URTeC: 2903086

Oilfield Data Analytics: Linking Fracturing Chemistry and Well Productivity

Daria Khvostichenko*[1], Sergey Makarychev-Mikhailov[1]; 1. Schlumberger.

This paper was prepared for presentation at the Unconventional Resources Technology Conference held in Houston, Texas, USA, 23-25 July 2018.

Abstract

Hydraulic fracturing is the method of choice for well stimulation in North America. Large-scale analysis of the effects of various well stimulation parameters on production is of great interest for elucidating production trends and treatment efficacy. The objective of this work is to create a suite of tools for fracturing fluid chemistry and production data analytics based on mining and statistical analysis of data for large datasets of wells in the US. Chemistry data in FracFocus Chemical Disclosure Registry coupled with monthly production data from a commercial production database vendor were used as inputs for data analytics tools. Data parsing methods and algorithms were developed to extract information on fracturing job types, volumes of used stimulation materials, and nature of chemical additives from noisy FracFocus data. Statistical methods were employed to infer the individual and combined effects of several chemical technologies on production. The unique analysis covered the majority of wells available in FracFocus for the Eagle Ford shale and the Permian-Midland Basin and was not limited to one operator or service company. Results are presented in the form of user-friendly dashboards that enable elucidating the main factors that affect production by basin and play and on a subplay level in the US. Detailed geographical granularity was pursued to focus on the effects of completion parameters on production and to minimize the effects of reservoir quality in the analysis. Three production data metrics, namely, first full month of production (M1F) and aggregated cumulative production for first three months (C03) and first 12 months (C12), were considered. Various types of production normalization (perforated interval, proppant, etc.) in combination with proper statistical analysis tools were applied to data. The uniqueness of this work is the combination of fluid chemistry data, production data, and statistical analysis techniques and application to a large dataset of wells. On the fracturing technology side, examples include the analysis of the effects of job type (crosslinked, hybrid, or slickwater), channel fracturing techniques, and flowback additives on production. The results allow us to evaluate the most important fluid variables for effective production to enable informed decision making in asset development.

Introduction

Development of unconventional plays in North America has demonstrated remarkable growth over the last decade, governed by many technological advances in horizontal drilling and hydraulic fracturing. With the steep learning curve, the E&P industry accumulated a significant amount of various reservoir, well drilling, and completion data, which, if properly utilized, can optimize the efficiency of hydrocarbon recovery and lead to better performing and/or more economical wells in unconventional plays. Not surprisingly, oilfield data analytics is arguably the most dynamically growing area for the industry, and many operators and service companies now invest considerable resources in data science. The field of oilfield data analytics is still relatively young, and there is a long way to go before our industry can catch up with others in terms of data utilization.

The high-level challenges of oilfield data analytics are far from being unique: a) data availability, b) data quality, and c) data processing algorithms and analytical workflows. *Data availability* can be an issue, when certain information is missing, but more crucially, when operators and service providers possessing just pieces of a puzzle, struggle to collaborate on creating joint datasets to see a big picture. *Data quality* is one of hindrances in this issue, because data can be available in not easily ingestible forms (e.g., on paper or buried within proprietary software) and further contaminated with errors and artifacts. Finally, although *data processing algorithms and workflows*, both for data cleansing and analysis, have advanced significantly in recent years and reached an exceptional level of power and sophistication, proper data processing and analysis still heavily rely on in-depth application knowledge, and the selection, adoption, and validation of a data analysis methodology is labor intensive.

In our recent work (Khvostichenko and Makarychev-Mikhailov 2018), major pitfalls in data analysis were elucidated and an initial data analysis workflow, avoiding those pitfalls, was proposed. Production data from about 4,500 wells, stimulated by one service company, were analyzed, and effects of two chemical additives were carefully evaluated. The present paper is a continuation of that study, which is now extended to a significantly increased number of wells and to completion parameters, extracted from public sources, and focused on further development of the statistically solid workflow for data analysis.

Methodology

Fracturing Job Data

The FracFocus database (FracFocus 2018), managed by the Ground Water Protection Council and Interstate Oil and Gas Compact Commission, is a publicly available source of hydraulic fracturing information, with approximately 130,000 wells in the US reported since 2011. The database contains information about wells, chemical additives, and volumes used, plus some additional information. The major challenge of using FracFocus is the data quality, because numerous typos and errors in most fields require careful and often manual data cleansing and processing. Data processing algorithms were developed to identify the use of fracturing fluid components and classify job types, as shown in Fig. 1. Cross-field checks of chemical data help to improve the accuracy of assignments. Proper well data processing was only possible for wells reported in 2013 and beyond. For this period, it is estimated that 70 to 80% of all wells stimulated in the US are reported in FracFocus. Of these, about 15% of the wells still either lacked data or had poor data quality. This fraction dropped below 10% in 2017 as reporting quality improved. The total number of wells with processed treatment data was about 73,000, which is believed to be sufficient for analysis.

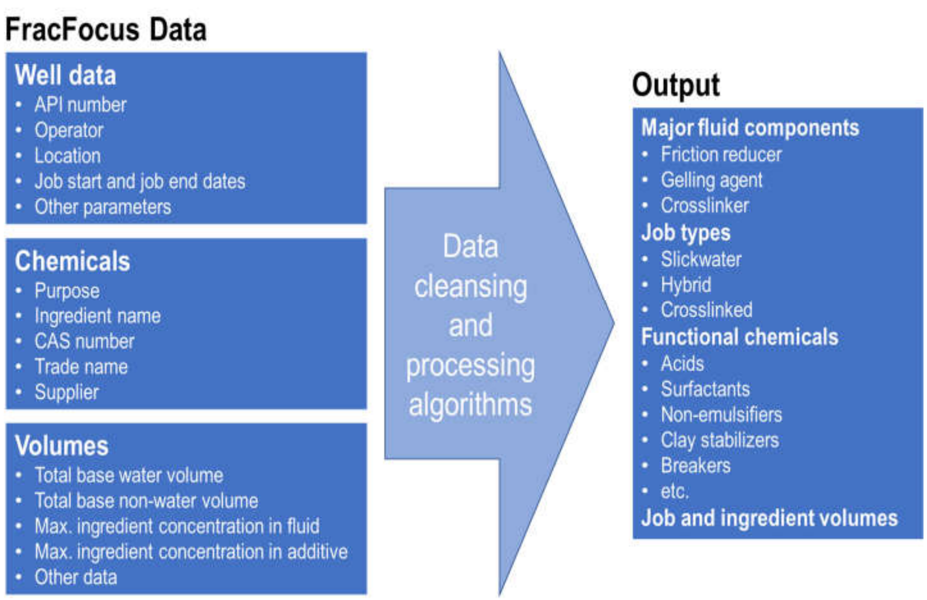

Fig. 1. Schematic of FracFocus data processing.

Completion and Production Data Selection

Production data and well data were taken from a commercial database source (IHS 2018) that compiles and verifies data retrieved from public sources. As IHS processes data using proprietary algorithms, only routine data quality checks were applied to data (e.g., removal of zero values and outliers), as described below. Production data were used either directly, or after conversion, normalization, and transformation to normal distribution (Fig. 2). Production data for first full month of production (M1F), cumulative production in the first three months (C03), and cumulative production in the first 12 months (C12) were selected for analysis to track the evolution of trends in production with respect to predictor variable effects. For wells producing both oil and gas, production was converted to barrels of oil equivalent (BOE) using 20:1 conversion factor (20 Mscf of natural gas equal to 1 bbl of oil), based on the relative cost of oil and gas.

