Using PCA and PLS on Publicly Available Data to Predict the Extractability of Hydrocarbons from Shales

E. Gallmeier^a, S. Zhang^a, J. McFarlane^a

^aOak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN, USA, 37831

etgallmeier@gmail.com, shichen98@gmail.com, mcfarlanej@ornl.gov²

ABSTRACT

Prediction of hydrocarbon extraction from shale requires specialized knowledge of shale play characteristics and analysis to assess effective, economical, and sustainable implementation of oil and natural gas production. In this work, we present a statistical approach that can be used as a preliminary investigation into the hydrocarbon resource potential of a shale play based on limited data. Statistical algorithms for Principal Component Analysis (PCA) and Partial Least Squares Regression (PLS) were used to determine if depositional environments and lithographic boundary characteristics of different plays allowed prediction of specific production parameters. This project characterizes Eagle Ford and Utica formations—two high-producing shale plays in the United States—and Banff/Exshaw and Colorado formations—two recently assessed shale plays in Alberta, Canada. Partial Least Squares Regression models were unable to model gas production parameters from predictor variables, highlighting the complexity of gas formations and the need for data on microscale petrophysical characteristics. In contrast, oil production

¹ This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

² Joanna McFarlane, corresponding author. Oak Ridge National Laboratory, Bldg 4500N, MS 6181

parameters were better predicted, because bulk variables such as mineral composition appeared to correlate with oil location in mineral interfaces. As expected, a PLS model's predictive capabilities increased with specificity of data sets to particular regions of a shale play. This study indicates how PCA and PLS modeling could assist stakeholders to make preliminary decisions regarding hydrocarbon extraction, especially when limited to publicly available petrophysical data.

KEYWORDS

statistical analysis; Principal Component Analysis; Partial Least Squares Regression; predictive modeling; hydraulic fracturing for hydrocarbon recovery; nanopores in shales

ABBREVIATIONS

AGS	Alberta Geological Survey
HI	hydrogen index
MINC	mineral carbon
NIPALS	nonlinear iterative partial least squares
OI	oxygen index
PI	production index
PLS	partial least squares regression
PCA	principal component analysis
PRESS	predictive residual sum of squares
RMSE	root mean square error
RPSEA	Research Partnership to Secure Energy for America

SANS	small angle neutron scattering
TOC	total organic carbon
VIP	variable importance in projection

1. Introduction

Oil and natural gas companies have been turning to unconventional resources and using production techniques such as hydraulic fracturing and horizontal drilling to meet increasing demands for natural gas [1, 2]. Although hydraulic fracturing has been used for almost a century, application of the technique in some areas has raised questions related to water waste, groundwater contamination, and induced seismicity [1, 3, 4]. To combat these consequences, scientists and engineers are searching for ways to maximize natural gas production while minimizing environmental impact [2].

Natural gas in shale is formed from fossilized plants and animals that are buried, compressed, and heated along with clay-sized sediments over millions of years [5, 6]. The maturity of shale dictates how far into a three-stage petroleum generating process the shale has progressed [7]. Diagenesis, catagenesis, and metagenesis constitute the petroleum generating process and correspond to immature, mature, or post-mature shale respectively, depending on the type of hydrocarbon formed: kerogen, oil and wet gas, and dry gas respectively [7].

Because shale plays are inherently complex due to lithological boundary formation and depositional history, geologists use specific shale play characterizations to determine which areas to target for economical and efficient extraction of oil and natural gas [3, 8, 9]. However, creating mineralogical, lithographical, geochemical, and production characterizations for multiple shale plays from direct measurements is expensive and time-consuming. For instance, tests run at wells evaluate hydrocarbon potential and extractability of a reservoir's shales. However, core data do not exist for all wells, since coring is an expensive and slow procedure

[10]. Thus, few current or past studies directly compare these data, though critical, to production analysis [11].

Statistical methods may offer an economical and efficient way to characterize and accurately model shale plays. Numerous studies have explored the use of Neural Network models for accurate prediction of permeability, porosity, and organic geochemical properties of shales [10, 12]. Multivariate analysis has also been used in petroleum studies to predict chemical properties of oil [13]. Further, Monte Carlo simulations have been implemented to predict reservoir production parameters [14]. These methods are discussed relative to the statistical approach that we present.

The goals of this project were to characterize selected shale plays through linear modeling by applying Principal Component Analysis (PCA) and Partial Least Squares Regression (PLS) and to evaluate these methods' abilities to predict shale play production [15, 16]. We then comprehensively analyzed four high producing shale plays in the United States and Canada using publicly available data. Our approach allowed for direct analysis of critical rock characteristics that may affect production of unique shale plays.

2. Material and Methods

Mineralogical, geochemical, and petrophysical data were used from publicly available scientific articles, doctoral theses, and research papers, as well as from private companies. Two high-producing shale plays in the United States were analyzed: the Eagle Ford and the Utica [11, 17]. Two recently assessed formations in Canada, the Banff/Exshaw and Colorado Group, were also analyzed [18].

2.1 Shale Plays Used as Basis for PLS Models

2.1.1 *Banff/Exshaw, Alberta, Canada*

The Banff shale is a Mississippian age formation that spans southern Alberta. The Banff shale overlies the Exshaw Formation, which dates from the Late Devonian to Early Mississippian age [19, 20]. While the Banff shale has only recently been identified as a potential source rock, particularly in gas production, the Exshaw Formation is known to have high total organic carbon (TOC) and is a well-established source rock in the Western Canadian Sedimentary Basin [19]. The Banff has a low percentage of clay minerals and low TOC.

2.1.2 *Colorado Group, Alberta, Canada*

The Colorado shale spans almost the entirety of Alberta and dates from the Cretaceous age. Because little lithographic and stratigraphic research has been published on the Colorado Group, not much is publicly known about the hydrocarbon potential of the shale. Both upper and lower sections have highly variable mineralogy, which, along with the play's large span, makes analysis difficult [21].

2.1.3 *Eagle Ford, Texas*

The Eagle Ford shale, located in Texas, is a uniquely heterogeneous shale play due to re-elevation of the Sabine uplift during formation [22]. Deposition dates from the Late Cretaceous age [23]. Eagle Ford is classified into an upper, progradational and a lower, retrogradational unit, based on the movement of water and sediments during formation [24]. Thus, distinct differences characterize these regions; natural gas bearing rock is typical of the upper unit while

oil bearing rock characterizes the lower unit [24]. Lower Eagle Ford also has a more diverse mineralogical character than Upper Eagle Ford [22, 24].

2.1.4 Utica; Ohio and Pennsylvania

The Utica shale is Middle to Late Ordovician in age and is linked to Marcellus and Huron shale by the Appalachian Basin [25]. Utica is an emerging shale play recently discovered to contain up to 20 times more extractable natural gas than previously predicted [26]. Utica shale is rich in calcite and is poor in clay minerals [17].

2.2 Predictor and Response Data Characterizing Shale Plays

Search options for data were limited by a need for comprehensive data sets that included both predictor variables and response variables. Predictor variables included depth, maturity, density, porosity, and shale mineral fractions in weight percent; response variables included petrophysical gas chromatography analyses and petroleum fractions that characterize the nature of hydrocarbons in the shale. Table 1 lists specific predictor and response variables used in our PCA analysis and PLS regression models.

Thermal maturity and depth of core samples were used as predictor variables. Maturity of shale was represented by vitrinite reflectance (%Ro), which is a measure of the percentage of incident light reflected from the surface of sedimentary rocks [27] and generally corresponds to stages of hydrocarbon generation in shales [27].

