

Pre-processing noisy production data using outlier detection methods reduces uncertainty in identifying flow regimes

Felix A. Munthe

felixmunthe@gmail.com

Texas A&M University <https://orcid.org/0009-0001-1122-7919>

W. John Lee

Texas A&M University <https://orcid.org/0000-0003-3748-0653>

Research Article

Keywords: outlier, rate-normalized pressure, flow regime, machine learning, rate transient analysis

Posted Date: April 10th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9361761/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Pre-processing noisy production data using outlier detection methods reduces uncertainty in identifying flow regimes

Felix A. Munthe ^{a,1,*} and W. John Lee ^{a,2}

^a Harold Vance Department of Petroleum Engineering, Texas A&M University, 3116 TAMU, Texas, USA

ARTICLE INFO

Keywords: outlier, rate-normalized pressure, flow regime, machine learning, rate transient analysis

Authorship contribution statement

Author 1: Conceptualization, data curation, formal analysis, funding acquisition, methodology, resources, software, visualization, writing – original draft. Author 2: Supervision, validation, writing – review & editing.

* Corresponding author

ABSTRACT

This study presents machine-learning-based outlier-removal techniques as a preprocessing step for noisy production data (rate and pressure). Applying outlier detection (OD) methods helps reduce uncertainty in flow-regime identification and preserve the unique profiles of these regimes. These methods are essential for addressing anomalies in production data and preventing inaccuracies in the extraction of reservoir properties.

Flow-regime identification involves constructing diagnostic tools such as the rate-normalized pressure (RNP) plot, which is commonly used in rate-transient analysis (RTA). In this study, we simulated a synthetic log-log RNP profile for a fractured, low-permeability gas reservoir and grouped the RNP data by flow regime. We then added noise and outliers to selected RNP data groups across seven noise-presence scenarios using additive white Gaussian noise (AWGN) with three noise-contamination levels, followed by the OD process. We evaluated 20 OD methods using performance metrics including mean absolute error (MAE), accuracy, precision, recall, and F1 score.

The F1 score combines precision and recall to assess the effectiveness of OD methods. Although a method may perform well in one scenario and less effectively in another, the Gaussian mixture model (GMM) consistently outperformed the other methods, achieving an average F1 score of 0.9. The optimal input parameters for GMM were a threshold factor between 0.8 and 1.1 and a component count between 7 and 10. We also performed sensitivity calculations for these input parameters using field data. In the synthetic cases, GMM produced cleaner RNP profiles and achieved an accuracy of 93.99%.

This work expands previous studies of OD methods applied to RNP data by evaluating 20 methods across various noisy data scenarios stratified by flow regime. Our findings highlight robust OD methods that can reduce subjectivity and time consumption while improving the precision of flow-regime profiles, leading to more accurate production forecasting and EUR estimation.

1. Introduction

This study investigates the ability of machine-learning-based outlier detection (OD) methods to detect outliers in a major diagnostic tool, the log-log rate-normalized pressure (RNP) vs. time plot. Removing unnecessary or anomalous data is essential in preliminary production analysis because it reduces the uncertainty in identifying flow regimes from the log-log RNP plot. This work expands previous studies of outlier-removal techniques applied to production data, particularly RNP data in rate-transient analysis (RTA). It also develops different approaches to evaluate the robustness of various OD methods using 210 synthetic production cases to identify a more suitable model for handling field production data. These approaches can reduce uncertainty in determining the start and end of flow regimes, which in turn can improve the accuracy of Arps b -factor values (Pratikno et al., 2021) by limiting distortion in the curve-fitting process of the decline profile. In this paper, we explain the concept of OD methods, evaluate each method's performance on synthetic data, and then apply the proposed method to field data.

We integrate outlier detection into production analysis to support production data preprocessing. Production analysis, such as RTA, relies on production data, such as pressure, rate, and time, to provide information that engineers can interpret to infer reservoir behavior or characteristics. Therefore, data quality strongly affects the clarity and uniqueness of the interpretation. For example, in log-log diagnostic plots, an RNP vs. time plot may yield a straight-line profile with a 0.5 slope followed by an upward-bending trend, indicating linear flow followed by intermediate flow and boundary-dominated flow (BDF) regimes (Lee, 2021).

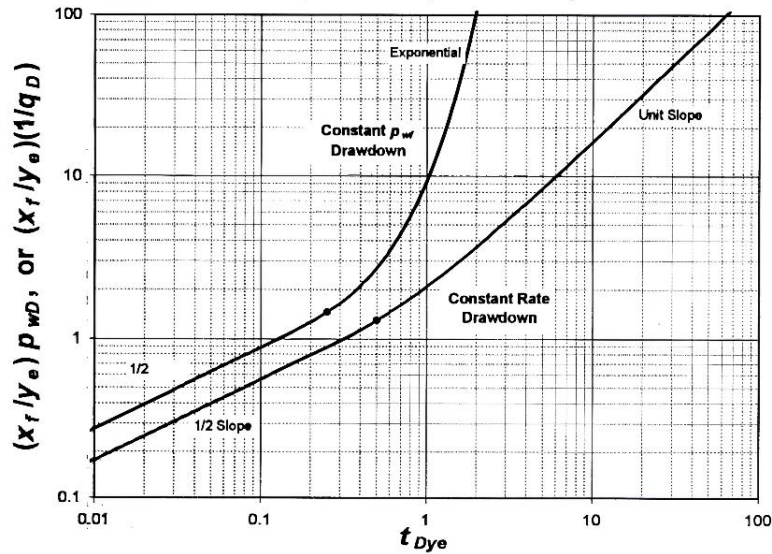


Fig. 1. Linear flow profile for a fractured tight gas reservoir under constant-rate and flowing-bottomhole-pressure (FBHP) conditions (Wattenbarger et al., 1998).

In production analysis, particularly in RTA, data quality plays a significant role (Clarkson, 2013). However, in practice, engineers often work with poor-quality production records due to operational issues. These data-quality problems generally involve two types of anomalies: noise and outliers. Noise typically arises from minor measurement errors, sensor limitations, or environmental variations such as temperature changes or minor shifts in reservoir conditions. In contrast, outliers are often caused by significant operational factors such as well shut-ins, equipment failures, or human error. Noise causes data to deviate within an acceptance range of the population mean, whereas outliers deviate far from the overall data population range (Chaudhary and Lee, 2016).

Compared with pressure-transient analysis (PTA), which has more controlled conditions, the continuous nature of RTA data collection makes it more susceptible to these anomalies. The presence of numerous outliers can reduce the accuracy of the interpretation and make the results uncertain and non-unique. In such cases, detecting and removing outliers can help reduce uncertainty in the analysis. Our study addresses the challenges posed by noise and outliers by examining seven scenarios in which they may occur in our production data. The proposed OD method is particularly well-suited to handle outliers caused by operational anomalies, but it is generally less effective at handling continuous noise, which tends to remain within the expected data range, or complex outliers caused by human error or mislabeling, which may not exhibit sufficient statistical separation for detection.

Chaudhary and Lee (2016) evaluated outliers in production-decline profiles and proposed the local outlier factor (LOF) method to detect and remove outliers. This proximity-based OD technique combines k-nearest neighbors (k-NN) and density-based approaches to identify data points that are less dense than their neighbors and are therefore classified as outliers. Shekhar Jha et al. (2021) conducted a comparative study of five different OD methods on the log-log RNP plot and found that the angle-based outlier detection (ABOD) method performed best for removing outliers from the RNP profile. Their work introduced noise by dividing RNP data into 32 time-based groups, randomly selecting 16 and adding noise to 40% of the data in each selected group. Building on this foundation, our study expands the noise percentage in the data to represent three levels of production-data uncertainty: low, moderate, and high.

We found that the approach of Shekhar Jha et al. (2021) can still introduce uncertainty because of the way noise was selected and added. In this work, we propose grouping data by flow regime for noise selection and adding additive white Gaussian noise (AWGN). These techniques provide a more systematic basis for evaluating the robustness of OD methods in removing outliers from the RNP profile for flow-regime identification. This combination of techniques advances the application of OD methods and establishes a new basis for preprocessing production data quality in RTA.

The following section reviews the four top-performing OD methods from different categories and demonstrates their application to both synthetic and field data.

2. Method

This section presents the best-performing outlier detection (OD) methods from the different model categories among the twenty methods evaluated in this study (Agarwal and Gupta, 2021). The selected techniques achieved relatively high F1 scores for outlier detection in the synthetic RNP data. In addition, most of these algorithms are widely used,

easy to implement, and commercially available. At the end of this section, we demonstrate the application of these methods to the synthetic data.

2.1. Angle-Based Outlier Detection (ABOD)

Angle-based outlier detection (ABOD) is a proximity-based OD method. Its main objective is to identify outliers as data points that are isolated from the main data population or located in sparsely populated neighborhoods (Aggarwal, 2017). Kriegel et al. (2008) introduced ABOD as a method for detecting anomalies in general data distributions by evaluating the angles formed between an observed point and different pairs of surrounding points. As shown in **Fig. 2**, an outlier (point A) exhibits less variation in its spectrum of angles than an inlier (point B).

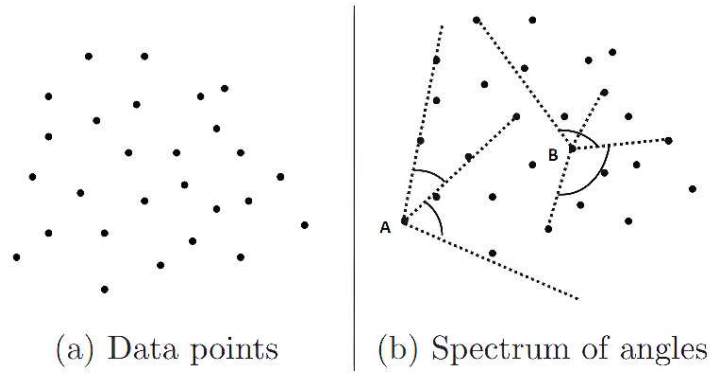


Fig. 2. ABOD concept. (a) A random set of example data points. (b) The outlier (point A) has limited variation in the spectrum of angles compared to the inlier (point B) when measured between different pairs of surrounding points (Aggarwal, 2017).

