Using the IGI ML tools to classify oils in the South Viking Graben

IGI TECHNICAL NOTE

Dan Cornford[1*], Marianne Nuzzo[1], Aimee Whatling[1], Jesper Kresten Neilsen[2]

[1] IGI Ltd, The Clock House, The Old Stables Business Park, Bideford, Devon, EX39 3QW, UK
[2] OKEA ASA, Oslo, Norway
** Corresponding author: dan@igiltd.com*

27 May 2025

This technical note outlines a workflow using the IGI Machine Learning (ML) tools to reproduce the oil-family framework reported by Holger Justwan, Birger Dahl and Gary Isaksen in their 2006 study of oils and condensates from the Norwegian South Viking Graben. This work was undertaken with Jesper Kresten Nielsen from OKEA as part of a mentoring week, in which we explored several elements of the IGI ML tools associated with the p:IGI+ software while building an oil family and PVT interpretation in two specific regions of interest to the company (not discussed here).

Introduction

In the Justwan et al. (2006) paper, the authors identified seven hydrocarbon families from molecular and isotopic data and linked them to three principal source horizons, demonstrating that multivariate geochemical patterns can be used to distinguish both endmember and mixed charges.

The investigation proceeded in a number of steps:

1. The IGI data team swiftly imported the data from the original paper into p:IGI+
2. We explored the loaded data in p:IGI+ reproducing the Justwan et al (2006) results – you can obtain the p:IGI+ v3.0 project and an artefact collection built to explore the data by emailing us at info@igiltd.com.
3. We built a ML classifier to learn the Justwan et al. (2006) oil families. You can also obtain the model via email.
4. We used the IGI Metis Global Data Service to download all publicly available oils and condensates from the South Viking Graben region, including those on the UK side.
5. We ran the ML classifier on these oils and condensates to explore the generalisation of the Justwan et al. (2006) groupings.

In this technical note we describe the steps and highlight the challenges and thinking behind the decisions we made.

Ingesting the Justwan et al. (2006) data

This step should be simple, but even with an expert data team, the subtleties in the naming of geochemical ratios meant care was necessary. Both reporting units and whether the ratios were normalised (a/(a+b)) or

direct (a/b, b/a) had to be considered. As we did not have access to the raw measurements, getting the detail correct was essential to be able to apply the learnt model elsewhere.

Reproducing the Justwan et al. (2006) results

Again, in theory this should be relatively straightforward, however details of which samples were used to produce the published statistical model, and which were excluded, was challenging to reproduce. Also, some of the ratios used on the plots were not reported in the data tables, so without the direct measurement data we were limited in what we could do. Figure 1 shows a reproduction of Figure 11 from Justwan et al. (2006), although we have had to substitute some properties shown in the original paper with the closest equivalents calculated with data provided in the paper.

