URTeC: 2897507

Statistical Controls on Induced Seismicity

Saurabh Sinha*, Yunjie Wen, Rafael Pires De Lima, and Kurt Marfurt; University of Oklahoma

Abstract

Seismicity in Oklahoma has shown a sharp increase since 2010 and is mostly attributed to wells used to dispose wastewater from hydraulically fractured production wells (McClure, Gibson, Chiu, & Ranganath, 2017). Many studies are conducted so far to include / eliminate various causes and solutions to this problem. The recent studies in this research area (Holland, 2011), show a general consensus on the main cause of the seismic events (earthquakes) as the high volume disposal in the Arbuckle formation causing the critically stressed faults in the basement rock to fail.

Some of these studies include the modeling and simulation of the physical processes whereas some studies delve into the statistical analyses of the relationships between disposal wells and induced seismicity. However, most of the previous attempts on the statistical analysis of this dataset use a more qualitative view of the problem (Langenbruch & Zoback, 2016) instead of quantifying the impact of various parameters such as injection rate, volumes, pressures, etc. Other work like Gogri et al. (2017) uses geo-modeling and simulation approach, but this constraint the modeling to a smaller section due to limitations on the seismic data extent.

In this work, we use various data analytics methods to quantify the impact of different injection well parameters and rock properties on the earthquake event magnitude and intensity. Our models show that hierarchical and K-means clustering are able to group the wells into clusters that conforms with the earthquake event density.

Including more clusters in our analysis refined the results but in general, four clusters are enough to capture the trends in our dataset. Fuzzy clustering, which is a soft clustering yields good results only after number of clusters exceed five. For predictive modeling part, Gradient boost regression and random forest work better than least absolute shrinkage and selection operator (LASSO), elastic nets and linear models.

Introduction

One of the key challenges in the quantification of the seismic events is the spatial and temporal relationships between the rock and well parameters. As the injection parameters change from one year to the next, the inputs are non-stationary. The second challenge is the definition of the events that will be modeled. Adding to the complexity, there are other more involved facets of this problem like determining the exact location of the seismic events. In this work, we have assumed the location of the events given by Oklahoma Corporate Commission (OCC) is exact, and we formulate our problem-based on these.

In this work we have attempted a suite of clustering and predictive modeling techniques. We present the results from the models that best fitted our dataset and are able to incorporate the variables interpreted as key indicators based on our expert knowledge. Our model is primarily based on the waste water disposal well information from OCC. The major challenge that we faced in our work is that the number of data points for the analysis is not optimal and the well operating conditions are not constant throughout the well injection cycle.

The time window of analysis in our study is from year 2010 to year 2016. Some wells start in 2010 and stop injecting for a while. Some new wells are added and some wells are abandoned. Also in the later period the “traffic light” (McNamara et al., 2015) approach adopted by OCC makes the analysis difficult. To resolve this, we take the median injection values and peak values of different input parameters.

Methodology

Input parameters

For this study we selected the following input parameters:

a) Average of total depth of the Arbuckle: This parameter takes care of highly irregular Arbuckle and basement topology. Guglielmi et al., (2015) studied the effect of the injection depth and found out if the wells are injecting closer to the basement which contains the critically stressed faults, they may trigger the induced seismicity. Including depths help us take this factor into account indirectly as the wells which are in close proximity to each other show great difference in injection depths. We do not have enough well logs that penetrate the basement to create a new parameter which show the "mid-perforation" depth from the top of the basement.
b) Zone thickness: This is the thickness of Arbuckle formation measured in feet
c) Median injected volume: This parameter helps us identify the wells which inject the highest volumes in a given period.
d) Peak Volume: This is the maximum monthly volume that a well injects in the reservoir.
e) Peak pressure: This is the maximum tubing head pressure (THP) encountered in a month for an injection well
f) Standard deviation in peak pressure and volume
g) Pressure gradient: This is the initial reservoir pressure gradient. We calculate this by first computing bottomhole pressure (BHP) using the wellbore configuration and using Beggs and Brill correlation. We then use Silin slope analysis (Silin, Holtzman, Patzek, Brink, & Minner, 2005) to estimate the virgin reservoir pressure and hence reservoir pressure gradient. We used evenly spaced 60 wells in the area of interest and then computed the BHP for these wells. We then generate a map of the BHP for the whole region using Kriging as the interpolation method and then resampled them on every well location.

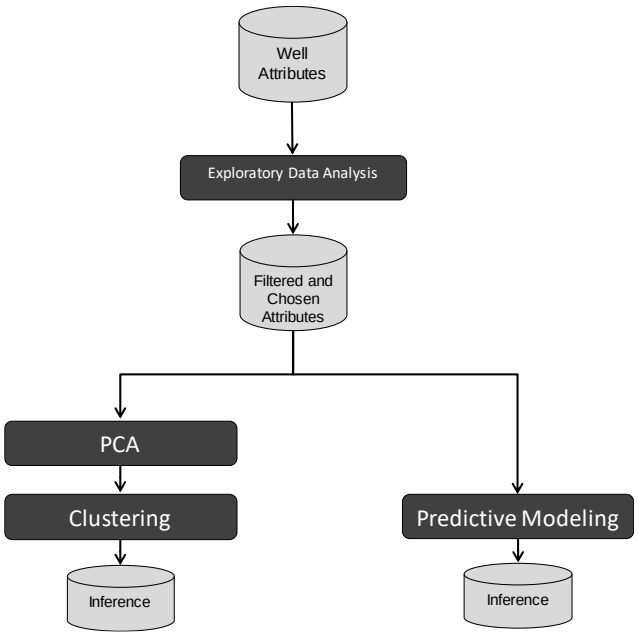


Figure 1. Workflow used in our study. We first perform the exploratory data analysis and the filtered values are then used for clustering and predictive modeling. Before clustering we use PCA to reduce dimensionality of our data.