The models used available mineralogical data for shale samples, including the composition of non-clay minerals such as quartz, pyrite, calcite, dolomite, plagioclase, mica, and siderite and clay minerals such as smectite, illite, and kaolinite. Entrained minerals (SiO_2 or

1
2
3
4 Fe₂O₃) can make shale brittle and thus more conducive to fractures, while hydrated clay tends to
5
6 be ductile and deform instead of shattering, leading to relatively few induced fractures [6, 28,
7
8 29]. The properties of clay-bearing rocks are also a complex function of burial conditions and
9
10 diagenesis, and cannot be determined from clay composition alone [30]. Understanding clay
11
12 composition and location is critical to understanding the shale-pore network because clays are
13
14 often associated with micropores in which organic material may reside and have the capability
15
16 of sorbing gas into their structure [31, 32] . Existing knowledge about clay mineral composition
17
18 in shales lags behind that of other sedimentary rocks, but recent research suggests that the
19
20 presence of clays have a significant impact in stress-porosity-permeability relations, having
21
22 been shown to decrease permeability by up to 5 orders of magnitude at fixed porosity [33].
23
24
25
26
27

28
29 Geochemical data, particularly Rock Eval Pyrolysis analyses, comprised most response
30
31 parameters. Rock Eval Pyrolysis is a common method of evaluating a source rock's
32
33 hydrocarbon potential. The method analyzes gas produced by samples by inducing kerogen
34
35 breakdown at high temperatures- up to 600 °C [19]. The method quantifies the 'S1' fraction
36
37 (free hydrocarbon), the 'S2' fraction (the remaining hydrocarbons generated through thermal
38
39 cracking of nonvolatile organic matter), and the 'S3' fraction (the amount of CO₂ released
40
41 during thermal breakdown of kerogen) [19]. In the PLS models, S2 was used as a marker of the
42
43 hydrocarbon potential of shale. T_{max} , the temperature at which the maximum amount of S2 was
44
45 generated, was taken as an indicator of maturity and type of hydrocarbon [19, 34]. The TOC
46
47 parameter represents a rock's organic loading as a fraction of mineral composition and is
48
49 usually positively correlated with the hydrocarbon potential of shale [34]. S1, S2, S3, and TOC
50
51 were used to calculate the Hydrogen Index (HI), Oxygen Index (OI) and Production Index (PI),
52
53 which provide information about the shale's production generating potential, kerogen
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 maturation and type, and the evolution of organic matter respectively [27]. TOC was used as a
5
6 predictor variable, while all other Rock Eval parameters were used as response variables.
7
8 Mineral carbon (MINC %) was included as a response variable in some data sets, although it is
9
10 not directly related to hydrocarbon production.
11
12
13

14 Parameters from adsorption isotherms were also used as response variables. In
15
16 adsorption isotherm analyses, rock samples are crushed and equilibrated with a gas over a range
17
18 pressures to mimic reservoir conditions [19]. Maximum gas capacity at particular conditions
19
20 can then be predicted from modelling experimental data with a Langmuir model of adsorption.
21
22 Langmuir volume is the theoretical limit of gas adsorption at infinite pressure, and Langmuir
23
24 pressure is the pressure at half of Langmuir volume [35]. Reservoir pressure (volume at
25
26 reservoir pressure) and maximum gas capacity then can be calculated using the isotherm curves.
27
28 Derived gas capacity from adsorption isotherms may underestimate the total amount of gas
29
30 present, since gas may also be present in pore space and fractures in shale [19, 35]. Langmuir
31
32 parameters as well as calculated reservoir characteristics were included as response variables.
33
34
35
36
37

38 Available petrophysical data consisted of bulk density and porosity in percent of bulk
39
40 volume, the measure of pore space in a sample in relation to total sample volume [36]. Intrinsic
41
42 permeability for some Colorado Group samples was also included as a response variable.
43
44 Density was used as a predictor variable for Utica models but as a response variable for
45
46 Colorado and Banff/Exshaw models because density's relationship depended on the methods of
47
48 measurement and association within each analyzed data set. Other petrophysical parameters
49
50 such as permeability were not publicly available, since these characteristics are often the most
51
52 difficult to measure and the information is often proprietary [35].
53
54
55
56
57
58
59
60
61
62
63
64
65

3. Calculation

Algorithms used in assessing the geological data include Principal Component Analysis (PCA), Partial Least Squares Regression (PLS), and associated tests of validity such as cross validation by predictive residual sum of squares (PRESS) and the F-ratio test of the null hypothesis. These tests are presented in more detail in Supplemental Information. Maple™ 18 was used as the platform to develop comprehensive PCA and PLS programs, verify program validity, run statistical tests, and create graphical representations.

Programs for PCA, PLS, and statistical tests were verified on independent data sets before creating a comprehensive program for the shale data. Principal Component Analysis was implemented using the covariance method [37]. Program functionality was confirmed comparing results from a separate PCA program based on the nonlinear iterative partial least squares (NIPALS) algorithm PCA [38]. The PLS program was adapted from the algorithm outlined in Geladi's paper as shown in blue in Figure 1 [38]. The PLS program was verified using Abdi's sample data set. Functionality was confirmed by comparing results from PLS algorithms outlined in Abdi's paper and Wold's iterative algorithm [39, 40].

Figure 1 shows a basic outline of the approach. Principal Component Analysis was first implemented to identify influential variables and characterize each shale play [41]. Partial Least Square regression models that accurately represent existing data from a variety of shale plays were created in order to find optimal predictor variables that described production parameters. Model analysis led to characterization of the shale plays according to predictor variables, which are given by specific shale characteristics, and response variables, which take the form of parameters that can be used to indicate hydrocarbon and production potential.

3.1 Data for PLS Models

Data were used from a variety of sources. Geochemical and mineralogical measurements for 118 Eagle Ford samples were obtained from Harbor's master thesis [22]. Similar measurements and petrophysical data for 10 different Eagle Ford samples were provided by Chesapeake Energy Corporation (private communication, Steve Chipera, 2015) [42]. Geochemical, mineralogical, petrophysical, and saturate/aromatic measurements for the Utica shale were also obtained from Chesapeake Energy Corporation. A production estimate was calculated as an additional response variable by multiplying porosity and gas saturation.

Data for Banff/Exshaw and Colorado were taken from the Alberta Geological Survey (AGS) online database. Rock Eval, adsorption isotherms, TOC, petrophysical measurements, and mineralogy were taken from reports also obtained from the AGS website [43-46].

Data used in PCA and in PLS, denoted $\mathbf{X}_0$ and $\mathbf{Y}_0$, were mean centered and variance scaled to remove variable units and ease calculations [16, 41].

3.2 Data Pre-processing with PCA

Primarily, PCA was used to choose correlated response variables for a PLS model. PCA was first employed to separately analyze sets of predictor variable and response variable data for each individual shale play, as shown in red in Figure 1. PCA can accentuate relations among data and variables. PCA was implemented using the covariance method [37]. Shale plays with similar sets of predictor and response variables were combined to create multi-play models as well as separate models for each shale play to enable inter-play comparison.

3.3 Partial Least Squares Model Building and Testing

After building the PLS model, cross validation was performed according to the descriptions of Geladi and Kowalski[47], Wold and coauthors [16], and Arlot and Celisse[48] to assess the model construction, as shown in green in Figure 1. Additional details of the validation are given in the supplemental information. Model analyses were successively repeated with truncated data sets applying a “leave one out” technique for data sets under 25 samples and a “leave a group out” technique for larger data sets. In addition to a cumulative PRESS statistic for every model, PRESS and Q^2 were calculated for each response variable separately to determine how the addition or removal of response variables was affecting the model. The goal was to optimize PRESS within the set of available response variables to create a model with a minimum total PRESS. Hence, the number of latent variables chosen for the analysis was based on PRESS statistics, convergence on a PRESS value, and Q^2 .

