TECHNICAL UNIVERSITY OF CRETE

Mineral Resources Engineering School

Performance of multivariate clustering methods in oil families' identification

By

Christina Karavoulia

Scientific Advisor

Prof. Nikos Pasadakis

Examination Committee:
Prof. N. Pasadakis
Prof. D. Christopoulos
Prof. V. Gaganis

Diploma thesis

Submitted in part fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN PETROLEUM ENGINEERING

Abstract

As science progresses, the need for analyzing multivariate data sets is growing by the minute. Multiple disciplines, either scientific or not, require the examination of large amounts of data, in a short period of time, in order to obtain useful information. During the recent few decades, multivariate statistical analysis methods have been developed, aiming to satisfy such purposes.

This dissertation deals with the implementation of multivariate data analysis methods on a given data set, derived from oil family affiliations, which originate from Williston Basin of North America. In particular, Hierarchical Clustering, k-means and Principal Component analysis have been applied on four independent models, in an attempt to extract information regarding the oil-oil correlations among the samples under study. The models used on the exploration of the compositional information were the Saturated Fraction Compositional Model, the Saturated Fraction Ratios Model, the Gasoline Range Compositional Model and the Biomarkers Compositional Model.

These standard statistical methods were found to be quite insufficient in classifying the sample set into distinct familial affiliations. For this reason, the need to examine the nature of the data set arose. Compositional data represent a category on their own as they are characterized by specific numerical properties which present significant consequences when being analyzed by standard multivariate techniques. The analysis of such type of data represents a whole new chapter in the world of statistics and the need for further examination on this matter is constantly growing.

Acknowledgments

Foremost, I would like to express my sincerest gratitude to my supervisor Prof. Nikos Pasadakis for his continuous support, the valuable comments, remarks and engagement through the learning process of this master thesis. His patience, motivation, enthusiasm, and immense knowledge helped me throughout the whole research and writing of this project.

Most importantly, I must express my very profound love and gratitude to my parents Theoni and Thanasis as well as my beloved sister Angeliki, for keeping me harmonious, providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. To my dearest friends outside the academic world, Elena and Nikos for being there whenever I was on the verge of losing my strength and for helping me putting pieces together, a simple *thank you* might never be enough. Polyanthi, I truly thank you for keeping me awake all these endless nights and "yelling" at me every morning to get out of bed. Studying would have been a torture without your positive and cheerful spirit. Apart from the honor to meet a great colleague, I am deeply thankful for making a friend like you.

This accomplishment would not have been possible, however, without the generous contribution of my uncle Tasos, who supported me in this step, in multiple ways. His encouragement to pursue this career path, has been the core of my persistence in carrying through this work, even at times when I felt like losing my courage.

Chania, May 2017

Christina Karavoulia

Table of Contents

List of Figures

List of Tables

1. Introduction

Over the last decades an overwhelming amount of data is poured into our lives and obtaining meaningful information out if them is an imperative task for people. Multiple disciplines such as chemistry, biology, medicine etc. demand the analysis of huge amounts of data and sometimes their multivariate nature makes it difficult to analyze. For this reason, special statistical techniques have been developed in order to process information in a meaningful fashion.

In this project, multivariate clustering methods have been implemented on geochemical data concerning oil family affiliations that exist in Williston Basin, North America, in order to explore the oil-oil correlations. The methods which have been utilized consider both Supervised and Unsupervised learning phases. These include Hierarchical Clustering, k-means clustering, as well as Principal Component Analysis. The ultimate goal of this project is to test how well such multivariate analysis methods perform, as far as classification of the compositional data is concerned.

The thesis project is organized into seven chapters. In Chapter 2 a detailed description of the geological setting of Williston Basin is presented. The stratigraphy and the tectonic regime are thoroughly described and special focus is placed upon the geochemical classification of oil families which have been recognized in the area.

Chapter 3 raises the subject of Multivariate Data Analysis (MDA). It provides a brief presentation of the principles of Hierarchical Clustering, k-means clustering as well as Principal component analysis. All the main concepts that characterize each method are included.

In Chapter 4 we discuss the matter of the existing Family Affiliations of Williston Basin. In this chapter, there is an attempt to test the criteria under which the classification of the oil families was determined.

Chapter 5 deals with the application of multivariate data analysis methods on two different models; the Saturated Fraction Component Model and the Saturated Fraction Ratios Model. All MDA methods were implemented on both models and the results are discussed briefly.

In the final Chapter (6) the subject of Compositional Data, as a special type of data, is introduced. In this chapter, we analyze the properties of Compositional Data as well as the methodology with which, such kind of data should be treated.

2. Geological Setting of Williston Basin

The Williston Basin is an intracratonic, sub-circular sag basin that comprises main part of the North American craton. In particular, it forms a large depression in the western edge of the Canadian shield, occupying much of North Dakota, northwestern South Dakota, the eastern quarter of Montana, a significant part of southern Saskatchewan, and a portion of southwestern Manitoba. Among these regions major production of oil and gas occurs. Williston Basin is characterized by Phanerozoic, carbonate and clastic sedimentation of more than 16,000ft strata thickness in its central part, near Watford City, North Dakota [1, 2]. Having undergone episodic and prolonged subsidence rates, it comprises a preservational basin and it is composed by six major depositional sequences, each bounded by larger structural features [2, 3, 4] (Fig. 1). The basin is neither considered structurally complex nor tectonically active and its well -established petroleum provinces, clearly described rock succession, modest burial history and simple tectonics make this an uncomplicated area to study.

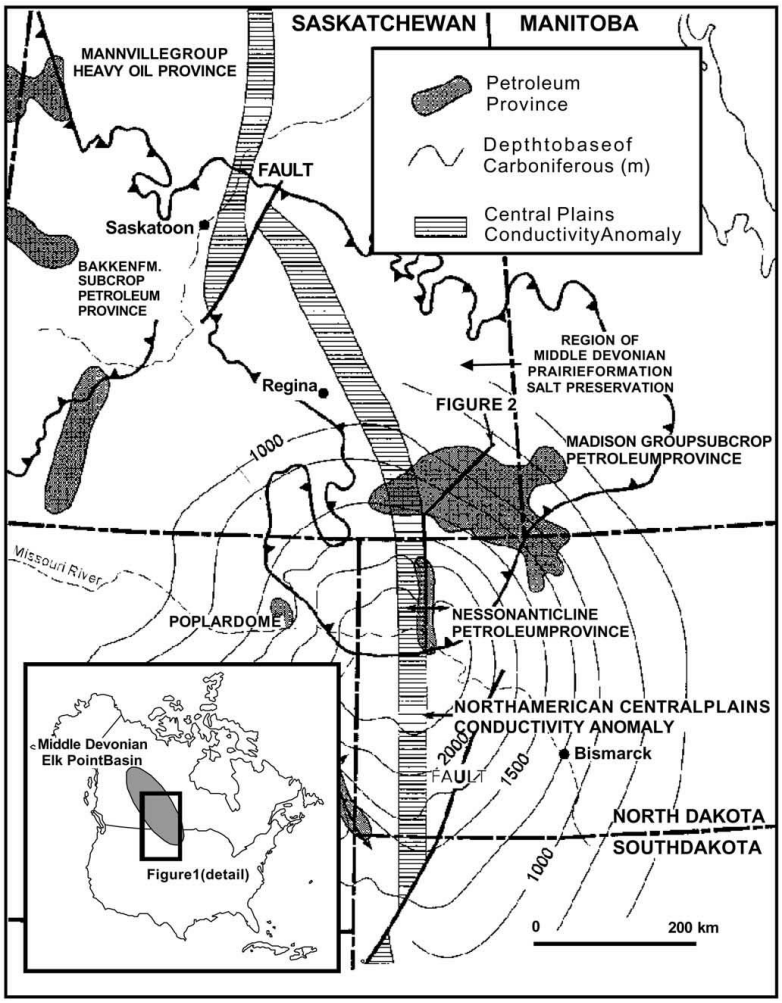


Fig. 1 Location map showing the main geological and geophysical elements of Williston Basin and environs. The region of anomalous subsidence that is Williston Basin proper (Ahern and Mrkvicka, 1984) is generally coincident with the 1 km depth contour on Carboniferous strata. The region of preserved Middle Devonian Prairie Formation salt deposited in Elk Point Basin is illustrated. The inset shows the location of Williston Basin and the extent of Elk Point Basin. Samples from petroleum pools entrapped at the subcrop of the upper Paleozoic succession in southeastern Saskatchewan and southwestern Manitoba, as well as American samples constitute the sample set for this study (following Burrus et al., 1996a).

Williston Basin is discretized into the American and the Canadian portions. The American portion of the basin is influenced by major deformational features, mainly anticlines (Fig. 2). The Canadian part of Williston Basin forms a petroleum province where oil production is quite active. Petroleum accumulations mainly occur in stratigraphic traps within the Phanerozoic succession [5]. There is, however, variety of trapping features which are structurally linked to Precambrian basement [6, 1, 7]. In southwestern Manitoba and southeastern Saskatchewan, oil exists around the Mississippian subcrop. In southwestern and west-central Saskatchewan, oil exists in stratigraphic traps within latest Devonian to Mississippian, Jurassic, and Lower Cretaceous formations.

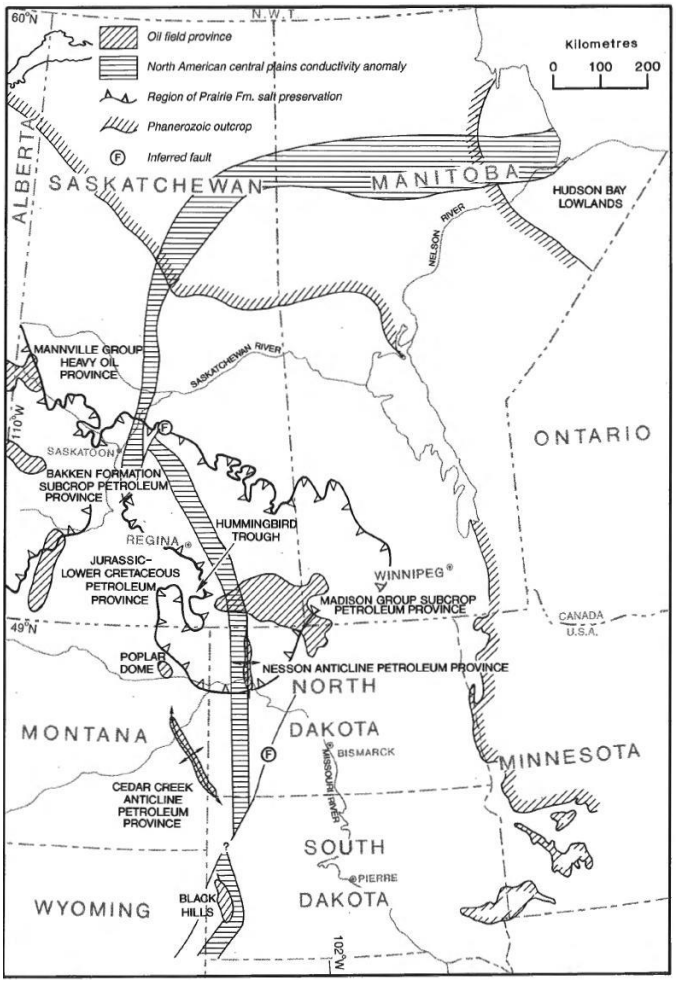


Fig. 2 Petroleum region and crucial tectonic elements in the Williston Basin and adjacent area. Only generalized outlines of the Mississippian Madison Group Subcrop Petroleum Province and other Williston Basin petroleum provinces are indicated.

2.1 Stratigraphy of Williston Basin

The Williston Basin forms a large, roughly circular depression on the North American Craton. Its sedimentology is characterized by Paleozoic and Cenozoic – Mesozoic carbonate and clastic deposition, accordingly with a thickness of strata that exceeds 16,000 ft in the basin's core (Fig. 3).

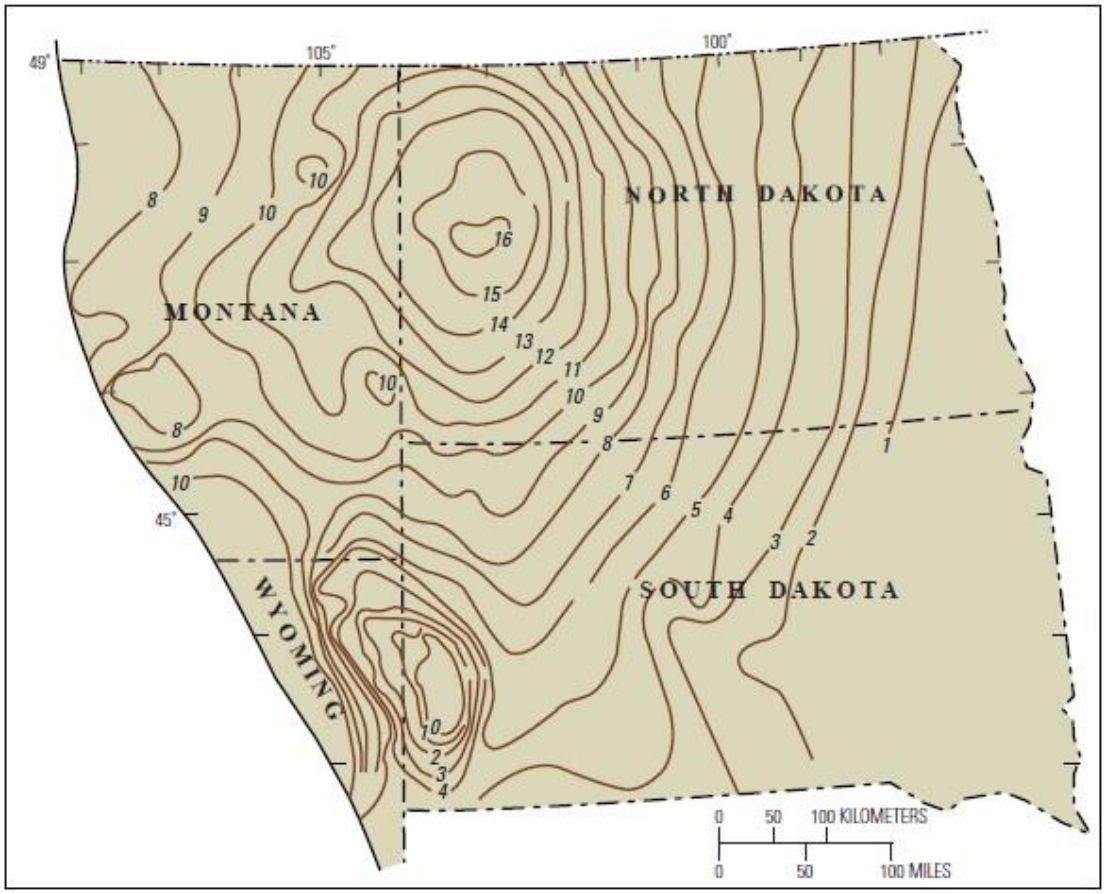


Fig. 3 Contour map pf Williston Basin presenting the thickness of sediments. Contour interval is 1,000 ft. [8]

There are six main depositional sequences, each bounded by major unconformities [3], which can be distinguished within the Phanerozoic succession of North American portion of the basin. The formulation of unconformities resulted in numerous processes affecting its final structure, such as primary and secondary dissolution, deposition of salt and anhydrite beds, and secondary dolomitization of limestone. Clastic deposition initiated in Mesozoic and Cenozoic Eras, including mudstone, sandstone, siltstone, coal and shale. All depositional sequences are briefly described in the following paragraphs.

Sauk Sequence (Middle Cambrian – Lower Ordovician)

The Sauk sequence was deposited on the early Paleozoic miogeocline of western North America [7, 9], and is composed of Upper Precambrian sediments, interrupted by minor transgressions and regressions, which create several sub-members within the formation [10]. Saul deposition, mainly represented by Deadwood formation, includes shallow marine, coastal and alluvial plain sediments along with sandstone, mudstone and siltstone successions and finalizes due to the activity of an unconformity.

Tippecanoe Sequence (Ordovician – Silurian)

The Tippecanoe sequence marks the beginning of Ordovician clastic, carbonate and evaporitic sedimentation. From bottom to top, it consists of Winnipeg, Red River, Stony Mountain and Stonewall Formations, each unconformably overlying the other (Fig. 4). Upper Ordovician

rocks of this sequence contain important petroleum sources. Depositional processes terminate at the end of the Silurian due to major regression activity.

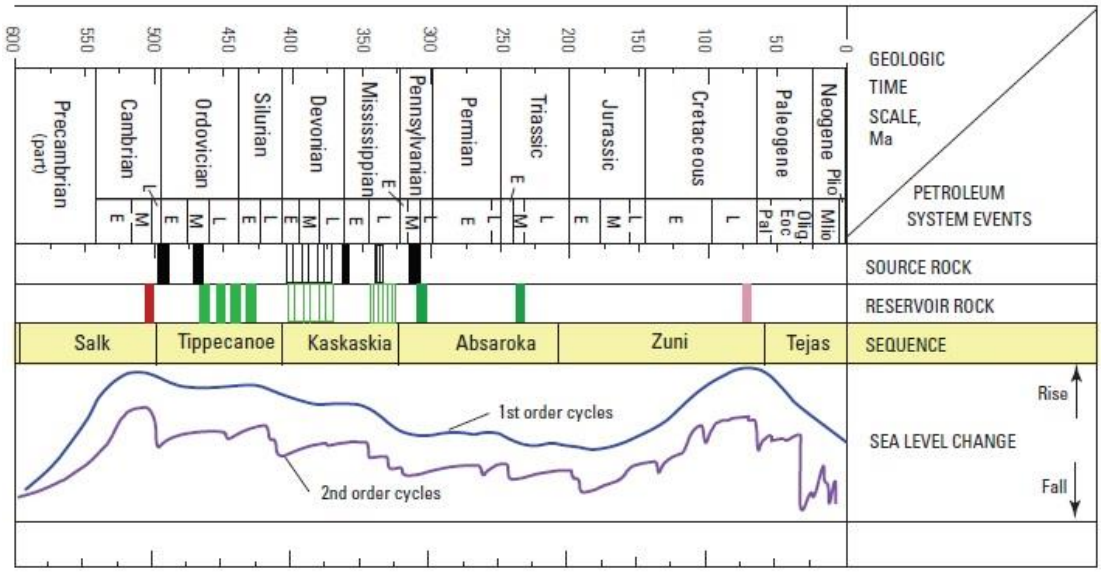


Fig. 4 Diagram showing geologic time scale, major stratigraphic sequences of [3], first- and second order sea level curves from [11], and ages of petroleum source and reservoir rocks in the Williston Basin. Solid black interval in source rock column are for thick accumulations; thin lines indicate association with carbonate depositional cycles. In reservoir rock column, green is for oil and red is for gas; thin lines indicate generalized reservoir rock and do not necessarily represent the full spectrum of possible reservoirs. E, Early; M, Middle; L, Late; Pal, Paleocene; Eoc, Eocen; Olig, Oligocene; Mio, Miocene; Plio, Pliocene (following Lawrence , et al., 2013).

Kaskaskia Sequence (Devonian - Mississippian)

The Kaskaskia sedimentation cycle initiated in Ordovician, continued to Jurassic and concluded due to transgressional activity. Three main transgressional events impacted on the depositional history of the sequence, during which several formations were deposited. The most significant is the Bakken Formation which represents the first major input of clastic material into the Williston Basin since the Cambrian Deadwood and Winnipeg Formations. Bakken marks a change in Kaskaskia sequence depositional patterns and sedimentation style [12, 13] and it is the most important interval for petroleum source rocks in the Williston Basin. In general, the Kaskaskia Sequence is stratigraphically characterized by subtidal, intertidal and rare supratidal depositional environments.

Absaroka Sequence (Pennsylvanian - Triassic)

The Absaroka Sequence includes the Tyler and the Minnesula formations and mainly occurs in the American portion of the Williston Basin. It is vastly affected by major unconformities, occurring near the end of Pennsylvanian, Permian and Triassic [14] and contains effective oil source rocks [15, 16].

Zuni Sequence (Jurassic – Early Tertiary [Eocene])

Two major transgressional events influenced the depositional history of the Zuni Sequence, which is characterized by shallow marine and clastic sediments. Sedimentation terminated during early Paleocene and the sands of the Dakota Group are likely the most significant targets for sequestration in the Zuni Sequence. This sequence can be locally subdivided into

two other sequences. The first includes the Jurassic, when Williston Basin changed from a large reentrant on the craton margin into an orogenic foreland [17, 18]. The lower sequence contains a time equivalent succession to the last cratonically derived miogeoclinal succession.

Tejas Sequence (Tertiary - Quaternary)

Latest Jurassic and Cretaceous successions of the Columbian and Laramide orogenic forelands [19] form the final significant depositional episode [20, 21]. Thick shales of this final sequence include significant probable source rocks, but they are all immature in the Canadian Williston Basin. The first produced hydrocarbons in North Dakota were from the youngest strata in the state, glacial drift of the Tejas Sequence. However, there is no production from glacial drift today.

2.2 Tectonic Regime of Williston Basin

In order to understand the Williston Basin's evolution, structural configuration, sedimentation, and thermal patterns, one must refer to the geological history of the Precambrian basement underlying the basin.

Two critical structures have influenced the evolution of the basin; the Trans-Hudson orogenic belt [22] and the northeast–southwest trending Proterozoic lineament and structural zones [23]. The Trans-Hudson belt sutured the Archean Superior craton to the Archean Wyoming craton (Fig. 5A, B); the resulting collision created a north–south trending strike-slip fault and shear belt. A basin center was created, caused in part by later folding of the Trans-Hudson orogenic belt and rifting [24], although Nelson et al., [25] stated that there is a lack of direct evidence of a rift.

The northeast–southwest trending Proterozoic lineament and structural zones were renamed as the Transcontinental arch, Brockton-Froid fault zone, Great Falls tectonic zone, Poplar fault, and Hinsdale fault. These Precambrian structures were reactivated during the Neoproterozoic, which resulted in the creation of new north–south and northwest–southeast trending structures.

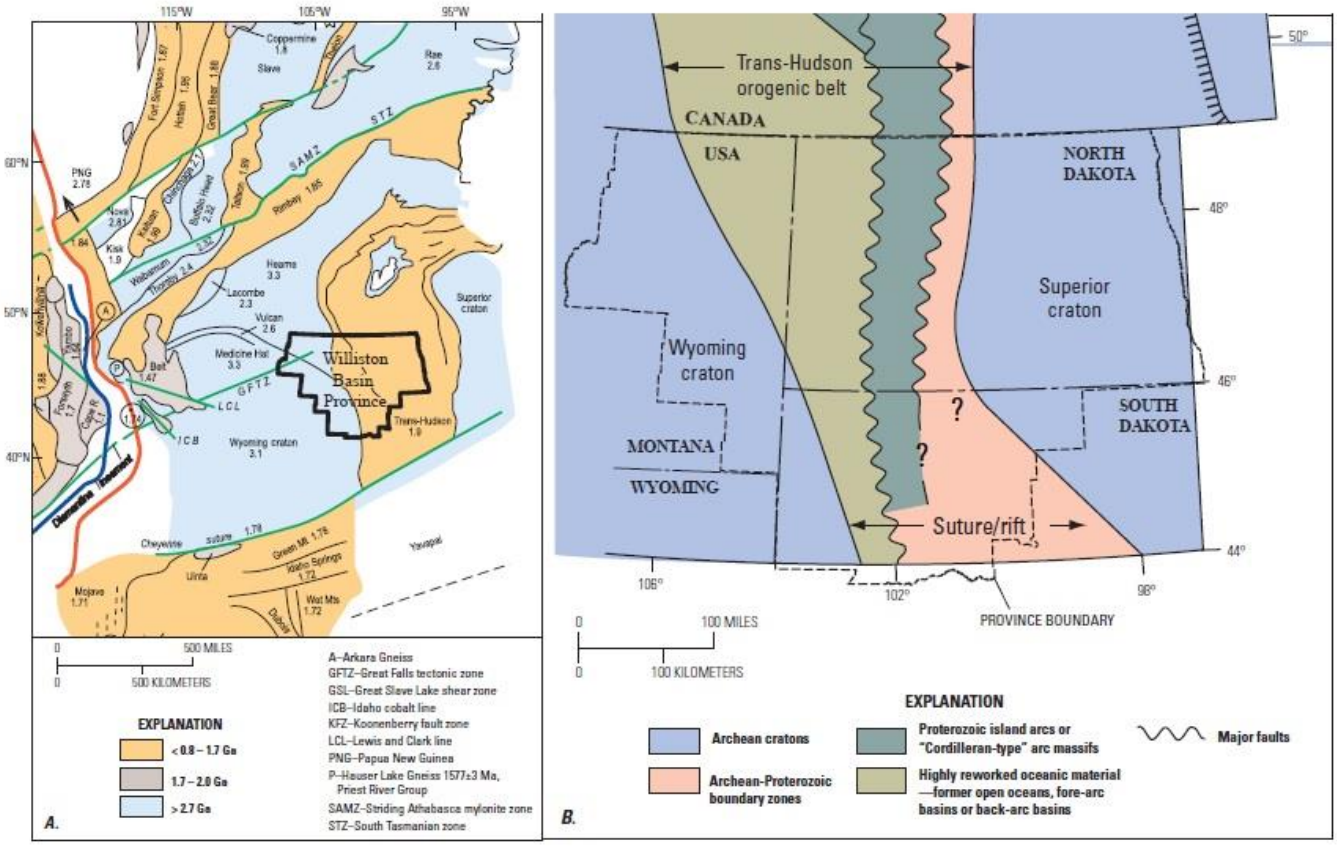


Fig. 5 Precambrian structural configuration of the Williston Basin and surrounding area. A: Tectonic map of the northern Great Plains region [23] showing northeast-southwest strike slip faults; Williston Basin province outline is shown for scale. Ga, billion years ago. B: Map showing the configuration of Trans-Hudson orogenic belt and associated north-south trending structures of the Williston Basin (modified Nelson et al., 1993).

Numerous studies have shown that surface lineament patterns in the Northern Great Plains region, including the Williston Basin, are a result of the aforementioned reactivation of Precambrian faults during the Phanerozoic [26, 27, 28, 29]. These studies show pervasive northeast–southwest and northwest–southeast trends that are parallel to major lineaments of Proterozoic terrane. North–south trending lineaments that are parallel to the Trans-Hudson structural system are less prominent, although north–south thermal patterns are evident from present-day subsurface temperature measurements.

Based on several observations, it is believed that Precambrian tectonic events and their recurrent movement along preexisting zones of weakness played a major role in the development of most of the major fault and shear systems in the Williston Basin. Although the basin is generally reported as a depression and tectonically inactive, its final structure is thought to be mostly formed as a result of structural deformation and down-to-the-basin block faulting from Precambrian rooted structures, as well as from deformation related to the Trans-Hudson orogenic belt.

2.3 Geochemical Classification of Oil Families in Williston Basin

Classification of oil families in the Canadian portion of the Williston Basin has been attempted by a number of investigators over the past decades. Dow and Williams, in their 1974 papers,

were the first researchers to apply the 'petroleum system' concept, identifying three oil systems in the Williston Basin, relying mainly on stable isotopic and gasoline range hydrocarbon composition: Tyler, Bakken, and Winnipeg [15, 16]. Each oil system is associated with a unique oil type. Type I refers to Ordovician and Silurian oils which originate from Middle Ordovician Winnipeg shale sources. Type II oils occur in Upper Devonian, Mississippian and Mesozoic reservoirs, and are probably linked to Fammenian – Tournaisian Bakken Formation Source rocks. Type III refers to Pennsylvanian oils which originate from Tyler Formation source rocks.

Most recent studies, however, have defined at least nine oil systems in the area. Zumberge [30] and Leenheer and Zumberge [31] defined five oil families based on the study of samples from the American part of the Williston Basin, while, Osadetz et al., [32] categorized oils from the Canadian part of the Basin (southeastern Saskatchewan and southwestern Manitoba) into four compositional families (Table 1). The criteria under which the classification of the latter was conducted, include pristane/phytane (Pr/Ph) ratio, n-alkane predominance, C_{23} tricyclic/C_{30} pentacyclic terpane ratio and prominence amongst extended hopanes.

In particular, Family A oils occur in Ordovician to Middle Devonian and Upper Ordovician formations and match solvent extracts from kukersites (marine Type I rocks) of the Late Ordovician Binghorn Group [32, 33], rather than, as initially suggested, extracts from Winnipeg shales [15, 16]. Oils of this family present diagnostic saturate fraction gas chromatograms (SFGC), low C_{23}tricyclic/C_{30} pentacyclic terpane ratios (<0.20) and a strong C_{34} hopane prominence. They can be further subdivided into a group distinguished by low Pr and Ph, relative to faster eluting n-alkanes nC_{17} and nC_{18}, a strong odd-even predominance among n-alkanes between C_{15} and C_{20}, and a low relative abundance in higher carbon number n-alkane homologues [34].

Family B oils primarily occur in Bakken reservoirs [32, 35, 33], they are however, also found in early Cretaceous reservoirs. They are sourced from Type II marine organic matter in the Upper Devonian-Mississippian Bakken Formation shale members. Main characteristic of this family is that it displays the highest Pr/Ph (>1.50) and C_{23}/C_{30} (>0.80) ratios, accompanying n-alkane and hopane profiles, without any predominance and prominence respectively.

Table 1 Table showing all groups and oil families, in correlation with the according formations, present in Williston Basin (modified by Osadetz, 1994)

Williams, 1974	Zumberge, 1983; Leeheer and Zumberge, 1987	Osadetz et al., 1992, 1994	Source rocks
Type III (Pennsylvanian oils) not studied	Not studied	Not studied	Tyler Fm. (Pennsyl.)
		Family E (Bakken oils)	Exshaw/Bakken Fm. (U. Dev.-Miss.)
Type II (Devonian, Mississippian & Mesozoic oils)	Group 2 (Mission Canyon oils)	Family B (Bakken oils)	Bakken Fm. (U.Dev.-Miss.)
		Family C (Miss. & Jurassic oils)	Lodgepole Fm. (L. Miss.)
Not studied	Group 4 (Nisku oils) Group 3 (Duperow oils)	Family D (Winnipegosis oils)	Winnipegosis Fm. (M.Dev.)
Type 1 (Ordovician-Silurian oils)	Group 1 (Red River oils) Group 5 (Cambrian oil)	Family A (Red River oils) Not studied	Winnipeg Gr. (M. Ord.) and Bighorn Gr. (U.Ord.) unknown (?U.Cam.-Ord)

Family C oils occur the Mississippian Madison Group and Mesozoic formations and are sourced from Type II marine rocks in the Mississipian Lodgepole formation. They present high C_{23}/C_{30} (>0.20) ratio but, compared to Bakken sources, lower Pr/Ph ratio (<1.1), a pronounced (>nC_{20}) even n-alkane predominance and a strong C_{35} prominence.

Finally, Family D oils occur in Silurian to Mississippian sediments. They originate from Middle Devonian Winnipegosis Formation marine rocks, which vary in terms of depositional background. In particular, there are two kinds of settings; the platform depositional and starved basinal. Family D oils display similar terpane compositional characteristics to kukersite derived oils (abundant Pr, Ph and generally complex SFGCs), they differ however, in that they present greater relative acyclic isoprenoid and higher carbon n-alkane abundance. Oils of D Family, are further discretized into D_1 platformal and D_2 starved basinal, based on nC_{17}/Pr and nC18/Ph ratios. They display higher nC17/Pr ratios for a given nC18/Ph ratio compared to otherwise similar oils that occur in overlying Saskatchewan and Manitoba groups' strata, and they belong to the Elk Point Group of Winnipegosis reef formulations. Group D_1 predominantly occurs in younger Devonian reservoirs, lacking however, clear source definition. Suggested possible source rocks are thin organic-rich beds in Winnipegosis platform carbonates, the Birdbear Formation, and some Upper Devonian rocks. Group D_2 occurs in pinnacle reefs of the Middle Devonian Winnipegosis Formation and the Brightholme Member comprises the source rock. Oils having similar molecular compositions to D_2 oils have been found in the Upper Cambrian Deadwood Formation, Silurian pools of the Nesson Anticline, and new discoveries in the Middle Ordovician Winnipeg Formation. They have, however, very different isotopic compositions of carbon and sulphur, suggesting that a still-undescribed petroleum system exists in Paleozoic strata [36]. Family D oils correlate to Groups 3, 4, and 5 of Leenheer and Zumberge [31].

As previously mentioned, the compositional classification of the Williston Basin petroleum, relied much on terpane, sterane, and select n- and iso- alkane characteristics. The original classification by Williams [16], however, took into consideration the gasoline range fraction

(GRH) and later studies, based on that scheme, came to agree that families A - D and families B - C were inseparable and consistent with oil Types I and II [37]. Most recent work, depends on multivariate statistical methods, such as Principal Component Analysis (PCA), combined with geological information, in an attempt to enhance the independent interpretation of GRH and SFH fractions [32, 38]. Findings show that, while Family A oils can be uniquely classified, oils from Families B, C and D present insufficient characteristics for independent classification. Especially the composition of Family C seems to be quite heterogenous, often overlapping with families B and D [39].

This is attributed to the mixing of oils derived from different sources, without however, the extent to which this process occurs, having been defined [38]. A characteristic example of that mixing is the relative effectiveness of Bakken and Lodgepole petroleum systems [40, 41]. While part of the scientific community suggests that mixing is rare in the American portion of the basin [42, 43, 44], there is another part, proposing that major mixing is possible, without an impact on the biomarker traits [45, 46]. What is to account for the inability to precisely define the extend of mixing sources, is either the neglection of current interpretive techniques or the semi-quantitative confirmation of the biomarker based classification in the GRH and SFH [34, 39].

3. Exploratory Data Analysis

Analysis of Data (DA) constitutes the science of collecting, organizing and examining raw data under the purpose of obtaining useful and usable information for decision-making by users. The analysis may describe and summarize the data, identify relationships among variables, compare and identify differences between them as well as forecast outcomes. Data analytics is distinguished from data mining, which is a particular data analysis technique, by the scope, purpose and focus of the analysis. The target of Data Mining is rather predictive than descriptive. Data miners sort through huge data sets using sophisticated software to identify undiscovered patterns and establish hidden relationships. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher. Statistician John Turkey defined the term "Data Analysis" in 1961 as: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

Turkey [47] distinguished in Data Analysis techniques and procedures, two major groups: Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA). In EDA analysts make a few assumptions under the purpose of suggesting hypotheses and according to Turkey it is a rather detective work. In contrast, CDA "quantifies the extent to which deviations from a model could be expected to occur by chance" [48]. Confirmatory Data Analysis utilizes the traditional statistical tools of inference, significance, and confidence.

As a scientific tool, DA can be further subdivided in alternate groups. Therefore, based on the quantity of variables examined, Data Analysis can be dichotomized into Univariate (UDA) and Multivariate (MDA). Univariate data analysis is conducted when one variable is used for one observation. Subsequently, it makes sense to state that Multivariate data analysis is used when more than one outcome variables are measured and it is concerned with the study of association among sets of measurements. It is referred to as any statistical technique used to analyze data that arises from more than one variable.

3.1 Multivariate Data Analysis (MDA)

This project will focus on MDA techniques that will be implemented on the given data set and the outcomes will be examined thoroughly. Multivariate Data Analysis can fall into two phases: Unsupervised learning and Supervised learning. The goal of unsupervised learning is the detection of hidden structure in unlabeled data and encompasses many techniques that seek to summarize and explain key features of the data (i.e. Clustering Analysis, PCA). Supervised learning is a task of inferring a function from labeled training data. Each example on training data is a pair consisting of an input object (typically a vector) and a desired output value (i.e. Classification Analysis). In general, supervised methods are used when the aim is the construction of a model to be used to classify future samples [49].

There are several clustering techniques established by the scientific community, all governed by some kind of taxonomy [50, 51]. A major distinction among them involves the Hierarchical and the Partitional approaches, which are based on whether the set of produced clusters is nested or unnested. A Hierarchical clustering leads to a set of nested clusters that are

organized as a tree, whereas a Partitional clustering formulates a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. Characteristic examples of algorithms derived from the aforementioned approaches are agglomerative or divisive, deterministic or stochastic, incremental or non-incremental, monothetic or polythetic and hard or fuzzy [51].

3.1.1 Hierarchical Clustering

Hierarchical Clustering Analysis is an unsupervised technique that examines the interpoint distances between all of the data objects and generates a tree diagram or dendrogram on which, that information is visualized. It can be considered both as a sequence of nested partitions and the similarity levels at which these change [51, 52]. Hierarchical clustering algorithms are either bottom-up (agglomerative) or top-down (divisive). At each step of the agglomerative hierarchical approach each subject is treated as a singleton cluster which is successively merged into the closest cluster [51, 49, 53]. This process is repeated until all clusters have been merged into a singleton cluster that contains all subjects. The alternate divisive approach, begins with a single cluster containing all subjects, and at each step, the cluster splits until N clusters form (each with a single subject).

The criterion under which clusters are merged or split, differentiates at each case. Since the bottom-up approach agglomerates pairs of clusters with the minimum distance, measures of similarity and dissimilarity have to be taken into account. Those measures are defined by linkage functions which have a direct impact on the whole clustering procedure. They affect the way clusters are merged together and subsequently the final cluster solution. Therefore, linkage measures will be discussed extensively in the process.

The following notation is given in order for the various linkages to be described:

- Cluster r is formed from clusters p and q.
- n_r is the number of objects in cluster r
- x_{ri} is the ith object in cluster r

Single Linkage (**Nearest Neighbor**) functions utilize the shortest distance between any two objects in a pair of clusters [54, 55]:

$$d(r,s) = \min\left(dist\left(x_{ri}, x_{sj}\right)\right), i \in (i, \dots, n_r), j \in (1, \dots, n_s)$$

The similarity under which two clusters merge is the similarity of their most similar objects and the merge criterion is local. Single linkage is a bottom-up (agglomerative) process where the number of clusters is reduced by one at each step.

Complete Linkage (**Furthest Neighbor/Maximum Method**) functions utilize the furthest distance between any two objects in a pair of clusters [55]:

$$d(r,s) = \max\left(dist\left(x_{ri}, x_{sj}\right)\right), i \in (i, \dots, n_r), j \in (1, \dots, n_s)$$

Accordingly, in complete linkage method, the similarity under which two clusters fuse, is the similarity of their most dissimilar objects and the merge criterion is non-local, that is, the entire structure of clustering can affect the way how clusters fuse.

Average linkage functions utilize the averaged distance between all pairs of the two clusters' members [55]:

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri} x_{sj})$$

There is also, an average linkage method within groups, proposed by Sokal & Michener [55], which takes into consideration the variability present within each cluster. This method will not be further discussed.

All the three methods mentioned above (single, complete and average) use a proximity matrix as input and the inter-cluster distances used are presented in Fig. 6.

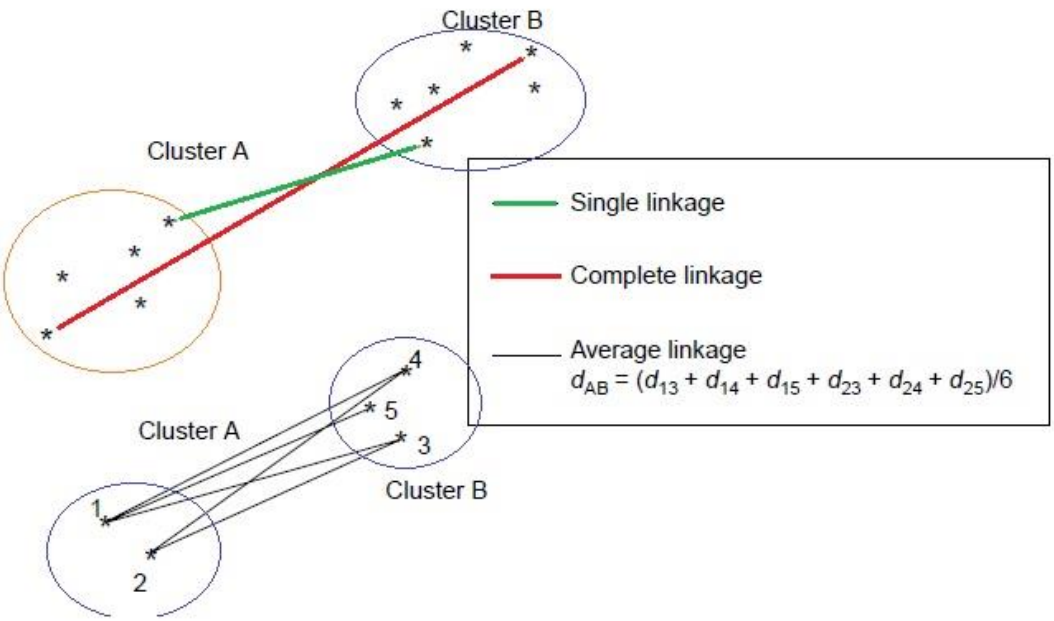


Fig. 6 Single, Complete and Average linkage graphical representations, modified after [56].

Centroid linkage (**Unweighted Pair-Group Method using the centroid approach- UPGMC**) utilizes the Euclidean distance between the centroids of the two clusters:

$$d(r,s) = \|\widetilde{x_r} + \widetilde{x_s}\|_2$$

where $\widetilde{x_r} = \frac{1}{n_r} \sum_{i=1}^{n_r} x_{ri}$

As single linkage method, centroid linkage also represents an agglomerative approach to hierarchical clustering. This approach uses a data matrix, in contrast to the previous ones, rather than a proximity matrix and involves merging clusters with the most similar mean vectors.

Median linkage (**Weighted Pair-Group Method using the centroid approach**) functions also utilize the Euclidean distance between the weighted centroids of the two clusters:

$$d(r,s) = \|\widetilde{x_r} - \widetilde{x_s}\|_2$$

where $\widetilde{x_r}$ and $\widetilde{x_s}$ are weighted centroids for the clusters r and s. If cluster r was created by combining clusters p and q, $\widetilde{x_r}$ is defined recursively as:

$$\widetilde{x_r} = \frac{1}{2}(\widetilde{x_p} + \widetilde{x_q})$$

Apart from the Euclidean distance, other proximity measures may be used for the Centroid and the Median linkage approaches, they would, however, lack interpretation in terms of the raw data [56]. The following table (Table 2) presents a brief description of various proximity measures used in linkages.

Table 2 Several Computational methods for distance

Distance measures	Formula
Euclidean Distance	$\|a-b\|_2 = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$
Squared Euclidean Distance	$\|a-b\|_2^2 = \sum_{i=1}^{n}(a_i - b_i)^2$
Manhattan/City block DIstance	$\|a-b\|_1 = \sum_{i=1}^{n}\|a_i - b_i\|$
Maximum Distance	$\|a-b\|_\infty = max\|a_i - b_i\|$
Mahalanobis Distance	$\sqrt{(a-b)^T S^{-1}(a-b)}$ *where S is the covariance matrix*

Ward's Method aims to minimize the variance between clusters by utilizing an incremental sum of squares: that is, the increase in the total within-cluster sum of squares as a result of joining two clusters [57]. The within-cluster sum of squares is defined as the sum of the squared distances between all objects in the cluster and the centroid of the cluster. The sum of squares measure is equivalent to the following distance measure d(r,s), which is the formula linkage:

$$d(r,s) = \sqrt{\frac{2n_r n_s}{(n_r + n_s)}}\|\widetilde{x_r} + \widetilde{x_s}\|_2$$

Where:

- $\|\widetilde{x_r} + \widetilde{x_s}\|_2$ is the Euclidean distance
- $\widetilde{x_r}$ and $\widetilde{x_s}$ are the centroids of the clusters r and s
- $\widetilde{n_r}$ and $\widetilde{n_s}$ are the number of elements in clusters r and s

In some references, factor of 2 multiplying $n_r n_s$ is not utilized by Ward's method. The linkage function uses this factor so that the distance between two singleton clusters is the same as the Euclidean distance. Ward's method differs from the centroid approach in clustering, in that centroids are weighted by $n_r n_s/(n_r + n_s)$ when computing distances between centroids, where n_r and n_s are the numbers of objects in the two clusters r and s.

