Data-based support for petroleum prospect evaluation

Summaya Mumtaz¹ · Irina Pene¹ · Adnan Latif¹ · Martin Giese¹

Received: 23 March 2020 / Accepted: 10 August 2020 / Published online: 5 September 2020
© The Author(s) 2020, corrected publication 2021

Abstract

We consider the challenging task of evaluating the commercial viability of hydrocarbon prospects based on limited information, and in limited time. We investigate purely data-driven approaches to predicting key reservoir parameters and obtain a negative result: the information that is typically available for prospect evaluation *and* is suitable for data-based methods, cannot be used for the required predictions. We can show however that the same information is sufficient to produce a limited list of potentially similar well-explored reservoirs (known as *analogues*) that can support the prospect evaluation work of human geoscientists. We base the proposal of analogues on similarity measures on the data available about prospects. Technically, the challenge is to define suitable similarity measures on categorical data like depositional environment or rock types. Existing data-based similarity measures for categorical data do not perform well, since they do not take geological domain knowledge into account. We propose two novel similarity measures that use domain knowledge in the form of hierarchies on categorical values. Comparative evaluation shows that the semantic-based similarity measures outperform the existing data-driven approaches and are effective in comparison to the human analogue selection.

Keywords Analogues · Prediction · Similarity · Machine learning

Introduction

One of the challenging tasks of commercial hydrocarbon exploration is that of *prospect evaluation*: a limited amount of information is available for a given location (the *prospect*). Based on substantial knowledge of previously explored petroleum reservoirs, one has to determine whether it will be worth investing in the further development of the prospect.

The usual method employed by geoscientists for prospect evaluation is based on *analogues*: a number of well-explored reservoirs are considered that share as many of

the given properties of the prospect as possible. These analogues are used (together with the available information about the prospect) to give estimates and uncertainty bounds for key parameters such as reservoir porosity, permeability, and size. The viability of the prospect is then determined from these parameters.

Prospect evaluation using analogues is challenging for a number of reasons:

1. the information about the prospect is often rather limited, which means that it is hard to say how good analogues really are, and therefore predictions will be very uncertain
2. the time available for prospect evaluation is also limited, so that it is not possible to reliably find and take into account *all* relevant information of *all* relevant analogues

The situation is particularly difficult for *early* prospect evaluation, i.e. supporting the initial decisions before the first wells are drilled. In practice, estimates are often based on as few as two or three analogues.

This article reports on work in the context of the SIRIUS centre for Scalable Data Access in the Oil & Gas Industry.¹ Equinor, a Norwegian petroleum company, tasked us to

Communicated by: H. Babaie

✉ Summaya Mumtaz
summayam@ifi.uio.no

Irina Pene
pene@ifi.uio.no

Adnan Latif
adnanl@ifi.uio.no

Martin Giese
martingi@ifi.uio.no

¹ SIRIUS Center for Scalable Data Access, University of Oslo, Gaustadalléen 23B, 0373 Oslo, Norway

¹<http://sirius-labs.no/>

find data-based techniques that could help explorationists in early prospect evaluation. Specifically, we tried to use information in data sets such as that curated by IHS² in combination with available parameters of the prospect to help in the evaluation.

Computer support for prospect evaluation presents its own set of challenges:

3. A considerable part of the knowledge about both prospects and well-explored reservoirs is in the form of seismic data, unstructured knowledge about geological processes, etc. None of these are readily digitized in a way that lends itself to common statistical methods. If the information is reduced to usable numerical and categorical parameters, we are typically restricted to fewer than ten parameters to describe reservoirs.
4. The information about well-explored reservoirs is rather small for the number of dimensions under consideration. The IHS dataset has 43000 reservoirs, but the information is often incomplete and tagging is not always consistent between reservoirs. Moreover, many of the available parameters of reservoirs are severely imbalanced: their distribution is such that the values are the same for very many of the reservoirs.

Despite these challenges, we set out to assist geoscientists in the task of prospect evaluation. At first, we tried to use interpolation/extrapolation methods to predict parameters like permeability and porosity directly from the given *numerical and categorical* parameters of the prospect.

Contribution 1: we show that this (numerical and categorical) information, together with the corresponding information about well-explored reservoirs can *not* be used to predict porosity or permeability.

This negative result is not a consequence of the used statistical methods, but of the available data; mostly the fact that much of the data is very similar on the given input parameters, while still varying widely on the output parameters of interest.

Since direct prediction of the key parameters needed for prospect evaluation is not possible, we concentrated on supporting geoscientists' search for analogues.

Contribution 2: a method to compute a ranked list of potential analogues based on the available numerical and categorical parameters.

A comparison to analogues selected by human geoscientists shows that many of the candidate analogues suggested by the algorithm are of interest to the human experts. Our method can therefore be combined with a further evaluation that includes seismic data, unstructured information, etc., to

produce a final selection of analogues in a shorter time than would usually be required to find suitable analogues based on the complete list of known reservoirs.

“[Related work](#)” discusses related work. In “[Relevant information for reservoir evaluation](#)”, we explain the different kinds of structured and unstructured information that are used in reservoir evaluation, not all of which are suitable for machine learning. “[Available data](#)” explains in detail the data we used in our experiments. “[Parameter prediction](#)” discusses the different statistical approaches we used on the data and concludes with the negative result of Contribution 1 above. “[An alternative approach: analogue recommendation](#)” then explains the alternative approach of analogue recommendation, and contains Contribution 2. “[Conclusion](#)” concludes the article.

Related work

Porosity and permeability prediction are important tasks for planning, drilling and production forecasting. The conventional method to determine permeability is to find a relationship between porosity and permeability based on core analysis. Many studies show the complexity of predicting porosity and permeability due to their non-linear relationship and wide variety in geometry and pore sizes (Lacentre and Carrica 2003; Jennings and Lucia 2003; Brigaud et al. 2014).

Several researchers have proposed machine learning models for porosity and permeability prediction based on petrophysical well logs and core data. The most common well logs used in the analysis are gamma ray, neutron porosity, total porosity, bulk density, water saturation, sonic time and micro-spherically focused log (Olatunji et al. 2011). (Zehui et al. 1996) applied artificial neural networks (ANN) on offshore Canada wireline data for porosity and permeability prediction. (Helle et al. 2001) followed a similar approach and developed two separate back-propagation artificial neural networks to predict porosity and permeability from well logs in the North Sea. (Abdideh 2012) found that ANN produce correlation coefficients comparable to regression analysis for predicting permeability from well logs. (Olatunji et al. 2011) used type-2 fuzzy logic to predict permeability and compared the results with type-1 fuzzy logic, support-vector machines (SVM) and ANN. The model yielded 11.7% improvement in the correlation coefficient.

It is evident from the literature that machine learning techniques can be used to estimate permeability and porosity. However, all of the successful work is based on using log data and the core samples as input. Such data is only available after drilling the first well. For early prospect evaluation, this information is not available. Our

²<https://ihsmarkit.com/products/oil-gas-field-reservoir-discoveries.html>

experiments are based only on information available during this early phase before wells are drilled.

(Perez-Valiente et al. 2014) attempted the prediction of reservoir parameters based on a similar set of input features as ours. They used various machine learning models, but with little success. Their mean absolute percentage error remained high for both average matrix porosity and average air permeability prediction. The only two parameters retrieved with high accuracy in their work are original pressure and pressure at top reservoir depth. The work uses these two parameters further to define a similarity function but this only retrieves 30% of analogous reservoirs.

Analogue recommendation in (Perez-Valiente et al. 2014) is based only on ML models and doesn't take into account domain knowledge. Most of the baseline methods we use for comparison in our work (see “[Experimental results](#)”) are much simpler (not ML based) and achieve comparable accuracy, although admittedly, the results are not directly comparable as they are based on different test cases. The main focus of our proposed methodology, in particular in “[An alternative approach: analogue recommendation](#)”, is to incorporate domain knowledge into early phase prospect evaluation.