Numerous statistical tests were applied to the chosen models. An F-ratio test, $R^2(Y)$ statistic, and Variable Importance in Projection (VIP) statistic were calculated for every model [47, 49-51]. Outlier detection used the DmodX, DmodY statistics, and Dmod critical value for each model’s optimal number of latent variables[52]. However, any samples above the Dmod critical value were also examined to assess if they were outliers of the data set in order to build a more accurate PLS model.

3.4 Graphical Representation

In addition to statistical tests, the calculation of $\mathbf{B}_{PLS}$, the linear relationship between any pair of predictor and response variables, gave insight to the importance of and relations among specific predictor variables in predicting specific response variables [39]. After building the

inner relation, the first two latent vectors in $\mathbf{T}$ and $\mathbf{U}$ were selected for a $\mathbf{t}$ versus $\mathbf{u}$ score plot, graphically exhibiting the strength of the linear relation between the $\mathbf{X}$ and $\mathbf{Y}$ matrices [16]. The final stage is depicted in yellow in Figure 1.

PLS was used in conjunction with PCA to determine the most sensitive and critical response variables for production parameters of each shale play. Variables contributing large variance in PCA were often important to PLS models. VIP and $\mathbf{B}_{\text{PLS}}$ were used to understand predictor variable influence on the PLS models. Critical predictor variables were further tested by removing them from PLS model building and analyzing the resulting model stability and predictive capabilities.

3.5 Predictions

Predictions were made for samples that were not included in model building and cross validation testing from the estimated $\hat{\mathbf{Y}}$. Predicted values were compared to measured values in the data set with a Root Mean Squared Error statistic [53], as is commonly done for assessing model predictions.

4. Results

Principal Component Analysis was performed for each shale play. The predictor and response variables are listed in Table 1. Results of the most significant PCs are displayed in Figures 2 (Banff/Exshaw), 3 (Utica and Eagle Ford), 4 (Utica and Banff/Exshaw), and 5 (Eagle Ford, Colorado, and Banff/Exshaw) for predictor and response variables respectively. The results were grouped to highlight commonalities in the sets of variables. The models chosen for further analysis, numerically labeled in Table 2, were primarily selected based on the PRESS and Q^2 as described in Section 3. Predictor and response variables were combined on the basis

of information from PCA and from VIP scores to produce the highest Q^2 values for each shale play. Q^2 for select models are displayed in Table 3 and VIP for select predictor variables of PLS models are shown in Table 4. B_{PLS} for important response variables from select models are shown in Tables 5 and 6. The results are grouped based on available predictor and response variables from each of the statistical models. Further statistical tests such as R^2 and the F ratio test were not extensively analyzed because R^2 ranged from 0.78 to 0.99 for all response variables and F-ratio statistics were so high that the null hypothesis was rejected for all models.

4.1 Banff/Exshaw

PCA results for Banff/Exshaw are displayed in Figure 2 and 5, VIP are displayed in Table 4, and B_{PLS} are displayed in Table 5. As seen from comparing the Banff/Exshaw samples in Figure 2, PCA results indicated the need for separate models for outcrop and core samples due to critical mineralogical differences in the clays. The mineralogy of the core sample was much more complete than that of the outcrop, with the exception of the key variable TOC. However, in examination of the first principle component, the relationships of the key minerals in the PCA model at least showed the same sign although there was variability in magnitude. PCA of response variables shown in Figure 5 also highlights the differences in core and outcrop samples based on Rock Eval parameters. Interestingly, while inorganic carbon, TOC, Tpeak, Tmax, and PI all tracked similarly, the variables associated with hydrocarbon content (H1, S1, S2, and S3) were opposite in sign for the core and the outcrop samples.

Three models were selected for analysis: Banff/Exshaw core 1 (#7) and Banff/Exshaw core 2 (#8) were used to predict Adsorption Isotherm (Figure 4) and Rock Eval parameters (Figure 5) respectively, and Banff/Exshaw outcrop (#9) that included only outcrop samples was also used to predict Rock Eval parameters (Figure 5). Because of the mineralogical difference,

no PLS predictions were made for any core samples based on outcrop models. The difference in mineralogy between core and outcrop samples led to a difference in Q^2 as seen in Table 3 and different optimal response variables as seen in Figures 4, 5, and Table 5. The results in Table 5 for model (#8) show that the response of the HI, associated with hydrocarbon production, is universally inversely related to that of the amount of inorganic carbon present in the shales, or MINC, a finding that is not unexpected. MINC, or inorganic carbon, had a significant response in both Banff/Exshaw core 2 (#8) and Banff/Exshaw outcrop (#9) models.

From the first principal components in Figure 5, while oxygen index (OI) was an optimal response variable for the outcrop model (#9), hydrogen index (HI) and other variables related to hydrocarbon production were all positively correlated outcrop variables and was used in the PLS analysis for the core model (#8). Banff/Exshaw outcrop (#9) appeared to have had lower uncertainty because of the limited range of the predictor variables as compared to the models built from more variable data. The only well predicted adsorption isotherm parameter for Banff/Exshaw was reservoir volume in Banff/Exshaw core 1 (#7).

The presence or absence of the variable TOC did not affect the Q^2 of either Banff/Exshaw core 2 (#8) or Banff/Exshaw outcrop 1 (#9) models, probably due to the fact that hydrocarbon recovery is associated with the mineralogy of the rock that is already included in the model. However, as stated previously, B_{PLS} statistics for predictor variables to MINC and HI have opposite signs. These findings support the theory that organic carbon is better associated with nanopores within the kerogen matrix or in larger pores within the rock matrix rather than the minerals [34].

4.2 Colorado Group

1
2
3
4 For the Colorado Group, Figures 2 and 5 show PCA results, Table 3 shows Q^2 , and
5
6 Table 5 shows B_{PLS} for model #10. Predictions for the Colorado Group shale are displayed in
7
8 Table 7. Figure 5 shows that response variables TOC, S2 and HI are closely correlated for the
9
10 Colorado data set, as is also the case with the Banff/Exshaw Core and the Eagle Ford plays.
11
12 This is expected since the S2 peak in Rock Eval Pyrolysis signifies the breakdown of organic
13
14 matter in a shale sample [44]. The regression analysis shown in Table 5 showed a similar
15
16 pattern for T_{max} and mineralogical response variables as HI in the Banff/Exshaw model.
17
18
19
20
21

22
23 Although gas adsorption and permeametry data were available, no accurate PLS models
24
25 could be created to predict gas adsorption or permeametry parameters from either shale mineral
26
27 composition or Rock Eval Pyrolysis parameters. The finding may imply that the Colorado
28
29 Group samples analyzed may be primarily immature and thus, have a relatively low gas/oil
30
31 ratio. This result is supported by analyzing the samples' range of T_{max} values; a majority of the
32
33 samples analyzed had values less than 435 or between 435 and 450°C, which indicate immature
34
35 and oil prone shale respectively [21]. The lack of model accuracy may be due to the gas
36
37 adsorption targeting micropores in mineral assemblages, not in organic matter. This is supported
38
39 by Allen's work in gas adsorption, which found that the primary site of gas adsorption in
40
41 immature Colorado shales is in illite clay mineral assemblages in the pore structures of the shale
42
43 matrix [54]. This finding is also supported by the relationship between T_{max} and constituent
44
45 minerals presented in Table 5.
46
47
48
49
50
51

52
53 VIP scores for Colorado Group outcrop (#10) in Table 4 imply that TOC does not
54
55 influence T_{max} or T_{peak} in the PLS model. In fact, when TOC was removed from the PLS model,
56
57 Q^2 values dropped only negligibly. This finding again suggests that the carbon content of the
58
59
60
61
62
63
64
65

