Revista GEOCIÊNCIAS

unesp

ISSN: 1980-900X (online)

DATA MINING IN ORGANIC GEOCHEMISTRY: CASE STUDY IN POTIGUAR BASIN

MINERAÇÃO DE DADOS NA GEOQUÍMICA ORGÂNICA: ESTUDO DE CASO NA BACIA POTIGUAR

Sarah BARRÓN TORRES[1], Ítalo de Oliveira MATIAS[1], Francisco Fábio de Araújo PONTE[1], Erica Tavares de MORAIS[2], Ygor dos Santos ROCHA[2], Mario Duncan RANGEL[2], Fabiano Galdino LEAL[2]

[1]Pontifical Catholic University (PUC-Rio), Informatics Division, Software Engineering Laboratory (LES). Rua Marquês de São Vicente, 225 - Gávea, Rio de Janeiro - RJ, Brazil. E-mails: sarah.barron@les.inf.puc-rio.br; italo.matias@les.inf.puc-rio.br; fabioponte@les.inf.puc-rio.br

[2]Petrobras Research and Development Center (CENPES). Avenida Horácio Macedo, 950 – Cidade Universitária, Ilha do Fundão, Rio de Janeiro – RJ, Brazil. E-mails: ericat@petrobras.com.br; ygor.rocha@petrobras.com.br; mduncan@petrobras.com.br; fabianoleal@petrobras.com.br

RESUMO - A quantidade de dados provenientes de análises geoquímicas de amostras coletadas em poços de petróleo cresce simultaneamente ao investimento no setor de exploração e produção. Por outro lado, o tratamento e a interpretação desses resultados ainda são muito dependentes de especialistas, e demandam tempo. Com a geração de extensas bases de dados, a mineração de dados se apresenta como uma boa alternativa para explorá-los por meio de métodos estatísticos e computacionais, proporcionando diferencial tecnológico e agilidade ao sistema. De forma experimental, com dados de 200 óleos da Bacia Potiguar, essas ferramentas foram implementadas, com a consequente sugestão de um fluxo de trabalho que, ao final, pôde retornar uma precisão razoável na previsão da classificação genética das amostras. Usando escalonamento multidimensional (MDS) e agrupamentos (dos tipos dendrograma e *k-means*), de 60 atributos iniciais, o conjunto ideal foi reduzido para 26. Aplicando aprendizado de máquinas, 92,50% de acurácia mediana foram obtidos no algoritmo de Árvore de Decisão, 95,00% na Floresta Aleatória e 87,50% em Rede Neural Artificial. Comparando a uma análise previamente apresentada na literatura pertinente, os benefícios em termos de eficiência podem ser percebidos com a adoção da metodologia aqui proposta.
Palavras-chave: Geoquímica orgânica. Mineração de dados. Estatística multivariada. Fluxo de Trabalho.

ABSTRACT - The amount of data from geochemical analysis using samples collected in oil wells grows simultaneously to the investment in the exploration and production sector. On the other hand, the treatment and interpretation of these results are still very dependent on experts and demand time. With the generation of extensive databases, data mining presents itself as a good alternative to explore them through statistical methods and computational algorithms, providing technological differential and agility to the system. In an experimental way, with data from 200 oils from the Potiguar Basin, these tools were implemented, with the consequent suggestion of a workflow that would, in the end, return a reasonable accuracy in predicting their genetic classification. Using multidimensional scaling (MDS) and clustering (dendrogram and k-means types), from 60 initial attributes, the optimal set was reduced to 26. Applying Machine Learning, 92.50% of median accuracy were obtained in the Decision Tree algorithm, 95.00% in Random Forest and 87.50% in Artificial Neural Network. Comparing to an analysis previously presented at the pertinent literature, the benefits in terms of efficiency can be realized with the adoption of the methodology herein proposed.
Keywords: Organic geochemistry. Data mining. Multivariate statistics. Workflow.

INTRODUCTION

The dynamics of the processes in the oil and gas industry in the search for new exploratory frontiers, reduction of risks and productivity gains, highlights the importance of having advanced technologies applied to the interpretation and treatment of a growing volume of data. Scientific contributions provide an environment with greater speed and reliability of information, which are essential to add value to the chain.

Organic geochemistry is the science that concerns the distribution of the carbon element in animals and plants, being, therefore, of unquestionable relevance to the oil sector. It includes the petroleum geochemistry, which studies the origin, generation, migration, accumulation and alteration of this fossil fuel (Navarro, 2008), using knowledge from organic chemistry to petroleum geology.

Petroleum is a mixture of hydrocarbons in solid, liquid and gaseous states, as well as of heteroatomic compounds and metals in lower

levels. It is from the study of these fractions that the geochemical variables of interest are extracted.

Among its components there are biomarkers, also called biological markers, related to molecules originally biosynthesized by living organisms and which kept their carbon skeletons practically unchanged during the stages of organic matter transformation (Peters & Moldowan, 1993; Philp, 1985). These are the species that provide the greatest number of analytical parameters. The chemical profiles obtained from biomarkers are unique because of the possibility of association with the respective biological precursors (Philp, 1985), and with the variety of geological conditions imposed on the oil generation.

