SPE-174914-MS

Performance Evaluation of Automatically Generated Statistical Type Curves

Heidi Anderson Kuzma, BetaZi LLC; Janette D. Conradson

This paper was prepared for presentation at the SPE Annual Technical Conference and Exhibition held in Houston, Texas, USA, 28–30 September 2015.

Abstract

Predicting the future production of Proved UnDeveloped (PUD) reserves is an important part of the evaluation of any asset. In the absence of production history, the most reliable way to make a PUD forecast is by analogy with production histories from similar nearby completions. This is most often done using a type curve. A traditional type curve is generated by normalizing to one the production histories of analogous wells and finding the shape of a curve which best fits their average decline. This is a tedious process when done by hand, and results are so difficult to test that in practice they rarely are. Our study introduces a methodology for testing statistical type curves based on 5-fold cross validation (5FXV), which rigorously evaluates the ability of an algorithm to produce accurate type curves, giving confidence that those type curves can be used to make careful and informed decisions. We offer two case studies to demonstrate the strength of the method and introduce a useful extension of the method.

A physio-statistical algorithm named BetaZi (BZ) automatically generates type curves which are considerably richer than a simple shape. They include information about actual production ranges abstracted as percentiles (90% of the time, actual production from a new well is expected to exceed the BZ p90 bound; 50% of the time it should exceed the p50, and so on.) 5-fold cross validation (5FXV) is used to test the performance of the algorithm. In 5FXV, a random 10% of the data from analogous wells is held out for testing and the remaining 90% is used to generate a type curve. A quantile-quantile (Q-Q) evaluation is made which compares the number of times the test data exceeded the bounds predicted by the type curve. The process is repeated 5 times on different subsets of data.

For this paper, we use 5FXV to estimate type curves for several plays in the Bakken and Kansas. The results show that excellent performance can be achieved if enough analogous wells are available to generate the type curve. Type curves that are made with wells that are not analogous can lead to overinflated estimations of value, as is shown with the Kansas wells. Finally, a *p-score* is computed for each well using the type curve, allowing production from wells of different histories to be conveniently plotted on one map or cross-plotted with other parameters.

Introduction

Being able to accurately predict the future production of oil and gas wells is critical to the evaluation of any oil and gas asset. Dozens of conferences and workshops are held every year devoted to the subject

and hundreds of papers have been published. Production forecasting and annual reserves estimation is a major activity across the industry. Publicly held companies must report reserves to the SEC. At the core of any production loan is a certified reserves report and an estimate of the Net Present Value of future production. Every company seeking investment must outline its expected future production and how it translates into cash flows.

For the purposes of economic evalution, *proved* production is production from wells that are either currently active [Proved Developed Producing (PDP)], inactive but with the potential to be re-opened [Proved Developed Non-Producing (PDNP)], or undrilled/unopened wells that are close enough to other, similar wells that their behavior can be predicted by the behavior of the neighboring wells [Proved UnDeveloped (PUD)]. Without insisting upon strict definitions for these terms (the P in PDP, PDNP and PUD can also stand for "Probable" or "Possible"), we can say that a production forecast made for wells in any of these categories is a production forecast made with *data*. A PDP, PDNP or PUD well has *production history*. There are many other categories of reserves for which there is no production history and where a production forecast must be made based on geologic or geophysical information alone, but this is risky and unproved production will be heavily discounted in financial transactions.

In this paper, we discuss the evaluation of production forecasts for PUD wells. We use the term "wells" for simplicity, but the analysis applies equally to the individual completions within a well that is producing from multiple reservoirs. A PUD well has not yet been opened, so there is no production data to use for forecasting, but a number of nearby *analogous* wells can typically be identified which are producing from the same reservoir and have had similar production treatments. We are therefore able to make the fairly safe assumption that the production from the new well will follow a similar path. This path is almost always quantified using a *type curve*.

It is important to distinguish between a *reservoir model* and a *production forecast*. A reservoir model is a picture of the underground arrived at by estimating all of the physical parameters which characterize a reservoir, such as thickness, drainage area, porosity, permeability and pressure. Using such a model, a purely theoretical type curve can be arrived at to predict production based on geologic and geophysical insight. If there is no production history from an area, this is the only way to forecast. At the other end of the spectrum, if there is data, then it is possible to arrive at a type curve by as simple a process as averaging the production from nearby wells, and this is often done. In this case, there is no reservoir model, only a production forecast. Finally, another way to arrive at a type curve is to tweak the parameters of a reservoir model until its forecast matches actual production. This is *history matching*, and it has the advantage of producing a forecast that is informed by data as well as a reservoir model that can be interpreted in terms of geologic parameters.

There is also a difference between a production forecast and *reserves*. Reserves are an estimate of how many hydrocarbons are "down there." A production forecast, on the other hand, gives an idea of how many of those hydrocarbons are likely to come up and how fast. The critical distinction, for our purposes, is that it is impossible to dig up a reservoir to find out its true reserves, whereas a production forecast can be tested. It is relatively simple to look at forecasts made some time ago and compare them to what actually happened. Similarly, holding back the final 10 or 20% of historic data and using it to validate (or invalidate) a production forecast made using the other 80-90% is a straightforward way to test the accuracy of a prediction. In this sense, production forecasts can be approached as hard numbers. When it comes to real world financial evalutions, it is nearly always forecasts (and short term forecasts in particular) that count, and not reserves. The Net Present Value of hydrocarbons produced 10 or 15 years in the future is hardly ever significant.