The following variables were considered as main potential predictors for well production: location (as a proxy for reservoir quality), perforated interval length, proppant quantity, and job type. Location was typically represented by county or groups of counties and assignment is provided for each example. To analyze the effect of surfactant/nonemulsifier and fiber technologies, further subclassification by the respective product was applied, as described in respective examples. Only data for horizontal wells that were put on production in 2013–2017 with sufficient completion information were analyzed.

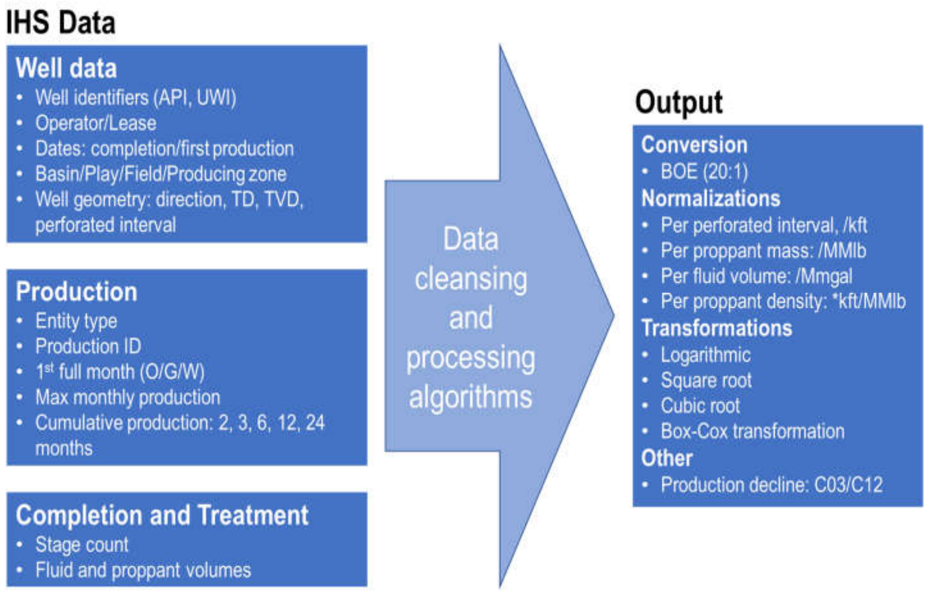

Fig. 2. Schematic of IHS data processing.

Proppant density (pounds of proppant per foot of perforated interval, lbm/ft) was selected as the proppant-related completion variable. Total proppant quantity per well was also considered but was found to correlate strongly with perforated interval length. Proppant density, however, was uncorrelated with the perforated interval length, which simplified statistical analysis.

Quantitative data from FracFocus and IHS such as proppant quantity and production are noisy, and data filtering was needed to exclude erroneous unrealistic values. For example, some proppant density values were as low as 50 lbm/ft or in excess of 8,000 lbm/ft, which is unlikely to be correct. In the analysis below, cutoffs were applied to proppant density and to perforated interval length. For proppant density, wells in the lowest 0.5% and the highest 0.5% of proppant density values were excluded from analysis for each basin. This produced the lower cutoff of proppant density of 100 to 200 lbm/ft and the upper cutoff of 4,000 to 4,400 lbm/ft. With respect to perforated interval lengths, wells under 500 ft were excluded. For production data, it was identified that even though some wells had seemingly very low or very high production values, they still were in line with the regression model obtained for the bulk of the dataset. Therefore, no cutoffs for production values were applied prior to initial regression analysis, and outliers were determined after constructing regression plots. With respect to categorical data, all wells with job type indicated as "Unknown" were excluded from the analysis because they could belong to any of the three job types investigated here and could bias results.

Statistical Analysis Workflow

A statistical analysis workflow has been developed to ensure reliable interpretation of the effects of well (e.g., perforated interval) and completion (proppant quantity, job type, and specific additive types) factors on production. The simplified schematic is shown in Fig. 3.

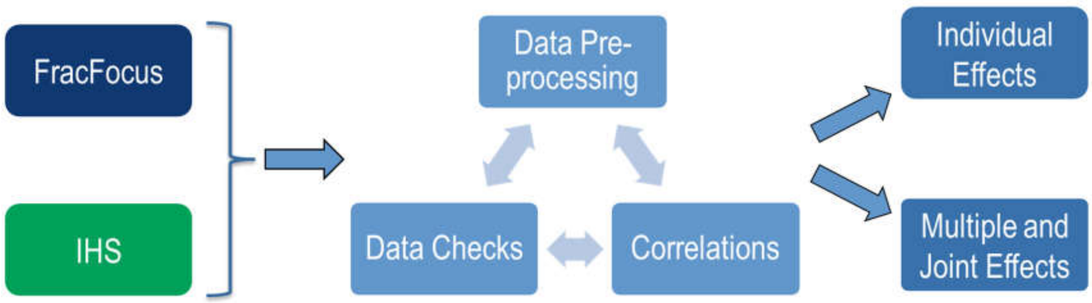

Fig. 3. Simplified schematic of the statistical analysis workflow.

Multiple regression was selected for simultaneous regression of a response variable against categorical and continuous predictor variables. The predictor and response variables selected in this work are listed in Table 1. It is fully appreciated that this relatively simple approach may not provide a comprehensive description of the complexity of factors affecting production. However, useful insights can still be gained with this approach provided certain requirements are met, as discussed in more detail below. All statistical analyses in this work were done using JMP software (V.14.0.0. SAS Institute Inc., Cary, NC).

Table 1. Response and predictor variables considered in the analysis of statistical effect of fracturing chemistries on production.

	Variable name	Variable type
Response variables	Production ("as is" or normalized)	Continuous
Predictor variables	Perforated interval	Continuous
	Proppant density	Continuous
	Job type	Categorical
	Location	Categorical
	Technology use (fibers, surfactants, etc.)	Categorical

Regression models and data transformation for analysis. Regression models were constructed to elucidate the effect of well and completion parameters on production in the geographical area of interest, such as a basin or play. Normalization of production data by perforated interval length and by proppant density in regression models was also explored.