Other tools are also used for characterizing oil systems in organic geochemistry, as follows: determination of stable isotopes of carbon, hydrogen and sulfur; determination of the total organic carbon (TOC) content in rocks and the corresponding hydrogen index; simulation of artificial maturation in open systems (pyrolysis); gas and liquid chromatography; optical microscopy to assess trapped fluid inclusions; organic petrography.

Considering direct and indirect methods of investigation, the petroleum geochemistry is important for the worldwide exploration and production of oil and gas, and also for the discovery and evaluation of deposits, as an aid to the recognition of the type of generating rock, quality, biodegradation, thermal maturation and depositional paleoenvironment.

The complete examination of oil samples is a time-consuming and complex task even for highly specialized professionals; consequently, the treatment and the geochemical interpretation of these results become more challenging and demanding. The high number of parameters produced, derived from multiple analytical techniques, requires a thorough knowledge of the available data.

Achieving a limited number of independent attributes (the key requirement of a regression model) without losing significant information is useful to facilitate data processing. This means that the inclusion of new attributes will return an insignificant improvement in results (Nasser Junior, 2009). In other words, using mathematical/statistical approaches to select some variables that represents the whole set gives greater assertiveness to the interpretations.

In this scenario, the need for data mining (DM), the prime objective of this paper, is reinforced, including the exploratory data analysis (EDA). Therefore, a case study is built focusing on reclassifying the origin of oils from the Potiguar Basin proposed by Morais (2007).

The obtained results are expected to encourage innovative digital transformation and to improve petroleum industry R&D methods to ensure broader understanding of Geochemistry.

METHODOLOGY

The geochemical databases contain data generated using different techniques applied to samples acquired as a result of oil exploration and production. The employed equipment varies according to the final data to be obtained in each analysis.

Oil samples and organic extracts from rocks can be analyzed by GC-FID (gas chromatography with flame-ionization detection) utilizing the stable carbon isotope technique (that obtains the parameter δ13C) and sulfur content.

Due to the chemical complexity of oil, fractions of compounds with similar properties are divided; liquid chromatography (LC) is one of the fractionation techniques that uses the differences in polarity between molecules to separate them and thus obtain the percentage distribution of classes of saturated hydrocarbons, aromatic hydrocarbons and polar compounds (represented by the acronym NSO, indicative of the heteroatoms commonly present). The saturated hydrocarbon fractions are analyzed by GC-MS (gas chromatography coupled to a mass spectrometry) to check saturated biomarkers.

In order to suggest a workflow for geochemical analysis, public data containing oil samples from the terrestrial portion of the Potiguar Basin were used, reproducing the research for the genetic classification of Morais (2007). The dataset consists of 200 samples (categorical or dependent attribute) representative of the different classes of oil listed there.

The geochemical characteristics of these samples reflect several factors, ranging from their origin to processes of secondary alteration and mixtures of oils of different origins in different proportions.

From the variables generated in the

geochemical study using the LC, GC, GC-MS techniques, total carbon isotopes, API and sulfur percentage, 60 predictive (or independent) attributes were selected, as shown in table 1.

The pertinent bibliography points to the existence of three main classes of oils in the emerged extension of the Potiguar Basin: Lacustrine, Siliciclastic Lacustrine and Mixed. The samples classified as mixed include marine-evaporitic oils and mixtures.

Table 1 - Selected geochemical variables for analysis grouped by color according to the used technique (Modified from Morais, 2007).

TYPE	VARIABLE
Gas Chromatography Parameters	PRI/PHY
	Pri/nC17
	PHY/nC18
	17/(17+C27)
	Total Height
Liquid Chromatography Parameters	%Saturated
	%Aromatic
	%NSO
Total Carbon Isotope	δ13C
Bulk Parameters	API
	% Sulfur
Biomarkers Parameters	%27ββS218
	%28ββS218
	%29ββS218
	%H31
	%H32
	%H33
	%H34
	%H35
	19/23TRI

TYPE	VARIABLE
Biomarkers Parameters	20S/(20S+20R) St
	21/23TRI
	21+22/STER
	23/24TRI
	24/25TRI
	25NOR/HOPANE
	26/25TRI
	26/28TRI
	27/29ββS218
	28/29ββS218
	29/30H
	30/29ββS218
	DIA/C27αα
	DIA30/C27αα
	DIAH/H30
	DITERP/H30
	GAM/H30
	GAM/TR23
	H28/H29
	H28/H30

TYPE	VARIABLE
Biomarkers Parameters	H28/TR23
	H29/C29TS
	H29/H30
	H30/C27αα
	H35/H34
	HOP/STER
	NOR25H/H29
	NOR25H/H30
	NORNEO/H29
	Total Esterane
	Total Hopane
	TET24/26TRI
	TET24/H30
	TPP
	TR23/H30
	TRIC/HOP
	TRIC/STER
	TS/(TS+TM)
	TS/TM
	αββ/(αββ+ααα)

The classes proposed by the mentioned author were subdivided as follows: samples preliminarily classified as Lacustrine into Lacustrine A and B; Siliciclastic Lacustrine into Siliciclastic Lacustrine A and B; Mixed into Marine Evaporitic, Mixed belonging to the Areia Branca Trend, and Mixed from the Carnaubais Trend.

The next methodological step consists of treating the data. In this sense, it is important, at first, to clean the database, checking for the existence of empty values and outliers, as well as deciding whether the best option is to delete or replace them. It is common to have attributes of some samples with missing values for different reasons.

