SPE-179958-MS

Detecting and Removing Outliers in Production Data to Enhance Production Forecasting

Nitinkumar L. Chaudhary, University of Houston; W. John Lee, Texas A&M University

This paper was prepared for presentation at the SPE/IAEE Hydrocarbon Economics and Evaluation Symposium held in Houston, Texas, USA, 17–18 May 2016.

Abstract

The objective of the study summarized in this paper was to establish a robust method to detect outliers in rate and pressure data used for decline analysis and forecasting, pressure and rate transient analysis, and similar workflows. Presence of outliers contributes significantly to the uncertainty and non-uniqueness associated with the rate/pressure transient analysis workflow.

In time-series data usually encountered in rate/pressure transient analysis workflows, we observed that the neighboring data points of an inlying data point will be closer to the point as compared to closeness of an outlier data point. The outlier detection method we developed exploits this observation. It relies on the concept of local density around a data point, and locality is quantified by a function of distance of the data point from its k-nearest neighbors. This concept further maps into a local outlier factor, which signifies the quantitative extent of a data point being an outlier. Outliers will have a higher value of local outlier factor than will inlying data points. A threshold value of local outlier factor is used to label the data points as outliers, and this threshold is set using the histogram of local outlier factor values.

Most commonly used outlier detection methods in the industry assume that rate/pressure data follow a known model (e.g., a Gaussian probability distribution model), but outliers deviate strongly from this model. In the usual parametric-model-based approach, an *a priori* assumption of the correct model needs to be verified, which can limit the validity of the approach. The newly proposed outlier detection method, being non-parametric, is more robust as it is independent of the assumption that data follows a known model.

We present a synthetic case study demonstrating the importance of outlier detection and outlining the limitations of currently used outlier detection methods. We then validate our method through synthetic examples generated using numerical models of multi-stage hydraulically fractured wells in unconventional reservoirs. Upon validation we demonstrate application of our method using field examples from four major shale plays.

What are outliers and how can we detect them?

We begin with a short background on parametric and non-parametric methods for outlier detection and its importance in production data analysis. Parametric methods currently used to detect outliers in rate/pressure time-series data have inherent assumptions and often these methods are applied without making

sure that the assumptions are satisfied. We outline these assumptions demonstrating synthetic and field cases where they are violated.

An exact definition of an outlier often depends on hidden assumptions regarding the data structure and the detection methods applied. Yet, some definitions are regarded general enough to cope with various types of data and methods. Hawkins (1980) defines an outlier *as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.* Barnet and Lewis (1994) indicate that *an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs*. Similarly, Johnson (1992) defines an outlier *as an observation in a dataset which appears to be inconsistent with the remainder of that set of data*.

From these definitions it could be inferred that the data has an underlying 'normal' model, and outliers can be identified as deviations from this normal model. The nature of the normal data model depends on the process generating the data and on the type of data being generated. Intuitively, we might further infer that choice of the normal data model is crucial in the detection of outliers. However, an incorrect choice of normal data model may lead to poor results (as demonstrated later in the paper). Also, it is often subjective as to what constitutes a sufficient deviation for a point to be considered an outlier. This subjective nature of the problem becomes amplified in presence of noise in the data set. Rate and pressure data, collected during the productive life of a well, is usually embedded with a significant amount of noise, and detecting outliers as significantly important deviations from the actual signal becomes challenging.

To illustrate this, we construct a synthetic rate dataset, using an underlying Stretched Exponential Production Decline (SEPD) Model. We add synthetic outliers with varying amounts of synthetic noise in this dataset as shown in Figure 1. The situation in Figure 1(a) demonstrates obvious detection of synthetic outliers as the observations deviating significantly from general trend of the data. The situation in Figure 1(b) is much more subjective as few synthetic outliers lie on the boundary of the normal observations and, therefore, it is much harder to qualitatively describe a significant deviation from the normal data points. The two situations presented in Figure 1 are common in field datasets where one might observe any degree of spread of noise and, therefore, qualitative ways of detecting outliers can lead to incorrect results.

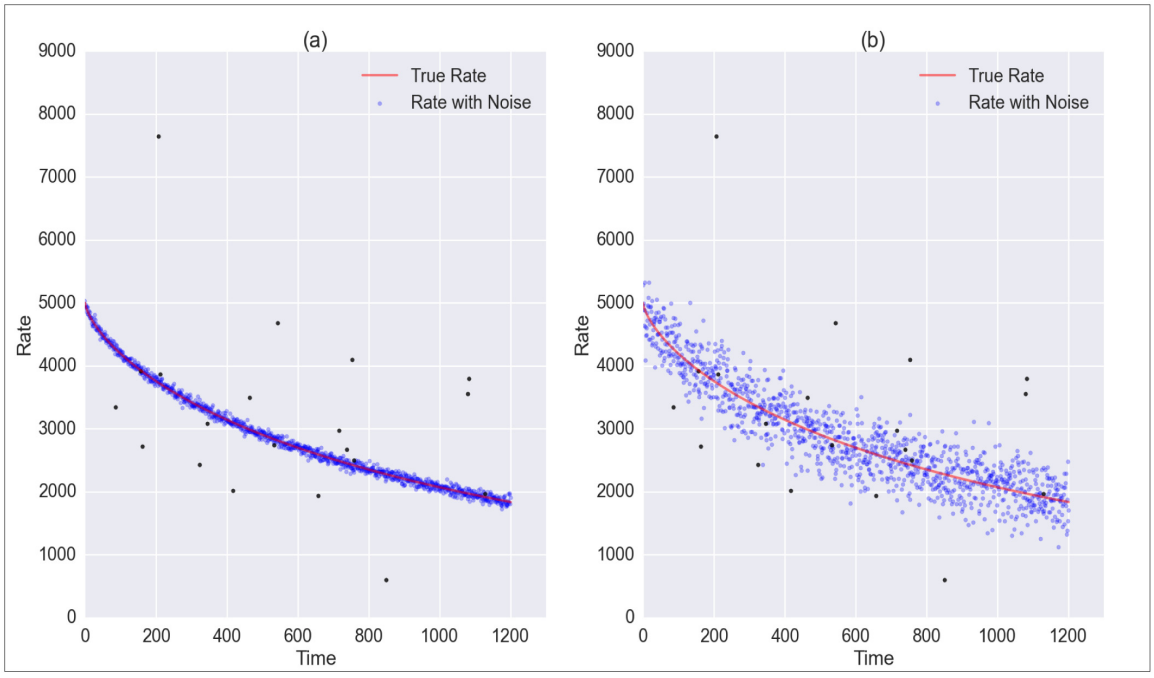

Figure 1—The difference between noise and outliers. Black dots are synthetic outliers. Notice the variation in amount of noise between (a) and (b). Outliers are less evident in (b)

In field cases it could be implied that measured time-series (rate/pressure) data could be represented as *True Model + Noise + Outliers.* In this approach, noise represents a boundary between normal data and outliers. This boundary could be quantified by some measure of *outlierness* of a data point. In terms of *outlierness,* data points could be represented as lying on the continuous spectrum from normal data to outliers as shown in Figure 2.

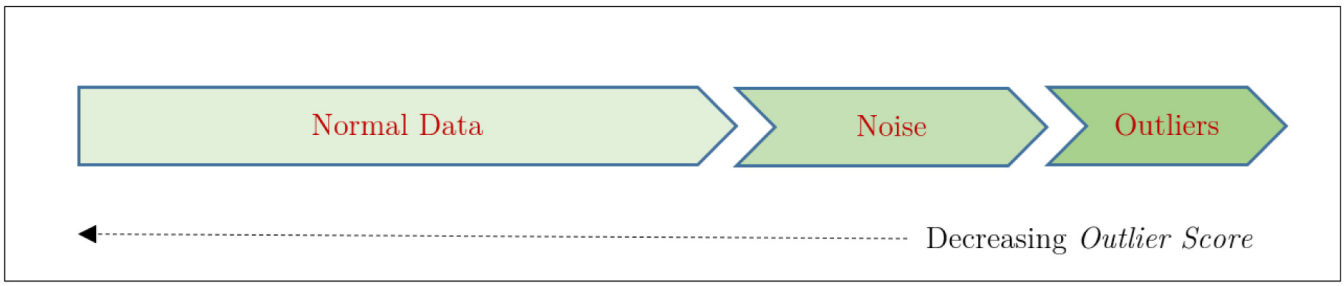

Figure 2—Data spectrum from normal to outliers

Outlier detection methods can be divided between supervised, semi-supervised and unsupervised scenarios. In the supervised scenarios, training data with labels of both normal and outlier points is available and from these labels a ″normal data model″ is trained to predict the label of unseen (future) data points. In this scenario, the problem of outlier detection is often posed as a classification problem. In the semi-supervised scenario, training data with labels of either normal or outlier points is available and from these labels a ″normal data model″ is trained in order to predict the label of unseen (future) data points. In the unsupervised scenario, no training data is available and outliers are detected based on hidden structure of the dataset. The problem of outlier detection in production data analysis falls under the unsupervised scenario.

All the outlier detection methods in the unsupervised scenario create a model of the normal data points and then compute an outlier score of a given data point. In the statistical parametric methods, normal data is modelled as having a known underlying probability distribution. The outlier score is computed by evaluating quality of the fit between the data point and the model. The downside of these methods is the requirement of *a priori* information about the probability distribution that normal data follows.

In the non-parametric methods, normal data is modelled based on the proximity of each data point to other data points in the data space. The outlier score is computed as the deviation of the proximity of a data point from the proximity of other data points. The proximity of a data point can be defined as its distance to its *k-nearest neighbor* or by comparing the density around it with the density around its local neighbors. The basic assumptions here are that the normal + noise data points have a dense neighborhood and the outliers have considerably less dense neighborhood. In other words, outliers are assumed to be isolated at large distances from most of the data. Considering the field cases of pressure and rate data (which is the focus of our discussion here), this assumption seems to be reasonably valid almost always.

In regression-based methods, a known underlying parametric model (linear, exponential, etc.) is fit to all data and the noise in the data is assumed to be following a particular probability distribution. Outlierness is scored based on the standard deviation of a data point from the fit. These methods have two *a priori* assumptions: the parametric model to be fit into the data and probability distribution of the noise in the data. Again, in field cases, this information is almost never available.

The various underlying assumptions associated with the normal data model make the choice of the model crucial. In case of the wrong choice of model, data points may be incorrectly reported as outliers because of a poor fit to the erroneous assumptions of the model. In the workflow of production data analysis, the currently dominant methods are often used without giving proper consideration to the validity of the underlying assumptions. Even worse cases arise where the results seem convincingly accurate enough, even though one or more of the assumptions are eventually invalid. In the next two sections we demonstrate some of the limitations in the current production data analysis workflow of removing outliers and we also demonstrate the importance of removing outliers in this workflow.

How wrong can things go if we avoid detecting outliers?

The analysis of production data in petroleum reservoirs to determine well/reservoir properties, completion effectiveness, and estimate future production has become popular in recent decades. The challenging aspect of production data analysis is the non-uniqueness and uncertainty associated with the end results. Production data diagnostics help identify the causes of these uncertainties and address them systematically. Production data diagnostics are utilized for two main purposes. First, measured data are reviewed to check for correlations between pressure and rate data. This is very important because data having no correlation will not provide any diagnostic value. Also, in this way we can detect the features or events which should be filtered or discarded prior to analysis. Next, characteristic features, which are exhibited by production data, are identified with the use of diagnostic plots. These diagnostic plots help to identify flow regimes and compare data to a well/reservoir model. We can infer that production data diagnostics is a very important part of production data analysis and should be given its due diligence to avoid unrealistic end results.

A very important aspect of production data diagnostics is outlier removal from the rate/pressure data and it should be performed before any of the diagnostic analysis is carried out on the data. Presence of outliers in the data could lead to incorrect identification of flow regimes, well/reservoir model and incorrect estimate of model parameters. This could in turn lead to an incorrect estimate of ultimate recovery from the reservoir. Especially for unconventional reservoirs, where there is a heavy reliance on empirical techniques for production data analysis, the presence of outliers adds more uncertainty to an already problematic workflow.

To demonstrate the effect of outliers, we constructed a synthetic case to represent well, reservoir, and fluid characteristics typically found in unconventional tight/shale oil and gas plays. Horizontal wells with multiple hydraulic fracture stages have become the norm in the majority of unconventional plays and, therefore, a multi-fractured horizontal well was chosen as the well model. An analytical model (Brown et al. 2009; Stalgorova et al. 2012; Stalgorova et al. 2013) was constructed using Fekete Harmony®, and 30 years of dry gas production from a multi-fractured horizontal well in a homogeneous reservoir was simulated, keeping bottomhole flowing pressure constant. The reservoir, well and fluid properties are provided in Table 1.

