USING CLUSTER ANALYSIS TO ENHANCE COMPLETION OPTIMIZATION STUDIES OF THE BAKKEN PETROLEUM SYSTEM

Final Report

Prepared for:

North Dakota Industrial Commission Oil and Gas Research Program
Partners of the Bakken Production Optimization Program Consortium (BPOP 3.0)

Contract No. G-051-98

Prepared by:

Alexander Chakhmakhchev
Nicholas A. Azzolina
Bethany A. Kurz
Xue Yu
Justin T. Kovacevich
Kyle A. Glazewski
James A. Sorensen
Charles D. Gorecki
John A. Harju
Edward N. Steadman

Energy & Environmental Research Center
University of North Dakota
15 North 23rd Street, Stop 9018
Grand Forks, North Dakota

2021-EERC-04-20

April 2021

EERC DISCLAIMER

LEGAL NOTICE This research report was prepared by the Energy & Environmental Research Center (EERC), an agency of the University of North Dakota, as an account of work sponsored by the Bakken Production Optimization Program. Because of the research nature of the work performed, neither the EERC nor any of its employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement or recommendation by the EERC.

NDIC DISCLAIMER

This report was prepared by the Energy & Environmental Research Center (EERC) pursuant to an agreement partially funded by the Industrial Commission of North Dakota, and neither the EERC nor any of its subcontractors nor the North Dakota Industrial Commission nor any person acting on behalf of either:

(A) Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

(B) Assumes any liabilities with respect to the use of, or for damages resulting from the use of, any information, apparatus, method, or process disclosed in this report.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the North Dakota Industrial Commission. The views and opinions of authors expressed herein do not necessarily state or reflect those of the North Dakota Industrial Commission.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

USING CLUSTER ANALYSIS TO ENHANCE COMPLETION OPTIMIZATION STUDIES OF THE BAKKEN PETROLEUM SYSTEM

EXECUTIVE SUMMARY

Since 2016, through the Bakken Production Optimization Program (BPOP), the Energy & Environmental Research Center (EERC) has been evaluating Bakken petroleum system (BPS) oil production and well completion optimization possibilities using statistical and machine learning (ML) methods (Pekot and others, 2016[1]; Dalkhaa and others, 2019[2]; Chakhmakhchev and others, 2020[3]). An expanded analysis was conducted in 2020 focused on applying ML methods to a data set of over 12,000 BPS wells to predict well performance using completion design parameters. The results of this effort showed that the model could accurately explain the variations in production from completion parameters when applied to the same subarea of the Bakken from which the training data originated; however, it lost strength when applied to wells located in other areas of the basin. The loss of performance when applied to other regions of the basin suggested that variables not included in the model, specifically geologic factors, were important and that further work incorporating these factors would reduce prediction errors. Subsequent work under BPOP 3.0 was therefore aimed at quantifying geologic variation across the BPS using available data.

Researchers from industry and academia generally agree that both geology and completion design parameters control well production performance in unconventional oil and gas plays like the BPS. Geologic properties of reservoirs on a basin or play scale are often not directly measured and consequently omitted from many analyses of production performance. Ignoring a major source of variation may result in weak or even inaccurate prediction models, which are used in completion optimization evaluations. The purpose of this work was to develop a tool to provide a quantitative means for integrating geologic factors into completion optimization analysis of the BPS.

Two data collection efforts were conducted to create a database using publicly available, internal, and commercial data sets. First, a completion and production master database containing about 14,700 wells located in the BPS was updated, cleaned, and organized for further analysis. Second, geologic and geochemical information was compiled for the purpose of implementing

---

[1] Pekot, L.J., Dalkhaa, C., Musich, M.A., and Martin, C.L., 2016, Bakken production analysis: EERC report, 35 p.

[2] Dalkhaa, C., Azzolina, N.A., Pekot, L.J., Kurz, B.A., Kalk, B.P., and Harju, J.A., 2019, Bakken production evaluation using multivariate statistical analysis: Final report for North Dakota Industrial Commission, North Dakota Oil and Gas Research Program, and Members of the Bakken Production Optimization Program Consortium, EERC Publication 2019-EERC-04-16, Grand Forks, North Dakota, Energy & Environmental Research Center, April.

[3] Chakhmakhchev, A.V., Azzolina, N.A., Kurz, B.A., Yu, X., Dalkhaa, C., Glazewski, K.A., Sorensen, J.A., Gorecki C.D., Harju, J.A., and Steadman, E.N., 2020, Advanced analysis of Bakken data to optimize future production strategies: Report for North Dakota Industrial Commission, North Dakota Oil and Gas Research Program, and Members of the Bakken Production Optimization Program (BPOP) Contract No. G-040-080, EERC Publication 2020-EERC-06-14, Grand Forks, North Dakota, Energy & Environmental Research Center, June.

cluster analysis for the wells in the master database. The geologic data include depth and formation thicknesses, such as true vertical depths of the Bakken and Three Forks Formations (feet) and thicknesses of the total Bakken Formation, Upper/Middle/Lower Bakken Members, and Three Forks Formation (feet). Additional geologic data include formation temperature (°F) and pressure gradient (psi/ft) for the Bakken Formation and reservoir porosity (%) and permeability (mD). Other geochemical information integrated into this analysis included source rock characteristics such as HI (hydrogen index), pyrolysis temperature (T_{max}, °C), total organic carbon (TOC, %) content for both the Upper Bakken and Lower Bakken, bulk oil and gas properties, and brine chemistry data.

Using the aforementioned data sets, a web-based BPS cluster analysis calculator was built to allow the user to interactively group portions of the BPS into clusters based on sets of geologic and geochemical input variables. The calculator allows the user to select which geologic and/or geochemical variables to include, cluster calculation method, and number of clusters. The calculator provides a convenient tool to interpret results including an interactive map showing the cluster outlines, variable importance charts, cluster number determination plots, data location map, and variable distribution plots. The cluster calculator was used in multiple iterations with various combinations of numbers of clusters, statistical engines, and types of variables to evaluate the performance of the tool. The BPS cluster analysis calculator will be incorporated into future BPOP completion optimization calculations by including geologic clusters in the list of production performance predictors or through the stratification of analyses by geologic cluster.

USING CLUSTER ANALYSIS TO ENHANCE COMPLETION OPTIMIZATION STUDIES OF THE BAKKEN PETROLEUM SYSTEM

1.0 INTRODUCTION

Researchers from industry and academia generally agree that both geology and completion design parameters control well production performance in unconventional oil and gas plays like the Bakken petroleum system (BPS). While completion design parameters are measured quantities and therefore represent "hard data." the geologic properties of the reservoir characteristics on a basin or play scale are often not directly measured and consequently omitted from many analyses of production performance. However, ignoring geologic factors and heterogeneity fails to adequately capture the effect that geology may have on production. Ignoring a major source of variation may result in weak or even faulty prediction models, which are used in completion optimization evaluations.

Frequently, the question asked is, What is the best completion design strategy in each geologic setting? However, the lack of detailed data covering the whole area of investigation forces researchers to substitute measurements of geologic/reservoir properties with proxies such as 1) depth for thermal maturity, overpressure, and fracture intensity; 2) water cut for reservoir quality; 3) API (American Petroleum Institute) gravity for oil viscosity; and 4) surface well geographic location (X- and Y-coordinates) for geologic similarities. Over the past two decades, there have been several approaches for incorporating geologic proxies into data analyses seeking to evaluate well production performance across broad regional areas.

Earlier optimization studies in unconventional reservoirs using data-mining techniques included geographic coordinates of the wells and absolute reservoir depths as proxies for reservoir qualities such as pressure, thickness, and maturity level of organic-rich shale (LaFollette and others, 2012). In these earlier modeling studies, geographic coordinates, true vertical depth, and completion parameters were integrated to predict gas production in wells.

In more recent studies, similar approaches of using well geographic coordinates along with completion parameters and limited reservoir characteristics data were used in prediction models by Garjan and Ghaneezabad (2020), Zhao and others (2020), Nabors and others (2020), and Porras and others (2020). These studies demonstrated that geographic coordinates or other geologic proxies ranked as one of the top variables of importance and could explain a significant portion of the variance in production (Table 1).