Finally, ***Weighted Average Linkage*** (**WPGMA**) utilizes a recursive definition for the distance between two clusters [58]. If cluster r was created by combining clusters p and q, the distance between r and another cluster s is defined as the average of the distance between p and s and the distance between q and s:

$$d(r,s) = \frac{(d(p,s) + d(q,s))}{2}$$

There are several other hierarchical approaches, related to the ones described above. There is the Sum-of-Squares Approach [59, 60] which differs from Ward's method in that it is based on the sum of squares within each cluster rather than the increase in sum of squares in the merged cluster. Another flexible method defined by values of the parameters of a general recurrence formula has also been introduced by Lance and Williams [61] but in this project, it will not be discussed any further.

3.1.2 k - means Clustering

The k-means algorithm is one of the most used clustering algorithms and it was first described by Macqueen [62]. It was designed to cluster numerical data in which each cluster has a center called the mean. k-means belongs to the partitional (non-hierarchical) clustering methods [50], which are fundamentally different from the hierarchical ones. Partitional clustering methods generate a single partition of the data in an attempt to recover natural groups in the data. While hierarchical clustering methods require only the proximity matrix among the data points, partitional techniques expect the data in the form of a pattern matrix.

k-means [62] is one of the simplest unsupervised learning algorithms, which is used to solve the well-known clustering problem. The goal of k-means method is to divide the data into k distinct groups (clusters) so that observations within a group are similar, whilst observations between groups are different. The value of k (number of clusters) may or may not be specified. In most cases, it is assumed to be fixed. As an algorithm, it is rather iterative than hierarchical, which means that at each stage of the algorithm data points are assigned to a fixed number of clusters (whereas in hierarchical clustering, the number of clusters ranges from the number of data points down to a single cluster). The method allows the reallocation of data objects from one cluster to another, which is not the case at hierarchical clustering.

There is an error function behind this reallocation of data objects. It proceeds, for a given initial k clusters, by allocating the remaining data to the nearest clusters and then repeatedly changing the membership of the clusters according to the error function until the error function does not change significantly or the membership of the clusters no longer changes. The conventional k-means algorithm [63, 64] is briefly described below.

Let D be a data set with n instances, and let $C_1, C_2, \ldots, C_k$ be the k disjoint clusters of D. Then the error function is defined as

$$E = \sum_{i=1}^{k} \sum_{x \in Ci} d\big(\boldsymbol{x}, \mu(Ci)\big),$$

where μ(Ci) is the centroid of cluster Ci. d(**x**,μ(Ci)) denotes the distance between **x** and μ(Ci), and it can be one of the many distance measures, a typical choice of which is the Euclidean distance.

Given a set of observations, k-means clustering aims to partition n observations into k clusters so that the total distance between the group's members and its corresponding centroid, representative of the group, is minimized. The component to be minimized is the within-cluster sum of squares (WCSS):

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^j - c_j\right\|^2$$

where the term $\left\|x_i^j - c_j\right\|^2$ provides the distance between any data point and the cluster's centroid.

Each cluster is associated with a centroid, which is the mean of the points in the cluster. Each point is assigned to the cluster with the closest centroid. The first step of k-means is to select as initial cluster centers K, randomly selected documents, the *seeds* (initialization phase). The algorithm then moves the cluster centers around in space in order to minimize WCSS (iteration phase). This is accomplished iteratively by repeating the following steps until a stopping criterion is met: reassigning documents to the cluster with the closest centroid; and re-computing each centroid based on the current members of its cluster. Firstly, WCSS decreases in the reassignment step, since each vector is assigned to the closest centroid, so the distance it contributes to WCSS decreases. Secondly, it decreases in the re-computation step because the new centroid is the vector $\vec{v}$ for which $WCSS_k$ reaches its minimum. Ultimately, k-means converges for the common similarity measures to a local minimum point after a finite number of iterations (normally the first few) [65]. Convergence and some probability properties regarding the k-means algorithm are also discussed in Pollard [66, 67], and Serinko & Babu, [68]. García-Escudero and Gordaliza [69] discussed the robustness properties of the k-means algorithm.

The complexity of the whole procedure is summarized in the following expression:

O(n*K*I*d)

Where: n= number of points

K= number of clusters

I= number of iterations

d= number of attributes

Choosing the right initial number of centroids is very important as it controls the performance of the algorithm. If there are K 'real' clusters (especially when K is large), then the probability of selecting one centroid from each cluster is relatively small. Particularly, if clusters are of the same size, n, then the aforementioned probability is as follows:

$$P = \frac{number\ of\ ways\ to\ select\ one\ centroid\ from\ each\ cluster}{number\ of\ ways\ to\ select\ K\ centroids} = \frac{K!\,n^K}{(Kn)^K} = \frac{K!}{K^K}$$

There are several approaches to this problem such as multiple runs, sampling and usage of hierarchical clustering to determine the initial centroid number, selection of more than k initial centroids and re-selection among these (the most widely separated), postprocessing and/or bisecting k-means. Some methods for selecting good initial centers are proposed in Babu and Murty [70] and Bradley and Fayyad [71]. Pena et al. [72] provide a comparison of four initialization methods: a random method, Forgy's approach [56], Macqueen's approach [62], and Kaufman's approach [73]. Other initialization methods are presented in Khan and Ahmad [74].

Silhouette analysis is a method for selecting the number of clusters for k-means clustering. It can be used as a tool to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster. In other words, a value of +1 is ideal and -1 is the least preferred. Hence, the higher the value, the better is the cluster configuration.

The silhouette value for the ith point, Si, is defined as

$Si = (b_i - a_i) / \max(a_i, b_i)$

where a_i is the average distance from the ith point to the other points in the same cluster as i, and bi is the minimum average distance from the ith point to points in a different cluster, minimized over clusters.

A disadvantage of k-means algorithm is that it is sensitive to the presence of outliers and when clusters are of different size, different density or non-globular it might be disfunctional. For this reason, pretreatment and postprocessing of data is essential when implementing k-means, especially on high-dimensional data. Also, working only on numerical data restricts some applications of the k-means algorithm.

All in all, k-means is a greedy, computationally efficient technique, being the most popular representative-based clustering algorithm.

3.1.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) constitutes a multivariate statistical technique, probably one of the most popular in the chemometric literature, used by various scientific disciplines, in order to identify patterns and relationships within a data set [75, 76]. It is an unsupervised learning method which aims to reduce the dimensionality of a high-dimensional data set consisting of a large number of interrelated variables and at the same time to retain as much as possible of the variation present in the data set. In mathematical terms, this is accomplished by manipulating a data matrix in such a way that the variation or spread of data objects (i.e. the description of their interpoint distances) is described by as few dimensions as possible. In addition to data reduction, Principal Component Analysis forms a transformation technique

of data, also used for simplification, modelling, outlier detection (identification of their class membership), variable selection, classification, prediction and unmixing of constant sum mixtures (curve resolution) [77, 78, 79].

The information that PCA extracts from the mathematical manipulation of the data matrix, is expressed by a new orthogonal set of variables (PC axes), known as the Principal Components (PCs) [76]. These are new variables that are uncorrelated and ordered such that the first few retain most of the variation present in all of the original variables. Principal components are obtained as linear combinations of the original variables and each one of them is characterized by certain properties. For example, the first PC contains the maximum amount of possible variance in the data set, in one direction and successive PCs describe decreasing amounts of variation. Each data object has coordinates, defined by the original variables, which are relative to the new principal component axes (scores). What is more, PC axes are influenced by variables and this is because the formulation of each axis is based on combinations among the original measurement variables. Variables' contribution to PC axes depends mainly on the relative orientation between those two elements. Hence, parallel arrangement (in space) of the variable and PC axes, means that minimum variation is contained in the PC and accordingly, orthogonal arrangement of the two, means maximum variation. Finally, the maximum PC quantity to be calculated, is at the same time, the minimum quantity of data objects or variables (six habits).

The PCs are defined as follows. Let $\mathbf{v} = (v_1, v_2, \ldots, v_d)'$ be a vector of d random variables, where ′ is the transpose operation. The first step is to find a linear function $\mathbf{a}'_{1\mathbf{v}}$ of the elements of $\mathbf{v}$ that maximizes the variance, where $\mathbf{a}_1$ is a d-dimensional vector $(a_{11}, a_{12}, \ldots, a_{1d})'$, so

$$a'_1 v = \sum_{i=1}^{d} a_{1i} u_i$$

After finding $\mathbf{a}'_1\mathbf{v}, \mathbf{a}'_2\mathbf{v}, \ldots, \mathbf{a}'_{j-1}\mathbf{v}$, we look for a linear function $\mathbf{a}'_j\,\mathbf{v}$ that is uncorrelated with $\mathbf{a}'_1\mathbf{v}, \mathbf{a}'_2\,\mathbf{v}, \ldots, \mathbf{a}'_{j-1}\mathbf{v}$ and has maximum variance. Then we will find d such linear functions after d steps. The j th derived variable $\mathbf{a}'j\,\mathbf{v}$ is the j th PC. In general, most of the variation in $\mathbf{v}$ will be accounted for by the first few PCs. To find the form of the PCs, we need to know the covariance matrix ∑ of $\mathbf{v}$. In most realistic cases, the covariance matrix ∑ is unknown, and it will be replaced by a sample covariance matrix. For j = 1, 2, . . . , d, it can be shown that the jth PC is given by $z_j = \mathbf{a}'_j\,\mathbf{v}$, where $\mathbf{a}_j$ is an eigenvector of ∑ corresponding to the jth largest eigenvalue λ_j .

4. Family Affiliations of Williston Basin Oils

The sample set under study consists of four compositional families, A, B, C and D, each containing 44, 11, 38 and 27 oil samples, respectively (a total of 120 oil samples – see Appendix). Family A oil samples belong to Red River and Yeoman formations. Family B oil samples belong to Bakken and Lodgepole formations while samples of family D belong to Winnipegosis formations. Oil samples of family C belong to various formations, such as Midale, Tilston, Bakken, Frobisher, Ratcliffe, Lodgepole, and Madison formations. The exploration of the compositional data was conducted on the main hydrocarbons of the gasoline range, the n-alkanes in the saturated fraction of the oils, as well as the biomarker's content of this sample set.

As far as the gasoline range is concerned, it represents the number of hydrocarbons containing less than twelve carbon atoms, and are often referred to as light hydrocarbons. In highly thermally mature oils, this range constitutes almost the 100% of the oil composition and therefore geochemical characterization of such oils is carried out based on these compounds.

The saturated fraction of hydrocarbons (SFH) is comprised of either the linear, branched or cyclic hydrocarbons. SFH contains the structural group of n-alkanes (usually between C_{12}-C_{35}) as well as the pristane (Pr) and phytane (Ph) isoprenoid compounds, measured in geochemical studies along with n-alkanes, due to their geochemical significance. In the analysis, the lighter n-alkanes were excluded and only the C_{13}-C_{32} alkanes were considered.

Biomarkers are a group of compounds, found in oils and rock extracts. They have a variety of applications in petroleum exploration. Such applications are in source-rock correlation and/or in the inference of characteristics of the source rock that generated an oil, without examining the source rock itself. Specifically, biomarkers in an oil can reveal the relative amount of oil-prone vs. gas-prone organic matter in the source kerogen, the age of the source rock, the environment of deposition, the lithology of the source rock (carbonate vs. shale), and the thermal maturity of the source rock during generation. Such data may be key inputs to effective basin modelling of a prospect or block. In this study, the sterane and hopane parts of the biomarkers' range have been examined thoroughly.

Before performing Multivariate Data Analysis (MDA) on the given oil sample set, an attempt was made in order to test the criteria under which the classification of the four family affiliations of Williston Basin, was determined in previous studies. The biomarker based classification of the four families relies on various compositional criteria, including Pr/Ph ratio, tricyclic to pentacyclic C_{23}/C_{30} ratio, n-alkane predominance and prominence amongst extended hopanes and many other, extensively described in the following paragraphs. Empty spaces on the barcharts presented below, correspond either to zero component values for specific samples, or to infinite numbers, generated during the calculative ratio calculations.

The compositional character of each family is unique and this is evident from their n-alkane distributions, biomarker signature as well as their gasoline range characteristics, in general [39, 32]. Family A oils display diagnostic saturate fraction gas chromatograms (SFGC) and are fairly distinguishable from the other families by their overall n-alkane profile (centered at C_{13}-C_{17}) and CPI values (average CPI: 1.59) [39, 32]. According to Obermajer et al., (2000), they

also present a smooth extended hopane profile with a steady decrease in the concentration of C_{31}+ homologues with increasing carbon numbers [39]. In addition, Family A oils, are characterized by a C_{34} hopane prominence, according to Osadetz et al., [32]. Homohopane distributions have been used to distinguish oils from different organic facies of the same source rock. Such distributions are sensitive and may be altered due to various factors such as thermal maturity and API gravity. Judging from the barchart (Fig. 7), Family A oils display a high C_{34} homohopane distribution, but, in addition, oils from Family D, present an even stronger prominence on this compound (Fig. 7). The behavior of C_{34} for Families B and C is similar to that of Family A.

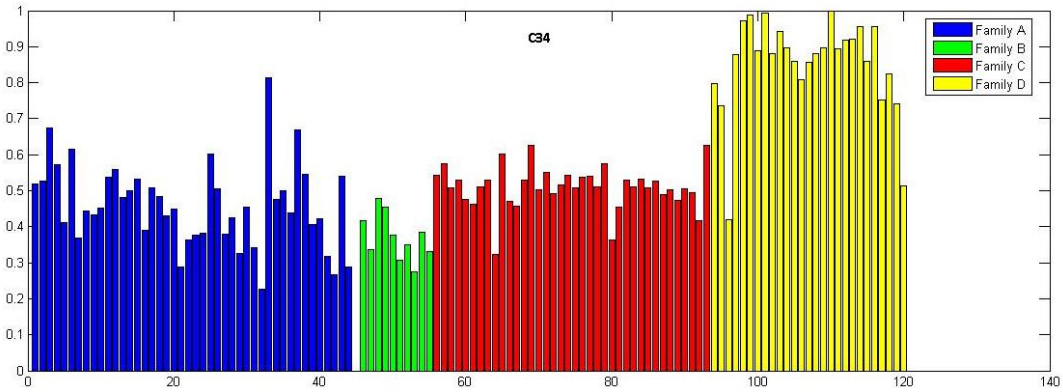


Fig. 7 C_{34} barchart for the whole sample set.

Another diagnostic feature of this group is its very low concentration of acyclic isoprenoids relative to n-alkanes, presenting the lowest Pr/C_{17} and Ph/nC_{18} ratios among all families [39]. The corresponding barcharts (Fig. 14, Fig. 15), in which these ratios have been plotted, is in agreement with this fact. According to Osadetz et al., [32], the C_{23} tricyclic/C_{30} pentacyclic terpanes ratio, especially for Families A and B, is very distinct, differentiating them from the rest family groups. From the corresponding barchart, it is indeed observed that Family A oils display very low values of C_{23}/C_{30}, whereas Family B displays the highest peaks for the same ratio (Fig. 8). What is also noticeable from the C_{23}/C_{30} barchart, is that Family D oils, similarly to Family A, present very low values for this ratio.

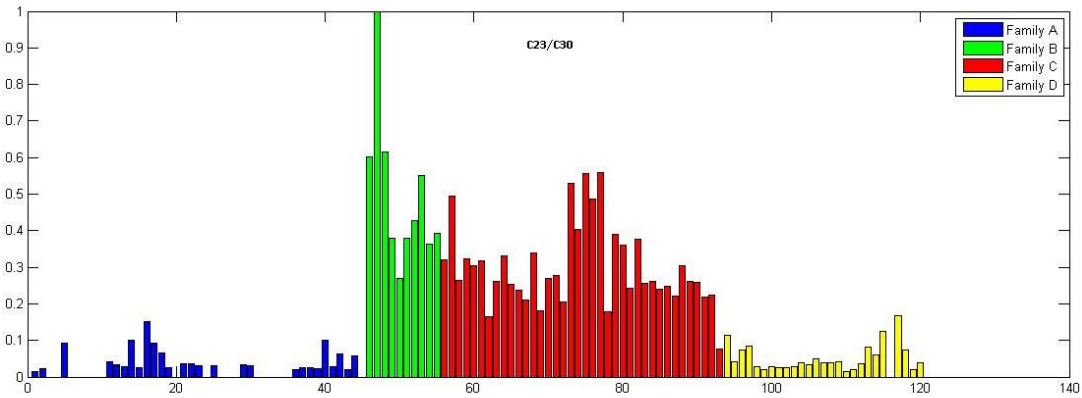


Fig. 8 Barchart presenting C_{23}/C_{30} ratios for the whole sample set.

Family B oils, according to Obermajer et al., [39], differ from the rest in that they present a smooth n-alkane distribution with a maximum in the C_{13}-C_{17} range, lacking any homohopane prominence, which is in agreement with Osadetz et al., [32]. According to Obermajer et al., [39], there are variations in 17a(H)-trisnorhopane (Tm) over 18a(H)-trisnorhopane (Ts), compared to the rest oils. From the respective barchart (Fig. 10) we observe that there are indeed, intense variations within this Family affiliation, the density of the specimens, however, is not adequate enough in order to confirm the clear distinction of this family from the rest. The calculative process of the code has produced the NaN notation, resulting in non-plotted samples. The Ts/Tm ratio profiles of the rest Families (A,C and D) show almost equivalent variations.

Another characteristic of Family B oils, is that they obtain values above unity for the Pr/Ph ratio [32]. This ratio is one of the most common correlation parameters, utilized as an indicator of depositional environment [80]. Variations may reflect multiple degrees of oxidation during the early stages of chlorophyll degradation. It is one of the most commonly utilized correlation parameters, indicative of the source rock's depositional environment [80]. Being sensitive to diagenetic conditions, values of Pr/Ph ratios substantially below unity are considered to indicative of petroleum origin and/or highly reducing depositional environments. Very high Pr/Ph ratios (> 3) reflect source material of terrestrial origin. Pr/Ph ratios ranging between 1-3 reflect oxidizing depositional environments [81]. According to Lijmbach [82] low Pr/Ph values (<2) reflect aquatic depositional environments including marine, fresh and brackish water (reducing conditions), intermediate values (2–4) reflect fluviomarine and coastal swamp environments, whereas very high values (up to 10) are related to peat swamp depositional environments (oxidizing conditions). From the corresponding barchart (Fig. 9), we observe that, contrary to Family C, Families A, B and D present similar, above unity values for this ratio, which is in agreement with Osadetz et al., [32]. At the same time, however, Family B oils display the highest peaks (Fig. 9).

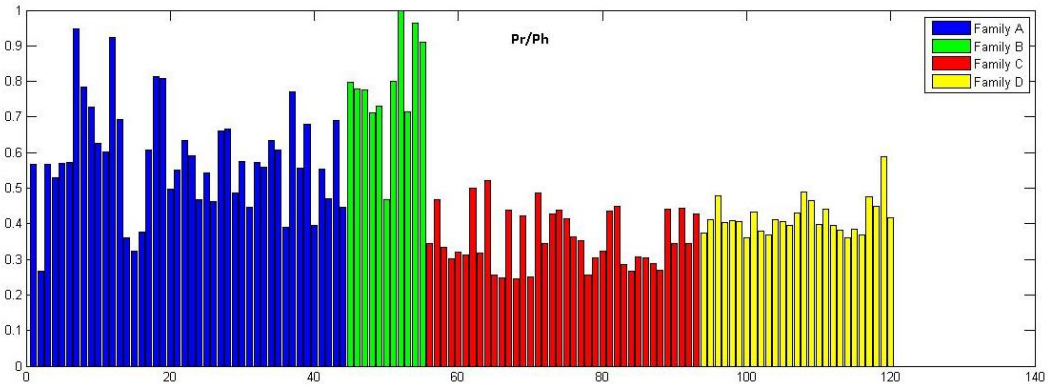


Fig. 9 Pr/Ph ratios barchart for the whole sample set.

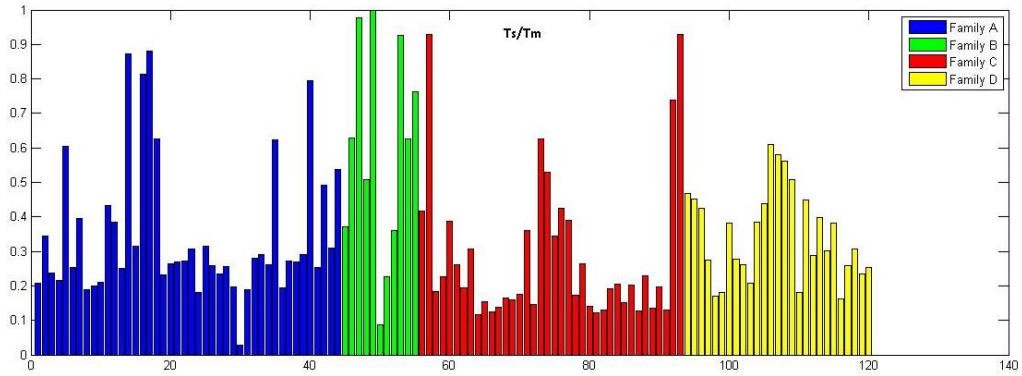


Fig. 10 Ts/Tm ratios barchart for the whole sample set.

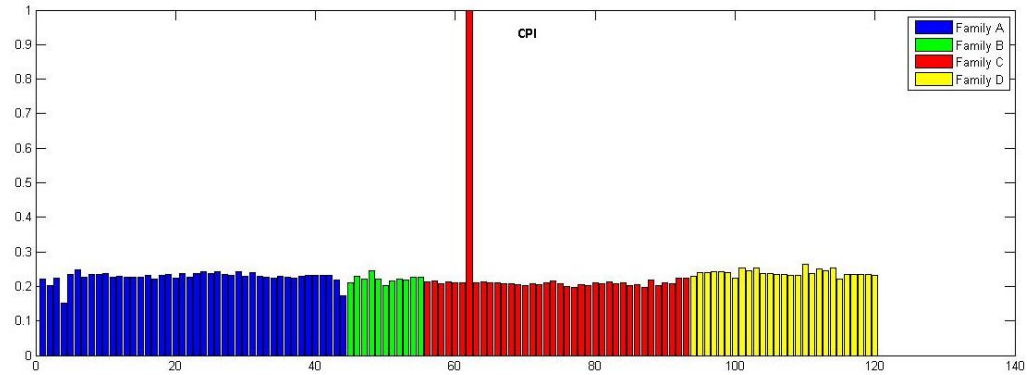


Fig. 11 CPI profile for the whole sample set.

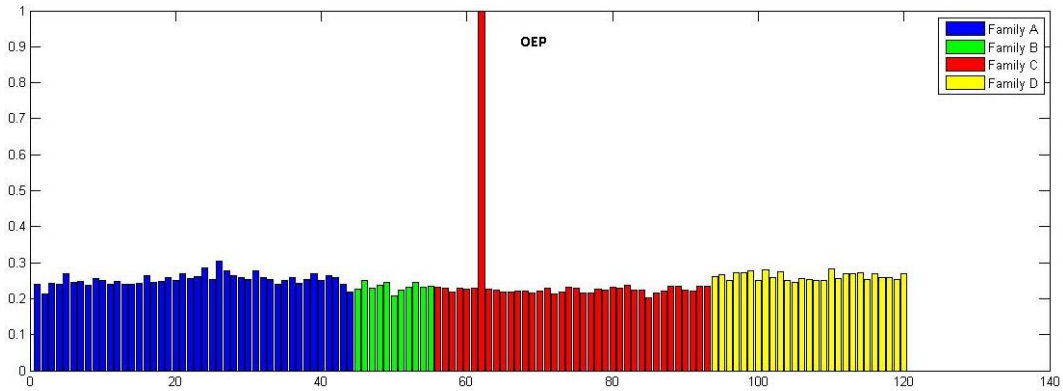


Fig. 12 Odd/Even predominance for the whole sample set.

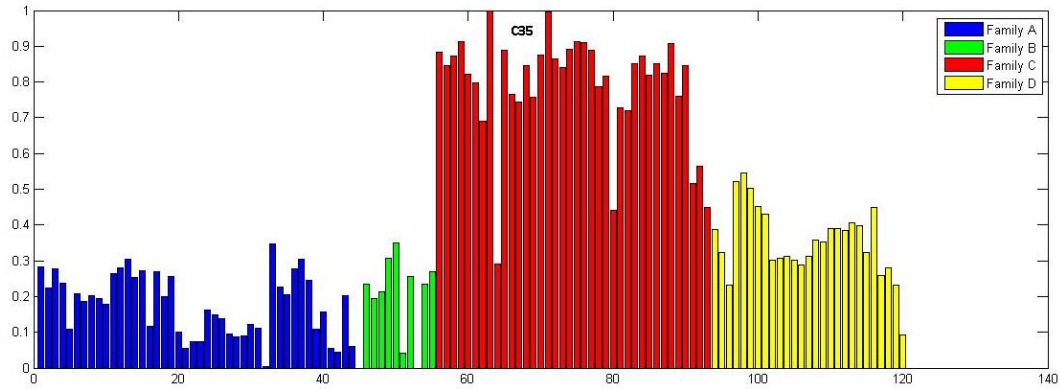


Fig. 13 C_{35} barchart for the whole sample set.

Main characteristic of Family C is the strong C_{35} prominence [32], which is confirmed by the respective barchart (Fig. 13). The lowest C_{35} homohopane distribution is indicative of Family A oils, as shown. According to Osadetz et al., [32] and Obermajer et al., [39], these oils obtain lower Pr/Ph values in comparison to the rest, and in particular, less than unity. This fact holds true, as we observe from the corresponding barchart (Fig. 9), confirming at the same time that Family C oils display a strong and consistent predominance of Ph/Pr ratio.

Additionally, oils of this familial group also display an even/odd n-alkane predominance [32]. The composition and distribution of n-alkanes carbon numbers reflect the source of kerogenic organic matter, sedimentary environment, and maturity of the rocks. Traditional geochemists feel that the odd/even carbon number predominance of n-alkane decreases as rocks mature. The OEP (odd/even predominance) of mature source rocks is close to 1. However, the odd carbon number predominance appears in Upper Ordovician source rocks, and an even carbon number predominance is found in Cambrian - Lower Ordovician source rocks. Family C oils are characterized by an even/odd n-alkanes predominance and this is confirmed by both the CPI and OEP, respective barcharts (Fig. 11, Fig. 12).

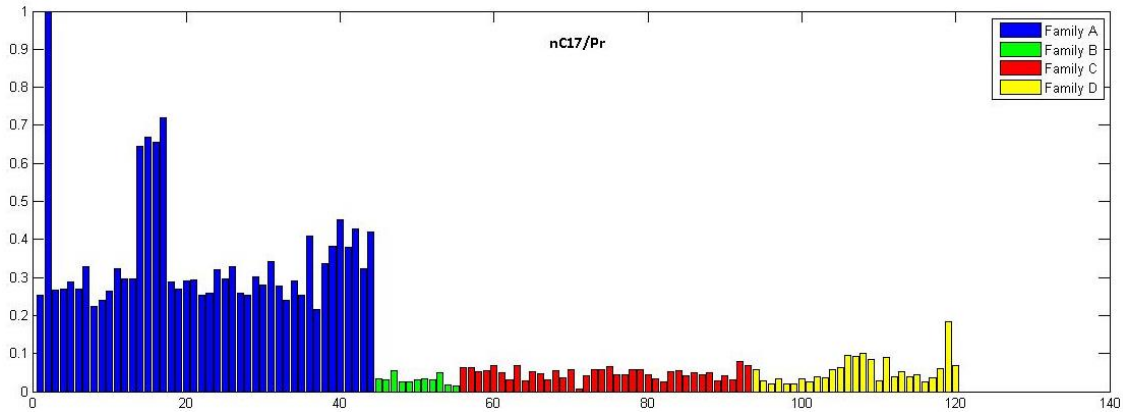


Fig. 14 nC_{17}/Pr barchart for the whole sample set.

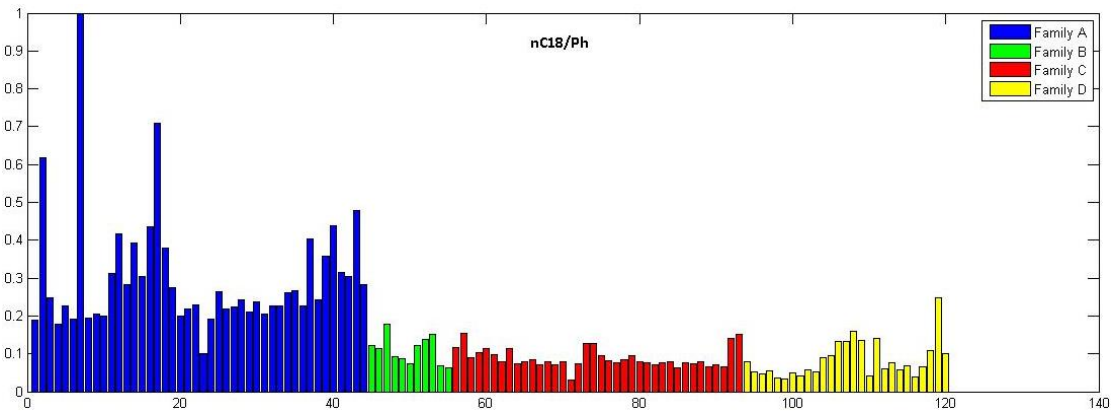


Fig. 15 nC_{18}/Ph barchart for the whole sample set.

Oils from Family D display a distinctive stratigraphic occurrence and have been subdivided into two separate groups D_1 and D_2, based on nC_{17}/Pr and C_{18}/Ph ratios [32]. The corresponding barcharts (Fig. 14, Fig. 15) present the distributions of these ratios amongst the whole sample set. What is more, D_1 and D_2 oils, depending on the pools they occur, either in Madison or Birdbear, they display Pr/Ph ≤ 1.0 and Pr/Ph> 1.1, respectively. This is indeed, evident, from the corresponding barchart (Fig. 9).

Reviewing the barcharts presented before, it would be important to state that based on individual geochemical characteristics, the four families can be indeed uniquely identified at a great extent. However, it would be a challenge to investigate if a clear classification can be obtained, by applying this time, multivariate data analysis (MDA) on raw data.

In the next chapters, we implement several multivariate methods on the given data set and examine the results, that each method produces. Hierarchical clustering, k-means and Principal Component Analysis are applied on four independent models that were developed for this purpose; the Saturated Fraction Compositional Model, the Saturated Fraction Ratios Model, the Gasoline Range Compositional Model and the Biomarkers Compositional Model. All of the steps that were followed are extensively described.

5. Application of MDA methods; inputs and results

The core of this project is the investigation of the oil-oil correlations among compositional data of a sample set from Williston Basin, by using multivariate statistical analysis methods. Oil-oil correlations are based on compositional criteria and examine whether a genetic relationship exists among a group of oil samples. In particular, Hierarchical Clustering, k-means and PCA have been employed in order to explore compositional data from the gasoline range (GRH), saturated hydrocarbons (SFH) and biomarker traits of 120 oil samples from the Williston Basin Petroleum province. The samples examined in this study are from four, previously defined, compositional families (A-D) [34].

For the application of MDA methods on the sample set, a MATLAB code created in the "Hydrocarbons Chemistry and Technology Research Unit", of the School of Mineral Resources of the TUC, was utilized. All necessary adjustments and modifications were applied in order for the code to work.

From the sample set under study, four independent models were developed in order to explore different compositional information. The models used for the identification of petroleum systems were: a) Saturated Fraction Compositional Model (SFCM) b) Saturated Fraction Ratios Model (SFRM), c) Gasoline Range Compositional Model (GRCM) and d) Biomarkers Compositional Model (BCM). SFCM embodies original variables derived from the gas chromatographic analysis of the Saturated Fraction of Hydrocarbons (SFH). It takes into account peak areas of n-alkanes, nC_{13}-nC_{24}, pristane (Pr) and phytane (Ph). The SFRM contains the most commonly utilized compositional ratios and factors derived from the gas chromatographic analysis of the SFH (Pr/Ph, n-C_{17}/Pr, n-C_{18}/Ph, CPI n-C_{14-20}, CPI n-C_{22-32}). GRCM includes variables derived solely from GRH compositional data. The parameters reflect internal variations for compounds with the same number of carbon atoms to minimize possible variations due to sample handling and experimental conditions. Finally, BCM contains all variables derived from biomarkers' traits of the oil sample set.

The approach under which all statistical methods were applied, was that of trial and error in order to achieve a "clear clustering" (if possible) of the oil samples. A pretreatment scheme of the sample set was considered necessary in order to reformat the original data file and prepare data for clustering. This is because the data set consists of peak areas that are analysis-dependent. As a consequence, only by preprocessing the data, we get meaningful statistical results, since all components are put under the same scale. The idea is that if different components of data (features) have different scales, then derivatives tend to align along directions with higher variance, which leads to poorer/slower convergence. The chemometric software package that was utilized, offered various pretreatment options, all of which were originally applied on the sample set, in order to examine which one produces the best classifying solution. While only the results from one preprocessing option will be presented, all pretreatment schemes which were utilized, are briefly described below.

Command "pre_scaling_0_1": It refers to the subtraction of the minimum value and the division of each column by the range. The results of this pretreatment scheme are going to be presented in the upcoming chapters.

Command “norm_variables_0_1”: It refers to the subtraction of the minimum value and the division of each variable by the range.

Command “pre_minusMean”: It concerns the subtraction of the mean value from each variable.

Command “pre_PQN” (Probabilistic quotient normalization): It refers to the division of each sample with the sum of the sample’s variables. The calculative process takes into consideration the median value of each column.

Command “pre_CLR” (Centered log-ratio normalization): It concerns the division of each sample with the sum of the sample’s variables. It differs, however, from “pre_PQN” in that it takes into consideration the geometric mean of each column.

Command “Subtract_sample_min”: It refers to the subtraction from each sample of its minimum value.

Command “pre_TSN” (total sum normalization): It concerns the division of each sample with the sum of the samples’ variables.

Command “pre_max”: This matlab command refers to the division of each sample with the maximum value of the samples’ variables.

5.1 Saturated Fraction Compositional Model (SFCM)

5.1.1 Hierarchical Clustering on SFCM

The subtraction of the minimum value from the subset and division of each variable by the range (“pre_scaling_0_1” command) resulted in the following dendrogram (Fig. 16). Average linkage with a correlation coefficient were combined

It is evident that the oil samples from all four family affiliations overlap, presenting no clear distinction. In particular, there is a slight overlap of samples from Families B (B1014, B1993, B2121, B2179, B1879, B1874) and D (D1275, D1276, D1289, D1313, D1288, D1290, D1291) with Family A. The original clustering solution detected outlier values (samples C599, D2595 and C566), removing which from the sample set and reprocessing it under the same pretreatment, made no difference on the clustering solution.

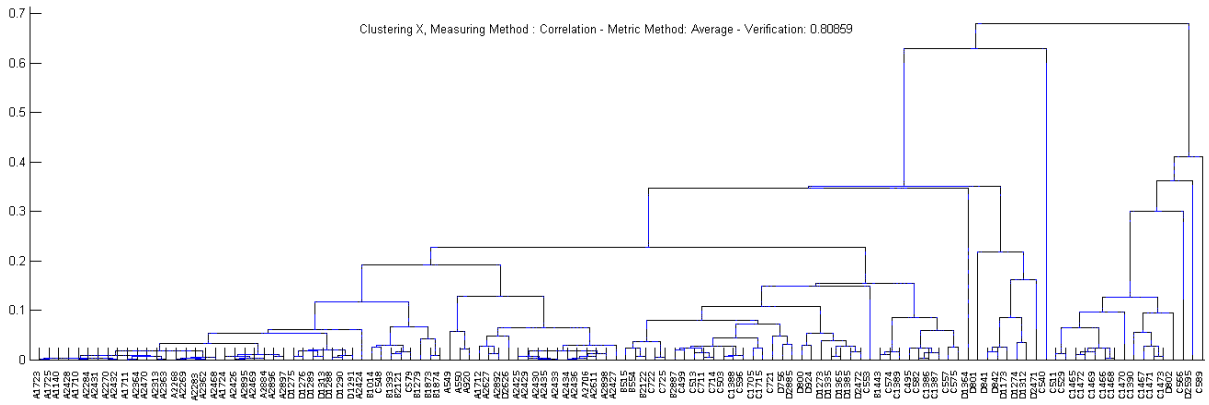


Fig. 16 Resulting Dendrogram under the command "pre_scaling_0_1" for the Saturated fraction compositional model (SFCM).

As observed, the algorithm failed to discriminate distinct familial affiliations among the given oil sample set, under this pretreatment scheme. In order to test how Hierarchical Clustering would offer the best clustering solution, many other pretreatment schemes were also applied on the data set and as a procedure, this was also followed in the upcoming MDA methods. The following dendrogram resulted from the division of each sample with the sum of the samples' variable - Total Sum Normalization ("pre_TSN" command of the chemometric software package). City block distance and Centroid linkage were combined and the produced dendrogram displays a relatively good distinction of Family A. It fails, however, to distinguish amongst Families B, C and D, which, once again, overlap one another (Fig. 17).

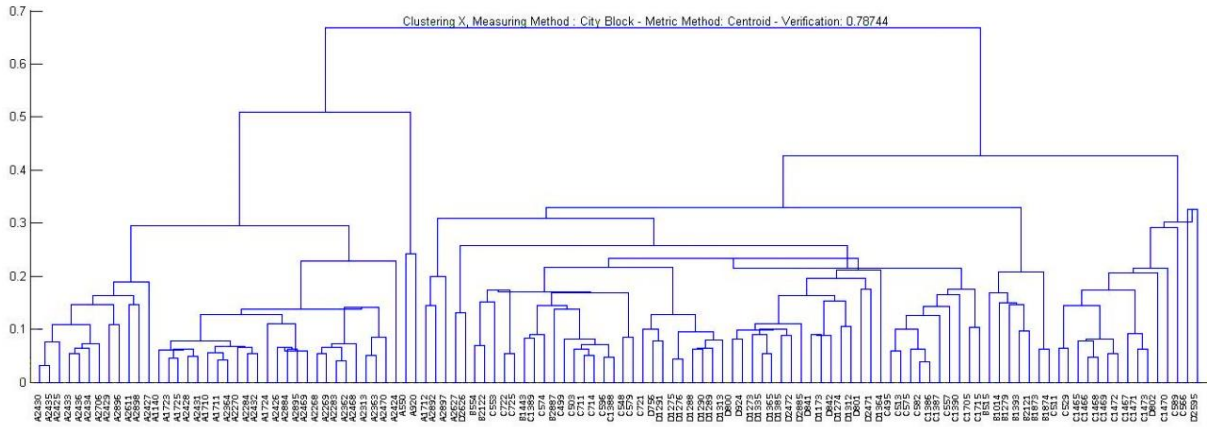


Fig. 17 Dendrogram under the "pre_TSN" command for the Saturated fraction compositional model (SFCM).

An interesting feature of the dendrogram in Fig. 17, is that it displays a non-monotonic tree. This occurs when the distance from the union of two clusters, r and s, to a third cluster is less than the distance between r and s. In this case, in a dendrogram drawn with the default orientation, the path from a leaf to the root node takes some downward steps. Usually, the centroid and median methods (as in this case) can produce a cluster tree that is not monotonic and if this happens, it is better to utilize another linkage method. In our case, however (Fig. 17), the centroid linkage, which was automatically chosen by the chemometric software package, produced a dendrogram which classified sufficiently samples of Family A. All other pretreatment options failed in this task significantly.

5.1.2 k – means algorithm on SFCM

k-means clustering was then performed under the same pretreatment option ("pre_scaling_0_1") resulting in the following features (Fig. 18, Fig. 19, Table 3).

Table 3 Summary of k-means clustering under the " pre_scaling_0_1" pretreatment option for the Saturated fraction compositional model (SFCM).

K-values	Best distances sums	Average silhuette values
K=2	594.077	0,566689
K=3	41.787	0,539449
K=4	345.548	0,503212
K=5	290.114	0,467112

The silhouette plots for K=2, K=3, K=4 and K=5 clusters are shown in the following figure (Fig. 18). An insufficient choice of an initial K value would result in clusters below average silhouette scores or even wide fluctuations in the size of the silhouette plots. This is the criteria under which, each clustering solution is evaluated as sufficient or insufficient.

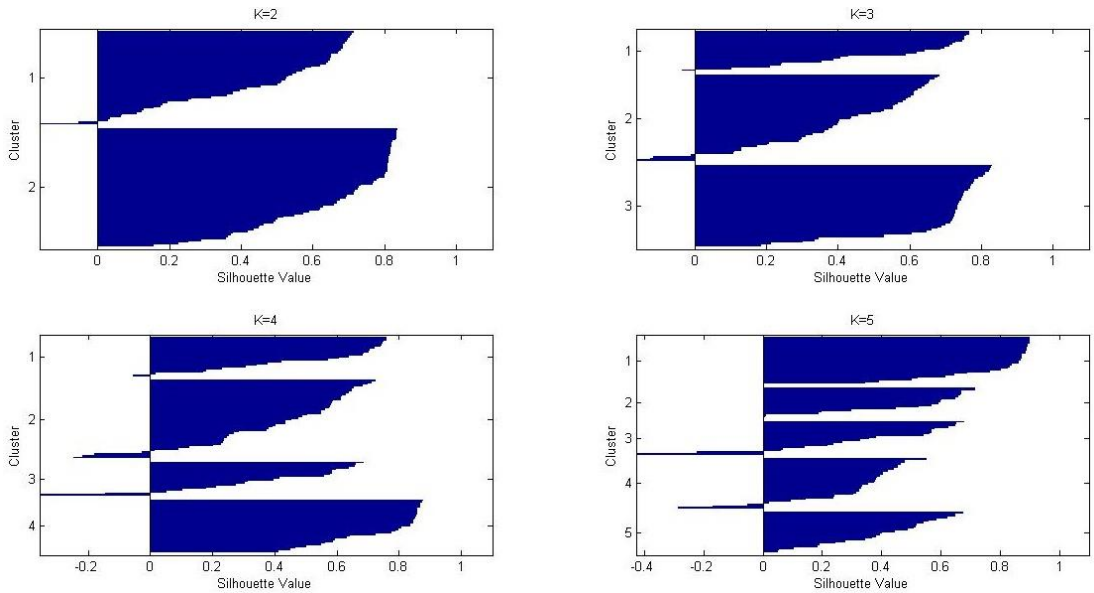


Fig. 18 Silhouette plots for k=2, k=3, k=4 and k=5 clusters under the " pre_scaling_0_1" pretreatment option for the Saturated fraction compositional model (SFCM).

From the silhouette plots (Fig. 18), we observe that, in general, the obtained silhouette values fall in the range of 0.1-0.9. The size of the silhouette plots does not present wide fluctuations for each case, and negative values are present in all clustering solutions. The two - cluster solution has an average silhouette value of 0.566689, being the highest amongst the others (Table 3). This is an indication that grouping into two clusters using k-means is more efficient compared to grouping into three, four or five clusters. It is not, however, sufficient enough, as we would expect, grouping into four clusters to be the best solution. In Fig. 20 we can observe, which cluster each sample is assigned to.

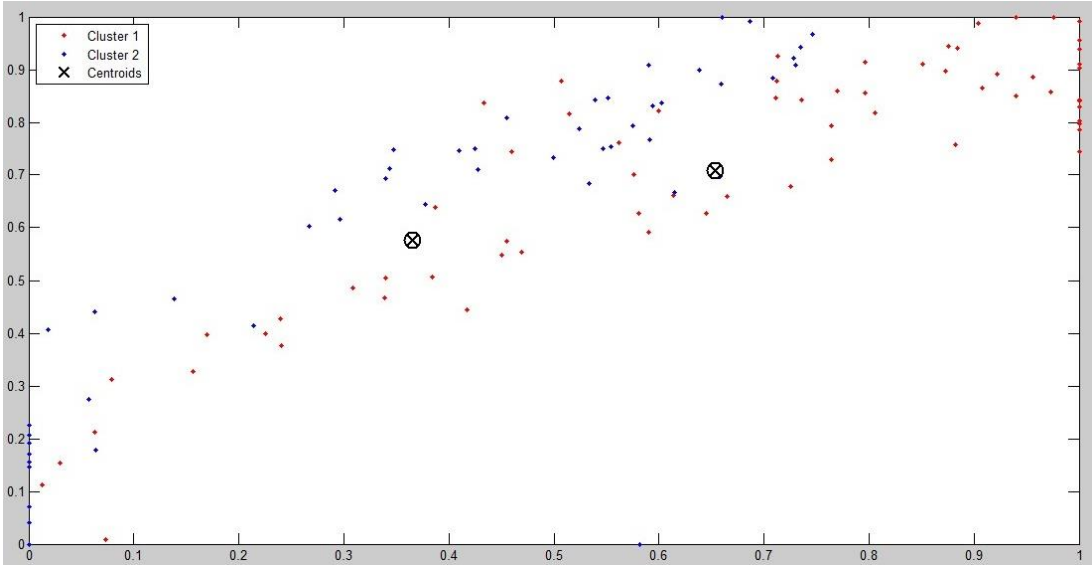


Fig. 19 The plot of k-means clustering for k=2 under the "pre_scaling_0_1" pretreatment option for the Saturated fraction compositional model (SFCM). The ⊗ symbol represents the centroid of each cluster.