Workflow

Fig. 1 summarizes our workflow for this study. We have divided our analysis in three parts:

a) Exploratory data analysis: In the exploratory data analysis section, we perform outlier detection using Grubb's test (Grubbs, 1950) and boxplots. We use principal component analysis (PCA) to reduce the dimensionality of the dataset to be used in simpler models. We test for multicollinearity in the data using cross plots. We include/ exclude the data based on these tests as well as their physical significance in the analysis. We test the correlation between the parameters. In our study, none of the variables show more than 95% correlation.

b) Clustering: In the clustering method, we test different clustering such as K-means: testing between elbow and Hartigan's method (Hartigan & Wong, 1979) to choose number of clusters, hierarchical clustering with Ward's (Ward, 1963), complete and centroid methods and Fuzzy clustering.
c) Predictive modeling: In this section we use Random forest (Ho, 1995), Gradient Boosting Regressor (Friedman, 2001), Ridge Regression (Fu, 1998), and LASSO (Tibshirani, 1996). Simpler models such as elastic net did not yield good results in our study.

Output parameters

The output parameter in our case is the event density of the earthquake of all magnitude reported by USGS from year 2010 to year 2016. **Fig. 2** shows the earthquake event density in the area of interest.

We also attempted using the same methodology to model earthquake magnitude as an output parameter, but the correlations that we obtained with the earthquake magnitude were weak just using the well parameters as predictors. It remains our future work to correlate magnitude of the event with more input parameters such as seismic attributes and well logs derived petrophysical properties.

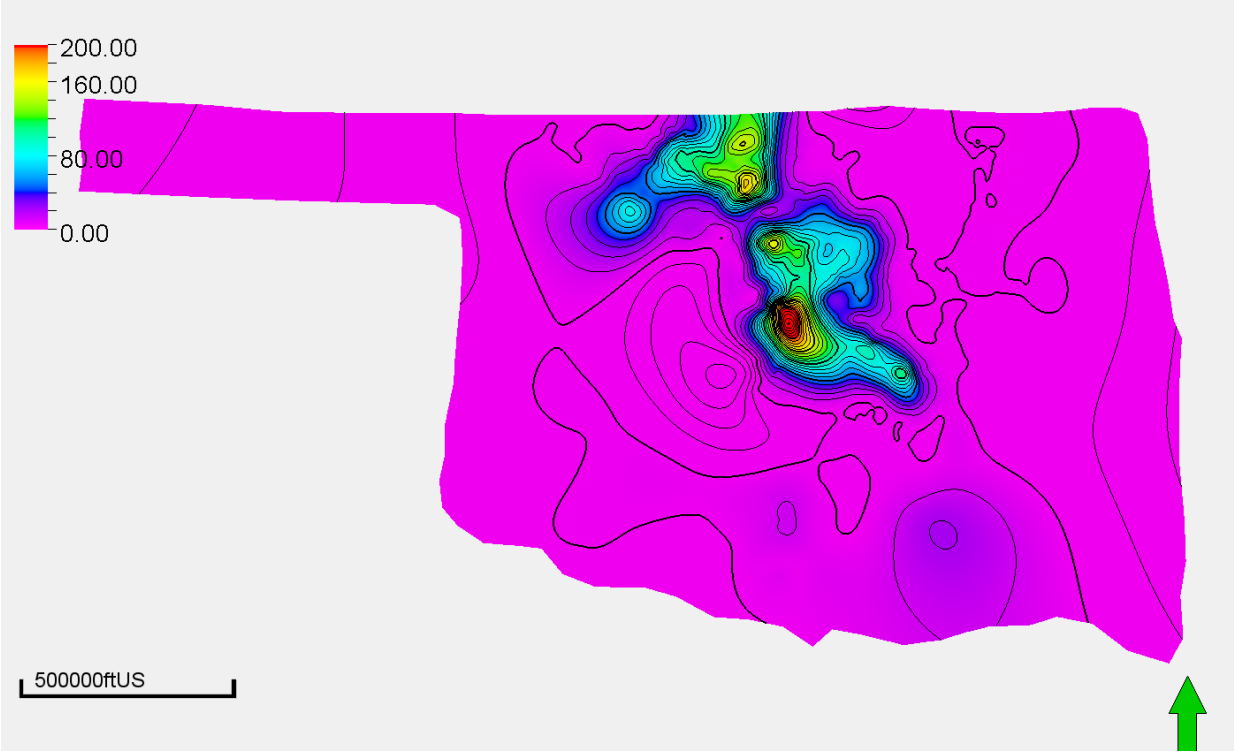


Figure 2. The earthquake event density. We use earthquake event density as an output parameter in our study.

Results

Exploratory Data Analysis

From Figure A1 in Appendix A, it can be observed that some of the variables show certain degree of multicollinearity but none of the variables are more than 95% correlated. Figure A2 shows distribution of different parameters in the reservoir. The THP and monthly volumes show log normal distributions with some outliers. The thickness of reservoir and reservoir pressure show the normal distribution.

Figure A3 shows the boxplots for the input parameters and the outliers. We use these plots and Grubb's test to remove few data points for which the THP is reported erroneously.

Clustering

We have a total of 231 wells after the exploratory data analysis for further modeling. We create a total of 12 attributes and first attempted the clustering with all 12 attributes listed in Table A1. Our clustering did not yield good results as the high number of input attributes is too large when compared to the number of data points. To reduce the number of variables and still keep the variability in the data, we first use PCA to reduce the number of variables. **Fig. 3** show the cumulative proportion of variance in our dataset with the number of principal components. It can be observed from the Figure 3, around 90% of the variability in our data can be represented using five principal components. Hence, we use the first five components to cluster our data.