1
2
3
4 samples appears to be independent of gas adsorption, which either could be explained by the
5
6 low maturity of the shales, or possibly by the migration of hydrocarbons from the outcrop
7
8 samples [55].
9

10
11 Colorado outcrop (#10) data and the model based on Rock Eval Pyrolysis response
12
13 variables were selected for closer analysis. Predictions of $T_{\max}$ and T_{peak} were made for the four
14
15 core samples and seven outcrop samples based on the Colorado Outcrop model (#10). The
16
17 samples used for predictions were from separate sample sets (from both core and outcrop
18
19 samples) than the one used in model building, as described in supplemental materials. Table 7
20
21 displays the resulting root mean square error (RMSE) for the predictions. Surprisingly, RMSE
22
23 was lower for core samples than for outcrop samples, perhaps because the data characterizing
24
25 the predictor variables of some of the outcrop samples were not within the range of the PLS
26
27 model.
28
29
30
31
32
33
34
35
36

37 *4.3 Eagle Ford*

38
39 PCA results for Eagle Ford are shown in Figures 3 and 5, Q^2 are shown in Table 3, VIP
40
41 are shown in Table 4, and B_{PLS} are shown in Table 6. Figure 5 shows that HI and S2 are
42
43 strongly correlated (both variables have similar magnitudes and signs for PC1 and PC2), and
44
45 therefore, may contribute to the high predictive quality of HI for Eagle Ford models. Five Eagle
46
47 Ford models were chosen for final analysis: two models taking all of Harbor's samples (Eagle
48
49 Ford Harbor 1 and Eagle Ford Harbor 2), two models for Harbor's division of data into Upper
50
51 Eagle Ford and Lower Eagle Ford, models denoted as such respectively, and one model
52
53 containing both Harbor's samples and those from Chesapeake (Eagle Ford All) [22].
54
55
56
57
58
59
60
61
62
63
64
65

All Eagle Ford models have high predictive validity. Eagle Ford Harbor 1 (#1) and Eagle Ford Harbor 2 (#2) have identical Q^2 values, (>0.7), although Eagle Ford Harbor 1 has more predictor variables. The similarity between Q^2 for Eagle Ford Harbor 1 (#1) and Eagle Ford Harbor 2 (#2) shown in Table 3 suggests that overall clay mineral and dolomite fractions do not contribute much to the analysis of Eagle Ford production, since these fractions were not included in the Eagle Ford Harbor 2 (#2) model. Although Lower Eagle Ford (#3) has 37 samples, the model has a lower Q^2 value than Upper Eagle Ford (#4), suggesting more variability in the Lower Eagle Ford data. This finding is indicative of the diverse mineralogical character of Lower Eagle Ford [22, 23].

The varying importance of depth and maturity, as displayed in Table 6, appears to depend on the depth profile of the sample set. For instance, Eagle Ford #5 includes depths from 592 to 4214 m, twice the range of the more specialized Upper Eagle Ford data set (#4, with a maximum depth of 2345 m). Thus, using the Eagle Ford Formation's unique highly variable thicknesses and sharp lithological boundaries to prepare separate models for the Upper and Lower Eagle Ford provides a greater predictive validity than using all of the Eagle Ford samples (#5) [22]. This is especially the case for Lower Eagle Ford (#3), which has a diverse mineralogical character including variable amounts of plagioclase and dolomite that may explain the model's lower Q^2 and wide spread of variance in PCA.

4.4 Utica

PCA analyses for Utica (#6) are shown in Figures 3 and 4. Table 4 displays VIP and Table 6 shows B_{PLS} . Figure 4 shows that the oil saturate fraction does not correlate with the other response variables (mobile gas saturation, production and porosity). The inverse

1
2
3
4 correlation with mobile gas saturate is as would be expected in consideration of the kerogen
5
6 maturation process [34].
7

8
9 Although oil saturate did not correlate directly with any predictor variables, the best
10
11 predictive PLS model included combination of porosity and oil saturate response variables.
12
13 Hence, both predictor variables were retained for PLS modeling.
14
15

16 Only one Utica model was selected for further analysis because of low Q^2 values. The
17
18 Utica model's low Q^2 can be explained by the small sample size. The poor prediction by PLS
19
20 signified that gas producing shales may be too complex to model with the limited inputs and
21
22 that the data needs to include variables more responsive to gas production.
23
24

25
26 Table 6 shows that the predictor variables' influences on oil saturate and porosity appear
27
28 inversely related. This result may be a consequence of the Utica's large illite and mica fraction
29
30 versus expandable smectite portion, the latter being more closely associated with the amount of
31
32 oil stored in pore space [56]. Another explanation may be the Utica's inclination for mobile gas
33
34 saturate, which constitutes 80-95% of pore volume in the samples. As suggested in recent small
35
36 angle neutron scattering (SANS) research performed on shale rock, hydrocarbons related to
37
38 shale gas may reside in nanopores while those related to larger carbon chains may reside in
39
40 mineral interfaces [34]. In the Utica model, porosity (measured in % bulk volume) is correlated
41
42 with mobile gas saturate, suggesting that nanopores make up most of the volume in samples. Oil
43
44 saturate may be located at mineral interfaces not considered in sample method for calculating
45
46 porosity. Furthermore, DiStefano's finding that oil may reside at larger scale mineral interfaces
47
48 and microfractures may also explain why oil saturate is well predicted by mineralogy. Our
49
50 results thus demonstrate the importance of determining the location of hydrocarbons within the
51
52 rock in order to predict feasibility of production.
53
54
55
56
57
58
59
60

5. Discussion

5.1 Data Selection

Results from PCA provided vital information for data selection. Generally, models with acceptable predictive abilities were those that contained correlated response variables as determined by PCA prior to building the PLS model [16]. In fact, uncorrelated response variables were detrimental to the PLS model and produced low Q^2 values because in these cases predictor variables could not produce an adequate range of response. Therefore, analyzing PRESS for each individual response variable was found to be critical to selecting optimal response variables for a final PLS model. Correlated predictor variables can be helpful, as seen with depth and maturity. Because predictive validity depends on correlations between the $\mathbf{X}$ and $\mathbf{Y}$ matrices, predictor variable selection is of upmost importance to PLS modeling. Unlike many other statistical techniques, arbitrarily removing predictor variables from a PLS model is not advised. For instance, limited data and number of variables affected the Utica, Colorado and Banff/Exshaw models.

PLS models were assessed on their predictive abilities and their stability, characterized by Q^2 and convergence on a single PRESS value respectively. Hence, these values were also used to analyze the data sets. For example, the Eagle Ford Harbor 1 (#1) and Eagle Ford Harbor 2 (#2) models gave identical Q^2 , despite the deletion of five predictor variables in Eagle Ford Harbor 2, thereby showing which variables were redundant. However, the PRESS values for Eagle Ford Harbor 2 did not converge as well for Eagle Ford Harbor 1, indicating that although predictive quality may not decrease with the exclusion of redundant information, PLS models rely on redundant information for stabilization.