Fig. 19 represents the plot of k-means clustering, for the case of k = 2. Taking into consideration the average silhouette value, k=2 is the most efficient clustering solution. However, by observing the plot, we could say that there is no clear boundary between the two clusters and samples overlap with each other.

		A549	A550	A920	A1140	A1710	A1711	A1712	A1723	A1724	A1725	A2268	A2269	A2270	A2283	A2284	A2313	A2362	A2363	A2364	A2424	A2425	A2426
prescale	idx2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	idx3	1	1	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	idx4	4	4	4	2	2	2	4	2	2	2	2	2	2	2	2	2	2	2	2	2	4	2
	idx5	5	5	5	1	1	1	5	1	1	1	1	1	1	1	1	1	1	1	1	1	5	1
		A2427	A2428	A2429	A2430	A2431	A2432	A2433	A2434	A2435	A2436	A2468	A2469	A2470	A2611	A2627	A2706	A2884	A2892	A2895	A2896	A2897	A2898
prescale	idx2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	idx3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	idx4	4	2	4	4	2	2	4	4	4	4	2	2	2	4	4	4	2	4	2	4	2	4
	idx5	5	1	5	5	1	1	5	5	5	5	1	1	1	5	5	5	1	5	1	5	1	5
		B515	B554	B1014	B1279	B1393	B1443	B2121	B2122	B2887	B1873	B1874	C495	C499	C503	C511	C513	C529	C540	C548	C553	C557	C566
prescale	idx2	2	1	2	2	2	1	2	1	1	2	2	1	1	1	1	1	1	2	2	1	1	1
	idx3	3	3	1	3	3	3	3	3	3	3	3	3	3	3	2	3	2	3	3	3	2	2
	idx4	1	1	2	1	1	1	1	1	1	1	1	1	1	1	3	1	3	4	1	1	3	3
	idx5	2	2	1	2	2	3	2	3	3	2	2	3	3	3	4	3	4	5	2	2	4	4
		C574	C575	C579	C582	C589	C596	C711	C714	C721	C722	C725	C1386	C1387	C1388	C1389	C1390	C1465	C1466	C1467	C1468	C1469	C1470
prescale	idx2	1	1	2	1	1	1	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1
	idx3	3	2	3	2	2	3	3	3	3	3	3	2	2	3	3	2	2	2	2	2	2	2
	idx4	1	3	1	3	3	1	1	1	1	1	1	3	3	1	1	3	3	3	3	3	3	3
	idx5	3	3	2	3	4	3	3	3	2	2	2	3	3	3	3	4	4	4	4	4	4	4
		C1471	C1472	C1473	C1705	C1715	D756	D800	D801	D802	D841	D842	D924	D1173	D1273	D1274	D1275	D1276	D1288	D1289	D1290	D1291	D1312
prescale	idx2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
	idx3	2	2	2	3	3	3	3	2	2	3	3	3	3	3	3	3	3	3	1	3	3	3
	idx4	3	3	3	1	1	1	1	3	3	1	1	1	1	1	1	1	1	2	2	2	1	1
	idx5	4	4	4	3	3	2	3	4	4	2	2	3	2	2	2	2	2	2	2	2	2	2
		D1313	D1335	D1364	D1365	D1385	D2471	D2472	D2595	D2626	D2885												
prescale	idx2	2	1	1	1	1	1	1	1	2	1												
	idx3	1	3	2	3	3	3	3	2	1	3												
	idx4	2	1	3	1	1	1	1	3	4	1												
	idx5	1	2	3	2	3	3	2	4	5	2												

Fig. 20 Table displaying to which cluster each sample belongs, for each K value of the SFCM (idx2 = k:2, idx3 = k:3, etc.)

Taking into consideration that the most sufficient clustering solution is that of k=2 (idx=2) and according to Fig. 20, all samples from Family A oils are assigned to one cluster. The vast majority of Family C oil samples are assigned to a different cluster with a few exemptions (C540, C543). The discretization of these two families is relatively sufficient, but as far as Family D and B oil samples are concerned, they overlap with A and C considerably, as samples from both families are assigned to both clusters.

5.1.3 Principal Component Analysis on SFCM

Sample scores describe a position in principal component space, and each original variable has loadings that describe their contribution to each principal component. The sample score of the first two principal components and the respective loading diagrams are presented in figure (Fig. 21a, b). The percentages of variation attributed to each of the Principal Components are shown in Fig. 21c.

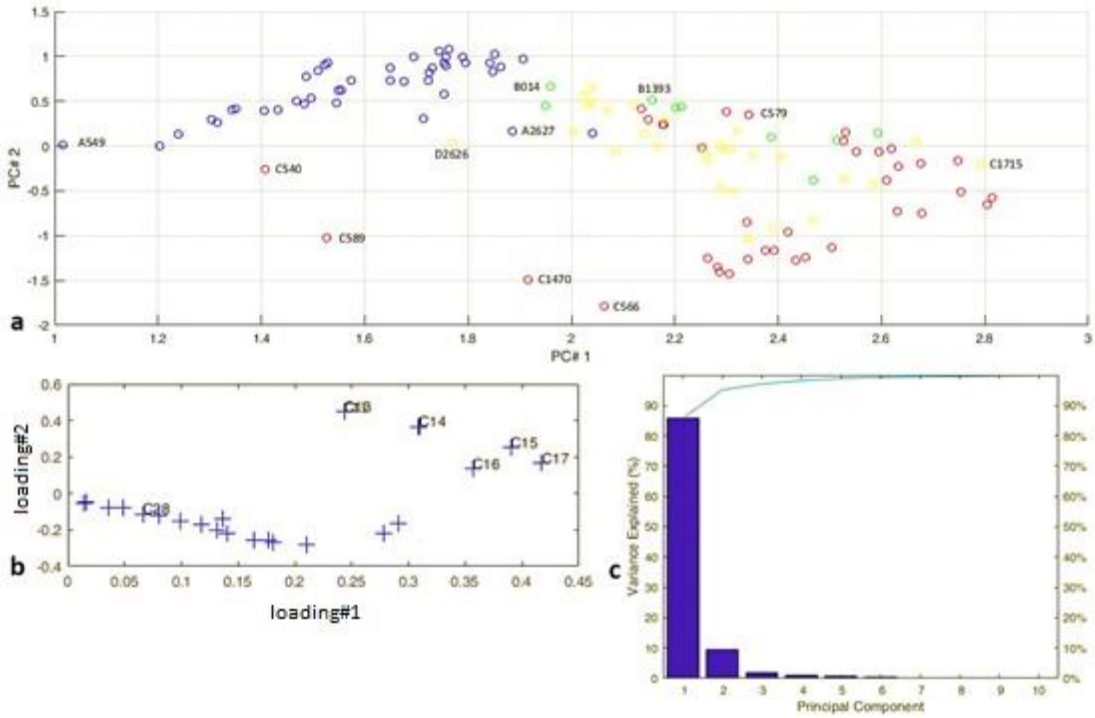


Fig. 21 a) Sample scores for the first to Principal Components resulting from the Saturated Fraction Compositional Model (SFCM) of selected Williston Basin petroleum oils. Colors on sample symbols indicate compositional families determined by independent analysis. Blue color represents oils of Family A, green applies for Family B oils, red for Family C oils and yellow for Family D oils. "Pre_scaling_0_1" command was used on the data set. b) Original Variable loadings for the first to Principal Components resulting from the Saturated Fraction Compositional Model (SFCM) of selected Williston Basin petroleum oils. c) Percentage of variance explained by each Principal Component.

SFCM sample scores of the first two PCs explain almost 85 per cent of the variance (Fig. 21c). There are linear gradients observed in the data by comparing the sample scores of the first two principal components of the SFCM. These gradients indicate that both distinctive family characteristics and linear compositional variations of the original variables exist within each family. Sample scores of Family A oils exhibit the most coherent grouping and are characterized by a positive gradient defined by positive PC1 scores and positive PC2 scores. Family C oils are also characterized by a positive gradient whereas Family B and D are defined by a negative gradient. The mild gradients of Families B, C and D exhibit positive PC1 scores, as Family A, but negative PC2 scores. There is a considerable overlap of Family B with Family D and a slight overlap of Family C with Family D. What is more, for a given value of PC1 Families C and D have more negative PC2 scores but this is not enough to be uniquely distinguishable.

As far as variable loadings are concerned, they are a tool used for the understanding of the role and importance of the original variables. The original variable loadings for the SFCM distinguish between a preponderance of lighter versus heavier n-alkanes (Fig. 21b). C_{13}-C_{17}

alkanes are characterized by strongly positive PC1 and PC2 loadings but C_{16} and C_{17} exhibit negative PC3 values. Probably all these variable loadings control the gradients that separates independently defined oil families. The variable with the higher weight (0.1034) among the 22 variables of the SFCM, is alkane C_{13} with strongly positive PC1, PC2 and PC3 loadings.

5.1.4 Discussion on the performance of MDA on the SFCM

To summarize, Hierarchical Clustering, k-means and Principal Component analysis were applied on the Saturated Fraction Component Model. Both in Hierarchical Clustering and PCA, Family A oils presented the most coherent group, being sufficiently separated from the rest familial affiliations. Families B and D overlapped significantly while also in both cases there appeared a slight overlap between families C and D. The method which completely failed to distinguish among the four oil families (A, B, C and D) was k – means clustering. The clustering solution produced only two clusters and according to which cluster each sample was assigned, k-means presents only 25% of success. Out of the three statistical methods, k-means was the one to produce the most insufficient results.

5.2 Saturated Fraction Ratios Model (SFRM)

5.2.1 Hierarchical Clustering on SFRM

The following dendrogram is the outcome of the “pre_scaling_0_1” command (Fig. 22). Average linkage along with Euclidean distance as a measure of proximity, were combined.

Family A is clearly distinguished from the rest. Family C considerably overlaps with Families B and D. All pretreatment schemes that were applied on the data set, behaved similarly producing almost the same results when Hierarchical Clustering was performed; all distinguished Family A quite sufficiently, but exhibited a slight overlap amongst Families B, C and D.

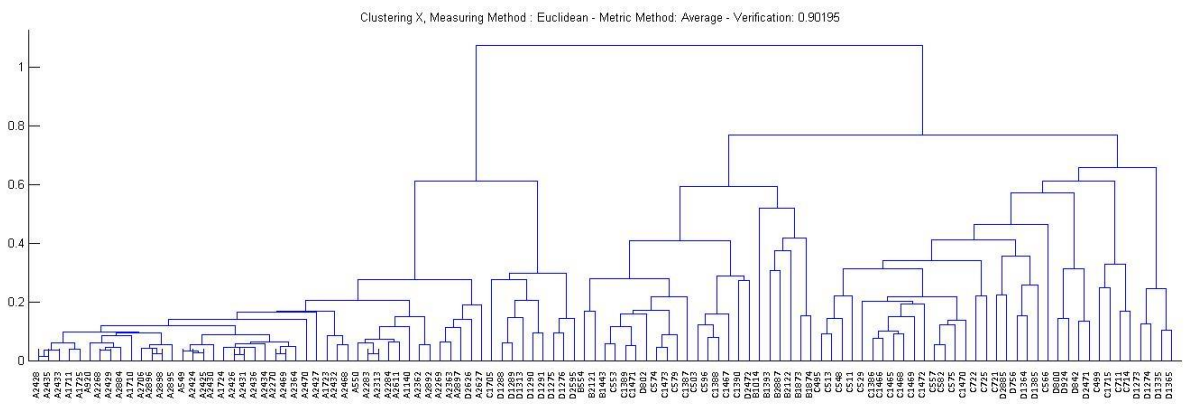


Fig. 22 Resulting Dendrogram under the command “pre_scaling_0_1” for the Saturated fraction ratios model (SFRM).

5.2.2 k – means algorithm on SFRM

Under the same pretreatment scheme ("pre_scaling_0_1" command), k-means algorithm was applied and below we present the results.

Table 4 Summary of k-means clustering under the "pre_scaling_0_1" pretreatment option on the Saturated Fraction Ratios Model (SFRM).

K-values	Best distances sums	Average silhuette values
K=2	184.117	0,727318
K=3	115.198	0,691804
K=4	832.004|831.205	0,715499
K=5	649.877	0,702406

The silhouette plots for K=2, K=3, K=4 and K=5 clusters are presented in the following figure (Fig. 23).

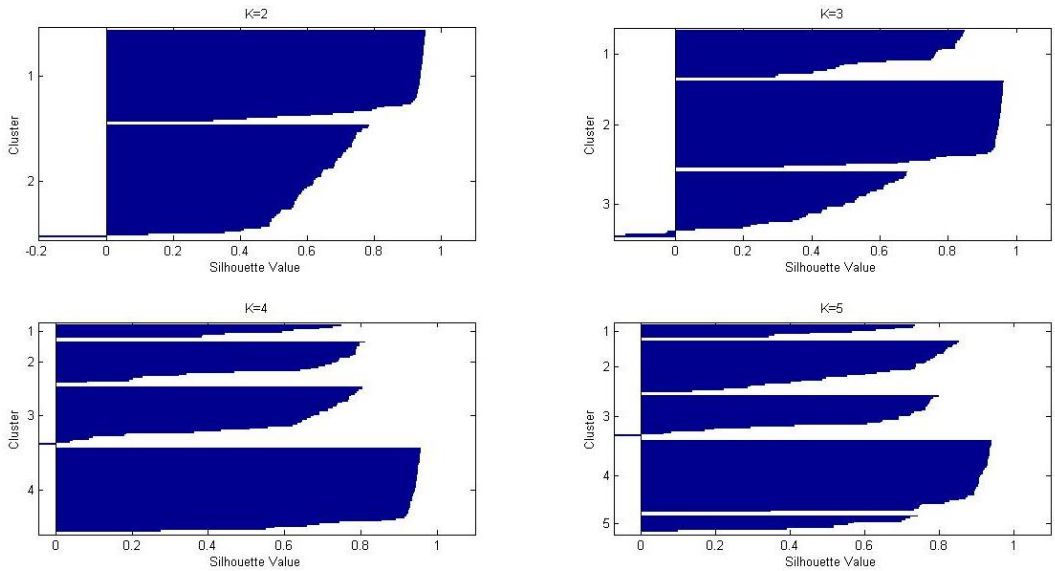


Fig. 23 Silhouette plots for k=2, k=3, k=4 and k=5 clusters under the " pre_scaling_0_1" pretreatment option for the Saturated Fraction Ratios Model (SFRM).

From the silhouette plots (Fig. 23), we observe that in all cases we obtain silhouette values above 0.6 and negative silhouette coefficients are always present. Average silhouette values are similar for all clustering solutions, with a maximum of 0,727318 for K=2 (Table 4). This is an indication that under the "pre_scaling_0_1" pretreatment scheme, grouping into two clusters using k-means is more efficient compared to grouping into three, four or five clusters. In Fig. 25 we can observe, which cluster each sample is assigned to.

In Fig. 24 the plot of k-means clustering, for the case of k = 2 is presented with different colors for sample members that belong to different clusters.

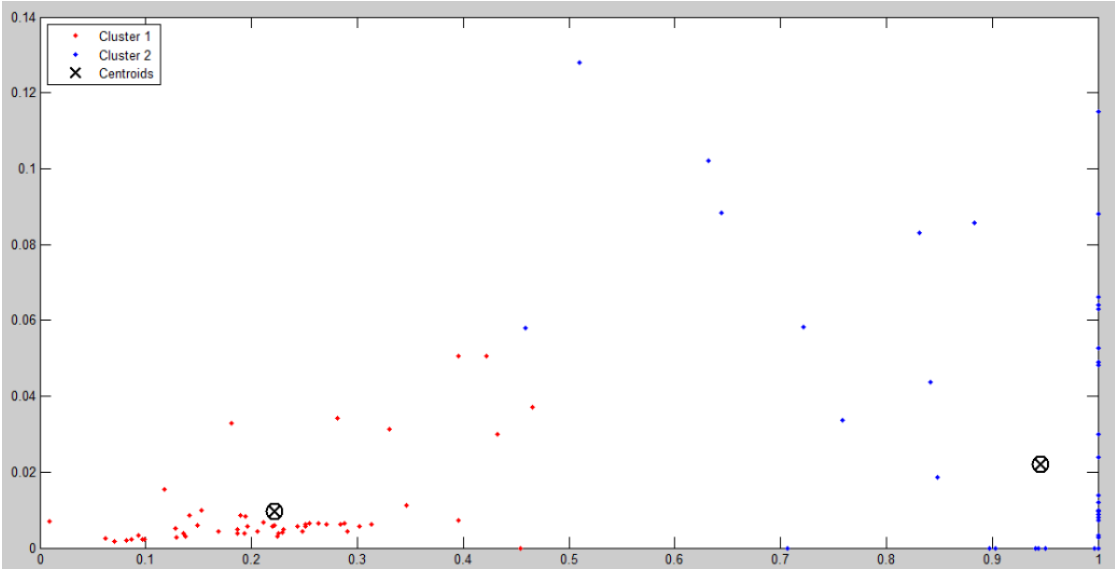


Fig. 24 The plot of k-means clustering for k=2, of the Saturated Fraction Ratios Model (SFRM). The ⊗ symbol represents the centroid of each cluster.

		A549	A550	A920	A1140	A1710	A1711	A1712	A1723	A1724	A1725	A2268	A2269	A2270	A2283	A2284	A2313	A2362	A2363	A2364	A2424
pre_scaling_0_1	idx2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	idx3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	idx4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
	idx5	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5

		A2425	A2426	A2427	A2428	A2429	A2430	A2431	A2432	A2433	A2434	A2435	A2436	A2468	A2469	A2470	A2611	A2627	A2706	A2884	A2892
pre_scaling_0_1	idx2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	idx3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	idx4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
	idx5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

		A2895	A2896	A2897	A2898	B554	B1014	B1279	B1393	B1443	B2121	B2122	B2887	B1873	B1874	C495	C499	C503	C511	C513	C529
pre_scaling_0_1	idx2	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	idx3	2	2	2	2	1	1	1	1	1	1	1	1	1	1	3	3	1	3	3	3
	idx4	3	3	3	3	4	4	4	4	4	4	4	4	4	4	1	1	4	1	1	1
	idx5	5	5	5	5	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1

		C548	C553	C557	C566	C574	C575	C579	C582	C589	C596	C711	C714	C721	C722	C725	C1386	C1387	C1388	C1389	C1390
pre_scaling_0_1	idx2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	idx3	3	1	3	3	1	3	1	3	1	1	3	3	3	3	3	3	1	1	1	1
	idx4	1	4	1	1	4	1	4	1	4	4	1	1	1	1	1	1	4	4	4	4
	idx5	1	2	1	1	2	1	2	1	2	1	1	1	1	1	1	1	2	1	2	2

		C1465	C1466	C1467	C1468	C1469	C1470	C1471	C1472	C1473	C1705	C1715	D756	D800	D801	D802	D841	D842	D924	D1173	D1273
pre_scaling_0_1	idx2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2
	idx3	3	3	1	3	3	3	1	3	1	2	3	3	1	3	1	3	1	3	3	3
	idx4	1	1	4	1	1	1	4	1	4	1	1	1	4	2	4	2	4	1	2	2
	idx5	1	1	2	1	1	1	2	1	2	4	1	1	2	3	2	3	2	1	3	3

		D1274	D1275	D1276	D1288	D1289	D1290	D1291	D1312	D1313	D1335	D1364	D1365	D1385	D2471	D2472	D2595	D2626	D2885
pre_scaling_0_1	idx2	2	2	1	1	1	1	1	2	1	2	2	2	2	2	2	1	1	2
	idx3	3	3	3	2	2	2	2	3	2	3	3	3	3	1	1	3	2	3
	idx4	2	1	1	3	3	3	3	2	3	2	1	2	1	4	4	1	3	1
	idx5	3	4	4	4	4	4	4	3	4	3	1	3	1	2	2	4	5	1

Fig. 25 Table displaying to which cluster each sample belongs, for each K value of the SFRM (idx2 = k:2, idx3 = k:3, etc.)

Based on the average silhouette values, the most efficient clustering solution is that of k=2 (idx=2). According to Fig. 25, all samples from Family A oils are assigned to cluster one. Almost all of Family C oil samples are assigned to cluster two (only sample C1705 is assigned to cluster 1). Oil samples from family B are all assigned to cluster 2, whereas family D oil samples are assigned in both clusters.

5.2.3 Principal Component Analysis on SFRM

The original variables used in the Saturate Fraction Ratios Model (SFRM) include the compositional factors Pr/Ph, nC_{17}/Pr, nC_{18}/Ph and the carbon preference indices for both lighter (nC_{14}-nC_{20}) and heavier (nC_{22}-nC_{30}) alkanes of the saturated fraction hydrocarbons. The

sample scores of the first two principal components and the respective loading diagrams are presented in Fig. 26. The percentages of variation attributed to each of the Principal Components are shown in Fig. 26c.

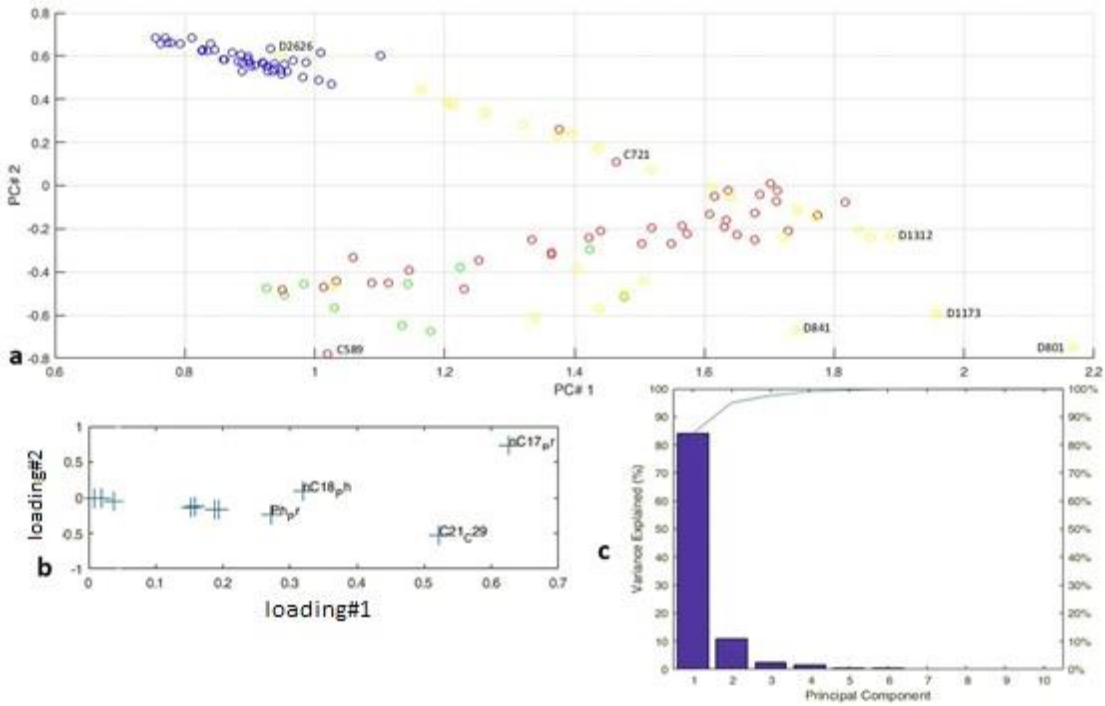


Fig. 26 a) Sample scores for the first to Principal Components resulting from the Saturated Fraction Ratios Model (SFRM) of selected Williston Basin petroleum oils. Colors on sample symbols indicate compositional families determined by independent analysis. Blue color represents oils of Family A, green applies for Family B oils, red for Family C oils and yellow for Family D oils. "Pre_scaling_0_1" command was used on the data set. b) Original Variable loadings for the first to Principal Components resulting from the Saturated Fraction Ratios Model (SFRM) of selected Williston Basin petroleum oils. c) Percentage of variance explained by each Principal Component.

SFRM sample scores of the first two PCs explain 83 per cent of the variance (Fig. 26c). There are two linear gradients observed in the data by comparing the sample scores of the first two principal components of the SFRM; a dispersed positive gradient displayed by samples with positive PC1 scores and PC2 scores less than 0, and a relatively tight negative gradient consisting of both positive PC1 and PC2 scores. As in the SFCM, these gradients also indicate that distinctive family characteristics and linear compositional variations of the original variables exist within each family. When the samples are compared to the biomarker-based oil families, Family A is once again clearly distinguished by consistently positive PC1 and PC2 scores and a linear variation between them. Only sample D2626 overlaps with this group, however. Family D oils are also characterized by a general positive gradient, while Families B and D are defined by mainly a positive gradient. All gradients exhibit high positive PC1 scores but, Families B, C and D exhibit negative PC2 scores. As in the SFCM, the fields of PC1 and PC2 in Family C overlap those of Families B and D, effectively obscuring their separation. However, Family C samples appear to fall along a positively correlated gradient in PC1 vs PC2 space.

The original variable loadings for the SFRM indicate a lack of discriminating power of the nC_{17}/Pr and nC_{18}/Ph with respect to Families B and C, which opposes to Osadezt et al. [32], who claim that this biomarker parameter is highly effective as far as the discrimination among these affiliations is concerned.

5.2.4 Discussion on the performance of MDA on the SFRM

MDA methods on the Saturated Fraction Ratios Model seemed to perform in a similar manner as in the Saturated Fraction Compositional Model. In all three methods Family A was significantly distinguished in contrast to the rest familial affiliations. Only sample D2626, in PCA overlapped with family A samples. As far as k-means is concerned, even though it discretizes family A, as a whole, it failed in considerably in separating families B, C and D. It produced a two-cluster solution.

5.3 Gasoline Range Compositional Model (GRCM)

5.3.1 Hierarchical Clustering on GRCM

Applying the Hierarchical Clustering algorithm on GRCM, produced the following dendrogram (Fig. 27). Single linkage with Euclidean distance were combined this time.

From the figure, we notice that oil samples from all four family affiliations overlap, presenting no clear distinction. In this case, we also observe that a few samples from C and D are excluded from the clustering solution (samples B1873, B1874, B1014, C1390, and D842).

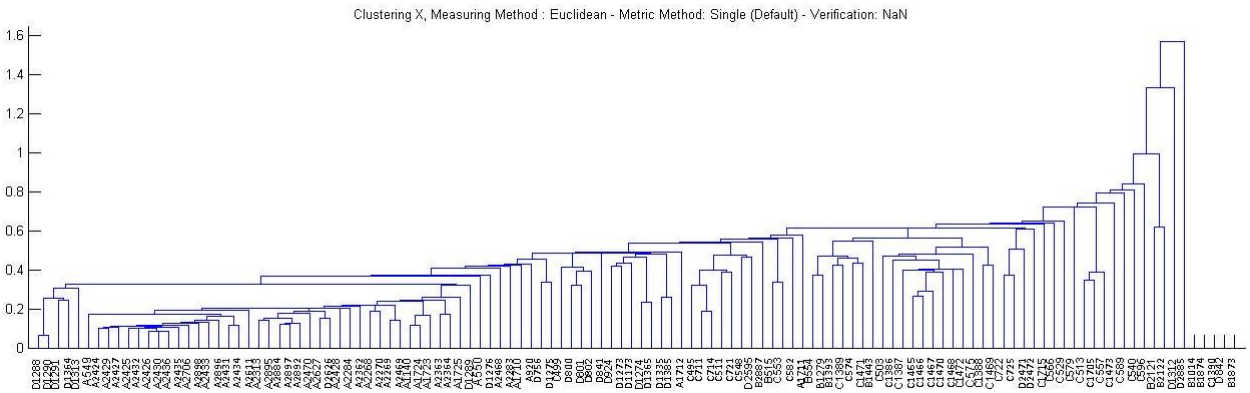


Fig. 27 Resulting Dendrogram under the command "pre_scaling_0_1" for the Gasoline range compositional model (GRCM).

These components presented zero values for all variables. To examine how the model would perform without these values, they were removed from the data set and then hierarchical clustering was implemented again. The following dendrogram is the result.

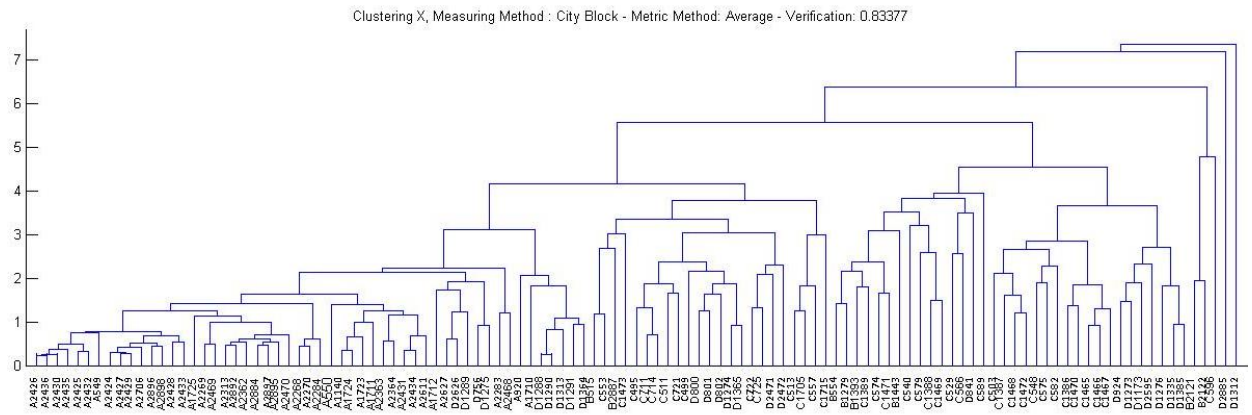


Fig. 28 Resulting Dendrogram under the command "pre_scaling_0_1" for the Gasoline range compositional model (GRCM) after removing zero values.

Implementing the algorithm produced another two outlier samples from family D (samples D1312 and D2885). Family A oil samples, however, seem to distinguish from the rest, but not sufficiently enough, as there is a slight overlap with samples from family D. As far as families B, C and D are concerned, there is a considerable overlap among them.

5.3.2 k-means algorithm on GRCM

Implementing the k-means algorithm on the Gasoline range compositional model produced the following results. Components with zero values (as mentioned before) were kept out of the analysis.

Table 5 Summary of k-means clustering under the "pre_scaling_0_1" pretreatment option on the Gasoline Range Compositional Model (GRCM).

K-values	Best distances sums	Average silhuette values
K=2	60,1718	0.4438
K=3	49,6733	0.4572
K=4	43,5093	0.4510
K=5	38,3703	0.4271

The silhouette plots for K=2, K=3, K=4 and K=5 clusters are presented in the following figure (Fig. 23).

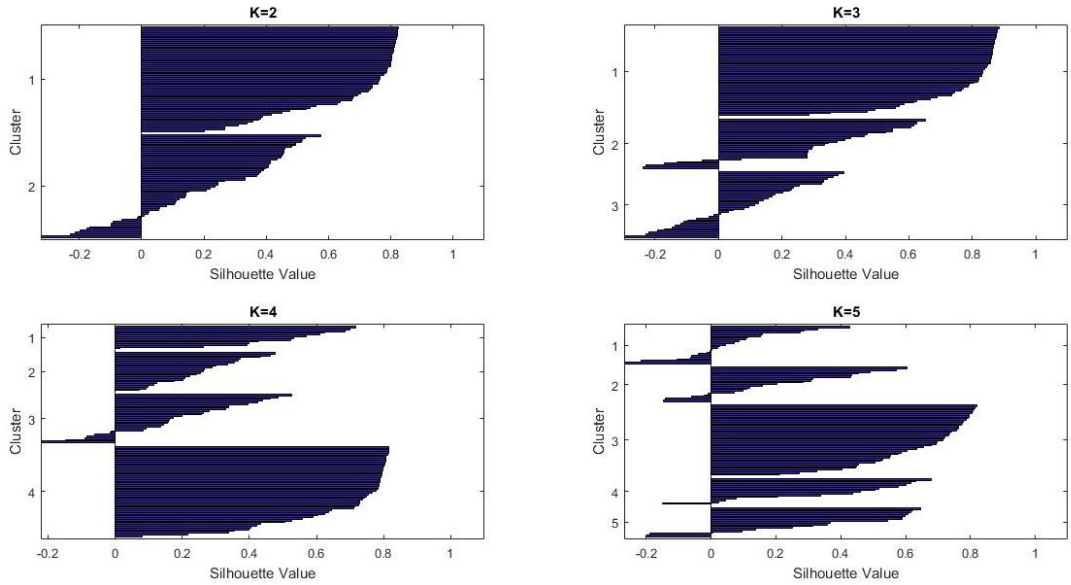


Fig. 29 Silhouette plots for k=2, k=3, k=4 and k=5 clusters under the " pre_scaling_0_1" pretreatment option for the Gasoline Range Compositional Model (GRCM).

From the silhouette plots (Fig. 29), we observe that generally we obtain silhouette values in the range of 0.01-0.8. Negative silhouette coefficients are present in all cases. Average silhouette values are close for all clustering solutions, with a maximum of 0.4572 for K=3 (Table 5). The outcome of this analysis, infers that grouping into three clusters using k-means is more efficient compared to grouping into two, four or five clusters. In Fig. 31 we can observe, which cluster each sample is assigned to.

In Fig. 30 we observe the clustering solution of k-means for k=3. The figure shows the three clusters along with their centroids.

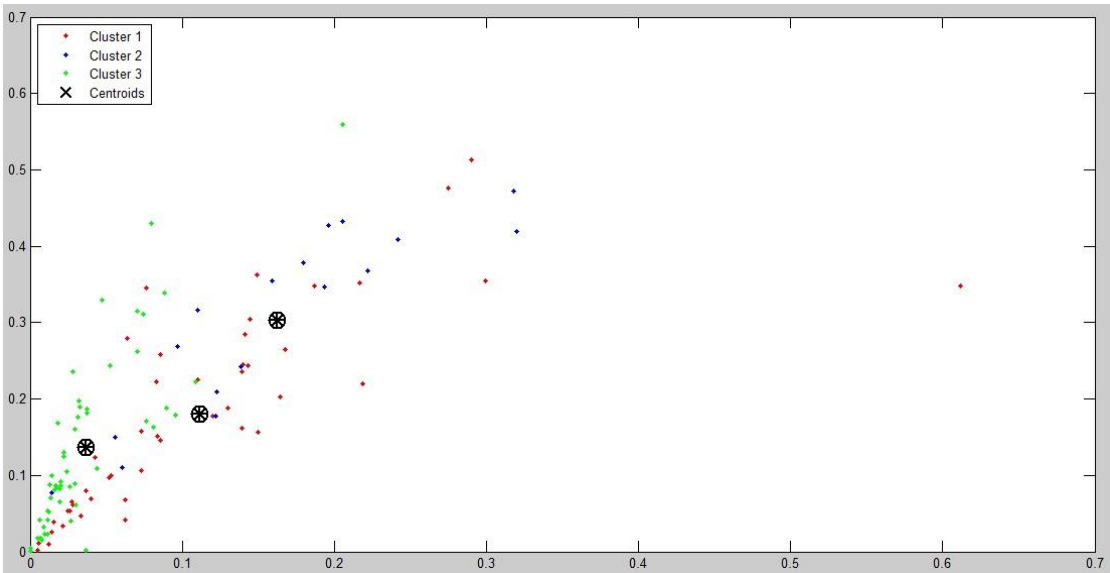


Fig. 30 Plot of k-means clustering for k=3, of the Gasoline Range Compositional Model (GRCM). The ⊗ symbol represents the centroid of each cluster.

		A549	A550	A920	A1140	A1710	A1711	A1712	A1723	A1724	A1725	A2268	A2269	A2270	A2283	A2284	A2313	A2362	A2363	A2364	A2424
pre_scaling_0_1	idx2	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	idx3	3	3	3	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
	idx4	1	1	4	1	4	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1
	idx5	1	1	5	1	5	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1

		A2425	A2426	A2427	A2428	A2429	A2430	A2431	A2432	A2433	A2434	A2435	A2436	A2468	A2469	A2470	A2611	A2627	A2706	A2884	A2892
pre_scaling_0_1	idx2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	idx3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
	idx4	1	1	1	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1
	idx5	1	1	1	1	1	1	1	1	1	1	1	1	4	1	1	1	1	1	1	1

		A2895	A2896	A2897	A2898	B515	B554	B1279	B1393	B1443	B2121	B2122	B2887	C495	C499	C503	C511	C513	C529	C540	C548
pre_scaling_0_1	idx2	1	1	1	1	1	2	2	2	2	2	2	2	2	1	2	1	2	2	2	2
	idx3	3	3	3	3	1	1	1	1	1	1	1	1	1	1	2	1	2	1	2	1
	idx4	1	1	1	1	4	2	2	2	2	2	2	4	4	4	3	4	4	2	3	4
	idx5	1	1	1	1	4	2	2	2	2	2	2	4	4	4	3	4	3	2	3	5

		C553	C557	C566	C574	C575	C579	C582	C589	C596	C711	C714	C721	C722	C725	C1386	C1387	C1388	C1389	C1465	C1466
pre_scaling_0_1	idx2	1	1	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2
	idx3	1	3	1	1	2	2	1	1	1	1	1	1	1	1	2	2	2	1	2	2
	idx4	4	4	2	2	3	3	3	2	2	4	4	4	4	4	3	3	3	2	3	3
	idx5	4	4	2	2	3	2	5	2	4	4	4	4	4	4	3	3	2	2	3	3

		C1467	C1468	C1469	C1470	C1471	C1472	C1473	C1705	C1715	D756	D800	D801	D802	D841	D924	D1173	D1273	D1274	D1275	D1276
pre_scaling_0_1	idx2	2	2	2	2	2	2	2	1	2	1	2	2	2	2	2	2	2	2	1	2
	idx3	2	2	1	2	1	2	1	2	2	3	1	1	1	1	1	1	1	1	3	1
	idx4	3	3	3	3	2	3	2	4	3	1	4	4	4	2	4	2	2	4	1	4
	idx5	3	3	2	3	2	3	4	3	3	1	5	4	5	2	5	5	5	5	1	5

		D1288	D1289	D1290	D1291	D1312	D1313	D1335	D1364	D1365	D1385	D2471	D2472	D2595	D2626	D2885
pre_scaling_0_1	idx2	1	1	1	2	2	2	2	2	2	2	1	2	2	1	1
	idx3	3	3	3	1	1	1	1	1	1	1	1	1	1	3	3
	idx4	4	1	4	4	2	4	3	4	4	3	4	4	4	1	1
	idx5	5	1	5	5	2	5	5	5	5	5	4	4	5	1	4

Fig. 31 Table displaying to which cluster each sample belongs, for each K value of the GRCM (idx2 = k:2, idx3 = k:3, etc.)

Even though the three-cluster solution seems to be the most efficient out of the analysis, from the plot we observe that the clusters present no clear boundaries from one another. The overlapping among samples is evident. Fig. 31 confirms this fact as it presents in which of the three clusters, each sample is assigned to.

5.3.3 Principal Component Analysis on GRCM

The sample scores of the first two principal components and the respective loading diagrams are presented in Fig. 32. The percentages of variation attributed to each of the Principal Components are shown in Fig. 32c.

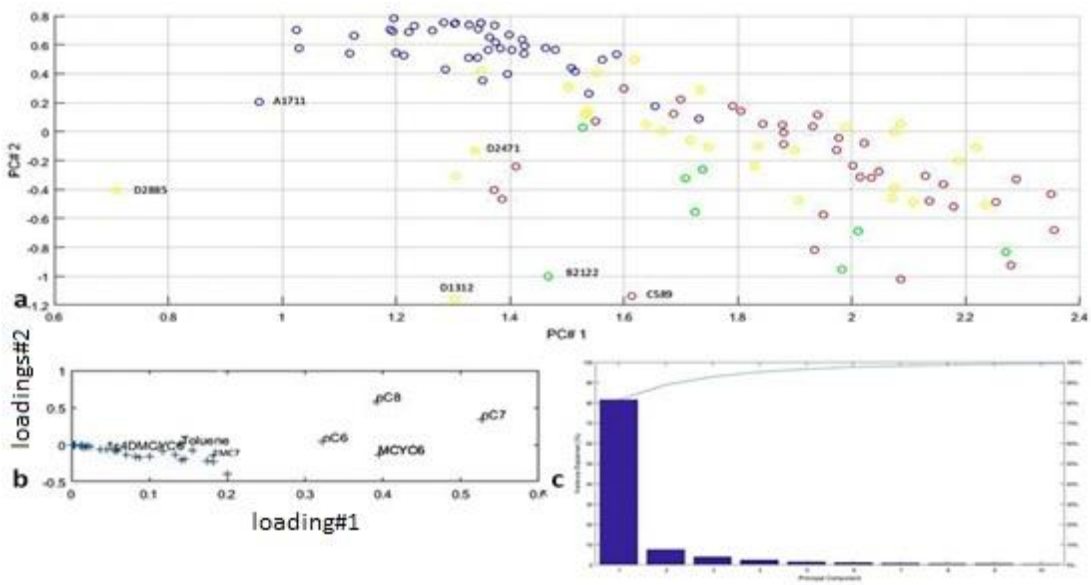


Fig. 32 a) Sample scores for the first to Principal Components resulting from the Gasoline Range Compositional Model (GRCM) of selected Williston Basin petroleum oils. Colors on sample symbols indicate compositional families determined by independent analysis. Blue color represents oils of Family A, green applies for Family B oils, red for Family C oils and yellow for Family D oils. "Pre_scaling_0_1" command was used on the data set. b) Original Variable loadings for the first to Principal Components resulting from the Gasoline Range Compositional Model (GRCM) of selected Williston Basin petroleum oils. c) Percentage of variance explained by each Principal Component.

GRCM sample scores of the first two PCs explain 82 per cent of the variance (Fig. 32c). There is generally one linear gradient observed in the data by comparing the sample scores of the first two principal components of the GRCM; a dispersed negative gradient displayed by samples with positive PC1 scores and PC2 scores both positive and negative. Family A is once again clearly distinguished by consistently positive PC1 and PC2 scores and a linear variation between them. The gradient exhibits high positive PC1 scores but for Families B, C and D exhibits also negative PC2 scores. The gradients of Families B, C and D overlap each other's scores resulting in the obscureness of their separation.

The variable loadings for the GRCM indicate that PC1 is controlled strongly by loadings attributed to the relative concentration of n-alkanes and branched and cyclic alkanes. High negative PC1 loadings are characteristic of the GRH n-alkanes, while the cyclic and branched alkanes with 6 to 8 carbon atoms are characterized by strong positive values. In our case the GRCM fails in the task of classifying the four family affiliations.

5.3.4 Discussion on the performance of MDA on the GRCM

Although in several studies (e.g. [38]) the Gasoline Range Compositional Model appears to be successful in classifying efficiently oil samples of the four family affiliations recognized in Williston Basin, in our case it substantially fails. All statistical methods that were implemented on this model, classified relatively sufficiently only family A. Families B, C and D presented a significant overlap, both one to another, but also with Family A. This is evident from the dendrogram of Fig. 28 as well as from Fig. 32a. The overlapping of oil families is incredibly apparent in the k-means plot (Fig. 30), where there is no distinct cluster.

5.4 Biomarkers Compositional Model (BCM)

The biomarkers of the given sample set were examined in multiple ways; firstly as a whole and secondly in their separate parts of steranes and hopanes. The results that each model produced were similar, as far as the classification of oil families, is concerned. For this reason, only the results from BCM will be presented in the upcoming paragraphs, as the most characteristic.

5.4.1 Hierarchical Clustering on BCM

Applying the Hierarchical Clustering algorithm on BCM, produced the following dendrogram (Fig. 27). Average linkage with Euclidean distance were combined this time.