Consider three data points as $\bar{X}$, $\bar{Y}$, and $\bar{Z}$, then we can calculate the angle (weighted cosine) of $\bar{X}$ as:

$$W \cos(\bar{Y} - \bar{X}, \bar{Z} - \bar{X}) = \frac{\langle \bar{Y} - \bar{X}, \bar{Z} - \bar{X} \rangle}{\|\bar{Y} - \bar{X}\|_2^2 \cdot \|\bar{Z} - \bar{X}\|_2^2}. \quad (1)$$

The $\|\cdot\|_2$ denotes the L_2 -norm, and the $\langle \cdot \rangle$ denotes the scalar product. **Eq. (1)** shows that inverse distance weighting reduces the weighted angle values for outlier points, which in turn limits the variance of their angle spectrum (Aggarwal, 2017). Kriegel et al. (2008) referred to this variation as the angle-based outlier factor (ABOF), where smaller ABOF values indicate a higher likelihood that a point is an outlier:

$$ABOF(\bar{X}) = Var_{\{Y,Z\}} W \cos(\bar{Y} - \bar{X}, \bar{Z} - \bar{X}). \quad (2)$$

2.2. Gaussian Mixture Model (GMM)

The Gaussian mixture model (GMM) is a probabilistic-based OD method. It is one of the earliest statistical approaches used for anomaly detection. The fundamental principle of probabilistic methods is to represent data as a probability

distribution, in which model parameters are learned iteratively using algorithms such as expectation-maximization (EM) or maximum a posteriori (MAP) estimation (Aggarwal, 2017; Reynolds, 2009). **Fig. 3** illustrates the probability density function (PDF) of a dataset comprising two mixtures, in which the separation between outliers and the main distribution is sufficiently clear.

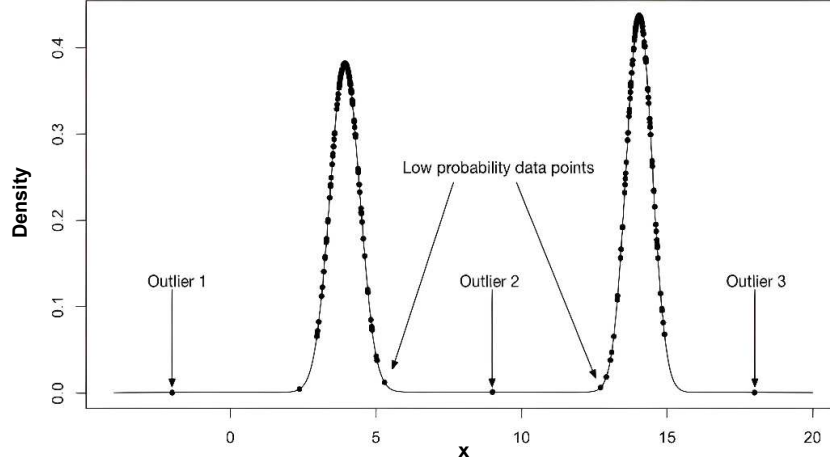


Fig. 3. GMM concept. The PDF separates the normal distribution of the two mixtures from outliers (Reddy et al., 2017).

GMM represents the data as a weighted sum of M Gaussian component densities, where each component corresponds to a class or subpopulation within the data (Reddy et al., 2017; Reynolds, 2009),

$$p(x|\lambda) = \sum_{i=1}^M \omega_i g(x|\mu_i, \sigma_i). \quad (3)$$

where μ_i is the mean, σ_i is the variance, and λ is the mixture weights for all densities. Each Gaussian density is defined as:

$$g(x|\mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}, \quad (4)$$

Reddy et al. (2017) applied a logarithmic transformation to enhance the contrast between outliers and low-permeability data in Fig. 3.

$$OS_x = (\log(p(x)))^{2f}, \quad (5)$$

where OS_x is the outlier score term, $p(x)$ is the PDF of the data point x , and f is the scale factor for the log probability. In this study, however, we define outlier score using the negative log-likelihood:

$$p(\lambda|x) = -\log \left(\sum_{i=1}^M \omega_i g(x|\mu_i, \sigma_i) \right) / p(x), \quad (6)$$

Then we applied a threshold parameter to the negative log-likelihood distribution to classify data points as outliers.

2.3. Isolation Forest (iForest)

Isolation forest (iForest) is an ensemble-based OD method. Its main principle is to identify anomalous points by isolating them through a collection of randomly generated partitioning structures, thereby improving detection performance through ensemble learning (Aggarwal, 2017; Cheng et al., 2019). Liu et al. (2008) introduced iForest as a method for detecting anomalies by separating a data point from the rest of the dataset using random partitions, referred to as isolation trees (iTree). The number of partitions required to isolate a point corresponds to the path length in the tree structure. As illustrated in **Fig. 4**, an outlier (point ' x_o ') generally requires fewer random partitions and therefore has a shorter path length than an inlier (point ' x_i ').

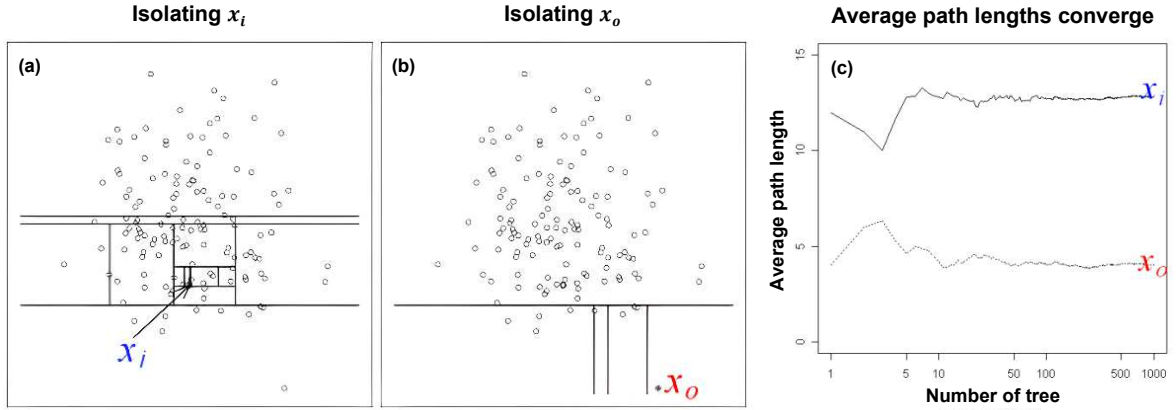


Fig. 4. iForest concept. (a) Isolating an inlier (point x_i) requires more random partitions. (b) Isolating an outlier (point x_i) requires fewer random partitions. (c) Inlier (point x_i) produces a greater average path length, whereas outlier (point x_o) produces a smaller average path length (after Liu et al., 2008).

Liu et al. (2008) showed that the average path length of an iTree ($h(x)$) is equivalent to the average path length of an unsuccessful search in a Binary Search Tree (BST) ($c(n)$), and we can calculate it as follows:

$$c(n) = 2H(n-1) - (2(n-1)/n), \quad (7)$$

where n is the number of dataset points, and $H(i)$ is the harmonic number equivalent to $\ln(i) + 0.5772156649$ (Euler's constant). We can calculate the anomaly score (s) of a point x as:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}, \quad (8)$$

where $E(h(x))$ is the average of $h(x)$ from an iTree structure. Data points with anomaly scores close to 1 indicate anomalies, whereas points with scores below 0.5 indicate normal points. Scores near 0.5 indicate that the points are not clearly distinguishable as either anomalies or normal (Liu et al., 2008).

2.4. Minimum Covariance Determinant (MCD)

Minimum covariance determinant (MCD) is a linear-based OD method. Its main principle is to represent the data structure using linear correlations and to identify outliers based on robust estimates of multivariate location and scatter (Aggarwal, 2017; Fauconnier and Haesbroeck, 2009). Compared with the classical covariance-based approach, MCD searches for the subset of observations with the minimum covariance determinant, resulting in more robust parameter estimates in the presence of outliers (Hubert et al., 2018). **Fig. 5** illustrates the difference between the classical covariance ellipse and the robust ellipse obtained using MCD, where points lying outside the MCD ellipse are classified as outliers.

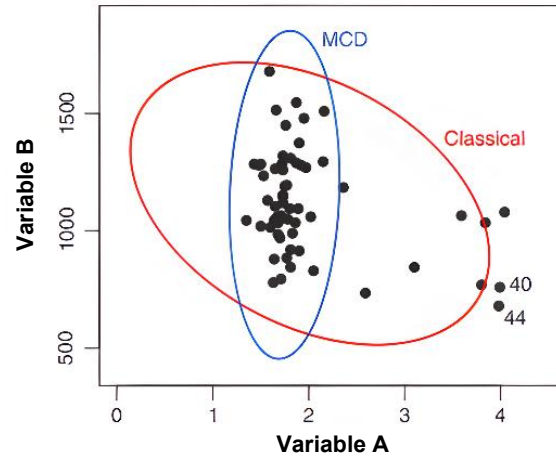


Fig. 5. MCD concept. Comparison between the classical covariance ellipse and the robust MCD ellipse (after Hubert et al., 2018).

In the classical approach, elliptical tolerance is defined using the Mahalanobis distance:

$$MD(x) = d(x, \bar{x}, Cov(X)) = \sqrt{(x - \bar{x})^T Cov(X)^{-1} (x - \bar{x})}, \quad (9)$$

where $\bar{x}$ is the sample mean, and $Cov(X)$ is the sample covariance matrix. In contrast, MCD defines elliptical tolerance using robust distances (RD):

$$RD(x) = d(x, \hat{\mu}_{MCD}, \hat{\Sigma}_{MCD}), \quad (10)$$

where $\hat{\mu}_{MCD}$ is the MCD estimate of location and $\hat{\Sigma}_{MCD}$ is the MCD estimate of covariance. Based on the asymptotic distribution of robust distances, we use the cutoff value of $\sqrt{\chi^2_{p,0.975}}$, with an alternative cutoff based on the F -distribution for a small number of objects (n) and p is the number of variables. Rousseeuw and Driessen (1999) used the distance-distance plot to compare robust distances derived from MCD with classical Mahalanobis distances. As shown in **Fig. 6**, this comparison helps distinguish points identified as anomalous under the robust framework from those evaluated using the classical covariance structure.