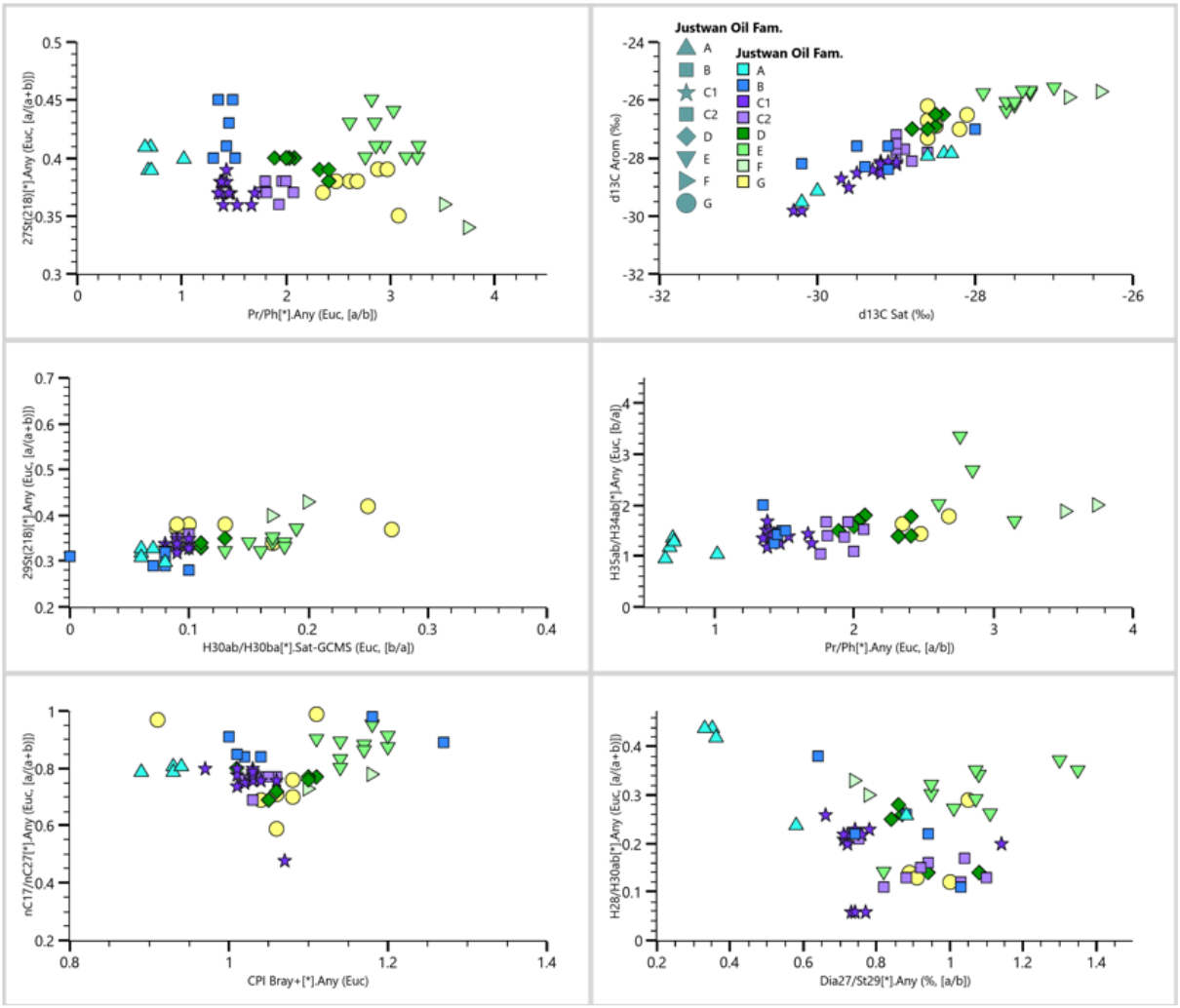

Figure 1. A reproduction of Figure 11 from the original Justwan et al. (2006) paper.

We were able to reproduce their dimension reduction (PCA) results and explore cluster models using the IGI ML tools. These were very similar, but not identical, to the results presented in the paper. We attribute this to potentially using slightly different learning sets (we do not know exactly the data used in training their model) – this also highlights the sensitivity of hierarchical cluster models to the exact data used in training them.