Nevertheless, the fact is that production forecasts are almost never regarded as hard numbers and subjected to validation, even though the mechanism exists to do so and vast amounts of money ride on those forecasts. By extension, although generating type curves may be scientifically interesting, unless it can be proved that a type curve really can predict production from a new well, it is useless (or worse) as

a predictive tool. Yet the need to test type curves as a means of forecasting new production is not typically explored except through case studies [Fetkovitch et al., 1987], [Yang et al., 2015] after the fact. This paper presents a rigorous methodology for testing production forecasts for PUD wells based on a statistical type curve that is arrived at using data from analogous wells. Such testing should be a standard part of forecasting workflows and is critical to the success of financial decisions made based on the returns that can be expected from new wells.

Semantics

Reserves and production forecasts often make use of ambiguous terminology. The first P in PUD, while generally meaning "Proved" (in the sense that at least that much production should be expected), can also mean "Probable" (likely to be the outcome but not guaranteed) or "Possible" (the given value is an estimate of the upside potential of a well). In this paper, we keep the P designation (because people are used to it), but define the likely outcome of production much more rigorously using percentile values such as p90, p50 and p10. With 90% probability, production is expected to exceed the value for p90. With 50% probability it should exceed the p50, and with 10% exceed the p10. By doing this, we establish not only that forecasting production is an uncertain science, but that a good deal of effort should be spent in quantifying that uncertainty.

Traditonal Curve Fitting

Although type curves can be found using reservoir models and history matching, by far the most common way to establish one is to fit a curve to the normalized, averaged production from a group of analogous wells, usually using the Arps equation [1945]. A number of refinements have been made to the general equation over the last 70 years [Fetkovitch, 1980]. For purposes of this discussion, classic Arps still stands.

The Arps equation can be written:

$$q(t) = \frac{q_i}{(D_i bt+1)^{1/b}}.$$

In this formulation, q is a production volume (bbl/month or mcf/mo), t is time (months), q_i is the initial production rate, D_i is the initial decline rate (vol/month/month) and b is a unitless quantity sometimes called the Reservoir Factor, decline exponent or hyperbolic factor. In any case, the interesting thing about b, from the point of view of finding a curve to fit to data, is that it essentially adds a knob which can tune the tail of an exponential function up or down. The parameters of this equation play against each other, making it very difficult to find a unique solution for all three coefficients.

Figure 1 shows an example of two curves that fit the same data. In this case, it is production from a single well, but it could equally be the averaged production from a group of wells. One curve, found using least squares function fitting, has a b factor of 0.7. The other, which by eye at lease seems to fit the data equally well, has a b factor of 0.2. It is for this reason that forecasters often fix the value of b before solving for the other parameters of the equation.

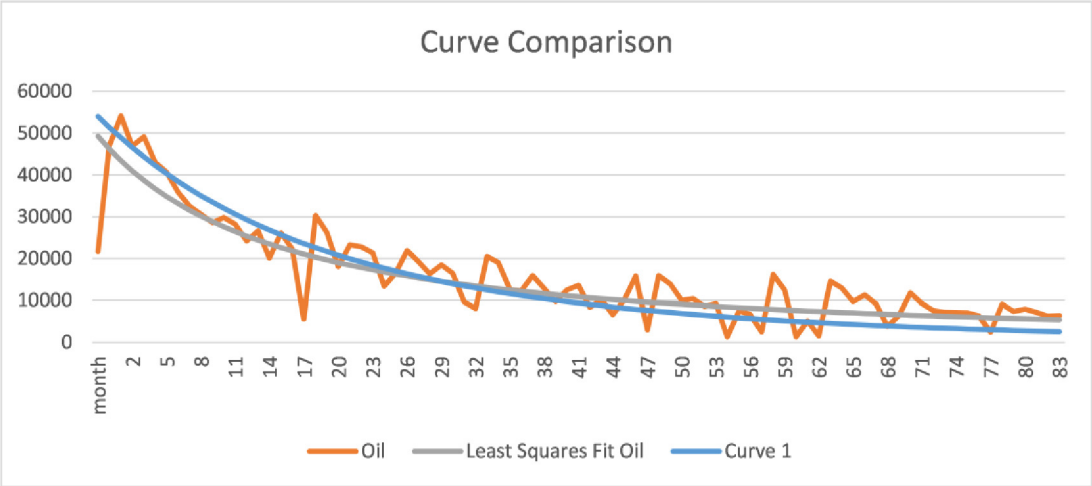

Figure 1—Least squares Fit has coefficients qi,Di and B = 42,000; 0.06; 0.7. Curve 1 has coeffieicnts 54,000; 0.05; 0.2.

This particular well is producing out of the Bakken formation. In a 2013 USGS study, [Cooke et al] used least squares analysis to find Bakken b-factors that ranged all the way from 0.3 to 2.5 with a mean around 1.2. They also found that the mean b-factor changed with the age of wells. To a student of curve fitting, this suggests less that overall reservoir characteristics change with time (which is certainly true) than that the Arps equation is not ideal for capturing the full behavior of production histories.

(We have already discussed the use of type curves as a means of forecasting production from undrilled wells. One of the other common uses of a type curve is to fix the b factor so that the other parameters of the Arps equation can be found for a particular well history without ambiguity.)

Methods

Physio-statistics

BetaZi (BZ) is a physio-statistical production forecasting algorithm which is gaining industry acceptance while being used as a forecasting engine for several commercial products [Kuzma et al, 2014]. The algorithm combines a physical model with a set of statistical distributions learned from big datasets. The physical model expresses production volumes in terms of simple curves which mimic eigenvectors of differential fluid flow equations. (Since they are not exactly eigenvectors, we will call them *eigencurves;* the actual formula for them is proprietary.) This allows BZ to make accurate forecasts regardless of the number of workovers or stages a well has had or number of different fluid flow reqimes it has gone through and without the ambiguity of fixed equations such as Arps.

