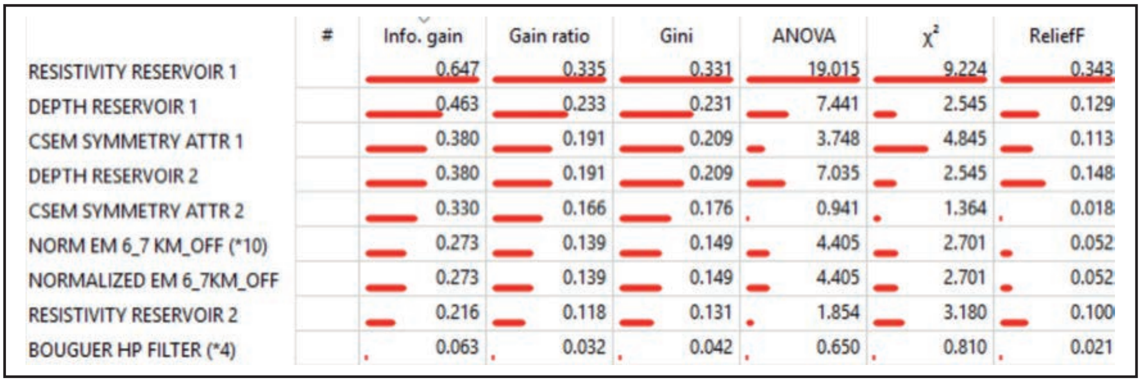
In our approach, we used many different learners mostly based on the paradigm of supervised learning. We remind again that in the case of supervised learning algorithms, ML techniques are used to train on model examples. These examples are built using a set of features, or attributes, diagnostic for the classification purposes. In our approach, we combined the diagnostic power of different categories of features. The expression "diagnostic power", is here intended as the capability to distinguish between oil-bearing sands and brine-sands. A first type of feature includes attributes extracted directly from surface (or sea bottom) measurements (data-space attributes). For instance, the main CSEM features in the data-space include amplitudes and phase of electric and magnetic fields, observed at 10 different offsets, for 10 different frequencies (including the main harmonics), at each receiver position. Furthermore, other CSEM features are the "symmetry attributes" calculated, at each CSEM receiver, for the same range of offset and frequencies. Other non-seismic attributes in the data space are the Bouguer anomaly, its filters, and its spatial derivatives.

Furthermore, we used attributes retrieved at reservoir depth through the modelling and inversion workflow described above (model-space attributes). These include electrical resistivity and density. That category of model-space features includes the picked horizons interpreted from seismic data (in depth, in this case). Where available, we used also other attributes retrieved from seismic amplitude analysis. Finally, we used the well logs for calibration purposes in order to prepare the "labelled data set" necessary for training the classification learners. Our approach was to use the wells drilled in both reservoirs in order to create a “multi-attribute labelled matrix”. This includes all the available seismic, CSEM, and gravity information assigned to a specific known scenario (“oil” or “brine” class), with a certain probability. In such a way, we created a borehole-calibrated data set to be used for training. Line 02 is particularly suited for calibration purposes because it runs along (or close to) several wells. For that reason, in the first part of this paper, we discussed the modelling and inversion work along that line. We are going to show that, after training the learners along Line 02, it is possible to classify all the unlabelled seismic, CSEM and gravity data in terms of oil probability and brine probability in the entire investigated area.

4.1. Feature engineering

The different attributes show variable sensitivity for distinguishing the oil scenario from the brine scenario. In order to understand the diagnostic power of the various features, we first performed a ranking using various types of indexes. Table 1 shows an illustrative example of the indexes that are frequently used and their corresponding values for this test.

Table 1 - Example of feature ranking based on various indexes. The length of the horizontal bar is proportional to the sensitivity of the corresponding feature in separating the two classes (“oil” and “brine”) in the labelled data set.



	#	Info. gain	Gain ratio	Gini	ANOVA	χ^2	ReliefF
RESISTIVITY RESERVOIR 1		0.647	0.335	0.331	19.015	9.224	0.343
DEPTH RESERVOIR 1		0.463	0.233	0.231	7.441	2.545	0.129
CSEM SYMMETRY ATTR 1		0.380	0.191	0.209	3.748	4.845	0.113
DEPTH RESERVOIR 2		0.380	0.191	0.209	7.035	2.545	0.148
CSEM SYMMETRY ATTR 2		0.330	0.166	0.176	0.941	1.364	0.018
NORM EM 6_7 KM_OFF (*10)		0.273	0.139	0.149	4.405	2.701	0.052
NORMALIZED EM 6_7KM_OFF		0.273	0.139	0.149	4.405	2.701	0.052
RESISTIVITY RESERVOIR 2		0.216	0.118	0.131	1.854	3.180	0.100
BOUGUER HP FILTER (*4)		0.063	0.032	0.042	0.650	0.810	0.021