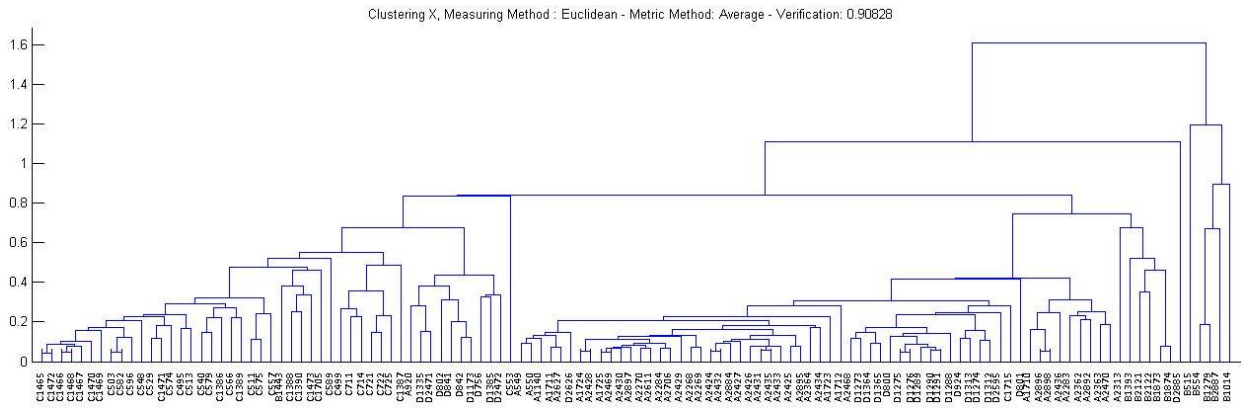


Fig. 33 Resulting Dendrogram under the command "pre_scaling_0_1" for the Biomarkers compositional model (BCM).

Hierarchical clustering on BCM seems to separate relatively well Family C oils. Only sample B1443 (of Family B) overlaps with family C. The dendrogram illustrates an overlapping of Family D with Family A and the rest of Family B samples form a small group which is interrupted by sample D2885.

5.4.2 k-means algorithm on BCM

Implementing the k-means algorithm on the Biomarkers compositional model produced the following results.

Table 6 Summary of k-means clustering under the "pre_scaling_0_1" pretreatment option on the Gasoline Range Compositional Model (GRCM).

K-values	Best distances sums	Average silhuette values
K=2	250.993	0.5503
K=3	15.438	0.6665
K=4	131.334	0.5865
K=5	113.741	0.5425

The silhouette plots for K=2, K=3, K=4 and K=5 clusters are presented in the following figure (Fig. 23).

From the silhouette plots (Fig. 34), we observe that generally the highest silhouette values we obtain almost reach the value of 0.9. Fluctuations in the width of clusters is present in all cases and so are negative silhouette coefficients. Average silhouette values fall in the range of 0.5425-0.665, with 0.665 being the maximum for K=3 (Table 6). The outcome of this analysis, infers that grouping into three clusters using k-means is more efficient compared to grouping into two, four or five clusters. In Fig. 36 we can observe, which cluster each sample is assigned to.

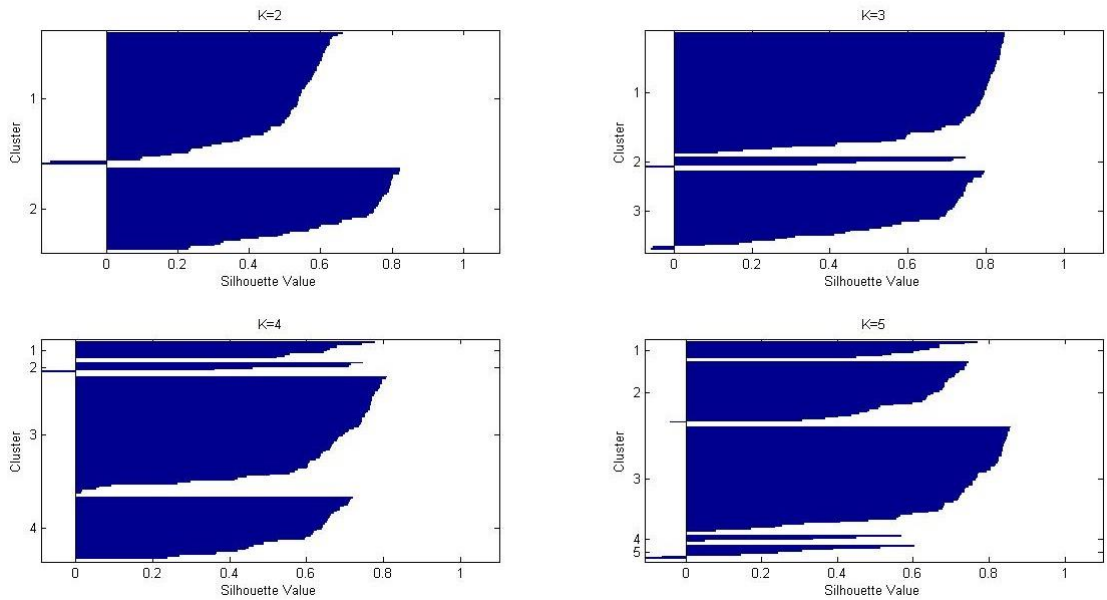


Fig. 34 Silhouette plots for k=2, k=3, k=4 and k=5 clusters under the " pre_scaling_0_1" pretreatment option for the Biomarkers Compositional Model (BCM).

Fig. 35 illustrates the clustering solution of k-means for k=3. The figure shows the three clusters along with their centroids.

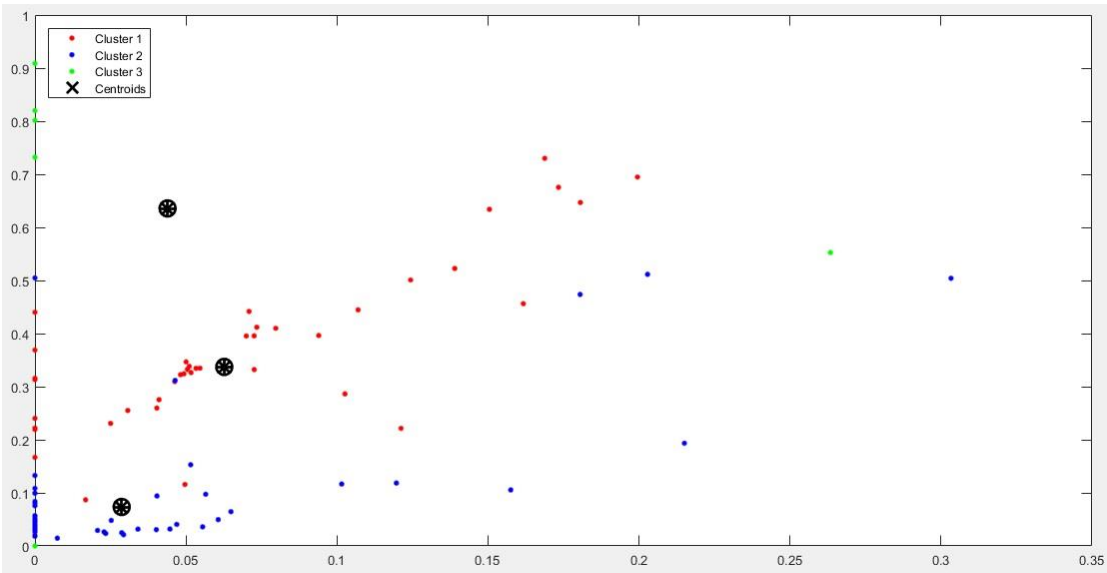


*Fig. 35 Plot of k-means clustering for k=3, of the Biomarkers Compositional Model (**B**CM). The ⊗ symbol represents the centroid of each cluster.*

		A549	A550	A920	A1140	A1710	A1711	A1712	A1723	A1724	A1725	A2268	A2269	A2270	A2283	A2284	A2313	A2362	A2363	A2364	A2424
pre_scaling_0_1	idx2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	idx3	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	idx4	3	3	4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
	idx5	4	4	1	4	3	4	4	4	4	4	4	4	4	3	4	3	3	3	4	4

		A2425	A2426	A2427	A2428	A2429	A2430	A2431	A2432	A2433	A2434	A2435	A2436	A2468	A2469	A2470	A2611	A2627	A2706	A2884	A2892
pre_scaling_0_1	idx2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	idx3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	idx4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
	idx5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	3	4	4	4	4	3

		A2895	A2896	A2897	A2898	B515	B554	B1014	B1279	B1393	B1443	B2121	B2122	B2887	B1873	B1874	C495	C499	C503	C511	C513
pre_scaling_0_1	idx2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2
	idx3	2	2	2	2	3	3	3	3	2	2	2	3	3	2	2	1	1	1	1	1
	idx4	3	3	3	3	1	1	1	1	1	3	1	1	1	3	3	2	2	2	2	2
	idx5	4	3	4	3	2	2	2	2	3	4	3	2	2	3	3	5	5	5	5	5

		C529	C540	C548	C553	C557	C566	C574	C575	C579	C582	C589	C596	C711	C714	C721	C722	C725	C1386	C1387	C1388
pre_scaling_0_1	idx2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	idx3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	idx4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
	idx5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

		C1389	C1390	C1465	C1466	C1467	C1468	C1469	C1470	C1471	C1472	C1473	C1705	C1715	D756	D800	D801	D802	D841	D842	D924
pre_scaling_0_1	idx2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	1	1	2	2	2	1
	idx3	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	2	1	1	1	2
	idx4	2	2	2	2	2	2	2	2	2	2	2	2	3	4	3	3	4	4	4	4
	idx5	5	5	5	5	5	5	5	5	5	5	5	5	4	1	4	4	1	1	1	4

		D1173	D1273	D1274	D1275	D1276	D1288	D1289	D1290	D1291	D1312	D1313	D1335	D1364	D1365	D1385	D2471	D2472	D2595	D2626	D2885
pre_scaling_0_1	idx2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	1	1	1
	idx3	1	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	1	2	2	2
	idx4	4	3	4	3	3	3	3	3	3	4	3	4	3	3	4	4	4	3	3	3
	idx5	1	4	4	4	4	4	4	4	4	4	4	1	4	4	1	1	1	4	4	3

Fig. 36 Table displaying to which cluster each sample belongs, for each K value of the BCM (idx2 = k:2, idx3 = k:3, etc.)

Fig. 35 illustrates the three-cluster solution that silhouette analysis produced as the most efficient. The clusters do not exhibit clear boundaries and overlapping is evident. Fig. 36 supports the overlapping fact as it illustrates in detail in which cluster each sample is assigned to.

5.4.3 Principal Component Analysis on BCM

The sample scores of the first two principal components and the respective loading diagrams are presented in Fig. 32. The percentages of variation attributed to each of the Principal Components are shown in Fig. 32c.

BCM sample scores of the first two PCs explain almost 90 per cent of the variance (Fig. 37c). By comparing the sample scores of the first two principal components of the BCM, we observe no clear distinction among families. All families exhibit high positive PC1 scores and all of them present both negative and positive PC2 scores. Family B (green symbols on the PC plot) exhibit solely negative PC2 scores. Family A overlaps here mainly with family D and a few samples of family D overlap with family C. Scores of family B are quite dispersed in the plot.

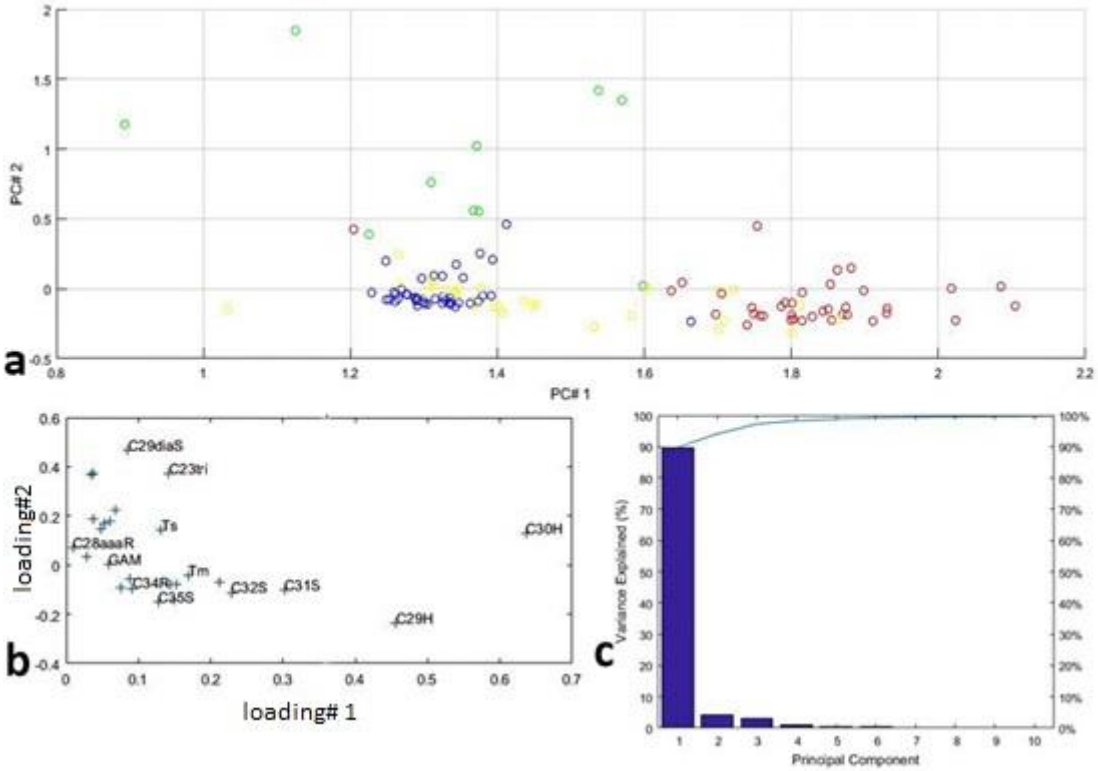


Fig. 37 a) Sample scores for the first to Principal Components resulting from the Biomarkers Compositional Model (BCM) of selected Williston Basin petroleum oils. Colors on sample symbols indicate compositional families determined by independent analysis. Blue color represents oils of Family A, green applies for Family B oils, red for Family C oils and yellow for Family D oils. "Pre_scaling_0_1" command was used on the data set. b) Original Variable loadings for the first to Principal Components resulting from the Biomarkers Compositional Model (BCM) of selected Williston Basin petroleum oils. c) Percentage of variance explained by each Principal Component.

5.4.4 Discussion on the performance of MDA on the BCM

The Biomarkers Compositional Model appears to be successful in classifying relatively well oil family C. All methods produced similar results as far as this classification pattern is concerned. Families A and D overlap significantly, while family B overlaps slightly with family D.

All in all, the performance of MDA methods was insufficient, failing in all models to classify the samples into four familial affiliations. Based on common compositional information, it seems that unsupervised methods fail to cluster these oils. They cannot be implemented blindly without additional information. For this reason, in the next chapter we examine the compositional character of the given data set in an alternative approach.

6. Compositional Data

As discussed in the previous chapters, MDA methods failed in the task of classifying the data set into distinct oil family affiliations. This applies to all the compositional models and is probably attributed to the nature of the data, which fall into a special category of data; the Compositional Data. The Saturates Fraction Ratios Model is excluded from this category and none of the following information concerns this model.

Compositional Data (CoDa) are a type of multivariate data, the components of which represent proportions or fractions of a whole. Such data come in a closed form, meaning that they sum to a constant value (e.g. one if measured in parts per unit or 100 if measured in percentages). However, the term Compositional Data, covers all those vectors representing parts of a whole which only carry information on the relative (and not the absolute) frequencies, with which different and positive components occur.

Typical examples of Coda are geochemical elements in geology, data corresponding to categories of sedimentary particle-size distributions, proportions of fossil species in two or more assemblages, body composition (fat, sugar, etc.) in medicine, nutrient-balance ionomics (measurement of the total elemental composition of an organism to address biological problems) in agriculture, genotype frequency in genetics, chemical compositions in chemistry, and many more other. This type of data is generally widespread in disciplines supporting modeling, classification or discrimination and is characterized by specific numerical properties that have significant consequences for any statistical analysis [85] [86] [87] [88] [89] [90] [91]. Their fundamental properties are briefly reviewed in the upcoming paragraphs.

6.1 The Constant Sum Constraint (CSC) – Impacts on the Analysis

As mentioned before, Compositional Data only convey relative information as they represent part of a whole, and their unique properties are a corollary of this fact. They concern data consisting of vectors of always positive components, often subject to a constant (unit-) sum constraint; they must sum to one because they are proportions. Their main difference to unconstrained variables is that they are never free to vary independently, which in turn imposes constraints upon their variance-covariance structure (Aitchison 1986, chapter 3). The constant sum constraint forces at least one of their covariance to obtain a negative value. The result is at least one correlation or coefficient between elements, is also negative. This is explained as a consequence of the Euclidean Foundation of classical statistics, where the scale is absolute and not relative.

In particular, for a D-part composition $[x_1, \ldots, x_D]$ with the component sum $x_1 + \ldots + x_D = 1$, since

$$\mathrm{cov}(x_1, x_1 + \ldots + x_D) = 0$$

we have

$$\mathrm{cov}(x_1,x_2) + \ldots + \mathrm{cov}(x_1,x_D) = -\mathrm{var}(x_1) .$$

The right-hand side here is negative except for the trivial case where the first component is constant.

The fact that data are closed, induces invalid correlations and as a result, all methods based on the covariance or correlation matrix of vectors of observations, are inappropriate to examine and analyze Compositional Data in crude or raw form (e.g. as simple percentages) [92]. Conventional statistical methods present uncertainty in the analysis of compositional data, as far as the results are concerned. The main reason is because it is not possible to distinguish between the spurious effects caused by the constant sum constraint and the effects that would be attributed to natural processes. Rock [90] in his paper describes some of the problems: trends and clusters on petrological ternary and principal components diagrams can have little or no geological significance; dendrograms produced by cluster analysis can be severely biased; results from discriminant analysis are likely to be illusory; any correlation coefficient will be affected to an unknown degree by spurious effects induced by the constant sum constraint, etc. In general, problems appear with all methods based on regression and multivariate analysis which rely on an assumption of multivariate normality. Such methods refer to Factor Analysis, Discriminant Analysis or Principal Component Analysis and they seem to perform better on unconstrained random variables.

6.2 Approaches in the Statistical Analysis of CoDa

In the early 1980's the analysis of Compositional Data began to obtain a more efficient form. The key to such analysis is the relative magnitudes and variations of the parts in a D-part composition, rather than their absolute values. Thus, the information provided is essentially about ratios and any meaningful function (scale-invariant) of a composition should be expressed under such terms. The principal justification for using ratios of components is the Sub-Compositional coherence, which is a fundamental property of Aitchison's approach to compositional data analysis. Ratios are unaltered in the process of forming sub-compositions ($s_i/s_j=u_i/u_j$) which should mean that there exists some form of covariance structure based upon them.

However, mathematically and statistically speaking, ratios are somewhat difficult to handle. For example, between $var(u_i/u_j)$ and $var(u_j/u_i)$ there does not exist any simple relationship. Therefore, in order to overcome this difficulty, Aitchison was the first to introduce the log-ratio method, because of the simplicity of relationships such as

$$var\{log(x_i/x_j)\}=var\{log(x_j/x_i)\}.$$

Since there is also a one-to-one correspondence between compositions and a full set of log-ratios, for example,

$$[\, y_1 \ldots y_{D-1}] = [\log(\, x_1/\, x_D) \ldots \log(\, x_{D-1} / \, x_D)]$$

with inverse

$$[\, x_1\, x_2 \ldots x_D] = [\, \exp(\, y_1) \ldots \exp(\, y_{D-1}) + 1\,] \,/\, \{\, \exp(\, y_1) + \ldots + \exp(\, y_{D-1}) + 1\, \}$$

any problem or hypothesis concerning compositions can be fully expressed in terms of log ratios and vice versa.

The proposed methodology is simple; first transform each of the compositions (ui,…,ud) to their log-ratio vectors and then apply standard multivariate procedures upon them. The conclusions of the unconstrained multivariate analysis can then be translated back into conclusions about the compositions, and the analysis is complete.

The aforementioned methodology represents a transformation technique, widely utilized in statistics. Starting with McAlister [93] and his logarithmic transformation, the lognormal distribution and the significance of the geometric mean, the log-ratio transformation comes in line with a long tradition of statistical methodology.

6.3 The Simplex S^D – Fundamental Properties of CoDa Analysis

There has been much debate against transformation techniques over the scientific community [94, 95, 96, 97, 98, 99, 100, 101, 102, 103]. However, while most of them are still valid, new approaches have been developed towards the statistical analysis of compositional data. *Staying-in-the-simplex* approach, represents part of them, offering the advantage of keeping the analysis free of dependence upon transformations and results in unconstrained multivariate analysis. Therein, compositional data analysis is conducted within a simple algebraic-geometric structure on the simplex. At this point, the term simplex has to be defined.

One of the main differences between compositional and unconstrained data, is the sample space within which, each type lies. The natural sample space of CoDa is the (restricted) unit simplex $\mathbf{S^d}$ (while unconstrained data belong to the real space R). The simplex is a basic geometric element in a Euclidean space, and is defined as

$$S^D = \{x = [x_1, x_2, \dots, x_D] \in \mathbb{R}^D | x_i > 0, \iota = 1,2 \dots, D; \sum_{\iota=1}^{D} x_i = \varkappa\}$$

The constant $\varkappa$ simplex is positive and arbitrary. Frequent values for $\varkappa$ are 1 (per unit), 100 (percent, %), 1000, etc. The simplex S^D is a line segment in one dimension (D=1), a triangle in two dimensions (D=2), a tetrahedron in three dimensions (D=3), and so on. As far the superscript in the S^D is concerned, it accounts for the effective dimension of D-part compositions and is often reduced to D-1, due to the unit-sum constraint. A unit-simplex is defined as

$$S^D = \{[x_1, \dots, x_D]: x_i > 0(\iota = 1,2 \dots, D) | x_1 + \cdots + x_D = 1\}$$

With this representation, scale invariance is an element to be ensured by formulating all statements concerning compositions in terms of ratios of components.

Scale invariance is one the fundamental principles governing the compositional data analysis according to Aitchison. What scale variance addresses, is that statistical inferences about compositional data should not depend upon the scale of the data.

More specifically, two vectors of D positive real components x, y ∈ R D + (x_i , y_i≥ 0 for all i = 1, 2, . . . , D), are compositionally equivalent if there exists a positive scalar λ∈ R + such that x = λ· y and, equivalently, C(x) = C(y). It is highly reasonable to ask our analyses to yield the same result, independently of the value of λ. This is what Aitchison (1986) called scale invariance.

A function f(·) is scale-invariant if for any positive real value λ∈ R + and for any composition x ∈ SD, the function satisfies f(λx) = f(x), i.e. it yields the same result for all vectors compositionally equivalent. This can only be achieved if f(·) is a function only of log-ratios of the parts in x (equivalently, of ratios of parts) [102, 104]. According to Aitchison, apart from scale invariance, there are also two other conditions that should be satisfied in order for any statistical method to be performed on compositional data; permutation invariance and sub-compositional coherence.

A function is permutation-invariant if it yields equivalent results when the order of parts of the composition is changed. Two examples might illustrate what "equivalent" means here. If we are computing the distance between our initial sandstone and our final sand compositions, this distance should be the same if we work with [Q, F, R] or if we work with [F, R, Q] (or any other permutation of the parts). On the other side, if we are interested in the change occurred from sandstone to sand, results should be equal after reordering. A classical way to get rid of the singularity of the classical covariance matrix of compositional data is to erase one component: this procedure is not permutation-invariant, as results will largely depend on which component is erased.

Before examining the topic of sub-compositional incoherence, the definition of sub-composition must be given. A composition only representing some of the possible components is called a sub-composition and most of real compositional data is actually representing a sub-composition, as we never analyze each and every possible component of our samples. Sub-compositions represent the marginals of compositional data analysis. Two compositions (a greater and a smaller one) sharing common parts (therefore, the smaller is the sub-composition) should produce common correlations for these parts, regardless of whether we analyzed only that sub-composition or a larger composition containing other parts. This is what coherence means. If this is not the case, then there is what is expressed as sub-compositional incoherence.

6.4 Perturbation and Powering

In any sample space there is, only certain operations can be performed. For example, in real space R^D translation and scalar multiplication are the most commonly used operations. However, the typical algebraic/geometric operations (addition/translation, product/scaling, scalar product/orthogonal projection, Euclidean distance) used to deal with conventional real vectors are neither sub-compositionally coherent nor scaling invariant. The simplex is a sample space characterized by a different, compositional geometry and such operations would not be adequate for any analysis within it. Two fundamental groups of operations for the simplex are the perturbation operations, analogous to translation in the real space, and power transformation, analogous to multiplication by a scalar in the real space. These operational sets were introduced by Aitchison [89], they underpin the complete algebraic – geometric structure of the simplex and both require in their definition the closure operation [104, 105]. Closure is nothing but the operation responsible for the constant sum constraint as it divides each component of a vector by the sum of the components and represents the projection of a vector with positive components onto the simplex.

For any two equivalent compositions x and X, in the same compositional class, there is a scale relationship (X_1, …, X_D)= (ax_1, …, ax_D) for some a >0, where each component of x is scaled by the same factor a to obtain the corresponding component of X. For any two compositions x and X in different compositional classes c and Ca similar, but differential, scaling relationship (x_1, …, X_D)= (p_1x_1, …, p_Dx_D) can always be found, simply by taking $p_i = X_i/x_i$(i= 1, …,D).Denoting the operation between the positive perturbing vector p= (p_1, … ,p_D)and the composition x by $\oplus$ we have p $\oplus$ x =(p_1x_1, … , p_Dx_D) and X = p $\oplus$ x. Such a perturbation operator is then easily adapted to the simplex simply by defining p$\oplus$u =(p_1u_1, … , p_Du_D)/(p_1u_1+ … + p_Du_D). Note that the roles of p and u are interchangeable in this definition and we can conveniently restrict p to lie in the simplex S^d. Perturbations thus defined form a group, with p^{-1}, the inverse of p, defined as (p_1^{-1}, … , p_D^{-1})/(p_1^{-1}+… + p_D^{-1})and the identity perturbation as (1/D, … ,1/D).Moreover, for any two compositions u, U there is a unique perturbation p $\in S^d$ such that U = p $\oplus$ u and u = $p^{-1}\oplus$ U, where p = U $\oplus$ u^{-1}.Thus,the perturbation U $\oplus$ u^{-1}, or equivalently X $\oplus x^{-1}$ characterizes the change from c to C; the change from X to x is simply the inverse perturbation U $\oplus$ u^{-1}.

Powering or power transformation, as mentioned before, is the second fundamental operational group in the simplex. First, we define the power operation and then consider its relevance in compositional data analysis. For any real number $a \in R^1$ and any composition $x \in S^D$, we define:

$$X = \alpha \otimes x = C\,[\,x_1^{\alpha} \ldots x_D^{\alpha}]$$

as the *a*-power transform of *x*. Such an operation arises in compositional data analysis in two distinct ways. First it may be of relevance directly because of the nature of the sampling process. More indirectly the power transformation can be useful in describing regression relations for compositions.

It is clear that powering $\otimes$and perturbation $\oplus$ play a significant role as far as the geometry of S^D is concerned. Powering is an external operation whereas perturbation is an internal one, and it would be meaningless to establish that they define a vector or linear space structure on S^D. In particular, the $\oplus$operation defines an abelian group with identity e = [1, . . . ,1] / D. Both operational groups are marked by certain properties, which will now be addressed.

$$x \oplus y = y \oplus x,\ (x \oplus y) \oplus z = x \oplus (y \oplus z),\ a \otimes (x \oplus y) = (a \otimes x) \oplus (a \otimes y).$$

The operator Θ is the inverse of $\oplus$and is defined by:

$$x \Theta y = C[x_1/y_1 \ldots x_D/y_D]$$

and plays an important role in the construction of compositional residuals.

The structure can be extended by the introduction of the simplicial metric

$$\Delta: S^D x S^D \rightarrow R \geq 0$$

Defined as follows:

$$\Delta\,(x,y) = \left[\sum_{i=1}^{D}\left\{log\frac{x_i}{g(x)} - log\frac{y_i}{g(y)}\right\}^2\right]^{1/2} = \left[\sum_{i<j}^{D}\left\{log\frac{x_i}{x_j} - log\frac{y_i}{y_j}\right\}^2\right]^{1/2} (x, y \in S^D)$$

where g() is the geometric mean of the components of the composition. The metricΔ satisfies the usual metric axioms:

- *Positivity*: $\Delta(x, y) > 0 (x \neq y)$, $\Delta(x, y) = 0\ (x = y)$
- *Symmetry*: $\Delta(x, y) = \Delta\ (y, x)$
- *Power relationship:* $\Delta(a \otimes x, a \otimes y) = | a | \Delta(x, y)$
- *Triangular inequality:* $\Delta(x, z) + \Delta\ (z, y) \geq \Delta(x, y)$

The fact that this metric has also desirable properties relevant and logically necessary, such as scale, permutation and perturbation invariance and sub-compositional dominance, for meaningful statistical analysis of compositional data is now well established and the relevant properties are recorded briefly here:

- *Permutation invariance*: $\Delta(xP, yP) = \Delta\ (x, y)$, for any permutation matrix *P*.
- *Perturbation invariance*: $\Delta(x \oplus p, y \oplus p) = \Delta(x, y)$, where *p* is any perturbation.
- *Sub-compositional dominance*: if s_x and s_y are similar, say (1, . . . , C)-Sub-compositions of *x* and *y*, then $\Delta_S^C\ (s_x, s_y) \leq \Delta_S^D\ (x,y)$.

6.5 The Log Ratio Methodology

The constant-sum constraint is a mathematical property embedded in any compositional data set, causing problems on the analysis of such a type of data. Aitchison [106, 107, 89] showed that the effects of this constraint on the covariance and correlation matrices disappear, if the raw percentage data are expressed as logarithms of ratios, where the denominator is the geometric mean of the percentages in each sample.

For applying statistical methods designed for the Euclidean geometry on compositional data, as wells as for representing them in the Aitchison geometry on the simplex, some kind of transformations are first necessary. The main idea that leads to such transformations is to find a basis (or a generating system) and to express compositions in coefficients of such a basis (coordinate system). This class of mappings is widely known under the term log ratio transformations. There are three types to be presented in the upcoming paragraphs: a) the additive log ratio transformation (alr) and inverse b) the centered log ratio transformation (clr), and finally, c) the isometric log-ratio transformation (ilr). All of them move the operations of perturbation and power transformation to the usual vector addition and scalar multiplication. However, only the latter two transformations move the whole Aitchison geometry to the Euclidean one, i.e. including the Aitchison inner product. As the proposed transformations are one-to-one transformations, the obtained results are usually back-transformed to the simplex in order to simplify the interpretation.

6.5.1 Additive Log Ratio Transformation (alr)

The additive log ratio (alr) transformation transforms raw compositional data from simplex to real (Euclidean) space. Alr transformation is also capable of performing its inverse

transformation (from real space to simplex) with its inverse *ALR*-1 (Aitchison, 2003). *ALR* differs from other transformations in that it maps a composition in the D-part simplex none isometrically to a D-1, dimensional Euclidean vector. As it maps, the last part is treated as a common denominator to the others, which means that in case the denominator changes, then the *ALR* transformations obtained, would be different. The additive log ratio transformation follows the idea to construct a (non-orthonormal) basis which is very easy to interpret, since the relation to the original D-1 first parts is preserved. Thus, for a composition x, a special case of the additive log ratio (alr) transformation [89] to R^{D-1} , is defined as:

$$alr(x) = \left(ln\frac{x_1}{x_D}, \dots, ln\frac{x_{D-1}}{x_D}\right)'.$$

In this equation, there is a division of each of the first D-1 components by the final component. It is easy to see that also another part can be used as ratio part in the denominator. It is usually chosen in such a way that the interpretation of the result is facilitated. Note that different alr transformations are related by linear transformations (see, e.g., Filzmoser and Hron, 2008).

The inverse transformation ALR^{-1}: $R^{D-1} \rightarrow S^D$ is

$$x = alr^{-1}(x) = C[\exp(y_1), \exp(y_2)\dots\exp(y_{D-1})1]$$

,where C is the closure operation. When data are in their transformed state, they can be analyzed by all those statistical methods not relying on a distance. The drawback of alr transformation is that it is not an isometric transformation from the simplex. It lacks symmetry and orthogonality dew to the use of a common numerator or denominator. This weakness could be solved by use of an appropriate metric with oblique coordinates in real ALR-space, but that is not a standard practice [91].

6.5.2 Centered Log Ratio Transformation (clr)

Taking a generating system on the simplex leads to the centered log ratio (clr) transformation (Aitchison, 1986) to R^D,

$$clr(x) = \left[ln\frac{x}{g(x)} - \dots - ln\frac{x_D}{g(x)}\right]$$

,where g(x) is the geometric mean of the parts involved:

$$g(x) = \left(\prod_{i=1}^{D} x_i\right)^{1/D} = exp\left(\frac{1}{D}\sum_{i=1}^{D} lnx_i\right)$$

,or with the inverse transformation (clr^{-1}), from real space (clr coefficients) to the simplex (raw data) (Aitchison, 1986). The clr coordinates represent a generating system, not a basis, and therefore clr coordinates sum up to zero [108], i.e. we get a constrained transformed vector. As a result, correlations and covariances between clr parts are not sub-compositionally coherent.

6.5.3 Isometric Log Ratio Transformation (ilr)

The calculation of ilr coordinates is more complex and the generation of specific expressions is dominated by different rules. With ilr the data are transformed from the simplex to real space, as ilr coordinates, or conversely applying the inverse ilr^{-1}. Both features are defined by a sequential binary partition [108, 109]. The ilr transformation is defined as:

$$ilr(x)=(y1, y2,\ldots, yD-1) \in R^{D-1},$$

where $y_i = \sum_{j=1}^{D} y_{ij} lnx_j, i = 1, 2, \ldots D-1$ and

$$\psi_{i,j} = \sqrt{\frac{s_i}{r_i(s_i + r_i)}} \; if\ at\ step\ i\ the\ part\ j\ is + 1$$

or

$$\psi_{i,j} = -\sqrt{\frac{s_i}{r_i(s_i + r_i)}} \; if\ at\ step\ i\ the\ part\ j\ is - 1$$

or

$$\psi_{i,j} = 0 \;\; if\ at\ step\ i\ the\ part\ j\ is\ 0$$

with r_i the number of parts at step i as +1, and s_i the number of parts at step i as -1.

The ilr^{-1} transformation is defined as:

$X=ilr^{-1}(y)=(x_1, x_2, \ldots, x_D) \in S^D$, where $[x_1, x_2, \ldots, x_D]=Cexp[z_1, z_2, \ldots, z_D]$, $z_j = \sum_{j=1}^{D-1} \psi_{ij} y_{,}$, C stands for the closure operation [89].

6.6 The CoDaPack v2 Software Package

Over the last years, a new methodological approach has been developed for the statistical analysis of compositional data, based on the approach introduced in the early eighties by John Aitchison. This methodology is not straightforward to use with standard statistical packages. For this reason, in this project, we examine a new freeware software, The Compositional Data Package, which implements at this moment the most elementary of mentioned statistical methods. The features of this new software are very wide:

- Transformations between the real space to the simplex or vice versa such as the alr, clr and ilr transformations.
- Operations inside the simplex like centering, perturbation, power transformation, amalgamation, subcomposition (closure) or rounded zero replacement.
- 2-D and 3D graphical outputs like ternary diagrams, alr plots, clr plots, biplots, plots of principal components.

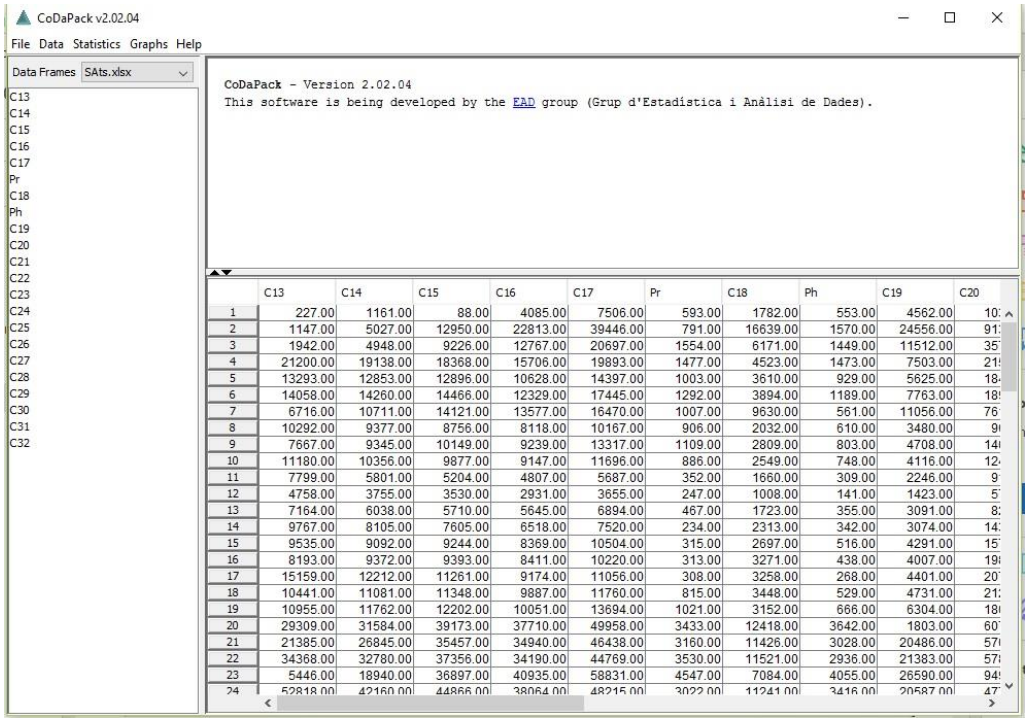
- Compositional Descriptive Statistics.

The software has been developed by members of the Research Group on Compositional Data Analysis at the Dept. Informàtica, Matemàtica Aplicada i Estadística (IMAE-UdG) under the projects Compositional Data Analysis and Related methods (CODA-RETOS) and Compositional and Spatial Data Analysis (COSDA). The core of the group belongs to the University of Girona (UdG), and includes members from the Technical University of Catalonia (UPC), and Biomathematics & Statistics Scotland (BioSS).

6.6.1 Interface of the CoDaPack software

This time the analysis will be conducted only on a small part of the data set, in order to examine briefly, how a different treatment approach would impact on the data. There will be a comparison of the results between the classical statistical analysis and the compositional statistical approach. For this attempt, the Saturate Fraction Compositional Model (SFCM) was selected, and in the next paragraphs there will be a presentation of the interface of the software package.

Data could be imported from Excel files or recovered from previous sessions. The observations are organized in rows and the variables in columns. CoDaPack v2 main window (Fig. 38) has four parts. On the very top there are the menus, on the left the active data frame and the name of its variables. The bigger part is the right side. On top of this part there is the place where alphanumerical results are placed, and on bottom there is the data.



	C13	C14	C15	C16	C17	Pr	C18	Ph	C19	C20
1	227.00	1161.00	88.00	4085.00	7506.00	593.00	1782.00	553.00	4562.00	10
2	1147.00	5027.00	12950.00	22813.00	39446.00	791.00	16639.00	1570.00	24556.00	91
3	1942.00	4948.00	9226.00	12767.00	20697.00	1554.00	6171.00	1449.00	11512.00	35
4	21200.00	19138.00	18368.00	15706.00	19893.00	1477.00	4523.00	1473.00	7503.00	21
5	13293.00	12853.00	12896.00	10628.00	14397.00	1003.00	3610.00	929.00	5625.00	18
6	14058.00	14260.00	14466.00	12329.00	17445.00	1292.00	3894.00	1189.00	7763.00	18
7	6716.00	10711.00	14121.00	13577.00	16470.00	1007.00	9630.00	561.00	11056.00	76
8	10292.00	9377.00	8756.00	8118.00	10167.00	906.00	2032.00	610.00	3480.00	9
9	7667.00	9345.00	10149.00	9239.00	13317.00	1109.00	2809.00	803.00	4708.00	14
10	11180.00	10356.00	9877.00	9147.00	11696.00	886.00	2549.00	748.00	4116.00	12
11	7799.00	5801.00	5204.00	4807.00	5687.00	352.00	1660.00	309.00	2246.00	9
12	4758.00	3755.00	3530.00	2931.00	3655.00	247.00	1008.00	141.00	1423.00	5
13	7164.00	6038.00	5710.00	5645.00	6894.00	467.00	1723.00	355.00	3091.00	8
14	9767.00	8105.00	7605.00	6518.00	7520.00	234.00	2313.00	342.00	3074.00	14
15	9535.00	9092.00	9244.00	8369.00	10504.00	315.00	2697.00	516.00	4291.00	15
16	8193.00	9372.00	9393.00	8411.00	10220.00	313.00	3271.00	438.00	4007.00	19
17	15159.00	12212.00	11261.00	9174.00	11056.00	308.00	3258.00	268.00	4401.00	20
18	10441.00	11081.00	11348.00	9887.00	11760.00	815.00	3448.00	529.00	4731.00	21
19	10955.00	11762.00	12202.00	10051.00	13694.00	1021.00	3152.00	666.00	6304.00	18
20	29309.00	31584.00	39173.00	37710.00	49958.00	3433.00	12418.00	3642.00	1803.00	60
21	21385.00	26845.00	35457.00	34940.00	46438.00	3160.00	11426.00	3028.00	20486.00	57
22	34368.00	32780.00	37356.00	34190.00	44769.00	3530.00	11521.00	2936.00	21383.00	57
23	5446.00	18940.00	36897.00	40935.00	58831.00	4547.00	7084.00	4055.00	26590.00	94
24	52818.00	42160.00	44866.00	38064.00	48215.00	3022.00	11241.00	3416.00	20587.00	47

Fig. 38 CoDaPack v2 main window.

In order to run a CoDaPack routine we first import the data. The software stores a set of data on Data Frames or Tables. It is possible to have opened more than one Data frame. A set of

Data frames could be saved as a Workspace and also it could be recovered by means of the item button Open Workspace (Fig. 39).

Each Data frame contains the name of variables and its numerical values. As far as the missing values are concerned, there are two kinds; non-detected or non-available data and there is a specific symbol to distinguish them. Non-detected data should begin with a character prefix, for example <, followed by the value of low detection limit while Non-Available data should use a symbol, for example "NA".

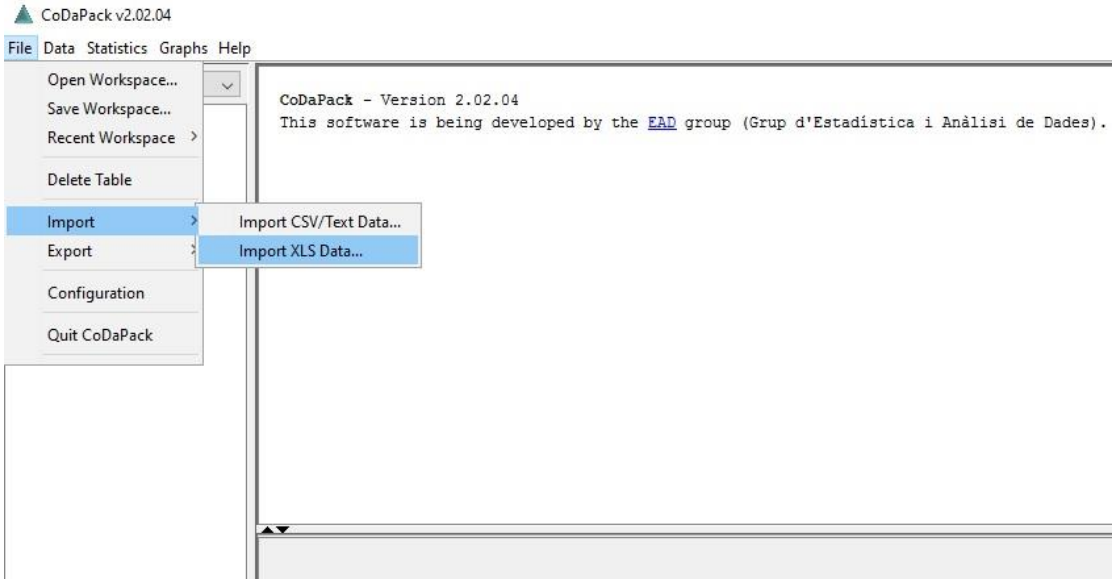


Fig. 39 Menu File

Data frames may be imported and exported from Excel files. After data are imported, (Fig. 40) we must indicate in which row starts the data, if there are labels, non-available symbol and non-detected prefix. At any time, we may can delete a Data Frame from the active workspace. The exportation saves the names of the variables into the first row of an Excel file and the data in rows below variable names.

