Skoltech
Skolkovo Institute of Science and Technology

Thesis Changes Log

Name of Candidate: Viktor Duplyakov

PhD Program: Petroleum Engineering

Title of Thesis: Machine learning on field data for hydraulic fracturing design optimization

Supervisor: Professor Andrei Osiptsov
Co-supervisor: Professor Evgeny Burnaev

The thesis document includes the following changes in answer to the external review process.

Egor Dontsov

1. It is quite apparent that most of the effort is spent on data mining, which is a crucial step to success. However, from reader's perspective, the emphasis should be on the results. I found that the first part of the thesis is written better that the second one that deals with the results. It is harder to follow and the key points are not well emphasized. One reason, perhaps, is that figure captions are too cryptic. In other words, it is practically impossible to understand what is plotted on the figures by just reading captions. The reader needs to go back and forth to the text in order to have a full picture. To sum up, the first major comment is to make chapter 4 more clear and to better highlight the main results that are relevant to field applications.
Figures' captions are rewritten, some figures rearranged, a few remarks introduced. Chapter 4 has been restructured a bit.

2. The second issue, which is partly discussed in future work, is the coupling with economics. I think that there should be at least a qualitative discussion on how economics affects the result. Because otherwise, the answer is the bigger the better. The more fracs are out there and the bigger the fracs are, the more oil they are going to drain. There are no constraints. And economics provides these constrains. Here is an example how this is done in ResFrac: https://www.resfrac.com/blog/resfracsautomated-economic-optimization-tool. Once you add economics, then typically there is a clear maximum or optimal value.
Subsection 5.3.3 "Optimal target selection" is written, dealing with the question with economics criteria in target

3. List of your publications and conference proceedings: use the same format for all entrees.
Corrected

4. "Pipeline presented in this study" – consider rewording as "The approach presented in this study".
Rephrased

5. "The activation of the natural fractures network by hydraulic fracturing is a key issue in the commercial production of shale reservoirs." - this was a line of thought a few years ago and it was based mostly on microseismic results. This applies to US at least. Right now, there is a lot of evidence from fiber optic measurements in the offset wells that fractures are predominantly planar in shales. That's why nowadays people rarely discuss stimulation of natural fractures.
Rephrased

6. "special dimensionality reduction" – did you mean "spatial"?
Corrected

7. “yields an maximum” – typo.
Corrected

8. “good“ – check all quotation marks for correct formatting.
Corrected

9. “euclidean distance” – I think “e” should be capitalized, at least in english version. Fix throughout the whole thesis.
Fixed

10. Figure 4-2. Update caption so that readers can understand what is actually plotted. Is it prediction of production for various parameters for two models? Please check all figure captions to make sure that it is possible to understand what is plotted by just reading the caption.
Captions updated

11. Regarding the field test described in 4.2.5. How different were the parameters from the training set? Can they be plotted somehow on the parametric diagram?
The wells originated from the identical oilfield as the training data, functioning within the same geological layers. Consequently, all key parameter values fell within the distribution observed in the training dataset.

12. Make sure that all references at the end of the dissertation are formatted to the same style. Also fix “booktitle=SPE Oil others”.
Corrected

Clément Fortin

1. Some figures are hard to read such as Fig. 2.1 and 2.4. In general, the font sizes in figures are too small.
Figures corrected

2. Some terms like “pad share” and “pad volume” are never properly defined and the document need to be checked carefully to improve its precision and clarity. Many acronyms are not properly defined either such as the t-SNE method which is used extensively. It is required to explain its importance and value for the research work. The same comment applies to the acronym “NaNs”. The author must be more careful to make sure readers can follow his logic and understand the real value of his contributions to the field.
Acronyms definitions were introduced with their first occurrence in the text

3. In 5.1, first phrase : “The resulting gathered during the study database…” needs to be modified.
Corrected

Dmitry Garagash

1. Thesis researches two sets of problems with regard to design of ML approach to reservoir forecasting: a) what constitutes the optimum set of data (for a given well) and number of wells in given groups in the database for the method traning purposes; b) what constitutes a reliable ML methodology and corresponding optimal set of parameters to design a successful HF job. The preferred trained ML model in this thesis, when tested on hold-out sample of the database, yielded R^2=0.64. Is this a good result for an ML forward model?

Does “heterogeneity” refers to fundamentally different reservoir and well/fracture classes? (for example: vertical, multi-layer well vs horizontal presumably single layer contained well. Or different type of reservoir: conventional vs unconventional?)
Would an ML-method, given sufficiently representative training database, is expected to learn the structure of heterogeneity and adapt and recognize it accordingly, allowing for accurate forward ML forecasting fore distinct input setups?
The fact that the accuracy is “not ideal” - does this mean that the database perhaps is still under-represented for some subclasses of input conditions?
Was there an attempt to select a more ‘homogeneous’ sub-set of the database and provide forward ML forecasting for this subset - perhaps in some ways similar to previous studies which looked at much smaller data samples? If so, would that allow for a more accurate ML prediction? And if so, should one then suggest to have individually trained ML forecasters for defined subclasses of data ? In other words, it is not clear to this reviewer - is there an advantage in ML trained on very large but heterogeneous dataset vs. a set of MLs trained for subsets of the data?
Added considerations on this matter in 5.2

2. What is the distinction between ‘hold-out’ and ‘field test’ wells? From the description, it follows that 21 field test wells belong to the same field (and ML trained on the subset of the database is used) - is the correct?
Explained in 4.1
While hold-out wells are a wider set (which would include the ‘field test’ ones) from the full-database. The fact that similar prediction error is observed for sub-set (field test) and full population of hold-out wells - would that be an indication that algorithm works equally well (or not) irrespective of how heterogeneous are the field conditions. And the prediction error is rather dominated by poor prediction/modeling of multi-lateral vertical wells (again irrespective of a field).
Added in 5.2