Training a ML classifier to learn the oil families

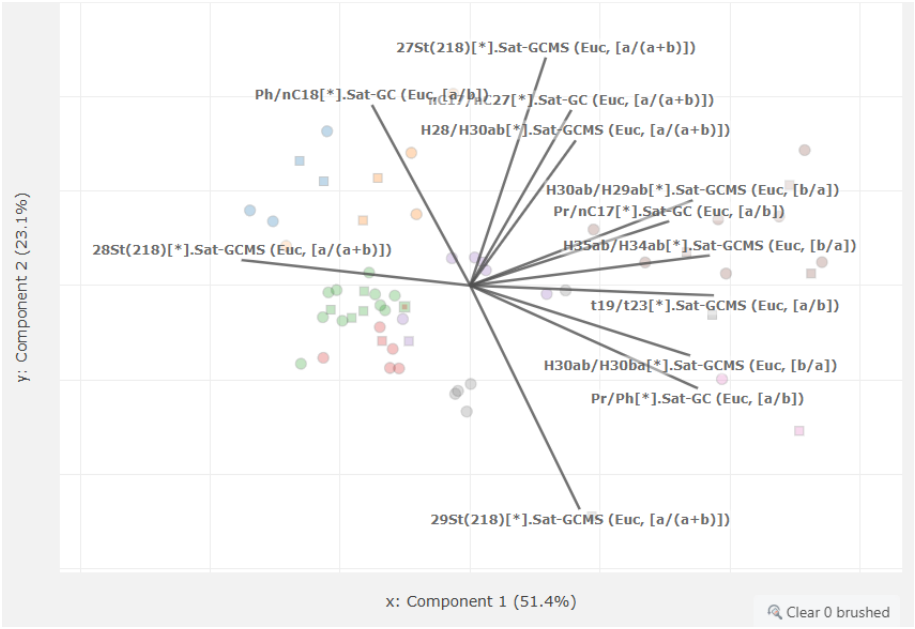

Figure 2. The classifier, showing the target points projected into the first two principal components (feint colours) with the component loadings shown over the top.

Using the data from the paper, we could train a classifier to reproduce their labelling. Since a classifier could be applied beyond the training data, we could then classify other oils in the region using their labels. To make sure the classifier was as general as possible, we only used the ratios from the paper that are commonly reported and that they identified as useful. We explored several model choices. We were especially keen to ensure that the model generalised well, so the degree of smoothing was always kept relatively high.

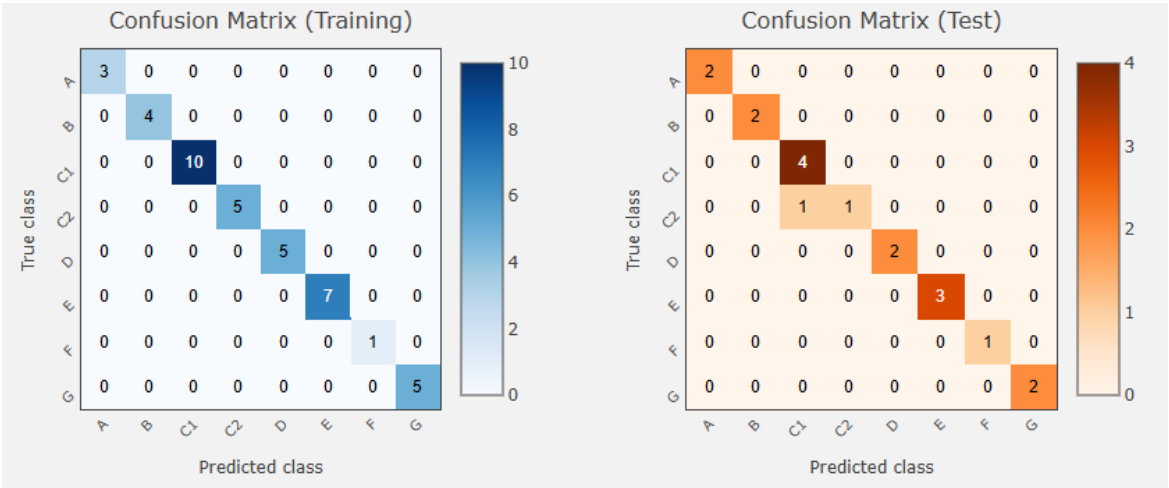

Figure 3. The confusion matrix for the training and test data for the neural network model selected.