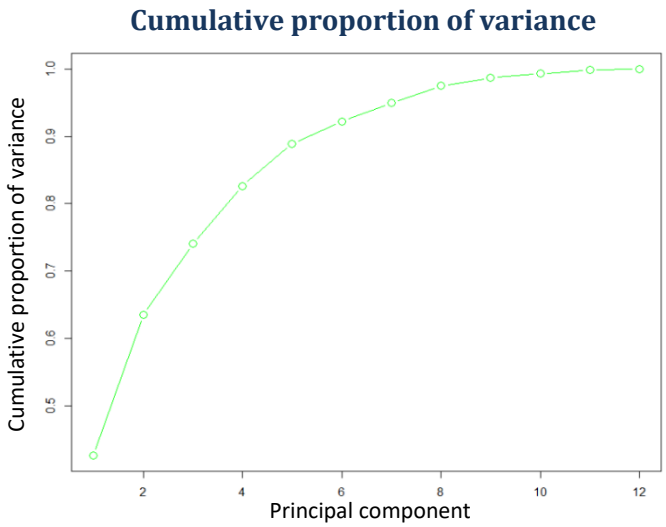
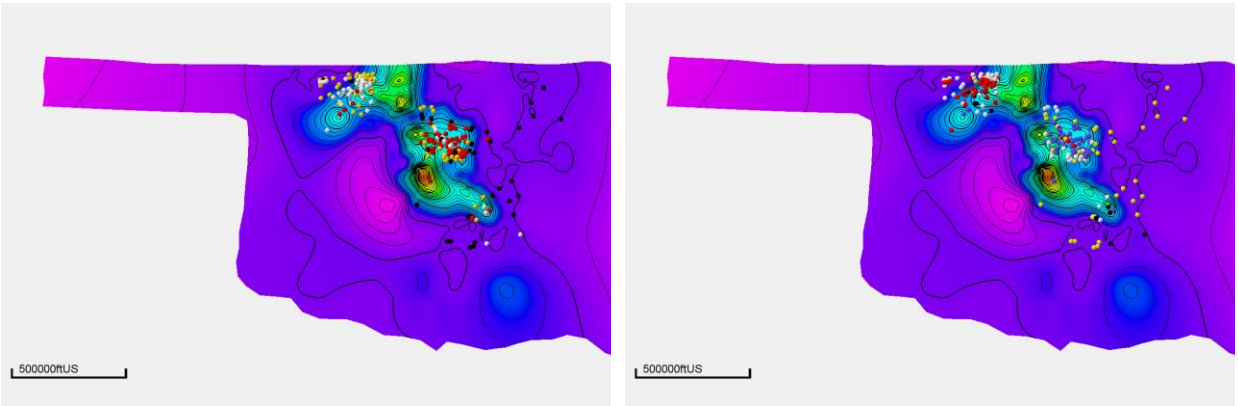


Figure 3. Cumulative proportion of variance in the data. It can be observed that almost 90% of the variance can be represented by first five principal components.

Clustering with Principal Components

K-means

We use elbow method and Hartington's rule to identify the number of clusters in our data using five principal components. The optimal number of clusters for our data is four. We also repeat the analysis for four and five clusters. The results from K-means clustering is shown in **Fig. 4**.



(a) **(b)**

Figure 4. K-means clustering on the dataset with (a) four clusters, and (b) five clusters overlaid on the event density map (Figure 2). It can be observed that the wells clustering conforms with the event density map. This suggests the correlation between wells and event densities. With four or five clusters, enough resolution in the trend can be observed.

Hierarchical Clustering

For hierarchical clustering we use three methods i.e. Ward's, centroid and complete. Ward's method yielded best results in our case. Centroid and complete method failed to cluster the data and lumped it majorly into one single cluster. **Fig. 5** shows these clusters on the dendogram for number 3, 4 and 5 clusters respectively. **Fig. 6** shows these clusters spatial distribution.

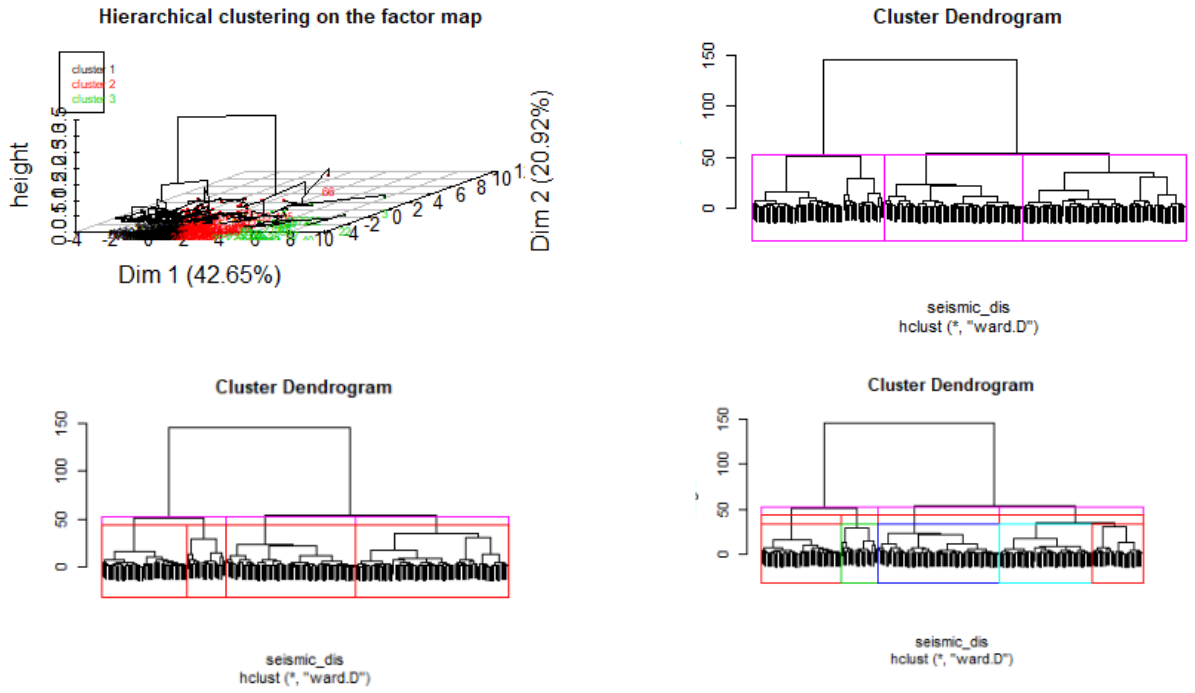


Figure 5. Sketch of the clusters obtained using Ward's method for hierarchical clustering. From left to right: tree pruned at three, four and five clusters.