Forms of regression models investigated in this work are presented in Eqs. 1–3:

$$\mathbf{log_{10}(Production)} = \boldsymbol{a_0} + \boldsymbol{a_{PL}} \cdot \mathbf{PL} + \boldsymbol{a_{PD}} \cdot \mathbf{PD} + \boldsymbol{a_{Loc}} \cdot \mathbf{Loc} + \boldsymbol{a_{JT}} \cdot \mathbf{JT} + \boldsymbol{a_{Loc \cdot JT}} \cdot (\mathbf{Loc} \cdot \mathbf{JT}) \quad (1)$$

$$\mathbf{log_{10}(Production/PL)} = \boldsymbol{a_0} + \boldsymbol{a_{PD}} \cdot \mathbf{PD} + \boldsymbol{a_{Loc}} \cdot \mathbf{Loc} + \boldsymbol{a_{JT}} \cdot \mathbf{JT} + \boldsymbol{a_{Loc \cdot JT}} \cdot (\mathbf{Loc} \cdot \mathbf{JT}) \quad (2)$$

$$\mathbf{log_{10}}(\boldsymbol{Production/PD}) = \boldsymbol{a_0} + \boldsymbol{a_{PL}} \cdot \mathbf{PL} + \boldsymbol{a_{Loc}} \cdot \mathbf{Loc} + \boldsymbol{a_{JT}} \cdot \mathbf{JT} + \boldsymbol{a_{Loc \cdot JT}} \cdot (\mathbf{Loc} \cdot \mathbf{JT}) \quad (3)$$

where PL, PD, Loc, and JT stand for perforated length, proppant density, location, and job type, respectively; a_{PL}, a_{PD}, a_{Loc}, and $a_{Loc \cdot JT}$ are regression coefficients or "effects" associated with each predictor variable; and a_0 is the value of the intercept. As can be noted from Eqs. 1–3, all models contained an interaction term (Loc · JT) in addition to separate location and job type effects. The logarithmic transformation of production, rather than production "as is", was used in the regression. These are explained in detail below.

Model description. All models in Eqs. 1–3 contain terms to describe the response of the production variable to continuous predictor variables PL and PD, categorical predictor variables Loc and JT, and an interaction term for location and job type. The interaction term for location and job type was necessary to elucidate location-specific effects of job types. In a given location, relative performance of job types is determined by the sum of terms [$a_{JT} \cdot$ JT + $a_{Loc \cdot JT} \cdot$(Loc · JT)]. In the absence of this term, only the average ranking of job types for the entire dataset can be inferred in the analysis, even though local results may be drastically different from the average ranking. For example, if the interaction terms $a_{Loc \cdot JT}$ are sufficiently large, it may change the direction of job type effects in the location relative to the average for the data set. The absence of interaction terms for PL and PD variables implies the assumption that the effects of perforated interval length and proppant density are independent of location and job type.

The output of regression model fitting is the values of coefficients and their *p*-values for each term in the model, including all values of categorical variables, and all combinations of values of categorical variables in the interaction term. The *p*-values are related to the uncertainty of determining the coefficients from noisy data and signify the probability that a given predictor variable has no effect on the response variable. Therefore, a small *p*-value indicates high probability that a given variable is important for production.

Three layers of interpretation are possible based on the output of regression models:

(1) A quick summary of *p*-values for the effect on production for each predictor variable in general (i.e., JT, Loc, JT · Loc) provides a quick summary of the relative importance of different predictor variables on the response variable.
(2) For a more detailed examination, a detailed regression report provides regression coefficients for each value of each variable (for example, for each job type) and for each combination in the interaction effects, and associated *p*-values are calculated in regression analysis. Continuous variables only have one associated regression coefficient and *p*-value that is the same as in the quick summary described in (1) above. However, for categorical variables, each value of the variable (for example, each job type), has an associated regression coefficient and a corresponding *p*-value that is distinct from those in the quick summary report. Even though the predictor variable overall can have a small *p*-value in the quick summary, it may be due to the large deviation of just one group in the categorical variable from the rest of the groups. Detailed information about each group provided in the regression report is of interest in certain cases.
(3) Of particular interest in this work are comparisons between specific groups such as jobs of different types in a given location to identify whether the differences between them are statistically significant. This is not directly related to *p*-values in levels (1) and (2) above and is achieved via post hoc pairwise comparison tests that take into account data scatter and number of observations in each group. The Tukey "honestly significant difference" test for post hoc comparisons was used here. The value of $\alpha = 0.1$ was selected as a cutoff for statistical significance.

Only overall *p*-values for predictor variables from (1), regression coefficients for continuous variables from (2), and results of pairwise comparison tests from (3) are reported in the analyses below. The large number of variables and *p*-values in (2) preclude their reporting in the manuscript in the interest of brevity.

Each model in Eqs. 1–3 was tested for each example below. The predictive power of the models was inferred from the correlation coefficient R^2. Only results for the model with the highest value of R^2 are discussed in each example. The model with scaling by proppant density (Eq. 3) had a consistently lower R^2 than the other two models and is not presented in any of the examples.

Data transformation. The logarithmic transformation of production was applied to meet the requirement of the homogeneity of variance required for statistical interpretation of linear regression results. In simple terms, this means that the spread of production values should be similar in magnitude regardless of the values of the predictor variables. However, production data are known to be distributed log-normally (Beliveau 1995; O'Sullivan and Montgomery 2015), and this assumption for production "as is" does not hold. Although production does show some correlation with predictor variables such as PL and PD, the spread of production data tends to increase with increasing values of PL and PD, as shown for C03 plotted as a function of PL for the Spraberry zone in the Permian basin (Fig. 4, left). Using the logarithm of production data rectifies this issue (Fig. 4, right).

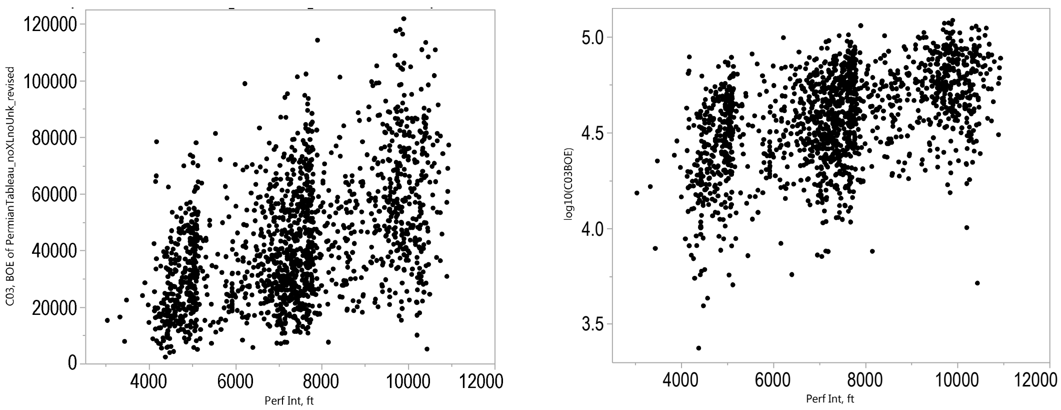

Fig. 4. Production C03 BOE (left) and $\log_{10}$(C03 BOE) for the Spraberry producing zone of the Permian-Midland basin as a function of perforated interval length.

When normalizing production by perforated length or by proppant density, the respective terms were removed from predictor variables on the left-hand side of Eqs. 2 and 3, respectively. The logarithmic transformation, however, was still needed to meet the requirement on the homogeneity of variance. It should be noted that with the logarithmic transformation, quantitative interpretation of results is straightforward. A specific incremental increase in the predictor variables corresponds to a specific percent change in production that is determined by the value of the regression coefficient *a* of the respective effect. Other transformations were also tested, following the Box-Cox methodology. Although in certain cases the square root (λ=0.5) and the cubic root (λ=0.33) transformations provided better normality and homogeneity of variance than the logarithmic one, interpretation of thus transformed data was cumbersome, and therefore logarithmic transformation was selected.