1
2
3
4 Although PLS creates the best predictive models for large data sets, Upper Eagle Ford's
5
6 (#4) high Q^2 demonstrates how PLS models can converge and offer accurate predictions for
7
8 small sample sizes and specific data sets [57]. The same result was also seen with the
9
10 core/outcrop divisions in Banff/Exshaw models. Furthermore, the differences in Q^2 between the
11
12 models, shown in Table 3, highlight a need for careful assessment of data for PLS. Predictive
13
14 models are most effective when samples are chosen from subsets of a population instead of a
15
16 wide sampling, as demonstrated by the Banff/Exshaw core and outcrop models.
17
18
19
20
21
22

23 24 *5.2 Model Development and Optimization*

25
26 Optimal response variables described depended greatly on the predictor variable data
27
28 available. In data sets comprising depth, maturity, and mineral composition, the Rock Eval
29
30 Pyrolysis parameters of HI, OI, S1, S2, and S3 yielded the highest Q^2 models. Thus, a
31
32 combination of mineral composition and indicators of the maturity of the rock are necessary to
33
34 predict accurately hydrocarbon extraction. Parameters that were not well predicted by PLS
35
36 models included gas saturate, intrinsic permeability and other gas containment evaluation
37
38 parameters, such as reservoir capacity in the Colorado and Banff/Exshaw. Both mineralogical
39
40 data and Rock Eval Analysis data were used as predictor variables, but no successful models
41
42 were created except for Banff/Exshaw core 1 (#7), which even then yielded only poor predictive
43
44 abilities according to the cross validation results in Table 3. The failure of reservoir capacity
45
46 and gas saturate to be modeled may be attributed to the incapability of mineral composition to
47
48 describe adequately nanopore space in which gases reside. Gas-producing shales may be too
49
50 complex to predict gas production parameters without petrophysical parameters such as pore
51
52 size distribution and open pore volume. On the other hand, oil saturate and other oil parameters
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 may be better predicted by macroscopic variables such as mineral composition, as suggested by
5
6 the Utica model, due to the oil's location at mineral interfaces.
7
8

9 The maturity of shale in each data set may have also played a factor in the inability of
10
11 PLS models to predict for gas adsorption data. For example, Colorado shale samples were
12
13 determined to be immature or oil-prone while Banff/Exshaw samples were mature and gas
14
15 prone. Additionally, the Colorado shale samples may have a majority of pore volume being
16
17 closed and inaccessible, as determined by Allen [54]. This may explain why no PLS models
18
19 containing gas adsorption response variables could be built for the Colorado shale samples,
20
21 while it was possible to build PLS models for Banff/Exshaw samples despite the models' weak
22
23 predictive abilities.
24
25
26
27
28
29
30
31

32 *5.3 Tools for Analyzing Models*

33

34 PCA primarily proved to be an invaluable tool for data pre-selection. PCA for individual
35
36 shale plays made the process of choosing and combining complementary response variables that
37
38 would produce an optimized, accurate predictive model during testing more feasible.
39
40

41 PRESS and Q^2 were critical to PLS because model predictive capability and stability
42
43 were judged on these statistics. Before in-depth analysis or drawn conclusions from PLS
44
45 models, cross-validation or similar model testing must be performed in order to establish model
46
47 validity and determine the quality of a data set [16]. Model quality was assessed using Q^2 , the
48
49 higher the value the better the fit. Individual PRESS values calculated for each response
50
51 variable were especially useful for determining how the PRESS of individual response variables
52
53 affects the cumulative PRESS and for detecting predictor variables that contribute to a high
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 predictive validity of a response variable. Thus, individual PRESS calculations were critical for
5
6 creating accurate predictive PLS models.
7

8
9 The F-ratio test and R^2 statistics on the other hand were not as useful in shale play
10
11 analysis because PLS was able to model the existing data so well that both F-statistics and R^2
12
13 statistics were high for all models and could not be used to select optimal predictor and response
14
15 variables. R^2 calculations are generally not highly regarded in PLS model analysis because R^2
16
17 can only describe goodness of fit of the PLS model, and therefore, by definition continuously
18
19 increases as more latent variables are added to the model [50]. Additionally, the continuous
20
21 decrease of F-statistics with the addition of latent vectors is to be expected because of the
22
23 breakdown of the inner relation [58]. Therefore, the F-statistic did not provide any new
24
25 information. For other data sets, however, these statistics may prove a better indicator of a data
26
27 set's ability to be modeled in PLS.
28
29
30
31
32

33
34 The outlier detection program employed using Dmodx and Dmody statistics was a
35
36 valuable addition to the PLS program. Not only did calculations inform of unrealistic samples,
37
38 but the graphical display also showed the variability of each samples' predictor and response
39
40 variable values, therefore offering further insight into the nature of the data set.
41
42

43
44 VIP statistics and $\mathbf{B}_{PLS}$ rendered the most valuable information concerning predictor and
45
46 response variables and included a wealth of information that was used to analyze and
47
48 characterize the shale plays themselves. VIP statistics, however, were somewhat redundant and
49
50 more limiting than $\mathbf{B}_{PLS}$ because the statistics did not provide information about predictor
51
52 variable correlations or their specific influence on each response variable. Thus, VIP can
53
54 support $\mathbf{B}_{PLS}$ and create an easily understood picture of the importance of variables to the
55
56 model, but does not add any information not already given by $\mathbf{B}_{PLS}$. Meanwhile, $\mathbf{B}_{PLS}$ values
57
58
59
60
61
62
63
64
65

were critical to shale play analysis because $\mathbf{B}_{\text{PLS}}$ included information about the predictor variables' relative importance and intercorrelations as indicated by magnitude and sign. Thus, $\mathbf{B}_{\text{PLS}}$ offers the best method to analyze predictor and response variables and characterize a data set.

5.4 Model Performance

We conclude that PCA and PLS are valuable methods for analyzing certain shale play production parameters on the basis of extensive PLS model analysis, statistical testing, and comparisons of results to literature. The $>0.78 R^2$ values for response variables of all models signify that PLS accurately describes the existing data while the range of Q^2 demonstrates that PLS models may attain high predictive abilities. However, in some models with $Q^2 < 0.55$, the range between R^2 and Q^2 was quite large, thus indicating that these models were unstable [52]. This may also be a contributing factor to the high RMSE of the Colorado shale model (#10) as shown in Table 7. The inclusion of a variety of geochemical response variables resulted in models with high Q^2 , indicating that PLS can accurately predict shale play hydrocarbon production parameters from mineralogy, as seen by the prediction of HI in most models. However, the inability of PLS to model gas-specific production parameters from predictor variables signifies the complexity of gas formation in shale and a need for petrophysical characteristics, particularly pore volume, pore size distribution, and permeability. Predictor variables with the greatest range in PCA were the most important for the PLS model. That fact, along with the importance of VIP and $\mathbf{B}_{\text{PLS}}$, demonstrates the utility of coupling the statistical techniques. This modeling approach should apply to other hydrocarbon containing shale data sets, particularly those that are oil-producing.

Information about predictor and response variables, the data sets, and general trends in the shale plays gleaned from statistical tests such as VIP and $\mathbf{B}_{PLS}$ can be explained and supported by scientific literature. The consistency of our statistical results with other reported research on the shale plays [19, 21-23, 34, 43, 54] demonstrates that PCA and PLS are valid methods of analyzing shale play production .

5.5 Predictive Abilities of Core versus Outcrop Samples

Because the environmental conditions experienced by the samples are very different, and even the mineralogical profiles can be different as discussed earlier with respect to the Banff/Exshaw shales, confounding core and outcrop data sets is not generally a good practice. However, the Colorado Outcrop model was used to generate predictions for both the core and outcrop sample sets, due to deficiencies in the core data set, being a limited number of samples and a lack of matching input/output variables. That both core and outcrop predictions generated reasonable Q^2 values supported this cross validation method as an indicator of a model's predictive capability.

It is significant that the PLS model made predictions for core samples more accurately than for outcrop samples as shown in Table 7. This may imply that despite location and weathering, the mineralogy of the Colorado outcrop rock samples, including quartz, pyrite, kaolinite, muscovite, illite, and the TOC value, may still be representative of core samples. However, this will be dependent on many factors not discussed in this study and should not be generalized to other outcrop/core samples. In fact, this is contrary to what was observed with the Banff/Exshaw models.