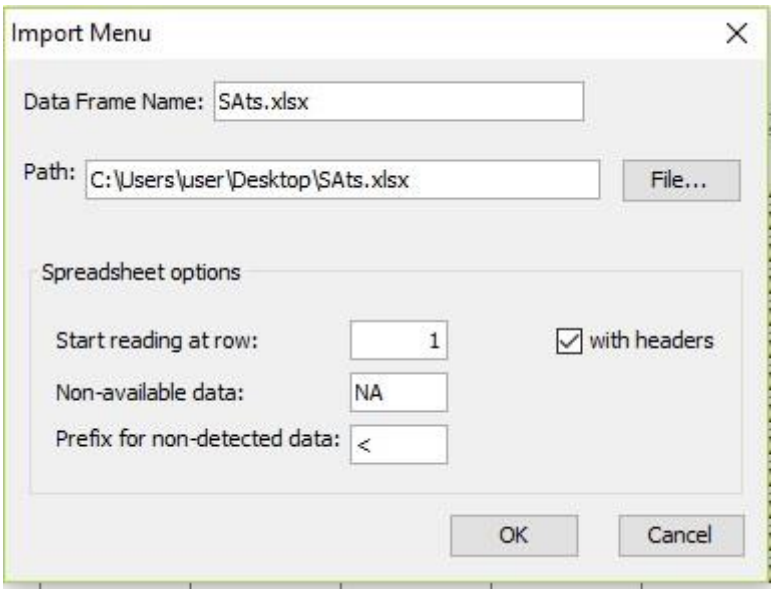


Fig. 40 Importing Data

Another part of the menu which is utilized in this project, is the Data menu (Fig. 41). In general, this menu manages three kinds of routines: 1) transformations of the data from the simplex

to the real space and vice versa, 2) operations inside the simplex and 3) management of variables.

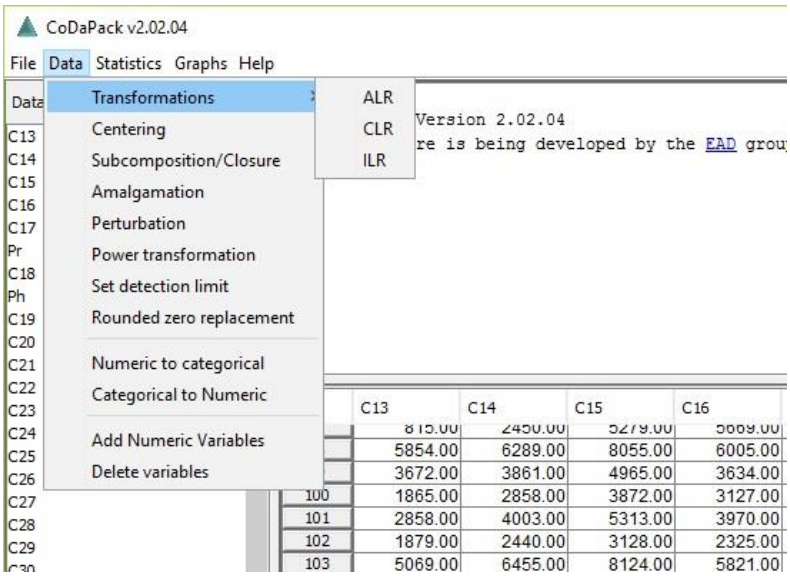


Fig. 41 Menu: Data

The software package offers various options as far as the data analysis is concerned (Fig. 41). Beginning with the Data Menu, *Centering* is a feature with which the data are centered, that is, they are perturbed by the center or closed geometric mean of the data (Fig. 42).

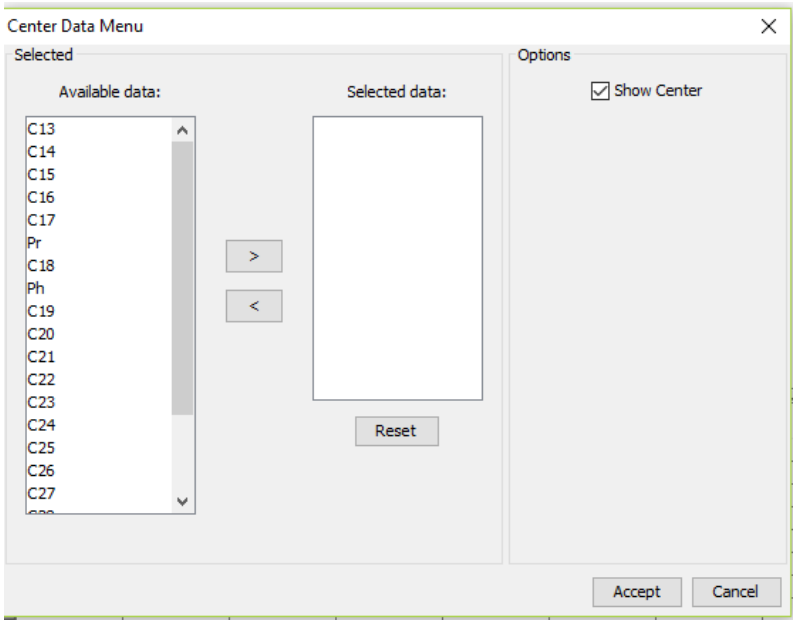


Fig. 42 Data: Centering

This routine centers the data set, that is, it returns the data set Y formed by the D-part compositions $y = gN(X)^{-1} \otimes X$, where

$$gN(X) = C\left[\left(\prod_{k=1}^{N} x_{k1}\right)^{1/N}, \dots, \left(\prod_{k=1}^{N} x_{kD}\right)^{1/N}\right]$$

is the closed geometric mean of the data set X. The center of the set Y is e, the barycenter of the simplex; e.g. for D = 3 the geometric center of a ternary diagram is [0:333; 0:333; 0:333]. If Show Center is activated this routine writes the center of the parts selected on the output window.

The feature *Subcompostion/Closure* the data is closed, i.e. data are converted into parts of some whole summing to a given constant, Y = C(X) : This constant is, by default 1:0 but could be entered by the user by means of the Closure form. If S parts, S < D; are selected, a subcomposition with S-parts is obtained (Fig. 43).

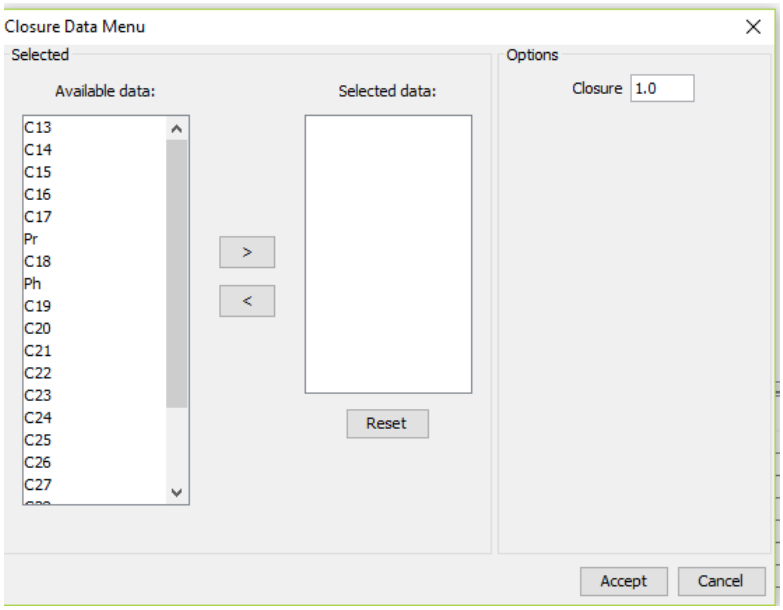


Fig. 43 Data : Subcomposition/Closure

The *Amalgamation* feature amalgamates some columns of the data (Fig. 44). The result of amalgamation of some of the parts of a D-composition selected by the user is the sum of those parts. Amalgamation should be used only as a first step in preparing the data for further analysis, as this operation is non-linear in the Aitchison geometry and might lead to inconsistent results if compared to analysis made without amalgamation.

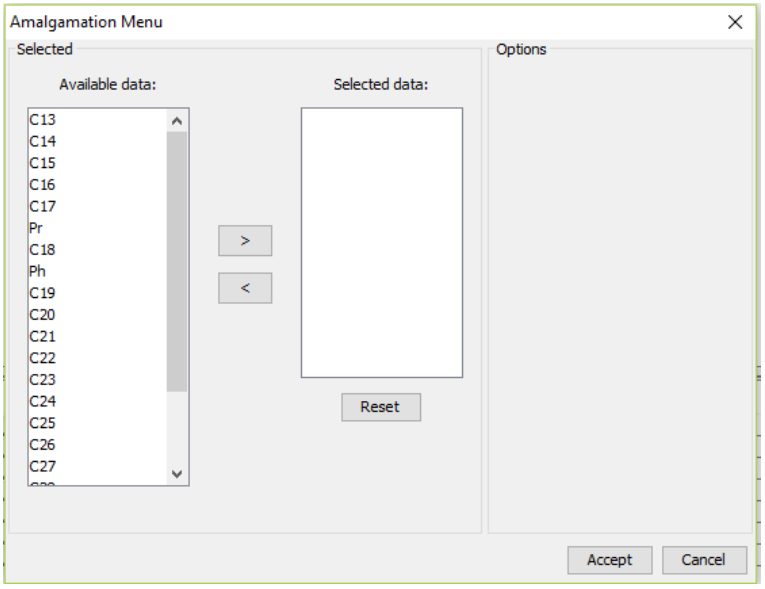


Fig. 44 Data: Amalgamation

With the *Perturbation* feature a vector perturbs the data. The output is a matrix of D-part compositions

$$y = \boldsymbol{p} \quad \boldsymbol{x} = C\left[p_1 x_1 p_2 x_2 ,\ldots, p_D x_D\right],$$

where C stands for the closure operation, and p is a given D-part composition. The user has to indicate on Perturbation box the vector p, which has to be the same length as the compositions x.

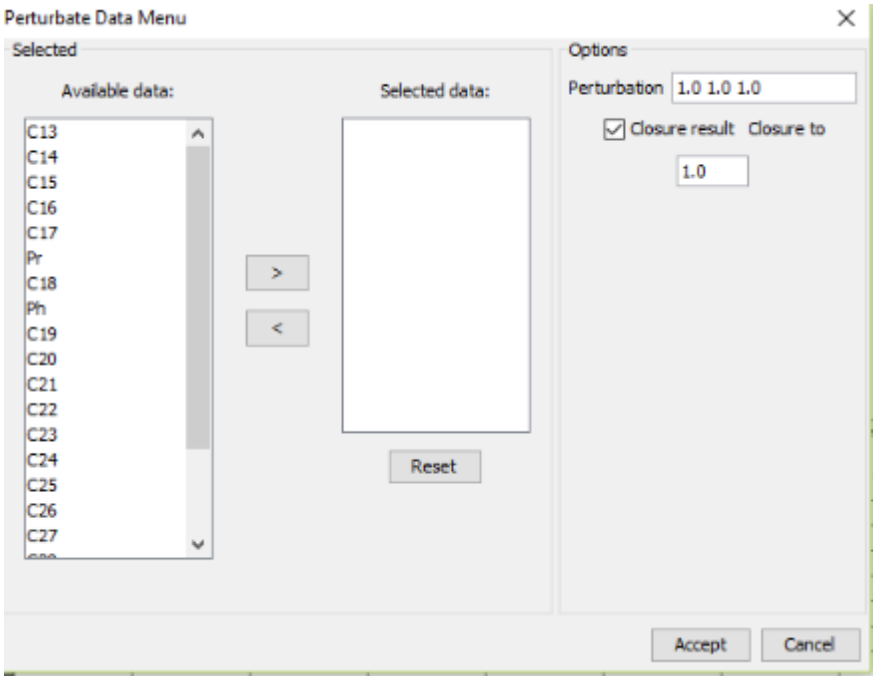


Fig. 45 Data : Perturbation

The *Power Transformation* feature applies a power transformation to the data. For a ∈ R; the power transformation returns

$$a \otimes x = C\ [x_1^a, x_{2,\ldots,}^a x_D^a]$$

In this option, we have to indicate the constant of the operation on the *Power* box.

The *Rounded Zero Replacement* applies a transformation to the data to avoid zeros (Fig. 46). This transformation involves substituting an observation x, with zeros in some parts, by an observation y using the expression:

$$y_i = \begin{Bmatrix} \delta_i, & if\ x_i = 0 \\ x_i\left(1 - \frac{\Sigma x_j = 0^{\delta_j}}{C_x}\right), & if\ x_i > 0 \end{Bmatrix}$$

where δi is the replacement value for the i-th part defined by the user and Cx the components sum of observation x. This routine applies to non-detected data (the software distinguishes between non-available and non-detected data). There is an individual constant δi for each non-detected value, that is stored on the data frame.

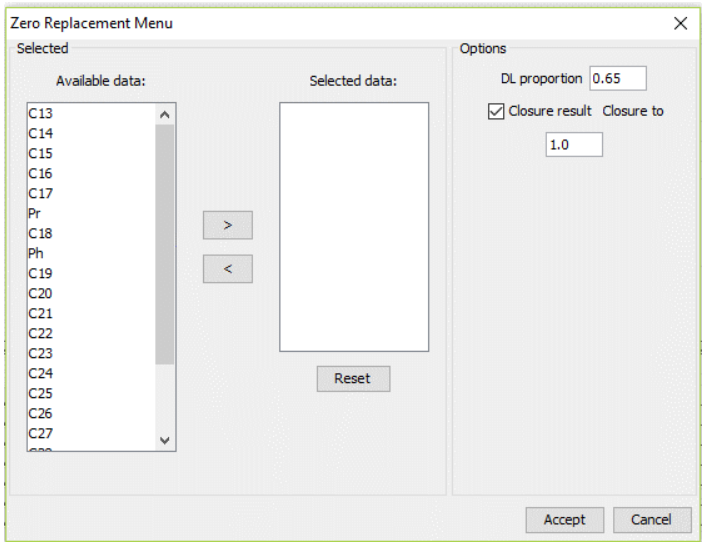


Fig. 46 Data : Rounded Zero Replacement

The *Numeric to Categorical* feature transforms the selected variables into strings and overwrites the results on the same variables.

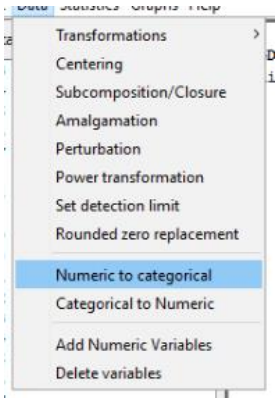


Fig. 47 Data : Numeric to Categorical

The *Numeric to Categorical* feature, on the other hand, transforms the selected variables coded with a string into numerical ones, and overwrites the result on the same variables.

The *Add Numeric Variables feature*, imports date to the data set by a simple copy-paste action (Fig. 48).

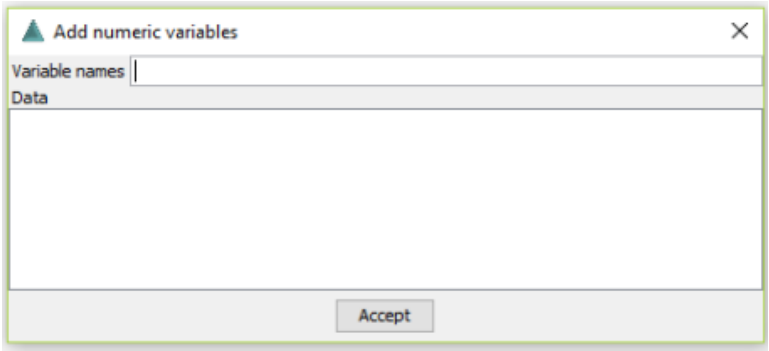


Fig. 48 Data : Add numeric variables

Finally, the *Delete Variables* routine deletes the variables the user selects from the workspace (Fig. 49).

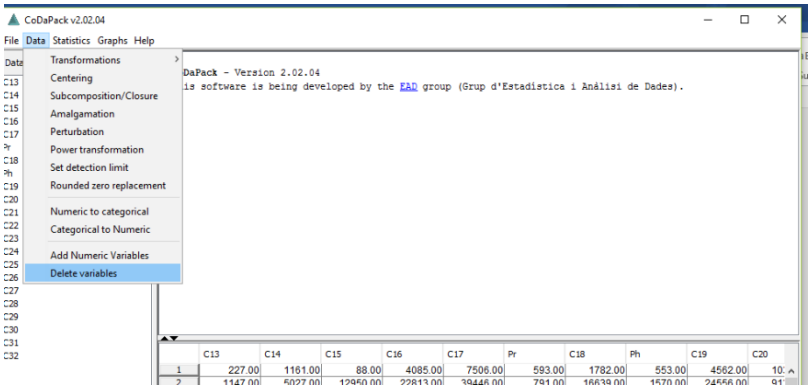


Fig. 49 Data : Delete Variables

The CoDaPack software includes a *Statistics* Menu. The first option is the Compositional Statistics Summary (Fig. 50). This menu produces two types of descriptive statistics: the first related to logratios (Variation Array, CLR variance and Total Variance) and the second related

to compositional descriptive statistics (Centre, Min, Max and quartiles). This routine is utilized and the results are presented in the next chapter.

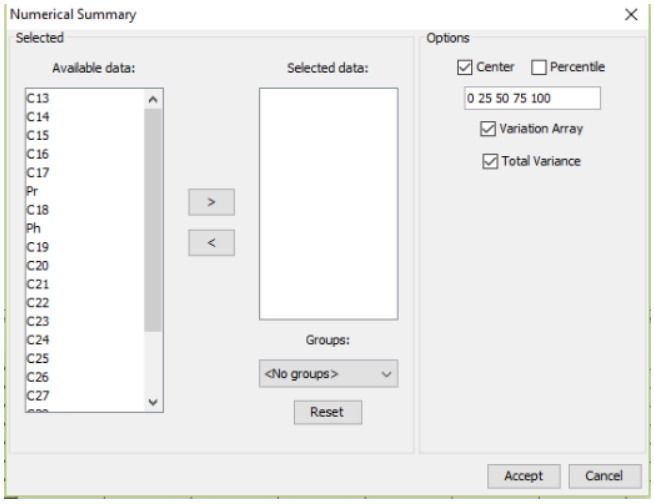


Fig. 50 Statistics : Compositional Statistics Summary

1. Variation Array: Returns a matrix where the upper diagonal contains the logratio variances and the lower diagonal contains the logratio means. That is, the ij-th component of the upper diagonal is var [ln(X_i=X_j)] ; and the ij-th component of the lower diagonal is E[ln(X_i=X_j)], where i, j= 1 ,2, …, D.

2. CLR Variances: Returns, for each part, the sum of logratio variances that involve it. Thus, for the i-th clr component ξi we have

$$var(\xi_i) = \frac{1}{2D} \sum_{i=1, j\neq 1}^{D} var\left[ln\left(X_i/X_j\right)\right].$$

3. Total Variance: The sum of all clr Variances is the Total Variance totvar.

4. Centre: Returns the center of the data set, that is, $\widehat{\xi} = C[g_1 g_2, \dots, g_D]$, where $g_i = \left(\prod_{k=1}^{N} x_{ki}\right)^{1/N}$ stands for the geometric mean of part Xi in data set X. The data set X has been previously closed.

5. Minimum and Maximum: For each part of the data set X it returns the maximum and the minimum of the closed data set.

6. Quartiles: For each part of the data set X it returns the first quartile Q1, the median Q2 and the third quartile Q3 of the closed data set. The user has to select the columns to close and where to put the results. There are two buttons in this routine:

The output of the routine is placed on the output part. It includes a color classification of the logratio variances (elements of the upper diagonal of Variation Array). It is assumed that the logarithm of the logratio variances follow a t-student distribution, then dark blue colores those elements below percentile 5, light blue from percentile 5 to 25, light red form percentiles 75 to 95 and dark red up to percentile 95.

The menu *Classical Statistics Summary* produces standard descriptive statistics, including mean (arithmetic), standard deviation, covariance matrix, Min, Max and quartiles). The output of the routine is placed on the output part.

The *Additive-Logistic normality test* feature allows the user to perform a test for logistic normality of a D-part composition (Fig. 51). It includes all marginal, univariate distributions (with a total of (D - 1) tests); all bivariate angle distributions (with a total of D(D-1)/2 tests); and the (D-1)-dimensional radius distribution. For each kind of test the Anderson-Darling, Cramer-von Misses and Watson statistics are computed and their significance is given.

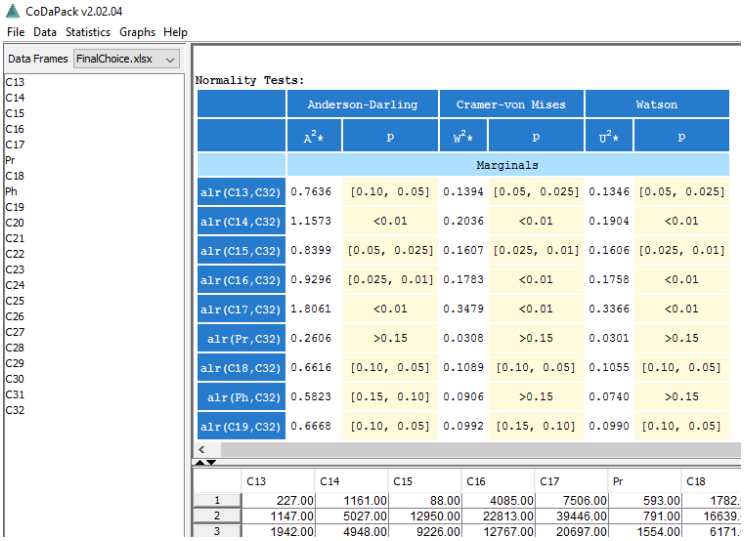


Fig. 51 Statistics: Logistic Normality tests

The *Atypicality Indices* feature obtains the atypical observations and their indices under the assumption of Additive Logistic Normal distribution of the selected parts (Fig. 52). The user has to select the columns to calculate its atypical observations and the threshold of atypicality (usually 0:95) has to be given.

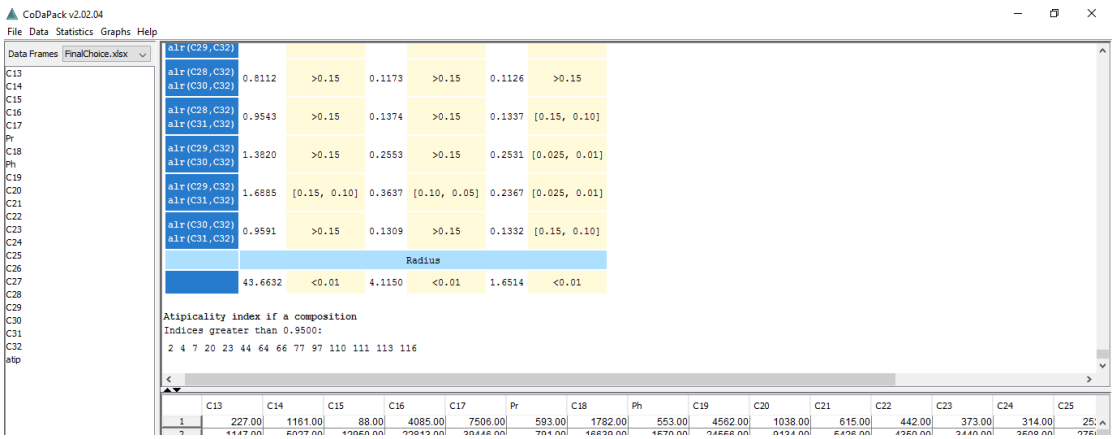


Fig. 52 Statistics : Atypicality indices

The last part in the Menu section is the *Graphs* Section (Fig. 53). The options this software offers, enable the user to create graphs in independent windows. The can customize the appearance of each graph and, in some cases, plot the observations in the graph according to a previous classification. These graphs can be zoomed and, in 3D, rotated.

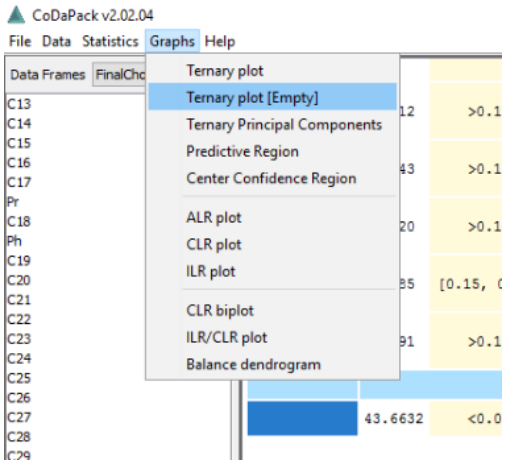


Fig. 53 Graphs Menu

To perform a zoom in a graph it is possible to use the slider scroll at the bottom of the graph or just using the scroll wheel of the mouse. It is also possible to rotate a figure by means of the left button of the mouse. Holding the left mouse button and moving it the graph rotates following the direction of the mouse. If the graph is 2D then the figure just moves inside the windows without rotation. To move the graph inside the window holding the left mouse button and simultaneously holding the ALT key. Furthermore, the graphs can be saved by means of snapshots of what windows have each moment. This can be done with the menu File-Snapshot and the files produced could be in jpeg, eps, png and bitmap formats. The same menu File includes a submenu Configuration that allows to customize the elements of the graph like lines and labels by means of changing size and colors.

The Graphs menu will not be further presented here, as many of the options will be used straight on the data set, and the outcome will be discussed.

6.6.2 Application of the CoDaPack's routine on the Saturates' fraction

To examine how compositional data behave when treated according to Aitchison, only a part of the whole data was used; the Saturates' fraction (see Appendix). Components with zero values were removed from this data set, as they would cause problems to the transformation operations. In particular, samples A549, A1711,A1724, A2268, A2283, A2284, A2468, A2469, B515, B554, B014, B1279, B2121, B2122, C540, C1465 and finally, C1473 were removed.

The first step is to use the *Amalgamation* option. As mentioned before, amalgamation should be applied on the data to prepare them before further analysis. Amalgamation is equal to addition in R. The results are presented in the following table (Table 7).

Table 7 Amalgamation of all variables for each component

	A549	A550	A920	A1140	A1710	A1711	A1712	A1723	A1724	A1725	A2268	A2269	A2270	A2283	A2284	A2313	A2362	A2363	A2364	A2424
amalg	24574.0	164351.0	95894.0	118237.0	84688.0	96204.0	130134.0	58253.0	65913.0	66642.0	39317.0	24519.0	41602.0	53986.0	62386.0	65691.0	79984.0	76814.0	78769.0	239513.0
	A2425	A2426	A2427	A2428	A2429	A2430	A2431	A2432	A2433	A2434	A2435	A2436	A2468	A2469	A2470	A2611	A2627	A2706	A2884	A2892
amalg	229793.0	252557.0	255369.0	288556.0	147814.0	212470.0	152544.0	103844.0	142061.0	202393.0	284439.0	171464.0	94280.0	255283.0	348520.0	51185.0	115613.0	64207.0	137398.0	121527.0
	A2895	A2896	A2897	A2898	B515	B554	B1014	B1279	B1393	B1443	B2121	B2122	B2887	B1873	B1874	C495	C499	C503	C511	C513
amalg	239946.0	141455.0	191406.0	104755.0	63637.0	85889.0	37508.0	24213.0	83605.0	63678.0	18410.0	22286.0	37639.0	39058.0	40661.0	73354.0	80288.0	124331.0	300295.0	200714.0
	C529	C540	C548	C553	C557	C566	C574	C575	C579	C582	C589	C596	C711	C714	C721	C722	C725	C1386	C1387	C1388
amalg	286154.0	74231.0	152070.0	253089.0	268887.0	181418.0	133703.0	233839.0	124065.0	124764.0	49701.0	70429.0	69213.0	95710.0	201581.0	69094.0	53147.0	124722.0	162462.0	224475.0
	C1389	C1390	C1465	C1466	C1467	C1468	C1469	C1470	C1471	C1472	C1473	C1705	C1715	D756	D800	D801	D802	D841	D842	D924
amalg	93596.0	135161.0	33821.0	26273.0	16547.0	20901.0	77669.0	18958.0	22813.0	19675.0	15327.0	101468.0	80685.0	101365.0	164064.0	206299.0	87376.0	89224.0	56791.0	45470.0
	D1173	D1273	D1274	D1275	D1276	D1288	D1289	D1290	D1291	D1312	D1313	D1335	D1364	D1365	D1385	D2471	D2472	D2595	D2626	D2885
amalg	54842.0	33217.0	85176.0	47918.0	49561.0	52781.0	52667.0	41060.0	31760.0	82846.0	66946.0	45664.0	79499.0	117301.0	244337.0	403272.0	93609.0	83301.0	79110.0	139966.0

The next options utilized are the Compositional Statistics Summary and the Classical Statistics Summary (Table 8, Table 9).

Table 8 Compositional Statistics Summary

Compositional NA's:
25
Sample size:
95
Statistics

Center					
C13	**C14**	**C15**	**C16**	**C17**	**Pr**
0.0653	0.0869	0.1131	0.1035	0.1227	0.0265
C18	**Ph**	**C19**	**C20**	**C21**	**C22**
0.0676	0.0305	0.0771	0.0463	0.0375	0.0355
C23	**C24**	**C25**	**C26**	**C27**	**C28**
0.0300	0.0290	0.0266	0.0233	0.0200	0.0168
C29	**C30**	**C31**	**C32**	**C31**	**C32**
Variation array: 0.0145	0.0116	0.0086	0.0072	0.0086	0.0072

Variance ln(Xi/Xj)

Xi\Xj	**C13**	**C14**	**C15**	**C16**	**C17**	**Pr**	**C18**	**Ph**	**C19**	**C20**	**C21**	**C22**	**C23**	**C24**	**C25**	**C26**	**C27**	**C28**	**C29**	**C30**	**C31**	**C32**	**clr variances**	
C13		0.2668	0.2975	0.4900	0.5675	15.260	0.9115	20.073	0.9388	12.839	15.900	14.883	16.916	16.127	16.077	16.570	16.381	17.318	15.840	16.146	15.135	18.519	10.099	
C14	0.2860		0.1872	0.2844	0.3442	11.968	0.5762	15.872	0.5804	0.8622	11.456	10.586	12.674	12.097	12.433	12.637	12.676	13.312	12.123	12.002	11.353	14.289	0.6816	
C15	0.5492	0.2632		0.0518	0.0812	0.8739	0.2740	12.048	0.2701	0.5247	0.7707	0.6867	0.8709	0.8133	0.8331	0.8605	0.8596	0.9301	0.8117	0.8116	0.7528	10.275	0.3700	
C16	0.4608	0.1748	-0.0884		0.0372	0.9882	0.2583	12.921	0.1937	0.4852	0.7392	0.6621	0.8605	0.8143	0.8593	0.8682	0.8736	0.9294	0.8137	0.7888	0.7329	10.108	0.3809	
C17	0.6304	0.3444	0.0812	0.1696		11.495	0.3356	14.719	0.2185	0.5932	0.8582	0.7788	0.9760	0.9261	0.9645	0.9806	0.9811	10.527	0.9213	0.8987	0.8330	11.264	0.4747	
Pr	-0.9007	-11.866	-14.498	-13.615	-15.310		0.4274	0.0982	0.8334	0.4168	0.4618	0.3732	0.4566	0.3901	0.3871	0.4090	0.4426	0.4735	0.4574	0.4543	0.4949	0.5272	0.3265	
C18	0.0353	-0.2507	-0.5139	-0.4255	-0.5951	0.9360		0.5457	0.1672	0.0679	0.2097	0.1376	0.2684	0.2278	0.2878	0.2702	0.2959	0.3095	0.2770	0.2563	0.2596	0.4260	0.0516	
Ph	-0.7596	-10.456	-13.088	-12.205	-13.900	0.1410	-0.7950		10.365	0.4776	0.4970	0.3924	0.4639	0.3790	0.3856	0.3866	0.4261	0.4401	0.4556	0.4437	0.5201	0.5050	0.4256	
C19	0.1660	-0.1200	-0.3831	-0.2948	-0.4644	10.667	0.1307	0.9257		0.2502	0.4559	0.3874	0.5627	0.5349	0.6184	0.5950	0.6189	0.6455	0.5700	0.5345	0.5160	0.7796	0.2570	
C20	-0.3448	-0.6308	-0.8940	-0.8056	-0.9752	0.5558	-0.3801	0.4148	-0.5109		0.1028	0.0385	0.1521	0.1318	0.2198	0.1831	0.2251	0.2148	0.2254	0.1878	0.2187	0.3730	0.0719	
C21	-0.5536	-0.8396	-11.028	-10.144	-11.840	0.3471	-0.5889	0.2060	-0.7196	-0.2088		0.0698	0.0337	0.1423	0.2067	0.1793	0.2167	0.2159	0.2446	0.2074	0.2532	0.3730	0.1509	
C22	-0.6104	-0.8964	-11.596	-10.712	-12.408	0.2902	-0.6457	0.1492	-0.7765	-0.2656	-0.0568		0.0699	0.0444	0.1110	0.0794	0.1145	0.1138	0.1347	0.1079	0.1522	0.2674	0.0734	
C23	-0.7775	-10.635	-13.267	-12.383	-14.079	0.1231	-0.8128	-0.0179	-0.9436	-0.4327	-0.2239	-0.1671		0.0706	0.1153	0.0865	0.1093	0.1201	0.1534	0.1427	0.1841	0.2731	0.1489	
C24	-0.8121	-10.981	-13.613	-12.729	-14.425	0.0886	-0.8474	-0.0524	-0.9781	-0.4673	-0.2585	-0.2017	-0.0346		0.0437	0.0122	0.0282	0.0406	0.0665	0.0742	0.1147	0.1903	0.1006	
C25	-0.8986	-11.846	-14.478	-13.594	-15.290	0.0021	-0.9339	-0.1390	-10.646	-0.5538	-0.3450	-0.2882	-0.1211	-0.0865		0.0409	0.0469	0.0761	0.0853	0.1047	0.1423	0.2210	0.1339	
C26	-10.294	-13.154	-15.786	-14.902	-16.598	-0.1288	-10.647	-0.2698	-11.955	-0.6846	-0.4758	-0.4190	-0.2519	-0.2173	-0.1308		0.0079	0.0186	0.0351	0.0513	0.0843	0.1559	0.1169	
C27	-11.851	-14.711	-17.343	-16.459	-18.155	-0.2844	-12.204	-0.4254	-13.511	-0.8403	-0.6315	-0.5747	-0.4076	-0.3730	-0.2865	-0.1557		0.0157	0.0238	0.0492	0.0712	0.1421	0.1273	
C28	-13.571	-16.431	-19.063	-18.179	-19.875	-0.4564	-13.924	-0.5974	-15.231	-10.123	-0.8035	-0.7467	-0.5796	-0.5450	-0.4585	-0.3276	-0.1720		0.0221	0.0424	0.0631	0.1343	0.1485	
C29	-15.070	-17.930	-20.562	-19.678	-21.374	-0.6063	-15.423	-0.7473	-16.730	-11.622	-0.9534	-0.8966	-0.7295	-0.6949	-0.6084	-0.4776	-0.3219	-0.1499		0.0297	0.0334	0.1249	0.1195	
C30	-17.317	-20.176	-22.808	-21.925	-23.620	-0.8310	-17.670	-0.9720	-18.977	-13.868	-11.781	-11.212	-0.9541	-0.9196	-0.8331	-0.7022	-0.5466	-0.3746	-0.2247		0.0322	0.1415	0.1145	
C31	-20.328	-23.188	-25.820	-24.936	-26.632	-11.321	-20.681	-12.731	-21.988	-16.880	-14.792	-14.224	-12.553	-12.207	-11.342	-10.034	-0.8477	-0.6757	-0.5258	-0.3011		0.1263	0.1173	
C32	-22.047	-24.906	-27.538	-26.655	-28.350	-13.040	-22.400	-14.450	-23.707	-18.598	-16.511	-15.942	-14.271	-13.926	-13.061	-11.752	-10.196	-0.8476	-0.6977	-0.4730	-0.1719		0.2524	
	Mean ln(Xi/Xj)																						56.537	**Total Variance**

The Menu Compositional Statistics Summary, as mentioned before, includes two types of descriptive statistics. On Table 8 we observe the Variation Array, CLR variance and Total Variance as well as the Center, Min, Max and quartiles. The sample size is 95 in this case, due to the fact that in the data set, there exist zero values. The inadvertency introduced by a log-ratio variance (here clr variance) is that the logarithm of zeros does not exist, so if there are such observations in the data

On the other hand, the Menu Classical Statistics Summary includes the arithmetic mean, standard deviation, covariance matrix, Min, Max and quartiles (Table 9). The first step in analyzing multivariate data is computing the mean vector and the variance-covariance matrix. The mean vector consists of the means of each variable and the variance-covariance matrix consists of the variances of the variables along the main diagonal and the covariances between each pair of variables in the other matrix positions. The variance and the standard deviation are important in data analysis because of their relationships to correlation and the

normal curve. Correlation between a pair of variables measures to what extent their values co-vary. The term covariance is undoubtedly associatively prompted immediately. There are numerous models for describing the behavioral nature of a simultaneous change in values, such as linear, exponential and more. Observing Table 9, it is evident that all variables are correlated positively. The strongest positive correlation forms between C14 and C13 (0.9434). What is interesting here, is that in the classical statistics summary, the sample size remains at each original form of 120 samples. This is contrast to the Compositional Statistics summary, where sample size reduces, due to the exclusion of zero values.

Table 9 Classical Statistics Summary

Clasical statistics summary:

NA's: 0

Sample size: 120

Statistics

	Mean	Std.Dev	0	25	50	75	100
C13	82.765.083	88.639.087	0.0000	28.380.000	54.460.000	104.410.000	528.180.000
C14	97.541.417	84.559.280	0.0000	42.280.000	76.800.000	133.040.000	421.600.000
C15	121.133.500	101.799.595	0.0000	53.530.000	92.440.000	144.660.000	461.170.000
C16	113.407.583	99.571.357	7.800.000	46.660.000	81.820.000	135.770.000	449.950.000
C17	142.907.667	135.388.242	14.280.000	56.110.000	94.930.000	180.720.000	592.610.000
Pr	31.463.083	32.061.154	2.340.000	11.090.000	21.010.000	41.090.000	232.300.000
C18	69.584.167	51.639.032	10.080.000	30.240.000	56.290.000	96.550.000	226.860.000
Ph	39.796.000	45.425.844	1.410.000	11.890.000	24.180.000	53.310.000	332.640.000
C19	84.766.167	70.033.547	9.580.000	29.900.000	59.420.000	114.730.000	275.920.000
C20	50.742.083	46.542.324	0.0000	18.190.000	39.300.000	61.150.000	239.940.000
C21	42.336.167	40.446.243	1.070.000	15.320.000	30.870.000	52.330.000	208.640.000
C22	39.475.083	38.984.371	2.720.000	14.510.000	27.610.000	49.230.000	206.220.000
C23	34.644.500	34.646.541	1.030.000	12.480.000	24.740.000	45.240.000	177.670.000
C24	32.257.667	32.915.131	2.770.000	10.930.000	21.540.000	40.650.000	184.340.000
C25	29.748.917	28.974.933	2.290.000	9.380.000	21.880.000	39.380.000	176.900.000
C26	25.860.417	26.808.898	1.040.000	8.090.000	18.480.000	33.200.000	158.100.000
C27	22.054.167	22.940.668	1.750.000	7.380.000	15.070.000	28.450.000	143.090.000
C28	18.874.667	19.582.024	1.440.000	6.410.000	12.350.000	23.310.000	97.930.000
C29	15.620.417	15.579.848	0.0000	5.260.000	11.190.000	19.470.000	79.610.000
C30	13.196.333	13.438.383	730.000	4.390.000	9.790.000	16.330.000	71.050.000
C31	9.331.583	9.881.062	0.0000	3.360.000	6.390.000	11.380.000	52.020.000
C32	9.408.833	18.428.670	0.0000	2.730.000	5.430.000	9.000.000	184.820.000

Correlation:

	C13	C14	C15	C16	C17	Pr	C18	Ph	C19	C20	C21	C22	C23	C24	C25	C26	C27	C28	C29	C30	C31	C32
C13	10.000	0.9434	0.8505	0.7750	0.7316	0.1084	0.3769	0.0310	0.5192	0.1427	0.0897	0.0689	0.0309	0.0319	0.0439	0.0321	0.0402	0.0024	0.0281	0.0054	0.0418	0.0148
C14	0.9434	10.000	0.9604	0.9007	0.8646	0.2542	0.5544	0.1844	0.6985	0.3212	0.2635	0.2370	0.1840	0.1874	0.1920	0.1798	0.1789	0.1345	0.1618	0.1328	0.1713	0.1180
C15	0.8505	0.9604	10.000	0.9599	0.9485	0.3125	0.6306	0.2415	0.7903	0.3870	0.3303	0.3023	0.2688	0.2505	0.2820	0.2426	0.2433	0.1980	0.2303	0.2289	0.2413	0.1689
C16	0.7750	0.9007	0.9599	10.000	0.9789	0.1943	0.6044	0.1533	0.8272	0.3911	0.3277	0.2994	0.2445	0.2416	0.2455	0.2340	0.2323	0.2016	0.2363	0.2166	0.2640	0.1388
C17	0.7316	0.8646	0.9485	0.9789	10.000	0.1779	0.5905	0.1267	0.8324	0.3590	0.2900	0.2620	0.2186	0.1995	0.2183	0.1905	0.1885	0.1584	0.1922	0.1913	0.2250	0.1168
Pr	0.1084	0.2542	0.3125	0.1943	0.1779	10.000	0.6878	0.9485	0.4244	0.6665	0.7037	0.6977	0.7359	0.7382	0.7940	0.7458	0.7500	0.6854	0.6873	0.6799	0.6575	0.7282
C18	0.3769	0.5544	0.6306	0.6044	0.5905	0.6878	10.000	0.7424	0.8479	0.9215	0.8922	0.8728	0.8494	0.8214	0.8012	0.7920	0.7658	0.7739	0.7466	0.7647	0.7445	0.5363
Ph	0.0310	0.1844	0.2415	0.1533	0.1267	0.9485	0.7424	10.000	0.4478	0.7715	0.8127	0.8167	0.8470	0.8572	0.8836	0.8576	0.8525	0.8048	0.7838	0.7748	0.7580	0.7944
C19	0.5192	0.6985	0.7903	0.8272	0.8324	0.4244	0.8479	0.4478	10.000	0.7398	0.6854	0.6606	0.6024	0.5925	0.5589	0.5687	0.5472	0.5455	0.5475	0.5295	0.5701	0.3413
C20	0.1427	0.3212	0.3870	0.3911	0.3590	0.6665	0.9215	0.7715	0.7398	10.000	0.9888	0.9802	0.9314	0.9259	0.8545	0.8921	0.8542	0.8896	0.8478	0.8336	0.8486	0.5529
C21	0.0897	0.2635	0.3303	0.3277	0.2900	0.7037	0.8922	0.8127	0.6854	0.9888	10.000	0.9965	0.9649	0.9660	0.9071	0.9386	0.9079	0.9307	0.8921	0.8750	0.8843	0.6171
C22	0.0689	0.2370	0.3023	0.2994	0.2620	0.6977	0.8728	0.8167	0.6606	0.9802	0.9965	10.000	0.9714	0.9763	0.9203	0.9527	0.9238	0.9460	0.9071	0.8887	0.9002	0.6351
C23	0.0309	0.1840	0.2688	0.2445	0.2186	0.7359	0.8494	0.8470	0.6024	0.9314	0.9649	0.9714	10.000	0.9760	0.9675	0.9599	0.9443	0.9514	0.9130	0.9400	0.9014	0.7147
C24	0.0319	0.1874	0.2505	0.2416	0.1995	0.7382	0.8214	0.8572	0.5925	0.9259	0.9660	0.9763	0.9760	10.000	0.9687	0.9918	0.9787	0.9736	0.9458	0.9131	0.9245	0.7296
C25	0.0439	0.1920	0.2820	0.2455	0.2183	0.7940	0.8012	0.8836	0.5589	0.8545	0.9071	0.9203	0.9675	0.9687	10.000	0.9735	0.9760	0.9500	0.9359	0.9367	0.9076	0.7775
C26	0.0321	0.1798	0.2426	0.2340	0.1905	0.7458	0.7920	0.8576	0.5687	0.8921	0.9386	0.9527	0.9599	0.9918	0.9735	10.000	0.9950	0.9843	0.9711	0.9245	0.9484	0.7566
C27	0.0402	0.1789	0.2433	0.2323	0.1885	0.7500	0.7658	0.8525	0.5472	0.8542	0.9079	0.9238	0.9443	0.9787	0.9760	0.9950	10.000	0.9802	0.9751	0.9280	0.9513	0.7832
C28	0.0024	0.1345	0.1980	0.2016	0.1584	0.6854	0.7739	0.8048	0.5455	0.8896	0.9307	0.9460	0.9514	0.9736	0.9500	0.9843	0.9802	10.000	0.9860	0.9512	0.9663	0.7007
C29	0.0281	0.1618	0.2303	0.2363	0.1922	0.6873	0.7466	0.7838	0.5475	0.8478	0.8921	0.9071	0.9130	0.9458	0.9359	0.9711	0.9751	0.9860	10.000	0.9438	0.9764	0.6980
C30	0.0054	0.1328	0.2289	0.2166	0.1913	0.6799	0.7647	0.7748	0.5295	0.8336	0.8750	0.8887	0.9400	0.9131	0.9367	0.9245	0.9280	0.9512	0.9438	10.000	0.9364	0.6922
C31	0.0418	0.1713	0.2413	0.2640	0.2250	0.6575	0.7445	0.7580	0.5701	0.8486	0.8843	0.9002	0.9014	0.9245	0.9076	0.9484	0.9513	0.9663	0.9764	0.9364	10.000	0.7392
C32	0.0148	0.1180	0.1689	0.1388	0.1168	0.7282	0.5363	0.7944	0.3413	0.5529	0.6171	0.6351	0.7147	0.7296	0.7775	0.7566	0.7832	0.7007	0.6980	0.6922	0.7392	10.000

The CoDaPack softaware offers the option of a Ternary Principal Component Graph. This feature calculates the two (or three) compositional principal components for a 3-part (or 4-part) composition and displays the result in a ternary diagram. What is more, it returns, as a numerical result, the Principal Components and the cumulative proportion explained with each component.

