TECHNICAL UNIVERSITY OF CRETE

MASTER THESIS

Identifying oil families using Malcom (SLB) software package

Author:
Anna Koukounya

Supervisor:
Prof. Nikos Pasadakis

A thesis submitted in fulfillment of the requirements for the Master's degree in Petroleum Engineering

in the

School of Mineral Resources Engineering

Chania, October 2015

"Where oil is first found is in the minds of men"

Wallace Pratt (1885-1981)

ABSTRACT

Petroleum system definition and analysis use compositional links between petroleum and source kerogen. Petroleum families are commonly classified using qualitative or semi-quantitative methods based on compound presence or relative abundance, while petroleum-source rock correlation links a petroleum family to a stratigraphic unit, facies and/or locality containing the source kerogen. In this study, we apply multivariate statistical analysis to explore the oil family classification in Williston Basin using the Malcom (SLB) software package.

In this study twenty oil samples, that belong to five previously defined compositional families (oil families A, B, C, D and E) obtained from the Williston Basin are analyzed. The objective is to utilize biomarkers distribution and apply classification methods using the Malcom Interactive Fluid Characterization Software to identify the oil families. Principal Component Analysis (PCA), Ascending Hierarchical Classification (AHC) and Sammon mapping were employed to explore compositional data from saturated hydrocarbon fractions and hopane biomarkers of the oil samples from Williston Basin petroleum province.

The results indicate that an efficient classification of the oil families may be obtained when the AHC was used, with Ward's aggregation method. In addition PCA also revealed a good classification scheme of the oil families, especially when the n-alkanes are used. Finally the Malcom (SLB) software package was found to be a useful tool in geochemical data treatment.

AKNOWLEDGEMENTS

I wish to thank Professor Nikos Pasadakis for his devoted guidance, his enlightening discussions and endless patience made this work possible. I also wish to thank Mrs Eleni Hamilaki for her help.

Finally my deep gratitude goes to my family and friends for their encouragement and support.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

In this study twenty oil samples, that belong to five previously defined compositional families (oil families A, B, C, D and E) obtained from the Williston Basin are analyzed. The objective is to utilize biomarkers distribution and apply classification methods using the Malcom Interactive Fluid Characterization Software of Schlumberger to identify the oil families.

Principal Component Analysis (PCA), Ascending Hierarchical Classification (AHC) and Sammon mapping were employed to explore compositional data from saturated hydrocarbon fractions (SFH) and hopane biomarkers of the oil samples from Williston Basin petroleum province. In PCA the samples are displayed in PC space, and subsequently assigned to oil families. In AHC the samples are classified using four different aggregation methods, specifically Ward's, Link's Complete,Link's Simple and Link's Average aggregation method. Lastly in Sammon mapping the samples were projected in a 2D space while the distances of the samples are kept the same.

Finally, the results regarding the classification of the oil families for all three methods are presented.

1 WILLISTON BASIN

1.1 Introduction

The Williston Basin is a structural basin located in North America, specifically in North Dakota, South Dakota, Montana, Saskatchewan and Manitoba. It is a major producer basin of oil and gas (Figure 1). The petroleum system concept was first applied by Dow and Williams who defined three oil systems in the Williston Basin: Tyler, Bakken, and Winnipeg (Dow, 1974; Williams, 1974).

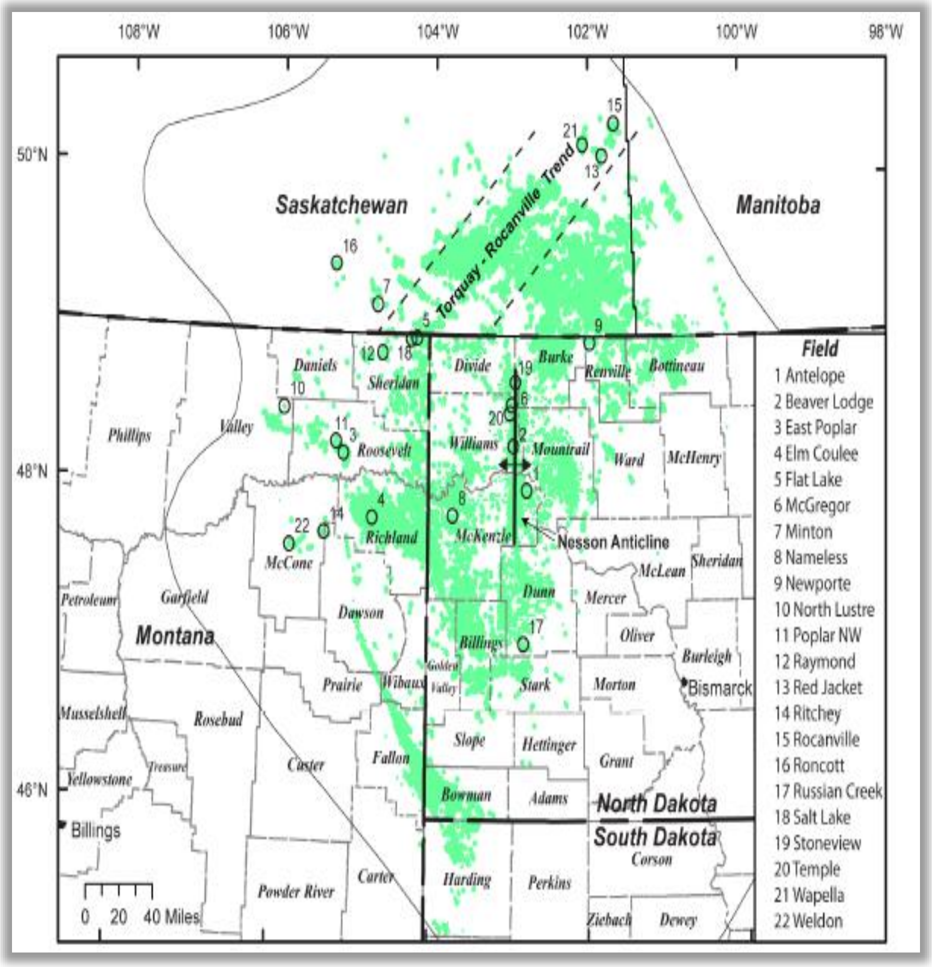


Figure 1: Map of the Williston Basin, United States and Canada (oil fields in green) after Lillis (2012)

Since then, the petroleum system concept has evolved and recent work has defined at least nine oil systems in the basin (Figure 2).

Figure 2 shows the stratigraphic distribution of the identified petroleum systems in the Williston Basin. The stratigraphic distribution of the oil families from each system is generally limited to the same source rock due to efficient seals and a paucity of vertical migration pathways (Lillis, 2012).

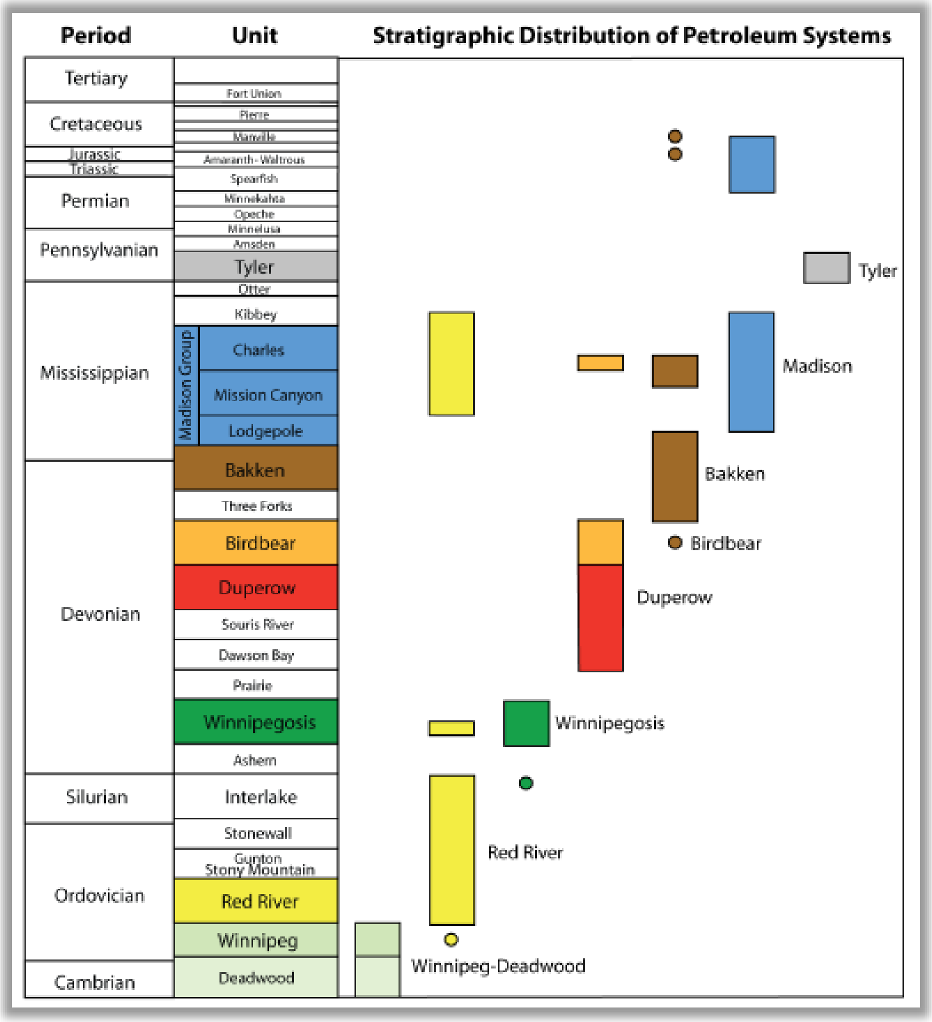


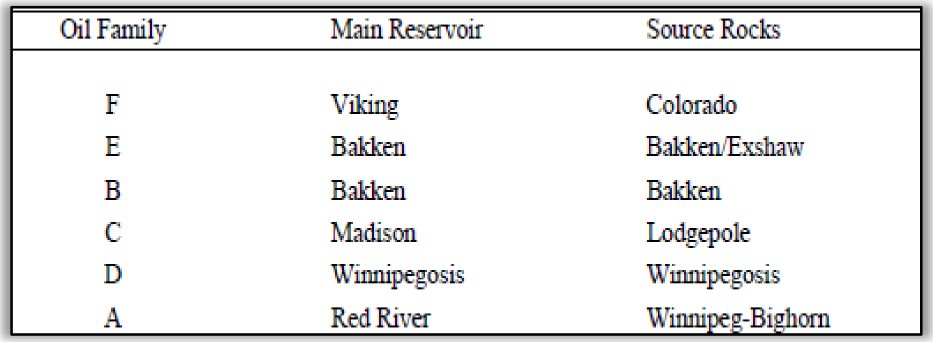
Figure 2: Stratigraphic column of the Williston Basin showing the petroleum systems in color and the stratigraphic distribution of the petroleum system fluids, after Lillis (2012)

1.2 Geological Backround

The Williston Basin is located in the United States and Canada with an area of approximately 300 000 square miles. The United States side is comprised of states of Montana, North Dakota and part of South Dakota.

The sedimentation in the Williston Basin started during the Cambrian and continued up to the Quaternary. The stratigraphic section in this basin has an approximate thickness of up to 16000 ft. in the central part of the basin. During deposition during the Middle Devonian, the Williston Basin was tilted to the north and connected with the Elk Point Basin. With this new configuration the sediments in the basin thicken from south to north (Gerhard et al., 1987). Subsequent deposition during Late Devonian showed a dominance of marine conditions, but this time the sedimentation comes from west through the Montana Trough (Gerhard et al., 1990).

According to Obermajer (2003) the regional petroleum systems are relatively well defined due to previous studies, making this setting ideal for developing alternative means of identifying genetic families of oils and characterizing petroleum systems. Six of the compositionally distinctive oil families that have been recognized in the Williston Basin, are shown in Table 1 (Obermajer et al., 2003), while the stratigraphic range of the families is presented in Figure 3.



Oil Family	Main Reservoir	Source Rocks
F	Viking	Colorado
E	Bakken	Bakken/Exshaw
B	Bakken	Bakken
C	Madison	Lodgepole
D	Winnipegosis	Winnipegosis
A	Red River	Winnipeg-Bighorn

Table 1: Generalized Williston Basin oil family classification (after Obermajer 2003)

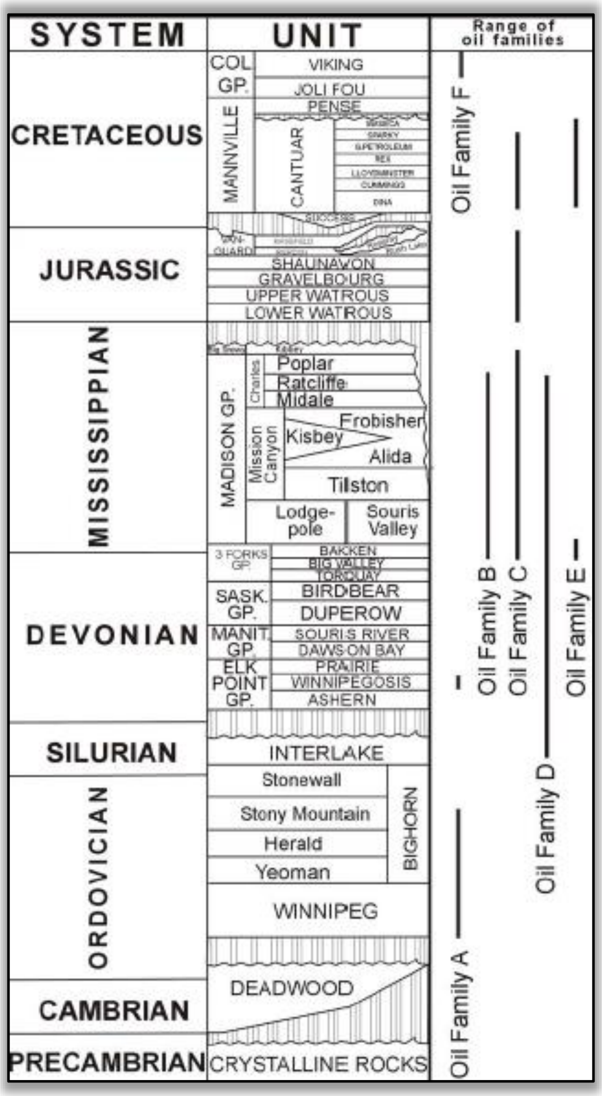


Figure 3: Stratigraphic range of oil families in the Williston Basin (after Osadetz 1992, 1994)

1.3 Williston Basin Petroleum Systems

In this study, classification is applied in five different oil families, specifically in Families A, B C, D and E. Twenty samples, containing all five oil familes, were analyzed with GC-MS analysis (more details of the analysis are presented in the chapter "RESULTS"). Following the petroleum systems of each family are analysed, Also the distributions of the saturate hydrocarbons and hopane biomarkers of those samples are presented.

1.3.1 Red River Petroleum System (Oil Family A)

The Red River oil family was first identified by Williams (1974) as "Type I" oils and was also confirmed in other geochemical studies (Thode, 1981; Zumberge, 1983; Leenheer and Zumberge, 1987; Brooks et al., 1987) The gas chromatogram signature is characterized by an odd carbon number predominance in the C9 to C19 n-alkanes and unusually low concentrations of C20+ n-alkanes and acyclic isoprenoids, particularly pristane and phytane. Red River oils have a low S content and display a wide range of maturities.

Most of the Red River formations consist of marine limestone and dolomite with TOC values ranging from 0.14 to 0.54 wt % according to Williams (1974), but Kohm and Louden (1982) reported kerogenite beds in the lower Red River Formation with TOC values between 9 and 14 wt %. The distribution of the n-alkanes (Figure 4) and the hopanes (Figure 5) of the oil family A is presented below, where all the samples of the specific family appear.

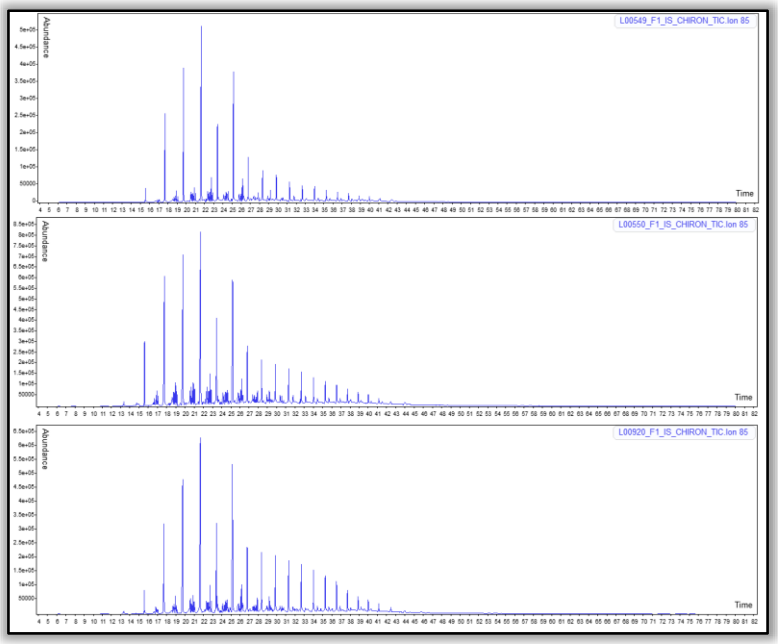


Figure 4:Distribution of the n-alkanes of the oil family A

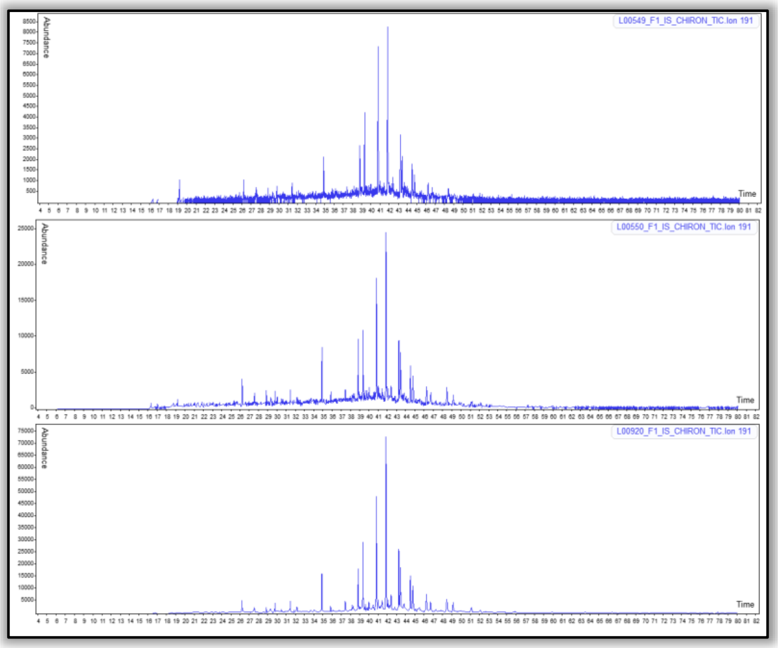


Figure 5: Distribution of the hopanes of the oil family A

1.3.2 Bakken Petroleum System (Oil Families B & E)

The Upper Devonian and Lower Mississippian Bakken Formation has long been known to be a world-class source rock in the Williston Basin (Murray, 1968; Williams, 1974; Dow, 1974) In his study, Williams (1974) identified the Bakken as a major source rock for the oils found in Mississippian Madison Group reservoirs in the Williston Basin, and based on this study Dow (1974) proposed the Bakken-Madison oil system (petroleum system).Later on the Bakken-Madison petroleum system of Williams (1974) was supported by Thode (1981) who used sulfur isotopes to correlate the Madison oil family to the Bakken oil family.

In 1987 Brooks was the first to recognize the genetic difference between the oils produced from Madison and Bakken reservoirs based on biomarker geochemistry. The low pristane/phytane and diasterane/sterane values in addition to the high norhopane/hopane values in the Madison-reservoired oils, indicate a carbonate source deposited in an anoxic water column. On the other hand Bakken-reservoired oils have high diasterane/sterane values indicating an argillaceous source rock. So they proposed that the oils found in the Madison are not Bakken-sourced but are derived from source rocks within the Madison Group. In addition, Osadetz (1992) correlated the Bakken reservoired oils in the Canadian Williston Basin to Bakken source rocks based on their biomarker signatures. All subsequent geochemical studies have supported the Canadian Bakken oil-source correlation of Osadetz (1992) and extended the correlation into the U.S. Williston Basin (e.g. Price and LeFever, 1992, 1994; Obermajer et al., 1998; Jarvie, 2001).

The Bakken Formation was defined in the subsurface of Dakota by Nordquist (1953) and subdivided into three parts: lower and upper part of organic-rich mudstones deposited in a restricted marine basin under largely anoxic conditions, and a middle part consisting of various lithologies, such as sandstone, siltstone, dolomite, and mudstone deposited in a shallow marine environment (LeFever et al., 1991; Smith and Bustin, 2000). Numerous studies have shown that the Upper and Lower parts have very similar organic richness and kerogen quality, while the Middle part has very low organic carbon content. The kerogen type in the Upper and Lower Bakken parts is considered to be mostly Type II (Osadetz et al., 1992).

In the Bakken Formation in Saskatchewan (Canada) two oil families can be found, the oil Families B and E. The organic facies in the Bakken Formation can be distinguished from the Upper Devonian and Lower Mississippian Exshaw Formation, with the former (Oil Family B) containing deep to intermediate water depth organic facies, and the latter (Oil Family E) shallow-water organic facies, as was mentioned by Stasiuk (2004). The distribution of the n-alkanes and the hopanes of the oil families B and E is presented in Figures 6-9, where all the samples of the family appear.

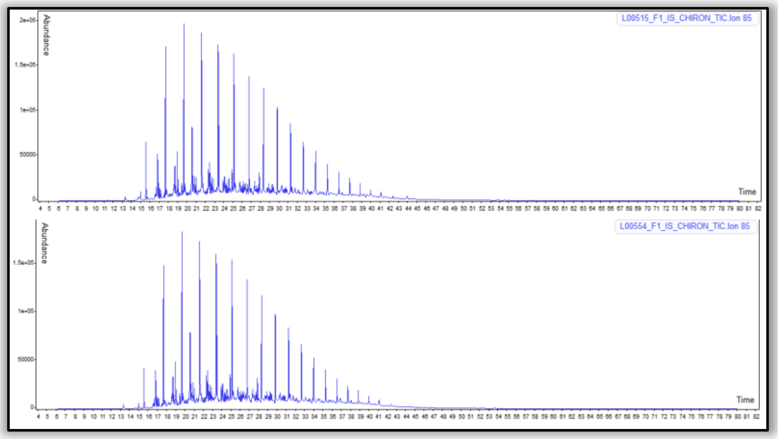


Figure 6: Distribution of the n-alkanes of the oil family B

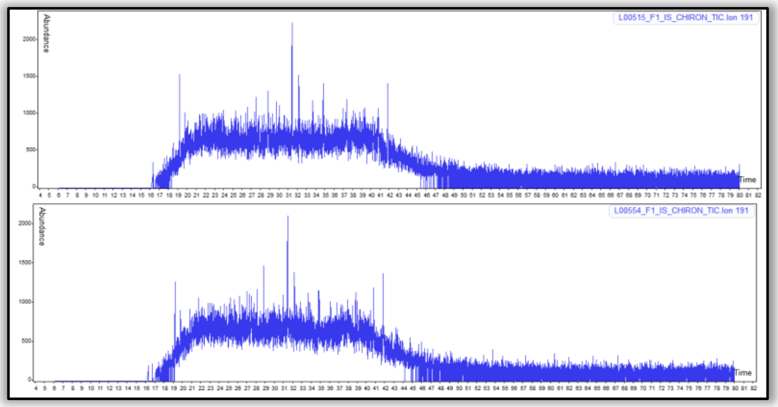


Figure 7: Distribution of the hopanes of the oil family B

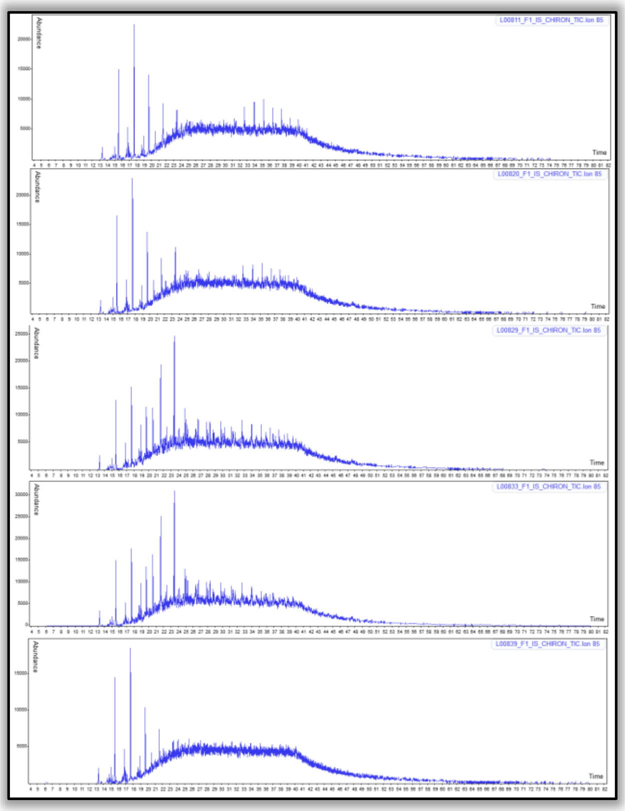


Figure 8: Distribution of the n-alkanes of the oil family E

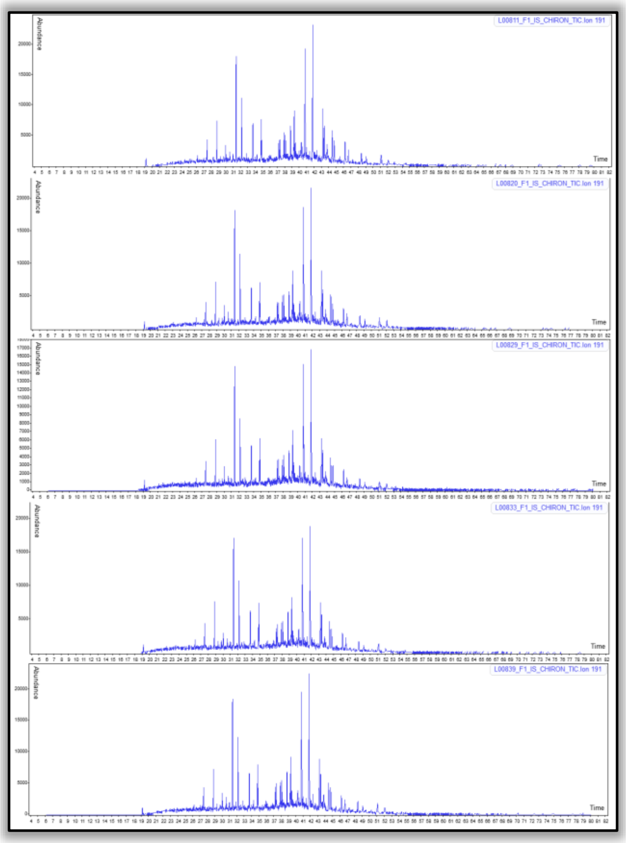


Figure 9: Distribution of the hopanes of the oil family E

1.3.3 Madison Petroleum System (Oil Family C)

Most of the Williston oil families have unique compositions and are easily correlated with a specific source rock but oil family C has been shown to be more heterogeneous than the other families, according to Obermajer (2000). Family C was initially defined in the Madison Group of eastern Saskatchewan by Osadetz (1992) and later on was extended into the western area of Saskatchewan again by Osadetz (1994), and into the American portion of the basin by Price (1994). As Jarvie (2001) reported, the compositional variation within the family C oils has been attributed to heterogeneity of their sources and mixing of oils derived from Bakken and Madison systems.

“The geochemical composition of Madison oils, is characterized by high S content, low pristane/phytane and diasterane/sterane values and high norhopane/hopane values, indicating a carbonate source rock”, mentioned Brooks (1987).

The distribution of the n-alkanes and the hopanes of the oil family A is presented in Figure 10 and Figure 11 respectively, where all the samples of the specific family appear.