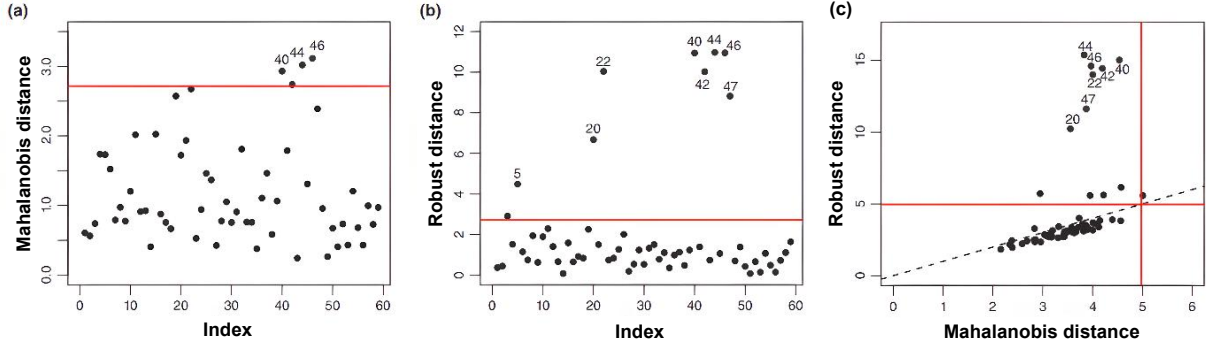


Fig. 6. (a) Example of Mahalanobis (classical) distances. (b) Example of robust distances based on MCD. (c) Distance-distance plot combining robust and Mahalanobis distance results (Hubert et al., 2018).

The following subsections describe the simulation of synthetic RNP data, including procedures for adding noise and outliers. The final subsection then demonstrates the application of the OD methods described above.

3. Synthetic Data

We constructed a synthetic RNP profile based on simulated production data from a homogenous single-layer gas reservoir in a low-permeability formation produced from a single-fractured vertical well. **Table 1** summarizes the simulation property values we used, and **Fig. 7** illustrates our reservoir model and the resulting RNP profile over production time (t) and material-balance time (t_{mb}) (Palacio and Blasingame, 1993).

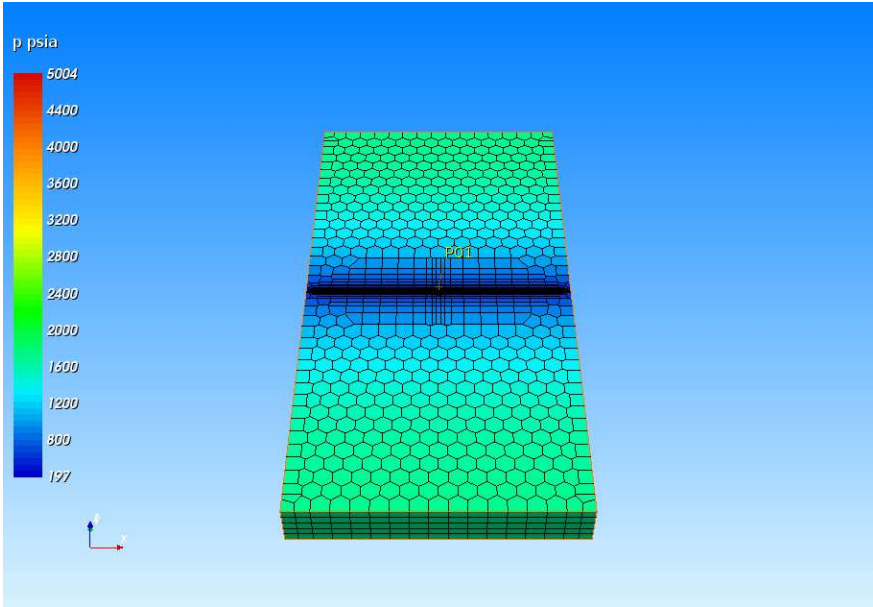
Table 1

Property of the synthetic reservoir model.

Property	Value	Unit
Flowing bottomhole pressure	300	psia
Fracture conductivity	500	md-ft
Fracture half-length	350	ft
Fracture height	100	ft
Fracture porosity	7	%
Gas gravity	0.7	dimensionless
Initial gas saturation	90	%
Initial reservoir pressure	5000	psia
Initial water saturation	10	%

Length	1400	ft
Permeability	0.001	md
Porosity	5	%
Production time	200	year
Reservoir temperature	212	°F
Skin	0	dimensionless
Thickness	100	ft
Well radius	0.4	ft
Width	700	ft

218



219

220

Fig. 7. Synthetic reservoir grid model at $t = 525600$ hrs. (21900 days).

221

222

223

224

225

226

227

228

229

230

We identified three flow regimes using **Fig. 8**: linear flow for $t \leq 4,000$ days, transition flow for $4,000 < t \leq 40,000$ days, and boundary-dominated flow (BDF) for $t > 40,000$ days. Although the simulation assumes constant flowing bottomhole pressure or constant rate for simplicity, this does not limit the method's applicability to actual conditions. Our proposed method can handle various production scenarios because the fundamental flow patterns will still emerge in the diagnostic plot regardless of data complexity or variations. The primary difference in practice is that real-world data often contains more anomalies due to operational factors, which can be filtered using the outlier detection (OD) method. Once outliers and anomalies are addressed, the remaining data will align more closely with the expected flow regime patterns, as shown in **Fig. 8**.

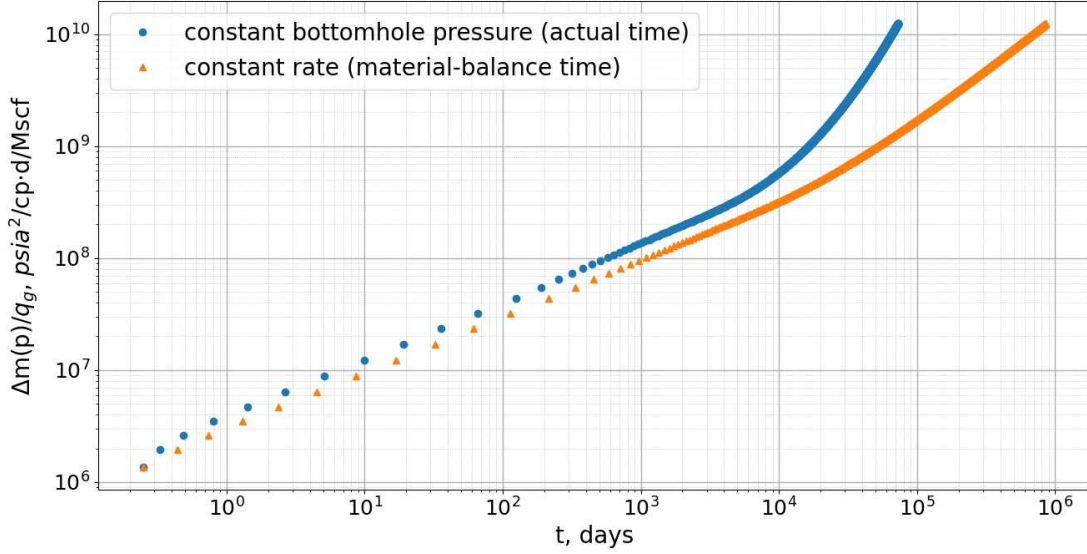


Fig. 8. RNP profile results from the synthetic production simulation.

Fig. 8 represents our synthetic RNP data, in which we label each data value as 0 for our noise-adding and outlier-adding processes.

While using a simplified single-layer model provides clear insights into the performance of the OD method, it does present limitations regarding the generalizability of the results. In more complex reservoir systems, such as multilayer or heterogeneous formations, interactions between layers may cause deviations from the simplified behavior observed in this study. Additionally, multiple wells or variable production rates could introduce further complexities, such as interference effects or changes in reservoir pressure dynamics, which are not captured in this model. Despite these limitations, the proposed OD method analyzed using this synthetic data remains applicable in more complex scenarios, as the underlying principles of flow regime behavior, characterized by a half-slope for linear flow and a unit-slope for BDF in the log-log RNP-time plot, are consistent and experienced by wells across reservoir types. Future work could focus on testing in more complex synthetic models to add additional validation.

3.1. Additive White Gaussian Noise

The additive white Gaussian noise (AWGN) technique selects random RNP data points and replaces them with noise or outliers. This technique takes a random standard deviation multiplier, ranging from -3 to 3, and multiplies it by the mean data value. Since each time step contains only one RNP data point, the mean RNP value of each time point is its RNP data point. So, there is no correlation between individual data each time with others, meaning we applied the AWGN procedure independently point by point. This adjustment eliminates biased results when applying the GMM method by generating random noise or outliers, avoiding any specific distribution patterns, such as Gaussian or normal, in the synthetic noisy production data. We chose the range of -3 to 3 for the standard deviation multiplier because, in a normal distribution, 99.7% of the data points fall within 3 standard deviations of the mean. This ensures that the

noise added is within a reasonable range, reflecting typical deviations without introducing extreme or unrealistic values with enough variability to test the robustness of the OD method effectively.

In this study, we used actual time as the basis for the time variable (Jha and Lee, 2017). We considered three levels of noise contamination, 25%, 50%, and 75%, and repeated the process, resulting in ten noisy RNP data sets for each level. We selected these noise levels based on the desire to simulate more challenging conditions that reflect the operational anomalies potentially encountered in real-world scenarios. **Table 2** shows the flow regime combination scenarios to add noise and outliers to our true RNP data. Each scenario includes three noise-contamination levels, with ten repetitions for each level. While we did not explicitly test noise levels below these defined ranges, such as 5%, these cases were indirectly included in each contamination category. The 25%, 50%, and 75% levels serve as upper limits to test the robustness of the proposed method. This means that when the OD method can handle RNP data with a 25% noise level, it will also perform well for noisy data under this level.