It would be meaningful if the Principal Components of this 3-part composition, is based on variables that present a bigger weight, related to the rest variables of the data set. For this reason, the command ***var(X)*** was utilized on Matlab to examine which are the three variables with the most impact. It was found that C13, C14 and phytane obtain the biggest weights (0.1034, 0.0796 and 0.924 respectively). Below we present both the ternary principal

component graph for all components produced by CoDaPack, as well as the PC plot that is produced by matlab (using raw compositional data).

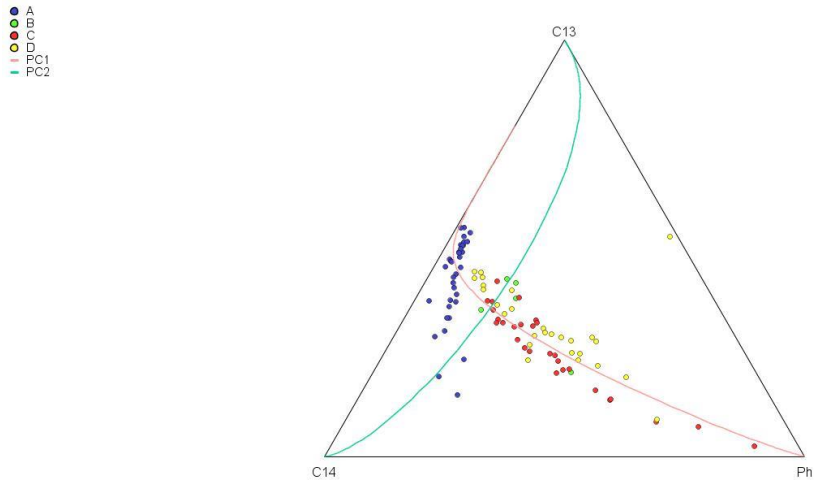


Fig. 54 Ternary Principal Component Graph for C13, C14 and phytane.

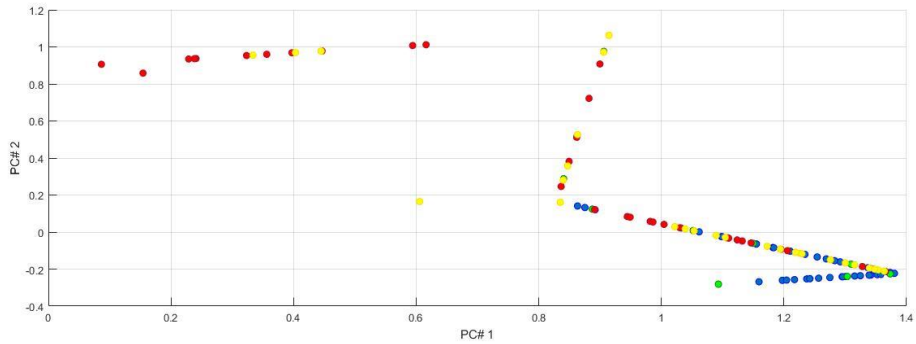


Fig. 55 Plot of the first two Principal Components for C13, C14 and phytane.

By examining the ternary principal component graph (Fig. 54), we observe that Family A oils are distinctively separated from the rest, presenting a sub-parallel alignment to the first principal component axis (PC1). Samples from families B, C and D follow a linear trend along the PC2 axis overlapping each other. In Table 10 we observe the numerical representation of the principal components for each variable, as well as, the cumulative proportion explained with each PC. Both PC1 and PC2 are positively correlated to the three variables. C13 is the most important in explaining PC1, whereas C14 is the most important in explaining PC2

Table 10 Principal Components as Numerical results and the Cumulative proportions explained with each principal component.

	C13	C14	Ph	Cum. Prop. Exp.
PC1	0.4715	0.3999	0.1287	0.9074
PC2	0.1465	0.5970	0.2565	1.0000

On the other hand, Fig. 55 displays a completely different principal component analysis result. As far as the discrimination of the four family affiliations is concerned, it is evident that there is no clear distinction among them. All samples follow strictly linear gradients, overlapping

significantly, at the same time. PC1 scores for all samples are positively high, whereas for PC2, the majority obtains negative scores. The first Principal Component in this case explains 81% of the total variance, and PC2 follows with 15% of the total variance.

A simple ternary plot of C13, C14 and phytane is displayed on Fig. 56. As in the ternary principal component graph, in this plot there is a significant overlapping among oils B, C and D. Family A oils form a quite distinct group along the C13-C14 axis. Along the C13-phytane axis there is a sample (number) which displays a different behavior from the rest and it is D1338. Fig. 57 displays the centered version of the same plot. It offers a better understanding of how oil samples exist in the ternary plot's space.

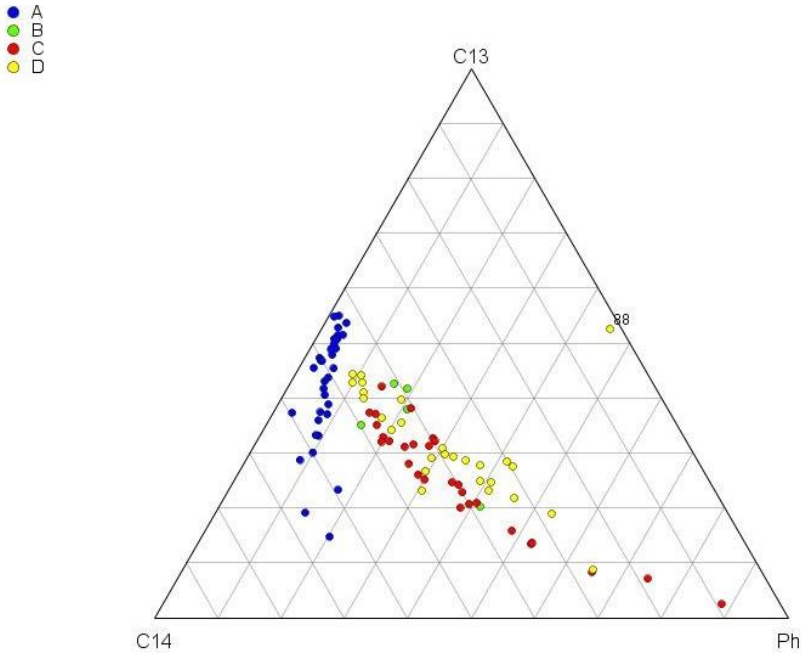


Fig. 56 Ternary Plot of C13, C14 and phytane

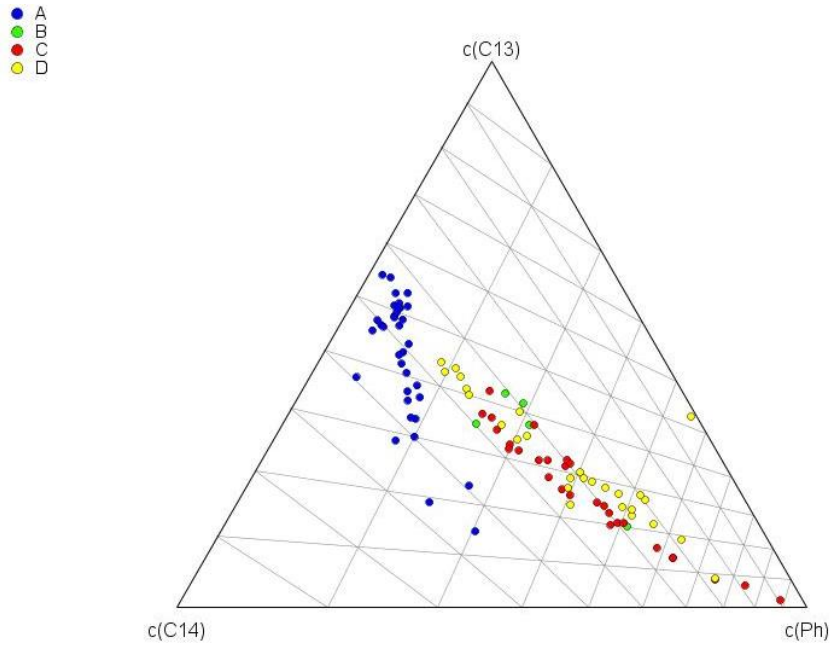


Fig. 57 Centered ternary plot with grid on

The ALR plot represents a plot of three (four in 3D) alr-transformed parts (Fig. 58). The new variables obtained with the ALR transformation are displayed in an orthogonal coordinate system to visualize how the plot changes when permuting the components or initial columns. Nevertheless, care is required when interpreting the plot, as the axis are not really orthogonal, but at 60°.

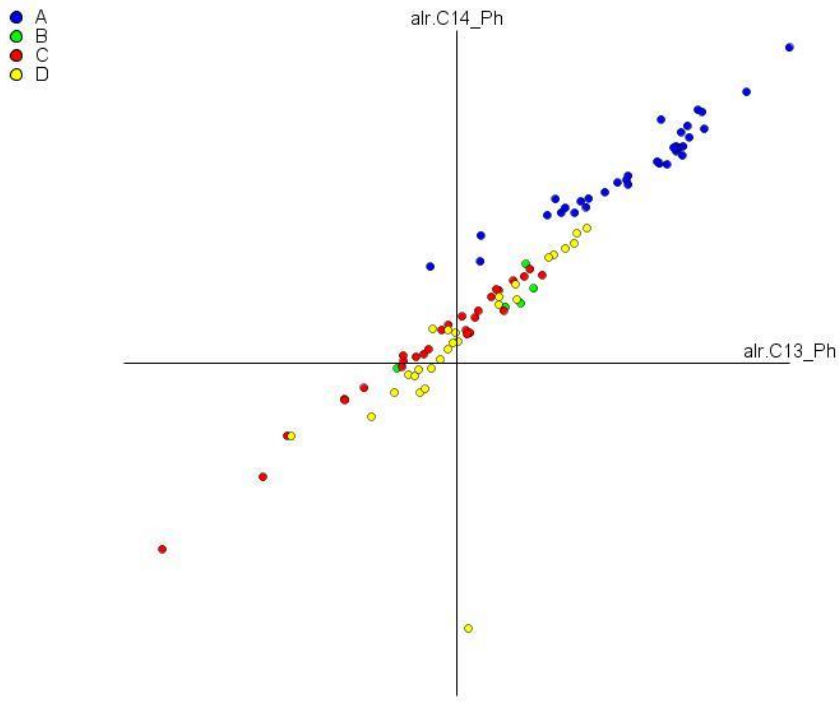


Fig. 58 ALR plot of C13, C14 and phytane

What is observed in the ALR plot, is that oil samples form a positive gradient of 30° along the intersection of alr.C14_Ph and alr.C13_Ph axes. The additive logratio transformation seems to reveal a linearity embodied in oil families. Once more, the most distinct group is that of family A oils. The overlapping still holds among the other oil families.

The CLR plot feature represents a plot in an orthogonal coordinate system of the data, after the centred logratio transformation (clr) of two (three in 3D) selected parts. It has the same capabilities as the ALR Plot.

The ILR plot feature displays a plot in an orthogonal coordinate system of the data after the isometric logratio transformation (ilr) of three (four in 3D) selected parts according to a sequential binary partition. The way to select the partition is the same as in Transformation-ILR routine. The partition selected in our case is the default (Table 11).

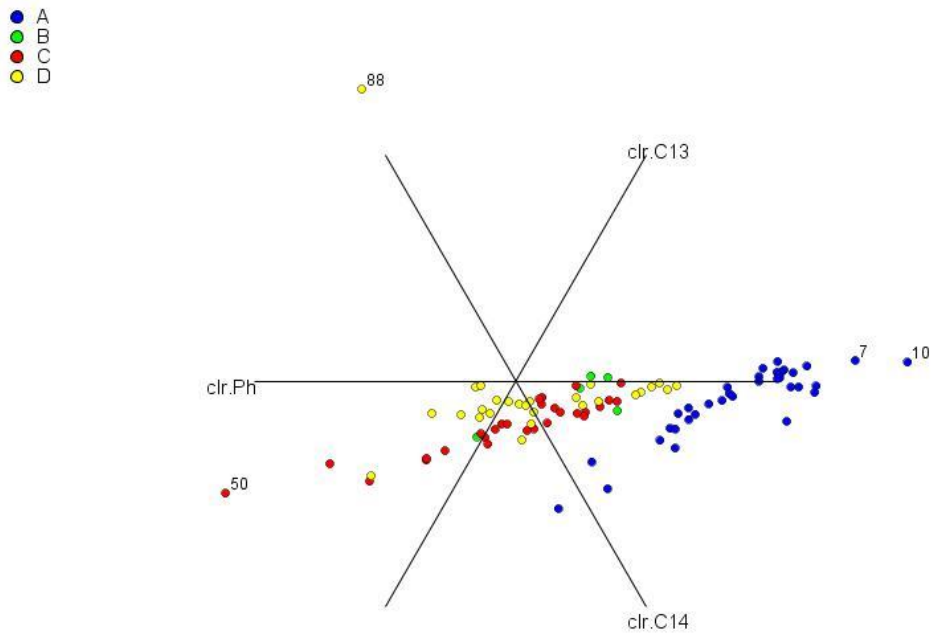


Fig. 59 CLR plot of C13, C14 and phytane

Table 11 Binary partition for ILR transformation

C13	C14	Ph
1	1	-1
1	-1	0

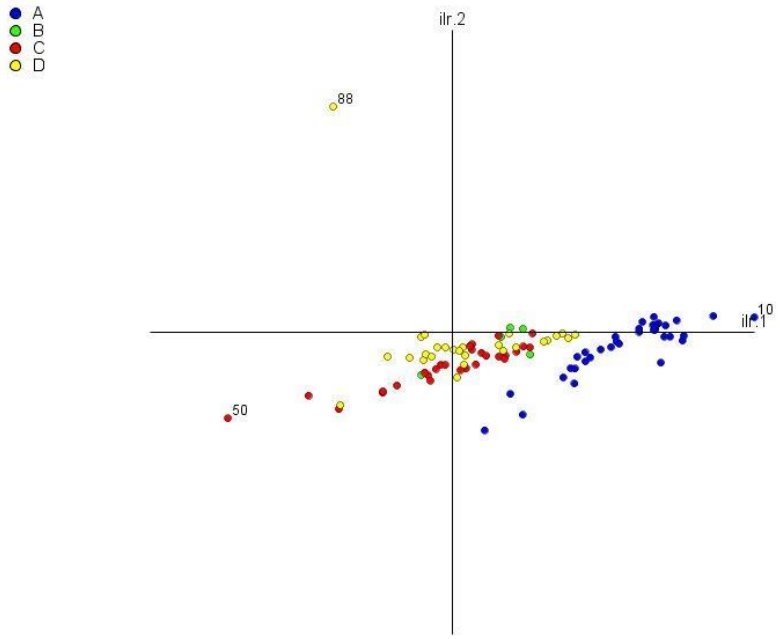


Fig. 60 ILR plot of C13, C14 and phytane

In the ILR plot there are two distinct positive gradients sub-parallel to and below the ilr.1 axis (Fig. 60). One of the two gradients, consists of oil samples solely from Family A and the other consists of oil samples from families B, C and D. The projections of sample points of family A oils do not overlap with any of the other, in contrast to the rest that overlap significantly.

The CLR biplot includes the selected variables C13, C14 and phytane. Once the graph is performed, we may choose 1) which 2D view we prefer (axes XY, YZ or XZ), 2) to display observations or not, and 3) which biplot display depending on the Form value; α = 0 corresponds to a Covariance Biplot, α = 1 Form Biplot, and α = 0:5 Symmetric Scaling Biplot, which is the default value. In Fig. 61 the biplot is a Form Biplot.

What is more this routine returns, as a numerical result, the Principal Components and the cumulative proportion explained with each component (Table 12). Biplot consists on the decomposition of clr matrix, X = UDV'. If numerical output is desired the routine writes three matrices: UD, D and V. UD are the ilr coordinates of the original data.

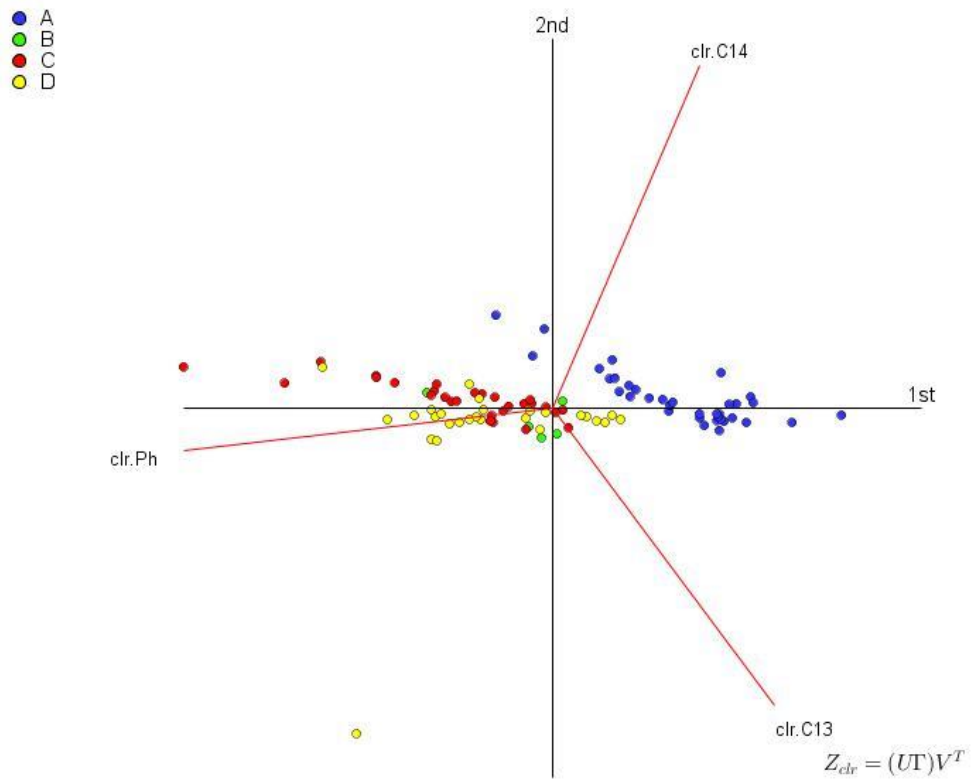


Fig. 61 CLR biplot of C13, C14 and phytane

As far as the distinction of the families is concerned, more or less, the CLR Biplot presents the same results, as in the previous graphs.

Table 12 Principal Components explained by clr.13, clr.14 and phytane

	clr.C13	clr.C14	clr.Ph	Cum.Prop.Exp.
PC1	0.4878	0.3231	-0.8109	0.9074
PC2	-0.6548	0.7498	-0.0951	1.0000

Table 12 displays with which variable each principal component is explained along with the cumulative proportion explained. PC1 is positively correlated with clr.C13 and clr.C14, but negatively with clr.ph. PC2 is negatively correlated with clr.C13 and clr.ph, but positively with clr.C14.

Lastly, the Balance Dendrogram represents a dendrogram by means of a sequential binary partition of selected parts (Fig. 62). The way to select the partition is the same as in Transformation-ILR routine. Here the default partition is chosen (Table 14). As a numerical output, this routine returns on the output window the sequential binary partition used, the mean and the variance of each balance (Table 13). Also on the Data window are the ilr coordinates produced with this partition.

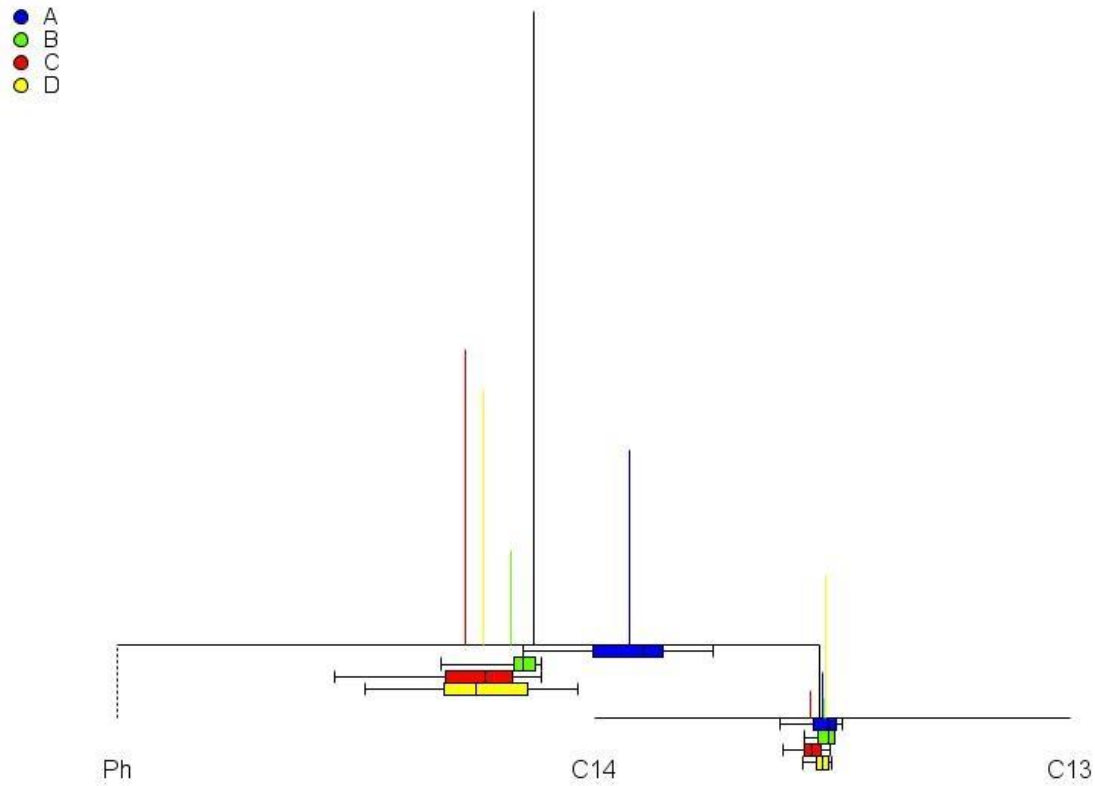


Fig. 62 Balance dendrogram of C13, C14 and phytane

Table 13 Numerical output of Balance Dendrogram routine, including the mean and variance

	Mean		Variance	
	Balance 1	**Balance 2**	**Balance 1**	**Balance 2**
	0.7370	-0.2022	1.1537	0.1334

Table 14 Default partition for the Balance Dendrogram routine

C13	C14	Ph
1	1	-1
1	-1	0

7. Conclusions

The aim of this project has been the examination of the way multivariate clustering methods perform on the classification of oil family affiliations. The methods implemented hereby include Hierarchical clustering, k-means clustering and Principal Component Analysis. The data set under study contained raw compositional information of four distinct oil families present at Williston Basin of Canada. For the needs of the study, four different models were developed out of the given geochemical information; the Saturates' Fraction Compositional Model, the Saturates' Fraction Ratios Model, the Gasoline Range Compositional Model. Focus was not placed on how the models would perform under the aforementioned statistical analysis, but the exact opposite. The effort was on the examination of the data set through a manifold manner.

Taking into consideration the performance of each method separately we conclude as follows:

- Hierarchical Clustering performed relatively well on all models. Family A oils were classified sufficiently and in some cases Family C oils appeared to form fair clusters. However, there was always considerable overlapping among families B, C and D.
- k-means failed in the task of classifying the given data set into distinct groups. In the SFCM and SFRM, it produced a two-cluster solution, with one cluster including mainly samples from Family A, and another cluster containing the rest. Judging, however, from the k-means plots, the clusters produced, did not present clear boundaries between them. In the GRCM and BCM, k-means produced a three-cluster solution, but significant overlapping among all families was observed. This was also evident from the respective k-means plots.
- Principal Component Analysis performed similarly to hierarchical clustering. It mainly distinguished Family A samples and presented significant overlapping among the rest oil samples. In BCB especially, there was an overlapping between families A and D, as well as with families C and D. Family B oil samples were dispersed in the plot.

All in all, the geochemical information under study, contains complex compositions of different oils. A blind application of multivariate data analysis methods on such data seems to be unable to classify them into distinct groups. Compositional data require probably different approaches concerning their analysis. Their special properties cause problems when analyzed with standard multivariate methods and a whole new chapter has been introduced by the scientific community on the way to examine them. The final chapter of this project deals with an alternative approach towards the analysis of compositional data, and results are compared to previous approaches. Principal Component Analysis in particular, presents a completely different picture when approached in a different manner. Further investigation, however, should be conducted on this type of data in order to understand their behavior and obtain meaningful information through their analysis.

References

[1] L. C. Gerhard, S. B. Anderson, J. A. LeFever and C. G. Carlson, "Geological development, origin and energy mineral resources of Willinston Basin, North Dakota,," *Bulletin of American Association of Petroleum Geologists,* vol. 66, pp. 989 - 1020, 1982.

[2] J. L. Ahern and S. R. Mrkvicka, "A mechanical and thermal," *Tectonics,* vol. 3, pp. 79-102, 1984.

[3] L. Sloss, "Comparative anatomy of cratonic unconformities. In: Schlee, J.S. (Ed.), Interregional Unconformities," *American Association of,* vol. 36, pp. 1-6, 1984.

[4] Mossop, G.D. and Shetsen, I. (Eds.),, "Geological Atlas of the Western Canada Sedimentary Basin," *Canadian Society of Petroleum Geologists and Alberta Research Council,* 1994.

[5] Podruski, J. A., Barclay, J. E., Hamblin, A. P., Lee, P. J., Osadetz, K. G., Procter, R. M. and Taylor, G. C., "Conventional Oil Resources of Western Canada (Light and Medium), Part 1: Resource Endowment," *Geological Survey of Canada, Paper 87-26,* p. 149, 1988.

[6] Clement, J. H., "Cedar Creek: a significant paleotectonic feature of the Williston Basin. In: Longman, M.W. (Ed.), Williston Basin: Anatomy of a Cratonic Oil Province," *Rocky Mountain Association of Geologists,* pp. 323-336, 1987.

[7] LeFever, J.A., LeFever, R.D. and Anderson, S.B.,, "Structural evolution of the central and southern portions of the Nesson Anticline, North Dakota. In: Carlson, C.G., Christopher, J.E. (Eds.)," in *Proceedings of the Fifth International Williston Basin Symposium*, 1987.

[8] A. O. Lawrence , R. Pollastro and S. B. Gaswirth, "Williston Basin Province—Stratigraphic and Structural Framework to a Geologic Assessment of Undiscovered," in *U.S. Geological Survey Williston Basin Province Assessment Team, Assessment of undiscovered oil and gas resources of the Williston Basin Province of North Dakota, Montana, and South Dakota,*, U.S. Geological SurveyDigital Data Series 69–W, 17p., 2013.

[9] G. C. Bond and M. A. Kominz, "Construction of tectonic subsidence curves for the early Paleozoic miogeocline, southern Candian Rocky Mountains: implications for subsidence mechanisms, age of breakup; and crustal thinning," *Geological Soviety of Amreica Bulletin,* vol. 95, pp. 155-173, 1984.

[10] R. D. LeFever, "Sedimentology and stratigraphy of the Deadwood-Winnipeg interval (Cambro-Ordovician), Williston basin," *Paleozoic systems of the Rocky Mountain region: Rocky Mountain Section SEPM,* pp. 11-28, 1996.

[11] P. R. Vail, R. M. Mitchum and S. Thompson, "Seismic stratigraphy and global changes of sea level, Part 3:," in *Seismic stratigraphy—Application to hydrocarbon*, C. E.

Payton, Ed., American Association of Petroleum Geologists Memoir 26, 1977, pp. 63-81.

[12] R. W. Eddie , "Mississippian sedimentation and oil fields in southeastern Saskatchewan," *Bulletin,* vol. 42, pp. 94-126, 1958.

[13] C. A. Sandberg, R. C. Gutschick , J. G. Johnson, F. G. Poole and W. J. Sando, "Middle Devonian to late Mississippian history of the overthrust belt region, western United States," *Geologic Studies of Cordilleran Thrust Belt ,* vol. 2, pp. 691-719, 1983.

[14] L. C. Gerhard, S. B. Anderson and D. W. Fischer, "Petroleum geology of the Williston Basin," in *Interior cratonic basins: American Association of Petroleum Geologists Memoir 51*, 1990, pp. 507-557.

[15] W. G. Dow, "Application of oil-correlation and source-rock data to exploration in Williston Basin," *AAPG Bulletin,* vol. 58, pp. 1253-1262, 1974.

[16] J. A. Williams, "Characterization of oil types in Williston Basin," *AAPG Bulletin,* vol. 58, pp. 1243-1252, 1974.

[17] C. G. Carlson, "Triassic-Jurassic of Alberta, Saskatchewan, Manitoba, Montana, and North Dakota," vol. 52, pp. 1969-1983, 1968.

[18] T. P. Poulton, "The Jurassic of the Canadian Western Interior, from 49°N latitude to Beaufort Sea," vol. Memoir 9, D. F. Stott and D. J. Glass, Eds., Canadian Society of Petroleum Geologists, 1984, pp. 15-41.

[19] D. F. Stott, "Cretaceous Sequences of foothills of the Canadian Rocky Mountains;," in *The Mesozoic of Middle North America*, Canadian Society of Petroleum Geologists, Memoir, 1984, pp. 67-105.

[20] J. E. Christopher , "The Lower Cretaceous Mannville Group, northern Williston Basin region, Canada," vol. Memoir 9, pp. 109-126, 1984a.

[21] J. E. Christopher , "Depositional patterns and oil field trends in the Lower Mesozoic of the northern Williston Basin, Canada," *Oil and Gas in Saskatchewan,* vol. 7, pp. 83-102, 1984b.

[22] A. G. Green, W. Weber and Z. Hajnal, "Evolution of Proterozoic terranes beneath the Williston Basin," *Geology,* vol. 13, pp. 624-628, 1985b.

[23] C. Burret and R. Berry, "Proterozoic Australia–Western United States (AUSWUS) fit between Laurentia and Australia," *Geology,* vol. 28, no. 2, pp. 103-106, 2000.

[24] A. G. Green, Z. Hajnal and W. Weber, "An evolutionary model of the western Churchill Province and western margin of the Superior Province in Canada and the north-central United States," *Tectonophysics ,* vol. 116, pp. 281-332, 1985a.

[25] K. Nelson, D. Baird, J. Walters, M. Hauck, L. Brown, J. Oliver, J. Ahern, Z. Hajnal, A. Jones and L. Sloss, "Trans-Hudson orogen and Williston Basin in Montana and North Dakota-New COCORP deep profiling results," *Geology,* vol. 21, pp. 447-450, 1993.

[26] G. E. Thomas, "Lineament-block tectonics—Williston-Black Creek Basin," vol. 58, no. 7, pp. 1305-1322, 1974.

[27] L. O. Anna, "Geologic framework of the ground water system in Jurassic and Cretaceous rocks in the Northern Great Plains, in parts of Montana, North Dakota, South Dakota, and Wyoming," Vols. 1402-B, p. 36, 1986.

[28] D. L. Brown and D. L. Brown, "Wrench-style deformation and paleostructural influence on sedimentation in and around a cratonic basin," *Williston Basin: Anatomy of a cratonic oil province,* pp. 57-70, 1987.

[29] E. K. Maughan and W. J. J. Perry, "Lineaments and their tectonic implication in the Rocky Mountains and adjacent plains region," vol. 41, pp. 41-53, 1986.

[30] J. E. Zumberge , "Tricyclic diterpane distributions in the correlation of Palezoic crude oils from the Wlliston Basin," in *Advances in Organic Geochemistry*, New York, John Wiley, 1981, pp. 738-745.

[31] M. J. Leenheer and J. E. Zumberge, "Correlation and thermal maturity of Williston Basin crude oils and Bakken source rocks using terpane biomarkers," in *Williston Basin: Anatomy of a cratonic oil province*, Rocky Mountain Association of Geologists, 1987, pp. 287-298.

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

[33] K. G. Osadetz and L. R. Snowdon, "Significant Paleozoic petroleum source rocks in the Canadian Williston Basin: their distribution, richness, and thermal maturity (Southeastern Saskatchewan and Southwestern Manitoba)," *Geological Survey of Canada, Bulletin ,* vol. 487, p. 60, 1995.

[34] K. G. Osadetz, P. W. Brooks and L. R. Snowdon , "Oil families and their sources in Canadian Williston Basin, (southeastern Saskathewan and sothwestern Manitoba)," *Bulletin of Canadian Petroleum Geology,* vol. 40, no. 3, pp. 254-273`, September 1992.

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

[36] K. G. Osadetz, N. Pasadakis and M. Obermajer, "Definition and characterization of petroleum compositional families using principal component analysis of gasoline and saturate fraction compositional ratios," *Summary of Investigations 2002,* vol. 1, 2002.

[37] L. R. Snowdon and K. G. Osadetz, "Geological processes interpreted from gasoline range analysis of oils from southeast Saskatcewan and Manitoba," *Current Research,* 1988.

[38] N. Pasadakis, M. Obermajer and K. G. Osadetz, "Definition and characterization of petroleum compositional families in Williston Basin, North America using principal component analysis," *Organic Geochemistry,* vol. 35, pp. 453-468, 2004.

[39] M. Obermajer , K. G. Osadetz, M. G. Fowler and L. R. Snowdon, "Light hydrocarbon (gasoline range) parameter refinement of biomarker-based oil-oil correlation studies: an example from Williston Basin," *Organic Geochemistry,* vol. 31, pp. 959-976, 2000.

[40] J. Burrus , K. G. Osadetz, S. Wolf, B. Doligez, K. Visser and D. Dearborn, "A two-dimensional regional basin model of Williston Basin hydrocarbon systems," *APPG Bulletin ,* vol. 80, pp. 265-291, 1996a.

[41] J. Burrus, K. Osadetz, S. Wolf and K. Visser , "Physical and numerical modelling constraints on oil expulsion and accumulation in the Bakken and Lodgepole petroleum systems of the Williston Basin (Canada-USA)," *Bulletin of Canadian Petroleum Geology ,* vol. 44, pp. 429-445, 1996b.

[42] D. M. Jarvie and R. F. Inden, "Re-Evaluation of Williston Basin potential Paleozoic source rocks and petroleum systems," *AAPG Annual Meeting Expanded Abstracts,* vol. 6, p. 55, 6-9 April 1997.

[43] D. M. Jarvie and P. R. Walker , "Correlation of oils and source rocks in the Williston Basin using classical correlation tools and thermal extraction high resolution C7 gas chromatography," *8th International Meeting on Organic, Oral presentation,* 1997.

[44] D. M. Jarvie, "Williston Basin petroleum systems: inferences from oil geochemistry and geology," *The Mountain Geologist,* vol. 38, pp. 39-41, 2001.

[45] C. Jiang, M. Li, K. Osadetz, L. R. Snowdon, M. Obermajer and M. G. Fowler, "Bakken/Madison petroleum systems in the Canadian Williston Basin; Part 2: molecular markers diagnostic of Bakken and Lodgepole source rocks:," *Organic Geochemistry,* vol. 32, pp. 1037-1054, 2001.

[46] C. Jiang and M. Li, "Bakken/Madison petroleum systems in the Canadian Williston Basin; Part 3: geochemical evidence for significant Bakken-derived oils in Madison Group reservoirs," *Organic Geochemistry,* vol. 33, pp. 761-787, 2002.

[47] J. W. Tukey, Exploratory Data Analysis, Addison-Wesley Publishing Co., 1977.

[48] A. Gelman, "Exploratory Data Analysis for Complex Models," *Journal of Computational and Graphical Statistics,* vol. 13, no. 4, p. 755–779, 2004.

[49] A. C. Rencher, Methods of Multivariate Analysis, 2nd ed., John Wiley & Sons, Inc., 2002.

[50] A. K. a. D. R. C. Jain, Algorithms for Clustering Data, New Jersey: Prentice Hall, Englewood Cliffs, NJ, 1988.

[51] A. K. M. M. N. a. F. P. J. Jain, "Data Clustering: A Review," *ACM Computing Surveys,* vol. 31, pp. 264-323, 1999.

[52] F. J. J. a. P. J. Husson, "Principal Component Methods-hierarchcal clustering-partitional clustering: why would we need to chose for visualizing data?," *Agrocampus Quest,* 2010.

[53] Y.-Y. Chi, "Multivariate Methods," *WIREs Computational Statistics,* pp. 35-47, 2012.

[54] K. S. a. L. S. Y. Fu, "A Clustering procedure for syntactic patterns," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 7, pp. 734-742, 1977.

[55] R. R. a. M. C. D. Sokal, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin,* vol. 38, pp. 1403-1438, 1958.

[56] M. R. Anderberg, Cluster Analysis for Applications., New York: Academic Press, 1973.

[57] J. H. Ward, "Hierarchical Groupings to optimize an objective function," *Journal of the American Statistical Asscoiation,* vol. 58, pp. 236-2244, 1963.

[58] L. L. McQuitty, "Similarity analysis, by reciprocal pairs for discrete and continuous data," *Educational and Psychological Measurement,* vol. 27, pp. 21-46, 1966.

[59] M. Jambu, Classification Automatique pour l'Analyse des Donnees, vol. 1, Paris: Dunod, 1978.

[60] J. Podani, "New Combinatorial SAHN Clustering Methods," *Vegetatio,* vol. 81, pp. 61-77, 1989.

[61] G. N. W. W. T. Lance, "Note on a new information-statistic classificatory," *Computer Journal,* vol. 11, no. 2, p. 195, 1968.

[62] J. B. McQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1967.

[63] J. Hartigan, Clustering algortihms, Toronto: John Wiley & Sons, 1975.

[64] J. Hartigan and M. Wong, "Algorithm AS136: A k-means clustering algorithm," *Applied Statistics,* vol. 28, no. 1, pp. 100-108, 1979.

[65] Z. S. Selim and M. A. Ismail, "k-means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality," *IEEE Transactions on Pattern Analysis and Machine Intelligence ,* Vols. PAMI-6, no. 1, pp. 81-87, 1984.

[66] D. Pollard , "Strong Consistency of KK-means Clustering," *The Annals of Statistics,* vol. 9, no. 1, pp. 135-140, 1981.

[67] D. Pollard, "A Central Limit Theorem for k-means Clustering," *The Annals of Probability,* vol. 10, no. 4, pp. 919-926, 1982.

[68] R. Serinko and G. Babu, "Weak limit theorems for univariate k-means clustering under a non-regular condition," *Journal of Multivariate Analysis,* vol. 41, pp. 273-296, 1992.

[69] L. García-Escudero and A. Gordaliza, "Robustness properties of k-means and trimmed k-means," *Journal of the American Statistical Association,* vol. 94, no. 447, pp. 956-969, 1999.

[70] G. Babu and M. Murty, "A near-optimal initial seed value selection in k-means algorithm using a genetic algorithm," *Pattern Recognition Letters,* vol. 14, no. 10, pp. 763-769, 1993.

[71] P. Bradley and U. Fayyad, "Refining initial points for k-means clustering.," in *The fifteenth international conference on machine learning*, San Francisco, 1998.

[72] J. M. Peña , J. A. Lozano and P. Larrañaga , "An empirical comparison of four initialization methods for the K-means algorithm," *Pattern Recognition Letter,* vol. 20, pp. 1027-1040, 1999.

[73] L. Kaufman and P. Rousseeuw, "Finding Groups in Data-An introduction to Cluster Analysis," in *Wiley Series in Probability and Mathematical Statistics*, New York, John Wiley & Sons, Inc., 1990.

[74] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for k-means clustering," *Pattern Recognition Letters,* vol. 25, no. 11, pp. 1293-1302, 2004.

[75] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine,* vol. 2, no. 11, pp. 559-572, 1901.

[76] H. Hotelling, "Analysis of a Complex of Statistical Variables Into Principal Components," *Journal of Educational Psychology,* vol. 24, pp. 417-441 and 498-520, 1933.

[77] S. Wold, K. Esbensen and P. Geladi, "Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems,* vol. 2, pp. 37-52, 1987.

[78] W. Lawton and E. Sylvestre, "Self modelling curve resolution," *Technometrics,* vol. 13, pp. 617-633, 1971.

[79] W. Full, R. Ehrlich and J. Klovan, "Extended Q model-Objective definitionof external end members in the analysis of mixtures," *Journal of Mathematical Geology,* vol. 13, pp. 331-334, 1981.

[80] K. Peters, C. Walters and J. Moldowan, The Biomarker Guide: Biomarkers and Isotopes in Petroleum Systems and Earth History, 2 ed., vol. 2, New York: Cambodge University Press, 2005.

[81] J. Hunt, Petroleum Geochemistry and Geology, 2 ed., New York: Freeman and Company, 1996, p. 743.

[82] G. Lijmbach, "On the origin of petroleum," in *Applied science publishers*, London, 1975.

[83] L. B. Magoon and W. G. Dow, "The petroleum system," in *The Petroleum System-From Source to Trap*, vol. 60, L. B. Magoon and W. G. Dow, Eds., American Association of Petroleum Geologists Memoir, 1994, pp. 3-24.

[84] J. A. Curiale, "Correlation of oils and source rocks-a conceptual and historical prospective," in *The Petroleum System-From Source to Trap*, vol. 60, L. B. Magoon and W. G. Dow, Eds., American Association of Petroleum Geologists Memoir, 1994, pp. 251-260.

[85] K. Pearson, "Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs," in *The Royal Society of London*, 1897.

[86] O. V. Sarmanov and A. B. Vistelious, "On the correlation of percentage values," in *Doklady Akademii Nauk SSSR*, 1959.

[87] C. Krumbein, "Open and closed number systems: stratigraphic mapping," *Bulletin of American Association of Petroleum Geologists,* vol. 46, pp. 322-37, 1962.

[88] J. C. Butler, "The effect of closure on the measure of similarity between samples," *Journal of Mathematical Geology,* vol. 11, pp. 73-84, 1979.

[89] J. Aitchison, The Statistical Analysis of Compositional Data, London: Chapman and Hall, 1986.

[90] N. M. S. Rock, "Numerical Geology. Lecture Notes in Earth Sciences," vol. 18, 1988.

[91] J. Aitchison and J. J. Egozcue, "Compositional data analysis: where are we and where should we be heading?," *Journal of Mathematical Geology,* vol. 37, no. 7, pp. 829-850, 2005.

[92] V. Pawlowsky-Glahn and J. J. Egozcue, "Compositional data and their analysis: an introduction," *The Geological Society of London, Special Publications,* vol. 264, pp. 1-10, 2006.

