TECHNICAL UNIVERSITY OF CRETE

MINERAL RESOURCES ENGINNERING SCHOOL

FAMILY AFFILIATIONS OF DEVONIAN WESTERN CANADA OILS, USING CHEMOMETRIC METHODS

By

Papadakis Dimitrios (S.N.: 2014028008)

Diploma thesis

Submitted in part fulfillment of the requirements
for the degree of

MASTER OF SCIENCE IN PETROLEUM ENGINEERING

Scientific Advisor: Prof. Nikos Pasadakis

Examination Committee:

1) Prof. N. Pasadakis
2) Prof. D. Christopoulos
3) Prof. N. Varotsis

Chania

October 2015

To my family
Manolis, Georgia, Maria, Katerina

ACKNOWLEDGEMENTS

I would want firstly to thank my professor and supervisor of this work, Nikos Pasadakis for the guidance, the precious advices and the interest that he showed at all the duration of this thesis. Furthermore, I would like to thank professors and co-advisors of this thesis Nikolaos Varotsis and Dionisios Christopoulos for their time to read my master thesis.

At this point I also want to thank my parents for the moral and material support at all the duration of my studies at Technical University of Crete. Special thanks to my sister Katerina for her supporting at many difficulties during this thesis.

Contents

LIST OF FIGURES..6

LIST OF TABLES...8

ABSTRACT..9

1. Geological setting 11

1.1 Devonian Petroleum Systems and Exploration Potential 11

1.2 Source rock formations in Western Canada 12

1.3 Summary of previous studies on family affiliation of oils 14

2. Chemometric Exploratory data analysis 17

2.1 Dimensionality reduction 17

2.2 Principal Component Analysis – PCA... 17

2.3 Kernel Principal Component Analysis - KPCA 19

2.4 Unsupervised Clustering.. 21

2.4.1 Hierarchical Clustering 22

2.4.2 Visualizing Hierarchical Clustering Using the Dendrogram 27

2.4.3 Optimization Methods – k-Means.. 28

2.4.3.1 k-means 28

2.4.3.2 Silhouette Plot 29

3. Description of the Devonian oils composition data 31

3.1 Oil Samples 31

3.2 Compositional data 34

3.3 Gasoline and saturated range geochemical indices 40

4. Chemometric models for families' affiliation of Devonian oils 43

4.1 Model 1 Gasoline range with all compositional variables...................................... 44

4.1.1 PCA analysis for model 1 46

4.1.2 Kernel PCA for model 1 51

4.1.3 Silhouette with k-means for model 1 54

4.1.4 Clustering for model 1 55

4.2 Model 2 Saturate range with all variables 58
4.2.1 PCA analysis for model 2 58
4.2.2 Kernel PCA for model 2 65
4.2.3 Silhouette – k-means for model 2 67
4.2.4 Clustering for model 2 68
4.3 Model 3 Calculation of nine geochemical indexes from Gasoline range 69
4.3.1 PCA analysis for model 3 70
4.3.2 Kernel PCA for model 3 74
4.3.3 Silhouette – k-means for model 3 77
4.3.4 Clustering for model 3 78
4.4 Model 4 Calculation of seven indexes from saturated fraction 79
4.4.1 PCA analysis for model 4 80
4.4.2 Kernel PCA for model 4 85
4.4.3 Silhouette – k-means for model 4 89
4.4.4 Clustering for model 4 90
4.5 Summary of the results 92
4.6 Conclusion 93
5. REFERENCES 95
APPENDIX 97

List of figures

Figure 1.1: Table for lithostratigraphic units for the Devonian formations of the Western Canada Sedimentary Basin (source: Fowler et al 2001). 12
Figure 1.2: The distribution of source rocks and the major depositional facies in Western Canada Sedimentary Basin (source: Fowler et al 2001). 12
Figure 3.1: The geographical location of oil samples 33
Figure 3.2: Normalized histogram of the chromatographic peak areas from gasoline range of sample L01677 35
Figure 3.3: Normalized histogram of the chromatographic peak areas from gasoline range of sample L01822 36
Figure 3.4: Normalized histogram of the chromatographic peak areas from gasoline range of sample L02045 36
Figure 3.5: Normalized histogram of the chromatographic peak areas from saturated range of sample L01677 38
Figure 3.6: Normalized histogram of the chromatographic peak areas from saturated range of sample L01822 38
Figure 3.7: Normalized histogram of the chromatographic peak areas from saturated range of sample L02045 39
Figure 4.1: The interface of the chemometric software package 43
Figure 4.2: The Editor menu in MATLAB program 44
Figure 4.3: The dialog menu with the available options of samples' import 45
Figure 4.4: The dialog menu for sample selection 45
Figure 4.5: The workspace section with the variables for model 1 46
Figure 4.6: Plot of the two major principal components for model 1 46
Figure 4.7: The subplot of five principal components of model 1 47
Figure 4.8: The subplot for five principal components with different color for each formation of model 1 47
Figure 4.9: Plot of the first principal component of PCA analysis vs the geographical location of the samples. 48
Figure 4.10: Plot of the second principal component of PCA analysis 49
Figure 4.11: Original variable loadings for the first five principal components for the model 1. 49
Figure 4.12: Percentage of variance explained by each principal component 50
Figure 4.13: The subplot of five kernel principal components of model 1 51
Figure 4.14: The subplot of kernel principal components with different color for each formation of model 1 51
Figure 4.15: Plot of the first kernel principal component of KPCA analysis in relationship vs the geographic location of the samples. 52
Figure 4.16: Plot of the second kernel principal component of KPCA analysis in relationship vs the geographical location of the samples. 53
Figure 4.17: Silhouette plots for k=2, k=3, k=4 and k=5 clusters for model 1 54
Figure 4.18: The plot of the first two PCs of k-means clustering for k=2 of model 1 55
Figure 4.19: The dialog menu with the available choices for clustering 55
Figure 4.20: The dialog menu with the available choices for distance and linkage methods 56
Figure 4.21: The best choices for distance and linkage method are marked for model 1. 57
Figure 4.22: The hierarchical clustering dedrogram for model 1 57
Figure 4.23: The workspace section with the selected variables for model 2 58
Figure 4.24: Plot of the two major principal components of model 2 59
Figure 4.25: The subplot of five principal components of model 2 60

Figure 4.26: The subplot for five principal components of model 2 with different color for each formation 60
Figure 4.27: Plot of the first principal component of PCA analysis vs the geographical location of the samples. 61
Figure 4.28: Plot of the second principal component of PCA analysis vs the geographical location of the samples. 62
Figure 4.29: Original variable loadings for the first five principal components from the model 2. 63
Figure 4.30: Percentage of variance explained for each principal component 64
Figure 4.31: The subplot of five kernel principal components of model 2 65
Figure 4.32: The subplot for five kernel principal components of model 2 with different color for each formation 65
Figure 4.33: Plot the values of the first kernel principal component of KPCA analysis in relationship with the location of the samples. 66
Figure 4.34: Plot the values of the second kernel principal component of KPCA analysis in relationship with the location of the samples. 66
Figure 4.35: Silhouette plots for k=2, k=3, k=4 and k=5 clusters for model 2 67
Figure 4.36: The plot of the first two PCs of k-means clustering for k=2 of model 2 68
Figure 4.37: The hierarchical clustering dedrogram for model 2 69
Figure 4.38: The workspace section with the importing variables for model 3 70
Figure 4.39: Plot of the two major principal components for model 3 70
Figure 4.40: The subplot of five principal components of model 3 71
Figure 4.41: The subplot for five principal components with different color for each formation of model 3 71
Figure 4.42: Plot of the first principal component of PCA analysis vs the geographical location of the samples. 72
Figure 4.43: Plot of the second principal component of PCA analysis vs the geographic location of the samples. 72
Figure 4.44: Original variable loadings for the first five principal components from the model 3 73
Figure 4.45: Percentage of variance explained of each principal component 73
Figure 4.46: Plot of the two major kernel principal components for model 3 74
Figure 4.47: The subplot of five kernel principal components of model 3 75
Figure 4.48: The subplot for five kernel principal components with different color for each formation of model 2 75
Figure 4.49: Plot of the first principal component of PCA analysis vs the location of the samples. 76
Figure 4.50: Plot of the second kernel principal component of KPCA analysis vs the location of the samples 76
Figure 4.51: Silhouette plots for k=2, k=3, k=4 and k=5 clusters for model 3 77
Figure 4.52: The plot of the first two PCs of k-means clustering for k=5 of model 3 78
Figure 4.53: The hierarchical clustering dedrogram for model 3 79
Figure 4.54: The workspace section with the selected variables for model 4 80
Figure 4.55: Plot of the two major principal components for model 4 80
Figure 4.56: The subplot of five principal components of model 4 81
Figure 4.57: The subplot for five principal components of model 4 with different color for each formation 82
Figure 4.58: Plot of the first principal component of PCA analysis vs the geographical location of the samples. 82
Figure 4.59: Plot of the second principal component of PCA analysis vs the location of the samples. 83
The Keg River samples in location coordinates of longitude and latitude have values greater than 1.4 in first principal component and below zero for the second principal component as illustrated in figures 4.58 and 4.59. 83
Figure 4.60: Original variable loadings for the first five principal components from the model 4 83
Figure 4.61: Percentage of variance explained of each principal component 84
Figure 4.62: Plot of the two major kernel principal components for model 4 85

Figure 4.63: The subplot of five kernel principal components of model 4 86
Figure 4.64: The subplot for five kernel principal components of model 4 with different color for each formation 86
Figure 4.65: Plot of the first principal component of PCA analysis vs the geographical location of the samples. 87
Figure 4.66: Plot of the second kernel principal component of KPCA analysis vs the geographic location of the samples 88
The Keg River samples in location coordinates of longitude and latitude have near zero values in first and second principal components according to figures 4.65 and 4.66..88
Figure 4.67: Silhouette plots for k=2, k=3, k=4 and k=5 clusters for model 4..89
Figure 4.68: The plot of the first two PCs of k-means clustering for k=3 of model 4 90
Figure 4.69: The hierarchical clustering dedrogram for model 4 91

List of tables

Table 1.1: Comparison for the different source rock formations 16
Table 3.1: The geographical location and the source formation of sampled oils 32
Table 3.2: The connection of the first letter in sample names with source formation of sampled oils 34
Table 3.3: The commonly identified hydrocarbons in the gasoline range fraction of oil 35
Table 3.4: The commonly identified hydrocarbons in the saturated fraction of oils 37
Table 4.1: Color representation of oils with respect to their reservoir formation origin. 48
Table 4.2: Summary of k-means clustering for model 1 54
Table 4.3: Summary of k-means clustering for model 2 67
Table 4.3: Summary of k-means clustering for model 3 77
Table 4.4: Summary of k-means clustering for model 4 89

Abstract

The gasoline range and saturate fractions compositional data of oils carry significant geochemical information. In this master thesis we investigate the ability of chemometric methods in revealing of the oil families for the Devonian petroleum systems. The multivariate statistical methods were applied to the gasoline range and saturate fraction data for 146 oil samples from the Western Canada Sedimentary Basin.

Master thesis is organized in four chapters. Chapter 1 focuses on the introduction of the geological setting of the sampling area. Chapter 2 covers the principles for the methods of dimensionality reduction. The methods employed here are: hierarchical clustering, principal component analysis (PCA), kernel principal component analysis (KPCA), k-means clustering and hierarchical clustering. In chapter 3 we focus in a more detailed examination of geochemical information contained in gasoline and saturated fractions and we introduce the characteristics indices for these fractions. Finally in chapter 4 we examine four models aiming to separate the oils into families and we investigate the performance of each model in family affiliation. For the first two models normalized values of original component data for the gasoline and saturated fractions were used, while in models 3 and 4 calculated characteristic geochemical indices were used. Due to geochemical similarity among oil samples, the first two models did not give satisfactory results for oil affiliation. Model 3 gives the best results and reveals two distinct families in oil samples and finally model 4 give results for three different oil families.

1. Geological setting

1.1 Devonian Petroleum Systems and Exploration Potential

The Devonian is a geologic period of the Paleozoic. This era starts at the end of the Silurian period, about 420 million year ago and it ends with the beginning of the Carboniferous period, about 359 million year ago. Furthermore the Devonian period is divided further in three subdivisions, the Early, Middle and Late. For the needs of master thesis we focus on the formation rocks that corresponding in middle Devonian geological period.

In Southern Alberta and especially in the Devonian formations exist numerous porous and permeable reservoirs, but a few major hydrocarbon traps have revealed and drilling activities are very limited. Previous studies have shown that source rock exist and hydrocarbons have been generated in the Nisku, Winnipegosis, and Exshaw/Lower Banff in Southern Alberta. There is also associated hydrocarbon production in the Lower Banff to Big Valley, in the Nisku and in the Winnipegosis as it is shown in figure 1.2.

The previous work of Dr. Andy Mort of the Geological Survey of Canada (GSC) on source rock evaluation of selected samples from cores in the Beaverhill Lake and Winnipegosis shows the presence of oil prone source rocks within these intervals as referred in [Mort et al 2015]. The evaluation of oils in the Leduc, Nisku, Beaverhill Lake and Winnipegosis reveals an evaporite related source.

Oil is being produced from a dolomite in the Winnipegosis formation in the Rich area immediately to the south of the Big Valey Stettler Leduc platform. The Winnipegosis depositional environment in this area was in the evaporitic interior of a carbonate platform. The dolomite is lied above the salt of the Prairie Evaporite. The oil that had trapped in the reservoir is heavy with gravity of 25 API, and the geochemical analysis indicates an evaporitic algal source, which should be common in a platform interior.

In figure 1.1 the table of lithostratigraphic units for the Devonian subsurface of the Western Canada Sedimentary Basin is presented.

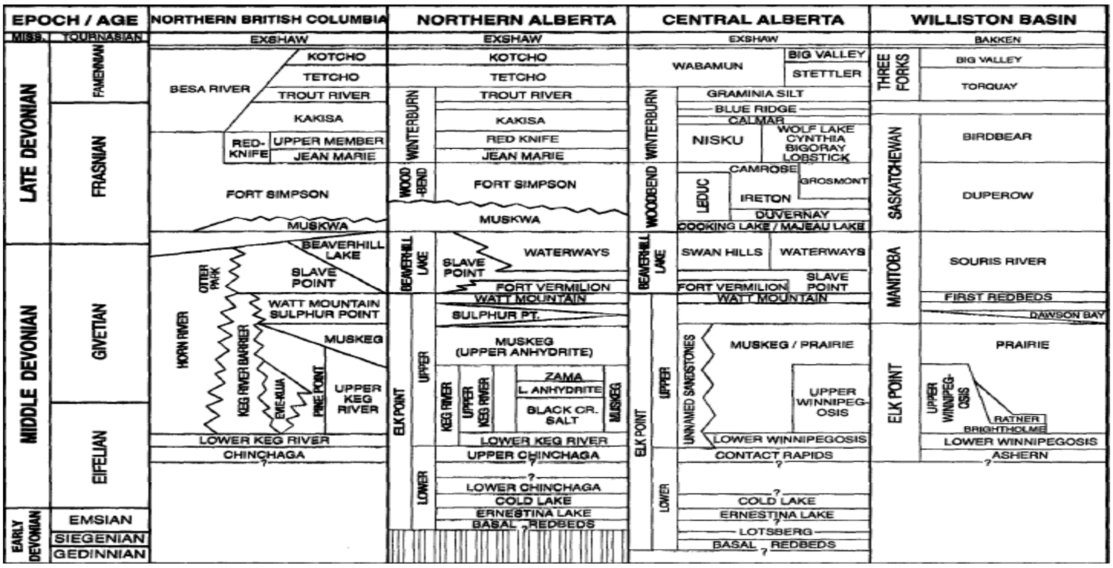


Figure 1.1: Table for lithostratigraphic units for the Devonian formations of the Western Canada Sedimentary Basin (source: Fowler et al 2001).

1.2 Source rock formations in Western Canada

In the following map (Figure 1.2) the distribution of Elk Point Group basinal source rocks and the major depositional facies and paleoenvironments in Western Canada Sedimentary Basin are shown.

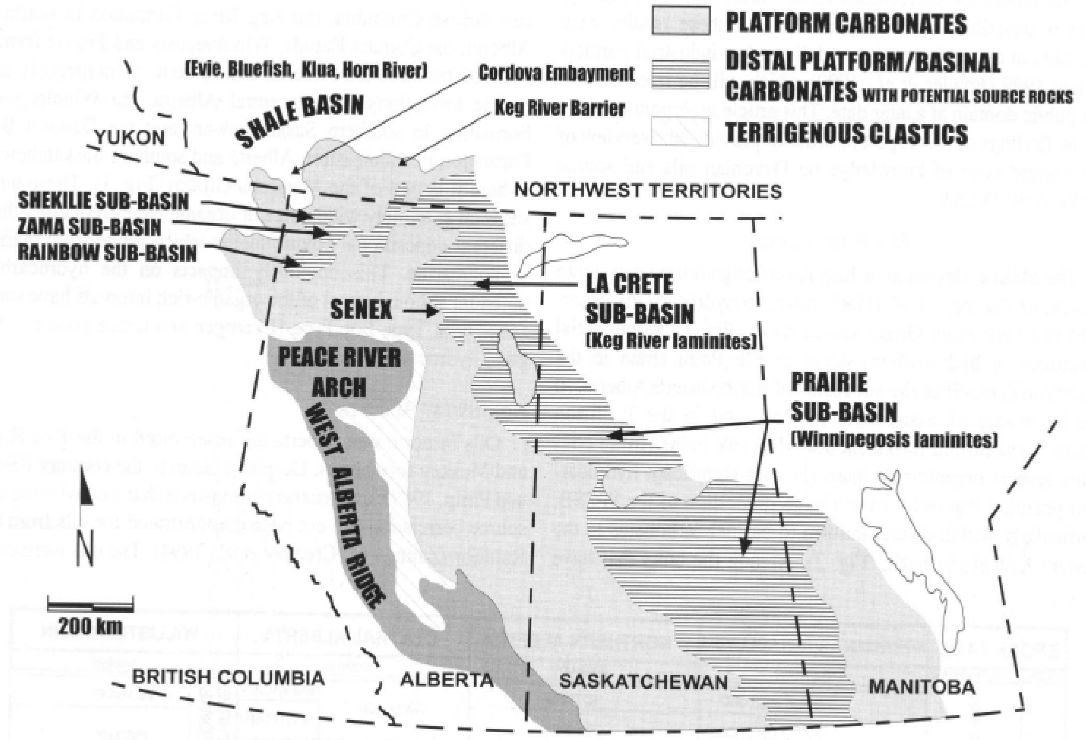


Figure 1.2: The distribution of source rocks and the major depositional facies in Western Canada Sedimentary Basin (source: Fowler et al 2001).

At this point of our analysis, we will briefly describe the major formation systems that existing in the area under study.

Elk Point System

In the Elk Point System clastic dominated shoreline to offshore carbonate platform capped with evaporates exist. There is one major oil pool at the location Rich with Original Oil in Place (OOIP) 2.6 mmbbl and 23 API gravity. Source material is mainly basal algal laminites.

Beaverhill Lake System

In the Southeastern Alberta, the group of Beaverhill Lake was deposited as prograding series of carbonate ramps of the North West carbonate ramps in the Souris River Platform. The ramp complex is time equivalent to the alluvial and backstopping Swan hills platform, which were developed on the West Alberta Ridge to the north and west. The Slave Point equivalent at the base of the system had a salt basin deposited that is surrounded by evaporitic platform interior sediments. At this point, the geochemical data from previous studies indicates also an evaporitic source. There is no production from the Beaverhill Lake although dolomitized reservoirs are present and there are many indications of source rocks.

Leduc System

The depositional pattern of the Leduc formation is the following:
The margins of the carbonate platform are placed in direction to north and west. The margins have well developed dolomite porosity. Mainly peritidal carbonates and evaporates consist the interior of the platform. Vertical seals are present but lateral seal would have to due to structural high adjacent to a fault as mention in [Mort et al 2015].

Furthermore Leduc has oil and gas production and this production of oil and gas reveals, challenge in trapping because there are internal barriers to flow and structural and diagenetic traps.

Nisku System

The Nisku system is the main proven productive formation in Southern Alberta. An early Nisku prograding carbonate platform is present in the South with a biogenic carbonate accumulations barrier and a platform interior evaporitic basin. These evaporitic dolomites have variable thickness due to the occurrence of salt dissolution, which results in structural closures, charged with light oil. In the EnChant area over forty million barrels have been produced from Nisku/Arcs dolomites. Geochemical data indicates that Nisku is most likely charged from self-sourcing evaporate related algal laminates.

Wabamun System

At the area of the Wabamun Stettler the depositional environment consist of tight evaporates and dolomites, deposited in a platform inner of evaporate basin. The overlying Upper Big Valley Wabamun member is a limy limestone, but it become dolomitized in many areas and finally is producing oil and gas from horizontal wells. The probable source is the overlying Exshaw source rork, the Alberta Bakken, made up of Big Valley dolomites, Bakken dolomitic siltstones and, the most productive member, the Lodgepole lower Banff sandstone to siltstone.

The structural history of Southern Alberta is very complex. It was the result of the development of a horst and graben system. The existence of the Sweet Grass arch results in the dip changing from down to the west-southwest to being down to the northwest as referred in [Mort et al 2015]. As a result many of the fault blocks remain open to the south. Conventional traps could also occur due to porosity pinch out, within the carbonate platforms.

In addition there is potential for unconventional regional hydrocarbon traps that were developed in porous, but low permeable dolomites within the evaporitic interiors of the carbonate platforms. The dolomites could be charged from inter-bedded evaporitic algal laminates. In order to develop these reservoirs, it is necessary to use horizontal drilling and completion techniques.

1.3 Summary of previous studies on family affiliation of oils

The main phase of oil generation and migration from Devonian Strata took place during the late Cretaceous – Early Tertiary for the majority of the Western Canada areas. Briefly the source formations have the following characteristics:

Keg River formation: In La Crete Sub-basin

In this sub-basin there are Upper and lower Keg River members with 1-5 meters thickness. The Keg River formation is considered as an excellent potential source rock. The organic matter is regarded as type II to type II-I, with the value of Hydrogen Index (HI) in the range of 500-600. It is assumed that has low level of maturity, with mainly algal bloom organic facies. Keg River kerogen samples behaves as immature to marginally mature with most Tmax values in the range of 420-430 ^{o}C as referred in [Flower et al 2001].

In Upper and Lower Keg River members the available studies reveal different biomarkers characteristics. The Saturated Fraction Gas Chromatogram (SFGCs) of the lower members contains: lower molecular weight n-alkanes with low pristane/phytane ratios. In Upper members: higher amounts of C20+ n-alkanes are present.

Beaverhill Lake

The source rock is mainly carbonate and especially in southern Alberta appears as a good enough hydrocarbon source rock. The organic matter is categorized as type II, with maturity ranges from 0.55% to 0.65% RoVE.

Slave Point Formation

The organic matter of Slave Point formation is kerogen type II/III with values of Hydrogen Index in the range of 170-390. The basic depositional environment is lagoonal. The Nisku oils have higher maturity compared to Leduc oils. The thickness of the Lower Nisku formation is 1-7 meters and there are mainly in basinal platform areas. The depositions were mainly open-marine and classified as type II organic matter with Hydrogen Index (HI) values in the range of 400-600 in East-Central Alberta and these oils are considered as immature. Furthermore, the depositional environment was marine-derived and especially unicellular Prasinophyte alginites. The SFGs are dominated by C15-C21 normal alkanes. The Pristane to Phytane ratios vary between 0.73-1.73. The Nisku formation is more mature in the area of Cynthia Shale Basin with 1.0-1.1 % RoVE.

Carmose formation

For the Carmose formation, previous studies identify four possible source units with the following characteristics:

Unit 1, potential source rock is of type I organic matter with high Hydrogen Index values.

Unit 2, potential source rock is 1-2.3 m in gross thickness and contains type I organic matter.

Unit 3, potential source rock exists in the middle of Nisku formation with 3-4 m thickness and also contains type I organic matter.

Finally the unit 4 potential source rock contains an isolated type I organic matter of terrestrial-influenced organic facies B and C [2]. In addition to, the maturity of organic matter varies from immature in Eastern region to late oil window.

Carmose/Nisku formation:

For the Carmose/Nisku formation members the Saturated Fraction Gas Chromatograms are dominated by lower molecular weight n-alkanes,. The Pristane to Phytane ratio varies between 0.6-1.20 (for the most samples in previous studies the value of this index is less than one (1)). In addition to, similar results in maturity levels found between Carmose Member and Nisku formation.

Birdbear Formation

The Birdbear formation is a very thin (usually a few millimeters) potential source rock, with organic matter of Type I and III. The values of Hydrogen Index are in the range of 138-802 mg HC/g TOC, which is very similar to the Carmose member. The pattern of depositional environment ranges from water lagoonal to tidal flat.

Wabanum group

Wabanum group is reported in very few reports, one of these is [Flower et al 2001]. In the area of British Columbia a Basinal laminate facies exist at the base of Wabanum.

In table 1.1 a brief comparison among the different source rock formations are presented:

Formation	Organic matter	HI value and Tmax	Additional Information
Keg River	Type II to Type II-I	500-600 420-430 °C	Upper and Lower Keg River members. In lower molecular weight n-alkanes with low pristine/phytane ratios. In Upper members: higher amounts of C20+ n-alkanes.
Beaverhill Lake	Type II	high HI values	Source rock is carbonate. Maturity ranges ~0.55% to 0.65% RoVE. Immature
Slave Point	Type II/III	170-390 HI value	Lagoonal type settings
Nisku	Type II	400-600 HI value	SFGs are dominated by C15-C21 n-alkanes. Marine-derived, unicellular Prasinophyte alginites. Relatively low thermal maturity.
Leduc			Immature in comparison with Nisku.
Carmose	Type I	HI high	Units 1, 2, 3 and 4 potential source rocks. Unit 4 is isolated and has terrestrial-influenced organic facies. Maturity of organic matter varies from immature (Eastern region) to late oil window.
Carmose / Nisku			SFGC dominated by lower molecular weight n-alkanes, low amount of C17 compounds.
Birdbear	Type I and III	138-802	Very similar to the Carmose. Few millimeters potential source rock. Organic facies C and E. Water lagoonal to tidal flat paleoenviroments.
Wabaunum			Thermal over-mature difficult to access its original hydrocarbon potential.

Table 1.1: Comparison for the different source rock formations

2. Chemometric Exploratory data analysis

2.1 Dimensionality reduction

The term ***Dimensionality reduction*** is used for methods aiming to find a suitable lower dimensional space in order to represent the original multivariate data.

The exploration of low-dimensional data is easier as the discovery of structure or patterns and the creation and checking of statistical hypotheses is more convenient. Dimension reduction enables the visualization of the data in an appropriate form with the use of the scatter plots, especially if dimensionality of the original data is reduced to 2-D or 3-D as mention in [Martinez 2011].

One common method for dimensionality reduction could be the process of selection subsets of the variables in order to process and analyze them in groups. However, in some cases, with this approach we could eliminate a lot of useful information. An alternative of this first approach could be the creation of new variables that are functions (e.g., linear or nonlinear combinations) of the original variables. Dimensionality reduction methods lead to a mapping from the higher dimensional space to a lower-dimensional one, while we keep all the information of all available variables. In general, this mapping can be linear or nonlinear.

2.2 Principal Component Analysis – PCA

The main purpose and goal of ***principal component analysis*** (PCA) is the attempt to reduce the dimensionality from *p* to *d*, where *d* < *p*, **while** at the same time we try to explain as much as the variation of the original data set as possible. With the PCA technique, we transform the original data to a new set of coordinates or variables. These coordinates or variables can be a linear combination of the original variables. In addition, the observations that are transformed in the new principal component space are uncorrelated. The aim of this method is to achieve better information and understanding of the original data by looking at the observations in the new space.

The PCA methodology can be briefly presented as follows: We start with centered data matrix X_C with dimensions n X p. This matrix contains the observations that are centered about the mean. With other words, the sample mean has been subtracted from each row. Thus we form the sample covariance matrix **S** as

$\boldsymbol{S} = \frac{1}{n-1} \cdot \boldsymbol{X}_C^T \cdot \boldsymbol{X}_C$, where the superscript T denotes the matrix transpose. The jk-th element of the sample covariance matrix S is given by

$$S_{jk} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_{ij} - \overline{x_j}) \cdot (x_{ik} - \overline{x_k}),\ j, k = 1, 2, ..., p,$$

with

$$\overline{x_j} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{ij}.$$

The following step is the calculation of the eigenvalues and eigenvectors of the S matrix, the eigenvalues can be found by solving the following equation for each l_j , $j = 1,2,..., p$

$|S - l \cdot \boldsymbol{I}| = 0$ (Equation 2.1), where I is an identity matrix with dimensions p x p and $|\bullet|$ denotes the determinant. The result of equation 2.1 is to produce a polynomial equation of degree p. The eigenvectors are obtained if we solve the following set of equations for a_j

$$(\boldsymbol{S} - l_j \cdot \boldsymbol{I}) \cdot \boldsymbol{a}_j = 0 \quad j = 1,2,..., p,$$

subject to the condition that all the set of eigenvectors is orthogonal. The previous is equal to that the magnitude of each eigenvector is equal to one, and they are orthogonal to each other:

$$\boldsymbol{a}_i \cdot \boldsymbol{a}_i^{\boldsymbol{T}} = 1$$
$$\boldsymbol{a}_j \cdot \boldsymbol{a}_i^{\boldsymbol{T}} = 0, \ \ for \ i, j = 1,2,..., p \ and \ i \neq j.$$

At this point, we must recall from the matrix algebra the following statement: any square, symmetric and nonsingular matrix could be transformed to a diagonal matrix using the following transformation:

$\boldsymbol{L} = \boldsymbol{A}^{\boldsymbol{T}} \cdot \boldsymbol{S} \cdot \boldsymbol{A}$, where the columns of **A** contain the eigenvectors of matrix **S**, and the **L** matrix is a diagonal matrix which has the eigenvalues along the diagonal.

The final step in the Principal Component Analysis (PCA) is to use the eigenvectors of **S** in order to obtain new variables called principal components (PCs). The Principal Components (PCs) are obtained by solving the following equation:

$z_j = \boldsymbol{a}_j^{\boldsymbol{T}} \cdot (\boldsymbol{x} - \overline{\boldsymbol{x}}) \ \ j = 1,2,..., p,$ (Equation 2.2) where the elements of **a** vector provide the weights or the old variables coefficients in the new PC coordinate space.

We transform the observations of the initial data to the PC coordinate system with the following equation:

$\boldsymbol{Z} = \boldsymbol{X}_{\boldsymbol{C}} \cdot \boldsymbol{A}$ (Equation 2.3)

The principal component scores are in the matrix **Z**. An important characteristic of these PC scores is that they have zero mean and are uncorrelated. We could also use a different transformation of the original observations in X. In this case the PC scores will have mean $\bar{z}$. But we are able to invert this transformation in order to get an expression relating the initial or original variables as a function of the Principal Components (PCs), which is given by the following equation:

$$x = \bar{x} + A \cdot z.$$

To sum up, Principal Components (PCs) are the transformed variables and Principal Components (PCs) scores are the individual transformed data values.

The dimensionality that has the principal component scores in equation (2.3) is also p, so no dimensionality reduction has been done. But from the linear algebra we know that, the sum of the variances of all original variables is equal to the summation of the eigenvalues. On the other hand, the general idea of dimensionality reduction with the technique of PCA is the following. We could include in our analysis only the PCs with the highest eigenvalues, thus we explained the highest amount of variation with the fewest dimensions or PC variables.

Thus we can reduce the dimensionality to d with the following equation:

$$\boldsymbol{Z_d} = \boldsymbol{X_C} \cdot \boldsymbol{A_d}$$

where $\boldsymbol{A_d}$ contains the first d eigenvectors or columns of **A**. $\boldsymbol{Z_d}$ is an n x d matrix because now each observation has only d elements and $\boldsymbol{A_d}$ is a p x d matrix.

To make a conclusion, ***the main purpose of principal component analysis (PCA) is to analyze the data in order to identify patterns that represent the data “well”. The principal components can be seen as new axes of the dataset that maximize the variance along those axes. Theses axes are not anything more, but the eigenvectors of the covariance matrix. In other words, the target of PCA is to find the axes with maximum variances along which the data is most spread.***

2.3 Kernel Principal Component Analysis - KPCA

The Kernel Principal Component Analysis is a powerful technique for extracting structure from data and extends **conventional** principal component analysis (PCA) to a high dimensional feature space. With KPCA we are able to extract up to N, where N is the number of samples, nonlinear principal components without expensive computation.

Furthermore the conventional PCA extracts principal components in the input space, while the extension of KPCA is the extraction of principal components of variables (or features) that are nonlinearly related to the input variables, **the nonlinear principal components**.

The computation procedure of KPCA is the following:

The first step is to project samples from the input space to a high dimensional feature space

$$x \in R^d \to \Phi\ (x) \in R^D,\ D >> d$$

In kernel PCA an arbitrary Φ function is selected. This function is never calculated explicitly, thus we have the possibility to use high-dimensional Φ's because we have never evaluate in that space the data.

The widely used kernels are the linear, polynomial and Gaussian kernel that given by:

$$linear:\ K(\boldsymbol{x_i}, \boldsymbol{x}_j) = \boldsymbol{x}_i \cdot \boldsymbol{x}_j$$

$$polynomial:\ K(\boldsymbol{x_i}, \boldsymbol{x}_j) = (1 + \boldsymbol{x}_i \cdot \boldsymbol{x}_j)^P$$

$$Gaussian:\ K(\boldsymbol{x_i}, \boldsymbol{x}_j) = \exp(-\frac{-|\mathrm{x}_i - x_j|^2}{2 \cdot \sigma^2}$$

We also assume that the data has been centered, thus $\sum_{i=1}^{N} \Phi(x_i) = 0,$ the covariance matrix in $\boldsymbol{R}^D$ is

$\overline{C} = \frac{1}{N} \cdot \sum_{i=1}^{N} \Phi(x_i) \cdot \Phi(x_i)^T$ (Equation 2.4) and the eigenvalue problem becomes

$$\overline{C} \cdot \boldsymbol{V} = \lambda \cdot \boldsymbol{V}$$

All solutions **V** lie in the span of $\Phi\ (x_1), \ldots, \Phi\ (x_N)$,

$V = \sum_{j=1}^{N} a_j \cdot \Phi(x_j)$ (Equation 2.5) and

$(\Phi(x_k) \cdot \overline{C} \cdot \boldsymbol{V} = \lambda \cdot (\Phi(x_k) \cdot \boldsymbol{V})\ \forall k,\ k = 1, \ldots, N.$ (Equation 2.6)

Combining equations (2.4), (2.5) and (2.6), we take

$$\frac{1}{N} \cdot \sum_{j=1}^{N} a_j \cdot (\Phi(x_k) \cdot \sum_{i=1}^{N} \Phi(x_i)) \cdot (\Phi(x_i) \cdot \Phi(x_j)) = \lambda \cdot \sum_{j=1}^{N} a_j \cdot (\Phi(x_k) \cdot \Phi(x_j))\ \ \forall k = 1, \ldots, N.$$

(Equation 2.7)

At this point if we define an N x N matrix K by

$K_{ij} = (\Phi\ (x_i) \cdot \Phi\ (x_j)).$, the 2.7 equation can be written in matrix form

$K^2 \cdot \alpha = N \cdot \lambda \cdot K \cdot \alpha$ (Equation 2.8), where α is a column vector of $\alpha_1, \ldots, \alpha_N$.

Thus we solve the following eigenvalue problem to obtain solution for (2.8)

$$K \cdot \alpha = N \cdot \lambda \cdot \alpha$$

The next step is to let $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_N$ denote the eigenvalues of K, and $\alpha^1,\ldots,\alpha^N$ be the corresponding set of eigenvectors with λ_p be the first nonzero eigenvalue. We can normalize $\alpha^p,\ldots,\alpha^N$ by requiring the following relationship:

$$\boldsymbol{V}^k \cdot \boldsymbol{V}^k = 1,\ \forall k = p,\ldots,N.$$

Finally, we compute the projection onto $V^k \in R^D$ $(k = p,\ldots,N)$. Let x be a test sample with an image $\Phi\ (x) \in R^D$.

$$V^k \cdot \Phi(x) = \sum_{j=1}^{N} \alpha_j^k \cdot (\Phi(x_j) \cdot \Phi(x))$$

2.4 Unsupervised Clustering

Clustering is the technique of organizing a set of data into groups. This organization is based on the fact that observations that are within a group are more similar to each other than the observations belonging to different clusters. We can assume that the data represent features that allow the investigator to distinguish or separate group from the others. A fundamental point in the process is to choose a way for representation of the objects to be clustered.

Many methods are available in order to group or cluster data and many representation schemes could be used. In the literature the clustering is also known as ***unsupervised learning***. In order to understand clustering, we will compare it to the discriminant analysis or supervised learning. In supervised learning the set of observations has a class label associated with it. Thus for data the true and real number of groups is known, as well as the number of members that belongs to every actual group of data. The next step is to use the data with the class labels in order to create a classifier. Therefore a new, unlabeled observation; may be classified using this function.

In contrast, in clustering (or unsupervised learning), we do not have class labels for the observations. Furthermore there is no a priori knowledge about how many groups exist within the dataset.

The basic steps of clustering are the following:

1. Pattern representation: This initial step includes the preparation and initial work in our dataset, such as making a decision of the number of clusters to look for and picking what measurements to use in the analysis. This process is known as ***feature***

selection. After this we must determine how many observations we use for the process, and choose the appropriate scaling or other transformations of the data. This step is known as ***feature extraction***.

2. Pattern proximity measure: The majority of clustering methods require a measure of distance or proximity between observations and between clusters. As it is expected, different measure of distances result in different partitions of the data.

3. Grouping: The definition of the grouping process is the partitioning of the data into clusters. The grouping can be ***hard***, which means that an observation only belongs to a group or not. On the other hand can be ***fuzzy***, in where each data point has a **degree of membership** in each of the clusters. It can also be *hierarchical* in which we have nested sequence of partitions.

4. Data abstraction: This step represents an optional process in order to obtain a simple and compact representation of the partitions. One possible solution for this process could be a description of each cluster in words (e.g., one cluster represents oils, while another corresponds to gases). It can also be a quantitative description such as a representative pattern, e.g., the centroid of the cluster.

5. Cluster Assessment: This process involves the examination whether the data contains any clusters. However, in the majority of cases it means an examination of the algorithm result in order to determine whether or not the clusters possess a physical meaning.

2.4.1 Hierarchical Clustering

The hierarchical method is one of the most common approaches in clustering data. This method is very important in the areas of data mining and gene expression analysis. In ***hierarchical clustering***, the investigator does *not have to know a priory the number of groups* and the data do not need to be divided into a predetermined number of partitions.

The process consists of a sequence of steps, where two groups can either be merged (in the ***agglomerative clustering***) or divided (in the ***divisive clustering***) with the use of some optimum criterion.

In the simplest and most commonly used form, the hierarchical methods have *n* observations in their own group (i.e., *n* total groups) at one end of the process and one group with all *n* data points at the other end. The difference between these two types of clustering is the point of the grouping process.

In agglomerative clustering, we have *n* single clusters and end up with one group in which all points belong to. In divisive methods we take just the opposite output; we start with all observations in one group and keep splitting the initial group until we have *n* single clusters.

The agglomerative clustering requires several selections, such as, how to measure the proximity (distance) between data points and how to define the distance between two clusters. The choice of the type of distance depends mainly on the type of data (continuous, categorical or a mixture of the two). A major role plays the kind of the features the analyst wants to emphasize.

