2º CONGRESSO BRASILEIRO DE P&D EM PETRÓLEO & GÁS

# APLICAÇÃO DO *K-MEANS CLUSTER* A DADOS DE PERFIS DE POÇOS PETROLÍFEROS

A. C. B. da Rocha [1], F. A. M. de Souza [2]

[1] UFCG/CCT/DME – Av. Aprígio Veloso, 882, Bodocongó, Caixa Postal 10044, cep 58109-970, Campina Grande – PB – Brasil, e-mail: chrisrrocha@yahoo.com.br.
[2] UFCG/CCT/DME, e-mail: fsouza@dme.ufcg.edu.br.

**Resumo** – A Análise de Agrupamentos é uma técnica bastante utilizada na realização de análises estatísticas. Isso decorre pelo fato de que este método facilita sobremaneira o trabalho do pesquisador, visto que reduz o volume de dados sem que haja perda significativa de informações. Esta pesquisa tem como objetivo fazer uma análise estatística de dados dos perfis *GR*, *ILD*, *NPHI* e *RHOB* relativos a 44 poços pertencentes ao Campo Escola de Namorado, utilizando, em particular, as técnicas de agrupamento. Como técnica de agrupamento, utilizou-se o *K-Means Cluster*, tendo em vista a sua eficiência computacional. Este método consiste em dividir os diferentes casos de uma matriz de dados em *K* grupos homogêneos, cada grupo constituindo, em princípio, uma população bem definida. No nosso caso, a utilização de um procedimento estatístico tem por finalidade tornar possível a identificação das semelhanças existentes entre os poços alocados no mesmo agrupamento, como também das disparidades existentes entre poços alocados em agrupamentos diferentes, objetivando uma maior facilidade na avaliação das formações a que pertencem.Considerando outras informações fornecidas pela ANP, sobre as características de cada poço, e um mapa geográfico que ilustra a proximidade dos mesmos, foi constatado que as análises apresentaram resultados bastante satisfatórios.

Palavras-Chave: Análise de Agrupamentos; *K-Means Cluster*; Perfis de Poços

**Abstract** – For reducing the volume of data without lose information, the Cluster Analysis is one technique sufficiently used in the several fields. Using the clustering methods, data of well logs such that GR, ILD, NPHI and RHOB relative to 44 wells located in Campo Escola de Namorado had been analyzed with the purpose to become easier to accomplish the evaluation of the formations that each one belongs. For its well-known computational efficiency, we used K-Means Cluster, which consists separate the different cases of the data matrix in K homogeneous groups, each one constituting, in principle, a well defined population. The coherence of the results gotten from the accomplishment of the analyses was evidenced with the consideration of additional informations supplied by the National Agency of Oil, regarding the characteristics of each well, and a geographic map showing the proximity between the wells.

Keywords: Cluster Analysis, K-Means, Well Logs

## 1. Introdução

Em exploração de petróleo, existem diversos tipos de perfis, com aplicações as mais variadas, todos com o objetivo de melhor avaliar as formações geológicas quanto à ocorrência de uma jazida comercial de hidrocarbonetos.

Este trabalho tem como objetivo fazer uma análise estatística de dados relativos a alguns destes tipos de perfis, utilizando técnicas de agrupamento.

Estas técnicas são de grande utilidade, visto que facilitam o estudo de grandes grupos de dados. A análise de agrupamento tem como base um conjunto de *n* indivíduos para os quais existe informação sobre *p* variáveis. O método faz o agrupamento dos indivíduos em função da informação existente, de tal modo que os indivíduos pertencentes a um mesmo grupo sejam tão semelhantes quanto possível e sempre mais semelhantes do que a indivíduos dos outros grupos. No nosso caso, os indivíduos correspondem a 44 poços pertencentes ao Campo Escola de Namorado e as variáveis são dados dos perfis *GR*, *ILD*, *NPHI* e *RHOB* relativos a estes poços, de maneira que, após o agrupamento, seja possível identificar as semelhanças existentes entre poços alocados no mesmo grupo e as disparidades existentes entre os poços alocados em grupos diferentes, facilitando a realização da avaliação das litofácies a que pertencem.

## 2. Avaliação das Formações

Entende-se por "Avaliação de Formações" as atividades e estudos que visam determinar, em termos qualitativos e quantitativos, o potencial de uma jazida petrolífera. Para tanto, são necessários alguns procedimentos, tais como: a perfuração do poço pioneiro; a verificação dos indícios que podem proporcionar a suspeição da presença de hidrocarbonetos na formação; a perfilagem a poço aberto.