There is no scientific consensus on how to proceed in such cases. However, it is surely relevant to assess the impact of this operation on the whole dataset.

It is possible to extract complementary information and to reduce the dimensionality of the dataset using EDA (Exploratory Data Analysis), which comprises a wide range of univariate and multivariate techniques, as described below.

Correlation matrix – corresponds to a chart containing the correlation coefficients (r) between the variables in a range from -1 to 1. It is built to describe how the data are associated, with 0 being the indicator of no relationship; as the value approaches 1 or -1, the stronger the dependency (Härdle & Simar, 2015). The use of a color scale makes the matrix visually more pleasant.

Multidimensional scaling (MDS) – consists of the spatial representation of a proximity matrix between a series of objects; the greater the distance between two observations, the less the similarity. The objective is to find a set of points in low dimension that approximates the larger dimension configuration to the original matrix (Hair Junior, 2009; Härdle & Simar, 2015; Manly & Alberto, 2016).

Hierarchical cluster analysis (HCA) – is represented by a dendrogram type graph that shows the relationships between the variables. From the initial matrix, pairs of similar cases are grouped (according to their correlation coefficients) and one dimension is decreased; this step is repeated until all points are brought together in a single group in a two-dimensional space (Hair Junior, 2009; Härdle & Simar, 2015; Manly & Alberto, 2016).

K-Means partitioned clustering – simpler non-hierarchical algorithm that assigns the

observations to the group to which they are most similar. K random objects are chosen to form the base of the groups (centroid) and the existing objects are associated with the nearest centroid. Within each group, the centroid is recalculated until it is stable (Hair Junior, 2009; Härdle & Simar, 2015; Manly & Alberto, 2016).

MDS acts as a pre-processing step for cluster analysis. This combination of techniques allows for a better grouping of objects, as it eliminates possible noise present in high-dimensional data (Husson et al., 2010; Husson et al., 2017; Kassambara, 2017).

In addition to the aforementioned procedures, further techniques include the representation in histograms (pie and bar graphs) and ternary and binary diagrams, which are suitable to visualize trends. Finding the ideal number of attributes (optimum point) to work on is important for the subsequent phase of interpretation. Another relevant issue is the uniformity of the set (normalize/standardize/scale) to avoid bugs in the results due to redundancies or to different units of measurement. A transformation is performed on the values, applying the same rule to the entire database, so that all variables fall under a single numeric range.

Additionally, Machine Learning (ML), an integral part of Artificial Intelligence (AI), is used for analysis, modeling, and data visualization in order to solve problems and to support decision making. ML is the field of study in which computers achieve the ability to simulate human reasoning by means of algorithms, when trained to do so, allowing data to be correlated and classified without explicit programming.

Decision Trees (DT), Random Forests (RF) and Artificial Neural Networks (ANN) are some of the ML methods with good applicability and established use in the relevant literature for geochemical investigations. The definition of a classification model begins by separating the samples for training and testing as a percentage of the whole dataset. All the steps of this work methodology are performed using the Jupyter environment for Python language.

RESULTS

Considering appropriate the results obtained by Morais (2007), the classes therein proposed were taken as a starting point to develop a sequence of procedures that would allow to create a workflow for geochemical analysis, from familiarization with the database to the classification model stage.

Each of the 200 oil samples from the Potiguar Basin is described by up to 60 parameters (Table 1). Table 2 shows that 22 variables have empty data, of which 10 present more than 30% nulls, reaching almost 90% in 30/29ββS218.

Table 2 - Empty values in each attribute (considering all classes). Percentages different from zero are highlighted in blue.

VARIABLE	% NULLS	VARIABLE	% NULLS	VARIABLE	% NULLS
PRI/PHY	22,00%	20S/(20S+20R) St	0,00%	H28/TR23	40,50%
Pri/nC17	33,50%	21/23TRI	0,00%	H29/C29TS	0,00%
PHY/nC18	23,50%	21+22/STER	0,00%	H29/H30	0,00%
17/(17+C27)	33,50%	23/24TRI	0,00%	H30/C27αα	0,00%
Total Height	1,50%	24/25TRI	0,00%	H35/H34	1,50%
%Saturated	0,50%	25NOR/HOPANE	43,50%	HOP/STER	0,00%
%Aromatic	1,00%	26/25TRI	0,00%	NOR25H/H29	36,00%
%NSO	1,00%	26/28TRI	0,00%	NOR25H/H30	36,00%
δ13C	0,50%	27/29ββS218	0,00%	NORNEO/H29	0,00%
API	20,00%	28/29ββS218	0,00%	Total Esterane	0,00%
%Sulfur	43,50%	29/30H	0,00%	Total Hopane	0,00%
%27ββS218	0,00%	30/29ββS218	88,50%	TET24/26TRI	0,00%
%28ββS218	0,00%	DIA/C27αα	0,00%	TET24/H30	0,00%
%29ββS218	0,00%	DIA30/C27αα	0,00%	TPP	0,00%
%H31	0,00%	DIAH/H30	6,50%	TR23/H30	0,00%
%H32	16,00%	DITERP/H30	0,00%	TRIC/HOP	0,00%
%H33	0,00%	GAM/H30	0,00%	TRIC/STER	0,00%
%H34	0,00%	GAM/TR23	0,00%	TS/(TS+TM)	0,00%
%H35	1,50%	H28/H29	40,50%	TS/TM	0,00%
19/23TRI	0,00%	H28/H30	40,50%	αββ/(αββ+ααα)	0,00%

A decision was made to replace the missing values with the medians of the proposed classes in each parameter. Despite this, the variables 30/29ββS218 and 25NOR/HOPANE persisted

presenting empty data (revealing that in one or more classes it was not possible to obtain a central measure) and were therefore discarded. A parameter with 70-80% of nulls, *i.e.* about 2/3 of the dataset, will most likely be suppressed in future phases of the analysis because the resulting uniformity of the values will not give weight to the representativeness of the sample.