[93] D. McAlister, "The law of the geometric mean," in *Royal Society of London*, 1879.

[94] D. F. Watson and G. M. Philip, "Measures of variability for geological data," *Journal of Mathematical Geology,* vol. 21, pp. 233-54, 1989.

[95] J. Aitchison, "Letter to the Editor. Comment on "Measures of Variability for Geological Data"by D. F. Watson and G. M. Philip," *Journal of Mathematical Geology,* vol. 22, pp. 223-6, 1990a.

[96] J. Aitchison, "Relative variation diagrams for describing patterns of variability of compositional data," *Journal of Mathematical Geology,* vol. 22, pp. 487-512, 1990b.

[97] D. F. Watson, "Reply to Comment on "Measures of variability for geological data" by D. F. Watson and G. M. Philip," *Journal of Mathematical Geology,* vol. 22, pp. 227-31, 1990.

[98] D. F. Watson , "Reply to "Delusions of uniqueness and ineluctability" by J. Aitchison," *Journal of Mathematical Geology,* vol. 23, p. 279, 1991.

[99] J. Aitchison, "Letter to the Editor. Delusions of uniqueness and ineluctability.," *Journal of Mathematical Geology,* vol. 23, pp. 275-277, 1991a.

[100] J. AItchison, "A plea for precision in Mathematical Geology," *Journal of Mathematical Geology ,* vol. 23, pp. 1081-1084, 1991b.

[101] A. Woronow, "The elusive benefits of logratios," in *IAMG97, The Third Annual Conference of the International Association for Mathematical Geology*, Barcelona, 1997a.

[102] J. Aitchison , "The one-hour course in compositional data analysis or compositional data analysis is easy," in *Third Annual Conference of the International Association for Mathematical Geology*, Barcelona, 1997.

[103] U. Rehder and S. Zier, "Comment on “Logratio analysis and compositional distance by Aitchison et al. (2000)”," *Journal og Mathematical Geology,* vol. 32, 2001.

[104] C. Barceló-Vidal, J. A. Martín-Fernandez and V. Pawlowsky-Glahn, "Mathematical foundations of compositional data analysis," in *IAMG01*, 2001.

[105] D. Billheimer, P. Guttorp and W. Fagan, " Statistical interpretation of species composition," *Journal of the American Statistical Association,* vol. 96, no. 456, pp. 1205-1214, 2001.

[106] J. Aitchison, "A new approach to null correlations of proportions," *Journal of Mathematical Geology,* vol. 13, pp. 175-189, 1981a.

[107] J. Aitchison, "Distributions on the simplex for the analysis of neutrality," *Statistical distributions in scientific work,* vol. 4, pp. 147-156, 1981b.

[108] J. J. Egozcue and V. Pawlowsky-Glahn, "Groups of parts and their balances in compositional data analysis," *Journal of Mathematical Geology,* vol. 37, no. 7, pp. 795-828, 2005.

[109] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras and C. Barceló-Vidal, "Isometric logratio transformations for compositional data analysis," *Journal of Mathematical Geology,* vol. 35, no. 3, pp. 297-300, 2003.

[110] G. Demaison and R. J. Murris, "Petroleum Geochemistry and Basin Evaluation," vol. 35, p. 426, 1984.

[111] Sokal, R. R. and Michener, C. D., "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin,* vol. 38, pp. 1403-1438, 1958.

[112] J. Aitchison, "Some distribution theory related to the analysis of the subjective performance inferential tasks," *Statistical distributions in Scientific Work,* vol. 5, pp. 363-385, 1981c.

[113] J. Aitchison, "The statistical analysis of compositional data (with discussion)," *Journal of the Royal Statistical Society: Series B,* vol. 44, pp. 139-177, 1982.

[114] J. Aitchison, "The triangle in statistics," in *The Art of Statistical Science. A Tribute to G. S. Watson*, V. M. K., Ed., New York, Wiley, 1992a, pp. 89-104.

[115] J. Aitchison, "On criteria for measures of compositional differences," *Journal of Mathematical Geological,* vol. 24, pp. 365-380, 1992b.

[116] J. Aitchison, "Logratios and natural laws in compositional data analysis," *Journal of Mathematical Geology ,* vol. 31, pp. 563-89, 1999a.

[117] J. Aitchison, C. Barcelo-Vidal, J. A. Martin-Fernandez and V. Pawlowsky-Glahn, "Logratio analysis and compositional distance," *Journal of Mathematical Geology,* vol. 32, pp. 271-275, 2000.

[118] J. Aitchison, C. Barceló-Vidal and V. Plawlowsky-Glahn, "Reply to Letter to the Editor by S. Rehder and U. Zier on 'Logratio analysis and compositional distance' by J.Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández and V. Pawlowsky-Glahn," *Journal of Mathematical Geology,* vol. 33, 2001.

[119] A. Woronow, "Regression and discrimination analysis using raw compositional data - is it really a problem?," in *IAMG97, The Third Annual Conference of the International Association for Mathematical Geology*, Barcelona, 1997b.

APPENDIX

Below we present the data set under study. The next tables include all raw data concerning the Biomarkers, the Gasoline range and the Saturated fraction. All models that were examined by multivariate statistical were derived from these three parts of the data set.

Biomarkers (Hopanes and Steranes) of the sample set

	A549	A550	A920	A1140	A1710	A1711	A1712	A1723	A1724	A1725	A2268	A2269	A2270	A2283	A2284	A2313	A2362	A2363	A2364	A2424
C21tri	0	0	0	0	5970000	0	0	0	0	0	5450000	4610000	5660000	5860000	8030000	1660000	2720000	2660000	2350000	0
C23tri	228000	351000	0	0	6780000	0	0	0	0	0	4560000	4040000	5300000	5810000	6330000	1500000	1930000	2650000	2610000	0
Ts	1739000	2310000	3690000	5412000	26000000	5525000	158000	4008000	5385000	3689000	21100000	20700000	23100000	22100000	35800000	4800000	7970000	12600000	9580000	15700000
Tm	3484000	2700000	6290000	10050000	17300000	8793000	161000	8605000	10870000	7074000	19600000	21700000	37100000	10200000	45900000	2380000	3650000	8100000	16700000	24000000
C29H	6940000	6220000	12200000	20080000	35700000	20390000	261000	17780000	25400000	15160000	46900000	48300000	84100000	25200000	########	4450000	8030000	12700000	37400000	51000000
C30H	12390000	11200000	17800000	35390000	54400000	38870000	625000	32470000	40200000	28010000	81000000	90000000	########	43600000	########	7460000	15500000	30600000	75200000	91300000
C31S	5500000	5060000	10900000	16410000	18200000	17460000	287000	14280000	18130000	11730000	31600000	32800000	56700000	16000000	69800000	2570000	6150000	11700000	32100000	33600000
C31R	4080000	3450000	7990000	10660000	11800000	11950000	199000	9508000	12140000	7828000	21400000	22500000	38100000	11700000	50100000	1850000	4270000	7550000	14700000	22200000
GAM	0	0	1800000	2497000	2630000	1854000	95700	1629000	1609000	1366000	4210000	3960000	7250000	3090000	9280000	522000	1130000	1840000	2680000	3740000
C32S	3990000	3660000	8790000	11430000	10600000	12300000	188000	9274000	11230000	7625000	24000000	24200000	39700000	12300000	55200000	1610000	4490000	7860000	22000000	20700000
C32R	2560000	2520000	5990000	7324000	6780000	8477000	162000	6205000	7783000	5245000	16300000	15800000	27200000	8110000	37700000	1040000	3400000	5140000	14700000	14000000
C33S	2270000	2110000	5570000	6089000	4550000	7306000	106000	4868000	5845000	3922000	12500000	14000000	21400000	6040000	32100000	708000	2410000	3990000	11300000	10100000
C33R	1420000	1280000	3620000	4079000	2840000	4664000	50500	3000000	3563000	2448000	8850000	8900000	13900000	4490000	19600000	407000	1800000	2530000	6630000	6400000
C34S	1990000	1940000	6170000	6683000	4330000	7886000	83400	4113000	5086000	3463000	13000000	13900000	19000000	6010000	29500000	666000	2440000	3770000	8580000	9340000
C34R	1420000	1170000	3930000	4013000	2540000	4907000	29300	2486000	2895000	2059000	7540000	8270000	12200000	3580000	17200000	306000	1340000	2230000	5390000	5460000
C35S	1020000	770000	2220000	2417000	1130000	2487000	34400	1718000	2112000	1278000	5940000	6410000	11700000	2840000	14000000	189000	1130000	1490000	4730000	2180000
C35R	640000	412000	1460000	1532000	496000	1345000	16200	956200	1066000	666600	3010000	3440000	5780000	1450000	7260000	71000	648000	721000	2630000	781000
C21S	126000	184000	171000	760700	3110000	140100	0	230400	263500	350400	1870000	1420000	1660000	2160000	2530000	1290000	2110000	2470000	1510000	883000
C27diaS	537400	566000	591000	1955000	5450000	703600	0	877400	1119000	768600	2850000	2650000	3200000	3340000	4810000	1880000	2660000	4350000	3810000	1590000
C29diaS	1034000	1110000	1250000	3849000	11000000	1438000	0	1812000	2301000	1463000	4560000	4330000	5450000	5530000	7800000	2840000	4060000	7040000	7160000	3170000
C27aaaR	220700	263000	331000	1054000	3150000	324200	0	6165000	708200	424200	1090000	873000	1600000	1390000	1770000	386000	633000	1570000	1760000	657000
C28aaaR	103900	128000	117000	520700	1580000	485800	0	237800	297800	144000	570000	403000	669000	730000	830000	182000	330000	779000	644000	321000
C29aaaS	550800	554000	931000	2598000	4480000	1046000	0	1223000	1504000	983300	2590000	2410000	3310000	2470000	4620000	875000	1660000	2480000	3610000	1590000
C29abbR	551000	706000	1030000	2938000	4730000	1462000	0	1627000	2090000	1365000	3290000	3030000	4050000	3180000	5890000	1330000	2390000	3340000	4880000	2110000
C29abbS	475600	578000	956000	2344000	3860000	1181000	0	1106000	1648000	1080000	2710000	2390000	3300000	2570000	4730000	953000	1840000	2730000	3810000	1630000
C29aaaR	481400	462000	931000	2408000	3280000	1057000	0	1297000	1643000	1063000	2390000	2410000	3110000	2090000	4890000	808000	1610000	2460000	3110000	1470000

	A2425	A2426	A2427	A2428	A2429	A2430	A2431	A2432	A2433	A2434	A2435	A2436	A2468	A2469	A2470	A2611	A2627	A2706	A2884	A2892
C21tri	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3650000	1110000	1930000	2990000	0
C23tri	2380000	2310000	1400000	0	872000	0	0	0	944000	1890000	0	0	0	0	0	3650000	1460000	1730000	2130000	1990000
Ts	9930000	8780000	6560000	4610000	3360000	7500000	4590000	3350000	3010000	784000	3060000	8900000	37240000	34840000	19540000	17700000	6410000	9650000	12500000	7760000
Tm	14900000	13000000	8590000	10200000	4280000	11700000	7860000	5270000	6190000	11500000	6500000	12800000	51820000	53980000	12650000	36500000	9500000	14400000	17300000	3930000
C29H	30700000	28500000	19500000	22400000	11000000	26700000	15700000	12200000	13500000	25700000	12900000	22600000	135900000	116800000	25630000	74700000	21200000	31800000	38400000	8790000
C30H	48500000	48000000	34700000	36900000	21900000	46600000	26900000	22600000	20800000	46400000	21000000	33300000	259700000	222300000	49010000	129000000	41400000	53400000	66900000	15000000
C31S	17700000	19200000	11600000	16100000	8210000	19200000	10700000	8920000	8610000	18100000	8750000	10700000	99660000	92850000	16760000	54700000	17800000	22800000	25200000	5270000
C31R	11300000	12400000	7550000	11100000	5770000	12400000	7120000	5470000	5740000	12000000	5680000	6530000	68410000	59630000	11480000	37800000	12200000	15000000	16600000	3850000
GAM	1630000	1800000	1570000	1670000	1220000	2350000	1200000	1010000	965000	2300000	904000	849000	14840000	10560000	2731000	5890000	2050000	1870000	0	0
C32S	9630000	11200000	7020000	10500000	5760000	12300000	6510000	5480000	5290000	11800000	5300000	4960000	70540000	58890000	11570000	36600000	12800000	15700000	16500000	3710000
C32R	6000000	7330000	4700000	7080000	4090000	8270000	4420000	3520000	3430000	7850000	3500000	3150000	48970000	42700000	8838000	24000000	8950000	10500000	10700000	3560000
C33S	3950000	5230000	3280000	5200000	3340000	6350000	3150000	2620000	2300000	5920000	2360000	1860000	38360000	32930000	6148000	19100000	7880000	9230000	7130000	2250000
C33R	2100000	2760000	1870000	3150000	2060000	4010000	1770000	1520000	1350000	3690000	1450000	1020000	23290000	20630000	3837000	12000000	4980000	5500000	4770000	1260000
C34S	2760000	4030000	2640000	3810000	3560000	6170000	2450000	2280000	1700000	5320000	1760000	1210000	65720000	28830000	5905000	16000000	9380000	8670000	6480000	1690000
C34R	1490000	2280000	1430000	2400000	2210000	3880000	1400000	1280000	877000	3060000	1000000	597000	41110000	18250000	3576000	10100000	5830000	5520000	3560000	923000
C35S	471000	828000	501000	1550000	849000	1670000	618000	479000	448000	1410000	568000	20100	25600000	13250000	2297000	9570000	3820000	3630000	1630000	635000
C35R	230000	291000	200000	784000	408000	757000	228000	179000	186000	599000	227000	10000	14750000	6785000	1134000	5100000	2340000	2020000	761000	220000
C21S	653000	560000	519000	381000	368000	486000	311000	207000	298000	530000	401000	1030000	3841000	2877000	2492000	1600000	1050000	880000	1080000	1290000
C27diaS	1040000	968000	710000	753000	423000	692000	551000	372000	424000	908000	614000	1300000	6462000	5582000	3957000	3840000	1490000	1780000	2170000	1950000
C29diaS	2170000	1720000	1520000	1610000	799000	1430000	1110000	787000	810000	1840000	1140000	2370000	10820000	9838000	6672000	6200000	2740000	3250000	3810000	3330000
C27aaaR	536000	235000	283000	439000	124000	348000	241000	155000	149000	413000	307000	475000	4360000	2580000	1530000	1570000	708000	727000	809000	333000
C28aaaR	319000	184000	187000	240000	118000	213000	133000	109000	118000	243000	199000	245000	1760000	1040000	718000	642000	258000	408000	287000	236000
C29aaaS	1200000	1000000	908000	961000	517000	878000	650000	451000	514000	966000	738000	1210000	8594000	6405000	2794000	3220000	1550000	1820000	2010000	1180000
C29abbR	1540000	1290000	1230000	1240000	683000	1280000	879000	630000	672000	1220000	944000	1530000	11220000	6871000	3647000	4170000	1840000	2170000	2350000	1650000
C29abbS	1210000	1060000	926000	950000	538000	1030000	712000	485000	518000	898000	725000	1160000	8935000	5362000	2897000	3390000	1360000	1730000	1940000	1310000
C29aaaR	1100000	931000	894000	924000	489000	873000	628000	421000	492000	854000	672000	1010000	8251000	5234000	2464000	3290000	1460000	1540000	1730000	1150000

	A2895	A2896	A2897	A2898	B515	B554	B1014	B1279	B1393	B1443	B2121	B2122	B2887	B1873	B1874	C495	C499	C503	C511	C513
C21tri	4730000	0	2050000	0	0	0	0	0	0	1680000	1010000	7250000	0	1720000	1320000	1927000	1959000	1063000	2119000	1800000
C23tri	3060000	2210000	1570000	2470000	0	713400	1024000	2312000	213800	5482000	1670000	14100000	1640000	4260000	3140000	7834000	6310000	4719000	9069000	6670000
Ts	16100000	10300000	10700000	13900000	231100	257400	351500	643900	156000	1086000	500000	3560000	713000	1520000	1130000	5106000	5777000	1943000	4091000	5090000
Tm	25500000	8430000	13900000	10400000	250900	164900	145000	511800	62860	4973000	888000	3970000	310000	978000	598000	4939000	2506000	4280000	7294000	5310000
C29H	52700000	16900000	31000000	21500000	0	405300	0	1192000	153700	15010000	1870000	10800000	911000	3990000	2760000	16580000	7859000	13290000	20800000	16370000
C30H	85200000	26400000	58400000	32600000	1226000	890200	769000	2820000	423600	15340000	3300000	24700000	2240000	8820000	6020000	18390000	9606000	13400000	21140000	16390000
C31S	28900000	8990000	23400000	10400000	0	350800	372100	903800	133800	8302000	1140000	8040000	594000	2590000	1640000	10040000	5414000	8007000	12410000	9473000
C31R	19800000	5630000	15800000	7100000	0	245500	189300	678400	108000	5655000	838000	8280000	607000	2090000	1310000	6935000	3574000	5854000	9553000	6814000
GAM	0	0	2910000	0	0	178600	189900	504300	0	1482000	205000	2360000	0	772000	406000	2251000	1345000	2140000	4093000	2368000
C32S	17900000	5260000	15900000	6620000	0	219500	182900	748500	130500	5674000	658000	5100000	509000	1530000	938000	8455000	4632000	6697000	11400000	7430000
C32R	11000000	3480000	11000000	4540000	0	145900	116100	443320	85620	3731000	434000	3760000	341000	1210000	715000	5593000	2922000	4299000	7440000	4865000
C33S	7150000	2100000	9440000	2600000	0	135800	133000	528200	103000	3371000	357000	3770000	283000	1250000	773000	5977000	3224000	4785000	8079000	5043000
C33R	3920000	1120000	5500000	1520000	0	84820	74310	324900	55780	2150000	175000	2960000	128000	635000	423000	3704000	2042000	3076000	5182000	3062000
C34S	5510000	1330000	8950000	1830000	0	99420	63970	369900	51350	2141000	198000	2390000	118000	700000	370000	5606000	3190000	4065000	7127000	4298000
C34R	2750000	712000	5370000	929000	0	56750	45770	189000	40400	1362000	124000	1080000	75500	419000	230000	3394000	1923000	2602000	4603000	2471000
C35S	886000	206000	3200000	370000	0	47700	32800	138000	30600	1880000	25400	1240000	0	395000	238000	8054000	4406000	6263000	10990000	6528000
C35R	400000	100000	1540000	150000	0	30700	23860	82060	24200	1015000	13000	1010000	0	214000	193000	4965000	2271000	3880000	6974000	3842000
C21S	1590000	1270000	823000	1480000	454100	408900	1114000	1624000	67580	1557000	1020000	8970000	1370000	1090000	792000	1039000	951500	512200	952000	731000
C27diaS	2570000	1720000	1730000	2100000	627700	592900	809100	1961000	121100	1193000	768000	7910000	846000	1510000	1050000	575800	906900	350100	610700	302000
C29diaS	4370000	3330000	3210000	4010000	915100	764000	1127000	2388000	153300	2575000	928000	9090000	1070000	2130000	1540000	2305000	1867000	1962000	3130000	1500000
C27aaaR	960000	667000	682000	645000	512300	319700	408200	1032000	33720	3036000	574000	7030000	387000	514000	305000	1425000	788800	1219000	2148000	789000
C28aaaR	366000	302000	346000	508000	230000	124200	124600	365100	18980	1141000	228000	2540000	144000	156000	124000	500600	273600	453800	827800	287000
C29aaaS	2210000	1320000	1830000	1930000	421000	290200	330800	806100	48980	2337000	401000	5350000	318000	651000	461000	1625000	1085000	1411000	2362000	975000
C29abbR	2460000	1830000	2230000	2450000	456900	348400	665300	1247000	74660	2936000	511000	6500000	678000	1020000	702000	2252000	1327000	1645000	2325000	1650000
C29abbS	1870000	1480000	1750000	1970000	386000	291400	574300	955800	64740	2392000	366000	5210000	555000	841000	590000	2321000	1398000	1936000	2866000	1580000
C29aaaR	1820000	1200000	1470000	1640000	539800	314600	303100	1055000	46680	3249000	491000	8860000	359000	658000	463000	1843000	1099000	1478000	2593000	1050000

	C529	C540	C548	C553	C557	C566	C574	C575	C579	C582	C589	C596	C711	C714	C721	C722	C725	C1386	C1387	C1388
C21tri	1180000	0	327600	0	1945000	0	633600	2336000	0	1028000	0	1142000	1891000	1700000	3679000	3057000	2850000	0	7534000	1883000
C23tri	5266000	4076000	1705000	4127000	9556000	7040000	2818000	12240000	3100000	5828000	1780000	5860000	5902000	5536000	13490000	10930000	9689000	8534000	24780000	6858000
Ts	2469000	2476000	1124000	873400	4172000	2850000	1099000	4068000	1540000	2148000	1050000	2223000	3653000	3736000	5880000	5885000	3600000	5214000	14080000	1533000
Tm	3815000	5104000	1472000	3050000	10940000	9180000	3190000	9866000	3860000	4913000	1170000	6095000	2350000	2843000	6867000	5588000	3728000	12140000	21470000	4392000
C29H	12860000	15620000	5080000	8709000	27690000	21970000	10980000	25150000	11660000	16400000	3867000	20550000	7601000	8856000	18110000	16330000	14110000	38440000	42650000	15040000
C30H	12410000	18570000	4882000	9377000	28270000	22300000	10120000	27150000	12900000	16190000	4825000	21520000	8373000	10310000	18210000	16890000	13000000	36060000	47640000	14260000
C31S	7263000	10810000	2793000	5187000	18190000	13280000	6104000	16650000	7624000	9512000	2489000	12970000	4971000	6227000	12720000	11230000	8357000	25660000	33750000	7472000
C31R	5252000	8168000	1941000	3651000	13960000	9910000	4585000	13130000	5353000	7169000	2	9068000	3150000	3901000	8577000	7351000	5876000	18140000	28310000	4966000
GAM	1955000	2675000	683800	1053000	5409000	4012000	1670000	5190000	1592000	2869000	691900	3352000	1087000	1408000	3550000	2613000	2132000	5808000	11180000	1922000
C32S	5690000	9532000	2526000	3163000	17020000	10370000	4782000	14880000	6348000	8144000	1991000	10540000	3769000	4805000	9945000	9009000	6780000	20900000	28210000	5796000
C32R	3698000	6466000	1536000	2063000	10620000	6581000	3189000	9487000	4412000	5189000	1309000	7042000	2582000	3249000	7154000	6301000	4991000	13810000	23850000	3651000
C33S	3901000	6267000	1911000	1607000	12260000	7082000	3166000	10180000	4225000	5627000	1457000	7473000	2717000	3664000	8055000	6884000	5654000	14460000	22860000	3234000
C33R	2383000	3856000	1130000	1061000	8114000	4369000	2062000	6395000	2607000	3465000	865400	4689000	1773000	2326000	5086000	4453000	3631000	9174000	15830000	1953000
C34S	3078000	5325000	1675000	1061000	12100000	5618000	2497000	9181000	4800000	4765000	1164000	6216000	2401000	3301000	6451000	6166000	4748000	12430000	21140000	2063000
C34R	1914000	3357000	976000	624200	8086000	3612000	1602000	5826000	2912000	3089000	736500	3858000	1471000	2046000	4179000	3822000	3047000	7852000	14320000	1196000
C35S	4651000	6550000	2672000	905900	15810000	8189000	3681000	12590000	5145000	7428000	1889000	9774000	3460000	4756000	10160000	8955000	6848000	17340000	25970000	2289000
C35R	2990000	3841000	1772000	442400	10640000	5117000	2237000	8637800	3138000	4727000	1169000	5988000	2144000	3053000	6783000	6039000	4523000	10400000	18750000	1214000
C21S	567600	702100	235000	1257000	993500	968000	429030	917800	407000	492800	324000	718400	1008000	952700	1745000	1773000	1308000	1024000	2654000	1137000
C27diaS	301200	888600	68550	1681000	579600	604000	377600	442500	432000	218300	321000	495900	640600	706200	728000	792100	655100	639100	1837000	306700
C29diaS	1628000	2838000	458000	3782000	4848000	3020000	1611000	3869000	1770000	2065000	751000	2853000	1792000	1950000	3672000	3199000	2306000	4579000	9029000	1520000
C27aaaR	1000000	1787000	325700	2930000	3777000	2080000	983000	2896000	1140000	1486000	450000	2082000	856000	985400	2560000	2067000	1421000	3455000	6385000	1090000
C28aaaR	350600	665100	81900	1057000	1278000	786000	367800	985100	447000	508100	164000	725800	275500	310000	748800	606800	488700	1078000	2295000	755400
C29aaaS	1044000	2050000	313400	2321000	3831000	2160000	1043000	3009000	1300000	1518000	489000	2208000	939100	1136000	2451000	2184000	1637000	4104000	7327000	1113000
C29abbR	1411000	2916000	629400	2952000	3846000	2280000	1436000	3002000	1640000	1787000	711000	3190000	1527000	1801000	4193000	3364000	2304000	6109000	11000000	1966000
C29abbS	1510000	2560000	545800	2501000	4755000	2640000	1558000	3749000	1900000	2126000	640000	3033000	1401000	1696000	3721000	3313000	2157000	5598000	9603000	1647000
C29aaaR	1135000	2430000	459800	2883000	4389000	2290000	1164000	3499000	1410000	1754000	525000	2497000	1160000	1406000	3259000	2917000	2569000	4344000	7833000	1343000

	C1389	C1390	C1465	C1466	C1467	C1468	C1469	C1470	C1471	C1472	C1473	C1705	C1715	D756	D800	D801	D802	D841	D842	D924
C21tri	0	996100	664200	367300	1020000	894600	199500	1552000	1007000	459900	441300	5290000	2470000	1236000	0	0	199900	0	0	0
C23tri	93200	2573000	3650000	1999000	5607000	4926000	1014000	7321000	4607000	2369000	2190000	13400000	5220000	2123000	1117000	676600	509600	1410000	758000	884000
Ts	28760	508600	1380000	768900	1882000	2138000	324600	2747000	1317000	858700	713600	16300000	19900000	6729000	3547000	899000	1294000	4495000	3340000	4210000
Tm	94580	1578000	2896000	1519000	4999000	4262000	1017000	4814000	3928000	1746000	2198000	8890000	8620000	5809000	3171000	850800	1897000	10580000	7490000	4431000
C29H	297400	5484000	9874000	5447000	17220000	14010000	3765000	16910000	13870000	6530000	7594000	34600000	28300000	10070000	9151000	3033000	3272000	26180000	20500000	12370000
C30H	287700	5134000	10720000	5759000	17630000	14880000	3445000	18110000	13170000	6888000	7570000	44900000	51600000	14000000	19660000	6808000	4543000	36490000	29100000	23270000
C31S	172000	2787000	5617000	3207000	10060000	8297000	2167000	10020000	7452000	3693000	4248000	20900000	18400000	7918000	7455000	2446000	2809000	22620000	16600000	9931000
C31R	122800	1886000	3862000	2139000	6932000	5746000	1420000	6785000	5101000	2478000	2946000	18300000	13600000	6367000	5651000	1684000	2452000	17950000	12500000	8031000
GAM	40200	678200	1334000	750200	2670000	2017000	563300	2236000	1909000	878600	561900	3520000	3440000	3051000	1209000	430400	1115000	5758000	2490000	1791000
C32S	127100	2188000	4994000	2797000	8656000	6968000	1756000	8478000	5736000	3125000	3462000	18300000	17000000	6284000	5760000	1713000	2384000	18260000	13300000	7937000
C32R	87500	1329000	3119000	1732000	5396000	4487000	1116000	5468000	3691000	1981000	2139000	11800000	12000000	4696000	3734000	1149000	1864000	12870000	9180000	5505000
C33S	80080	1542000	3574000	2029000	5950000	5235000	1253000	6365000	4050000	2323000	2352000	11000000	9860000	3885000	2878000	993900	1882000	13430000	9240000	4619000
C33R	51080	899400	2092000	1201000	3556000	3124000	728500	3801000	2455000	1306000	1414000	6510000	6430000	2630000	1832000	639300	1393000	9333000	6340000	2961000
C34S	65830	1470000	2997000	1796000	5101000	4505000	999000	5156000	3099000	1865000	1807000	8040000	11200000	5602000	4624000	728500	2537000	22420000	16300000	8541000
C34R	42410	678800	1721000	1018000	2979000	2617000	616600	3152000	2013000	1112000	1080000	4570000	6310000	4006000	2663000	417500	2066000	17030000	12000000	5334000
C35S	94340	1564000	4329000	2494000	7028000	6284000	1485000	8158000	4463000	2737000	2516000	9930000	7340000	2472000	1787000	353900	1393000	11600000	7830000	3955000
C35R	59770	1032000	2661000	1608000	4567000	3948000	930200	5150000	2818000	1699000	161000	5260000	3800000	1683000	1040000	205400	1036000	8020000	4950000	2306000
C21S	9194	339300	259500	146900	408300	366000	79450	550800	490100	165100	216700	2590000	2200000	859500	403900	243600	150600	359300	204000	357800
C27diaS	9683	249800	129800	76990	213600	186400	49580	281400	410600	114500	194100	1970000	1860000	1477000	1381000	1028000	244900	927200	674000	860700
C29diaS	60480	945300	867500	476900	1409000	1143000	300900	1452000	1497000	589600	806600	3930000	4190000	2831000	3244000	1944000	602300	2546000	1740000	2450000
C27aaaR	39250	446400	492300	271800	908300	708500	182900	904700	839700	329900	442500	2730000	1330000	1052000	744100	407500	291900	1721000	1180000	649900
C28aaaR	9622	130400	153200	82050	257700	194200	58080	256200	0	97880	134800	759000	402000	568900	495900	199000	125800	616000	419000	527100
C29aaaS	36570	421900	538300	282100	914800	690600	195200	904500	829100	338800	437800	3640000	3360000	2591000	1584000	732900	656000	3212000	2230000	1788000
C29abbR	64220	717100	985000	533200	1629000	1228000	364100	1573000	1426000	610200	789500	6340000	5740000	3021000	2335000	1224000	834900	3594000	2510000	2365000
C29abbS	57320	705700	866600	483500	1464000	1137000	310100	1412000	1239000	543200	713700	5080000	4580000	2654000	1930000	1002000	809400	3027000	2150000	1849000
C29aaaR	40610	616400	591200	341400	1140000	836900	221600	1071000	954600	399800	508800	4110000	3790000	2307000	1275000	682500	837200	4359000	2930000	1603000

	D1173	D1273	D1274	D1275	D1276	D1288	D1289	D1290	D1291	D1312	D1313	D1335	D1364	D1365	D1385	D2471	D2472	D2595	D2626	D2885
C21tri	0	0	0	0	0	1495000	0	0	0	0	0	0	0	0	0	0	40240000	38060000	3400000	21500000
C23tri	2190000	1940000	1640000	1307000	459800	2240000	904000	1119000	2694000	1060000	196000	2025000	4857000	2389000	6462536	0	73510000	65530000	4560000	63800000
Ts	9977000	6860000	4978000	4885000	1898000	9307000	4411000	5231000	11540000	4271000	1320000	7922000	7404000	3293000	11889165	50850000	89300000	139300000	27000000	102000000
Tm	14520000	10620000	9615000	5100000	1745000	6141000	3061000	3757000	9172000	9517000	1181000	11050000	7504000	4405000	12547264	125300000	139100000	182500000	46400000	279000000
C29H	45700000	32840000	26540000	13890000	5381000	16950000	9189000	10670000	24290000	31990000	3712000	26620000	23280000	14890000	23644468	348400000	240600000	442100000	104000000	417000000
C30H	63720000	54950000	44070000	26110000	10030000	33350000	17940000	21790000	47750000	54910000	6955000	41960000	44790000	29790000	38740612	552500000	332000000	673300000	164000000	154000000
C31S	36610000	22970000	20290000	9667000	3723000	11340000	5947000	7574000	17290000	23630000	2828000	21910000	15660000	10590000	22531618	284000000	202800000	289700000	76800000	103000000
C31R	28010000	16420000	15270000	7236000	2673000	8950000	4633000	5936000	13170000	17620000	2020000	17170000	12570000	7637000	18345696	212800000	180500000	215400000	52500000	0
GAM	7810000	3384000	3396000	1776000	662400	2399000	1325000	1194000	3544000	3882000	681700	4549000	3594000	2066000	0	72150000	65770000	65370000	9690000	90100000
C32S	28900000	14970000	13630000	7168000	2663000	8066000	4279000	5559000	13170000	16350000	2291000	16750000	12420000	7901000	18272410	206100000	174500000	200300000	49900000	54600000
C32R	21280000	9832000	9388000	4816000	1702000	5853000	3025000	3959000	9060000	11460000	1613000	11760000	8750000	5247000	12608585	148000000	130100000	142400000	33900000	39200000
C33S	18790000	8439000	8503000	3872000	1391000	4576000	2642000	3274000	7628000	10660000	1462000	10620000	7382000	4446000	11472005	125100000	107600000	110000000	27200000	21500000
C33R	12180000	5193000	5504000	2241000	919800	2832000	1630000	1935000	4779000	6740000	1052000	6802000	4743000	2133000	6666641	84690000	66680000	72580000	17200000	48500000
C34S	34940000	16050000	16040000	7450000	2648000	7631000	4477000	5898000	14080000	21470000	2319000	18760000	13790000	8989000	17794250	252200000	139400000	194600000	40400000	26000000
C34R	25150000	9978000	10610000	4583000	1574000	4769000	2685000	3724000	8596000	14020000	1643000	12250000	8843000	5707000	11573888	165400000	92250000	118400000	26900000	8580000
C35S	14280000	4726000	4820000	2345000	798300	2485000	1434000	2182000	4902000	7427000	938000	7112000	5508000	3217000	6087510	107900000	45730000	59790000	11700000	3430000
C35R	8887000	3159000	2920000	1371000	513600	1428000	884800	1290000	3035000	4824000	601900	4412000	3393000	2222000	3711206	66260000	25050000	34610000	6900000	7110000
C21S	904700	641700	600700	512700	221200	971000	438600	553700	1209000	382500	33090	522400	656000	325000	3345943	4722000	14900000	19500000	3630000	7050000
C27diaS	1935000	1094000	973700	995200	426400	1970000	934400	1150000	2304000	655000	151500	980100	1113000	492500	2659136	5227000	14590000	22400000	5780000	10800000
C29diaS	5681000	3036000	3427000	2672000	1058000	3794000	1868000	2042000	4782000	2386000	430600	2973000	3139000	1486000	4962737	12800000	29960000	0	9690000	7520000
C27aaaR	2282000	957900	1383000	763900	357500	1238000	627700	667400	1901000	1163000	150500	1056000	1179000	586400	4106019	8598000	16490000	19300000	5480000	2700000
C28aaaR	1211000	626100	795700	594100	219400	669400	328200	402400	976700	680900	95550	677800	681600	344800	938203	0	0	8210000	2200000	15800000
C29aaaS	6729000	2368000	3045000	1955000	727900	2829000	1460000	1686000	4045000	2564000	354600	2675000	2723000	1519000	5821991	18730000	36190000	39800000	10100000	19400000
C29abbR	8488000	3671000	3897000	2754000	1117000	3526000	1861000	2112000	5146000	3790000	401100	3013000	3748000	2160000	6429155	24330000	37490000	53200000	10300000	14900000
C29abbS	7452000	3087000	3228000	2155000	841000	3138000	1638000	1888000	4315000	2846000	372800	2702000	2767000	1659000	5344605	18040000	37380000	39100000	7440000	15100000
C29aaaR	6180000	1952000	2956000	1636000	617500	2068000	1091000	1245000	2959000	2443000	304400	2239000	2338000	1339000	5281395	18500000	30330000	40100000	10100000	0

Gasoline Range of the sample set

	A549	A550	A920	A1140	A1710	A1711	A1712	A1723	A1724	A1725	A2268	A2269	A2270	A2283	A2284	A2313	A2362	A2363	A2364	A2424
iC5	55542	21546	27212	52715	58241	89114	30	131515	230926	6817	545	428	101	1593	175	6435	2717	111323	40486	7585
nC5	318243	116509	73721	289413	262769	826245	28	914993	1956355	25043	1344	1409	527	3244	620	19441	9057	411988	155665	38142
22DMC4	2280	516	631	2003	2986	1006	29635	50	5804	38	100	87	72	370	0	417	215	6041	1740	554
CYC5	25896	10887	11854	0	0	12	0	0	0	0	523	396	579	984	209	2383	1121	42928	17353	5260
23DMC4	10381	2857	3685	29858	34352	68099	88006	70526	151376	4919	423	429	413	1484	148	2143	1258	24039	6070	3050
2MC5	104041	29987	30193	130583	144297	349068	574334	243729	616055	37218	1917	1898	1950	5424	896	17067	9976	208920	65037	26639
3MC5	57370	16372	16339	68078	84412	170457	359261	126369	310620	19821	1402	1311	1607	4157	602	10460	6092	118336	36167	15407
nC6	755000	250732	104057	1076489	718615	2898181	1862030	1986983	5500576	247261	13223	14177	14621	22934	7454	99348	58365	1091204	341645	198422
22DMC5	1891	0	861	0	0	0	61849	12	0	0	202	171	230	679	98	908	518	6476	1902	1106
MCYC5	177747	56634	54513	160175	170548	469354	372664	347952	913015	47310	4719	3919	7500	9320	2058	25540	14805	257563	87580	46504
24DMC5	6214	1597	2098	9583	14367	21718	110874	13769	34681	4968	437	370	525	1176	164	2592	1560	20774	4397	2661
223TMC4	387	0	139	97	1246	1070	18499	1032	2114	188	76	53	73	156	16	48	0	1688	322	251
Benzene	36379	4339	8832	100171	2565	315731	457909	167030	341217	5856	1516	1564	1775	2320	966	5303	2965	54941	18588	34255
33DMC5	2622	466	463	2649	5926	5245	45954	3572	9001	1423	234	149	385	567	58	879	544	6627	1678	901
CYC6	140977	50378	32873	147881	285327	336075	554182	277736	665707	45318	10456	7710	19002	21779	3824	36256	21264	310517	105647	55768
2MC6	80288	17270	17071	123392	126597	314466	833427	204345	489543	67886	3159	2849	4376	5750	1655	23167	15460	167589	43523	31012
23DMC5	27618	5899	9942	23328	42453	80481	129360	48532	113191	15707	1912	1551	3388	4058	804	7193	4385	55425	17532	10610
11DMCYC5	10061	2332	2841	6884	21341	21966	107501	15351	26539	3932	1161	854	2012	2852	362	3714	2474	27201	7197	4279
3MC6	88461	19536	20187	120454	133077	321055	672129	210296	520409	72318	3786	3205	5612	6676	1854	23647	15136	168824	48303	34016
1c3DMCYC5	38226	8408	11149	39359	42131	117031	173511	78990	197499	24434	1640	1319	3000	2885	775	8058	5051	60087	22731	12987
1t3DMCYC5	35584	8066	9976	36634	39469	107489	161391	72349	181987	21527	1629	1297	2870	2733	693	7573	4694	56311	21275	12308
1t2DMCYC5	102216	22469	26603	80145	119600	270822	315403	198282	494526	65106	4545	3403	8854	7768	1903	15819	9777	118927	50860	32376
nC7	1353245	270972	131842	1801696	761812	4907136	3788353	2772286	8284335	597473	39644	37911	56200	47810	25606	217421	138014	1569122	458845	463322
MCYC6	391026	104210	86498	450542	732219	1081369	2300511	802157	2036684	288509	33755	24332	62692	52316	14069	113459	70468	723970	240097	199695
22DMC6	17788	2248	4719	14934	29403	37406	194400	30226	53264	12634	1442	969	2742	2286	374	4898	3201	30604	9648	6945
ECYC5	56089	13502	9762	54875	31502	185275	72605	139816	367601	35469	1553	1259	3457	1625	785	6188	3587	43587	21024	20727
25DMC6	7636	562	1984	8861	11126	22779	154208	22502	27954	7467	561	268	996	563	255	2555	1664	14777	3798	3978
24DMC6	9164	943	2644	11967	18490	28892	166566	26321	37947	10110	999	614	1488	1141	384	3791	2583	22443	5152	5212
223TMC5	24469	2550	6321	19889	30655	59890	103696	57437	95113	21303	1611	1055	3182	1748	650	4818	3006	29715	14304	9161
234TMC5	293	2410	166	18236	73	57600	2708	1991	376	98	3	85	74	32	34	3922	9671	460	0	8661
Toluene	72971	7214	4036	110717	4667	383304	2145944	270875	203431	1983	10977	8427	15357	8478	6671	12410	7836	108013	34141	89931
2MC7	109847	9356	15756	107398	100914	302095	944763	247382	400086	86082	8633	5031	10842	5455	3687	26373	19389	144962	44023	50292
3MC7	162823	17047	9553	161171	230850	404304	1239310	346441	631610	153603	6100	3344	9163	4656	2428	13872	31420	226731	72620	26838
1c4DMCYC5	15152	2651	4779	21517	48731	65471	162605	55087	89679	25964	2771	1328	3896	2435	807	7332	5017	36448	13255	16262
nC8	1673016	113780	76085	1113303	320167	155836	4055131	1810897	4930636	383090	76062	42656	102014	32583	35416	157615	119479	917433	319316	533826

	A2425	A2426	A2427	A2428	A2429	A2430	A2431	A2432	A2433	A2434	A2435	A2436	A2468	A2469	A2470	A2611	A2627	A2706	A2884	A2892
iC5	16341	10018	10765	43413	9442	22561	95953	2266	2936	39029	10510	13920	394	301	408	16844	16651	17756	5326	1098
nC5	92756	48391	54252	216734	42000	127901	428005	9637	19771	179472	64385	68086	721	1309	1689	70143	33413	103126	25222	5113
22DMC4	707	363	608	1560	511	767	2600	179	107	1372	348	625	147	0	32	664	1548	1250	495	158
CYC5	8367	4670	5650	29252	5133	16674	41891	848	1491	15484	7010	5684	475	529	406	11658	5787	15113	3980	925
23DMC4	4025	1824	3301	7976	2611	4302	12201	785	719	7137	1594	2845	457	167	289	2425	5341	4757	2205	801
2MC5	44781	18946	28010	86068	20719	47159	115980	5943	8111	72075	18224	27066	1983	1132	1530	23428	31955	40117	20232	6004
3MC5	24180	10122	16640	49261	12358	26038	64108	3482	4313	38050	9889	14892	1652	834	979	13618	20357	23983	11994	3723
nC6	343844	142964	240212	722805	153400	396428	853373	42360	81979	475606	180164	216829	9998	8444	10880	143167	108120	290472	132793	39326
22DMC5	1102	369	1203	1767	852	974	2578	262	91	1587	256	794	285	52	73	512	2374	1219	696	392
MCYC5	67453	29513	45164	193611	34186	87562	190296	7422	13072	105856	35189	39730	5686	3999	2802	50730	39954	85393	38801	10188
24DMC5	3356	1218	3211	5998	2234	3097	7115	688	571	4724	948	2030	600	141	199	1314	4539	3276	2068	964
223TMC4	229	58	237	453	228	49	639	54	52	344	41	172	47	29	37	0	536	304	153	22
Benzene	52515	21707	37014	52639	23992	70986	111690	7989	8421	44821	11999	29485	1407	994	1005	7365	4522	10269	11275	3314
33DMC5	1054	289	982	2236	702	956	2258	157	208	1453	326	625	344	83	97	568	1593	1526	805	332
CYC6	62638	25126	63191	160297	45366	84645	175447	8436	10892	87497	29512	39283	13857	5900	4813	43191	41291	97689	43120	15693
2MC6	43227	16070	37086	94101	24834	41128	83535	6641	8837	52053	15612	23918	3628	1499	1694	16102	32475	41274	22452	8881
23DMC5	14655	5312	11073	29321	8497	13379	26331	1875	2296	18973	4268	7077	3345	986	841	7001	11015	16173	10382	3371
11DMCYC5	4151	1598	4377	9092	3384	4058	9410	696	703	6179	1367	2617	1622	390	434	2365	5287	5647	3532	1651
3MC6	47201	17175	39142	107705	26683	44895	90712	6640	9455	56216	17074	24761	4702	1841	1812	18644	32800	48170	25811	9444
.c3DMCYC	17440	6650	13128	45838	9347	17238	35801	2105	3481	22795	6729	8828	2425	1100	741	9664	10864	19564	11269	3414
Lt3DMCYC	16209	6190	12412	43773	9010	16274	33911	1935	3245	21376	6345	8271	2321	1057	683	9238	10176	18635	10624	3184
Lt2DMCYC	46548	16483	29309	133458	20801	43598	93596	4395	9101	60097	18349	21622	8159	3465	1742	25305	23441	56260	29538	6761
nC7	709806	261409	592292	1863803	369986	723152	1355517	89974	183587	735507	326982	364779	38572	22137	20446	225600	204950	658959	298457	94256
MCYC6	199564	67947	203596	546108	139642	194682	419141	29413	38458	234862	77059	102379	45723	16058	12037	96417	124084	265162	144572	49193
22DMC6	6860	2251	6584	17189	5083	5829	11936	1254	1128	8357	1839	3252	2370	451	203	3246	7227	9099	7075	2428
ECYC5	32659	11029	20818	106141	13675	34569	65124	2859	7569	33462	15132	14233	2344	1465	602	13439	7313	32225	14287	2774
25DMC6	3894	1169	3704	9610	2895	3411	6417	530	592	4213	945	1698	627	248	120	1612	4603	4498	3643	1122
24DMC6	5133	1611	5217	12426	4159	4368	8492	1077	705	5505	1318	2260	1036	248	314	1905	6017	6101	4656	1750
223TMC5	10989	3619	8332	33909	6412	9761	18620	1283	2122	12422	3539	4525	2214	812	343	5217	6588	13864	9223	3345
234TMC5	11244	3610	7774	36127	5850	796	18679	1106	2216	12024	3818	4572	78	54	13	0	86	84	44	77
Toluene	131055	54314	110538	246063	81310	152880	206502	22545	26355	103829	36868	62526	7778	4195	3176	42668	23271	78609	66343	16134
2MC7	52259	19055	47609	162528	37423	47382	90043	8425	11564	50097	17912	24368	5869	2416	1845	22885	33912	61223	40634	14924
3MC7	22203	6718	17185	62045	19689	22898	42586	2134	5955	26469	8009	10747	5089	1765	778	4426	16626	30843	11520	6963
.c4DMCYC	12658	4014	16668	37844	11112	11042	25227	2129	2455	11801	3697	5558	2150	535	477	4223	8068	16712	7251	2353
nC8	737743	266551	621238	2669269	453351	668177	1224282	91295	198923	609769	307479	352505	42353	24975	15746	193130	184156	792047	276712	77921