5.6 Comparison with Other Methods of Predictive Analysis

In comparison to other statistical methods, PLS regression does not render the same accuracy of prediction. Several studies indicate the value of neural networks in areas such as prediction of water saturation in oilfields, TOC, and porosity [10, 12]. For example, RMSE for artificial neural networks workflows predicting water saturation consistently ranges from around 3 to 4 [59]. This is significantly lower than for comparable PLS regression models. Predictions reported here were for one PLS model with a Q^2 value of 0.55, based on Colorado Outcrop data. Predictions based on models with a higher Q^2 should have less error, depending on the quality of the data set. However, an advantage of PLS is its much shorter model building and prediction run-time, ranging from 20 seconds to a few minutes, as compared to extensive time taken to build neural network models. Thus, PLS may be of more practical use to industry for pressing production decisions or to stakeholders for understanding issues relevant to hydrocarbon production.

Despite the aforementioned shortcomings, PLS may be able to contribute to further studies because of the sensitivity of the models to predictor variables and outlier samples. While other modeling techniques may offer superior predictive capabilities, PLS offers the ability to extensively analyze a shale play, assess a variety of input data, and identify or confirm important geological parameters in shale samples analyzed, especially due to the quick model development and statistical testing. In this study, depth, maturity, TOC and quartz were determined to be the four main predictor variables for most models, while HI, PI, and OI were the best modeled response variables.

6. Conclusions

1
2
3
4 This project used PCA and PLS to characterize four different shale plays in the U.S. and
5
6 Canada and evaluate the use of these statistical methods. Results show that PLS can be used for
7
8 preliminary research of optimal shale play production parameters; $\mathbf{B}_{\text{PLS}}$ especially can help
9
10 geologists identify specific shale play characteristics conducive to hydrocarbon extraction. The
11
12 PLS models were generally optimized when samples were chosen from specific subsets or
13
14 regions of a population instead of a wide sampling of the shale play.
15
16
17

18
19 The inability of PLS to model gas production parameters from predictor variables
20
21 signifies the complexity of gas formation in shale and a need for petrophysical characteristics
22
23 for potential predictive models to be built. This study's lack of petrophysical characteristics,
24
25 especially accessibility porosity, pore size distribution, and open pore volume, thus may have
26
27 hindered the modeling of gas production [8]. Oil saturate and other oil parameters may be better
28
29 predicted by macroscopic variables such as mineral composition, as suggested by the Utica
30
31 model, due to the oil's location at mineral interfaces. The inverse relationship of predictor
32
33 variables modeling porosity and oil saturate in Utica shale suggests a difference in pore space
34
35 location of heavy and light hydrocarbons, which may correspond to different locations of oil
36
37 and gas within the rock matrix.
38
39
40
41

42
43 Prediction error was larger than for other statistical techniques, but may yield better
44
45 results with other data sets. Petrophysical data, such as porosity, were publicly largely
46
47 unavailable because of difficulties and expenses accompanying these analyses, would provide
48
49 valuable input to the prediction of gas formation and production [2, 34, 36].
50
51
52
53
54
55
56

57 **7. Acknowledgements**

58
59
60
61
62
63
64
65

The authors wish to thank Tammy Carneim, Oak Ridge High School, for her guidance; Steve Chipera, Chesapeake Energy LLC, for providing rock eval analyses of Eagle Ford samples; and Dr. Shar Anderson, Alberta Energy Research Institute, for his help regarding the www.ags.gov.ca database, the source for the Colorado and Banff/Exshaw data.

8. Funding Sources

This work was supported by the U.S. Department of Energy, Office of Fossil Energy Research Partnership to Secure Energy for America (RPSEA), grant number NFE-11-03260, at Oak Ridge National Laboratory under contract number DE-AC05-00OR22725. T

Table 1: Predictor and Response Variables
for PCA and PLS regression models

Predictor Variables	Response Variables
Depth (m)	<u>Rock Eval Pyrolysis</u> Parameters: S1, S2, S3, HI, OI, PI, T _{max} , T _{peak}
<u>Mineralogy (wt%):</u> Quartz, Calcite, Dolomite, Pyrite, Plagioclase, Illite and Mica, Kaolinite, Albite, Orthoclase, Gypsum, Ankerite, Clinocllore, Palygorskite, Muscovite, Sepiolite	<u>Adsorption Isotherm</u> Parameters: Langmuir Pressure, Langmuir Volume, Calculated Reservoir Pressure, Calculated Reservoir Volume at reservoir pressure, Maximum Gas Capacity
Density	Density
Thermal Maturity (%Ro)	Intrinsic Permeability (KI)
Porosity (%bv)	MINC (%)
Total Organic Carbon (TOC)	

Table 2: Basic Model Details

Model	Number of samples	Number of predictor variables	Number of response variables
1. Eagle Ford Harbor 1	53	10	2
2. Eagle Ford Harbor 2	53	5	2
3. Lower Eagle Ford Harbor	37	5	2
4. Upper Eagle Ford Harbor	7	5	2
5. Eagle Ford	63	6	2
6. Utica	9	5	2
7. Banff/Exshaw core 1	6	6	1
8. Banff/Exshaw core 2	45	8	2
9. Banff/Exshaw outcrop	13	13	2
10. Colorado Outcrop	62	9	2

Table 3: Highest Q^2
for select PLS Models

PLS Model	Highest Q^2
1	0.74
2	0.74
3	0.70
4	0.89
5	0.64
6	0.30
7	0.47
8	0.63
9	0.72
10	0.55

Table 4: VIP statistics for the most common predictor variables of PLS models

Model	Depth	Maturity	Quartz	TOC	Pyrite	Calcite
1	1.37	1.39	1.22	1.27	1.20	1.15
2	1.48	1.34	1.16	1.05	-	0.99
3	1.11	1.15	1.07	1.15	-	1.00
4	0.94	1.02	1.02	1.08	-	0.99
5	1.40	1.44	1.23	1.12	1.05	0.99
6	1.39	1.30	1.07	0.99	-	-
7	-	-	1.25	1.00	-	0.95
8	-	-	1.09	0.99	0.93	0.98
9	-	-	1.17	0.99	0.78	0.94
10	-	-	1.47	1.01	1.03	1.05

Table 5: B_{PLS} for select predictor variables of Banff/Exshaw core model (#8) and Colorado Outcrop model (#10)

Model	Response Variable	Quartz	Albite	Pyrite	Kaolinite	Illite	Muscovite	Sepiolite	Calcite	Dolomite
8	HI	0.09	0.19	0.12	-	-	0.21	0.27	-0.16	-0.17
8	MINC	-0.14	-	-0.17	-	-	-0.20	-0.08	0.18	0.17
10	T _{max}	0.46	0.21	0.07	0.27	0.32	-	-	-0.07	-

Table 6: B_{PLS} for select predictor variables of Eagle Ford and Utica PLS models

Model	Response Variable	Depth	Maturity	TOC	Quartz	Calcite	Density
3	HI	-0.23	-0.23	-0.69	-0.18	-0.03	-
3	OI	-0.44	-0.44	0.29	0.10	0.03	-
4	HI	-0.27	-0.05	-0.57	0.67	0.67	-
4	OI	-0.31	0.20	-0.48	-0.41	-0.32	-
5	HI	-0.43	-0.48	0.14	-0.02	-0.04	-
6	Porosity	0.57	-0.06	-0.01	0.48	-	-0.62
6	Oil Saturate	-0.79	0.07	0.57	-0.29	-	0.56

Table 7: Root Mean Squared Error for each response variable in the Colorado Outcrop PLS model (#10)