Completion design optimization studies in the BPS have utilized a wide range of parameters to predict well production performance (Male and others, 2018). This broad set of parameters has included completion design details as well as geologic and reservoir characteristics of the area of interest. Male and others (2018) predicted estimated ultimate recovery (EUR) using nine parameters including water cut, production start date, API oil gravity, true vertical depth, depth of the Three Forks Formation top, initial reservoir pressure, and other completion design

Table 1. Utilization of Geologic Parameters and Their Proxies in Predictive Modeling

Study and Oil/Gas Play	Geologic, Rock Properties, and Proxy Variables Used in the Study	Results
Male and others (2018) BPS	Water cut, total vertical depth, depth of the Three Fork Formation top, initial reservoir pressure, and oil API gravity	Used geologic and completion design parameters in a gradient-boosting statistical model to predict EUR and create single-variable dependency plots. Found that water cut was the most important parameter, followed by completion date and injection volumes. Single-variable dependency plots indicated optimal volumes of injected fluid and proppant.
Luo and others (2019) BPS	True vertical depth, measured depth, formation thickness, formation depth, porosity, and water saturation, selected from 16 predictors	Used an interpolation method to extend data from 300 wells to a larger population of 3000 wells. Applied random forest to predict first-year production normalized by stage. Showed different impacts of proppant amount on production, depending on formation thickness and porosity.
Garjan and Ghaneezabad (2020) Motney Formation	Well latitude and longitude and true vertical depth	A random forest model was applied to predict production performance in 184 wells using 13 predictors. The analysis revealed that well production performance did not improve despite continuous enhancement of hydraulic fracture parameters.
Zhao and others (2020) Eagle Ford	Well latitude and longitude, oil type, total organic carbon (TOC), vitrinite reflectance equivalent, hydrocarbon pore volume, compressive strength, Young's modulus, pore pressure gradient, Upper Eagle Ford thickness, Lower Eagle Ford thickness, well depth, and oil API gravity	Random forest was used to predict EUR in the Eagle Ford play. The most import features were well depth, hydrocarbon pore volume, API gravity, and formation thickness. With a reduced number of predictors and by limiting the investigation to a smaller area, the authors managed to reduce the mean square error of prediction by 26%–52%.
Nabors and others (2020) Eagle Ford	Well latitude and longitude, true vertical depth, API gravity, and elevation	Geologic and completion parameters were used to predict 12-month production and create variable importance and dependency graphs to optimize well operation.
Porras and others (2020) Viking Formation, Canada	Well latitude and longitude, reservoir true vertical depth, net pay, and average gas:oil ratio (GOR)	Random forest was applied to predict 12-month production and create variable importance and partial dependency plots for completion optimization. Concluded that completion length, well geographic location, and net pay were the most important features that contributed to oil production.

characteristics such as volumes of fluid and proppant injected and perforated length. Luo and others (2018) combined normalized completion parameters with measured depth, well true vertical depths, total formation thickness, porosity, and water saturation to predict first-year annual production. In this context, normalized completion parameters included volume of proppant per

stage, volume of fluid per stage, normalized stage length, and first-year production per stage. Luo and others (2018) retrieved the geologic and reservoir characteristics from a limited number of vertical wells (300 wells) and then interpolated the results for the 3000 horizontal wells within their focus area. They demonstrated the impacts of completion practices on well performance for different values of porosity and thicknesses of reservoir rock. They concluded that well performance improved from an increased volume of proppant in thicker, low-porosity reservoirs and showed that thickness of the Middle Bakken and structural depth most significantly influenced the first year of production (Table 1).

Like many other petroleum resource plays, the BPS demonstrates basin-scale geologic variability. Numerous investigations of the BPS have documented differences in the source rocks, reservoir rock, reservoir pressure (P) and temperature (T), formation thickness, formation depth, fluid characteristics, and well production performance across the basin. Companies operating in the BPS ("operators") have learned by experience that these variations in geologic properties affect the optimal completion design, and therefore operators tune their well completion strategies depending on their location within the basin.

Since 2016, through the Bakken Production Optimization Program (BPOP), the Energy & Environmental Research Center (EERC) has been evaluating BPS oil production and well completion data using statistical and machine learning (ML) methods (Pekot and others, 2016; Dalkhaa and others, 2019; Chakhmakhchev and others, 2020). An expanded analysis conducted in 2020 using over 12,000 BPS wells and ML methods to predict well performance using completion design parameters demonstrated an overfitting problem on a basin scale (Chakhmakhchev and others, 2020). Stated differently, while the ML-based models performed well on the training data set and could accurately explain the variation in production from completion parameters, the models did not perform equally well on the test data set for a different set of wells that were not included in the model training and tuning. The overfitting suggested that variables not included in the model (e.g., geologic factors) were important and that further work incorporating geologic factors would reduce prediction errors. Subsequent work under BPOP 3.0 was therefore aimed at quantifying geologic variation across the BPS using available data.

To account for geologic heterogeneity across the BPS, the current work investigated categorizing the BPS into several subareas or clusters characterized by similar geology and geochemistry. In-house and publicly available data sets describing the geology and geochemistry of the BPS were compiled and interpolated. Cluster analysis was used to create subareas within the BPS characterized by similar geologic and reservoir characteristics. The remainder of this report describes the input data used in the clustering analysis, the data-processing steps used to input data into the clustering algorithms, the different clustering algorithms considered, and the results of the clustering analysis applied to the BPS. The outcomes from the clustering analysis will provide inputs to subsequent analyses conducted under BPOP to enhance completion optimization studies that account for both geologic and completion design parameters.

2.0 METHODS

Three processes were used to cluster the BPS into groups of wells with similar geologic and reservoir fluid properties: 1) data acquisition, which acquired geologic and fluid properties data from different sources; 2) data processing, where the geologic and fluid properties data were interpolated and extracted to form a tabular data set; and 3) cluster analysis, where the interpolated geology data from Step 2 were classified into groups based on clustering algorithms (Figure 1).

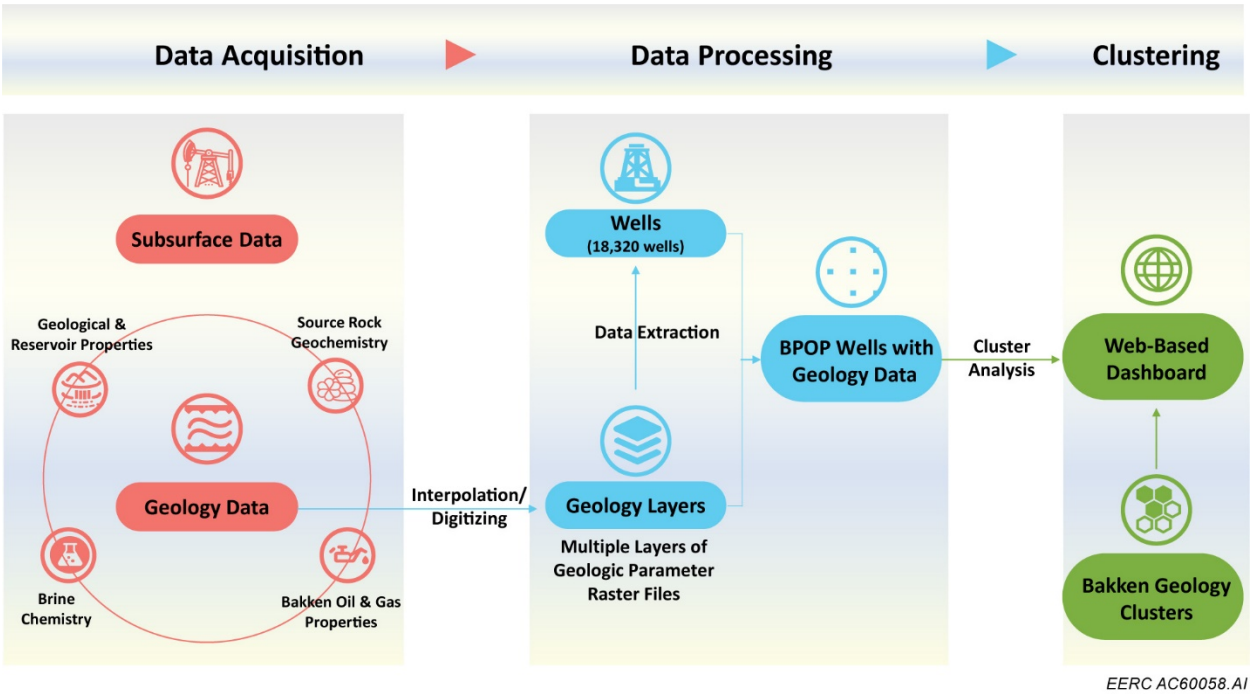


Figure 1. Flow chart showing the three processes used to cluster the BPS into groups of wells with similar geologic and reservoir fluid properties.

3.0 DATA ACQUISITION

The EERC project team compiled available geologic and reservoir fluid data for wells completed in either the Bakken or Three Forks Formation of the BPS. These data were compiled into a master data set (see Appendix A) and used as input data to the cluster analysis. The master data set was compiled from both the North Dakota Industrial Commission (NDIC) and through a DrillingInfo (now Enverus, Inc.) subscription (DrillingInfo, 2019). The master data set contains 14,691 wells. To achieve a reasonable speed of calculation and efficient visualization in the online dashboard, a fraction of the total number of wells (1832 wells, or 12%) was randomly selected from the master data set across the entire BPS. These 1832 wells were used as the base locations for the cluster analysis. The spatial distribution of the final set of 1832 wells which were used to create the classification is shown in Figure 2.

