SPE-176820-MS

Drilling and Completion Optimization in Unconventional Reservoirs with Data-Driven Models

Keith R. Holdaway, and Moray L. Laing, SAS Institute Inc.

Copyright 2015, Society of Petroleum Engineers

This paper was prepared for presentation at the SPE Asia Pacific Unconventional Resources Conference and Exhibition held in Brisbane, Australia, 9–11 November 2015.

This paper was selected for presentation by an SPE program committee following review of information contained in an abstract submitted by the author(s). Contents of the paper have not been reviewed by the Society of Petroleum Engineers and are subject to correction by the author(s). The material does not necessarily reflect any position of the Society of Petroleum Engineers, its officers, or members. Electronic reproduction, distribution, or storage of any part of this paper without the written consent of the Society of Petroleum Engineers is prohibited. Permission to reproduce in print is restricted to an abstract of not more than 300 words; illustrations may not be copied. The abstract must contain conspicuous acknowledgment of SPE copyright.

Abstract

We shall discuss the following objectives: the use of Multivariant Analysis (MVA) to identify and classify important performance variables implemented in the hydraulic fracture treatment strategies. We shall discuss field optimization workflows initially performed in the Pinedale asset in Wyoming. Data were collated from multiple fluvial sand layers that aggregate across the anticline structure under study. Initial exploratory data analysis and bivariate analyses fell short of a comprehensive appreciation of the multivariate, stochastic and multivariant nature of this complex heterogeneous system. The MVA approach addressed the complexity ″inherent in the data's coincident variation in multiple parameters. Data clustering was used to create different models and assess different parameters. Models were able to identify the relative impact of the most significant variables affecting stage production performance, and develop probability distributions for potential outcomes at different categories of production. A neural network was chosen to evaluate both reservoir parameters as well as variables that are controlled by the operator such as proppant volume and flowback methods.″[1]

″Evaluation was conducted on 195 stages of which 49 were identified as candidates for increase in proppant volume. Through this process the authors identified a need to update the model to include the impact of pressure depletion from down spacing. Even in the absence of accounting for pressure depletion the team experienced excellent results.″[1]

The goal is to design as efficient a completions strategy as possible across the anticline as more wellbores are drilled. Capturing the knowledge garnered from the geological parameters to maximize reservoir contact and the proppant volumes deemed appropriate across each stage of the wellbore, it is feasible to implement a function that identifies the values of operational parameters to maximize production. We can also identify which stages are most productive and shut down those stages that are under performers to reduce operational expenditure.

Introduction

Designing and then implementing a well completion is a complex process fraught with uncertainties. Much of the environment is heterogeneous and the physics often do not adequately model the behavior and conditions of the subsurface environment for optimized completions design. In this paper we demonstrate how a data-driven approach ″that addresses the multivariate, multivariant, multidimensional and stochastic″[1] nature of

the subsurface can be created; see Figure 1. The resultant model is then transformed from a descriptive model of the subsurface, into a model which can be used in the design and implementation stages of constructing an effective completion that maximizes well production rates.

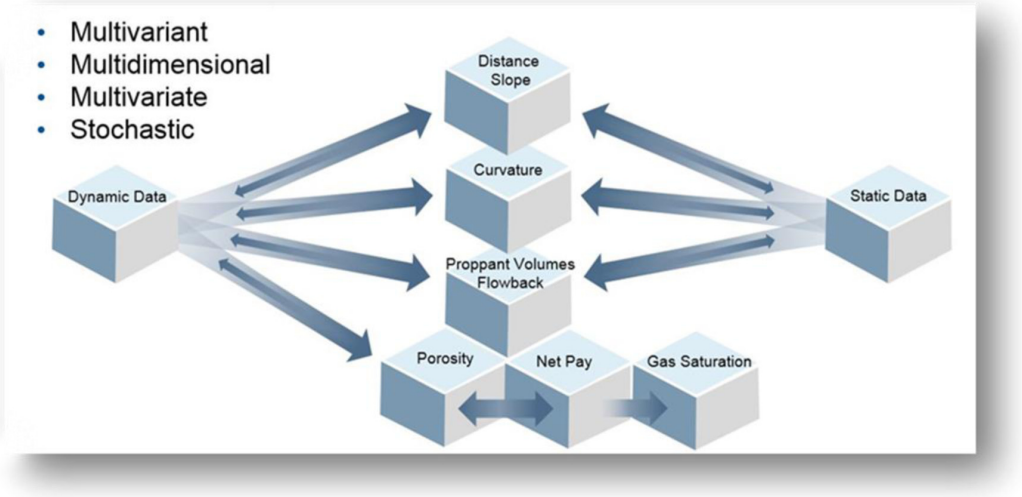

Figure 1—Subsurface Multivariant, Multidimensional, Multivariate and Stochastic challenge

Analytical Methodology Explained

Let us detail the analytical methodology followed in this paper. We have adopted the methodology detailed by SAS Institute as implemented across multiple industries.[2] The acronym SEMMA – **S**ample, **E**xplore, **M**odify, **M**odel, and **Assess** – refers to the core process of generating data mining workflows, see Figure 2. Beginning with a statistically representative sample of your data, SEMMA makes it easy to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy.

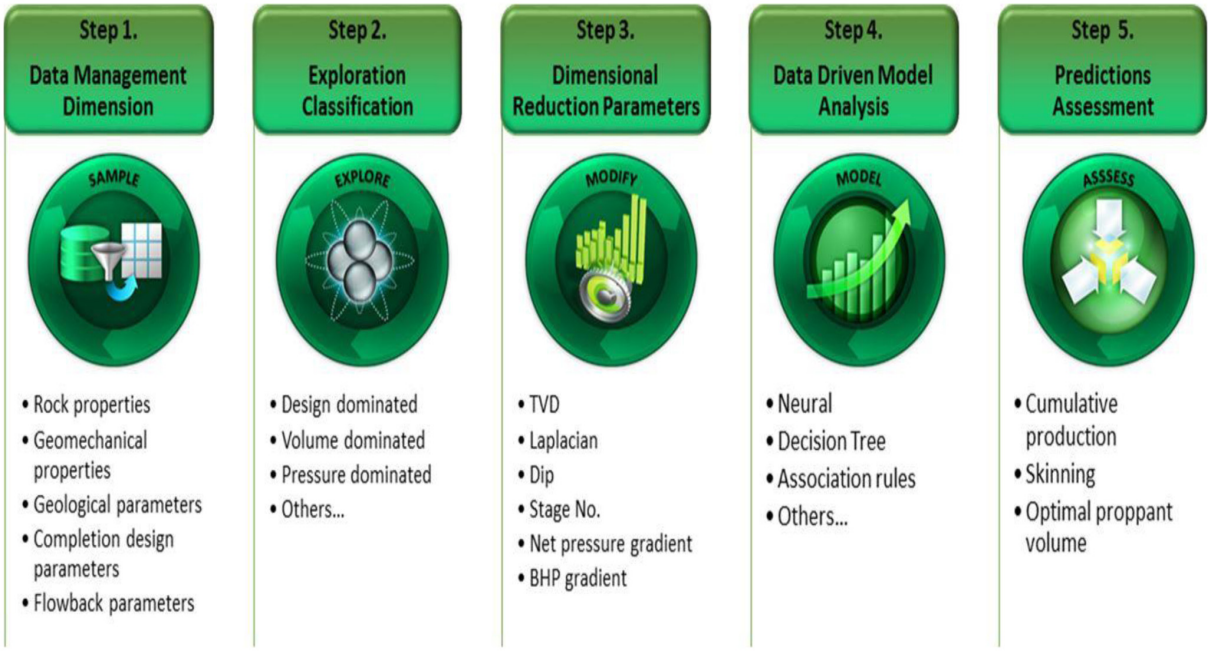

Figure 2—SEMMA Process for unconventional reservoir completion optimization

Sample your data by extracting windows, be they spatial or temporal, from a larger data set. We must be able to efficiently surface the sampled data tends and patterns without losing any inherent business significance. Reducing the data input dimensions is an important step as long as the critical patterns are still represented.

