Using machine learning to reduce ensembles of geological models for oil and gas exploration

Oliver Thomson Brown[1] and Anna Roubíčková[2]

EPCC, The University of Edinburgh

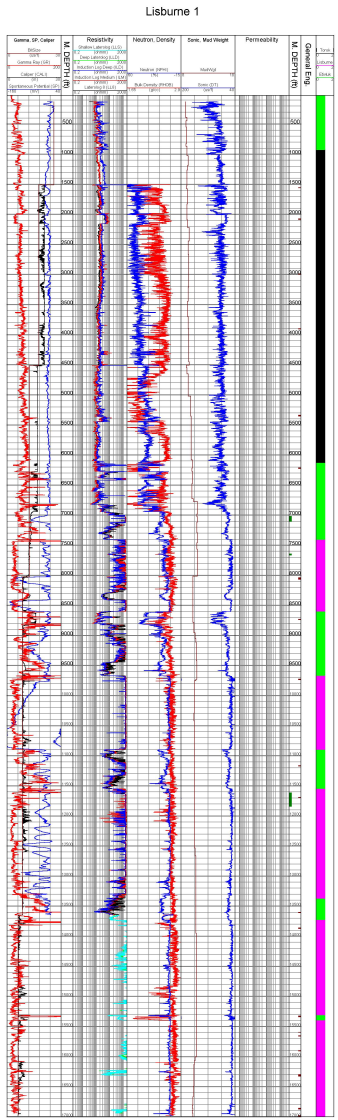
The 5th International Workshop on Data Analysis and Reduction for Big Scientific Data at SC19

[1] o.brown@epcc.ed.ac.uk
[2] a.roubickova@epcc.ed.ac.uk

Background

Background

- Have an oilfield.
- Where should we place wells?
  - Drill a bunch of boreholes.
  - Get a bunch of well logs.



Background

- Well logs are very detailed...
  - $\sim$ 20000 rows of measurements.
- ...but sparsely distributed across the oilfield.
  - Drilling a borehole and putting detectors down it costs time and money.
- So the challenge is to predict the subsurface structure of the entire oilfield from these sparse borehole measurements.
- Specifically of course, we want to know where the richest oil wells are!

Properties

- Principle target property, OIP.
  - The total oil content of an oil reservoir.
  - Cannot be measured directly, has to be estimated.
  - We will use just three properties. This is greatly simplified, but helps keep the number of models manageable!
- Three properties derived from well logs used as estimators.
  - Porosity.
    - Ratio of pore volume to volume of rock.
  - Net to Gross.
    - A slightly more complicated metric, but roughly the ratio of volume of rock that can store hydrocarbons, to volume of rock.
  - Saturation.
    - The fraction of effective porosity which is filled with a specific fluid (like oil!).

Trends

- Petroleum industry concerned with four ‘trends’ of these properties.
  - Depth.
  - Stratigraphy.
  - Strike.
  - Dip.
- How a particular property evolves with a trend defines a function, represented by its knot points.
  - Three knot points for Depth, Dip, Strike, 35 for Stratigraphy.
- We can concatenate the description of these trends for a particular property to a single 44 knot vector but, order matters.

Genes

- We call the concatenated knot vector for a particular property a gene.
- A gene with Depth-Dip-Strike-Stratigraphy, is different from one with Stratigraphy-Strike-Dip-Depth.
- To define geological model of an oilfield, identify all possible genes which are in line with well logs.
- We will be less sophisticated, and just consider all possible genes. (For now!)
  - ${}^4P_4 = 24$

Genome

- As noted before we are considering just three properties of our oilfield.
- A sequence of the three properties' genes defines a geological model for the oilfield, and we will call it a genome.
  - Defined by 132-element vector.
- For each property we have 24 equally valid explanations (genes).
- In total then we generate 24^3 valid geological models.
  - Quick reminder that we are looking at a reduced number of properties, so exponent would typically be much larger.
- Each genome can be uniquely identified either by a triplet of numbers identifying its genes, or by an identification number generated treating the triplet as a three-digit base-24 number, and converting it to base 10.