Sample elimination was not performed due to the insufficient size of the database. Thus, the methodology was followed with only 58 (out of 60 initials) descriptive attributes. Although substitution by average values is also common, the choice of median is justified because it is less biased towards outliers. However, outliers were not removed since they can characterize some of the variables.

After treating the empty values, univariate statistical techniques were used to visualize the data according to the dependent variable, the oil samples. Table 3 describes the relative and absolute frequencies of the seven classes proposed by Morais (2007).

Ternary (Figure 1) and binary (x versus y) diagrams are useful to interpret the quality, origin, and maturation of the investigated oils.

Table 3 - Number of samples by class.

CLASS	INITIALS	SAMPLES	% OF TOTAL
Lacustrine A	LAC A	48	24.00
Lacustrine B	LAC B	22	11.00
Siliciclastic Lacustrine A	SIL LAC A	25	12.50
Siliciclastic Lacustrine B	SIL LAC B	10	5.00
Marine Evaporitic	MAR EVP	20	10.00
Mixed – Areia Branca Trend	MIX T ARB	40	20.00
Mixed – Carnaubais Trend	MIX T CAR	35	17.50
Total		**200**	**100.00%**

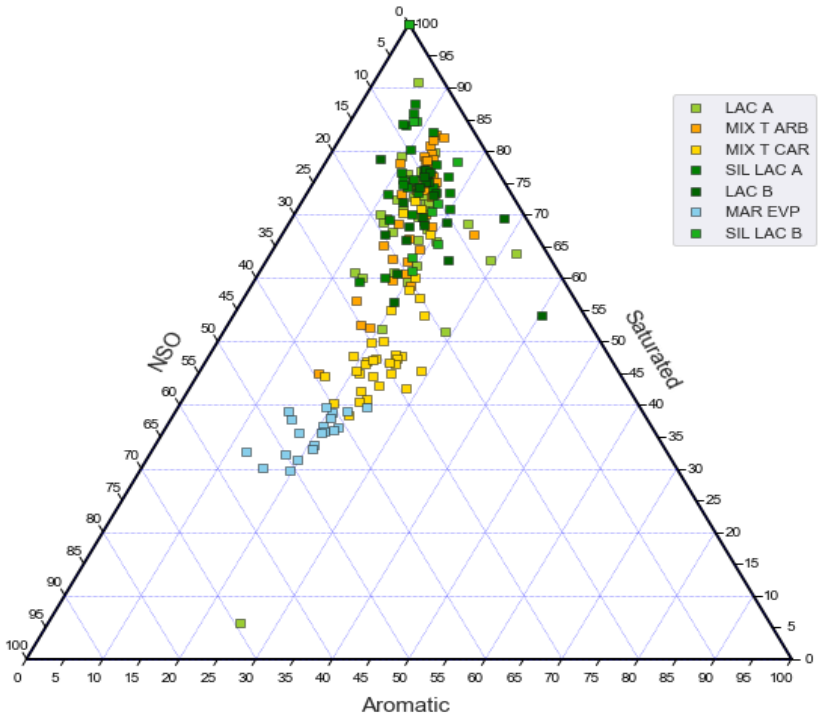

Figure 1 - Oil quality indicated by the presence (in %) of NSO compounds, saturated and aromatic hydrocarbons.

The next step includes data standardization so as not to impact their statistical distribution. In multivariate statistics, the correlation matrix (Figure 2) shows that many attributes are strongly associated (extremes of the color scale). The excess of similar information can be understood as redundancy and the cause of instabilities in the model. Therefore, dimensionality reduction is required.

In multidimensional scaling (MDS), the similarity between the objects is observed through the distances in the matrix (Figure 3). Indeed, it is possible to infer the existence of 5 to 6 groups with the visual examination of the diagram.

However, techniques such as clustering, dendrogram and k-means were used for this purpose.