Relevant information for reservoir evaluation

In the process of evaluating hydrocarbon prospects, making development plans or undertaking economic evaluations, geoscientists use reservoir analogues, together with their own knowledge and experience for better decision-making. This method of analogy has played a significant role in the assessment of resource potential and reserves estimation. In this context, analogue reservoirs have similar rock and fluid properties but are at a more advanced stage of development than the reservoir of interest and thus may provide concepts to assist in the interpretation of more limited data. Analogue reservoirs are formed by the same or very similar processes as regards sedimentation, diagenesis, pressure, temperature, chemical and mechanical history and structural deformation. Selecting an appropriate analogue requires matching the target reservoir with its potential analogue with respect to a number of parameters.

The relevant parameters for our work are the geological parameters, such as geological age, lithology, depositional environment, diagenetic and structural history of the reservoir, reservoir continuity and heterogeneity, reservoir size and column height etc. The data associated with these parameters is of two types. The first type is the data available only in human-readable format, thus requiring a high level of domain knowledge for understanding

and interpreting it (e.g. diagenetic and structural history). The second type is the data that can be structured, by dividing it into groups and sub-groups (e.g. geological age), thus making it usable in a machine learning model. Table 1 summarizes all the available information and shows its division into viable and non-viable, per-reservoir and generic background knowledge.

Informal context/background knowledge

Capturing knowledge about the target reservoir requires a vast amount of “non-viable” data (uncodified knowledge) for the area and the reservoir, e.g. geological and tectonic history of the area, basin evolution, play concepts, geochemical processes, and lessons learned by experience. This knowledge is subjected to the experience level of the geoscientists and has associated uncertainty as the conceptual models might vary between different schools of thought. It is very difficult to transform this knowledge into a form that can be further used by the machine learning algorithm for finding a good reservoir analogue. An example of such a scenario is using the influence of diageneses on the reservoir as one of the input parameters.

For instance, it has been shown (Lin et al. 2017) that a reservoir's history of diagenetic processes has an important influence on reservoir quality (porosity and permeability). However, it is extremely difficult to codify this history and its influence on the parameters we would like to predict, in such a way that it can be used by computer programs.

Table 1 Types of available information

	Viable	Non-viable
per reservoir	<ul style="list-style-type: none"> – geological age – lithology – depositional – environment – diagenetic history – structural history – reservoir continuity – reservoir heterogeneity – ... 	<ul style="list-style-type: none"> reports publications paleographic maps seismic – ...
generic background knowledge	<ul style="list-style-type: none"> Hierarchies of – chrono-stratigraphy – lithology – ... 	Knowledge of Geology

Formal knowledge/hierarchies

Even though much of the background knowledge required for prospect evaluation is non-viable, as discussed in the previous section, there is some generic domain knowledge that can be formalised in such a way that it can easily be used by a computer program. We call such information “viable.”

Several of the features that have the highest impact on reservoir properties such as porosity and permeability are commonly systematised into a hierarchy of groups and sub-groups:

1. Geological age of the rock units,
2. Rocks and their lithology, representing the reservoir, source and seal rock units,
3. Depositional environment in which these rock units have been deposited.

Geological age, for example, has been divided into a hierarchical set of subdivisions, corresponding to similar geological time periods, describing the Earth’s history. The generally accepted subdivisions are eon, era, period, epoch and age. Eons and eras are larger subdivisions than periods while periods themselves may be divided into epochs and ages.

An example of viable background knowledge about geological ages, based on this hierarchy, would be “The Lower Jurassic epoch and the Middle Jurassic epoch are both parts of the Jurassic period. They are therefore more closely related to each other than to epochs of the Cretaceous period.”

To define and classify different types of rocks and their lithology, the following hierarchy concept has been developed:

- Rocks are separated into three main groups of entities, based on their mode of formation: igneous, metamorphic and sedimentary rocks. Each of these rock groups contains many different types of rock, and each can be identified from its physical features.
- The above mentioned groups are further divided into sub-groups (e.g. sedimentary rocks are divided into three sub-groups: clastic sedimentary rocks, biologic sedimentary rocks and chemical sedimentary rocks), representing the second level of hierarchy.
- This sub-division was continued until all the values of the lithology feature in our data sets were organized inside the hierarchy.

This lithological hierarchy provides viable background knowledge like for instance “Carbonate is a kind of Sedimentary Rock.”

In “[Knowledge-based feature augmentation](#)” and in particular “[Semantic similarity](#)”, we discuss how to use these hierarchies. “[Available data](#)” provides a detailed

Table 2 Different types of features

	Numeric	Categorical
Input Features		Age
		Lithology
	Top Depth	Depositional Environment
	Net Thickness	Trap Type
Output Features		Basin Area
	Porosity	
	Permeability	

description of all the available features that have been used for reservoir parameters prediction and analogues selection.

Available data

In order to conduct statistical experiments, a quantitative description of key parameters is required. This quantitative description is often based on the interpretation of the depositional environment, the geological history of the basin, processes that formed the rock, experimental simulations, and well logs, etc. Due to the complexity of interpretation and degree of freedom, only a limited number of datasets are available in this domain.

For the main experiments presented in this paper, we have used a dataset licensed by IHS Markit.³

EDIN-IHS

The EDIN-IHS dataset consists of international E&P data (seismic, drilling, development & production) and Basins (Basin Outlines, Geological & Exploration History & scanned images), covering oil fields in more than 200 countries and a total of 43 000 reservoirs. Based on input from Equinor, the important key features suitable for input and output of prediction and recommendation tasks are listed in Table 2.

The key features are extracted and stored in a dataset, where each row corresponds to a case (a reservoir in this scenario) and the column represents features related to each reservoir.

Data definition & mapping to ontologies

Lithology: A lithology gives a detailed description of the composition or type of rock and its gross physical characteristics such as the macroscopic nature of the mineral

³<https://cdn.ihs.com/www/pdf/EDIN-Upstream-English.pdf>

content, grain size, texture, and volume of sand and clay in the rocks (Schlumberger 2019).

In the given dataset, lithology is a categorical feature and has values like sandstone, quartzite, limestone, silty limestone, algal limestone, etc. In order to better understand lithology, we carefully analyzed all unique values for lithology in the given dataset. The analysis showed that the values belong to different levels of the hierarchy described in “[Formal knowledge/hierarchies](#)”. For example, the values ‘silty limestone’ and ‘algal limestone’ both can be generalized as limestone. Furthermore, based on prior geological knowledge, lithologies are divided into three high-level classes: sedimentary rocks, metamorphic rocks and igneous rocks. These main concepts are further divided into sub-levels until we reach rock names defined at fine-grained levels such as a specific kind of limestone, e.g. ‘silty limestone.’

However, there is no standard naming convention or formally defined taxonomy that explains all kinds of lithologies and their hierarchical groupings. For utilizing this information in “[An alternative approach:analogue recommendation](#)”, we have developed a taxonomy that groups all the values based on similar physical characteristics of rocks. Fig. 1 shows part of the taxonomy.

Depositional Environment: A depositional environment describes the combination of physical, chemical, and biological processes associated with the deposition of a particular type of sediment and, therefore, the rock types that will be formed after lithification, if the sediment is preserved in the rock record (Wikipedia 2020).

Fig. 1 Part of taxonomy for Lithology

The depositional environment is a categorical feature and examples of typical values are ‘aeolian,’ ‘fluvial,’ etc. For formalizing the depositional environments, four high-level classes have been defined: continental, transitional, marine and others. Then, these high-level classes have been specialized into sub-classes and instances. As an example, the transitional depositional environments have been divided into beach, deltaic, and lagoonal depositional environments. Furthermore, the delatic depositional environments have been subdivided into prodelta, delta front, and delta plain. Based on all unique depositional environments occurring in the dataset, we have formalized the classification hierarchy in the form of the taxonomy shown in Fig. 2.