4.1.1. Brief description of the ranking indexes

Before continuing the discussion, we provide a basic description about the indexes included in Table 1. These are exhaustively discussed in dedicated works, like for instance in the book of Raschka and Mirjalili (2017) as well as in the book of Russell and Norvig (2016).

a) Information Gain tells us how important a given attribute of the feature-vectors is. For instance, we can understand the meaning of this index when we use it at the nodes of a Decision Tree. This is a flow chart resembling a tree structure, where an attribute value is tested at each node. Each branch represents an outcome of the test. The tree leaves represent classes or class distributions. The Information Gain is related to the decrease in “entropy” (this will be defined below) after a data set is split on an attribute. In order to estimate the relevance of a certain attribute used in the Decision Tree, we need to define an objective function to be optimised via the tree learning algorithm. That objective function corresponds to the Information Gain: it must be maximised at each split, and is defined as follows:

$$IG\left(D_p, f\right) = I\left(D_p\right) - \sum_{J=1}^{m} \frac{N_j}{N_p} I\left(D_j\right). \tag{1}$$

In Eq. 1, *f* is the feature to perform the split, D_p and D_j are, respectively, the data set of the parent and *j*-th child node, *I* is our impurity measure (defined below), N_p is the total number of samples at the parent node, and N_j is the number of samples in the *j*-th child node. The Information Gain is the difference between the impurity of the parent node and the sum of the child node impurities. The lower the impurity of the child nodes, the larger the information gain. We remind that the impurity of a node *t* is a measure of the homogeneity of the labels at the node. There are various impurity measures for classification, like Gini impurity and entropy (Shannon, 1948), and one impurity measure for regression (variance).

b) Gain Ratio is a ratio of the Information Gain (defined above) and the attribute's *Intrinsic Information*.

The *Intrinsic Information* of a split represents the information generated by splitting the data set *D* into *N* partitions and is defined as below:

$$Intrinsic\ Information \ = \ -\sum_{j=1}^{N} \frac{|D_j|}{|D|} log_2\left(\frac{|D_j|}{|D|}\right). \tag{2}$$

D_j is data set of the *j*-th child node. High *Intrinsic Information* means that partitions have more or less the same size. Intuitively, Gain Ratio corrects the Information Gain by taking the Intrinsic Information of a split into account. In this way, it reduces a bias towards multi-valued attributes by considering the number and size of branches when choosing an attribute (Raschka and Mirjalili, 2017).

c) Gini index can be considered like a criterion to minimise the probability of misclassification. It can be understood better after defining the intuitive concept of entropy, $I_H(t)$, for all non-empty classes ($p(i|t) \neq 0$):

$$I_H(t) = -\sum_{i=1}^{C} p(i|t) log_2\big(p(i|t)\big). \tag{3}$$

In Eq. 3, $p(i|t)$ is the proportion of the samples that belongs to class *C* for a particular node *t*. The entropy is 0 if all samples at a node belong to the same class. Instead the entropy is maximal if we have a uniform class distribution.

The Gini index is defined as follows:

$$I_G(t) = \sum_{i=1}^{C} p(i|t)\big(-p(i|t)\big) = 1 - \sum_{i=1}^{C} p(i|t)^2. \tag{4}$$

In practice, the Gini index and entropy have a similar meaning and generally they produce comparable results.

Additional indexes frequently used in the process of features engineering are ANOVA (the name means "Analysis of Variance"), Chi-Square, and Relief (Zani, 1994; Zaffar *et al.*, 2018).

d) ANOVA index is very simple and intuitive: it is the difference between average values of the feature in different classes in which we want to classify our data set.

e) Chi-Square test is used in statistics to test the independence of two events. Given a data set about two "events", we can compare the observed count *O* and the expected count *E*. Chi-Square measures how much the expected counts *E* and observed Count *O* derivate from each other, using the following formula:

$$\chi_c^2 = \sum_i^N \frac{(O_i - E_i)^2}{E_i} , \tag{5}$$

where c is the degree of freedom, O_i is the i-th observed value, and E_i is the i-th expected value. In feature selection, the two "events" are, respectively, occurrence of the feature and occurrence of the class. We want to test whether the occurrence of a specific feature and the occurrence of a specific class are independent. If the two events are dependent, we can use the occurrence of the feature to predict the occurrence of the class. We aim to select the features, of which the occurrence is highly dependent on the occurrence of the class.

f) Relief is the ability of an attribute to distinguish between classes on similar data instances. A weight is assigned to each feature depending on its ability to distinguish among the class values. The features are ranked by weight. The weights are determined on the basis of how well a certain attribute is able to differentiate the instances of similar data samples (Kononenko, 1994; Sharma *et al*., 2017).