Results and Discussion

Data Visualization

Dealing with complex multidimensional datasets is always laborious. A series of dashboards were built to simplify this task for preliminary assessment of major trends and selection of subsets of data for further statistical analysis. As an example, an overview dashboard in Fig. 5 shows operator activity in the Eagle Ford shale and wells stimulated with major job types and with or without surfactants/nonemulsifiers. Applying filters on the right allows us to change the scope to other basins or plays, or smaller geographical areas, timeframes, etc.

High-level trends in well completion and production can be further quickly assessed using the next dashboard (Fig. 6). Historical data for horizontal wells in the Eagle Ford are shown for the variety of completion parameters. Non-normalized and normalized production M1F and C12 are further shown. As evident from the figure, over time, operators have drilled longer laterals, as shown by the increase in median lateral lengths in the Eagle Ford from ~5,150 ft to 6,200 ft between 2013 and 2017. Relatively stable values of true vertical depths and reservoir fluid properties (median gas-to-oil ratio, GOR, and water-to-oil ratio, WOR), indicate that no significant changes occurred in drilling practices or in the reservoir. However, volumes of stimulation materials dramatically increased. For example, median proppant mass per well more than doubled in 5 years. Apparently, the increased proppant density per perforated interval is due to the combined effect of the increased number of stages and bigger jobs. There are obviously other factors and reasons, not reflected on the dashboard, which can explain the observed production trends, for example, the increase in the last years of infill wells in the Eagle Ford, which are usually poorer producers (in normalized production) compared to parent wells (Lindsay et al. 2018).

Median well production demonstrates steady growth over time, although it becomes less pronounced for production normalized per perforated interval (BOE/thousand feet). Furthermore, production normalized per mass of proppant (BOE/million pounds denoted as BOE/MMlb) or per proppant density per perforated interval (BOE·thousand feet/million pounds, denoted as BOE·kft/MMlb) show negative trends. The current practice of increasing proppant density (and volumes per well) is probably not sustainable in the long run because returns from stimulation treatments are diminishing with larger volumes. More detailed analysis on proppant volumes and recent trends in different basins and plays is reported elsewhere (Srinivasan et al. 2018).

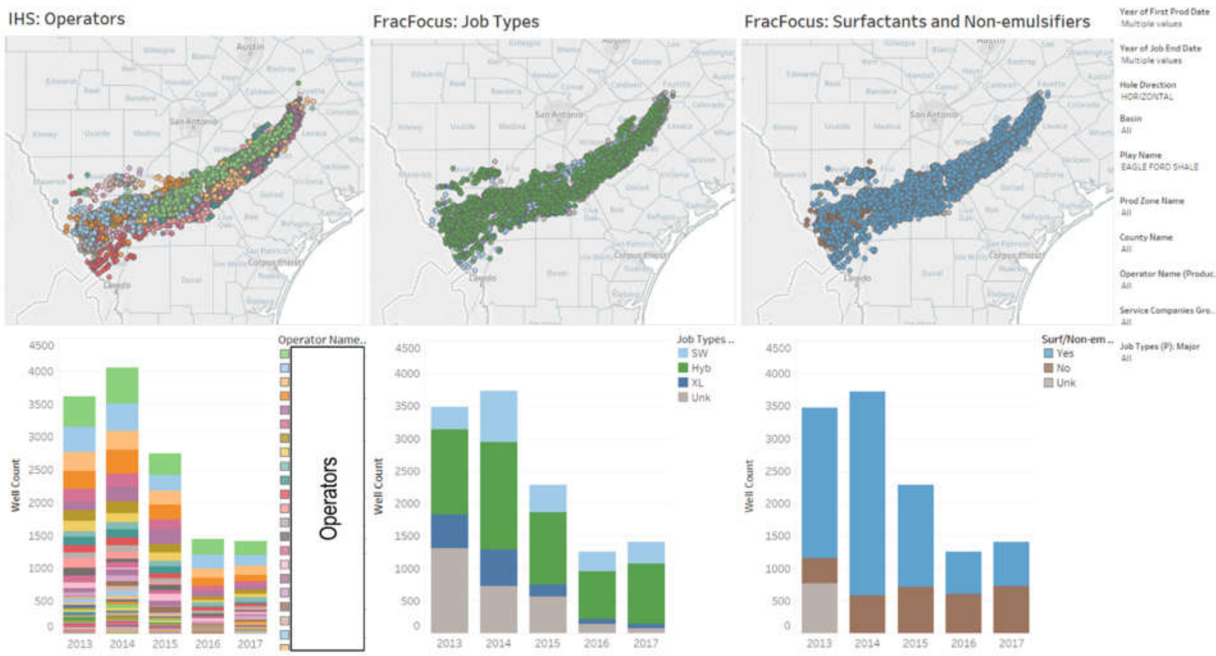

Fig. 5. Dashboard: basin/play/technology.

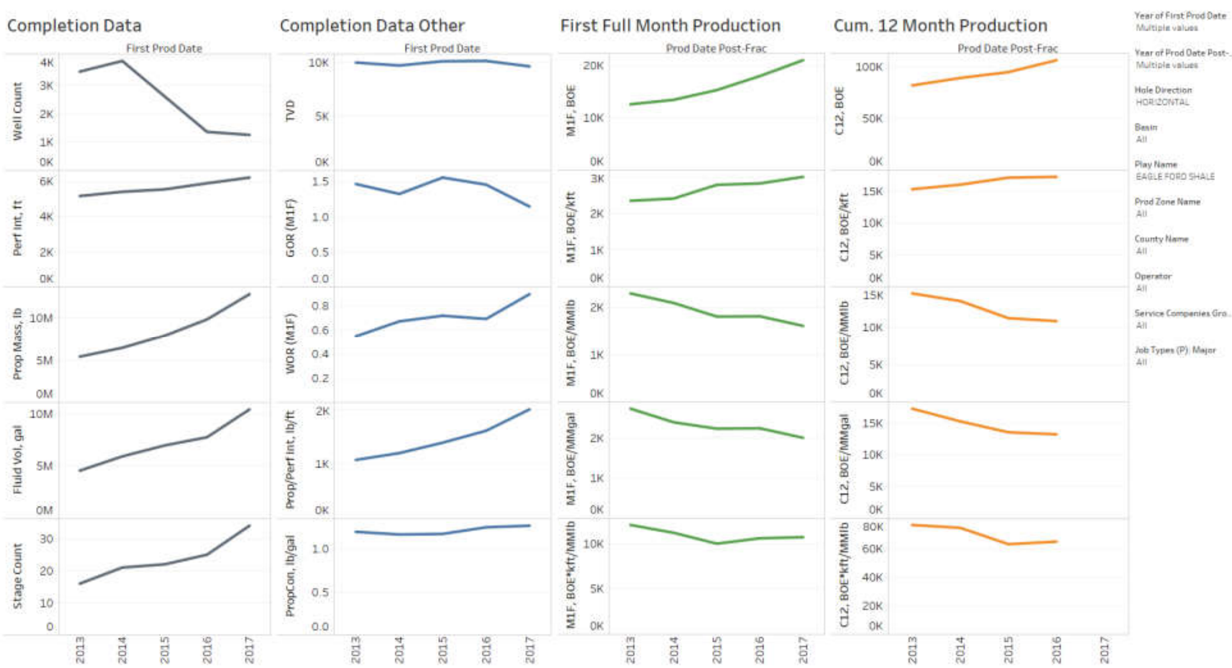

Fig. 6. Dashboard: general trends in completion and production (median values).