Table 2

Scenarios to add noise and outliers in RNP data, with associated numbering data set.

Scenario	Noise Level (Respectively)	Noisy RNP Data Set Numbering
Linear flow	25%, 50%, 75%	1-10, 11-20, 21-30
Transition flow	25%, 50%, 75%	31-40, 41-50, 51-60
Boundary-dominated flow (BDF)	25%, 50%, 75%	61-70, 71-80, 81-90
Linear + transition flow	25%, 50%, 75%	91-100, 101-110, 111-120
Transition + BDF	25%, 50%, 75%	121-130, 131-140, 141-150
Linear + BDF	25%, 50%, 75%	151-160, 161-170, 171-180
Linear + transition + BDF	25%, 50%, 75%	181-190, 191-200, 201-210

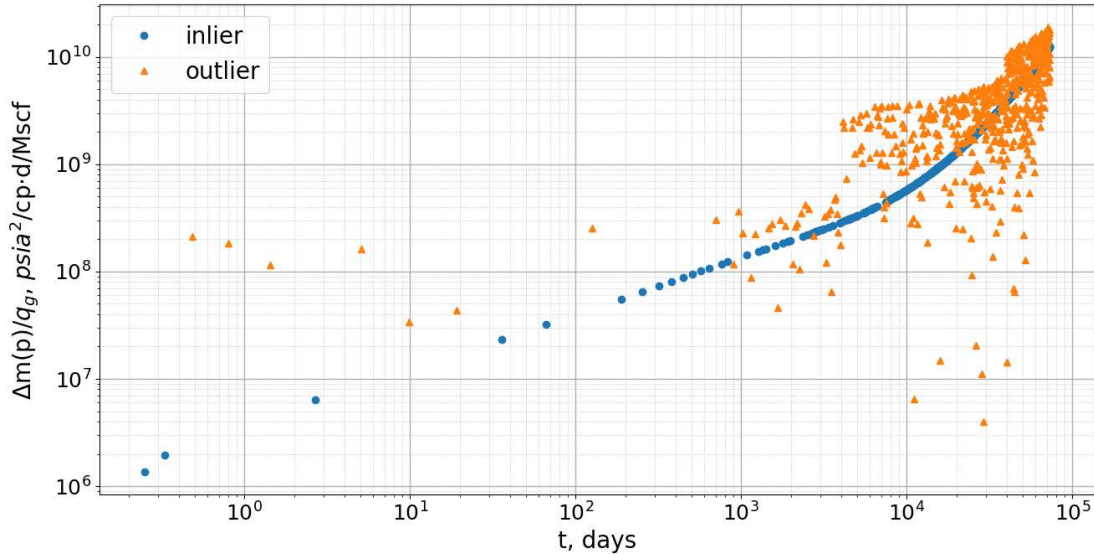


Fig. 9. Noisy RNP data set sample 197, scenario linear + transition + BDF with 50% noise contamination level.

We labeled each noise and outlier added to the true RNP data with a value of 1. We used the 210 noisy RNP data set to evaluate twenty OD methods and compared the performance of each OD result using performance metrics.

3.2. Hyperparameter Tuning.

Defining the hyperparameter values for each OD method before the learning process of machine learning begins is a crucial step. Hyperparameter controls the learning process itself and can significantly affect the performance of the OD models, unlike model parameters, which are learned during training. Tuning OD hyperparameters involves finding the optimal set of hyperparameters that results in the best model performance. **Table 3** shows the set of hyperparameter combinations for each OD described in the previous section. We provide a complete set of hyperparameters for all OD methods in Appendix A—Complete Set of Hyperparameter Combinations of Outlier Detection (OD) Methods.

Table 3

Set of hyperparameter combinations for the four OD methods described in the previous section.

OD Method	Hyperparameter	Optimal Value
ABOD	n_neighbors; threshold_factor	4; 0.1
GMM	threshold_factor; n_components	0.8 – 1.1; 7 – 10
iForest	contamination	0.1 – 0.5
MCD	support_fraction	0.4 – 0.9

Selecting the right combination of hyperparameters is essential for accurately predicting outliers in our RNP data, as incorrect settings may lead to underfitting or overfitting. We implemented the grid search technique to optimize hyperparameter selection, systematically exploring a given range of input values to identify the best-performing set. This method enhances model performances by ensuring the hyperparameters are finely tuned to the data.

Based on a trial and error approach, we initially selected a broad range of values for each hyperparameter, such as the number of components in GMM and contamination levels in iForest. We started with a wide range and iteratively narrowed the ranges to more specific values that can still deliver the best performance result for each method. For example, the number of components in GMM was initially tested from 1 to 30, eventually settling on 7 to 10 components after observing significant improvements in performance within this range. Similarly, we fine-tuned the threshold factor based on the model's sensitivity to anomalies in the RNP data. This trial and error process allowed us to achieve a balance between precision and recall, ensuring that the models accurately identified outliers while minimizing false predictions.

3.3. Performance Metrics.

We calculated performance metrics such as mean absolute error (MAE), accuracy, precision, recall, and F1 score after identifying the number of correct and incorrect detections. Precision ignores false negative values (outliers classified as inliers), whereas recall ignores false positive values (inliers classified as outliers) (Jha et al., 2022). Evaluating precision or recall alone can result in biased interpretations, as high values may indicate significant false positives or false negatives, respectively. Hence, we utilized the F1 score as a comprehensive model evaluation index because it balances both metrics, providing a single metric that summarizes its performance. The F1 score is the harmonic mean of precision and recall, which is one of the important evaluation metrics for scoring and comparing various OD method performances (Powers, 2020). An F1 score reaches its best value at 1 and worst at 0.

$$MAE = \sum_{i=1}^n \frac{|y_d - y_r|}{n}, \quad (11)$$

$$Accuracy = \frac{TP + TN}{Total\ number\ of\ samples}, \quad (12)$$

$$Precision = \frac{TP}{TP + FP}, \quad (13)$$

$$Recall = \frac{TP}{TP + FN}, \quad (14)$$

$$F1 - score = 2 \times \frac{1}{\frac{1}{precision} + \frac{1}{recall}}, \quad (15)$$

where n is the number of samples, y_d is the data label value detected, y_r is the data label value reference, true positive (TP) is the sum of correct outliers detected using the OD method, false positive (FP) is the sum of incorrect outliers detected using the OD method, true negative (TN) is the sum of correct inliers detected using the OD method, and false negative (FN) is the sum of incorrect inliers detected using the OD method.

We successfully evaluated all twenty OD methods. However, for comparison and analysis purposes, we focused only on the most optimal performance of the OD method from each category. This selection is in line with the concepts discussed in the previous section. The following section will detail the performance results of these top OD methods.

4. Results and Discussion

We utilized the F1-score result from 210 noisy data set scenarios to compare OD methods within the same model category. The F1-score helps summarize performance metrics results where a value close to one means the OD method has high precision and recall values. These values are associated with the ability of the OD method to detect outliers from a given data set correctly. **Table 4** shows the performance metric results of the four OD methods, with the highest F1-score from each model category. We provide complete results for all OD methods in Appendix B—Complete Performance Metric Results of Outlier Detection (OD) Methods.

Table 4

Performance metric results for the four OD methods with the highest F1-score in each model category.

OD Method	TP (%)	FN (%)	TN (%)	FP (%)	MAE	Accuracy	Precision	Recall	F1-score
ABOD	52.759	47.241	76.495	23.505	0.278	0.722	0.475	0.528	0.490
GMM	90.227	9.773	87.848	12.152	0.060	0.940	0.920	0.902	0.901
iForest	68.191	31.809	82.129	17.871	0.219	0.781	0.558	0.682	0.582
MCD	82.984	17.016	63.661	36.339	0.289	0.711	0.475	0.830	0.563

Note that the results shown in **Table 4** are the values after averaging the 210 results. The comparison of the F1-scores in **Table 4** shows that GMM has the highest F1-score value, indicating that this method provides high confidence in

detecting outliers in RNP data. GMM also provides high-accuracy results for predicting outliers and inliers compared to other methods. From given outliers in the RNP data set, GMM can detect them correctly an average of 90.23% of the time. This method also performs well in detecting inliers, averaging 87.85% correct determinations. This number is the best score for TN results compared to other methods.

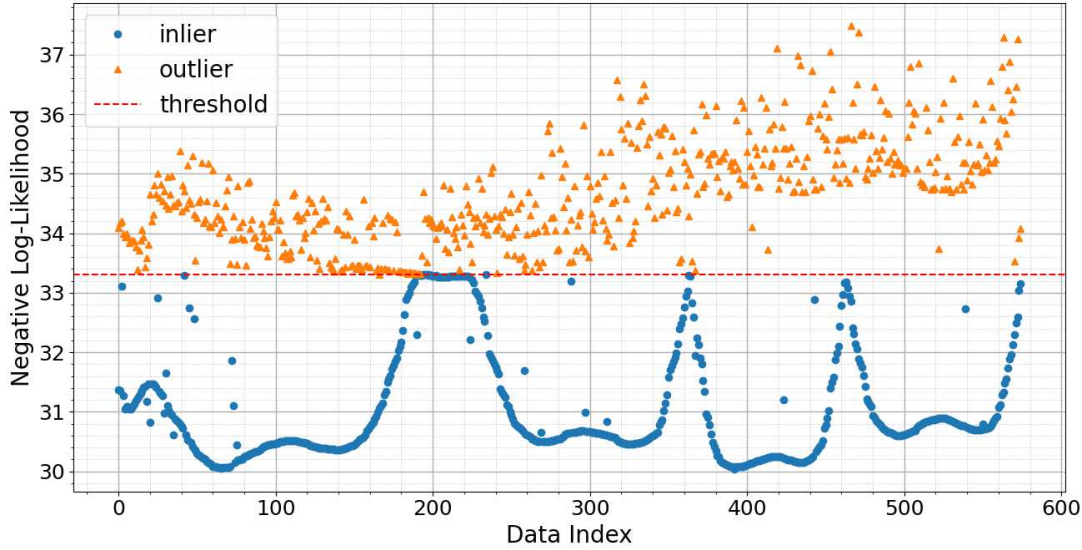


Fig. 10. Negative log-likelihood of GMM score for data set sample 197 with F1-score 0.937.