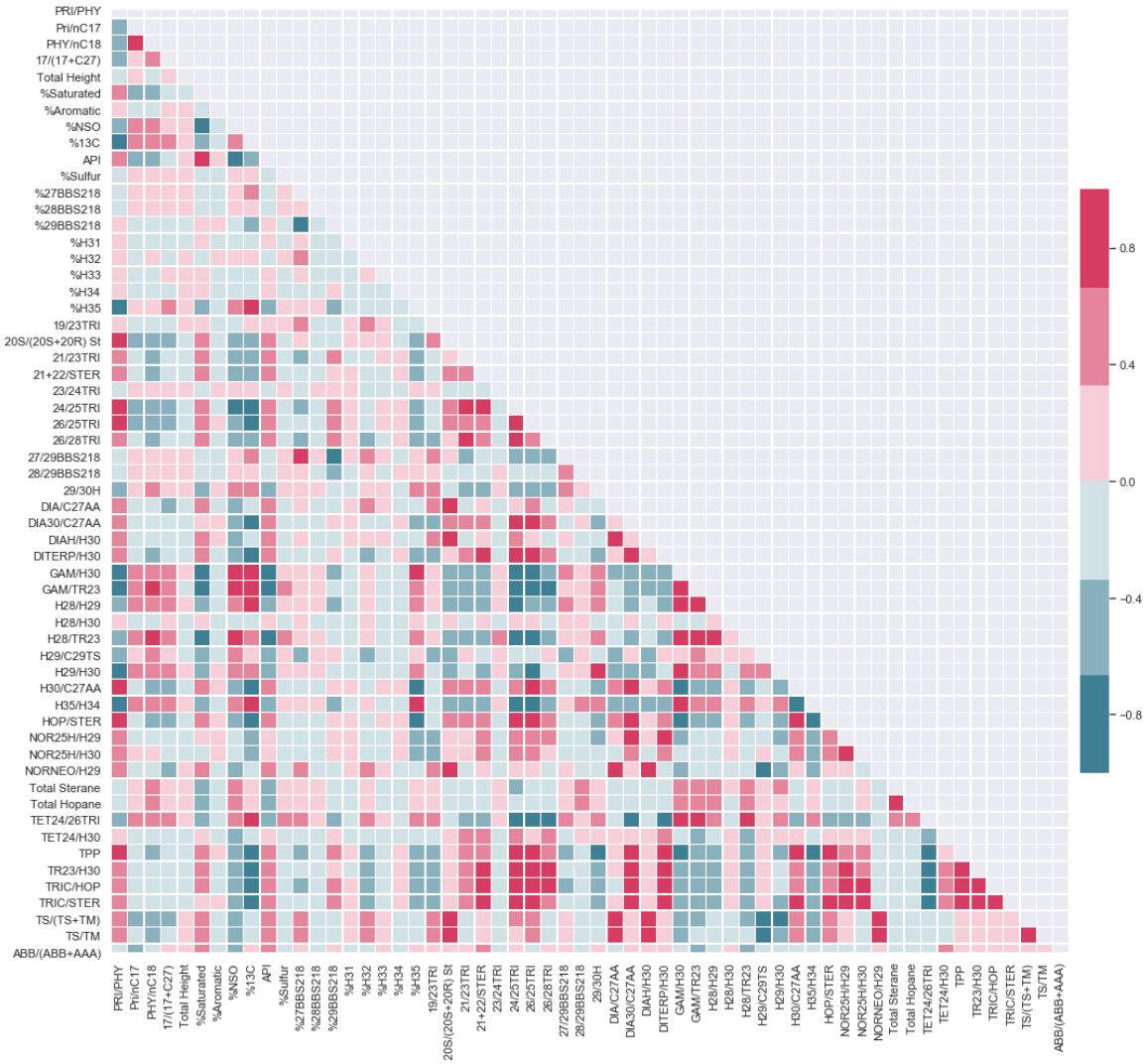

Figure 2 - Correlation matrix.

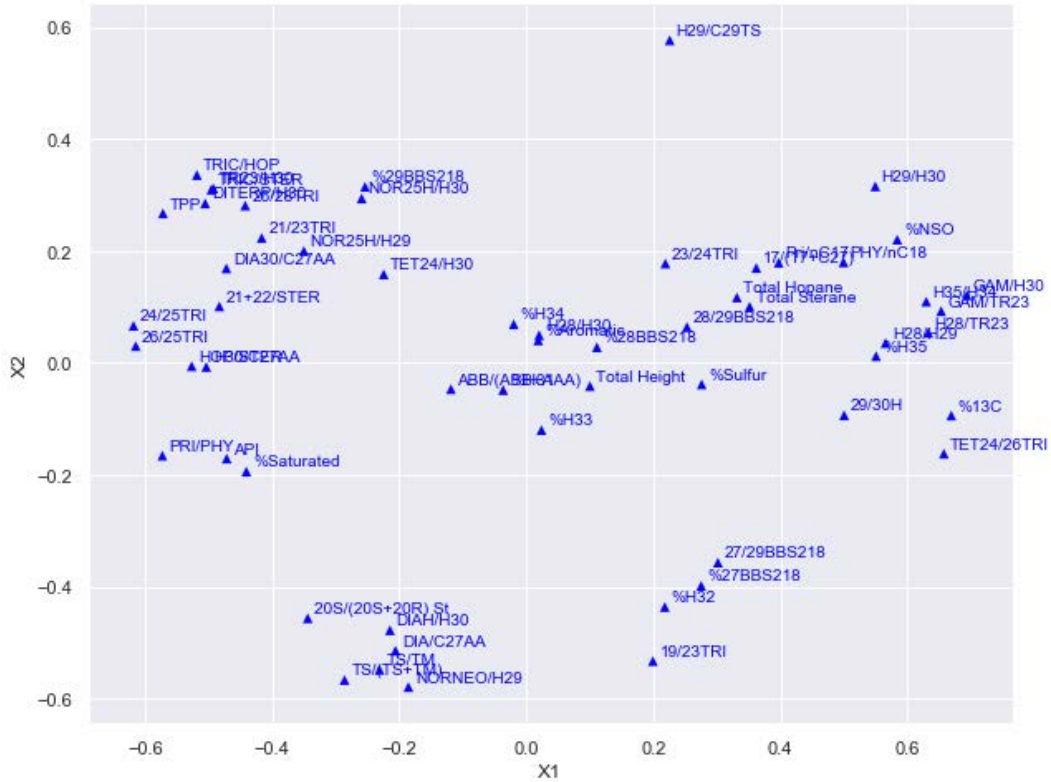

Figure 3 - Multidimensional Scaling (MDS).