In this selection the main aim of hierarchical methods and clustering algorithms is to find “good” clusters in the data using an appropriate computationally efficient technique. A dataset of n items can be partitioning with the number of ways into g clusters. It is given by the relationship:

$N(n,g)=\frac{1}{g!}\cdot\sum_{k=1}^{g}\binom{g}{k}\cdot(-1)^{g-k}\cdot k^{n}$ where N (n,g) is the number of ways of partitioning a given dataset. Thus if the numbers of g and k are high is not feasible to examine all clustering possibilities for a given dataset as mentioned in [Rencher 2009].

In agglomerative hierarchical approach of clustering at each step, an observation or a cluster of observations is merged into another cluster. With the evolution of this process, the number of clusters shrinks and the clusters themselves become larger. We start with n clusters and end with one single cluster that contains the whole dataset.

At each step, an agglomerative hierarchical procedure combines the two closest clusters; we must take seriously into account how to measure the similarity or dissimilarity of two clusters. There are different approaches to measure the distance between clusters. At this point, it is very important to describe the following different distance metrics, the Euclidean, City Block and Pearson correlation.

1) Euclidean Distance

In a Euclidean n-space the position of a point is a Euclidean vector.

In Cartesian coordinates, if $\boldsymbol{p}=(p_1,p_2,\dots,p_n)\ and\ \boldsymbol{q}=(q_1,q_2,\dots,q_n)$ are two points in an Euclidean n-space, then the definition of distance (d) from p point to q point or vice versa is given by the Pythagorean formula:

$$d(\boldsymbol{p},\boldsymbol{q})=d(\boldsymbol{q},\boldsymbol{p})=\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\dots+(q_n-p_n)^2}=$$

$$=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}$$

2) City Block Distance

The City Block distance between two points, assume a and b, with k dimensions is calculated as: $\sum_{j=1}^{k} | a_j - b_j |$

The City block distance is always greater than or equal to zero. For identical points the value of distance would be zero and it is high for points that have little similarity. With City block distance, the effect of a large difference in a single dimension is weakening because the distances are not squared like as in Euclidean distance. The City block distance very often referred to as Manhattan distance is defined as follows: In the xy-plane, the hypotenuse is the shortest distance between two points, which is the Euclidean distance. But the City block distance is calculated as the distance in x coordinate plus the distance in y coordinate, which is identical to the way that someone is moving in a city where someone must move around the buildings instead of going straight through.

3) Pearson correlation distance

Pearson Correlation measures the similarity between the shapes of two profiles. The mathematic formula for the Pearson Correlation distance is the following: $d = 1 - r$ where

$r = \frac{Z(x) \cdot Z(y)}{n}$, is the dot product of the vectors x and y that are contained the z-scores of these vectors. Z-score of x is calculated by subtracting from x its mean and dividing by its standard deviation.

Continuing, the main methods for linkage of clusters are the following as referred in [Martinez 2011].

1) Single Linkage

Single linkage is maybe the most common used method in agglomerative clustering, and it is the default method in the MATLAB linkage function. Single linkage is also known as ***nearest neighbor***, because the distance between two clusters is given by the **smallest distance between objects**, where each distance is measured from one of the two groups. Thus, we have the following distance between clusters

$d_c(r,s) = \min\{d(x_{ri}, x_{sj})\} \quad i = 1,\ldots,n_r;\ j = 1,\ldots,n_s,$ (Equation 2.9)

where $d(x_{ri}, x_{sj})$ is the distance between observation i from group r and observation j from group s. This is the interpoint distance (e.g. an Euclidean), which is the input to the clustering procedure. In the single linkage method at each step, the distance in equation (2.9) is found for every pair of clusters, and the two clusters are merged. The previous action has the consequence that the number of clusters is reduced by one. After two clusters are merged, the procedure is repeated in the next step: the distances between all pairs of the new formed clusters are calculated again, and the pair that has the minimum distance is merged into a single cluster.

The problem of ***chaining*** is the major drawback of the single linkage clustering. This problem exists when clusters are not well separated, and snake-like chains can form. Thus observations at opposite ends of the chain can be very dissimilar, but yet they end up in the same cluster. Furthermore the single linkage does not take into account the cluster' structure.

2) Complete Linkage

Complete linkage is also known as the furthest neighbor method, because it uses the largest distance between the observations, one in each group, as the distance between the clusters. The distance between clusters is given by

$d_c(r,s) = \max\{d(x_{ri}, x_{sj})\} \quad i = 1,\ldots,n_r;\ j = 1,\ldots,n_s,$ (Equation 2.10)

At each step, the distance in equation (2.10) is found for every pair of clusters, and the two clusters that have the smallest distance are merged.

Complete linkage is not sensitive to the problem of chaining. Additionally, the created clusters have the tendency to be spherical, and for complete linkage are difficult to recover no-spherical groups. The same as single linkage, complete linkage does not account for cluster structure.

3) Average Linkage

The distance in ***average linkage*** method between clusters is the average distance from all observations in one cluster to all of the samples in another cluster. Thus, we have the following distance

$d_c(r,s) = \frac{1}{n_r \cdot n_s} \cdot \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} d(x_{ri}, x_{sj}),$ (Equation 2.11), where the sum is over all x_i in r and all x_j in s. At each step, we join the two clusters with the smallest distance, as measured in equation (2.11).

This method has the tendency to combine clusters that have small variances, and also tends to produce clusters with approximately equal variance. It is relatively robust method and does take the cluster structure into account.

4) Centroid Linkage

In the centroid linkage method, the distance between the two clusters r and s is defined as the Euclidean distance between the mean vectors of the two clusters. Thus:

$d_c(r,s) = d(\overline{x_r}, \overline{x_s}),$ where $\overline{x_r}$ and $\overline{x_s}$ are the mean vectors for the observation vectors in r and the observation vectors in s, and $d(\overline{x_r}, \overline{x_s})$ defined in the below equation:

$$d(\overline{x_r}, \overline{x_s}) = \sqrt{(\overline{x_r} - \overline{x_s})^T \cdot (\overline{x_r} - \overline{x_s})}$$

The definitions of $\overline{x_r}$ and $\overline{x_s}$ are the following:

$$\overline{x_r} = \sum_{i=1}^{n_r} x_i / n_r$$
$$\overline{x_s} = \sum_{i=1}^{n_s} x_i / n_s$$

Further the two clusters with the smallest distance between centroids are merged at each step. When two clusters r and s are joined, the centroid of the new cluster rs is given by the weighted average

$$\overline{x_{rs}} = \frac{n_r \cdot \overline{x_r} + n_s \cdot \overline{x_s}}{n_r + n_s}$$

5) Median linkage

If two clusters r and s are combined using the centroid method, and if r contains a larger number of items than s, then the new centroid $\overline{x_{rs}} = \frac{n_r \cdot \overline{x_r} + n_s \cdot \overline{x_s}}{n_r + n_s}$ with high probability be much closer to $\overline{x_r}$ than to $\overline{x_s}$. In order to avoid weighting the mean vectors according to cluster size, we can use the median or the midpoint of the line that joins r and s as the point for computing new distances to other clusters:

$$m_{rs} = \frac{1}{2} \cdot (\overline{x_r} + \overline{x_s}).$$

The two clusters that has the smallest distance between medians are merged in every step.

A weakness for both centroid and median linkage methods is the possibility of reversals. This is possible to happen if the distance between one pair of cluster centroids is less than the distance between the centroid of the other pair that was merged earlier. In other words, if the distances between clusters are not monotonically increasing, reversals can be created. The consequence of this is that the results making very confusing and difficult to interpret.

6) Ward's Method

In the Ward's method, the merging of two clusters is determined by the size of the incremental sum of squares during the agglomerative hierarchical clustering. It examines the increase in the total within-group sum of squares when clusters *r* and *s* are joined. In the Ward's method the distance between two clusters is given by the relationship:

$d(r,s)=\dfrac{n_r \cdot n_s \cdot d_{rs}^2}{n_r+n_s}$, where d_{rs}^2 is the distance between the r-th and s-th cluster as defined previous in the centroid linkage definition.

In other words, during the procedure of each merging, the within-cluster sum of squares is minimized over all possible partitions that could be obtained, if we are combing two clusters from the current set of groups.

Ward's method has the tendency in combining clusters that have a small number of observations. It also tends to locate clusters that are of the same size and spherical. This method is very sensitive to the presence of outliers in the dataset because it uses the criterion of sum of squares.

2.4.2 Visualizing Hierarchical Clustering Using the Dendrogram

A ***dendrogram*** is a simple tree diagram that shows the structure of the partitions and how the number of groups is linked at each stage. The dendrogram can be drawn horizontally or vertically.

A dendrogram is a mathematical, as well as a visual representation of a hierarchical procedure that, as mentioned in the previous chapter, can be divisive or agglomerative. The **root** is the starting point of the tree, which can be either at the top for a vertical tree or on the left side for a horizontal tree. The clusters are represented with ***nodes*** in the dendrogram, and they can be ***internal*** or ***terminal***. Based on the linkage type and distance metric that used, the internal nodes contain or represent all observations that are grouped. For the majority of dendrograms, the terminal nodes contain a single observation.

The **stem** or **edge** shows "children" of internal nodes and the connection with the clusters below it. The distances at which clusters are joined are represented by the length of each edge. The dendrograms for hierarchical clustering are ***binary trees,*** so they have two edges emanating from each internal node. The arrangement of stems and nodes is referred to as the tree **topology.**

To sum up, the dendrogram illustrates the process for constructing the hierarchy, and the internal nodes are describing particular partitions, if the dendrogram has been cut at a given level.

2.4.3 Optimization Methods – k-Means

Another group of clustering methods uses techniques that optimize some criterion when partitioning the observations into a *specified* or *predetermined* number of groups. These methods are referred to as ***partition*** or ***optimization methods*** differ in the nature of the **objective function**, and in the optimization algorithm used to create the final clustering.

One important issue that must be addressed when implementing these methods (as is also the case with the hierarchical methods) is the determination of the number of clusters in the dataset. On the other hand, one of the major advantages of the optimization-based methods is that they use only the data as input, not the interpoint distances, as in hierarchical methods. Thus, these methods are more suitable and highly recommended for large datasets.

2.4.3.1 k-means

The k-means clustering is one of the most commonly used optimization-based methods. The goal of *k*-means clustering is to partition the data into *k* groups such that the within-group sum-of-squares is minimized. The definition of within-class scatter matrix is given by the following equation

$S_W = \frac{1}{n} \cdot \sum_{j=1}^{g} \sum_{i=1}^{n} I_{ij} \cdot (x_i - \overline{x_j}) \cdot (x_i - \overline{x_j})^T$, where I_{ij} is one (1) if x_i belongs to group j and zero (0) otherwise, and g is the number of groups.

The criterion or objective function that is minimized in k-means is the sum of the diagonal elements of S_W , that it is the definition of the trace of the matrix, as follows

$$Tr\ (S_W) = \sum_i S_{W_{ii}}$$

When the trace is minimized, it is equivalent to minimization of the total within-group sum of squares about the group means. But the minimization of the trace of S_W is equivalent to minimizing the sum of the squared Euclidean distances, which are calculated between each point (individuals) and their group mean.

At this point we briefly describe the procedure for obtaining clusters via *k*-means. The basic algorithm for *k*-means clustering has two major steps in the whole procedure. In the first step, each observation is assigned to its closest group, with the use of the Euclidean distance between the observation and the cluster centroid. The second step of the procedure is to find the new centroids using the assigned observations. These steps are repeated until there are no changes in cluster membership or until the centroids do not change.

The k-means algorithm is implemented as follows:

1. The number of clusters *k must be specified*.

2. Determination of the initial cluster centroids. These can be randomly chosen or the analyst can specify them.

3. Calculation of the distance between each observation and each cluster centroid.

4. Every observation must be assigned to the closest cluster.

5. Calculation of the centroid, (the *d*-dimensional mean) of every cluster using the observations that are assigned to him.

6. Repeat the steps 3 through 5 until the centroid remain constant according to a specified criterion.

By the k-means algorithm empty clusters may be created.

2.4.3.2 Silhouette Plot

A different way of estimating the group number in a dataset is the silhouette statistic. Given the observation i, we define the average dissimilarity to all other points in its own cluster as a_i .

For any other cluster c, we use $d(i,c)$ that represents the average dissimilarity of i observation to all objects in cluster c. Finally, with b_i denote the minimum of these average dissimilarities $\overline{d}(i,c)$.

The silhouette width for i-th observation is given by the following relationship:

$$sw_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

At this point we can find the average silhouette width by averaging the sw_i for all observations:

$$\overline{sw} = \frac{1}{n} \cdot \sum_{i=1}^{n} sw_i$$. (Equation 2.12)

A large silhouette width indicates efficient clustering, but the observations with small values have the tendency to be scattered between clusters. The silhouette width sw_i in equation (2.12) ranges from -1 to 1.

If an observation has a value of silhouette width close to one (1), then this data point is closer to its own cluster than to a neighboring one. If it has a silhouette width close to -1, it means that it is not very well-clustered. A silhouette width close to zero is indicator that the observation could just belong to its current cluster or in one that is near to it.

Furthermore, we can use the average silhouette width in order to estimate the number of clusters in the dataset. This is done, with the use of the partition with two or more clusters that gives the largest average silhouette width. If an average silhouette width is greater than 0.5, is a strong enough indicator for reasonable partition of the data, and a value of less than 0.2 may would indicate that the data do not have cluster structure.

3. Description of the Devonian oils composition data

3.1 Oil Samples

In this work compositional data from a sample set consisting of 146 oils from western Canada were used. These oils are considered to originate from sources located in Devonian formations. In Table 3.1 the geographical location and the reservoir source formation of the oils are shown.

Sample	Lat	Long	Formation	Sample	Lat	Long	Formation
L00794	54.51052	-115.498	Beaverhill Lake	L02080	52.62858	-113.357	Leduc
L00858	54.54708	-116.821	Beaverhill Lake	L02081	52.64824	-113.337	Leduc
L01143	52.0215	-112.769	Nisku	L02082	52.63602	-113.346	Leduc
L01144	51.69111	-111.296	Arcs	L02084	52.51058	-113.188	Nisku
L01277	53.94367	-117.594	Wabamun	L02086	52.47948	-113.182	Nisku
L01350	53.37903	-115.27	Nisku	L02098	51.99341	-114.053	Leduc
L01354	55.64275	-118.161	Wabamun	L02099	52.38876	-113.343	Leduc/Nisku
L01420	54.01195	-116.421	Nisku	L02100	52.32724	-112.876	Leduc
L01453	54.46445	-117.745	Leduc	L02103	52.79822	-113.084	Nisku
L01556	51.66458	-111.273	Arcs	L02106	51.61418	-113.764	Crossfield
L01557	52.11556	-112.99	Nisku	L02108	51.48797	-112.746	Nisku
L01558	52.14638	-112.773	Nisku	L02109	52.25888	-114.588	Leduc/Nisku
L01559	50.96459	-112.019	Arcs	L02110	56.91252	-114.684	Keg River
L01576	52.22727	-113.331	Nisku	L02112	56.31595	-116.087	Slave Point
L01598	50.71718	-111.591	Jefferson	L02151	51.54145	-112.838	Nisku
L01638	53.17386	-115.726	Nisku	L02152	51.54145	-112.838	Leduc
L01639	53.1208	-115.789	Nisku	L02153	51.54145	-112.838	Leduc
L01641	53.04243	-115.974	Nisku	L02154	52.41677	-113.3	Nisku
L01644	53.02228	-116.209	Nisku	L02155	52.39959	-113.301	Nisku
L01645	53.10956	-115.896	Nisku	L02156	52.40738	-113.338	Nisku
L01646	53.11573	-115.698	Nisku	L02157	52.40819	-113.349	Leduc/Nisku
L01647	53.02817	-115.964	Nisku	L02158	52.32835	-113.368	Nisku
L01648	53.14161	-115.782	Nisku	L02159	51.58022	-113.278	Leduc/Nisku
L01650	53.03202	-116.091	Nisku	L02160	52.5109	-113.426	Leduc
L01651	53.18903	-115.74	Nisku	L02161	52.51126	-113.426	Nisku
L01652	53.09129	-115.868	Nisku	L02162	52.27013	-113.346	Nisku
L01654	53.12568	-115.883	Nisku	L02163	52.28353	-113.335	Nisku
L01655	53.15396	-115.87	Nisku	L02164	52.27629	-113.347	Nisku
L01656	53.15396	-115.87	Nisku	L02165	52.26258	-113.33	Nisku
L01658	53.09544	-116.067	Nisku	L02166	52.2702	-113.323	Nisku
L01664	51.50473	-112.673	Arcs	L02167	52.2914	-113.342	Nisku
L01667	51.50136	-112.668	Nisku	L02168	52.07925	-112.746	Nisku
L01676	56.71672	-114.451	Keg River	L02169	52.55221	-113.145	Leduc
L01677	56.73744	-114.383	Keg River	L02170	52.58097	-113.151	Leduc/Nisku
L01679	56.72312	-114.52	Keg River	L02171	52.54484	-113.133	Leduc

L01680	56.75265	-114.537	Keg River	L02177	51.61497	-113.764	Leduc
L01684	56.71129	-114.447	Keg River	L02178	51.62427	-113.791	Leduc
L01686	56.73082	-114.395	Keg River	L02182	52.59544	-113.169	Leduc
L01687	56.73672	-114.46	Keg River	L02183	52.75407	-113.127	Leduc
L01688	56.69365	-114.379	Keg River	L02184	52.8009	-113.086	Nisku
L01690	56.69621	-114.597	Keg River	L02190	52.83786	-113.066	Nisku
L01691	56.84899	-114.767	Keg River	L02191	52.83815	-113.062	Leduc
L01692	56.82601	-114.724	Keg River	L02192	52.33846	-112.849	Leduc
L01693	56.70099	-114.601	Keg River	L02196	51.81394	-113.586	Leduc
L01810	53.21439	-115.656	Nisku	L02197	51.83151	-113.577	Nisku
L01816	51.98092	-112.787	Leduc/Nisku	L02198	51.81826	-113.588	Leduc
L01819	51.99535	-112.788	Leduc/Nisku	L02199	51.99535	-112.788	Nisku
L01820	52.10105	-112.752	Leduc/Nisku	L02200	52.59156	-113.295	Nisku
L01821	52.19548	-112.77	Leduc/Nisku	L02201	52.61111	-113.265	Nisku
L01822	52.2029	-112.776	Leduc	L02202	52.62592	-113.265	Nisku
L01823	51.98398	-112.783	Nisku	L02203	52.64631	-113.253	Nisku
L01824	52.60546	-114.196	Leduc	L02205	52.33768	-112.864	Leduc
L01825	52.19268	-112.772	Leduc	L02206	52.33048	-112.91	Leduc
L01827	52.62012	-114.184	Leduc	L02207	53.52126	-113.723	Nisku
L01828	52.05289	-112.745	Leduc	L02208	53.54302	-113.73	Nisku
L01831	53.18291	-115.637	Nisku	L02209	53.56794	-113.729	Leduc
L01832	52.75428	-114.108	Leduc	L02210	53.56116	-113.724	Leduc
L01833	52.02095	-112.758	Nisku	L02211	53.51028	-113.741	Leduc
L01834	51.99059	-112.761	Leduc	L02212	53.5357	-113.735	Nisku
L02032	53.26003	-113.797	Nisku	L02213	53.60476	-113.704	Leduc
L02034	53.26687	-113.668	Leduc	L02215	53.48118	-113.766	Leduc
L02035	53.27047	-113.687	Leduc	L02219	52.15582	-112.77	Leduc/Nisku
L02038	53.27063	-113.765	Leduc	L02220	52.13719	-112.758	Nisku
L02039	53.27097	-113.753	Leduc	L02221	52.10469	-112.752	Leduc/Nisku
L02040	53.27406	-113.747	Leduc	L02223	52.24668	-112.794	Bearspaw
L02041	53.27409	-113.705	Leduc	L02224	52.28313	-112.835	Leduc
L02042	53.27408	-113.722	Leduc	L02225	52.28664	-112.776	Nisku
L02043	53.82364	-113.481	Nisku	L02226	52.25394	-112.788	Nisku
L02044	53.82008	-113.493	Nisku	L02254	56.81703	-115.453	Granite Wash
L02045	53.81636	-113.475	Nisku	L02255	56.8946	-115.513	Keg River
L02077	52.51099	-113.26	Nisku	L02257	52.4624	-112.167	Camrose
L02078	52.51794	-113.261	Nisku	L02290	54.77219	-116.68	Swan Hills
L02079	52.5206	-113.267	Nisku	L02291	55.10916	-117.661	Leduc

Table 3.1: The geographical location and the source formation of sampled oils

In the map of figure 3.1 the sampled oils are presented:

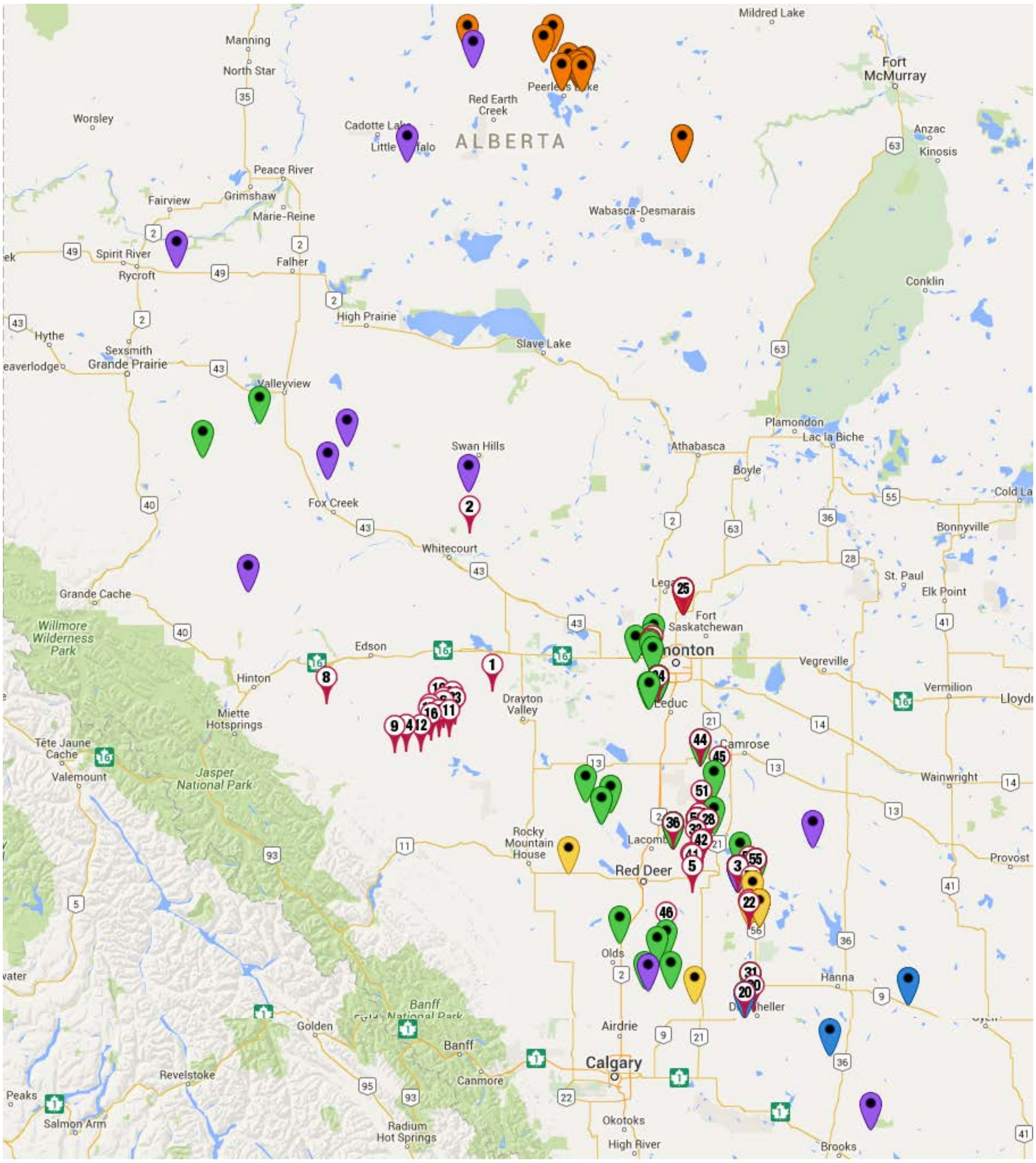


Figure 3.1: The geographical location of oil samples

The following notation of the symbol markers for oil samples is presented:

🟠 Keg River samples

🟡 Nisku_Leduc samples

🟢 Leduc samples

🔵 Arcs samples

# Nisku samples

🟣 Remaining samples

In order to differentiate the oil samples a letter was added to their names as it is shown.

Formation	First letter in sample name
Nisku	N
Leduc	L
Keg River	K
Leduc_Nisku	E
Arcs	A
Remains	G

Table 3.2: The connection of the first letter in sample names with source formation of sampled oils

Thus the KL01677 sample name means that the source formation of this sample is Keg River and in addition is the 1677 sample in our dataset.

3.2 Compositional data

The data was provided by the Geological Survey of Canada. In this work two compositional data sets were employed, namely the composition of the main hydrocarbons of the gasoline range and the composition of n-alkanes in the saturated fraction of the oils.

Petroleum hydrocarbons with number of carbon atoms less than twelve (<C12) are usually referred to as **light hydrocarbons** or **gasoline range** hydrocarbons. They constitute a significant amount of oils. In highly thermal mature oils, these hydrocarbons constitute almost the 100% of the oil composition and therefore geochemical characterization of mature oils is carried our based on these compounds, since they lack intermediate and heavy compounds.

The available identified components in the gasoline range are shown in Table 3.3 together with their abbreviations.

iC5	i-Pentane	MCYC5	Methylcyclopentane	3MC6	3-Methylhexane	24DMC6	2,4-Dimethylhexane
nC5	n-Pentane	24DMC5	2,4-Dimethylpentane	1c3DMCYC5	1,cis-3-Dimethylcyclopentane	223TMC5	2,2,3-Trimethylpentane
22DMC4	2,2-Dimethylbutane	223TMC4	2,2,3-Trimethylbutane	1t3DMCYC5	1,trans-3-Dimethylcyclopentane	234TMC5	2,3,4-Trimethylpentane
CYC5	Cyclopentane	BEN	Benzene	1t2DMCYC5	1-trans-2-Dimethylcyclopentane	TOL	Toluene
23DMC4	2,3-Dimethylbu	33DMC5	3,3-Dimethyl	nC7	n-Heptane	2MC7	2-Methylheptane

	tane		pentane				
2MC5	2-Methylpentane	CYC6	Cyclohexane	MCYC6	Methylcyclohexane	3MC7	3-Methylheptane
3MC5	3-Methylpentane	2MC6	2-Methylhexane	22DMC6	2,2-Dimethylhexane	1c4DMCYC6	1,cis-4-Dimethylcyclohexane
nC6	n-Hexane	23DMC5	2,3-Dimethylpentane	ECYC5	Ethylcyclopentane	nC8	n-Octane
22DMC5	2,2-Dimethylpentane	11DMCYC5	1,1-Dimethylcyclopentane	25DMC6	2,5-Dimethylhexane		

Table 3.3: The commonly identified hydrocarbons in the gasoline range fraction of oil

In figures 3.2, 3.3 and 3.4 characteristic normalized histograms of the chromatographic peak areas from the usual gasoline range components are presented.

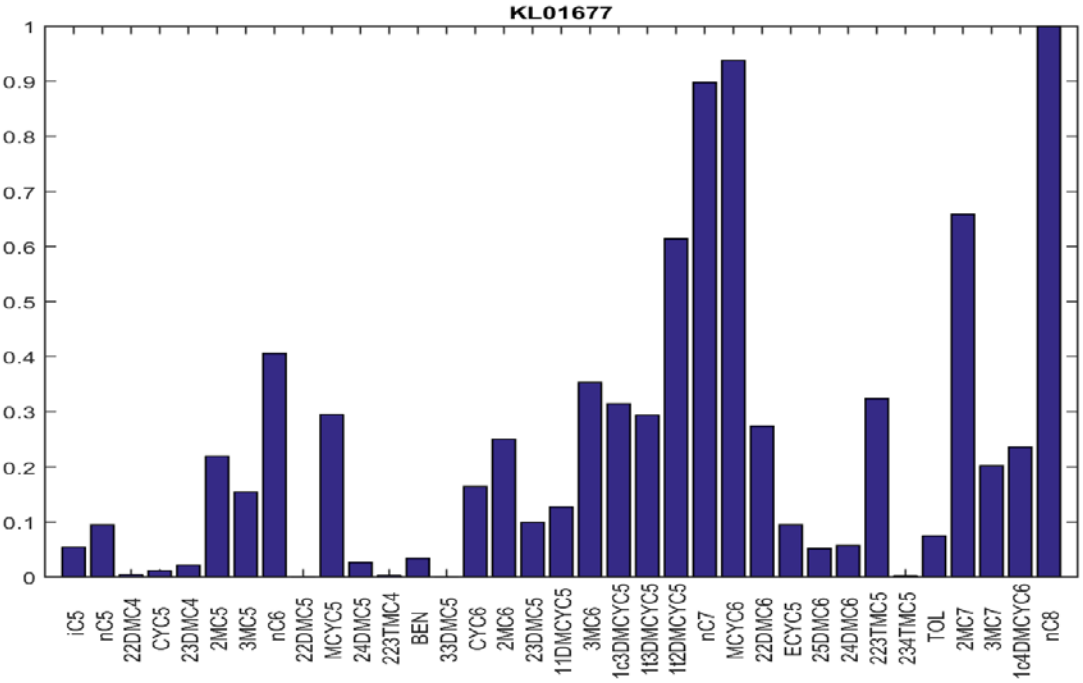


Figure 3.2: Normalized histogram of the chromatographic peak areas from gasoline range of sample L01677

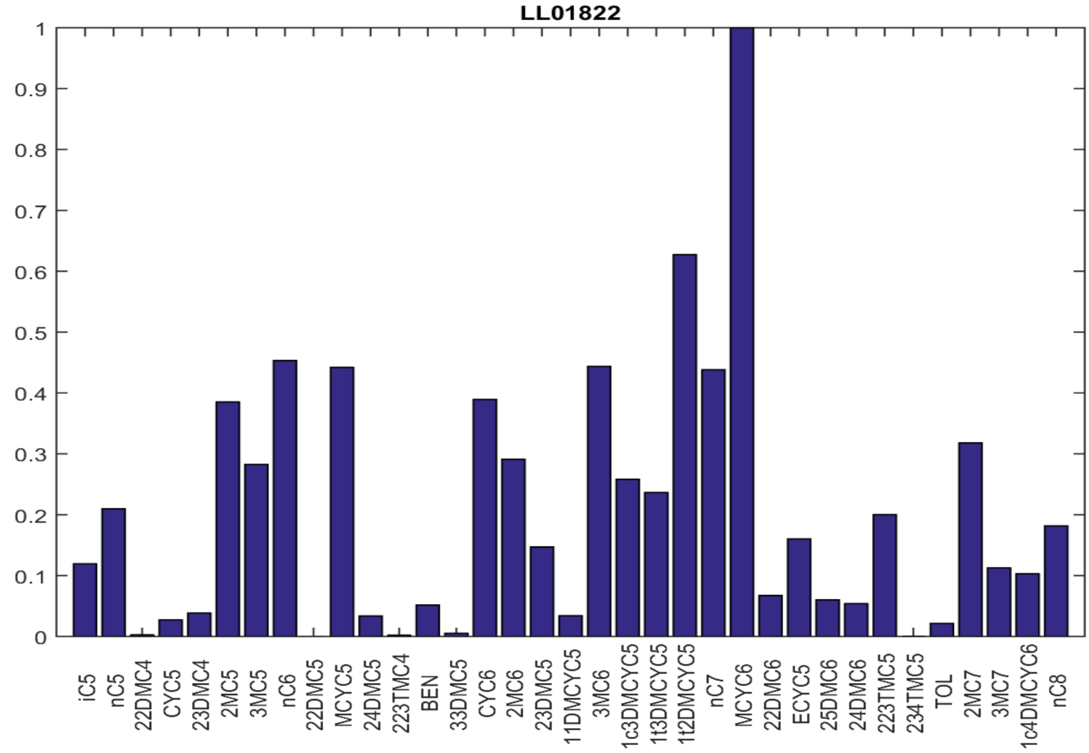


Figure 3.3: Normalized histogram of the chromatographic peak areas from gasoline range of sample L01822

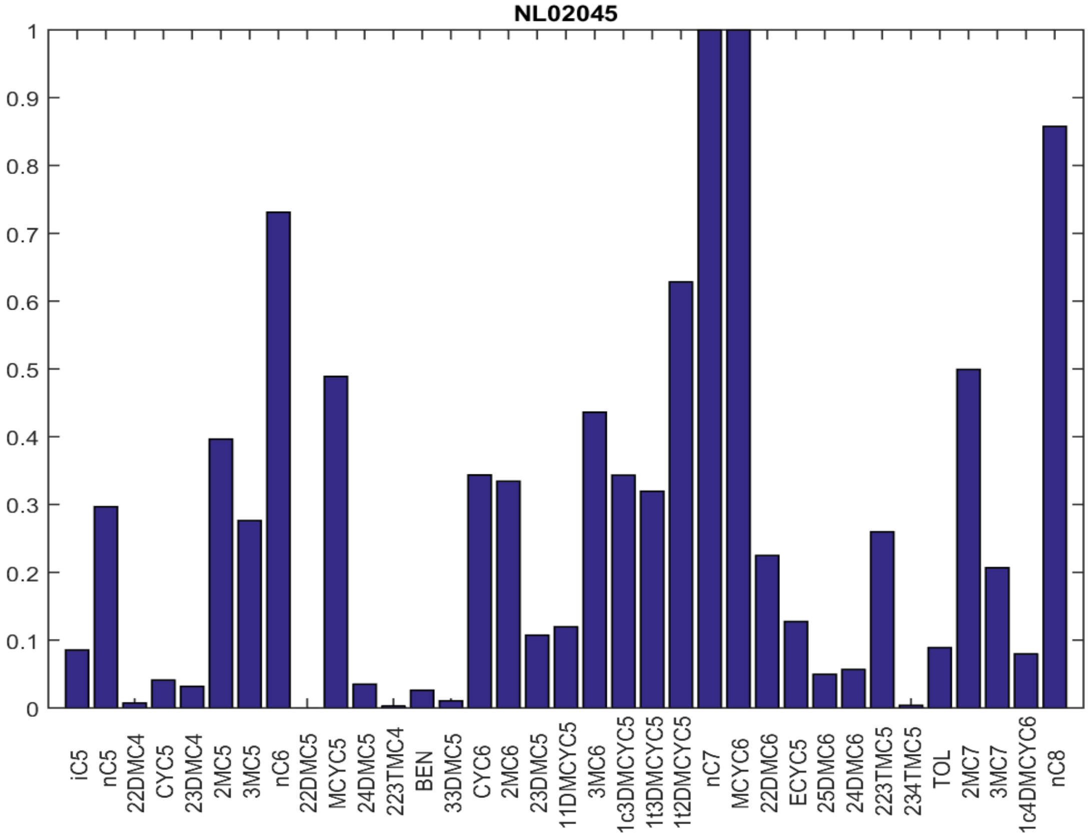


Figure 3.4: Normalized histogram of the chromatographic peak areas from gasoline range of sample L02045

The above three normalized histograms for gasoline ranges have significant differences. In Keg River oil sample the component with the highest peak area is the normal octane and follows the cyclohexane and the normal heptane. In Leduc oil sample the component with the highest peak area is the cyclohexane and follows the 1-trans-2- Dimethylcyclopentane and the normal hexane. Finally in Nisku sample oil the component with the highest peak area is the cyclohexane with the normal heptane and follows the normal octane.

Saturated hydrocarbons are found in petroleum with linear, branched or cyclic structure. The n-alkanes are the simplest structural group. They are formed entirely of single bonds in linear chains between carbon atoms, which are saturated with hydrogen. The n-alkanes have the general formula: C_nH_{2n+2}. The usually measured n-alkanes in the saturated fraction of oils are those with carbon atoms between 12 and 35. These components are shown in Table 3.4 together with their abbreviations used in this text. Two more compounds, pristane (Pr) and phytane (Ph) are included in this list. They are isoprenoids compounds, which are measured in geochemical studies together with n-alkanes, due to their geochemical significance.