Table 1—Synthetic Dataset: Model Input Parameters

Reservoir Properties:	
Reservoir Temperature, T_r	350 °F
Initial Reservoir Pressure, p_i	5000 psia
Net Pay, h	300 ft.
Porosity, ϕ	10 %
Initial Gas Saturation, S_{gi}	70 %
Initial Oil Saturation, S_{oi}	0 %
Initial Water Saturation, S_{wi}	30 %
Reservoir Length, x_e	3500 ft.
Reservoir Width, y_e	600 ft.
Permeability, k	0.00007 md
Area of SRV, A	48 acres
Hydraulically Fractured Well Model Parameters:	
Fracture Half Length, x_f	300 ft.
Effective Hz Well Length, L_{ex}	3500 ft.
Wellbore Radius, r_w	0.328 ft.
Fracture Conductivity, F_{cd}	350 (dimensionless)
Number of Fracture, n_f	20
Production Parameters:	
Bottomhole Flowing Pressure, P_{wf}	500 psia
Producing Time, t	30 years

We choose 5.5 years of production data as history and to add realism we added noise and outliers to the data. To demonstrate the influence of outliers on model parameter estimation we further created 3 cases by adding 10 outliers (Case B), 20 outliers (Case C) and 30 outliers (Case D) to the 5.5 years of production data.

Next, we forecast production (up to 30 years) using the Duong + Arps hybrid model. A separate fit based on the data classified as history was generated for all three cases described above and for a fourth case with no outliers (Case A). In all the fits the switch from transient to boundary dominated flow model was made at the same time based on our *a priori* knowledge about the synthetic reservoir system, this primarily achieves the purpose of not letting the switch time impact the fitting accuracy in either of the cases and makes the fitting results discussed below purely demonstrative of the impact of outliers on them. Fits generated for all the cases are shown in Figures 3 – 6. Note that the outliers were not removed while obtaining the fit for any of the cases.

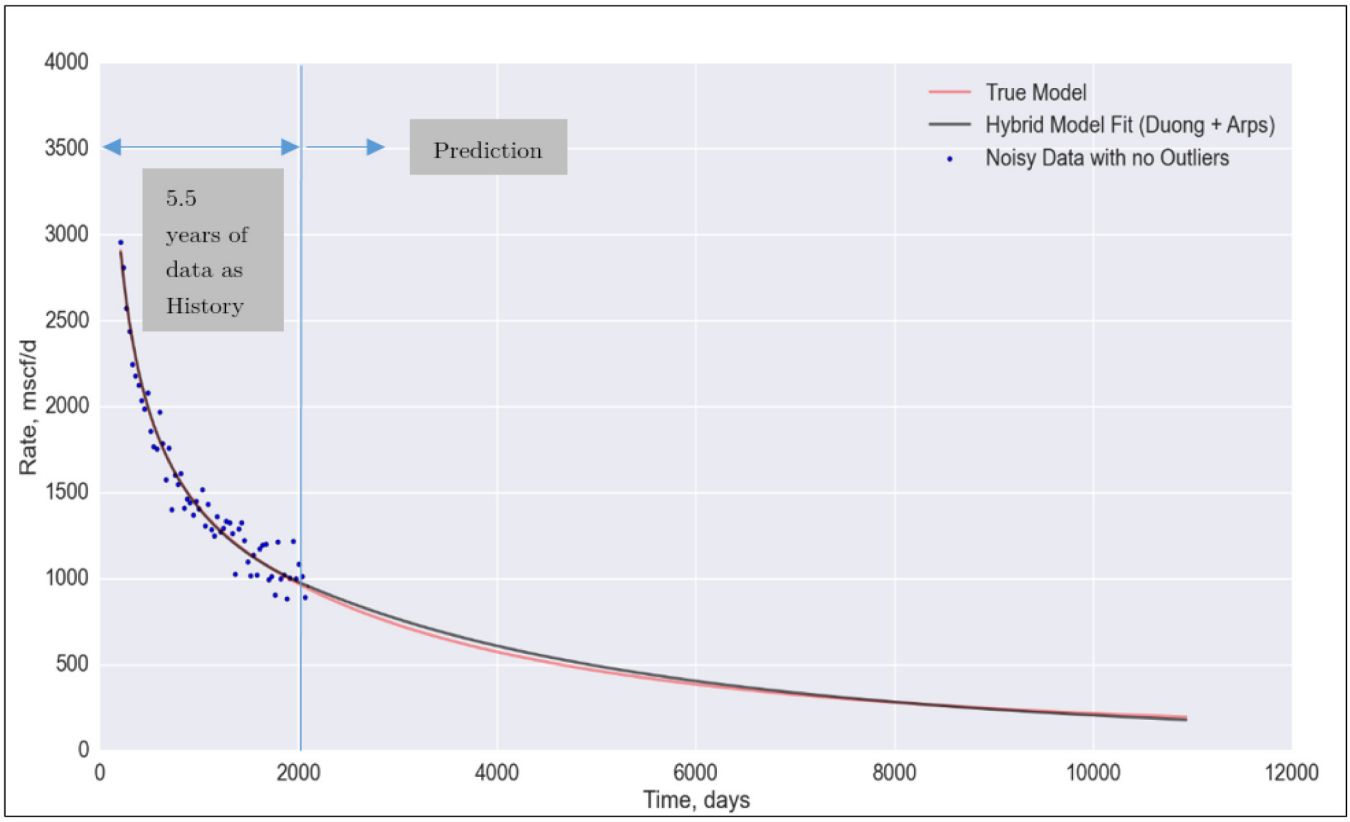

Figure 3—Case A (No Outliers) – Hybrid model fit generated using noisy data with no outliers

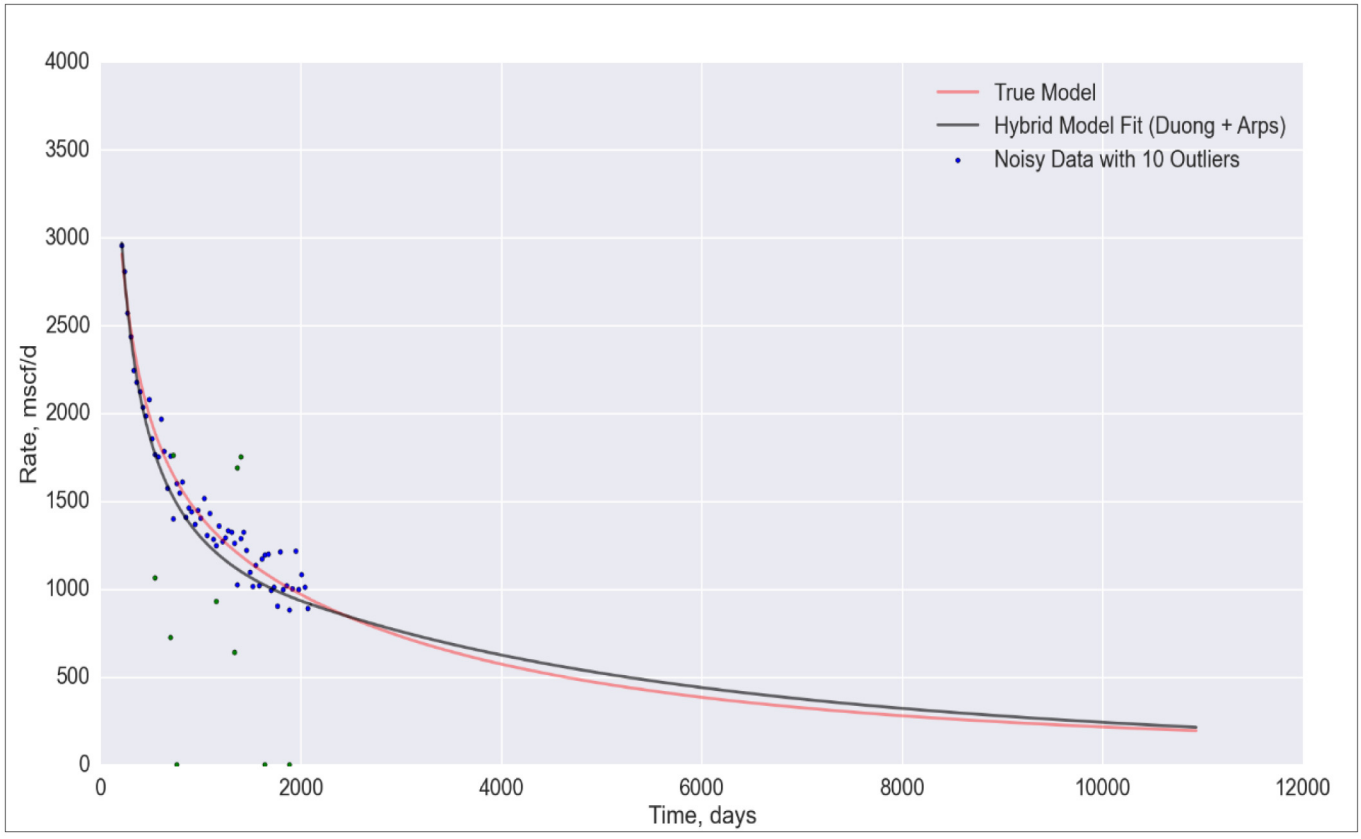

Figure 4—Case B (10 Outliers) – Hybrid model fit generated using noisy data with 10 outliers

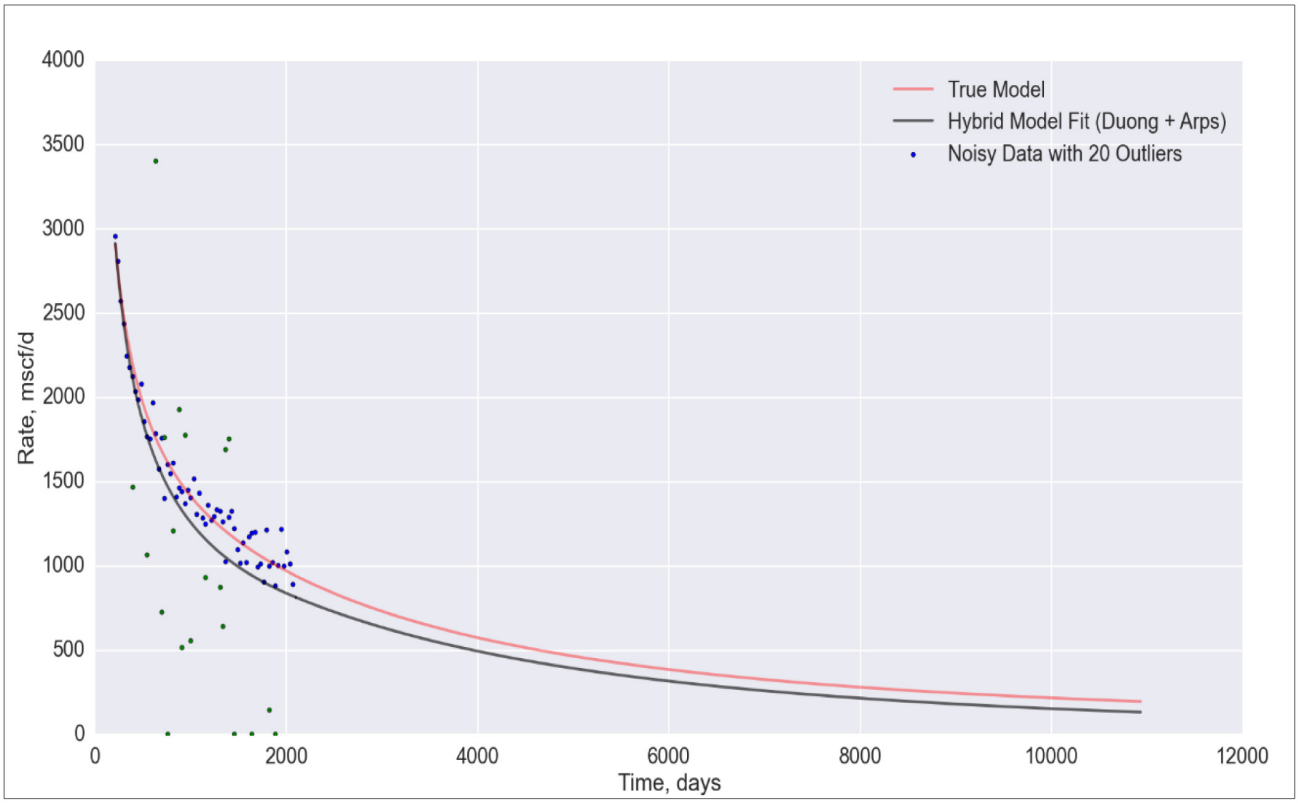

Figure 5—Case C (20 outliers) – Hybrid model fit generated using noisy data with 20 outliers

