Statistical learning and causal inference for oil and gas production analysis and forecasting

A. Bertoncello J. Garnier M. Asch

April 17, 2018

Context

The forecast of oil and gas production decline, at well and field levels, is an issue of major importance. Traditionally this is done by exponential decline curve fitting. More recently a *random forest* based statistical learning approach was used to analyze well production. The results, though promising, did not give entire satisfaction. Uncertainty was not accurately quantified and in particular the extreme forecasts were badly estimated–when compared to field data they were underestimated. In addition, influent variables were incorrectly identified.

Production modeling depends on a large number of parameters and variables. Standard statistical inference approaches, and even machine learning (as seen above), cannot distinguish between association (correlation) and cause. Thus, in a multivariate context, it is very difficult to make predictions that are reliable and whose uncertainties can be quantified. The optimal production strategy for exploiting wells and fields could have extensive *economic* and *environmental* benefits. Today, companies rely on a trial-and-error approach to find an optimal completion practice, modifying one or two design characteristics at a time and comparing production with previously drilled and competitors' wells. This approach is inefficient because it does not account for other impacting effects. Competitors' wells may be drilled in slightly different geology. Older wells may have been drilled with different strategies. Service companies may vary between operators, etc.

However, large quantities of *field data* are available (more than 700 wells in one field) that include production data time-series, exploitation conditions and the associated geological data. This thesis will thus develop a data-driven approach, coupling *causality* with *machine learning*, to study causal inference for production analysis and forecasting.

Approach

Initially, the thesis will thoroughly explore the available field data and perform *unsupervised learning* in order to identify clusters of influential variables that

should be related to the problem physics, based on correlation analysis, principal component analysis (PCA) and clustering methods [1]. Based on this exploration, predictive, *supervised learning* models will be investigated using for example GLM (Generalized Linear Models) or GAM (Generalized Additive Models) techniques. The predictive capacities of these models will be compared against test data and compared with the random forest results of the earlier study.

This initial study will lay the basis for an in depth study of *causality* [4, 5]. A causal model will be specified in two parts: a statistical model, and a causal graph that describes the causal relations between the variables. Causal models belong to two families: (1) Bayesian causal networks, (2) structural equation models (SEM). Causal inference approaches are seldom used in industry, even though such an approach could be of tremendous value. A causal graph enables both to *predict* the effects of manipulating some of the variables, or to make *backward inferences* from effects to causes. The ultimate objective of the thesis is to construct a causal workflow that will lead to optimal production strategies including complete uncertainty quantification [2, 3].

Work program

Years 1 and 2. The first part of the thesis will be dedicated to familiarization with the physical context and exploratory data analysis. The second stage will explore unsupervised followed by supervised learning approaches for well and field modelling and produce verifiable forecasts with associated uncertainty quantification. The objective is to predict the field's production at a given time horizon, and further, based on already producing wells whose decline can be estimated. This will provide an estimate based on the future *potential* of new wells and of neighbouring wells. For this task, the quantification of uncertainty is a critical aspect.

Years 2 and 3. The third, and most important stage, will be dedicated to a causal study and the development of an end-to-end approach that will provide a decision making tool. We will investigate how the previous prediction models can be used as input to the causal analysis. In particular, we will identify which variables, or clusters of variables, provide the most informative and physically relevant forecasting power. Initially, two-by-two causality analysis will be performed (see justification in [3]). This will then be extended, if possible to the elaboration of causal graphs, and hopefully lead to development of "analogues" that can be used to transfer knowledge from known instances, to unknown ones.

The overall workflow should be split into two major branches: the predictive part, and the causal part. Each part of the model (predictive or causal) should be dealt with separately for both estimation and inference.

Requirements

We are looking for a candidate (holding an engineering or masters degree) with a strong background in Probability and Statistics. Experience with statistical simulation codes, such as R or Python environments (e.g. scikit-learn), is desirable, as is basic knowledge and experience with machine learning algorithms. Fluency in English is required and the willingness to travel.

Conditions

The thesis will be based on a collaboration industry-university between Total SA and Ecole Polytechnique. The industrial director is Dr. Antoine Bertoncello, researcher in geosciences at Total-Pau, the academic director is Prof. Josselin Garnier, CMAP, Ecole Polytechnique-Palaiseau.

The thesis work will begin in Pau, and then can continue in Paris/Saclay or in Pau. A research visit at Stanford University will be programmed during the thesis. The remuneration is of the order of 30 000 euros (annual, gross), plus travel allowances for regular Pau-Paris trips and congress participations, and bonuses. The projected starting date will be September-October 2018.

For further information and for submission of your candidacy please contact:

- antoine.bertoncello@total.com
- josselin.garnier@polytechnique.edu
- mark.asch@total.com

The application dossier should include:

1. A full CV.
2. Detailed academic transcripts (grade sheets) of the last 3 years.
3. Internship reports, or at least a résumé.
4. Letters of recommendation, if available.
5. A letter of motivation.

References

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer. 2013. (document)

[2] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. *Annals of Statistics*, 41, 2324–2358, 2013. (document)

[3] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmark. *J. of Machine Learning Research,* 17, 1–102, 2016. (document)

[4] J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys,* 3, 96–146, 2009. (document)

[5] P. Spirtes. Introduction to causal inference. *J. Machine Learning Research*, 11, 1643–1662, 2010.

(document)