C10	n-Decane	C18	n-Octadecane	C26	n-Hexacosane
C11	n-Undecane	Ph	Phytane	C27	n-Heptacosane
C12	n-Dodecane	C19	n-Nonadecane	C28	n-Octacosane
C13	n-Tridecane	C20	n-Eicosane	C29	n-Nonacosane
C14	n-Tetradecane	C21	n-Heneicosane	C30	n-Triacontane
C15	n-Pentadecane	C22	n-Docosane	C31	n-Hentriacontane
C16	n-Hexadecane	C23	n-Tritosane	C32	n-Dotriacontane
C17	n-Heptadecane	C24	n-Tetracosane		
Pr	Pristane	C25	n-Pentacosane		

Table 3.4: The commonly identified hydrocarbons in the saturated fraction of oils

In figures 3.5, 3.6 and 3.7 characteristic normalized histograms of the chromatographic peak areas from the saturated range components are presented:

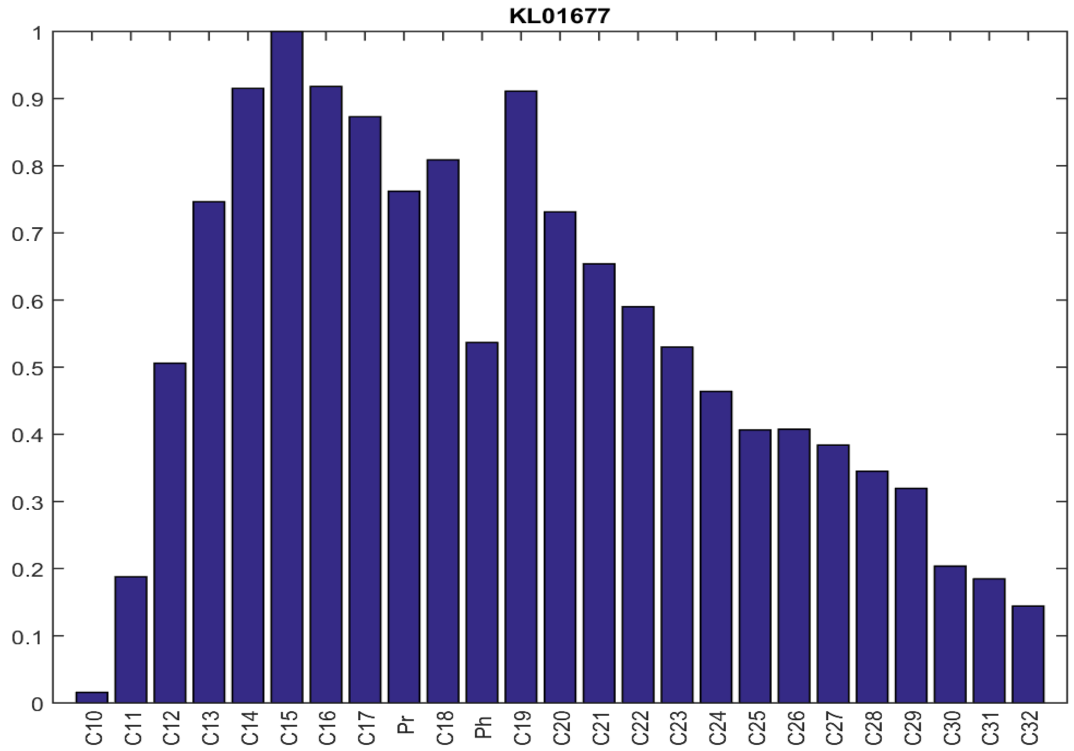


Figure 3.5: Normalized histogram of the chromatographic peak areas from saturated range of sample L01677

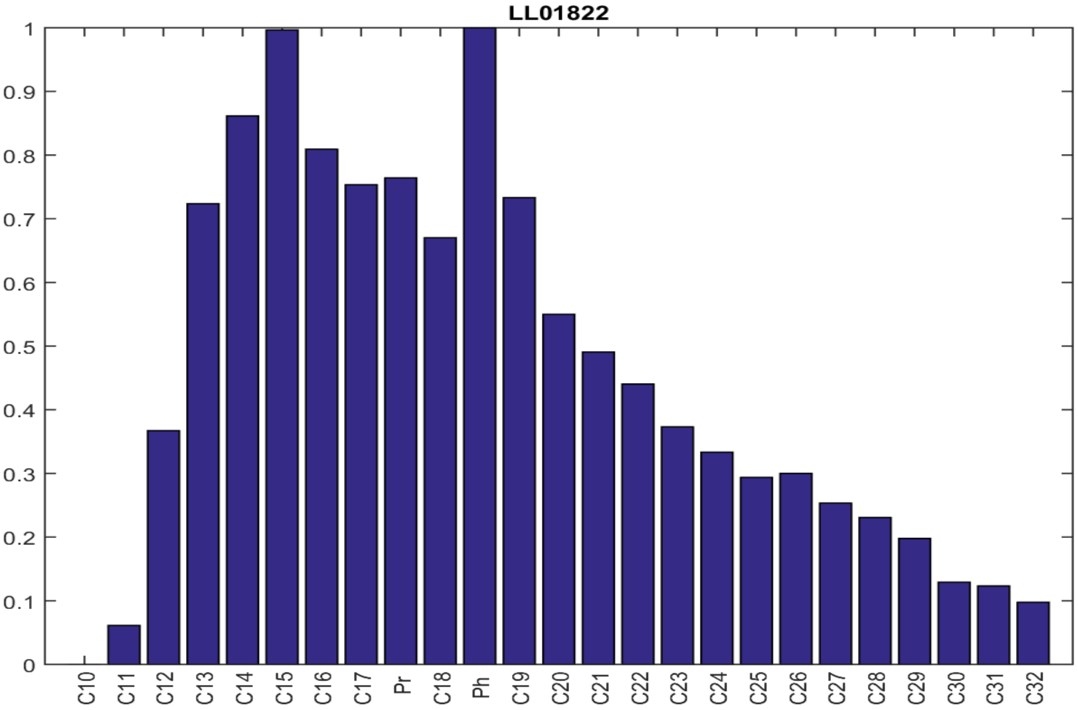


Figure 3.6: Normalized histogram of the chromatographic peak areas from saturated range of sample L01822

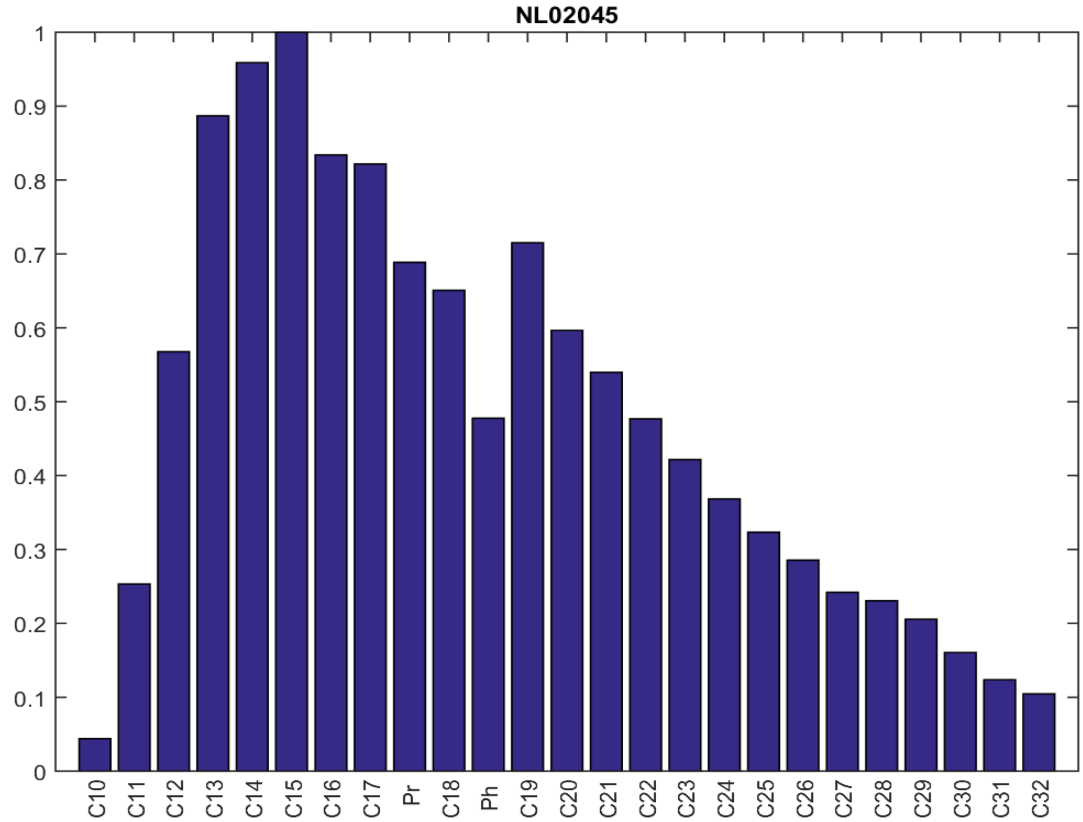


Figure 3.7: Normalized histogram of the chromatographic peak areas from saturated range of sample L02045

The above three normalized histograms of the chromatographic peak areas have the normal-pentadecane as the component with the maximum peak area. Furthermore some variations in others components are existing but in general terms, they carry almost the same geochemical information.

Often in petroleum geochemistry studies instead of pure compositional data, ratios calculated based on the concentrations of selected compounds are used. These ratios, referred to as geochemical indices. They provide significant combined information about the oils. Based on the gasoline and saturated fraction compositional data presented above, a series of such indices were calculated and used in the family affiliation models that described below.

3.3 Gasoline and saturated range geochemical indices

Based on the available compositional data a series of geochemical indices have been introduced. Each one of them carries specific geochemical information [Thompson, 1983], which is briefly summarized below.

For the gasoline range compositional data the most commonly used indices are the following:

K1, A, B, C, I, F, R, U, H

The definitions of these indexes are:

$$K_1 = \frac{2\text{ - }methylheptane + 2{,}3\text{ - }dimethylpentane}{3\text{ - }methylheptane + 2{,}4\text{ - }dimethylpentane}$$

$$A = \frac{benzene}{n\text{ - }hexane}$$

$$B = \frac{toluene}{n\text{ - }heptane}$$

$$C = \frac{n\text{ - }hexane + n\text{ - }heptane}{cyclohexane + methyl\text{ - }cyclohexane}$$

$$I = \frac{2 + 3\text{ - }methylhexanes}{1c3 + 1t3 + 1t2\text{ - }dimethylcyclopentanes}$$

$$F = \frac{n\text{ - }heptane}{methylcyclohexane}$$

$$R = \frac{n\text{ - }heptane}{2\text{ - }methylcyclohexane}$$

$$U = \frac{cyclohexane}{methylcyclohexane}$$

$$H = \frac{100 \times n\text{ - }heptane}{\sum(cyclohexanes + C7HCs)}$$

The K1 index is for confirmation if exists a common creation mechanism of light hydrocarbons from the heavier ones. The following eight indexes are known as Thompson indexes. To be more specific, the A and B indexes indicated the aromaticity property. The C, I and F indexes are indicators for the paraffinicity of each sample. The U index indicated the extent of napthene branching and the R index is an indicator for paraffin branching property.

For the saturated range compositional data the most commonly used indices are the following:

Pr/Ph, Pr/nC17, Ph/nC18, CPI25-33, nC24+/nC24-, nC19/nC31, R22

The definitions of these indexes are:

$$CPI_{25\_33} = [\frac{(C25+C27+C29+C31)}{(C24+C26+C28+C30)} + \frac{(C25+C27+C29+C31)}{(C26+C28+C30+C32)}] \cdot \frac{1}{2}$$

$$nC24+ \big/ nC24- = \frac{C25+C26+C27+C28+C29+C30+C31+C32}{C17+C18+C19+C20+C21+C22+C23+C24}$$

$$R22 = \frac{2 \cdot C22}{C21+C23}$$

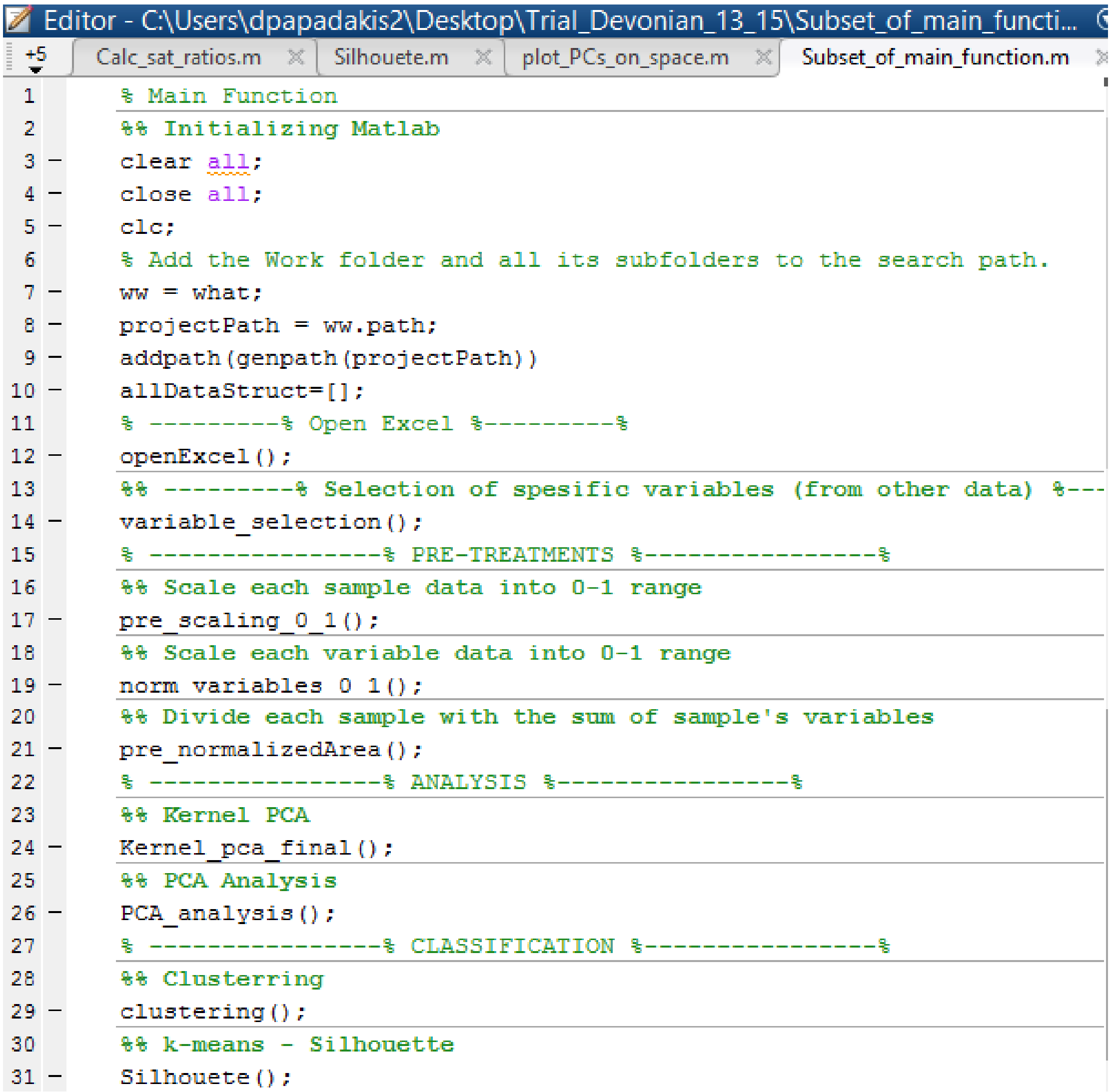
The pristane/phytane ratio is used as an indicator of how much oxidized is an environment. The ratio pristane/nC17 is very useful for differentiated organic matter from swamp environment from those that formed under marine environment. The phytane/nC18 index refers to marine organic input. The CPI_25_33 index is a Carbon Preference Index which is defined as the ratio of sum of concentration areas of odd to even carbon number of n-alkanes. This index is specific for maturity, but also affected by other processes such as biodegradation. The index nC24+/nC24- alkanes is the ratio of heavy hydrocarbons above nC24 to the light hydrocarbons below nC24. Carbon number of C31 is used as an indication of terrestrial biogenic hydrocarbon while C19 presents the marine biogenic sources. Thus the ratio of C31/C19 is used to identify the predominance of hydrocarbon input from land or marine environments.

4. Chemometric models for families' affiliation of Devonian oils

In this chapter we use MATLAB code that is created in the Hydrocarbons Chemistry and Technology Research lab. The mentioned MATLAB code is used with the necessary additional ads and adjustments in order to taking the functionality and the results in an efficient way.

A brief presentation of the graphical environment of the chemometric software package is shown below. In this example chromatographic data from the gasoline range or saturated fraction compound of the analyzed oils are used.

The interface of the chemometric software package is the following:



Editor - C:\Users\dpapadakis2\Desktop\Trial_Devonian_13_15\Subset_of_main_functi...
+5
Calc_sat_ratios.m
Silhouete.m
plot_PCs_on_space.m
Subset_of_main_function.m

```
% Main Function
%% Initializing Matlab
clear all;
close all;
clc;
% Add the Work folder and all its subfolders to the search path.
ww = what;
projectPath = ww.path;
addpath(genpath(projectPath))
allDataStruct=[];
% ---------% Open Excel %---------%
openExcel();
%% ---------% Selection of spesific variables (from other data) %---
variable_selection();
% ----------------% PRE-TREATMENTS %----------------%
%% Scale each sample data into 0-1 range
pre_scaling_0_1();
%% Scale each variable data into 0-1 range
norm_variables_0_1();
%% Divide each sample with the sum of sample's variables
pre_normalizedArea();
% ----------------% ANALYSIS %----------------%
%% Kernel PCA
Kernel_pca_final();
%% PCA Analysis
PCA_analysis();
% ----------------% CLASSIFICATION %----------------%
%% Clusterring
clustering();
%% k-means - Silhouette
Silhouete();
```

Figure 4.1: The interface of the chemometric software package

To be more specific, in figure 4.1 the first section with title "Main Function" is for the initialization of the MATLAB environment and for importing the dataset from the appropriate Excel file. The second choice with the title "Selection of specific variables (from other data)", gives to us the opportunity to exclude original variables from the loaded dataset.

Furthermore, with the three following choices in the section of Pre-Treatments, we have the ability to enforce some pretreatments in the dataset. The pre_scaling_0_1 () scales each sample data in the range of 0 to 1. The norm_variables_0_1 () scale each variable data in the range of 0 to 1 and finally the pre_normalizedArea () divide the component area of each sample with the area summation of all variables for this sample, thus it gives the % percentage of each variable in the sample.

Continuing in the analysis section we have two options, as first to run kernel Principal Component Analysis (KPCA) and secondly to run the conventional Principal Component Analysis (PCA). Finally, in classification section we have the ability to run hierarchical clustering with the choice of clustering () and k-means with the silhouette plots if we select the choice of Silhouete (). The MATLAB code for kernel principal component analysis is based on dimensionality reduction toolbox of Laurens van der Maaten from Tilburg University.

In MATLAB environment and especially in Editor Menu if someone push the button Run section has the ability to run each possible choice separately as it is shown in figure 4.2.

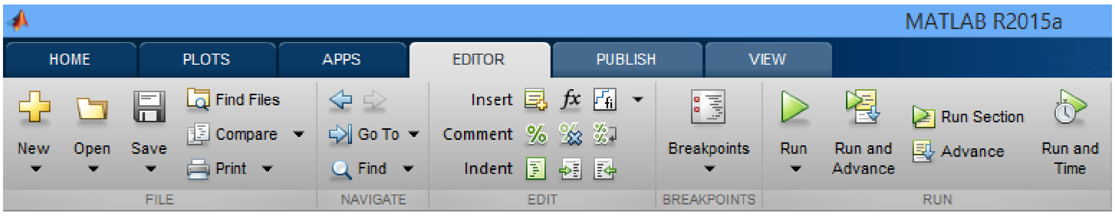


Figure 4.2: The Editor menu in MATLAB program

4.1 Model 1 Gasoline range with all compositional variables

The process of running the model 1 is the following:

First of all, you must load the values of chromatographic peak areas for the gasoline range from the Excel file. The name of Excel file that used in order to load the appropriate data in the MATLAB environment is All_Devonian_data.xlsx.

As we mention in page 43 we need to initialize the MATLAB program with the gasoline range chemical data.

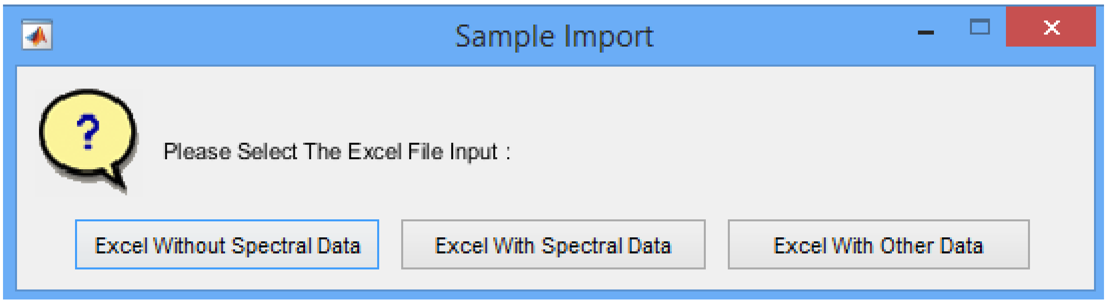
We have the dialog menu that is presented in figure 5, with three possible options for samples import:

1) Excel without spectral data

2) Excel with spectral data

3) Excel with other data

For the needs of this diploma thesis, we select in Sample Import menu the option three which is: Excel with Other Data. This selection has the consequence of importing the necessary data from Excel file in the MATLAB environment.



Please Select The Excel File Input :

Figure 4.3: The dialog menu with the available options of samples' import

A further step is to decide how many samples someone wants to import in the MATLAB program. The default choice is loading all samples that are presented in the first sheet of the Excel file that used as input. But there is also the option to load the samples manually by picking them. The sample loading are become with the press of Import Sample button as illustrated in figure 4.4.

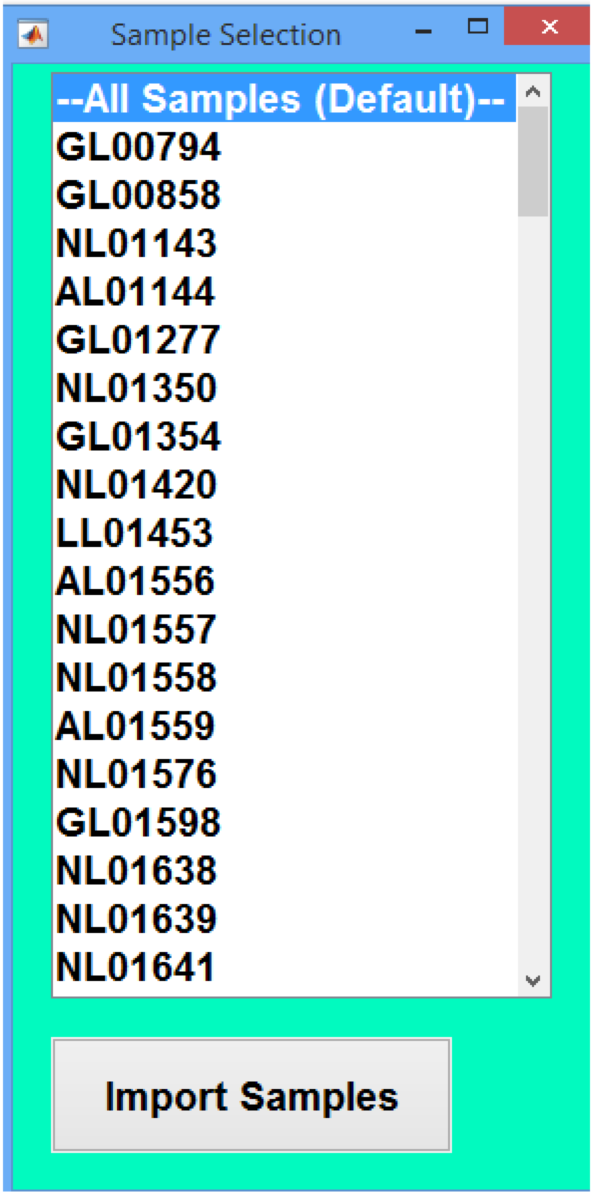
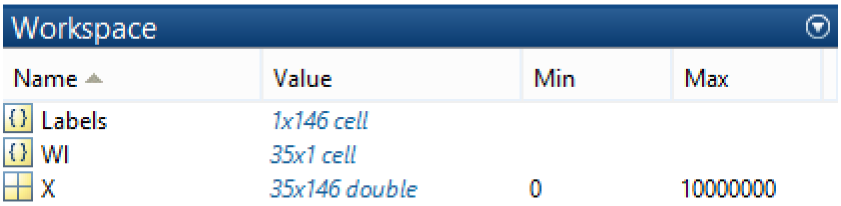


Figure 4.4: The dialog menu for sample selection

Now in MATLAB environment and especially in the workspace section we have created the following variables: Labels cell array that contains the sample names. The Wl that is a cell array containing the variable names and finally the dataset X is a 35 X 146 matrix that has the values of chromatographic peak areas for the gasoline range for each variable and sample as illustrated in figure 4.5.



Workspace

Name	Value	Min	Max
Labels	1x146 cell		
Wl	35x1 cell		
X	35x146 double	0	10000000

Figure 4.5: The workspace section with the variables for model 1

The next step is with the use of the available option pre_scaling_0_1 () to normalize the 146 samples in the range of 0 to 1. The result of this pretreatment is that we take for chromatographic peak area values between 0 and 1 for all samples.

The final step in pretreatments is the normalization of the 35 variables that contains the gasoline range with the MATLAB function norm_variables_0_1 () in the range of 0 to 1 for all samples.

4.1.1 PCA analysis for model 1

In order to perform Principal Component Analysis, we run the option PCA_analysis () in the section Analysis at the interface of the chemometric program and we take the following results.

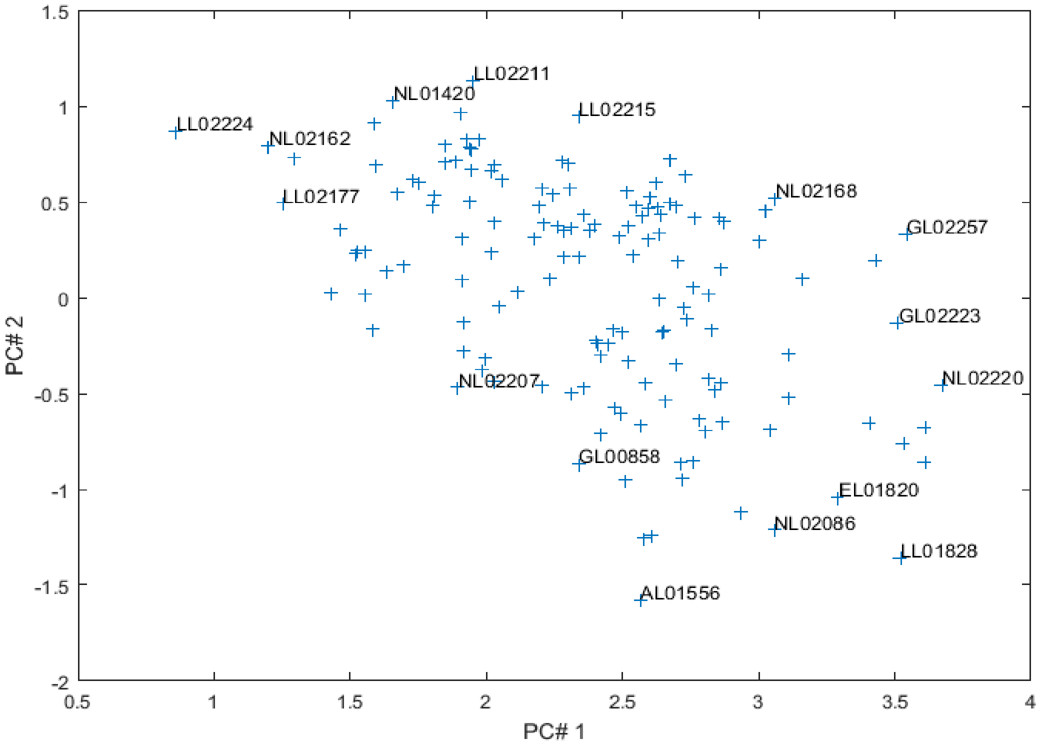


Figure 4.6: Plot of the two major principal components for model 1

In figure 4.6 the plot of the first two major principal components is shown. We do not see a clear separation for samples in distinguish clusters because the behavior of the majority of samples is the almost the same. But samples LL02224, NL02162 and LL02177 have significant differences with the samples EL01820, NL02086, LL01828 and AL01556 according to figure 4.6.

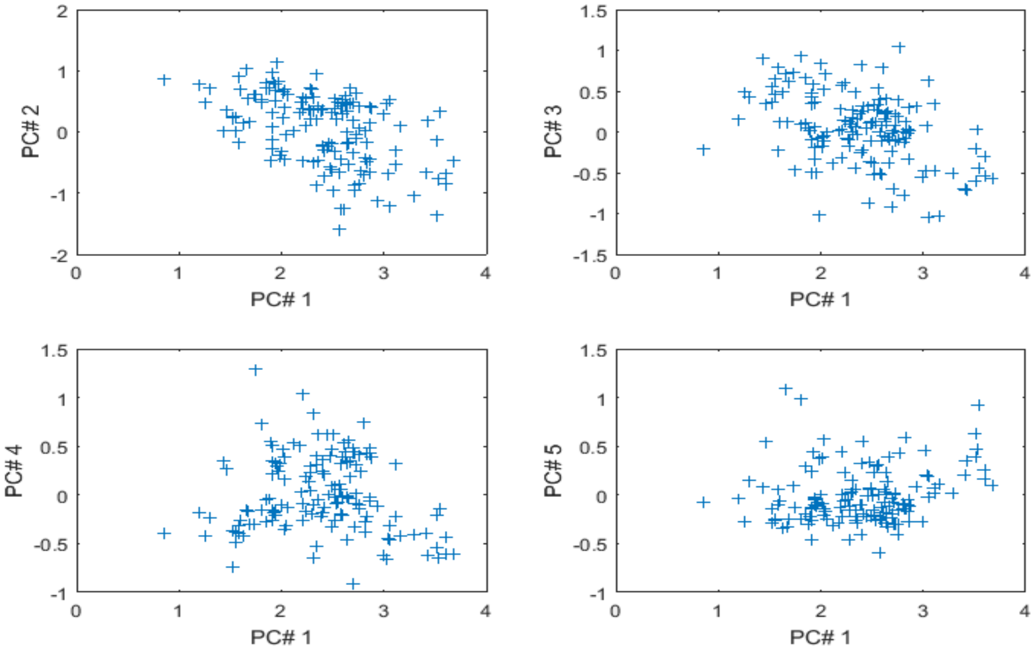


Figure 4.7: The subplot of five principal components of model 1

The figure 4.7 illustrates in four subplots, the first principal component versus the second, third, fourth and fifth principal component for all samples in our data.

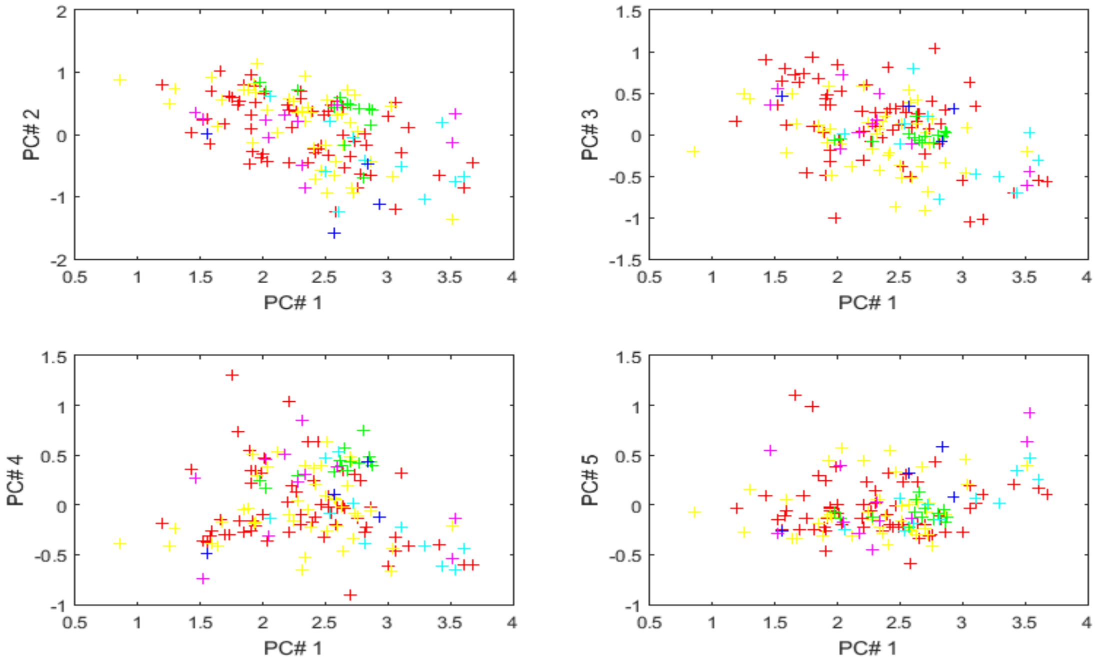


Figure 4.8: The subplot for five principal components with different color for each formation of model 1

The figure 4.8 illustrates in four subplots, the first principal component versus the second, third, fourth and fifth principal component for all our data but in this figure each color depict different formations. The mapping between reservoir rock formation and color is shown in the Table 4.1.

Formation	Color
Nisku	Red
Leduc	Yellow
Keg River	Green
Leduc_Nisku	Blue
Arcs	Dense Blue
Remains	Pink

Table 4.1: Color representation of oils with respect to their reservoir formation origin.

The plot of the first principal component in the space of longitude and latitude coordinates of our samples is presented in figure 4.9.

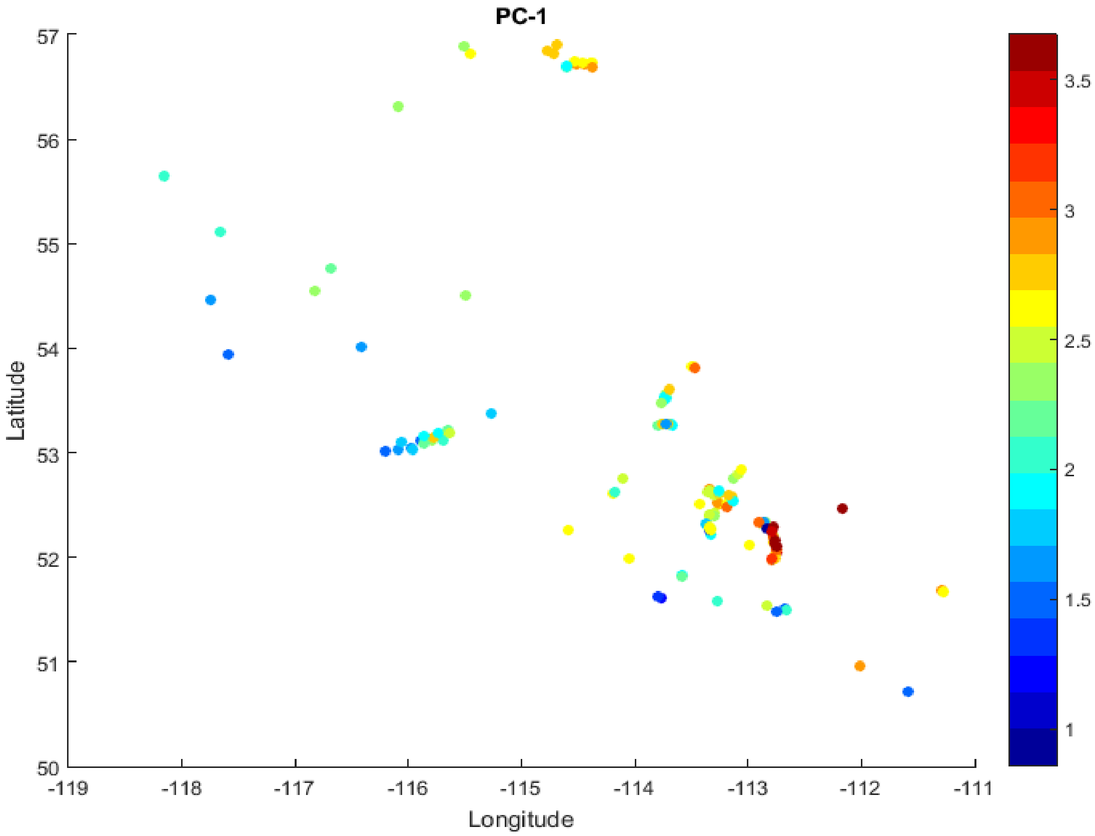


Figure 4.9: Plot of the first principal component of PCA analysis vs the geographical location of the samples.

Continuing, figure 4.10 presents the plot of the second principal component according to the latitude and longitude coordinates of our samples.

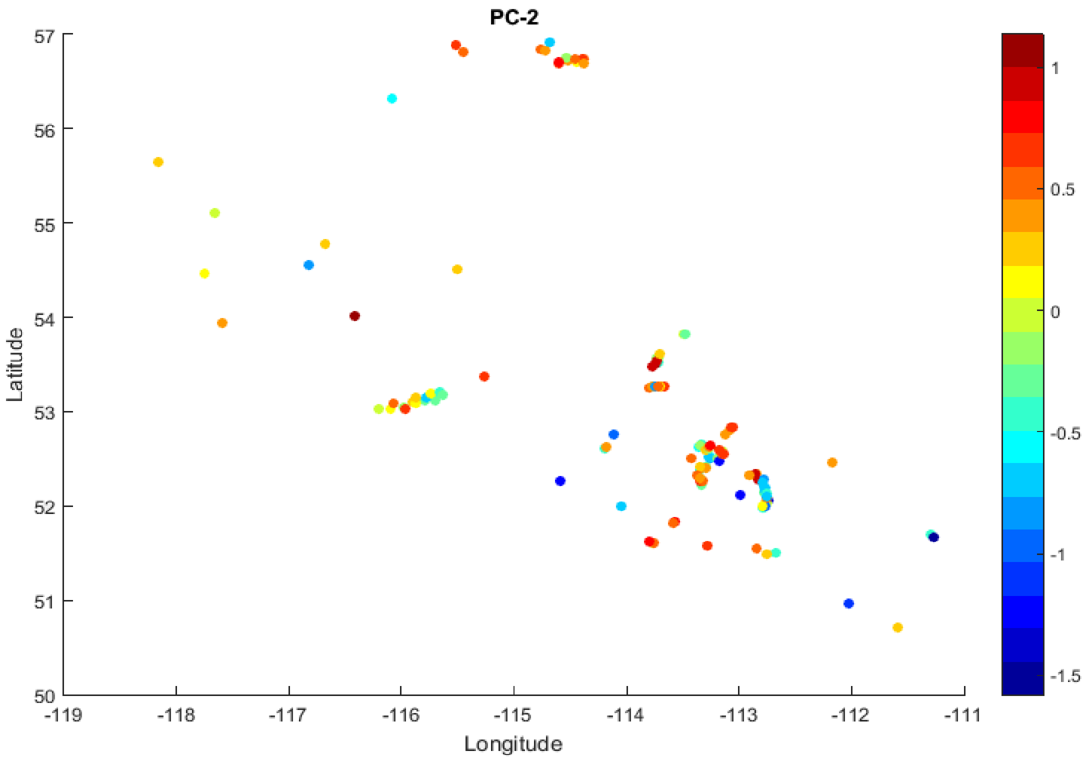


Figure 4.10: Plot of the second principal component of PCA analysis vs the geographical location of the samples.

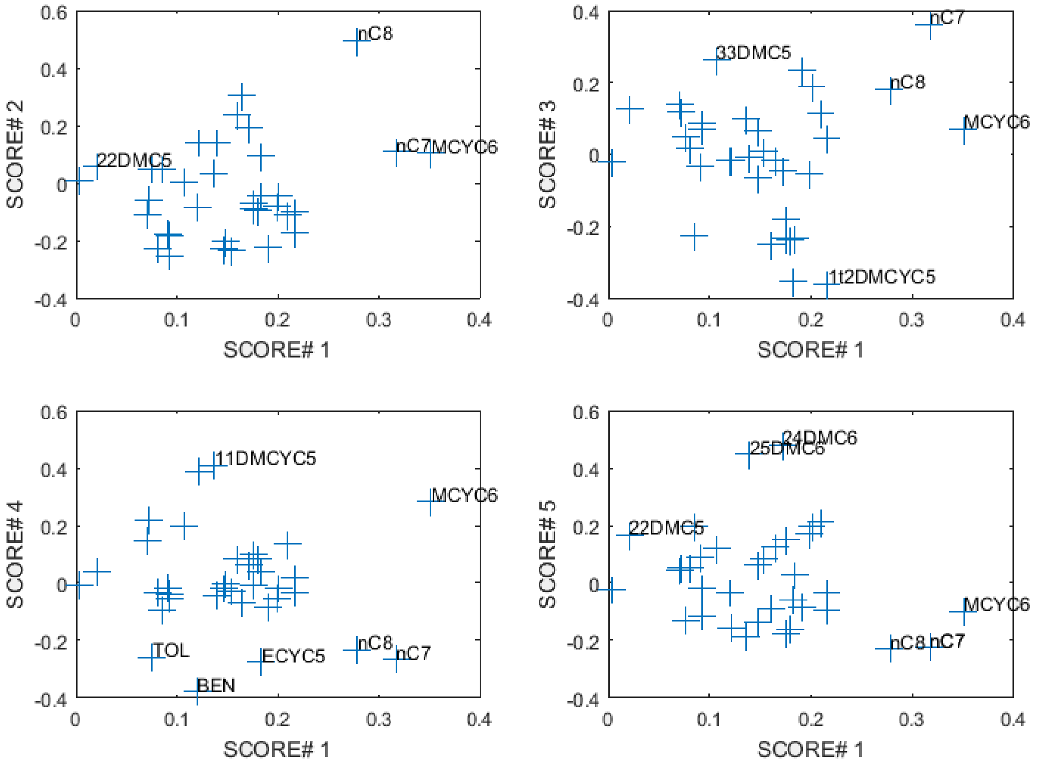


Figure 4.11: Original variable loadings for the first five principal components for the model 1.

The original variable loadings are presented in figure 4.10. In this subplot we have four instances that depict the original variable loadings for the first principal component versus the original variable loadings for the second, third, fourth and fifth principal component.