We also advocate creating partitioned data sets for:

- Training – to improve model fitting.
- Validation – to assess and to prevent potential over fitting.
- Test - to determine fitness validity for each model.

Explore your data by searching for unanticipated trends and anomalies in order to gain understanding and ideas. Exploration helps refine the discovery process. If visual exploration doesn't reveal clear trends, you can explore the data through statistical techniques including factor analysis, correspondence analysis, and clustering.

Modify your data by creating, selecting, and transforming the variables to focus the model selection process. Based on discoveries in the exploration phase, you may need to manipulate your data to include information such as the grouping of wells and significant events. Because data mining is a dynamic, iterative process, you can update data mining methods or models when new information is available.

Model your data by allowing searching for a combination of data that reliably predicts a desired outcome. Modeling techniques include neural networks, tree-based models, logistic models, and other statistical models, such as time-series analysis, memory-based reasoning, and principal components analysis. Each type of model has particular strengths, and is appropriate within specific data mining situations depending on the data. For example, neural networks are very good at fitting highly complex nonlinear relationships.

Assess your data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well the model performs. A common means of assessing a model is to apply it to a portion of data that is set aside during the sampling stage. If the model is valid, it should work for this reserved sample as well as for the sample used to construct the model. Similarly, you can test the model against known data. For example, if you know which Bottom Hole Assembly (BHA) runs have suffered from stuck-pipe issues and the model is predicting stuck-pipe, you can check to see whether the model predicts the events accurately based on the historical data. By assessing the results gained from each stage of the SEMMA process, you can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data.[2]

First Principles vs. Data Observations

As we have touched on above, there is a fundamental difference between measurements and our understanding of their use depending on whether we consider the value within the context of a measurement of a specific piece of process or equipment, or as an observation of how a system responds to various inputs and desired outputs. The first principles of any system are important, but in order to create valuable analytical models we must also understand that there are relationships within the data that can be observed from a systematic viewpoint and can be further developed to define models that describe and allow control over desired outcomes. We would suggest that the best solutions utilize a hybrid model that is constrained by first principles but defined by data observations. These bounded solutions can provide an analytically sound model of the system, which can be operationalized within the real-world context of the operation. Empirical observations and indefinite, more or less plausible, speculations attempt to explain initial conditions of any system, including the reservoir, well and surface facilities systems. Dynamic equations based on mathematics and experimental physics shed light on the dynamics inherent in systems. Thus a data-driven approach provides much needed insight.

Building the Descriptive Model

In this paper our intent is not to spend time on the minutia of how to build a data-driven model, but rather on how to use the model and obtain real operational value from it. As such we will consider the model built in SPE 135523[3], see Figure 3, as the starting point of the exercise. In this case a Neural Network has been created that describes the relationships between three key sets of parameters and a categorized objective output of cumulative gas production over the first 100 days of production.

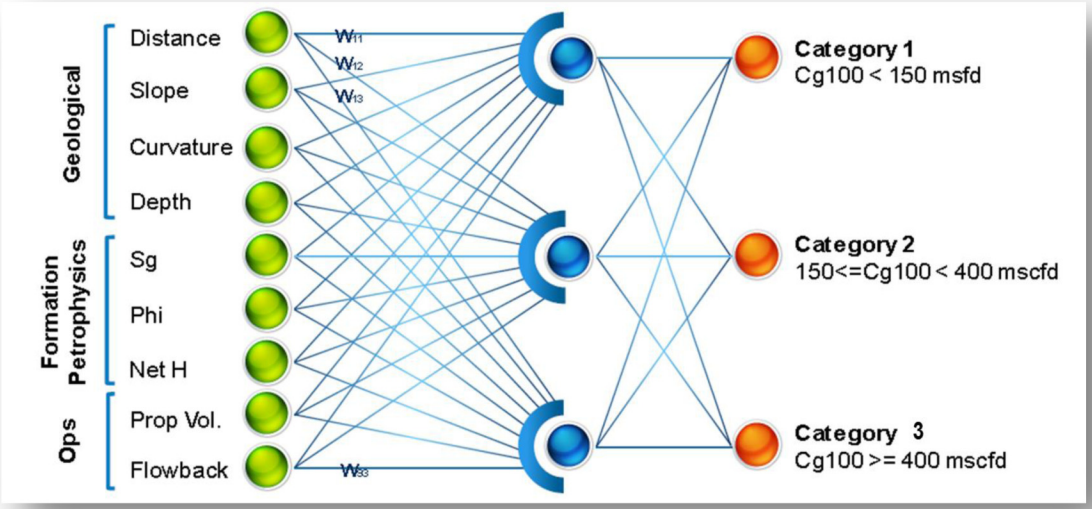

Figure 3—Neural Network of Pinedale anticline from SPE 135523

1. Geological parameters
   a. Distance measured from ″global maximum location at the peak of the anticline″[3]
   b. Slope of structure gradient: 1st derivative
   c. Curvature [Convex, Concave]: 2nd derivative or Laplacian. In mathematics Laplacian is a differential operator given by the divergence of the gradient of a function on Euclidean space.
   d. ″True Vertical Depth (TVD) from the top surface of the structure; Gamma Ray Marker″[3]
2. Formation Petrophysical Properties
   a. Sg = Gas saturation
   b. Ø = Porosity
   c. Net H= Net feet of Petrophysical pay
3. Operational Properties
   a. Proppant Volume placed in formation per stage
   b. Flowback Regime

Owing to the complex nature of earth sciences this neural network model must be validated and modified for each well and reservoir across each field under study. In this paper we will walk through its use in an operational context for a specific well. What are we trying to achieve? In this scenario we are providing input to the decision-making process for the completions design, as well as for the operational parameters used during fracturing to maximize gas production. In order to do this we must operationalize the model depicted in Figure 3 such that it can advise on those elements of the input we can control.

- Petrophysical properties – optimal pay thickness
- Geological structure in relation to the completion – placement of the wellbore
- Completions Design – Number of stages
- Operational parameters – Proppant types and volumes by stage

Initial Exploratory Analysis

First let us begin by performing some rudimentary exploration of the relationships within the petrophysical data. In our first example, see Figure 4, we have plotted gas saturation and porosity against the cumulative production by individual stages. We already begin to gain some insight to our operations by seeing a clear delineation in performance as represented by Qg100 between stages 4 through 9 and stages 1 through 3. Qg100 is a calculated variable determined by the engineers and analysts to represent the cumulative gas production measured 100 days post the collection of the Production Logging Tool (PLT) data. This logic is part of the **M**odify step in the SEMMA process; normalizing the data across all the wells under study exhibiting different spud dates.