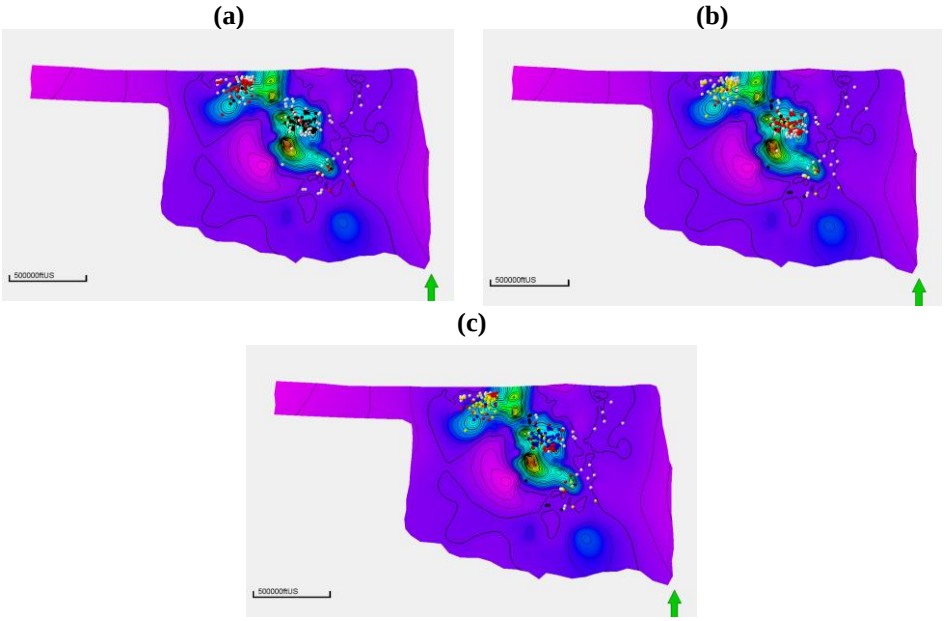


Figure 6. Hierarchical clustering with (a) 3 clusters, (b) 4 clusters, and (c) 5 clusters on the well data. The earthquake trends are captured efficiently with less number of clusters than K –means.

Fuzzy Clustering

We use the Fuzzy clustering with silhouette width (Campello & Hruschka, 2006) to select the optimal number of clusters. We obtain a total of 5 clusters using this method. **Fig 7** shows the distribution of the clusters and the average silhouette width. **Fig. 8** show the spatial distribution of these clusters.

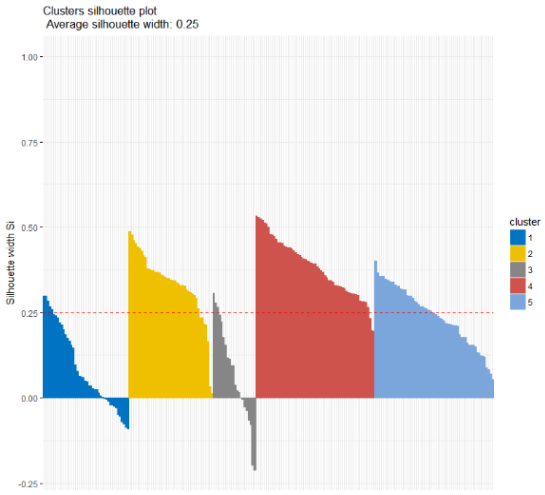


Figure 7. The five clusters and the average silhouette width obtained from the Fuzzy clustering

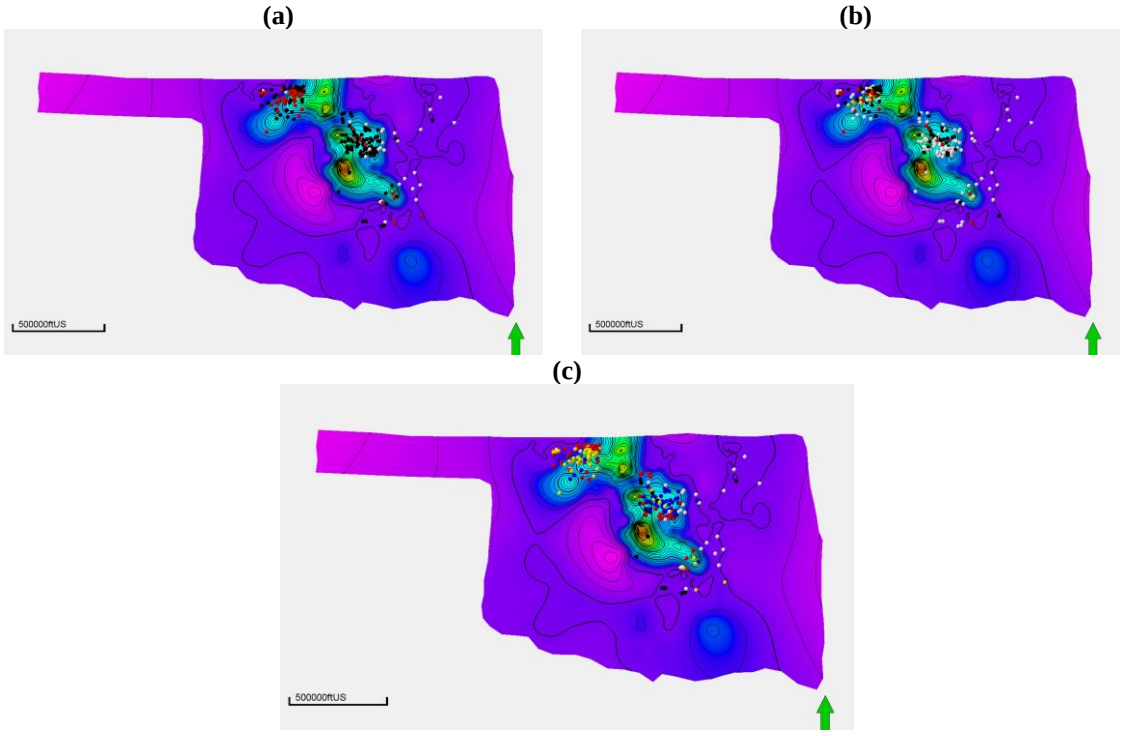


Figure 8. Fuzzy clustering on the well data for (a) 3 clusters, (b) 4 clusters, and (c) 5 clusters. Being a "soft" clustering method, Fuzzy clustering provides enough resolution only after 5 clusters.

Modeling

We split our data into two parts: training and validation set. The training data contains 70% of available data whereas the validation set holds the remaining 30%. The total data sample for the model is 231. We used machine learning supervised algorithms (Random forest, Gradient Boosting, Ridge Regression, and Lasso) to generate predictive

models. Based on the model performance from preliminary results, we choose Random forest and Gradient Boosting as final model for our analysis.