GMM uses the negative log-likelihood to measure how well each data point fits into the GMM model. A higher value indicates that the data point is less likely a part of the model distribution. In addition, appropriate values of input parameters such as threshold factor and number of components (mixtures) are essential to obtain the best result. The threshold factor will constrain the data between inliers and outliers by multiplying the factor by the median negative log-likelihood value. The optimal parameters obtained for GMM are the threshold factor with values ranging from 0.8 to 1.1 and the number of components ranging from 7 to 10. **Fig. 10** shows the GMM score (negative log-likelihood) for data set sample 197 (Fig. 9) with optimal parameter values: threshold factor is one, and number of components is 10. **Fig. 11** presents the RNP data results associated with Fig. 10 after applying the GMM method, and **Fig. 12** illustrates the clean RNP data results.

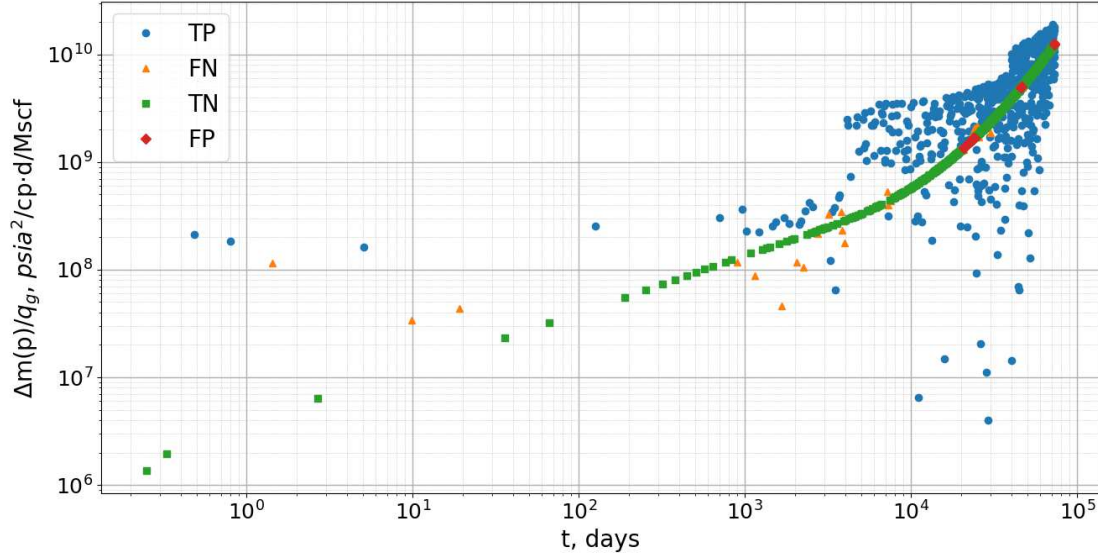


Fig. 11. RNP data results after detecting outliers using GMM for data set sample 197 with TP = 93.74%, FN = 6.26%, TN = 93.74%, FP = 6.26%, MAE = 0.063, Accuracy = 0.937, Precision = 0.937, Recall = 0.937.

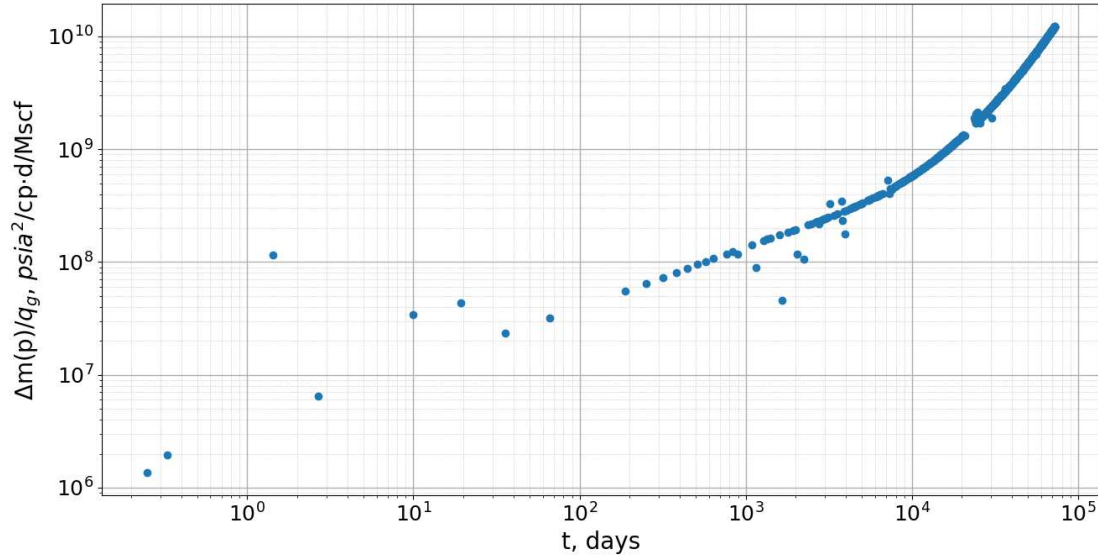


Fig. 12. Clean RNP data results after removing outliers using GMM for data set sample 197.

The GMM method offers a suitable approach for removing outliers from data with a form like an RNP data profile. GMM can have an F1-score of about 0.947 for low noise contamination (25%). GMM evaluates the data by computing the probability density function (PDF) as in Fig. 3. Converting the data into PDF allows GMM to identify the RNP data profile as the distributions representing flow regimes. RNP data from one flow regime corresponds to a PDF distribution. This method is suitable for detecting outliers in RNP data because the nature of the process is similar to that of grouping the data into several flow regime distributions.

Other methods, such as ABOD, iForest, and MCD, produce good results but are not superior with low noise contamination, unlike the GMM method. ABOD relies on the k-Nearest Neighbor and the angle threshold number to

compute the spectrum of angles. However, this method cannot provide high confidence for data such as an RNP profile because data showing a trend will have a limited spectrum of angles between inlier data rather than between sparse data. Like ABOD, the iForest method depends on the number of partitions and average path length, which means that for data sets containing trends, iForest might generate fewer partitions and path lengths than a sparse data set. MCD is a different approach that indirectly evaluates the correlation or interaction between the data and the surrounding data population using robust distances based on the MCD. This approach may perform poorly for RNP data forms because our calculation uses only a robust distance plot as the outlier threshold guide. Adding the Mahalanobis plot to a robust distance plot, also called the distance-distance plot, can improve the result of the MCD method.

After describing the results with each OD method above, with GMM as our best OD method, we quantify the impact of errors caused by noise and outliers on log-log RNP vs. time plots. For example, consider **Fig. 13** as our raw RNP data containing noise and outliers; it is Fig. 9 without identifying inliers and outliers.

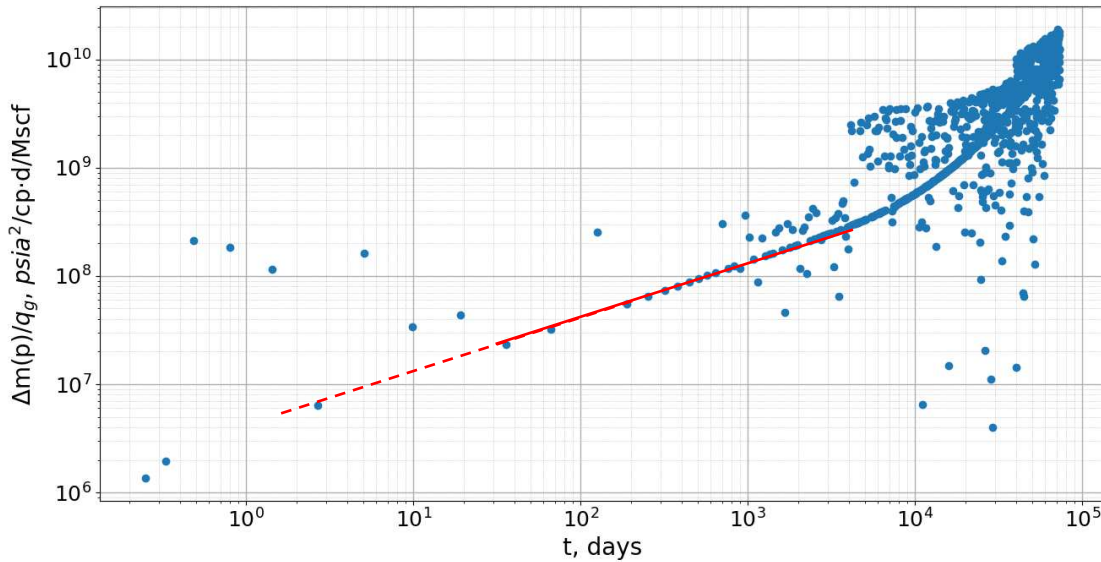


Fig. 13. An example of synthetic raw noisy RNP data points with noise and outliers before applying the GMM method to remove outliers. The red line shows the 0.5 slope that identifies linear flow on the RNP profile. The solid line fits the noisy data points, while the dashed line fits the clean data points in Fig. 12.