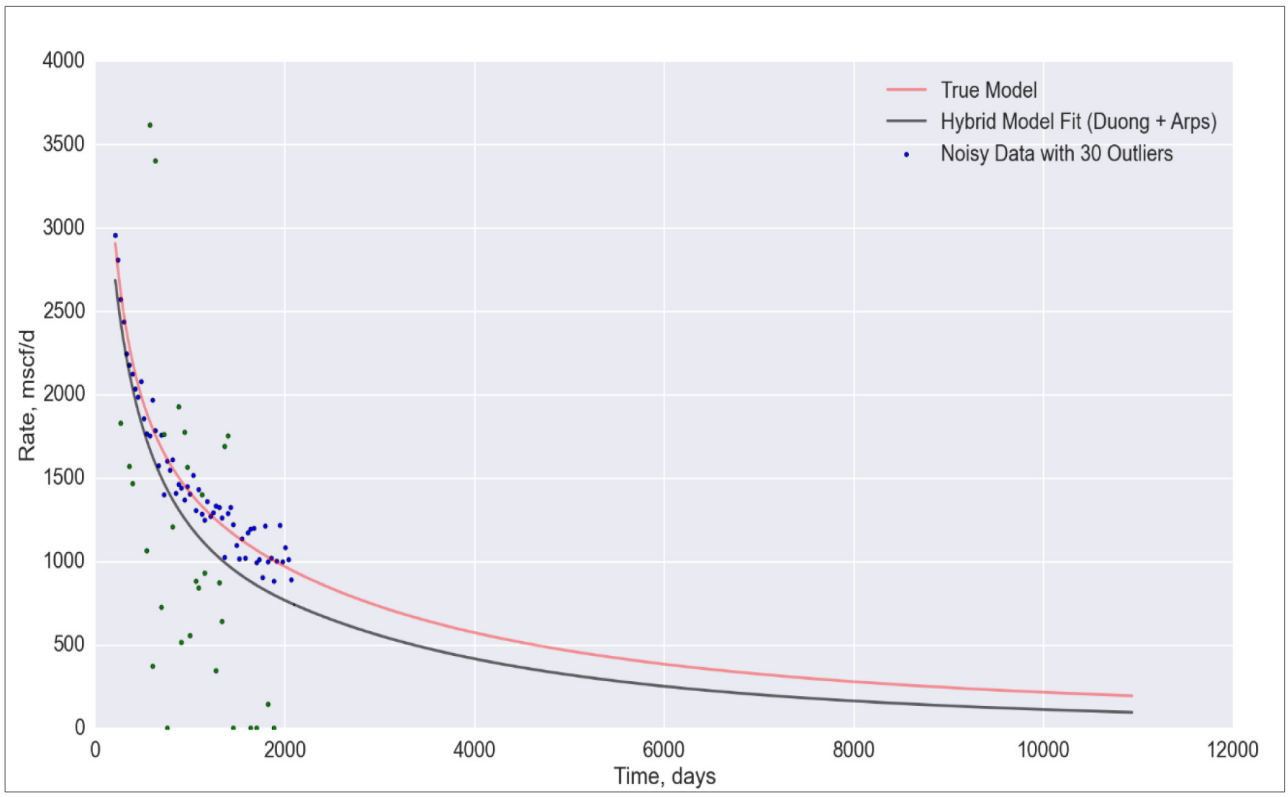

Figure 6—Case D (30 outliers) - Hybrid model fit generated using noisy data with 30 outliers

Comparing the hybrid model fit among all the cases, Case A with no outliers shows the best fit and Case D with 30 outliers shows the worst fit. Table 2 compares the estimated ultimate recovery (EUR) from all the cases to the EUR generated using the analytical model. We note that the calculated EUR for Case A has the least difference and Case D has the greatest difference from the analytical model EUR.

Table 2—Summary of EUR values for fit in all cases

Case	EUR (MMSCF)	% Difference
Analytical Model	7485.53	0
Case A (No Outliers)	7483.16	- 0.03
Case B (10 Outliers)	8146.81	8.83
Case C (20 Outliers)	6612.72	- 11.66
Case D (30 Outliers)	5593.30	- 25.28

From this synthetic case study, we can firmly infer that if outliers are not accurately identified and removed, they can drastically reduce the diagnostic value and reliability of rate/pressure transient analysis.

What are we doing wrong in outlier detection?

″Experience with real-world data, however, soon convinces one that both stationarity and Gaussianity are fairy tales invented for the amusement of undergraduate classes.″ – Thompson (1994)

In the petroleum industry, not much literature has been published on the problem of outlier detection. We demonstrated the need for outlier detection in the previous section. In this section we outline the shortcomings of methods currently used in the industry for outlier detection and contemplate the need for a more robust method.

The most basic form of outlier detection used in the industry is visual detection. Data points which visually appear to deviate from the general trend of the rate/pressure data are classified as outliers. The outliers thus detected are subjective, based on the analyst's perspective of what is enough deviation from the general trend for a data point to be declared an outlier. The limitations of this subjectivity have already been demonstrated in the previous sections (refer Figures 1 – 2). A limitation of this approach is the time consuming and intellectually draining nature of visually processing all the data points, especially if the analyst has to analyze hundreds of wells in a short time (which is usually the scenario for a reservoir engineer).

The second most frequently used methodology in the industry for detecting outliers is statistical tests. The general idea is that, given a statistical distribution, we compute the parameters assuming all data points have been generated by such a statistical distribution, and outliers are points that have a low probability to be generated by the overall distribution. The test most frequently used in the petroleum industry is an assumed Gaussian probability distribution, for which we calculate parameters like mean and standard deviation, and data points deviating more that x (usually $x = 1.5$) standard deviations from the mean are declared outliers. The basic inherent assumptions involved in this method are that:

- Normal data objects follow a (**known**) distribution and occur in a high probability region of this model.
- Outliers deviate significantly from this (**known**) distribution.

For rate/pressure data, the underlying probability distribution of the data is unknown and is usually assumed to be Gaussian. Even though some might argue that this is a reasonable assumption, the statistical test procedure is not robust to the violation of this assumption. Therefore, if the statistical test is performed with the assumption of data following a Gaussian distribution and instead the data follows a different distribution, then the outliers could be misidentified or go undetected. In production data diagnostics, the

large influence of outliers is dealt with by removing them completely from future steps of the analysis. This makes deletion of outliers using statistical tests methodologically unsound, especially when the underlying model of the process being measured and the usual distribution of measurement error are not known with a high degree of confidence.

Another limitation of statistical tests is that parameters of statistical distribution (mean and standard deviation for the Gaussian distribution) are calculated for the entire data set. For a non-stationary time series data these parameters are not constant but evolve over time. The production from an oil and gas reservoir is a non-stationary process; therefore, calculation of mean and standard deviation for the entire rate/pressure data set does not make any sense. To overcome this, Holdaway (2014) suggested the calculation of these parameters over a window of data which can be shifted along the data. This approach again requires analyst effort to estimate the optimum size of the window and ends up being as draining as the visual detection method. In addition, inside a window the parameter estimates become even more sensitive to outliers due to the smaller number of samples to offset the effect of outliers.

The third form of outlier detection used in the industry combines statistical tests with regression based methods. The entire rate/pressure dataset is fitted to a (**known**) physical model and the noise in the dataset is fitted to a Gaussian distribution. Outliers are points which deviate more than x (usually $x = 1.5$) standard deviations from the model fit over the entire dataset. This method, apart from the assumptions of the previous method, has one more assumption: that the true physical model generating the normal data is known. Neither the true physical model nor the noise following Gaussian distribution are known *a priori* in the case of rate/pressure data. This method suffers from the same limitation as the previous method plus adding the uncertainty of assuming a physical model generating the data.

From the above discussion, we might conclude that all the currently used methods in the industry could work well if the assumptions made during their application are satisfied. The validity of satisfaction of these assumptions is a time consuming, exhaustive process and is usually not performed in real world applications in the petroleum industry. The general practice in the industry is to blindly apply these methods on the real world data sets, which could lead to unfavorable scenarios. We need to introduce better methods to detect outliers with underlying assumptions that are almost always satisfied. We need to make the analyst's job easy by being able to apply the outlier detection universally to any data set and not worry about the underlying assumptions. In the next section we introduce a new method which satisfies all of these needs.

A new approach to detect outliers in rate/pressure data

Motivated by the limitations of the methods currently used in the industry, we introduce a new approach to detect outliers in this section. The basic idea is based on the concept of a local density, where locality is defined by the k nearest neighbors, whose distance is used to estimate the density. By comparing the local density of an object to the local densities of its neighbors, one can identify regions of similar density, and points that have substantially lower density than their neighbors are considered to be outliers. This approach was first suggested by Breunig et al. (2000) as the Local Outlier Factor method.

What is the Local Outlier Factor (LOF)?

To explain LOF, we will first define and explain notions which build up to LOF. Let us begin by defining the k – *distance* of an object (data point) p:

> The k – *distance*(p) of an object p is defined as its distance (any distance metric like Euclidean, Manhattan, etc. could be used) to the k^{th} nearest neighbor.

Now let us define the k – *distance neighborhood* of an object p:

The set of all objects belonging to the data set whose distance from the object p is less than the $k-distance(p)$, is known as the $k-distance\ neighborhood$ of object p. It is denoted as $N_k(p)$.

To present a visual demonstration of these definitions, let us pick data from the synthetic case created in the previous section. We add synthetic noise and outliers in the data set to make it look similar to a typical field data set. Figure 7 illustrates the idea of $k-distance$ and $k-distance\ neighborhood$ with $k = 5$, the green circle around the yellow data points represents the $k-distance\ neighborhood$ for these particular data points and the radius of the circle represents the $k-distance$. Notice that the k^{th}(here 5^{th}) nearest neighbor of data points in orange lies on the periphery of the circle. The distance metric used in this example was Euclidean distance, which will also be used for all future examples in this paper.

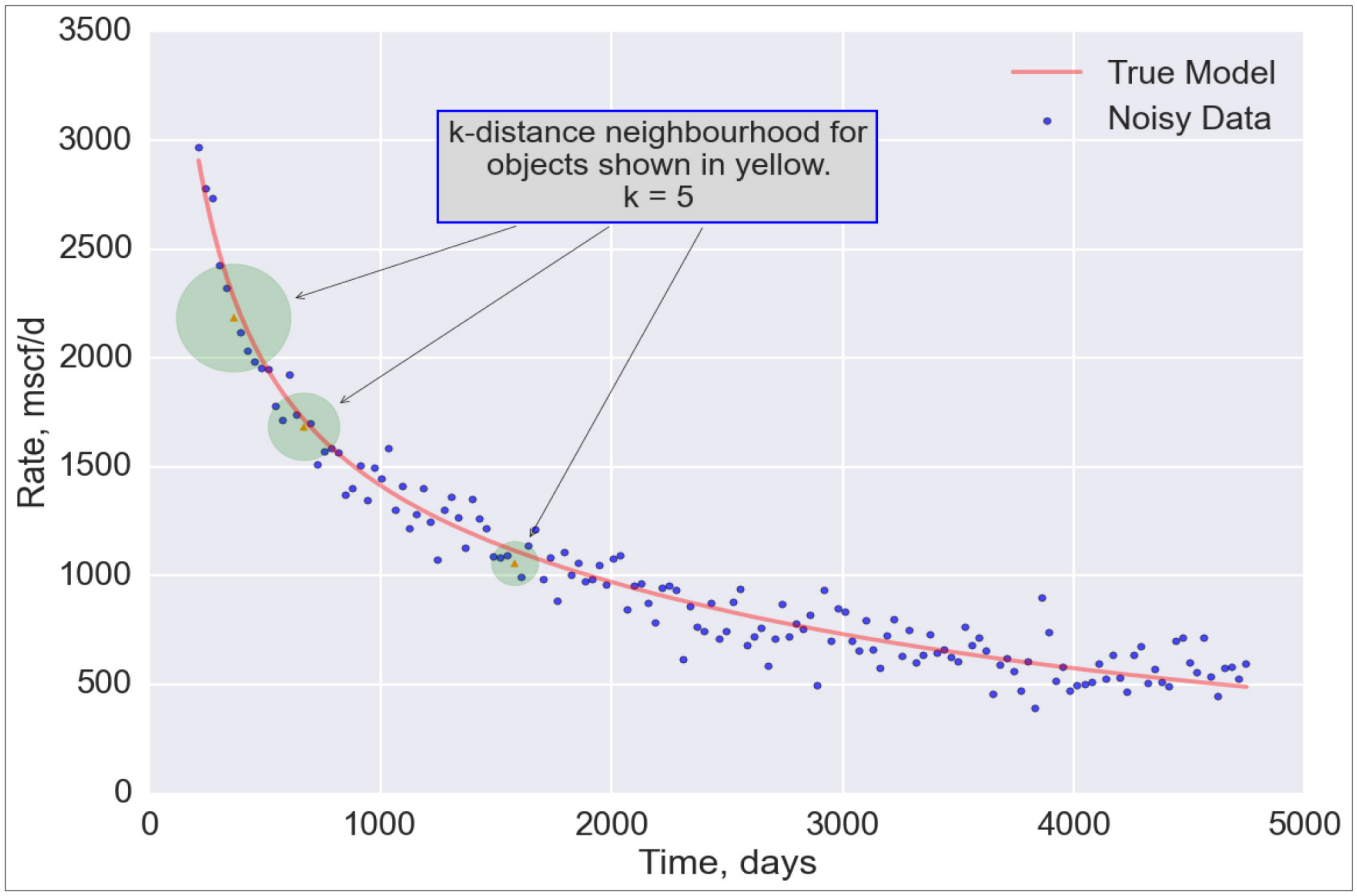

Figure 7—Demonstration of k – distance neighborhood, for k = 5

Next we define $reachability-distance$ and $local\ reachability-distance$ (lrd).