It must be emphasized that although the dashboards are a source of valuable high-level insights, they should be treated with care. Generalizations at the play level can be misleading. Assessment of individual factors in the simplest way, as on this intuitive dashboard, generally requires fixing all other parameters, which is difficult in geographically vast and heterogeneous formations. Selecting smaller datasets (e.g., confined by a geographical area such as county or *n*-

mile radius) can improve veracity of certain inferences. However, rigorous statistical analysis is required to make solid conclusions.

Statistical Analysis

The Eagle Ford shale and the Permian basin, the two major unconventional plays in the US, were selected for in-depth analyses beyond general trends. The below examples demonstrate the value of using a systematic approach and the developed statistical workflow. With the automated processing algorithms, analyses of the effects of many factors on production becomes straightforward and can be easily extended to other basins, plays, and subplays where sufficient completion and production data are available.

Example 1. Effect of different job types in Eagle Ford shale

The Eagle Ford shale is one of the most important hydrocarbon-producing formations in the lower 48 US states. The main producing area of the Eagle Ford shale is approximately 200 miles long, stretching from north of Gonzales down to Webb County in Texas with a total distance of approximately 200 miles (Fig. **7**). Oil, condensate, and gas are produced in the Eagle Ford shale depending on location.

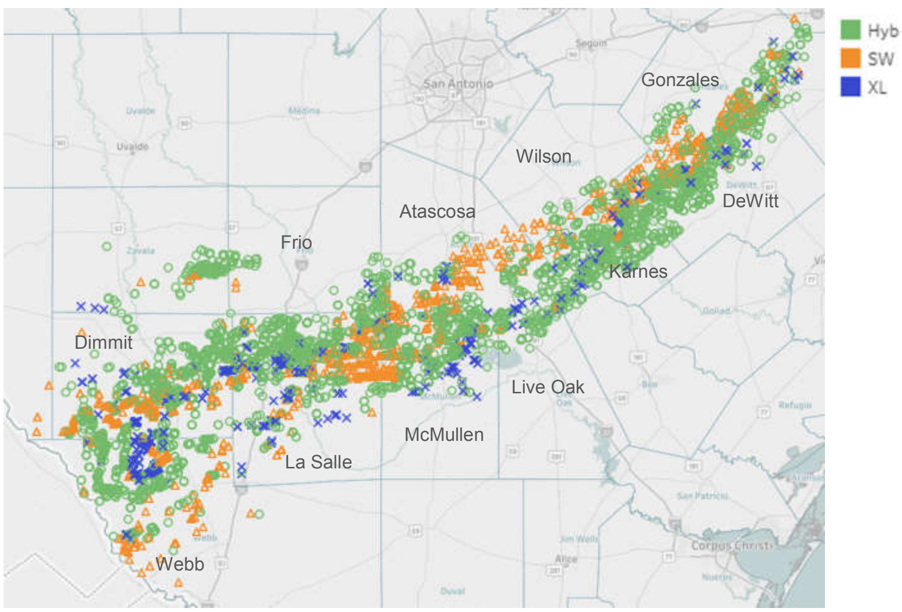

Fig. 7. Locations of different job types in the Eagle Ford shale.

Because of the large distance span of the Eagle Ford shale and significant regional disparities in average productivity, for analysis, the area was split into several locations in the east-west direction based on groups of counties: DeWitt-Gonzales, Karnes-Wilson, McMullen, La Salle, and Dimmit-Webb. All three main job types considered, namely, slickwater (SW), hybrid (Hyb), and crosslinked gel (XL), were well represented in the Eagle Ford. However, several counties were not included in the analysis because they did not have enough representation of all three job types. Even when combined with other counties, there were large disparities in the location of different job types. Jobs using channel fracturing (fiber) technology are analyzed in subsequent sections and were excluded from the analysis in this section.

The model with production scaled by perforated interval as the response variable (Eq. 2) provided the highest value of R^2 for the Eagle Ford shale meaning it had the highest predictive power. This model was used to elucidate effects of other variables on production. Results are presented and discussed below with detailed interpretation.

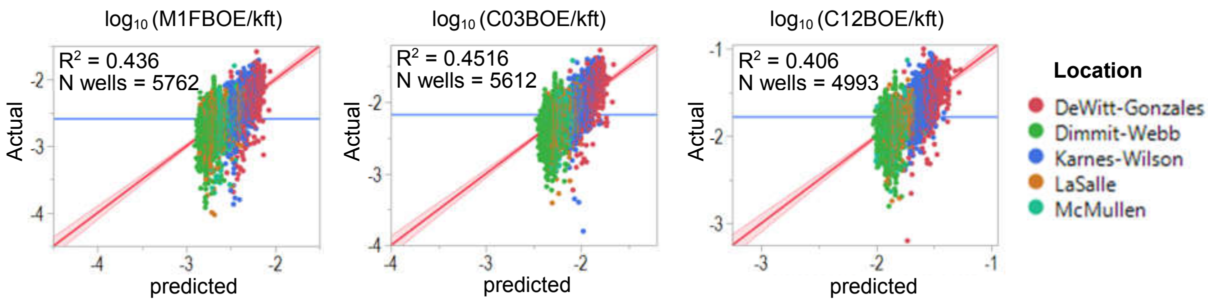

Fig. 8. Regression plots for production analysis in the Eagle Ford shale.

Table 2. *p*-values for regression model variables for production by job type in Eagle Ford shale.

Term	$\log_{10}$(M1FBOE/kft)		$\log_{10}$(C03BOE/kft)		$\log_{10}$(C12BOE/kft)	
	Regression coefficient	*p*-value	Regression coefficient	*p*-value	Regression coefficient	*p*-value
Location		<.0001*		<.0001*		<.0001*
Job type		<.0001*		<.0001*		0.0051*
Proppant density, lbm/ft	0.000121	<.0001*	0.000112	<.0001*	0.000101	<.0001*
Location*job type		<.0001*		<.0001*		<.0001*

Overall fit results are shown in Fig. 8. For all response metrics (M1F, C03, and C12), all predictor variables in the model had a statistically significant nonzero effect on production, as indicated by their low *p*-values (Table 2). Table 2 also reports the regression coefficients for proppant density, which tend to be small, indicating a relatively low effect of proppant loading on production. The values of regression coefficients in Table 2 correspond to 2.5% to 2.8% increase in average production for every additional 100 lbm proppant per foot.

Table 3 summarizes calculated relative effects of each job type on production (ΔProduction, in percent) as well as outcomes of pairwise comparison tests to elucidate which effects are important in a given location. ΔProduction (%) is relative to average for a given location. For example, for the M1FBOE/kft metric, average production from XL jobs in LaSalle County was 13.7% higher than average production from all job types in LaSalle County.

Table 3. Ranking of job types with respect to effect on production in different locations in the Eagle Ford shale.