Shale Type Analyzed	Number of Samples	Response Variable Predicted	RMSE
Core	4	T _{max}	9.9
Core	4	T _{peak}	11
Outcrop	7	T _{max}	19
Outcrop	7	T _{peak}	19

References

- [1] K.B. Gregory, R.D. Vidic, D.A. Dzombak, Water management challenges associated with the production of shale gas by hydraulic fracturing, *Elements*, 7 (2011) 181-186.
- [2] R.B. Jackson, B.R. Pearson, S.G. Osborn, N.R. Warner, A. Vengosh, Research and Policy Recommendations for Hydraulic Fracturing and Shale-Gas Extraction, in, Duke University, Durham, NC, 2011.
- [3] J.D. Arthur, B. Bohm, M. Layne, Hydraulic Fracturing Considerations for Natural Gas Wells of the Marcellus Shale, in, 2009, pp. 49-59.
- [4] G.T. Llewellyn, F. Dorman, J.L. Westland, D. Yoxtheimer, P. Grieve, T. Sowers, E. Humston-Fulmer, S.L. Brantley, Evaluating a groundwater supply contamination incident attributed to Marcellus Shale gas development, *PNAS*, 112 (2015).
- [5] K.A. Bowker, Recent Developments of the Barnett Shale Play, Fort Worth Basin, West Texas Geological Society Bulletin, 42 (2007) 4-11.
- [6] R.M. Slatt, Important Geological Properties of Unconventional Resource Shales, *Central European Journal of Geosciences*, 3 (2011) 435-448.
- [7] J. Klentzman, Master Thesis: Geochemical Controls on Production in the Barnett Shale, Fort Worth Basin, in, Baylor University, Waco, TX, 2009.
- [8] C. Chen, D. Hu, D. Westacott, D. Loveless, Nanometer-scale characterization of microscopic pores in shale kerogen by image analysis and pore-scale modeling, *Geochemistry Geophysics Geosystems*, 14 (2013) 4066-4075.
- [9] F. O'Sullivan, Shale Gas Production: The GHG Emissions Question, in: U.S. E.P.A. Natural Gas Inventory Workshop, U.S. E.P.A., Boston, MA, 2012.
- [10] M. Shrimankosh, N. Kasiri, S.M. Mousavi, Improved Permeability Prediction of a Heterogeneous Carbonate Reservoir Using Artificial Neural Networks Based on the FLOW Zone Index Approach, *Petroleum Science and Technology*, 29 (2011) 2494-2506.
- [11] G. Wang, T.R. Carr, Organic-rich Marcellus Shale lithofacies modeling and distribution pattern analysis in the Appalachian Basin, *AAPG Bulletin*, 97 (2013) 2173-2205.
- [12] S. Zendehboudi, M.A. Ahmadi, L. James, I. Chatzis, Prediction of Condensate to Gas Ratio for Retrograde Gas Condensate Reservoirs Using Artificial Neural Network with Particle Swarm Optimization, *Energy and Fuels*, 26 (2012) 3432-3447.
- [13] K. Washburn, J. Birdwell, Multivariate analysis of ATR-FTIR spectra for assessment of oil shale organic geochemical properties, *Organic Geochemistry*, 63 (2013) 1-7.
- [14] S. Cao, A.E. Abbott, I. Lerche, Risk and Probability in resource assessment as functions of parameter uncertainty, in: Quantification and Prediction of Hydrocarbon Resources, Norsk Petroleum-forum, Amsterdam, 1996, pp. 199-218.
- [15] H. Abdi, L.J. Williams, Principal Component Analysis, *Wiley Interdisciplinary Reviews: Computational Statistics*, 2 (2010) 433-459.
- [16] S. Wold, M. Sjörström, L. Eriksson, Partial least squares projections to latent structures (PLS) in chemistry, in: Encyclopedia of Chemical Technology, John Wiley and Sons Inc., New York City, NY, 1998, pp. 2006-2021.
- [17] J. Harrington, Mineralogical and TOC Trends in the Ohio Utica Shale, in: Geological Society of America, Charlotte, NC, 2012.
- [18] AGS, Alberta Geological Survey Digital Databases, in, Alberta Geological Survey, 2015.

- [19] C.D. Rokosh, J.G. Pawlowicz, H. Berhane, S.D.A. Anderson, A.P. Beaton, Geochemical and Sedimentological Investigation of the Banff and Exshaw Formations for Shale Gas Potential: Initial Results, in, Alberta Geological Survey, Edmonton, 2009.
- [20] K.C. Stoyles, G.R. Davies, M.V. Spila, Regional Stratigraphic Study of the Exshaw/Bakken Formations: Insights from Sedimentology and Ichnology, in: Recovery – 2011 CSPG CSEG CWLS Convention Ene, Calgary, 2011.
- [21] C.D. Rokosh, J.G. Pawlowicz, H. Berhane, S.D.A. Anderson, A.P. Beaton, Geochemical and Sedimentological Investigation of the Colorado Group for Shale Gas Potential: Initial Results, in, Alberta Geological Survey, Edmonton, 2009.
- [22] R.L. Harbor, Facies Characterization and Stratigraphic Architecture of Organic-Rich Mudrocks, Upper Cretaceous Eagle Ford Formation, South Texas, (2011).
- [23] U. Ononogbu, Biomarkers: Review and Application to the Eagle Ford Formation, in, The University of Texas at Arlington, Arlington, TX, 2012.
- [24] A.P. Cardneaux, Mapping of the Oil Window in the Eagle Ford Shale Play of Southwest Texas Using Thermal Modeling and Log Overlay Analysis, (2012).
- [25] M.A. Kirschbaum, Schenk, C. C.J., T.A., R.T. Ryder, R.R. Charpentier, T.R. Klett, S.B. Gaswirth, M.E. Tennyson, K.J. Whidden, Assessment of Undiscovered Oil and Gas Resources of the Ordovician Utica Shale of the Appalachian Basin Province, 2012, in, U.S. Geological Survey, 2012.
- [26] M. Hohn, S. Pool, J. Moore, A Geologic Play Book for Utica Shale Appalachian Basin Exploration, in, West Virginia Geological Survey, 2015.
- [27] C.B. Enomoto, J.L. Coleman Jr., C.S. Swezey, P.W. Niemeyer, F.T. Dulong, Geochemical and Mineralogical Sampling of the Devonian Shales in the Broadtop Synclinorium, Appalachian Basin, in Virginia, West Virginia, Maryland, and Pennsylvania, in, U.S. Geological Survey, 2015.
- [28] D.M. Jarvie, R.J. Hill, R.M. Pollastro, Assessment of the Gas Potential and Yields from Shales: the Barnett Shale Model, in: Oklahoma Geological Survey Circular 110, 2005, pp. 37-50.
- [29] V.A. Kuuskraa, Shale Gas and Shale Oil Resource Assessment Methodology, in, Advanced Resources International Inc., Washington D.C., 2013.
- [30] E. Rybacki, T. Meier, G. Dresen, What controls the mechanical properties of shale rocks? Part II: Brittleness, Journal of Petroleum Science and Engineering, 144 (2016) 39-58.
- [31] D. Ross, R.M. Bustin, The importance of shale composition and pore structure upon gas storage potential of shale gas reservoirs, Marine and Petroleum Geology, 26 (2009) 916-927.
- [32] U. Kulia, M. Prasad, Specific surface area and pore-size distribution in clays and shales, Geological Prospecting, 61 (2013) 341-362.
- [33] I. Bourg, L. Beckingham, D. DePaolo, The Nanoscale Basis of CO₂ Trapping for Geologic Storage, Environmental Science and Technology, 49 (2015) 10265-10284.
- [34] V.H. DiStefano, J. McFarlane, L.M. Anovitz, A.G. Stack, A.D. Gordon, K.C. Littrell, K.C. Chipera, S.J. Hunt, R.D. Lewis, S.A.S. Hale, R.E. Perfect, Extraction of Organic Compounds from Representative Shales and the Effect on Porosity, Journal of Natural Gas Science and Engineering, (2016) 646-660.
- [35] T. Alexander, J. Baihly, C. Boyer, B. Clark, G. Waters, V. Jochen, J. Le Calvez, R. Lewis, C.K. Miller, J. Thaeler, B. Toelle, Shale Gas Revolution, Oilfield Review, 23 (2011) 40-55.
- [36] F.J. Lucia, Petrophysical Rock Properties, in: Carbonate Reservoir Characterization: An Integrated Approach, Springer-Verlag Kirchner Heidelberg, Berlin, Germany, 2007, pp. 1-27.