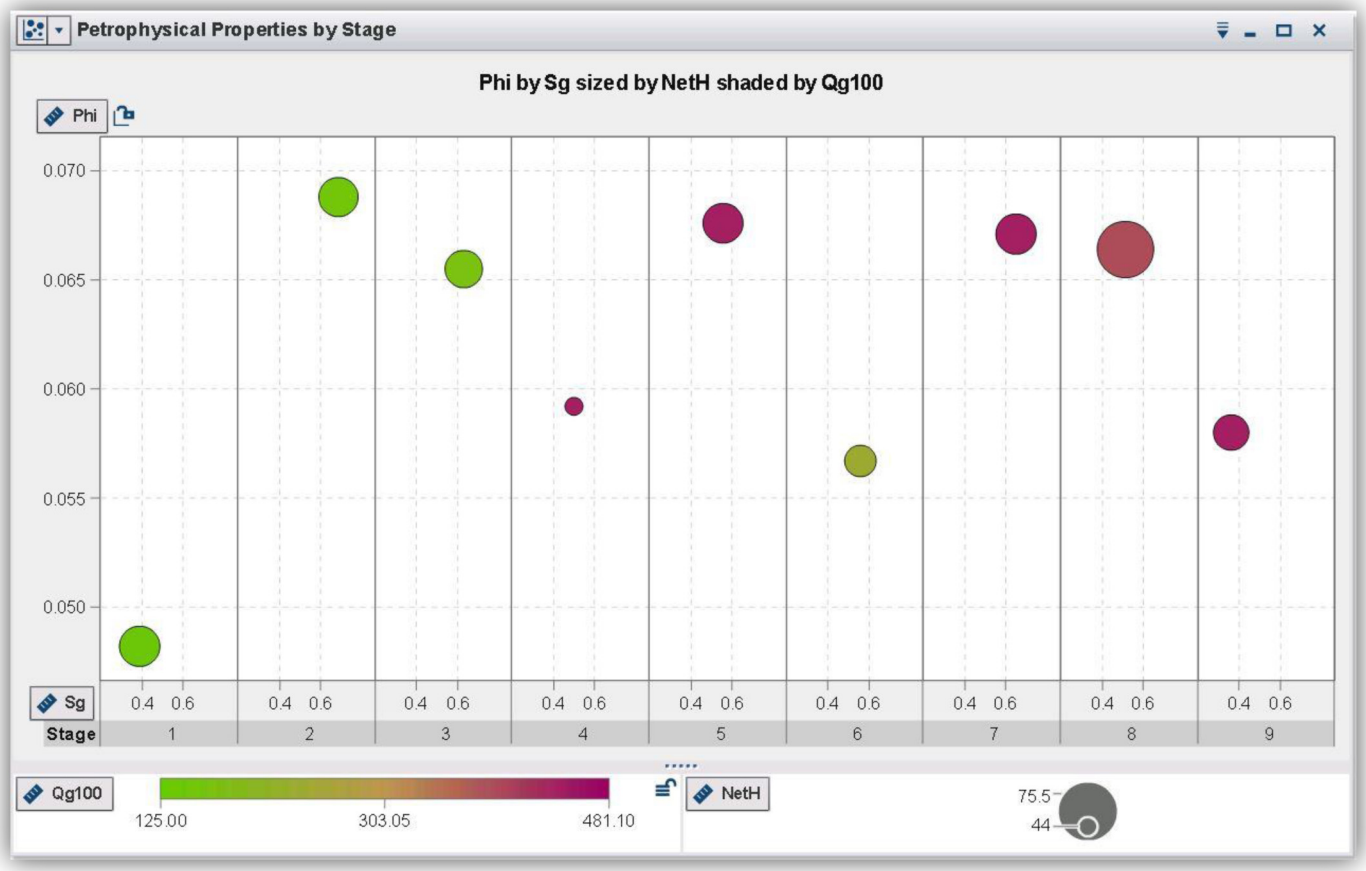

Figure 4—Gas saturation and porosity vs cumulative production by stage

Going further with this thread of exploration we further categorize the data and look at the gas saturation and porosity versus individual categories that represent different output buckets of production, see Figure 3. What we see in Figure 5 is that the same relationship of inputs to outputs exists regardless of the production rates. Although this renders no further insight into operational performance it does in

fact validate the neural network model in terms of which input variables influence cumulative production by showing the same distribution across all stages and categories.