Age: A geologic age is a subdivision of geologic time that divides an epoch into smaller parts. A succession of rock strata laid down in a single age on the geologic timescale is a stage. The geological time has been divided into high level classes, know as eras. Eras are further divided into periods and epochs. The naming convention of the geological times is based on fossil evidence (International Commission on Stratigraphy 2019).

In the given dataset, reservoir age is a categorical feature and comprised of values that belong to age, periods and epochs. Fig. 3 shows a part of the taxonomy developed to classify all values as geological times organized as Era, Period, Epoch and Age.

Trap Type: A trap is a configuration of rocks suitable for containing hydrocarbons and sealed by a relatively impermeable formation through which hydrocarbons will not

Fig. 2 Part of taxonomy for Depositional Environment

Fig. 3 Part of taxonomy for Geological Age

migrate. Traps are described as structural traps (in deformed strata such as folds and faults) or stratigraphic traps (in areas where rock types change, such as unconformities, pinch-outs and reefs) (Schlumberger 2019). This is a categorical feature and contains combinations of various trap types. We did not construct a hierarchy for the different trap types.

Porosity: The percentage of pore volume or void space, or that volume within the rock that can contain fluids. Porosity can be a relic of deposition (primary porosity, such as space between grains that were not compacted together completely). It can also develop through alteration of the rock (secondary porosity, such as when feldspar grains or fossils are preferentially dissolved from sandstones) (Schlumberger 2019). In the given data set, porosity is numeric attribute and has values in range 0–48.5.

Permeability: The ability, or measurement of a rock's ability to transmit fluids. It is typically measured in darcies or millidarcies (Schlumberger 2019). Formations that transmit fluids readily, such as sandstones, are described as permeable and tend to have many large, well-connected pores. Impermeable formations, such as shales, tend to be finer grained or of a mixed grain size, with smaller or less interconnected pores. Permeability is a numeric attribute and has values between 0 and 18500md.

Gross Thickness: It is the thickness of the stratigraphically defined interval in which the reservoir beds occur, including

such non-productive intervals as may be interbedded between the productive intervals (Wikiversity 2015). It defines the thickness of the whole reservoir. In the given dataset, the gross thickness is a numeric feature with values in the range 0–13000 feet.

Top Depth: Top depth is the depth at the top of the geological unit representing the reservoir rock. Top depth is a numeric feature and has values in the range 0–33000 feet.

Data validation & standardization

This step involves the in-depth process of analyzing and cleaning the data to ensure that the statistical model gets the correct input. The input features are mixed-type, comprising both categorical and numeric features. Figures 4 and 5 show the distribution of the categorical features 'Reservoir Age' and 'Reservoir Lithology.' As can be observed that, few values are frequent while others occur rarely, hence resulting in highly skewed data distribution for both features. The plot for Trap and Depositional Environment shows the same pattern. Thus, it is important to pre-process the data before statistical modeling. Detailed pre-processing steps for each feature are described below.

Age: There are about 250 unique categorical values that occur in the given dataset. As shown in Fig. 4, the data distribution for this feature is skewed. A detailed manual analysis of the actual values showed that the majority of the values that occur only once in the dataset are non-standard age names: the same geological age is denoted by different names. We replace the non-standard names by the International terms defined in the ICS.

Trap Type: Trap type consists of 36 unique raw values. These 36 values are based on different combinations of the following five traps types:

- Structural
- Stratigraphic
- Unconformity
- Continuous-type
- Hydrodynamic

The original trap column is replaced by five new columns, each column representing a unique trap category. The combination of these categories for each reservoir data point is shown by using 'True' values for the categories present in the actual trap column.

Lithology: There are 1731 raw categories of lithology before pre-processing the data. These values contain different abbreviations for the same lithology type, unofficial

Fig. 4 Histogram for the most common types of Reservoir Age

lithology names, and combinations of various lithologies. First, based on the feedback from experts, unofficial names and different abbreviations are identified and replaced by standard domain names. Then combinations are discarded and only the primary lithology for the respective reservoir is stored in the database. In addition, all the values that occur once and belong to the lowest level of the taxonomy, are replaced by the parent concept.

Depositional Environment: There are 32 types of unique categories for the Depositional Environment. Some of the categorical values occurring only once are replaced by the parent categories.

Porosity, Permeability, Gross Thickness & Top Depth For numeric features, we performed an outlier analysis. Outliers are extreme values that deviate from the majority of

Fig. 5 Histogram for the most common types of Lithology

the observations in the data and identify variability in measurements due to human or experimental error. Figure 6 shows the data distributions for all numeric features.

We used the interquartile range rule for the detection of outliers. The interquartile range (IQR), is a measure of statistical dispersion, and is equal to the difference between the 75th and 25th percentiles, or equivalently, between the upper and lower quartiles (Graham Upton).

$$IQR = Q_3 - Q_1 \quad (1)$$

Any number that is more than 1.5 times the IQR away from either the lower resp. upper quartiles for a feature is considered an outlier:

$$\begin{aligned} \text{Outliers} = \\ \{x \mid x > (Q_3 + 1.5 * IQR) \text{ or} \\ x < (Q_1 - 1.5 * IQR)\} \end{aligned} \quad (2)$$

All outliers according to this criterion were removed from the data.

Parameter prediction

Porosity and permeability are two of the fundamental parameters of a reservoir rock. While porosity refers to

the amount of empty space (pores) within a given rock, permeability is a measure of the ability of a porous material (often, a rock or an unconsolidated material) to allow fluids to pass through it.

Porosity is represented as the ratio between pore volume and the bulk rock volume. It does not give any information about pore sizes, their distribution, and their degree of connectivity. Thus, rocks of the same porosity can have widely different permeability. An example of this might be a carbonate rock and a sandstone. Each could have a porosity of 0.2, but carbonate pores are often very unconnected resulting in its permeability being much lower than that of the sandstone.

Permeability is related to the porosity, but also to the pore geometry (shapes of the pores) and their level of connectivity, making it the most difficult parameter to predict.

The complexity of the relationship between permeability and pore geometry has resulted in much research. No fundamental law linking the two has been found. Instead, many empirical approximations for calculating permeability have been developed. A good description and comparison of all these approaches can be found on (PetroWiki 2015). All of these approaches rely on other parameters beside porosity, such as pore throat radius, grain diameters, water saturation, which are not present in our data set.

Therefore we have concentrated our efforts in predicting porosity only, and not permeability.

Fig. 6 Histogram for Numeric Features

Fig. 7 Flow diagram for the proposed prediction method

The method we followed for porosity prediction, consists of five steps: Data visualization, categorical data encoding, knowledge-based feature augmentation, statistical learning and experimental evaluation. Figure 7 shows the overall prediction framework. After data visualization, categorical features are encoded using encoding schemes discussed in “[Encoding techniques](#)”. First the data is directly used for porosity prediction. Further to analyze the effect of domain knowledge in the prediction process, we added additional features in the dataset. Below, we provide a detailed overview of all steps.

Data visualization

To aid the prediction process and to understand the correlation between input and output features first, we tried to visualize the relationship between the numeric input features and porosity. Figure 8 shows the relationship between the porosity and the other numeric features like top depth and gross thickness. For porosity range 0–40, we can notice an entire range of values for gross thickness and top depth, with a significant amount of overlap between

values. The pairwise plots show that the relationships are both complex and noisy.

Further, to visualize the relationship between categorical features and ‘Porosity,’ we used box-plots. A box plot shows the distribution of categorical data in a way that facilitates comparisons between features or across levels of categorical features. Figure 9 shows the porosity distribution for the most frequent categories in the depositional environment. Generally, the same trend can be observed for the majority of depositional environments. The mean porosity lies in the range 10–20% and the overall significant amount of data falls into the same range of 5–25 for every single category. This shows a huge degree of overlap for porosity values in various depositional environments. The value ‘Restricted Marine’ appears to correspond to significantly lower porosity values, but it accounts for only 1.26% of the data points. Predictions for this case would therefore be neither very reliable nor very useful in terms of overall precision.