Parameter Tuning- Random forest and Gradient Boosting

Tuning is essentially the selection of the best parameters for an algorithm to optimize its performance to complete a learning task. In this work, we implement experiments to find out which parameters play important roles for the model. Two parameters are selected for performance improvement

1. ***n_estimators:*** the number of tree built within a Random Forest before aggregating the prediction (number of trees). Normally the higher the number the better, but its computation cost might come with it. Generally, adding more trees to the model can lead to overfitting.
2. ***max_depth:*** the selection of how deep you want to make your trees (maximum depth). You can split once, twice or none. The deeper trees are; more complex the model is. Generally, better results are seen with 4 – 8 levels.

Table 1. Parameter tuning values and selection for best fit.

	Tuning Values	Gradient Boosting	Random Forest
maxDepth	[1,2,…,20]	3	80
n_estimators	[50,100,150,200,…,950]	8	200

Random Forest Model

Random Forest algorithm is a supervised learning technique. The relationship between the number of tress in the forest and the results the model can get is strongly correlated. The Larger number of trees, the more accurate the result is. The best parameters are set based on the results of parameter tuning.

Fig. 9 shows in part (a) and (b) the model performance. It is observed that the model performance improves with increasing the maximum depth, but the performance slightly drops after optimal value eight. The figure also tells us that the model performs well on training data, which the best score is 89%, but the best score of test data is only 20%.

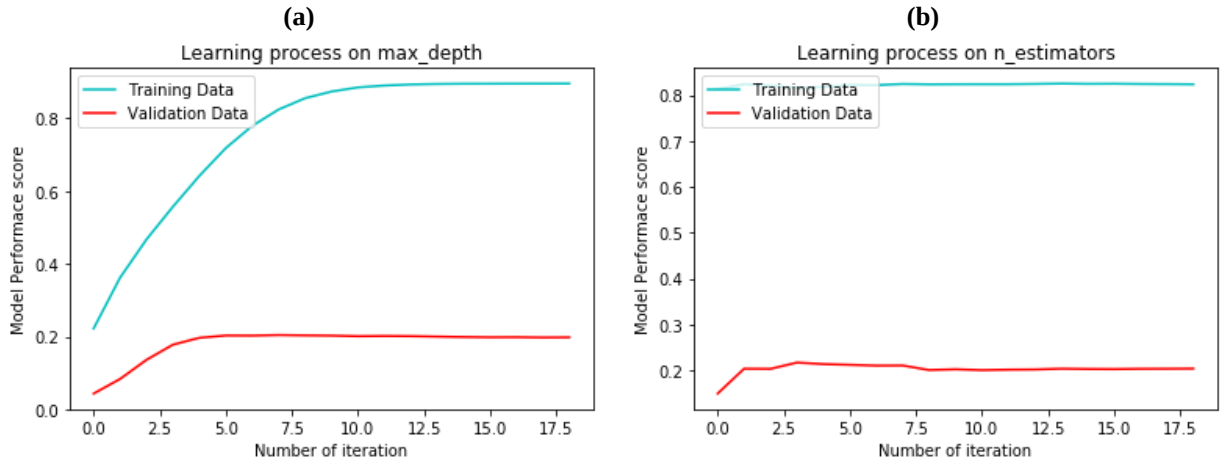


Figure 9. Model performance score with number of iterations. It can be observed that the model performance improves with increasing the maximum depth, but the performance slightly drops after optimal value 8. It can also be observed that the model performs well on training data, which the best score is 89%, but the best score of test data is only 20%.

Fig. 10 show that the model performance is measured by R square with result 0.2172, which mean about 22% data fit to the model. The red and blue line in the figure present predicted and true value for the seismic magnitude respectively. The closer distances between two values, the better the model performs.

Applying the architecture of the random forest with 8 max_depth and 200 estimator, the model performance is measured by R square with result 0.2172, which mean about 22% data fit to the model.

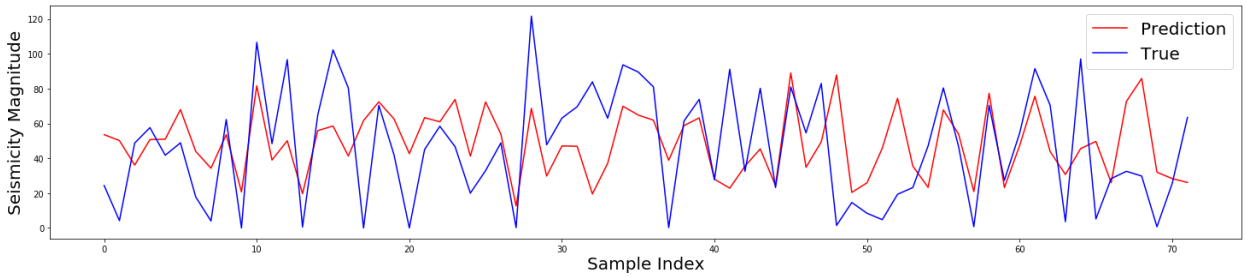


Figure 10. Prediction and the actual fit of the model

Gradient Boosting Regression

Gradient Boosted Regression Trees (GBRT) or shorter Gradient Boosting (Friedman, 1999) is a flexible non-parametric statistical learning technique. GBRT is tree-based model, therefore max_depth and n_estimators are still our focus for the parameter learning.

As compared to the Random Forest model, the learning rate is the third parameter that we train in GBRT. **Fig. 11(a)** shows the performance results from training data and testing data while the max_dept is tuned for the model. **Fig. 11(b)** shows the model performance score with number of iterations.

It can be observed from **Fig. 11**, that the GBRT model does not perform well on test data after max_dept reaches to optimal value, which is three. For training data, the model performance does not improve if the value of max_dept is bigger than three. The number of estimators does not impact the performance on testing data as we can see from the figure. In this case the optimal value for n_estimators is 80.