4.1.2. Statistical analysis of multidisciplinary features

After the brief summary about the ranking indexes used in this work, we can come back to the feature engineering process applied to our case history. We plotted and compared the statistical distributions of the various features extracted from our multidisciplinary data set. Figs. 9 to 12 show some examples of the probability density distribution of some key features along Line 02. The two classes (brine = blue; oil = red) have been assigned with the help of the wells. These labelled data are those used as training data set.

We started with the simplest attribute: the depth of the top of the reservoirs, as picked on the seismic data. In Fig. 9, we can see that the depth of the top of both the reservoirs is correlated in some way with the presence of oil or brine in the sandy reservoirs. Indeed, we have seen in the previous part of the paper that both reservoirs are dislocated by many faults, and oil tends to accumulate in the structural highs. These correspond with relatively shallow depths (here depth is

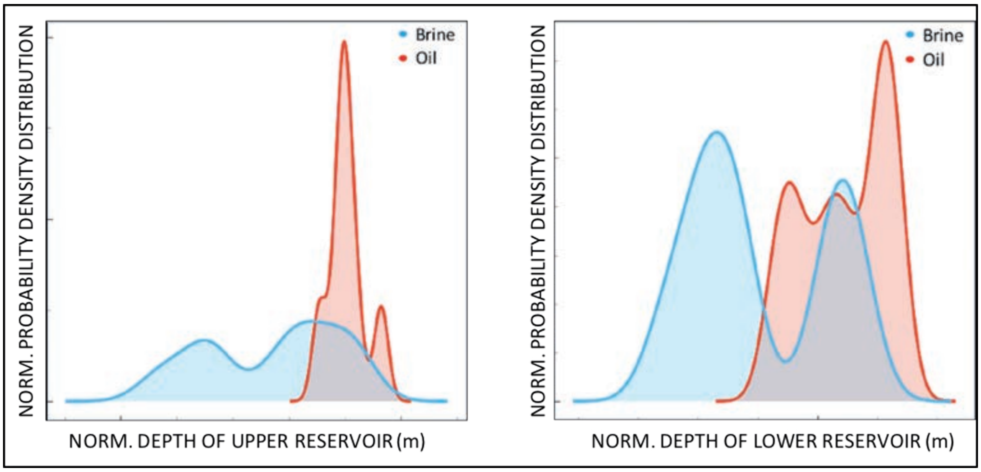


Fig. 9 - Density probability distributions of depth of the top of the upper reservoirs (left) and of the lower reservoir (right).

expressed with negative values). Of course, this is not a rigid rule. In fact, there is a certain degree of overlap between the distribution density areas.

Fig. 10 shows the distributions (along Line 02) of other important attributes: the normalised amplitude of the CSEM data at an offset between 6 and 7 km, at frequency of 0.50 Hz (left panel), and the symmetry CSEM attribute calculated at an offset of 5 km, for a frequency of 0.15 Hz (right panel). The left panel shows that the amplitude of the CSEM data is positively correlated with the presence of oil in the reservoir. Of course, also in this case, this is not any rule valid in every case. In fact, there is a certain overlap between the brine and the oil distribution curves. The same happens for the attribute of CSEM symmetry. We remind that this last attribute is useful for detecting lateral resistivity discontinuities. Used independently from any other information, it cannot be diagnostic about the presence of oil or brine. In a more general sense, the CSEM response does not represent a direct hydrocarbon indicator. Fig. 10 is just a simple confirmation of such well-known fact. CSEM data (and attributes) must be used jointly with other complementary information.