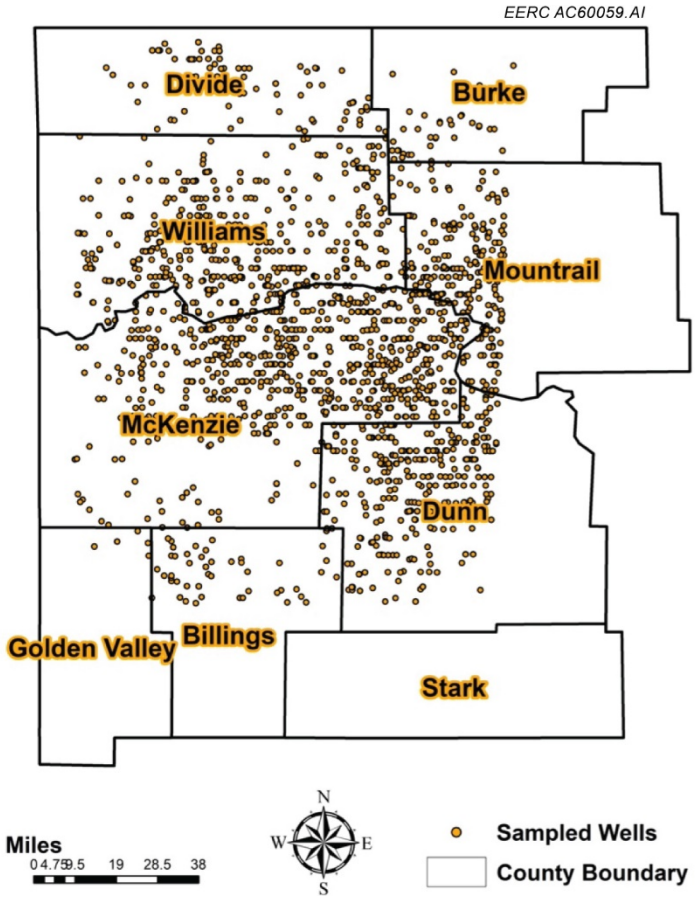


Figure 2. Spatial distribution of 1832 sampled wells completed in either the Bakken or Three Forks Formation of the BPS that were used in the online BPS cluster analysis calculator (https://eerc-ai-team.shinyapps.io/Bakken-Geology/).

Four types of subsurface data were acquired: 1) geologic and reservoir properties, 2) source rock geochemistry, 3) brine chemistry, and 4) oil and gas bulk property data. Each of these types of subsurface data are described in detail below in their respective sections.

3.1 Geologic and Reservoir Properties Data

Table 2 lists the type of geologic and reservoir data, measurement units, and their sources. The geologic data include depth and formation thicknesses, such as true vertical depths of the Bakken and Three Forks Formations (feet) and thicknesses of the total Bakken Formation, Upper/Middle/Lower Bakken Members, and the Three Forks Formation (feet). Additional geologic data include formation temperature (°F) and pressure gradient (psi/ft) for the Bakken Formation and reservoir porosity (%) and permeability (mD).

Table 2. Geologic and Reservoir Properties Data Used in the BPS Cluster Analysis

No.	Variable	Unit	Source
1	Middle Bakken temperature	°F	Sonnenberg and others (2017)
2	Lower Bakken thickness	ft	NDIC Department of Mineral Resources database query; LeFever (2008a)
3	Middle Bakken thickness	ft	Sonnenberg and others (2017)
4	Upper Bakken thickness	ft	NDIC DMR (database query)
5	Bakken thickness	ft	Calculated from the thicknesses of the three Bakken members
6	Three Forks thickness	ft	Sonnenberg and others (2017)
7	Bakken depth	ft	EERC basin model; Burton-Kelly and others (2018)
8	Three Forks depth	ft	EERC basin model. Burton-Kelly and others (2018)
9	Middle Bakken pressure gradient	psi/ft	Sonnenberg and others (2017)
10	Permeability	mD	North Dakota Geologic Survey (NDGS) database
11	Porosity	%	NDGS database

With the exception of porosity and permeability, the geologic data were interpolated using published contour maps (see Table 1 for references). While digital contour maps were interpolated directly to raster format, contour maps published as images were first georeferenced, digitized, and then interpolated to raster format. After raster versions were created for all data sets, the value of the raster layer was extracted for each wellhead location from the master data set, thereby using interpolated properties data (raster) to assign point properties data to each well.

The porosity and permeability data were collected from the NDGS database as point data from 123 wells. The sampling size was limited to 85 Middle Bakken wells, 43 Three Forks wells, and five wells had data from both formations. Vertical variations in porosity and permeability were not accounted for because there was only one data point available for each well and 2) the data for the Middle Bakken and Three Forks Formations were not distinguished, while instead data from both formations were lumped together to achieve larger spatial coverage. The porosity and permeability values were interpolated using the inverse distance weight (IDW) method to create a regional map of BPS porosity and permeability. The permeability data were skewed by larger values, thus they were log-transformed before IDW interpolation. It should be noted that because of the relatively small number of source data for porosity and permeability, the final interpolated maps of these two variables covered a smaller area than the interpolated maps for other variables.

3.2 Source Rock Geochemistry Data

Table 3 lists the source rock geochemical variables including variable name, unit, and data source. The geochemical variables include HI (hydrogen index), pyrolysis temperature (T_{max}, °C),

Table 3. Source Rock Geochemistry Data Used in the BPS Cluster Analysis

No.	Variable	Unit	Source
1	Upper Bakken HI	%	EERC and NDGS database
2	Lower Bakken HI	%	EERC and NDGS database
3	Upper Bakken T_{max}	°C	LeFever (2008b)
4	Lower Bakken T_{max}	°C	LeFever (2008b)
5	Upper Bakken TOC	%	LeFever (2008b)
6	Lower Bakken TOC	%	LeFever (2008b)

and TOC (%) for both the Upper Bakken and Lower Bakken. The sources of these parameters are point data from well measurements documented in the literature. The number of wells with measurements of HI, T_{max}, and TOC are 458, 413, and 458 wells for the Upper Bakken and 377, 332, and 377 wells for the Lower Bakken, respectively. The EERC performed interpolation using these point data to create raster files, after which the values were extracted to the master wells based on the interpolated source rock geochemical raster layers.

3.3 Brine Chemistry Data

The brine chemistry data include the concentrations of calcium (Ca), iron (Fe), magnesium (Mg), sulfate (SO_4), and sodium chloride (NaCl) in the reservoir fluids of the Bakken and Three Forks Formations (Table 4). These data were acquired from NDIC DMR. The water chemistry data from both the Bakken Formation and Three Forks Formation were grouped together for individual wells during the interpolation process. The total number of wells from both formations with measurements of each analyte were Ca (424), Fe (411), Mg (424), SO_4 (421), and NaCl (385).

Table 4. Brine Chemistry Data Used in the BPS Cluster Analysis*

No.	Variable	Unit
1	Bakken Ca	mol/L
2	Bakken Fe	mol/L
3	Bakken Mg	mol/L
4	Bakken SO_4	mol/L
5	Bakken NaCl	mol/L
6	Three Forks Ca	mol/L
7	Three Forks Fe	mol/L
8	Three Forks Mg	mol/L
9	Three Forks SO_4	mol/L
10	Bakken Ca	mol/L

* Data provided by NDIC DMR.

3.4 Oil and Gas Properties Data

Variables characterizing bulk oil composition include concentrations of sulfur and paraffin, API gravity, and viscosity. The gas composition variables include concentrations of methane (C1), ethane (C2), and the ratios of C1/C2, C1/C3, and gas wetness. Gas wetness is defined by Equation 1:

$$gaseous\ wetness = \frac{C2+C3+C4+C5}{C1+C2+C3+C4+C5} x100\% \quad \text{[Eq. 1]}$$

Where C3 is propane, C4 is butane, and C5 is pentane. The list of variables characterizing reservoir oil and gas data, measurement units, and sources are presented in Table 5. All the oil and gas properties data were interpolated from digitized raster maps derived from published contour maps.