	A2895	A2896	A2897	A2898	B515	B554	B1014	B1279	B1393	B1443	B2121	B2122	B2887	B1873	B1874	C495	C499	C503	C511	C513
iC5	4845	4957	7371	1539	7599	141883	0	45539	342289	482913	29221	12006	3578	0	0	7347	2660	144741	8116	27749
nC5	20384	34169	24186	10143	13617	240481	0	72005	833954	1458581	50608	17217	7572	0	0	13290	6311	214333	15143	49963
22DMC4	373	317	1212	175	356	1767	0	775	5551	0	2040	1723	273	0	0	273	131	1979	204	959
CYC5	4372	4703	3709	1683	44	0	0	0	0	0	0	0	1741	0	0	2094	1232	40151	2891	9200
23DMC4	1883	1638	4228	850	5955	30206	0	8783	171714	183375	18506	11852	1958	0	0	2531	1140	31702	2518	9951
2MC5	16039	16728	27787	7663	80278	397225	0	124546	1044306	1793018	185932	91014	18391	0	0	19427	11785	204480	22820	78476
3MC5	9989	9969	17731	4678	55498	265174	0	80879	754058	1263633	127402	66182	14705	0	0	14644	8346	167100	20038	60386
nC6	102389	173712	130954	75704	127412	626552	0	173698	1936610	3834118	256485	104466	34781	0	0	43945	32997	366012	57298	181917
22DMC5	510	437	2272	257	0	0	0	0	0	0	6385	2540	970	0	0	529	593	3470	238	1930
MCYC5	37599	34839	38087	15461	68978	518762	0	121077	1785196	2746982	117179	74191	41497	0	0	23590	17536	235884	36456	92202
24DMC5	1643	1453	4444	906	12148	30646	0	11860	66545	86916	27055	15932	3868	0	0	1533	1384	11281	1981	6378
223TMC4	131	0	588	0	360	1143	0	68	4914	1775	1544	1280	146	0	0	0	0	839	18	550
Benzene	6698	11041	8569	5201	712	10869	0	923	190841	118582	786	0	5798	0	0	35	12516	400516	12735	160795
33DMC5	698	616	1666	399	2531	4323	0	1479	12801	5911	3922	2667	1052	0	0	366	393	8563	452	2302
CYC6	38394	53586	44606	35024	14974	97039	0	31570	630693	416051	26411	17079	28161	0	0	42854	29517	249828	42423	200078
2MC6	17063	19591	33144	11662	131890	351133	0	129800	675349	1529421	221459	96281	35409	0	0	21624	18471	135428	33858	84386
23DMC5	8261	7141	13090	4468	44463	84183	0	29128	218486	421966	113878	70264	21362	0	0	10841	7530	62365	15388	42680
11DMCYC5	2648	2606	5485	1626	15780	65802	0	17485	273058	48779	33383	25167	12585	0	0	3865	4161	15767	6255	13861
3MC6	20011	22721	34634	13461	203935	519764	0	184399	1121140	2748085	369586	174328	60600	0	0	33188	26604	221169	53701	128459
.c3DMCYC	9478	7899	11688	4654	139146	458016	0	118631	1240709	1704722	240116	147135	36456	0	0	11172	10264	89413	26375	36688
Lt3DMCYC	8946	7420	10992	4404	127891	422281	0	110451	1184249	1586388	217023	135290	35146	0	0	10214	9504	84395	24944	33552
Lt2DMCYC	27455	22332	29221	13499	304098	925929	0	271563	2078723	3223733	565692	362145	94482	0	0	25230	22329	215540	63754	85011
nC7	234548	386974	282039	220821	377655	989254	0	233694	2297995	5642711	565128	176482	104999	0	0	82311	81316	453100	135910	350807
MCYC6	115546	146371	152989	94544	216008	651881	0	199562	2191021	2227259	350320	217369	138979	0	0	87528	83178	351573	112639	339551
22DMC6	4110	3667	8448	2558	66043	167291	0	47500	602395	107622	127114	74218	29382	0	0	7935	9921	29691	14585	23871
ECYC5	13861	17052	12006	9988	62425	169493	0	45660	242485	1251303	88764	51776	17603	0	0	9518	6524	76507	20309	39675
25DMC6	1853	1733	4673	1394	23377	31901	0	13216	68743	164907	40886	18383	6414	0	0	3793	3575	19643	6934	14593
24DMC6	2364	2414	6106	1830	37215	49668	0	22244	102817	203149	67225	31975	11889	0	0	4979	5348	22051	8933	18694
223TMC5	5881	4909	8018	3042	183293	290791	0	115388	638310	796597	332321	175155	47758	0	0	11772	13449	65428	30779	35562
234TMC5	60	1	2	50	4166	5994	0	88	8336	20651	498730	275481	1738	0	0	75	407	1540	1248	2186
Toluene	19591	36694	47320	25702	10542	23844	0	37	123071	159829	9649	4726	33386	0	0	3146	21435	373458	19929	451232
2MC7	20025	26815	39528	19029	280639	369957	0	121994	880252	1950240	438858	152187	68146	0	0	39238	46463	168290	83535	159433
3MC7	10816	14667	20019	10895	166640	624614	0	169080	1240921	1985330	732212	338384	2962	0	0	2109	65341	45473	23642	46299
.c4DMCYC	3030	7071	3682	3329	113010	106101	0	47034	238435	0	199979	106733	19734	0	0	8240	11001	32887	21136	24782
nC8	191893	373185	254510	246482	529524	443095	0	75260	928244	2894993	454999	77150	82922	0	0	65762	97256	293912	152355	325961

	C529	C540	C548	C553	C557	C566	C574	C575	C579	C582	C589	C596	C711	C714	C721	C722	C725	C1386	C1387	C1388
iC5	8957	43712	131989	2646	1051	42419	5858	7507	128450	109902	7092	108	5277	8516	1059	16159	3360	514115	243917	262484
nC5	9343	74421	190458	6676	1560	62352	9939	10869	230196	143764	8229	31	11152	15920	898	23249	5219	851472	411995	460672
22DMC4	336	749	3203	803	317	1063	92	207	1816	1544	245	0	375	339	0	733	331	4812	3724	3180
CYC5	3374	18392	0	936	540	10735	1642	2383	44018	22071	1349	94	2211	2819	0	4055	881	0	0	159723
23DMC4	5511	9307	79737	1520	549	19517	1555	4362	34074	29614	2437	268	3501	3629	941	6905	3458	253968	135115	37364
2MC5	21789	95337	397918	20721	3236	124051	12897	24019	272158	155447	18500	739	30507	31126	7860	49197	24743	974015	499412	703469
3MC5	19579	77747	290090	16263	3253	108582	11214	20146	225116	128543	16015	937	23438	23246	6561	38776	19929	784000	386393	623535
nC6	29867	174752	696022	42973	8060	200914	23939	44143	456607	261384	16306	1272	78933	73221	20762	72728	38791	1914212	932405	1370787
22DMC5	6581	0	0	367	93	1815	165	298	3848	1604	357	110	1323	1049	448	1694	1100	0	0	0
MCYC5	33150	160212	334499	31060	6832	176850	26351	19257	443630	121982	25262	3768	48819	47378	9723	64299	30266	905845	491547	1243485
24DMC5	1943	8062	26547	2146	282	10826	1032	1174	16552	6198	1956	281	4525	3256	1324	5704	3361	37644	22506	42097
223TMC4	191	309	2214	25	136	545	0	123	668	450	107	0	287	378	0	360	262	2476	2031	2502
Benzene	320	187465	110116	7736	10800	40088	6773	53034	408767	86216	65	511	15416	27186	1298	295	38	2314197	829395	1036419
33DMC5	7619	2498	6143	257	0	2515	147	623	7216	1849	308	120	1155	879	303	1406	1034	7079	5051	5335
CYC6	43051	110777	539298	9957	9446	121497	12094	32561	250624	196536	14389	4591	75545	63779	15577	108190	52203	1229279	659497	790381
2MC6	21976	95411	364706	31618	6651	124182	10430	19463	238474	108889	14835	2879	47242	38934	22456	53828	41895	699249	367819	658773
23DMC5	15155	38452	121458	17317	3912	68055	6211	10635	73411	39889	10771	3397	25042	21722	11736	31417	23627	209709	110626	196525
11DMCYC5	5030	18268	34969	4679	558	21908	2936	1019	24427	4131	3533	1100	11447	8577	2869	11615	8044	42461	20218	80935
3MC6	36571	151034	485096	61454	11713	196295	19888	29857	343765	162350	28587	5904	75393	64372	36750	87691	67169	987340	503623	980124
.c3DMCYC5	16914	106027	123305	34390	4204	130778	17457	4572	258874	31742	20468	4671	32858	29823	12577	34743	25442	255637	154043	644370
Lt3DMCYC5	18707	99539	111735	30569	3936	116841	16567	4161	245238	29789	19189	4140	29685	28001	11140	32186	23253	235480	142377	603365
Lt2DMCYC5	45895	248773	270261	85044	11272	306841	42131	12514	612454	84329	50444	13753	72368	70150	29142	79027	57427	579003	345509	1380585
nC7	58627	322588	1015137	117778	28836	349575	34410	60677	661926	342723	20320	6020	178477	164883	88918	109633	89743	2136244	1008065	1897831
MCYC6	47691	269413	812230	69888	24502	317286	36827	44179	567970	256834	46759	21671	173451	154150	72601	216382	156446	1553881	772063	1526450
22DMC6	807	45334	78720	9452	1365	50797	5564	1148	105358	11809	8002	4160	20996	19371	8567	19538	17016	51627	34624	165959
ECYC5	12492	62001	107008	25006	6488	69634	9012	9304	135793	54448	10288	3874	20156	20336	12121	25730	19504	344646	152931	386730
25DMC6	4787	17156	50284	7112	2018	20810	1286	3503	39097	20051	2157	1866	6361	6077	5066	7649	7536	82959	38504	74252
24DMC6	6175	21013	61146	9081	2045	25823	1878	3325	46232	20067	3787	2375	9482	9134	6444	10733	10314	87967	41479	91227
223TMC5	20653	94114	123671	35238	5018	117732	11899	2934	246912	23220	19282	9829	32372	32067	17169	29648	27331	151326	89837	414838
234TMC5	17139	3474	5624	1109	3809	2932	3446	6421	7160	1594	261	278	921	0	476	999	763	6405	3648	8289
Toluene	3877	355820	269444	42386	36387	69802	4498	22056	561640	130594	29979	356	40104	43288	1316	376	453	1607882	520902	721874
2MC7	43531	130435	478278	94308	10277	209009	13044	24922	450567	158616	11705	14725	83120	71118	53854	67170	36674	748001	322695	781250
3MC7	20498	118489	77483	69868	1424	81291	16339	7620	135757	177234	21339	8711	16923	105	1945	2395	24616	831282	106038	178269
.c4DMCYC6	12270	52757	86864	31247	3068	64507	4989	1992	4694	7749	8360	11661	16699	18715	11053	16382	18099	71684	39203	163993
nC8	56265	280082	710465	168343	38773	268618	13872	40195	141066	146919	3612	13818	145739	119322	79793	60801	74981	1155940	449886	1023797

	C1389	C1390	C1465	C1466	C1467	C1468	C1469	C1470	C1471	C1472	C1473	C1705	C1715	D756	D800	D801	D802	D841	D842	D924
iC5	113082	0	1196558	971382	474917	818401	233635	581452	278050	513158	3432	41027	95602	5402	4972	2432	13262	31613	0	8451
nC5	210604	0	2600008	2686237	1358678	1722929	379404	1294519	742486	1078195	6660	208001	250854	10667	21438	6900	22587	31714	0	16851
22DMC4	1538	0	3786	3380	980	5122	1244	2906	1322	2780	22	2599	4149	511	142	88	622	689	0	255
CYC5	0	0	579471	786094	239276	489570	108393	389408	176044	246974	4113	30916	46113	1942	2015	703	4735	9605	0	3924
23DMC4	46708	0	142786	112187	61489	106345	38563	77499	23824	66299	1891	27275	40748	3420	1556	1149	4961	6919	0	2439
2MC5	294803	0	2267821	2957983	1284815	2052294	482539	1188430	973071	1166211	27179	331941	486101	31000	18069	11545	49710	56859	0	24383
3MC5	229265	0	1728583	2267937	960820	1655125	421690	1061767	784602	915932	29544	226386	286124	19927	11124	7371	30886	37328	0	14715
nC6	460116	0	5217965	8133985	3208362	4462517	841034	3060950	2223910	2359148	88550	791434	947644	91203	36836	21435	78486	68170	0	44911
22DMC5	0	0	0	0	0	0	0	0	323	0	0	6682	13900	2414	2696	565	1587	1498	0	856
MCYC5	429871	0	2458811	3678022	1508511	2408116	797751	1648916	2146024	1543941	136769	363735	481765	37035	16145	12303	54612	77291	0	27710
24DMC5	18592	0	75573	103539	43498	84290	23955	53496	58755	49437	6234	23174	39319	5781	1808	1288	4472	5993	0	2179
223TMC4	839	0	3702	4037	1325	4486	1314	3583	1916	2838	0	4030	4743	634	201	155	53	605	0	182
Benzene	120512	0	6092301	9548080	4136692	2494797	530983	2847729	784891	1414900	83027	881955	1138601	34619	22394	17878	33035	47120	0	15580
33DMC5	2537	0	12822	25773	9908	13834	2889	10320	6875	7765	1474	8219	10157	2391	380	505	813	919	0	514
CYC6	161530	0	2636983	3895748	1075681	2860137	573071	2030298	734317	1252318	93343	846008	960986	40390	17503	10136	38647	34413	0	20401
2MC6	225563	0	1561413	2567840	1051743	1542819	398076	1019452	1010105	805435	149359	847034	393230	66370	23176	13878	49939	55163	0	20139
23DMC5	70503	0	462533	653686	253285	501786	143906	344214	271156	267776	46756	563603	203790	34620	13105	11777	38493	45105	0	14047
11DMCYC5	36234	0	49109	64131	31464	75101	40746	46881	121699	64498	21896	30939	48473	9407	3443	2221	7471	6049	0	3251
3MC6	347866	0	2263386	3756242	1546617	2268744	627831	1539088	1613518	1222403	249257	819019	461991	79042	27968	19504	63884	70166	0	25878
.c3DMCYC	285590	0	543569	878381	542555	701858	364721	427734	1408259	596603	197474	199864	133757	22427	9673	7783	27955	32000	0	8396
Lt3DMCYC	261537	0	503221	810070	494237	650558	343407	398164	1316310	556810	192158	160532	121386	21220	9316	7191	25190	29353	0	7958
Lt2DMCYC	604780	0	1036890	1529074	905517	1426052	776015	866385	2553040	1181387	438366	407121	266351	52303	22100	18807	68680	88219	0	20488
nC7	492933	0	5706334	10000000	4302291	4557135	1077448	3647116	3340468	2495998	601919	2144461	1176033	323343	76658	49575	153370	114553	0	58695
MCYC6	466706	0	2834614	4504933	1772435	3142540	1020456	2252173	2231750	1705678	519397	2091279	1648796	228625	76332	46989	152640	144094	0	56494
22DMC6	84150	0	90940	130251	89035	127023	86799	82551	310754	123825	84951	73195	69240	22221	7571	5692	13209	15915	0	4877
ECYC5	110914	0	800801	1398972	604422	722786	266275	559915	697779	437922	146581	271073	97776	15186	4659	4545	18184	21607	0	4240
25DMC6	23758	0	176626	282292	140321	152926	48972	117264	96924	80953	33571	86762	53659	16749	4347	3757	10672	15634	0	3308
24DMC6	32721	0	169698	267855	136275	166759	57372	121577	131868	91432	42471	95492	63173	20347	5299	3615	11140	15393	0	3807
223TMC5	193400	0	257145	394611	286521	359116	229209	230357	826541	325875	232521	118049	82199	22805	8167	7431	21282	28822	0	5606
234TMC5	3079	0	11004	16079	8568	12463	5336	9068	12476	7367	4546	11178	5708	349	67	174	569	77	0	227
Toluene	62401	0	4871332	8742651	4149725	1433743	248074	2227438	545930	766321	182539	2643579	880246	52657	23701	20740	24856	48125	0	18940
2MC7	227620	0	1666263	3004334	1663098	1351206	490874	1118179	1211391	786463	444076	964572	451412	149927	41323	36040	102987	101967	0	27841
3MC7	362127	0	322287	387331	216886	371093	391009	164334	1048553	948323	627319	242720	154052	191992	48696	38684	112443	116823	0	31146
.c4DMCYC	69913	0	111321	161702	116174	139798	92342	103817	285370	120309	117589	117265	85255	34791	10203	8152	22354	25549	0	5246
nC8	165114	0	3008788	5779488	3268254	1781692	573637	1852102	1270570	1056009	582878	2453420	877285	439879	72413	55308	153891	121880	0	42449

	D1173	D1273	D1274	D1275	D1276	D1288	D1289	D1290	D1291	D1312	D1313	D1335	D1364	D1365	D1385	D2471	D2472	D2595	D2626	D2885
iC5	94538	134276	45552	7000	45704	96470	59498	104537	200217	258406	351958	461452	100508	5797	595958	3172	165	36759	7501	6862
nC5	116792	235331	92009	16739	93538	215553	146041	220380	421189	146856	661085	800296	217707	12654	1055444	2195	180	43322	15365	386
22DMC4	2515	4568	2953	622	4378	6791	3653	7086	9028	9360	20688	14151	8503	363	11516	186	13	2150	483	2221
CYC5	0	0	10559	0	0	0	0	0	0	0	0	0	0	0	0	1174	64	5611	2134	2761
23DMC4	44669	60570	25450	6876	30431	38533	32194	35707	78452	89184	170147	182316	53121	3494	222786	1280	51	8003	1790	18787
2MC5	265428	465151	432984	71264	188824	313104	226835	299757	453538	422297	1058997	1004955	458063	44591	1291374	6930	417	53887	12362	13183
3MC5	156943	259735	254604	41884	111036	176607	130229	168276	258305	277420	587529	564294	258889	24816	702321	5441	294	33019	7415	52873
nC6	345932	722921	787223	210766	414372	877412	693981	829083	1171081	122451	2544993	1655343	1049220	83160	2055890	7232	841	90242	39545	1004
22DMC5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	490	30	2633	547	27588
MCYC5	239458	424697	429296	66332	141408	247979	193259	219692	381911	270838	944988	949440	389869	38365	1120450	11762	746	43886	15779	2980
24DMC5	28710	38661	49526	14362	20405	36637	30263	36797	45738	26002	101115	69578	48869	5345	79447	1059	60	5293	1433	429
223TMC4	1956	5060	6311	1253	1840	3730	2783	3673	4571	5961	11342	8631	5080	515	8098	331	25	781	174	22035
Benzene	64017	166134	347830	55992	46494	127127	103313	94026	120542	20455	441902	293092	213658	22240	297055	5042	453	26046	3783	779
33DMC5	4509	8040	9764	3635	5112	10187	8461	10634	12336	8326	26625	15520	13219	1427	15251	307	31	1781	358	26200
CYC6	131683	297777	242082	57485	114367	237965	186537	216561	342043	197081	770259	688593	319845	28979	740591	13894	751	38833	12248	25441
2MC6	281712	433597	648771	184262	132689	307062	292969	320913	405652	85512	938656	679974	461418	63716	797576	6588	571	36876	10183	18820
23DMC5	121186	166379	297520	48655	63430	59628	78598	66707	102452	184987	277863	220472	135735	23898	300049	10858	473	25265	4784	4145
11DMCYC5	25677	37231	45047	14598	22614	41615	32950	34699	48640	29721	106732	83961	31760	7817	83181	2714	113	7139	1494	32172
3MC6	294136	447217	713919	179412	160804	300621	289237	291614	383217	153759	928245	729612	451893	67533	824554	10650	762	43139	11110	13002
.c3DMCYC	106089	175473	301995	46609	49822	82200	76414	71568	109551	82302	323577	296911	141932	21760	321661	6350	339	16401	4955	12391
Lt3DMCYC	99094	165827	283042	43329	46334	78546	72990	68034	105280	84163	311929	285277	135534	20524	302644	6110	331	15320	4562	35575
Lt2DMCYC	227195	387425	684343	93617	100909	158404	146447	136985	217112	309621	697897	660559	304431	49132	674999	22170	851	41754	11735	148223
nC7	497407	955606	1739653	729702	410079	1257250	1343104	1168100	1385452	48925	3682525	1614640	1374752	158996	1757874	16678	2207	120856	68748	104587
MCYC6	575866	959877	1315354	393576	361503	790420	766847	727803	962672	333469	2422692	1679247	1024496	143107	1710729	47789	2635	111824	39098	7216
22DMC6	66692	82710	128525	39064	23745	48816	52109	46875	58899	36277	185419	118203	73669	13468	111621	4971	225	9005	2293	11180
ECYC5	52465	93598	201564	29034	19975	40437	42712	32772	51875	13598	213665	147661	69919	12147	157453	3430	218	8831	3785	5291
25DMC6	55512	55007	110217	36966	15461	33889	41118	35036	40978	12806	111060	71207	50935	10991	75440	1997	106	6103	1399	5980
24DMC6	59170	60490	106272	41296	18879	43335	51022	44356	51718	13309	129160	80640	59325	11850	82997	2231	116	6744	1828	10558
223TMC5	83314	103959	201611	42293	22458	42391	46989	37410	52603	37278	206961	136469	76302	16886	132962	6370	0	10089	3183	8952
234TMC5	2699	3696	6949	154	482	106	408	140	1230	277	7665	4874	2346	79	4649	260	0	13978	44	62556
Toluene	88237	68240	151332	84057	18126	121548	146084	81012	87322	17104	1383005	99678	147039	16991	86862	6110	0	25514	8275	46430
2MC7	348425	419138	941268	332090	93286	247708	343356	237813	290217	17944	830669	505404	363326	87462	529579	13261	230	44236	11982	18220
3MC7	429471	455596	887111	327854	117778	305775	392986	294141	353295	52356	894599	583929	415171	93297	572842	2827	39	13579	2475	5349
.c4DMCYC	71211	75179	148247	45491	17892	44760	54770	41632	51108	13193	124735	94990	64581	15773	85925	6037	0	9959	2624	186021
nC8	339542	460709	1192059	675552	114918	578157	900782	508473	555092	18597	1662119	585710	582640	100817	554837	35353	0	99851	66224	0

Saturate fraction of the sample set

	A549	A550	A920	A1140	A1710	A1711	A1712	A1723	A1724	A1725	A2268	A2269	A2270	A2283	A2284	A2313	A2362	A2363	A2364	A2424
C13	227	1147	1942	21200	13293	14058	6716	10292	7667	11180	7799	4758	7164	9767	9535	8193	15159	10441	10955	29309
C14	1161	5027	4948	19138	12853	14260	10711	9377	9345	10356	5801	3755	6038	8105	9092	9372	12212	11081	11762	31584
C15	88	12950	9226	18368	12896	14466	14121	8756	10149	9877	5204	3530	5710	7605	9244	9393	11261	11348	12202	39173
C16	4085	22813	12767	15706	10628	12329	13577	8118	9239	9147	4807	2931	5645	6518	8369	8411	9174	9887	10051	37710
C17	7506	39446	20697	19893	14397	17445	16470	10167	13317	11696	5687	3655	6894	7520	10504	10220	11056	11760	13694	49958
Pr	593	791	1554	1477	1003	1292	1007	906	1109	886	352	247	467	234	315	313	308	815	1021	3433
C18	1782	16639	6171	4523	3610	3894	9630	2032	2809	2549	1660	1008	1723	2313	2697	3271	3258	3448	3152	12418
Ph	553	1570	1449	1473	929	1189	561	610	803	748	309	141	355	342	516	438	268	529	666	3642
C19	4562	24556	11512	7503	5625	7763	11056	3480	4708	4116	2246	1423	3091	3074	4291	4007	4401	4731	6304	1803
C20	1038	9134	3578	2159	1849	1898	7612	961	1401	1247	917	579	829	1438	1571	1986	2078	2126	1800	6074
C21	615	5426	2769	107	1350	1367	6686	674	1028	878	759	440	657	1131	1118	1620	1659	1679	1206	4446
C22	442	4350	2256	1135	1001	984	5750	483	748	712	603	361	474	1028	955	1412	1451	1488	977	4047
C23	373	3440	2455	103	875	955	5138	452	649	614	581	323	449	927	831	1281	1314	1363	877	2792
C24	314	3508	2487	1086	809	697	4583	326	537	475	510	277	404	795	701	1078	1216	1118	704	2397
C25	252	2759	2182	793	663	695	3835	291	461	423	441	229	332	720	617	938	1032	1025	622	2580
C26	221	2507	2124	757	533	660	3210	284	432	401	399	201	290	660	548	809	902	901	557	1869
C27	193	2027	1916	641	543	580	2783	230	389	360	343	175	262	576	477	806	829	799	546	1496
C28	180	2242	1700	575	472	516	2303	195	320	305	305	144	244	515	417	646	745	677	456	1235
C29	159	1526	1429	591	476	526	1718	190	347	297	248	117	194	401	341	498	556	523	386	1059
C30	131	995	1206	404	332	328	1135	160	228	177	194	73	130	317	247	378	439	416	284	1078
C31	99	860	835	332	308	302	885	151	227	198	152	71	132	0	0	336	371	337	299	754
C32	0	638	691	273	243	0	647	118	0	0	0	81	118	0	0	285	295	322	248	656

	A2425	A2426	A2427	A2428	A2429	A2430	A2431	A2432	A2433	A2434	A2435	A2436	A2468	A2469	A2470	A2611	A2627	A2706	A2884	A2892
C13	21385	34368	5446	52818	11053	14303	25091	16490	10133	10434	20542	9057	16221	31823	47419	2838	7088	2676	15745	8557
C14	26845	32780	18940	42160	14391	22309	22118	14926	13825	18065	30177	15778	13680	31541	40905	4145	8210	5827	16910	10251
C15	35457	37356	36897	44866	20169	32431	24593	13984	20096	29226	42761	24112	13857	36703	46117	6084	10740	9134	20934	11842
C16	34940	34190	40935	38064	20166	32758	21008	13655	20687	30383	44995	25631	11684	37679	43993	7143	10455	9707	19120	13362
C17	46438	44769	58831	48215	27760	45418	25776	17535	29076	44615	59261	36278	13441	47857	53529	10207	14091	14130	25405	18016
Pr	3160	3530	4547	3022	1872	2771	2007	1393	1931	3189	3483	2617	1120	3300	4258	499	1314	844	1332	798
C18	11426	11521	7084	11241	8217	11799	6161	4601	7554	11941	14602	9456	4099	12321	16933	2607	6211	3325	6341	8026
Ph	3028	2936	4055	3416	1821	3154	1602	1103	2097	2932	4122	2418	1058	2748	3697	674	899	802	1034	1064
C19	20486	21383	26590	20587	12850	20526	10809	7057	14686	22585	27592	17473	6223	22824	25494	5942	9007	6319	11907	10412
C20	5701	5784	9490	4776	4977	5700	2845	2381	4132	5759	7474	5415	2428	6064	10747	1397	4920	1659	3396	5128
C21	4019	4162	7384	3367	4004	4058	1959	1958	3087	4283	5233	4028	1863	3966	8469	1110	4892	1394	2466	4431
C22	3490	3880	6525	2945	3455	3594	1787	1730	2579	3825	4634	3636	1620	3047	7239	1097	4969	1291	2210	4313
C23	2474	2792	5294	2139	2860	2521	1314	1393	2020	2716	3237	2634	1467	2679	6567	846	4581	1043	1769	3974
C24	2098	2450	4367	1854	2598	2154	1123	1241	1786	2313	2804	2361	1270	2202	5697	798	4524	981	1602	3703
C25	2331	2543	4386	2012	2701	2287	1043	1220	2057	2472	3175	2531	1088	2029	5209	775	4245	970	1611	3485
C26	1560	1918	3320	1474	2085	1418	746	831	1427	1895	2329	1902	930	2030	4732	798	4029	817	1227	3027
C27	1244	1507	2772	1351	1642	1403	639	673	1128	1415	1978	1462	819	1831	4479	788	3631	738	1105	2718
C28	896	1142	2156	944	1288	931	453	488	892	1029	1340	1068	640	1592	3869	704	3264	641	833	2322
C29	940	1150	2233	1119	1379	979	561	392	992	1156	1582	1169	442	1172	3139	842	2887	636	761	2076
C30	803	1071	1716	909	1086	903	427	361	861	991	1367	1071	330	1017	2384	775	2444	546	738	1688
C31	589	743	1544	799	827	616	278	256	602	677	1106	870	0	858	2028	608	1871	387	561	1332
C32	483	582	857	478	613	437	204	176	413	492	645	497	0	0	1616	508	1341	340	391	1002

	A2895	A2896	A2897	A2898	B515	B554	B1014	B1279	B1393	B1443	B2121	B2122	B2887	B1873	B1874	C495	C499	C503	C511	C513
C13	30812	15904	19295	3980	5227	5682	5330	2407	10008	2071	1970	1404	2572	4309	4134	2676	5809	7574	3262	10151
C14	28845	15942	18796	7913	7161	7936	5288	2969	9573	3944	2126	1950	3638	4058	4490	4647	7358	10373	8157	14615
C15	33834	17093	22973	11991	8710	9558	4820	3378	9616	5529	2250	2315	3464	4231	4543	5769	7590	11459	13731	16419
C16	31975	18994	22106	15456	7545	8459	3658	2371	7796	6195	1808	2136	3423	3115	2922	6193	7253	10907	18209	16952
C17	42396	26529	29212	23413	6575	8189	3443	2170	6644	5611	1584	2330	3665	2931	3126	5948	6555	9813	20203	15975
Pr	2245	1244	1814	1118	3919	5415	1263	1628	5395	3747	930	1478	1495	3588	3930	1922	2101	3794	7274	4732
C18	11566	7277	11347	6444	5384	7101	2624	1917	5894	5256	1279	1863	2845	2297	2421	5865	6263	9207	22369	15396
Ph	2136	1392	1385	1322	2590	3668	860	1205	3900	4228	613	780	1103	1967	2278	2937	2371	6019	12715	7790
C19	20784	12440	16462	11473	4492	6584	2284	1201	5411	5001	1400	1720	3269	2373	1920	5585	5572	9212	25010	15915
C20	6123	4101	6682	3930	3393	5749	1783	970	4053	4199	924	1297	2294	1783	1819	5030	4939	7939	23780	13403
C21	4424	3068	5587	2998	2263	3870	1343	677	3332	3545	777	1039	1940	1485	1572	4176	4006	6306	20864	10905
C22	4207	2761	5395	2655	1691	3277	1041	601	2542	3089	615	841	1702	1236	1327	3827	3597	5980	19452	10109
C23	3329	2308	4524	2199	1190	2605	854	485	2048	2169	502	692	1364	1039	1137	3179	2993	4907	17767	8411
C24	3130	2154	4323	2008	912	1865	675	409	1602	1915	406	561	1106	863	947	2874	2632	4431	16119	7796
C25	3137	2202	4197	375	731	1569	501	484	1176	1783	319	443	891	791	837	2394	2228	3569	12732	6358
C26	2510	1815	3713	1650	589	1198	423	316	1088	1531	268	363	753	684	730	2265	2036	3347	12969	6011
C27	2250	1552	3192	1395	428	948	318	251	917	993	188	284	573	548	591	1863	1648	2475	10781	4712
C28	1768	1167	2881	1082	377	736	300	269	745	966	166	269	425	508	551	1703	1566	2331	9793	4291
C29	1715	1113	2285	1035	275	610	301	258	671	655	129	230	366	442	503	1416	1288	1669	7961	3414
C30	1446	1031	2124	956	185	458	209	149	517	578	89	163	348	318	327	1269	1100	1350	7105	2919
C31	1138	783	1808	766	0	412	190	98	349	342	67	128	227	284	297	980	817	965	5182	2258
C32	176	585	1305	596	0	0	0	0	328	331	0	0	176	208	259	836	566	704	4860	2182

	C529	C540	C548	C553	C557	C566	C574	C575	C579	C582	C589	C596	C711	C714	C721	C722	C725	C1386	C1387	C1388
C13	1740	921	15825	18484	5409	229	4465	7219	12002	4671	536	4651	3760	5521	12007	4292	2875	4082	4307	14228
C14	5568	95	16192	24206	12127	841	8596	13304	12711	7680	1396	5626	5468	7952	18016	6995	5102	7136	8950	17572
C15	11039	12546	15105	27607	17886	2718	11191	16893	12320	9343	2191	6133	6360	8656	21681	7881	6000	9145	11552	18933
C16	15510	1384	13554	8182	12442	5774	10855	13564	10727	9948	2571	5888	6020	8340	20941	7154	5652	9783	12930	17639
C17	19461	10799	11548	23546	21078	8937	10625	18072	9609	9405	2044	5325	5798	8019	18963	6157	4993	9493	13429	16661
Pr	7988	6820	3356	16204	7980	3794	7004	6447	5406	3271	5155	2560	1999	2742	5801	2743	2286	3233	4723	7659
C18	22578	9655	10865	20992	22686	11701	10148	18606	8137	9445	3059	4974	5330	7195	12020	5608	4524	9506	13485	16890
Ph	13514	7208	5595	16374	16509	8065	8450	13940	6759	6922	5588	3924	2460	3302	7410	3999	3430	6652	8213	12493
C19	24071	958	10018	20123	23127	14583	10043	17560	7396	9565	3731	4524	5510	7113	10649	5001	3711	9351	14424	16908
C20	23994	0	8588	15810	20831	15304	8935	17109	6945	8842	3235	4315	4383	5380	9697	3230	2664	8803	12631	13995
C21	20800	442	6765	12266	16992	13703	7391	14065	5464	7185	2895	3419	3683	4616	8797	2706	2123	7408	10856	11770
C22	20622	272	6296	10769	16573	14399	7011	13755	4923	7055	3092	3250	3226	4222	8431	2281	1788	6965	9763	10555
C23	17080	7775	5200	8179	13966	12812	5603	11467	3925	5679	2193	2712	2704	3696	7181	1498	1086	5620	7901	8814
C24	15406	532	4768	6855	12360	11732	5103	10422	3520	5334	2018	2533	2422	3369	7434	1333	950	5550	7889	8131
C25	12536	6214	3802	5779	10263	9841	4019	8250	2864	4185	1821	2120	1953	2932	6016	1162	859	4456	5657	6474
C26	11991	104	3633	4874	9508	9851	3787	8087	2797	4373	1580	2032	1866	2744	5884	1266	901	4148	4952	5928
C27	10086	620	2845	3505	7404	8277	2715	6183	2020	3268	1368	1512	1458	2332	5085	1117	802	3392	3519	4991
C28	9437	506	2568	3082	7174	8647	2419	5746	1878	2810	1399	1505	1464	2187	4956	1290	931	3222	2849	4689
C29	6981	0	1926	2358	4877	6446	1913	4266	1478	1985	1109	1090	1171	1739	3745	1119	773	2375	1787	3516
C30	6306	4071	1633	1848	4210	5711	1535	3715	1326	1679	1033	969	1008	1610	3268	1046	728	2070	1357	3123
C31	4949	0	1146	1207	2913	4198	1006	2637	971	1115	866	729	602	1032	1850	614	481	1182	675	1877
C32	4497	3309	842	839	2572	3855	889	2532	887	1004	821	638	568	1011	1749	602	488	1150	613	1629

	C1389	C1390	C1465	C1466	C1467	C1468	C1469	C1470	C1471	C1472	C1473	C1705	C1715	D756	D800	D801	D802	D841	D842	D924
C13	4227	2415	0	0	0	0	0	0	137	0	0	5757	4960	7269	7853	5533	815	5854	3672	1865
C14	6970	5903	0	0	311	132	1016	0	382	263	274	7751	6206	8859	9855	8059	2450	6289	3861	2858
C15	8454	8644	1974	904	1022	767	3651	0	1069	1030	787	7729	6556	9905	14084	14665	5279	8055	4965	3872
C16	8145	8962	2451	1643	1349	1364	5360	780	1659	1361	1099	7756	6041	7581	11473	12789	5669	6005	3634	3127
C17	7899	10326	3005	2274	1504	1718	5954	1563	2135	1656	1428	7422	5938	8711	12114	13454	7879	7051	4320	3808
Pr	4883	7873	1147	831	716	713	2710	642	1474	808	959	1872	1772	3082	8277	14000	4597	6993	4109	2261
C18	7902	11223	2769	2218	1348	1636	6274	1717	1987	1529	1282	6910	5629	5998	9670	12353	5709	5457	3068	2833
Ph	5928	9276	2117	1647	1231	1238	4981	1252	1763	1241	1140	2878	2183	4363	10640	15474	6003	9051	5331	3314
C19	7086	9558	2946	2381	1470	1890	6923	1958	1808	1695	1228	6885	5668	6351	10900	10190	5804	3089	2527	2381
C20	6115	10100	2551	2067	1231	1583	6036	1704	1848	1397	1168	6094	5003	4521	7869	9710	5214	2977	2149	2080
C21	4955	8034	2199	1773	1049	1370	4869	1458	1561	1141	977	5287	4595	4419	7818	9242	5550	2634	2161	1980
C22	4145	7039	2151	1747	1059	1384	4511	1306	1489	1154	958	4964	4406	4142	7610	8780	5294	2702	2146	1956
C23	3212	5490	1826	1510	853	1191	3464	1233	995	1003	771	4340	3912	3962	7205	8447	5209	2812	2121	1874
C24	2967	5195	1637	1336	696	1069	3664	1093	970	894	662	3931	3635	4065	7475	8820	5152	3148	2352	2047
C25	2284	4558	1303	1096	550	827	2906	850	845	744	537	3473	3189	3938	8039	10010	5213	3973	2677	2171
C26	1967	3621	1174	983	601	791	2846	748	684	697	463	3326	2905	3392	6204	9280	3338	2974	1884	1722
C27	1509	3190	1016	807	405	666	2396	635	559	592	364	3045	2376	3143	5575	8466	2655	2783	1703	1502
C28	1386	3228	1063	822	388	724	2493	594	515	615	364	3074	2055	2383	4092	7368	1713	2088	1183	1209
C29	1106	2749	737	608	293	513	1965	464	366	490	269	2840	1607	1846	3125	7491	1177	1654	901	777
C30	995	3028	758	637	295	550	2119	421	355	523	276	2251	867	1683	2170	5359	979	1794	970	846
C31	703	2252	500	451	176	377	1679	263	212	392	168	1899	639	852	1287	3989	386	1088	587	621
C32	758	2497	497	538	0	398	1852	277	0	450	153	1984	543	900	729	2820	1291	753	470	366

	D1173	D1273	D1274	D1275	D1276	D1288	D1289	D1290	D1291	D1312	D1313	D1335	D1364	D1365	D1385	D2471	D2472	D2595	D2626	D2885
C13	2858	1879	5069	4195	3989	4665	5499	3818	2463	4347	6753	2733	4602	6611	12827	15786	4616	1733	4348	7943
C14	4003	2440	6455	4335	5027	5287	5648	4228	2878	6303	7186	3461	158	8740	18449	23301	7563	3447	4567	10509
C15	5313	3128	8124	5353	5386	6199	6244	4872	3354	8070	7823	4427	7798	10739	22939	28660	9830	4917	5985	13650
C16	3970	2325	5821	3866	4044	4666	4587	3488	2514	5827	5991	3253	6420	7959	17488	23549	7426	4598	6210	9843
C17	4675	2853	7799	4530	4750	5098	4912	4122	3002	7398	6724	3881	7556	9196	21138	27983	8365	5988	9171	12688
Pr	3877	1495	4325	1565	1499	1070	1080	816	709	5126	1500	2010	2875	4712	9384	23230	4582	1952	995	3723
C18	3458	2029	5578	3059	3223	3276	3024	2419	1854	4982	4301	2733	5308	6812	15243	22205	5710	4244	3802	8166
Ph	4730	2081	6181	2011	1952	1426	1323	879	805	6815	1794	2674	3979	6919	12882	33264	5076	2296	893	4704
C19	2061	2156	2892	2797	2990	2956	2892	2346	1850	1797	3974	2492	5095	6471	16050	22667	5515	5107	6243	8627
C20	1784	1164	2615	1864	1748	1838	2346	1753	1454	1553	2848	1625	3822	4821	11586	17432	4306	4139	3123	6337
C21	1818	1107	2689	1532	1538	1769	1435	1627	1393	1797	1348	1753	3744	5069	11272	17542	4165	4428	3166	6499
C22	1760	1054	2434	1452	1491	1786	1508	1491	1306	2182	1529	1784	3617	4992	10650	17295	3943	4580	3432	6756
C23	1835	1149	2679	1432	1502	1806	1609	1445	1248	2667	1668	1848	3632	4931	10023	17715	3863	4622	3261	6464
C24	2185	1264	3242	1531	1573	1835	1721	1369	1233	3251	1909	1994	3731	5237	10618	18434	3642	4959	3425	6794
C25	2569	1391	3790	1619	1640	1813	1689	1295	1162	4075	2006	2188	3926	5809	9621	17690	3384	4991	3537	6629
C26	1848	1256	3127	1518	1552	1655	1601	1151	1051	3242	2005	1728	3109	4380	8263	15810	2852	4495	3285	5472
C27	1772	1131	3037	1335	1367	1568	1526	1063	936	3176	1952	1571	2919	4013	7363	14309	2604	4168	3055	4698
C28	1296	901	2328	1124	1200	1299	1281	913	779	2271	1652	1168	2221	3000	5833	9392	1974	3441	2700	3029
C29	1210	883	2306	1120	1188	1092	1078	729	691	2300	1521	1082	1947	2923	4620	7537	1426	3326	2799	2510
C30	773	501	1342	680	761	728	714	521	465	1172	1018	563	1184	1409	4586	5787	1015	2575	2222	2437
C31	639	324	1147	563	630	550	551	415	346	1280	799	414	959	1533	1795	5202	987	1847	1697	1438
C32	408	706	2196	437	511	399	399	300	267	3215	645	282	897	1025	1707	18482	765	1448	1194	1050