A similar analysis for the lithology and age features also shows a significant overlap for related porosity values. Overall, the visualizations show the complex

Fig. 8 Pairwise Relationships: Porosity vs. Gross Thickness & Porosity vs. Top Depth

Fig. 9 Box plot for Depositional Environment

nature of the domain and the challenges inherent in the dataset.

Encoding techniques

As most of the machine learning algorithms accept only numeric input, it is necessary to first encode categorical features into numerical values using encoding techniques. We experimented with one-hot, binary, ordinal, hashing and target encoding (Potdar et al. 2017). What is common to all of these encodings is that they do not take into account the semantic similarity between the categorical values. No ‘domain knowledge’ is captured in the encodings.

Knowledge-based feature augmentation

The key idea here is to combine formal knowledge in the sense of “[Formal knowledge/hierarchies](#)” with standard machine learning approaches (the ML techniques we used are given in “[Statistical learning & evaluation](#)”). Our intuition here is simply to enhance the performance using domain knowledge. Therefore, in this problem setting, each model is provided with an additional form of information as input. We say that we use ‘knowledge-based features.’ For each categorical feature, additional features are extracted from the taxonomy, based on the value’s ancestors at the level above the leaf nodes. Table 3 shows one such example where two additional features ‘Epoch’ and ‘Period’ are added for feature ‘Age’. The values for feature ‘Age’ are leaf

nodes in the taxonomy, while ‘Epoch’ and ‘Period’ shows the grouping of various ages based on their similarity.

These additional categorical features were then encoded in the same way as the original categorical features, see “[Encoding techniques](#)”.

Statistical learning & evaluation

We attempted to use a variety of machine learning models to predict porosity. Among the discussed methods, the Random forest, Bayesian regression and multi-layer perceptron algorithms performed best for the given dataset and are included in this section. The performance of these models is compared and measured using the explained variance. The explained variance indicates a discrepancy between a model and actual data. Higher values of explained

Table 3 Knowledge-based Features for ‘Age’

Age	Epoch	Period
Bashkirian	Pennsylvanian	Carboniferous
Bathonian	Middle Jurassic	Jurassic
Berriasian	Lower Cretaceous	Cretaceous
Burdigalian	Miocene	Neogene
Callovian	Middle Jurassic	Jurassic
Cambrian	Cambrian	Cambrian
Campanian	Upper Cretaceous	Cretaceous
Capitanian	Guadalupian	Permian

Table 4 Explained Variance(%) for Porosity Prediction

Encoding Scheme	Multilayer Perceptron Regression	Random Forest	Bayesian Regression
One-Hot	37	41	40
Target	38	41	42
Binary	34	39	14
Hashing	8	7	3

variance indicate strong associations and better predictions. At least 60% of explained variance should be achieved to make quality predictions (Joseph et al. 2014).

Tables 4 and 5 show the explained variance for the algorithms mentioned above. Based only on input features, the maximum explained variance for neural networks is 38%, using target encoding. A slight improvement in the variance is observed using random forest and Bayesian regression. The maximum explained variance is 42%, achieved using Bayesian regression with target encoding.

Based on the given input data, these models only account for about 40% of the variance in the data. Statistical models perform well for modeling relationships between numeric input parameters and output variables. However, for the given dataset, the high cardinality of the categorical features and the missing semantic context makes it extremely difficult to train statistical models that can capture the inherent complexities of the oil and gas domain.

A certain improvement can be achieved by using domain information as described in “[Knowledge-based feature augmentation](#)”, see Table 5. For multi-layer perceptrons and Bayesian regression, we can observe approximately 3% increase in explained variance. The random forest gives a maximum variance of 53% with a 12% increase as compared to the previous results. A significant improvement in the explained variance for random forest is justified in this scenario as the domain information aids the model by explicitly specifying the grouping of categorical variables that have similar characteristics. However, the results show that based on the given input features, none of the methods provide good predictions of porosity.

Table 5 Explained Variance(%) for Porosity Prediction with Domain Knowledge

Encoding Scheme	Multilayer Perceptron Regression	Random Forest	Bayesian Regression
One-Hot	42	50	45
Target	41	53	44
Binary	23	51	40
Hashing	21	37	21

Analysis: porosity cannot be predicted from the given input

The main reason for the low variance of the above-mentioned models is the high degree of overlap for the porosity values for different combinations of input parameters, see the visualization shown in “[Data visualization](#)”. The influence of the input parameters on the numeric output variable is dwarfed by the variance of the output for almost identical inputs.

In particular, in the most common cases (sandstone, albian & sandstone, campanian), identical input parameters are associated with very different porosities. Figure 10 shows that this variance is so large that even for these two cases, predicting porosity based on this input data is not possible with interesting precision. Theoretically, this could be a consequence of unlucky sampling, but since we use a database of all discovered reservoirs, that seems very unlikely. The situation only gets worse by the fact that there are many other combinations of input parameters that occur rarely or never in the data.

From the point of view of conventional geoscience, this finding is not very surprising. Our analysis confirms that data-based prediction of porosity (from the parameters we take into account) cannot add anything to the conventional methods used by geologists.

The quality of a reservoir rock is mainly defined by its lithology, porosity, permeability, thickness, and lateral distribution. Among these parameters, porosity and permeability are considered to be the most critical, but uncertain, parameters in the prospect evaluation.

Porosity and permeability of rock are strongly affected by post-depositional physical and chemical processes. These processes are constrained by the fundamental variables such as pressure, temperature, time, depositional environment, composition, grain size and texture.

Precise prediction of present-day porosity and permeability through statistical machine learning requires the parameters mentioned above in a format that can be input to the model. The dataset used in this study consists of only a few of these parameters, such as depositional environment, lithology and geological age. Beside the incompleteness and sparseness of the data, there is not sufficient correlation between these parameters and porosity, making the prediction of porosity for this particular dataset an impossible task.

Alternative approaches to support early prospect evaluation

To summarise, the initial idea of estimating crucial reservoir parameters like porosity from the immediately viable input

Fig. 10 Porosity Distribution for Young Age: Campanian and Albian

parameters fails due to available viable parameters and the nature of the problem itself.

What can be done about this? Evidently, more information needs to be included in the task. Going back to the discussion in “[Relevant information for reservoir evaluation](#)”, this would be data that was not considered viable for our experiments so far:

- unstructured data like seismic, natural language resources, well logs if available
- vast geological background knowledge

When it comes to seismic data, approaches from signal processing might be applied, but it is not clear what features in seismic data would be useful for parameter prediction. The limited number of reservoirs for which data is available precludes using methods like deep learning.

Natural language resources are also difficult to use: in order to be useful, the content of the documents has to be reliably connected to the available structured data. Moreover, the documents are in an extremely technical language specific to the geological domain, so that standard natural language processing tools are not applicable without modification. Again, the available corpus is limited by the number of well-documented reservoirs.

It seems that the only way around these difficulties is to involve the human experts in the task: rather than trying to automate parameter prediction completely, we should help geologists and make their task easier.

The following section describes how a computerised method can be used to augment the current standard way of doing reservoir evaluation.

An alternative approach: analogue recommendation

When experts evaluate prospects today, they base their estimates on their knowledge about other, similar areas

that can be expected to have similar properties. Such areas are known as *analogues*. We will see that the search for suitable analogues is challenging, but can be supported by a computer.

Finding and using analogues: state of the art

A global survey conducted by (Sun and Wan 2002) indicates that Explorationists consider the use of geological analogues to be an import tool for improving exploration success, but still, the workflows currently in use lack a systematic approach for selecting and using analogues.

Finding the best reservoir analogues is a complex process and is based on matching at several levels (location or proximity, geological, petrophysical, fluid and engineering properties). With the workflows currently in use, selection of analogues is solely dependent on the evaluator’s judgment and experience in various domains of geoscience. In most cases, geoscientists rely on using only a single analogue, which carries inherent risk and is insufficient to benchmark or characterize an opportunity. Such *ad hoc* methods usually result in inappropriate analogues being applied for comparative purposes. Nowadays, extensive analogue databases are available to improve the understanding of geological processes, to assess the relationships between different reservoir features, and to quantify the distribution of reservoir rocks and reservoir characteristics. Due to various sources available for retrieving information (geo-applications, databases, knowledge bases, reports, images), their variety (structured, unstructured, semi-structured information) and veracity (contradictory information from different sources, duplication, information based on interpretation rather than facts), the selection and usage of analogues is an extremely sophisticated and time-consuming process.