Recap

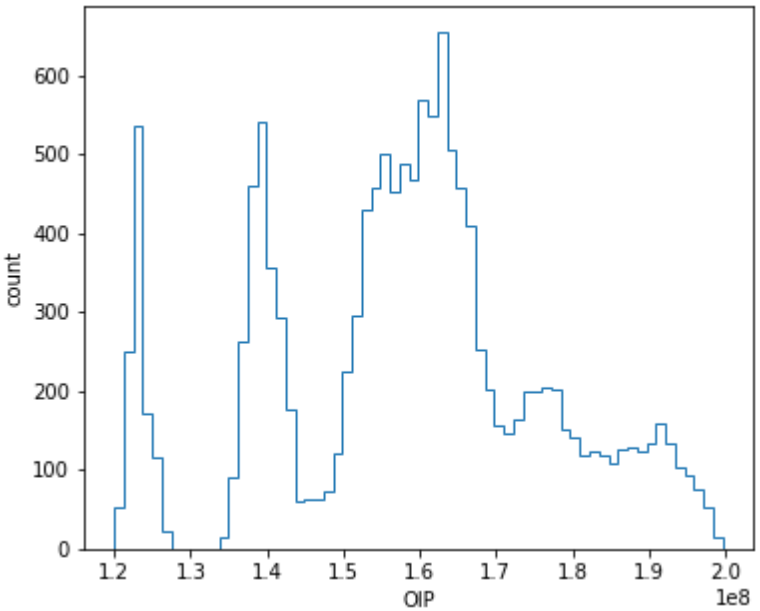
- That's a lot of background!
- Well logs $\rightarrow$ Derived properties $\rightarrow$ Trends $\rightarrow$ Genes $\rightarrow$ Genomes.
  - Side note: Our collaborators at Cognitive Geology have developed software to automate this process of going from well logs to an ensemble of models.
- Some data reduction has taken place, as we consider trends of derived properties of the well logs, but nothing interesting...

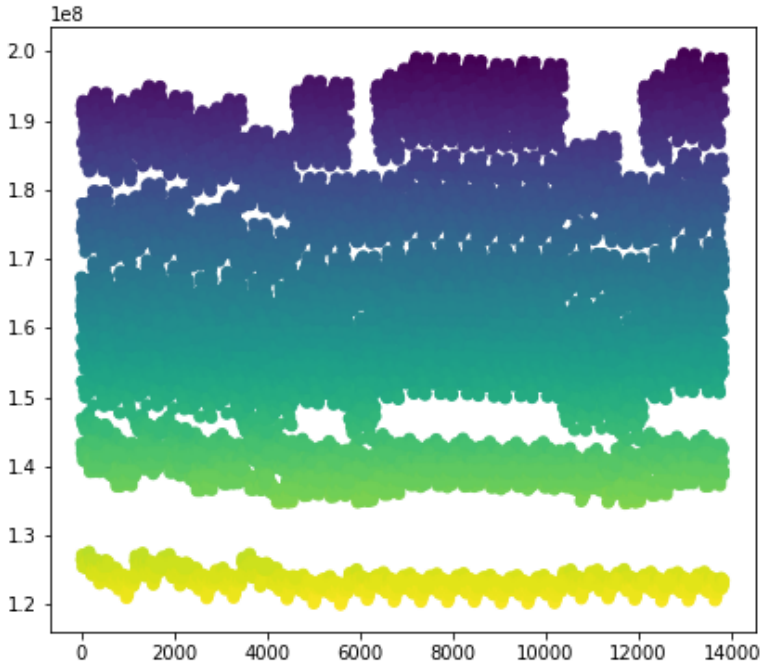
Ensemble reduction

- An ensemble of 24^n models for n properties is infeasible to evaluate.
- Ideally we would like to group together all models which give broadly the same result, and only consider one model from each group.
- Two questions:
  - What is the result we are interested in?
  - How similar is 'broadly the same'?
- Answer to the first question is of course OIP.
- Unfortunately, determining OIP requires an evaluation of the model which we have to do for every model...