The statistical part of the algorithm contains probability distributions that describe the probability of the number of eigencurves that are typically needed to fit a curve, as well as their parameters. In addition, there are distributions describing other, non-physical and operational parameters such as the typical number of workovers and shutins and other deviations from a well history that could be described purely by reservoir dynamics.

All of the probability distributions taken together constitute a “prior” in statistical parlance: a best guess based on analogous information. In other words, they are the statistical equivalent of a type curve. By drawing random samples from this prior it is possible to create an infinite number of synthetic production histories that have the same statistics as the dataset used to train the prior and, when inspected, look very much like actual production histories, including all of the noise and operational artifacts that keep actual histories from having smooth declines. The sampling can be focused using Markov Chain

Monte Carlo (MCMC) techniques in order to explain a particular well history. The result is a number of samples which are consistent with the history and which can be pushed into the future for forecasting. Instead of a single forecast, however, the result is a million samples. To make the results more palpable, statistics are computed over the forecasts resulting in percentile values that range from 1 to 99. As shown in Figure 2, 99% of the samples fall above the p99 line. 98% of the samples fall above the p98, and so on. This leads to a median forecast (the p50) and two important bounds: the p10 above and the p90 below.

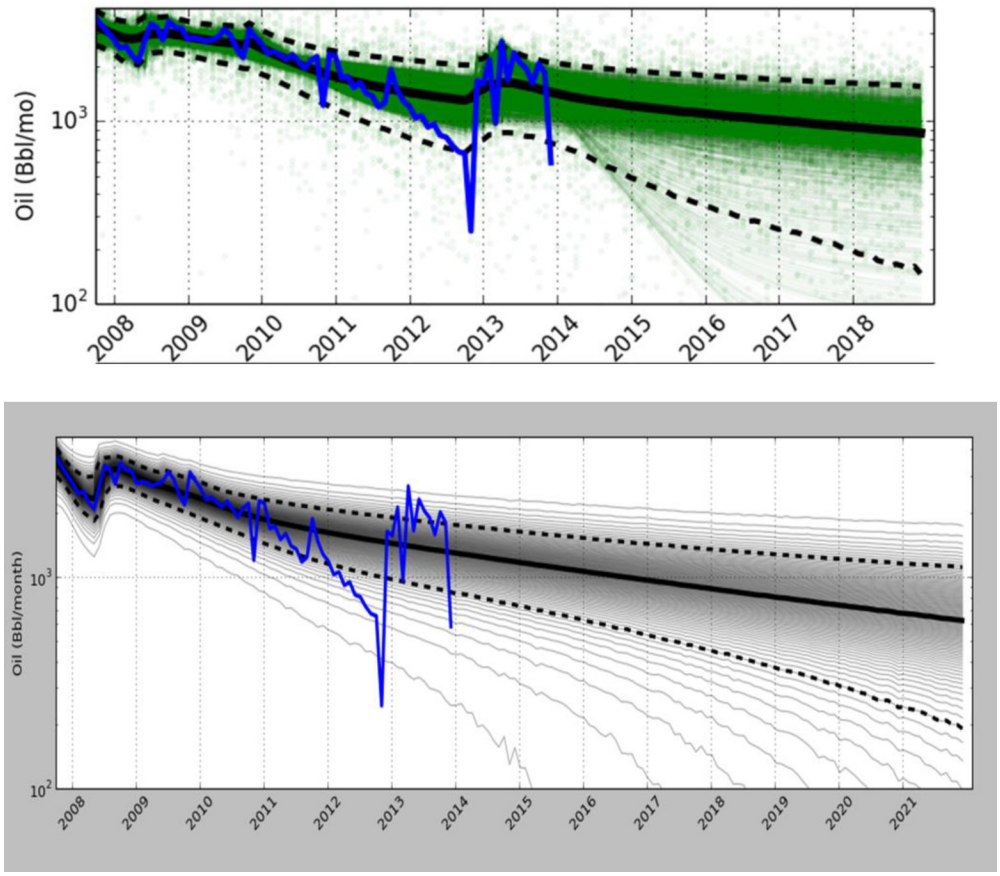

Figure 2—Top: A million samples (green) are computed to explain the production history (blue). Bottom: Statistics are run over these samples to produce p-values ranging from p99 to p1. The most important bounds are the p90 (lower dotted line), p50 (solid), and p10 (upper dotted line).

Physio-statistical type curves

BZ is not exactly a black box because the equations that it is solving are transparent. However, the parameters of its internal probability must be "learned" by estimating their values using the production data from actual wells.

The standard BZ prior was learned from several hundred thousand wells. However, other priors can be learned using much smaller subsets of wells-- as few as 10 or 20. If these wells are carefully chosen to represent a particular reservoir, the resulting prior is equivalent to a type curve, accurately representing the full range of production statistics for those wells.

An example of such a statistical type curve is shown in Figure 3. It is similar to the forecast in Figure 1, since it is also expressed as a series of curves representing percentiles, but this time there is no data on the graph. (Only every 10th percentile is plotted instead of all 100 percentiles in order to make the plot cleaner.) This prior is a statistician's best guess of production from a new well based on knowledge of the behavior of the well's neighbors. It is a very rich guess, containing not only an estimate of how the well

will decline, but estimates of when it will reach its peak production (which is usually some months after the well is opened); estimates of opening volumes; and an automatic estimate of high, medium and low scenarios with rigorous definitions (p10, p50 and p90).