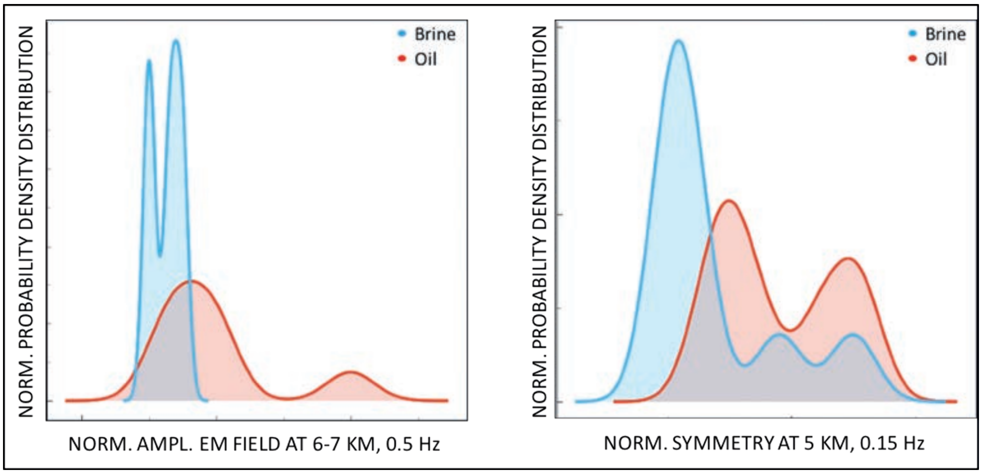


Fig. 10 - Normalised amplitude of the CSEM data at an offset between 6 and 7 km, at frequency of 0.50 Hz (left panel); CSEM symmetry attribute calculated at an offset of 5 km, for a frequency of 0.15 Hz (right panel).

Fig. 11 is a further confirmation of the same concept. It shows the probability density distribution along Line 02 of the resistivity obtained by CSEM constrained inversion, for both reservoirs. Consistently with Table 1, the left panel effectively shows that high resistivity is well correlated with the presence of oil in the upper reservoir. Instead, the right panel shows that high resistivity of reservoir 2 (the lower one) does not correspond necessarily with the presence of oil. Indeed, if we look at Fig. 6 showing the CSEM resistivity model, we can see that comparable values of resistivity can be associated with some parts of the lower reservoir, to carbonates and to another shallow layer. In summary, both Figs. 10 and 11 tell us that CSEM information used without any other independent data can be misleading. That is true both in case we use CSEM amplitudes (observed data at sea floor), and when we use the results of CSEM inversion (inverted parameters at target depth).

Fig. 12 shows the probability distribution of the first vertical derivative of the Bouguer anomaly. As indicated in Table 1, we cannot expect that the gravity method has enough sensitivity

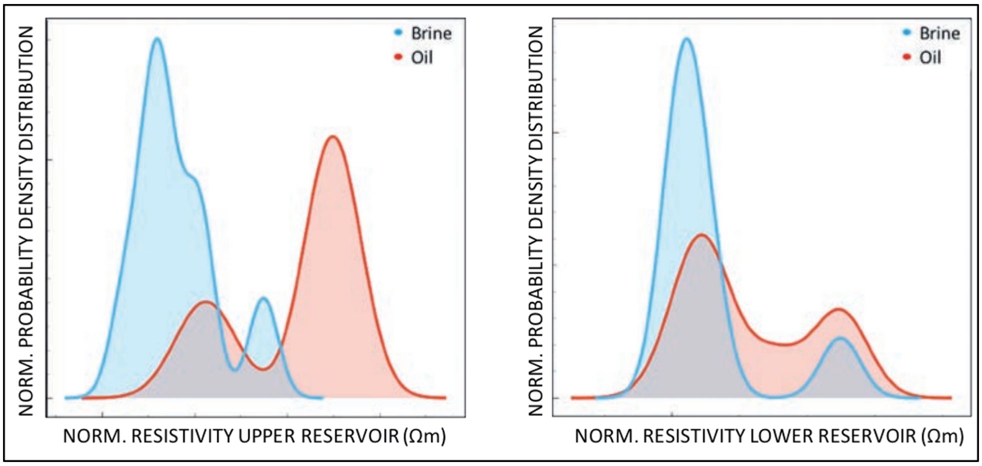


Fig. 11 - Probability density distribution in the upper reservoir (left panel) and in the lower reservoir (right panel).

for distinguishing, from the surface, between oil and brine saturated sands. In fact, the two curves are partially overlapped in the figure. However, as we have already discussed in previous papers (Dell'Aversana *et al*., 2016), the gravity anomalies can be helpful for interpreting CSEM anomalies and for supporting seismic interpretation. In some circumstances, the degree of correlation between the different types of responses (like Bouguer and electromagnetic responses) can help distinguishing resistive CSEM anomalies associated with hydrocarbons from resistive CSEM anomalies associated with "high density/high resistive" geological formations. This is the case of the real example discussed in this paper. In fact, in the studied area, we observe a general consistency at large spatial scale (of the order of 10 km or more) between CSEM, first vertical derivative of Bouguer anomaly and seismic horizons. This correlation indicates, approximately, the regional area where hydrocarbons have been trapped (a geological structural high). However, the situation is complicated by many fault systems in that region. Consequently, the spatial distribution of hydrocarbons is conditioned by the presence of these faults. The effect of such complexity is that, at a smaller spatial scale (2-3 km or less), the CSEM response, the first vertical derivative of Bouguer anomaly and the seismic response are often not correlated