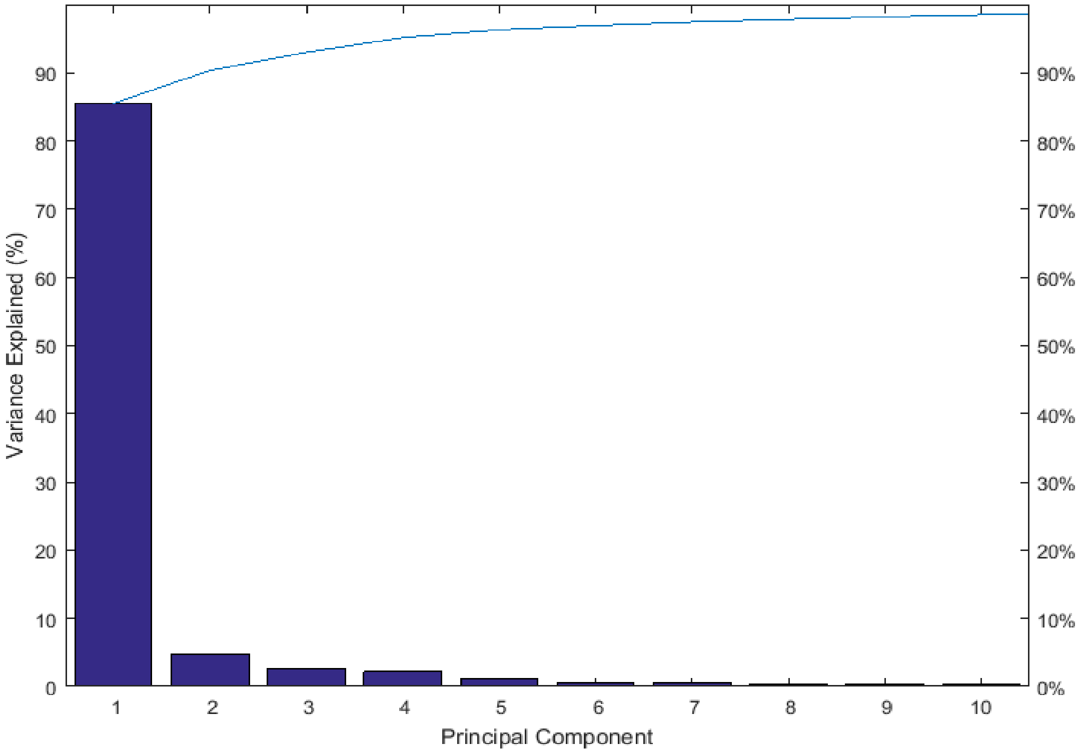


Figure 4.12: Percentage of variance explained by each principal component

Figure 4.12 reveals the contribution of each principal component in the total variance that is explained from the PCA model. As it is shown the first PC explains 85% of variation and the second PC explains 6% of the remaining variation.

4.1.2 Kernel PCA for model 1

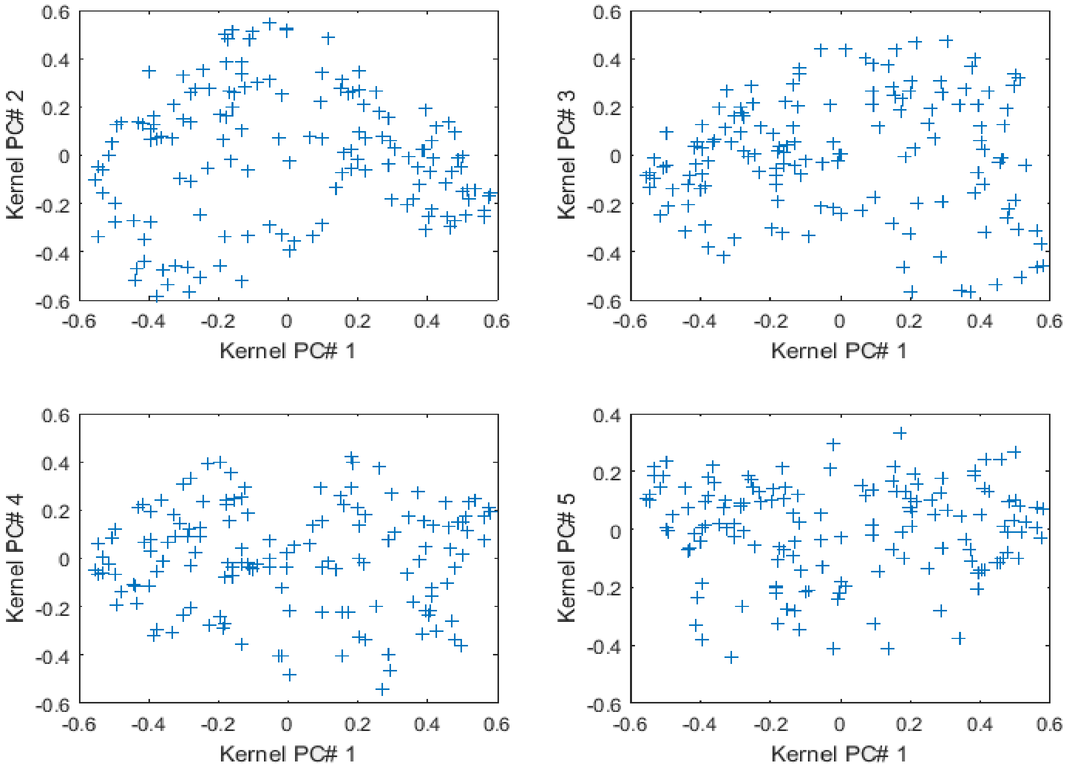


Figure 4.13: The subplot of five kernel principal components of model 1

The figure 4.13 illustrates in four instances, the first kernel principal component versus the second, third, fourth and fifth kernel principal component for all our data.

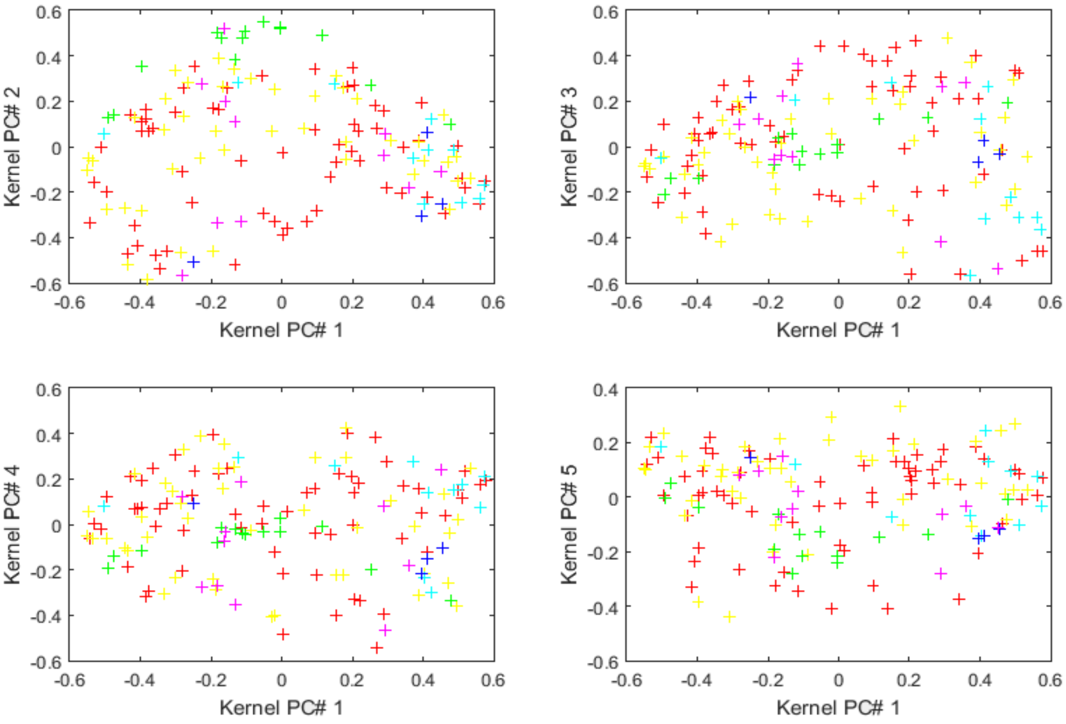


Figure 4.14: The subplot of kernel principal components with different color for each formation of model 1

The figure 4.14 illustrates in four instances, the first kernel principal component versus the second, third, fourth and fifth kernel principal component for all our data. In addition to, in this figure each color depict different formations (For mapping see table 4.1).

The plot of the values of first kernel principal component in the space of longitude and latitude coordinates of our samples are presented in figure 4.15.

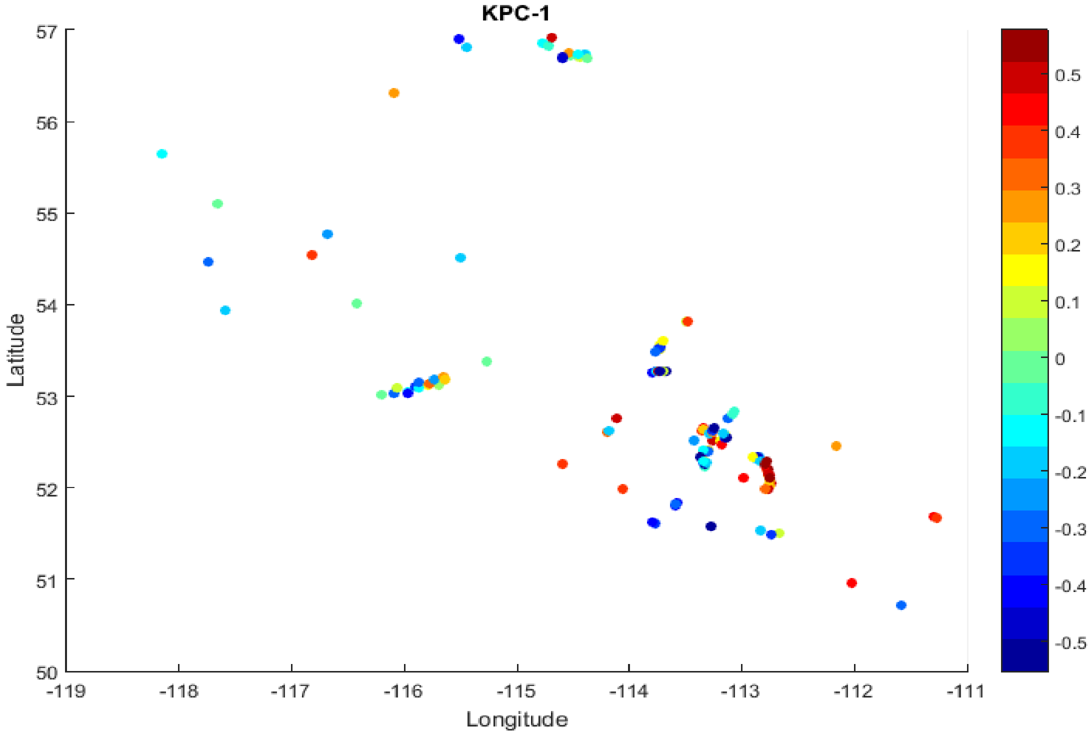


Figure 4.15: Plot of the first kernel principal component of KPCA analysis in relationship vs the geographic location of the samples.

Continuing, figure 4.16 illustrates the plot of the second kernel principal component according to the latitude and longitude coordinates of our samples.

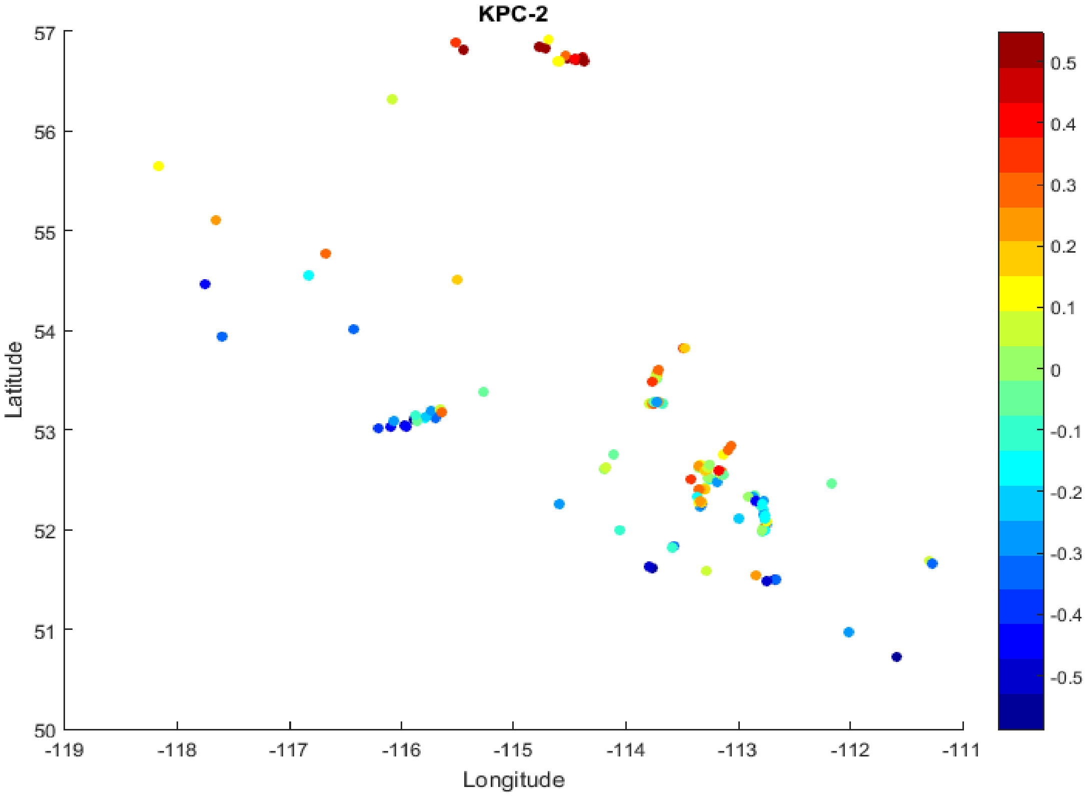


Figure 4.16: Plot of the second kernel principal component of KPCA analysis in relationship vs the geographical location of the samples.

4.1.3 Silhouette with k-means for model 1

With the use of the available option Silhouete () at the interface of the chemometric program in the section of clustering, we get a k-means clustering using two, three, four and five clusters. The k-means clustering is repeating five times for each case. This is achieved with the use of replicates as an argument in kmeans MATLAB function.

In table 4.2 we summarize the results that are taken:

k-means clustering	Best Total sum of distances	Average silhouette value
K=2	140.814	**0.3864**
K=3	122.207	0.3680
K=4	105.674	0.3075
K=5	99.5886	0.2677

Table 4.2: Summary of k-means clustering for model 1

The silhouette plots for K=2, K=3, K=4 and K=5 clusters are shown in figure 4.16.

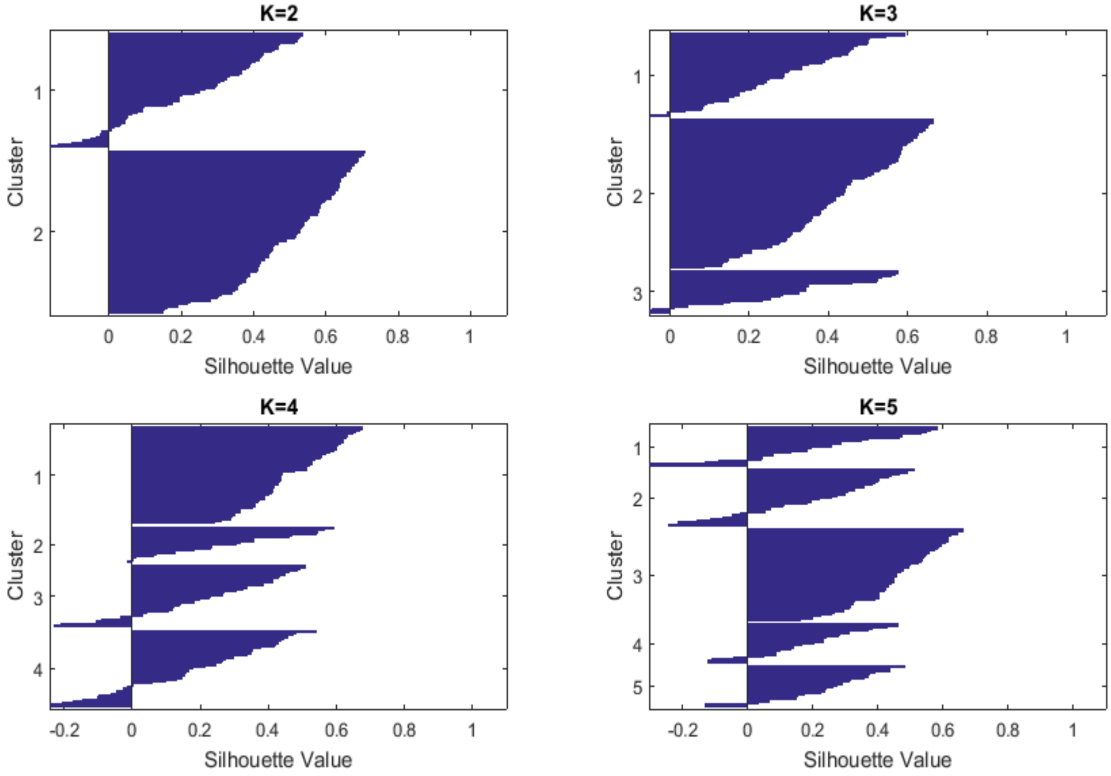


Figure 4.17: Silhouette plots for k=2, k=3, k=4 and k=5 clusters for model 1

We see that in the case of two clusters we have the mostly large silhouette values and very few negative values in cluster one. A one-number summary in order to describe the performance of each clustering is the average of the silhouette values. The two cluster solution has an average silhouette value of 0.3864 and this value is the maximum among the others cases. Thus it is an indicator that the grouping into two clusters using k-means is better than the one with three or four or five groups.

In figure 4.18 the plot of the first two Principal Components (PCs) that were taken from running the option PCA_analysis in the section Analysis at the interface of the chemometric program combined with k-means clustering, for the case of k =2 for our dataset of 146 oils are presented with different color for samples members that belongs to a different cluster:

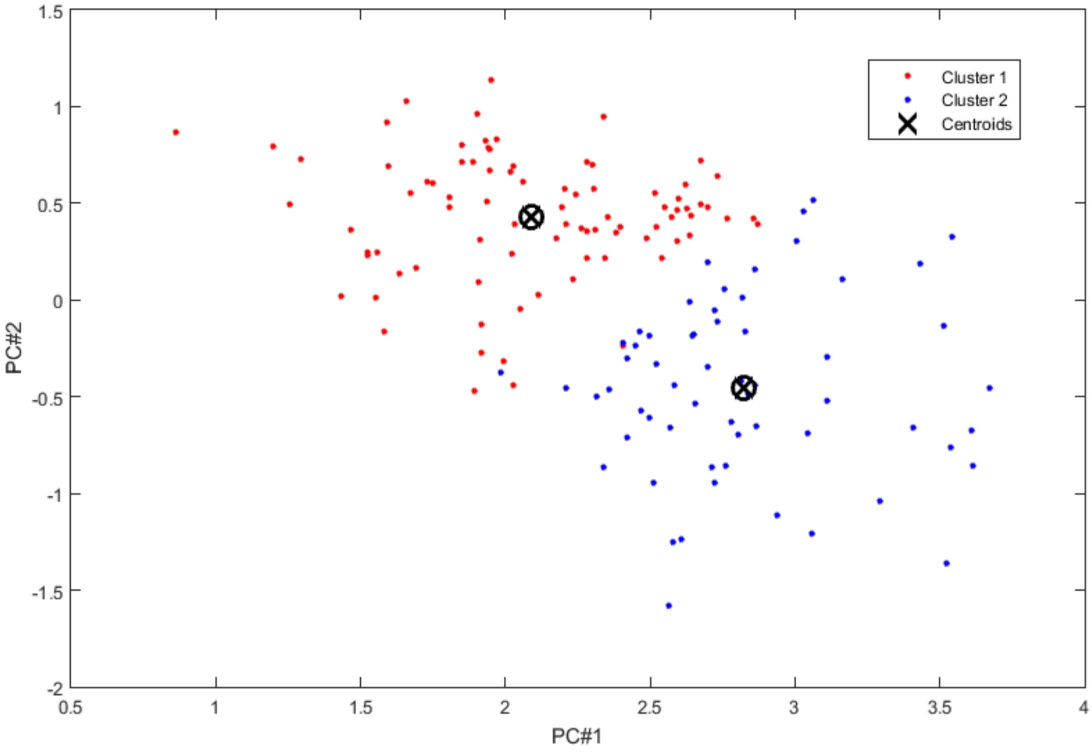
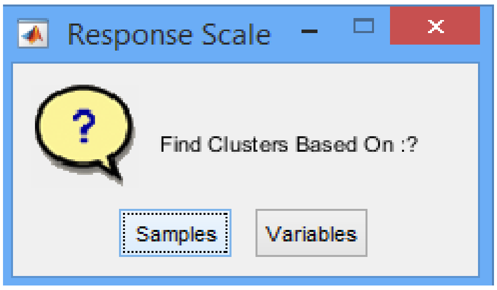


Figure 4.18: The plot of the first two PCs of k-means clustering for k=2 of model 1

4.1.4 Clustering for model 1

In order to perform clustering, we use the option clustering () at the interface of the chemometric program in the section of Clustering. After running the previous option a new dialog menu query us with which response scale we want to find clusters. The first choice is clustering based on samplers and the second choice is based on variables as it is shown in figure 4.19:


Response Scale

Find Clusters Based On :?

Figure 4.19: The dialog menu with the available choices for clustering

For the needs of our models we choose to push the Samples button because we want our clustering to be based on samples.

In the new dialog window that opens: we have the choice to find the best clustering method in terms of Distance and Linkage. This is done if we press the Find Best button as it is depicted in figure 4.20.

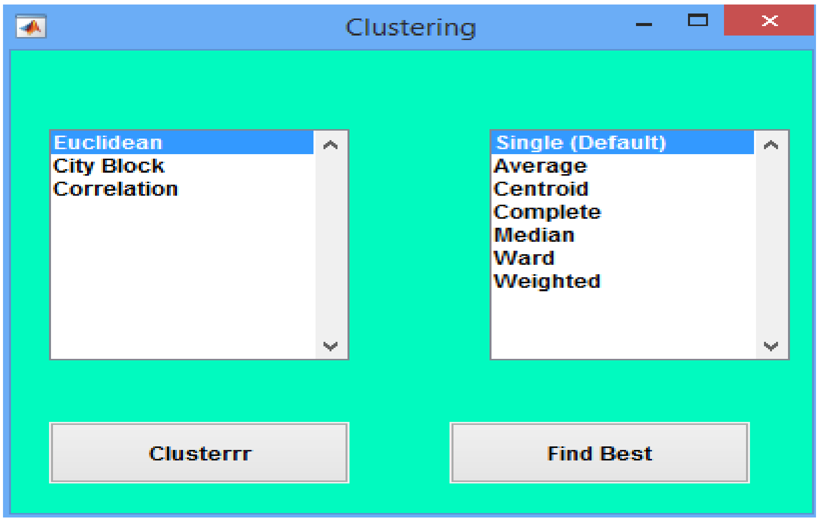


Figure 4.20: The dialog menu with the available choices for distance and linkage methods.

The result for the Find Best choice, in command window of MATLAB program is the following:

```
c =

   0.6868   0.7671   0.7444   0.5382   0.7468   0.4895   0.5018

   0.6541   0.7587   0.7623   0.5095   0.4701   0.5774   0.4201

   0.5656   0.7104   0.7261   0.5264   0.5741   0.4956   0.5775

bestLinkage =  2

 bestDistance = 1
```

The explanation of which is that the best metric for distance is the Euclidean and the best linkage method is the Average. Thus we select the above options in the dialog window as it is shown in figure 4.21:

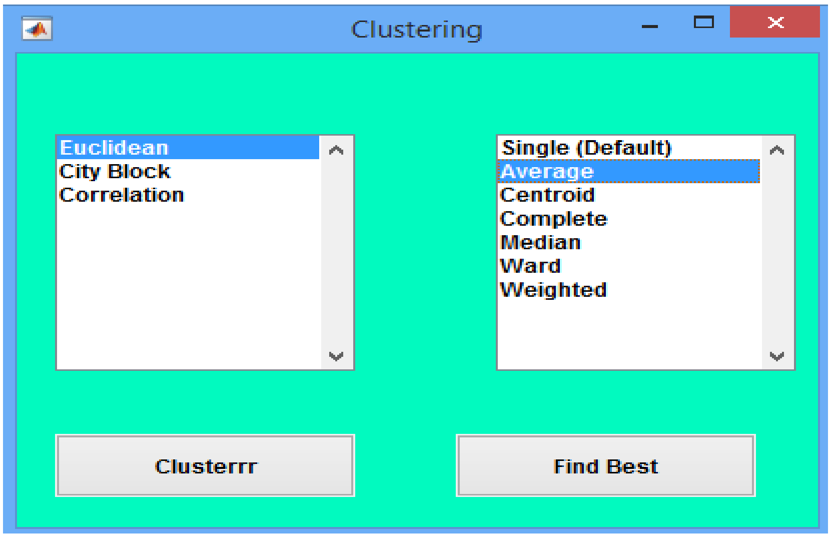


Figure 4.21: The best choices for distance and linkage method are marked for model 1.

The result for hierarchical clustering is presented in figure 4.22 in which the verification is 76.0449 %. This figure illustrates four major clusters for the oil samples.

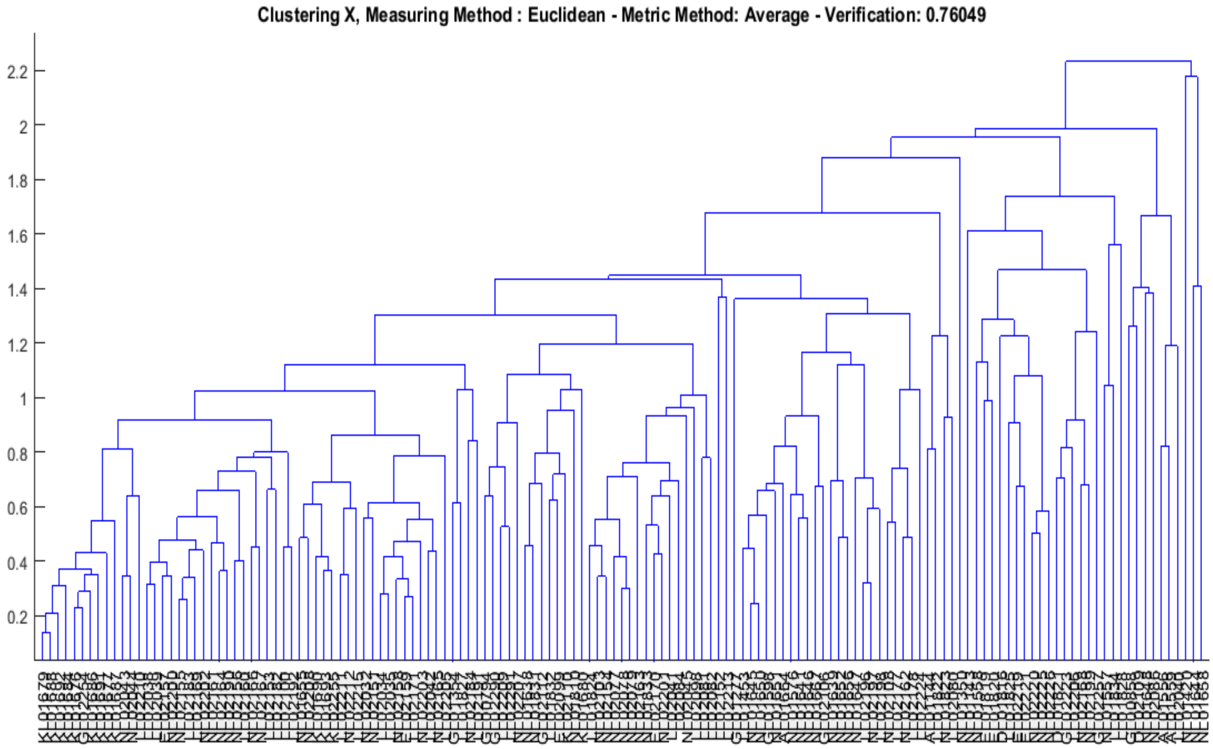


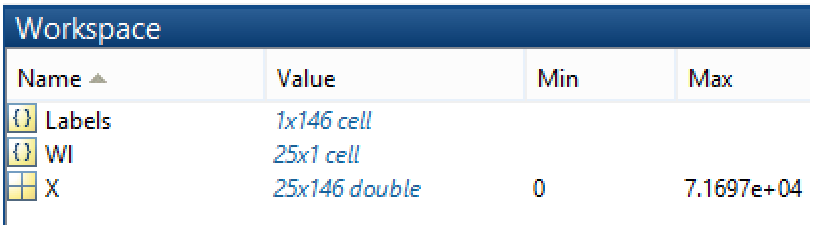
Figure 4.22: The hierarchical clustering dedrogram for model 1

4.2 Model 2 Saturate range with all variables

The process that is followed in order to run the model 2 is very similar with the procedure steps that was followed in model 1. To be more specific, the next steps are followed in order to load the data for saturate range:

First of all, we must load the values of chromatographic peak areas for the saturated fraction from the Excel file that are existed in the spreadsheet with name Complete_Saturate in Excel file All_Devonian_data.xlsx.

In MATLAB environment and especially in the workspace section we have created the following variables: Labels cell array that contains the sample names. The Wl that is a cell array containing the variable names and finally the dataset X is a 25 X 146 matrix that has the values of chromatographic peak areas for the saturated fraction for each variable and sample as illustrated in figure 4.23.



Name	Value	Min	Max
Labels	1x146 cell		
Wl	25x1 cell		
X	25x146 double	0	7.1697e+04

Figure 4.23: The workspace section with the selected variables for model 2

The next step is with the use of the available option pre_scaling_0_1 () to normalize the 146 samples in the range of 0 to 1. The result of this pretreatment is that we take for chromatographic peak area values between 0 and 1 for all samples.

The final step in pretreatments is the normalization of the 25 variables that contains the saturated fraction with the MATLAB function norm_variables_0_1 () in the range of 0 to 1 for all samples.

4.2.1 PCA analysis for model 2

In order to perform Principal Component Analysis, we run the option PCA_analysis () in the section Analysis at the interface of the chemometric program and we take the following results.

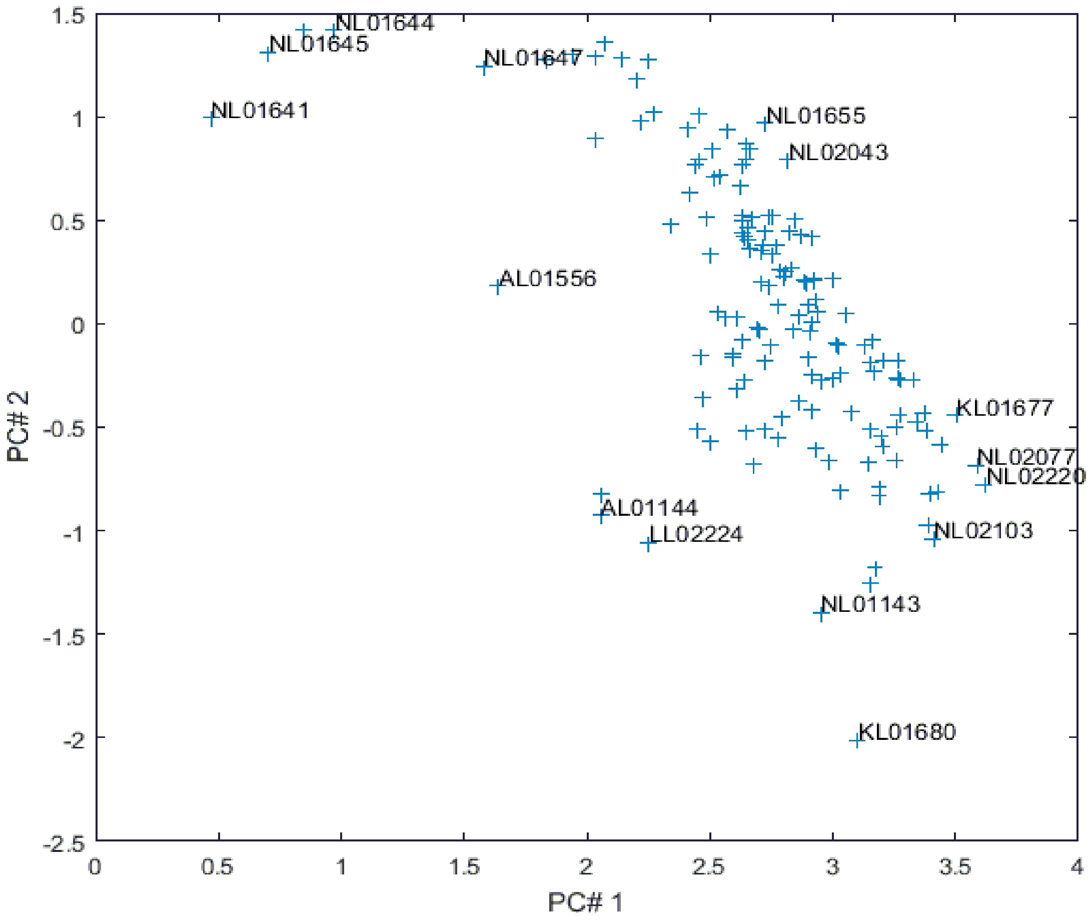


Figure 4.24: Plot of the two major principal components of model 2

In figure 4.24 the plot of the first two major principal components for the case of saturate range is illustrated. We also do not see a clear separation for samples in distinguish clusters because the behavior of the majority of samples is the almost the same. But the samples that are marked in figure have different behavior in comparison with the samples' majority. These samples are the following: NL01641, NL01645, NL01644, NL01647, NL01655, NL02043, AL01556, AL01144, LL02224, KL01677, NL02077, NL02220, NL02103, NL01143 and KL01680.

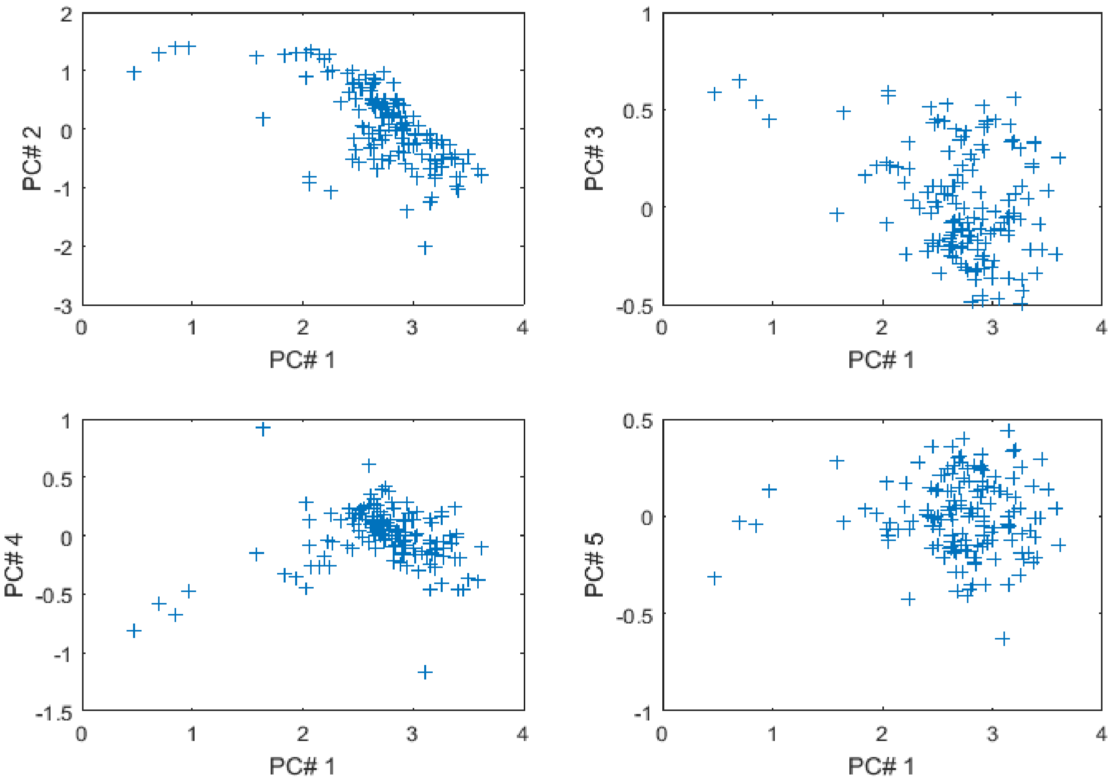


Figure 4.25: The subplot of five principal components of model 2

The figure 4.25 illustrates in four subplots, the first principal component versus the second, third, fourth and fifth principal component for all samples in our data.

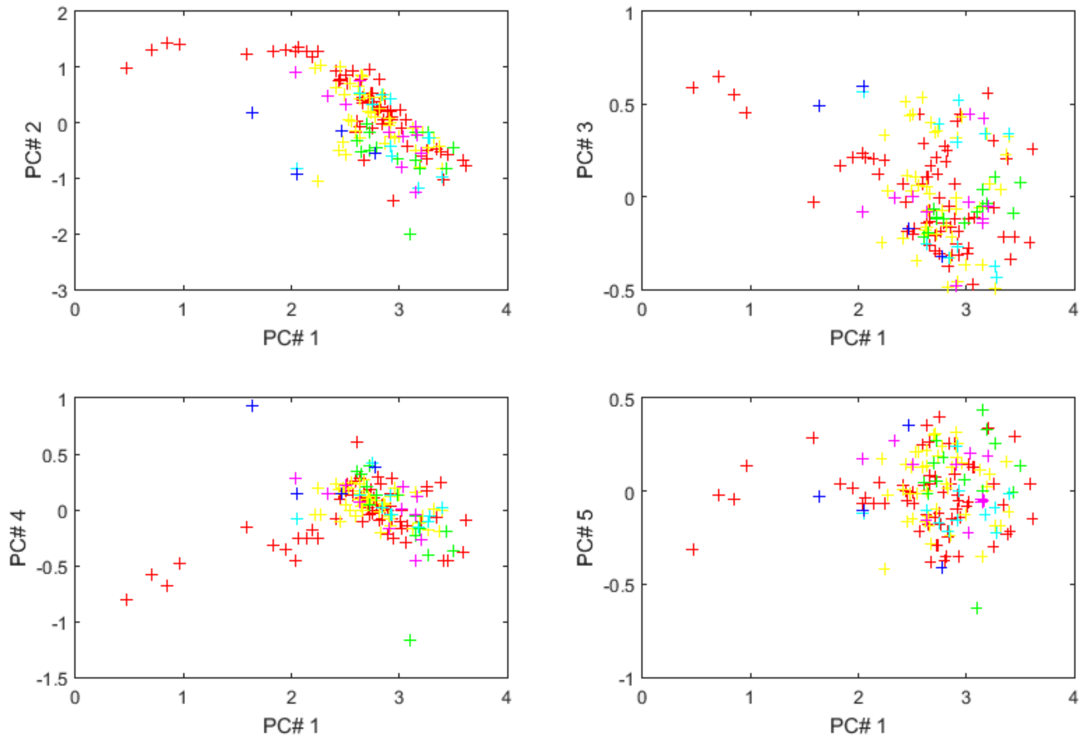


Figure 4.26: The subplot for five principal components of model 2 with different color for each formation

The figure 4.26 illustrates in four instances, the first principal component versus the second, third, fourth and fifth principal component for all our data but in this figure each color depict different formations (For mapping see table 4.1).

At this point I would quote the plot of the first principal component in the space of longitude and latitude coordinates of our samples in figure 4.27.