The hierarchical grouping of the tree type, which is also known as dendrogram, gave rise to seven large collections (shown as different colors in Figure 4). For this data, the distance of 0.15 was chosen as a criterion for selecting the most representative variables. Below this limit, there would not be considerable reduction in the number of variables; above it, underfitting could be a problem.

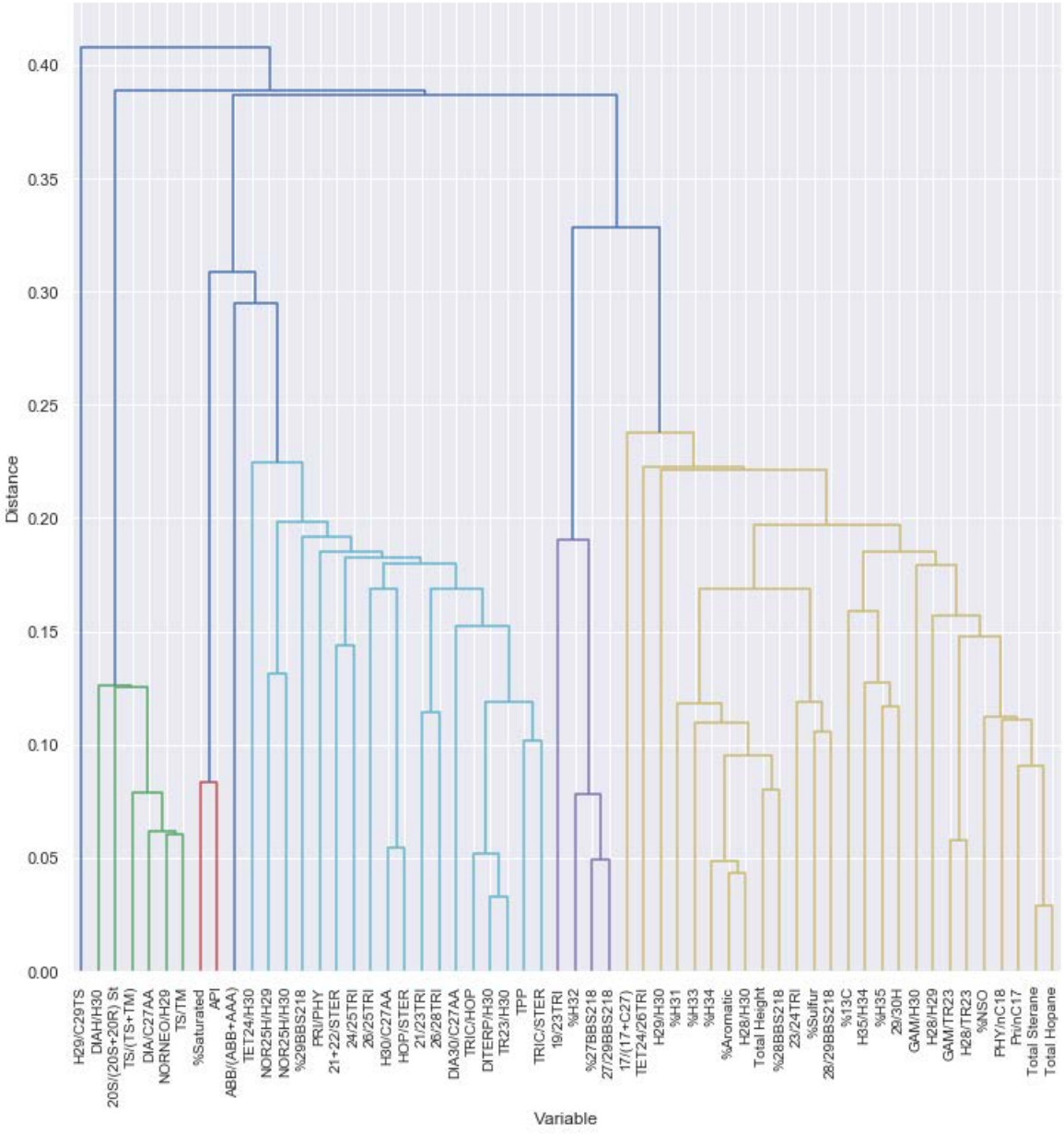

Figure 4 - Hierarchical cluster analysis (dendrogram).

There were some groups that indicated more than one variable, so the distance to the centroid was the factor considered to decide between them.

The partitioned grouping (k-means), when applying the Elbow method, identified 7 as the ideal number of clusters for this dataset. Figure 5 shows the achieved distribution in a three-dimensional perspective.