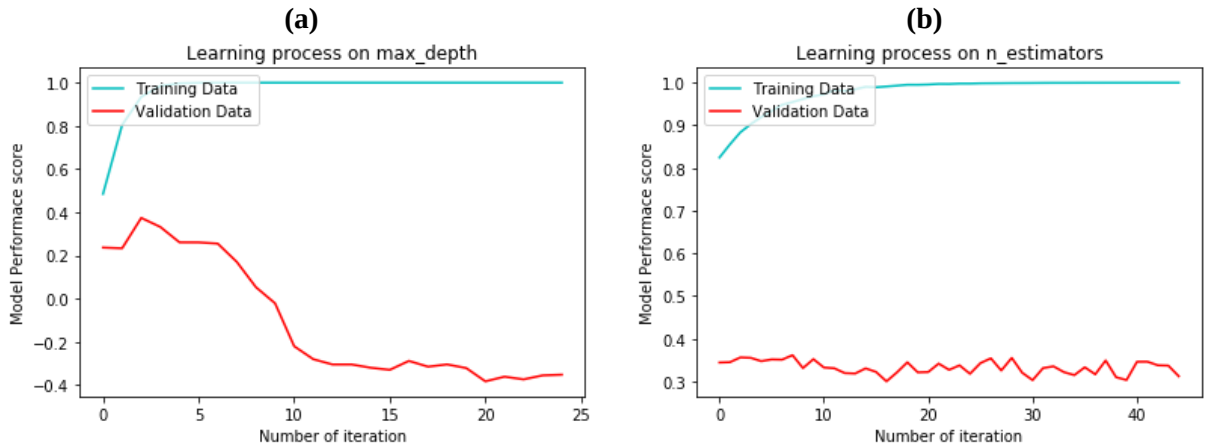


Figure 11. (a) Learning process of the model: It is based on the change of the tree depth. The model does not perform well on test data after max_dept reaches to optimal value, which is 3. For training data, model performance does not improve if the value of max_dept is bigger than 3, (b) Number of estimators: They do not impact the performance on testing data. In this case the optimal value for n_estimators are 80.

The architecture of GBRT model is max_dept = 3, n_esitmator = 80, and learning_rate = 0.1. The score of test data from this model is 35% and for the training data is 90%. Figure 12 Show the prediction and true value for this model.

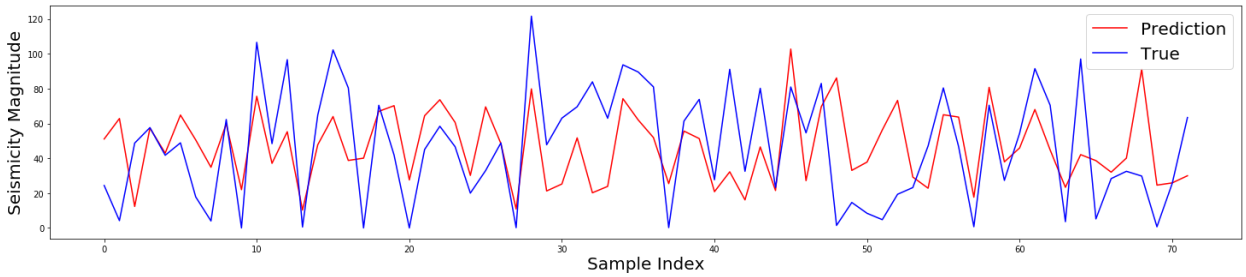


Figure 12. The architecture of GBRT model is max_dept = 3, n_esitmator = 80, and learning_rate = 0.1. The score of test data from this model is 35% and for the training data is 90%. The red and blue line in the figure present predicted and true value for the seismic magnitude respectively. The closer distances between two values, the better the model performs.

Models Inferences

Both model predict well on the training data, but poorly on the test data. The reason for the overfitting is the size of the dataset. The data size is limited by the number of injection wells and the wells having all the parameters. In future, we intend to repeat the same process with PCA component and/ or including more injection wells from neighboring states to Oklahoma as well.

Conclusions and Future Work

In our study, the clustering techniques show that the wells cluster together and conforms with the earthquake event density showing a correlation between induced seismicity and injection wells. For clustering, hierarchal clustering yields the best results with minimum number of clusters and high resolution. PCA analysis before clustering improved the results in our case to a great extent, removing the extra parameters without losing the variability in the dataset.

For modeling, we obtained an excellent correlation with the training data, however we obtained moderate correlation with the validation set. Simpler models like linear, LASSO and elastic net failed to capture the trends in the dataset. Gradient Boost regressor and random forest work best in our case. Major limitation in our study is the limited amount of data and including more data points can help alleviate the overfitting of complicated models to a great extent. Also, our analysis does not incorporate the complexities of the physical processes underlying the induced seismicity. We do not consider the different stress fields in the region or the flow path of the disposed water. Incorporating data from more geologically constrained regional models could help improve the accuracy of our predictions, however we do not have access to such models at this time.

Our next immediate work focuses on characterizing faults and fractures in the area (Ghosh et al., 2018; Milad et al., 2018) and incorporate them in our modeling along with fluid properties in the reservoir (Mehana et al., 2017; Salahshoor et al., 2018)

References

Campello, R. J. G. B., & Hruschka, E. R. (2006). A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets and Systems*, *157*(21), 2858–2875. https://doi.org/10.1016/j.fss.2006.07.006

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*. Institute of Mathematical Statistics. https://doi.org/10.2307/2699986

Fu, W. J. (1998). Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, *7*(3), 397–416. https://doi.org/10.1080/10618600.1998.10474784

Ghosh, S., Galvis-Portilla, H. A., Klockow, C. M., & Slatt, R. M. (2018). An application of outcrop analogues to understanding the origin and abundance of natural fractures in the Woodford Shale. *Journal of Petroleum*

Science and Engineering, *164*, 623–639. https://doi.org/10.1016/J.PETROL.2017.11.073

Gogri, M. P., Rohleder, J. M., Kabir, C. S., Pranter, M. J., & Reza, Z. A. (2017). Prognosis for Safe Water-Disposal-Well Operations and Practices Based on Reservoir Flow Modeling and Real-Time Performance Analysis. *SPE Annual Technical Conference and Exhibition*, *i*. https://doi.org/10.2118/187083-MS

Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics*, *21*(1), 27–58. https://doi.org/10.1214/aoms/1177729885

Guglielmi, Y., Cappa, F., Avouac, J. P., Henry, P., & Elsworth, D. (2015). Seismicity triggered by fluid injection-induced aseismic slip. *Science*, *348*(6240), 1224–1227. https://doi.org/10.1126/science.aab0476