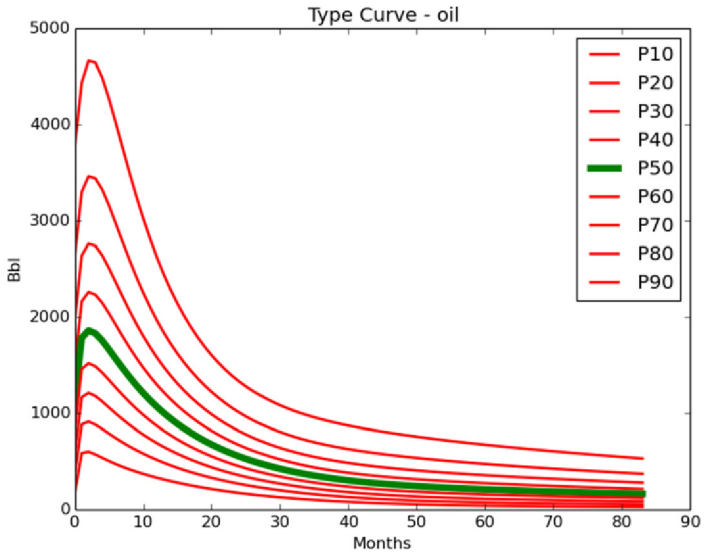

Figure 3—Example of a statistical type curve

Calibration

Cross validation is a standard method in computer science for testing the ability of an algorithm to predict data that the algorithm has not seen before (such as the data used to make the prior in BZ). To cross-validate, part of the data is held out, the algorithm is trained using the rest and then used to predict the test data. In N-fold cross validation, an Nth of the data is held out and the process is repeated N times so that all of the data has been used both for training and testing. The measure of success is how well the algorithm performs on some metric such as squared error or, in the case that follows, quantile-quantile (q-q) testing.

QQ testing

When evaluating an algorithm which produces statistical bounds such as BZ and its type curves, it is important that not just the overall error but the *bounds themselves* be tested. An easy way of doing this is simply to count them. Over a large number of test wells, does actual production really exceed its predicted p90 ninety percent of the time? And the p80 eight percent of the time? If yes, then the bounds are accurate on average. If not, then there is a problem. Usually the problem is that the wells that were used to make the type curve are not, in fact, similar to the well that is being tested. A quantile-quantile (QQ) plot such as the one show in Figure 4 (bottom) is a visualization of such a test. On it, percentiles are shown on the x axis against counts on the y. If the test results are perfect, the dots representing data counts will lie flat along the dashed black line of parity. (In this case results are excellent.) To illustrate the type curve itself, the plot above the qq plot shows one of the type curves produced from 5-fold cross validation against a background of all of the wells.

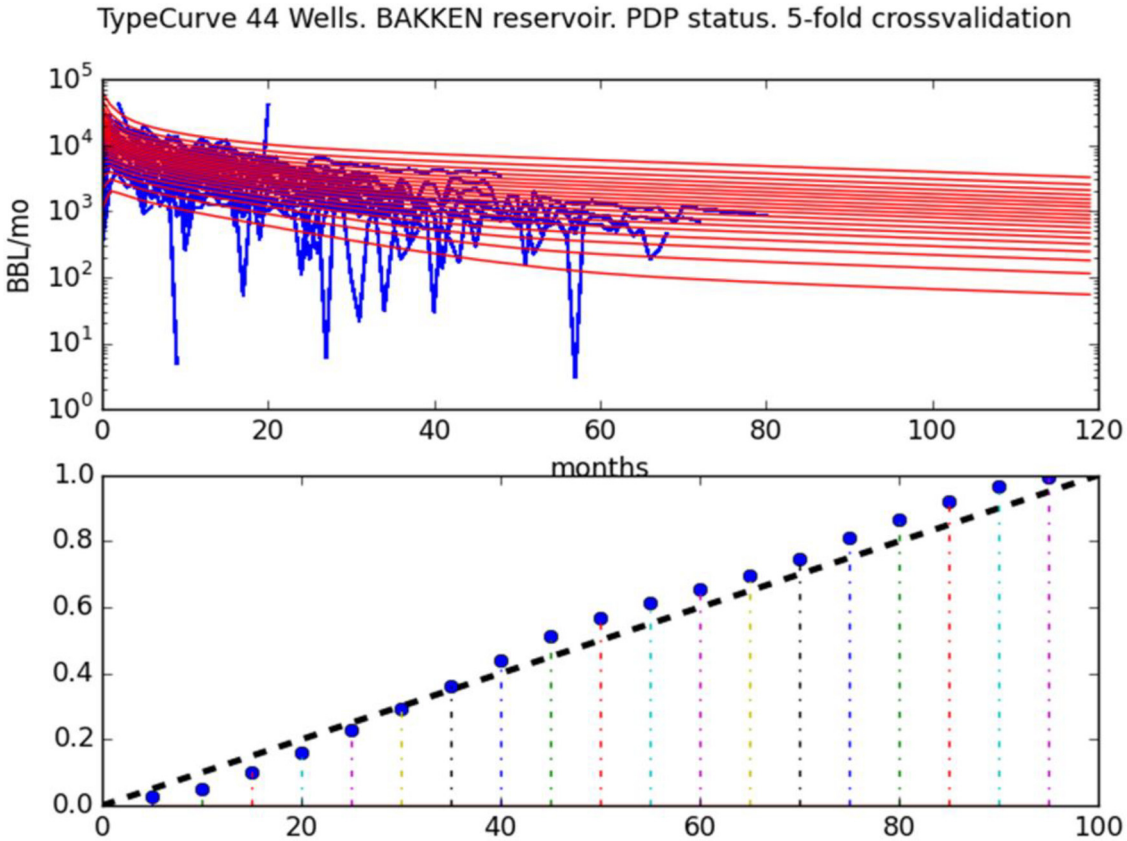

Figure 4—Type curve (red) and data used to generate it (blue). Bottom: Q-Q plot for cross validation. The y-axis is percent of wellsfor which production fell below percentile on the x-axis.