Within the Justwan et al. (2006) data there were no outliers. We used a 70:30% training test split and explored many model options. Our preferred model, which seemed to generalise well, used four principal components (of the 12 input ratios, see Figure 2), with standardised inputs and k-nearest neighbour imputer to replace missing values in samples that had at most four missing values. The model itself was a neural network model (multi-layer perceptron) with 8 hidden 'relu' (piecewise linear) units and a smooth strength of 0.4. The performance is illustrated in Figure 3, which shows the confusion matrices for the test and

training sets. The single misclassification on the test set is for a sample that could be in families C1 or C2 and, from visual inspection, could readily be misattributed by a human.

The model was saved from the machine learning tools into the p:IGI+ project. It is also available as a template, which can be loaded into the p:IGI+ model manager and used in your projects (email us at info@igiltd.com). Although the advanced neural network models can only be trained by users subscribing to the IGI ML tools, they can be applied by any user with access to p:IGI+ 3.0 and up. The built-in AI assistant can also help support model decision making outside of mentoring projects, as it has been created using IGI's machine learning expertise.

Including all oils from the region

Having trained the ML classifier, the next step was to download additional oil and condensate data for the region. We used our Metis system to query for all oil and condensate samples in the South Viking Graben region, from both the Norwegian and UK sectors.

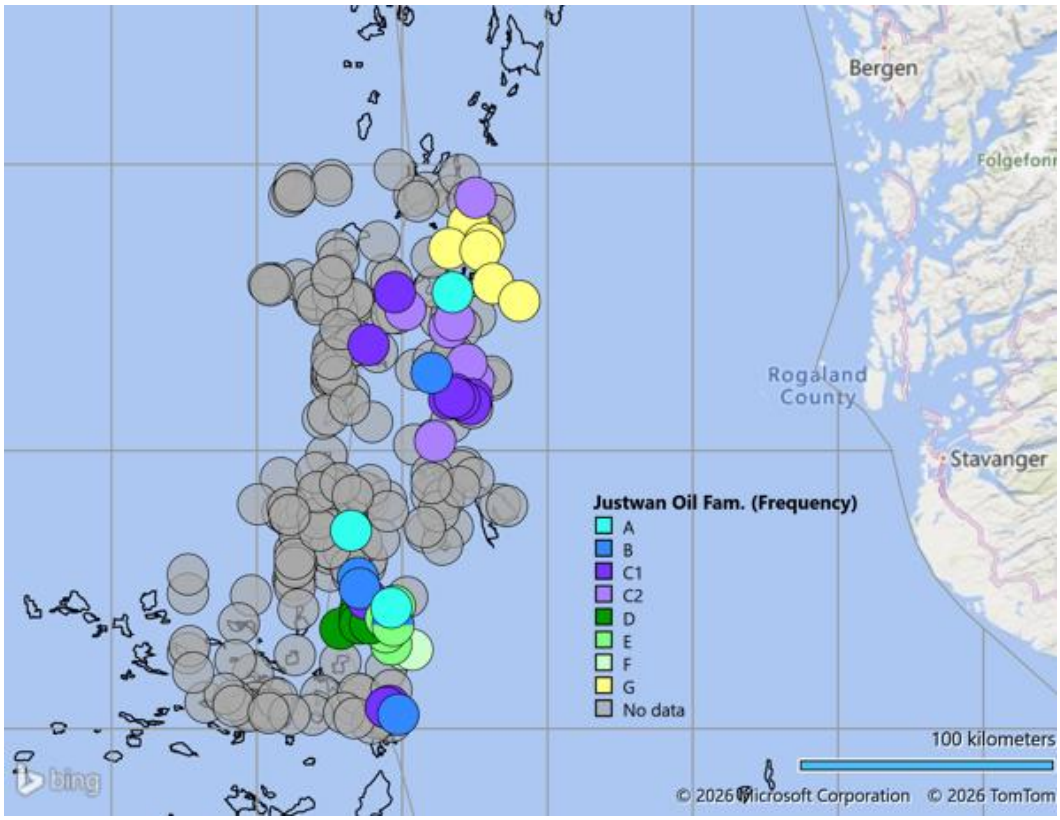

Figure 4. The original Justwan et al. (2006) oil families for the samples used to learn their model, and an overview of the location of all other oil / condensate samples from the region (light grey).