The $reachability-distance$ of an object p from o is defined as the true distance between the two objects, but at least the $k-distance$ of object o,

$$reachability-distance_k(p,o) = \max\{k-distance(o), d(p,o)\}. \quad (1)$$

Therefore, if an object p belongs to the $k-distance\ neighborhood$ of o, then $reachability-distance_k$ of p from o is the of $k-distance$ the object o, else the $reachability-distance_k$ of p from o is just the true distance of p from o. In terms of $reachability-distance$, the k nearest neighbors of object o are considered to be equally distant from object p.

Figure 8 illustrates the idea of $reachability-distance$. Notice object p_1 belongs to the $k-distance\ neighborhood$ of object o, whereas object p_2 does not belong to the $k-distance\ neighborhood$ of object

o. Therefore, $reachability - distance_k(p_1, o) = k - distance(o)$ and $reachability - distance_k(p_2, o) = d(p_2, o)$. The reason for introducing *reachability – distance* over the true distance is to reduce the statistical fluctuations of true distance for all the p's close to o. This smoothing effect of *reachability – distance* can be controlled by k. The higher the value of k, the more similar the reachability distances of objects within the same neighborhood (Breunig et al. 2000). Next we introduce the concept of density in the neighborhood of a data point to be able to differentiate between outliers and normal points.

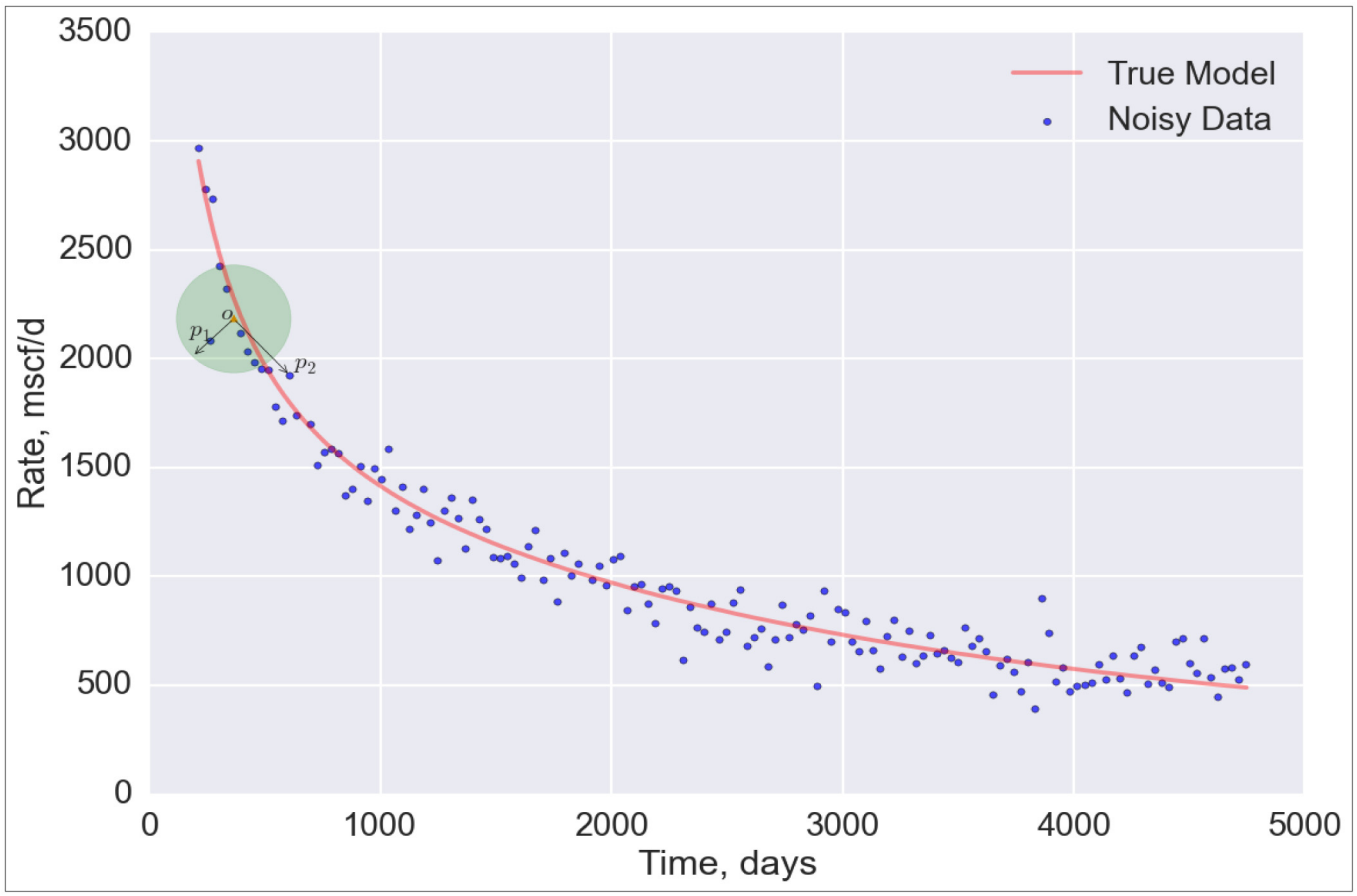

Figure 8—Demonstration of *reachability – distance*$_k$(p_1, o) and *reachability – distance*$_k$(p_2, o), for $k = 5$

The of *local reachability density* an object p is defined as the inverse of the average reachability distance of the object p from its k nearest neighbors,

$$lrd_k(p) = 1/\left(\frac{\sum_{o \in N_k(p)} reachability - distance_k(p,o)}{|N_k(p)|}\right). \quad (2)$$

Intuitively, we might notice that *local reachability distance* of the object p represents the inverse of average distance at which p can be 'reached' from its neighbors. Therefore, the *local reachability density* for an outlier point will be much smaller than that for an inlying point in a dense neighborhood, which is demonstrated by figure 9.

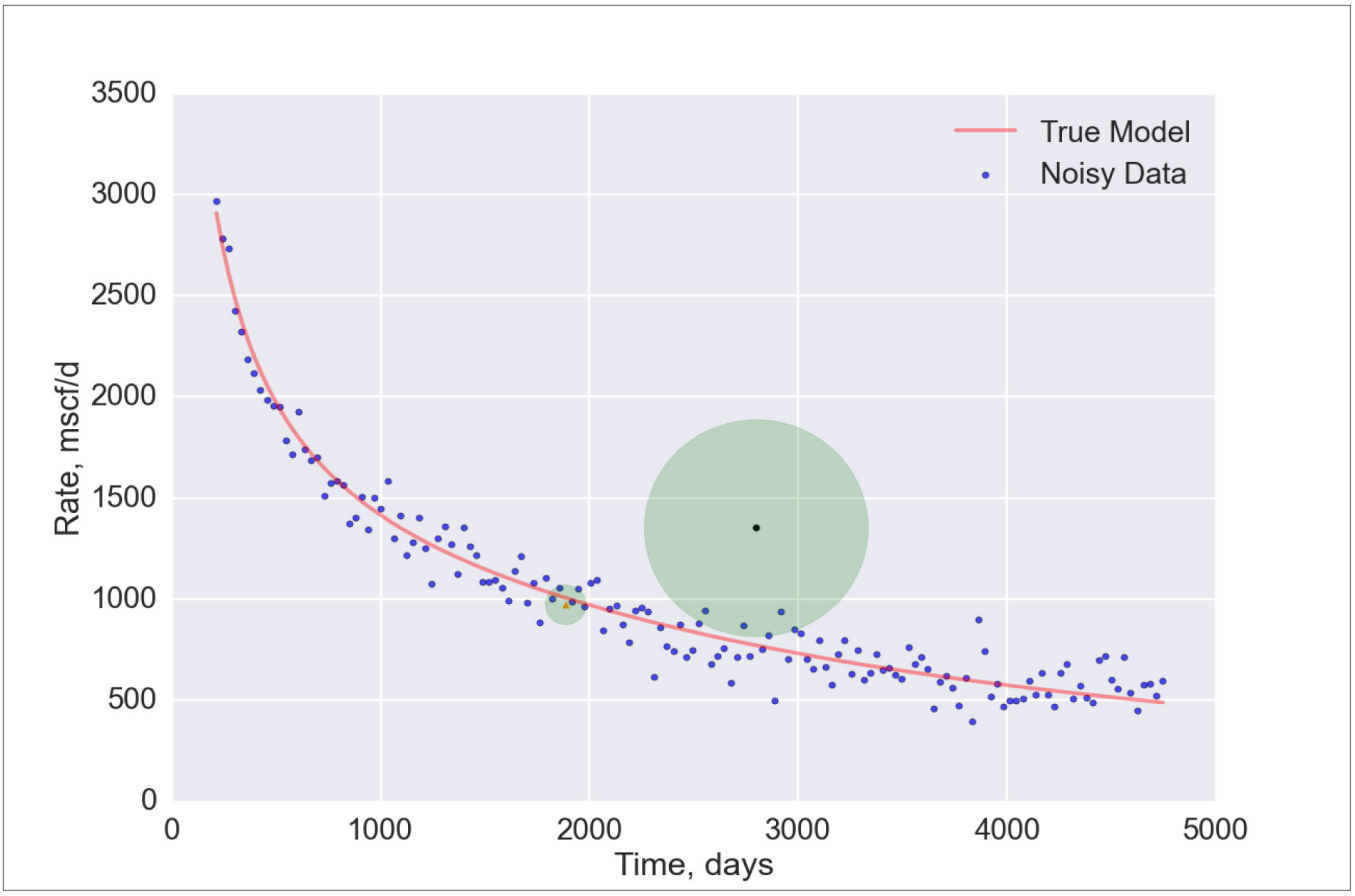

Figure 9—Demonstration of local reachability distance, for k - 5. Outlier is in black

We have created an idea of the *local reachability density* of an object and how it differs between the inlying objects and outlier objects. Now, we need to compare the *lrd* of an object to its neighbors to be able to identify outliers, as outliers are expected to have lower densities than their neighbors. We do this comparison by computing the *Local Outlier Factor* (*LOF*) of an object, defined as,

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \left(\frac{lrd(o)}{lrd(p)} \right)}{|N_k(p)|} = \left(\frac{\sum_{o \in N_k(p)} lrd(o)}{|N_k(p)|} \right) / lrd(p). \quad (3)$$

The *LOF* of an object is the average *lrd* of its neighbors divided by the object's own *lrd*. Therefore, *LOF* captures the degree to which we call the object an outlier.