Table 5. Oil and Gas Properties Data Used in the BPS Cluster Analysis

No.	Variable	Unit	Source
1	Bakken sulfur	mol/L	NDIC DMR
2	Bakken paraffin	%	NDIC DMR
3	Bakken oil gravity	API	NDIC DMR
4	Bakken oil viscosity	cSt	NDIC DMR
5	Bakken C1	%	NDIC DMR
6	Bakken C2	%	NDIC DMR
7	Bakken C3	%	NDIC DMR
8	Bakken C1/C2	-	NDIC DMR
9	Bakken C1/C3	-	NDIC DMR
10	Bakken wetness	%	NDIC DMR
11	Three Forks sulfur	mol/L	NDIC DMR
12	Three Forks paraffin	%	NDIC DMR
13	Three Forks oil gravity	API	Calculated from the EERC basin model (Burton-Kelly and others, 2018).
14	Three Forks oil viscosity	cSt	NDIC DMR
15	Three Forks C1	%	NDIC DMR
16	Three Forks C2	%	NDIC DMR
17	Three Forks C3	%	NDIC DMR
18	Three Forks C1/C2	-	NDIC DMR
19	Three Forks C1/C3	-	NDIC DMR
20	Three Forks wetness	%	NDIC DMR
21	Bakken sulfur	mol/L	NDIC DMR

3.5 Interpolations of Properties Data for Point Measurements

Many of the measurements were collected from individual wells and therefore represent point data with a unique x- and y-location in the BPS. The geographic locations of these points (well locations) can be found in the online BPS cluster analysis calculator (https://eerc-ai-team.shinyapps.io/Bakken-Geology/). EERC-interpolated maps were created for porosity and permeability (n = 123 wells), source rock geochemistry data, and brine chemistry data. As noted above, the other property data used existing contour maps that were published in the literature, which included geologic and reservoir properties (formation temperature, depth, thickness, and pressure gradient), and oil and gas properties data. The values of the parameters were extracted from these interpolated raster maps for each properties data type (geologic and reservoir properties, source rock geochemistry, brine chemistry, and oil and gas properties) and then compiled as the data set used in the data processing for cluster analysis.

4.0 DATA PROCESSING

After appending the interpolated data for the four types of property data to the master data set, a final tabular data set was created for input to the cluster analysis. This tabular data set contains 14,691 wells and 48 subsurface parameters from which 1832 wells (or 12%) were used in the online BPS cluster analysis calculator. The data set was preprocessed before performing cluster analysis to overcome the problems of 1) missing values, 2) outliers, and 3) scaling issue.

4.1 Missing Values

The point source wells for some variables such as porosity and permeability cover less spatial area than other variables. As a result, there are missing values in the final data set for these variables after the processes of interpolation and extraction. In this report, no procedure was conducted to address the issue of missing values due to limited data coverage.

4.2 Outliers

Very few issues of outliers among the variables were detected in the data set. All geologic parameters were standardized before cluster analysis, and the standardization process reduced the impact from potential outliers (very high or low values).

4.3 Scaling Issue

The scaling issue refers to the situation when variables used in the calculation have different scales. If the raw data were used in the cluster analysis, then some variables would have larger effects than others solely because of their measurement scale and not because of their importance in explaining variation. For example, Bakken thickness has a range of tens of feet, while permeability ranges several orders of magnitude. To avoid the scale issue, all the variables were standardized by subtracting the mean and dividing by two standard deviations (Equation 2). Dividing by two standard deviations means that a one-unit change in the scaled predictor corresponds to a change from one standard deviation below the mean to one standard deviation

above the mean (Gelman and Hill, 2007). For example, the standardized input for porosity was calculated as:

$$z.porosity = (porosity - mean[porosity])/(2*sd[porosity) \quad \text{[Eq. 2]}$$

Where:

z.porosity = The standardized value of porosity.
mean(porosity) = The average of porosity.
sd(porosity) = The standard deviation of porosity.

5.0 CLUSTER ANALYSIS

Three types of clustering algorithms were evaluated: k-means, hierarchical, and model-based. The k-means clustering algorithm was selected to run calculations in this study because it is relatively more straightforward to understand, faster in calculation, and more easily adapts to new examples with clusters of different shapes and sizes. However, all three methods are available in the BPS cluster analysis calculator (see Section 6).

5.1 Clustering Algorithm

5.1.1 K-Means Clustering Algorithm

The k-means algorithm is a relatively simple calculation that enables rapid assignment of individual points (e.g., wells) into clusters (groups). The k-means clustering method is an unsupervised ML algorithm that does not require prior information to conduct the calculation. However, the number of clusters (k number) must be specified by the user prior to calculation (i.e., the k number is a hyperparameter for the method). The selection of the k number can be determined by several different approaches such as the elbow method, which looks to minimize the within-cluster variation (within-group sum of squares, or WSS), or the silhouette approach, which measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation) (Boehmke and Greenwell, 2020). However, domain knowledge must also be utilized during k number selection, as the number of clusters and their geospatial distribution should align with domain knowledge.

Additional detail about the k-means clustering algorithm can be found in the literature (Boehmke and Greenwell, 2020; Hartigan and Wong, 1979) and is only briefly described here. As noted above, the user must specify k number of clusters. The k-means algorithm will then randomly choose k points as the starting centroids, calculate the distance of all the other points to these centroids, then group the other points to these centroids to form k clusters according to the distance values. The centroids serve as the center or mean of the cluster. In this context, distance is the similarity score in a matrix of variables, for example, the geologic and reservoir properties, source rock geochemistry, brine chemistry, and oil and gas properties data previously described. The grouping algorithm decides whether a point belongs to one cluster or another cluster based on the similarity score. The similarity score is the smallest within the cluster but is the largest between the clusters. In a series of runs, k-means will continue to randomly select k points as centroids and

perform the clustering again. The total distances of these runs of clustering are compared and the one with the smaller distance will be selected. This process of setting centroids and assigning data to the clusters repeats until the sum of the distances to the cluster centroids is minimized.

5.1.2 Hierarchical Clustering Algorithm

Hierarchical clustering is an algorithm that groups similar objects into groups (i.e., clusters) that have a predetermined order (i.e., hierarchy). There are two types of hierarchical clustering: agglomerative and divisive (Jain, 2010). In the agglomerative algorithm, each observation starts in its own cluster and merges into groups as it moves up the hierarchy until stop conditions (or a threshold in similarity) are reached. On the other hand, the divisive hierarchical algorithm starts with all observations in one cluster and divides observations into groups through successive iterations. A dendrogram, which is a treelike diagram that records the sequences of merges or splits of the data points, is used to select the number of clusters.

5.1.3 Model-Based Clustering Algorithm

The model-based clustering method assumes that the data are coming from a distribution representing a mixture of two or more components (i.e., clusters). Each component or cluster is modeled by Gaussian distribution, whereas each point is assigned a probability of belonging to each cluster. The most common approach to define the data distribution is the expectation maximum (EM) algorithm (Rodriguez and others, 2019). The EM algorithm attempts to model every point in each cluster by multivariate normal distributions, therefore, the distribution observed for the whole data set is a mixture of various normal distributions. The k-means algorithm is a special case of model-based clustering because the k-means considers the same variability for all mixed normal distributions. The key advantage of model-based clustering over standard clustering algorithms (k-means, hierarchical clustering, etc.) is that the method suggests the number of clusters in resulting calculations, while the standard algorithms require predetermination of cluster numbers. The main disadvantages of model-based clustering are 1) higher computational demand and 2) occasional inability to meet convergence criteria.

5.2 Variable Importance Charts

Variable importance charts were constructed to visualize the contribution of different variables in generating the clustering assignment. Two types of variable importance chart were explored: 1) one variable importance chart for all clusters and 2) variable importance charts for individual clusters, i.e., one chart per cluster.

5.2.1 Variable Importance Chart for All Clusters

The variable importance chart for all clusters illustrates how each variable contributes to the overall clustering. The variable importance chart can only be generated for supervised ML algorithms, while k-means clustering is an unsupervised ML algorithm. Hence, a proxy approach was used by making the calculated clustering identifier (cluster ID) the predicted variable (i.e., labeled data) and the geologic data as the independent data (or features in ML terminology). This approach enabled the supervised ML algorithms to be used to calculate the variable importance

chart. For this application, the random forest (RF) algorithm was used to generate a variable important plot. RF is a ML algorithm that utilizes a multitude of decision trees that randomly sample the features and the data and use the mean prediction of the individual trees (Breiman and others, 1984). The RF algorithm measures the importance of variables by evaluating the loss of prediction performance (e.g., root mean square error or accuracy) that occurs by eliminating independent variables one by one and permuting its values across all trees. The permutation procedure is carried out by randomly shuffling the values of the variable in the out-of-bag sample while keeping all other variables the same.

5.5.2 Variable Importance Charts for Individual Clusters

The variable importance chart built for individual clusters explains how the variables contribute to form that specific cluster. Unlike the variable importance chart for all the clusters, there is only one clustering ID, which makes using the proxy models such as RF classification algorithm meaningless. To overcome this issue, density plots of each variable were constructed. The variable contribution score was calculated as the area under the density curve. The area score approach is better than evaluation of the mean values of the variables to determine which variable is more important because the scales of the variables are different. To determine the variable importance for individual clusters, the variable with the larger area score was considered as having a larger contribution in forming the cluster.