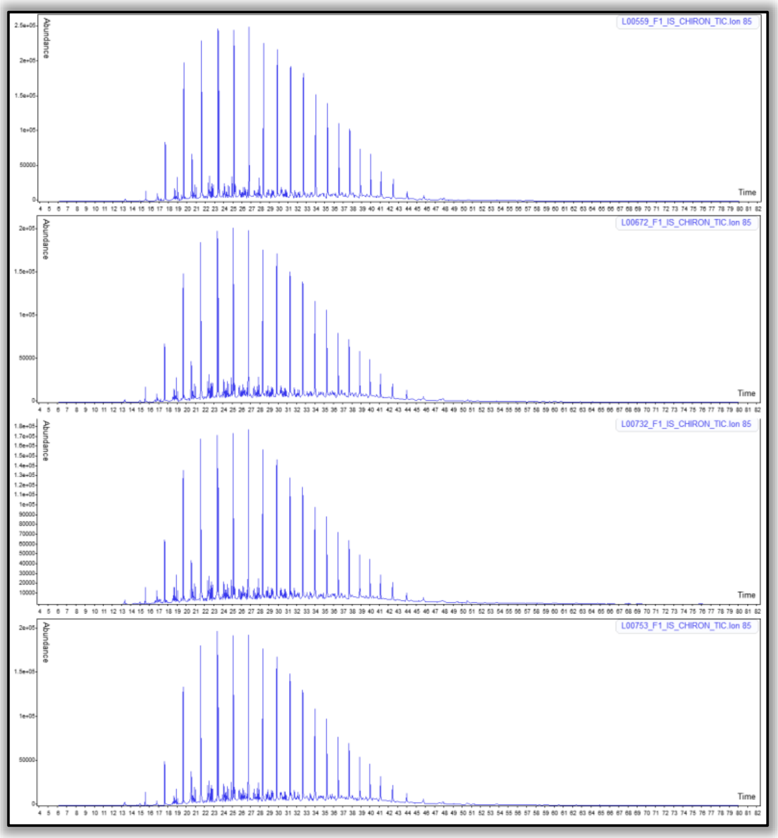


Figure 10: Distribution of the n-alkanes of the oil family C

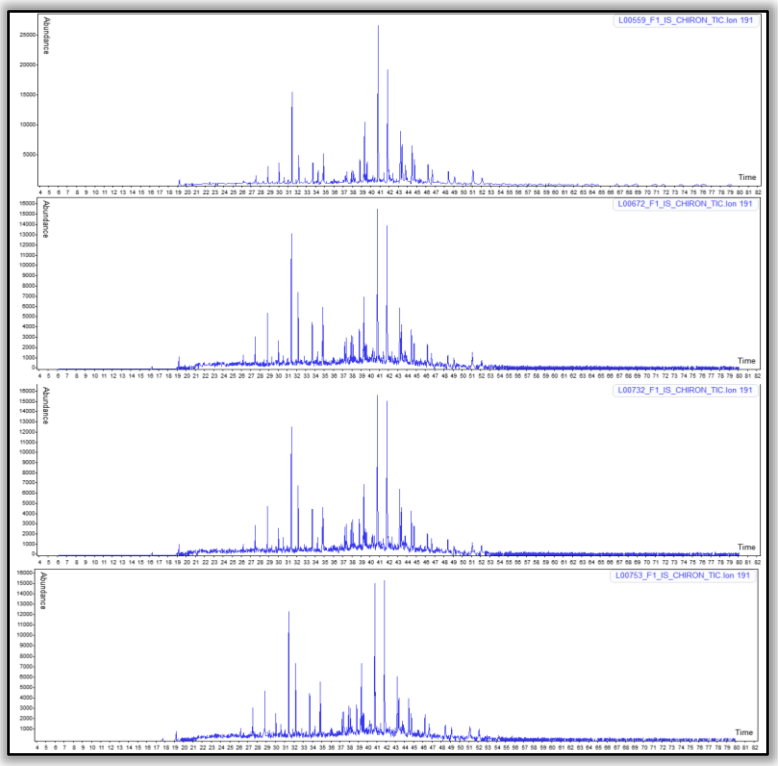


Figure 11: Distribution of the hopanes of the oil family C

1.3.4 Winnipegosis Petroleum System (Oil Family D)

Oil Family D was first recognized as a distinct genetic oil family amd is produced from the Devonian Winnipegosis Formation in the Canadian Williston Basin (Brooks, 1987). According to Dow (1974) the overlying Devonian Prairie Formation is a good regional seal and most likely prevented the Winnipegosis oil from migrating into younger reservoirs. Finally, Osadetz and Snowdon (1995) believe that the Winnipegosis source rocks contain Type I and Type II kerogen based on Rock-Eval and visual kerogen analyses. The distribution of the n-alkanes and the hopanes of the oil family A is presented in Figure 12 and Figure 13 respectively, where all the samples of the specific family appear in one picture.

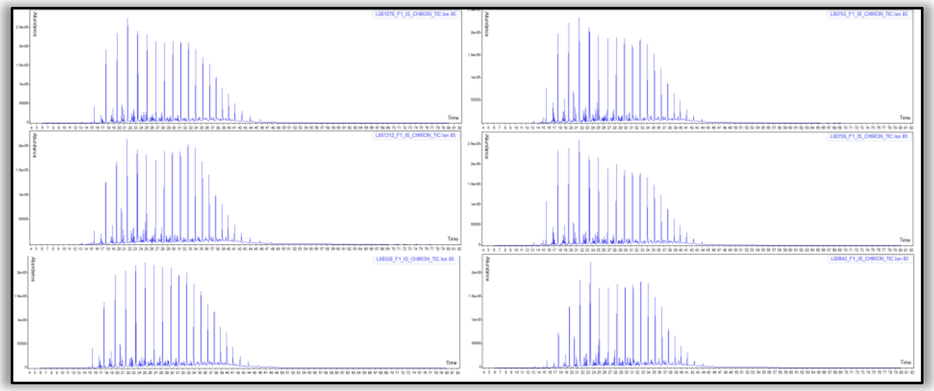


Figure 12: Distribution of the n-alkanes of the oil family D

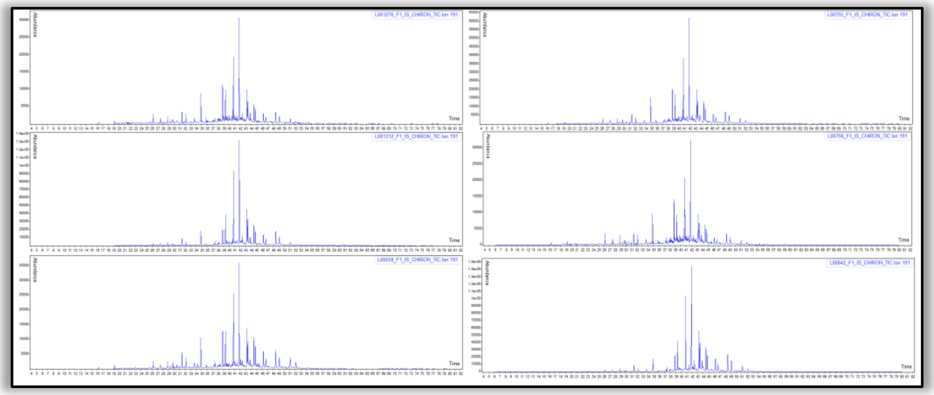


Figure 13: Distribution of the hopanes of the oil family D

2 CLASSIFICATION METHODS

In this study, we apply multivariate statistical analysis to explore the oil family classification in Williston Basin. Petroleum system definition and analysis use compositional links between petroleum and source kerogen. Petroleum families are commonly classified using qualitative or semi-quantitative methods based on compound presence or relative abundance (Dow, 1974), while petroleum-source rock correlation links a petroleum family to a stratigraphic unit, facies and/or locality containing the source kerogen (Curiale, 1994).

2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an exploratory multivariate statistical method that can be used to identify relations in large data sets influenced by multiple variables. Petroleum systems are well suited to such exploration because oil composition results from a complicated interaction of biological, environmental, geological, and physical processes. PCA has many applications to geological problems. It has been applied to organic geochemistry describing and classifying the petroleum generation and the secondary processes. It has also been used to identify petroleum families while characterizing alteration pathways and it has been shown to be efficient for the discrimination of petroleum. In this work we use PCA for variable reduction and classification purposes, maximizing the diagnostic characteristics of the saturated hydrocarbon and the hopanes biomarkers (Obermajer, 2003).

Principal Components (PCs) are the underlying structure of the data, specifically they are the directions where there is the most variance. In other words the directions where the data is more spread out. Principal Components are derived through PCA and represent a linear combination of original variables that account for the largest possible portion of the original data total variance. The first PC passes through the centroid of the standardized data set and explains the greatest amount of variance of any single PC, whereas successive PCs explain progressively less of the original variance. The number of principal components is less than or equal to the number of original samples. This transformation is defined in such a way that the first principal component has the largest possible variance and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. The principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric. Finally, it has to be noted that PCA is sensitive to the relative scaling of the original variables.

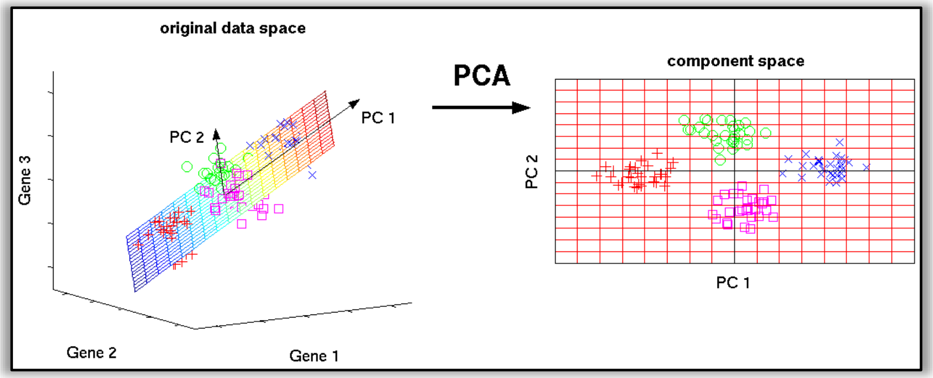


Figure 14: Data reduction using Principal Component Analysis (PCA)

A group of individuals is characterized by a number of p descriptive variables. These p descriptive variables are firstly standardized: each component is subtracted from its mean value and divided by its standard deviation (on the whole dataset). This enables to give the same weight to each variable. The principal component analysis consists in finding a space Ek (k<p) such as the inertia of the cloud points projected on Ek is at its maximum. This involves diagonalizing the correlation matrix and studying Eigen vectors' space:

- Eigen vectors represent the factor axes
- Eigen values correspond to the inertia related to each axis.

The correlation matrix is deducted from the covariance matrix.

$$\mathbf{M} = \begin{bmatrix} x_1 - \bar{x} & y_1 - \bar{y} \\ \vdots & \vdots \\ x_n - \bar{x} & y_n - \bar{y} \end{bmatrix} \Rightarrow \begin{gathered} \mathbf{Cov} = \frac{1}{n-1}\mathbf{M}^{\mathbf{T}}.\mathbf{M} \\ \mathbf{Corr}\ (X, Y) = \frac{\mathbf{Cov}\ (X, Y)}{\sigma_X . \sigma_Y} \end{gathered}$$

Equation 1: Correlation matrix

The correlation matrix is symmetrical with diagonal elements equal to 1. It gives access to correlation between variables from non-diagonal components.

If a non-diagonal element c_{ij}, representing links between variables i and j, is close to:

- -1, then the two variables are anti-correlated
- 0, then the variables have no evident correlation
- 1, then the variables are correlated

Eigen vectors and values exist in pairs: every Eigen vector hasa corresponding Eigen value. Eigen vectors give the direction of the axes of the orthonormal basis while Eigen values are numbers telling you how much variance there is in the data in that direction. The more their associated Eigen values are high, the more their corresponding axis contain information. The inertia, which can be associated to the quantity of information, is calculated from normalized Eigen values.

Results can be displayed on a two-dimensional graph, representing the values of the projected individuals on the more significant principal axes. The principal axes are defined by Eigen vectors of the correlation matrix and sorted by decreasing Eigen values. The first principal axes contain the highest information. If the inertia of a particular axis is too low, it is then possible to ignore the projected values on these axes, so the number of dimensionality can be reduced.

2.2 Ascending Hierarchical Classification (AHC)

The ascendant hierarchical classification method consists in building a series of clusters with n, n-1,…,1 classes stacked one to each other. Each iteration leads to the aggregation of two classes following an aggregation criteria, based on the measure of Euclidean distances or on dissimilarity of classes. The key point of this method relies on a non-random initialization. As all the parameters of this method are identical, a set of data can only have a unique solution.

For the aggregation, at the initial stage, each individual is considered as a whole group. As long as the individuals have not been merged inside one single group, the two nearest group are gathered, the distances between the newly created group and the other groups are updated and the process is repeated until all the individuals are gathered into one single group.

For the calculation of the distances between groups two different calculation methods can be applied:

- ***The Ward method***. It's the most commonly used method. This aggregation corresponds to the one with the minimal variance (i.e. the mean-square distance between each individual of a class and the center of gravity of this class is minimum).

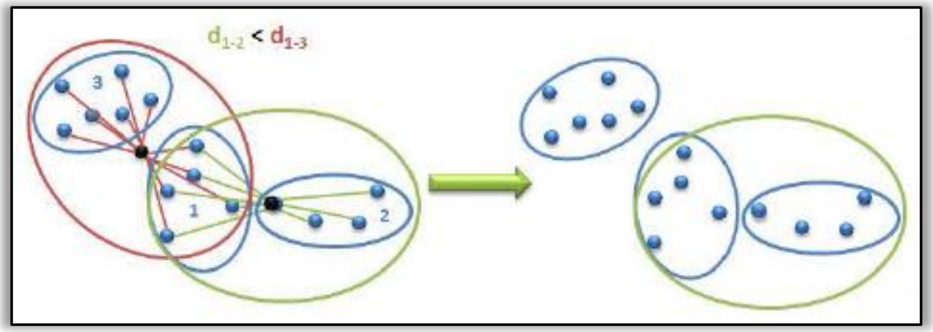


Figure 15: Calculation of dinstances – Ward Method

- ***The links method.*** Consists of three different aggregation methods, described below.

- *Single links method*. The calculation of the distance between two classes is achieved by choosing the closest individuals of each class. The aggregation is performed between the two classes having the lower distance.

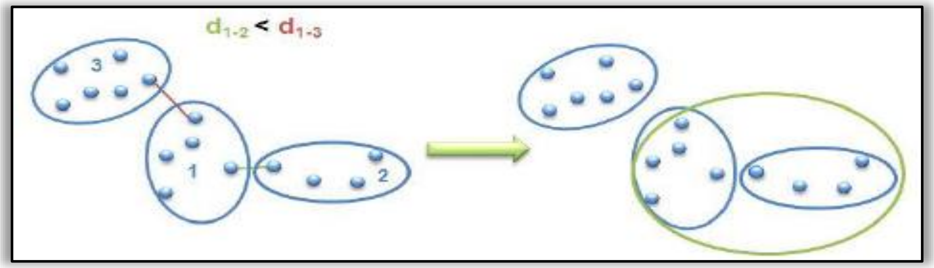


Figure 16: Calculation of distances – Single Links Method

- *Complete links method*. The calculation of the distance between two classes is achieved by choosing the farthest individuals of each class. The aggregation is performed between the two classes having the lower distance.

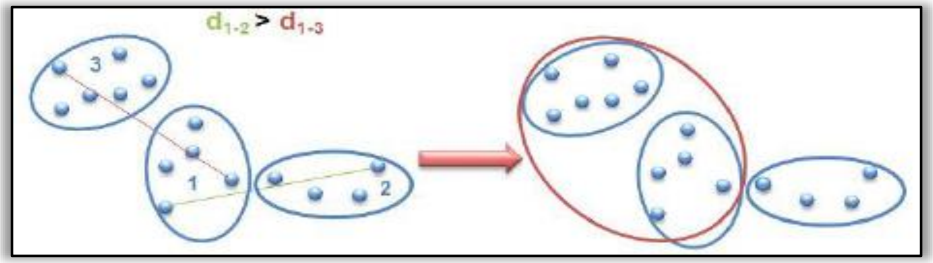


Figure 17: Calculation of distances – Complete Links Method

- *Average links method*. The calculation of the distance between two classes is achieved by evaluating the mean distance between all the individuals of each class, so the aggregation is performed between the two classes having the lower distance.

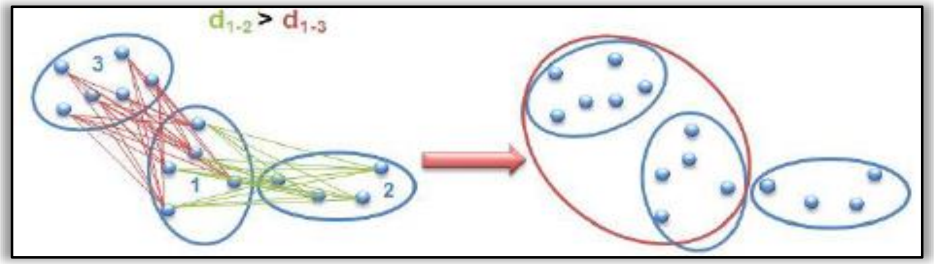


Figure 18: Calculation of distances – Average Links Method

For groups in which individuals have very different characteristics, these methods lead to the same results. On the other hand, for groups in which individuals are poorly differentiated, the results can be different depending of the chosen method:

- The single links method favors the aggregation of a high number of individuals,
- On the opposite, the complete links method favors the aggregation of groups containing a small number of individuals,
- The average links method is an intermediate between the two above methods

Finally, the Ward method relies on a minimal inertia criterion, nearly independent on the number of individuals in the group.

2.3 Sammon Method

It is known that individuals characterized by more than 3 variables can't be visualized in their variables space. However, it is possible to perform some projections in a two or three dimensions space, which keep the distance between individuals at best, in order to get an idea of the potential relative similarities between individuals.

The Sammon algorithm is particularly adapted to this situation. This method reduces the number of variables related to the individuals with the aim to keep the same distances between the individual before and after reduction. It provides a quick overview of the relative position of the samples in a 2D space but it has to be used cautiously, as part of the information is lost with the reduction of the data.

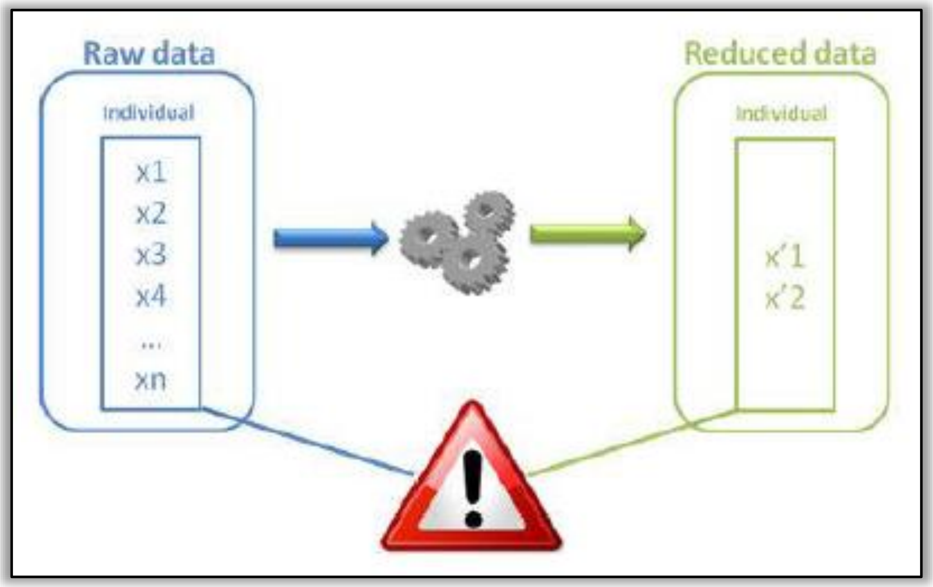


Figure 19: Data reduction using 2D sammon visualization

The Sammon method consists in minimizing the quadratic error distances between inter-individuals in the variables space and inter-individuals in the projected space.

$$C = \frac{1}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\left[D_{ij} - d_{ij}\right]^2}{d_{ij}}$$

Equation 2: Calculation of the total quadratic deviation

With:

D_{ij}: distance between individual I and individual j in the variables space

d_{ij}: distance between individual I and individual j in the projected space

C: total quadratic deviation

The coordinates of individuals in the projected space are optimized by an iterative process, so that the quadratic deviation converges towards a minimum. It is important to note that the initialization of the algorithm is performed randomly. It is then possible to get different results from one test to another, since the distances between individuals are important, and not the individual position in the projected space.

3 MALCOM SOFTWARE

In this study Malcom Interactive Fluid Characterization Software by Schlumberger was used to identify different oil families of the Williston Basin. Malcom Software is capable of providing comprehensive and rapid interpretation of geochemical rock and fluid properties. It stores, evaluates and processes geochemical data streams to facilitate the interpretation process, while it also enables a faster and more dynamic integration into the full chain of upstream exploration and production.

In more details the software includes project management to organize and store geochemical datasets as well as chromatographic peak identification and quantification and extracted ion analysis. It also features high-quality chromatogram extracting tools that provide easy exploration and quality control of geochemical datasets. It relies on comparison of the chemical composition of several chromatograms acquired under the same chromatographic conditions. The comparison of oil GC fingerprints provides information on biomarkers, reservoir connectivity, estimation of the size of reservoirs, and production allocation calculation (www.slb.com/malcom).

The software is used mostly for characterization of reservoirs, as well as studies of reservoir continuity. Characterization of reservoir continuity provides information for reducing the key uncertainties in a proposed oil field development and in planning and implementing the optimal development of petroleum reservoirs. “Various methods can be used to assess the compartmentalization of reservoirs, such as PVT measurements, gas chromatography (GC) ‘‘oil fingerprinting”, gas chromatography–mass spectrometry (GC/MS). Since the beginning of the 1980s, reservoir oil fingerprinting (ROF) has been widely used to determine reservoir continuity” was mentioned by Nouvelle (2010).

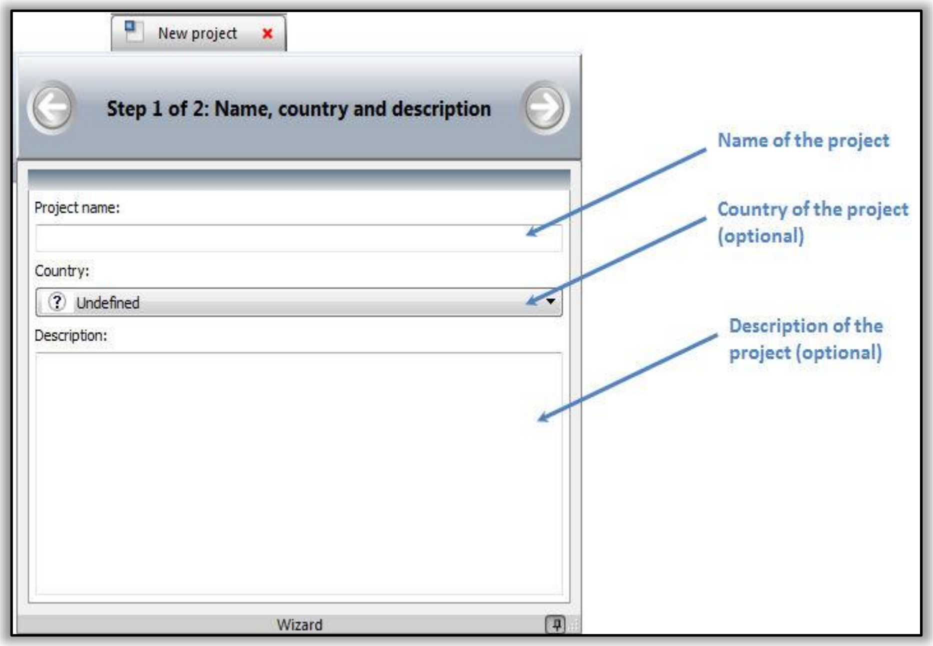
In this study the Malcom Software was used for the classification of the oil samples of the Williston Basin oil families. The procedure of identifying the oil families will be decribed in details below.

3.1 Creating a Project

Starting the program, the first thing to do is to create a New Project. A project is a group of data comprising analyses, numerical arrays and results of the data processing.

When the New Project is selected, a dialog box appears in a new dynamic tooltab, enabling the user to define the following three options:

- The name of the project (mandatory).
- The country: Greece was selected from the list (optional),
- A description of the project (optional).



New project

Step 1 of 2: Name, country and description

Project name:

Country:

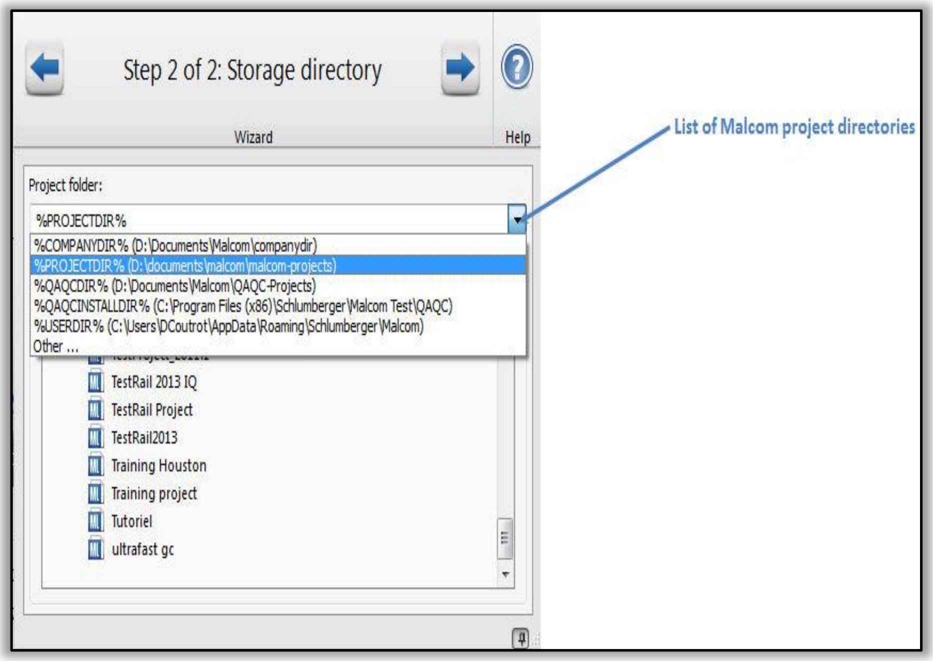
? Undefined

Description:

Wizard
Name of the project
Country of the project (optional)
Description of the project (optional)

Figure 20: Generate a New Project tab– step 1 (after Malcom's users manual)

The second step is to modify the project location wherever it is desirable for the user. Click on the drop drop down menu to access the list of Malcom project directories and choose another folder in the list of Malcom project directories.



Step 2 of 2: Storage directory

Project folder:

%PROJECTDIR%

%COMPANYDIR% (D:\Documents\Malcom\companydir)
%PROJECTDIR% (D:\documents\malcom\malcom-projects)
%QAQCDIR% (D:\Documents\Malcom\QAQC-Projects)
%QAQCINSTALLDIR% (C:\Program Files (x86)\Schlumberger\Malcom Test\QAQC)
%USERDIR% (C:\Users\DCoutrot\AppData\Roaming\Schlumberger\Malcom)
Other ...

TestRail 2013 IQ
TestRail Project
TestRail2013
Training Houston
Training project
Tutoriel
ultrafast gc

Figure 21: Generate a New Project tab– step 2 (after Malcom's users manual)

When clicking on the "Next" button, it creates the new project. This new project is empty, so the next step consists in importing analyses in the project.

3.2 Import of Data

Now the import of the data follows. In Malcom a lot of different file formats can be imported, but in this case only analysis files of GC/MS analysis are used.

It is possible to import some new analyses in a project at any time. Please note that data import into Malcom is a two steps process: first comes the import the data into the Import buffer, then from this buffer, we can pre-visualize data in the Preview tab. After check properties of the data in the Properties tab and choose the desirable data to be imported. Once the data are checked and selected to import, click on the Import button to load the data into the project. The "Project" dynamic tooltab will open the import wizard, which enables the user to drag and drop data from the file system explorer on the left to choose the file or files to import.