The smaller *p*'s *lrd*, and the larger the *lrd* of *p*'s *k* nearest neighbors are, the larger the *LOF* value of *p* is. Figure 10 compares the *lrd* of an outlying point to its *k* nearest neighbors. An *LOF* value close to 1 indicates that the object's density is comparable to its neighbors and thus it is not an outlier. A *LOF* value larger than 1 indicates that the object's density is much lower than its neighbors and thus it is an outlier.

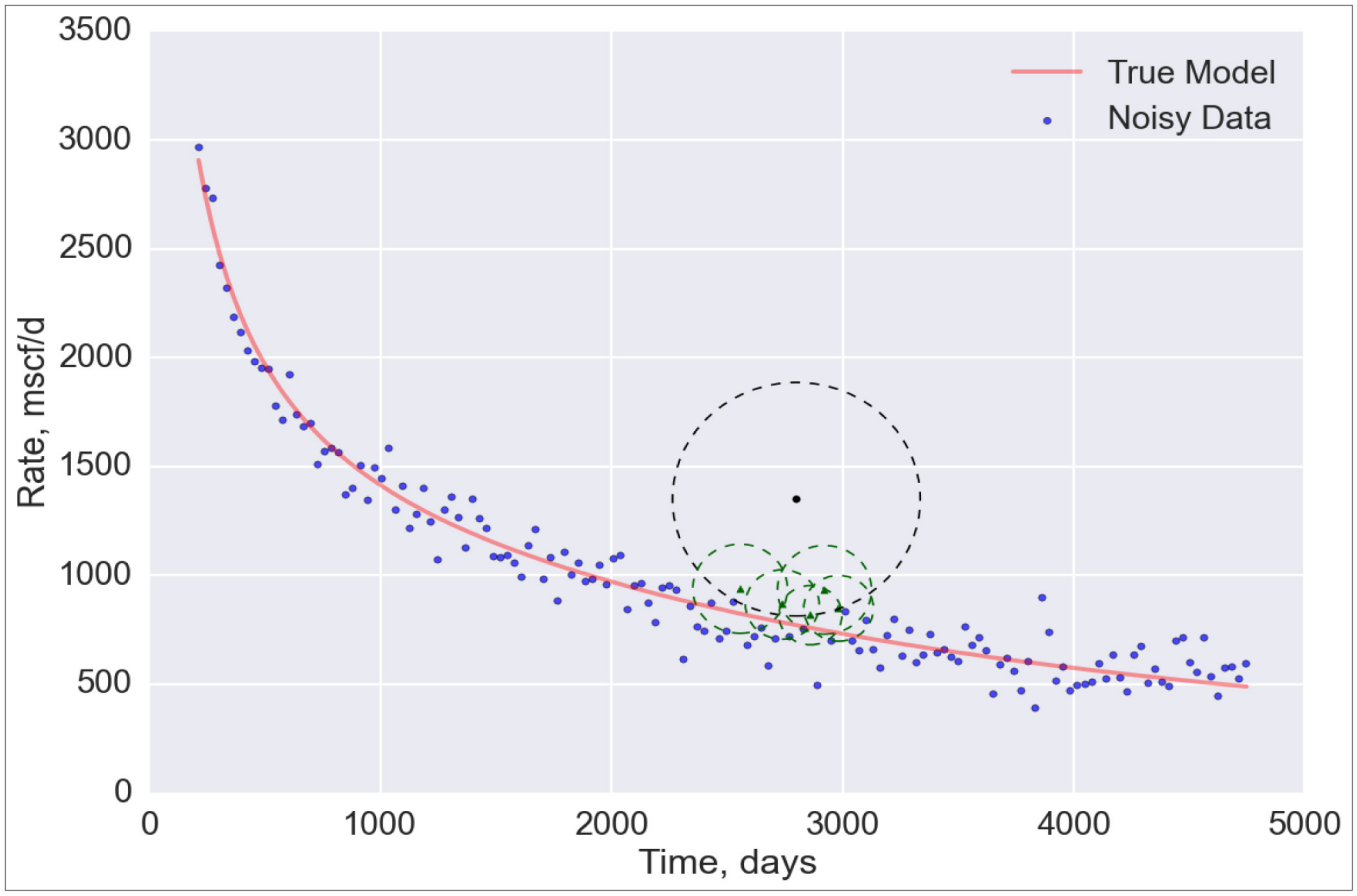

Figure 10—Comparing the local density of an object with the densities of it's k nearest neighbours, for k = 5

The only assumption made during the development of this method is that the density around a normal data object is similar to the density around its neighbors and the density around an outlier is considerably different than the density around its neighbors. This is a very intuitive assumption and should almost always be satisfied for the kind of outliers encountered in the rate/pressure data. Therefore, this method could be applied without being concerned about its underlying assumptions, as they are almost always satisfied.

LOF for all the data points is calculated and a threshold value of *LOF* can be used to label outliers and inlying points. In next section we validate this method and demonstrate it on real field examples. We also present a way to estimate the threshold value.

Validation

We will now validate the method described in previous section to detect outliers in the rate/pressure data. We understand that the method described can be challenging to understand for someone within the petroleum industry who has not been exposed to the outlier detection literature and therefore we created an Excel® based outlier detection tool.

The back end of the tool where the actual computations are performed has been written in python programming language for increased speed and Excel® acts as a front end to interact with the user of the tool. Figure 11 shows a screenshot of this tool. The input parameters are *number of neighbors* (*k*) and the *Threshold* value of *LOF* for a data point to be classified as an outlier. The tool, though specifically designed for rate and pressure, could also be modified for other kinds of data streams encountered in the petroleum industry. We will keep these modifications as future work recommendations.

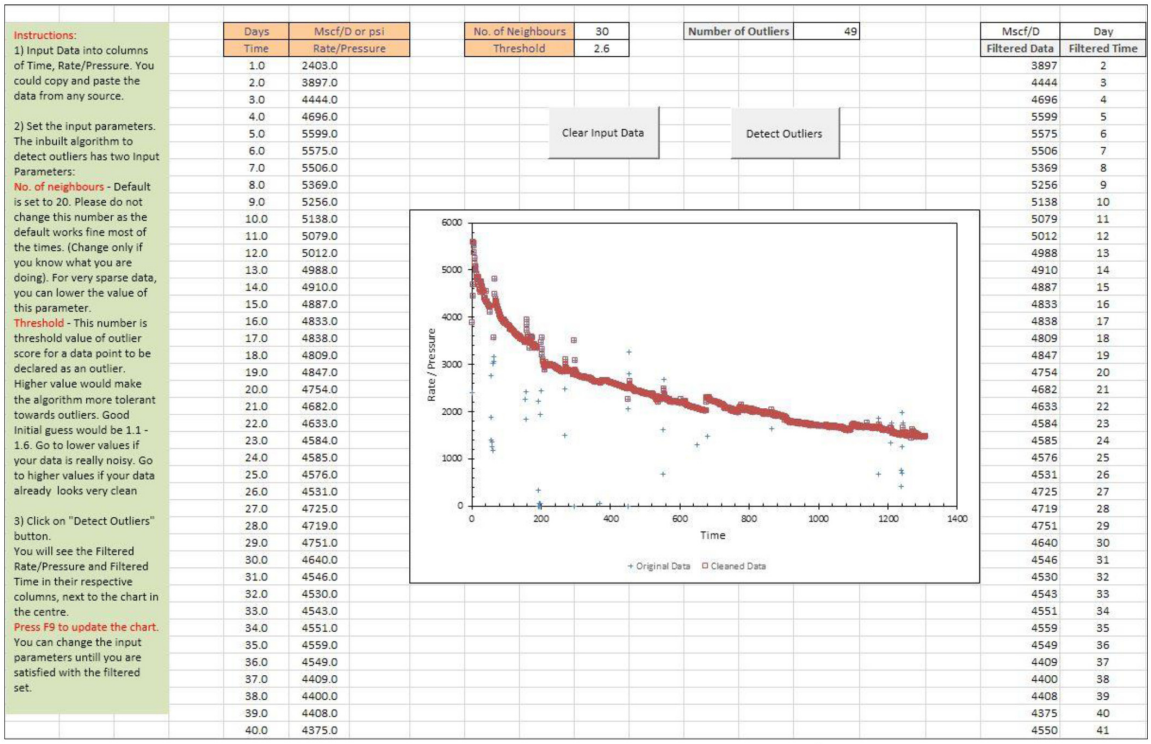

Figure 11—Screenshot of outlier detection tool built in Excel®

We created 13 years of simulated data for the synthetic case discussed earlier, and inserted synthetic noise and outliers into the data set to make it look more realistic. Figure 12 shows the performance of the method on the synthetic data set. As expected, the new method detected the outliers perfectly. The input parameters were set as $k = 20$ and $Threshold = 1.5$. Next we demonstrate a way to decide the appropriate value of the *Threhold* parameter.

Figure 12—Outlier detection demonstration on synthetic data set, for k = 20 and Threshold = 1.4

We have already discussed that the expected value of *LOF* for inlying points is close to 1, therefore an intital guess of *Threshold* = 1, should be a good start. But in our experience we have found this value is usually too conservative, and values slightly higher than 1 provide much better performance. But, this still does not tell us much about the *Threshold* parameter. We could see the most usual values of *LOF*, and get an idea of the range by plotting the histogram of *LOF* values. Figure 13 represents the histogram of *LOF* values for the synthetic case data set. From this histogram, we could infer that most of the data points have a *LOF* value in the range of 0.9 – 1.4, and therefore a good initial guess for the *Threshold* parameter is 1.4. As seen in the figure 12, this initial guess produces an excellent result. Thus we have validated our method. In the next section we demonstrate the method on five field data sets.

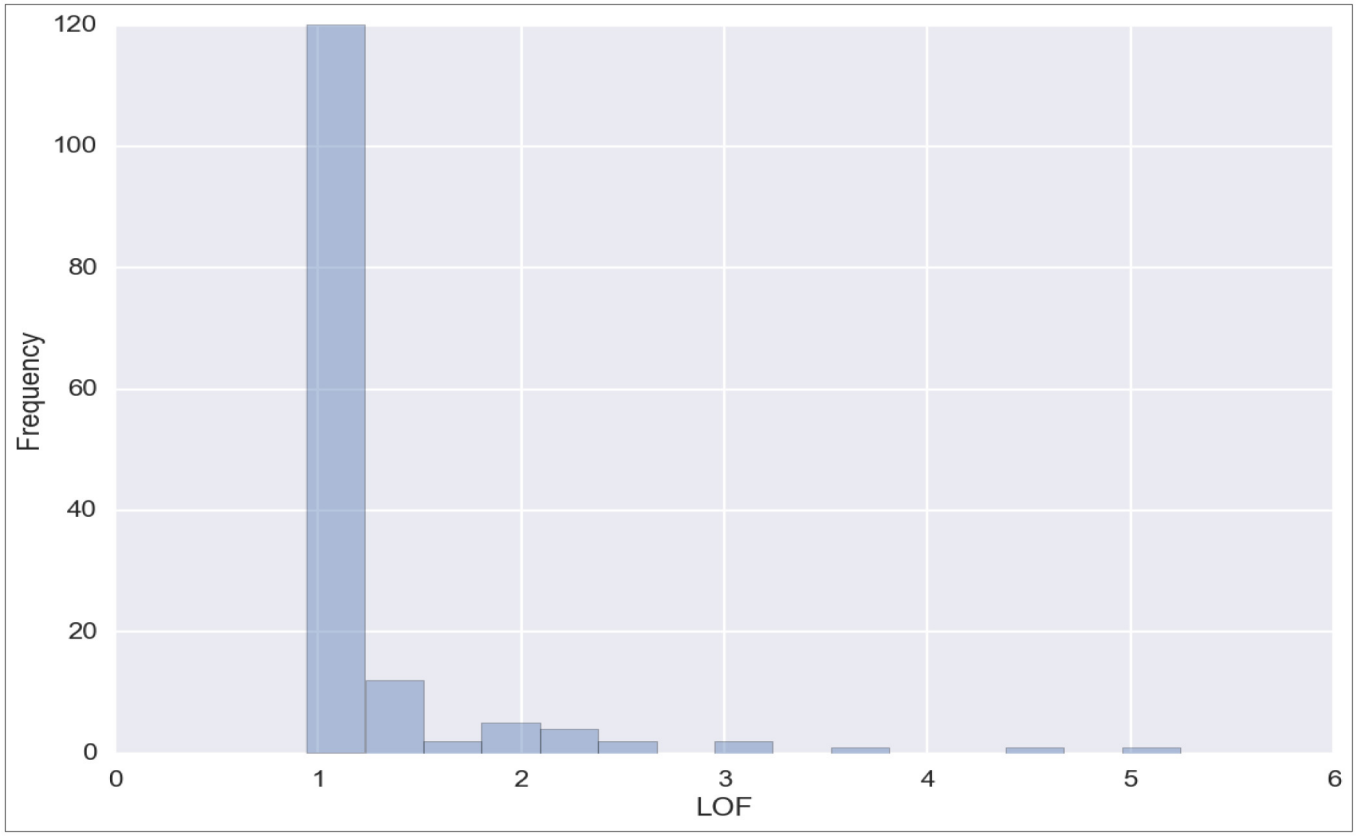

Figure 13—Histogram of LOF values in the synthetic data set, for k = 20

Field Examples

We have applied the new method in five field examples from different shale plays. These are:

Example 1: Marcellus Well
Example 2: Woodford Well
Example 3: Bakken Well
Example 4: Eagle Ford Well
Example 5: Barnett Well

The outliers detected in all the examples are shown in figures 14 to 21, along with the histogram of LOF values. The value of k used was 30 in all the examples. In our experience, $k = 30$ works well for most cases. If the data are very sparse (e.g., monthly or biannual data), then increasing the value of k might improve performance.