Reservoir analogues are not only used for estimating the resource potential, they are also important for reserves calculation and economic evaluation. For the economic

evaluation, strict guidelines are defined by the SEC and SPE (Hodgin and Harrell 2006), for when a reservoir may be considered an analogue. The general definition of the reservoir analogue is similar in both cases, but SPE includes several additional geological and petrophysical parameters. The matching criteria defined by SPE are considered as a reference in this study. The SPE defines an analogous reservoir as “one in the same geographical area that is formed by the same, or very similar geological processes as, a reservoir in question (or under study for reserves evaluation) as regards sedimentation, diagenesis, pressure, temperature, chemical and mechanical history, and structure. It also has the same or similar geologic age, geologic features, and reservoir rock and fluid properties. Analogous features and characteristics can include approximate depth, pressure, temperature, pay thickness, net-to-gross ratio, lithology, heterogeneity, porosity, and permeability. The development scheme for a reservoir (e.g. as reflected by well spacing) can also be important in establishing the relevance of the analogy”⁴

A study conducted by (Hodgin and Harrell 2006) compares the SEC and SPE criteria and recommends the following guidelines to reduce mistakes in using analogues.

- Give preference to analogues in the closest vertical and/or areal proximity to the target field or reservoir
- Follow a strict process whereby key parameters are tabulated and compared
- Accept an analogue only if a good match exists or if adjustments can be quantified to account for differences
- Account for and, if necessary, adjust for operational similarity, including the analysis of the costs and associated economic viability for replicating the same operational conditions in the subject reservoir
- Know the reserves definition and guidelines relevant to the assessment of the reserves
- Do not accept a case as compelling unless all the key parameters meet the statutory guidelines

In order to perform prospect evaluation based on the guidelines given by SPE and SEC, it is important to automate the process of prospect recommendation. The traditional machine learning approaches fail in this scenario because of the complexity and variety of the domain and the limited information. Our key intuition is to use the concept of semantic similarity in the recommendations, using a standard knowledge source and to develop a method that is flexible enough to produce analogues for reservoirs with no neighboring analogues.

Several of the relevant features of reservoirs are categorical, so part of our task is to choose suitable similarity measures for categorical data. Not taking into account any domain knowledge (e.g., the hierarchical structures on age, lithology, etc.), the only basis on which to define a similarity measure is the relative frequency of values in the data set. In the following, first we provide an overview of the existing frequency-based similarity measures, followed by a detailed explanation of semantic similarity measures (i.e., measures that also take into account domain knowledge) and the proposed techniques.

Frequency-based similarity

For unsupervised learning, the Hamming distance is used and overlap similarity is defined as a matching measure that assigns 1 if both values are identical, and 0 otherwise (Esposito et al. 2000; Ahmad and Dey 2007). The overall similarity between two data instances of multivariate categorical data is proportional to the number of attributes in which they are identical. The overlap measure does not distinguish different values of attributes, hence matches and mismatches are treated equally. Various similarity measures have been derived using this distance measure, e.g. the Jaccard similarity coefficient, the Sokal-Michener similarity measure, the Grower-Legendre similarity measure, etc. (Esposito et al. 2000). These measures are inherently quite coarse: in the absence of an ordering between the categorical values, the only possible distinction is whether two values are identical or not (Esposito et al. 2000).

To improve on these measures, *frequency-based* similarity measures have been proposed that take the frequency distribution of different attribute values into account. These measures are data-driven and hence are dependent on certain data characteristics such as the size of data, the number of attributes, the number of values for each attribute and the relative frequency of frequency of each value. Below we provide an overview of existing frequency-based measures.

The Occurrence Frequency (OF) measure assigns lower similarity to mismatches on less frequent values and a higher similarity to mismatches on more frequent items (Boriah et al. 2008) and is defined as follows:

$$OF(x, y) = \begin{cases} 1 & \text{if } x = y, \\ \frac{1}{(1 + \log(\frac{N}{f(x)+1}) + \log(\frac{N}{f(y)+1}))} & \text{otherwise} \end{cases} \quad (3)$$

where $f(x)$ is the number of occurrences of the attribute value x and N the total number of observations in the data set.

Eskin proposed a distance measure for categorical points. This measure assigns higher weight to mismatches of categorical attributes that contain many values. For a single

⁴January 2005 SPE/WPC/AAPG Glossary of Terms Used, <https://www.spe.org/en/industry/terms-used-petroleum-reserves-resource-definitions/>

categorical variable, the distance between two values is defined as $2/n^2$ where n is the number of unique values for that variable. (Boriah et al. 2008) later defined a similarity measure based on that distance as follows:

$$Eskin(x, y) = \begin{cases} 1 & \text{if } x = y, \\ \frac{n^2}{n^2+2} & \text{otherwise} \end{cases} \quad (4)$$

Lin proposed a similarity framework based on information theory (Lin 1998). According to Lin, similarity can be explained in terms of a set of assumptions. If the assumptions are considered true, the similarity measure is necessarily followed. Therefore, the similarity between the two values is calculated by the ratio between the amount of information required to state the commonality of both values and the information needed to fully describe both values separately. Lin's similarity measure is defined as:

$$Lin(x, y) = \begin{cases} 2 \log p(x) & \text{if } x = y, \\ 2 \log(p(x) + p(y)) & \text{otherwise} \end{cases} \quad (5)$$

where

$$p(x) = f(x)/N \quad (6)$$

is the relative frequency of the value x for the attribute in the dataset.

Jian, Cao, Lu, and Gao proposed coupled metric similarity(CMS) to capture the value-to-attribute-to-object heterogeneous coupling relationships (Jian et al. 2018). They defined intra- and inter-attribute couplings to learn overall similarity for categorical values. The similarity between two values of a categorical variable is defined as.

$$CMS(x, y) = \begin{cases} 1 & \text{if } x = y, \\ \frac{\log p(x) \cdot \log p(y)}{\log(p(x) \cdot p(y)) + \log p(x) \cdot \log p(y)} & \text{otherwise} \end{cases} \quad (7)$$

These data-driven measures perform well on simple datasets. However, these measures are unable to take into account semantic relationships and often don't perform well on complex datasets with hidden domain dependencies. Moreover, a concept of similarity that is based solely on how often values occur in the data cannot be expected to give reasonable results in all cases. Using frequencies seems more like a 'last straw' when frequencies are the only distinguishing feature between categorical values.

Semantic similarity

Semantic similarity refers to similarity based on meaning or semantic content as opposed to form (Neil and Paul 2001). Semantic similarity measures are automated methods for assigning a pair of concepts a measure of similarity and can be derived from a taxonomy of concepts arranged in is-a relationships (Pedersen et al. 2005). The

concept of semantic similarity has been applied in natural language processing for the past decade to solve tasks such as the resolution of ambiguities between terms, document categorization or clustering, word spelling correction, automatic language translation, ontology learning or information retrieval. Similarity computation for categorical data can improve the performance of existing machine learning algorithms (Ahmad and Dey 2007) and may ease the integration of heterogeneous data (Wilson and Martinez 2000).

Is-a relationships in a concept hierarchy provide a formal classification. An agreed-on concept hierarchy, e.g. as part of a domain ontology, provides us with a common understanding of the structure of a domain, and makes explicit domain assumptions as far as the hierarchy of concepts is concerned.

In order to achieve interpretable and good quality results in machine learning models, it is vital to take this information into account. This intuition motivates us to link the notion of similarity based on is-a relationships with the similarity measures for the analogue recommendation. We have developed a framework to use is-a relationships extracted from a concept hierarchy to quantify semantic similarity and propose a distance measure for categorical data.