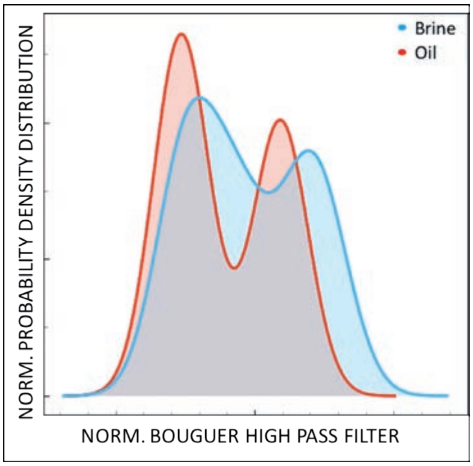


Fig. 12 - Probability distribution of the high pass filter of the Bouguer anomaly.

(Dell'Aversana *et al.*, 2012). A possible interpretation is that when gravity, seismic and CSEM responses are spatially correlated, they contribute to identify the regional structure of hydrocarbon accumulation. Instead, the "local lack of correlation" represents mainly the effect of fluids, that are sensed by the three methods with different sensitivity and variable resolution. Of course, this interpretative approach is not a rigid rule, and can fail if it is used without considering the real geological complexity of the area.

Unfortunately, there is not any deterministic formula that can express, analytically, the relationship between all three geophysical methods (seismic, CSEM, and gravity) and the presence of oil or brine. However, we can try to infer some type of "probabilistic rule" from the data using various ML algorithms. These are more efficient if they are trained on large and heterogeneous data sets. This is the reason why it is important to combine multidisciplinary geophysical data (and models) for assessing the exploration risk. Of course, many other types of information could be included in this integrated approach, like for instance magnetotelluric measurements, many types of seismic attributes, independent geological information, and so forth. In this paper, we limited the discussion to a restricted number of multidisciplinary data sets, because our scope is to show the key methodological aspects of our integrated methodology. However, the same approach can be expanded to every type of information, including other geophysical, geological, structural, and production data too. Finally, synthetic data can be helpful for improving the training process of the learning algorithms (Colombo *et al.*, 1997, 2020).

4.2. Performance of the different learning algorithms

4.2.1. Machine Learning methods

In the practice of ML, we can use many different algorithms (learners). We applied several algorithms addressed to supervised classification, including CN2 Rule Induction, Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, and Adaptive Boosting. We implemented a suite of open source Python libraries running in the same software platform. We linked this ML framework with the platform of integrated geophysical modelling and inversion (QUIS). In such a way, we created a compact and user friendly environment for Data Science and Interpretation of multidisciplinary geophysical/geological data. The following is just a brief and qualitative description of the algorithms that we used. For a full discussion about all these algorithms and their implementation into Python codes, I recommend the book of Raschka and Mirjalili (2017).

The CN2 Rule Induction consists of an algorithm designed for the efficient induction of simple rules of form "if condition, then predict class". Its main advantages are that it works properly even in presence of significant noise, and the classification rules can be easily understood.

The Naïve Bayes classifier works using a Bayesian approach. A probabilistic classifier estimates conditional probabilities of the dependent variable from training data. Then it applies the posterior probabilities for classification of new data instances. A key advantage offered by this approach is that it is fast for discrete features; instead, it is less efficient for continuous features.

Support vector machine (SVM) works on a different principle with respect to the previous algorithms. In fact, it splits the attribute space with a hyper-plane, and try to maximise the margin between the instances of different classes or class values.

The Decision Tree algorithm is a technique that works by separating the data into two or more homogeneous sets (or sub-populations). The separation criteria are based on the most significant features in input variables. It is a precursor to Random Forest.

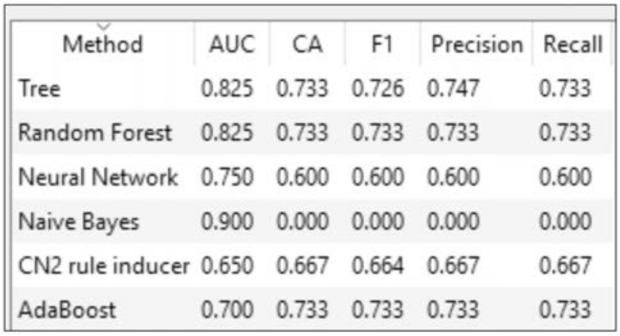
Random Forest is an "ensemble learning" method that uses a set of Decision Trees. Each Tree is developed from a sample extracted from the training data. When developing individual Trees, an arbitrary subset of attributes is drawn (hence the term "Random"). The best attribute for the split is selected from that arbitrary subset. The final model is based on the "majority vote" from individually developed Trees in the Forest.