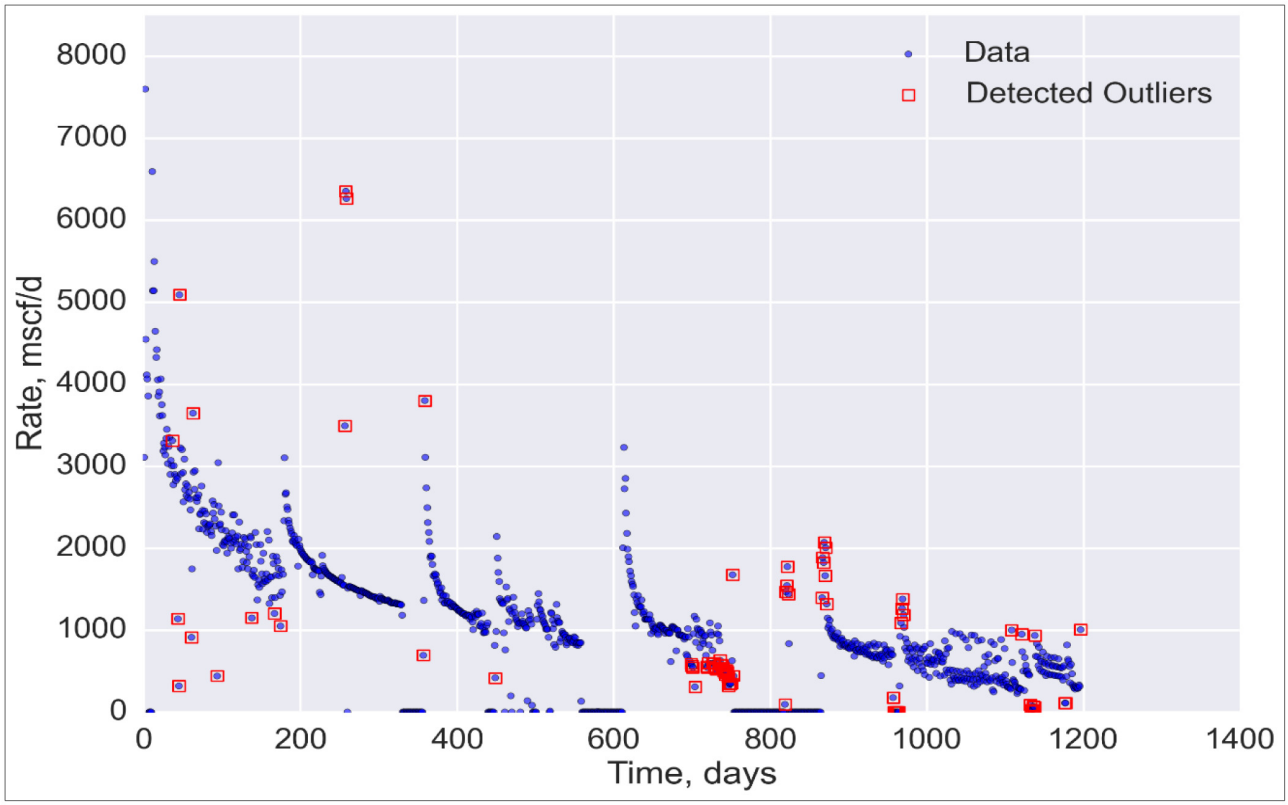

Figure 14—Field example 1 (Marcellus well)

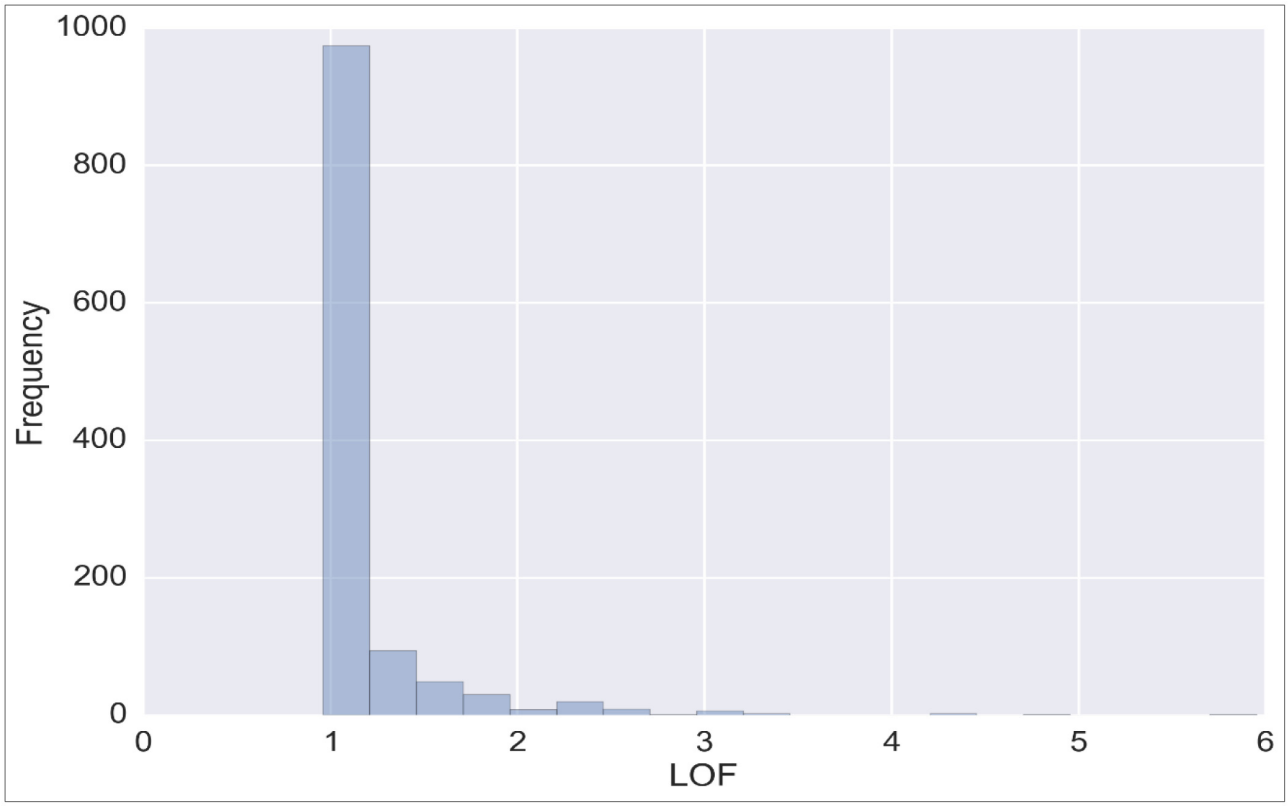

Figure 15—Field example 1 (Marcellus well) LOF histogram

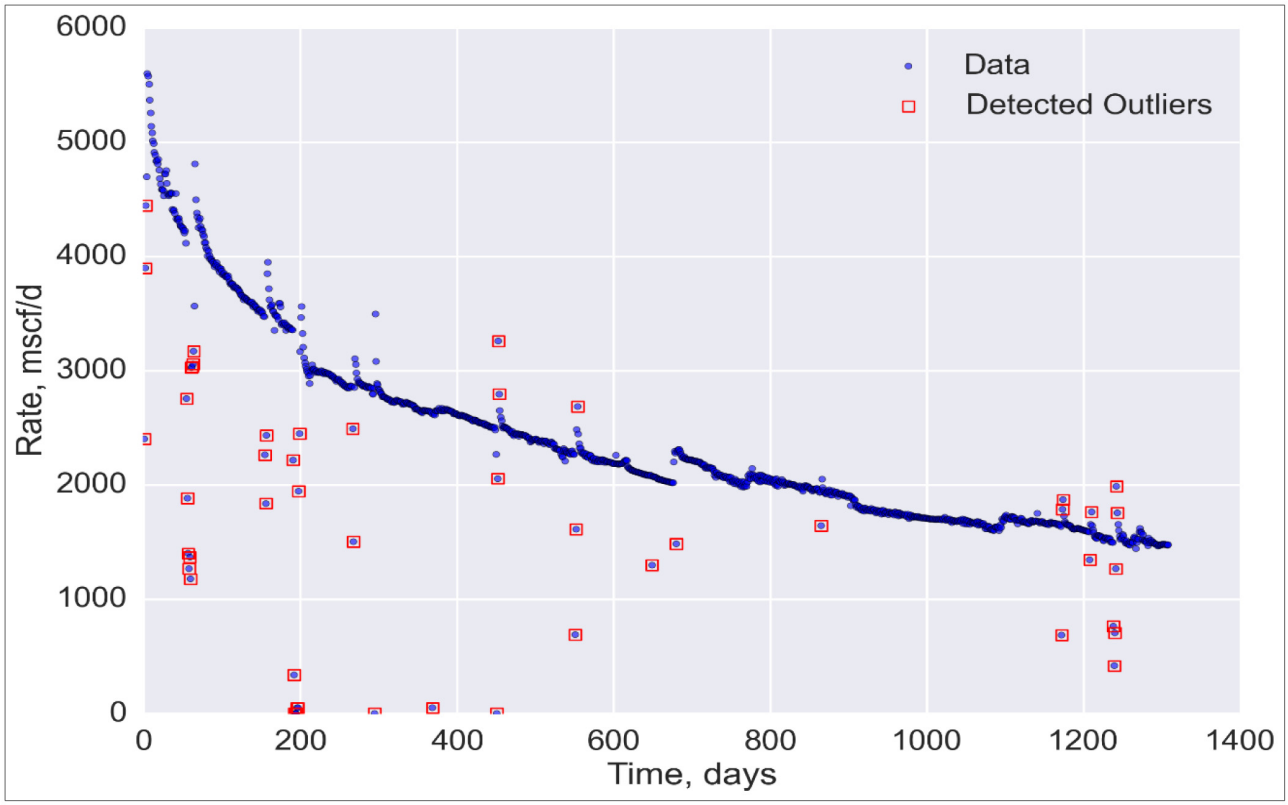

Figure 16—Field example 2 (Woodford well)