Location	Job Type	$\log_{10}$(M1FBOE/kft)		$\log_{10}$(C03BOE/kft)		$\log_{10}$(C12BOE/kft)	
		ΔProduction %	Pairwise comparison	ΔProduction %	Statistical comparison	ΔProduction %	Statistical comparison
DeWitt-Gonzales	Hyb	+11.6	Hyb ≈ SW ≈ XL	+11.2	Hyb > SW XL ≈ Hyb, SW	+20.1	Hyb > SW, XL
	SW	+8.8		-1.8		-6.7	
	XL	-17.6		-8.5		-10.7	
Karnes-Wilson	Hyb	-4.5	Hyb ≈ SW ≈ XL	+0.4	Hyb ≈ SW ≈ XL	+3.1	Hyb ≈ SW ≈ XL
	SW	+4.2		-3.5		-8.4	
	XL	+0.5		+3.2		+5.9	
McMullen	Hyb	-0.7	SW > XL Hyb ≈ SW, XL	+0.0	SW > Hyb, XL	+1.9	SW ≈ Hyb SW > XL
	SW	+17.8		+17.5		+12.3	
	XL	-14.5		-14.9		-12.6	
La Salle	Hyb	-28.7	SW ≈ XL SW, XL > Hyb	-17.5	SW, XL > Hyb	-5.4	SW > Hyb XL ≈ Hyb, SW
	SW	+23.4		+17.7		+8.1	
	XL	+13.7		+2.9		-2.2	
Dimmit-Webb	Hyb	-5.9	XL > Hyb SW ≈ Hyb, XL	-5.2	SW > Hyb XL ≈ SW, Hyb	-2.7	SW > Hyb, XL
	SW	-2.3		+3.1		+7.6	
	XL	+8.8		+2.3		-4.5	

Results in Table 3 highlight three important points:

1. Relatively large differences in average production in a given location do not necessarily translate into statistical significance. Statistical significance also depends on other factors, such as data variation (standard deviation) and number of observations. For example, in the DeWitt-Gonzales location, average production

for Hyb and XL job types differed by almost 30%. However, based on the pairwise comparison test, the differences among all three job types were not statistically significant, presumably due to the large spread of data. A similar difference in average production between SW and XL jobs in McMullen County, on the other hand, is statistically significant.

2. With three categories or more, it may not be possible to pick an "absolute" winner or to conclude that all treatments are the same. For example, in the DeWitt-Gonzales location for the C03BOE/kft metric, Hyb jobs outperformed SW jobs. However, there were no statistically significant differences in production from XL jobs compared to either Hyb or SW jobs.
3. Trends in performance of job types may change depending on the selected production metric. For example, Hyb jobs emerged as highest performers in the DeWitt-Gonzales location for C12BOE/kft, whereas for M1FBOE/kft there was no statistically significant differences between different job types.

Outliers required a specific approach. Fig. 9 shows regression plots based on Eq. 2 for various production metrics in the Eagle Ford with outliers for the $\log_{10}$ (M1FBOE/kft) metric highlighted in all plots. These outliers were excluded in the regression analysis above. As seen in Fig. 9, many points that appeared to be outliers in the M1F production-related metric fell in line with the bulk of the dataset for the C03 metric and even more so for the C12 production-related metric. The M1F outliers were consistently on the low side of production, and only a few cases in all analyses in this work were found with high production outliers. This indicates that the very low production reported for these outliers was unlikely to be related to the performance of the fracturing treatment. The more likely reasons for these outliers could be externally choked production and unreported or incorrectly reported production. The outliers had a marked effect on the values of the correlation coefficient R^2 calculated for the model but did not significantly affect pairwise comparisons of effects of job types, most likely due to the large total number of points in the dataset.

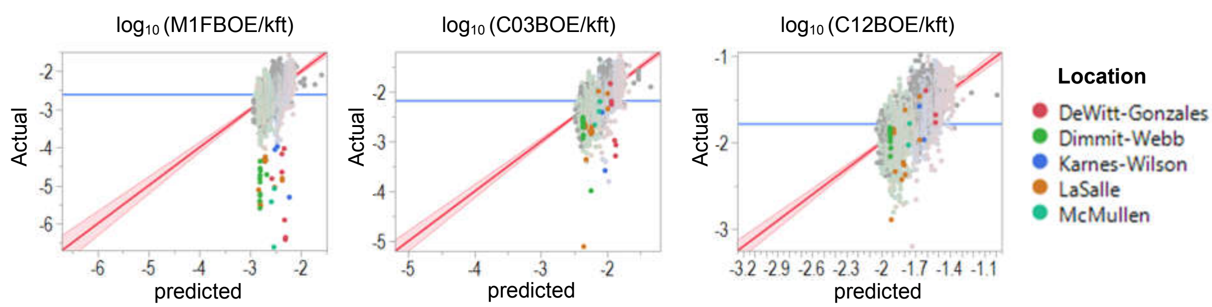

Fig. 9. Outliers in regression analysis of production in the Eagle Ford shale.

In summary, results of statistical analysis in this section demonstrate that depending on location, different job types may have different performance in the Eagle Ford shale. Hybrid treatments appear to have the highest benefit in the eastern section, whereas slickwater treatments appear to have the highest benefit in the western section. However, further analysis may be required to fully elucidate the effects of reservoir quality on production.

Example 2. Effect of surfactants/nonemulsifiers in Eagle Ford

Surfactants and nonemulsifiers are added to fracturing fluids to improve resultant production. The additives prevent damage to permeability due to water and emulsion blocks and facilitate fracture and matrix cleanup and it is of interest to assess their effectiveness. A data set of wells that differ by surfactant usage (with/without) for each job type was required for this analysis because different job types may have different productivity. The vast majority of wells in the Eagle Ford shale, however, were completed with the use of surfactants and/or nonemulsifiers, and only the SW job type in several counties was found to have a representative number of cases with "yes/no" surfactant usage. For statistical analysis of surfactant effects, the model of Eq. 2 was augmented:

$$\mathbf{log_{10}(Production/PL)} = \boldsymbol{a_0} + \boldsymbol{a_{PD}} \cdot \mathbf{PD} + \boldsymbol{a_{Loc}} \cdot \mathbf{Loc} + \boldsymbol{a_{Surf}} \cdot \mathbf{Surf} + \boldsymbol{a_{Loc \cdot Surf}} \cdot (\mathbf{Loc} \cdot \mathbf{Surf}) \quad (4)$$

where $a_{Surf} \cdot$ Surf and $a_{Loc \cdot Surf} \cdot$ (Loc $\cdot$ Surf) are, respectively, the terms related to surfactant/non-emulsifier usage (with/without) and the term for interaction effects to describe potentially different action of surfactants in different

10.15530/urtec-2018-2903086

locations. Location was represented by county, rather than groups of counties. The models with $\log_{10}$(Production) and $\log_{10}$(Production/PD) as response variables were also explored but produced lower values of the correlation coefficient compared to the model of Eq. 4.

Table 4. *p*-values for the regression model for surfactant effect on production in Eagle Ford shale.

Term	$\log_{10}$(M1FBOE/kft)		$\log_{10}$(C03BOE/kft)		$\log_{10}$(C12BOE/kft)	
	Regression coefficient	p-value	Regression coefficient	p-value	Regression coefficient	p-value
Location		<.0001*		<.0001*		<.0001*
Proppant density, lbm/ft	0.000228	<.0001*	0.000259	<.0001*	0.000242	<.0001*
Surfactant use (with/without)		0.0009*		<.0001*		<.0001*
Location · surfactant use		0.0153*		0.0111*		0.0805

Results of statistical analysis of the effect of surfactant/nonemulsifier are presented in Table 4 and Table 5. As per the *p*-values in Table 4, all predictor variables selected for analysis were statistically significant at the $\alpha = 0.1$ level. The interaction term (location · surfactant use) had the highest *p*-value, which means it had the lowest significance in the model. As per detailed regression report of calculated effects and corresponding *p*-values (not shown), the interaction term was statistically significant only for two counties, Atascosa and La Salle.