If the predicted bounds are too high, then the dots will be above the line. If they are too low, then the dots will be below. If the bounds are too wide, then the dots will start below and then go above the line. Similarly, if the bounds are too narrow then the dots will start above and finish below the dotted line.

Case Studies

Example 1: Bakken

The example in Figure 4 showed the cross-validation success of a type curve made using 44 Bakken wells which are very close together, have similar lateral lengths and are producing from the same formation. The excellent validation assures that if the same company expected to drill another 20 similar wells over the next three years, the type curve would do a good job of predicting the outcome.

The statistical type curve is shown again in Figure 5. A standard Arps curve has been fit to the p50 of the statistical distribution. It has coefficients q_i = 12,161; D_i = 0.18; and b =1.2. These factors were found using least squares curve fitting. However, there is much debate as to whether a b factor greater than one is realistic or even meaningful [Haskett, 2011]. Luckily, the curve can be equally well fit with a b factor of 0.9 if D_i is 0.10 and q_i is 10,901 — so well that the difference is hard to see on the graph.

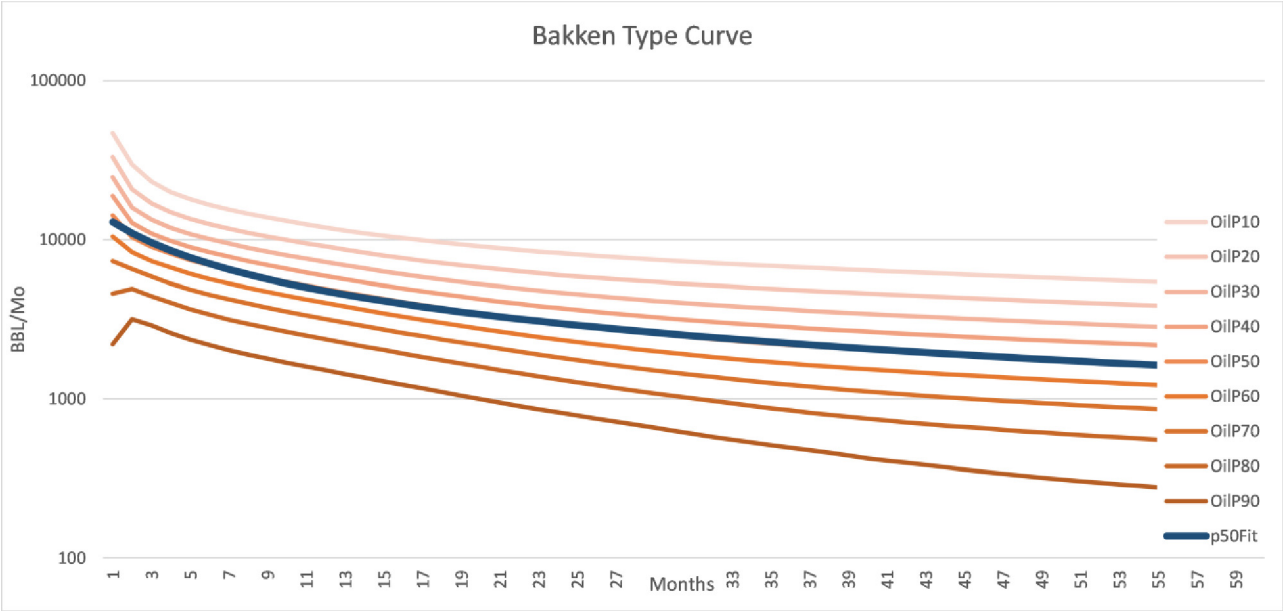

Figure 5—Bakken type curve. P50 is fit with harmonic factors qi = 12961, Di = 0.18, b=1.2.

Not only does the probabilistic type curve bypass the issue of b factors, but it rolls all assumptions about opening volumes into the type curve itself. Notice that the bounds on the type curve are quite wide, with opening volumes ranging from 1500 to 5000 BBL/mo. This means that the ratio of uncertainty to production is 1:1. The bounds tend to narrow, however, in relationship to total production as more wells are drilled because statistics do not add. It is unlikely that all of the new wells will be either high or low (unless there are geologic or operational considerations that can be taken into account). As more wells are statistically *rolled in* to an estimate of future production, the bounds can narrow by a factor of 10 or more. This is significant, but also a subject for a separate discourse.

Example 2: Kansas

In our second example run, a small number of wells in Kansas were selected by a petroleum engineer as representative of typical well production in an area that is often prospected using 3-D seismic. Indeed, the motivation for the study was to test the financial statements of a small company raising money to support seismic and the drilling of a dozen new wells. The type curves used by the company estimated single well economics by averaging the production from several nearby leases. Such production in Kansas is reported at a lease level and is easily downloaded from a state website. To make these type curves, the company simply downloaded the data, divided it by the number of wells in each of 7 leases, averaged the results and fit a curve. The result was a stated Internal Rate of Return (IIR) of 55%, which appeared *very* favorable.

However, when we made a statistical type curve using the individually allocated data from 28 nearby wells, a less rosy picture appeared. For starters, the type curve failed to pass validation, as shown in the bottom of Figure 6. The type curve clearly overpredicts in the upper percentiles. A glance at the top of the figure above (individual well) quickly illustrates why. A number of the wells, notably those with long production histories, have significantly higher production than new wells with short histories. Of course, it doesn't require cross-validation to see this, once it's been pointed out that the feature is there. But obviously the company and its investors did not look. A failed cross-validation is immediate cause to pay closer attention to the data.