A perfilagem a poço aberto, executada após a perfuração do poço, permite obter informações importantes a respeito das formações atravessadas pelo poço: litologia (tipo de rocha), espessura, porosidade, argilosidade, radioatividade total da formação, prováveis fluidos existentes nos poros e suas respectivas saturações, etc. Essas informações, denominadas de perfis, são obtidas a partir do deslocamento contínuo de um sensor de perfilagem (sonda) dentro do poço e são denominados genericamente de perfis elétricos. Alguns tipos de perfis mais utilizados são: Potencial Espontâneo (*SP*); Sônico (*DT*); Raios Gama (*GR*); Indução (*ILD*); Neutrônico (*NPHI*) e Densidade (*RHOB*).

Neste trabalho foram utilizados dados dos perfis *GR*, *ILD*, *NPHI* e *RHOB* de poços do Campo Escola de Namorado, cedidos pela ANP (Agência Nacional do Petróleo).

## 3. Análise de Agrupamentos

Análise de Agrupamentos é o nome dado às técnicas de análise que dividem os dados em grupos, classificando objetos ou indivíduos sem preconceitos, isto é, observando apenas as similaridades ou dissimilaridades entre eles. Os métodos de Análise de Agrupamentos tentam organizar um conjunto de indivíduos, para os quais é conhecida informação detalhada, em grupos relativamente homogêneos (*clusters*).

Genericamente, a análise de agrupamentos compreende seis etapas:

1. Escolha de variáveis e objetos a serem analisados;
2. Obtenção dos dados;
3. Tratamento dos dados;
4. Escolha de critérios de similaridade ou dissimilaridade;
5. Adoção e execução de um método (algoritmo) de agrupamento;
6. Validação e interpretação dos resultados.

Para este trabalho foram selecionados 44 poços do Campo Escola de Namorado, e as variáveis utilizadas foram dados dos perfis *GR*, *ILD*, *NPHI* e *RHOB* relativos a esses poços e dados de medidas das profundidades em que estes perfis foram obtidos. No intuito de obter resultados que contribuíssem na determinação do potencial da jazida a que os poços pertencem, foi aplicado, no início, um procedimento hierárquico baseado na distância euclidiana, tendo como critério de agregação dos casos o Critério do Vizinho mais Próximo (*Single Linkage*), como base para uma aplicação posterior do método *K-Means,* também baseada na distância euclidiana, tendo como critério de agregação dos poços o Critério do Centróide. Os resultados obtidos serão mencionados mais adiante.

## 4. O Método *K-Means*

Este método consiste em dividir os diferentes casos de uma matriz de dados em *k* grupos mais ou menos homogêneos, cada grupo constituindo, em princípio, uma população bem definida. Em outras palavras, este método baseia-se diretamente na escolha antecipada de um número de agrupamentos que englobarão todos os casos. Procede-se, em seguida, a uma divisão de todos os casos pelos *k* grupos preestabelecidos e a melhor partição dos *n* casos será aquela que otimiza o critério escolhido.

Ao aplicar um critério de otimização que divida uma amostra em *k* grupos homogêneos, pretende-se que,

dentro de cada grupo, os elementos sejam o mais semelhante possível entre si, ao passo que as semelhanças entre os elementos de grupos distintos sejam as menores possíveis.[5]

Cada grupo é representado por um ponto central, denominado centróide. O termo *K-Means* é sugerido para descrever um algoritmo que aloca cada item no grupo cujo centróide esteja mais próximo. O processo é composto das seguintes etapas:

1. Divisão dos itens em *k* conjuntos iniciais.
2. Seguir prosseguimento através da lista de itens, alocando cada item ao grupo cujo centróide é mais próximo. (O procedimento geralmente é computado utilizando-se a distância euclidiana entre observações padronizadas ou não padronizadas.).
3. Realização do cálculo do centróide para o grupo que recebe o novo item e do cálculo do centróide para o grupo de onde o item é excluído.
4. Repetição das etapas 2 e 3 até que o reagrupamento dos itens não seja mais necessário.