Predicting the families of new oils

To use the ML model for predictions involving the new dataset, we first wanted to ensure we used samples with sufficient data for reliable prediction. We used the "Finding good data set" option from the p:IGI+ Model menu.

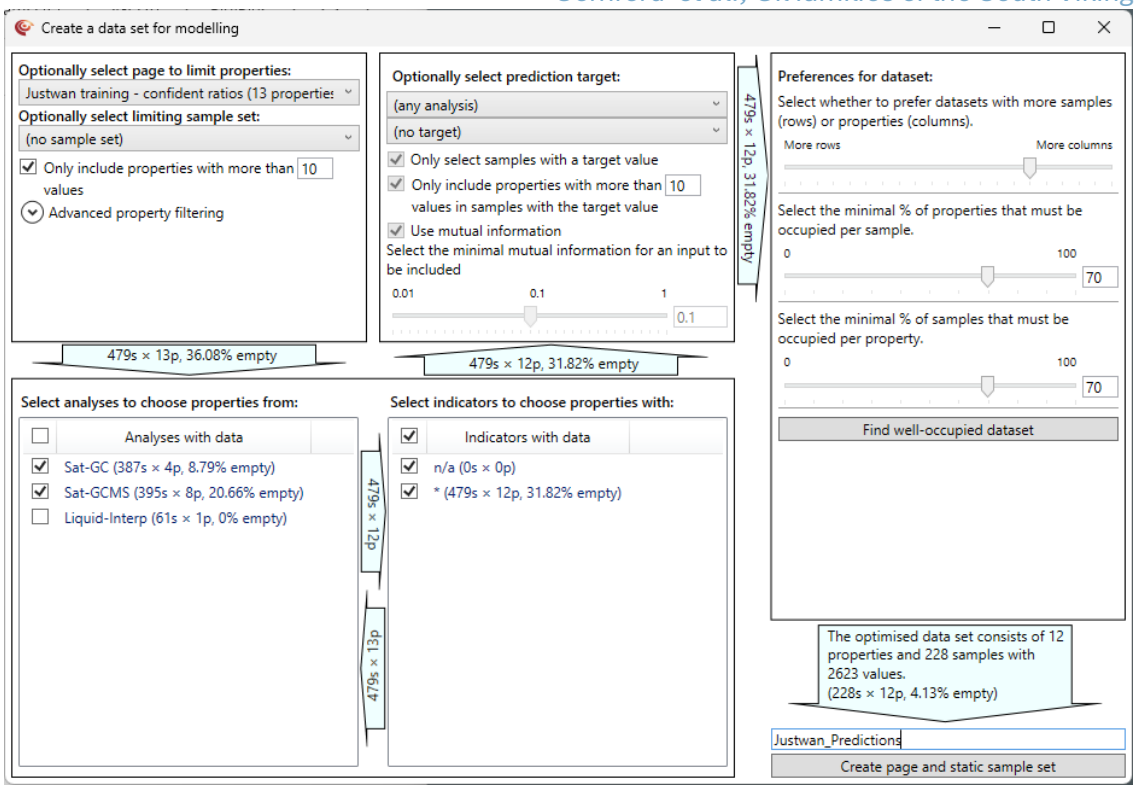

Figure 5. Illustrating the use of the tool to find a suitable data set in p:IGI+.

Figure 5 shows the options used in the tool – basically, we used the page with the relevant input properties (in the Sat-GC and Sat-GCMS analyses) to select a prediction data set that includes all 12 properties used to build the classifier (Figure 2).