6.0 BPS CLUSTER ANALYSIS CALCULATOR

The final geologic clustering results for the BPS were integrated into a web-based interactive dashboard called the BPS cluster analysis calculator, a stand-alone, web-based dashboard created using an application developed by Rshiny. The calculator is available to BPOP members through a link on the members-only website. This online calculator is fully interactive and allows the user to select the clustering algorithm (k-means, hierarchical, or model-based), data types (geologic and reservoir properties, source rock geochemistry data, brine chemistry data, and oil and gas properties data), and number of clusters. A general outline of the calculator interface is shown in Figure 3.

The calculator is accessible via an online dashboard, where users can interactively build the clustering results according to their specific data needs. The dashboard consists of 1) a sidebar where the user can choose different cluster analysis scenarios and 2) a main page where the clustering results and associated variable properties analysis and control variables are presented. Details of the calculator dashboard are described as follows.

6.1 Sidebar

The sidebar is located on the left-hand side of the dashboard and has two functions available to users to interactively choose different geologic clustering scenarios. The first function is the clustering algorithm, which allows the user to select from a dropdown list of three different clustering algorithms: k-means, hierarchical, and model-based clustering. The second function allows users to specify different numbers of clusters ranging from three to 12.

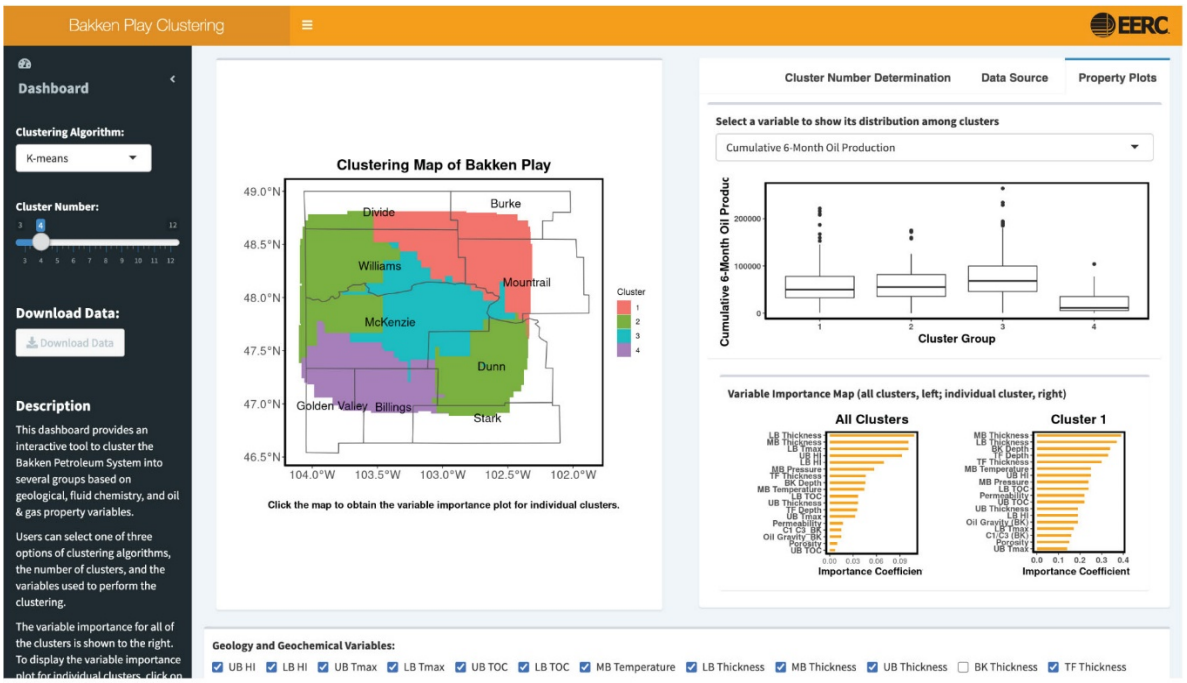


Figure 3. User interface for the web-based interactive BPS cluster analysis calculator.

In addition to these two functions, the sidebar has a Download Data button that provides users with the option to download the data they chose to include in the clustering as well as the final assigned cluster ID for each well. There is a short description on the dashboard explaining how to use different tools to interactively perform cluster analysis based on different scenarios.

6.2 Main Page

The main page of the dashboard shows the cluster analysis results and associated analysis. The main page consists of five tabs: 1) a map of the cluster analysis results, 2) property plots, 3) data source map, 4) cluster number determination, and 5) variable selection tools. Each of these components of the main page is described in greater detail as follows.

6.2.1 Cluster Analysis Results Map

The cluster analysis results map shows the nine BPS counties included in the database and polygons delineating the cluster analysis results based on user inputs. The user can click on a specific cluster to obtain the variable importance map for the individual cluster on the right tab under the Property Plots tab.

6.2.2 Property Plots

The property plots tab includes three plots: a boxplot on the top and two variable importance plots on the bottom. The boxplot shows the distribution of the selected variables among different

clusters. The boxplot helps to understand how different variables behave in different clusters. The whiskers of the boxplot show the minimum and maximum values, while the box shows the interquartile range (IQR) from the 25th percentile (P25) to the 75th percentile (P75). The horizontal line in the middle of the IQR shows the median, or 50th percentile (P50). Above the boxplot, there is a dropdown menu where users can select the variables they want to show in the boxplot.

There are two variable importance charts. The variable importance chart on the left shows the contributions of the selected variables to the classification of all clusters. The variable importance chart on the right shows the importance of the selected variables within a single cluster that the user selects from the map.

6.2.3 Data Source

The data source tab shows a map with the spatial distribution of all geologic variables in the data set used for the cluster analysis. There is a dropdown menu on top of the map where users can select the variable they want to visualize in the map. The map shows the spatial distribution of locations of the original point data sources as well as the number of wells of the point sources.

6.2.4 Cluster Number Determination

The cluster number determination tab helps users to determine the optimum number of clusters to include in the cluster analysis. This tool is only a reference for users who may also apply their own domain knowledge to select the optimum cluster number from three to 12. The tool is designed to be interactive with the specific clustering algorithm. For example, the elbow approach plotting WSS versus number of clusters is used to guide the selection of the k number for k-means clustering. A dendrogram plot is used to select the k number for hierarchical clustering. Lastly, a Bayesian information criterion (BIC) plot of various models is used to select the k number for the model-based clustering algorithm. It should be noted that the calculation of the BIC plot for the model-based clustering can take additional time because of the complexity of the calculations.

6.2.5 Variable Selection Tool

At the bottom of the main page is the variable selection tool. There are three sets of checkbox selection tools available to the user to customize the list of variables or subsurface parameters used in the cluster analysis. The default selection of variables includes all geologic and geochemical variables except for Bakken thickness as well as the variables of Bakken oil gravity and the ratio of C1/C3 of Bakken produced gas. To customize variable selection, users can check or uncheck any variables listed in the three main categories. The cluster analysis will automatically update to reflect the checked variables.

7.0 EXAMPLE APPLICATION OF THE BPS CLUSTER ANALYSIS CALCULATOR

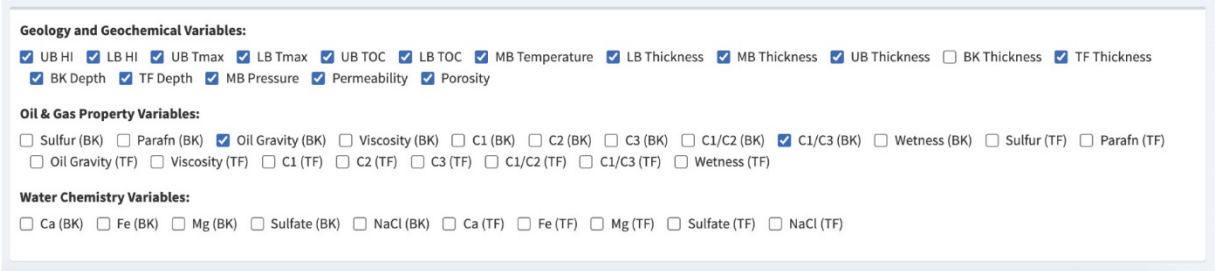
This section provides an example application of the BPS cluster analysis calculator and walks the user through variable selection, cluster analysis implementation, and interpretation of

results. The calculator is an interactive tool that allows the user to select literally hundreds of different combinations of variables; therefore, the example shown here is solely for illustration purposes.

7.1 Variable Selection Tool

The first step in the cluster analysis is to select the input variables that the tool will use to group the wells into clusters. These selections are made under the variable selection tool at the bottom of the main page.

For the current example, two categories of variables were used in the cluster analysis. All the geologic and geochemical variables were selected. In addition, two of the oil and gas property variables were selected: oil API gravity (BK) and C1/C3 (BK) gas compositional ratio, where BK refers to the Bakken Formation (Figure 4).