Thereby, out of the 60 initial attributes, the 26 listed in Table 4 remained, what constitutes about 57% of reduction to reach the so-called optimal set. These are the Machine Learning input data, a tabular file with 200 lines (classes of oils, the categorical attributes) and 26 columns (geochemical parameters, the predictive attributes).

Transforming (standardizing) the data was useful to make the orders of magnitude equivalent, in addition to not overestimating any information, thus facilitating the processing of ML algorithms. About 80% of the samples (160) were used for training so that the accuracy of the method was tested on the remaining 20% (40). This 80/20 ratio for training/testing was adopted in order not to compromise the learning process, since data are few.

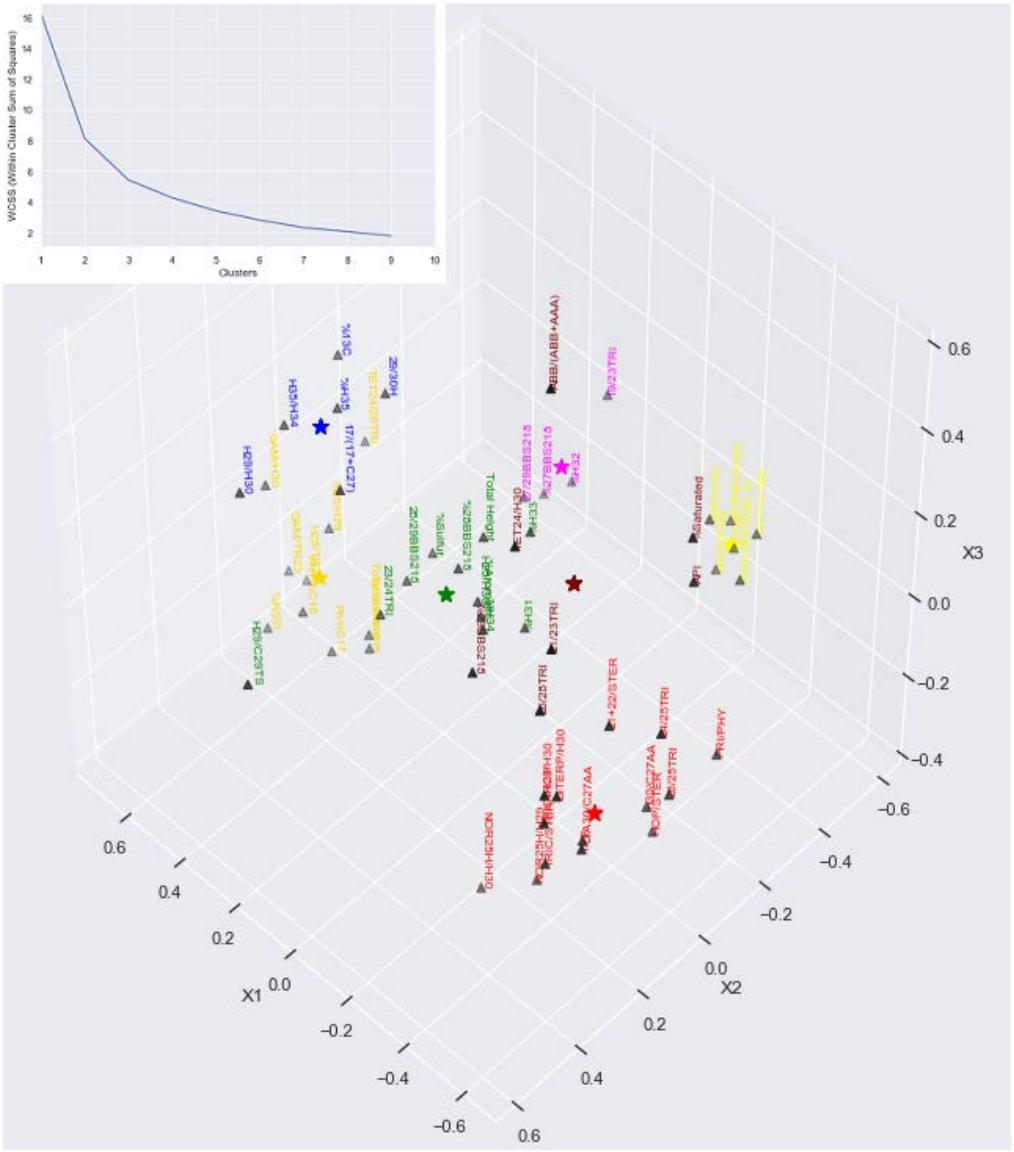

Figure 5 - 3D clusters using the k-means technique (and Elbow method above left).

Table 4 - Optimal set of attributes.

TYPE	VARIABLE
Gas Chromatography Parameters	PRI/PHY
	Pri/nC17
	PHY/nC18
	17/(17+C27)
Liquid Chromatography Parameters	%Saturated
	%Aromatic
Biomarkers Parameters	%27ββS218
	%28ββS218
	%29ββS218
	%H32
	%H35
	21/23TRI
	21+22/STER
Biomarkers Parameters	26/25TRI
	DIA30/C27αα
	H28/H30
	H29/C29TS
	H29/H30
	H30/C27αα
	H35/H34
	NOR25H/H29
	TET24/26TRI
	TET24/H30
	TPP
	TS/TM
	αββ/(αββ+ααα)

The Decision Tree (DT), Random Forest (RF) and Artificial Neural Network (ANN) algorithms were all run with default settings, achieving median accuracy of 92.50%, 95.00% and 87.50%, respectively.

