Machine learning versus physics-based modeling in petroleum reservoir modelling

Jose Andino Saint Antonin

Introduction

This article compares data-based modelling vs physical modelling for the specific challenges faced by geoscientists in the oil industry.

We begin with definitions:

Data-based models can be quite simply described as surfaces that interpolate available data in some way. Figure 1 shows a simple data-based model, describing income of professionals as a function of years of education and seniority.

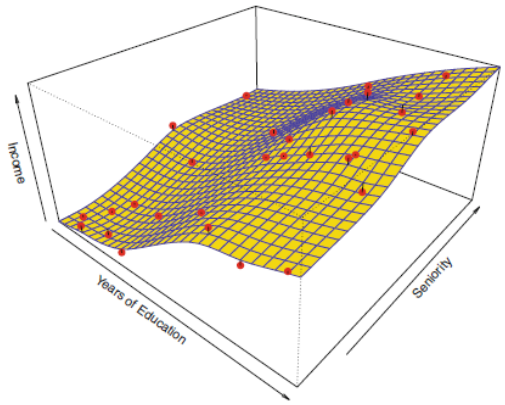


Figure 1 simple data-based model, describing income as a function of education and seniority. The red dots indicate observations and the surface is the data-based models

Our model in this example can predict, based on this interpolation or fitted surface, what will be a reasonably expected income of an individual.

Data based models are built without having any knowledge of laws that regulate the processes under study, yet they can be very useful and predictive, and above all, easy to construct if data is available.

Data based models can be very simple, such as a linear fit using least squares of a very complex neural network such as the one shown in Figure 2. Either way, they are just functions that can produce a certain output based on an input and they honor the data in which they were trained.

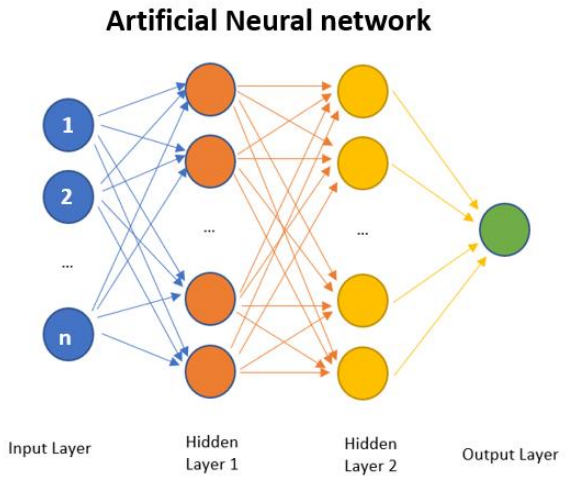


Figure 2 artificial neural network scheme used to generate a data based model

The type of modelling described above falls into the category called supervised learning (we know input and output and the data is structured). Other disciplines within data science focus on unstructured data or interacting with outside world or a program to achieve something (reinforcement learning) and are not covered in this article.

Conversely, physics based models begin with the knowledge of the underlying physics (e.g. see Figure 3 for the constitutive equations of continuum mechanics) and require perfect knowledge of the properties of the domain and its boundary conditions.

$$\begin{cases} \dfrac{D\rho}{Dt} + \rho\boldsymbol{\nabla}\cdot\boldsymbol{v} = 0 \\ \rho\dfrac{\partial \boldsymbol{v}}{\partial t} + \rho(\boldsymbol{v}\cdot\boldsymbol{\nabla})\boldsymbol{v} = \rho\boldsymbol{b} + \boldsymbol{\nabla}\cdot\boldsymbol{\sigma} \\ \rho\dfrac{\partial e}{\partial t} + \rho\boldsymbol{v}\cdot\boldsymbol{\nabla}e = \boldsymbol{\sigma} : \boldsymbol{\nabla}\boldsymbol{v} - \boldsymbol{\nabla}\cdot\boldsymbol{q} \end{cases}$$

Figure 3 system of coupled PDEs used in continuum mechanics (conservation of mass, momentum and energy, from top to bottom).

An additional, often non-trivial hurdle is that the equations that need to be solved with highly specialized software and often making very significant simplifications.