É melhor especificar *k* centróides iniciais (pontos de origem), e então prosseguir às etapas 2 e 3, do que começar com uma divisão de todos os itens em *k* grupos preliminares na etapa 1. A alocação final dos itens aos agrupamentos é, de certo modo, dependente da divisão inicial ou da seleção inicial de pontos de origem. A experiência sugere que as principais mudanças na atribuição ocorram com a primeira etapa do reagrupamento. [2]

Outra etapa que merece bastante cuidado é a seleção das variáveis, um dos fatores que mais influenciam o resultado de uma análise de agrupamento. Variáveis que assumem praticamente o mesmo valor para todos os objetos são pouco discriminatórias, e sua inclusão pouco contribui para a determinação da estrutura do agrupamento. Por outro lado, a inclusão de variáveis com grande poder de discriminação, porém irrelevantes ao problema, pode mascarar os grupos e levar a resultados equivocados. Freqüentemente, o número de variáveis medidas é grande, dificultando a análise. Deve-se, então, procurar diminuir este número, de forma que a seleção de variáveis contemple tanto a sua relevância como seu poder de discriminação face ao problema em estudo. [1]

## 5. Aplicação da Análise de Agrupamentos a Dados da Indústria de Petróleo e Gás

Com o objetivo de facilitar a realização do trabalho, foi utilizada uma nomenclatura particular para os poços do Campo Escola de Namorado. Segue abaixo uma tabela com alguns dos poços estudados e suas respectivas nomenclaturas para ilustrar as mudanças feitas.

Tabela 1. Nomenclatura dos Poços do Campo Escola de Namorado

| Nome do poço | Nome por extenso | Nova nomenclatura |
|---|---|---|
| 3NA0001A RJS | NAMORADO.1A | 1 |
| 3NA0002 RJS | NAMORADO.2 | 2 |
| 3NA0003 RJS | NAMORADO.3 | 3 |
| 3NA0003D RJS | NAMORADO.3D | 4 |
| 3NA0004 RJS | NAMORADO.4 | 5 |
| . | . | . |
| . | . | . |
| . | . | . |

Inicialmente, os dados foram transformados para o formato utilizado pelo *software* SPSS (*Statistical Package for the Social Sciences*), para que pudessem ser analisados, como podemos observar na Figura 1. Como o volume de dados era muito grande (aproximadamente 50000 casos, isto é, 50000 linhas pertencentes à matriz de dados) e o objetivo inicial era utilizar um procedimento hierárquico, optou-se por fazer uma divisão em faixas de profundidade, a fim de obter mais eficiência.

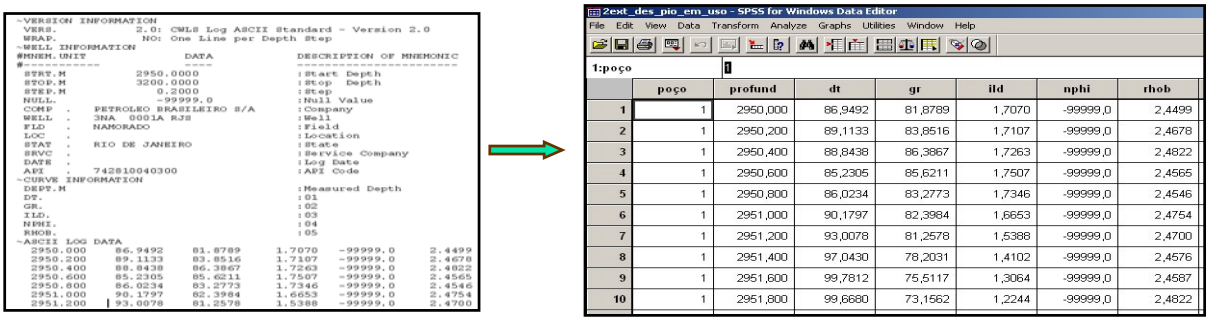


Figura 1. Transformação dos dados para o formato utilizado pelo *software* SPSS.

Foi necessária uma divisão do grupo de dados inicial em faixas de profundidade. O número de faixas escolhido foi 50, visto que a cada faixa pertenceriam aproximadamente 1000 casos, valor de razoável tolerância pelo *software* SPSS. Neste caso, seria feita a análise em cada faixa separadamente e, em seguida, a análise conjunta dos resultados para uma interpretação. Porém, como a análise utilizando o procedimento hierárquico estava se dando de maneira bastante lenta, buscou-se utilizar um procedimento mais eficiente quando se conhece o número de agrupamentos a serem formados e bastante utilizado em análise de agrupamentos em se tratando de grande volume de dados: o método *K-Means.* Para tanto, a partir de análises feitas com a aplicação do procedimento hierárquico, pode-se observar que o número ideal de agrupamentos a serem formados estaria entre 7 e 10, informação imprescindível para a aplicação do *K-Means*, visto que este procedimento requer, de início, uma definição do número de agrupamentos a serem formados.

## 6. Realização das Análises

Utilizando o método *K-Means*, foram feitas análises para *K*=7 e *K*=10 (isto é, 7 e 10 agrupamentos, respectivamente). Estes valores de *K* foram obtidos a partir do procedimento hierárquico, o qual foi utilizado apenas no estágio anterior, por motivos já citados.