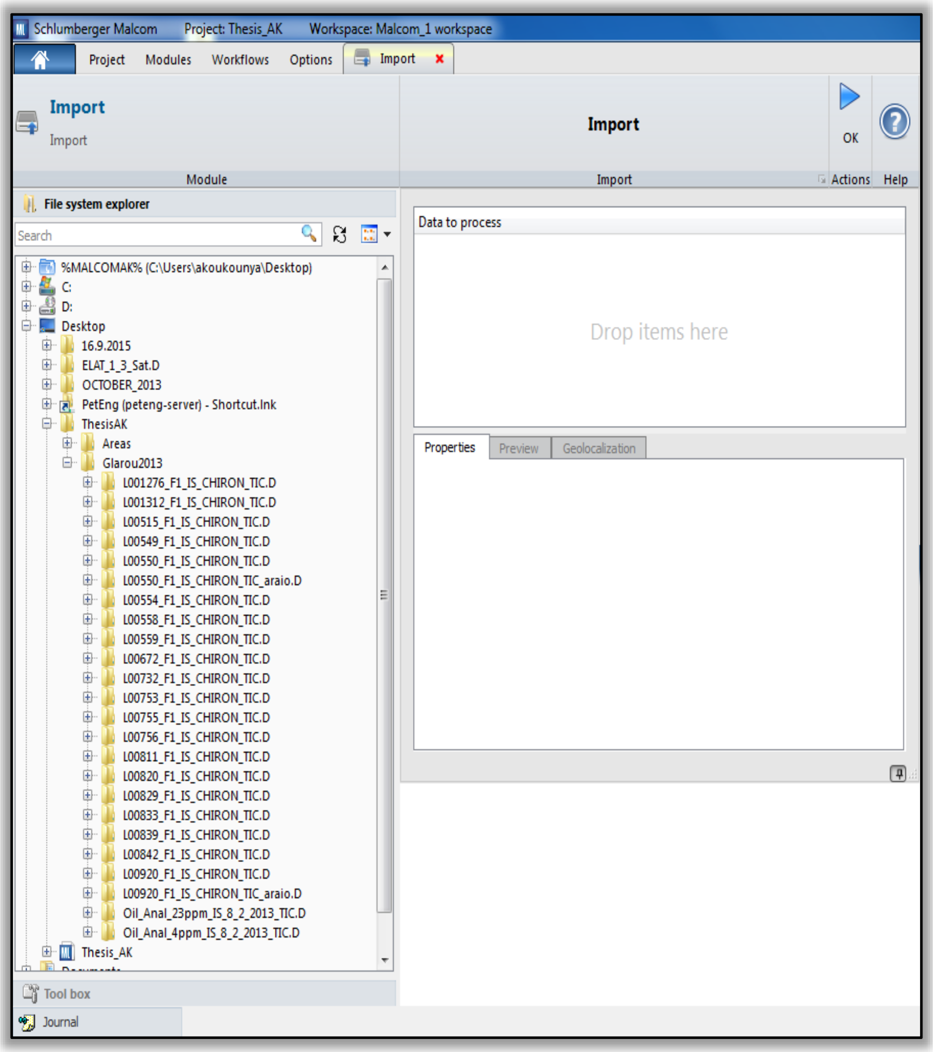
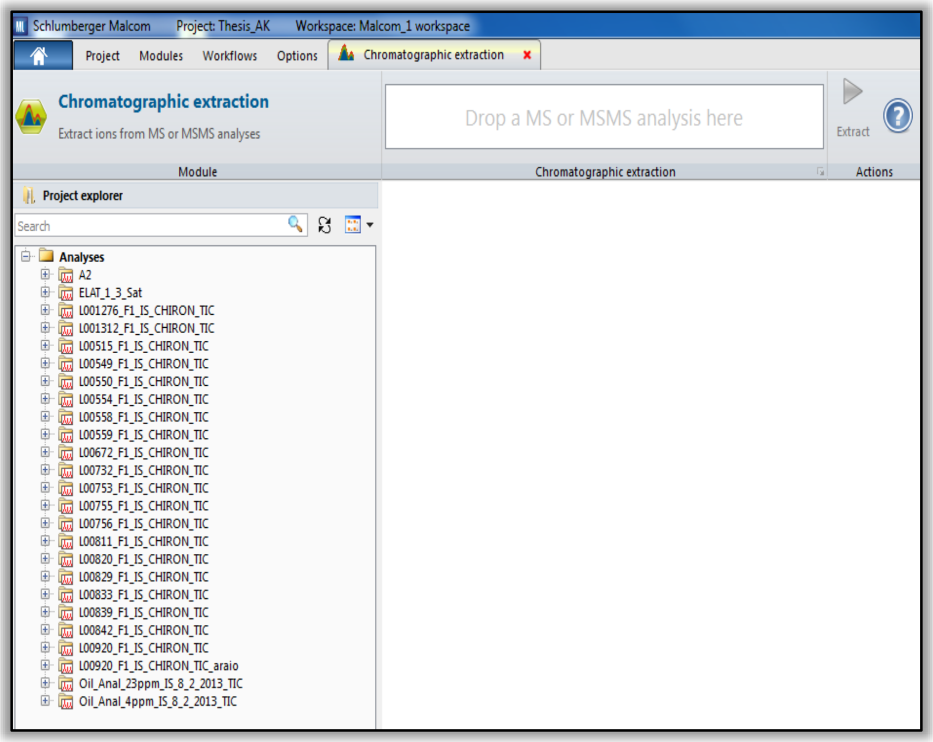

Schlumberger Malcom
Project: Thesis_AK
Workspace: Malcom_1 workspace
Project
Modules
Workflows
Options
Import
Import
Import
Import
OK
Module
Import
Actions
Help
File system explorer
Search
%MALCOMAK% (C:\Users\akoukounya\Desktop)
C:
D:
Desktop
16.9.2015
ELAT_1_3_Sat.D
OCTOBER_2013
PetEng (peteng-server) - Shortcut.lnk
ThesisAK
Areas
Glarou2013
L001276_F1_IS_CHIRON_TIC.D
L001312_F1_IS_CHIRON_TIC.D
L00515_F1_IS_CHIRON_TIC.D
L00549_F1_IS_CHIRON_TIC.D
L00550_F1_IS_CHIRON_TIC.D
L00550_F1_IS_CHIRON_TIC_araio.D
L00554_F1_IS_CHIRON_TIC.D
L00558_F1_IS_CHIRON_TIC.D
L00559_F1_IS_CHIRON_TIC.D
L00672_F1_IS_CHIRON_TIC.D
L00732_F1_IS_CHIRON_TIC.D
L00753_F1_IS_CHIRON_TIC.D
L00755_F1_IS_CHIRON_TIC.D
L00756_F1_IS_CHIRON_TIC.D
L00811_F1_IS_CHIRON_TIC.D
L00820_F1_IS_CHIRON_TIC.D
L00829_F1_IS_CHIRON_TIC.D
L00833_F1_IS_CHIRON_TIC.D
L00839_F1_IS_CHIRON_TIC.D
L00842_F1_IS_CHIRON_TIC.D
L00920_F1_IS_CHIRON_TIC.D
L00920_F1_IS_CHIRON_TIC_araio.D
Oil_Anal_23ppm_IS_8_2_2013_TIC.D
Oil_Anal_4ppm_IS_8_2_2013_TIC.D
Thesis_AK
Documents
Tool box
Journal
Data to process
Drop items here
Properties
Preview
Geolocalization

Figure 22: Import data in the New Project

3.3 Chromatographic Extraction

Subsequently, the chromatographic extraction is the next step to follow. In Malcom the extraction of ion chromatogram tool enables the user to generate Extracted Ion Chromatogram from GC/MS data. To start this tool it is necessary to click on the “Extracted Ion Chromatogram” icon in the "Modules" dynamic tooltab.


Schlumberger Malcom
Project: Thesis_AK
Workspace: Malcom_1 workspace
Project
Modules
Workflows
Options
Chromatographic extraction
Chromatographic extraction
Extract ions from MS or MSMS analyses
Drop a MS or MSMS analysis here
Extract
Module
Chromatographic extraction
Actions
Project explorer
Search
Analyses

A2
ELAT_1_3_Sat
L001276_F1_IS_CHIRON_TIC
L001312_F1_IS_CHIRON_TIC
L00515_F1_IS_CHIRON_TIC
L00549_F1_IS_CHIRON_TIC
L00550_F1_IS_CHIRON_TIC
L00554_F1_IS_CHIRON_TIC
L00558_F1_IS_CHIRON_TIC
L00559_F1_IS_CHIRON_TIC
L00672_F1_IS_CHIRON_TIC
L00732_F1_IS_CHIRON_TIC
L00753_F1_IS_CHIRON_TIC
L00755_F1_IS_CHIRON_TIC
L00756_F1_IS_CHIRON_TIC
L00811_F1_IS_CHIRON_TIC
L00820_F1_IS_CHIRON_TIC
L00829_F1_IS_CHIRON_TIC
L00833_F1_IS_CHIRON_TIC
L00839_F1_IS_CHIRON_TIC
L00842_F1_IS_CHIRON_TIC
L00920_F1_IS_CHIRON_TIC
L00920_F1_IS_CHIRON_TIC_araio
Oil_Anal_23ppm_IS_8_2_2013_TIC
Oil_Anal_4ppm_IS_8_2_2013_TIC

Figure 23: Chromatographic extraction Wizard

The wizard shown above appears in a new dynamic tooltab dedicated to "Chromatographic extraction" and the necessary steps are:

- Drag and drop the GC/MS analysis from the project explorer in the specified box of the wizard,
- Choose one or several m/z ratios of the ion(s) to extract from the list located on the left side of the wizard (in this case m/z 85 and 191)

Finally click on the Extract icon ▶, the specified ion chromatograms are displayed in the appearing window. You can choose how to display the results, either in one window or in separated windows. You can also choose not to display the resulting chromatograms and just save it in the project explorer.

3.4 Create a Compound Database

After the Chromatographic Extraction is applied on all the analysis for each different ion all it’s left is to identify and quantify each compound.

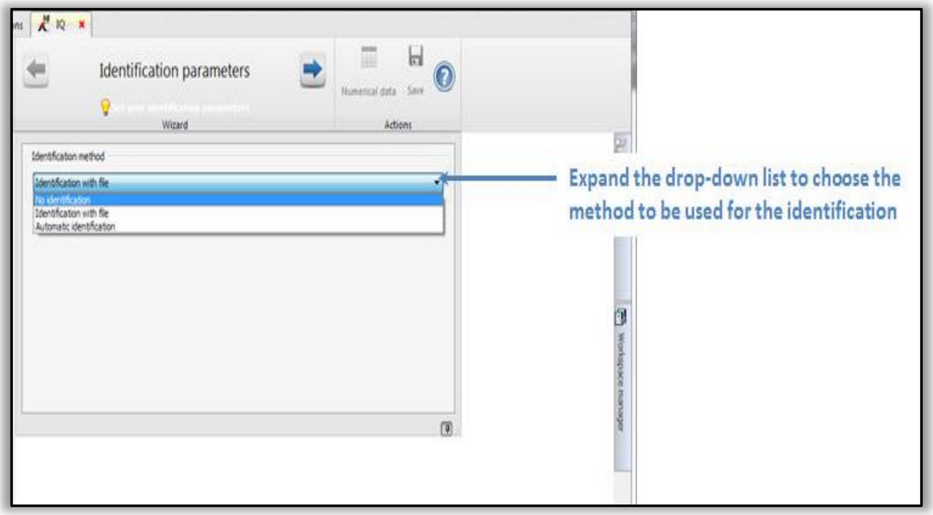
Malcom is delivered with a compounds database, which doesn't contain any compounds when the the software is first opened. In order to enter some compounds in the database, the user needs to perform some "manual" identification, either by entering the name of the compounds on the quantified chromatogram or by using a file containing information on retention time and name of the corresponding compounds. As soon as compounds have been referenced, they will be stored in the database and will be available for any further automatic identification.

3.5 Identification – Quantification

Since the desirable ions have been exracted it isimportant now to identify the containing compounds. In Malcom the objective of the identification tool is to automatically identified compounds of analyses. All new identification achieved "manually" by the user is automatically saved in a database.

The identification tool can be started by clicking on the IQ (Identification/Quantification) module icon in the "Modules" dynamic tooltab.The identification wizard opens in a new dynamic tooltab where the steps needen to be follow are described below:

The first step is to define signals to be processed by the "IQ" module. Drag and drop analyses to process from the project explorer into the "Working signal setup" window and click on the right arrow to go to the next step.



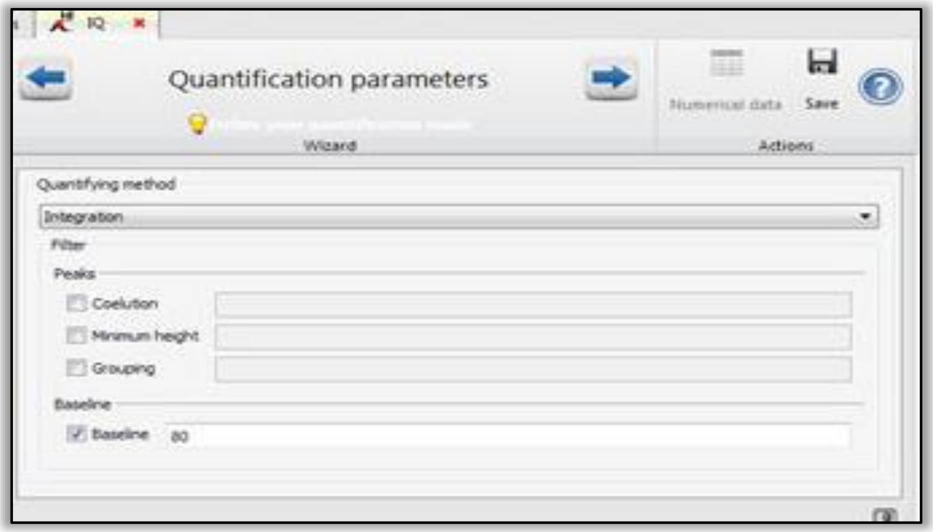
Expand the drop-down list to choose the method to be used for the identification

Figure 24: IQ wizard step 1 (after Malcom’s users manual)

Three different choices are available as identification method:

- No identification: this method has to be used when no identification is required
- Identification with file: this method has to be used when an external identification file is available for a specific chromatogram.
- Automatic identification: this method enables the user to perform identification on several chromatograms simultaneously and automatically by using a reference analysis where compounds have already been identified.

After choosing "No identification", click on to go to the next step, the user is now asked to define the analyses to process. The analyses is dragged and dropped to be process in the corresponding window and continue by clicking on .



Quantifying method

Integration

Filter

Peaks

□ Coelution

□ Minimum height

□ Grouping

Baseline

☑ Baseline 80

Figure 25: IQ wizard step 2 (after Malcom's users manual)

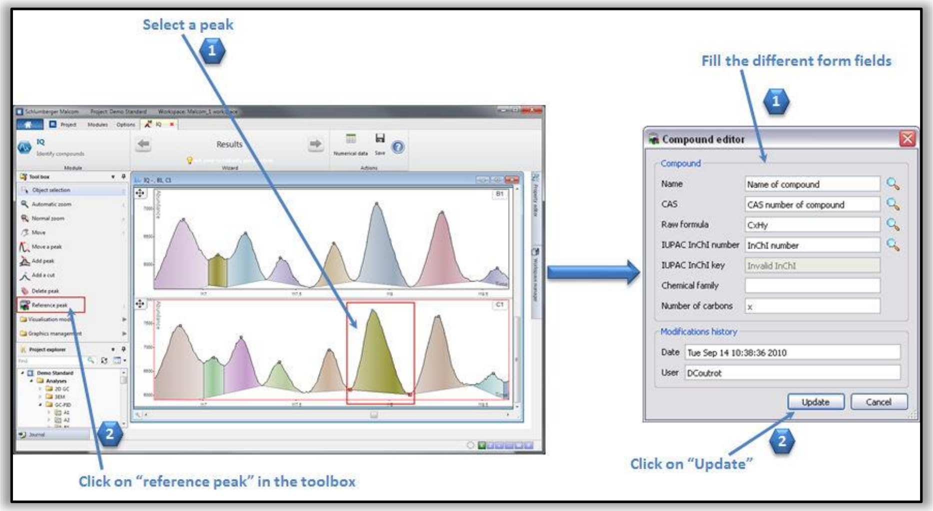
The second step of this process is to define the quantification parameters:

- integration
- integration with discontinuous baseline or
- deconvolution

With additional information about the filters (coelution, minimum height and grouping) and the baseline sensitivity.

The quantification parameter used at first is integration with discontinuous baseline, which provides the user integrated chromatograms. As opposed to the classical integration method, this method provides integration of the peaks with a discontinuous baseline, each peak is integrated with independent anchors from the neighboring peaks. In that case, there is no continuity in the baseline.

To identify peaks from this integration, select a peak and choose "Reference peak" in the toolbox. The compound editor automatically opens and the different fields of the forms can be filled. When clicking on "update", the label of the peak appears on the chromatogram. The compound is automatically added in the current database and will be searched in any future identification.



Select a peak

Click on "reference peak" in the toolbox

Fill the different form fields

Click on "Update"

Compound editor

Compound

Name: Name of compound

CAS: CAS number of compound

Raw formula: CxHy

IUPAC InChI number: InChI number

IUPAC InChI key: Invalid InChI

Chemical family:

Number of carbons: x

Modifications history

Date: Tue Sep 14 10:38:36 2010

User: DCoutrot

Figure 26: Identify a peak manually (after Malcom's users manual)

This "IQ" can be saved by clicking on the "Save" icon of the "IQ" dynamic tooltab. This type of identification is represented with the icon , while the automatic identifications is represented with the icon . This analysis can be reloaded by dragging and dropping it in the workspace and additional peaks can be referenced on this analysis if needed.

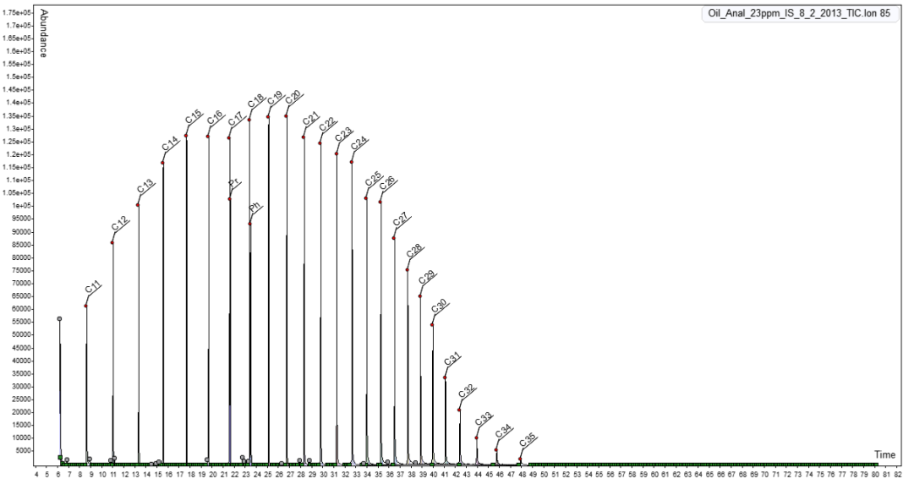
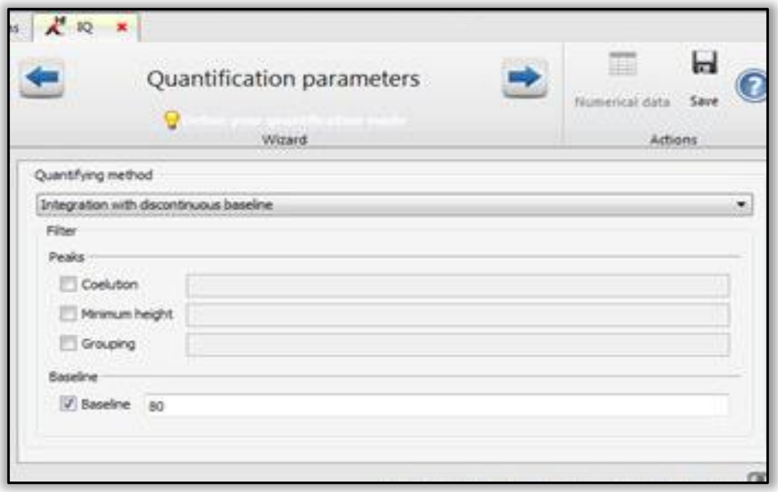


Figure 27: Identification of the Standard sample of the n-alkanes (m/z 85)

After indentifying and quantifying them, for the rest of the samples, the "identification with file" method was used. The identification with file relies on the analysis of the similarity between already processed chromatograms (called reference chromatograms) and one or several chromatograms to process (analyses to treat). The reference chromatograms include a list of peaks, which have been identified. The software will use this list to match the compounds of the chromatogram to process with the list of peaks previously identified in the reference chromatogram.

The similarity analysis is performed by comparing the relative retention times and abundances of all detected compounds in the reference chromatograms and in the chromatograms to process.

In order to get relevant results in the similarity analysis, it is necessary that the chromatograms are not too different in terms of retention time and abundance as the method relies on these values, which are the only values that can be used with single-channel chromatograms.



Quantifying method: Integration with discontinuous baseline

Filter

Peaks

□ Coelution

□ Minimum height

□ Grouping

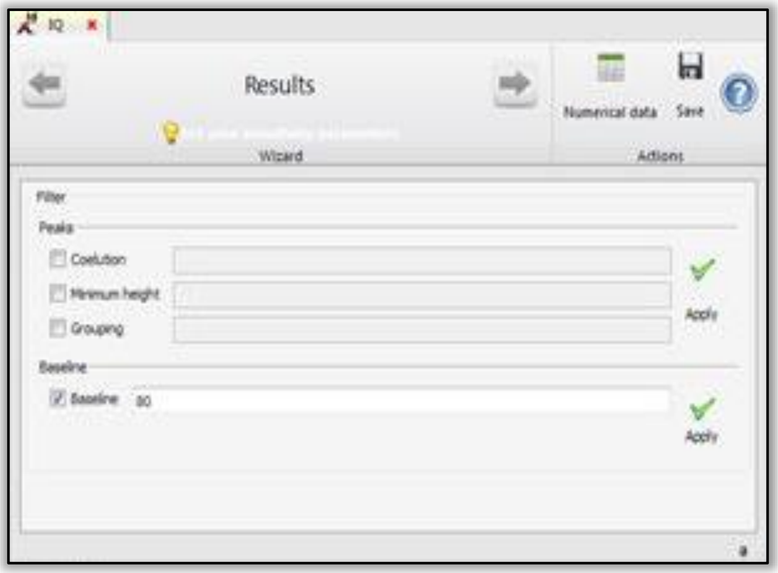
Baseline

☑ Baseline 80

Figure 28: Quantification parameters in the IQ wizard (after Malcom's users manual)

This process provides the user an integrated chromatogram together with an identification of the peaks which were described in the identification dataset. In this step, the user can define several sensitivity parameters: coelution, minimum height, grouping and baseline sensitivity. Once the values have been defined, click on "Apply" (Figure 29) to visualize the updated integration. In this study all the parameters were set to the default ones, specifically neither the coelution, nor the minimum height or grouping parameter were selected and the baseline was set to 80.

Using the identification file of the standard sample that has already been identified manually, the n-alkanes of the samples are identified automatically as presented in Figure 30.



Filter

Peaks

☐ Coelution

☐ Minimum height

☐ Grouping

Baseline

☑ Baseline 80

Figure 29: Application of IQ parameters

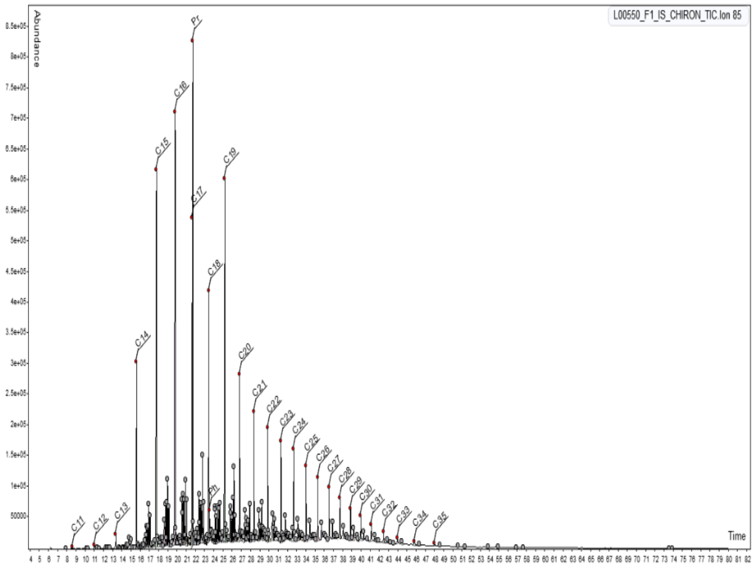
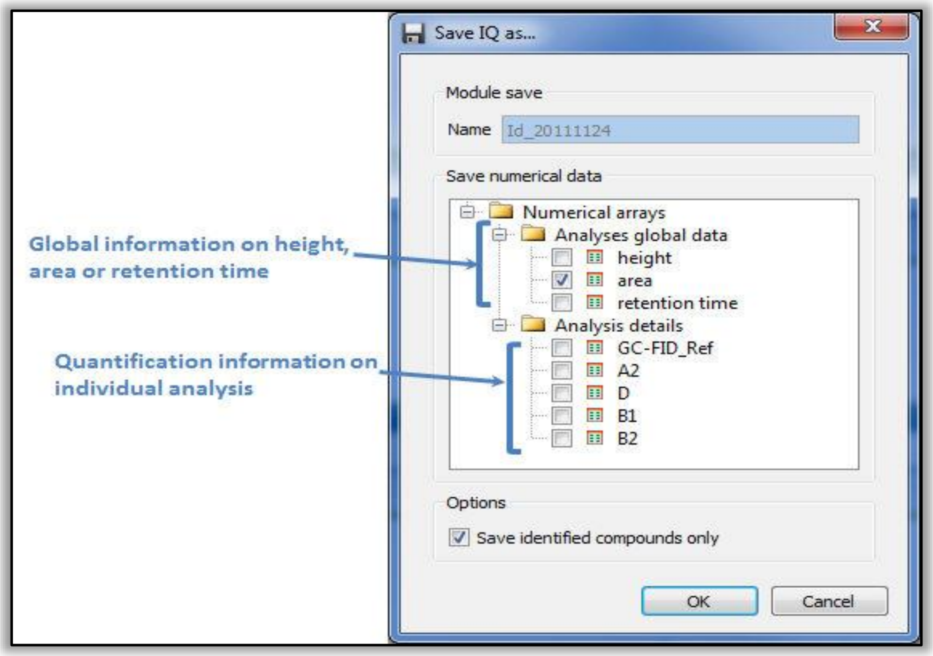


Figure 30: Identification with file of the n-alkanes

Once the targeted compounds have been identified on each signal, simply click on the right arrow icon ➡ in the "Results" identification wizard. All the results of the identification are gathered and displayed in the results tables, which are saved at the end of the identification process.

The identification will be automatically saved when clicking on the red cross of the identification dynamic tooltab. When closing the identification process, the identification assistant will ask you if you want to save your identification. You can save only the identified compounds or you can save all the detected peaks (by checking the "Save all compounds" box). The identification is saved as a graphical way under the identification tree and as numerical tables in the project explorer, under the "numerical arrays" tree.



Save IQ as...

Module save

Name: Id_20111124

Save numerical data

- Numerical arrays
  - Analyses global data
    - □ height
    - ☑ area
    - □ retention time
  - Analysis details
    - □ GC-FID_Ref
    - □ A2
    - □ D
    - □ B1
    - □ B2

Options

☑ Save identified compounds only

Global information on height, area or retention time
Quantification information on individual analysis
OK
Cancel

Figure 31: Save IQ numerical arrays (after Malcom's users manual)

The identification will be automatically saved when clicking on the red cross of the identification dynamic tooltab. When closing the identification process, the identification assistant will ask you if you want to save your identification. You can save only the identified compounds or you can save all the detected peaks (by checking the "Save all compounds" box). The identification is saved as a graphical way under the identification tree and as numerical tables in the project explorer, under the "numerical arrays" tree.