Proposed framework

In the following, we propose two techniques for measuring similarity based on domain knowledge expressed as a concept hierarchy. First, we present a framework for calculating semantic similarity using information content and a concept hierarchy by adapting an idea of (Resnik 1970). This measure is based on both the hierarchy and the relative frequency of values, and we were interested in seeing to which extent the relative frequency is useful at all. We therefore defined a second measure that is based only on the hierarchy.

Further, we are interested in computing global semantic similarity in a multi-dimensional setting where we have several hierarchically structured features. We define the global similarity between two data objects X and Y in a d -dimensional attribute space as

$$\delta(X, Y) = \sum_{i=1}^d w_i \delta(x_i, y_i) \quad (8)$$

where $\delta(x_i, y_i)$ is the similarity between two values x_i and y_i in the i -th dimension and w_i is the weight associated with each dimension. The following section presents both proposals for calculating a semantic based similarity $\delta(x_i, y_i)$.

Information Content Semantic Similarity (ICS)

This approach is as adaptation of an idea presented by Resnik (Resnik 1970). Resnik proposed a measure for finding semantic similarity in an is-a taxonomy based on information content and defined similarity between two nodes in a hierarchy as the extent to which they share common information.

In order to formulate the semantic similarity of two given categorical values, the key intuition is to find the common information in both values. This information is represented by the lowest common ancestor in the hierarchy that subsumes both values (Lin 1998). If the lowest common ancestor of two values is close to the leaf nodes, that implies that both values share many characteristics. As the lowest common ancestor moves up in the hierarchy, fewer commonalities exist between a given pair of values.

For the analogue dataset, we can map the given categorical features to their relevant taxonomies (“[Data definition & mapping to ontologies](#)”) by placing all the values at the corresponding leaf nodes in the defined taxonomies whereas intermediate nodes represent the lowest common ancestors for given pairs. In Fig. 3, ‘Carnian,’ ‘Norian,’ and ‘Rhaetian’ are all subsumed by the lowest common ancestor ‘Upper Triassic’, whereas the lowest common ancestor that subsumes ages ‘Norian’ and ‘Aptian’ is ‘Mesozoic’ (close to the root node of the hierarchy). Hence, taking into account the lowest common ancestor, we expect the similarity between Norian and Carnian to be significantly greater than the similarity between the Norian and Aptian ages.

Our intuition about semantic similarity is that for two categorical values x and y that share a lowest common ancestor c , farthest from the root node, are always considered to be more semantically similar than to two categorical values x and z that share a lowest common ancestor c' close to the root node. In addition, identical values should have a maximum similarity of 1.

For categorical data, we can find the information content I of the lowest common ancestor c by finding the information content of all the leaf values subsumed by c in the hierarchy.

$$I(c) = -\log \sum_{n \in \text{leaf}(c)} p(n) \quad (9)$$

where $\text{leaf}(c)$ is the set of all leaf values in $x \in S$ with $x \sqsubseteq c$. The probability of leaf values may be estimated by the relative frequency using Eq. 6.⁵ Based on the above definitions, we formulate information content based

semantic similarity(ICS) between two categorical values x and y as

$$\text{Sim}(x, y) = \begin{cases} 1 & \text{if } x = y, \\ k \cdot I(x \sqcup y) & \text{otherwise,} \end{cases} \quad (10)$$

where $I(x \sqcup y)$ denotes the information content of the lowest common ancestor of x and y , as defined by Eq. 9. k is used for normalization and defined as

$$k = \frac{1}{\max_{x \in C} I(x)} \quad (11)$$

where $\max_{x \in C} I(x)$ is the maximum information content of all leaves.

Hierarchy-based Semantic Similarity (HS)

As explained earlier, the main intuition of semantic similarity is based on the idea that any two values having the lowest common ancestor close to leaf nodes, should have high similarity and vice versa. Hence, we quantify semantic similarity by considering the level of the lowest common ancestor in the hierarchy. The level of a node is defined as 1 more than the number of connections between the node and the root.⁶ The greater the level of the lowest common ancestor of any given pair of values in the hierarchy, the more similar the values are. We formulate the similarity as

$$\text{Sim}(x, y) = \begin{cases} 1 & \text{if } x = y, \\ \lambda^{d - \text{level}(x \sqcup y)} & \text{otherwise,} \end{cases} \quad (12)$$

where $0 < \lambda < 1$ is a fixed decay parameter, $\text{level}(n)$ is the distance of n from the root in the hierarchy, and $d = \max_{n \in X} \text{level}(n)$ is the maximum depth of the hierarchy.

The main advantage of HS as given by Eq. 12 over ICS from the previous section, is that the calculation of semantic similarity no longer requires any input from training data. Once the concept hierarchy is formalized, we can measure the similarity between any two values, even including categorical values not observed in the training data.

Below, we perform an evaluation for only one-dimensional data, followed by a detailed evaluation for the complete analogues use case in “[Evaluation of similarity measures for analogue recommendation](#)”.

Comparative evaluation of similarity measures

We evaluate the performance of frequency-based and hierarchy-based similarity by considering an example of reservoir age only. We take the *Upper Triassic* epoch as the main data point and compute its similarity to other values in

⁵Probabilities may also be known from other sources, e.g. known priors for the specific domain.

⁶The level of the root is 1, and children have higher levels than their parents.

the dataset. Table 6 shows the similarity score for different pairs of epochs using the methods discussed in the previous section. It is evident from the knowledge source or hierarchy shown in Fig. 3, that *Upper Triassic* has greater similarity to *Middle Triassic* and *Lower Triassic* than to other epochs such as *Jurassic* or *Devonian*. This information is taken as the base criterion to analyse the performance of similarity measures.

First, we analyse the similarity scores calculated using frequency-based methods (OF, CMS, Eskin). In Table 6, the pair (*Upper Triassic*, *Upper Jurassic*) has a higher similarity score than the pairs (*Upper Triassic*, *Lower Triassic*) and (*Upper Triassic*, *Middle Triassic*). Further, we see random similarity scores between various pairs given by frequency-based methods (OF, CMS), with no intuition of similarity based on actual semantics of terms as defined by the knowledge source. These results are not surprising as the similarity in frequency-based methods is purely based on the occurrence frequency of different values. If any two pairs occur more frequently in the given data set, it will result in higher similarity. Frequency-based techniques provides various strategies to take frequency into account, however it is evident from the results that the concept of similarity between two values has no correlation with how many times these values occur in the given data set.

For Eskin, similarity score depends on the cardinality of the given variable. For high cardinality variables, similarity between any given pair will always be close to 1 irrespective of the actual similarity. This is obvious in the Table 6, all pairs of unequal values have similarity score of 0.9 except (*Upper Triassic*, *Upper Triassic*) that are identical and have similarity score of 1. This explains the poor performance for in our use case, as we have all variables with high cardinality.

The hierarchy-based information content method *ICS* also assigns higher similarity to the pair (*Upper Triassic*, *Upper Jurassic*) as compared to (*Upper Triassic*, *Lower Triassic*). The main reason for the poor performance of *ICS* is that it relies on the calculation of information content

for the common ancestors in the hierarchy. The occurrence probability of all children of the common ancestor has a significant impact on the calculated information content. Nodes that have enough representation in the actual dataset, they will end up with a lower IC value. This ultimately results in lower similarity scores for the sibling nodes. Hence, combining information content with hierarchy may not yield best results in similarity scores.

The only method that assigns a higher similarity to the pair (*Upper Triassic*, *Lower Triassic*) than to the others is the *hierarchy-based similarity HS*. Hierarchy-based similarity does not rely on the occurrence frequency of different values, instead similarity is calculated based on the similarity information embedded in the hierarchy/knowledge source.

Evaluation analysis by geologist

To consider two reservoirs as geological analogues depends on matching them at multiple levels. No single matching criterion is enough, and there is very little probability of finding an analogue identical to the candidate reservoir.