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282). IEEE Comput. Soc. Press. https://doi.org/10.1109/ICDAR.1995.598994

Holland, A. (2011). Examination of possibly induced seismicity from hydraulic fracturing in the Eola Field, Garvin County, Oklahoma. *Oklahoma Geological Survey*, (OF1-2011), 1–31. https://doi.org/10.1017/CBO9781107415324.004

Langenbruch, C., & Zoback, A. M. D. (2016). How will induced seismicity in Oklahoma respond to decreased saltwater injection rates? *Science Advances*, *2*(11), 1–10. https://doi.org/10.1126/sciadv.1601542

McClure, M., Gibson, R., Chiu, K. K., & Ranganath, R. (2017). Identifying potentially induced seismicity and assessing statistical significance in Oklahoma and California. *Journal of Geophysical Research: Solid Earth*, *122*(3), 2153–2172. https://doi.org/10.1002/2016JB013711

McNamara, D. E., Rubinstein, J. L., Myers, E., Smoczyk, G., Benz, H. M., Williams, R. A., … Earle, P. (2015). Efforts to monitor and characterize the recent increasing seismicity in central Oklahoma. *The Leading Edge*, *34*(6), 628–639. https://doi.org/10.1190/tle34060628.1

Mehana, M., Fahes, M., & Huang, L. (2017). System Density of Oil-Gas Mixtures: Insights from Molecular Simulations. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers. https://doi.org/10.2118/187297-MS

Milad, B., Ghosh, S., & Slatt, R. M. (2018). Comparison of rock and natural fracture attributes in karsted and non-karsted Hunton Group Limestone: Ada and Fittstown area, Oklahoma, *69*(2), 70–86. Retrieved from http://archives.datapages.com/data/ocgs/data/069/069002/70_ocgs690070.htm

Oklahoma Corporate Commission. Oil and Gas Info.
https://apps.occeweb.com/RBDMSWeb_OK/OCCOGOnline.aspx (accessed 14th April, 2018)

Salahshoor, S., Fahes, M., & Teodoriu, C. (2018). A review on the effect of confinement on phase behavior in tight formations. *Journal of Natural Gas Science and Engineering*, *51*, 89–103. https://doi.org/10.1016/j.jngse.2017.12.011

Silin, D., Holtzman, M., Patzek, T., Brink, J., & Minner, M. (2005). Waterflood Surveillance and Control: Incorporating Hall Plot and Slope Analysis. *Proceedings of SPE Annual Technical Conference and Exhibition*, 1–15. https://doi.org/10.2523/95685-MS

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. WileyRoyal Statistical Society. https://doi.org/10.2307/2346178

Appendix A. Exploratory Data Analysis

Table A 1. Variable description used in this study.

Variable	Description	Variable	Description
API	API ID of the well	PEAK_VOL	Peak Volume
YEAR	Year in which Well came online	SD_MONTHLY_VOL	Standard deviation in in the monthly volume
LAT_Y	Y coordinate of the well in CRS	VAR_MONTHLY_VOL	Variance in monthly volume
LAT_X	X coordinate of the well in CRS	MONTHY_THP	Monthly average THP
TD	Total Depth	MAX_THP	Maximum THP in Well life
ZT	Zone thickness (upper perforation - lower perforation)	SD_THP	Standard deviation in in the THP
CUMM_VOL	Total Cummulative volume	PR_GRAD	Reservoir Pressure gradient
MONTHLY_VOL	Monthly Volume	EVENT_DENSITY(OUTPUT)	Earthquake event density at a location

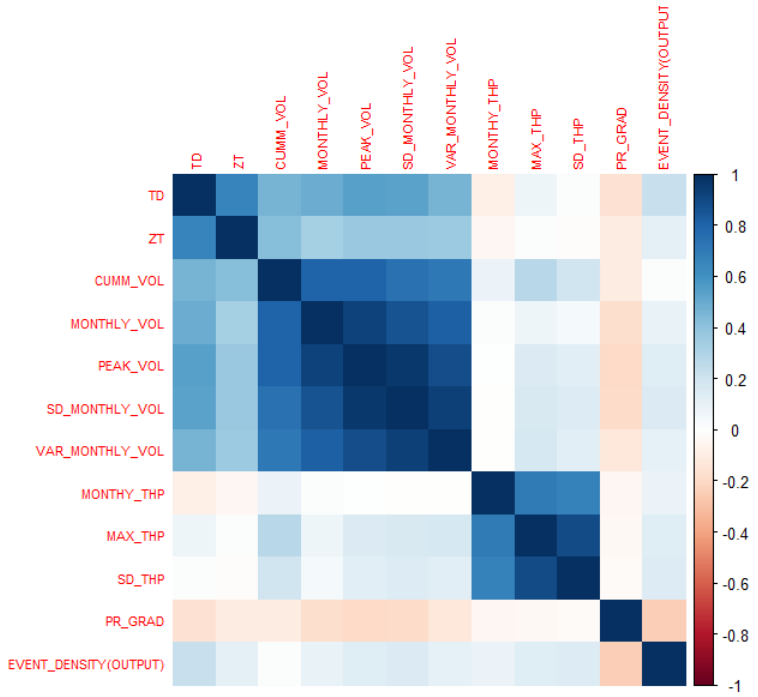


Figure A 1. Correlation plot between different variables in the study. With respect to the output variable, total depth, Zone thickness, Standard deviation in monthly volume and max THP show moderate correlation. There is multicollinearity between cumulative volume/pressure and its standard deviation, peak monthly volume/pressure and the variance of volume/pressure.