The same method was followed also for the hopanes (m/z 191). The identification of the hopanes was made manually using one of the given samples.Then it was used as the file to identify the hopanes for the rest of the samples. An indicative identification of the hopanes is presents in Figure 32.

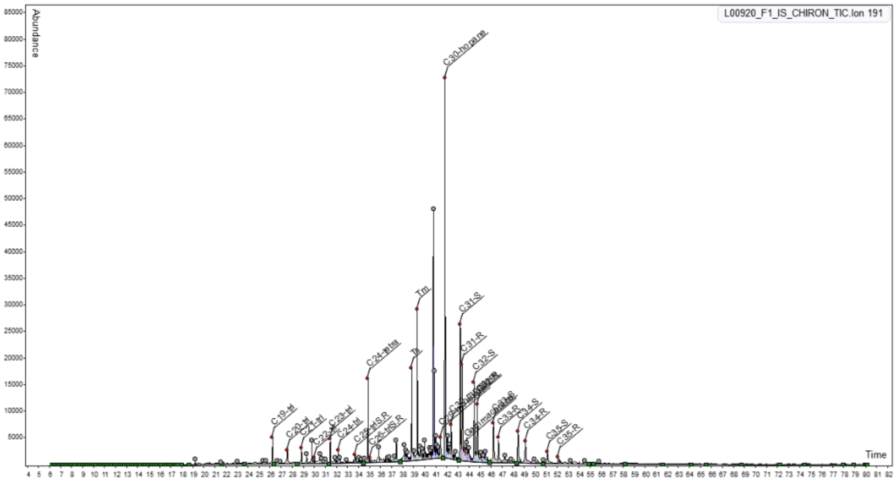


Figure 32: Identification with file of the hopanes

3.6 Editing the Numerical Arrays

After creating the numerical arrays for all samples and all the ions for each sample, the arrays were exported in Excel files, where they were modified. Specifically, summary tables were generated for all the samples in each different ion (see Appendix “Identification – Quantification”). These arrays were used later on in the chemometric tools.

3.7 Chemometric Tools

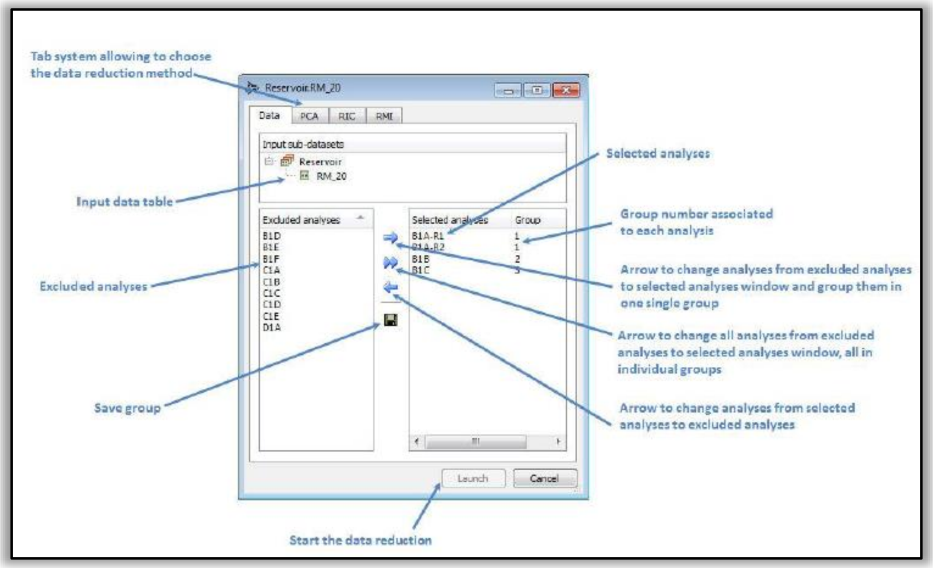
The Malcom Chemometrics tab enables the user to perform some statistical analysis from the data of the editor.

3.7.1 Data reduction - Principal Component Analysis –

The data reduction tool enables the user to reduce the number of descriptive variables to extract a few significant variables to simplify the interpretation of a complex dataset.

The "data reduction" tool can be started from the "Data editor" dynamic tooltab, which is available as soon as a data editor in opened in the workspace. This is achieved by dragging and dropping a numerical array from the project explorer into the workspace. Once a data editor is opened, click on the "Chemometrics" icon and choose data reduction in the drop-down list.

The user can work on specified individuals. For this, select the individuals on which you want to perform the data reduction in the excluded window and click on. If you want to select several individuals simultaneously, click on the analyses while holding down the "Ctrl" key. In that case, all the analyses selected at the same time will be considered in the same group



Excluded analyses
B1D
B1E
B1F
C1A
C1B
C1C
C1D
C1E
D1A

Selected analyses	Group
B1A-R1	1
B1A-R2	1
B1B	2
B1C	3

Reservoir.RM_20
Data
PCA
RIC
RMI
Input sub-datasets
Reservoir
RM_20
Launch
Cancel
Tab system allowing to choose the data reduction method
Input data table
Excluded analyses
Save group
Selected analyses
Group number associated to each analysis
Arrow to change analyses from excluded analyses to selected analyses window and group them in one single group
Arrow to change all analyses from excluded analyses to selected analyses window, all in individual groups
Arrow to change analyses from selected analyses to excluded analyses
Start the data reduction

Figure 33: Data Reduction Wizard (after Malcom's users manual)

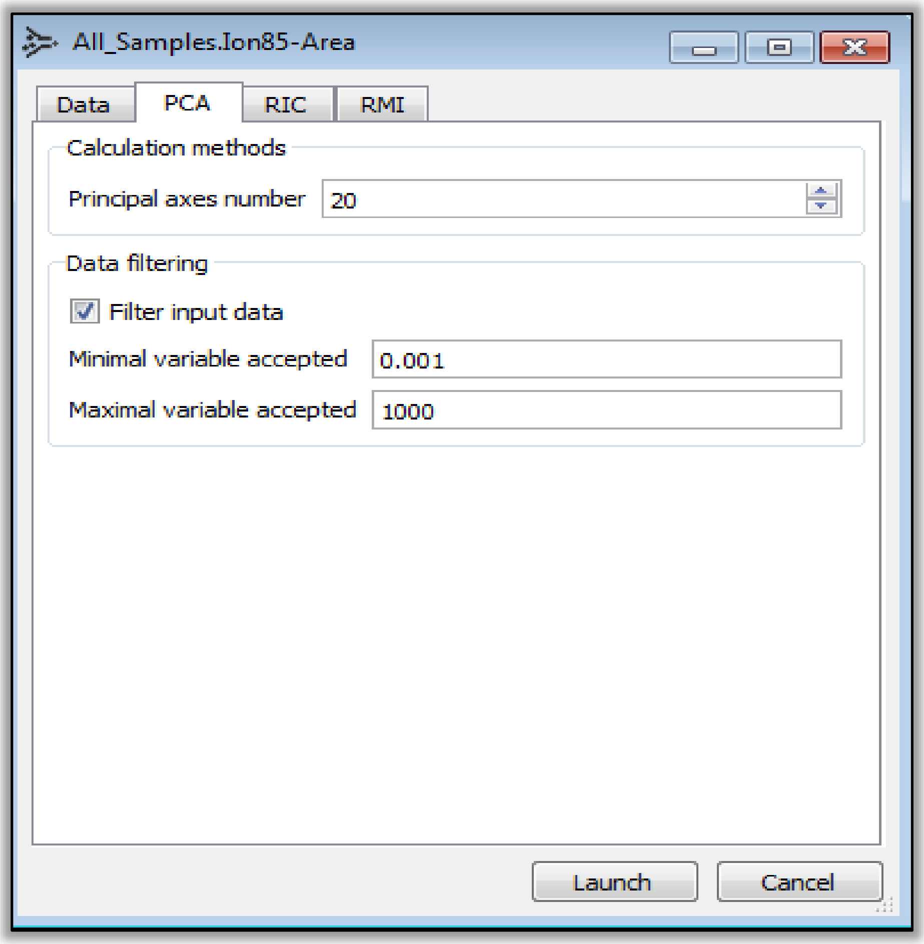
The groups that have been defined can be saved to be reused later by clicking on the save icon of the data reduction window. The group are saved as numeric arrays in the project explorer under the folder corresponding to the original input dataset.

Three different data reduction method are proposed in Malcom:

- Principal Component Analysis (PCA),
- Reduction by Inertia Constraining (RIC),
- Reduction by Maximization of the Inertia (RMI).

In this study, PCA was applied on our data and one single parameter is needed in this method. It concerns the number of principal components on which the individuals will be projected as it is presented in Figure 34, along with the method's parameters.

By clicking on "Launch" button, the results are displayed in the data editor. The last column displays the quality associated to each principal component and the results are all saved as arrays.



Calculation methods

Principal axes number: 20

Data filtering

☑ Filter input data

Minimal variable accepted: 0.001

Maximal variable accepted: 1000

Figure 34: Data reduction –PCA parameters

3.7.2 Ascending Hierarchical Classification (AHC)

In Malcom the classification tool enables the user to group individuals according to similarities of their descriptive variables. Three classification tools are implemented in Malcom:

- Ascendant Hierarchical Classification (AHC)
- Kohonen Neuronal Map
- Fuzzy logic

To start the analyses classification tool, the "Data editor" dynamic tooltab needs to be activated. This dynamic tooltab is activated as soon as a data editor is opened in the workspace. All the classification tools are accessible from the "Statistical tools" icons in the "Data editor" dynamic tooltab.

Fisrtly, click on "Analyses classification" icon in the drop down menu of the statistical tools icon :

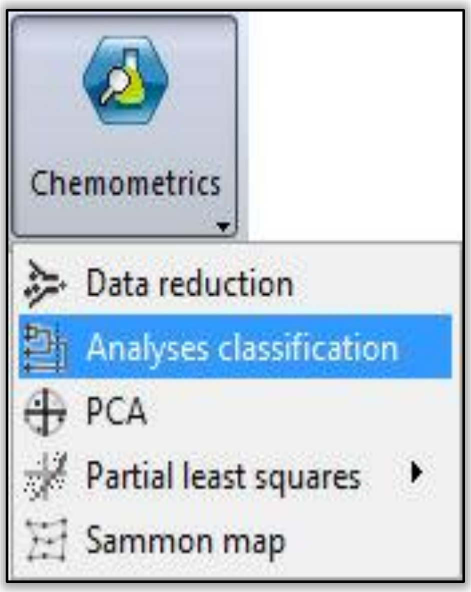
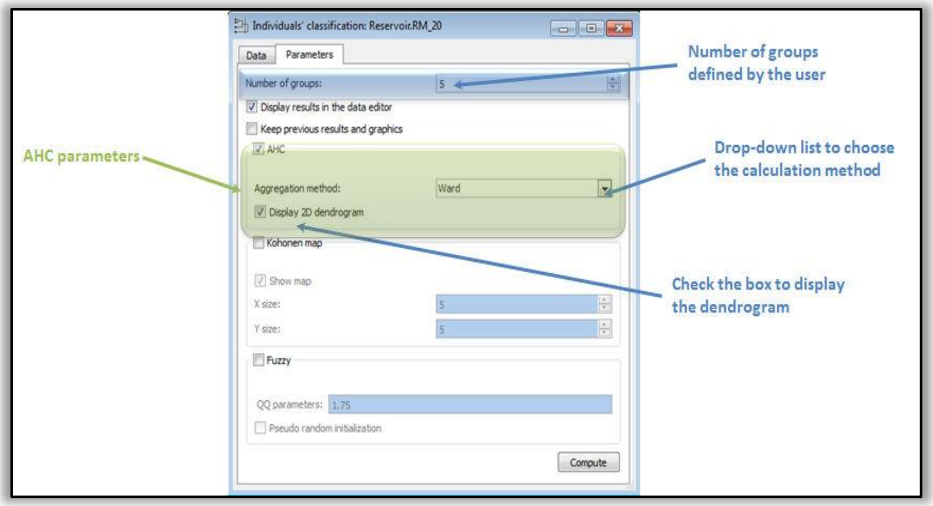


Figure 35: Chemometric tools selection window

The three methods can be started simultaneously in order to facilitate comparisons between the different methods, but in this study only the AHC method is used. The parameters need to be set for the method and are available in the "parameters" tab of the classification tools (Figure 36).

For the AHC, the number of groups is just an indication of the maximum number of groups. The display of the 2D dedrogram is also selected, while the “previous results and graphic” is not necessary in this study.



Individuals' classification: Reservoir.RM_20

Data | Parameters

Number of groups: 5

☑ Display results in the data editor

☐ Keep previous results and graphics

☑ AHC

Aggregation method: Ward

☑ Display 2D dendrogram

☐ Kohonen map

☑ Show map

X size: 5

Y size: 5

☐ Fuzzy

QQ parameters: 1.75

☐ Pseudo random initialization

Compute

AHC parameters

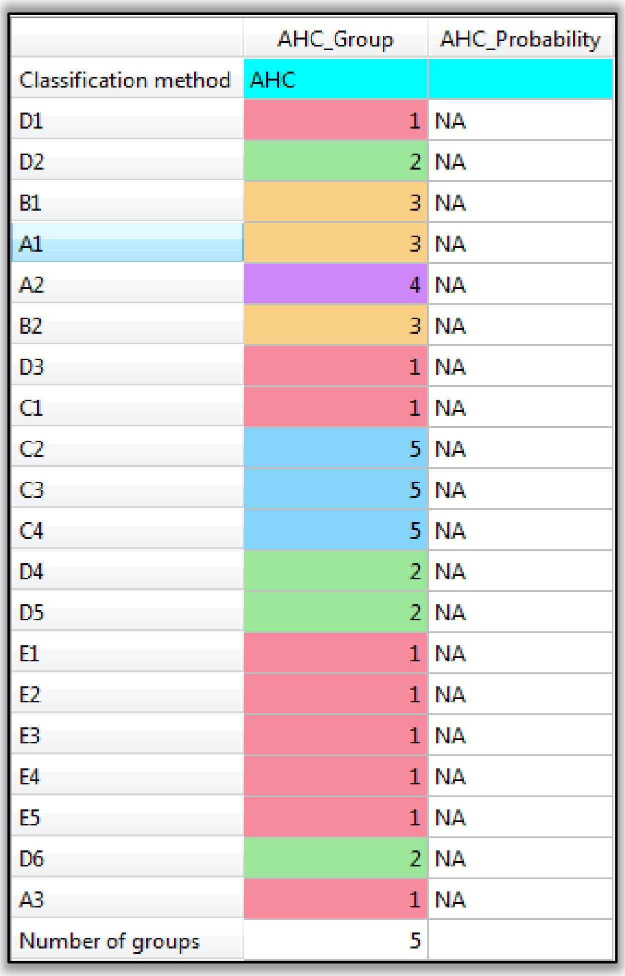
Number of groups defined by the user

Drop-down list to choose the calculation method

Check the box to display the dendrogram

Figure 36: AHC parametrs window (after Malcom's users manual)

The resulting groups are displayed as an additional tab in the data editor. This results can be saved by clicking on the "Save" icon in the menu bar.



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
B1	3	NA
A1	3	NA
A2	4	NA
B2	3	NA
D3	1	NA
C1	1	NA
C2	5	NA
C3	5	NA
C4	5	NA
D4	2	NA
D5	2	NA
E1	1	NA
E2	1	NA
E3	1	NA
E4	1	NA
E5	1	NA
D6	2	NA
A3	1	NA
Number of groups	5	

Figure 37: Indicative Results of AHC

3.7.3 Sammon Mapping

To visualize the Sammon projections from a data editor, expand the "Chemometrics tools" icon select the "Sammon map" tool.

All the individuals are represented on a two dimensions graph, while the distances between individuals are fitted in the most accurate way to the initial variables space.

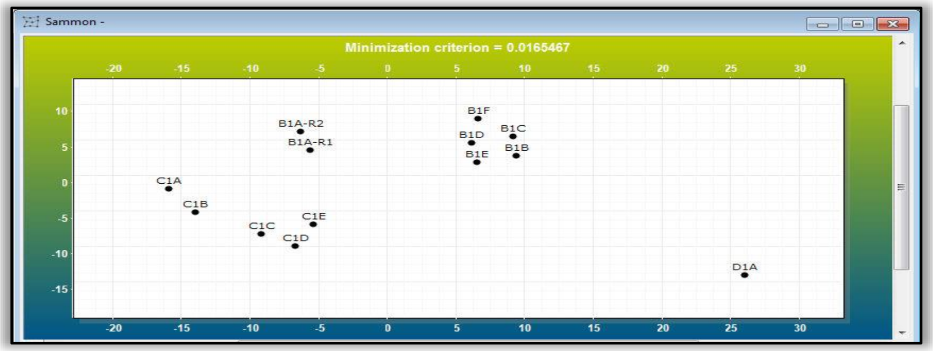


Figure 38: Representation of data in 2D sammon map

4 RESULTS

The data used in this study were obtained from the GC-MS analysis of the saturates fractions that was performed using an Agilent HP 7890/5975C system, with an HP-5 5% phenyl methylsiloxane column (30m x250 mm x0.25 mm), with the initial oven temperature set at 60°C, followed by a temperature ramp of 6°C/ min up to 300°C. The samples (1 μl) were injected through a split-splitless injector (pulsed-splitless mode, at 250°C) diluted (1/200) in ultra-pure hexane (SupraSolv®, Merck). The transfer line, MS source and quantrupole temperatures were set at 280°C, 230°C and 150°C respectively. The analysiswas carried-out in full scan ion detection mode (50-500 amu). Peak areas of n-alkanes (m/z 85) were determined from the Malcom Software as well as the respective areas for hopanes biomarkers (m/z 191).

4.1 Available Data

It was necessary to create a Compound Database for all the compounds that will be used for identification. For the n-alkanes a standard sample of a concentration of 23ppm was used for the identification. In addition, one of the 20 samples, specifically the sample A2 was used for the hopane identification. The identification method we used at first is the “no identification” one, since neither a compound database nor a reference file are available yet. The database will consist of all the compounds identified in the standard sample as well as the sample A2, thus it consists of the n-alkanes, isoprenoids and hopanes identified manually and they will later on be used as reference files for the IQ of the rest of the samples. Specifically :

- A standard analysis of 23ppm concentration, containing n-alkanes and isoprenoids, was used as a reference file for the identification of m/z 85.
- An oil sample, specifically the sample A2 was used as a reference file to identify the hopanes (m/z 191)

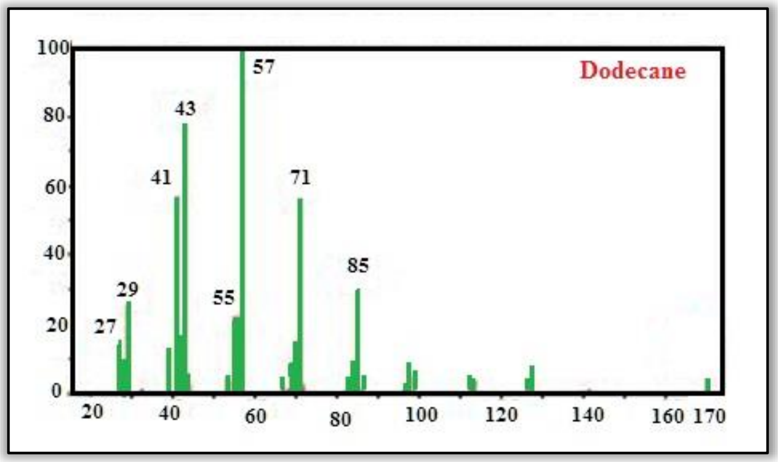


Figure 39: Mass spectra of dodecane

The n-alkanes are important for geochemical characterization because they are the first compounds used as biomarkers due to the analytical easiness and their high concentration in bitumens and oils. Also their distribution can provide information about their origin. An indicative mass spectra of a n-alkane, specifically the dodecane, is presented above in Figure 39.

On the other hand the identification of the hopanes is based on their characteristic ion (m/z 191) which is created by the separation of the A and B rings in their molecule (Figure 40). In Table the biomarkers used for the hopanes identification are presented.

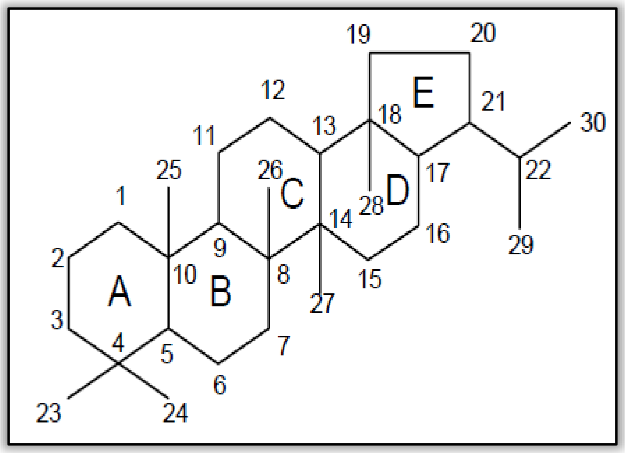


Figure 40: Hopane characteristic structure

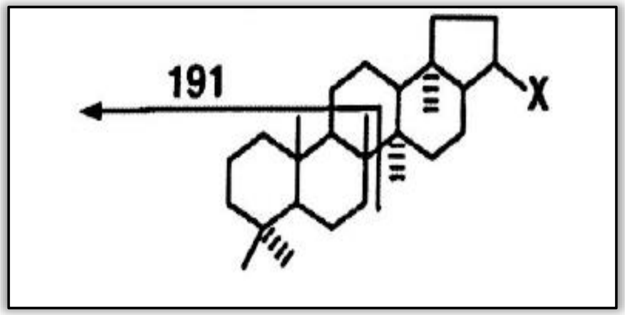


Figure 41: Creation of m/z 191

Biomarker	*Abbreviation*
C19-tricyclic terpane	C19-tri
C20- tricyclic terpane	C20-tri
C21- tricyclic terpane	C21-tri
C22- tricyclic terpane	C22-tri
C23- tricyclic terpane	C23-tri
C24- tricyclic terpane	C24-tri
C25- tricyclic terpane S,R	C25-triS,R
C24-tetracyclic terpane	C24-tetra
C26- tricyclic terpane S,R	C26-triS,R

18a(H)-21β,29,30-trisnorhopane	Ts
17a(H)-22,29,30-trisnorhopane	Tm
C29-moretane	C29-moretane
C30-hopane	C30-hopane
C30-moretane	C30-moretane
17α(H), 21β(H)-22S-homohopane	C31-S
17α(H), 21β(H)-22R-homohopane	C31-R
Gammacerane	Gammacerane
17α(H), 21β(H)-22S-bishomohopane	C32-S
17α(H), 21β(H)-22R-bishomohopane	C32-R
17α(H), 21β(H)-22S-trishomohopane	C33-S
17α(H), 21β(H)-22R-trishomohopane	C33-R
17α(H), 21β(H)-22S-tetrakishomohopane	C34-S
17α(H), 21β(H)-22R-tetrakishomohopane	C34-R
17α(H), 21β(H)-22S-pentakishomohopane	C35-S
17α(H), 21β(H)-22R-pentakishomohopane	C35-R

Table 2: List of the hopane biomarkers used in this study

The manually identified n-alkanes of the standard sample and the hopanes of the A2 sample are presented in the following figures.

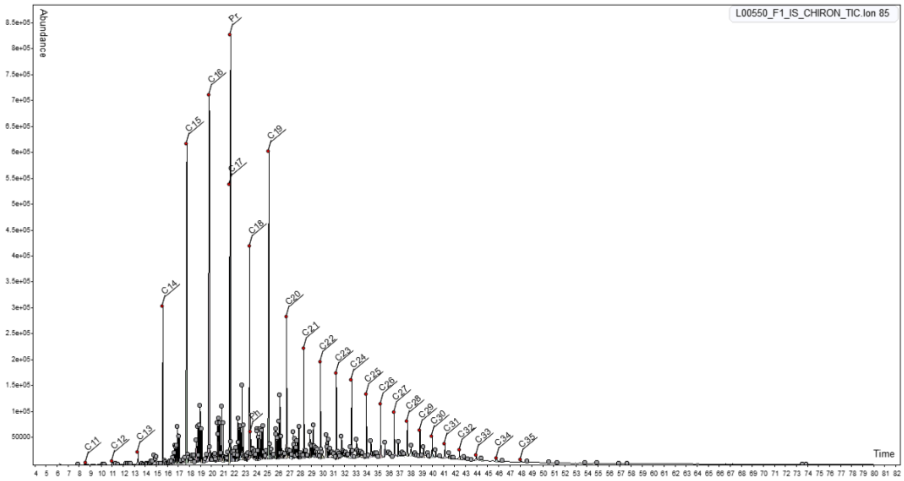


Figure 42: Identified n-alkanes of sample A2

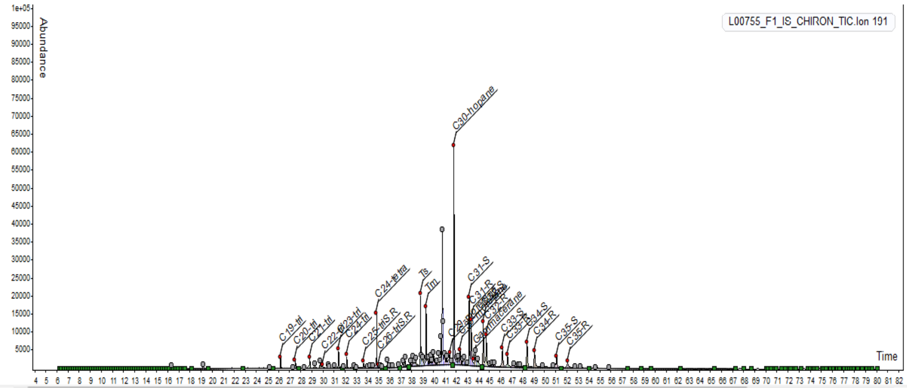


Figure 43: Identified hopanes of sample D2

As mentioned before, the results of the Identification – Quantification step of the Malcom, were exported in Excel files and summary tables of all the samples' area were generated. On these files several classification methods were applied. Following the results of each method used in Malcom are presented.

4.2 Principal Component Analysis (PCA)

Initially PCA was applied in three different datasets that were created. The values that were used are the original ones (the areas of the compounds), since PCA uses its own normalization: the mean value is subtracted from the data and then division of the data follows with the standard deviation. The three datasets (datasets 1-3) include:

1. Variables derived from SFH compositional data, specifically the areas of the whole range of the n-alkanes (C_{11}-C_{35}) and isoprenoids (m/z 85).
2. Variables derived from SFH compositional data, specifically the areas of the range of C_{15} to C_{30} n-alkanes and isoprenoids (m/z 85).
3. Variables derived from SFH compositional data, specifically the areas of the hopanes (m/z 191). It has to be noted that the oil family B was excluded because of its lack on biomarkers.

At first PCA was applied in the n-alkanes (m/z 85) data (dataset 1) and the classification of the oil families was sufficient, as presented in Figure 44.