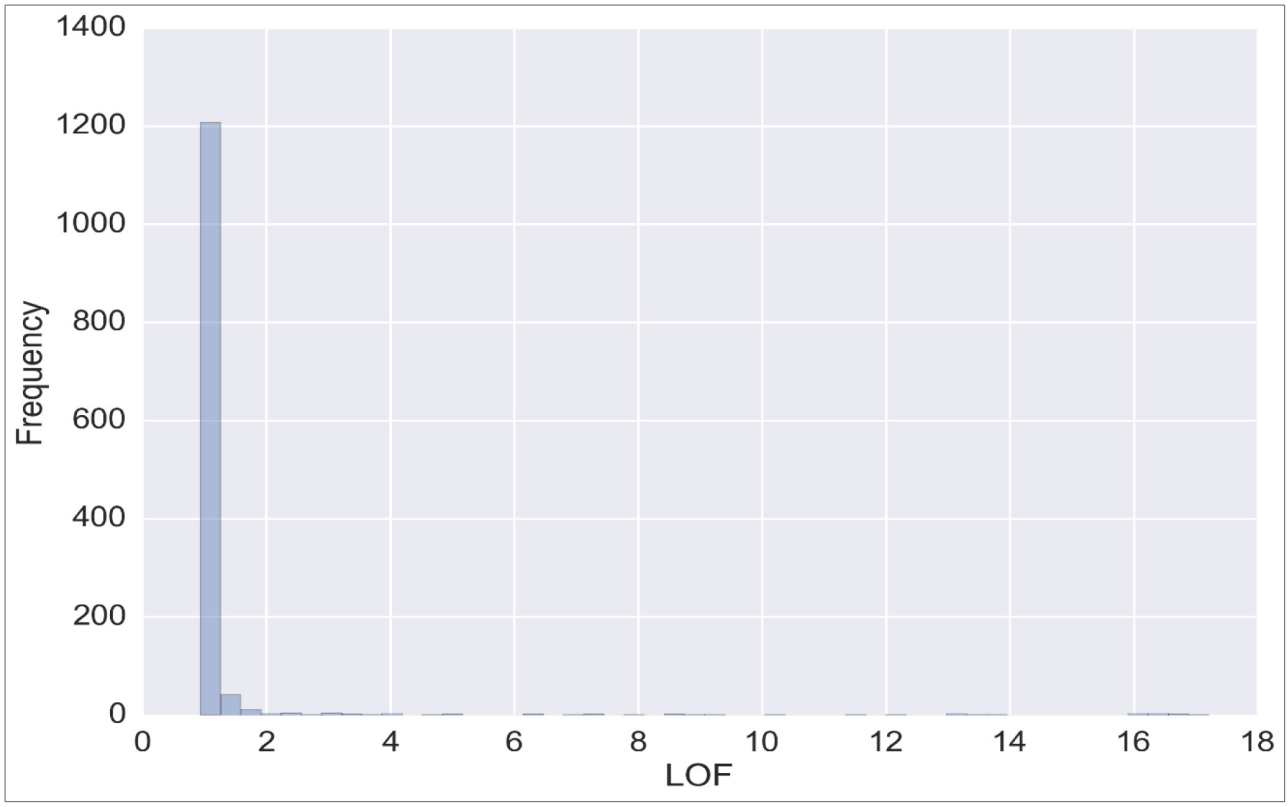

Figure 17—Field example 2 (Woodford well) LOF histogram

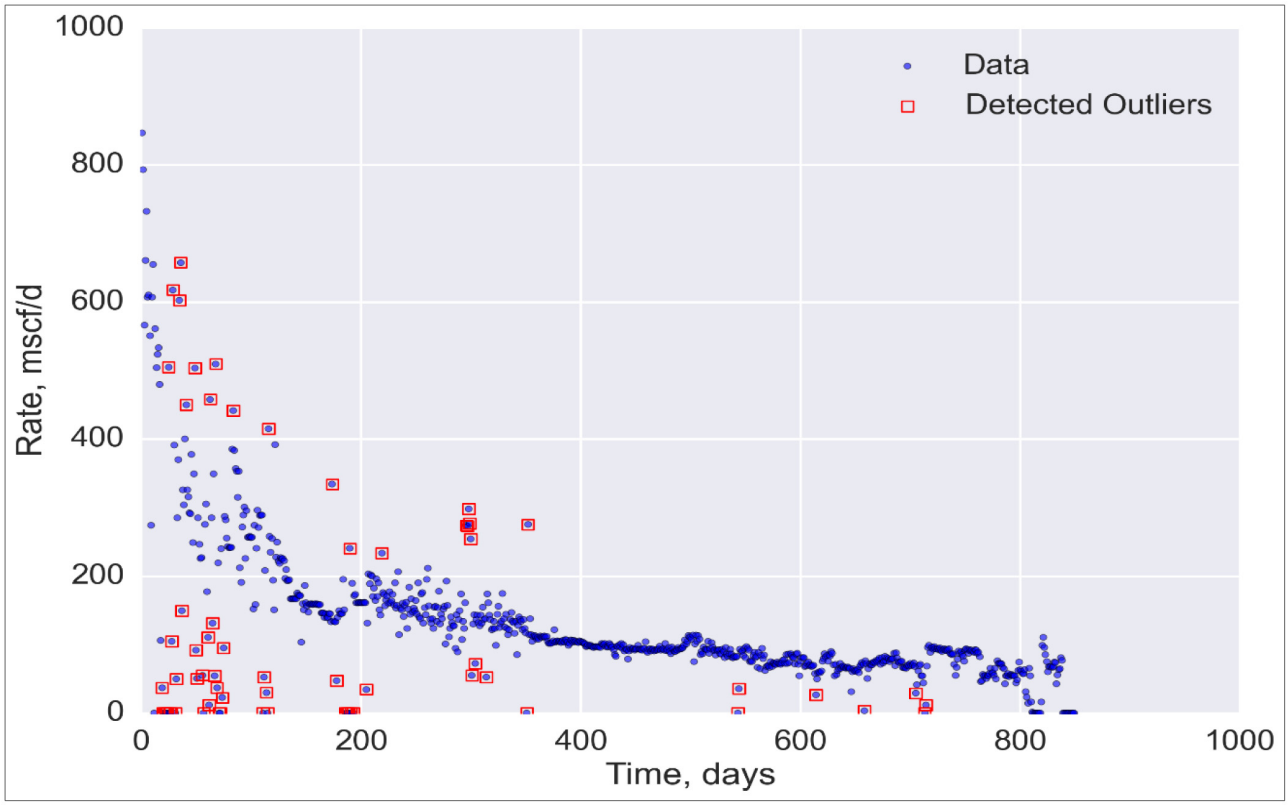

Figure 18—Field example 3 (Bakken well)

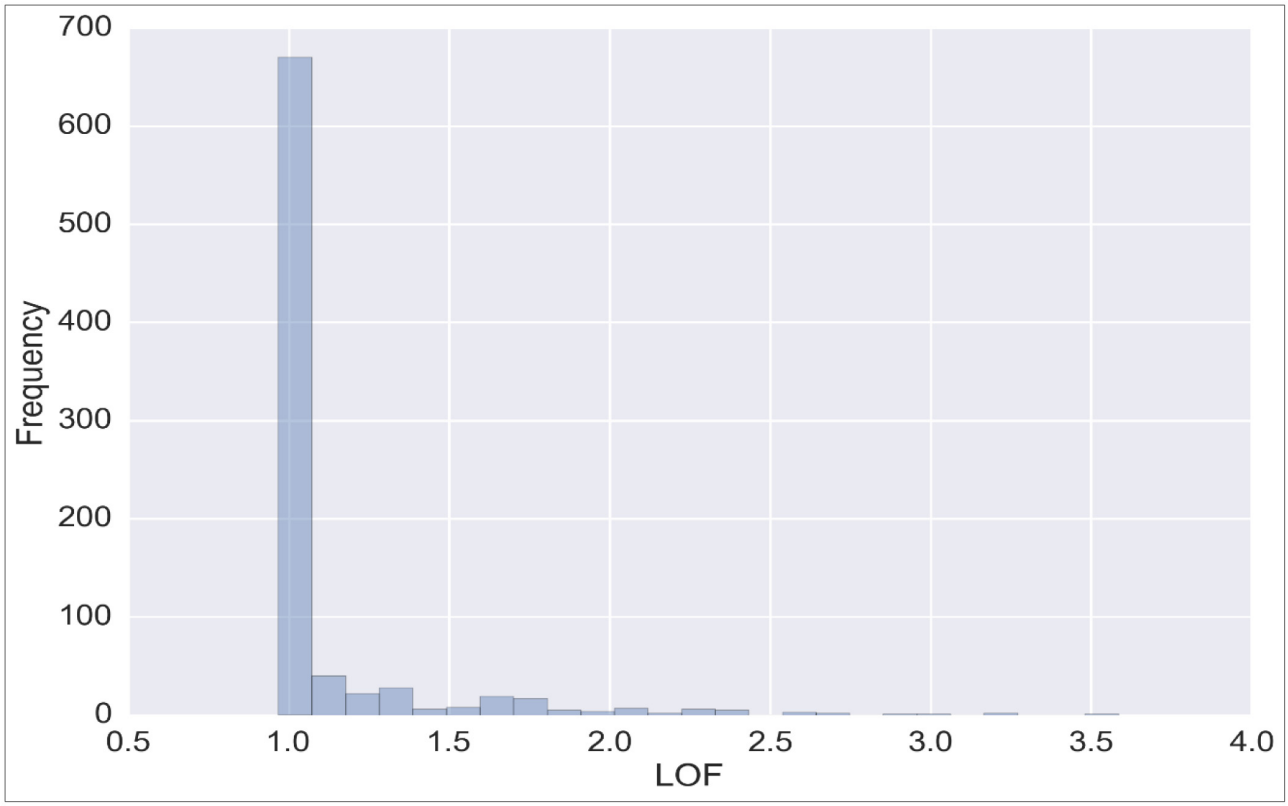

Figure 19—Field example 3 (Bakken well) LOF histogram

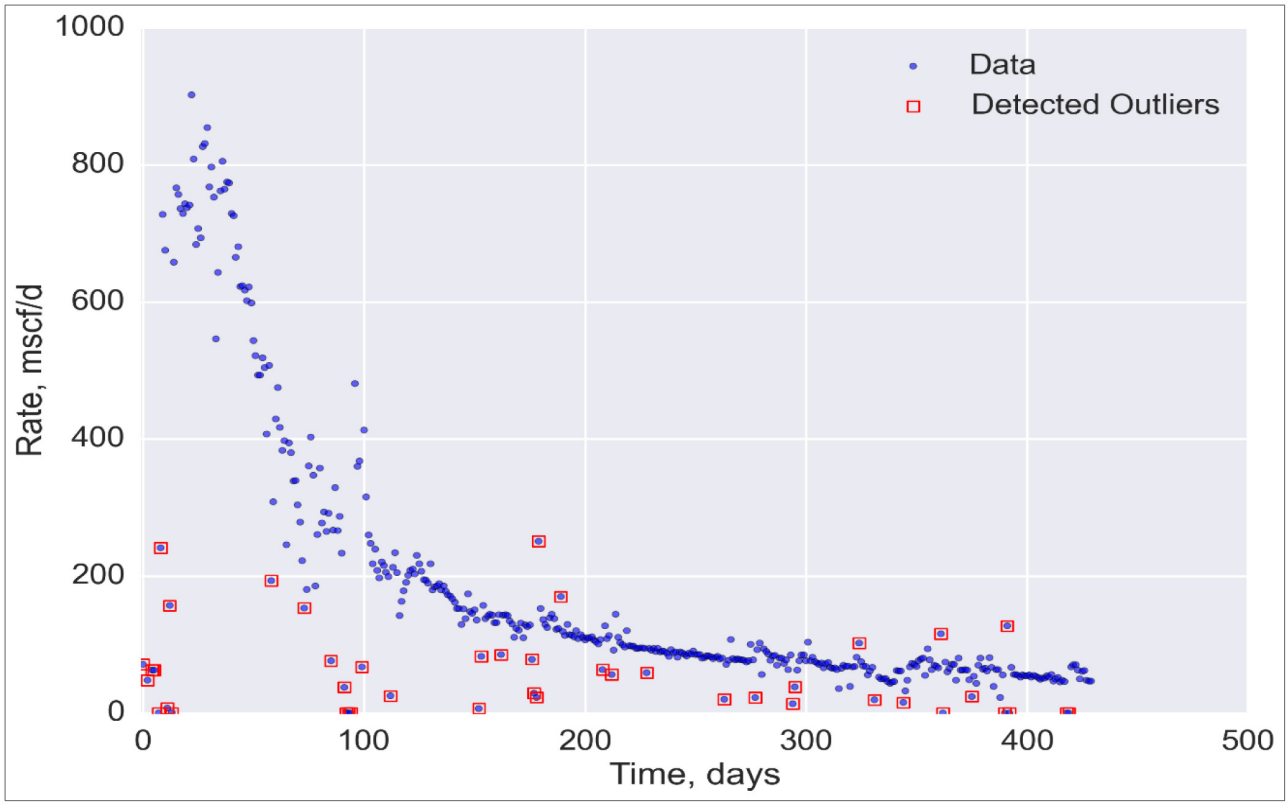

Figure 20—Field example 4 (Eagle Ford well)