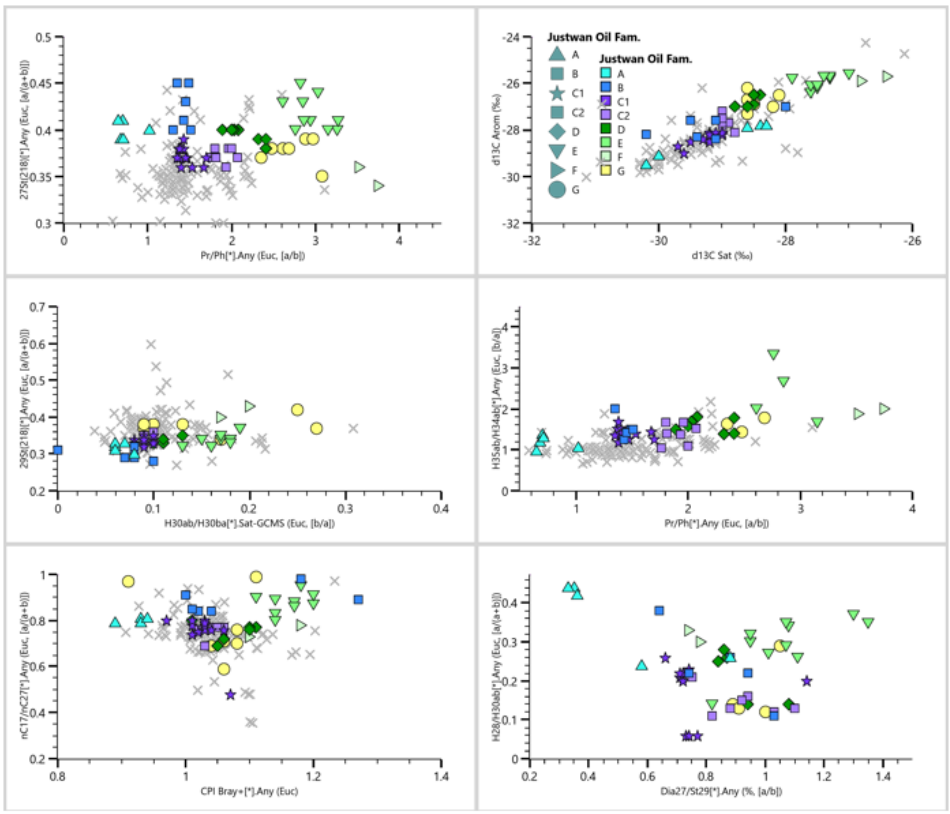

Figure 6. Illustration of the Justwan et al. plots with data from the original paper (colours) and the new data (grey crosses) from the region. Note no new data is shown for the lower right plot as we did not precisely know the ratio used.

However, only samples with at least 70% of their values populated were considered. This decreased the number of samples from 479 samples with at least one property with a value and 32% missing values to only 228 samples with only 4% missing values. The tool has a range of uses and is included in version 3.0 of p:IGI+ for all users. A preview of the new data from the region is shown by the grey crosses on Figure 6. For most properties, this lies in similar regions of input space to the original Justwan et al. (2006) data.

A more complete view of the data can be seen using the ML tool dot plots and brushing the original Justwan et al. (2006) samples to highlight them. Figure 7 shows that, for the most part, the other oils in the region are similar to the samples used the train the model. There are, however, some examples that are significantly different.

Figure 7. Distribution dots plots from the ML tool univariate outlier detection stage, with the brushed Justwan et al. (2006) points shown in pink, and the new oil data for the region in green.

Using the sample set created in the data set finding tool, we predicted the Justwan et al. classes on the new data using the ML model manager in p:IGI+.

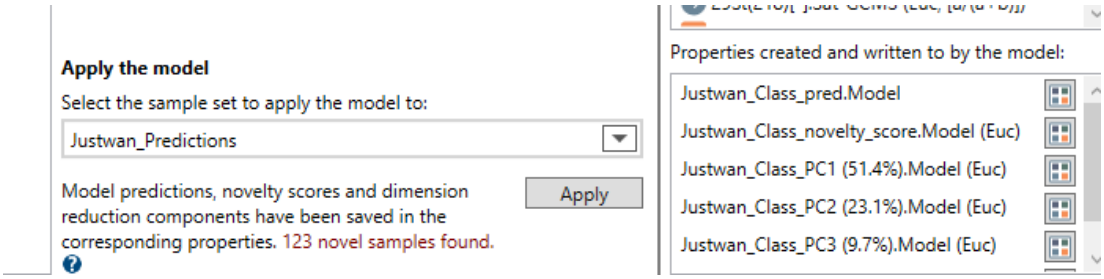

Figure 8. Example of applying a machine learning model to new data, with warning about novel samples.