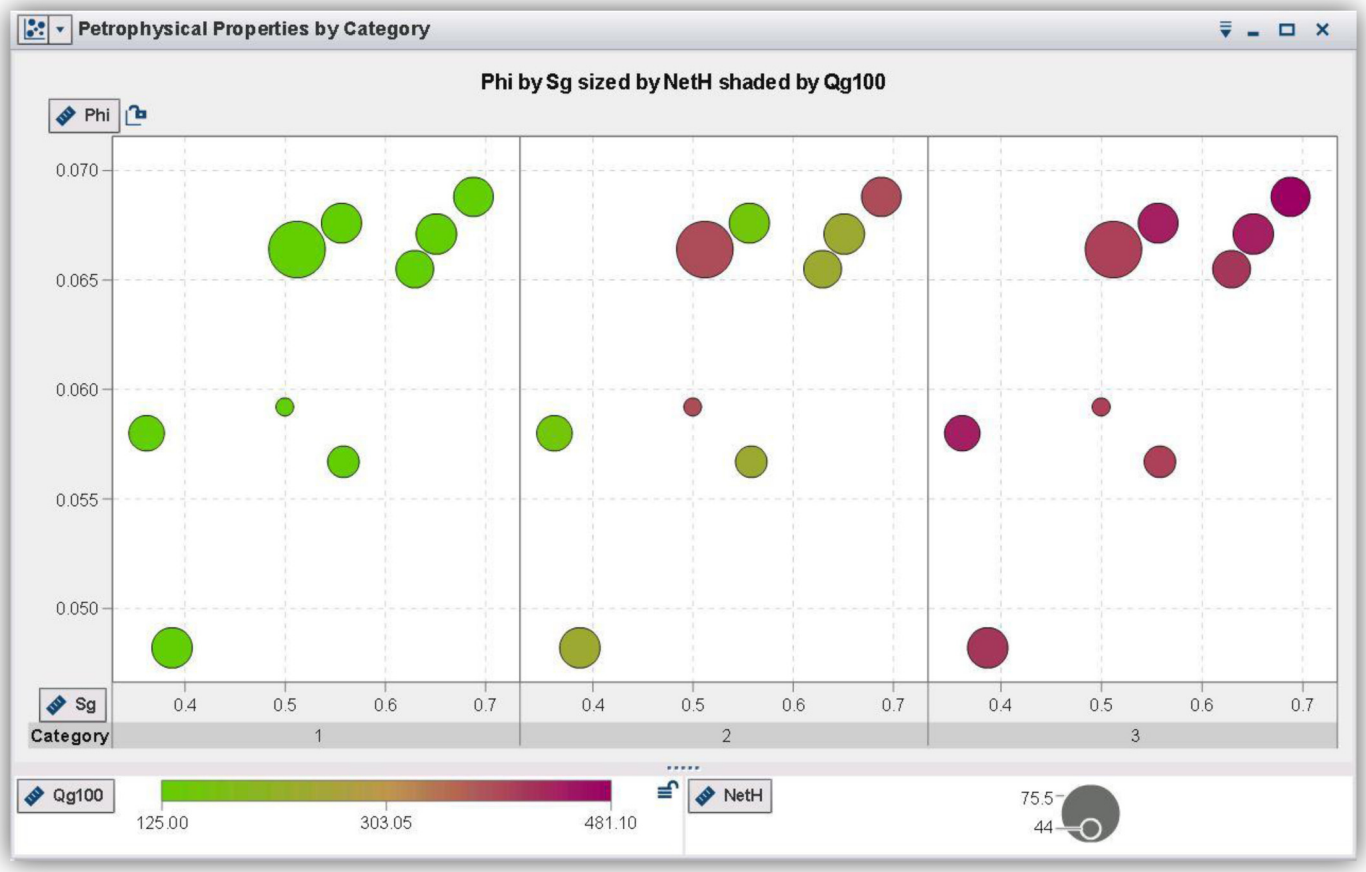

Figure 5—Gas saturation and porosity against cumulative production by stage and by category

We now perform a similar analysis of the geological parameters: Laplacian and distance from anticline peak by stage against cumulative production (Color) and the operational parameter proppant volume (Size) as depicted in Figure 6. What is immediately noticeable is that cumulative production does not appear to be a function only of the proppant volume. But there is still clearly a relationship between the multiple dimensions of cumulative production to the wellbore design and completions activities. So what is new? First let's dig a little deeper on those proppant volumes and their relationship to production. In Figure 7 we have mapped cumulative gas production by wellbore stage and proppant volume. It can be deduced that wellbores 1 and 10 are our best producers and yet the latter has used much less proppant to attain high production. Does this mean that proppant volume has no impact on production rates? Of course not, so we need to drill down one level further and study the individual wellbores. Figure 8 shows the cumulative production on each stage for wellbore 10, sized by proppant volume. This clearly shows that stage 1 in green is producing ineffectively.

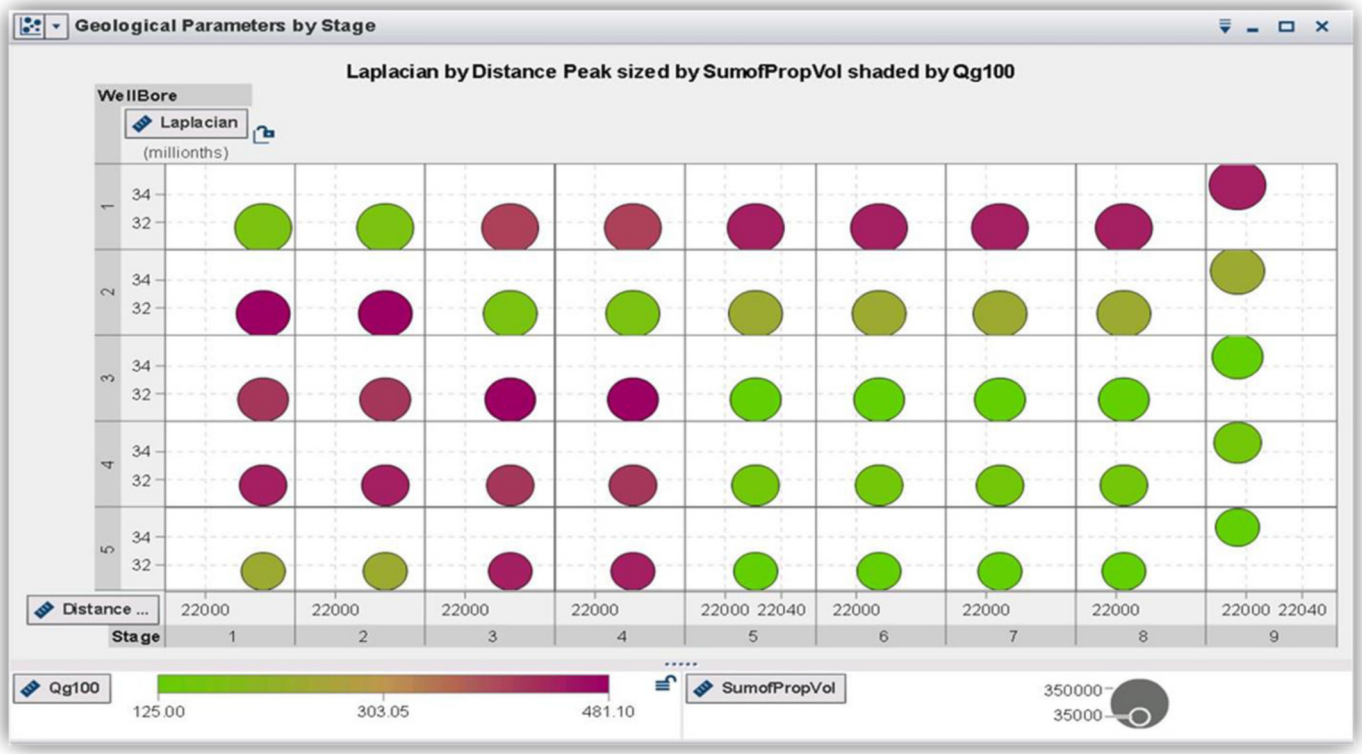

Figure 6—Wellbore placement data by wellbore, stage, proppant volume and cumulative production

Figure 7—Cumulative gas production by wellbore proppant volume impact

Figure 8—Cumulative production by stage and proppant volume on wellbore 10 Stage 1

So what? And that really is the question associated with all data-driven model studies. We have to be able to operationalize these insights in some effective manner to gain more value from our assets. From the analysis above can we now create a solution that adds real value by increasing our cumulative production rates for both existing and new wellbores?

Data Model Types

Let's look at the proppant volume in the next phase. Can we operationalize the volume required based on the completions design, petrophysical properties recorded, and geologic subsurface structure in relation to the completions system?

Let's enumerate some of the soft-computing techniques available to optimize our methodology:

- Linear
- Integer and mixed integer
- Non-linear
- Quadratic
- Network
- Stochastic
- Dynamic

There are many different flavors of optimization. Linear programming is the process of identifying how one of several variables have a direct 1 to 1 relationship with the value that is being predicted. In this example a 1 to 1 relationship does not exist. Looking at the neural net implemented for proppant

optimization, it is a classic example of a non-linear relationship. As such, a non-linear optimization program was implemented in order to find how non-obvious relationships have effect on the proppant volume variable.