- [37] D. Lay, Linear Algebra and Its Applications, Addison Wesley Longman Inc., New York, NY, 2000.
- [38] P. Geladi, B.R. Kowalski, Partial Least-Squares regression: A tutorial, *Analytica Chimica Acta* 185 (1986) 1-17.
- [39] H. Abdi, Partial Least Squares (PLS) Regression, in: M. Lewis-Beck, A. Bryman, T. Futing (Eds.) *Encyclopedia for Research Methods for the Social Sciences*, SAGE, Thousand Oaks, CA, 2003.
- [40] OriginLab, Partial Least Squares Method, in: originlab.com, Origin Lab Corporation 2015.
- [41] J. Shlens, A Tutorial on Principal Component Analysis: Derivation, Discussion, and Singular Value Decomposition, in: www.cs.princeton.edu, Cornell University Library, 2003.
- [42] P.C. Hackley, R.T. Ryder, M.H. Trippi, H. Alimi, Thermal maturity of northern Appalachian Basin Devonian shales: Insights from sterane and terpane biomarkers, *Fuel*, 106 (2013) 455-462.
- [43] A.P. Beaton, J.G. Pawlowicz, S.D.A. Anderson, C.D. Rokosh, Rock-Eval, Total Organic Carbon, Adsorption Isotherms and Organic Petrography of the Banff and Exshaw Formations: Shale Gas Data Release, in, Alberta Geological Survey, Edmonton, 2009.
- [44] A.P. Beaton, J.G. Pawlowicz, S.D.A. Anderson, Rock Eval, Total Organic Carbon, Adsorption Isotherms and Organic Petrography of the Colorado Group: Shale Gas Data Release, in, Alberta Geological Survey, Edmonton, 2009.
- [45] J.G. Pawlowicz, S.D.A. Anderson, C.D. Rokosh, A.P. Beaton, Mineralogy, Permeametry, Mercury Porosimetry and Scanning Electron Microscope Imaging of the Colorado Group: Shale Gas Data Release, in, Alberta Geological Survey, Edmonton, 2009.
- [46] J.G. Pawlowicz, S.D.A. Anderson, C.D. Rokosh, A.P. Beaton, Mineralogy, Permeametry, Mercury Porosimetry and Scanning Electron Microscope Imaging of the Banff and Exshaw Formations: Shale Gas Data Release, in, Alberta Geological Survey, Edmonton, 2009.
- [47] P. Geladi, B.R. Kowalski, An example of 2-block predictive partial least-squares regression with simulated data, *Analytica Chimica Acta*, 185 (1986) 19-32.
- [48] S. Arlot, A. Celisse, A Survey of Cross-validation Procedures for Model Selection, *Statistics Survey*, 4 (2010) 40-79.
- [49] N. Akarachantachote, S. Chadcham, K. Saithanu, Cutoff threshold of variable importance in projection for variable selection, *International Journal of Pure and Applied Mathematics*, 94 (2013) 307-322.
- [50] R. Todeschini, Useful and unuseful summaries of regression models, in: www.molecularDescriptors.eu, Milano Chemometrics & QSAR Research Group, 2010.
- [51] S. Wold, M. Sjöström, L. Eriksson, PLS regression: A basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 58 (2001) 109-130.
- [52] Umetrics, Multivariate Data Analysis and Modelling: Basic Course, Umetrics, 2003.
- [53] S. Holmes, Susan Holmes Lab: RMS Error, in, Stanford University, 2000.
- [54] N. Allen, Pores, porosity and pore size distribution of some Draupne Formation and Colorado Group shales and kerogens, in, Newcastle University: School of Civil Engineering and Geosciences, Newcastle upon Tyne, 2014.
- [55] M.M. El Nady, F.S. Ramadan, M.M. Hammad, N.M. Lotfy, Evaluation of organic matters, hydrocarbon potential, and thermal maturity of source rocks based on geochemical and statistical methods: Case study of source rocks in Ras Gharib oilfield, central Gulf of Suez, Egypt, *Egyptian Journal of Petroleum*, 24 (2015) 203-211.

- [56] D.R. Pevear, Illite and hydrocarbon exploration, *Proceedings of the National Academy of Science*, 96 (1999) 3440-3446.
- [57] J. Henseler, T.K. Dijkstra, M. Sarstedt, C.M. Ringle, A. Diamantopoulos, D.W. Straub, D.J.J. Ketchen, J.F. Hair, G.T.M. Hult, R.J. Calantone, Common Beliefs and Reality about PLS: Comments on Rönkkö and Evermann (2013), *Organizational Research Methods*, (2014) 1-28.
- [58] P. Geladi, B.R. Kowalski, An example of 2-block predictive partial least squares regression with simulated data, *Anal Chim Acta*, 185 (1986) 19-32.
- [59] N.I. Al-Bulushi, P.R. King, M.J. Blunt, M. Kraaijveld, Artificial neural networks workflow and its application in the petroleum industry, *Neural Computing and Applications*, 21 (2012) 409-421.
- [60] S. Wold, K. Esbenson, P. Geladi, Principal component analysis, *Chemometrics and Intelligent Laboratory Systems*, 2 (1987) 37-52.
- [61] A. Hyvärinen, J. Karhunen, E. Oja, Principal Component Analysis and Whitening, in: *Independent Component Analysis*, John Wiley and Sons Inc., 2001, pp. 125-137.
- [62] R.I. Cramer, *Partial Least Squares: Its Strengths and Limitations*, ESCOM, 1993.

Figure Captions

Figure 1: The flowchart outlines the PCA and PLS program. The color coding means: Red- Data preprocessing; Blue- PLS model building; Green-PLS model testing and Yellow- Graphical representation. Centered and scaled predictor and response data are denoted **X** and **Y** respectively.

Figure 2: Principal Components consisting of predictor variables from the Colorado (model #10) and Banff/Exshaw (core models #7, #8, and outcrop model #9 respectively) shale plays are displayed. The data sets have been grouped based on common input variables, although each shale play does not have a complete set.

Figure 3: Principal Components consisting of predictor variables from the Utica (model #6) and Eagle Ford (models #1-5) shale plays are displayed. The data sets have been grouped based on common input variables, although each shale play does not have a complete set.

Figure 4: Principal Components consisting of response variables from the Utica (model #6) and Banff/Exshaw (core models #7 and #8) shale plays are displayed. The data sets have been grouped based on common output variables, although each shale play does not have a complete set.

Figure 5: Principal Components consisting of response variables from the Eagle Ford (models #1-5), Colorado (model #10), and Banff/Exshaw (core models #7, #8 and outcrop model #9) shale plays are displayed. The data sets have been grouped based on common output variables, although each shale play does not have a complete set.

Figure 1

Figure 1.

Figure 2

Figure 2.

Figure 3

Figure 3

Figure 4

Figure 4

Figure 5

Figure 5

Shale Data Input

Principal Component Analysis

Analysis of Shale
Play Production

Partial Least Squares Regression