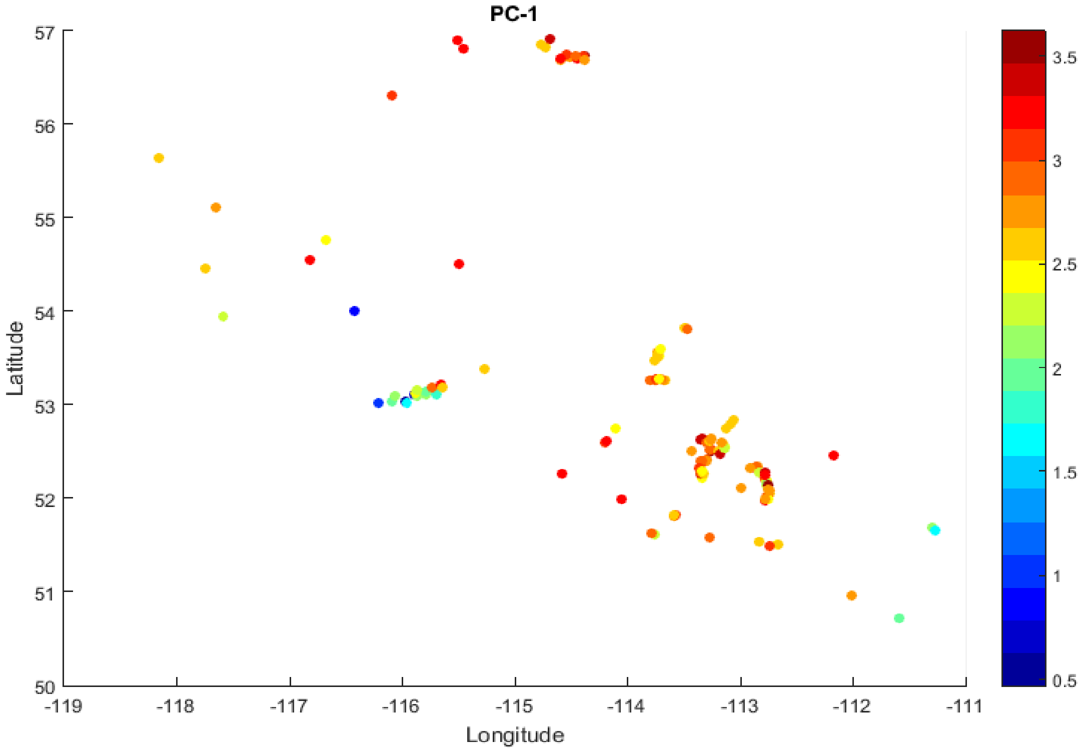


Figure 4.27: Plot of the first principal component of PCA analysis vs the geographical location of the samples.

Continuing, figure 4.28 presents the plot of the second principal component according to the latitude and longitude coordinates of our samples.

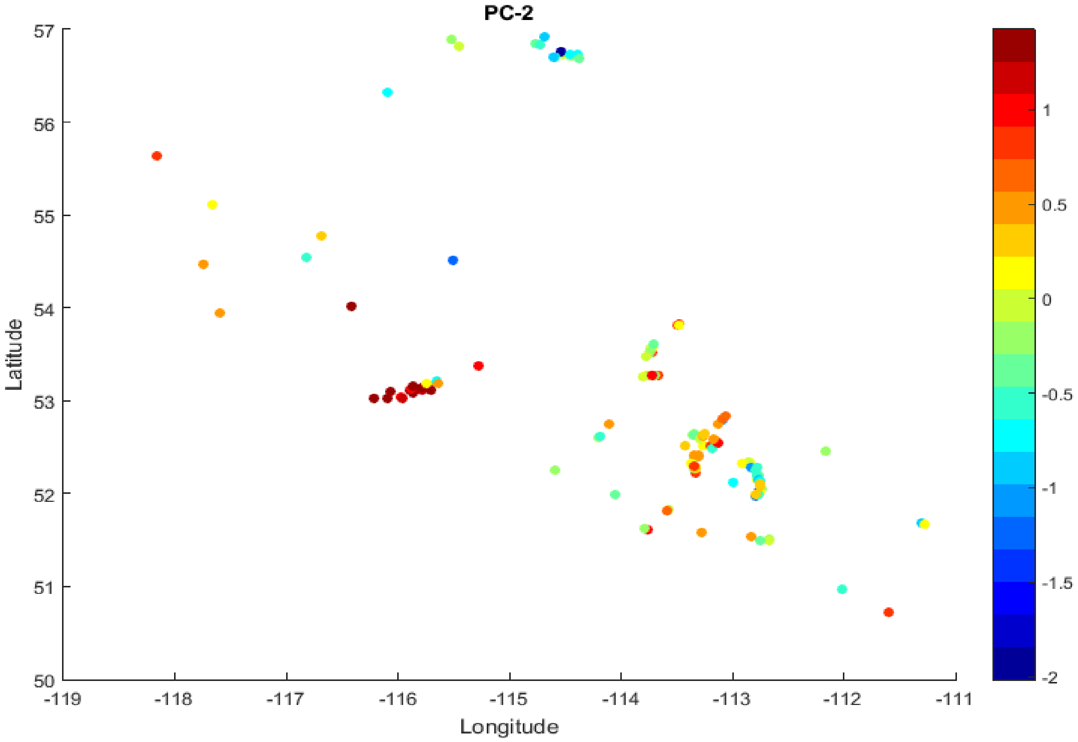


Figure 4.28: Plot of the second principal component of PCA analysis vs the geographical location of the samples.

The location plots of the first Principal Component and the second Principal component of the group of Nisku oils that are placed with longitude from -115.50 to -116.50 and latitude from 52.70 to 53.20 carries important information. The values for this group of oils for the first principal component (PC1) are low and high for the second principal component (PC2) as it is shown in figures 4.27 and 4.28.

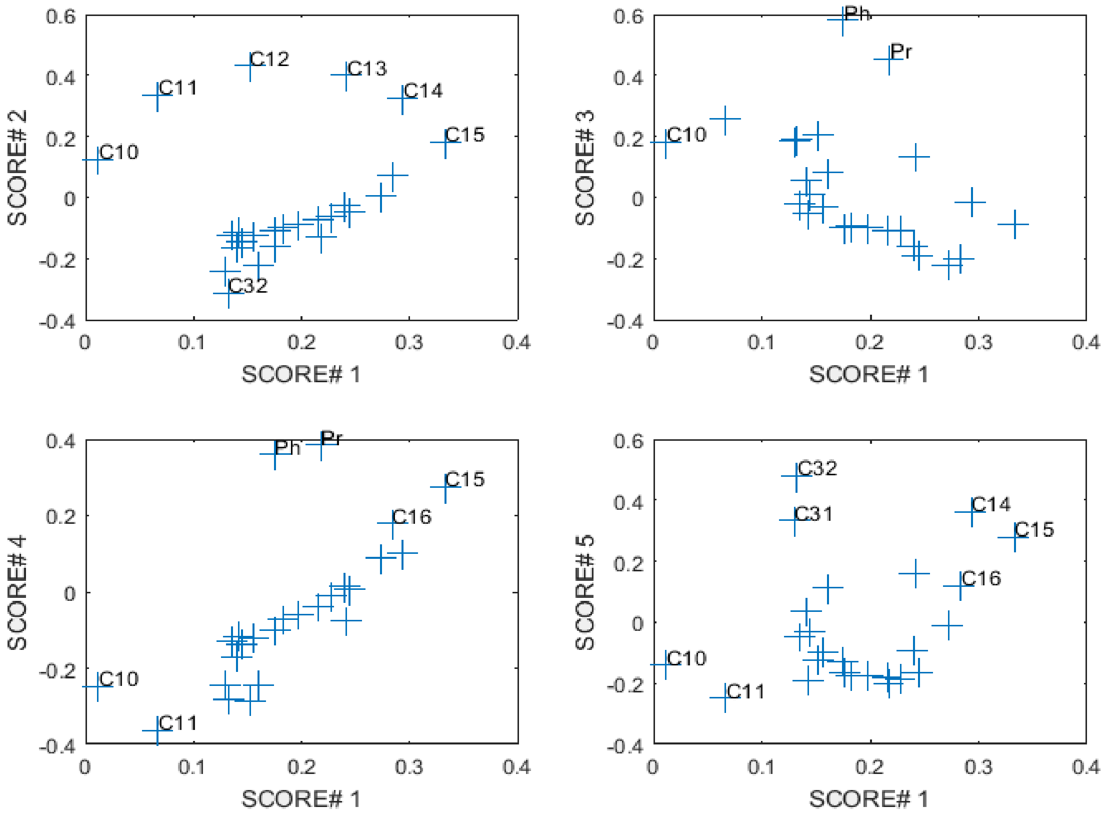


Figure 4.29: Original variable loadings for the first five principal components from the model 2.

The original variable loadings are presented in figure 4.29. In this subplot we have four instances that depict the original variable loadings for the first principal component versus the original variable loadings for the second, third, fourth and fifth principal component.

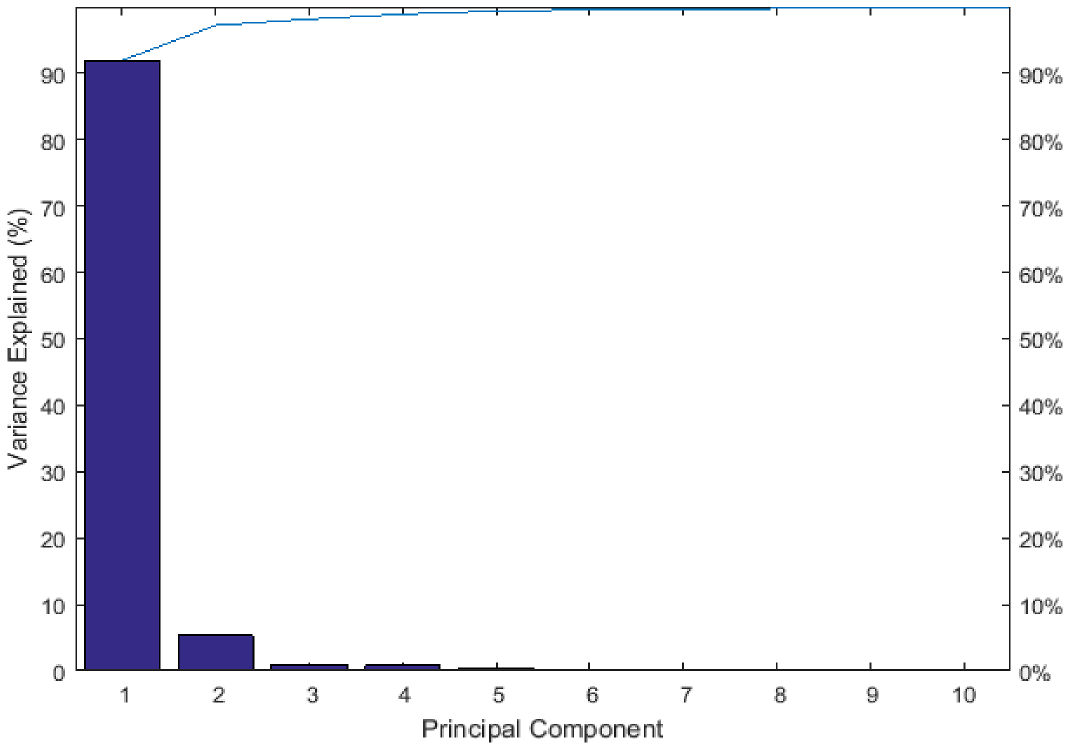


Figure 4.30: Percentage of variance explained for each principal component

Figure 4.30 reveals the contribution of each principal component in the total variance that is explained from the PCA model. As it is shown the first PC explains 91% of variation and the second PC explains 6% of the remaining variation.

4.2.2 Kernel PCA for model 2

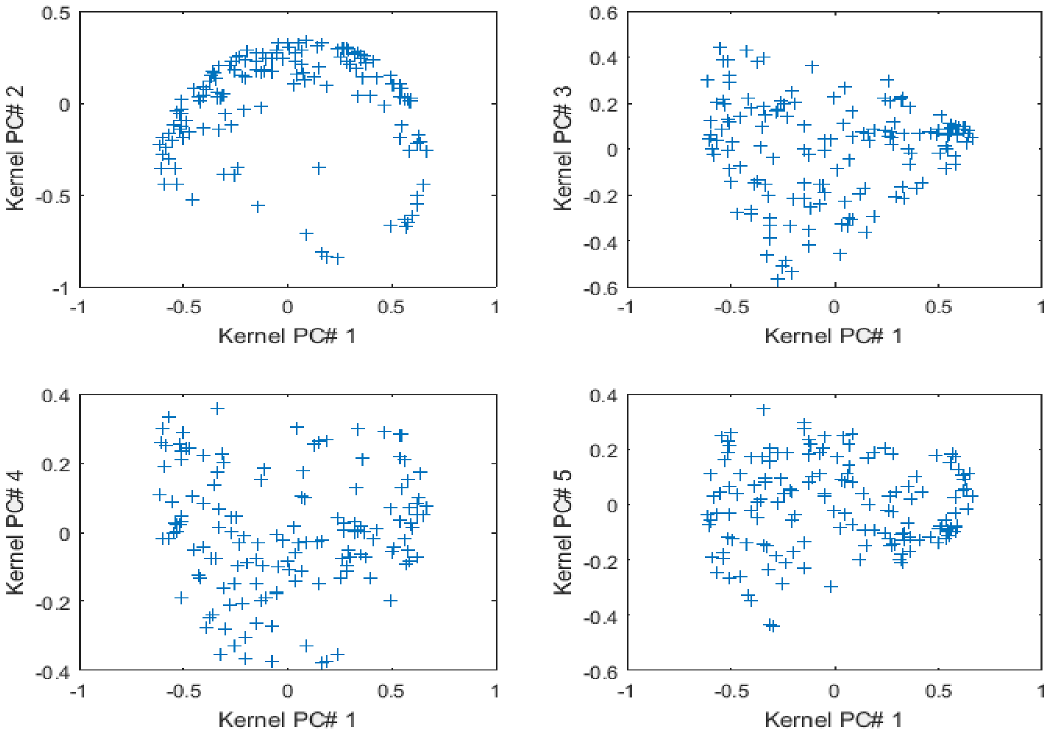


Figure 4.31: The subplot of five kernel principal components of model 2

The figure 4.31 illustrates in four subplots, the first kernel principal component versus the second, third, fourth and fifth kernel principal component for all our data.

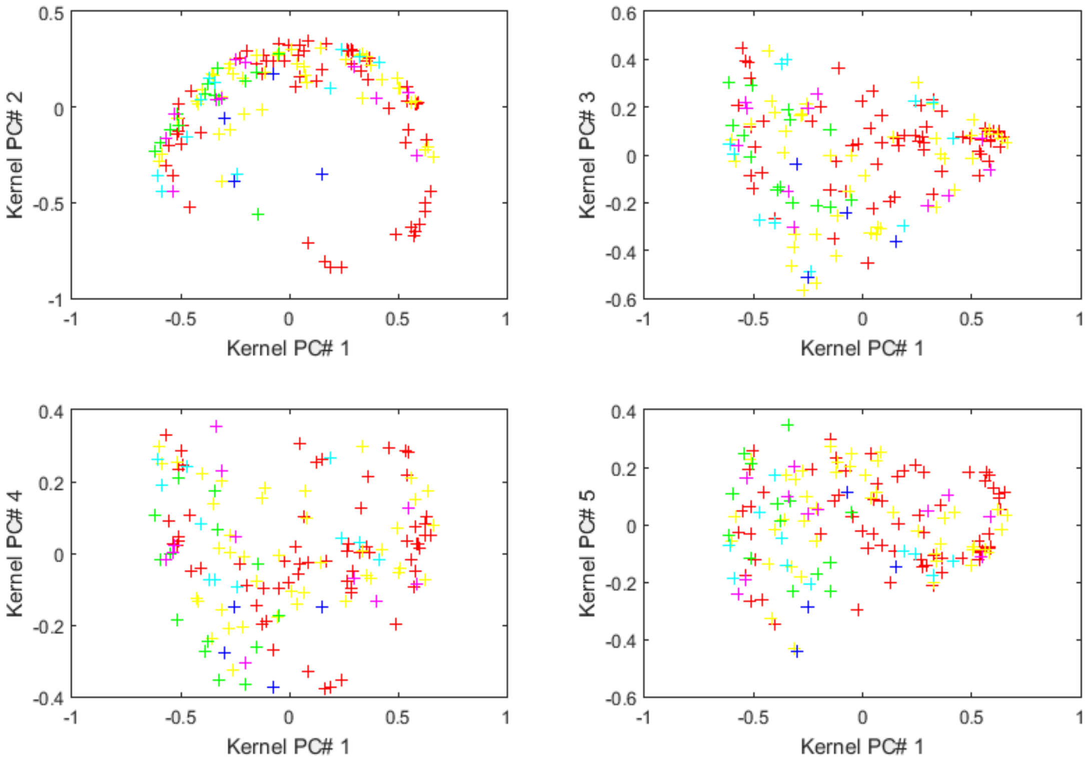


Figure 4.32: The subplot for five kernel principal components of model 2 with different color for each formation

The figure 4.32 illustrates in four instances, the first kernel principal component versus the second, third, fourth and fifth kernel principal component for all our data. In addition to, in this figure each color depict different formations (For mapping see table 4.1).

The plot of the values of first kernel principal component in the space of longitude and latitude coordinates of our samples are presented in figure 4.33.

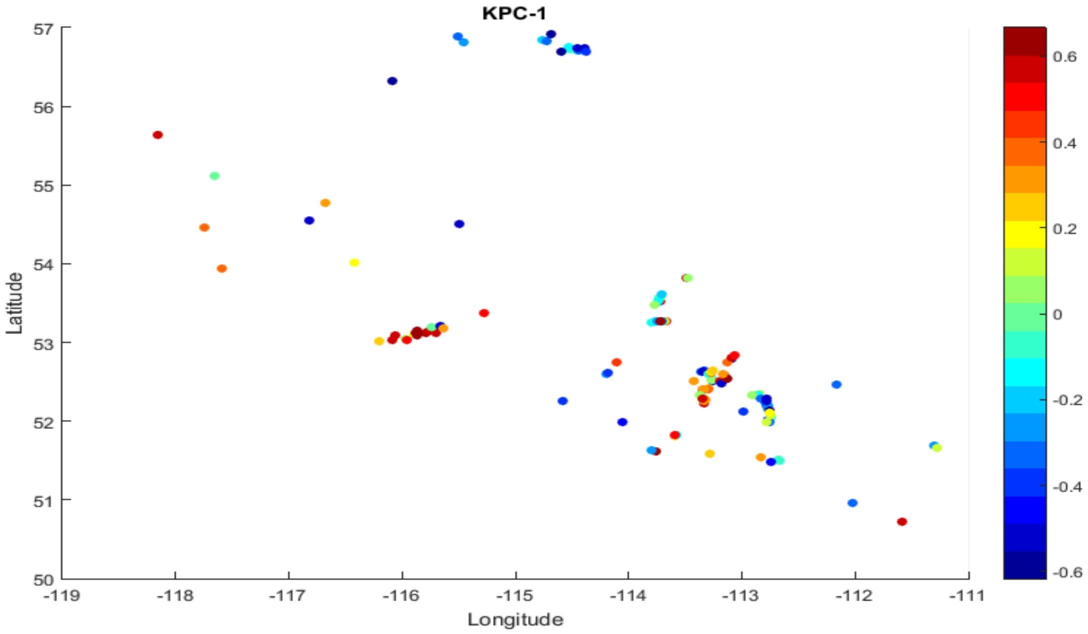


Figure 4.33: Plot the values of the first kernel principal component of KPCA analysis in relationship with the location of the samples.

Continuing, figure 4.34 illustrates the plot of the second kernel principal component according to the latitude and longitude coordinates of our samples.

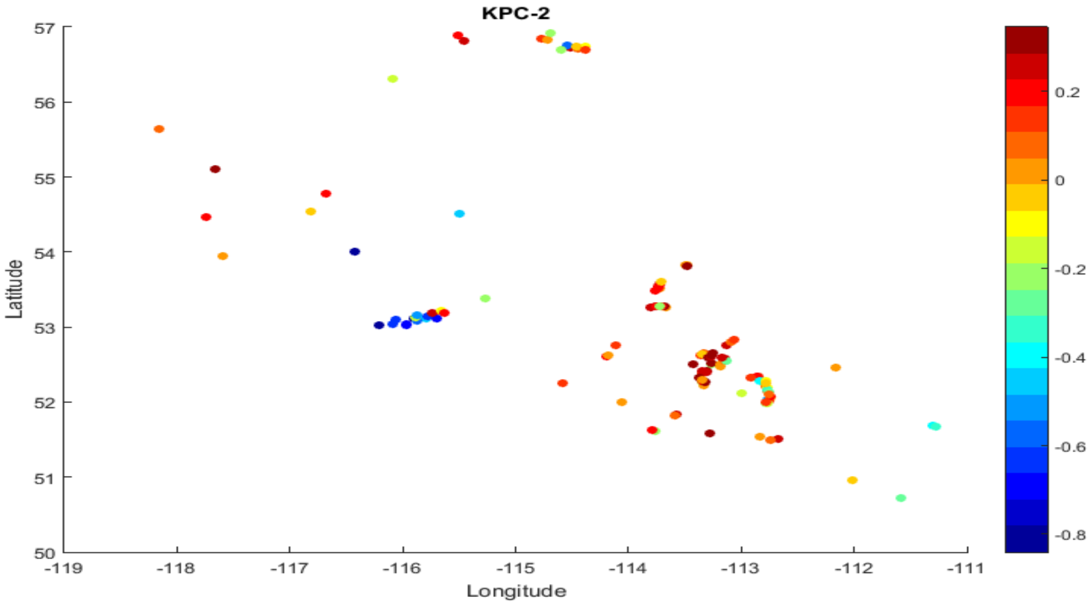


Figure 4.34: Plot the values of the second kernel principal component of KPCA analysis in relationship with the location of the samples.

4.2.3 Silhouette – k-means for model 2

With the use of the available option Silhouete () at the interface of the chemometric program in the section of clustering, we get a k-means clustering using two, three, four and five clusters. The k-means clustering is repeating five times for each case. This is achieved with the use of replicates as an argument in kmeans MATLAB function.

In table 4.3 we summarize the results that are taken:

k-means clustering	Best Total sum of distances	Average silhouette value
K=2	81.6495	**0.5191**
K=3	60.5961	0.4868
K=4	51.2501	0.3812
K=5	45.1028	0.3911

Table 4.3: Summary of k-means clustering for model 2

The silhouette plots for K=2, K=3, K=4 and K=5 are shown in following figure.

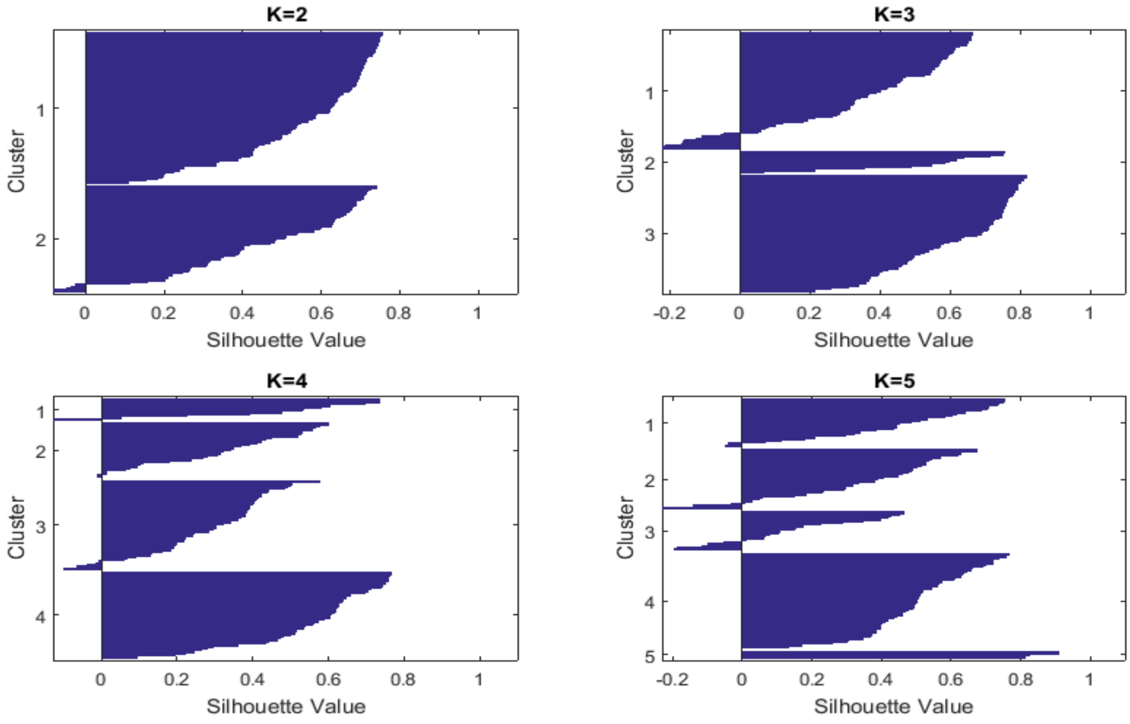


Figure 4.35: Silhouette plots for k=2, k=3, k=4 and k=5 clusters for model 2

We see that in the case of two clusters we have the mostly large silhouette values and very few negative values in cluster two. A one-number summary in order to describe the performance of each clustering, is the average of the silhouette values. The two cluster solution has an average silhouette value of 0.5191 and this value is the maximum among the others cases. Thus it is an indicator that the grouping into two clusters using k-means is better than the one with three or four or five groups.

In figure 4.36 the plot of the first two Principal Components (PCs) of k-means clustering, for the case of k =2 for our dataset of 146 oils are presented with different color for samples members that belong to a different cluster:

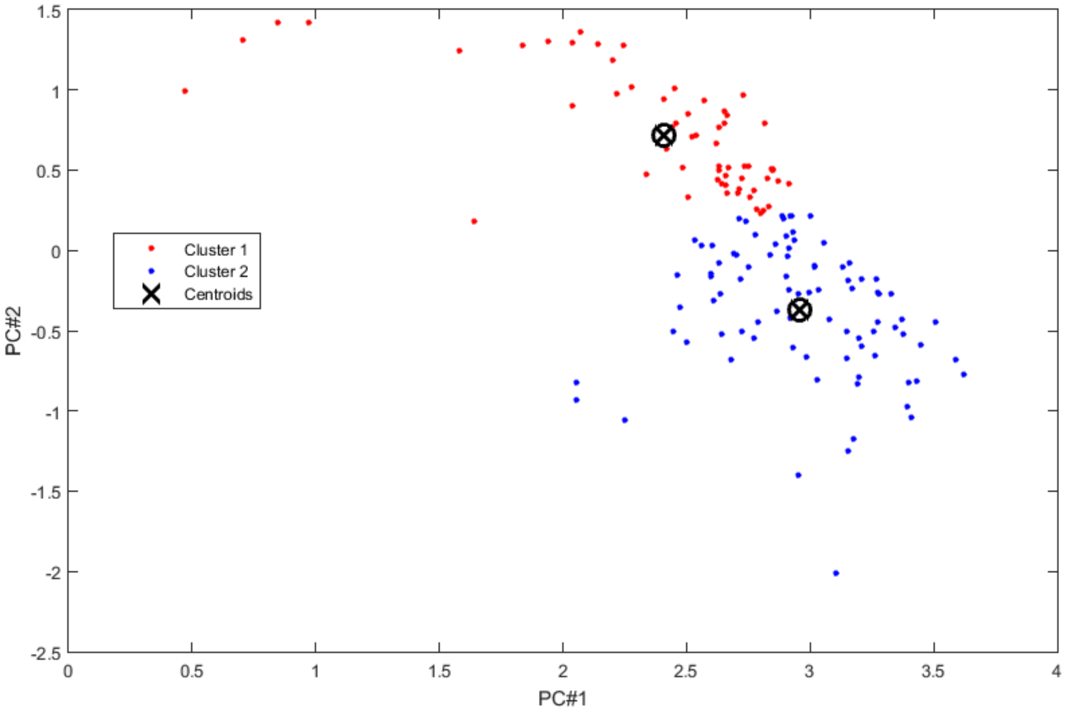


Figure 4.36: The plot of the first two PCs of k-means clustering for k=2 of model 2

4.2.4 Clustering for model 2

In order to perform clustering, we use the option clustering () at the interface of the chemometric program in the section of Clustering. If we follow the same steps as that in model 1, thus firstly select clustering based on samples and secondly use the Find Best choice in window dialog the following results are obtained:

```
c =

   0.6868    0.7671    0.7444    0.5382    0.7468    0.4895    0.5018

   0.6541    0.7587    0.7623    0.5095    0.4701    0.5774    0.4201

   0.5656    0.7104    0.7261    0.5264    0.5741    0.4956    0.5775


bestLinkage = 2

bestDistance = 1
```

The explanation of which is that the best metric for distance is the Euclidean and the best linkage method is the Average. If we choose the above options in the dialog window that is opened the following dedrogram obtained as it is shown in figure.

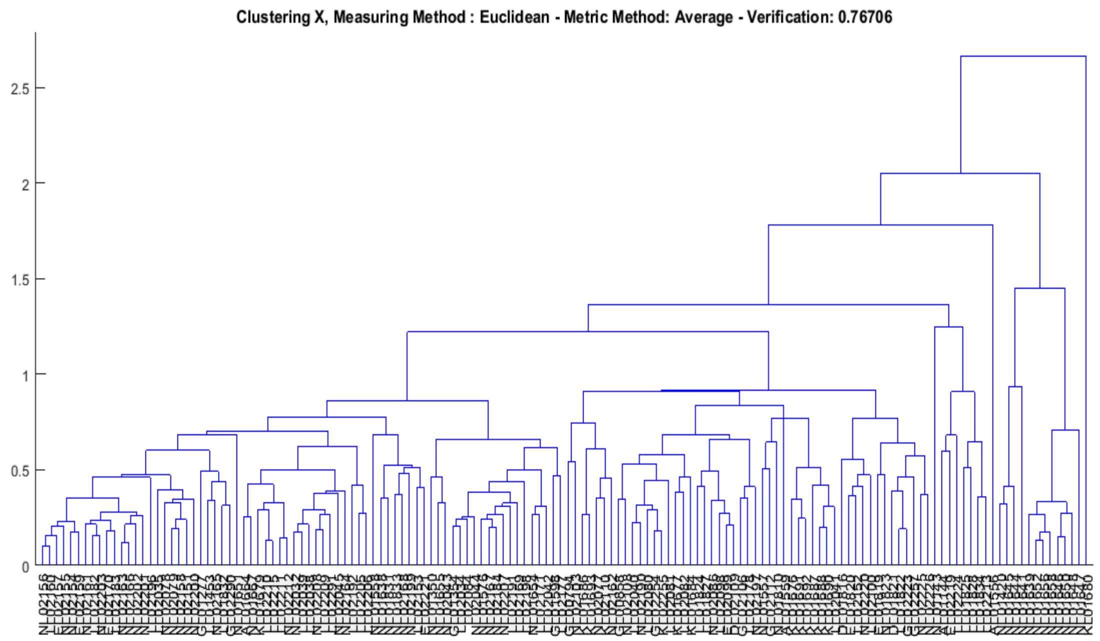


Figure 4.37: The hierarchical clustering dedrogram for model 2

The result for hierarchical clustering is presented in figure 4.37 in which the verification is 76.706 %. To be more specific, two major clusters of oil samples are illustrated with a small third one that is contained from the more dissimilar samples in this model.

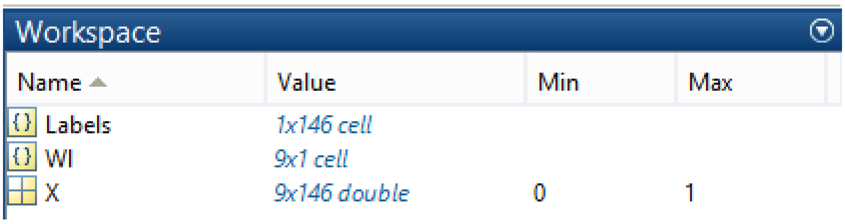
4.3 Model 3 Calculation of nine geochemical indexes from Gasoline range

The process that is followed in order to run the model 3 is very similar with the procedure steps that was followed in model 1 and 2. To be more specific, the next steps are followed for loading the data for gasoline range:

First of all, we must load the values of chromatographic peak areas for the gasoline fraction from the Excel file that are existed in the spreadsheet with name Complete_Gasoline_out in Excel file All_Devonian_data.xlsx.

Continuing, the next step is the calculation of nine geochemical indices that were described in chapter 3.3. This obtained with the run of Calc_gasoline_ratios.m MATLAB file.

In MATLAB environment and especially in the workspace section we have created the following variables: Labels cell array that contains the sample names. The Wl that is a cell array containing the variable names and finally the dataset X is a 9 X 146 matrix that has the values of nine geochemical indices that were created from the chromatographic peak areas for the gasoline fraction for each variable and sample as illustrated in figure 4.38. These indexes are: K1, A, B, C, I, F, R, U, H.



Workspace

Name ▲	Value	Min	Max
{} Labels	1x146 cell		
{} WI	9x1 cell		
X	9x146 double	0	1

Figure 4.38: The workspace section with the importing variables for model 3

The next step in pretreatments is the normalization of the 9 variables that contains the gasoline range indices with the MATLAB function norm_variables_0_1 () in the range of 0 to 1 for all samples.

In the final step, with the use of the available option pre_scaling_0_1 () to normalize the 146 samples in the range of 0 to 1. The result of this pretreatment is that we take for the nine indices variables, values between 0 and 1 for all samples.

4.3.1 PCA analysis for model 3

In order to perform Principal Component Analysis, we run the option PCA_analysis () in the section Analysis at the interface of the chemometric program and we take the following results.

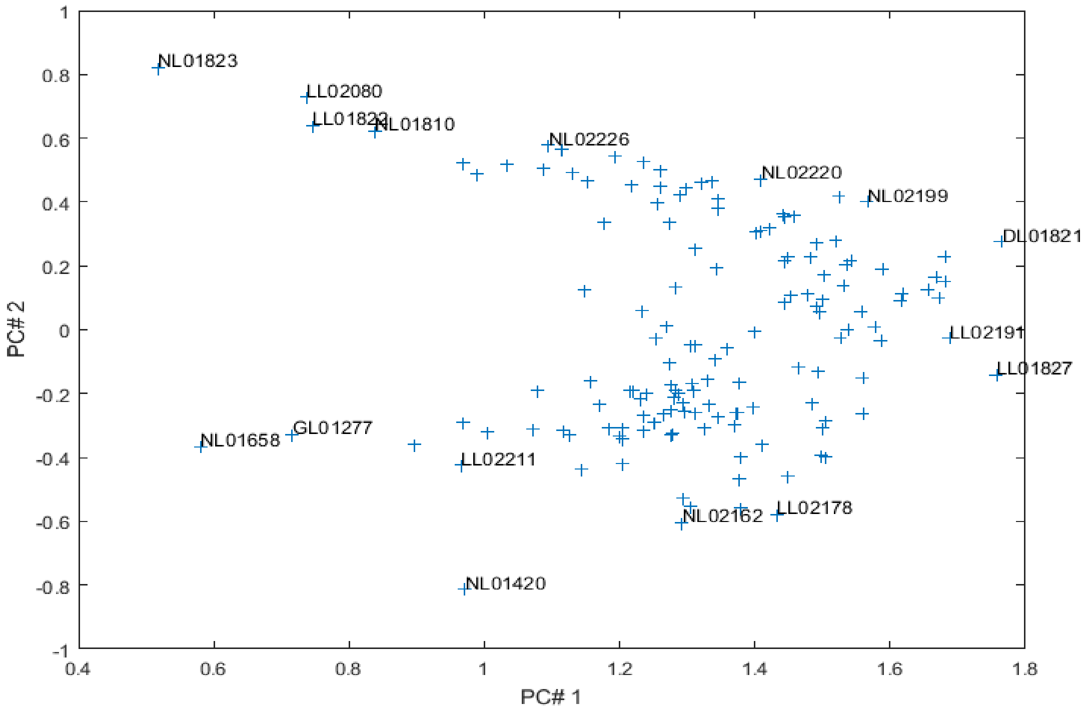


Figure 4.39: Plot of the two major principal components for model 3

In figure 4.39 is shown the plot of the first two major principal components. In this picture illustrated a clear enough separation for our samples in two distinguish clusters with some samples that have a little extreme values. These samples are the following: NL01823, LL02080, DL01821, LL01827, GL01277, NL01658 and NL01420.

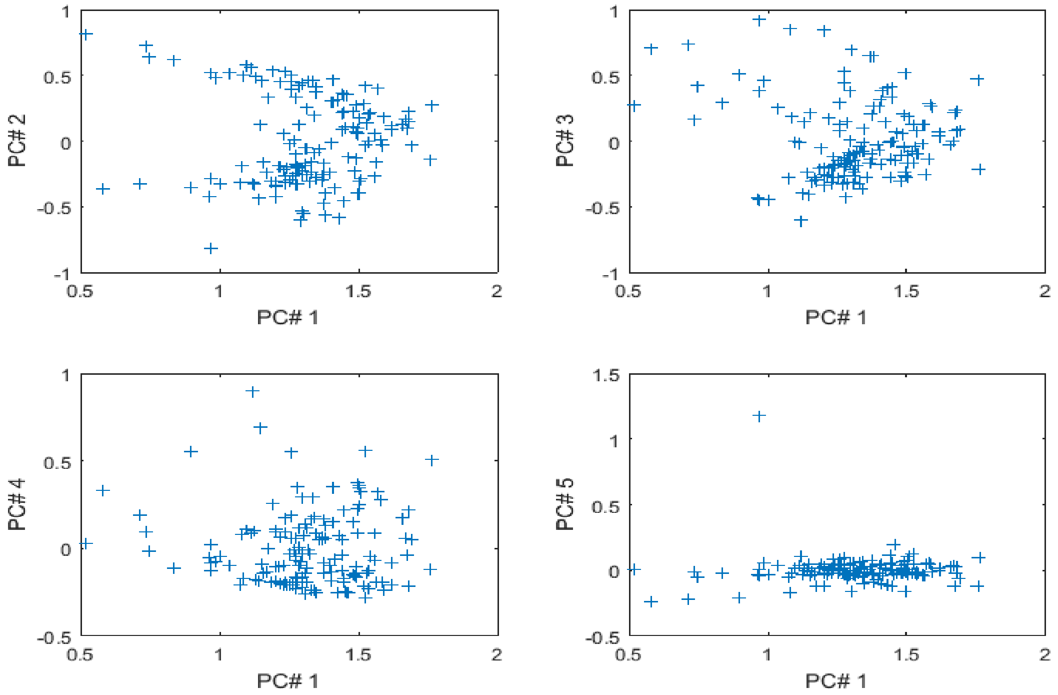


Figure 4.40: The subplot of five principal components of model 3

The figure 4.40 depicts in four subplots, the first principal component versus the second, third, fourth and fifth principal component for all samples in our data. As is mentioned in previous figure we have a clear enough separation in subplot PC#1 versus PC#2 and in subplot PC#1 versus PC#3. The above clear separation is not continuing in subplots PC#1 versus PC#4 and PC#1 versus PC#5.