There are opportunities to adopt a clustering workflow in this study, especially if the input data set is very large. SAS Institute Inc. describe clustering as most appropriate for this study. ″Variable clustering is a useful tool for data reduction and can remove collinearity, decrease variable redundancy, and help reveal the underlying structure of the input variables in a data set. When properly used as a variable-reduction tool, variable clustering can replace a large set of variables with the set of cluster components with little loss of information. Use the cluster workflow to also perform observation clustering, which can be used to segment databases. Clustering places objects into groups or clusters suggested by the data. The objects in each cluster tend to be similar to each other in some sense, and objects in different clusters tend to be dissimilar.″[4] If the data was sufficiently high in dimensionality we could have clustered to identify the characteristics or profiles of wellbores or stages across a wellbore as determined by the independent variables such as geological parameters, petrophysical formation properties and operational strategies and tactics.

Figure 9 illustrates a correlation matrix, detailing the measures of correlation coefficients. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. We can see the elements that are marked as top predictors to predict the optimal amount of proppant to use in a given stage. This information enables us to model the Non-Linear Program Optimization methodology.

Observations	90
Variable	1
Partial Variables	11

Simple Statistics

	SumofPropVol	Stage	NetH	Phi	Sg	Distance_from_Jn1_15	Laplacian	Dip	Delta_Height	Qg100	Category	WellBore
Mean	192500.0000	5.000000000	57.11111111	0.0619444444	0.5380333333	22020.65556	0.0000319233	1.675377778	4972.244444	287.9735556	2.177777778	5.500000000
StD	101093.0431	2.596453934	7.89245756	0.0065132727	0.1062097353	14.84543	0.0000009801	0.001664418	610.106775	139.0136879	0.772821734	2.888372661

Correlation Matrix

	SumofPropVol	Stage	NetH	Phi	Sg	Distance_from_Jn1_15	Laplacian	Dip	Delta_Height	Qg100	Category	WellBore
SumofPropVol	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-.0151	0.0403	-1.000
Stage	0.0000	1.0000	0.2769	0.2173	-.2150	-.9920	0.5466	-.5304	-.9957	0.0623	0.0560	0.0000
NetH	0.0000	0.2769	1.0000	0.3267	0.0388	-.2544	-.1178	0.1306	-.2317	-.0073	-.0097	0.0000
Phi	0.0000	0.2173	0.3267	1.0000	0.7623	-.1751	-.2161	0.2082	-.2443	0.0868	0.1000	0.0000
Sg	0.0000	-.2150	0.0388	0.7623	1.0000	0.2718	-.5916	0.5875	0.1730	0.0590	0.0746	0.0000
Distance_from_Jn1_15	0.0000	-.9920	-.2544	-.1751	0.2718	1.0000	-.6421	0.6263	0.9807	-.0578	-.0507	0.0000
Laplacian	0.0000	0.5466	-.1178	-.2161	-.5916	-.6421	1.0000	-.9993	-.5088	0.0140	0.0102	0.0000
Dip	0.0000	-.5304	0.1306	0.2082	0.5875	0.6263	-.9993	1.0000	0.4936	-.0150	-.0117	0.0000
Delta_Height	0.0000	-.9957	-.2317	-.2443	0.1730	0.9807	-.5088	0.4936	1.0000	-.0672	-.0617	0.0000
Qg100	-.0151	0.0623	-.0073	0.0868	0.0590	-.0578	0.0140	-.0150	-.0672	1.0000	0.8877	0.0151
Category	0.0403	0.0560	-.0097	0.1000	0.0746	-.0507	0.0102	-.0117	-.0617	0.8877	1.0000	-.0403
WellBore	-1.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0151	-.0403	1.0000

Figure 9—Correlation matrix identifying top predictors

Looking at the plot in Figure 10 of the predicted optimal proppant volume against actuals previously used, we can see a close fit to what has been used historically. However, we do have deviation. Rarely do we see an optimal amount of proppant usage. This can lead to many detrimental issues. Our aim in creating an optimized model is to have the proppant used match the ideal amount in all cases.

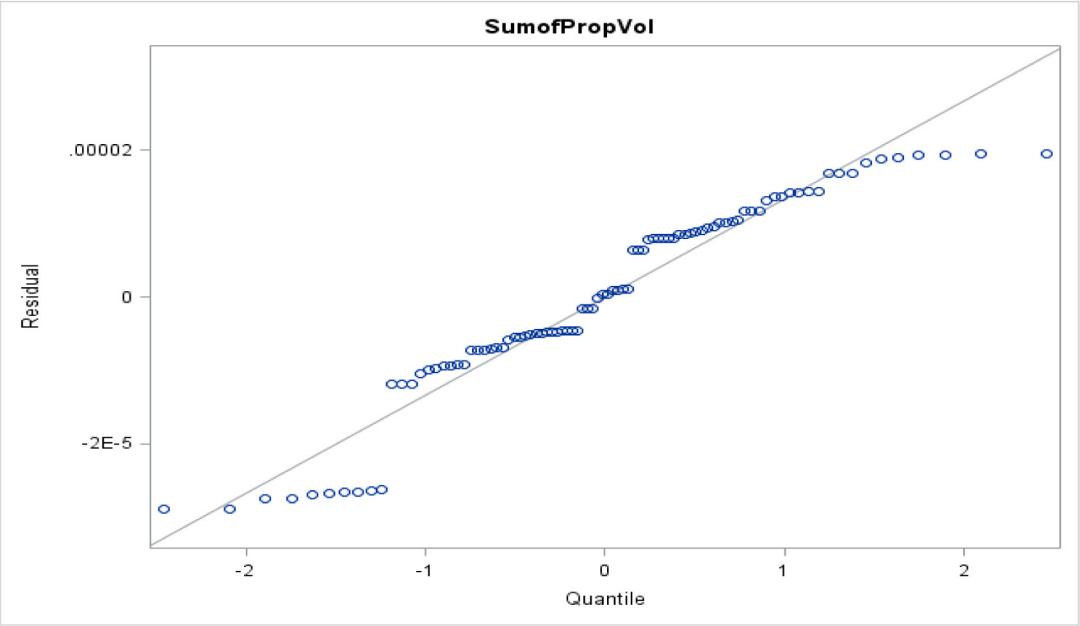

Figure 10—Predicted optimal proppant volume against actuals

Creating an operational Solution

First let's run through a basic workflow that demonstrates the techniques underpinning a predictive model. We begin, see Figure 11, by adding the data as a node into the workflow and by specifying which data should be included as inputs to the model (independent variables) and what the target variable (dependent variable) is that we wish to predict.