Fig. 13 indicates the most likely linear flow regime we can identify spans, in terms of raw noisy RNP data points, in the interval $30 < t \leq 4,000$ days (solid red line). We would probably define the data points earlier than 30 days as early off-trend data because these data are sparse, and we cannot identify clearly when these data points begin to fall on a 0.5-slope trend. Moreover, it is impossible to determine the start and end of BDF, which is usually associated with a hyperbolic decline of $0.3 < b \leq 0.5$. Even if we transform the time axis into material-balance time, it is difficult to detect the start of the BDF regime with a unit slope due to noise and outliers. This incorrect identification of the start and end of flow regimes will cause errors in Arps b -factor determination. In contrast, Fig. 12 is the clean RNP profile of the data in Fig. 13, and we see that the early sparse data is part of the linear trend, starting at $t = 3$ days. After removing outliers, we obtained the most likely linear flow regime that occurs in the interval $3 < t \leq 4,000$ days

(dash red line). Hence, outlier detection, specifically using GMM, significantly helps us to reduce the uncertainty in identifying the start and end time of flow regimes, in this case, the linear flow regime. The BDF regime is also more clearly defined, allowing us to determine an appropriate Arps b -factor value. This Arps b -factor from the BDF regime is crucial as we use it to forecast production and Estimated Ultimate Recovery (EUR). EUR represents the total amount of petroleum expected to be extracted from a well over its lifespan, including already produced and anticipated to be recoverable. This EUR can be predicted using decline curve analysis (DCA), a method introduced by Arps (1945) for analyzing declining production rates based on a graphical procedure (empirical) using the Arps b -factor.

The improved accuracy in flow regime identification can have substantial implications for reservoir management. For instance, by identifying correct flow regimes, the operator can optimize production strategies for existing or undrilled wells, such as adjusting choke settings, modifying stimulation plans, or implementing secondary recovery at the right time. It also helps pinpoint when wells have transitioned to BDF, which is crucial for determining when to initiate enhanced oil recovery (EOR) operations, drill infill wells, or decide about abandonment. These actions lead to more efficient resource allocation and extended well life, thereby impacting the reservoir's overall productivity. Additionally, the economic benefit of using the proposed OD method for production data pre-processing is that the production forecasts are more reliable, resulting in confidence for reserve estimates and financial projections. This can lead to better capital investment decisions, more structured budgeting, and improved financial planning for both short-term and long-term production strategies. Furthermore, better production forecasts help set more reliable well performance benchmarks, reducing risks of overestimating or underestimating production, which could have economic consequences.

5. Field Application

We applied the GMM method to RNP data from the field production of a singly fractured vertical gas well in a low-permeability reservoir. Fig. 14 shows the RNP profile of this field production data.

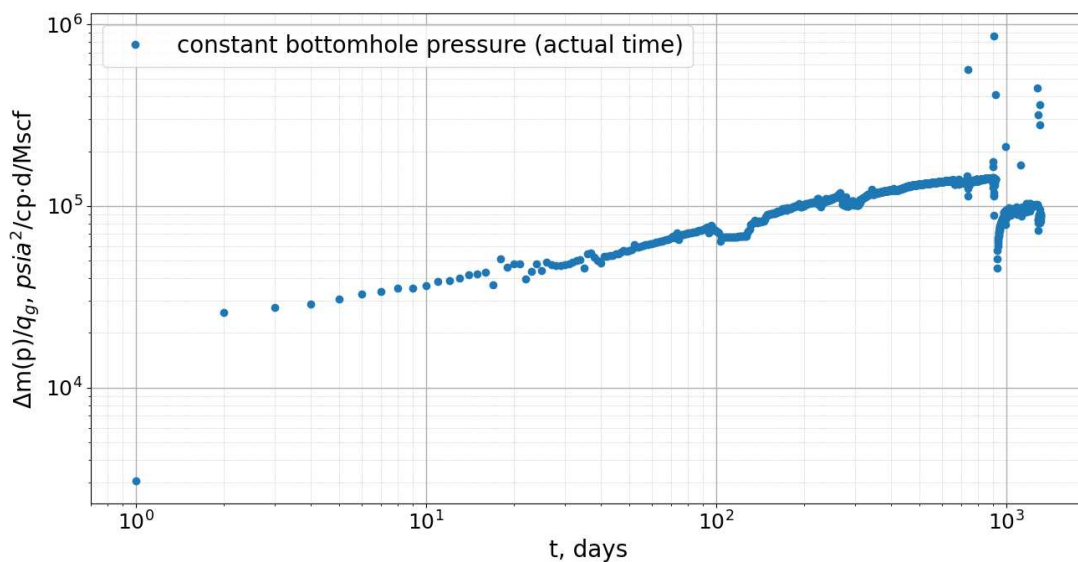


Fig. 14. RNP profile for field application data.

We computed the sensitivity of results to input parameter values. **Table 5** summarizes the sensitivity values of the GMM hyperparameter in terms of threshold factor and number of components.

Table 5

Sensitivity of GMM hyperparameter for field application.

Hyperparameter	Optimal Values
Threshold factor	0.8 – 1.1
Number of components	7 – 10

Fig. 15 illustrates the RNP result after detecting the outliers with input parameters of the threshold factor 0.8 and seven components. The RNP results shown Fig. 15 and **Fig. 16**, which has a threshold factor of 0.9 and the eight components, do not differ significantly. **Fig. 17** shows that GMM predicts some RNP data in the trend profile are outliers, which is unlikely. **Fig. 18** shows the result for the threshold factor of 1.1 and the ten components, which is much better than the others but still leads to some errors by not removing the two data points in the upper right corner. Since the data is limited, only three years of production, we determine that production is still in an early off-trend flow regime.

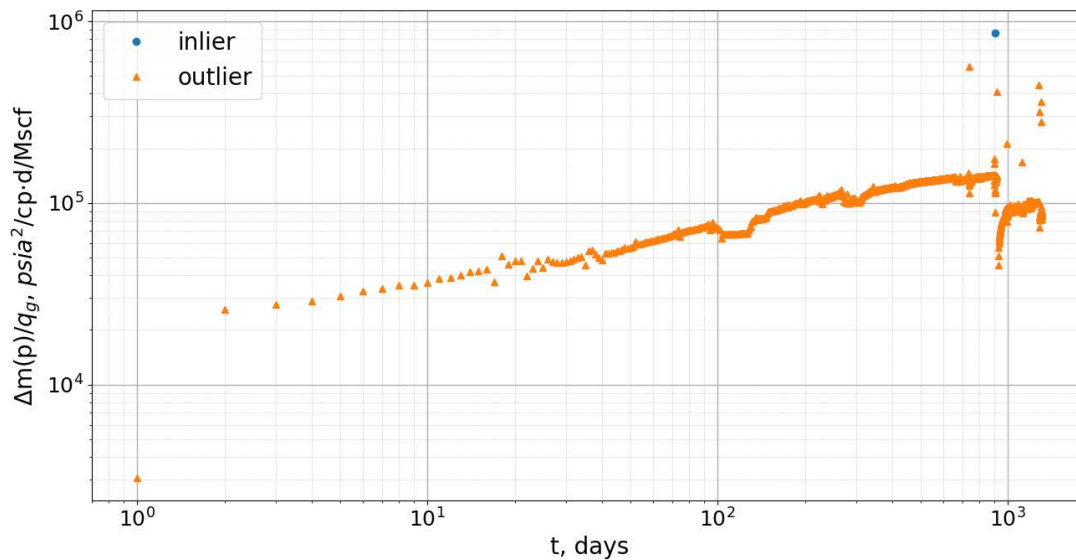


Fig. 15. Sensitivity of input parameter result for threshold factor = 0.8 and seven components.

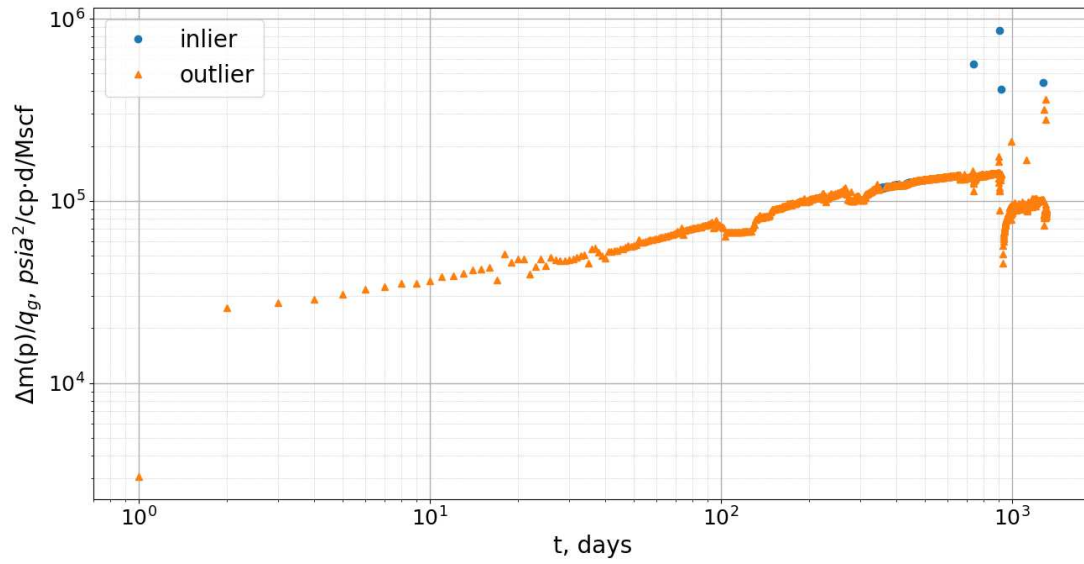


Fig. 16. Sensitivity of input parameter result for threshold factor = 0.9 and eight components.