The similarity between values of a single reservoir parameter can be defined by using hierarchies. The closer the parameters are in the hierarchy, the higher their similarity in terms of geological signature. For example, sediments deposited during upper triassic and middle triassic ages tend to be more similar than sediments deposited during upper jurassic.

The similarity (and dissimilarity) in this case arises from the fact that the physical, chemical and biological processes associated with the upper and middle triassic ages are relatively similar but are very different from the processes associated with the upper jurassic. These processes result in relatively similar reservoir properties for the two reservoirs deposited during the triassic as compared to the reservoirs deposited in jurassic. Similar relationships exist for the other geological parameters, e.g., proximity, depositional environment, rock type, lithology, etc.

The (manual) workflows currently in use in the geoscience community use a very similar approach for finding and using analogs in the subsurface evaluation. Geoscientists start their search by looking into nearby areas where the geological age, depositional environment, and other geological parameters are similar to the candidate reservoir. Search is then expanded by including information from other areas and by increasing tolerance on the matching criteria.

Most of the time, the volume of the information available to the geoscientists is enormous, while the information itself is unreliable. The absence of machine-assisted workflows and tools for simultaneously matching hundreds of reservoirs at multiple levels with flexible and hierarchical

Table 6 Similarity Scores for Upper Triassic & other epochs

	HS	ICS	OF	CMS	Eskin
Upper Triassic, Upper Triassic	1.00	1.00	1.00	1.00	1.00
Upper Triassic, Lower Triassic	0.064	0.078	0.025	0.62	0.9
Upper Triassic, Middle Triassic	0.064	0.078	0.035	0.71	0.9
Upper Triassic, Lower Jurassic	0.008	0.086	0.039	0.73	0.9
Upper Triassic, Middle Jurassic	0.008	0.086	0.051	0.75	0.9
Upper Triassic, Upper Jurassic	0.008	0.086	0.053	0.75	0.9
Upper Triassic, Lower Devonian	0.008	0.068	0.040	0.73	0.9
Upper Triassic, Upper Devonian	0.008	0.068	0.035	0.72	0.9

matching criteria makes it extremely challenging to utilize the available information fully.

Evaluation of similarity measures for analogue recommendation

In this section, we compare the ICS and HS approaches to other similarity measures for the recommendation of reservoir analogues of a target reservoir, given the dataset of known reservoirs.

Baseline Methods

The following four state-of-the-art similarity/distance measures are compared with the proposed techniques: occurrence frequency (Eq. 3), Eskin similarity measure (Eq. 4), Lin Similarity measure (Eq. 5) and Coupled Similarity Matrix (Jian et al. 2018).

We compare the performance of the different similarity measures in a recommendation scenario: given a query item, we compute its similarity to each item in the ‘training’ dataset using Eq. 8, and determine the top k items with highest similarity. For our evaluation, we do this for all of the different similarity measures, and compare the outcome to a fixed ‘gold standard’ list of items to determine the average precision.

Evaluation method

For the given task, we will evaluate the similarity measure on two main objectives.

- Retrieving the 15 most similar analogues to the target reservoir.
- Producing the result in a ranked order such that the first retrieved analogue corresponds to the most similar reservoir to the target reservoir.

Mean Average Precision (MAP) is the mostly commonly used evaluation metric in information retrieval and object detection (Baeza-Yates and Ribeiro-Neto). MAP is the arithmetic mean of the average precision (AP) values for an information retrieval system over a set of n query topics (Ling and Özsu 2009). It can be expressed as follows:

$$MAP = \frac{1}{n} \sum_n AP_n \quad (13)$$

The precision for a classification task is defined as

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (14)$$

Based on Eq. 14, recommender system Precision (P) is defined as,

$$P = \frac{\# \text{ of our recommendations that are relevant}}{\# \text{ of items we recommended}} \quad (15)$$

For evaluating the performance of recommender systems, we are only interested in recommending the top- N items to the user. Usually, the higher the number of relevant recommendations at the top, the more positive is the impression of the users. Therefore, it is sensible to compute precision and recall metrics in the first N items instead of all the items. Thus the precision at a cutoff k is introduced in order to evaluate ranking, where k is an integer that is set by the user to match the objective of the top- N recommendations. The average precision at cutoff k , is the average of all precisions in the places where a recommendation is a true positive and is defined as follows:

$$AP_q @ K = \frac{1}{K} \sum_{i=1}^K P(i) \cdot Rel(i) \quad (16)$$

where K represents the top K recommendations for the given query q and $Rel(i)$ shows the relevance of the recommendation. $Rel(i)$ is 1 if the recommended item was relevant (true positive) otherwise 0.

Usually, the performance of a recommendation system is calculated by considering a set of queries. Therefore, given a set of queries Q , the mean average precision ($MAP_Q @ K$) of an algorithm is defined as

$$MAP_Q @ K = \frac{1}{Q} \sum_{q=1}^Q AP_q @ K \quad (17)$$

where $AP_q @ K$ is calculated by using Eq. 16

Experimental results

There is no standard way to evaluate similarity measures for semantic similarity. Resnik uses human expert similarity ranking to judge similarity (Resnik 1970). We have followed the same approach. In order to perform this evaluation, we selected two target reservoirs ‘Snorre’ and ‘Snøhvit.’ We then asked our domain experts to produce a gold set for each reservoir. This gold set contains a set of reservoirs identified by our experts as most similar to the target reservoir based on their hindsight knowledge about the target reservoir. Furthermore, the gold set is produced in a ranked manner, the first item in the list corresponds to the most similar analogue and the last item corresponds to the lowest similar reservoir.

After acquiring the gold dataset, we perform an experimental evaluation to compare the performance of

Table 7 Average Precision for the selected target reservoirs (%)

Reservoir	ICS	HS	OF	CMS	Eskin
Snorre	39	59	39	40	34
Snøhvit	57	66	15	29	27

the proposed semantic similarity measures of “**Semantic similarity**” (ICS, HS) to the three frequency-based measures of “**Frequency-based similarity**” (OF, Eskin, CMS) for finding reservoir analogues. For each selected target reservoir, all the remaining reservoirs in the dataset are given as input to each similarity measure and the similarity between the target and all remaining reservoirs is calculated. The top 15 reservoirs with maximum similarity are retrieved and are now referred to as analogues to the target reservoir.

In order to penalize poor estimations, we are using Average Precision (Eq. 16) as a quality criterion for the ranked lists. For this metric, a higher value corresponds to better results. Table 7, shows the experimental result of each similarity measure separately for each target reservoir.⁷

Table 7 shows that ICS and HS outperform the frequency-based similarity measures for both selected reservoirs. For the target reservoir ‘Snøhvit,’ the average precision for ICS is 57% which is higher than the average precision of other similarity measures. For ‘Snorre,’ ICS gives the average precision of 39% which is higher than the Eskin score but comparable with the results of CMS and OF measures.

HS gives the average precision for ‘Snorre’ and ‘Snøhvit’ as 59% and 66%, much higher as compared to the all other measures. Further, Table 8 shows that the MAP (Eq. 17) for ICS and HS is 48% and 63% respectively, which significantly better than the MAP values of other algorithms. This evaluation supports the initial hypothesis that by adding domain information to the similarity measure, we can define more useful similarity measures on complex categorical data.

It is important to note that results obtained using ICS and HS are not directly comparable to the gold set provided by human experts. In order to produce a gold set, human experts take into account the geological history of the current basin, the analysis of geological time periods and overall processes of formation of reservoir rocks. Furthermore, they also use conceptual facies models,

Table 8 Mean Average Precision (%)

	ICS	HS	OF	CMS	Eskin
MAP	48	63	27	35	30

reservoir simulation models, core samples and well logs for selecting appropriate analogues. In contrast to this, our experimental evaluation of the proposed technique is based only on a limited part of this information. It is remarkable that the precision of 63% could be achieved in this scenario based only on the hierarchy-based semantic measure for such a limited number of parameters.