Ensemble reduction

- In an ideal world where we could quickly and easily cluster models based on OIP...





Ensemble reduction

- We considered clustering without calculating OIP, using Euclidean distance as the similarity metric. This did not work...

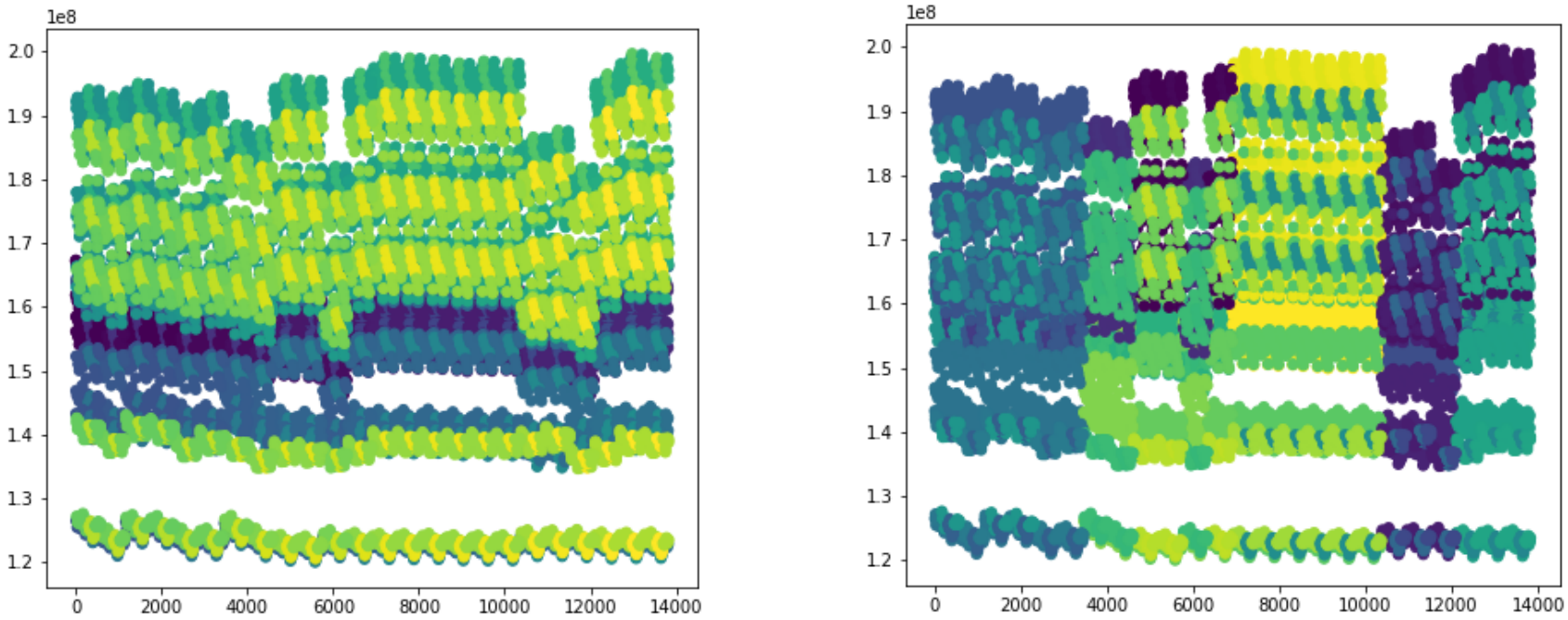


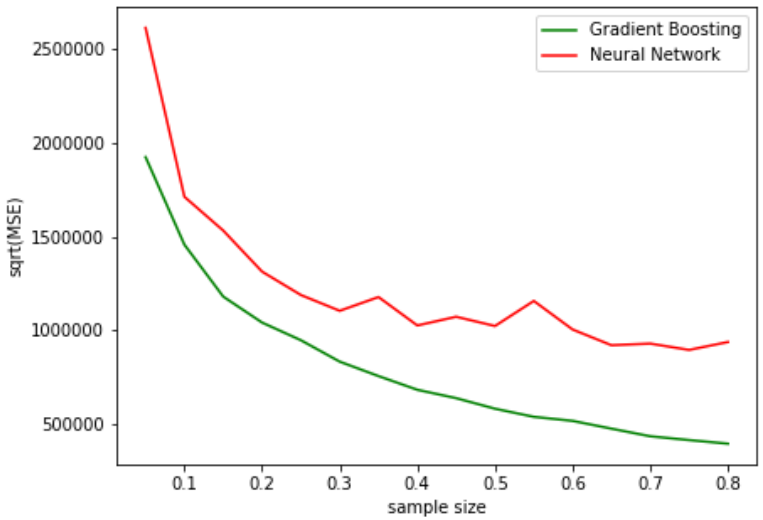
Figure: On the left, density based clustering was used, on the right, self-organising feature maps.

OIP Estimation

- Using OIP as the similarity metric is a necessary evil then.
- ...or a least a related metric.
- So why not train a regression model to estimate the OIP?
  - This will result in a far less computationally intensive metric for reducing the ensemble.
- We will try two off-the-shelf approaches, an artificial Neural Network (NN), and a Gradient Boosted regressor (GB), both implemented in SciKit-learn.

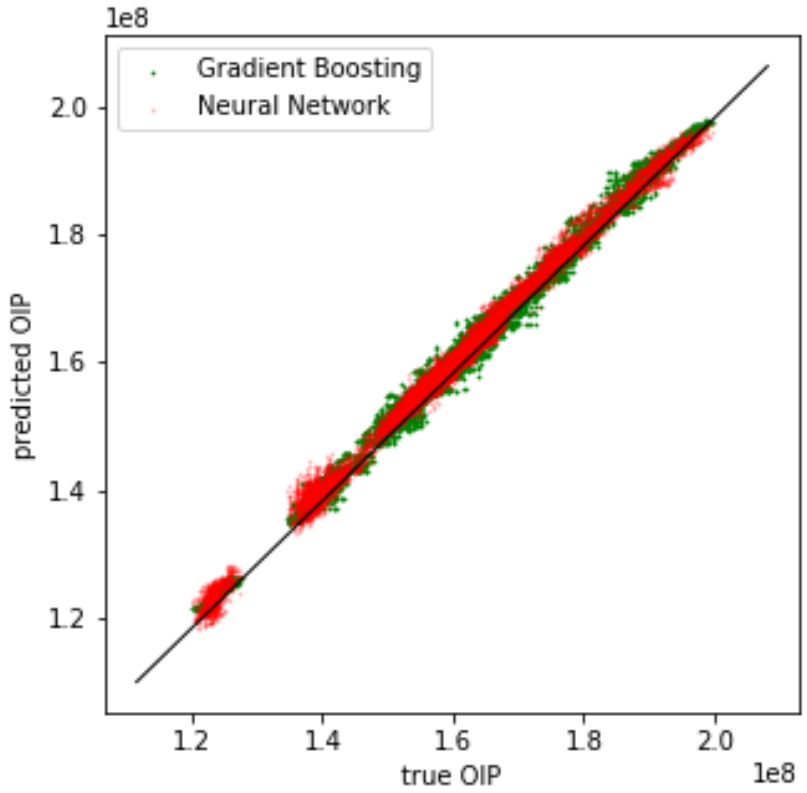
OIP Estimation

- Using 80% of the models for training and 20% for testing is computationally infeasible for real-world problems, but it does show the approach works.
  - Predictions were on average within 1% of the actual OIP value.
- We then experimented with reducing the training set size.



OIP Estimation

- Using 15% of the set for training gives an acceptable error.



Ensemble reduction revisited

- Let's try this again with our two-step approach.
  - Step one: Determine OIP estimator using gradient boosted regressor.
  - Step two: Cluster models with self-organising feature maps, using OIP estimator as the metric.
- Self-organising Feature Maps (SOFM/SOM).
  - A specialised form of artificial neural network.
  - Assumes a two-dimensional grid of neurons in the hidden layer.
  - Makes use of competitive learning, in which individual neurons 'compete' to respond to inputs.
  - Nodes are updated according to their Euclidean distance to the winning node.
  - Our models will be clustered according to which node they are mapped to.