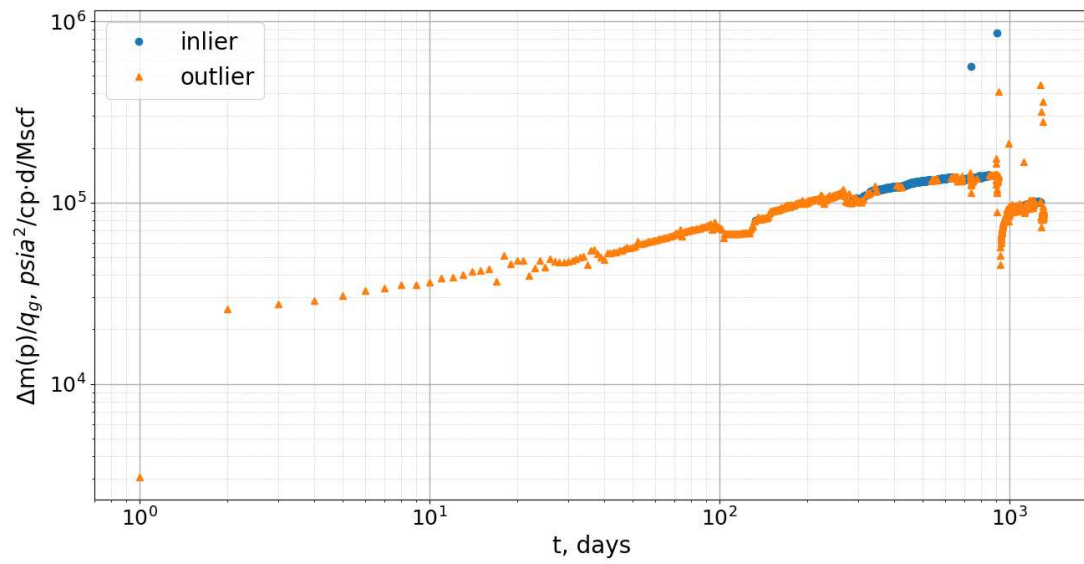


Fig. 17. Sensitivity of input parameter result for threshold factor = 1 and nine components.

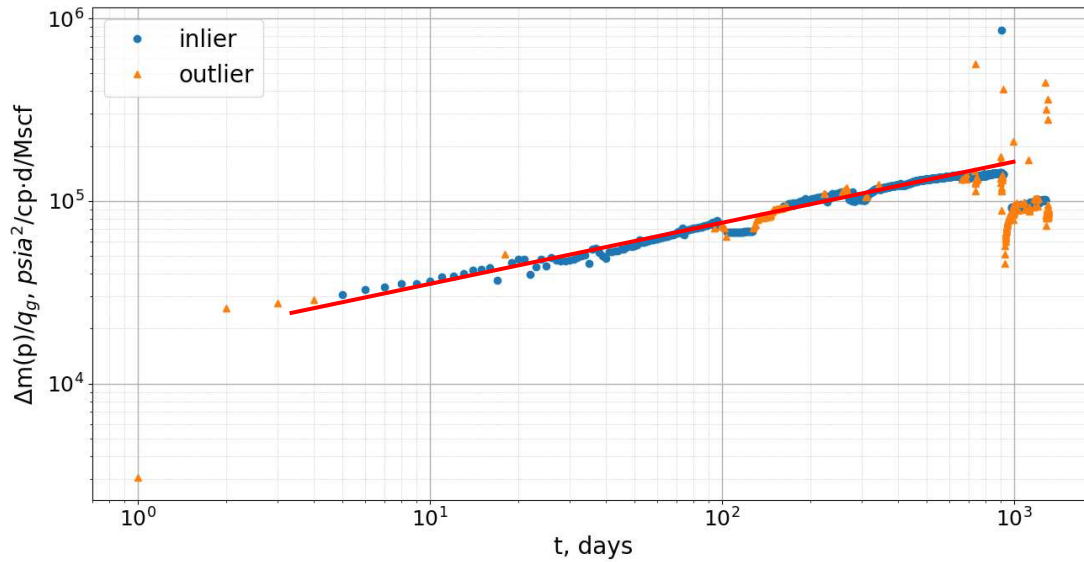


Fig. 18. Sensitivity of input parameter result for threshold factor = 1.1 and ten components. The red line has a slope of 1/3.

In Fig. 18, we can see that after applying the GMM method, the RNP data trend is firmer (without outliers), with a slope of 1/3 during the interval $30 < t \leq 300$ days. With the clearer start and end of the linear flow regime from Fig. 18, we can generate the Arps b -factor value more accurately than in Fig. 14. The comparison of the Arps b -factor extracted from the original data (Fig. 14) and clean data (Fig. 18), as summarized in **Table 6**, highlights why the cleaned data provide a more reliable basis for analysis. From Fig. 18, the removal of early data points identified as outliers makes the decline profile between them different. For the clean data, the fitted decline profile has a lower initial rate with a starting decline time shift to the right. Consequently, the extracted the Arps b -factor is higher.

Table 6

Comparison of decline parameters between original and clean field production data.

Decline Parameters	Original Data	Clean Data	Unit
Initial rate	8,715.2	8,100.1	Mscf/d
Arps b -factor	5.	9	fraction
Initial decline rate	0.0023	0.005	1/days

The reason we conclude that the Arps b -factor obtained from Fig. 18 is more reliable due to removing outliers, which results in a clear decline curve analysis. In Fig. 14, the presence of outliers distorts the decline curve, making it biased due to incorrect start and end times of each segment or flow regime. This improvement in accuracy is significant because the Arps b -factor is key for decline curve analysis (DCA) and production forecasting. By removing anomalous data, the GMM method enhances the clarity of the flow regimes, allowing for more reliable production predictions.

6. Conclusions

This work expands previous studies of OD methods applied to RNP data by developing various scenarios of adding noise and outliers based on flow regimes with the following conclusions:

- Of the twenty OD methods we evaluated, our top performance methods are ABOD (proximity-based), GMM (probabilistic-based), iForest (ensemble-based), and MCD (linear-based).
- GMM is the most suitable OD method to detect and remove outliers from RNP data because it provides the best performance with an F1-score of 0.9 and an accuracy of 93.99% when applied to our synthetic RNP data.
- Applying the GMM method requires a sensitivity analysis of input parameters (threshold factor and number of components); we found the optimal input parameters for the threshold factor ranging from 0.8 to 1.1 and the number of components ranging from 7 to 10.

We consider the GMM method a robust OD method for various conditions of outlier presence in RNP data; this will help us obtain reliable preliminary reservoir properties and improve production forecasts.

Nomenclature

Roman symbols:

$c(n)$	=	average path length of the unsuccessful search in Binary Search Tree (BST)
$E(h(x))$	=	average of $h(x)$ from an iTree structure
f	=	scale factor for the log probability
$h(x)$	=	average path length of iTree
$H(i)$	=	harmonic number
M	=	number of component Gaussian densities
n	=	number of objects or samples
OS_x	=	outlier score term
p	=	number of variables
$p(x)$	=	probability density function (PDF) of data point x
p_{wD}	=	dimensionless bottomhole pressure
p_{wf}	=	flowing bottomhole pressure, m/Lt ² , psia
q_D	=	dimensionless flow rate
s	=	anomaly score of a point x
t	=	time, t, day
t_{mb}	=	material-balance time, t, day
$\bar{x}$	=	mean of data
x_f	=	fracture half-length, L, ft
y_d	=	data label value detected
y_e	=	reservoir length, L, ft
y_r	=	data label value reference

Greek symbols:

λ	=	mixture weights from all densities of data
-----------	---	--

μ_i	=	mean of data
$\hat{\mu}_{MCD}$	=	MCD estimate of the data location
σ_i	=	variance of data
$\hat{\Sigma}_{MCD}$	=	MCD covariance estimate

Acknowledgment

This work was supported by the Indonesia Endowment Fund for Education (LPDP) Scholarship.

References

- Agarwal, A., Gupta, N., 2021. Comparison of Outlier Detection Techniques for Structured Data.
- Aggarwal, C.C., 2017. Outlier Analysis. Springer International Publishing, Cham. <https://doi.org/10.1007/978-3-319-47578-3>
- Arps, J.J., 1945. Analysis of Decline Curves. Transactions of the AIME 160, 228–247. <https://doi.org/10.2118/945228-G>
- Chaudhary, N.L., Lee, W.J., 2016. Detecting and Removing Outliers in Production Data to Enhance Production Forecasting, in: Day 2 Wed, May 18, 2016. SPE. <https://doi.org/10.2118/179958-MS>
- Cheng, Z., Zou, C., Dong, J., 2019. Outlier detection using isolation forest and local outlier factor, in: Proceedings of the Conference on Research in Adaptive and Convergent Systems. ACM, New York, NY, USA, pp. 161–168. <https://doi.org/10.1145/3338840.3355641>
- Clarkson, C.R., 2013. Production data analysis of unconventional gas wells: Review of theory and best practices. Int J Coal Geol 109–110, 101–146. <https://doi.org/10.1016/j.coal.2013.01.002>
- Fauconnier, C., Haesbroeck, G., 2009. Outliers detection with the minimum covariance determinant estimator in practice. Stat Methodol 6, 363–379. <https://doi.org/10.1016/j.stamet.2008.12.005>
- Hubert, M., Debruyne, M., Rousseeuw, P.J., 2018. Minimum covariance determinant and extensions. WIREs Computational Statistics 10. <https://doi.org/10.1002/wics.1421>
- Jha, H.S., Khanal, A., Seikh, H.M.D., Lee, W.J., 2022. A comparative study on outlier detection techniques for noisy production data from unconventional shale reservoirs. J Nat Gas Sci Eng 105, 104720. <https://doi.org/10.1016/j.jngse.2022.104720>
- Jha, H.S., Lee, W.J., 2017. Problems With Application of Material Balance Time to Transient Flow Data in Diagnostic Plots, in: Proceedings of the 5th Unconventional Resources Technology Conference. American Association of Petroleum Geologists, Tulsa, OK, USA. <https://doi.org/10.15530/urtec-2017-2697627>
- Kriegel, H.-P., Schubert, M., Zimek, A., 2008. Angle-based outlier detection in high-dimensional data, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08. ACM, New York, NY, USA, pp. 444–452. <https://doi.org/10.1145/1401890.1401946>
- Lee, J., 2021. Establishing the Basis for Multi-Segment Arps Decline Models, in: Proceedings of the 9th Unconventional Resources Technology Conference. American Association of Petroleum Geologists, Tulsa, OK, USA. <https://doi.org/10.15530/urtec-2021-5393>

- Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation Forest, in: 2008 Eighth IEEE International Conference on Data Mining. IEEE, pp. 413–422. <https://doi.org/10.1109/ICDM.2008.17>
- Palacio, J.C., Blasingame, T.A., 1993. UNAVAILABLE - Decline-Curve Analysis With Type Curves – Analysis of Gas Well Production Data, in: Proceedings of SPE Rocky Mountain Regional/Low Permeability Reservoirs Symposium. Society of Petroleum Engineers. <https://doi.org/10.2523/25909-MS>
- Powers, D.M.W., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.
- Pratikno, H., Lee, W.J., Torres, C.K., 2021. Application of Multiple Diagnostic Plots to Identify End of Linear Flow in Unconventional Reservoirs, in: Day 3 Thu, September 23, 2021. SPE, pp. 21–23. <https://doi.org/10.2118/205906-MS>
- Reddy, A., Ordway-West, M., Lee, M., Dugan, M., Whitney, J., Kahana, R., Ford, B., Muedsam, J., Henslee, A., Rao, M., 2017. Using Gaussian Mixture Models to Detect Outliers in Seasonal Univariate Network Traffic, in: 2017 IEEE Security and Privacy Workshops (SPW). IEEE, pp. 229–234. <https://doi.org/10.1109/SPW.2017.9>
- Reynolds, D., 2009. Gaussian Mixture Models, in: Encyclopedia of Biometrics. Springer US, Boston, MA, pp. 659–663. https://doi.org/10.1007/978-0-387-73003-5_196
- Rousseeuw, P.J., Driessen, K. Van, 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* 41, 212–223. <https://doi.org/10.1080/00401706.1999.10485670>
- Shekhar Jha, H., Mustafa Ud Din Sheikh, H., Lee, W.J., 2021. Outlier Detection Techniques Help Us Identify and Remove Outliers in Production Data to Improve Production Forecasting, in: Proceedings of the 2021 Asia Pacific Unconventional Resources Technology Conference. Unconventional Resources Technology Conference, Tulsa, OK, USA, pp. 1694–1706. <https://doi.org/10.15530/AP-URTEC-2021-208384>
- Wattenbarger, R.A., El-Banbi, A.H., Villegas, M.E., Maggard, J.B., 1998. Production Analysis of Linear Flow Into Fractured Tight Gas Wells, in: All Days. SPE. <https://doi.org/10.2118/39931-MS>

Appendix A—Complete Set of Hyperparameter Combinations of Outlier Detection (OD) Methods

Table A-1

Set of hyperparameter combinations for all OD methods

	OD Method	Hyperparameter	Search Value	Optimal Value
Ensemble	Ada Boost	• Learning rate	• 0.001, 0.03, 0.5, 1, 7	• 0.03
	Feature	• Max features	• 0.7, 0.8, 0.9, 1	• 0.8
	Bagging	• Max samples	• 0.7, 0.8, 0.9, 1	• 0.9
	Extra Trees	• Max depth	• 5, 7, 9, 11	• 5
		• Max features	• 0.1, 0.3, 0.7, 1	• 0.7
		• Min samples leaf	• 0.1, 0.3, 0.7, 1	• 0.1
		• Min samples split	• 0.1, 0.3, 0.7, 1	• 0.1
		• Min weight fraction leaf	• 0.1, 0.3, 0.5	• 0.1
	Gradient Boosting	• Max depth	• 5, 7, 9, 11	
		• Max features	• 0.1, 0.3, 0.7, 1	
		• Min samples leaf	• 0.1, 0.3, 0.7, 1	
		• Min samples split	• 0.1, 0.3, 0.7, 1	

Linear	Isolation Forest	• Min weight fraction leaf	• 0.1, 0.3, 0.5	
		• Contamination	• 'auto', 0.1, 0.2, 0.3, 0.4, 0.5	• 0.1
		• Max features	• 0.3, 0.5, 0.7, 0.8, 0.9	• 0.8
		• Max samples	• 'auto', 0.3, 0.5, 0.7, 0.8, 0.9	• 0.9
	Logistic Regression	• C	• 0.01, 0.3, 1, 5, 10	• 0.01
		• L1 ratio	• 0.001, 0.03, 0.5, 1	• 0.001
		• Penalty	• 'l1', 'l2', 'elasticnet'	• 'l1'
		• Solver	• 'lbfgs', 'liblinear', 'newton-cholesky', 'saga'	• 'saga'
	Minimum Covariance Determinant	• Contamination	• 0.1, 0.2, 0.3, 0.4, 0.5	• 0.1
		• Support fraction	• 0.1, 0.3, 0.5, 0.7, 0.9, 1	• 0.5
Probabilistic	One-Class Support Vector Machines	• Degree	• 1, 2, 3	• 1
		• Gamma	• 'scale', 'auto', 0.7, 0.8, 0.9, 1	• 0.7
		• Kernel	• 'linear', 'poly', 'rbf'	• 'rbf'
		• Nu	• 0.5, 0.7, 0.8, 0.9, 1	• 0.5
	Passive Aggressive	• C	• 0.01, 0.3, 1, 5, 10	• 0.01
	Ridge	• Alpha	• 0.3, 0.5, 0.7, 0.8, 0.9, 1	• 0.3
	Bayesian Gaussian Mixture	• Covariance type	• 'full', 'tied', 'diag', 'spherical'	• 'full'
		• Init params	• 'kmeans', 'k-means++', 'random', 'random_from_data'	• 'kmeans'
		• Number of components	• 5, 10, 30, 50, 100, 200	• 5
	Bernoulli Naive Bayes	• Alpha	• 0.1, 0.3, 0.5, 0.7, 0.8, 0.9, 1	• 0.1
Proximity	Gaussian Mixture	• Covariance type	• 'full', 'tied', 'diag', 'spherical'	• 'full'
		• Init params	• 'kmeans', 'k-means++', 'random', 'random_from_data'	• 'kmeans'
		• Number of components	• 10, 50, 100, 200, 300, 500, 700	• 10
	Gaussian Naive Bayes	• Var smoothing	• 1e-9, 1e-7, 1e-5, 1e-3	• 1e-9
	Kernel Density	• Bandwidth	• 0.3, 0.5, 0.7, 0.8, 0.9, 1	• 0.3
		• Leaf size	• 10, 20, 40, 70, 100	• 10
	Angle-Based	• Number of neighbors	• 3, 5, 10, 15, 20, 40	• 3
	Local Outlier Factor	• Contamination	• 0.1, 0.2, 0.3, 0.4, 0.5	• 0.1
		• Leaf size	• 10, 20, 30, 50, 70, 100	• 10
		• Number of neighbors	• 3, 5, 10, 15, 20, 40	• 40
		• P	• 1, 1.5, 2, 3	• 1
Proximity	K-Nearest Neighbors	• Leaf size	• 10, 20, 30, 50, 70, 100	• 10
		• Number of neighbors	• 3, 5, 10, 15, 20, 40	• 3
		• P	• 0.5, 1, 1.5, 2	• 1
		• Weights	• 'uniform', 'distance'	'distance'
	Radius Neighbors	• Algorithm	• 'auto', 'ball_tree', 'kd_tree', 'brute'	• 'auto'
		• Leaf size	• 10, 20, 30, 50, 70, 100	• 10

	• P	• 0.1, 0.5, 1, 1.5, 2, 3	• 1
	• Radius	• 0.1, 0.5, 0.7, 1, 2, 5	• 0.5
	• Weights	• 'uniform', 'distance'	• 'distance'
Rotation	• Contamination	• 0.1, 0.25, 0.3, 0.4, 0.5	• 0.5

Appendix B—Complete Performance Metric Results of Outlier Detection (OD) Methods

Table B-1

Performance metric results from all scenarios for combined noise contamination levels.

	OD Method	TP (%)	FN (%)	TN (%)	FP (%)	MAE	Accuracy	Precision	Recall	F1-score
Ensemble	Ada Boost	21.11	7.62	65.92	5.35	0.26	0.87	0.86	0.88	0.86
	Extra Trees	14.24	14.57	59.74	11.45	0.52	0.74	0.7	0.8	0.73
	Feature Bagging	25.64	3.09	69.74	1.53	0.09	0.95	0.95	0.96	0.95
	Gradient Boosting	21.59	7.22	67.48	3.71	0.22	0.89	0.88	0.91	0.89
	Isolation Forest	6.42	22.32	66.96	4.3	0.53	0.73	0.74	0.93	0.81
Linear	Logistic Regression	20.29	8.51	54.68	16.52	0.5	0.75	0.83	0.75	0.78
	Minimum Covariance Determinant	15.88	12.92	64.55	6.65	0.39	0.8	0.82	0.9	0.84
	One-Class Support Vector Machines	17.96	10.85	40.29	30.9	0.83	0.58	0.76	0.57	0.62
	Passive Aggressive	18.37	10.43	55.35	15.85	0.53	0.74	0.83	0.77	0.77
	Ridge	20.99	7.82	52.2	18.99	0.54	0.73	0.85	0.72	0.77
Probabilistic	Bayesian Gaussian Mixture	23.51	5.24	45.65	25.6	0.31	0.69	0.48	0.9	0.56
	Bernoulli Naive Bayes	21.91	6.9	51.78	19.42	0.53	0.74	0.8	0.67	0.72
	Gaussian Mixture	24.49	4.26	45.53	25.72	0.3	0.7	0.49	0.93	0.58
	Gaussian Naive Bayes	20.14	8.67	58.11	13.08	0.44	0.78	0.81	0.76	0.78
	Kernel Density	22.47	6.28	43.92	27.33	0.34	0.66	0.45	0.86	0.53
Proximity	Angle-Based	6.5	22.25	69.09	2.16	0.24	0.76	0.74	0.37	0.4
	Local Outlier Factor	13.56	15.25	65.45	5.74	0.42	0.79	0.79	0.9	0.83
	K-Nearest Neighbors	23.78	5.03	69.76	1.43	0.13	0.94	0.91	0.97	0.94
	Radius Neighbors	23.74	5.07	69.2	2	0.14	0.93	0.91	0.96	0.93
	Rotation	18.79	9.96	48.3	22.95	0.33	0.67	0.43	0.71	0.5