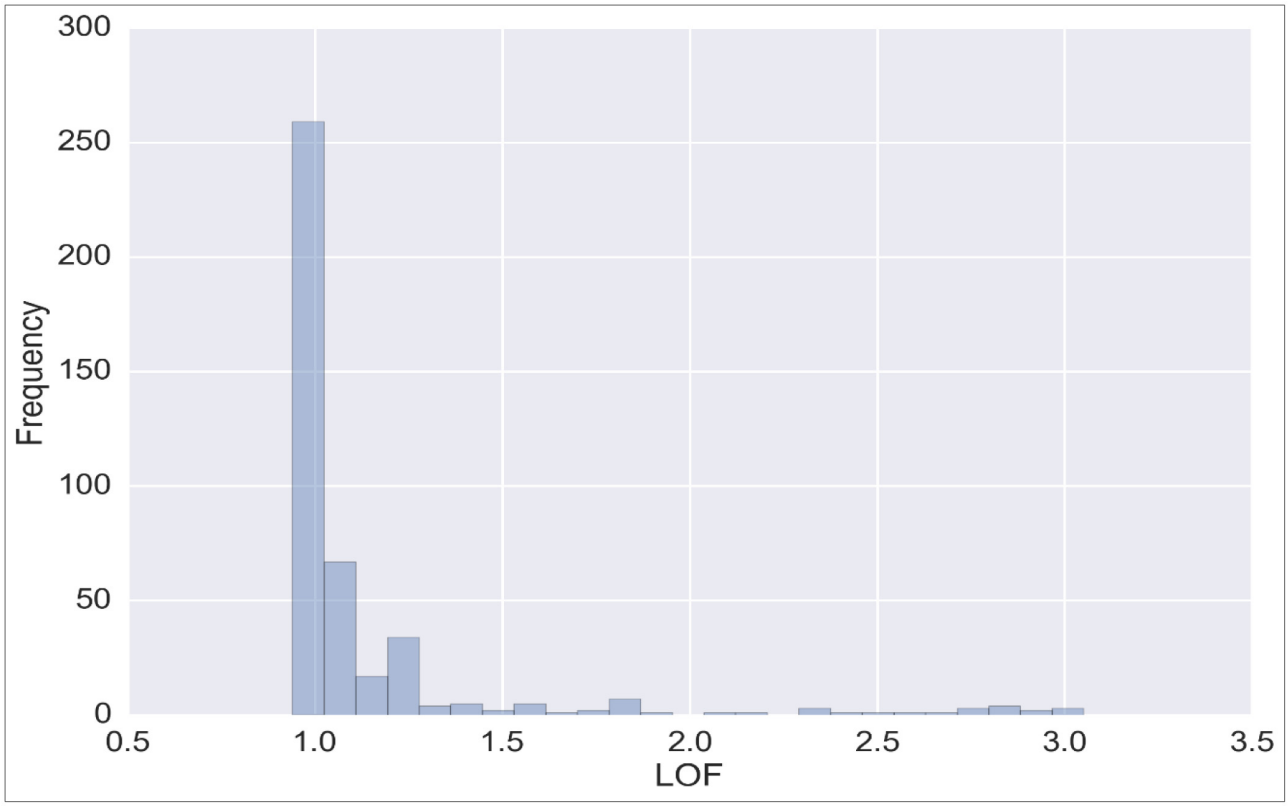

Figure 21—Field example 4 (Eagle Ford well) LOF histogram

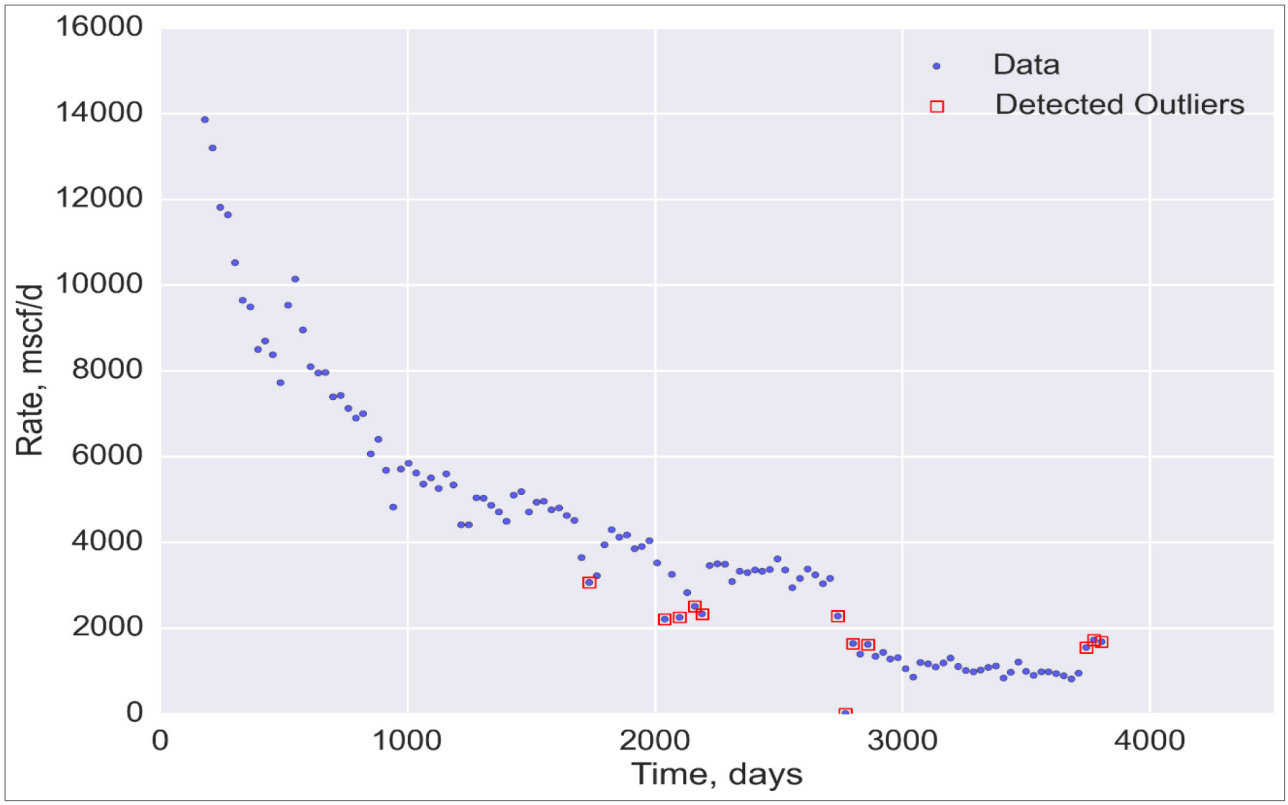

Figure 22—Field example 5 (Barnett well)

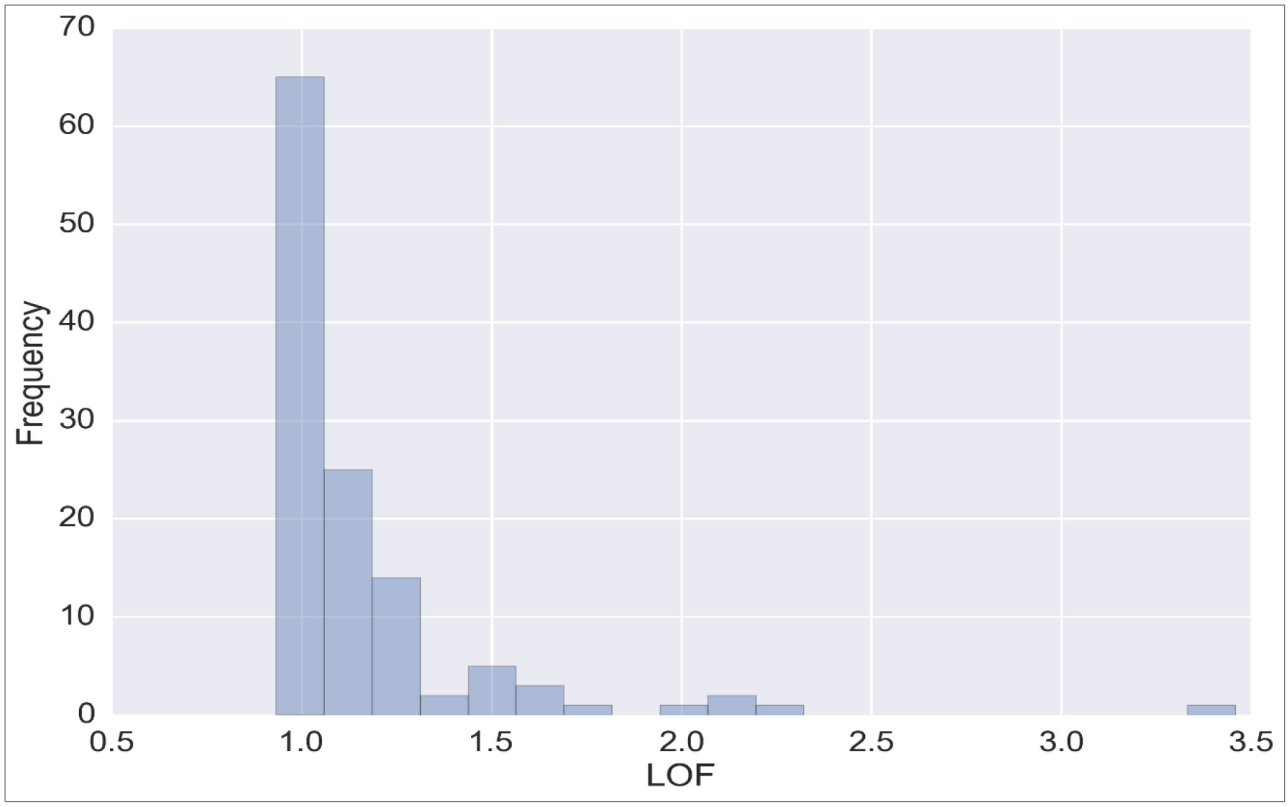

Figure 23—Field example 5 (Barnett well) LOF histogram

The *threshold* value for each case was picked using the histogram of LOF values. The range of values for *threshold* was 1.1 – 1.6. This range should work as a good initial guess for most field cases, modifying the value in steps of . This is required only if one does not want to go through the process of plotting a histogram. Doing this exercise in the tool we have provided is very fast and simple.

Conclusions

- Presence of outliers in the flow rate and pressure data can lead to incorrect identification of flow regimes, well/reservoir model and incorrect estimate of model parameters. This could in turn lead to incorrect estimate of ultimate recovery from the reservoir.
- Presence of outliers in the data decreases the diagnostic value and reliability of production data analysis workflow.
- None of the existing outlier detection methods are robust enough to be universally applicable without violating their underlying assumptions.
- We proposed a new method to detect outliers based on the Local Outlier Factor (LOF). This new method is almost always applicable to any kind of rate/pressure data set without violating its underlying assumptions.

Acknowledgements

We thank the companies, which must remain unnamed for reasons of data confidentiality, for providing the field data used in this research work.

References

Barnett, V. and Lewis, T. 1994. *Outliers in Statistical Data,* third edition, New Jersey: Wiley.

Breunig, M.M., Kriegel, H.P., Ng, R.T., and Sander, J. 2000. LOF: Identifying Density-based Local Outliers. *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* **29** (2): 93–104. https://dx.doi.org/10.1145%2F335191.335388

Brown, M., Ozkan, E., Raghavan, R., et al. 2011. Practical Solutions for Pressure-Transient Responses of Fractured Horizontal Wells in Unconventional Shale Reservoirs. *SPE Res & Eval Eng* **14** (06): 663–676. SPE-125043-PA. http://dx.doi.org/10.2118/125043-PA

Hawkins, D. 1980. *Identification of Outliers,* first edition: Springer Science+Business Media B.V. http://dx.doi.org/10.1007/978-94-015-3994-4

Holdaway, K.R. 2014. *Harness Oil and Gas Big Data Analytics,* first edition, New Jersey: Wiley.

Johnson R. 1992. *Applied Multivariate Statistical Analysis,* third edition: Prentice Hall.

Pedregosa, F., Varoquaux, G., Gramfort, A., et alet al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning and Research* **12**: 2825–2830.

Stalgorova, E. and Mattar, L. 2012. Practical Analytical Model to Simulate Production of Horizontal Wells with Branch Fractures. Presented at the SPE Canadian Unconventional Resources Conference, Calgary, Alberta, Canada, 30 October-1 November. SPE-162515-MS. http://dx.doi.org/10.2118/162515-MS

Stalgorova, E. and Mattar, L. 2013. Analytical Model for Unconventional Multifractured Composite Systems. *SPE Res Eval & Eng* **16** (03): 246–256. SPE-162516-PA. http://dx.doi.org/10.2118/162516-PA

Thompson, D.J. 1994. Jackknifing multiple window spectra. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing **VI**: 73–76. http://dx.doi.org/10.1109/ICASSP.1994.389899