Conclusion

In this paper, we provide a detailed overview of the complexity involved in directly predicting porosity and prediction using machine learning techniques and provide an alternate way of prospect evaluation by finding analogues using limited information and semantic similarity.

We propose two methods to calculate semantic similarity based on domain information extracted in the form of is-a links from the concept hierarchy and perform a comparative evaluation with existing frequency-based similarity methods. The experimental results show that by incorporating formal domain knowledge, we successfully retrieve similar analogues based on quite limited information about a prospect. Further, the results show that frequency-based methods do not represent the semantic similarities that are inherently present in the domain concepts and there is no link between occurrences of various concept pairs with their actual conceptual or domain-based similarities. Also the combination of information content with hierarchies does not produce improvements in the similarity score because the calculation of information content is highly influenced by the occurrence probability of various concepts. A significant improvement in the results by calculation of similarity purely based on the hierarchy is justifiable due to the fact that similarity is not dependent on the random count of the concepts in the given dataset. Instead, it is based on the criteria provided by the human experts as a formal representation of domain knowledge.

In our current work, we approach the problem by considering the lowest common ancestor in the concept hierarchy by considering mono-hierarchies only, and in an unsupervised setting. In the future, we want to extend the notion of semantic similarity for categorical input data in a supervised setting. Further, we are interested in finding

⁷The similarity measure proposed by Lin (Lin 1998) doesn’t retrieve any similar analogues in the top-*k* recommendations. Therefore, results are not included in Table 7.

similarities between terms in poly-hierarchies where each concept may have multiple parents.

Acknowledgements The authors would like to thank Equinor ASA, in particular Jens Grimsgaard for his support in negotiating access to reservoir data, and Vegard Sangolt for useful discussions. We are grateful to IHS Markit for supporting the project by granting access to their data set. This work was supported by the Norwegian Research Council via the SIRIUS Centre for Research Based Innovation, Grant Nr. 237898

Funding Open Access funding provided by University of Oslo (incl Oslo University Hospital).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abdideh M. (2012) Estimation of permeability using artificial neural networks and regression analysis in an iran oil field. *International Journal of the Physical Sciences* 7:09
- Ahmad A., Dey L. (2007) A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. *Pattern Recogn Lett* 28(01):110–118. <https://doi.org/10.1016/j.patrec.2006.06.006>
- Baeza-Yates R., Ribeiro-Neto B. *Modern Information Retrieval: The Concepts and Technology Behind Search*. Addison-Wesley Publishing Company, 2nd edition, 2008. ISBN 9780321416919
- Boriah S., Chandola V., Kumar V. (2008) Similarity measures for categorical data: A comparative evaluation. In: *proceedings of the SIAM International Conference on Data Mining*, volume 30, pages 243–254, 04. <https://doi.org/10.1137/1.9781611972788.22>
- Brigaud B., Vincent B., Durlet C., Deconinck J.-F., Jobard E., Pickard N., Yven B., Landrein P. (2014) Characterization and origin of permeability–porosity heterogeneity in shallow-marine carbonates: From core scale to 3d reservoir dimension (middle jurassic, Paris basin, France) . *Mar Pet Geol* 57:631–651. ISSN 0264-8172 <https://doi.org/10.1016/j.marpetgeo.2014.07.004>. <http://www.sciencedirect.com/science/article/pii/S0264817214002281>
- Esposito F., Malerba D., Tamma V., Bock H.-H. (2000) *Classical resemblance measures*. Springer Verlag, 12. ISBN 3-540-66619-2
- Graham Upton I. C. *Understanding Statistics*. Oxford University, 1996. ISBN 0-19-914391-9
- Helle H., Bhatt A., Ursin B. (2001) Porosity and permeability prediction from wireline logs using artificial neural networks: A north sea case study. *Geophys Prospect* 49(12):431–444. <https://doi.org/10.1046/j.1365-2478.2001.00271.x>
- Hodgin J. E., Harrell D. R. (2006) The selection, application, and misapplication of reservoir analogs for the estimation of petroleum reserves. In: *SPE Annual Technical Conference and Exhibition*, San Antonio, Texas, USA, 09. <https://doi.org/10.2118/102505-MS>
- International Commission on Stratigraphy (2019) *International chronostratigraphic chart*, <http://stratigraphy.org/index.php/ics-chart-timescale>
- Jennings J., Lucia F. (2003) Predicting permeability from well logs in carbonates with a link to geology for interwell permeability mapping. *SPE Reservoir Evaluation & Engineering* 6(08):215–225. <https://doi.org/10.2118/84942-PA>
- Jian S., Cao L., Lu K., Gao H. (2018) Unsupervised coupled metric similarity for non-iid categorical data. *IEEE Trans Knowl Data Eng* 30(9):1810–1823. <https://doi.org/10.1109/TKDE.2018.2808532>
- Joseph H., William B., Barry B., Rolphe A. (2014) *Multivariate Data Analysis*. Pearson Education Limited
- Lacetre P. E., Carrica P. M. (2003) A method to estimate permeability on uncured wells based on well logs and core data. In: *SPE Latin American and Caribbean Petroleum Engineering Conference*, Port-of-Spain, Trinidad and Tobago Society of Petroleum Engineers, 04. <https://doi.org/10.2118/81058-MS>
- Lin D. (1998) An information-theoretic definition of similarity. *ICML Madison* 1:08
- Lin W., Chen L., Lu Y., Hu H., Liu L., Liu X., Wei W. (2017) Diagenesis and its impact on reservoir quality for the Chang 8 oil group tight sandstone of the Yanchang formation (upper triassic) in southwestern Ordos basin, China. *Journal of Petroleum Exploration and Production Technology* 7:1–13. <https://doi.org/10.1007/s13202-017-0340-4>
- Ling L., Özsu M. T. (2009) *Encyclopedia of database systems* springer US
- Neil S., Paul B. (2001) *International encyclopedia of the social & behavioral sciences* elsevier
- Olatunji S., Selamat A., Abdulraheem A. (2011) Modeling the permeability of carbonate reservoir using type-2 fuzzy logic systems. *Comput Ind* 62(02):147–163. <https://doi.org/10.1016/j.compind.2010.10.008>
- Pedersen T., Pakhomov S., Patwardhan S. (2005) Measures of semantic similarity and relatedness in the medical domain *Journal of Biomedical Informatics - JBI* 01
- Perez-Valiente M. L., Rodriguez H. M., Santos C. N., Vieira M. R., Embid S. M. (2014) Identification of reservoir analogues in the presence of uncertainty. In: *SPE Intelligent Energy Conference & Exhibition*, Society of Petroleum Engineers. ISBN 978-1-61399-306-4 <https://doi.org/10.2118/167811-MS>
- PetroWiki (2015) Comparison of permeability estimation models, https://petrowiki.org/Comparison_of_permeability_estimation_models
- Potdar K., Pardawala T., Pai C. (2017) A comparative study of categorical variable encoding techniques for neural network classifiers. *Int J Comput Appl Technol* 175(10):7–9. <https://doi.org/10.5120/ijca2017915495>
- Sun S. Q., Wan J. (2002) Geological analogs usage rates high in global survey. *Oil Gas J* 11:46–49
- Resnik P. (1970) Using information content to evaluate semantic similarity in a taxonomy. *IJCAI* 95:02
- Schlumberger (2019) Schlumberger oilfield glossary, <https://www.glossary.oilfield.slb.com/>
- Wikipedia (2020) Depositional environment, https://en.wikipedia.org/wiki/Depositional_environment. Retrieved 2020-07-09

- Wikiversity (2015) Introduction to petroleum engineering, https://en.wikiversity.org/wiki/Introduction_to_petroleum_engineering
- Wilson D., Martinez T. (2000) Improved heterogeneous distance functions. *J of Artif Intell Res* 6:06
- Zehui H., John S., Mark W., John K. (1996) Permeability prediction with artificial neural network modeling in the venture gas field, offshore eastern canada *Geophysics* 61 <https://doi.org/10.1190/1.1443970>
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.