Similar to Random Forest, Adaptive Boosting consists of multiple classifiers: the final output is the combination of the outputs of those algorithms. The final goal is to create a strong classifier as linear combination of "weak" classifiers.

4.2.2. Cross-validation and performance analysis

The different learning algorithms work more or less effectively depending on many variables, such as the type and the quality of the data, the type of classification problems and so forth. A good approach for selecting the learning algorithm is to test the generalisation power of different methods and, finally, to select the ones showing the best performance. One criterion for selecting the learner is going through "Cross-validation tests". This approach requires partitioning the labelled data (the training data set) into complementary subsets. First, we perform the analysis on one subset (called the training subset), and, then, we validate the analysis on the other subset (called the validation subset or testing subset). Using various performance indexes, we can quantify the performance of each classification algorithm in the cross-validation test. Table 2 shows an example of these indexes calculated for the different classifiers (in one among the many cross-validation tests that we performed). In the table, the index "Area Under the Curve" (AUC) represents the degree or the measure of "separability". It tells how much a certain model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting classes. For instance, in medical applications, the higher the AUC, the better the model distinguishes between patients with disease and no disease. "Classification Accuracy" (CA) is another important index representing the proportion of correctly classified examples. The index "F1" is a weighted harmonic mean of "Precision" and "Recall". "Precision" is the proportion of true positives among instances classified as positive. "Recall" is the proportion of true positives among all positive instances in the data.

Table 2 - Examples of performance indexes.



Method	AUC	CA	F1	Precision	Recall
Tree	0.825	0.733	0.726	0.747	0.733
Random Forest	0.825	0.733	0.733	0.733	0.733
Neural Network	0.750	0.600	0.600	0.600	0.600
Naive Bayes	0.900	0.000	0.000	0.000	0.000
CN2 rule inducer	0.650	0.667	0.664	0.667	0.667
AdaBoost	0.700	0.733	0.733	0.733	0.733

4.3. Classification

First, we applied all the learners to the data set of Line 02, selecting the most sensitive seismic and non-seismic features extracted from our multi-disciplinary data set. In particular, we used the following information:

Features in the data space:

- picked seismic horizons of the top of both reservoirs;
- CSEM electric and magnetic fields observed in a range of offset between 4 and 8 km, and in a range of frequency between 0.15 and 0.50 Hz;
- CSEM attributes of symmetry in the same ranges of offsets and of frequency mentioned above;
- Bouguer anomaly;
- first vertical derivative of the Bouguer anomaly.

Features in the model space:

- CSEM: resistivity models (at target depth);
- gravity: density models (at target depth).

We obtained consistent classification maps of oil probability distribution using each one of the above-mentioned methods. The effectiveness of the approach here described was verified comparing the results of our probabilistic prediction at two wells not included in the calibration phase. In both cases the oil-or-brine prediction was consistent with the drilling results. Probably one of the best map is the one obtained with AdaBoost applied to the upper reservoir. "AdaBoost", short for "Adaptive Boosting", is an ensemble ML meta-algorithm that combines multiple learners. This technique allows you combining multiple "weak classifiers" into a single "strong classifier". A weak classifier is simply a classifier that performs poorly, but performs better than random guessing.

Fig. 13 shows, in colours, the probability to have oil at target depth (upper reservoir, in this case), limited to Line 02. It is obtained by interpolating all the probability values estimated by the "AdaBoost" learner at each CSEM receiver location along this line.

The oil distribution map is co-rendered with the depth map of the reservoir. The squared symbols in the maps indicate CSEM receivers.

Finally, we applied all the learners to the entire data set in our exploration area (including the remaining 10 CSEM lines). Fig. 14 shows in colours, the probability to have oil at target depth (upper reservoir, in this case), for the entire exploration area.