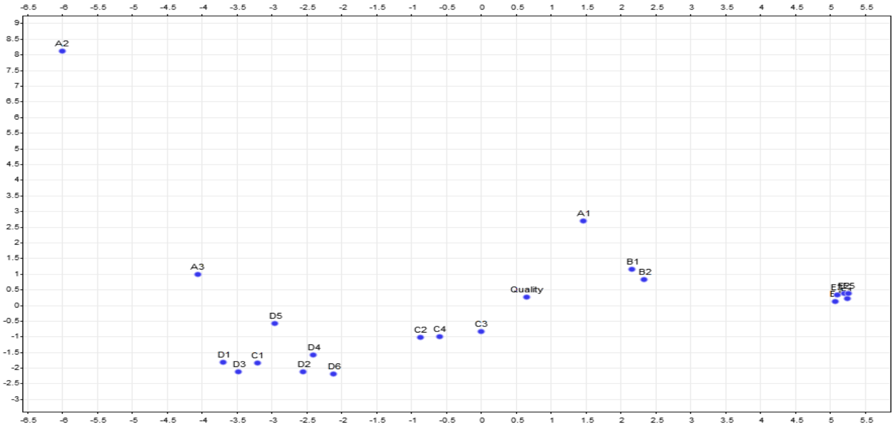
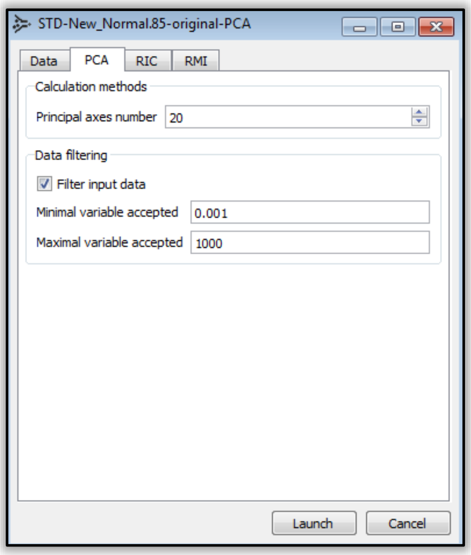


Figure 44: Crossplot of the first two PCs for dataset 1 (m/z 85)

The parameters chosen for this model are shown in Figure 45 and they remain the same for all the datasets used in PCA. For more details, all the Principal Components are presented in the Appendix "Principal Component Analysis (PCA)".



Calculation methods

Principal axes number 20

Data filtering

☑ Filter input data

Minimal variable accepted 0.001

Maximal variable accepted 1000

Figure 45: Parameters of Data reduction - PCA

Next the data of the n-alkanes were reduced, since this time we used only the selected areas of the hydrocarbon range of C_{15} to C_{35}. Likewise, the separation of the oil families was also sufficient but this time the samples on the crossplot seem to be more spreaded (Figure 46).

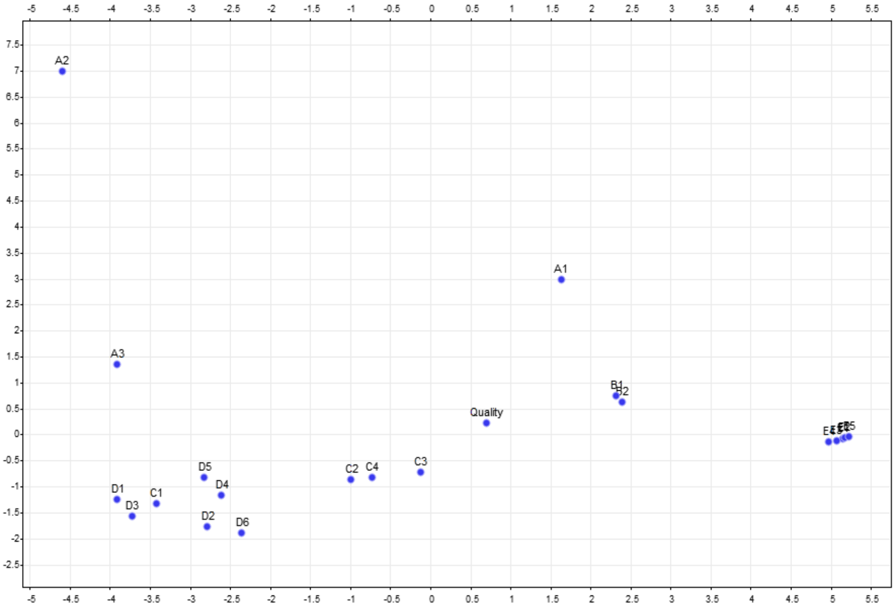


Figure 46: Crossplot of the first two PCs for dataset 2 (selected m/z 85)

The classification of the families in the third dataset (m/z 191) was efficient. The four different families were separated successfully as presented in Figure 47.

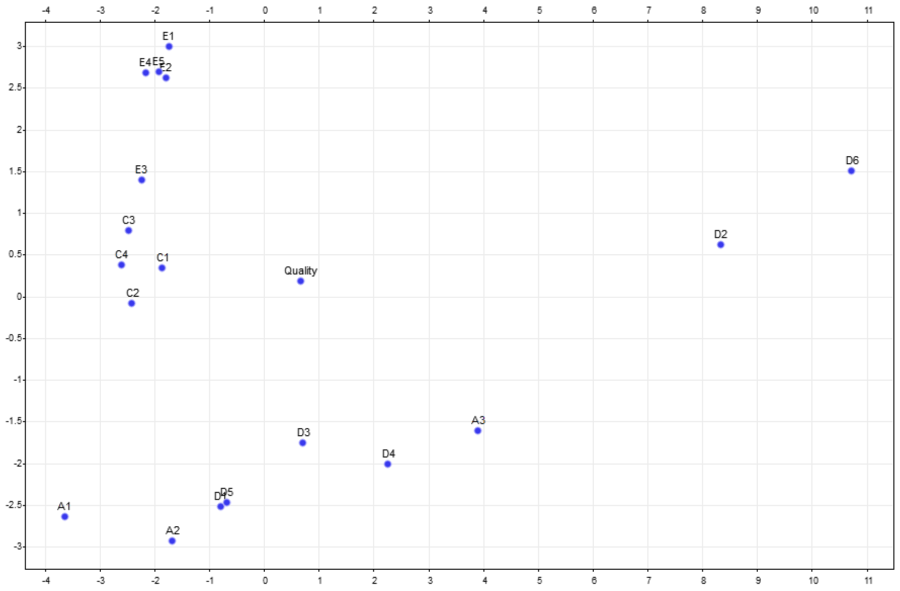


Figure 47: : Crossplot of the first two PCs for dataset 3 (m/z 191)

4.3 Ascendant Hierarchical Classification (AHC)

The AHC classification method was also applied on the samples. This classification method was repeated for 4 different aggregation methods: Ward, Link's Average, Link's Complete and Link's Simple aggregation methods. In this method the data were normalized manually, since the method itself does not provide any normalization. The normalization used is the same one that PCA uses, which means the mean value was subtracted from the data and later on the division with the standard deviation followed.

The next three datasets (datasets 4-6) created for the AHC method include:

4. Variables derived from SFH compositional data, specifically the normalized areas of the whole range of the n-alkanes (C_{11}-C_{35}) and isoprenoids (m/z 85).
5. Variables derived from SFH compositional data, specifically the normalized areas of the range of C_{15} to C_{30} n-alkanes and isoprenoids (m/z 85).
6. Variables derived from SFH compositional data, specifically the normalized areas of the hopanes (m/z 191). It has to be noted that the oil family B was excluded again because of its lack on biomarkers.

4.3.1 Ward Aggregation Method

At first AHC using the Ward's aggregation method was applied on the Datasets 4-6. For the Dataset 4 (m/z 85) the dedrogram as well as the results of the classification are presented in Figure 48 and Figure 49 respectively.

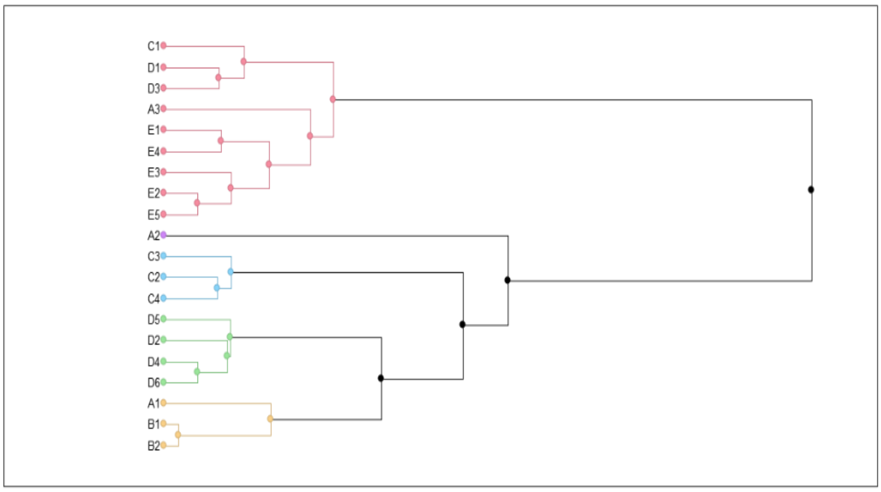
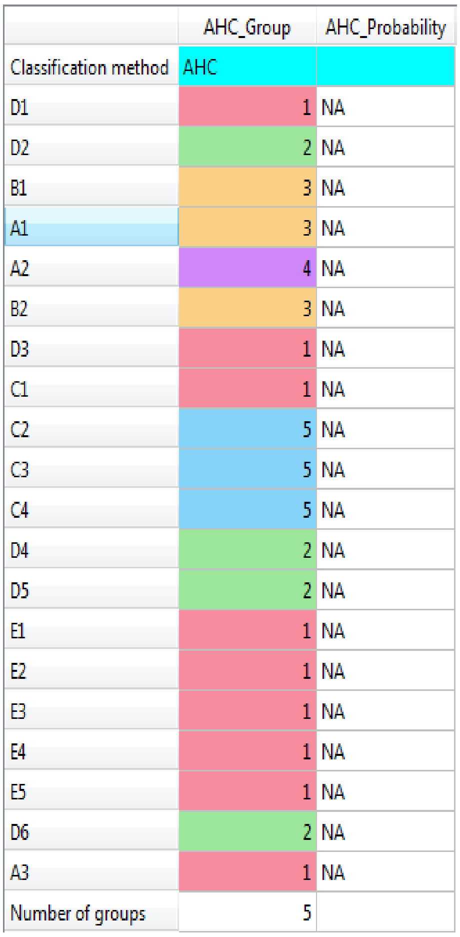


Figure 48: Dedrogram using the Ward aggregation method for Dataset 4



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
B1	3	NA
A1	3	NA
A2	4	NA
B2	3	NA
D3	1	NA
C1	1	NA
C2	5	NA
C3	5	NA
C4	5	NA
D4	2	NA
D5	2	NA
E1	1	NA
E2	1	NA
E3	1	NA
E4	1	NA
E5	1	NA
D6	2	NA
A3	1	NA
Number of groups	5	

Figure 49: Results of Ward aggregation method for Dataset 4

The separation of the oil families seems to be sufficient, but in a few samples there seems to be a discrepancy.

Next, AHC was applied on the Dataset 5 (selected m/z 85) and the dedrogram as well as the results of the method are presented in the following figures.

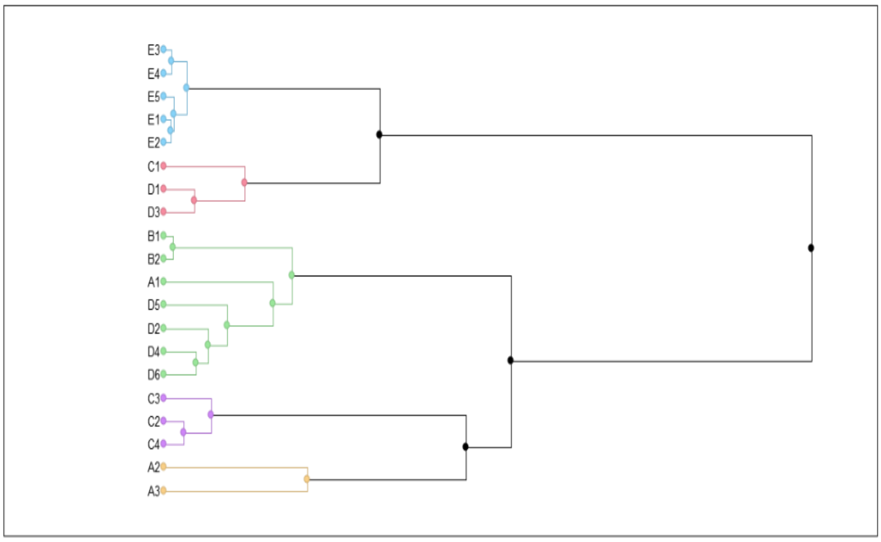
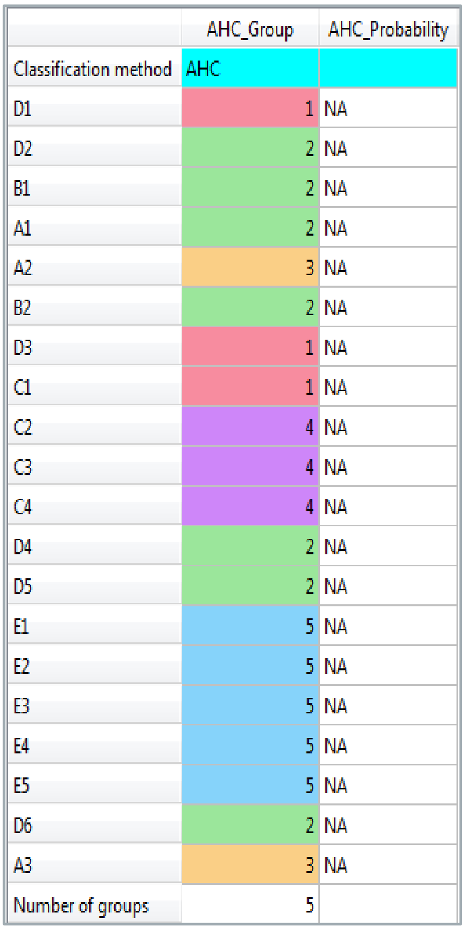


Figure 50: Dedrogram using the Ward aggregation method for Dataset 5



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
B1	2	NA
A1	2	NA
A2	3	NA
B2	2	NA
D3	1	NA
C1	1	NA
C2	4	NA
C3	4	NA
C4	4	NA
D4	2	NA
D5	2	NA
E1	5	NA
E2	5	NA
E3	5	NA
E4	5	NA
E5	5	NA
D6	2	NA
A3	3	NA
Number of groups	5	

Figure 51: Results of Ward aggregation method for Dataset 5

As well as before, the separation of the families seems to be sufficient, with some of the samples deviating from their original class.

Lastly, the AHC for the Dataset 6 (m/z 191) took place using again the Ward's aggregation method. As mentioned before, this dataset contains 4 out of the 5 oil families, since the oil family B has been removed.

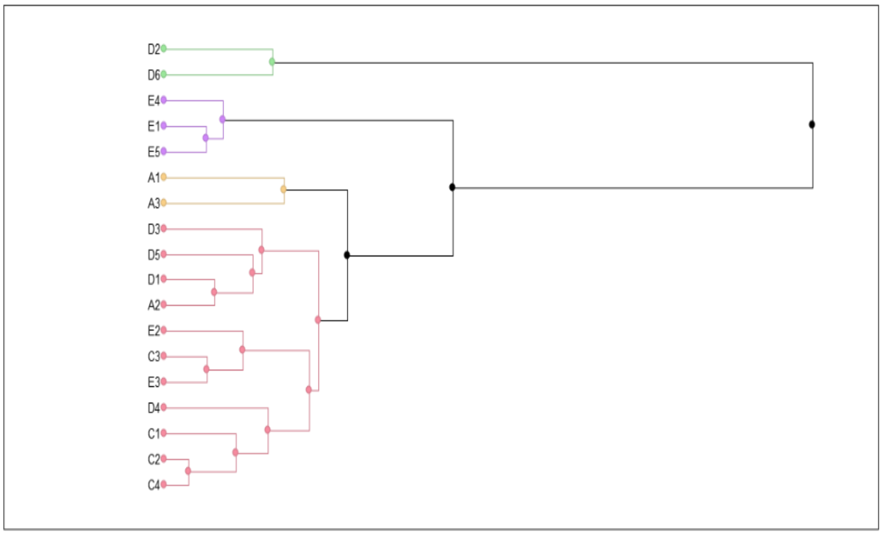
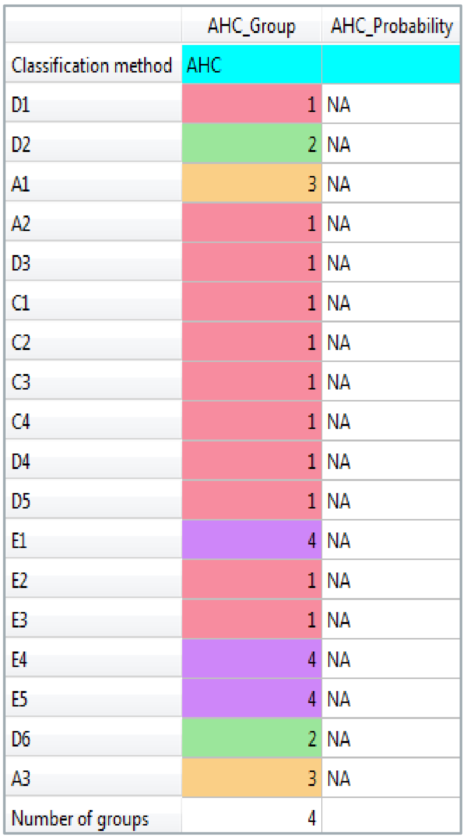


Figure 52: Dedrogram using the Ward aggregation method for Dataset 6



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
A1	3	NA
A2	1	NA
D3	1	NA
C1	1	NA
C2	1	NA
C3	1	NA
C4	1	NA
D4	1	NA
D5	1	NA
E1	4	NA
E2	1	NA
E3	1	NA
E4	4	NA
E5	4	NA
D6	2	NA
A3	3	NA
Number of groups	4	

Figure 53: Results of Ward aggregation method for Dataset 5

This time the classification of the oil families is not as sufficient as desired. That happens because the hopanes (m/z 191) contain less information than the n-alkanes (m/z 85).

4.3.2 Average Links Aggregation Method

Next AHC was applied on the Datasets 4-6. using the Link's Average aggregation method For the Dataset 4 (m/z 85) the dedrogram as well as the results of the classification are presented in Figure 54 and Figure 55 respectively.

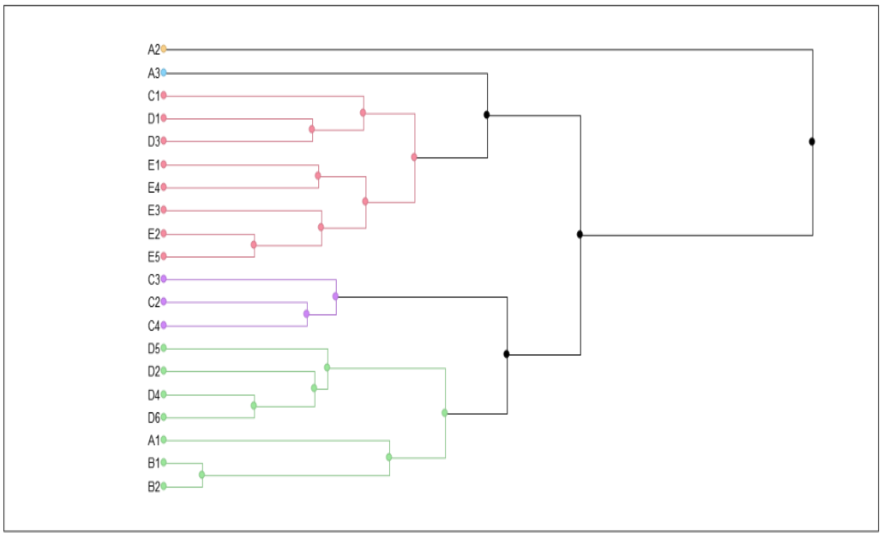
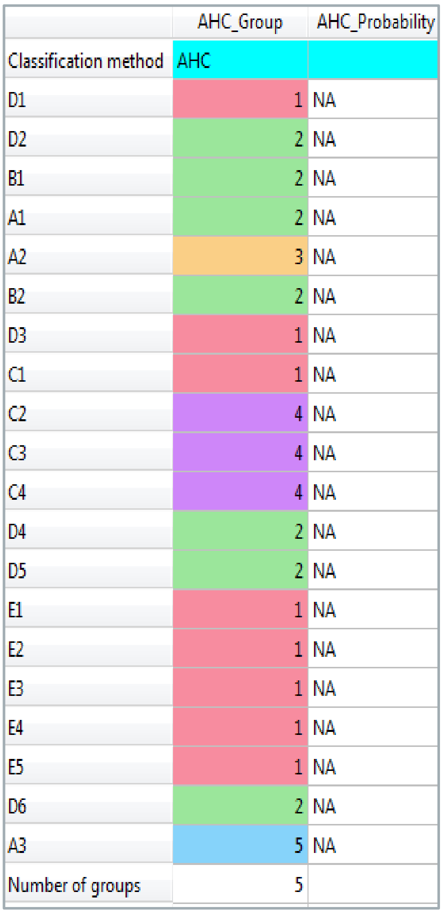


Figure 54: Dedrogram using the Link's Average aggregation method for Dataset 4



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
B1	2	NA
A1	2	NA
A2	3	NA
B2	2	NA
D3	1	NA
C1	1	NA
C2	4	NA
C3	4	NA
C4	4	NA
D4	2	NA
D5	2	NA
E1	1	NA
E2	1	NA
E3	1	NA
E4	1	NA
E5	1	NA
D6	2	NA
A3	5	NA
Number of groups	5	

Figure 55: Results of Link's Average aggregation method for Dataset 4

As we can see from the figures above, this method gives slightly worse results compared to the previous method. This happens because the Link's average aggregation method produced a higher number of inaccuracies in the oil families classification.

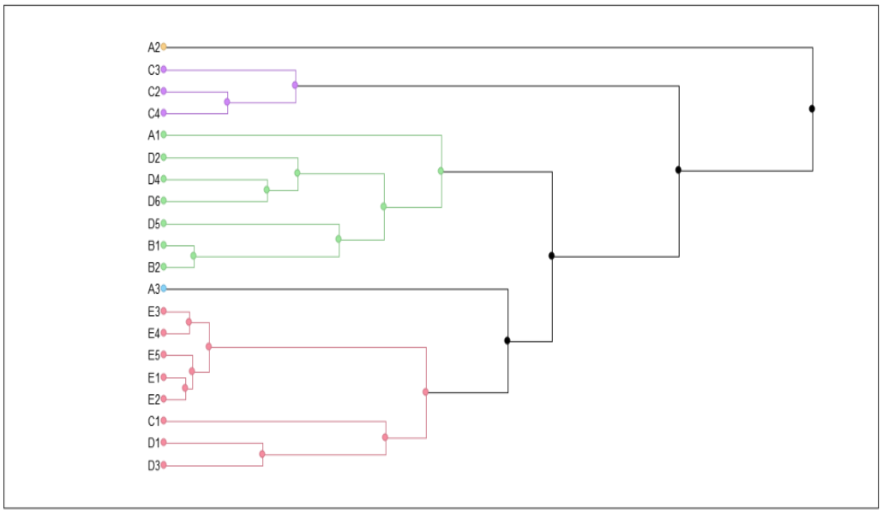
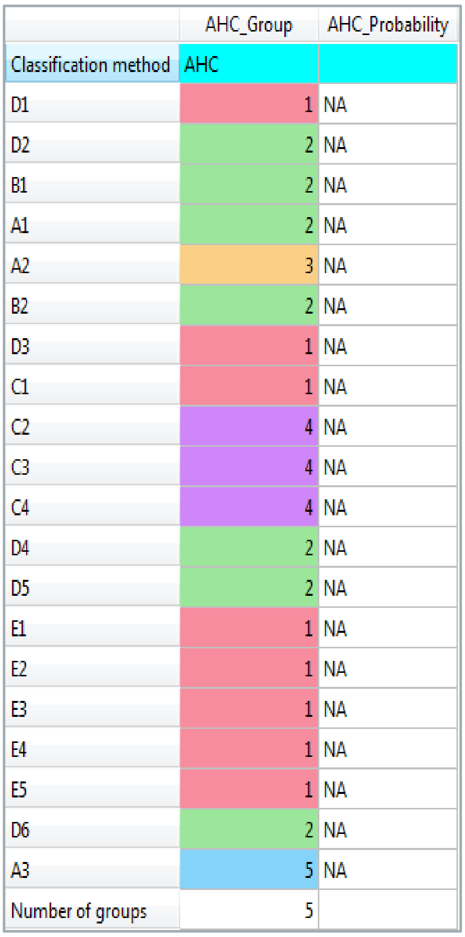


Figure 56: Dedrogram using the Link's Average aggregation method for Dataset 5



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
B1	2	NA
A1	2	NA
A2	3	NA
B2	2	NA
D3	1	NA
C1	1	NA
C2	4	NA
C3	4	NA
C4	4	NA
D4	2	NA
D5	2	NA
E1	1	NA
E2	1	NA
E3	1	NA
E4	1	NA
E5	1	NA
D6	2	NA
A3	5	NA
Number of groups	5	

Figure 57: Results of Link's Average aggregation method for Dataset 5

The same method, using the dataset of the selected n-alkanes (range of C_{15}-C_{30}), resulted in an identical classification to the classification of Dataset 4. It has to be noted that even though the classification lead to the same result, the dedrograms of the two datasets, are not the same.

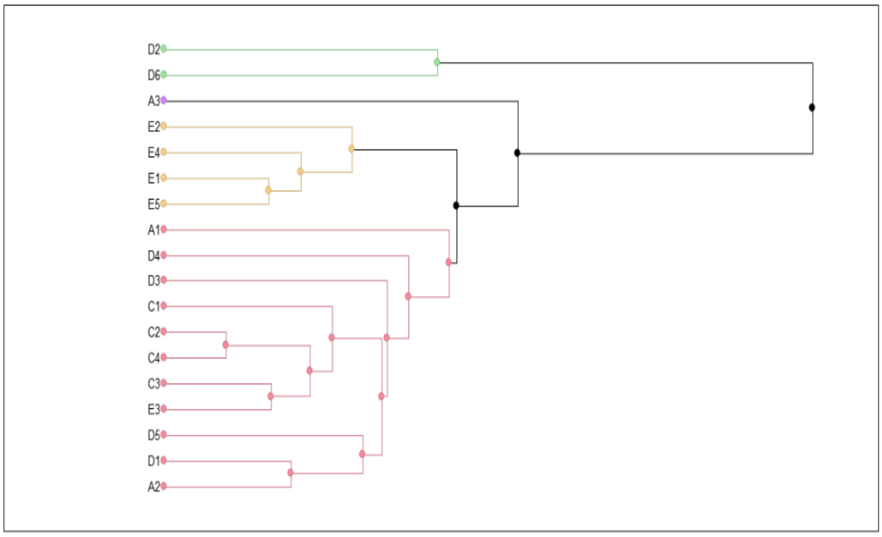
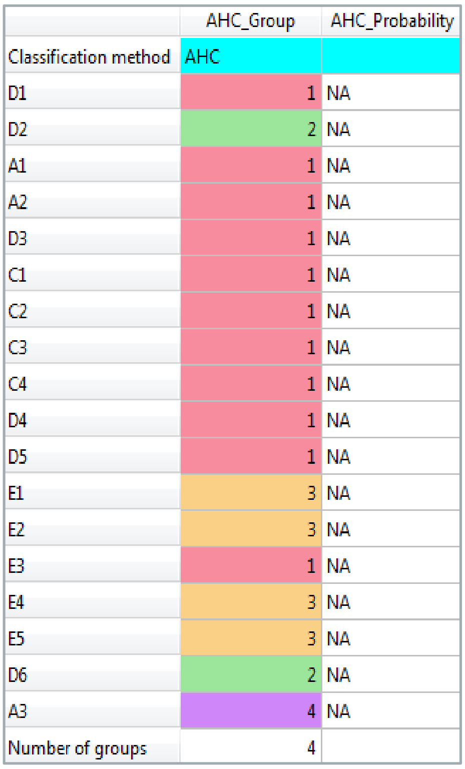


Figure 58: Dedrogram using the Link's Average aggregation method for Dataset 6



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
A1	1	NA
A2	1	NA
D3	1	NA
C1	1	NA
C2	1	NA
C3	1	NA
C4	1	NA
D4	1	NA
D5	1	NA
E1	3	NA
E2	3	NA
E3	1	NA
E4	3	NA
E5	3	NA
D6	2	NA
A3	4	NA
Number of groups	4	

Figure 59: Results of Link's Average aggregation method for Dataset 6

As we can see from the figures above, this method gives slightly worse results than the previous two datasets. This happens because the Link's average aggregation method produced a higher number of errors in the oil families classification.