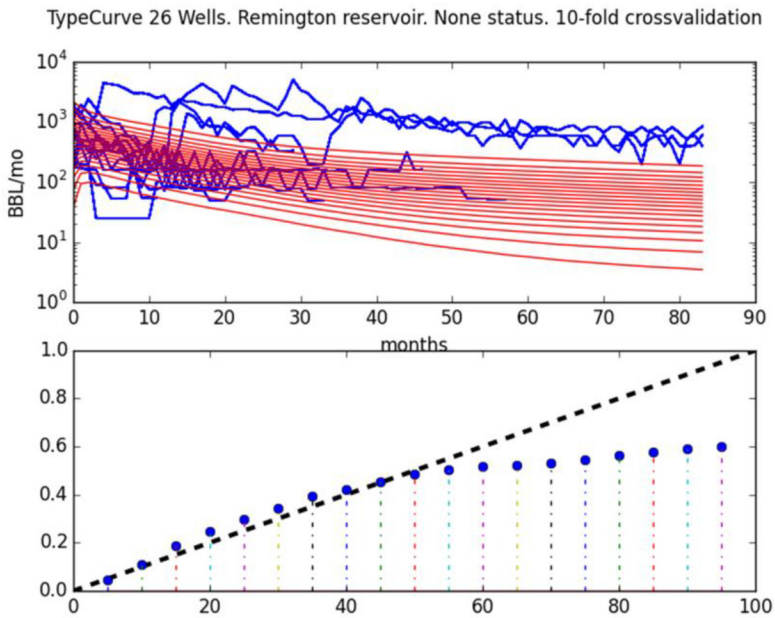

Figure 6—Kansas Type Curve Cross Validation.

The relationship between the age of the wells and their production volumes is further illustrated in Figure 7. The top shows a cross plot of opening volumes versus the first 18 months of production for each of the Kansas wells. It is clear that all of the low volume wells were opened recently, although some high volume wells were also opened. It is possible that the reason the older wells have higher volumes is because they are the wells that survive. At the time they were drilled, perhaps a number of low volume wells were also drilled which now are shut in. Or it could be that the better producing structures in the area were found and drilled early, leaving only smaller wells to be opened. Whatever the reason, it is clear that a type curve based on older wells would tend to overpredict production from the newer wells, and that is exactly what we're seeing in the failed type curve validation.