5. Conclusions

ML can support the integration workflow of heterogeneous geophysical data sets in the process of exploration risk evaluation. Our philosophy is to use ML in strict cooperation with advanced geophysical modelling and inversion. In this paper, we discussed how we combined statistical and automatic classification approaches with seismic pre-stack depth migration, with iterative modelling of CSEM and gravity data, and with optimised constrained inversion of CSEM data. This "hybrid approach" allows taking the benefits of automatic statistical and classification tools and, at the same time, to preserve all the advantages of the interactive geophysical data interpretation. We tested our methodology with complex multidisciplinary geophysical data sets recorded in a complicate

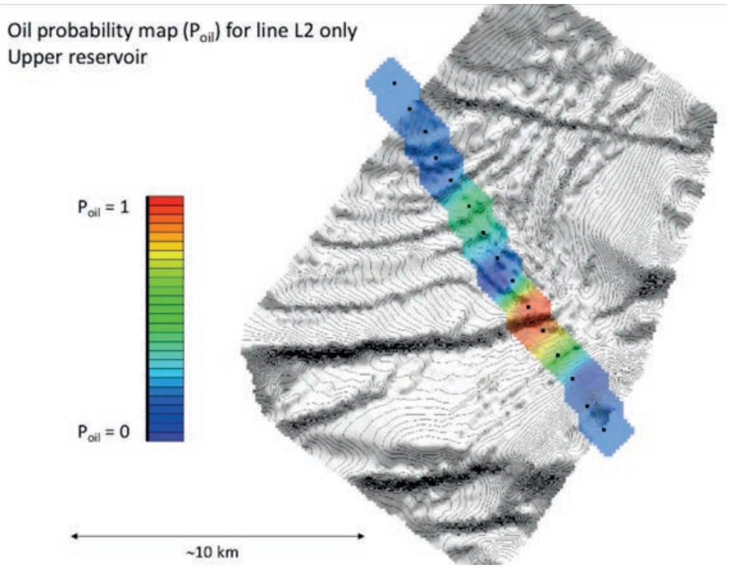


Fig. 13 - Probability map of oil distribution (P_{oil}) in the upper reservoir, only for Line 02. Legend: red: $P_{oil} = 1$; bleu: $P_{oil} = 0$.

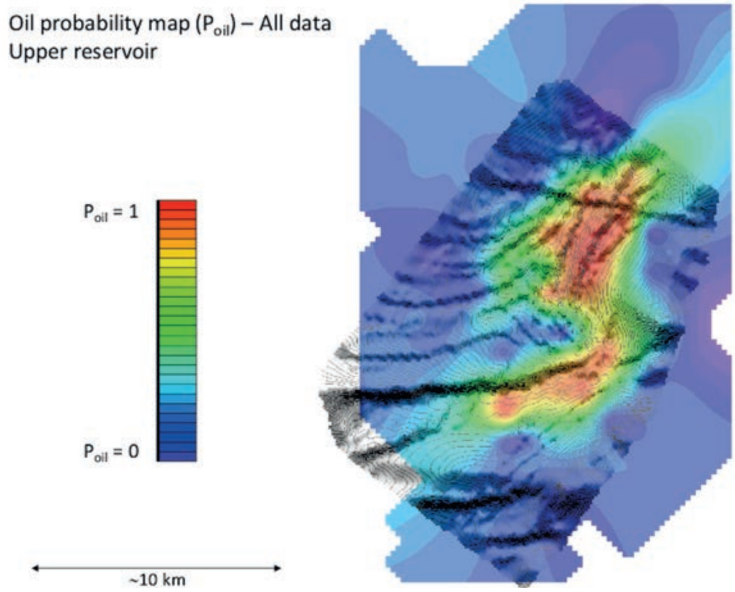


Fig. 14 - Probability map of oil distribution (P_{oil}) in the upper reservoir. Legend: red: $P_{oil} = 1$; bleu: $P_{oil} = 0$.

geological setting, obtaining encouraging results. Each one of the various geophysical data sets (seismic, gravity, and electromagnetic) gave its own contribution to the evaluation/mitigation of the exploration risk, allowing to produce reliable maps of hydrocarbon probability at target depth.

REFERENCES

Aminzadeh F. and de Groot P.; 2006: *Neural networks and other soft computing techniques with applications in the oil industry*. EAGE Publications, 161 pp.

Bishop C.; 2006: *Pattern recognition and Machine Learning*. Springer-Verlag, New York, NY, USA, 758 pp.

Cagniard L.; 1953: *Basic theory of the magneto-telluric method of geophysical prospecting*. Geophys., **18**, 605-635, doi:10.1190/1.1437915.

Castagna J.P. and Backus M.M. (eds); 1993: *Offset dependent reflectivity - Theory and practice of AVO analysis*. Society of Exploration Geophysicists, Tulsa, OK, USA, 357 pp., doi: 10.1190/1.9781560802624.