Figure 8 shows the results of applying the trained model to the new data. This highlights an important feature of the IGI ML tools – the novelty scores. Applying a model trained on local data, or even on a more regional data set, will always carry a risk that the model was trained on data that is different from the data we apply it to. When we train a model in p:IGI+, we also train a Gaussian mixture model to learn the underlying distribution of the training set data (using a Bayesian approach). This allows us to judge whether a new data point is similar to the training data, or different. We encapsulate that information in a 'novelty score' for each sample.

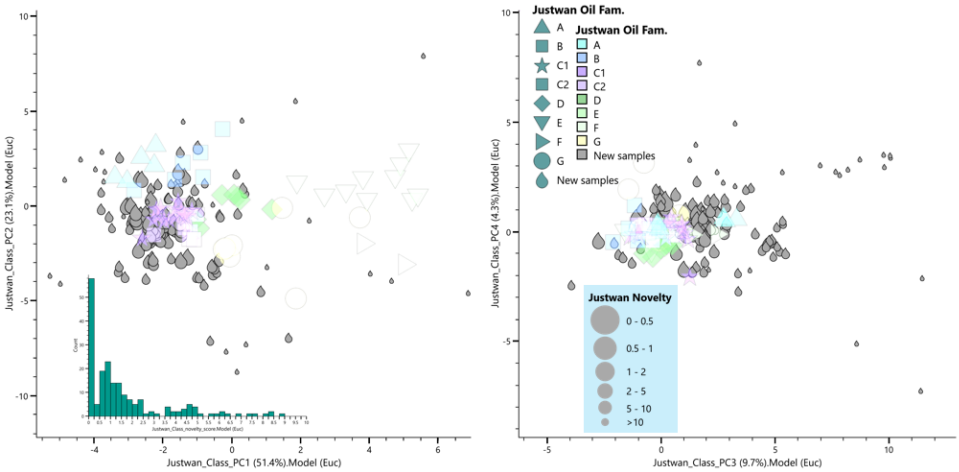

Figure 9. Results of projecting the new oils onto the Justwan et al. PC's, showing the original samples (high transparency) and the distribution of the new oils and their novelty scores. Higher novelty scores have smaller symbols size.

The novelty score is a positive numerical that reflects the probability of the new data under the distribution of the training data. The scaling of this score is arbitrary to some degree. A low score, especially less than one, represents data that is very similar to the learning set. A high score over 10 suggests the data could be quite different (when considering all variables together). Scores can get very high especially when using a large number of inputs, suggesting the model results should not be trusted for these samples.

Figure 9 shows the projection of the new oils on the principal component scores learnt from the original Justwan et al. (2006) samples. The plots show the learning sets (although with very high transparency to allow focus on the new data points). The new points are shown in grey and scaled inversely according to their novelty scores (low score = large size, trustworthy results). The inset histogram shows the distribution of the novelty scores across all data in the range 0-10. There are several points with higher novelty scores, some over 1000, suggesting these are very different from the training data.