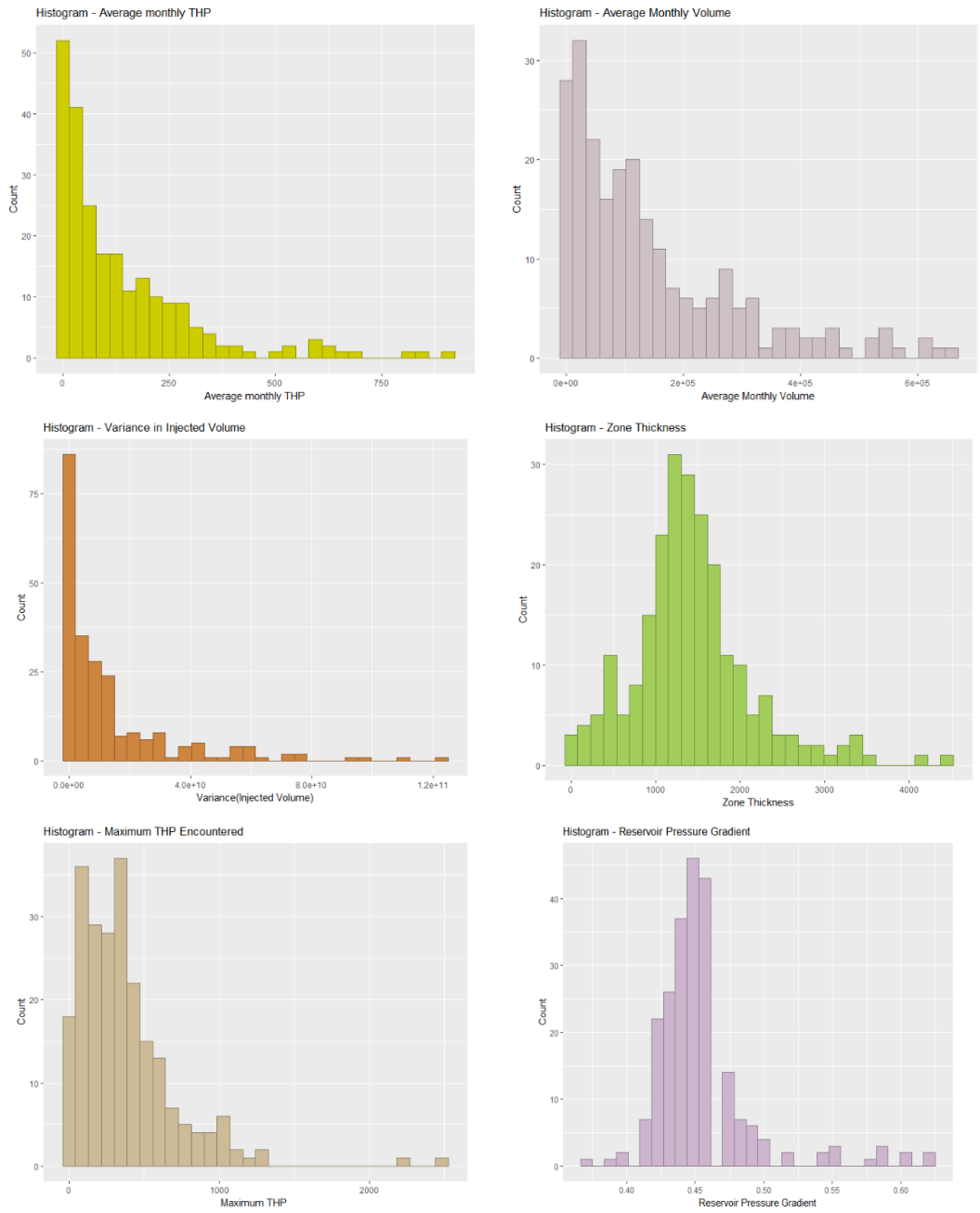


Figure A 2. Distribution of different parameters in the reservoir. The tubing head pressure and monthly volumes show log normal distributions with some outliers. The thickness of reservoir and reservoir pressure show the normal distribution.

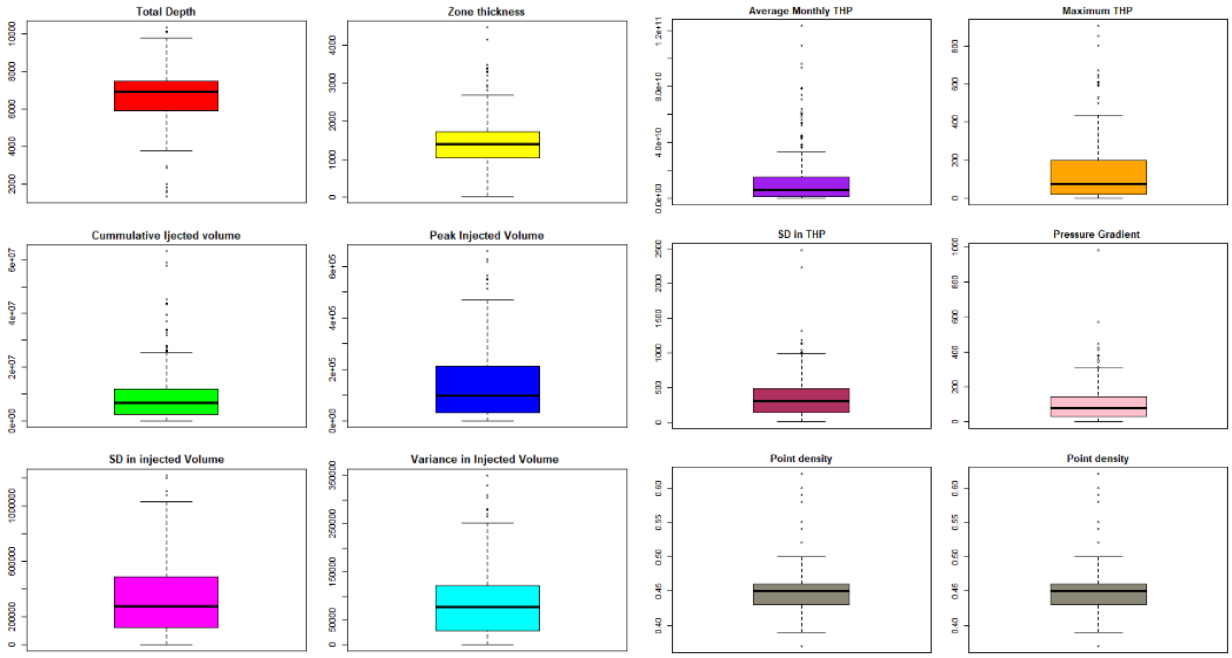


Figure A 3. Boxplots for different variables in the study. Some of the variables show outliers and after carefully examining the outliers, we delete a total of seven wells on account of erroneous data.