Chiappa F., Vandone V., Dell'Aversana P. and Bernasconi G.; 2017: *Sharp CSEM inversion by means of geological structural constraints*. In: Expanded Abstract, 79th EAGE Conference and Exhibition, vol. 2017, pp. 1-5, doi: 10.3997/2214-4609.201701359.

Colombo D., Gitis V. and De Franco R.; 1997: *Application of pattern recognition techniques to long-term earthquake prediction in central Costa Rica*. Eng. Geol., **48**, 7-18.

Colombo D., Li W., Sandoval-Curiel E. and McNeice G.W.; 2020: *Deep-learning EM monitoring coupled to fluid flow simulators*. Geophys., **85**, WA1-WA12, doi: 10.1190/geo2019-0428.1.

Condi F. and Talwani M.; 1999: *Resolution and efficient inversion of gravity gradiometry*. SEG Technical Program Expanded Abstracts, pp. 358-361, doi: 10.1190/1.1821022.

Dell'Aversana P.; 2014: *Integrated geophysical models: combining rock physics with seismic, electromagnetic and gravity data*. EAGE Publications, 244 pp.

Dell'Aversana P. and Zanoletti F.; 2010: *Spectral analysis of marine CSEM data symmetry*. First Break, **28**, 44-51.

Dell'Aversana P., Bernasconi G., Miotti F. and Rovetta D.; 2011: *Joint inversion of rock properties from sonic, resistivity and density well-log measurements*. Geophys. Prospect., **59**, 1144-1154.

Dell'Aversana P., Colombo S., Ciurlo B., Leutscher J. and Seldal J.; 2012: *CSEM data interpretation constrained by seismic and gravity data. An application in a complex geological setting*. First Break, **30**, 35-44.

Dell'Aversana P., Bernasconi G. and Chiappa F.; 2016: *A global integration platform for optimizing cooperative modeling and simultaneous joint inversion of multi-domain geophysical data*. AIMS Geosci., **2**, 1-31, doi: 10.3934/geosciences.2016.1.1.

Fairhead J.D.; 2015: *Advances in gravity and magnetic processing and interpretation*. EAGE Publications, 352 pp.

Kononenko I.; 1994: *Estimating attributes: analysis and extensions of RELIEF*. In: Bergadano F. and De Raedt L. (eds), Machine Learning: ECML-94, Lecture Notes in Computer Science, Springer, Berlin-Heidelberg, Germany, vol. 784, pp. 171-182, doi: 10.1007/3-540-57868-4_57.

Raschka S. and Mirjalili V.; 2017: *Python Machine Learning: Machine Learning and deep learning with python, scikit-learn, and tensorflow, 2nd ed*. PACKT Publishing, Birmingham, UK, 622 pp.

Russell S.J. and Norvig P. (eds); 2016: *Artificial intelligence: a modern approach, 3rd ed*. Prentice Hall Series in Artificial Intelligence, Upper Saddle River, NJ, USA, 1144 pp.

Samuel A.L.; 1959: *Some studies in Machine Learning using the game of checkers*. IBM J. Res. Dev., **3**, 535-554, doi: 10.1147/rd.33.0210.

Shannon C.E.; 1948: *A mathematical theory of communication*. Bell Syst. Tech. J., **27**, 379-423, doi: 10.1002/j.1538-7305.1948.tb01338.x.

Sharma R., Mantri A. and Dua S. (eds); 2017: *Computing, analytics and networks*. Revised Selected Papers, First International Conference, ICAN 2017, Chandigarh, India, Springer, Singapore, vol. 805, 227 pp., doi: 10.1007/978-981-13-0755-3.

Tarantola A.; 2005: *Inverse problem theory and methods for model parameter estimation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 353 pp.

Vozoff K.; 1990: *Magnetotellurics: principles and practice*. Proc. Indian Acad. Sci. (Earth Planet. Sci.), **99,** 441-471, doi: 10.1007/BF02840313.

Zaffar M., Hashmani M.A., Savita K.S. and Rizvi S.S.H.; 2018: *A study of feature selection algorithms for predicting students academic performance*. Int. J. Adv. Comput. Sci. Appl., **9**, 541-549, doi: 10.14569/IJACSA.2018.090569.

Zani S.; 1994: *Analisi dei dati statistici, vol. I: osservazioni in una e due dimensioni, 1st ed*. Giuffrè editore, Milano, 366 pp.

Corresponding author: Paolo Dell'Aversana
Eni S.p.A., Upstream and Technical services
San Donato Milanese (Milano), Italy
Phone: +39 02 52063217; fax: +39 02 520 63897; e-mail: paolo.dell'aversana@eni.com