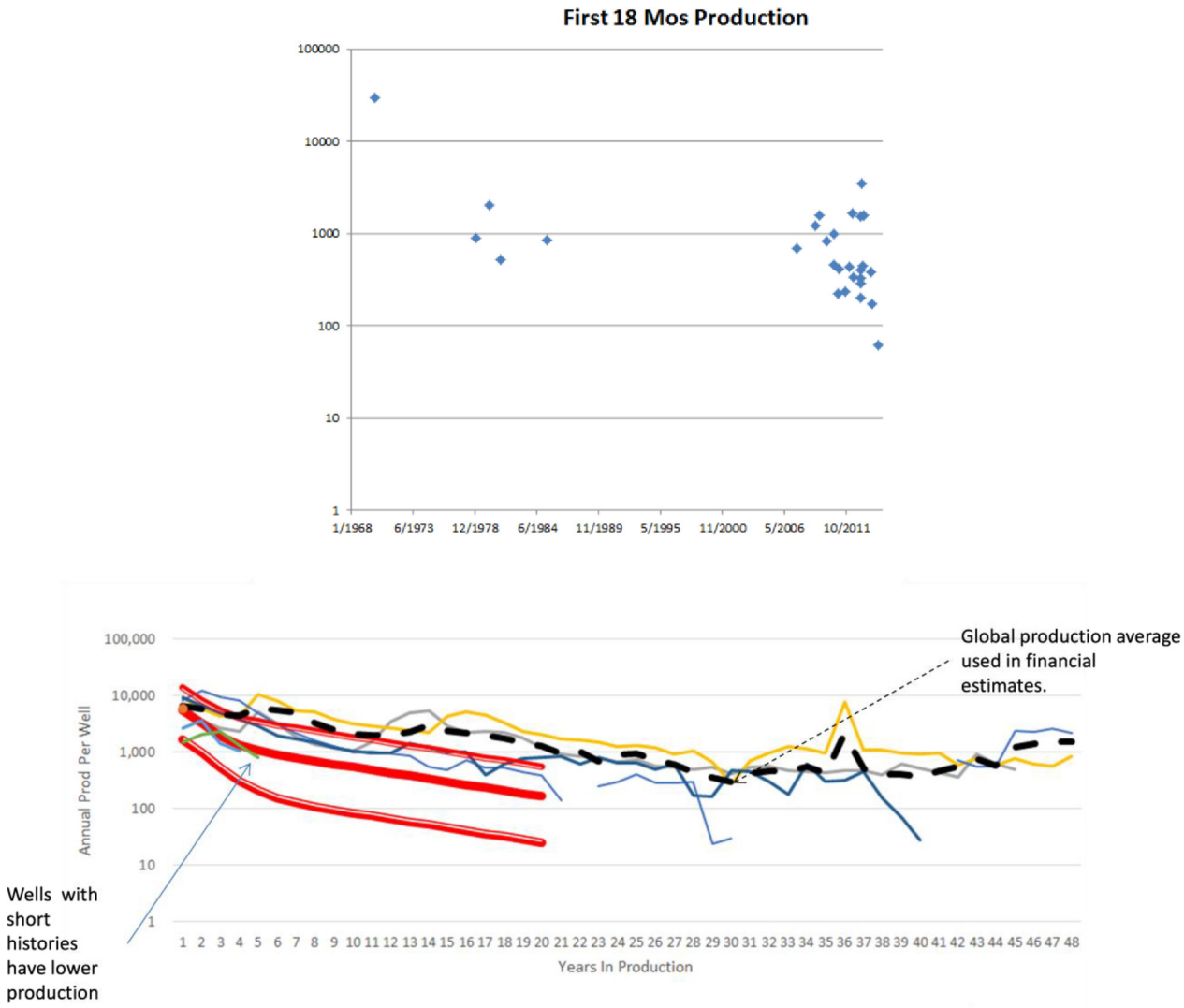

Figure 7—top: First 18 months of cumulative production (bbls) vs. opening dates for Kansas wells. Bottom: The company's type curve was based on the average production over 7 leases (multicolored). The average of the averages (black dotted) is considerably higher than the BZ type curve p10, p50 and p90 (red).

The bottom figure shows the average production from the leases which were presumably used to make the company's type curve. Even the p10 of the statistical type curve is lower than the average. It is worth noting that the p50 of the statistical type curve is close to the median of the data. As people involved in real estate valuations well understand, it is dangerous to use averages as a means of estimating typical house prices because it can be badly skewed by a few houses with extraordinarily high prices.

We can conclude from this study that one of the following is probably true:

1. The relationship between production and opening date is just chance
2. New well treatements and lateral lengths are less effective than older ones
3. High producing wells survive to be older and therefore are skewing the sample (more likely)
4. In this area where 3d seismic has been used since the 1990s, the better targets are taken.

In any case, revisiting of the type curves led to recomputation of the company's financial statements and the stated IRR was revised from 55% to a p90 of 2%, p50 of 18%, and the surprising upside with the p10 at 80%. This would be achieved if the new well was one of the big wells. Before putting money in

this company, it would be important that an investor be convinced of the operator's ability to reliably drill the bigger wells.

Extension of the method: p-scores

In a third study, which we use here to demonstrate a further benefit of the methodology, we generated a type curve for 36 horizontal wells belonging to a single operator producing from the Bakken formation near Fort Berthold. As shown in Figure 8, the type curve calibrated moderately well, although it appears to be slightly overestimating production at the top of the graph. A visual examination of the production data used to generate the type curve shows that a number of the wells show a great deal of shutin time that is probably causing the miscalibration. The initial goal of this study, however, was not to evaluate the type curve itself, but rather to use it as a metric for measuring individual well performance.

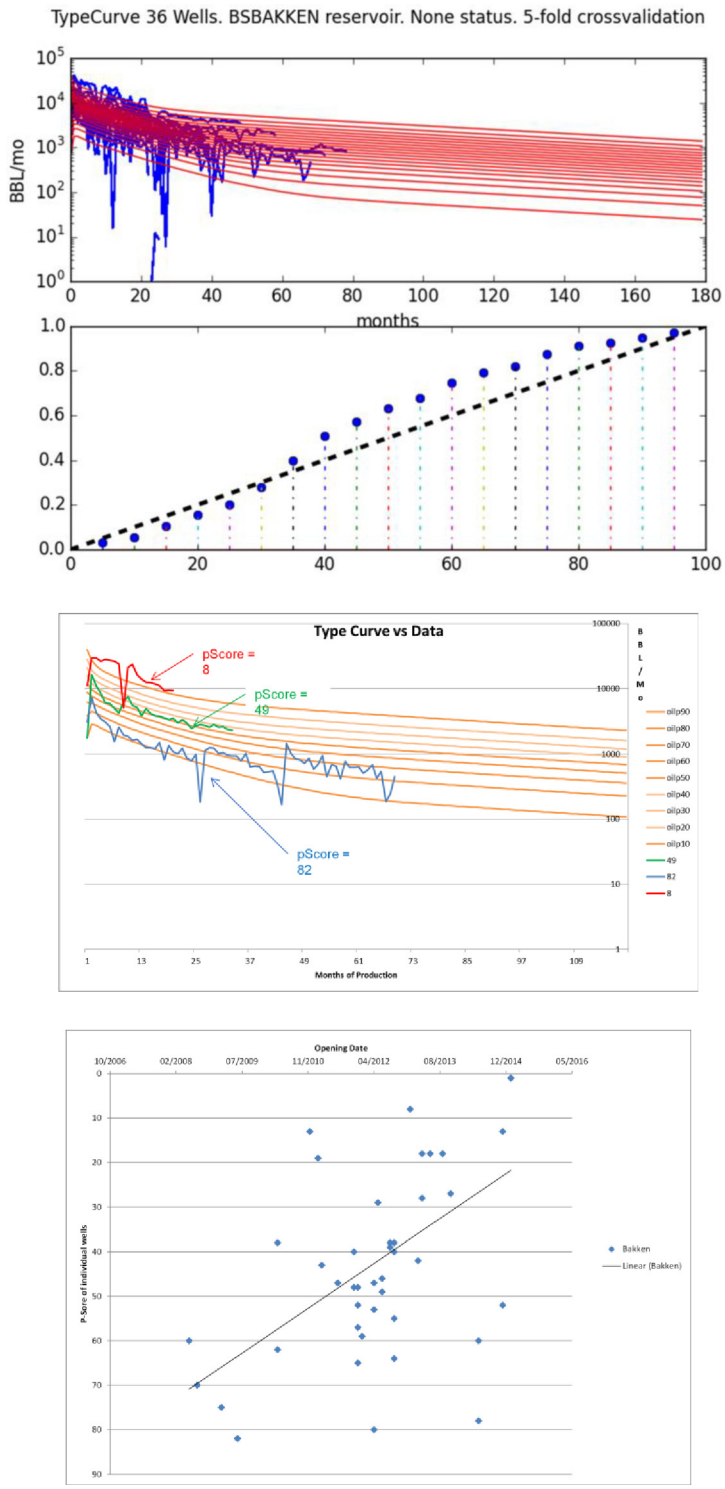

Figure 8—Top: Type curve calibration of 36 Bakken wells. Middle: pscores of three of the wells showing wells with high, medium and low production. Bottom: Cross plot showing P-scores vs. opening dates of wells.