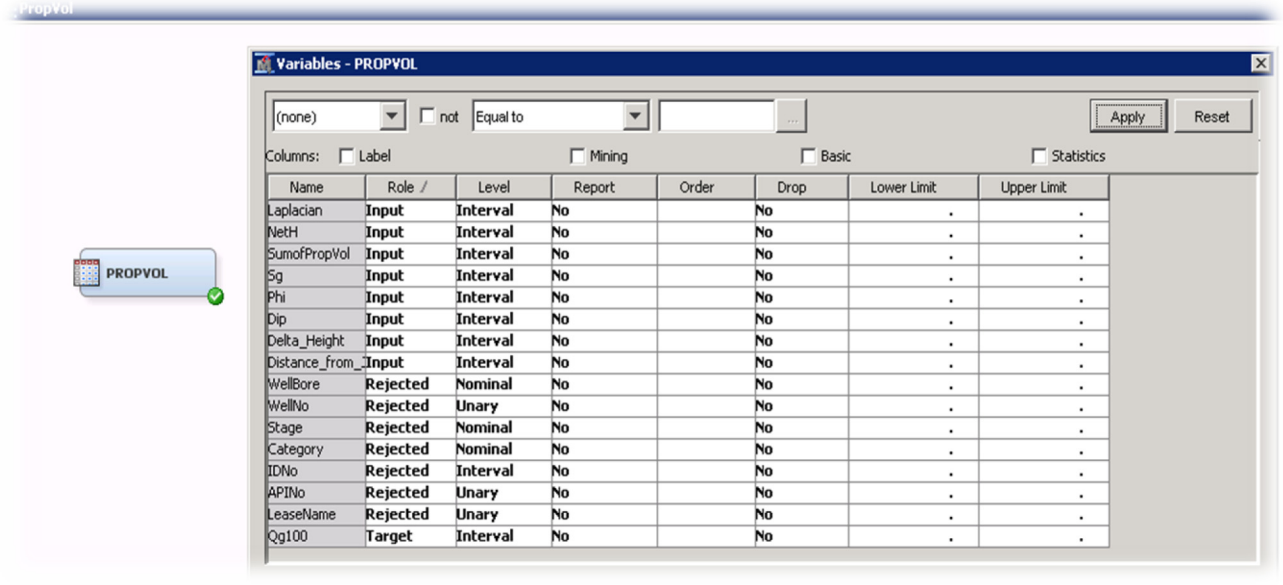

Figure 11—Setting the parameters for the model

Next we create a partition node, see Figure 12, to specify how much of the data is used to train the model, how much is used to validate the model, and how much is used to test the model. In the scenario we show we have split the partitions into 60% Training, 20% Validation and 20% Test.

- **Training data partition**: This data is where the initial learning of the data model takes place and an initial model is created.
- **Validation data partition**: This data partition is used to fine tune the model.
- **Test data partition**: This final partition is used to test the model and estimate the inherent error it may have.

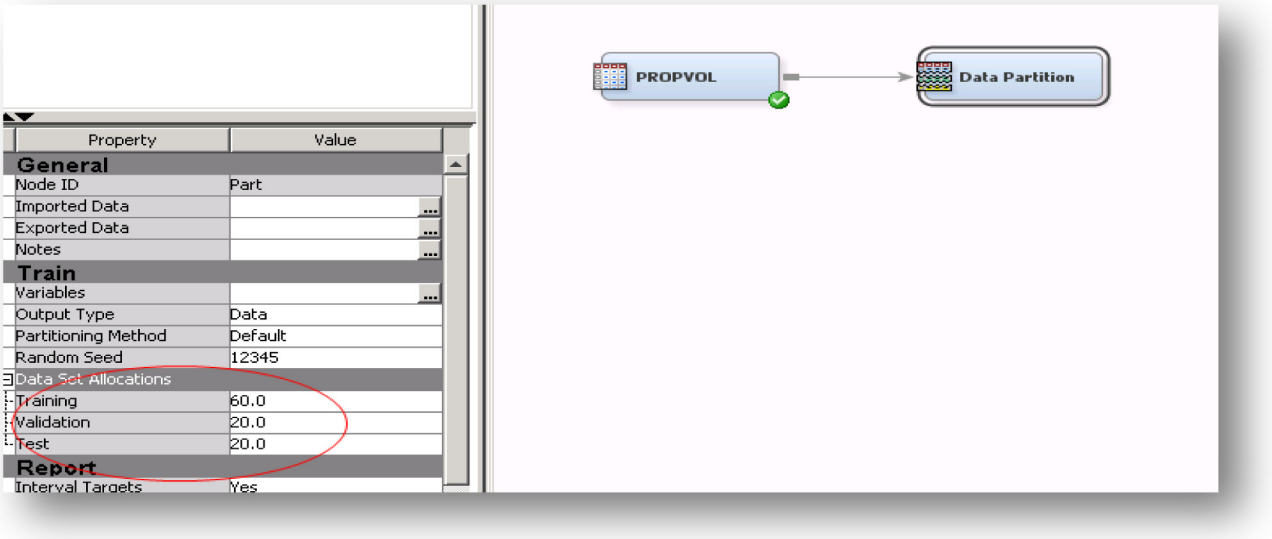

Figure 12—Partitioning the data into Train, Validate, and Test

Care should be taken in selecting how much data to assign for each of these partitions to prevent over fitting of the resultant model.

Over Fit Models

If too much data is partitioned to build and validate the model there is a likelihood of overfitting the model. Essentially the model becomes highly predictive in terms of the data supplied, but is a poor predictive model for any new data supplied. In short the model has been to closely fit to the training and validation partitions.

At this stage we don't know the best predictive model. What we need to do is build and compare the different types of model and thus ascertain which one wins, providing the best predictive fit to our partitioned data. To do this we add a selection of model types, see Figure 13, in our workflow and add a model comparison node to compare each of the results on the output side. In this example we have chosen a selection of neural networks and regression models.

