Introduction

Oil-oil and oil-source rock correlations are key targets in petroleum geochemistry. The former aims to determine whether the oils were generated and expelled from the same source facies, the latter to narrow down the specific source facies that gave rise to an oil or oil family, respectively. These studies endeavour to establish the original genetic relationship of geologic samples rather than a mere matching of bulk properties or chemical compositions.

Secondary processes (e.g. thermal maturation, migration, and in-reservoir alteration such as biodegradation, thermochemical sulphate reduction and phase separation) can alter the oil composition, and thus two or more petroleum samples with very different bulk, molecular, and isotopic composition may still be genetically related (Walters, 1998). Moreover, oils are often mixtures from different source rock facies or may have on their migration route taken up oil-components from (presumably) less mature source rocks (Farrimond, 2019).

Therefore, it is recommended to use all available compositional information to classify the oils, i.e. different hydrocarbon ranges (gases, volatiles, high boiling compounds) including biomarkers and stable carbon isotopic compositions, in order to decipher potential contribution from different source rocks. This list may also include more unusual parameters like the hydrogen, sulphur, and nitrogen stable carbon isotopic composition, metal contents and their respective isotopic composition, as well as results of other chemical characterisation methods.

Today a more or less defined set of parameters is applied in many cases for such oil-oil and oil-source rock correlations. Such classifications are often subject to human bias as data quantity and range is difficult to process, and in the case of conflicting parameters the data interpreter may face the dilemma to prioritize one set of results over another despite lack of clear scientific knowledge which parameter to give precedence. Especially in more complex petroleum plays processes as described above may become more prevalent.

Data driven approaches for oil typing become more feasible as both the amount and quality of available data increases. Roughly speaking there are two classes of machine learning approaches that can augment our oil typing capabilities: multi-dimensional clustering and supervised classification. Clustering techniques can provide a less unbiased grouping and classification of oil samples, compared to traditional approaches, and help identify unexpected or previously unknown subtypes of oil.

In many cases there are geographic areas with high confidence on the type and origin of oil samples. In such cases, supervised oil type classification can be used to make predictions in low confidence and frontier areas based on a priori assumptions on a subset of the oil samples available.

The purpose of this work is to compare both traditional and machine learning approaches for oil-oil and oil-source rock correlation on two large datasets on the Norwegian Continental Shelf.

Data quality

Data quality is a major concern when working with large data sets. From an instrumental analytical view all data have an associated measurement uncertainty. Apart from this over the last decades the main analytes are now readily determined by relatively standardized methods (NIGOGA, 2000).

Clearly, sound and standardized laboratory practices are indispensable to acquire good data, and for more advanced characterisations reliable methods must be stipulated.
Most relevant for this study are natural variations and especially natural modifications of the fluids. The former is often due to facies variations, or to varying petroleum constituents (including hetero compounds) expelled during different expulsion phases due to increasing maturation of the kerogen. The latter are often associated to migration processes and alteration in the reservoir, i.e. evaporative fractionation, secondary cracking and biodegradation. All these processes will inevitably change the petroleum composition as reflected by its different bulk, molecular, and isotopic composition.

The data driven approaches used in this study were enabled by both the amount and quality of the data. We performed a rigorous screening of all fluid data, resulting in the exclusion of a significant number of them as assessed to be dominated by oil-based mud.

**Methods**

**Biomarker approach**

In a classical biomarker practice a range of selected biomarker ratios are often applied in order to perform oil-oil and oil-source rock correlations. As described above this may lead to somehow inaccurate interpretation not only due to human bias but also because of data extent, underexplored ranking of parameters, and eventually restricted constrains of paleoenvironmental and depositional settings of the source rock as well as expulsion and fractionation processes in the petroleum system.

**Machine learning approach**

Both unsupervised clustering and supervised classification of oil types were applied to the data from both the Norwegian Sea and the Barents Sea. Two sets of features were used: Concentration of n-alkanes from GC, and calculated biomarker concentrations and parameters from mass spectroscopy data, respectively. However, we note that this approach is not limited to ‘classical’ geochemical data. In the future also other highly-resolved data like NMR or FT-ICR-MS maybe used as features in a similar way.

Each feature was pre-processed by first normalizing over the entire sample set by subtracting the mean and scaling to unit variance. This is followed by reducing the dimensionality of the problem by using the first five principal components of the normalized feature set. The first five principal components for each oil sample were clustered by fitting a Gaussian Mixture Model in the form

\[ p(x) = \sum_{k=1}^{K} \pi_k N(x | \mu_k, \Sigma_k) \]

where \( N(x | \mu_k, \Sigma_k) \) is the normal distribution over \( x \) with standard deviation \( \mu_k \) and covariance \( \Sigma_k \). The covariance matrix \( \Sigma \) is a full matrix for every component mixture component \( k \). The values for \( \pi_k, \mu_k \) and \( \Sigma_k \) are calculated iteratively using the expectation-maximization algorithm (Yu et al., 2011). Each sample \( x_i \) can then be assigned a class \( k \) based on which Gaussian component for which the likelihood of observing the data

\[ p_k(x_i) = \pi_k N(x_i | \mu_k, \Sigma_k) \]

is maximal.

Figure 1A shows the inferred clusters for the samples from the Norwegian sea based on three Gaussian components using concentrations of \( n \)-alkanes as features.
Figure 1A: Samples and inferred clusters for oil samples (n-alkanes) from the Norwegian Sea in the first two PCA components. Clusters are based on Gaussian Mixture Model on not two but five PCA components.

Figure 1B: Unsupervised clustering of oil samples in the Norwegian Sea based on Gaussian Mixture Models using n-alkane features on a geographic map. Distinct colours refer to distinct classes of oil samples.

The second approach used to infer oil types is based on supervised machine learning using the Random Forest algorithm (Ho, 1995). A subset of samples \( \{X_i, y_i\}, i = 1, \ldots, N \) was extracted where high confidence of the oil type could be inferred. The subset was further divided into a training set and a test set. The training set was used to train \( K \) shallow decision trees. An oil type classifier is made by computing the average prediction from all \( K \) decision trees, thereby using an ensemble of weak classifiers to produce one strong classifier.

The model trained on high-confidence oil samples was used to make predictions on the remaining oil samples, thereby providing oil typing based on both the data and a priori assumptions about a subset of the samples.

Results

In each of our two geographic areas (the Norwegian Sea and the Barents Sea) we classified samples both using geochemical signatures and GMM clustering. We compared the results geographically, as
shown in one realisation in Figure 1B. Furthermore, we used the biomarker-based classes to train a supervised model in part of the geographic area and test its accuracy on unseen samples.

Our results indicate that both supervised and unsupervised machine learning can differentiate between different petroleum provinces in a very consistent manner and may give additional insight into outlier samples. Unsupervised clustering of oil samples produces classes that are, overall, consistent with geological understanding of the origin of the oils. In some cases the subclasses identified by unsupervised clustering are different from those identified using the traditional approach, potentially indicating a more natural classification given the available data.

We demonstrate a workflow using supervised learning that enables an expert to infer oil types and source rock correlation in an immature (low confidence) area given available samples in a mature (high confidence) area.

Finally, we discuss limitations and challenges with data driven approaches to oil-oil and oil-source rock correlation. We emphasize through examples the importance of data quality and data consistency when using machine learning for automatic oil typing.

**Acknowledgement**

Most fluids in this study were analysed by Applied Petroleum Technology A.S. (Oslo), and we acknowledge their insight and willingness to discuss analytical uncertainties and problems.

**References**