EERC AC60061.AI

Figure 4. Example screenshot of the variable selection tool showing the selection of all geologic and geochemical variables and oil and gas property variables of oil gravity (BK) and C1/C3 (BK).

7.2 Cluster Analysis Results and Interpretation

Figure 5 shows the cluster analysis results map for selection of all geologic and geochemical variables, oil and gas property variables of oil gravity (BK) and C1/C3 (BK), and number of clusters set to 12 (the maximum) or four clusters. Figure 6 shows the associated elbow method plot from the cluster number determination tab. As shown in Figure 6, while 12 clusters provide the minimum WSS, there is a significant decrease in WSS from one to four clusters (from approximately 6500 to 3500), which asymptotically improves with each additional cluster from five to 12. The 12-cluster map is only shown for comparison purposes, and all subsequent results and interpretations are based on the four-cluster map. The four-cluster map was selected for two primary reasons. First, the boundaries of each cluster are well-defined and do not result in an overlapping mosaic of clusters like the 12-cluster example. Second, the geologic and geochemical variables driving classification, as shown in the variable importance charts, are consistent with domain expertise in the BPS, described as follows.

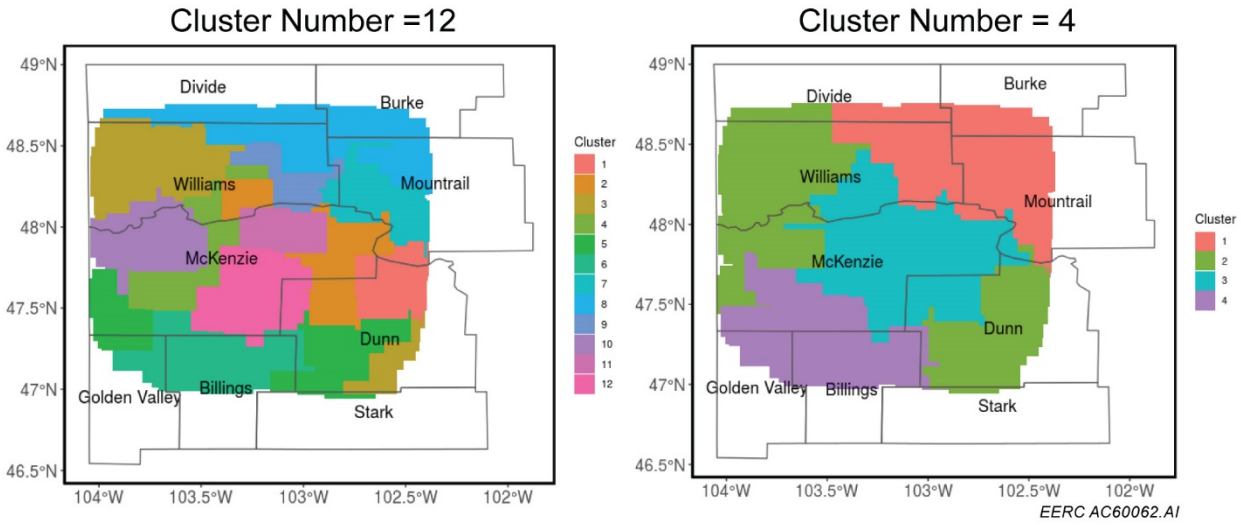


Figure 5. Cluster analysis results map using all geologic and geochemical variables, oil and gas property variables of oil gravity (BK) and C1/C3 (BK), and setting the cluster number equal to 12 (left) and four (right).

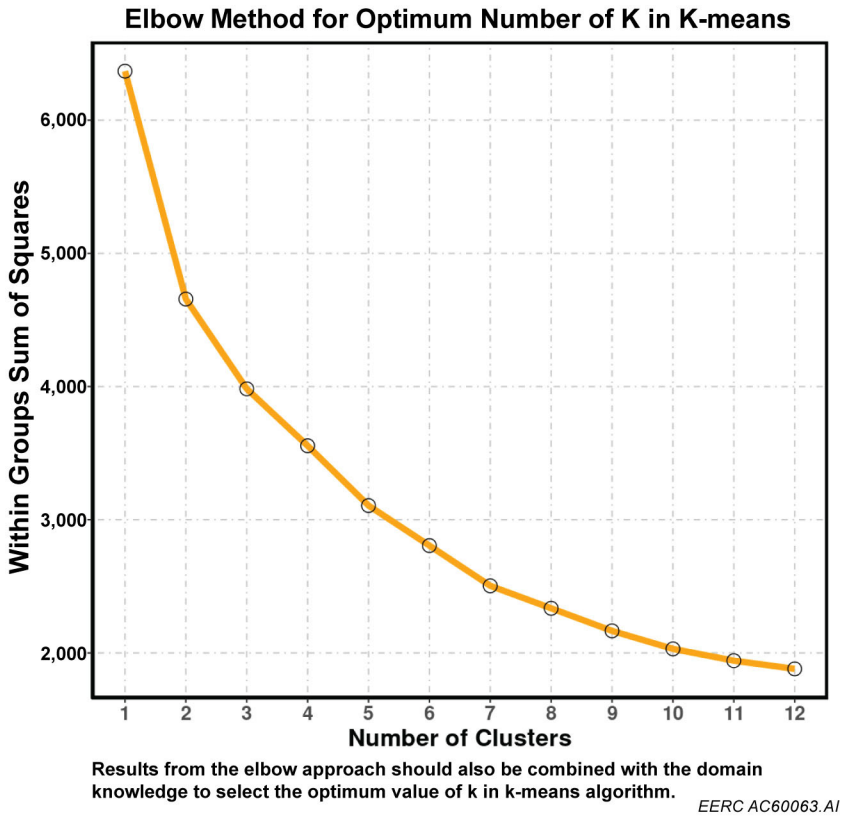


Figure 6. Elbow method plot for the k-means example using all geologic and geochemical variables and oil and gas property variables of oil gravity (BK) and C1/C3 (BK).

Figures 7 and 8 show the feature importance charts for Cluster N3 and Cluster N4, respectively. Cluster N3, which comprises central and eastern McKenzie County, northwestern Dunn County, and a portion of southern Williams County, has the highest average oil production performance, the greatest reservoir depth, and the highest reservoir temperature and pressure. These geologic settings translate into higher maturity levels of source rocks (Upper and Middle Bakken shales), as indicated by the highest T_{max} and lowest HI (or higher transformation ratio) values. The bulk hydrocarbon properties impacted by higher thermal maturity are characterized by the highest API gravity values, the lowest paraffin and sulfur contents, and dryer gas composition. The variable importance chart for Cluster N3 supports these interpretations and shows that gas composition, reservoir temperature, pressure, depth, and maturity indicators have the highest importance coefficients (Figure 7).

In contrast to Cluster N3, Cluster N4, which comprises northern Billings County, northeastern Golden Valley County, and southwestern McKenzie County, has the lowest average oil production performance, Bakken and Three Forks thickness, reservoir porosity, and TOC content in the source rocks (Upper Bakken Shale [UBS] and Lower Bakken Shale [LBS]). The variable importance chart for Cluster N4 supports these interpretations, as formation thicknesses and TOC content in the source rocks have the highest importance coefficients (Figure 8).

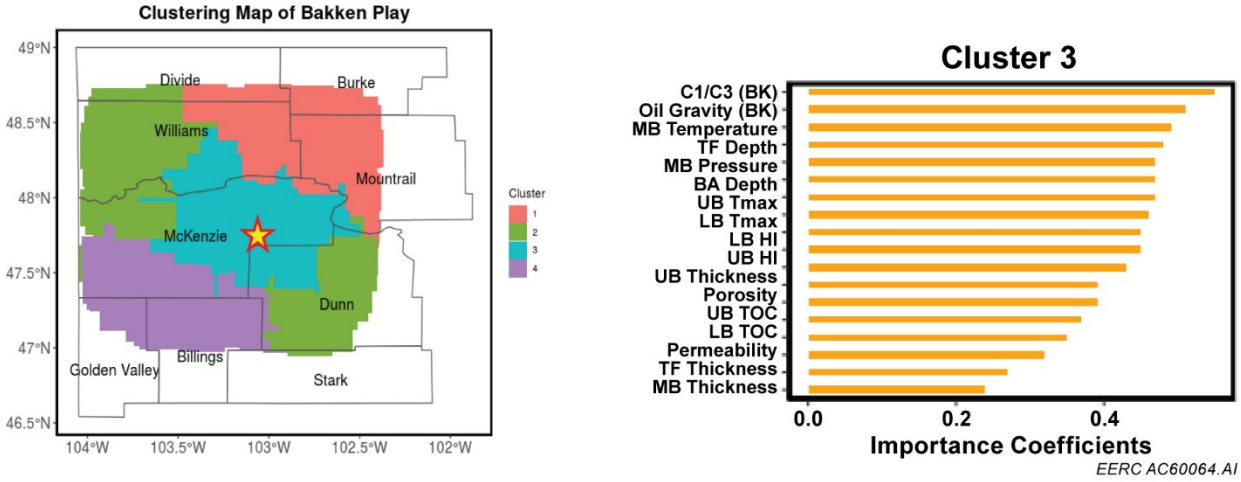


Figure 7. Cluster analysis results map (left) and variable importance chart for Cluster N3 (right) for the k-means example using all geologic and geochemical variables and oil and gas property variables of oil gravity (BK) and C1/C3 (BK).