4.3.3 Complete Links Aggregation Method

In this method, the number of errors in the oil family classification is the exact same as the Link's average classification methon. This occurs in both Datasets 4 and 5 as shown in Figures 60-63.

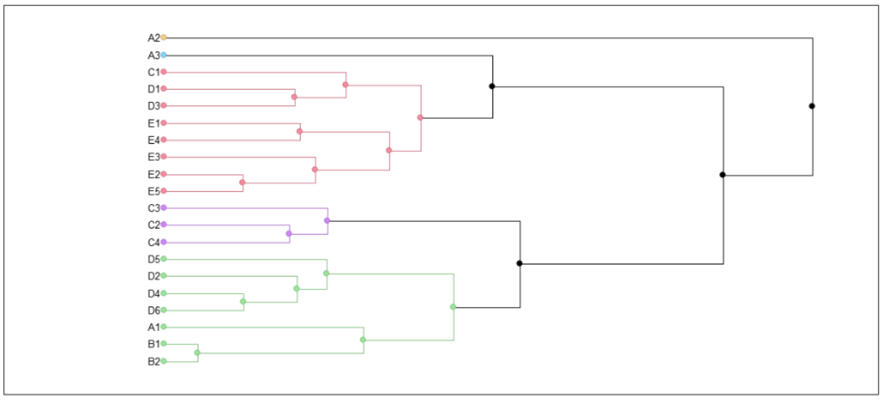
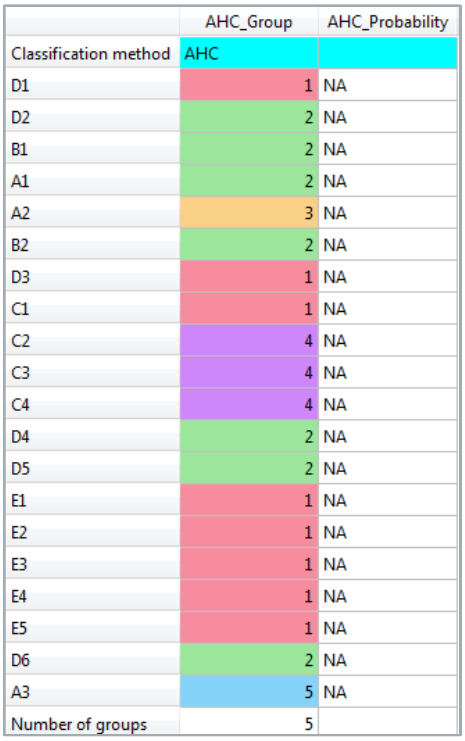


Figure 60: Dedrogram using the Link's Complete aggregation method for Dataset 4



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
B1	2	NA
A1	2	NA
A2	3	NA
B2	2	NA
D3	1	NA
C1	1	NA
C2	4	NA
C3	4	NA
C4	4	NA
D4	2	NA
D5	2	NA
E1	1	NA
E2	1	NA
E3	1	NA
E4	1	NA
E5	1	NA
D6	2	NA
A3	5	NA
Number of groups	5	

Figure 61: Results of Link's Complete aggregation method for Dataset 4

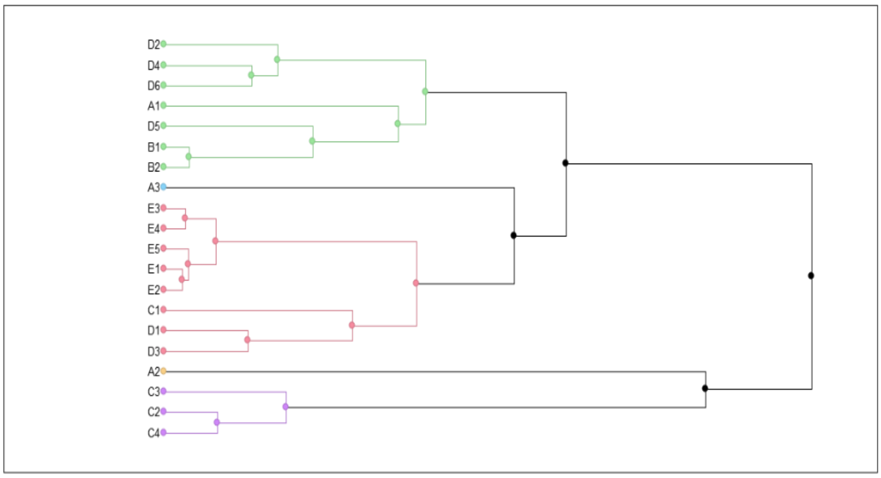
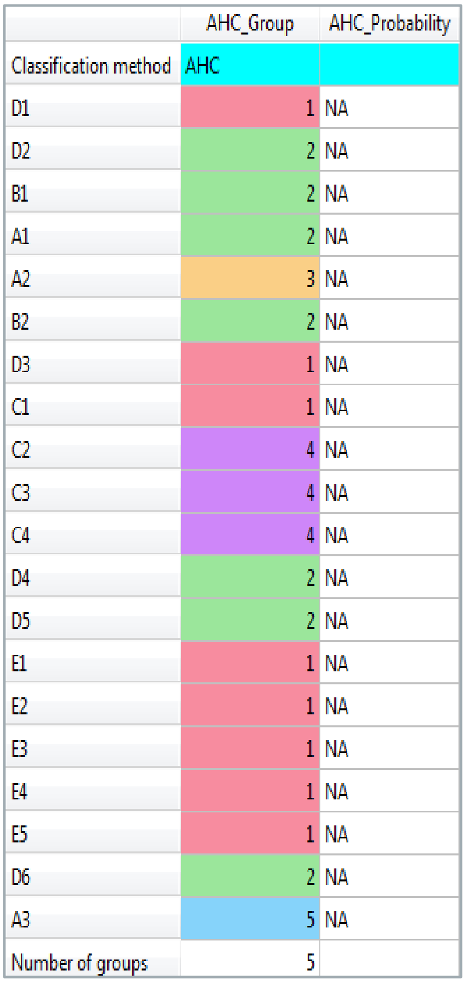


Figure 62: Dedrogram using the Link's Complete aggregation method for Dataset 5



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
B1	2	NA
A1	2	NA
A2	3	NA
B2	2	NA
D3	1	NA
C1	1	NA
C2	4	NA
C3	4	NA
C4	4	NA
D4	2	NA
D5	2	NA
E1	1	NA
E2	1	NA
E3	1	NA
E4	1	NA
E5	1	NA
D6	2	NA
A3	5	NA
Number of groups	5	

Figure 63: Results of Link's Complete aggregation method for Dataset 5

It is known that the Link's complete aggregation method works better with a small number of individuals. The result of the classification confirms this statements, since the number of errors is the second lowest (Figures 64-65).

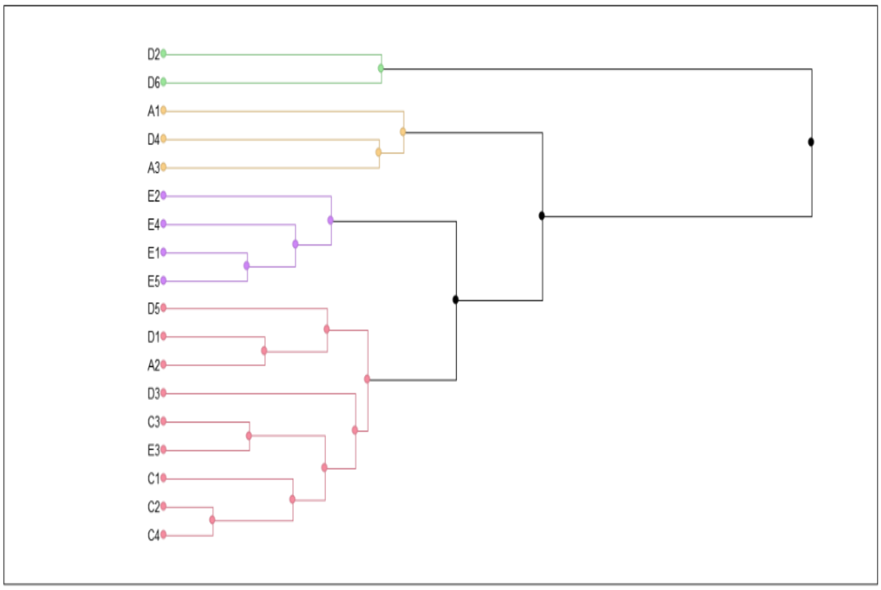
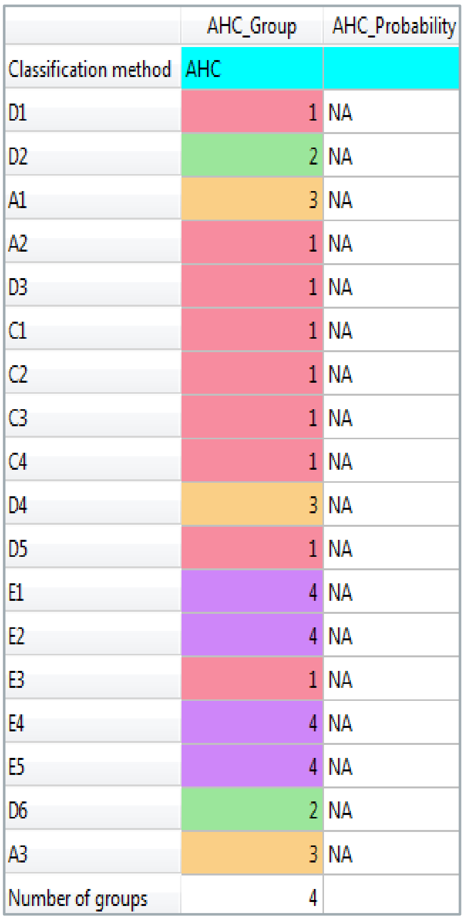


Figure 64: Dedrogram using the Link's Complete aggregation method for Dataset 6



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
A1	3	NA
A2	1	NA
D3	1	NA
C1	1	NA
C2	1	NA
C3	1	NA
C4	1	NA
D4	3	NA
D5	1	NA
E1	4	NA
E2	4	NA
E3	1	NA
E4	4	NA
E5	4	NA
D6	2	NA
A3	3	NA
Number of groups	4	

Figure 65: Results of Link's Complete aggregation method for Dataset 6

4.3.4 Simple Links Aggregation Method

The first dataset (4) used in the Link's simple aggregation method gave the same amount of errors as the previous Link's aggregation methods.

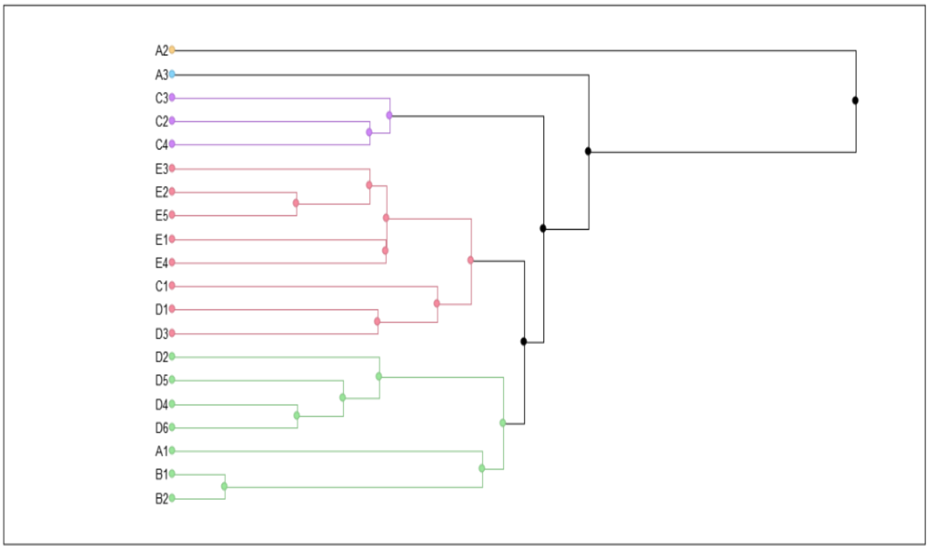
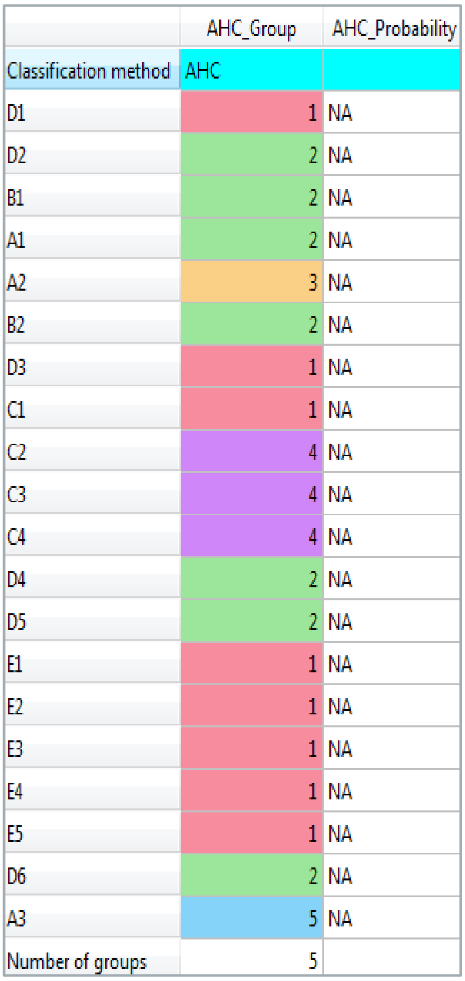


Figure 66: Dedrogram using the Link's Simple aggregation method for Dataset 4



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
B1	2	NA
A1	2	NA
A2	3	NA
B2	2	NA
D3	1	NA
C1	1	NA
C2	4	NA
C3	4	NA
C4	4	NA
D4	2	NA
D5	2	NA
E1	1	NA
E2	1	NA
E3	1	NA
E4	1	NA
E5	1	NA
D6	2	NA
A3	5	NA
Number of groups	5	

Figure 67: Results of Link's Simple aggregation method for Dataset 4

It should be underlined that this method works at its finest for a high number of individuals. This being said, it was expected to have more errors due to the fact that only 20 samples were used in this study.

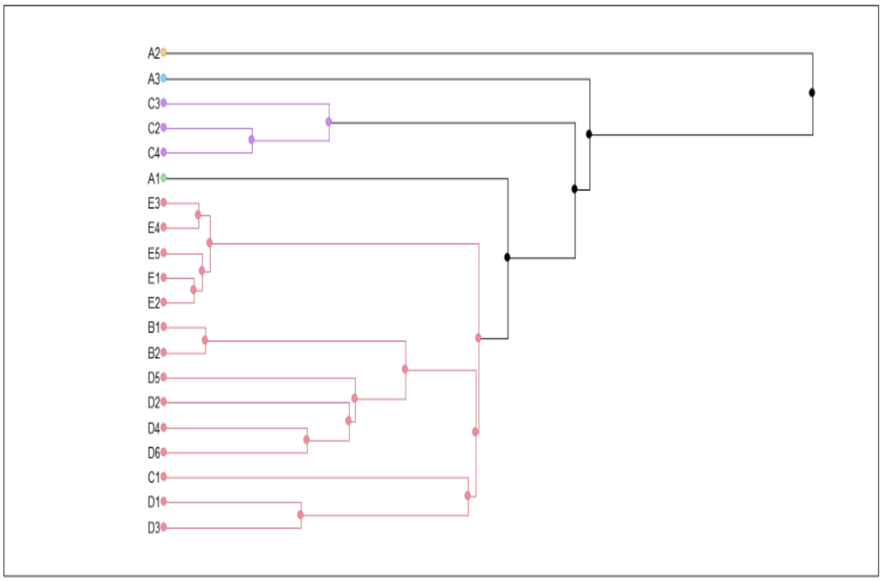
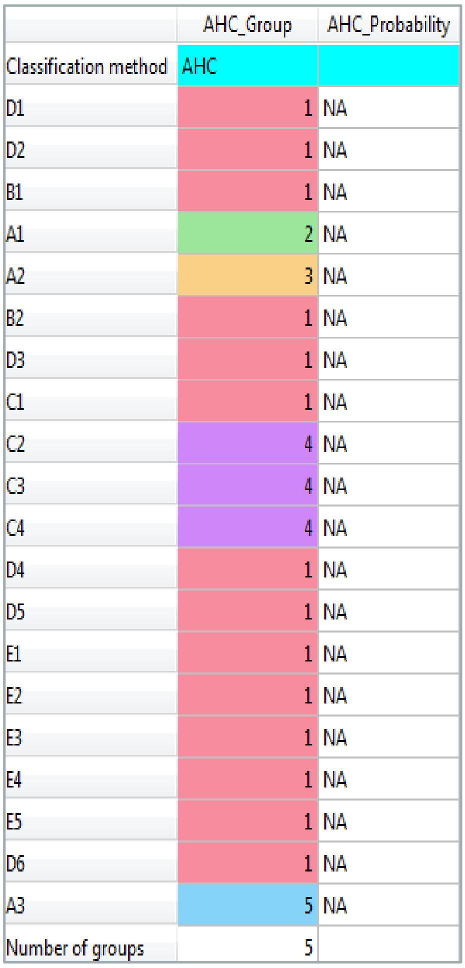


Figure 68: Dedrogram using the Link's Simple aggregation method for Dataset 5



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	1	NA
B1	1	NA
A1	2	NA
A2	3	NA
B2	1	NA
D3	1	NA
C1	1	NA
C2	4	NA
C3	4	NA
C4	4	NA
D4	1	NA
D5	1	NA
E1	1	NA
E2	1	NA
E3	1	NA
E4	1	NA
E5	1	NA
D6	1	NA
A3	5	NA
Number of groups	5	

Figure 69: Results of Link's Simple aggregation method for Dataset 5

Using this last dataset (6), Link's simple aggregation method produced the highest rate of errors. As displayed in the Figure 70 and Figure 71 it has the worst classification by far.

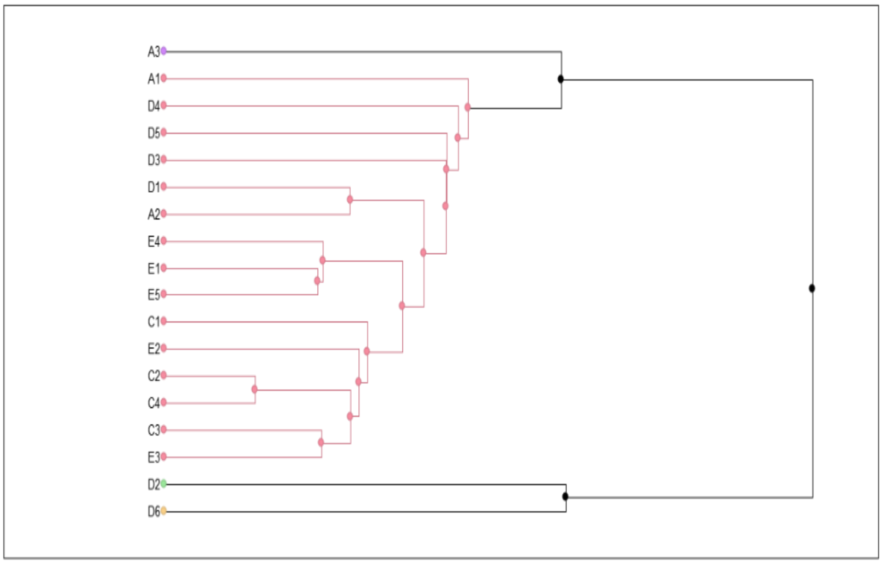
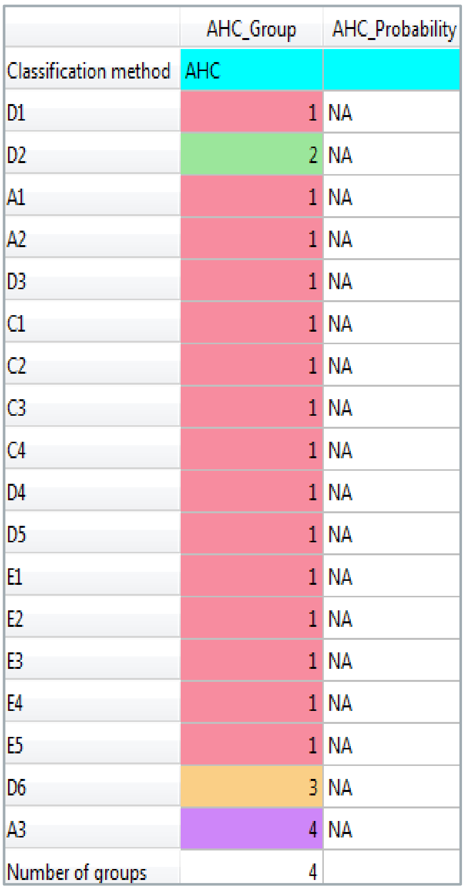


Figure 70: Dedrogram using the Link's Simple aggregation method for Dataset 6



	AHC_Group	AHC_Probability
Classification method	AHC	
D1	1	NA
D2	2	NA
A1	1	NA
A2	1	NA
D3	1	NA
C1	1	NA
C2	1	NA
C3	1	NA
C4	1	NA
D4	1	NA
D5	1	NA
E1	1	NA
E2	1	NA
E3	1	NA
E4	1	NA
E5	1	NA
D6	3	NA
A3	4	NA
Number of groups	4	

Figure 71: Results of Link's Simple aggregation method for Dataset 6

4.4 Sammon Mapping

Lastly the Sammon mapping method took place in this study. The datasets used in this method are the same ones used in the AHC method, which means datasets 4-6 were used. The projection of the data in a 2D space had a great result, since the oil families were well separated. That happened again in the n-alkanes successfully while failing in the hopane model.

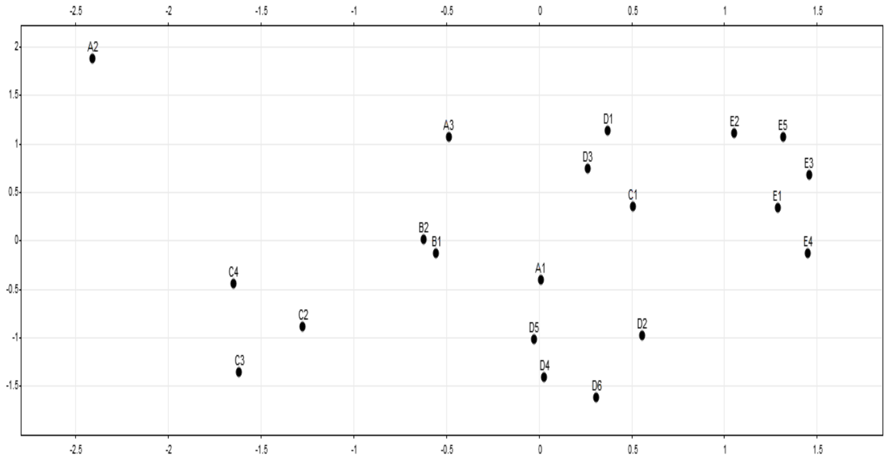


Figure 72: Sammon map of Dataset 4

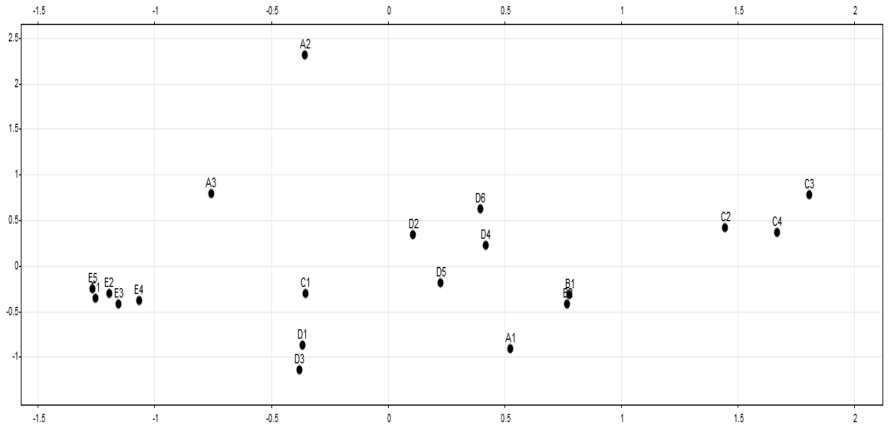


Figure 73: Sammon map of Dataset 5

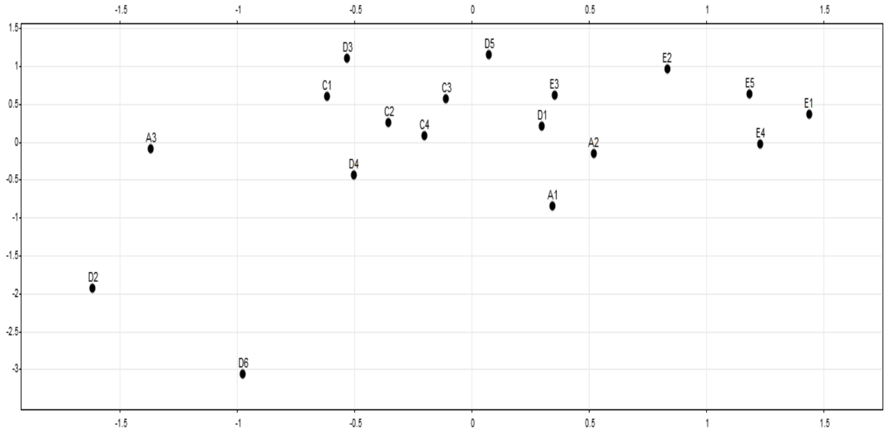


Figure 74: Sammon map of Dataset 6

5 CONCLUSIONS

After taking all of the above into consideration a comparison table can be easily built.

First of all the worst classification as presented in Table 3 is the Link's simple aggregation method. As mentioned before only 20 samples were used and this method favors the datasets composed of a high number of indivuduals. This is the reason that this method produces the worst results.

Link's average and complete aggregation method are similar, with one difference. Complete aggregation method shows a better result when dataset 6 is used. This kind of result is expected, since the complete aggregation method favors a small number of individuals, while the average aggregation method favors the intermediate ones.

In conclusion, the best method to be used in Ascending Hierarchical Classification is the Ward aggregation method with the minimum number of errors. That happens because the Ward method relies on a minimal inertia criterion, nearly independent of the number of individuals in the group.

	PCA	AHC				Sammon
		Ward	Average	Complete	Simple	
Dataset 1	✓	-	-	-	-	-
Dataset 2	✓	-	-	-	-	-
Dataset 3	✓	-	-	-	-	-
Dataset 4	-	✓	✓	✓	✓	✓
Dataset 5	-	✓	✓	✓	✗	✓
Dataset 6	-	✓	✗	✓	✗	✗

Table 3: Comparison table of the methods used

As far as the PCA is concerned, the classification is sufficient, with the n-alkanes datasets (dataset 1-2) to be more accurate than the hopane dataset (dataset 3). Even though dataset 3 is less accurate, the classification is sufficient.

Finally the Sammon Mapping method produced a good separation in the n-alkanes datasets (dataset 4-5) while using dataset 6, it lacked accuracy.

REFERENCES

Brooks, P.W., K.G. Osadetz, and L.R. Snowdon, 1988, Geochemistry of Winnipegosis discoveries near Tablelands, Saskatchewan, in Current Research, Part D: Geological Survey of Canada, Paper 88-1D, p. 11-20.

Brooks, P.W., L.R. Snowden, and K.G. Osadetz, 1987, Families of oils in southeastern Saskatchewan, in C.G. Carlson and J.E. Christopher, eds., Proceedings of the Fifth International Williston Basin Symposium: Saskatchewan Geological Society Special Publication 9, p. 253-264.

Curiale, J.A. 1994. Correlation of oils and source rocks—a conceptual and historical prospective. In: Magoon, L.B., Dow, W.G. (Eds.), The Petroleum System-From Source to Trap. American Association of Petroleum Geologists Memoir 60, pp. 251–260.

Dow, W. G. (1974) Application of oil-correlation and source rock data to exploration in Williston Basin. Bulletin of American Association of Petroleum Geologists 58, 1253-1262.

Gerhard, L. C., S. B. Anderson, and D. W. Fischer, 1990, Petroleum geology of the Williston Basin, in M. W.