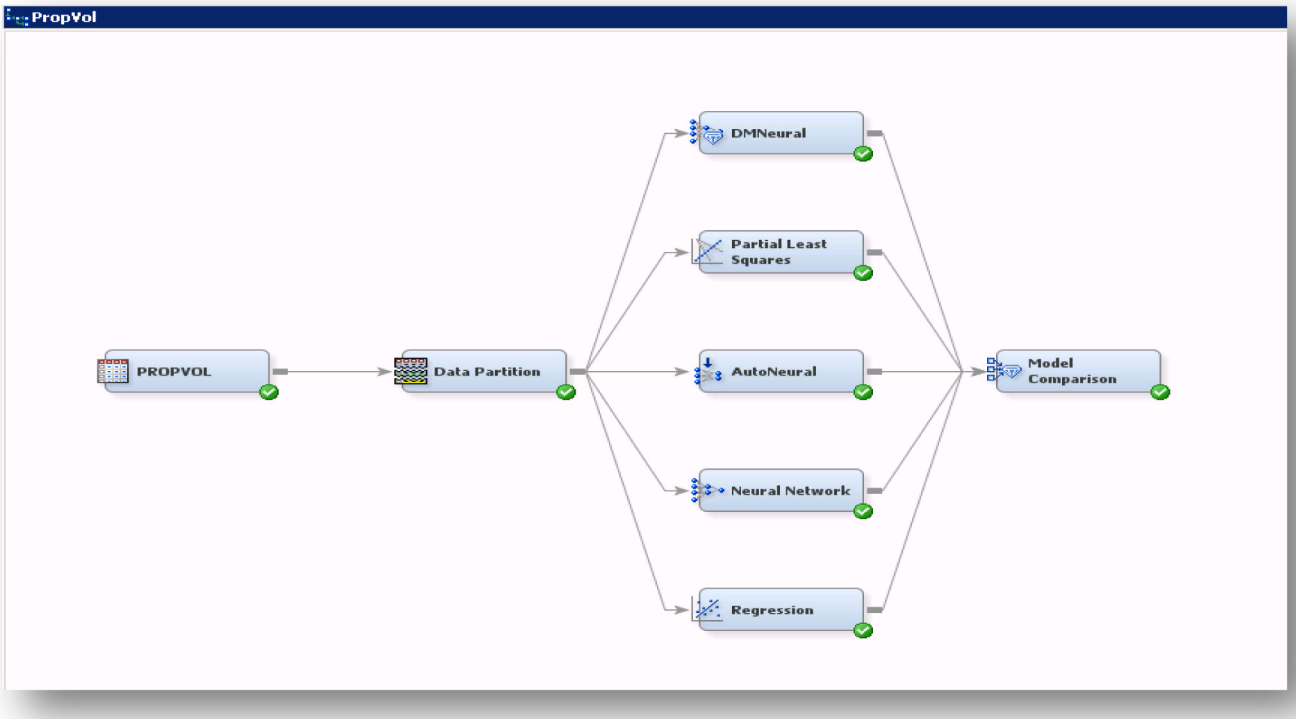

Figure 13—Build and compare the models

The results of the comparison node can now be reviewed as in Figure 14. We see that the Autoneural model is the winning model with the lowest Root Average Squared Error when the model was run on the test partition data. We also show the results of the model response from all 3 partitions in Figure 15. Normally we would expect a much higher fit, but with a small data set of 10 wellbores the resultant Autoneural model provides a good enough fit for our purposes.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target	Test: Root Average Squared Error ▲
Y	AutoNeural	AutoNeural	AutoNeural	Qg100	58.38669
	DMNeural	DMNeural	DMNeural	Qg100	82.66134
	Neural	Neural	Neural Net...	Qg100	84.10672
	Reg	Reg	Regression	Qg100	113.208
	PLS	PLS	Partial Lea...	Qg100	124.9589

Figure 14—Ranking models based on Root Average Square Error

Figure 15—Viewing the results of each partitioned model

Operationalizing the model

Having run through the SEMMA process we now have to implement the selected model. What is the inherent value of predicting the Qg100 parameter for the given inputs of our model? We need to operationalize the model to prescribe the optimal proppant volume on any given stage in order to maximize the expected Qg100. To do this we must first add another node to our workflow, see Figure 16, to score the winning Autoneural model. Once this workflow is run we obtain the mathematics that convert input values into the predicted Qg100.

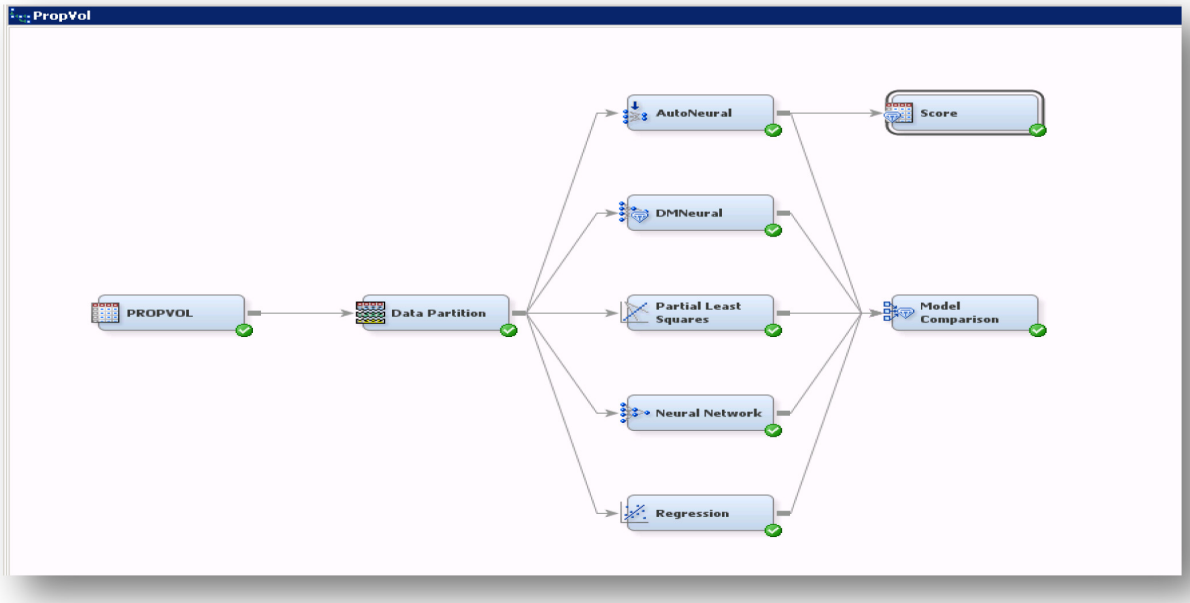

Figure 16—Completing the workflow by scoring the model

In order to compute the prescribed proppant volume we must convert the mathematics into a functional relationship in 2 phases. Ideally the software solution will have this capability built into the scoring node so that you can simply cut and paste the code as show in Figure 17 into any extant architecture and deployed application.

```
private double calculcateQG100(double propVol)
{
    double S_DELTA_HEIGHT = -7.65194194784069 + 0.00152629791116 * Convert.ToDouble(DeltaHeight.Text);
    double S_DIP = -1001.65499076188 + 597.866154469259 * Convert.ToDouble(Dip.Text);
    double S_DISTANCE_FROM_JN1_15 = -1413.45960565984 + 0.0641859229986 * Convert.ToDouble(DistanceFromPeak.
    double S_LAPLACIAN = -32.449484640268 + 1016481.71578578 * Convert.ToDouble(Laplacian.Text);
    double S_NETH = -7.46025943644986 + 0.12974364237304 * Convert.ToDouble(NetH.Text);
    double S_PHI = -8.41385196652734 + 137.568657823137 * Convert.ToDouble(Phi.Text);
    double S_SG = -4.72655651854879 + 8.95090117803797 * Convert.ToDouble(Sg.Text);
    double S_SUMOFPROPVOL = -8.15408207861868 + 0.00001999654094 * Convert.ToDouble(propVol);
    //Console.WriteLine(S_SUMOFPROPVOL);
    double H1X1_1 = 1.48140647743927 * S_DELTA_HEIGHT  -3.62825378838705 * S_DIP  -1.94510277952592 * S_DISTA
    double H1X1_2 = 6.30101973375499 * S_DELTA_HEIGHT  -15.6480900371772 * S_DIP  -7.90817462412412 * S_DISTA
    H1X1_1 = 0.51115960369446 + H1X1_1;
    H1X1_2 = -0.95566248160756 + H1X1_2;
    H1X1_1 = Math.Sin(H1X1_1);
    H1X1_2 = Math.Sin(H1X1_2);
    double P_QG100 = 112.817708299814 * H1X1_1 + -83.2054065161135 * H1X1_2;
    P_QG100 = 281.389412628665 + P_QG100;
    return P_QG100;
}
```

Figure 17—Code creation from the model for operationalization

This code can now be implemented as a prescriptive model by simply iterating through a specified range of proppant volumes to determine which volumes will deliver the maximum estimated Qg100. An example of this code is shown in Figure 18 where the engineer can specify the input parameters as well as their estimated range of proppant by stage. The program then outputs the models prediction of what proppant volume will deliver the optimized Qg100 production. In this example we have gone further and added a chart that also shows in Figure 19 the expected Qg100 by proppant volume for the given inputs.