Calculated overall effects of surfactant in each county and corresponding statistical significance are reported in Table 5. Production improvement ΔProduction in Table 5 was calculated as the difference between "With surfactant" and "Without surfactant". As seen previously, apparent large differences in average production were not necessarily statistically significant, as, for example, for the M1F metric in McMullen County. Typically, the use of surfactants and nonemulsifiers was associated with higher production, and the effect became more evident when moving from M1F to C03 and C12 production metrics.

Table 5. Calculated effects of surfactant/nonemulsifier in SW jobs in various counties in Eagle Ford shale.

Location	$\log_{10}$(M1FBOE/kft)		$\log_{10}$(C03BOE/kft)		$\log_{10}$(C12BOE/kft)	
	ΔProduction, With surfactant vs. without surfactant, %	Statistically significant?	ΔProduction, With surfactant vs. without surfactant, %	Statistically significant?	ΔProduction, With surfactant vs. without surfactant, %	Statistically significant?
Gonzales	+7.2	No	+16.8	**Yes**	+18.4	**Yes**
Karnes	+7.0	No	+21.2	**Yes**	+12.9	No
Atascosa	-1.8	No	7.1	No	+4.1	No
McMullen	+32.1	No	+19.5	No	+23.0	**Yes**
La Salle	+37.4	**Yes**	+36.6	**Yes**	+27.9	**Yes**

Example 3. Channel fracturing in Eagle Ford shale

Channel fracturing is a technology for fiber-enabled heterogeneous proppant placement that creates fractures with infinite conductivity (Gillard et al. 2010). The benefit of this technology has been reviewed in multiple case studies. Here, statistical analysis was applied to a section of the Eagle Ford shale to elucidate the effects of channel fracturing by incorporating corresponding effects in the regression model.

Inspection of locations of fiber-based jobs revealed that they were concentrated in De Witt and Karnes counties and were mostly represented in Hyb and XL jobs. Furthermore, there were distinct preferences for a specific job type in each county, with De Witt dominated by Hyb jobs and Karnes dominated by XL jobs. One of key requirements for the statistical model used in previous sections is representation of the technology in all job types and in all locations. Based on the availability of data for analysis, the terms for location, job type, and (location · job type) were omitted, and a term to account for effect of channel fracturing (fibers) was included in the model. The example presented below is for Hyb jobs in De Witt County. An identical analysis may be performed for XL jobs in Karnes County. The model for analysis had the form

$$\log_{10}(\text{Production}) = a_0 + a_{PL} \cdot \text{PL} + a_{PD} \cdot \text{PD} + a_F \cdot \text{F} \quad (5)$$

10.15530/urtec-2018-2903086

where $a_F \cdot$ F is the term to account for the effect of the fiber technology. Results are summarized in Table 6. These are also reported in Table 6. Statistically significant effects were associated with all model variables.

Table 6. Regression coefficients and p-values for the effect of channel fracturing in the Eagle Ford shale for hybrid jobs in DeWitt County.

Term	$\log_{10}$(M1FBOE)		$\log_{10}$(C3FBOE)		$\log_{10}$(C12BOE)	
	Regression coefficient	p-value	Regression coefficient	p-value	Regression coefficient	p-value
Intercept	4.073	<.0001*	4.453	<.0001*	4.814	<.0001*
Perf Int, 1,000 ft	0.0440	<.0001*	0.0558	<.0001*	0.0675	<.0001*
Proppant density, lbm/ft	0.000125	<.0001*	0.000108	<.0001*	0.000075	<.0001*
No fibers	-0.1100	<.0001*	-0.0814	<.0001*	-0.0255	0.0064*
With fibers	+0.1100	<.0001*	+0.0814	<.0001*	+0.0255	0.0064*

Beneficial effects were associated with channel fracturing in all three production metrics, with the greatest differences in the initial months of production. This is evidenced by the values of regression coefficient related to use of fibers in Table 6. Production for wells completed with channel fracturing was over 50% higher compared to wells with regular completions for the M1F metric.

As observed earlier, the effect of proppant density on production was relatively small and of the same order of magnitude as calculated for all-Eagle Ford analysis. For perforated length, incremental production increase for 1000 ft of lateral was approximately 11% to 16%, depending on the production metric. Note that the regression coefficient for perforated interval length reported in Table 6 is calculated per 1000 ft of interval.

Example 4. Production in the Permian-Midland basin

The Permian basin is one of the oldest-producing oil basins in the US. It spans west Texas and southeast New Mexico. The main producing subdivisions in the Permian are the Midland and the Delaware sub-basins. The stratigraphy of the Permian basin is highly complex, and various plays can be producing sources within a sub-basin.

According to the IHS producing zone assignment, from 2013 to 2017, the prevalent producing zones for horizontal wells in the Midland and the Delaware basins were Spraberry and Wolfcamp, respectively. The Spraberry producing zone was selected as an example for the analysis of job type effects in the Permian basin. The XL job type was poorly represented in the Permian basin and the analysis was limited to the effects of SW and Hyb job types.

The majority of wells producing from Spraberry were located in six Texas counties (Fig. 10)—Glasscock, Howard, Martin, Midland, Reagan, and Upton—which were used as proxies for the effect of location.

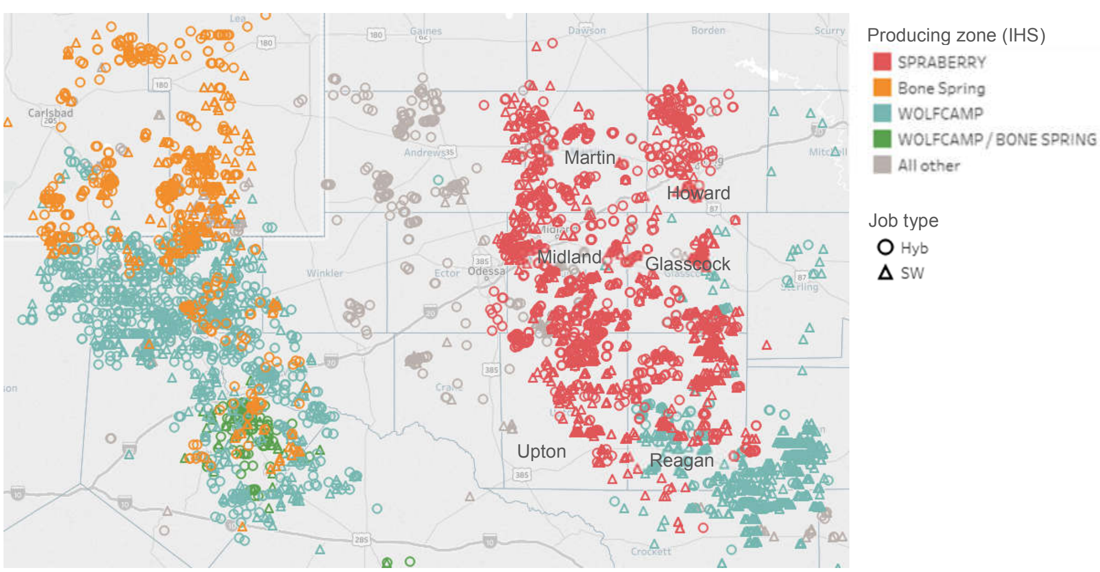

Fig. 10. Various producing zones in the Permian basin.