Gerhard, L. C., S. B. Anderson, and J. A. LeFever, 1987, Structural history of the Nesson Anticline, North Dakota, in M. W. Longman, ed., Williston Basin: Anatomy of a Cratonic Oil Province: Rocky Mountain Association of Geologists, p. 337-354.

Jarvie, D.M. (2001): Williston Basin Petroleum Systems: Inferences from oil geochemistry and geology; Mtn. Geol., v38, p19-41. Saskatchewan Geological Survey 10 Summary of Investigations 2003, Volume 1

Kohm, J.A., and R.O. Louden, 1982, Ordovician Red River of eastern Montana and western North Dakota: Relationships between lithofacies and production, in J.E. Christopher and J. Kaldi, eds., Fourth International Williston Basin Symposium: Saskatchewan Geological Society, Regina, p. 27-28.

Leenheer, M.J. and Zumberge, J.E. (1987): Correlation and thermal maturity of Williston Basin crude oils and Bakken source rocks using terpane biomarkers; in Longman, M.W. (ed.), Williston Basin: Anatomy of a Cratonic Oil Province, Rocky Mtn. Assoc. Geol., Denver, p287-298.

LeFever,J.A.,Martiniuk,C.D., Dancsok, E.F.R.,Mahnic, P.A.,1991. Petroleum potential of the Middle member, Bakken Formation,Williston Basin. In: Christopher,J.E., Haidl,F.M. (Eds.),The 6th International Williston Basin. Saskatchewan Geological Society Special Publication 11. Regina, Saskatchewan,Canada ,pp. 74–94.

Lillis P. G., 2012. Review of Oil Families and Their Petroleum Systems of the Williston Basin. In: U.S. Geological Survey, Denver

Nouvelle X., Coutrot D. (2010): The Malcom distribution analysis method: A consistent guideline for assessing reservoir compartmentalization from GC fingerprinting. Organic Geochemistry p. 981.

Obermajer M., Osadetz K.G., Fowler M.G., and Snowdon L.R. (2000): Light hydrocarbon (gasoline range) parameter refinement of biomarker-based oil-oil correlation studies: An example from Williston Basin; Org. Geochem., v31, p959-976.

Obermajer M., Osadetz K.G., Pasadakis N., 2003. Refining Compositional Affinity of Williston Basin Family C oils using multivariate statistical analysis of saturate biomarkers. In: Summary of Investigations 2003, Volume 1, Saskatchewan Geological Survey

Obermajer, M., K.G. Osadetz, and L.R. Snowdon, 1998, Familial association and sources of oil quality variation in the Williston Basin from gasoline range and saturated hydrocarbon parameters, in J.E. Christopher, C.F. Gilboy, D.F. Paterson and S.L. Bend, eds., Eighth International Williston Basin Symposium: Saskatchewan Geological Society Special Publication No. 13, p. 209-225.

Osadetz, K. G., Brooks, P. W. and Snowdon, L. R. (1992) Oil families and their sources in Canadian Williston Basin (southeastern Saskatchewan and southwestern Manitoba). Bulletin of Canadian Petroleum Geology 40, 254-273.

Osadetz, K. G., Brooks, P. W. and Snowdon, L. R. (1994) Oil families in Canadian Williston Basin (southwestern Saskatchewan). Bulletin of Canadian Petroleum Geology 42, 155-177.

Osadetz,K.G.,Snowdo n,L.R.,1995. Significant Paleozoic petroleum source rocks in the Canadian Williston Basin: their distributions,richness and thermal maturity (Southeastern Saskatchewan and Southwestern Manitoba). Geological Survey of Canada Bulletin 487.

Price L.C. and LeFever, J. (1994): Dysfunctionalism in the Williston Basin: The Mid-Madison/Bakken Petroleum System; Bull. Can. Petrol. Geol., v42, p187-218.

Price L.C., and J.A. LeFever (1992): Does Bakken horizontal drilling imply a huge oil-resource base in fractured shales? In J.W. Schmoker, E.B. Coalson, C.A. Brown, eds., Geological studies relevant to horizontal drilling: examples from Western North America: Rocky Mountain Association of Geologists, p.199-214.

Schlumberger website : www.slb.com/malcom (accessed in August 2015)

Smith M.G. and R.M. Bustin (2000): Late Devonian and Early Mississippian Bakken and Exshaw black shale source rocks, Western Canada Sedimentary Basin: a sequence stratigraphic interpretation: AAPG Bulletin, v. 84, p. 940-960.

Stasiuk, L.D., and M.G. Fowler, 2004, Organic facies in Devonian and Mississippian strata of Western Canada Sedimentary Basin: relation to kerogen type, paleoenvironment, and paleogeography: Bulletin of Canadian Petroleum Geology, v. 52, p. 234- 255.

Thode,H.G.,1981. Sulfur isotope ratios in petroleum research and exploration: Williston Basin. Bulletin of American Association of Petroleum Geologists 65,1527–15 35.

Williams, J. A. (1974) Characterization of oil types in Williston Basin. Bulletin of the American Association of Petroleum Geology 58, 1243-1252.

Zumberge, J.E. (1983): Tricyclic diterpane distributions in the correlation of Paleozoic crude oils from the Williston Basin; in Bjoroy, M. (ed.), Advances in Organic Geochemistry 1981, John Wiley & Sons Ltd., New York, p738-745.

APPENDICES

Identification – Quantification

Variable	*L001276*	*L001312*	*L00515*	*L00549*	*L00550*	*L00554*	*L00558*	*L00559*	*L00672*	*L00732*
C11	0	0	605	0	1930	0	0	485	522	461
C12	0	0	1485	0	4751	0	0	1074	1330	1228
C13	3235	2963	5599	4942.37	22956.5	4555.98	2345	4744	4361	4470
C14	43333.4	28720	65543.3	41964.5	303560	42080.4	42179.7	14737.8	17094.4	18158
C15	191463	127127	168852	259104	616341	145587	138025	87202.3	66941.2	66703.1
C16	236263	166837	191636	390742	706456	178222	194043	197506	147588	135008
C17	274188	213215	181809	513207	535983	166496	204005	227616	182816	165590
Pr	85047.9	118865	81952.1	10890.1	823679	81137.9	72712.3	82767.9	61407.2	61135.5
C18	237667	181755	165960	225786	408921	153014	213640	243024	194613	168609
Ph	125711	192270	75373.8	18757.3	48628.9	69652.9	93062.7	154307	116697	115305
C19	227992	178947	155631	377229	592696	146107	219651	241666	197801	170179
C20	210553	169723	130794	129153	271590	126307	213638	246184	194430	174180
C21	208210	186403	118574	91167.6	211365	110801	209914	222921	172097	153546
C22	211386	185024	97764.1	75951.3	186439	92299.8	209297	213421	167881	143184
C23	208881	185323	78613.9	60314.5	165332	77126.9	194045	188917	146122	124404
C24	206191	198082	59089.7	47216.2	150551	59750.3	197014	179320	134198	114662
C25	188924	192195	48315.4	42638.8	123047	45797.8	173540	148026	111442	93770.9
C26	166889	163696	34376.4	33140.2	104003	35475.3	157900	138023	101650	83976.9
C27	148862	137014	25899.9	29075.9	85749.3	25473.6	131283	107301	75736.1	68137.3
C28	114792	93886.8	18875.8	22222.8	65734	18368	114758	99334	67321.2	59964.7
C29	85940.2	75889.3	13287.8	16640.1	52126.2	14179.8	82137.2	72928.5	53704	47362.1
C30	69187.1	52588.1	9079.62	13221	39007.7	8976.31	68929.3	62834.7	44379.9	41125.5
C31	46686.3	34575.4	6331.89	9167.28	27029.5	6295.51	39487	38608.9	28992.4	26660.8
C32	29352.8	19476.3	3848.84	5619.29	17629.7	2972.65	22103.5	28617.2	18870	18472.9
C33	17823.8	12218	2970.72	3169.73	9145.45	2026.27	11555.6	13379.9	11359.9	10418.1
C34	9722.67	6857.49	1621.23	2004.82	5714.57	896.77	5754.91	7483.18	6197.48	6230.64
C35	5104.59	3850.75	903.543	1005.71	3625.13	734.18	2853.29	3755.82	3211.76	3234.18

Table 4: IQ summary table of areas of m/z 85 (part 1)

Variable	L00753	L00755	L00756	L00811	L00820	L00829	L00833	L00839	L00842	L00920
C11	0	0	0	0	0	0	0	0	0	1322
C12	1119	0	0	0	0	0	0	0	0	2534
C13	4288	5367	8619	1994	2366	2583	3750	2056	3661	8544
C14	16196	78645.5	109397	15073.9	16631.3	12838	15065.7	14547.5	16710.7	83187.6
C15	49230	198478	235162	22575.2	22890.2	14973.8	17409.4	18443.2	74537.4	174940
C16	133433	221333	239847	13680.5	13097.2	10559.1	12413.6	9534.15	129979	175771
C17	178830	232693	260480	3704.79	7048.16	8558.13	10650	5369.17	183902	298464

Pr	57030.5	111733	93264.9	1759.88	4186.44	16821.7	22146.6	1568.57	131587	628094
C18	193939	212961	218483	3551.59	3441.46	6680.34	7878.2	2634.68	176307	320513
Ph	107293	201758	140127	5403.52	7963.93	21654.7	27146.1	2830.14	221461	64359.6
C19	189219	191341	214636	2533.46	2876.69	5480.04	7067.82	2065.06	166367	530771
C20	188716	185002	187915	2771.25	2669.05	5472	6236.63	1759.5	167237	239357
C21	173447	186495	197923	2058.02	1761.28	5145.74	6015.51	1452.49	174516	215595
C22	163711	184687	184939	2832.11	1752.98	4946.15	5229.56	2118.67	170041	201554
C23	144862	170075	181965	3068.3	2276.5	4072.56	4670.63	1609.6	172365	185031
C24	125628	181663	176389	5039.76	3990.23	5397.22	5659.39	2165.11	180141	171010
C25	104031	171724	165024	5797.65	4253.9	4621.65	4944.34	1981.48	174600	149150
C26	92477.2	149918	145995	6137.07	4554.03	4554.94	4017.52	1642.09	145449	128474
C27	72262.2	117284	120084	4552.41	3982.39	3921.23	3234.68	1796.5	125008	107953
C28	65277.5	84221.9	87043	4573.14	3677.8	3695.61	2722.57	1766.77	77918.9	80992.8
C29	50113.4	61036.5	62575	3289.77	2737.85	3135.1	3011.15	1711.19	62916	60934.9
C30	41882.2	46339.9	45811.2	3053.11	2467.93	2292.28	2207.91	1726.63	41289	41806.8
C31	28615.6	26499.6	28353.7	2321.1	1822.11	2150.98	1415.16	0	28542	30397.3
C32	19853.5	13984.2	14567.3	1334.07	0	1723.96	1719.14	0	12674.7	17592.3
C33	11300.2	7152.71	7128.97	1218.97	0	0	1181.53	0	8050.39	9307.47
C34	6800.51	3441.24	4250.49	1015.94	1321.77	0	1371.38	856.811	3583.99	5560.81
C35	4488.56	1824.12	2119.86	903.626	0	0	1279.96	0	2373.29	3103.27

Table 5: IQ summary table of areas of m/z 85 (part 2)

Variable	*L001276*	*L001312*	*L00515*	*L00549*	*L00550*	*L00554*	*L00558*	*L00559*	*L00672*	*L00732*
C19-tri	155.794	147.612	0	69.8631	150.511	0	110.042	46.8137	137.374	78.5979
C20-tri	119.163	256.202	0	72.2067	145.634	0	128.158	129.168	173.545	292.637
C21-tri	129.876	195.026	0	195.703	111.451	0	119.437	156.147	214.194	236.429
C22-tri	83.8318	143.651	0	32.3876	65.9083	0	90.1404	165.508	118.977	119.002
C23-tri	171.633	381.981	15707.1	50.6325	98.6856	14642.3	251.071	668.948	573.44	548.186
C24-tri	121.428	237.588	0	66.2348	66.55	0	174.215	212.608	309.452	324.354
C25-triS,R	141.185	271.376	0		61.8986	0	187.544	293.552	310.679	305.673
C24-tetra	391.816	864.215	0	101.276	266.949	0	511.128	240.998	255.687	274.621
C26-triS,R	38.7337	142.149	0	18.5411	47.2241	0	0	0	0	0
Ts	512.03	965.077	0	110.556	412.221	0	625.203	211.612	219.653	218.677
Tm	490.479	1885.9	0	182.195	492.413	0	632.695	644.551	336.741	321.884
C29-moretane	244.817	768.58	0	111.238	167.091	0	215.259	113.201	125.819	121.013
C30-hopane	2067.01	8776.97	0	547.845	1363.61	0	2195.72	1152.06	854	956.9
C30-moretane	248.832	951.32	0	89.0156	201.831	0	224.638	126.117	133.345	129.933
C31-S	647.102	3151.84	0	209.69	544.065	0	941.461	618.423	385.653	419.455
C31-R	617.49	2535.83	0	172.59	414.646	0	727.492	537.822	319.126	324.777
Gammacerane	295.75	112.445	0	51.477	30.2502	0	541.177	394.846	276.903	133.363
C32-S	413.54	2037.67	0	151.59	379.217	0	840.547	521.895	315.505	307.132
C32-R	350.724	1639.04	0	133.57	328.178	0	721.989	374.079	224.543	251.588
C33-S	269.093	1146.64	0	98.8561	239.053	0	505.419	290.198	251.394	260.815
C33-R	229.677	861.165	0	93.6509	209.003	0	431.129	225.005	194.512	133.868

C34-S	450.673	2214.78	0	79.7592	276.725	0	610.096	253.192	166.056	209.029
C34-R	339.991	1637.08	0	72.5696	153.934	0	417.608	166.897	114.886	139.709
C35-S	220.827	680.326	0	31.3142	117.125	0	546.435	362.732	207.963	202.633
C35-R	175.132	560.964	0	0	57.5202	0	300.646	201.25	100.194	106.915

Table 6: IQ summery table of m/z 191 areas (part 1)

Variable	*L00753*	*L00755*	*L00756*	*L00811*	*L00820*	*L00829*	*L00833*	*L00839*	*L00842*	*L00920*
C19-tri	100.463	172.552	158.926	96.3988	84.3052	95.1054	102.509	64.5051	113.17	234.345
C20-tri	188.459	169.521	147.785	193.117	328.351	281.864	205.442	203.987	263.168	359.674
C21-tri	216.291	141.421	113.056	360.029	302.411	269.559	317.384	331.416	187.65	157.312
C22-tri	138.01	76.2091	85.4142	208.469	158.352	130.343	245.942	160.69	159.311	70.5241
C23-tri	540.953	249.361	188.691	800.843	778.582	638.31	736.463	814.138	458.953	217.432
C24-tri	319.875	170.288	133.528	518.015	496.223	401.176	450.514	524.301	240.203	140.198
C25-triS,R	300.433	184.396	122.987	454.419	435.138	373.487	460.434	481.443	305.893	207.301
C24-tetra	249.609	744.311	453.484	331.921	304.24	275.742	308.02	355.58	780.298	737.98
C26-triS,R	0	65.8766	133.301	0	0	0	0	0	0	94.6577
Ts	175.207	965.853	615.75	349.107	305.278	237.401	302.966	328.45	1025.87	810.538
Tm	328.02	790.802	466.112	435.901	400.521	324.464	375.658	442.111	2190.92	1407.16
C29-moretane	127.335	420.033	275.899	233.641	222.175	206.002	190.42	184.795	884.609	407.845
C30-hopane	942.794	3812.37	1974.68	1353.8	1287.38	981.41	1048.72	1299.69	9814.95	4381.77
C30-moretane	113.187	380.647	188.741	208.418	184.779	177.673	209.101	162.903	1216.97	636.068
C31-S	416.711	1268.19	651.343	650.87	583.747	439.476	509.134	584.605	3835.52	1789.34
C31-R	326.406	1128.64	520.964	493.499	450.847	321.33	401.569	457.685	2925.72	1405.32
Gammacerane	115.072	144.413	110.738	167.701	167.79	130.177	139.3	137.801	257.564	244.842
C32-S	299.45	947.868	475.989	409.117	405.388	333.231	350.353	360.065	2564.05	1223.38
C32-R	223.427	752.427	446.237	343.975	369.936	315.315	269.647	338.066	1984.61	1042.2
C33-S	249.536	479.624	288.776	389.087	363.638	277.75	276.872	351.048	1557.8	766.062
C33-R	205.516	462.076	276.112	288.131	297.892	252.584	233.366	291.264	1157.32	598.498
C34-S	149.319	815.179	433.301	257.094	254.01	149.43	213.773	258.982	2587.01	760.07
C34-R	115.423	753.899	270.244	201.871	156.556	170.574	190.586	203.317	1825.8	641.259
C35-S	185.271	364.732	170.482	273.795	199.751	165.799	214.151	313.943	977.006	401.11
C35-R	99.2128	341.525	167.312	170.191	179.104	138.904	97.149	214.839	761.541	168.696

Table 7: IQ summery table of m/z 191 areas (part 2)

Principal Component Analysis (PCA)

Variable	**A1**	**A2**	**A3**	**B1**	**B2**	**C1**	**C2**	**C3**	**C4**	**D1**	**D2**
PC1	1.4723	-5.99558	-4.05848	2.16299	2.33846	-3.20154	-0.86587	-0.00259	-0.59656	-3.68371	-2.54495
PC2	2.69186	8.1102	0.977267	1.13867	0.810176	-1.83901	-1.0343	-0.83767	-1.00064	-1.82368	-2.12557
PC3	1.93046	0.07309	-2.51394	0.11326	0.130251	-0.04337	-0.15796	-0.06739	-0.23754	0.355285	0.492905
PC4	-2.3018	1.10824	-1.30399	0.081657	-0.12666	-0.44452	-0.27658	-0.15202	-0.35506	-0.22383	0.66211
PC5	-0.10023	0.27834	-0.64846	-0.58252	-0.59599	0.048008	-0.02784	-0.14595	-0.07489	0.954835	-0.42331
PC6	-0.45387	0.061539	-0.65547	0.669037	0.547041	0.694169	0.646258	0.512324	0.717545	-0.31483	-0.62267

PC7	-0.11914	0.108454	-0.06861	-0.02281	0.153195	0.16699	-0.05205	0.069358	0.057203	-0.00536	-0.21386
PC8	0.068705	0.17914	-0.13826	-0.41193	-0.39282	0.231121	0.201719	0.243491	0.206502	-0.0899	0.155255
PC9	-0.03814	0.023492	-0.03664	0.045924	0.239052	-0.07334	0.002766	-0.11729	0.027053	0.141277	0.233771
PC10	-0.00017	0.007838	-0.00916	0.014091	0.060556	0.121911	-0.12664	0.040023	-0.15212	0.021029	-0.11643
PC11	0.002011	0.001548	-0.00237	-0.01132	0.015762	-0.01957	0.162362	-0.03951	-0.11289	-0.00084	-0.04953
PC12	-0.01463	-0.00225	0.005094	0.005446	0.000702	-0.01027	0.015438	0.052442	-0.03331	0.091207	-0.00416
PC13	0.001496	0.000148	-0.00171	0.043476	-0.03617	-0.07046	0.013124	0.054	0.008027	0.003652	-0.00829
PC14	0.002821	-0.00017	0.000696	-0.04092	0.036598	-0.01915	0.012555	0.042377	-0.02154	-0.0424	0.025319
PC15	0.000662	-0.00124	0.000817	0.034101	-0.02426	0.029784	0.000835	0.007349	-0.03533	-0.01548	0.034573
PC16	0.001586	-5.33E-05	-0.00078	0.012749	-0.00868	-0.00255	0.002985	-0.01218	0.005357	-0.00486	-0.00306
PC17	0.00024	-0.00029	2.86E-05	0.004001	-0.00077	-0.00435	-0.0029	0.005188	-0.00239	-0.00013	-0.00134
PC18	0.000324	-6.61E-05	-9.61E-06	0.003345	-0.00278	-0.00134	-0.00312	-0.0018	0.004571	-0.00179	0.002386
PC19	-4.31E-05	-6.28E-05	-1.16E-06	-0.00093	0.001257	0.00039	-0.00464	0.004468	0.00043	0.000348	6.86E-05

Table 8: Principal Components of Dataset 1 (part 1)

Variable	D3	D4	D5	D6	E1	E2	E3	E4	E5	Quality
PC1	-3.46697	-2.394	-2.94582	-2.1073	5.24722	5.19785	5.10156	5.07924	5.26374	0.652717
PC2	-2.12131	-1.57847	-0.57714	-2.1965	0.208889	0.373187	0.324639	0.117766	0.381627	0.256313
PC3	-0.23756	0.795071	0.418974	0.445131	-0.29483	-0.31201	-0.31354	-0.23364	-0.34264	0.030139
PC4	-0.23221	0.603861	0.638647	0.774002	0.244414	0.345915	0.366887	0.342815	0.248117	0.0277
PC5	1.40073	-0.4477	0.276021	-1.1616	0.371732	0.289255	0.125789	0.13161	0.332154	0.015692
PC6	-0.04875	-0.48567	0.177967	-0.38178	-0.30469	-0.1848	-0.11681	-0.3471	-0.10945	0.010597
PC7	0.150976	0.672526	-0.70719	-0.16997	0.175799	-0.14254	-0.17659	0.222601	-0.099	0.003212
PC8	-0.13435	-0.27205	-0.39696	0.246879	-0.02807	0.053319	0.148038	0.068462	0.061697	0.002382
PC9	0.063152	-0.20659	-0.20372	0.00896	-0.00883	-0.03791	-0.0367	-0.00475	-0.02155	0.000622
PC10	0.072827	-0.06363	-0.05065	0.163556	0.024976	0.054864	0.00368	-0.10882	0.04226	0.000324
PC11	0.021245	-0.0103	-0.00056	0.041524	0.038942	-0.0593	0.061708	0.030092	-0.06901	0.000152
PC12	-0.07075	0.016262	-0.01619	-0.0315	-0.02037	0.034631	0.023037	-0.02879	-0.01204	5.61E-05
PC13	0.025414	-0.01147	-0.01543	0.02697	0.032073	-0.0069	-0.04352	-0.01592	0.001491	4.09E-05
PC14	0.024773	0.011299	0.00463	-0.02209	-0.00902	0.024766	-0.00094	-0.02676	-0.00285	2.73E-05
PC15	0.001757	-0.00533	-0.00566	-0.02431	0.00592	-0.00268	-0.00685	0.005404	-6.11E-05	1.54E-05
PC16	0.011652	0.001869	-0.00767	0.004725	-0.01673	0.037556	0.002818	0.0028	-0.02755	7.88E-06
PC17	0.006013	0.000488	-0.00241	0.00089	-0.01614	-0.00673	0.008751	0.006654	0.005206	1.44E-06
PC18	0.002346	0.002839	-0.00257	-0.00178	0.005065	-0.0043	0.014125	-0.01227	-0.00318	1.23E-06
PC19	-0.00011	-0.00245	0.001901	-0.00018	0.002499	-0.00174	0.001631	0.004537	-0.00736	3.52E-07

Table 9: Principal Components of Dataset 1 (part 2)

Variable	A1	A2	A3	B1	B2	C1	C2	C3	C4	D1	D2
PC1	1.6333	-4.59546	-3.90682	2.3221	2.39399	-3.42024	-0.98898	-0.11966	-0.72878	-3.90542	-2.78176
PC2	2.97451	6.98946	1.35521	0.735069	0.617736	-1.33259	-0.85938	-0.72872	-0.81825	-1.25919	-1.77073
PC3	-1.76534	-0.10258	2.5936	-0.15096	-0.13952	0.033172	0.124038	0.039959	0.21448	-0.30667	-0.53105
PC4	-1.87364	1.16386	-0.76861	0.004566	-0.05363	-0.45774	-0.32525	-0.1726	-0.39209	-0.33371	0.708917
PC5	0.348749	-0.47637	0.808349	0.592524	0.604834	-0.01095	0.043197	0.148421	0.110364	-0.9057	0.304226

PC6	-0.60576	0.210864	-0.63444	0.625106	0.585604	0.653551	0.545152	0.465099	0.64937	-0.3124	-0.61621
PC7	0.103356	0.113175	-0.07552	-0.37672	-0.41676	0.234057	0.16307	0.259587	0.212852	-0.14187	0.025813
PC8	0.006112	-0.01272	0.022054	-0.00074	-0.04648	0.12146	-0.10955	0.080772	-0.16918	-0.03898	-0.26345
PC9	7.41E-05	0.007556	-0.01127	0.011523	0.067584	0.060551	-0.17695	0.025617	0.033516	0.03065	0.051776
PC10	-0.01424	-0.00824	0.013352	-0.01101	0.001778	-0.02317	-0.04133	0.033783	0.052215	0.051836	0.025801
PC11	-0.00339	0.001444	-0.00231	-0.02117	0.073056	-0.00011	0.056132	0.002968	-0.0821	0.041658	0.015906
PC12	-0.00456	-0.00024	-0.0003	0.048442	-0.03984	-0.06135	0.016117	0.063878	-0.00507	0.045563	-0.01332
PC13	-0.00662	0.000293	0.000114	0.025278	-0.01855	0.043385	-0.0072	-0.03409	-0.00671	0.062375	-0.01529
PC14	0.000954	-0.00139	0.001107	0.032154	-0.02516	0.025788	0.000168	0.008818	-0.03132	-0.0203	0.03402
PC15	0.002018	-0.0002	-0.00077	0.017483	-0.01248	-0.00324	0.001984	-0.01633	0.008265	-0.00675	-0.00366
PC16	0.000334	-8.68E-05	-7.43E-05	0.002789	-0.00096	-0.00103	0.000447	-0.00648	0.002434	0.000834	-0.00465
PC17	-0.00015	0.000271	-1.73E-05	-0.00306	-6.57E-05	0.003967	0.003309	-0.00638	0.002934	-0.00039	0.001413
PC18	-0.00031	0.000101	3.84E-05	-0.00268	0.00144	0.001148	0.004199	0.001055	-0.00401	0.000713	-0.0002

Table 10: Principal Components of Dataset 2 (part 1)

Variable	D3	D4	D5	D6	E1	E2	E3	E4	E5	Quality
PC1	-3.72518	-2.60372	-2.82282	-2.35121	5.15248	5.17186	5.07304	4.97446	5.22882	0.699042
PC2	-1.5719	-1.17021	-0.82482	-1.89384	-0.07698	-0.06991	-0.11204	-0.14803	-0.03538	0.214023
PC3	0.292678	-0.73707	-0.50249	-0.5209	0.317224	0.288425	0.275813	0.251465	0.325721	0.033106
PC4	-0.45702	0.620799	0.147334	0.916678	0.214924	0.244933	0.30076	0.355921	0.155577	0.023265
PC5	-1.31659	0.435468	-0.28257	1.00064	-0.38111	-0.32532	-0.18231	-0.16022	-0.35564	0.016968
PC6	-0.00909	-0.15869	-0.10924	-0.37918	-0.19329	-0.21479	-0.16237	-0.20825	-0.13103	0.010385
PC7	-0.12372	-0.00583	-0.38194	0.139854	0.003693	0.023142	0.089936	0.11857	0.03525	0.002123
PC8	0.04189	0.203127	0.028465	0.068456	0.05165	0.044589	0.004527	-0.0512	0.019185	0.000533
PC9	0.044955	-0.06053	-0.11217	0.045614	-0.01146	0.054599	-0.06484	-0.05774	0.060945	0.000222
PC10	-0.0644	0.125613	-0.02059	-0.12128	-0.03071	0.031858	-0.01753	0.01456	0.001699	0.000132
PC11	0.021626	0.022606	-0.10238	-0.00804	0.018702	-0.0071	0.012316	0.009347	-0.04916	8.97E-05
PC12	-0.01024	-0.01674	-0.016	0.021595	0.021223	0.005013	-0.02715	-0.02569	-0.00132	4.81E-05
PC13	-0.0492	-0.01917	-0.00505	0.009024	-0.0076	-0.00596	0.024406	0.014408	-0.00384	3.36E-05
PC14	0.003337	-0.0022	-0.00141	-0.02571	0.005764	-0.00505	-0.00746	0.003247	0.00464	1.61E-05
PC15	0.014783	0.004981	-0.00996	0.004937	-0.01385	0.032379	-0.00127	0.007408	-0.02572	8.89E-06
PC16	0.003261	0.00391	-0.00426	0.003125	0.003582	-0.00905	-0.01862	0.018723	0.005767	2.67E-06
PC17	-0.00532	0.000594	0.00168	-0.00103	0.016007	0.006114	-0.00843	-0.00757	-0.00389	1.55E-06
PC18	-0.00271	-0.00203	0.002407	0.000299	-0.00541	0.00641	-0.00504	0.000504	0.004069	4.89E-07