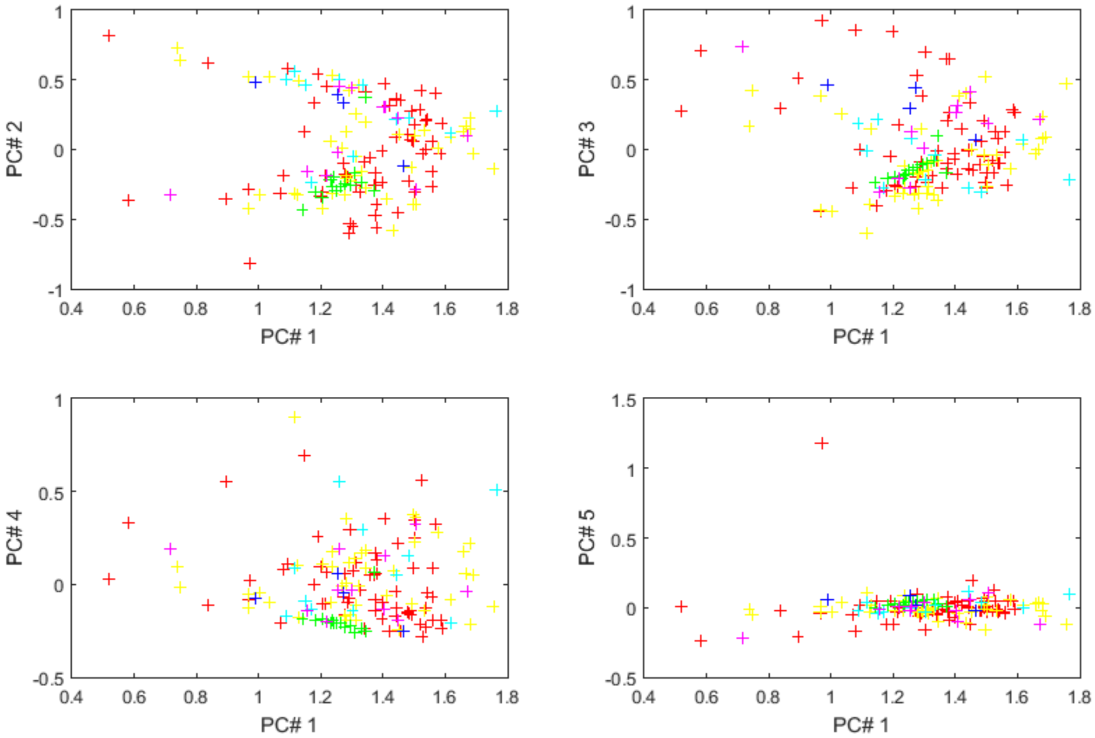


Figure 4.41: The subplot for five principal components with different color for each formation of model 3

The figure 4.41 illustrates in four instances, the first principal component versus the second, third, fourth and fifth principal component for all our data but in this figure each color depict different formations (For mapping see table 4.1).

The plot of the first principal component in the space of longitude and latitude coordinates of our samples is shown in figure 4.42.

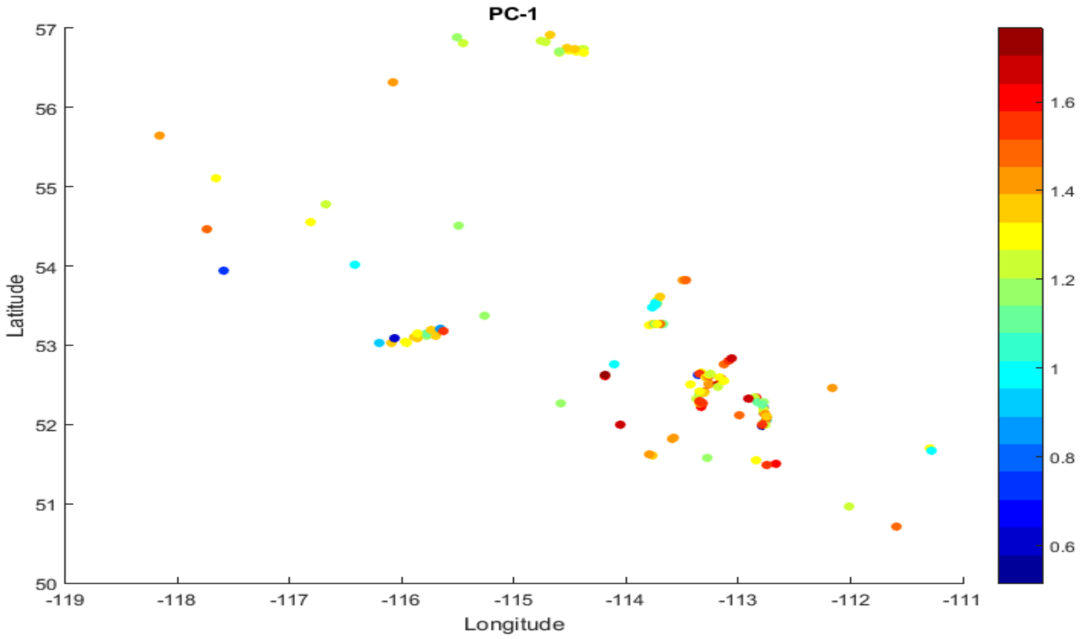


Figure 4.42: Plot of the first principal component of PCA analysis vs the geographical location of the samples.

Continuing, figure 4.43 presents the plot of the second principal component according to the latitude and longitude coordinates of our samples.

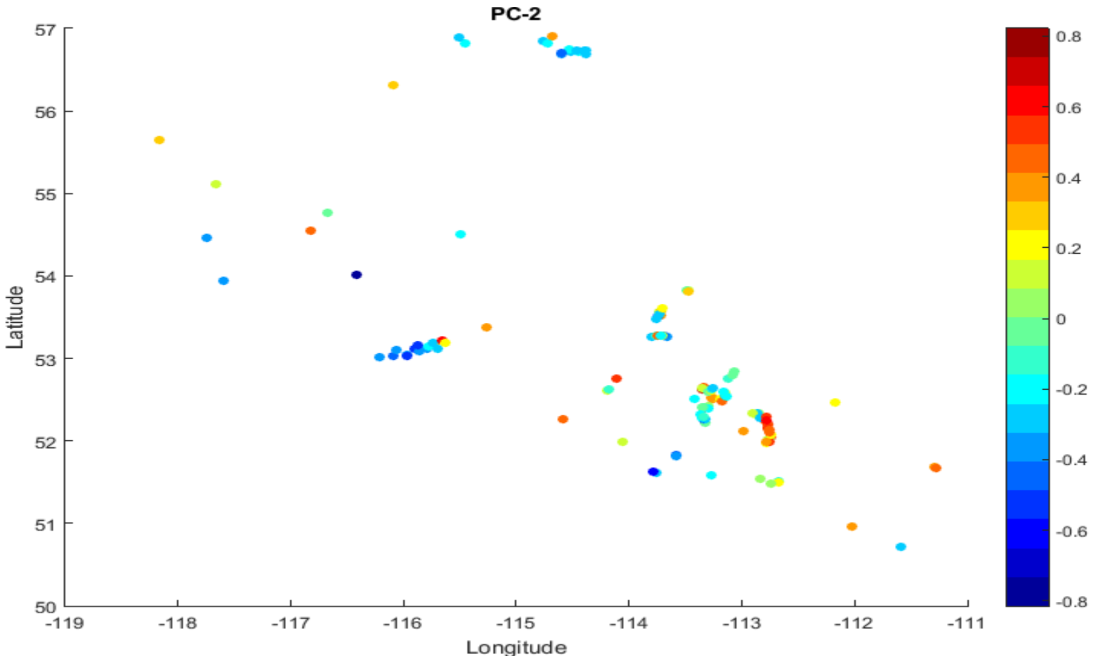


Figure 4.43: Plot of the second principal component of PCA analysis vs the geographic location of the samples.

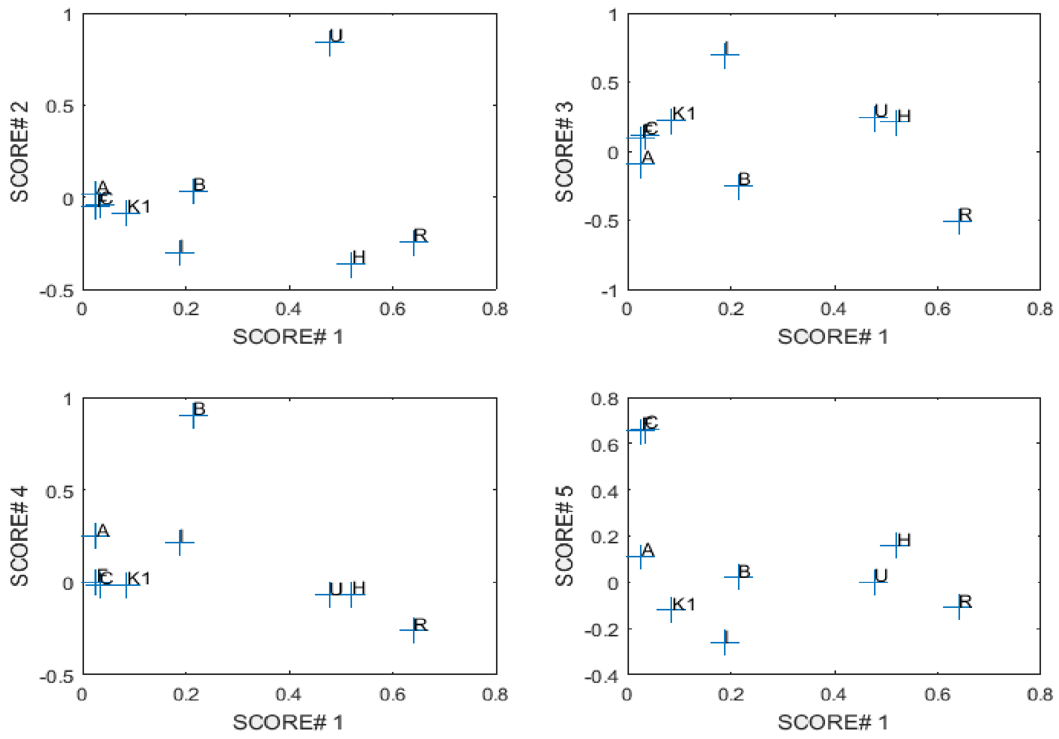


Figure 4.44: Original variable loadings for the first five principal components from the model 3

The original variable loadings are presented in figure 4.44. In this subplot we have four instances that depict the original variable loadings for the first principal component versus the original variable loadings for the second, third, fourth and fifth principal component. As it is shown the R index is very significant for the scores of first principal component and U and H indexes for the scores of second principal component.

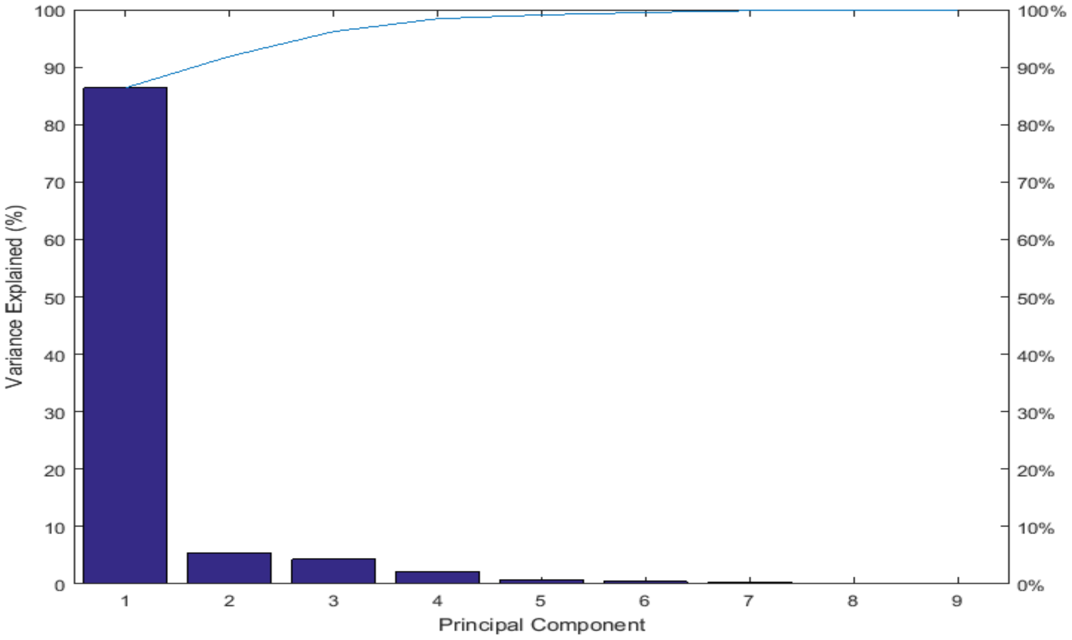


Figure 4.45: Percentage of variance explained of each principal component

Figure 4.45 reveals the contribution of each principal component in the total variance that are explained from the PCA model. As it is shown the first PC explains 87% of variation and the second PC explains 5% of the remaining variation.

4.3.2 Kernel PCA for model 3

In order to perform Kernel Principal Component Analysis, we run the option Kernel_pca_final () in the section Analysis at the interface of the chemometric program and we take the following results.

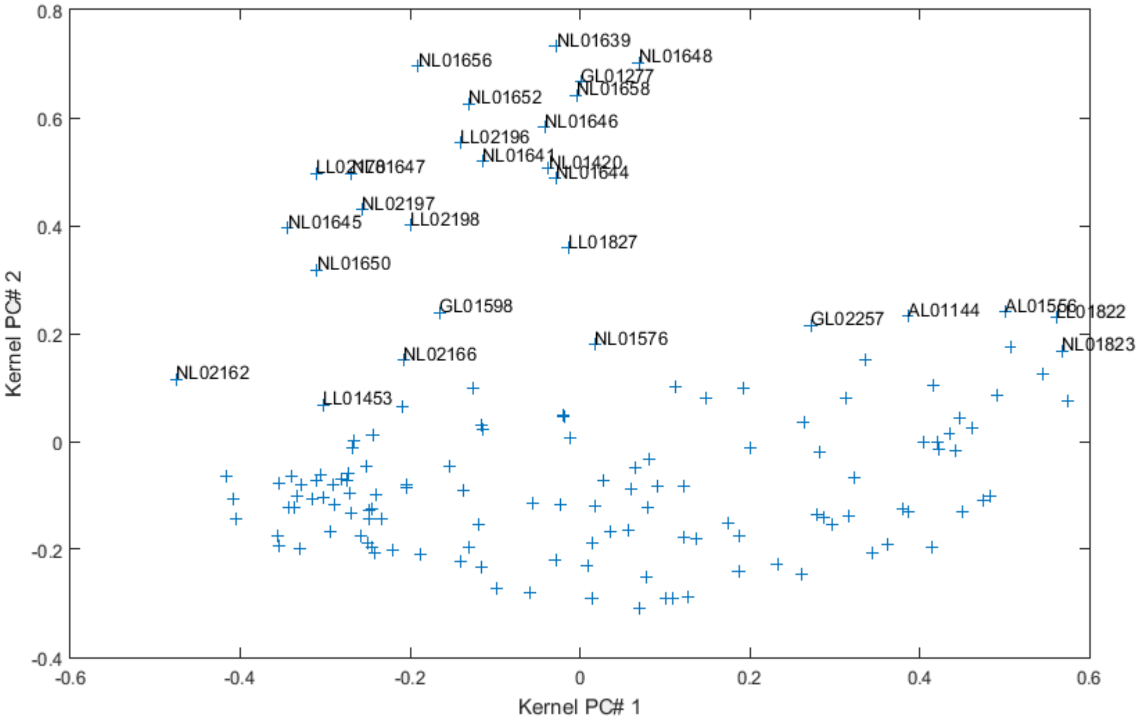


Figure 4.46: Plot of the two major kernel principal components for model 3

The figure 4.46 depicts clearly two different trends in our dataset. With the label names are the samples which are represented the first trend. The samples that are constructed the first trend have the following range in values for PC1: -0.5 to 0.6 and -0.8 to 0.1 for the second principal component. In second trend the samples have the same values in first principal component and in second PC: -0.4 to 0.1.

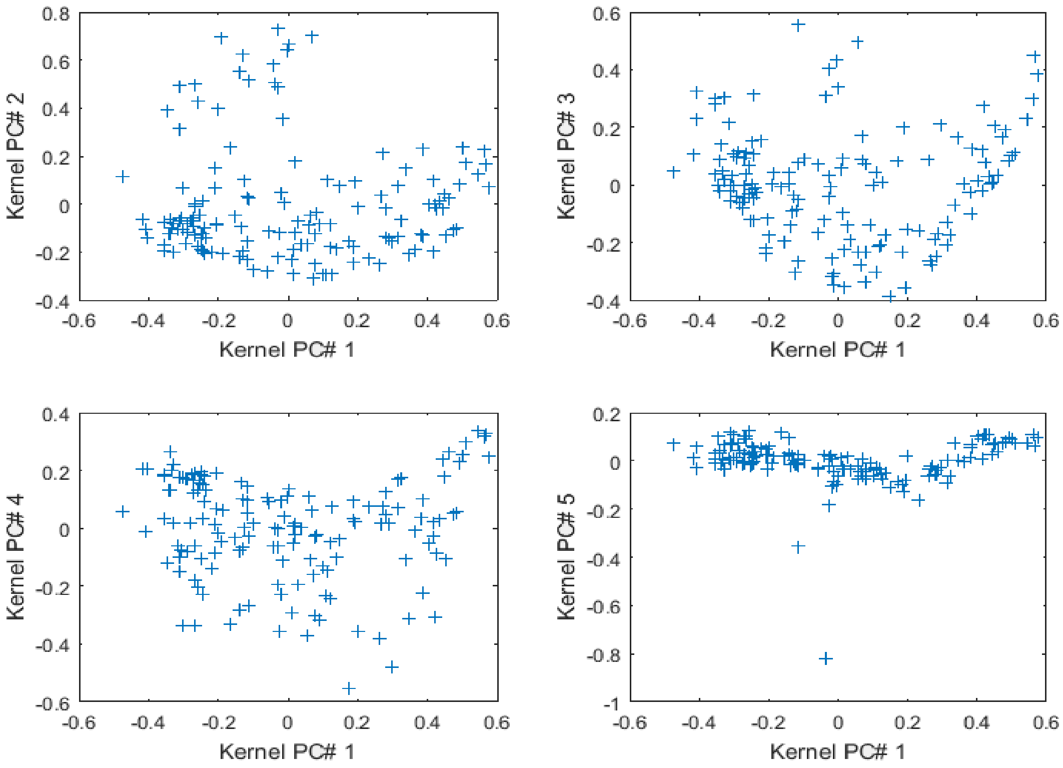


Figure 4.47: The subplot of five kernel principal components of model 3

The figure 4.47 illustrates in four subplots, the first kernel principal component versus the second, third, fourth and fifth kernel principal component for all our data. Two distinguish trends in kernel principal components are existed especially in subplot Kernel PC1 versus kernel PC2.

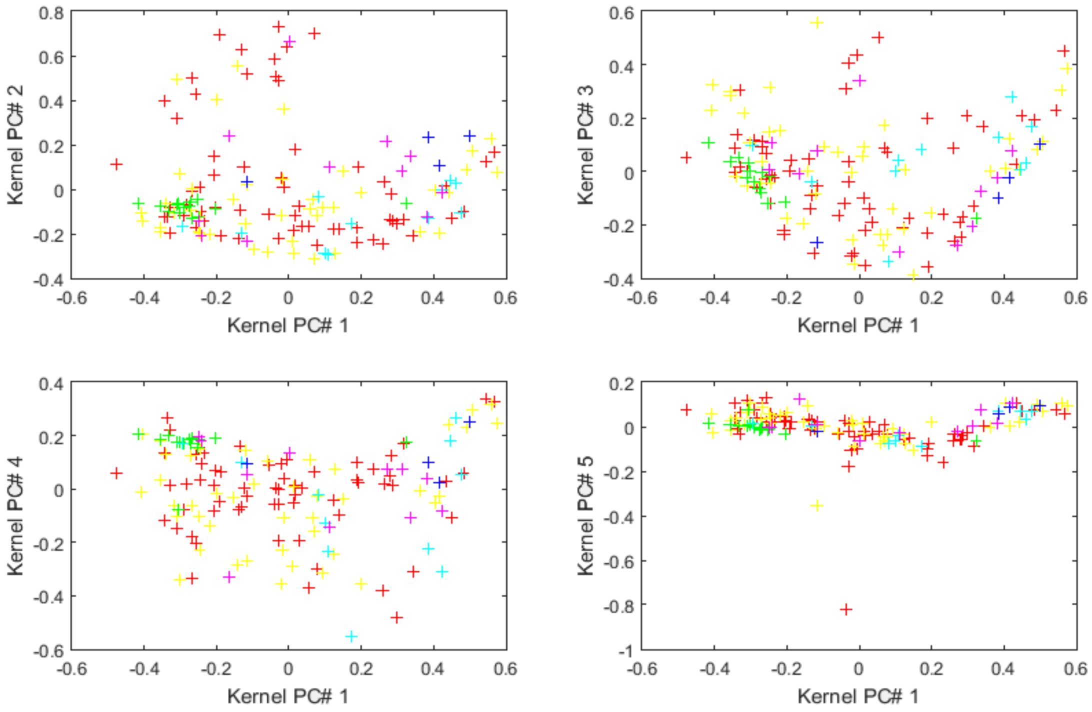


Figure 4.48: The subplot for five kernel principal components with different color for each formation of model 2

The figure 4.48 depicts in four instances, the first kernel principal component versus the second, third, fourth and fifth kernel principal component for all our data. In addition to, in this figure each color depict different formations (For mapping see table 4.1).

The plot of the values of first kernel principal component in the space of longitude and latitude coordinates of our samples are presented in figure 4.49.

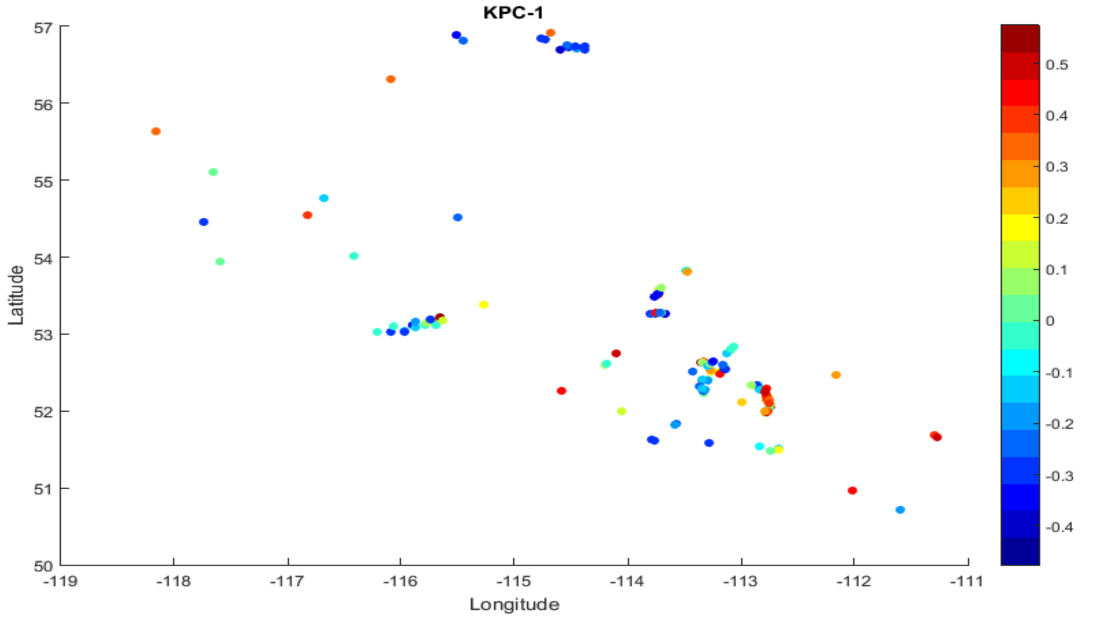


Figure 4.49: Plot of the first principal component of PCA analysis vs the location of the samples.

Continuing, figure 4.50 illustrates the plot of the second kernel principal component according to the latitude and longitude coordinates of our samples.

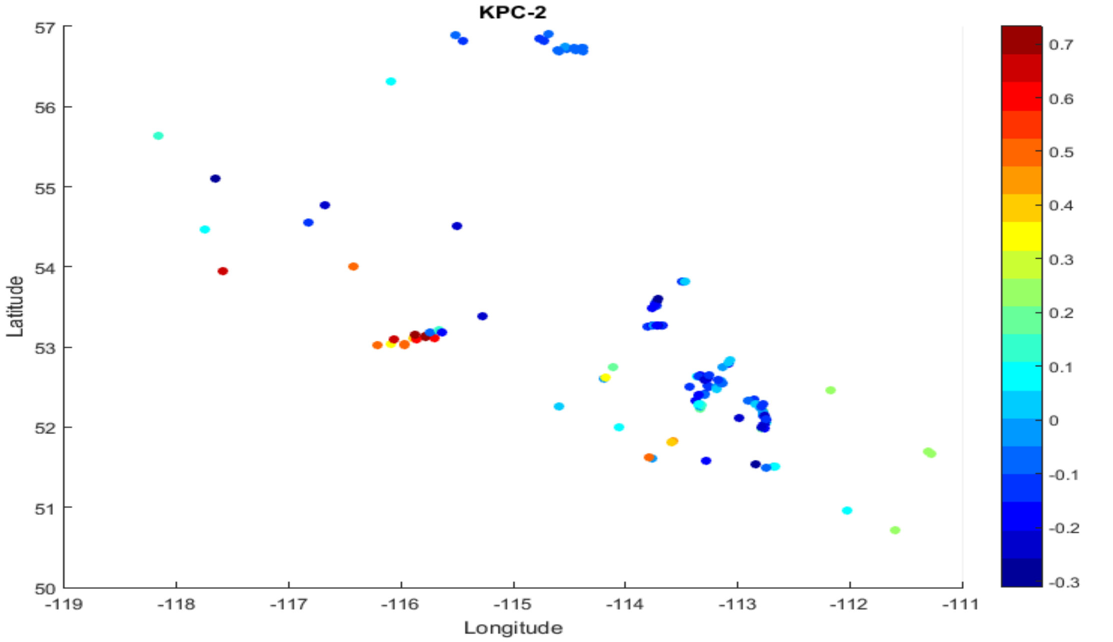


Figure 4.50: Plot of the second kernel principal component of KPCA analysis vs the location of the samples

4.3.3 Silhouette – k-means for model 3

With the use of the available option Silhouete () at the interface of the chemometric program in the section of clustering, we get a k-means clustering using two, three, four and five clusters. The k-means clustering is repeating five times for each case. This is achieved with the use of replicates as an argument in kmeans MATLAB function.

In table 4.5 we summarize the results that are taken:

k-means clustering	Best Total sum of distances	Average silhouette value
K=2	35.0922	0.3953
K=3	26.2149	0.4612
K=4	21.5088	0.4423
K=5	18.4307	**0.4741**

Table 4.3: Summary of k-means clustering for model 3

The silhouette plots for K=2, K=3, K=4 and K=5 are shown in following figure.

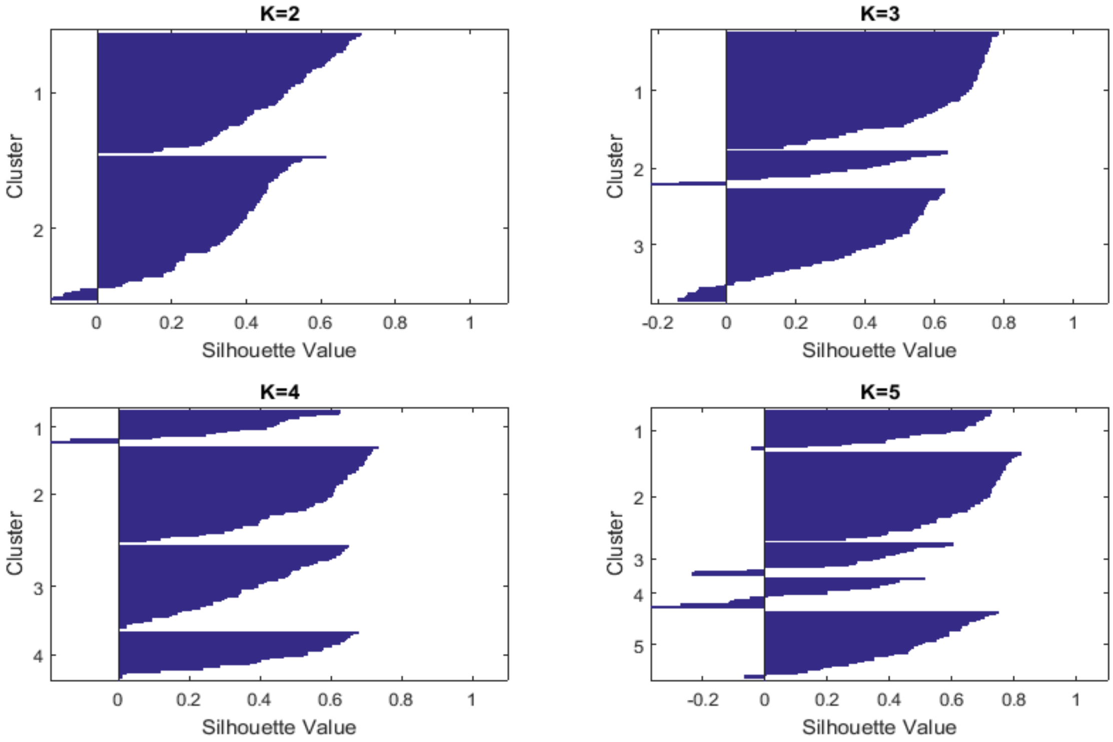


Figure 4.51: Silhouette plots for k=2, k=3, k=4 and k=5 clusters for model 3

We see that in the case of five clusters we have the mostly large silhouette values and few negative values in clusters one, three, four and five. A one-number summary in order to describe the performance of each clustering, is the average of the silhouette values. The five cluster solution has an average silhouette value of 0.4741 and this value is the maximum among the others cases. Thus it is an indicator that the grouping into five clusters using k-means is better than the one with two or three or four groups.

In figure 4.52 the plot of the first two Principal Components (PCs) of k-means clustering, for the case of k =5 for our dataset of 146 oils are presented with different color for samples members that belong to a different cluster:

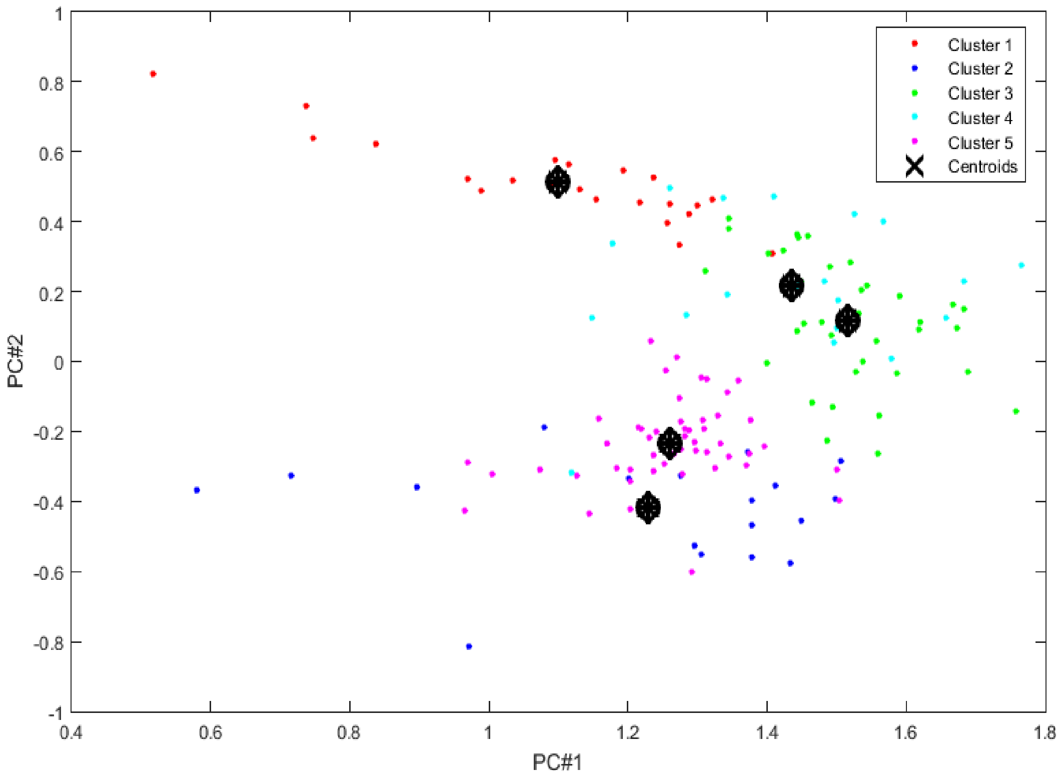


Figure 4.52: The plot of the first two PCs of k-means clustering for k=5 of model 3

4.3.4 Clustering for model 3

In order to perform clustering, we use the option clustering () at the interface of the chemometric program in the section of Clustering. If we follow the same procedure as this in model 1 and 2. Thus firstly select clustering based on samples and secondly use the Find Best choice in window dialog the following results are obtained:

```
c =

   0.6895   0.8198   0.8251   0.6905   0.7883   0.5061   0.6959

   0.7394   0.8331   0.8282   0.5510   0.8161   0.6180   0.6558

   0.6278   0.7520   0.7609   0.6770   0.7251   0.4515   0.7221

bestLinkage = 2

bestDistance = 2
```

The explanation of which is that the best metric for distance is the City Block and the best linkage method is the Average. If we choose the above options in the dialog window that is opened the following dedrogram obtained as it is shown in figure 4.53.

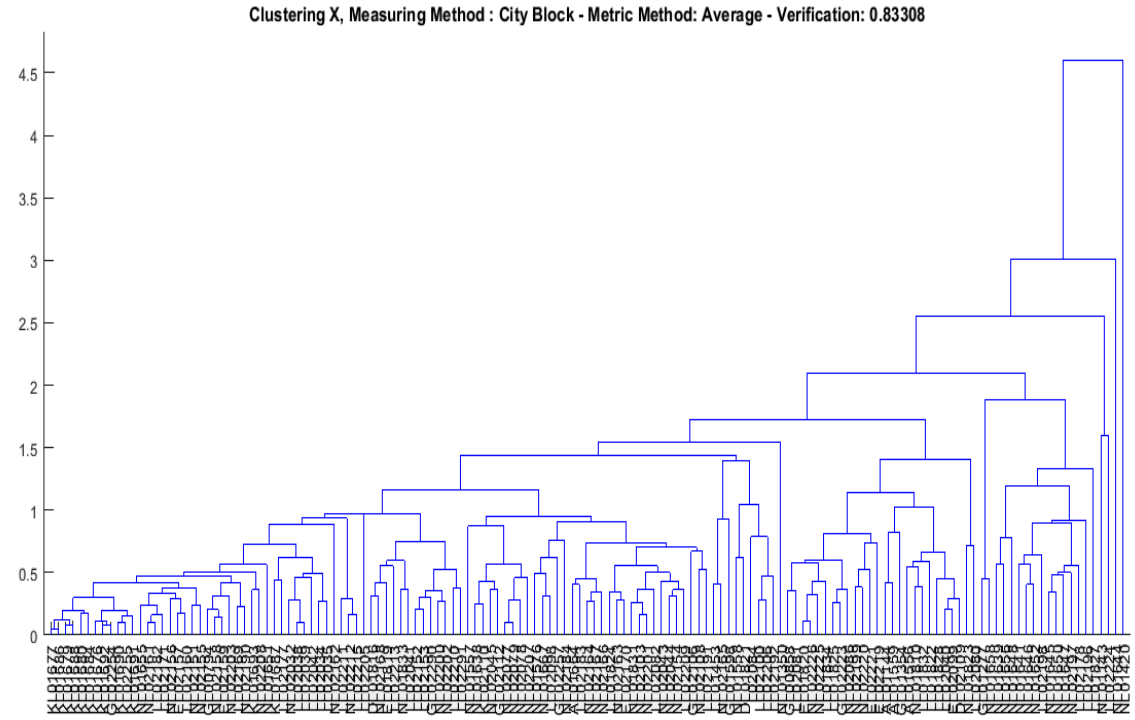


Figure 4.53: The hierarchical clustering dedrogram for model 3

The result for hierarchical clustering is presented in figure 4.50 in which the verification is 83.308 %. Clearly the result of clustering is a very big group of sample oils and secondly a big enough cluster of oil samples. It is important to mention a very small group of five oil samples that we could consider as outliers.

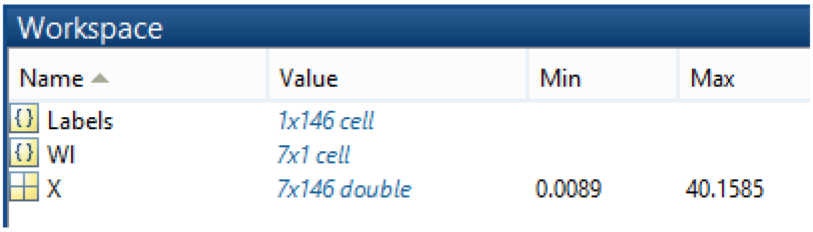
4.4 Model 4 Calculation of seven indexes from saturated fraction

The process that is followed in order to run the model 4 is very similar with the procedure steps that were followed in model 3. To be more specific, the next steps are followed for loading the data from the saturated fraction:

First of all, we must load the values of chromatographic peak areas for the saturated fraction from the Excel file that are existed in the spreadsheet with name Complete_Saturate in Excel file All_Devonian_data.xlsx.

Continuing, the next step is the calculation of seven geochemical indices that were described in chapter 3.3. This obtained with the run of Calc_sat_ratios.m MATLAB file.

In MATLAB environment and especially in the workspace section we have created the following variables: Labels cell array that contains the sample names. The Wl that is a cell array containing the variable names and finally the dataset X is a 7 X 146 matrix that has the values of seven geochemical indices that were created from the chromatographic peak areas for the saturated fraction for each variable and sample as illustrated in figure 4.54. These indexes are: Pr/Ph, Pr/nC17, Ph/nC18, CPI25-33, nC24+/nC24-, nC19/nC31, R22.



Workspace

Name	Value	Min	Max
Labels	1x146 cell		
WI	7x1 cell		
X	7x146 double	0.0089	40.1585

Figure 4.54: The workspace section with the selected variables for model 4

The next step in pretreatments is the normalization of the 7 variables that contains the saturated range indices with the MATLAB function norm_variables_0_1 () in the range of 0 to 1 for all samples.

In the final step, with the use of the available option pre_scaling_0_1 () to normalize the 146 samples in the range of 0 to 1. The result of this pretreatment is that we take for the seven indices variables, values between 0 and 1 for all samples.

4.4.1 PCA analysis for model 4

In order to perform Principal Component Analysis, we run the option PCA_analysis () in the section Analysis at the interface of the chemometric program and we take the following results.

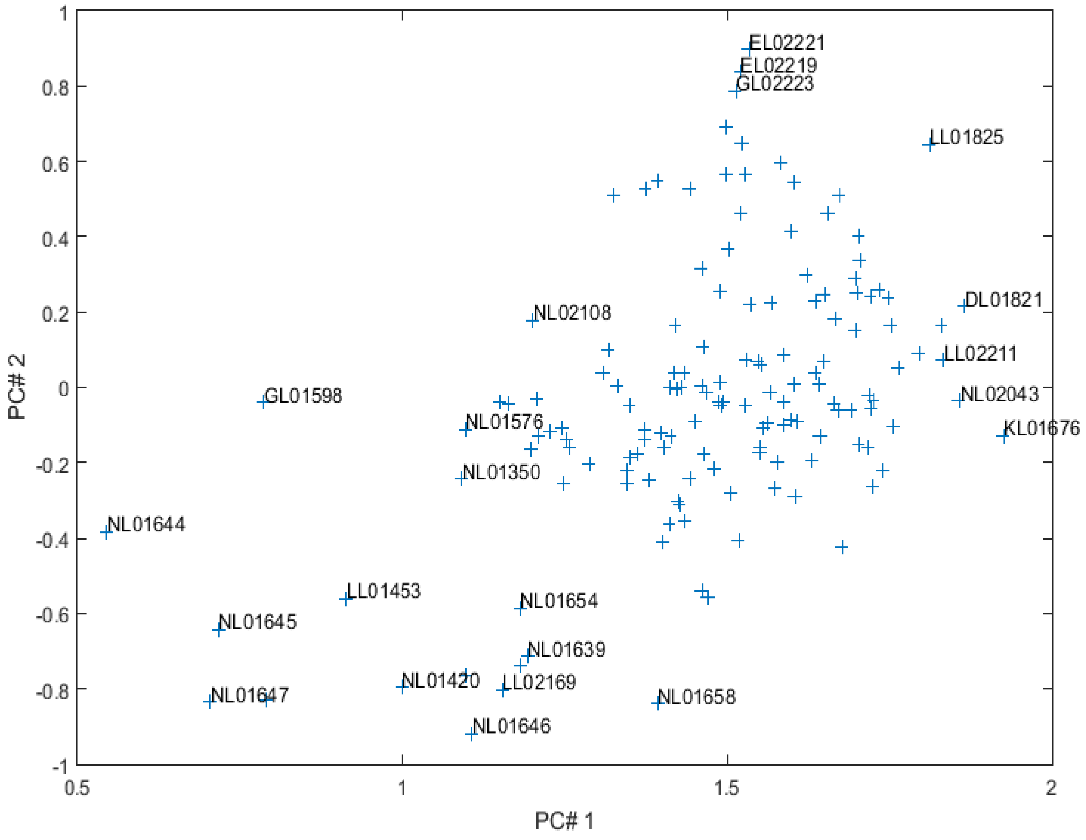


Figure 4.55: Plot of the two major principal components for model 4

In figure 4.55 is shown the plot of the first two major principal components. In this picture illustrated a clear separation for our samples in three distinguish clusters. The first cluster contains the labeled samples with values in first principal component in range of 0.5 to 1.3 and values in second principal component in range of -1 to 0.2. The second cluster contains the unlabeled samples and the third that is the final cluster consisted of the labeled samples with values in first principal component in range of 1.5 to 2 and values in second principal component in range of -0.2 to 1.