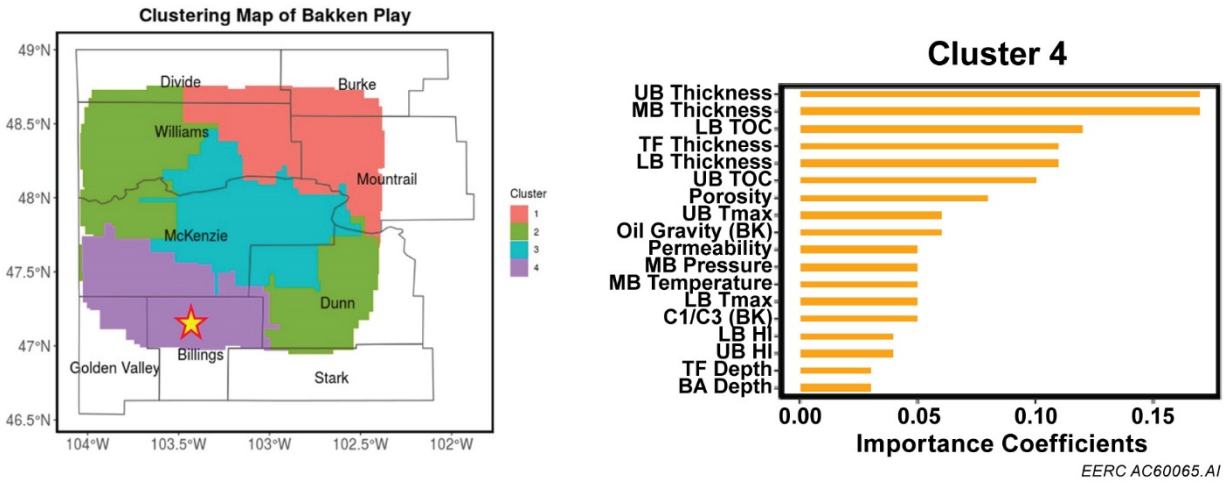


Figure 8. Cluster analysis results map (left) and variable importance chart for Cluster N4 (right) for the k-means example using all geologic and geochemical variables and oil and gas property variables of oil gravity (BK) and C1/C3 (BK).

Figure 9 depicts boxplots of 6-month cumulative oil production, Middle Bakken thickness, Upper Bakken TOC, and Oil API gravity for Clusters N1 through N4, which were obtained from the property plots tab and further support the interpretation of the variable importance charts and cluster assignments. As shown in the figure, there were significantly different distributions of these four variables among the clusters. These supporting figures and other information available from the data source and property plot tabs provide additional lines of evidence for incorporating cluster groupings into future work on production optimization. The BPS cluster analysis calculator therefore provides the user with an interactive experience for investigating the cluster analysis results and interpretations.

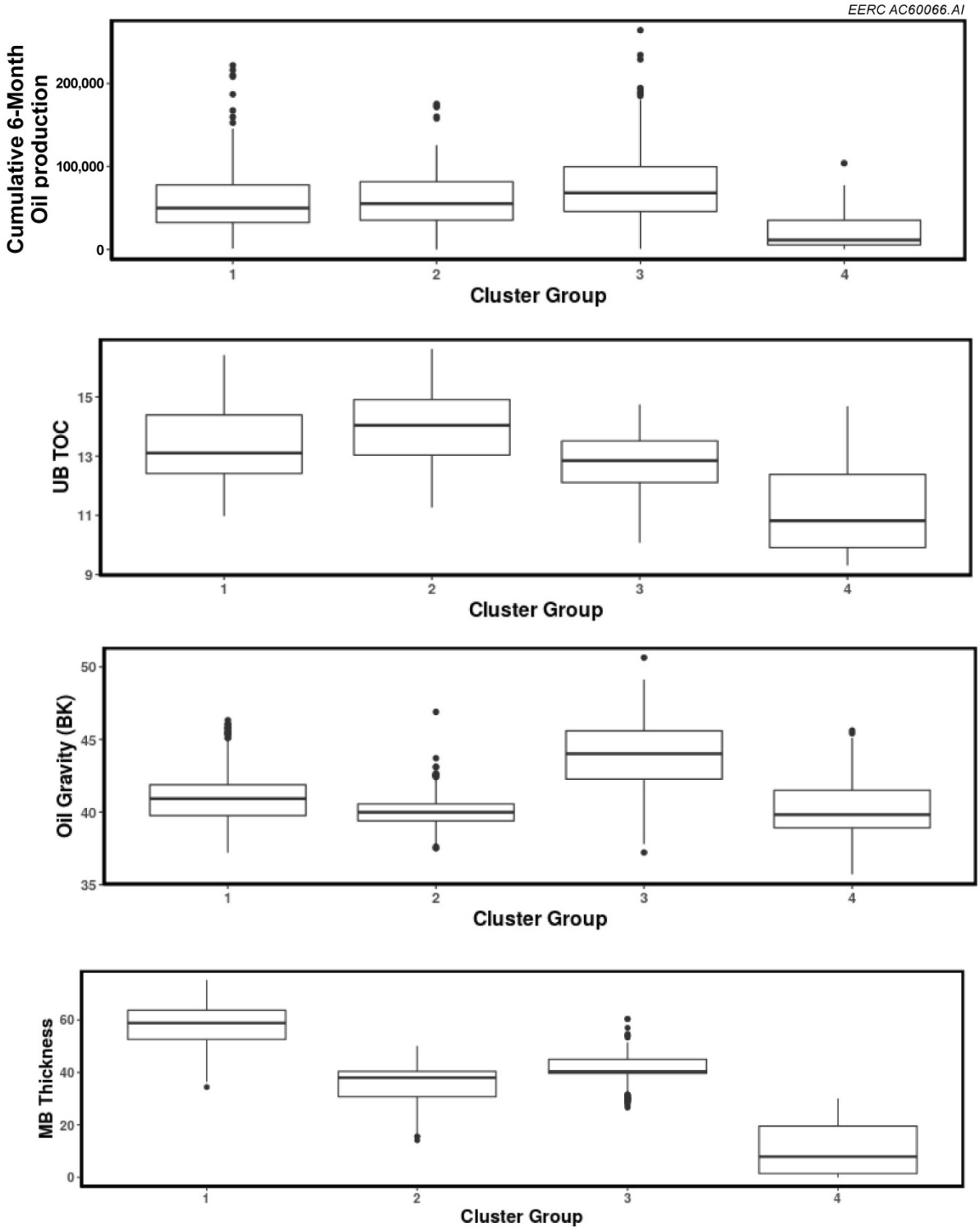


Figure 9. Boxplots of cumulative 6-month oil production, Middle Bakken thickness, Upper Bakken TOC, and Oil API gravity for the k-means example using four clusters and all geologic and geochemical variables and oil and gas property variables of oil gravity (BK) and C1/C3 (BK).

8.0 CONCLUSIONS

Researchers from industry and academia generally agree that both geology and completion design parameters control well production performance in unconventional oil and gas plays like the BPS. Geologic properties of the reservoir characteristics on a basin or play scale are often not directly measured and consequently omitted from many analyses of production performance. However, ignoring geologic factors and heterogeneity fails to adequately capture the effect that geology may have on the production result. Ignoring a major source of variation may result in weak or even inaccurate prediction models, which are used in completion optimization evaluations. The purpose of this work was to develop a tool to help provide a quantitative means for integrating geologic factors into analyses of the BPS.

Two data collection efforts were conducted to create a database using publicly available, internal, and commercial data sets. First, a completion and production master database containing about 14,700 wells located in the BPS was updated, cleaned, and organized for further analysis. Second, geologic and geochemical information was compiled for the purpose of implementing cluster analysis for the wells in the master database.

A web-based BPS cluster analysis calculator was built to allow the user to interactively group portions of the BPS into clusters based on sets of input variables. The calculator allows the user to select variables or subsurface characteristics, method of cluster calculations, and number of clusters. The calculator provides a convenient tool to interpret results including an interactive map showing cluster outlines, variable importance charts, cluster number determination plots, data location map, and variable distribution plots.