A comparison between the 26 selected attributes (Table 4) and the 25 ones (Table 5) chosen by Morais (2007) indicates that 11 of

them are coincident.

As a way to measure and validate the approach herein presented, with emphasis on data exploration rather than final classification, for the same three ML algorithms, with similar configurations, median accuracy of 87.50% in DT, 92.50% in RF and 92.50% in ANN was achieved for the dataset constructed by Morais (2007).

Table 5 - Attributes chosen by Morais (2007).

TYPE	VARIABLE
Gas Chromatography Parameters	PRI/PHY
	PHY/nC18
Liquid Chromatography Parameters	%NSO
Total Carbon Isotope	δ13C
Biomarkers Parameters	19/23TRI
	20S/(20S+20R) St
	21/23TRI
	24/25TRI
	26/25TRI
	27/29ββS218
	DIA30/C27αα
	DITERP/H30
	GAM/H30

TYPE	VARIABLE
Biomarkers Parameters	H28/29
	H28/H30
	H28/TR23
	H29/C29TS
	H29/H30
	HOP/STER
	TET24/26TRI
	TPP
	TR23/H30
	TRIC/HOP
	TS/(TS+TM)
	αββ/(αββ+ααα)

It is important to highlight that the methodology applied in this research achieved responses very close (and sometimes superior) to those made by the referred author.

Finally, a workflow was developed that includes all the phases of the data exploration process (Figure 6).

The positive results obtained with the proposed methodology attest the advantages of following a predefined workflow, without skipping steps or reversing the logical sequence. Obviously, future studies will require adjustments depending on research objectives and available databases (in terms of quantity and quality). Note that there is a loop for the selection of attributes if the classification model is not performing to expectations.

Figure 6 - Workflow proposed by this research.

CONCLUSIONS

The record number of exploratory wells drilled in the last decade in Brazil made inevitable the arrival of an increased number of samples to be prepared and examined in geochemistry laboratories, thus creating considerably large datasets.

The adoption of digital technologies in this process is, therefore, highly recommended, seeking to automate and speed up data treatment and interpretation.

The present work proposes a methodology that combines statistical analysis and Machine Learning for a case study using oil samples from the terrestrial portion of the Potiguar Basin. As a result, the following conclusions can be drawn: (i) from an initial number of 60 predictive attributes, an optimal set containing 26 variables was reached through dimensionality reduction by similarity (MDS); (ii) classification of the 200 oil samples achieved median accuracies of 92.50% in DT, 95.00% in RF and 87.50% in ANN; and (iii) the proposed workflow is essential to assist future projects in data mining applied to organic geochemistry.

ACKNOWLEDGMENTS

To PUC-Rio for the physical and technological structure and to CENPES/Petrobras for scientific support.

REFERENCES

HAIR JUNIOR, J.F.; BLACK, W.C.; BABIN, B.J.; ANDERSON, R.E. **Multivariate data analysis**. 7th. ed. Upper Saddle River, New Jersey: Prentice Hall, 2009.

HÄRDLE, W.K. & SIMAR, L. **Applied Multivariate Statistical Analysis**. 4th. ed. Berlin, Heidelberg: Springer, 2015.

HUSSON, F.; JOSSE, J.; PAGÈS, J. **Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data?** Technical Report – Agrocampus, 2010.

HUSSON, F.; LÊ, S.; PAGÈS, J. **Exploratory multivariate analysis by example using R**. 2nd. ed., CRC Press, 2017.

KASSAMBARA, A. **Practical Guide to Principal Component Methods in R: PCA, (M)CA, FAMD, MFA, HCPC, factoextra**. 1st. ed., STHDA, 2017.

MANLY, B.F.J. & ALBERTO, J.A.N. **Multivariate Statistical Methods: A Primer**. 4th. ed. Boca Raton: CRC Press, 2016.

MORAIS, E.T. **Aplicações de Técnicas de Inteligência Artificial para Classificação Genética de Amostras de Óleo da Porção Terrestre, Bacia Potiguar**, UFRJ, 2007.

NASSER JUNIOR, R. **Otimização das colunas de absorção da recuperação de acetona na produção de Filter Tow por meio de estudos fenomenológicos e análise estatística**, USP, 2009.

NAVARRO, M.E. Fundamentos de La Geoquímica del Petróleo. In: CONGRESO LATINOAMERICANO DE GEOQUÍMICA ORGÁNICA, XI, 2008, Isla de Margarita, Venezuela. **Actas**.... Isla de Margarita, Venezuela, 2008.

PETERS, K.E. & MOLDOWAN, J.M. **The Biomarker Guide: Interpreting Molecular Fossils in Petroleum and Ancient Sediments**. Englewood Cliffs, New Jersey: Prentice Hall, 1993.

PHILP, R.P. Fossil fuel biomarkers: Applications and spectra. Methods in geochemistry and geophysics. **Fontes de Energia**, v. 23, p. 294, 1985.

Submetido em 9 de setembro de 2021
Aceito para publicação em 20 de janeiro de 2022