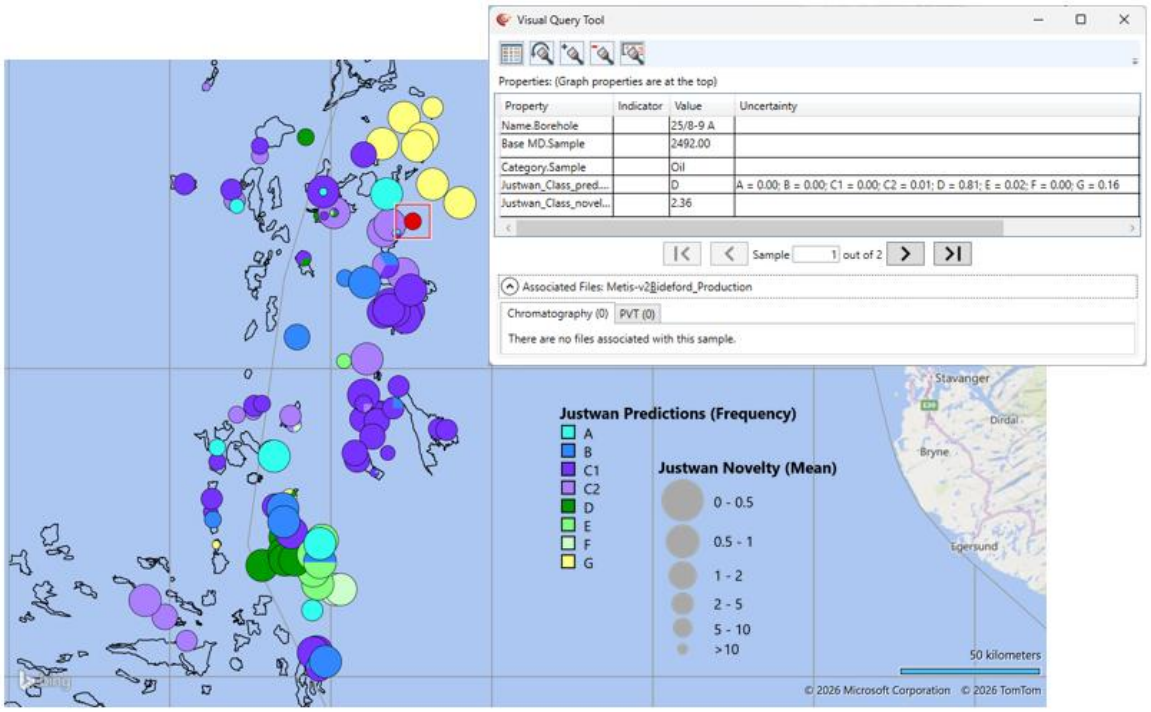

Figure 10. Map showing the classified oils in the South Viking Graben including the original and new oils, with visual query on highlighted sample.

Figure 10 illustrates the results spatially. We have classified the oils based on the neural network model and used the novelty scores to define the size of the points, with small points less trustworthy. The visual query tool illustrates the predictions for a new sample (red datapoint). Note that, while we predict the most likely class for each sample, the uncertainty retains the class probabilities based on the trained model. For instance, in the visual query tool, it indicates an 81% chance of the sample being associated with class D, but a non-trivial 16% probability that it could be class G, which might make more sense spatially. The novelty score of 2.36 is still within the acceptable threshold for this data.

Summary

We have shown how we can use an existing labelling of samples, in this case from an oft-cited paper, to learn a regional model. We've shown how the tools developed in p:IGI+ and the associated ML functionality can be applied effectively to the problem, highlighting some nice features along the way. As we used entirely public data, we have made the project and associated artefacts freely available. The application shows the importance of careful visualisation and validation of your learning set and model, and of taking care when applying the model. The novelty scores are always calculated for any model and should be checked to ensure you are not applying the model outside a sensible range – the model will always predict something, but it might not be very logical or useful.

With the release of p:IGI+ 3.0 the ML tools are available to all, but to realise the advanced features users will need to add the ML tool option. The tools include clustering, classification (this article), regression, spatial interpolation and unmixing and have comprehensive pre-processing options including outlier detection, dimension reduction and novelty scoring.

References

Justwan, H., Dahl, B., & Isaksen, G. H. (2006). Geochemical characterisation and genetic origin of oils and condensates in the South Viking Graben, Norway. *Marine and Petroleum Geology*, *23*(2), 213-239.