The model of Eq. 1 produced the best correlation coefficient for this dataset, and it was used for production analysis.

Table 7. *p*-values and regression coefficients for production analysis in the Spraberry producing zone of the Permian-Midland basin.

Term	$\log_{10}$ (M1FBOE)		$\log_{10}$ (C03BOE)		$\log_{10}$ (C12BOE)	
	Regression coefficient	*p*-value	Regression coefficient	*p*-value	Regression coefficient	*p*-value
Location		<.0001*		<.0001*		<.0001*
Job Type		<.0001*		0.0051*		0.0460*
Perforated interval, kft	0.0524	<.0001*	0.0629	<.0001*	0.0743	<.0001*
Proppant density, lbm/ft	0.000129	<.0001*	0.000145	<.0001*	0.000131	<.0001*
Location · Job Type		<.0001*		<.0001*		0.0080*

Table 8. Calculated effects of Hyb and SW treatments for the Spraberry producing zone.

Location	$\log_{10}$ (M1FBOE)		$\log_{10}$ (C03BOE)		$\log_{10}$ (C12BOE)	
	ΔProduction, Hybrid vs. SW, %	Statistically significant?	ΔProduction, Hybrid vs. SW, %	Statistically significant?	ΔProduction, Hybrid vs. SW, %	Statistically significant?
Martin	4.6	No	-4.9	No	0.6	No
Howard	50.3	**Yes**	19.0	No	3.2	No
Midland	-12.6	No	-10.6	No	-6.4	No
Glasscock	36.1	**Yes**	20.1	No	21.1	**Yes**
Upton	10.5	No	6.6	No	3.8	No
Reagan	27.4	No	22.6	**Yes**	13.1	No

All model variables had a statistically significant effect on production in the Spraberry producing zone as evidenced by low *p*-values (Table 7). Results of pairwise comparisons for the two job types are shown in Table 8.

Pairwise comparison tests of the performance of different job types in various locations showed that in the majority of cases, the difference between SW and Hyb was not statistically significant. Locations (counties) with statistically significant differences were different for different metrics. Overall, production in counties in the eastern section of the Spraberry producing zone (Howard, Glasscock, and Reagan), appeared more sensitive to the job type, with Hyb jobs outperforming SW jobs.

Conclusions

A methodology for processing fracturing treatment information available from a public source (FracFocus 2018) was developed. Based on data cleansing and processing algorithms, high-level job types and specific classes of chemical additives used for well stimulation were assigned. Well production data from a commercial database (IHS 2018) were processed by applying several carefully selected data conversion, normalization, and transformation operations and outlier detection techniques and coupled with chemistry data. Based on the processed data, several user-friendly and flexible dashboards were designed to quickly assess high-level trends in well completion and production in various basins and plays.

A rigorous statistical analysis workflow was developed and tested on the Eagle Ford shale and the Permian basin datasets, quantifying the effects of several main factors, as well geographical location, perforated interval, proppant mass, fracturing job type, and utilization of fracturing technologies on various production metrics. The analysis relied on multiple linear regression models and involved calculation of statistical significance for effects of particular factors, expressed through *p*-values, and for pairwise comparisons via specific post hoc pairwise comparison tests.

The Eagle Ford shale had the largest number of wells amenable for analysis of the effect of chemistry, although grouping of wells in different counties was required for analysis in several cases. With respect to job type, location-specific performance differences were found, with Hyb treatments preferable in the eastern section and SW treatments in the western section. Channel fracturing was associated with production increase for production from M1F through C12, as shown for hybrid jobs in DeWitt County. The effect of surfactants/nonemulsifiers was only investigated for SW jobs due to lack of data for other treatment types; the analysis showed that the effect of these additives was neutral or beneficial, depending on location.

In the Permian basin, the analysis was limited to SW and Hyb jobs in the Spraberry producing zone in the Permian-Midland section. Production in counties in the eastern section of the Spraberry producing zone (Howard, Glasscock, and Reagan), appeared more sensitive to the job type, with Hyb jobs outperforming SW jobs.

As the industry faces new scrutiny with respect to the of use publicly available data to claim improved production upon use of chemical products (Addison 2016), the need for a fully transparent, statistically solid, and reproducible data analysis workflow is obvious. The workflow under development, presented here, is a significant advancement in this direction, although further work is required to a) complement available datasets with more data, especially related to reservoir quality; b) further develop the classical statistical analysis methodology; c) expand utilization of modern machine learning and artificial intelligence algorithms in production analysis. These topics will be a subject of future publications.

References

Addison, V. 2016. Flotek Releases Report Defending Frack Fluid Technology. *E&P Magazine,* January 28. https://www.epmag.com/flotek-releases-report-defending-frack-fluid-technology-836876.

Beliveau, D., 1995, Heterogeneity, Geostatistics, Horizontal Wells, and Blackjack Poker: *J. Pet Technol* **47** (12): 1068–1074. SPE-30745-PA. https://doi.org/10.2118/30745-PA.

FracFocus 2018. FracFocus National Hydraulic Fracturing Chemical Registry. http://fracfocus.org/ (accessed 16 March 2018).

Gillard, M. R., Medvedev, O. O., Hosein, P. R. et al. 2010. A New Approach to Generating Fracture Conductivity. Presented at SPE Annual Technical Conference and Exhibition, Florence, Italy, 19-22 September. SPE-135034-MS. https://doi.org/10.2118/135034-MS.

IHS 2018. IHS Enerdeq Browser, IHS Markit, https://penerdeq.ihsenergy.com/thin2/secure/home/home.jsf (accessed 18 March 2018).

Khvostichenko, D. and Makarychev-Mikhailov, S. 2018. Effect of Fracturing Chemicals on Well Productivity: Avoiding Pitfalls in Big Data Analysis. Presented at SPE International Conference and Exhibition on Formation Damage Control, Lafayette, Louisiana, USA, 7–9 February. SPE-189551-MS. https://doi.org/10.2118/189551-MS.

O'Sullivan, F. M. and Montgomery, J. B., 2015. The Salient Distribution of Unconventional Oil and Gas Well Productivity: MIT Energy Initiative Working Paper, Cambridge, Massachusetts, USA. https://energy.mit.edu/wp-content/uploads/2015/12/MITEI-WP-2015-05.pdf (accessed 14 April 2018).

Lindsay, G., Miller, G., Xu, T. et al. 2018. Production Performance of Infill Horizontal Wells vs. Pre-Existing Wells in the Major US Unconventional Basins. Presented at the SPE Hydraulic Fracturing Technology Conference and Exhibition, The Woodlands, Texas, USA, 23-25 January 2018. SPE-189875-MS. https://doi.org/10.2118/189875-MS.

Srinivasan, K., MacKay, B., Alimahomed, F. et al. 2018. Is There Anything Called Too Much Proppant? Accepted for SPE Liquid-Rich Basin Conference – North America, Midland, Texas, USA, 5-6 September. SPE-191800.