The BPS cluster analysis calculator will be incorporated into future BPOP completion optimization calculations by including geologic clusters in the list of production performance predictors or through the stratification of analyses by geologic cluster. The EERC team believes that this will significantly improve the evaluation of optimal completion parameters at various locations within the BPS.

9.0 REFERENCES

Boehmke, B., and Greenwell, B., 2020, Hands-on machine learning with R: Boca Raton, Florida, CRC Press.

Breiman, L., Friedman, J., Olshen, R., and Stone, C., 1984, Classification and regression trees: New York, Wardsworth.

Burton-Kelly, M.E., Dotzenrod, N.W., Feole, I.K., Peck, W.D., He, J., Butler, S.K., Kurz, M.C., Kurz, B.A., Smith, S.A., and Gorecki, C.D., 2018, Identification of residual oil zones in the Williston and Powder River Basins: Final report for U.S. Department of Energy National Energy Technology Laboratory Cooperative Agreement No. DE-FE0024453, EERC Publication 2018-EERC-03-05, Grand Forks, North Dakota, Energy & Environmental Research Center, March.

Chakhmakhchev, A.V., Azzolina, N.A., Kurz, B.A., Yu, X., Dalkhaa, C., Glazewski, K.A., Sorensen, J.A., Gorecki C.D., Harju, J.A., and Steadman, E.N., 2020, Advanced analysis of Bakken data to optimize future production strategies: Report for North Dakota Industrial Commission, North Dakota Oil and Gas Research Program, and Members of the Bakken Production Optimization Program (BPOP) Contract No. G-040-080, EERC Publication 2020-EERC-06-14, Grand Forks, North Dakota, Energy & Environmental Research Center, June.

Dalkhaa, C., Azzolina, N.A., Pekot, L.J., Kurz, B.A., Kalk, B.P., and Harju, J.A., 2019, Bakken production evaluation using multivariate statistical analysis: Final report for North Dakota Industrial Commission, North Dakota Oil and Gas Research Program, and Members of the Bakken Production Optimization Program Consortium, EERC Publication 2019-EERC-04-16, Grand Forks, North Dakota, Energy & Environmental Research Center, April.

DrillingInfo, Inc., 2019, Website: www.enverus.com/ (accessed June 2019).

Garjan Y.S., and Ghaneezabad, M, 2020, Machine learning interpretability application to optimize well completion in Montney: SPE-200019-MS.

Gelman, A., and Hill, J., 2007, Data analysis using regression and multilevel/hierarchical models: New York, New York, Cambridge University Press.

Hartigan, J.A., ad Wong M.A., 1979, Algorithm AS 136—a k-means clustering algorithm: Journal of the Royal Statistical Society, Series C (Applied Statistics), v. 28, p. 100–108.

Jain, A.K., 2010, Data clustering—50 years beyond k-means: Pattern Recognition Letters, v. 31, p. 651–666.

LaFollette R.F., Holcomb W.D., and Aragon J., 2012, Practical data mining—analysis of Barnett shale production results with emphasis on well completion and fracture stimulation: Presented at SPE Hydraulic Fracturing Technology Conference, The Woodlands, Texas, February 6–8, 2012, SPE 152531.

LeFever, J., 2008a, Isopach of the Lower Bakken Shale: Geologic Investigations No. 59, Sheet 4, North Dakota Geological Survey.

LeFever, J., 2008b, S2 – T_{max} of the Upper Bakken Shale in North Dakota: Geologic Investigations No. 63, Sheet 6, North Dakota Geological Survey.

Luo, G., Tian, Y., Bychina, M., and Ehlig-Economides C., 2019, Production optimization using machine learning in Bakken shale: Presented at the Unconventional Resources Technology Conference, Houston, Texas, July 23–25, 2018.

Male, F., Chastity A., and Duncan I., 2018, Using data analytics to access the impact of technology change on production forecasting: Paper presented at the 2018 SPE Annual Technical Conference and Exhibition, Dallas, Texas, September 24–26, 2018, SPE 191536-MS.

Nabors, J., Voss, T., Bogdan, A., De Sario C., Fu D., and Khan S., 2020, Basin-specific machine learning models for efficient completions optimization: Presented at the Unconventional Resources Technology Conference, URTeC:2770.

Pekot, L.J., Dalkhaa, C., Musich, M.A., and Martin, C.L., 2016, Bakken production analysis: EERC report, 35 p.

Porras, L., Hawkes, C., and Islam A., 2020, Evaluation and optimization of completion design using machine learning in an unconventional light oil play: Presented at the Unconventional Resources Technology Conference, URTeC:2938.

Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L.F., and Rodrigues, F.A., 2019, Clustering algorithms—a comparative approach: PLOS ONE, https://doi.org/10.1371/journal.pone.0210236.

Sonnenberg, S.A., Theloy, C., and Jin, H., 2017, The giant continuous oil accumulation in the Bakken petroleum system, U.S. Williston Basin, *in* Merrill, R.K., and Sternbach, C.A. [eds.], Giant fields of the decade 2000–2010: AAPG Memoir, v. 113, p. 91–120.

Zhao, P., Dong, R., and Liang, Y., 2020, Regional to local machine-learning analysis for unconventional formation reserve estimation—Eagle Ford case study: SPE-201351-MS.

APPENDIX A

DATA COLLECTION PROCESS

DATA COLLECTION PROCESS

The Energy & Environmental Research Center (EERC) maintains an up-to-date master well database storing a variety of parameters in Petra, SQL Server, spreadsheet, and text files. North Dakota well data are primarily sourced from the North Dakota Industrial Commission (NDIC) website, FracFocus, and Enervus's DrillingInfo app. Other data sources include literature, maps, and internal interpretations. The ETL (extract, transform, load) process was applied to the available information with a goal of creating a data set with a single row of completion data for single lateral wells so that in-depth analysis can be performed on initial well performance.

Through the Bakken Production Optimization Program (BPOP), a master data set was created to include as much well completion data as possible. DrillingInfo was used as the primary data source as, at the time, it was the most comprehensive data source. The DrillingInfo well header is one big table that includes general well information, well completion data, cumulative production metrics, and estimated ultimate recovery (EUR) calculations. If a well has multiple completion entries, more than one lateral, or sidetracks, a singular API (American Petroleum Institute)/UWI (unique well identifier) will have multiple rows of data. To be able to do an in-depth analysis and avoid counting a well more than once, all well information was normalized (merged) into one row of data per API-reservoir.

The complexity of DrillingInfo well data records is shown below. The DrillingInfo well header data table consisted of 38,516 distinct UWI in API14 format. Of these, 5623 UWIs had more than one data entry. 4241 UWIs had two rows of data, 959 had three rows of data, and 423 had four or more rows of data. Wells may have more than one completion record because of recompletion or multilaterals. Other wells might have multiple rows of data based on DrillingInfo joining multiple tables together that had a one-to-many pairing. An example of this is when completion records were entered with more than one entry, mirroring what was input on NDIC Form 6 by the original operator. Another example is DrillingInfo including multiple survey data points, such as sidetracks, as part of its joined table.

The first step was to reduce the well information to only Bakken petroleum system wells. Because of the row duplication problem and other data errors in allocation of producing formation, data from NDIC monthly state production (www.dmr.nd.gov/oilgas/feeservices/stateprod.asp) and Bakken horizontal wells by producing zone (https://www.dmr.nd.gov/oilgas/bakkenwells.asp) were used to filter for Bakken, Three Forks, or Spanish wells by producing formation. This resulted in 15,703 total wells and 12,415 of which had a single row of data in DrillingInfo. 2629 wells had two rows of data, 498 had three rows of data, and 151 wells had four or more rows of data. The date range was then set to only wells that started to produce in 2008 or later, resulting in a 12,039 well count with one row of data and 2917 wells with two or more rows of data.

To reduce the number of wells that required manual review (from all source data), the 114 wells with four or more rows of data were removed from the data set. Of the 14,814 wells remaining, NDIC survey data for the wells were assessed and how many laterals and sidetracks each well has were determined. 123 wells were identified as having two to three laterals and removed. The final well list consisted of 14,691 single lateral wells.

The next step was the review of the completion parameters data, such as lateral length, perforation length, total fluid, total proppant, stages, etc. It was discovered that DrillingInfo's "perf interval" and "lateral length" values have deviated from the 2019 BPOP data set that used those values from the same source. DrillingInfo appears to source from NDIC Form 6, the operator drilling report in the well file, and FracFocus, but the changes seen from 2019 and 2020 are unexplained. To be complete and accurate, a data source priority list was used for well completion parameters. NDIC has shared with the EERC a data file that has some (not all) Form 6 data digitized (NDIC Stim). The data source priority list was NDIC Stim, BPOP 2019 data set, DrillingInfo 2020 download, FracFocus. NDIC Stim is not without flaws, and data were reviewed for any outliers for each variable. Most errors identified were typos by whomever originally digitized the data at the source or by the operator submitting the data.