Ensemble reduction revisited

- Does it work? Yes!

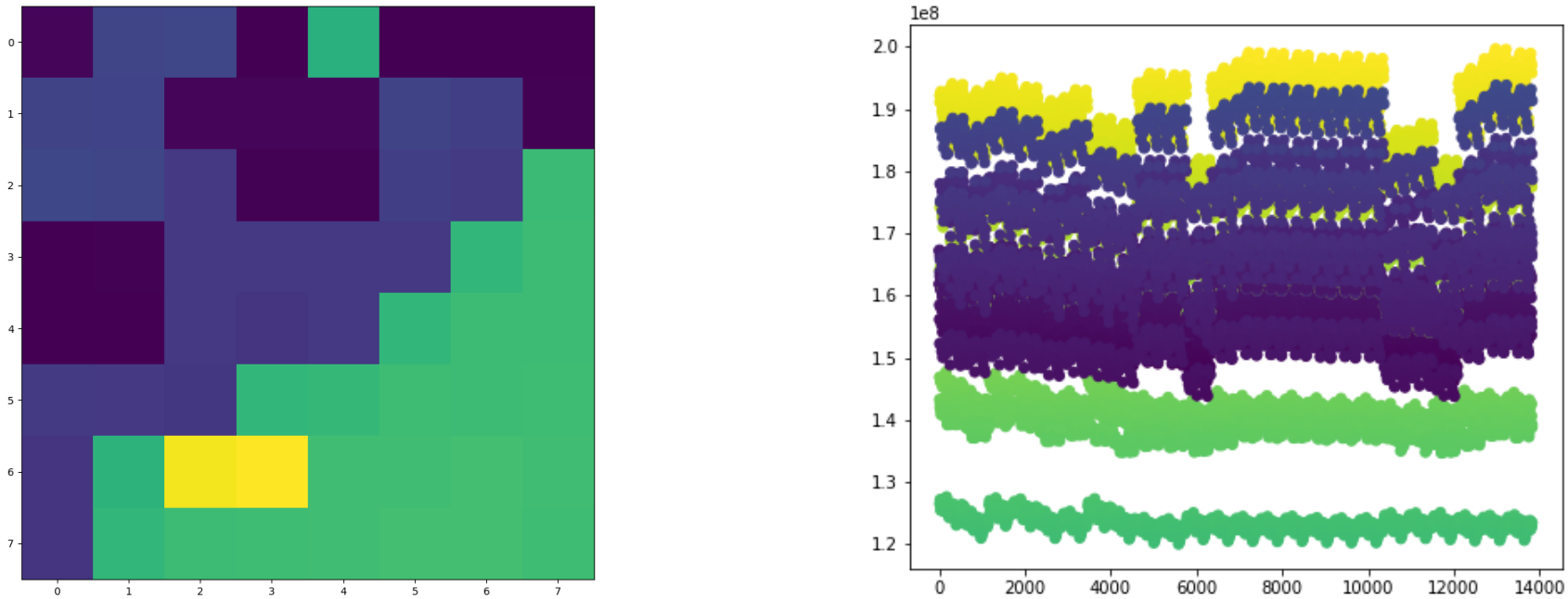


Figure: On the left our SOFM, and on the right, OIP, colour-coded by node/cluster.

Data Reduction

- Can choose one model from each cluster to form a representative model of the whole oilfield for further evaluation.
- Total reduction in ensemble size from $24^3 \rightarrow 64$.
  - How this scales with 24^n is an open question.
- Two step approach: using supervised learning to determine an estimator, and unsupervised learning to cluster models using that estimator.
- Only 15% (~ 2000) of the original ensemble required to train estimator model.

Thanks to...

...the people who actually did the work on this project but unfortunately couldn't be here:

- Anna Roubíčková
- Nick Brown (currently running the HPC for Urgent Deicision Making workshop)

and to our collaborators at Cognitive Geology:

- Lucy MacGregor
- Mike Stewart