One of the issues in trying to correlate production to geology or operations is that it is difficult to construct a single number which accurately reflects well performance and can be cross-plotted against other factors, especially when the wells have different lengths of history. The most commonly used metric is EUR, but that is an unstable number and may have more to do with forcasted than actual production if the well's history is short. Peak opening volumes are also used, but they too are unstable, particularly

for unconventional resources. Wells that open high might easily decline to much lower levels fast enough that the opening volume has very little to do with the overall cumulative production of the well. It is possible to substitute cumulative production over some period (such as the first 18 months used in Example 2), but again results can be heavily biased by wells that open high and decline quickly.

A statistical type curve solves these problems by allowing for the construction of a *p-score*. A well's p-score is defined as the average of the well's p-values as measured by the type curve. So, for example, in the middle of Figure 8, a well whose production values fall consistently above the p10 for the type curve might have a p-score of 8. A well which has overall average production will fall near the p50 and a poor performer which opened very low (below the p90) might be boosted by a workover to an overall p-score of 80. The p-score is a very natural measure of the performance of individuals with respect to their peers. Similar metrics are in common use in many fields, for example the percentile ranking of students taking the SAT test.

Use of a p-score has the further attraction of always ranking wells on a scale of 1 to 100, which makes them easy to plot. The bottom of Figure 8 shows a cross-plot of the p-score of the wells against their opening date. Unlike in the Kansas wells, newer wells tend to score better than older ones, which is consistent with the observation that operators have been learning to better produce from the shales as time has gone on. No control for lateral length was effectuated, so the effect might simply be because newer wells are longer. (Plotting the p-score against lateral length would clear up the mystery, but the data was not made available.)

Conclusions

Fundamentally, a type curve is a distillation of knowledge about a particular reservoir, resource, company, lease or play. A statistical type curve such as the one that the BetaZi alogorithm produces is more meaningful than a traditional type curve because it relies on carefully defined percentiles in establishing its bounds. As a tool for production forecasting, the primary economic function of any type curve is to predict the production of newly opened or undrilled wells. To do this effectively, it is imperative to be able to demonstrate that it is actually capable of such a prediction, which means that testing needs to be a part of the standard forecasting workflow. Cross-validation is a robust method for accomplishing just such a performance evaluation while still being able to use all available data for the final type curve.

If the type curve does not cross validate, then there is serious risk of drawing erroneous conclusions and seriously over (or under) valuing. Validation is a critical part of any mission to predict production for undrilled wells, particularly if financing is involved. Even when a type curve does not validate, it is still useful as a metric against which to measure the performance of individual wells and (in many cases) to determine why the type curve did not cross validate, which often involves useful reevaluation of data. A statistical type curve can be used to produce p-scores, which give an engineer a fast, convenient reference point to study correlations between production and any other factor.

Acknowledgements

The authors would like to thank BetaZi LLC for use of the BetaZi algorithm for statistical type curve evaluation and RPS Engineering for their expertise in interpreting results in terms of concrete petroleum engineering.

References

Arps, J. J., 1945. Analysis of Decline Curves. *Transactions of the AIME,* Vol **160,** No. 1 December. 945228-G SPE Journal Paper. DOI:10.2118/945228-G.

Cook, T. A., 2013, Procedure for calculating estimated ultimate recoveries of Bakken and Three Forks Formations horizontal wells in the Williston Basin: U.S. Geological Survey Open-File Report 2013–1109, 14 p., http://pubs.usgs.gov/of/2013/1109/.

Fetkovich, M. J., 1980. Decline Curve Analysis using Type Curves. *SPE Journal of Petroleum Technology*. **32**:06. DOI:10.2118/4629-PA.

Haskett, W. J., 2011. Unconventional Type Curves: Useful, or Sirens of Destruction. SPE Annual Technical Conference and Exhibition. DOI:10.2118/147059-MS.

Kuzma, H. A., N.S. Arora, Farid K., 2014. Generative Models for Production Forecasting in Unconventional Oil and Gas Plays. 2014-1928595-MS SPE Conference Paper.

McNulty, R. R., 1981. Statistical Decline Curve Analysis. 10279-MS SPE Conference Paper

Oghena, A., 2012. Pressure Decline Curve Analysis. SPE, North African Technical Conference and Exhibition. DOI:10.2118.1509606-MS

Raphael, S., C. P. Fuge, S. Guterierrez, H. A. Kuzma, N. S. Arora, 2015. Bit Data Every Day: Predicitve Analytics Used to Improve Produciton Surveillance. 173444-MS Conference Paper-2015.

Shirman, E., 1998. Universal Approach to the Decline Curve Analysis. Pet. Soc. Of Canada, Annual Technial Meeting. Doi:10.2118/98-50.

Yang, D., M. Hosseininejad M., Stewart, D. and S. Brand, 2015. Type Curves for Cyclic Steam Operations in the Grosmont Saleski Pilot and Their Implications for Recovery Mechanisma. 174448-MS Conference Paper.