Pros and cons

Data based models shine when:

- We have copious amounts of data about a specific process we want to model
- Our data is representative of the conditions under which we want to predict (the cover the whole range)
- The process is highly complex and we poor understanding of the underlying laws
- The process we wish to model is repetitive (or short term forecasts are required)
- We need a fast and cheap solution

Conversely, physics based models take the lead in the following circumstances:

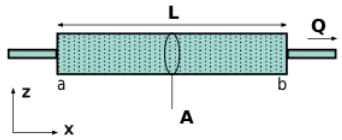
- We have a deep and established understanding of the underlying laws of the process we wish to model
- We have knowledge of the properties of the domain and the border conditions
- We wish to model long term processes that have not been observed yet (e.g. long term recovery in a field)
- We aim to optimize a design using the physics model as a substitute for constructing the real thing in multiple versions (which may be very expensive)
- We wish to change one variable at a time to assess its separate impact
- We have at our avail computing power and specialized software to carry out the simulations required.

It follows from this list what the disadvantages of each type of model is: Data-based models can

seldom be extrapolated into the future in time dependent problems and cannot be used for sensitivity analysis whiles physics-based models cannot be applied if we don't know the law or the domain properties in the problem of interest.

Oil industry specifics

The oil industry regularly models the process of depletion of its oil reservoirs. Physics based models are highly valuable as they allow forecasting of oil recovery over time and also help us answer 'what if?' questions helping us optimize large investment decisions. In particular, we use well established laws such as D'Arcy's law and mass conservation.



$$\boldsymbol{q} = -\frac{k}{\mu}\left(\boldsymbol{\nabla} p - \rho \boldsymbol{g}\right)$$

Figure 4 D'Arcy's law relates pressure losses to volumetric flow in porous media

However, the Achilles' heel of all these models (and the decision we make based on them) is the fact that, in the oil industry, **we seldom have good knowledge of domain properties or border conditions**. This may sound strange to people outside geosciences, however when you consider that reservoirs lie two to five kilometers underground, it's surprising how much we know about them when it is so costly just to reach them. Every day, geologists and reservoir engineers have to make reasonable assumptions about data they don't have: the strength of the aquifer, the permeability of an undrilled are, the orientation or permeability of fractures, etc.

The answer of the industry to this problem has been to 'history match' (a.k.a. validate) the models by varying the properties until the behavior exhibited by the models matches the one observed historically (in terms of pressures and rates in the wells). However, this is very challenging as **inverse problems** (finding the input to the model that produces the known output) don't have unique solutions and very different input can have the same historical behavior but very different predictions come out of them.

With the increase of computational power, the industry had moved in the direction of integrating uncertainty into the history match and forecasting process. Large ensembles of physics models are run with different domain parameters, and the ones that actually match the observe behavior within a certain threshold are also used for forecasting.

Development decisions therefore fall into the category of 'robust optimization' that is, optimizing under conditions of uncertainty, which makes the whole process lengthy and confusing to management.

In come data models….

There are multiple instances of processes in our industry for which we have a lot of data but modelling the process using physics is extremely challenging. A good example outside subsurface is using machine learning to predict multiphase flow in horizontal pipelines (see Alhashem 2019). However, subsurface data is both more expensive and therefore scarce. This means data-based models have encroached more slowly, mostly in shale oil applications where wells are being drilled constantly and data sets are more voluminous than in conventional reservoirs. In Feder (2019) for instance, they use machine learning to optimize completions and well designs with significant economic impact. In conventional oil, applications are starting to emerge of hybrid models, like for instance creating a proxy machine learning model of your reservoir that can be used as a substitute for very lengthy and expensive simulation runs, or applying reinforcement learning to optimize development decisions allowing a reinforcement learning script to interact with this proxy model.

Recently, Ma et. Al. (2019) have optimized the net present value of waterflooding under geological uncertainties by adjusting the water injection rate using machine learning under geological uncertainty. Application of this new technology abound and will continue to increase in number.

Peaceful cohabitation

As we discussed in this article, both types of modelling approaches have niches in which they are strongest. The choice is clear: we should use the strongest for the given problem. When both types of modelling are possible, data-based models are still quick and cheap relative to physics based models and can make an excellent starting point. If further and deeper understanding is required and the budget and time are available physical modelling can be carried out fruitfully. **We advocate here for peaceful cohabitation of both types of models based on a deep understanding of their applicability, strengths and limitations**.

References

Byeong (2017): “Data modeling versus simulation modeling in the big data era: case study of a greenhouse control system”

Darcy, H. (1856): “Les fontaines publiques de la ville de Dijon”

Alhashem (2019): “Supervised Machine Learning in Predicting Multiphase Flow Regimes in Horizontal Pipes” - SPE-197545-MS

Feder (2019) :”Machine Learning Optimizes Duvernay Shale-Well Performance” SPE-0519-0065-JPT

Ma (2018): “Waterflooding Optimization under Geological Uncertainties by Using Deep Reinforcement Learning Algorithms”