Table 11: Principal Components of Dataset 2 (part 2)

Variable	A1	A2	A3	C1	C2	C3	C4	D1	D2
PC1	-3.63489	-1.67111	3.90641	-1.85579	-2.40515	-2.46997	-2.60323	-0.79315	8.3351
PC2	-2.64059	-2.93923	-1.61596	0.338567	-0.08622	0.795522	0.374943	-2.51678	0.620051
PC3	0.060569	1.02538	1.71784	-2.74497	-0.44764	0.616833	0.222833	-0.61873	0.681038
PC4	2.30807	0.842453	-2.2816	0.06636	-0.70212	-0.06277	0.407275	-0.38198	0.904719
PC5	-0.64999	-0.03806	-0.70055	-0.37545	0.071638	-1.16263	-0.08947	0.496001	0.282514
PC6	0.392029	-0.54227	-0.34398	-0.663	-0.28974	0.133895	-0.32386	-0.16159	-0.04096

PC7	0.604717	-0.13117	0.409792	-0.58768	0.495229	-0.47071	-0.13819	0.312162	-0.17255
PC8	0.237941	-0.0569	0.136735	-0.07448	-0.6054	0.215983	-0.23705	-0.01142	-0.0426
PC9	-0.07015	-0.19818	-0.35256	-0.25242	0.156875	0.128385	-0.07634	0.436988	0.169262
PC10	-0.18079	0.316764	-0.1514	-0.29947	-0.05895	-0.11366	0.112557	-0.15747	-0.4117
PC11	-0.08877	0.060485	-0.04492	-0.20237	0.08226	0.236536	0.226957	-0.07348	0.164833
PC12	0.048691	-0.06607	0.0247	0.006068	-0.03809	-0.18889	0.111915	-0.11275	0.137155
PC13	0.052685	-0.07964	-0.00305	-0.0058	0.074618	-0.01179	0.064178	-0.10837	-0.05791
PC14	-0.0172	-0.01426	0.01811	-0.00784	-0.08027	-0.05255	0.077057	0.105301	-0.02615
PC15	-0.00301	0.08161	-0.01739	-0.00184	0.026367	-0.04184	-0.05005	-0.01873	0.029325
PC16	-0.01288	0.012674	-0.00217	0.004023	0.007428	0.007747	-0.06512	-0.00793	0.014719
PC17	0.002129	0.002868	0.002283	0.002505	0.001928	0.002335	0.002741	0.00257	0.002561

Table 12: Principal Components of Dataset 3 (part 1)

Variable	D3	D4	D5	D6	E1	E2	E3	E4	E5	Quality
PC1	0.712339	2.26491	-0.68203	10.7235	-1.73328	-1.79131	-2.23051	-2.15681	-1.91502	0.666869
PC2	-1.75436	-2.01061	-2.46856	1.50049	3.00277	2.62233	1.39271	2.68717	2.69775	0.185947
PC3	-2.87122	0.758988	0.686805	-0.96992	0.054464	0.775948	0.848392	0.204633	-0.00124	0.062165
PC4	-1.30697	-0.36194	0.086582	1.02545	-0.01279	-0.64331	-0.193	0.001664	0.303911	0.041203
PC5	-0.07052	1.13619	0.673141	-0.51907	0.996119	-0.76229	-0.69242	1.12059	0.284256	0.020591
PC6	0.643331	0.797998	-0.08274	-0.15419	0.100621	0.299272	0.192174	-0.78134	0.824325	0.009465
PC7	0.110598	-0.52483	-0.44552	0.189633	0.526258	-0.19003	-0.05655	0.033419	0.035419	0.005935
PC8	0.271107	-0.04997	-0.03737	-0.05067	0.083636	0.029635	-0.06907	0.482587	-0.22269	0.00235
PC9	0.024894	-0.07624	0.076819	0.007212	-0.11992	0.248039	0.199824	0.053836	-0.35631	0.001911
PC10	0.210679	-0.1014	0.259165	0.352914	0.011546	0.122755	0.081555	0.033457	-0.02655	0.001866
PC11	0.189417	-0.10214	-0.0898	-0.08865	-0.06403	-0.20698	-0.12117	0.057102	0.064722	0.00084
PC12	0.092458	-0.10149	0.113284	-0.10384	-0.02697	0.074351	0.070108	-0.00676	-0.03388	0.000362
PC13	0.003343	0.140186	-0.0661	0.03801	-0.04813	-0.08245	0.120586	0.079033	-0.10939	0.000256
PC14	-0.01655	-0.0053	-0.04986	0.01394	-0.08924	-0.02775	0.058528	0.031362	0.082682	0.000128
PC15	0.024744	0.019235	-0.07758	-0.02266	-0.04711	0.051108	0.002393	0.034991	0.010418	6.77E-05
PC16	0.007375	-0.02967	0.029564	-0.00665	-0.00327	-0.06081	0.075209	0.005971	0.027027	4.28E-05
PC17	0.002521	0.002528	0.001835	0.002233	0.003221	0.002267	0.002896	0.00185	0.001897	2.68E-07

Table 13: Principal Components of Dataset 3 (part 2)

Sammon Mapping

	Analyses	Before reduction	After reduction
0	D1 / D2	172.245	21.282
1	D1 / B1	1.472	157.342
2	D1 / A1	176.304	15.807
3	D1 / A2	388.964	287.502
4	D1 / B2	147.072	150.428
5	D1 / D3	0.78054	0.410142
6	D1 / C1	108.725	0.789999

7	D1 / C2	251.648	260.611
8	D1 / C3	291.945	318.524
9	D1 / C4	258.531	256.256
10	D1 / D4	223.947	256.183
11	D1 / D5	199.313	218.951
12	D1 / E1	120.208	121.383
13	D1 / E2	140.937	0.682404
14	D1 / E3	131.672	117.454
15	D1 / E4	149.677	166.657
16	D1 / E5	123.736	0.946394
17	D1 / D6	240.197	275.601
18	D1 / A3	196.928	0.860671
19	D2 / B1	126.004	139.596
20	D2 / A1	1.369	0.798179
21	D2 / A2	357.008	411.895
22	D2 / B2	13.219	154.038
23	D2 / D3	168.773	174.986
24	D2 / C1	182.719	134.269
25	D2 / C2	184.959	183.143
26	D2 / C3	212.309	220.219
27	D2 / C4	213.505	226.882
28	D2 / D4	0.796097	0.673361
29	D2 / D5	0.953093	0.584074
30	D2 / E1	188.274	151.227
31	D2 / E2	207.089	215.653
32	D2 / E3	203.957	18.948
33	D2 / E4	178.288	123.304
34	D2 / E5	214.997	219.071
35	D2 / D6	0.789079	0.68096
36	D2 / A3	195.899	229.912
37	B1 / A1	118.065	0.622294
38	B1 / A2	348.713	273.714
39	B1 / B2	0.197845	0.160017
40	B1 / D3	135.491	119.789
41	B1 / C1	136.191	116.936
42	B1 / C2	14.127	103.855
43	B1 / C3	17.611	161.497
44	B1 / C4	142.475	113.569
45	B1 / D4	151.633	139.325
46	B1 / D5	139.382	102.538
47	B1 / E1	202.502	190.143
48	B1 / E2	207.623	203.651
49	B1 / E3	197.597	216.987
50	B1 / E4	211.354	200.317

51	B1 / E5	207.813	222.515
52	B1 / D6	167.636	171.579
53	B1 / A3	186.472	120.387
54	A1 / A2	311.893	332.192
55	A1 / B2	119.376	0.753444
56	A1 / D3	176.139	117.098
57	A1 / C1	165.294	0.908258
58	A1 / C2	175.821	137.014
59	A1 / C3	204.179	188.231
60	A1 / C4	181.174	165.697
61	A1 / D4	147.134	100.031
62	A1 / D5	137.381	0.615355
63	A1 / E1	182.071	147.793
64	A1 / E2	195.102	184.151
65	A1 / E3	185.501	18.086
66	A1 / E4	189.944	146.518
67	A1 / E5	197.866	196.901
68	A1 / D6	15.947	125.341
69	A1 / A3	18.129	154.916
70	A2 / B2	349.041	258.442
71	A2 / D3	384.193	290.024
72	A2 / C1	356.841	328.535
73	A2 / C2	325.958	298.715
74	A2 / C3	332.924	332.729
75	A2 / C4	324.627	244.132
76	A2 / D4	331.846	408.539
77	A2 / D5	318.224	374.692
78	A2 / E1	338.814	400.114
79	A2 / E2	360.053	3.543
80	A2 / E3	347.842	404.203
81	A2 / E4	34.241	434.958
82	A2 / E5	364.288	380.833
83	A2 / D6	348.117	442.837
84	A2 / A3	260.528	208.485
85	B2 / D3	137.654	114.929
86	B2 / C1	133.865	118.226
87	B2 / C2	143.457	110.459
88	B2 / C3	177.215	168.436
89	B2 / C4	141.345	111.869
90	B2 / D4	158.196	155.324
91	B2 / D5	145.763	118.484
92	B2 / E1	202.443	193.897
93	B2 / E2	207.083	20.089
94	B2 / E3	195.792	218.556

95	B2 / E4	215.067	207.753
96	B2 / E5	206.906	221.101
97	B2 / D6	173.315	187.535
98	B2 / A3	18.512	106.665
99	D3 / C1	101.006	0.454501
100	D3 / C2	234.722	223.636
101	D3 / C3	286.511	281.253
102	D3 / C4	24.706	224.603
103	D3 / D4	204.909	215.621
104	D3 / D5	175.348	17.813
105	D3 / E1	132.557	110.126
106	D3 / E2	13.267	0.875307
107	D3 / E3	13.611	119.489
108	D3 / E4	142.267	147.568
109	D3 / E5	132.702	110.417
110	D3 / D6	225.564	236.061
111	D3 / A3	18.614	0.815411
112	C1 / C2	20.373	217.057
113	C1 / C3	255.142	27.256
114	C1 / C4	203.211	229.729
115	C1 / D4	208.641	182.402
116	C1 / D5	189.427	147.438
117	C1 / E1	12.143	0.781339
118	C1 / E2	119.318	0.933764
119	C1 / E3	129.084	100.354
120	C1 / E4	153.922	106.422
121	C1 / E5	113.673	107.864
122	C1 / D6	225.911	198.659
123	C1 / A3	167.362	121.891
124	C2 / C3	0.827937	0.579765
125	C2 / C4	0.751232	0.580364
126	C2 / D4	162.638	140.022
127	C2 / D5	165.584	125.251
128	C2 / E1	26.933	283.833
129	C2 / E2	266.905	306.792
130	C2 / E3	27.052	314.741
131	C2 / E4	255.175	282.429
132	C2 / E5	281.264	324.467
133	C2 / D6	180.291	174.487
134	C2 / A3	206.104	210.361
135	C3 / C4	0.980711	0.912068
136	C3 / D4	186.303	164.419
137	C3 / D5	191.958	162.314
138	C3 / E1	312.447	336.023

139	C3 / E2	326.586	363.488
140	C3 / E3	319.919	368.488
141	C3 / E4	303.025	329.831
142	C3 / E5	335.439	380.351
143	C3 / D6	199.735	194.357
144	C3 / A3	251.384	266.944
145	C4 / D4	200.381	19.314
146	C4 / D5	197.954	171.887
147	C4 / E1	280.654	303.699
148	C4 / E2	280.881	311.723
149	C4 / E3	269.628	330.061
150	C4 / E4	284.464	311.222
151	C4 / E5	286.392	332.683
152	C4 / D6	213.599	228.416
153	C4 / A3	219.747	19.028
154	D4 / D5	0.64845	0.389855
155	D4 / E1	216.001	214.943
156	D4 / E2	234.517	271.763
157	D4 / E3	235.052	252.757
158	D4 / E4	194.353	190.467
159	D4 / E5	249.995	278.822
160	D4 / D6	0.473808	0.355483
161	D4 / A3	19.786	252.102
162	D5 / E1	192.399	188.886
163	D5 / E2	22.013	238.943
164	D5 / E3	216.509	2.256
165	D5 / E4	173.583	172.033
166	D5 / E5	233.429	24.821
167	D5 / D6	0.975382	0.69057
168	D5 / A3	182.705	213.205
169	E1 / E2	0.900831	0.809234
170	E1 / E3	0.938033	0.38292
171	E1 / E4	0.812913	0.501218
172	E1 / E5	0.81484	0.731706
173	E1 / D6	23.454	21.889
174	E1 / A3	158.685	191.637
175	E2 / E3	0.905699	0.588983
176	E2 / E4	110.229	131.024
177	E2 / E5	0.471541	0.265671
178	E2 / D6	249.817	28.327
179	E2 / A3	160.156	154.058
180	E3 / E4	135.109	0.818312
181	E3 / E5	0.751963	0.411521
182	E3 / D6	249.005	257.181

183	E3 / A3	161.546	197.879
184	E4 / E5	126.897	121.279
185	E4 / D6	216.431	187.138
186	E4 / A3	159.921	227.769
187	E5 / D6	263.315	287.162
188	E5 / A3	171.114	180.189
189	D6 / A3	218.967	280.019

Table 14: Results of Sammon mapping using Dataset 4

	Analyses	Before reduction	After reduction
0	D1 / D2	117.514	129.885
1	D1 / B1	122.106	126.969
2	D1 / A1	142.788	0.892707
3	D1 / A2	303.721	31.822
4	D1 / B2	121.975	122.144
5	D1 / D3	0.404965	0.276466
6	D1 / C1	0.914786	0.564437
7	D1 / C2	21.911	222.446
8	D1 / C3	264.481	272.778
9	D1 / C4	236.073	237.667
10	D1 / D4	132.652	134.019
11	D1 / D5	0.992316	0.90224
12	D1 / E1	107.963	102.605
13	D1 / E2	107.277	100.502
14	D1 / E3	100.469	0.90724
15	D1 / E4	100.347	0.84949
16	D1 / E5	106.307	109.038
17	D1 / D6	15.844	167.993
18	D1 / A3	163.532	170.878
19	D2 / B1	0.956484	0.941563
20	D2 / A1	121.758	131.298
21	D2 / A2	294.835	202.723
22	D2 / B2	0.977881	100.345
23	D2 / D3	143.531	156.341
24	D2 / C1	132.498	0.793424
25	D2 / C2	178.536	133.893
26	D2 / C3	209.655	175.284
27	D2 / C4	194.576	155.793
28	D2 / D4	0.548038	0.332574
29	D2 / D5	0.739619	0.5417
30	D2 / E1	167.473	152.595
31	D2 / E2	16.995	145.167
32	D2 / E3	166.093	147.168
33	D2 / E4	165.707	13.833

34	D2 / E5	16.726	149.582
35	D2 / D6	0.55005	0.406879
36	D2 / A3	178.369	0.978622
37	B1 / A1	113.646	0.63656
38	B1 / A2	279.163	286.847
39	B1 / B2	0.120239	0.095833
40	B1 / D3	130.775	142.232
41	B1 / C1	122.949	113.321
42	B1 / C2	121.919	0.998042
43	B1 / C3	155.688	150.665
44	B1 / C4	138.368	112.014
45	B1 / D4	0.84658	0.648731
46	B1 / D5	0.71356	0.568253
47	B1 / E1	193.088	20.313
48	B1 / E2	193.831	197.376
49	B1 / E3	187.357	193.471
50	B1 / E4	183.604	184.746
51	B1 / E5	195.022	204.547
52	B1 / D6	113.897	102.368
53	B1 / A3	172.116	189.927
54	A1 / A2	238.394	333.685
55	A1 / B2	112.715	0.545579
56	A1 / D3	167.067	0.937162
57	A1 / C1	141.894	106.501
58	A1 / C2	163.548	161.463
59	A1 / C3	193.643	211.756
60	A1 / C4	175.925	170.532
61	A1 / D4	107.262	112.799
62	A1 / D5	10.183	0.776242
63	A1 / E1	1.677	186.201
64	A1 / E2	168.834	182.362
65	A1 / E3	164.587	174.825
66	A1 / E4	161.364	167.434
67	A1 / E5	169.847	19.063
68	A1 / D6	126.808	153.865
69	A1 / A3	164.847	212.976
70	A2 / B2	276.881	295.196
71	A2 / D3	314.594	345.838
72	A2 / C1	265.566	261.788
73	A2 / C2	245.569	261.136
74	A2 / C3	26.299	26.478
75	A2 / C4	246.816	280.979
76	A2 / D4	267.517	223.378
77	A2 / D5	259.378	25.689

78	A2 / E1	26.703	281.078
79	A2 / E2	26.758	274.416
80	A2 / E3	265.881	284.466
81	A2 / E4	261.166	279.409
82	A2 / E5	271.066	272.136
83	A2 / D6	283.456	184.522
84	A2 / A3	192.179	157.128
85	B2 / D3	130.104	136.011
86	B2 / C1	120.941	112.769
87	B2 / C2	121.705	107.768
88	B2 / C3	155.675	158.486
89	B2 / C4	138.076	118.891
90	B2 / D4	0.860744	0.724392
91	B2 / D5	0.72186	0.587534
92	B2 / E1	191.656	202.119
93	B2 / E2	192.341	196.616
94	B2 / E3	185.972	192.137
95	B2 / E4	182.146	183.548
96	B2 / E5	193.685	2.04
97	B2 / D6	114.692	110.881
98	B2 / A3	169.892	194.811
99	D3 / C1	0.900775	0.840827
100	D3 / C2	221.952	240.586
101	D3 / C3	270.044	291.372
102	D3 / C4	238.751	254.199
103	D3 / D4	153.469	15.794
104	D3 / D5	116.242	113.252
105	D3 / E1	122.513	117.925
106	D3 / E2	120.344	11.738
107	D3 / E3	113.829	106.158
108	D3 / E4	112.428	102.161
109	D3 / E5	121.707	125.745
110	D3 / D6	18.125	19.356
111	D3 / A3	17.005	197.584
112	C1 / C2	172.357	194.033
113	C1 / C3	223.281	2.417
114	C1 / C4	18.664	212.771
115	C1 / D4	116.348	0.93212
116	C1 / D5	0.923596	0.591915
117	C1 / E1	102.145	0.89918
118	C1 / E2	101.703	0.840569
119	C1 / E3	0.984025	0.807142
120	C1 / E4	0.931671	0.718082
121	C1 / E5	107.017	0.912742

122	C1 / D6	147.026	119.681
123	C1 / A3	132.269	11.709
124	C2 / C3	0.591035	0.509096
125	C2 / C4	0.258339	0.229281
126	C2 / D4	139.553	104.723
127	C2 / D5	137.358	136.318
128	C2 / E1	253.005	280.564
129	C2 / E2	25.323	273.651
130	C2 / E3	24.891	273.056
131	C2 / E4	242.407	263.984
132	C2 / E5	257.667	279.277
133	C2 / D6	161.693	106.916
134	C2 / A3	193.943	223.434
135	C3 / C4	0.488358	0.442228
136	C3 / D4	172.064	149.731
137	C3 / D5	179.754	185.317
138	C3 / E1	301.089	326.003
139	C3 / E2	302.089	318.849
140	C3 / E3	297.137	319.203
141	C3 / E4	290.882	310.151
142	C3 / E5	305.791	323.942
143	C3 / D6	186.271	141.722
144	C3 / A3	234.745	25.628
145	C4 / D4	156.018	125.681
146	C4 / D5	155.382	154.187
147	C4 / E1	266.871	300.432
148	C4 / E2	26.749	293.658
149	C4 / E3	262.901	292.551
150	C4 / E4	256.392	283.491
151	C4 / E5	272.296	299.526
152	C4 / D6	176.123	129.798
153	C4 / A3	205.357	246.266
154	D4 / D5	0.564867	0.449408
155	D4 / E1	168.558	176.411
156	D4 / E2	170.689	169.349
157	D4 / E3	167.018	169.472
158	D4 / E4	164.024	16.042
159	D4 / E5	169.922	174.741
160	D4 / D6	0.422901	0.410917
161	D4 / A3	159.755	130.978
162	D5 / E1	146.544	148.731
163	D5 / E2	147.014	142.534
164	D5 / E3	141.581	13.985
165	D5 / E4	137.282	130.897

166	D5 / E5	14.798	1.493
167	D5 / D6	0.893304	0.834275
168	D5 / A3	138.878	139.136
169	E1 / E2	0.0866322	0.0769755
170	E1 / E3	0.14444	0.11908
171	E1 / E4	0.20773	0.188713
172	E1 / E5	0.112301	0.10004
173	E1 / D6	189.646	191.674
174	E1 / A3	134.174	124.714
175	E2 / E3	0.135504	0.124815
176	E2 / E4	0.186221	0.155536
177	E2 / E5	0.1168	0.0853669
178	E2 / D6	19.233	184.124
179	E2 / A3	134.335	117.778
180	E3 / E4	0.101269	0.0907164
181	E3 / E5	0.180806	0.199981
182	E3 / D6	189.622	186.891
183	E3 / A3	131.303	12.745
184	E4 / E5	0.250993	0.240897
185	E4 / D6	187.789	178.172
186	E4 / A3	126.095	122.282
187	E5 / D6	190.488	18.797
188	E5 / A3	137.708	116.245
189	D6 / A3	18.487	11.659

Table 15: Results of Sammon mapping using Dataset 5

	Analyses	Before reduction	After reduction
0	D1 / D2	318.845	28.673
1	D1 / A1	130.581	105.627
2	D1 / A2	0.635635	0.422124
3	D1 / D3	0.964929	121.376
4	D1 / C1	100.561	0.994249
5	D1 / C2	0.931725	0.654317
6	D1 / C3	0.898564	0.541704
7	D1 / C4	0.995184	0.517765
8	D1 / D4	131.326	103.385
9	D1 / D5	0.965831	0.969782
10	D1 / E1	147.994	11.478
11	D1 / E2	123.313	0.922846
12	D1 / E3	0.8881	0.4073
13	D1 / E4	141.172	0.95976
14	D1 / E5	130.613	0.98316
15	D1 / D6	387.306	351.712
16	D1 / A3	169.002	169.475

17	D2 / A1	26.543	223.699
18	D2 / A2	305.676	277.653
19	D2 / D3	306.789	321.012
20	D2 / C1	286.845	2.713
21	D2 / C2	266.058	25.125
22	D2 / C3	266.777	290.961
23	D2 / C4	264.081	245.311
24	D2 / D4	244.352	185.354
25	D2 / D5	294.425	351.126
26	D2 / E1	347.985	38.134
27	D2 / E2	324.435	378.566
28	D2 / E3	287.004	321.238
29	D2 / E4	323.211	342.269
30	D2 / E5	337.315	379.507
31	D2 / D6	137.385	13.103
32	D2 / A3	221.935	184.474
33	A1 / A2	10.407	0.717421
34	A1 / D3	153.129	212.946
35	A1 / C1	161.883	173.356
36	A1 / C2	15.463	12.976
37	A1 / C3	145.188	148.183
38	A1 / C4	151.521	107.516
39	A1 / D4	136.607	0.939312
40	A1 / D5	156.475	201.682
41	A1 / E1	187.647	162.806
42	A1 / E2	15.687	187.271
43	A1 / E3	138.445	145.897
44	A1 / E4	158.657	120.919
45	A1 / E5	174.963	170.428
46	A1 / D6	319.159	258.492
47	A1 / A3	151.541	186.841
48	A2 / D3	110.131	162.954
49	A2 / C1	127.604	136.017
50	A2 / C2	114.447	0.960435
51	A2 / C3	119.521	0.952814
52	A2 / C4	11.776	0.758036
53	A2 / D4	135.952	106.445
54	A2 / D5	103.073	137.721
55	A2 / E1	135.648	104.987
56	A2 / E2	127.848	115.545
57	A2 / E3	104.506	0.781282
58	A2 / E4	124.853	0.72164
59	A2 / E5	119.841	103.018
60	A2 / D6	370.616	32.791

61	A2 / A3	179.027	188.836
62	D3 / C1	0.99841	0.506104
63	D3 / C2	108.056	0.866104
64	D3 / C3	118.591	0.677276
65	D3 / C4	112.842	106.738
66	D3 / D4	144.334	153.797
67	D3 / D5	12.877	0.607009
68	D3 / E1	184.854	209.964
69	D3 / E2	165.523	137.017
70	D3 / E3	121.228	100.697
71	D3 / E4	171.266	208.583
72	D3 / E5	174.088	17.749
73	D3 / D6	373.096	41.883
74	D3 / A3	191.481	145.778
75	C1 / C2	0.695341	0.436895
76	C1 / C3	0.850778	0.507606
77	C1 / C4	0.816152	0.662541
78	C1 / D4	120.128	104.515
79	C1 / D5	126.351	0.886458
80	C1 / E1	166.316	206.679
81	C1 / E2	143.821	149.511
82	C1 / E3	101.597	0.970199
83	C1 / E4	149.169	194.808
84	C1 / E5	152.563	180.135
85	C1 / D6	360.932	368.341
86	C1 / A3	160.764	102.304
87	C2 / C3	0.637978	0.400063
88	C2 / C4	0.309352	0.226272
89	C2 / D4	102.603	0.705757
90	C2 / D5	113.639	100.126
91	C2 / E1	159.215	17.945
92	C2 / E2	141.458	138.573
93	C2 / E3	0.817943	0.79643
94	C2 / E4	148.381	160.735
95	C2 / E5	140.736	1.587
96	C2 / D6	338.948	337.419
97	C2 / A3	166.695	10.707
98	C3 / C4	0.677558	0.49334
99	C3 / D4	116.529	10.809
100	C3 / D5	116.983	0.616284
101	C3 / E1	160.116	156.004
102	C3 / E2	109.772	102.358
103	C3 / E3	0.538275	0.466024
104	C3 / E4	150.252	146.335

105	C3 / E5	142.234	129.636
106	C3 / D6	338.665	373.502
107	C3 / A3	142.074	142.135
108	C4 / D4	100.473	0.602794
109	C4 / D5	116.142	110.781
110	C4 / E1	152.118	166.299
111	C4 / E2	139.945	136.025
112	C4 / E3	0.803096	0.770132
113	C4 / E4	137.797	143.622
114	C4 / E5	141.694	149.434
115	C4 / D6	335.102	324.226
116	C4 / A3	1.735	117.884
117	D4 / D5	123.072	169.623
118	D4 / E1	18.941	210.019
119	D4 / E2	17.535	193.875
120	D4 / E3	130.877	135.946
121	D4 / E4	173.403	178.393
122	D4 / E5	177.281	200.349
123	D4 / D6	328.635	26.688
124	D4 / A3	135.979	0.930082
125	D5 / E1	157.815	157.586
126	D5 / E2	143.965	0.782924
127	D5 / E3	113.081	0.607741
128	D5 / E4	152.756	164.892
129	D5 / E5	142.706	122.422
130	D5 / D6	386.735	435.002
131	D5 / A3	162.853	190.837
132	E1 / E2	0.969311	0.851331
133	E1 / E3	117.765	111.211
134	E1 / E4	0.542325	0.43616
135	E1 / E5	0.523849	0.373806
136	E1 / D6	409.935	419.302
137	E1 / A3	239.415	284.171
138	E2 / E3	0.666499	0.593217
139	E2 / E4	105.565	106.173
140	E2 / E5	0.814159	0.478245
141	E2 / D6	389.494	441.774
142	E2 / A3	188.922	244.272
143	E3 / E4	114.897	108.312
144	E3 / E5	0.999442	0.831184
145	E3 / D6	355.605	391.425
146	E3 / A3	164.554	186.211
147	E4 / E5	0.831077	0.66151
148	E4 / D6	381.851	376.033

149	E4 / A3	222.517	2.599
150	E5 / D6	400.916	4.289
151	E5 / A3	220.606	265.564
152	D6 / A3	310.507	299.728

Table 16: Results of Sammon mapping using Dataset 6