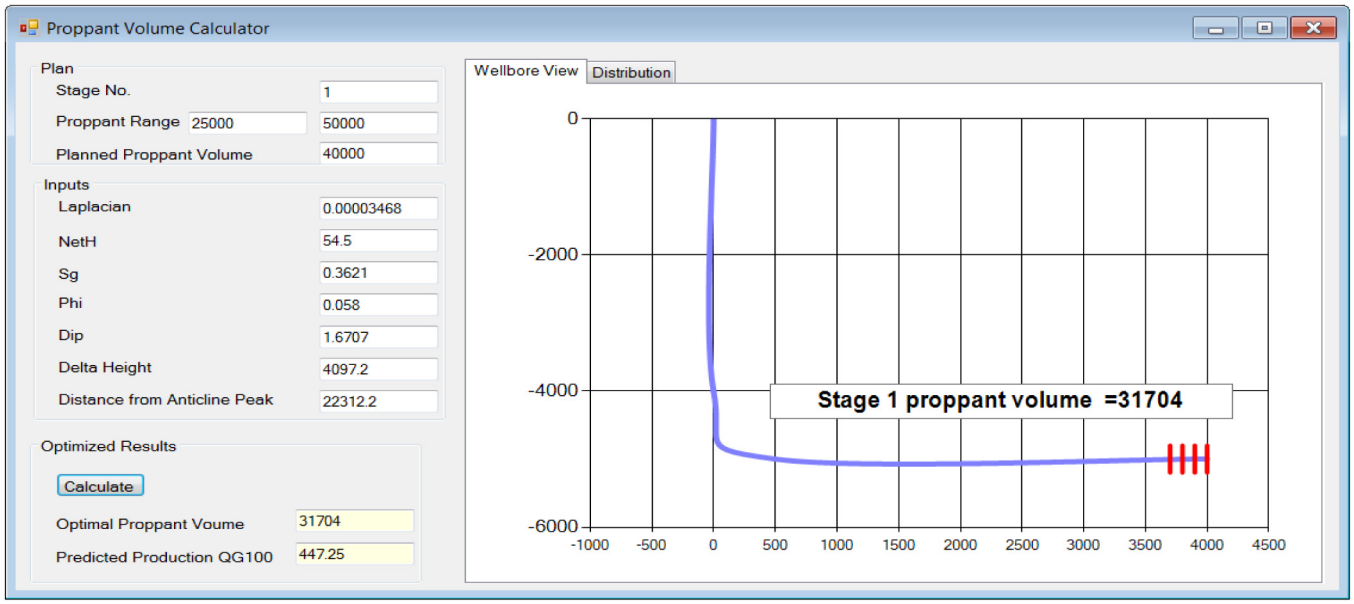

Figure 18—Example of an operationalized solution showing optimum solution

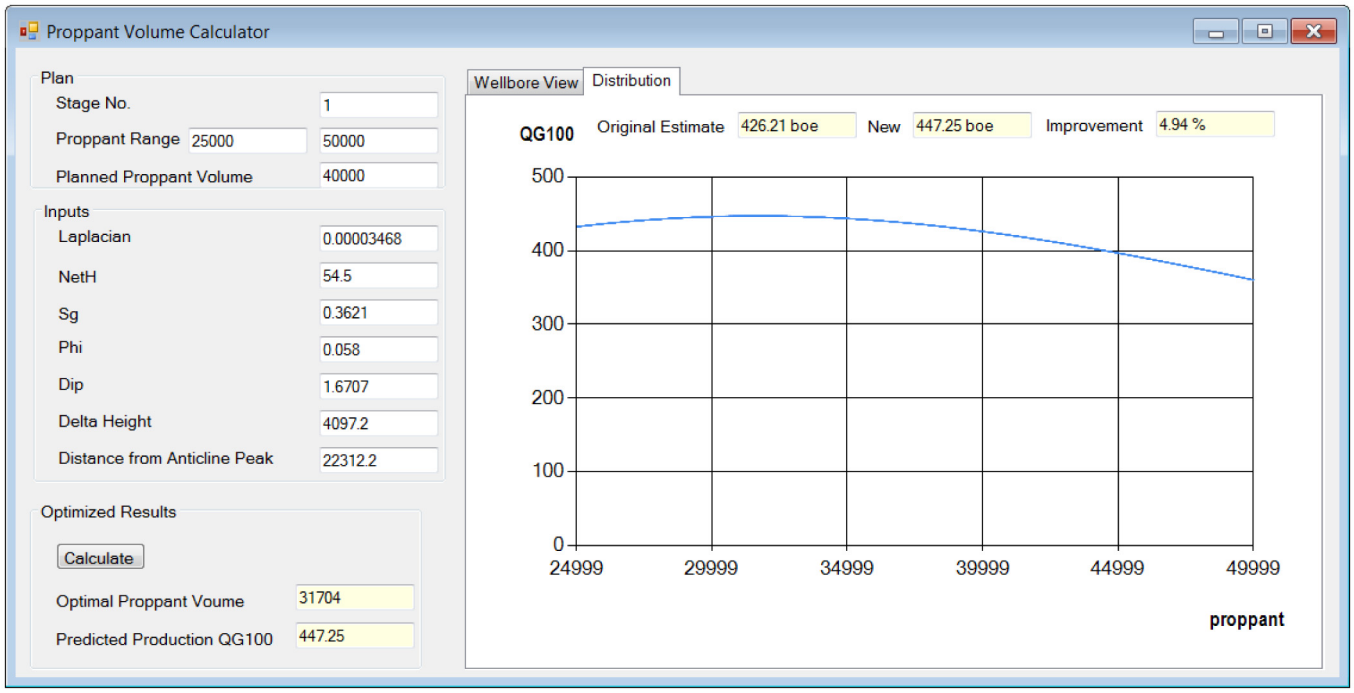

Figure 19—Example of an operationalized solution showing distribution of results

Conclusion

In conclusion we have demonstrated how efficiently, intuitively and easily data-driven insights can be achieved in not just the planning and design phases of an operation, but also in how these models can be rapidly implemented as prescriptive solutions utilized in the operational phase. The analytical workflows, models and exploratory data analysis carried out in this study are cornerstones to other data-driven methodologies you could implement to address a diverse array of business problems in the upstream. When marrying data-driven analytics with first principles we can establish a robust and repeatable suite of workflows to convert raw data into actionable knowledge. This paper illustrates one such example study.

References

1. Holdaway, Keith R., Harness Oil and Gas Big Data with Analytics, Optimize Exploration and Production with Data-Driven Models, May, 2014, Hoboken, NJ: Wiley Publishing.
2. SAS Institute Inc. Technical Documentation: http://www.sas.com/technologies/analytics/datamining/miner/semma.html
3. SPE 135523 Tight Gas Well Performance Evaluation with Neural Network Analysis for Hydraulic Propped Fracture Treatment Optimization, Paul Huckabee and Minquan Jin, Shell E&P Co., and Robert Lund, Dave Nasse and Kristin Williams, Shell Western E&P Inc. 2010, SPE ATCE, Florence, Italy, 19-22 September.
4. SAS Institute Inc. Technical Documentation: http://support.sas.com/documentation/cdl/en/emgsj/64144/PDF/default/emgsj.pdf