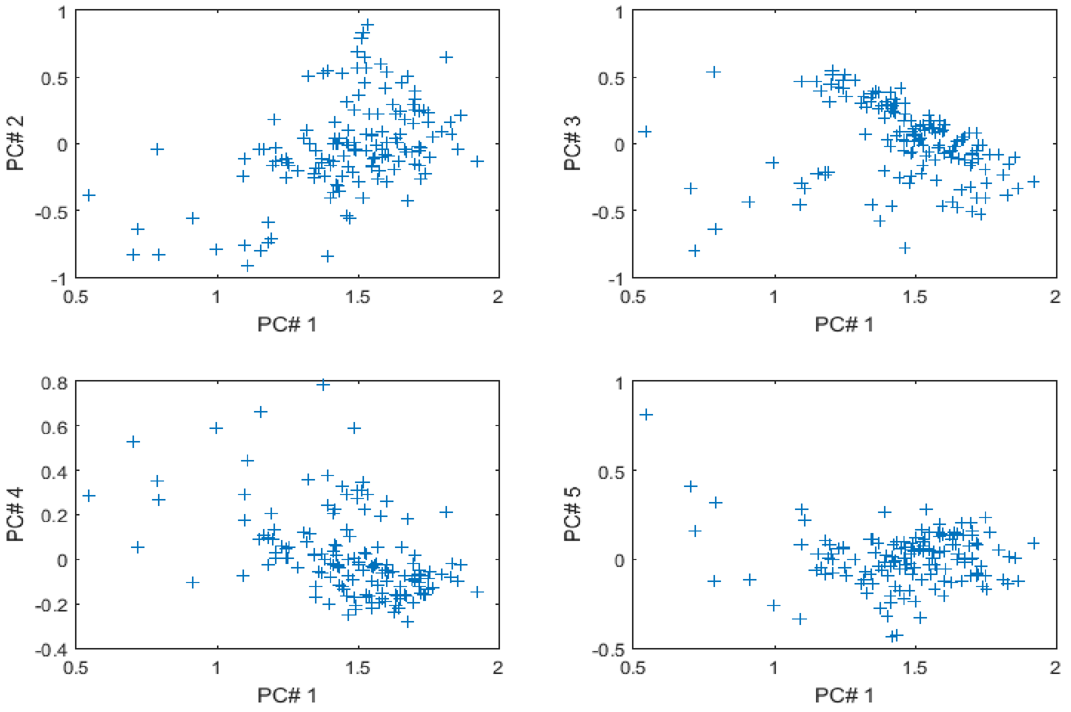


Figure 4.56: The subplot of five principal components of model 4

The figure 4.56 depicts in four instances, the first principal component versus the second, third, fourth and fifth principal component for all samples in our data. As we mention in previous figure we have a clear separation in subplot PC#1 versus PC#2 and in subplot PC#1 versus PC#3. The above clear separation is not continuing in subplots PC#1 versus PC#4 and PC#1 versus PC#5.

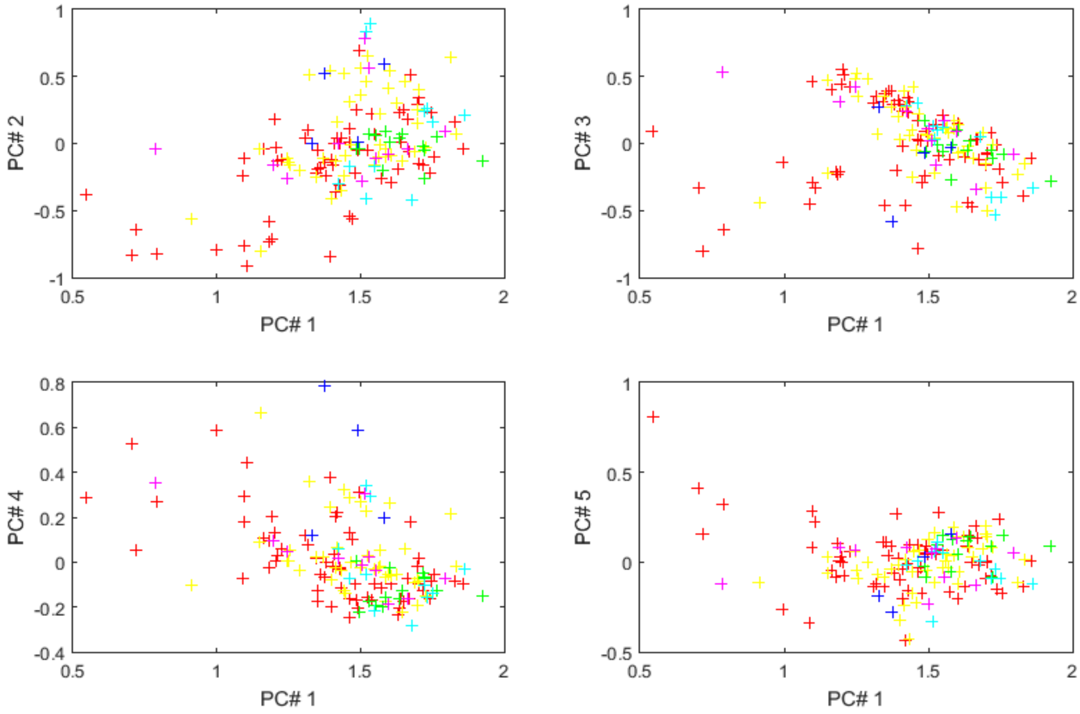


Figure 4.57: The subplot for five principal components of model 4 with different color for each formation

The figure 4.57 illustrates in four instances, the first principal component versus the second, third, fourth and fifth principal component for all our data but in this figure each color depict different formations (For mapping see table 4.1).

The plot of the first principal component in the space of longitude and latitude coordinates of our samples is shown in figure 4.58.

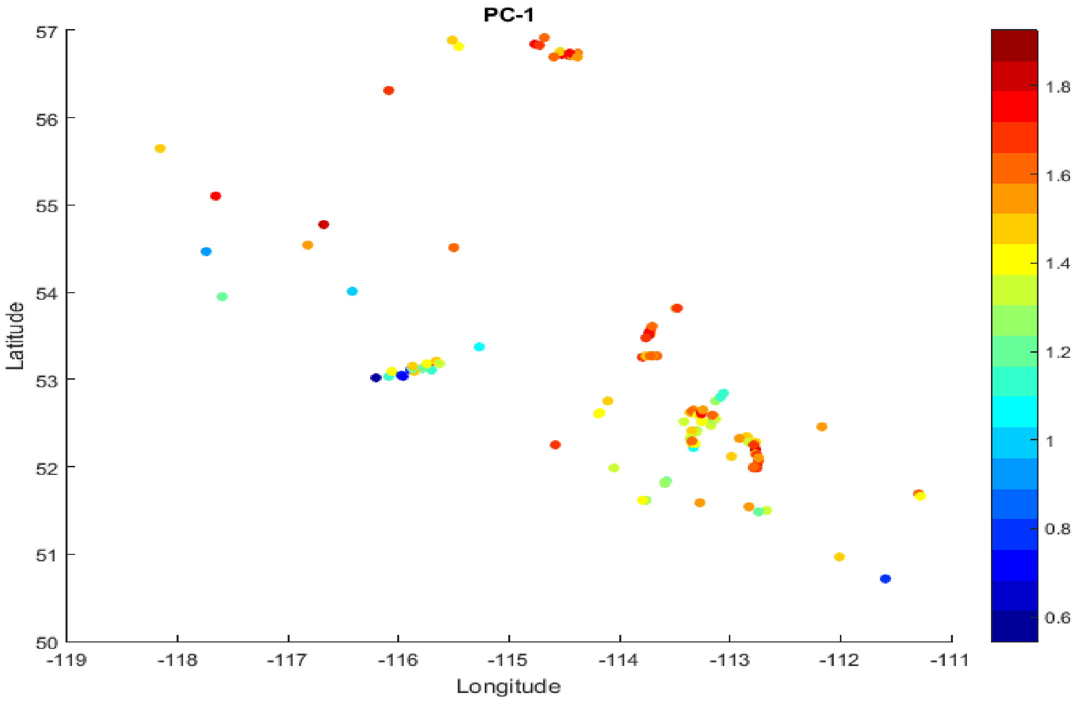


Figure 4.58: Plot of the first principal component of PCA analysis vs the geographical location of the samples.

Continuing, figure 4.59 presents the plot of the second principal component according to the latitude and longitude coordinates of our samples.

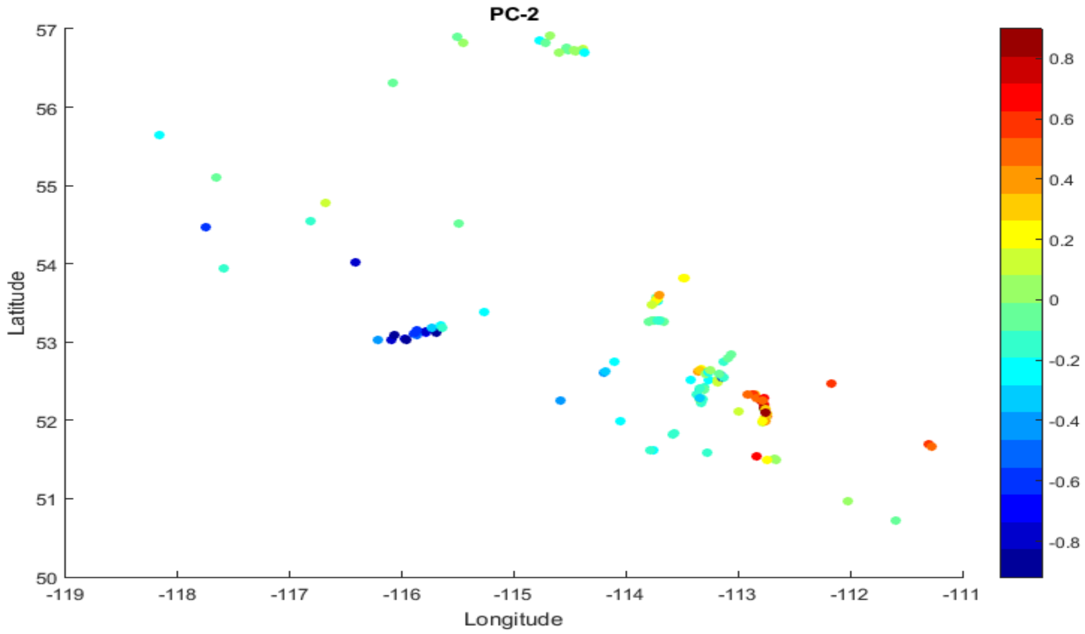


Figure 4.59: Plot of the second principal component of PCA analysis vs the location of the samples.

The Keg River samples in location coordinates of longitude and latitude have values greater than 1.4 in first principal component and below zero for the second principal component as illustrated in figures 4.58 and 4.59.

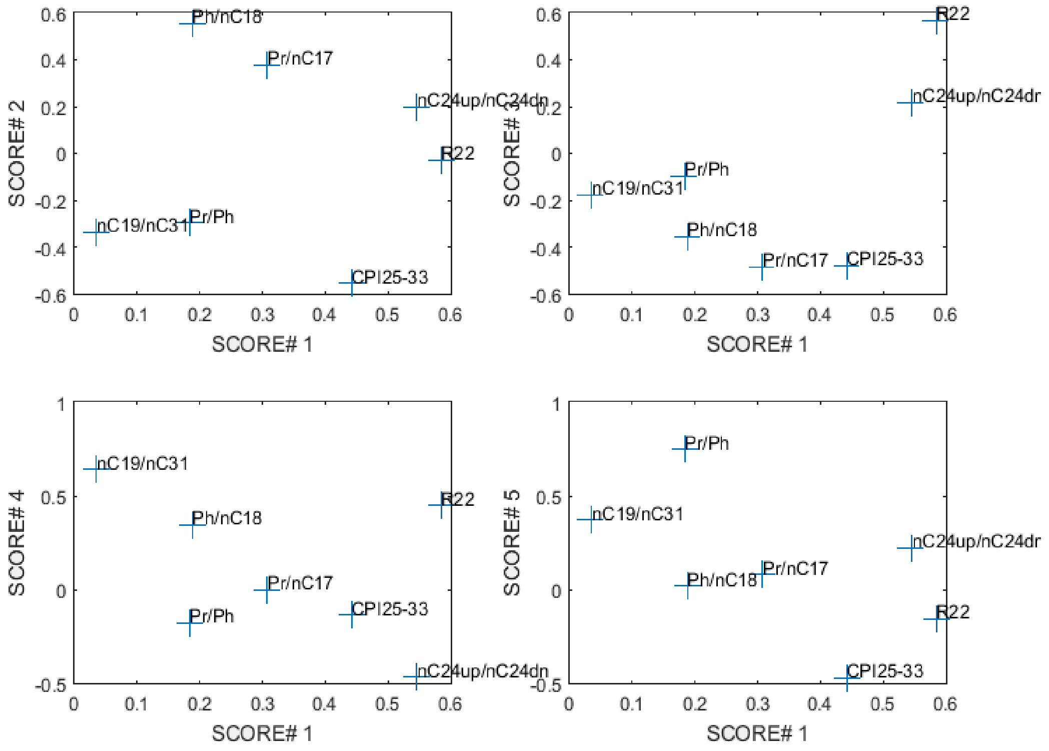


Figure 4.60: Original variable loadings for the first five principal components from the model 4

The original variable loadings are presented in figure 4.60. In this subplot we have four instances that depict the original variable loadings for the first principal component versus the original variable loadings for the second, third, fourth and fifth principal component. As it is shown the R22 index is very significant for the scores of first principal component and Ph/nC18, Pr/nC17 and CPI_25_33 indexes for the scores of second principal component.

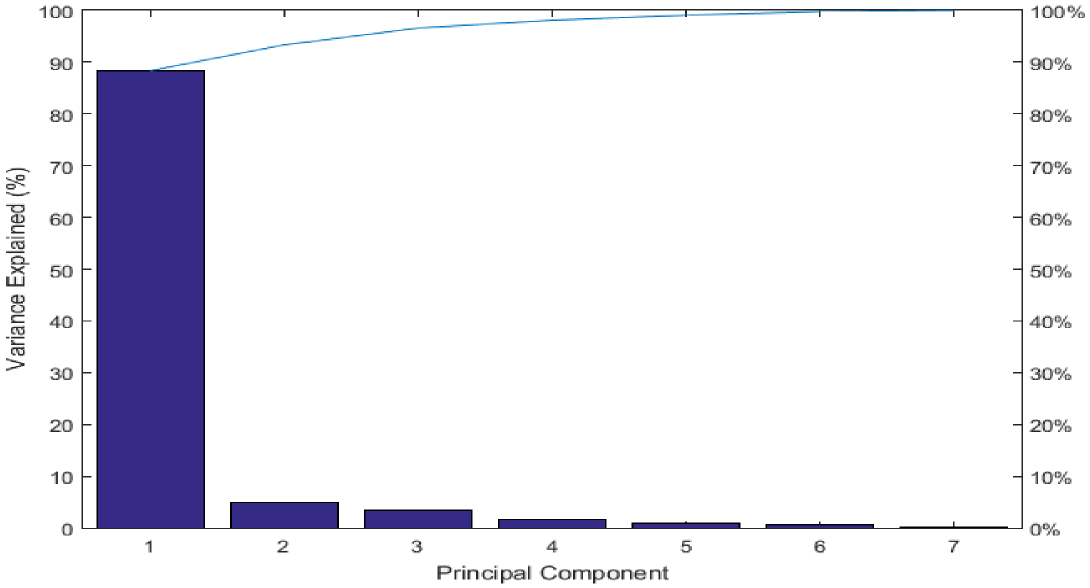


Figure 4.61: Percentage of variance explained of each principal component

Figure 4.61 reveals the contribution of each principal component in the total variance that is explained from the PCA model. As it is shown the first PC explains 88% of variation and the second PC explains 5% of the remaining variation.

4.4.2 Kernel PCA for model 4

In order to perform Kernel Principal Component Analysis, we run the option Kernel_pca_final () in the section Analysis at the interface of the chemometric program and we take the following results.

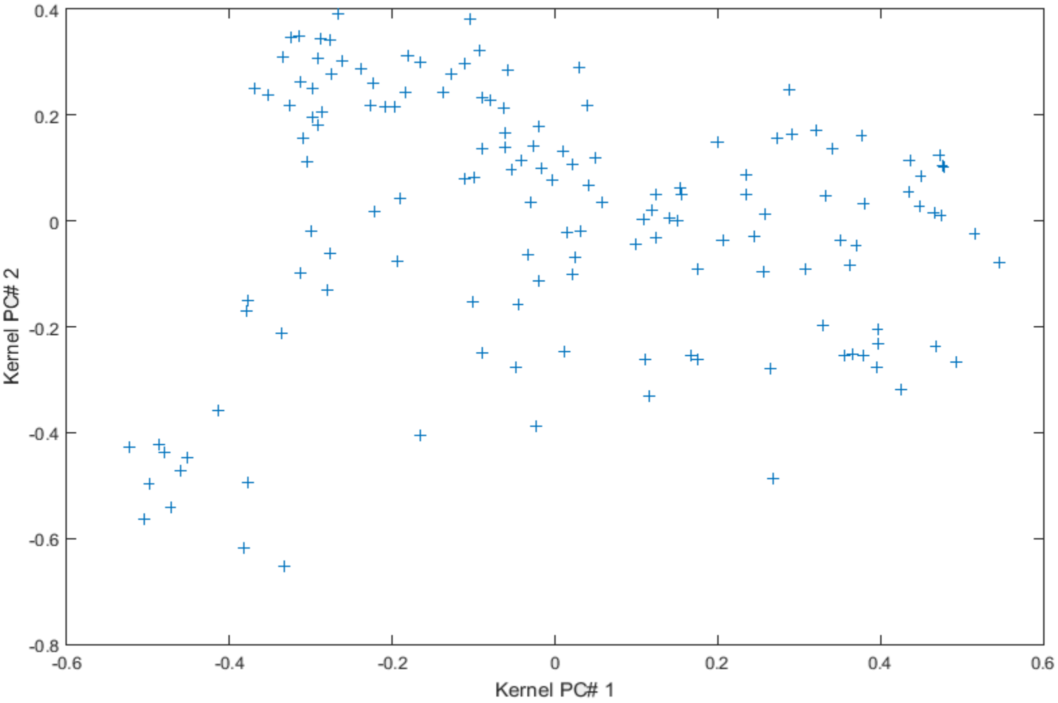


Figure 4.62: Plot of the two major kernel principal components for model 4

The figure 4.62 depicts clearly enough three different families in our dataset. For the first group the values for KPC1 are in the range of -0.6 to -0.2, and for the KPC2 are in the range of -0.8 to -0.3. For the second group the values for KPC1 are in the range of -0.3 to 0.1, and for the KPC2 are in the range of -0.4 to 0.4. For the final third group the values for KPC1 are in the range of 0.1 to 0.6, and for the KPC2 are in the range of -0.5 to 025.

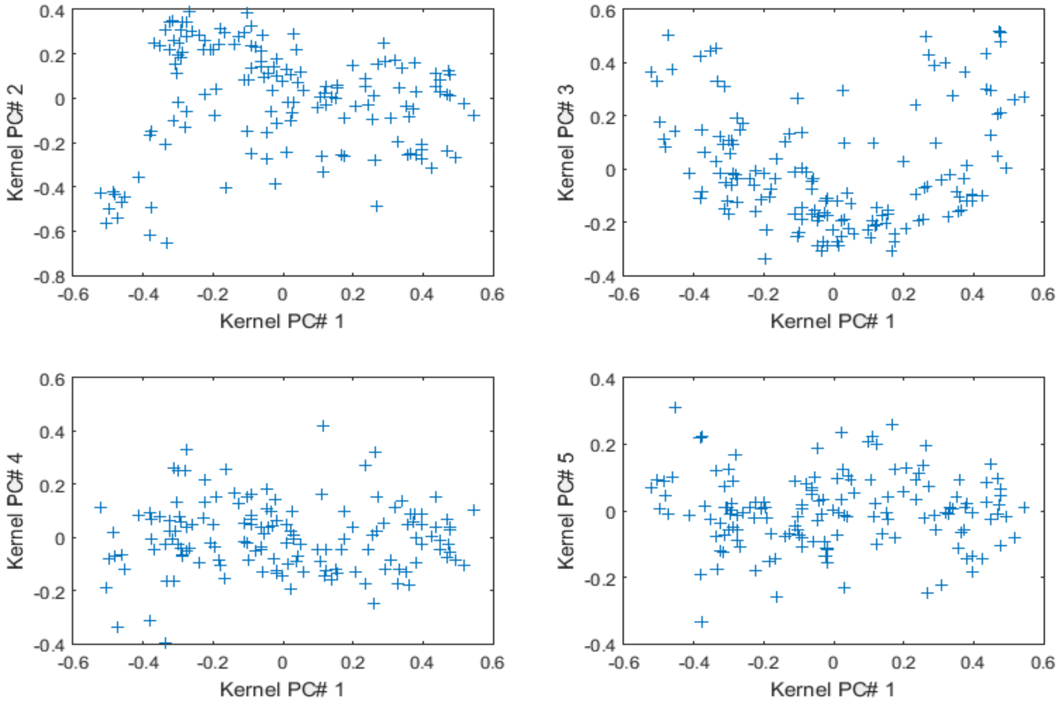


Figure 4.63: The subplot of five kernel principal components of model 4

The figure 4.63 illustrates in four instances, the first kernel principal component versus the second, third, fourth and fifth kernel principal component for all our data. Three distinguish clusters in kernel principal components are presented especially in subplot Kernel PC1 versus kernel PC2.

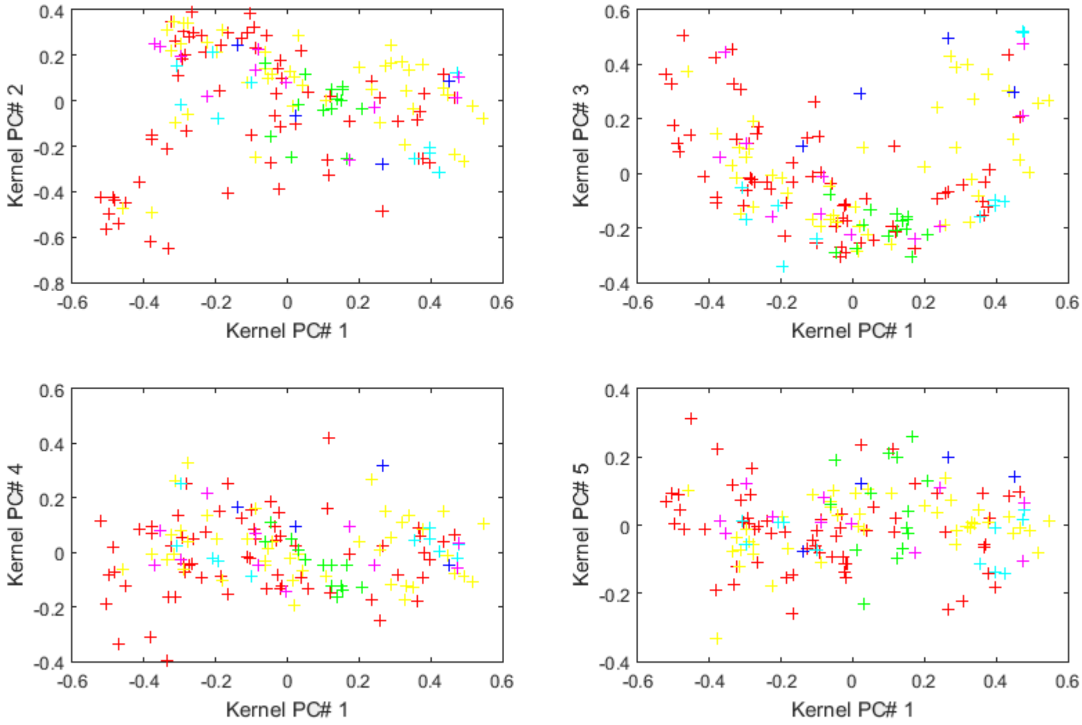


Figure 4.64: The subplot for five kernel principal components of model 4 with different color for each formation

The figure 4.64 depicts in four instances, the first kernel principal component versus the second, third, fourth and fifth kernel principal component for all our data. In addition to, in this figure each color depict different formations (For mapping see table 4.1).

The plot of the values of first kernel principal component in the space of longitude and latitude coordinates of our samples are presented in figure 4.65.

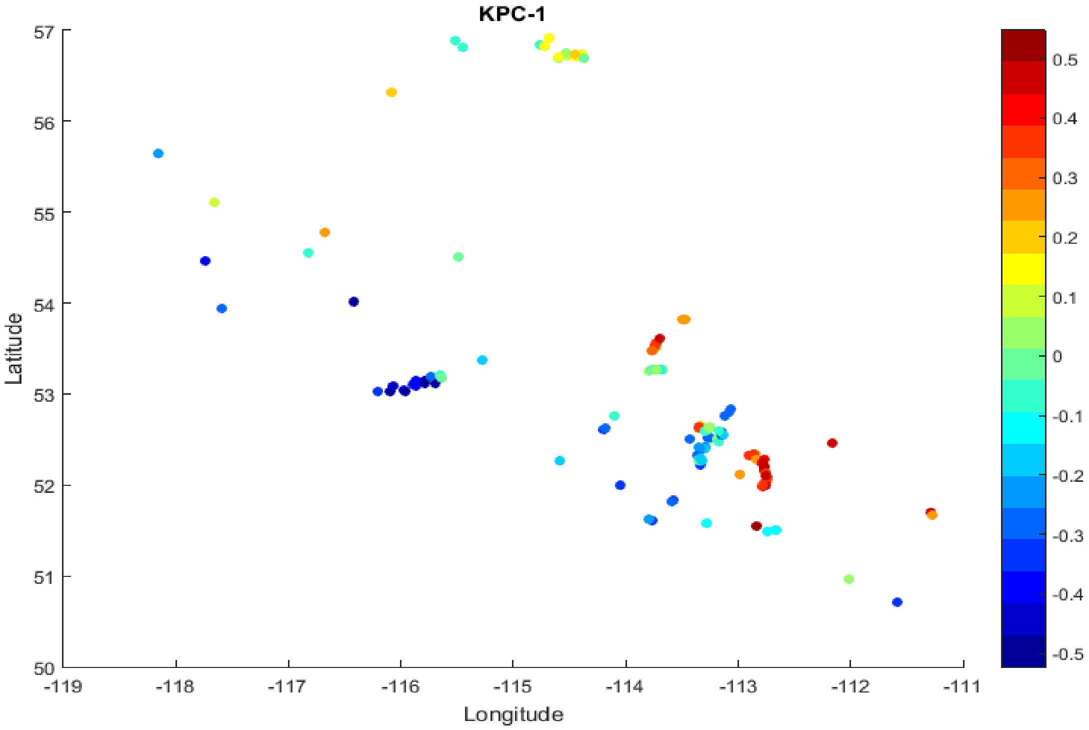


Figure 4.65: Plot of the first principal component of PCA analysis vs the geographical location of the samples.

Continuing, figure 4.66 illustrates the plot of the second kernel principal component according to the latitude and longitude coordinates of our samples.

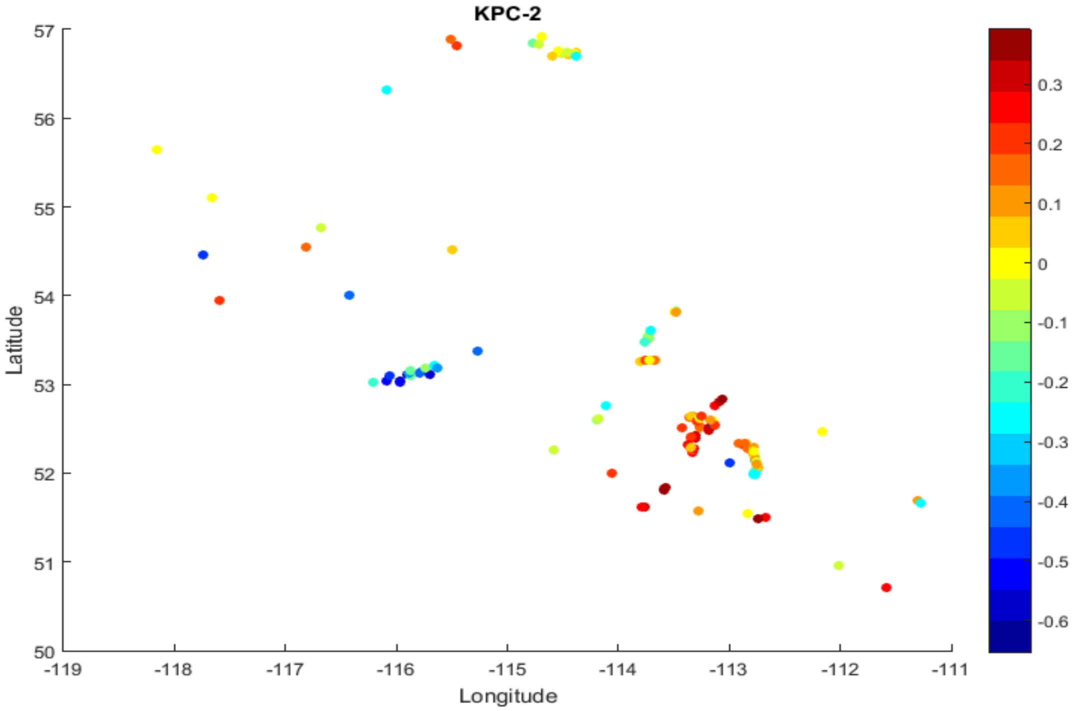


Figure 4.66: Plot of the second kernel principal component of KPCA analysis vs the geographic location of the samples

The Keg River samples in location coordinates of longitude and latitude have near zero values in first and second principal components according to figures 4.65 and 4.66.

4.4.3 Silhouette – k-means for model 4

With the use of the available option Silhouete () at the interface of the chemometric program in the section of clustering, we get a k-means clustering using two, three, four and five clusters. The k-means clustering is repeating five times for each case. This is achieved with the use of replicates as an argument in kmeans MATLAB function.

In table 4.4 we summarize the results that are taken:

k-means clustering	Best Total sum of distances	Average silhouette value
K=2	36.9291	0.3903
K=3	26.7421	**0.4915**
K=4	20.9558	0.4851
K=5	18.2235	0.4190

Table 4.4: Summary of k-means clustering for model 4

The silhouette plots for K=2, K=3, K=4 and K=5 are shown in following figure.

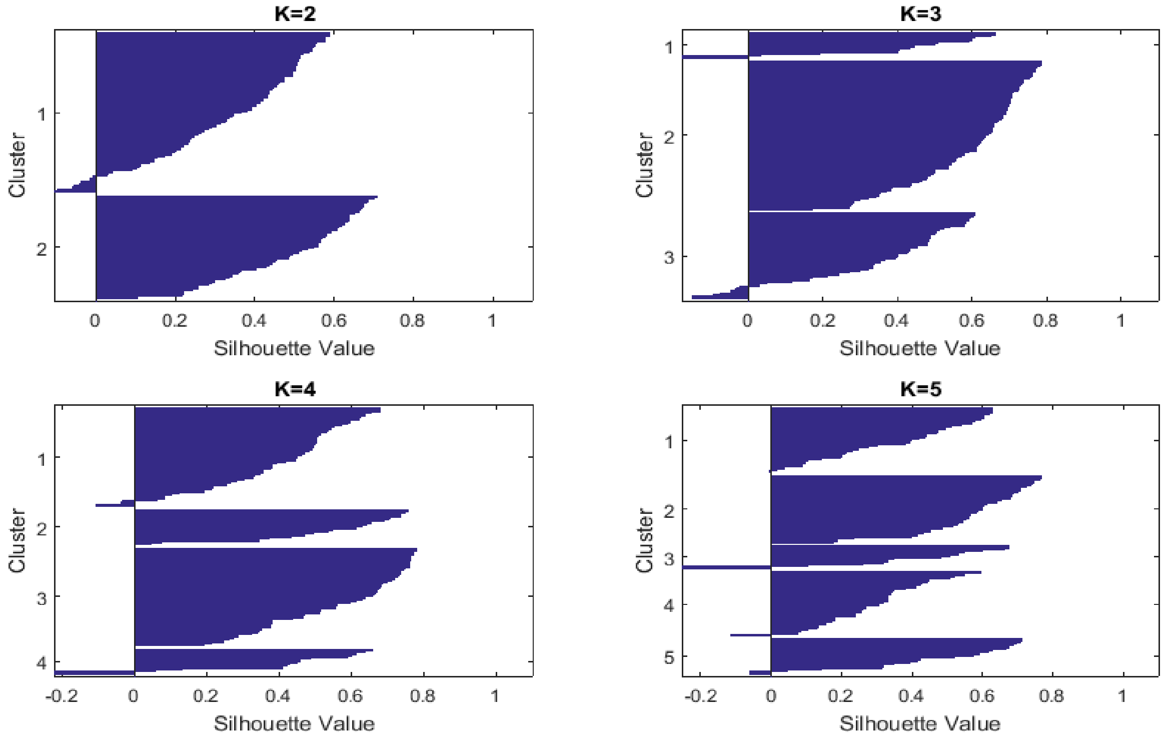


Figure 4.67: Silhouette plots for k=2, k=3, k=4 and k=5 clusters for model 4

We see that in the case of three clusters we have the mostly large silhouette values and few negative values in clusters one and three. A one-number summary in order to describe the performance of each clustering, is the average of the silhouette values. The three cluster solution has an average silhouette value of 0.4915 and this value is the maximum among the others cases. Thus it is an indicator that the grouping into three clusters using k-means is better than the one with two or four or five groups.

In figure 4.68 the plot of the first two Principal Components (PCs) for k-means clustering, for the case of k =3 for our dataset of 146 oils are presented with different color for samples members that belongs to a different cluster:

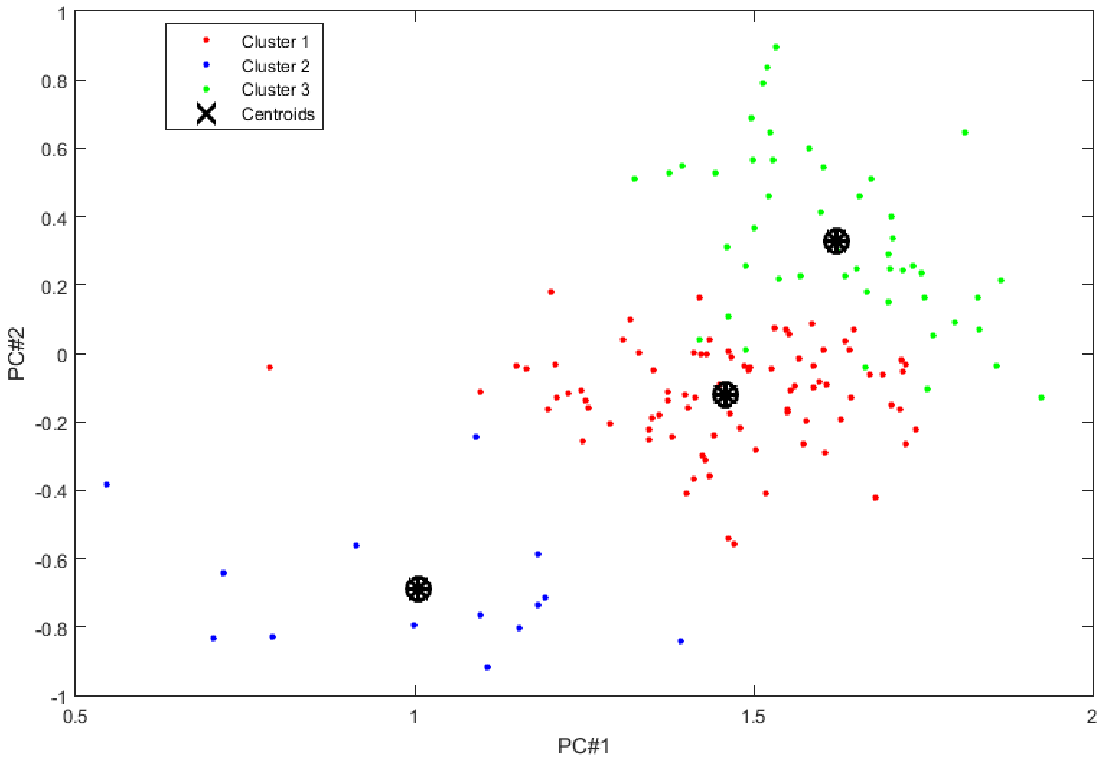


Figure 4.68: The plot of the first two PCs of k-means clustering for k=3 of model 4

4.4.4 Clustering for model 4

In order to perform clustering, we use the option clustering () at the interface of the chemometric program in the section of Clustering. If we follow the same procedure as this in model 3. Thus firstly select clustering based on samples and secondly use the Find Best choice in window dialog the following results are obtained:

```
c =

  0.6880   0.8101   0.8103   0.7088   0.7855   0.6457   0.7724

  0.7085   0.7992   0.8101   0.6916   0.7872   0.6847   0.7393

  0.6149   0.7907   0.7899   0.6607   0.7981   0.7687   0.7091


bestLinkage = 3

bestDistance = 1
```

The explanation of which is that the best metric for distance is the Euclidean and the best linkage method is the Centroid. If we choose the above options in the dialog window that is opened the following dedrogram obtained as it is shown in figure 4.69.

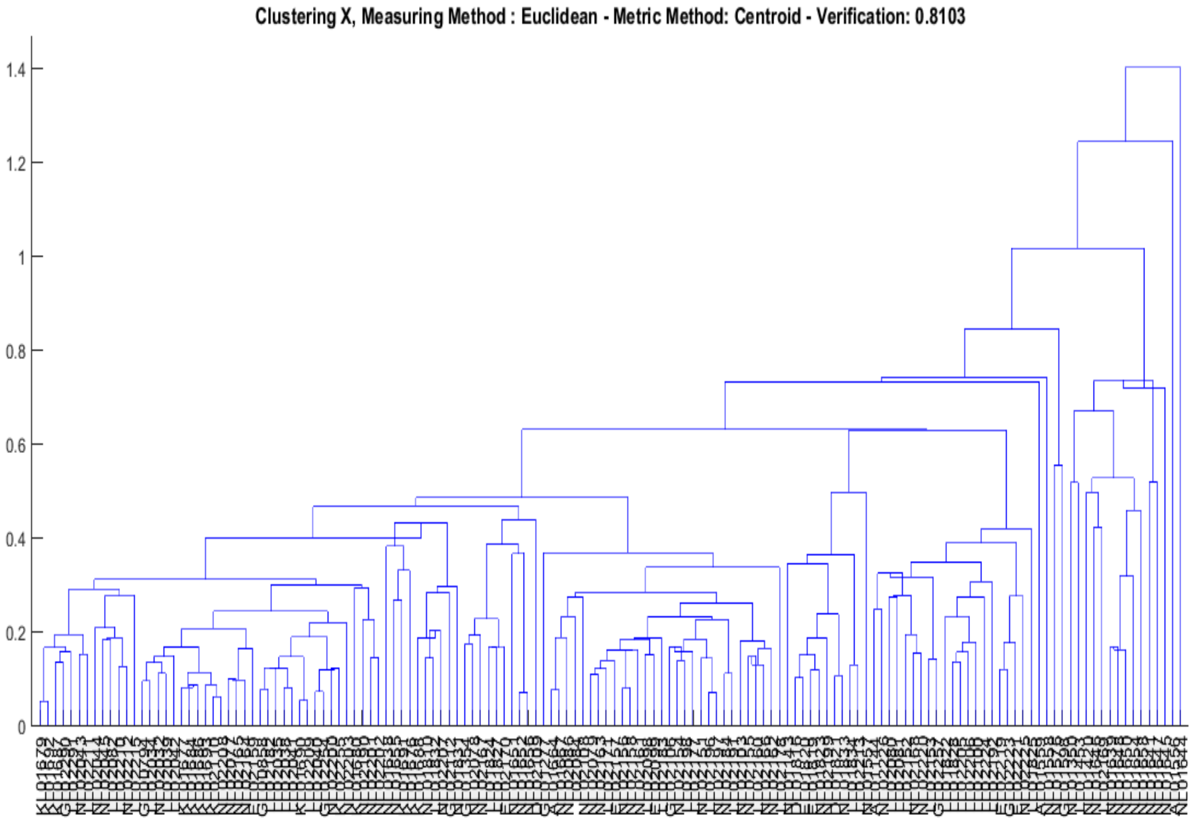


Figure 4.69: The hierarchical clustering dedrogram for model 4

The result for hierarchical clustering is presented in figure 4.69 in which the verification is 81.03 %. Three separate groups of sample oils are illustrated. The first group contains the majority of samples with the second one taking almost the remaining samples. But fifteen samples have significant different behavior and are cluster in a new third group.

4.5 Summary of the results

Model 1 uses the peak areas of the gasoline range chromatograms. Principal Component Analysis does not show a clear separation between the oil samples. Some samples have a very different behavior in contrast to the majority of the remaining. These samples are the following: LL02224, NL02162, LL02177, EL01820, NL02080, LL01828 and AL01556. The results of kernel principal component analysis are very close to the results of the linear principal component analysis. The k-means clustering reveals with higher possibility the existence of two clusters. Finally hierarchical clustering dedrogram indicates the existence of four major clusters.

Model 2 uses the peak areas of the saturated fraction chromatograms. As in model 1 PCA does not provide a clear separation between oil samples, except for the major deviations in the following Nisku samples: NL01641, NL01645, NL01644, NL1647 and NL01655. Kernel PCA technique reveals a significant different behavior for the following ten samples: AL01556, EL02219, AL01144, NL01143, NL02224, KL01680, NL01641, NL01645, NL01644 and NL01420. K-means show two separate groups of oils. Similarly the hierarchical clustering depicts two major clusters of oil samples..

Model 3 is based on nine geochemical indices from gasoline range components. The Principal Component Analysis provides sufficient separation between the samples into two clusters. The following samples are characterized as ‘extremes’ according to PCA model: NL01823, LL02080, LL01821, LL01827, GL01277, NL01658 and NL01420. The R index mainly distinguishes the samples along the first principal component (PC1) while U and H indices are important for the second principal component (PC2). Hierarchical clustering depicts clearly a big group of oils, while a second group is also shown together with several outliers.

Finally, model 4 is based on seven geochemical indices from the saturated fraction components. The Principal Component Analysis technique gives three distinct clusters. The R22 index mainly distinguish the samples along the first principal component (PC1) while the ratios Ph/nC18, Pr/nC17 and CPI_25_33 index are important for the second principal component (PC2). The kernel PCA identifies three different groups of oils. This result is compatible with the result of hierarchical clustering.

4.6 Conclusion

The models 1 and 2 did not provide a clear classification of oils. Principal Component Analysis in model 3 reveals two groups of oils. Hierarchical clustering gives a compatible result with PCA. The combination of silhouette statistic with k-means and the first two principal components of PCA from models 3 and 4 reveal clearly different behavior of the oil samples. This is a strong indicator of the existence of different oil families. The above combination of methods is a very promising and powerful tool for affiliation of oil's families. The kernel PCA did not provide a different classification pattern compared to the conventional linear PCA.

The gasoline range and the saturated fraction hydrocarbons carry significant geochemical information. The process of decoding it in our case found to be a difficult task possibly due to the significant compositional similarity of the oils. Further work should be carried out, including a more detailed analysis of compositional variations within each subgroup of the studied data set. Additionally the findings of this work have to be reevaluated taking into account the geochemical meaning of the compositional variables, used in the models.

“Before you begin a thing, remind yourself that difficulties
and delays quite impossible to foresee are ahead. If you
could see them clearly, naturally you could do a
great deal to get rid of them but you can’t.
You can only see one thing clearly and
that is your goal. Form a mental
vision of that and cling to it
through thick and thin.”
Kathleen Norris

5. REFERENCES

Brereton, R.G. Chemometrics for Pattern Recognition, 2009 John Wiley and Sons.

Flower, M.G., Stasiuk, L.D., Hearn, M. and Obermajer, M. (2001): Devonian hydrocarbons source rocks and their derived oils in the Western Canada Sedimentary Basin, Bulletin of Canadian Petroleum Geology, v49, p117-148.

Maowen, L., Flower, M.G., Obermajer, M., Stasiuk, L.D., and Snowdon, L.R. (1999): Geochemical characterization of middle Devonian oils in northwestern Alberta, Canada: possible source and maturity effect on pyrrolic nitrogen compounds, Organic geochemistry, v30, p1039-1057.

Martinez W.L., Martinez A.R, Solka J.L., Exploratory Data Analysis with MATLAB, Second edition 2011, CRC Press.

Mort, A.J. and Sanei, H. (2013): Investigating laboratory general pyrobitumen precursors for unconventional reservoir characterization: a geochemical and petrographic approach, GeoConvention 2013: Integration.

Mort, A.J., Stevens, L. and Wierzbicki, R. (2015): Devonian petroleum systems and exploration potential in southern Alberta, Part 3 Core conference, GeoConvention 2015: New horizons.

Obermajer, M., Osadetz, K.G., and Pasadakis, N. (2003): Refining compositional affinity of Williston basin family C oils using multivariate statistical analysis of saturate biomarker: in Summary of Investigations, Volume 1, Saskatchewan Geological Survey, p1-10.

Osadetz, K.G., Pasadakis, N., and Obermajer M. (2002): Definition and characterization of petroleum composition families using principal component analysis of gasoline and saturate fraction compositional ratios: in Summary of Investigations 2002, Volume 1, Saskatchewan Geological Survey, p3-14.

Rencher A.C., Methods of multivariate analysis, Second edition 2009, John Wiley and Sons.

Pasadakis N., Petroleum geochemistry, First edition 2014, Tziolas publications

Pasadakis, N., Obermajer, M. and Osadetz, K. (2004): Definition and characterization of petroleum compositional families in Williston basin, North America using principal component analysis, Organic Geochemistry, v35, p453-468.

Thompson, K.F.M. (1983): Classification and thermal history of petroleum based on light hydrocarbons; Geochim. Cosmochim. Acta, v47, p303-316.

APPENDIX

The calculated ratios that based on the concentrations of selected compounds from gasoline range hydrocarbons are presented in the next table. These indices are used in the analysis of the model 3.

	Ratios for model 3								
Sample	**K1**	**A**	**B**	**C**	**I**	**F**	**R**	**U**	**H**
GL00794	0.9277	0.1873	0.1888	1.1415	0.4385	0.9824	4.6655	0.2247	24.5151
GL00858	0.9618	0.1248	0.2967	1.5964	0.5639	1.5015	4.5238	0.6139	27.3379
NL01143	0.8166	0.8495	0.8917	0.9986	0.4015	1.0043	4.1135	0.3143	21.1871
AL01144	0.9346	0.0000	0.0617	0.9886	0.6732	0.7150	2.1203	0.2947	16.8838
GL01277	1.0467	0.0353	0.0937	0.5990	1.9616	0.5081	2.0336	0.1724	20.5261
NL01350	0.8808	0.0653	0.0737	0.1997	0.3102	0.1962	1.9811	0.1446	7.1890
GL01354	0.8748	0.2928	0.1283	0.5104	0.6499	0.4561	2.1296	0.2421	14.4956
NL01420	1.3926	0.2030	0.0082	19.2136	2.0109	17.0156	2.5558	0.0619	46.9568
LL01453	0.9951	0.2447	0.5266	2.2165	0.8231	1.7918	4.2334	0.2106	35.4217
AL01556	0.8936	0.0006	0.0333	1.6551	0.7571	0.9190	1.9669	0.4237	18.3902
NL01557	0.7802	0.0717	0.0754	1.9335	0.3001	1.3142	3.4773	0.3413	18.7973
NL01558	0.7942	0.5710	0.6999	1.5694	0.4992	1.4492	3.8659	0.4786	24.9728
AL01559	0.9434	0.0176	0.1883	1.6875	0.6771	1.1579	2.6456	0.4216	22.6051
NL01576	0.9935	0.1619	0.0891	1.7742	0.8269	1.4496	3.4794	0.3263	29.9297
GL01598	0.9251	0.1829	0.5180	2.8888	1.2184	2.5240	4.2237	0.3049	40.4153
NL01638	0.8893	0.0828	0.0385	0.9341	0.4717	0.6494	2.4292	0.2569	16.3913
NL01639	1.0944	0.0950	0.0806	1.3575	1.4110	0.9381	2.2129	0.2155	25.7441
NL01641	1.2903	0.1189	0.1798	3.0002	1.2671	2.2105	3.0652	0.2629	37.6798
NL01644	2.5784	0.0553	0.6343	0.9678	1.6166	0.6323	2.6573	0.1715	26.2942
NL01645	1.0242	0.1991	0.2851	1.8949	1.3383	1.5138	4.0593	0.1815	37.2240
NL01646	1.1419	0.1142	0.1326	2.0427	1.2749	1.4642	2.7581	0.2705	31.6895
NL01647	1.1426	0.2026	0.3760	2.1830	1.4107	1.8363	3.4410	0.1835	37.7313
NL01648	1.0311	0.2369	0.0855	1.5003	1.5705	0.9995	2.1824	0.2677	25.5599
NL01650	1.0307	0.1552	0.3054	1.9386	1.0877	1.5228	3.7227	0.1907	34.9181
NL01651	0.8626	0.1599	0.2092	1.8657	0.6227	1.4836	4.3717	0.2322	30.9095
NL01652	1.0486	0.1032	0.1008	1.2139	1.1388	0.8903	2.5116	0.1932	25.5423
NL01655	0.9098	0.1288	0.1229	0.6827	0.5168	0.5954	3.1266	0.1710	18.6641
NL01656	1.0183	0.1020	0.0806	0.9482	1.1823	0.7240	2.3858	0.1476	23.4057
NL01658	1.1290	0.1422	0.2294	0.5583	2.3333	0.4741	1.8909	0.1373	20.0562
AL01664	0.9973	0.2866	0.1016	1.7112	0.9503	1.6383	5.1102	0.3780	36.3595
NL01667	1.0195	0.1418	0.0731	1.4855	0.8856	1.2786	4.0275	0.4476	30.1105
KL01676	0.9130	0.0697	0.0925	1.1878	0.4676	0.9626	3.8920	0.1980	23.0483
KL01677	0.9206	0.0836	0.0833	1.1830	0.4933	0.9576	3.5923	0.1752	22.7223
KL01679	0.9175	0.0621	0.0669	1.2639	0.4972	1.0042	3.6839	0.1848	23.3728
KL01680	0.9236	0.0724	0.0648	1.2816	0.5147	0.9973	3.4562	0.1890	22.9452
KL01684	0.9155	0.0485	0.0505	1.2756	0.4838	0.9879	3.6347	0.2088	22.7619
KL01686	0.9253	0.0787	0.0827	1.1329	0.4935	0.9443	3.5844	0.1656	22.6501
KL01687	0.9473	0.1211	0.2322	1.0868	0.5631	0.8933	3.4813	0.1727	23.2499
KL01688	0.9267	0.0652	0.0682	1.1800	0.4979	0.9436	3.4592	0.1775	22.3311

KL01690	0.9175	0.0933	0.1005	1.1504	0.5023	0.9783	3.8449	0.1526	23.9104
KL01691	0.9233	0.1440	0.1035	1.0805	0.4890	0.9090	3.5884	0.1716	22.3160
KL01692	0.9117	0.1331	0.0969	1.1186	0.4595	0.9002	3.5926	0.1810	21.6907
KL01693	0.9179	0.1611	0.1106	1.1062	0.5140	0.9829	3.8796	0.1217	24.4104
NL01810	0.8859	0.0375	0.0223	0.5534	0.3992	0.3450	1.4883	0.2197	9.1400
EL01816	0.8556	0.4031	0.3221	1.2678	0.4042	1.0542	3.5712	0.3076	20.5812
EL01819	0.8370	0.3528	0.2512	1.5537	0.3472	1.2238	3.5480	0.2990	19.8980
EL01820	0.8345	0.3579	0.2695	1.2565	0.4306	0.9646	2.9628	0.4544	18.0999
EL01821	0.8342	0.6041	0.4168	1.2553	0.5032	1.0946	3.1691	0.3069	21.2610
LL01822	0.9173	0.1138	0.0488	0.6416	0.6549	0.4382	1.5056	0.3894	11.7847
NL01823	0.8931	0.0716	0.0507	0.4380	0.3267	0.3016	1.0713	0.3127	6.1917
LL01824	0.9753	0.1209	0.1659	1.3534	0.6049	0.9825	3.4840	0.3104	24.1326
LL01825	0.9001	0.3376	0.2526	1.1907	0.6681	0.9206	2.8366	0.3949	21.0841
LL01827	0.9652	0.0390	0.0485	0.6365	0.6454	0.4714	2.1852	0.1732	15.8352
LL01828	0.8636	0.1755	0.0837	1.3961	0.5476	0.9379	2.1471	0.3554	17.2983
NL01831	0.9053	0.2186	0.1558	0.9212	0.4618	0.7051	2.9167	0.2595	17.9687
LL01832	0.9736	0.0346	0.0215	0.9577	0.5415	0.5581	1.8451	0.3246	14.0378
NL01833	0.8819	0.3914	0.2939	1.3837	0.5059	1.1737	3.6003	0.2806	23.7788
LL01834	0.8322	0.3396	0.2445	1.3544	0.3355	0.9684	2.9650	0.3769	16.3648
NL02032	0.9823	0.2832	0.3266	0.9820	0.6098	0.8525	4.1238	0.1995	25.4050
LL02034	0.9281	0.3874	0.3719	0.9200	0.4925	0.8631	4.8087	0.1499	25.5189
LL02035	0.9706	0.4235	0.3511	1.1818	0.5707	1.0431	4.4071	0.1374	27.7110
LL02038	0.9540	0.3103	0.3682	1.1674	0.5172	1.0062	4.1904	0.2172	25.4704
LL02039	0.9478	0.3306	0.3691	1.1647	0.4964	0.9630	4.0188	0.2199	24.2291
LL02040	0.9440	0.0755	0.0557	1.0294	0.4559	0.7036	2.6601	0.3778	16.8972
LL02041	0.9385	0.2710	0.4030	1.1807	0.4967	0.9769	3.8578	0.3041	23.5760
LL02042	0.9536	0.3644	0.5666	1.0544	0.5024	0.9283	4.5727	0.2044	25.5186
NL02043	0.9105	0.0443	0.1909	1.1253	0.4543	0.8553	3.5651	0.2779	20.7915
NL02044	0.9150	0.0230	0.0936	1.0851	0.5035	0.8484	3.6040	0.2602	21.7083
NL02045	0.9376	0.0357	0.0890	1.2885	0.5968	1.0000	2.9898	0.3435	22.1000
NL02077	0.9729	0.2545	0.1872	1.5388	0.5339	1.1622	3.7203	0.4220	24.2481
NL02078	0.9624	0.2419	0.1429	1.2926	0.5060	1.0356	3.8353	0.3767	23.4572
NL02079	0.9676	0.2167	0.1459	1.3419	0.5173	1.0281	3.5924	0.3978	22.8150
LL02080	0.8295	0.3440	0.1204	0.3885	0.4012	0.3595	1.6077	0.3423	8.8982
LL02081	0.8444	0.2826	0.1949	1.0206	0.4367	0.7818	3.0334	0.3508	17.4250
LL02082	0.8521	0.2873	0.1082	0.9810	0.5102	0.7743	2.8064	0.2393	18.1940
LL02084	1.0004	0.4538	0.3537	1.5778	0.7287	1.3221	3.3589	0.3713	27.0558
NL02086	1.0105	0.3806	0.2816	1.7375	0.7024	1.2964	3.0980	0.4854	25.1188
LL02098	1.0080	0.1169	0.0859	1.5643	0.7211	1.1485	3.4974	0.3584	26.2177
EL02099	0.9821	0.1030	0.0507	0.8642	0.4810	0.6060	2.4081	0.3472	15.3781
LL02100	0.8402	0.3850	0.2660	1.4481	0.6909	1.2893	3.8091	0.3650	26.4915
NL02103	0.9621	0.2875	0.1890	1.3816	0.5615	1.1549	3.9848	0.3389	25.6598
GL02106	1.0918	0.2463	0.2436	1.2432	1.1146	1.1426	4.5529	0.4499	30.9909
NL02108	1.0337	0.1454	0.3739	1.2521	0.8397	1.1797	4.8243	0.4115	30.4813
EL02109	1.0015	0.1904	0.1233	1.1218	0.6582	0.7485	2.9183	0.4423	19.9378
KL02110	0.9259	0.0564	0.0367	0.9502	0.4360	0.6806	2.6553	0.2938	16.2643

GL02112	0.9134	0.0667	0.0597	0.8122	0.5339	0.5558	2.2081	0.2428	15.3336
NL02151	1.0098	0.2571	0.1903	1.2042	0.8116	1.0743	4.2865	0.2535	29.2643
LL02152	0.8611	0.3171	0.2003	0.9131	0.3893	0.8401	4.1054	0.2746	20.3008
LL02153	0.8721	0.3216	0.2489	0.9804	0.4156	0.8795	3.9370	0.2543	20.8577
NL02154	0.9825	0.2065	0.1573	1.3650	0.5798	1.0997	3.9103	0.3262	25.5029
NL02155	0.9862	0.2376	0.1685	1.3302	0.6376	1.1453	4.1377	0.2571	27.6334
NL02156	0.9624	0.1624	0.0817	0.9170	0.5457	0.7757	3.4982	0.2309	21.5622
EL02157	0.9576	0.2636	0.1927	1.2205	0.5158	1.0521	4.3084	0.2826	25.3259
NL02158	0.9618	0.2790	0.2203	1.1796	0.5338	1.0402	4.6407	0.2330	26.6091
EL02159	0.9450	0.3140	0.1900	1.1234	0.5168	1.0084	4.6151	0.2079	26.1436
LL02160	0.9501	0.1878	0.1165	0.9374	0.5112	0.8005	3.6890	0.2446	21.6332
NL02161	0.9581	0.2412	0.1597	1.1858	0.5384	1.0206	4.0826	0.2290	25.2343
NL02162	0.9784	0.2879	0.1929	1.3404	0.8200	1.2548	4.1406	0.1245	31.5836
NL02163	0.9721	0.1652	0.0886	1.1776	0.6723	0.9741	3.5719	0.2940	24.7763
NL02164	0.9850	0.1802	0.1245	1.2333	0.7268	1.0826	3.6602	0.2368	26.9491
NL02165	0.9842	0.2376	0.4437	1.2238	0.7303	1.0880	3.8570	0.2063	27.7400
NL02166	0.9965	0.2008	0.1495	1.2026	0.7539	1.0306	3.2641	0.2193	25.7218
NL02167	0.9895	0.1880	0.0984	1.1692	0.7257	1.0083	3.3902	0.2506	25.4651
NL02168	0.8175	0.4817	0.2825	0.8891	0.4264	0.7971	3.1122	0.2531	17.9548
LL02169	0.9627	0.3259	0.1627	1.4057	0.6368	1.2517	4.2186	0.2014	28.8723
EL02170	0.9740	0.2076	0.1147	1.3453	0.6109	1.0923	3.4208	0.3144	24.5017
LL02171	0.9463	0.2597	0.1568	1.1835	0.6086	1.0942	4.1470	0.2238	26.9824
LL02177	1.0676	0.3353	0.4205	1.2086	1.1152	1.1829	5.7370	0.2940	34.9503
LL02178	1.0599	0.1733	0.1995	1.3214	1.4239	1.2372	3.8160	0.1846	33.8402
LL02182	0.9494	0.2873	0.1317	1.1732	0.5413	1.0113	3.8992	0.2184	24.7493
LL02183	0.9650	0.2358	0.1864	1.5618	0.6521	1.3165	3.8710	0.2712	28.0666
NL02184	0.9391	0.2363	0.0080	0.9724	0.5654	0.8145	3.0409	0.2384	20.7530
NL02190	0.9555	0.4127	0.2315	1.2801	0.6231	1.1578	4.0041	0.1977	27.4210
LL02191	0.9732	0.2380	0.1660	0.9135	0.5998	0.7732	2.7124	0.2208	19.9010
LL02192	0.8233	0.5750	0.4080	1.3540	0.6825	1.2446	4.0379	0.3154	26.8311
LL02196	1.1469	0.3859	0.3289	1.1541	1.3434	1.0837	3.0201	0.2313	29.7509
NL02197	1.1359	0.3782	0.3344	1.4476	1.3383	1.3938	3.7229	0.2281	34.5698
LL02198	1.0638	0.2012	0.1243	1.5880	1.1053	1.3882	3.3928	0.2355	31.5787
NL02199	0.8487	0.4876	0.3222	1.1330	0.4061	0.9317	2.9297	0.2971	17.8185
NL02200	0.9378	0.4410	0.3325	1.3261	0.5284	1.1758	4.3470	0.2829	26.4465
NL02201	0.9492	0.3279	0.1892	1.3347	0.5331	1.0987	3.7295	0.3452	24.0028
NL02202	0.9214	0.3874	0.2443	1.2289	0.5109	1.0869	4.4097	0.2597	25.7331
NL02203	0.9214	0.4216	0.1696	1.0544	0.5318	0.9523	3.9441	0.1689	24.3273
LL02205	0.7787	0.9918	0.3322	0.9175	0.6305	0.9480	4.5908	0.1659	25.9012
LL02206	0.8223	0.4849	0.2763	1.1366	0.6576	1.0157	3.4478	0.3110	23.2509
NL02207	0.9435	0.0834	0.0539	0.6721	0.4026	0.5144	2.7679	0.2893	14.7768
NL02208	0.9079	0.1042	0.0777	0.6251	0.4507	0.5689	3.6066	0.1308	18.8382
LL02209	0.9365	0.0543	0.0511	0.6972	0.4613	0.5621	2.7774	0.2244	16.4697
LL02210	0.8950	0.1393	0.1622	0.7907	0.3387	0.6381	3.5909	0.2981	16.4990
LL02211	0.8898	0.3734	0.1771	0.6526	0.3913	0.6224	3.9992	0.0762	19.3326
NL02212	0.8671	0.2413	0.1793	0.6207	0.3491	0.5707	4.1786	0.1165	18.0290

LL02213	0.8288	0.2407	0.2100	0.6746	0.2979	0.5585	2.7675	0.2026	12.8964
LL02215	0.8384	0.2565	0.2068	0.7565	0.3785	0.7006	4.1143	0.1145	19.5528
EL02219	0.8879	0.4650	0.5139	1.2001	0.6668	0.9020	2.5738	0.4091	19.8267
NL02220	0.8428	0.4473	0.4634	1.2509	0.4445	1.0432	3.4021	0.4110	20.1745
EL02221	0.8835	0.4349	0.3442	1.2373	0.5015	0.9668	2.7760	0.3571	19.1378
GL02223	0.8323	0.3967	0.2564	1.0801	0.6188	0.8820	2.6309	0.3699	18.8353
LL02224	0.7900	5.0582	1.0104	0.7017	0.5615	0.7372	5.3396	0.1011	24.8135
NL02225	0.8339	0.5008	0.3828	1.1855	0.4578	0.9762	3.0135	0.4452	18.6703
NL02226	0.8346	0.4228	0.2729	1.0015	0.4445	0.8415	2.9145	0.4522	17.2177
GL02254	0.8946	0.1134	0.0959	1.0065	0.4616	0.8332	3.6581	0.1876	21.2346
KL02255	0.8914	0.0807	0.0969	1.0610	0.4904	0.9207	3.9062	0.1589	23.2246
GL02257	0.8948	0.0115	0.0231	1.2059	0.7371	1.0695	2.6293	0.3254	21.7858
GL02290	0.8833	0.0984	0.2076	0.7693	0.4364	0.6739	3.7875	0.2321	19.1561
LL02291	0.8754	0.2491	0.1750	0.6331	0.3590	0.5401	3.0593	0.2178	14.7658

The calculated ratios that based on the concentrations of selected compounds from saturate fraction hydrocarbons are presented in the next table. These indices are used in the analysis of the model 4.

	Ratios for model 4						
Sample	**Pr/Ph**	**Pr/nC17**	**Ph/nC18**	**CPI25-33**	**nC24+/nC24-**	**nC19/nC31**	**R22**
GL00794	1.6477	0.7245	0.5133	1.0793	0.4695	4.6056	1.0144
GL00858	1.3586	0.6834	0.6273	1.1080	0.4363	6.3139	1.0394
NL01143	1.1140	1.1254	1.1193	1.0892	0.5364	2.4613	0.9825
AL01144	1.1701	1.4952	1.4915	0.9541	0.4496	3.5011	1.0501
GL01277	1.6742	0.6456	0.4465	1.0143	0.2754	8.3020	1.0262
NL01350	1.3845	1.1292	0.9341	1.2413	0.2943	10.9406	0.9827
GL01354	1.2972	0.6333	0.5597	1.0898	0.2997	8.9641	1.0028
NL01420	0.7204	0.0400	0.0775	1.0787	0.0287	16.0000	0.9673
LL01453	1.3078	0.5577	0.4930	1.0919	0.2214	14.3138	0.9319
AL01556	1.2242	1.9459	2.4674	1.1303	0.1666	16.0000	1.0659
NL01557	1.4470	1.3298	1.0386	1.1329	0.3341	8.8325	0.9441
NL01558	1.2005	1.2820	1.4099	1.1945	0.2715	10.9984	1.0162
AL01559	0.9867	0.6820	0.7419	0.9867	0.2299	16.4164	0.9709
NL01576	1.3565	0.5091	0.4790	0.9415	0.2341	13.6064	0.9813
GL01598	1.1510	0.5066	0.5644	0.9602	0.2082	11.7938	1.0218
NL01638	1.5153	1.0284	0.7633	1.0555	0.2630	13.2169	0.9764
NL01639	1.5752	0.4847	0.3712	1.1057	0.1795	15.4156	0.9729
NL01641	0.9312	0.0865	0.1076	0.9865	0.0934	16.0000	0.8823
NL01644	3.3488	0.2052	0.0951	0.8499	0.0188	16.0000	0.9788
NL01645	2.2880	0.6492	0.4744	1.2500	0.0089	16.0000	0.8873
NL01646	1.6280	0.3177	0.2403	1.1164	0.1525	24.4153	0.9612
NL01647	1.9264	0.0541	0.0353	1.1202	0.0988	40.1585	0.9420
NL01648	1.7211	0.3797	0.2676	1.1169	0.1907	14.2481	0.9648
NL01650	1.5685	0.3708	0.2774	1.1080	0.1698	16.4359	0.9555
NL01651	1.4464	0.8606	0.6544	1.0845	0.2564	15.0345	0.9993

NL01652	1.5214	0.5557	0.4249	1.0581	0.2217	11.9785	0.9848
NL01654	1.6794	0.5996	0.4415	1.1308	0.2166	11.3271	0.9759
NL01655	1.5338	0.8765	0.6578	1.0356	0.2651	11.5703	0.9757
NL01656	1.5759	0.5646	0.4248	1.0768	0.2269	11.4742	0.9904
NL01658	1.7961	0.5103	0.3481	1.0784	0.1687	17.8718	0.9786
AL01664	0.9749	0.4896	0.7170	1.0370	0.3335	6.8375	1.0218
NL01667	1.0031	0.4963	0.7213	1.0227	0.3460	6.5735	1.0272
KL01676	1.5413	0.8689	0.6591	1.0391	0.2609	9.7611	0.9682
KL01677	1.4197	0.8729	0.6638	1.0440	0.4307	4.9291	0.9968
KL01679	1.5368	0.8506	0.6527	1.0381	0.2963	8.6423	0.9862
KL01680	1.1563	0.8562	0.6674	1.1697	0.5741	4.3915	1.0117
KL01684	1.3636	0.8230	0.6730	1.0562	0.4584	4.3999	0.9960
KL01686	1.4789	0.8515	0.6490	1.0561	0.4508	3.9134	0.9911
KL01687	1.5076	0.8806	0.6751	1.0385	0.3531	6.4919	0.9882
KL01688	1.5224	0.8457	0.6656	1.1447	0.3640	6.2763	0.9825
KL01690	1.4138	0.7894	0.6457	1.0516	0.3864	5.5271	1.0194
KL01691	1.5194	0.7970	0.6235	1.0769	0.2786	9.3176	0.9866
KL01692	1.5216	0.8431	0.6850	1.0365	0.3001	8.2874	0.9875
KL01693	1.4694	0.8424	0.6597	1.0421	0.4206	4.6062	1.0035
NL01810	1.4855	1.0249	0.7012	1.1760	0.3316	7.7868	0.9986
EL01816	1.1598	1.2194	1.1354	1.1572	0.4679	4.1442	0.9911
EL01819	1.1952	1.4059	1.2805	1.1539	0.4743	3.8691	0.9774
EL01820	1.0325	1.1170	1.1741	1.1177	0.4327	4.0042	0.9806
EL01821	1.0875	1.0916	1.1536	1.1187	0.4119	4.4855	0.9992
LL01822	0.7641	1.0142	1.4924	1.0104	0.3743	5.9417	1.0196
NL01823	1.2015	1.3167	1.2547	1.1421	0.4911	3.2874	0.9735
LL01824	1.3201	0.6323	0.5250	1.1338	0.3066	9.8376	0.9998
LL01825	0.7223	1.1709	1.8102	1.0419	0.4374	4.1587	1.0192
LL01827	1.2366	0.7155	0.6272	1.1151	0.3065	11.8256	0.9980
LL01828	0.5317	0.8498	1.7434	0.9664	0.4338	4.5060	1.0541
NL01831	1.5269	1.0435	0.8048	1.1671	0.2959	7.9939	0.9730
LL01832	1.3669	0.7757	0.6529	1.1470	0.3160	6.8652	0.9828
NL01833	1.0836	1.1386	1.1810	1.1583	0.4468	3.6429	0.9974
LL01834	1.4157	1.4529	1.2479	1.0658	0.4033	4.9461	0.9765
NL02032	1.4718	0.6709	0.5662	1.0333	0.3455	6.2654	0.9834
LL02034	1.5256	0.6743	0.5483	1.0146	0.3482	7.1533	0.9827
LL02035	1.4821	0.6147	0.5958	1.0617	0.3885	6.0975	1.0167
LL02038	1.4738	0.7232	0.6204	1.0556	0.3374	7.4672	1.0082
LL02039	1.4634	0.7035	0.5890	1.0454	0.3549	6.2440	0.9992
LL02040	1.4583	0.6988	0.6007	1.0259	0.3636	6.8145	1.0189
LL02041	1.4603	0.6863	0.5769	1.0670	0.3308	7.6655	1.0079
LL02042	1.4829	0.6943	0.5946	1.0439	0.3289	8.3973	0.9741
NL02043	1.4296	0.8435	0.7221	1.0879	0.3734	5.6240	1.0010
NL02044	1.4594	0.8424	0.7286	0.9874	0.3411	6.6814	0.9663
NL02045	1.4416	0.8381	0.7341	1.0007	0.3651	5.7804	0.9920
NL02077	1.3791	0.6240	0.5345	1.0378	0.3918	6.3026	0.9837

NL02078	1.4330	0.6201	0.5326	1.0827	0.3268	10.5714	1.0101
NL02079	1.4522	0.6193	0.5279	1.0497	0.3186	10.8950	1.0060
LL02080	1.1281	0.8173	0.8523	0.9698	0.3688	6.1936	0.9838
LL02081	1.2391	0.9753	0.8767	1.0273	0.3693	8.6803	1.0116
LL02082	1.2988	0.9892	0.9189	1.0117	0.4099	5.2133	0.9956
LL02084	1.1012	0.6612	0.7272	0.9926	0.3899	5.3607	1.0257
NL02086	1.0963	0.6740	0.7087	1.0093	0.3549	9.3575	1.0151
LL02098	1.2953	0.4790	0.4263	1.0495	0.3395	10.2446	1.0141
EL02099	1.3885	0.5655	0.5019	1.0492	0.3315	13.1226	1.0032
LL02100	0.7352	0.8899	1.4133	0.9434	0.4226	5.0665	1.0411
NL02103	1.2117	0.5901	0.5140	1.0184	0.3967	6.8505	1.0202
GL02106	1.5352	0.3282	0.2993	1.0262	0.3423	8.3233	1.0325
NL02108	1.1813	0.5934	0.7803	0.9617	0.4018	7.5935	1.0459
EL02109	1.5530	0.5846	0.4512	1.0787	0.3317	9.2553	0.9785
KL02110	1.4507	0.8043	0.6852	1.0471	0.3855	5.9973	0.9869
GL02112	1.3577	1.0063	0.8364	1.1248	0.3625	8.0242	0.9808
NL02151	1.2519	0.4836	0.4964	0.9809	0.3385	9.5571	1.0065
LL02152	1.0605	1.0552	1.3090	1.0206	0.4315	5.4546	1.0054
LL02153	1.1213	1.0767	1.3393	0.9690	0.3878	6.2847	0.9739
NL02154	1.4262	0.5632	0.4792	0.9973	0.3152	10.3310	1.0109
NL02155	1.3182	0.4913	0.4610	0.9880	0.3520	8.2670	0.9822
NL02156	1.5140	0.5569	0.4405	1.0124	0.3506	8.5202	1.0097
EL02157	1.4682	0.5636	0.4695	1.0513	0.3748	7.8569	1.0164
NL02158	1.4927	0.5426	0.4666	1.0276	0.3546	9.3498	1.0122
EL02159	1.4316	0.5724	0.4987	1.0501	0.3646	7.4469	0.9838
LL02160	1.3784	0.5668	0.5097	1.0050	0.3482	9.6630	0.9915
NL02161	1.6366	0.6184	0.4599	1.0288	0.3254	10.5338	1.0138
NL02162	1.2045	0.5025	0.4423	0.9890	0.4063	5.0392	1.0008
NL02163	1.2574	0.5334	0.4820	1.0353	0.3180	8.0617	1.0032
NL02164	1.2432	0.5205	0.4776	1.0074	0.3847	4.8145	0.9848
NL02165	1.2160	0.5056	0.4792	1.0233	0.3663	4.9926	0.9839
NL02166	1.2621	0.5206	0.4763	0.9969	0.3026	9.2813	0.9689
NL02167	1.2297	0.5267	0.4923	1.0746	0.3152	7.9464	0.9920
NL02168	1.0630	1.0636	1.0692	1.0391	0.4354	4.5723	1.0132
LL02169	1.3559	0.5922	0.4894	1.1525	0.2265	32.3788	1.0006
EL02170	1.2307	0.5746	0.5218	1.1023	0.3119	10.8569	0.9961
LL02171	1.3264	0.5798	0.5198	1.0260	0.3074	7.8153	1.0068
LL02177	1.2636	0.2706	0.2997	1.0041	0.3337	6.6454	1.0087
LL02178	0.7717	0.2645	0.3710	1.0155	0.3220	6.0700	0.9912
LL02182	1.2877	0.6325	0.5765	1.0505	0.3427	6.9965	0.9995
LL02183	1.3561	0.6167	0.5438	1.0268	0.2990	10.6386	1.0063
NL02184	1.3329	0.6444	0.5743	1.0003	0.2925	10.1979	1.0117
NL02190	1.2345	0.6267	0.5730	1.1113	0.3816	6.3608	0.9791
LL02191	1.2759	0.6123	0.5558	0.9964	0.3137	10.7992	1.0129
LL02192	0.7903	0.8618	1.2864	1.0409	0.2980	5.4513	1.0234
LL02196	1.0355	0.2462	0.2655	0.9880	0.3485	7.0510	1.0118

NL02197	1.0098	0.2628	0.2794	0.9903	0.3392	6.9023	1.0163
LL02198	1.3458	0.3562	0.3704	0.9875	0.3476	6.7741	1.0195
NL02199	1.4177	1.2479	1.0740	1.0984	0.4396	5.7197	0.9693
NL02200	1.2410	0.6663	0.5922	1.0165	0.3221	7.8139	1.0062
NL02201	1.3313	0.6508	0.5693	1.0706	0.3513	7.4580	0.9951
NL02202	1.3159	0.6677	0.5917	1.0998	0.3565	7.6491	0.9922
NL02203	1.2950	0.6959	0.6075	1.0032	0.3446	8.3007	0.9877
LL02205	0.7447	0.8771	1.3228	0.9599	0.3614	6.1208	1.0136
LL02206	0.7823	0.9127	1.3296	1.0015	0.3746	5.4360	1.0234
NL02207	1.3973	0.7283	0.6063	1.1260	0.3280	7.6400	0.9933
NL02208	1.4858	0.7551	0.6242	1.0386	0.3762	5.6598	0.9866
LL02209	1.3990	0.7335	0.6164	1.0317	0.3353	6.9950	1.0033
LL02210	1.4029	0.9007	0.9030	1.0166	0.3459	7.0018	0.9791
LL02211	1.4460	0.8791	0.9042	1.0666	0.3409	7.1659	0.9955
NL02212	1.5212	0.9021	0.8977	1.0127	0.3288	7.2520	0.9748
LL02213	1.4266	1.5371	1.2442	1.0557	0.3941	5.2827	0.9804
LL02215	1.4974	0.9428	0.8100	1.0366	0.3412	6.0648	0.9628
EL02219	0.6206	1.2177	1.9950	0.9070	0.4566	4.5596	1.0433
NL02220	1.0283	1.0787	1.1111	1.0762	0.4360	5.9022	1.0043
EL02221	0.7875	0.9526	1.6293	0.8985	0.4000	5.5684	1.0086
GL02223	0.8158	1.1018	1.5176	0.9249	0.3624	5.7050	1.0137
LL02224	0.8244	1.1863	1.5215	0.9742	0.3517	7.1881	1.0540
NL02225	0.9640	1.1175	1.2482	0.9483	0.3371	9.6595	1.0044
NL02226	1.0159	1.1246	1.2297	1.0029	0.3533	6.0579	1.0079
GL02254	1.5430	0.7839	0.6207	1.0290	0.3769	7.6352	1.0292
KL02255	1.4489	0.7909	0.6973	1.0503	0.3494	9.1854	1.0138
GL02257	0.9428	1.1472	1.4093	0.9927	0.4582	4.4864	0.9925
GL02290	1.3463	0.8254	0.7396	1.0339	0.3284	7.0811	0.9846
LL02291	1.3939	0.7335	0.6064	1.0182	0.2942	8.6451	0.9714