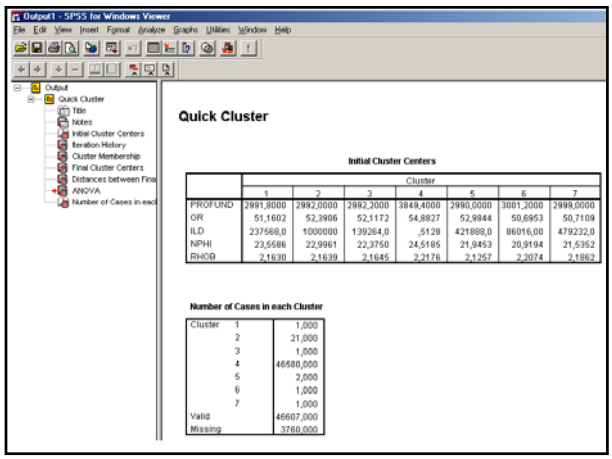


Figura 2. Resultado das Análises Utilizando o *K-Means* para *K*=7.

Com base nos resultados, resolveu-se excluir o poço 1 e refazer as análises, visto que os resultados estavam sendo muito discrepantes. Isso se deve ao fato de que a variável *ILD* do poço 1 apresenta valores altíssimos (da ordem de 1 milhão), os quais se distanciam muito dos valores de *ILD* dos demais poços, mascarando, assim, os resultados. Uma análise descritiva conduziu à hipótese da existência de *outliers* ou observações discrepantes, necessitando de uma discussão com especialistas.

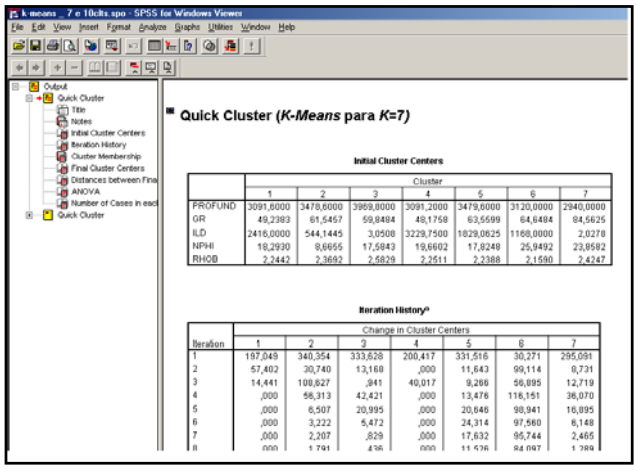


Figura 3. Resultado das Análises Utilizando o *K-Means* para *k*=7 após a exclusão do poço 1.

A partir das análises, optou-se por considerar 7 grupos.

Além disso, observou-se que os grupos 2 e 7 obtiveram maior concentração de dados. O que os caracteriza é o valor do *ILD*, por ser a variável que apresenta valores mais discrepantes entre os grupos. No grupo 2, o *ILD* assume valores bastante grandes enquanto que no grupo 7, assume valores bem pequenos.
O grupo 3 também obteve grande volume de dados. Os valores de *ILD* pertencentes a este agrupamento também são pequenos, próximos aos valores de *ILD* do grupo 7.

## 7. Resultados Obtidos

O *software* SPSS trabalha utilizando o conceito de casos. Cada caso corresponde a uma linha, que pode conter informações sobre várias variáveis. No caso dos dados do Campo Escola de Namorado, cada linha corresponde a um nível de profundidade do poço, de maneira que a cada poço estão relacionados vários casos, visto que cada poço tem medidas em vários níveis de profundidade.

Após as análises para $K$=7 (ou seja, 7 agrupamentos), foi necessário observar quais casos estavam relacionados com cada poço, para que se pudesse associar casos a poços e, assim, saber em que agrupamentos cada poço foi alocado. Para isso, foi feita uma tabela contendo a nomenclatura dos poços, o número de casos relacionados a cada um deles e os intervalos aos quais cada conjunto de casos relacionados a cada poço pertencia, como podemos observar na Tabela 2. Em seguida, foi feita outra tabela contendo todos os casos e o grupo a que cada um deles pertencia.

Tabela 2. Associação caso-poço

| Poço | Número de casos | Casos correspondentes |
|---|---|---|
| 1 | 1250 | 1 – 1250 |
| 2 | 1125 | 1251 – 2375 |
| 3 | 550 | 2376 – 2925 |
| 4 | 875 | 2926 – 3800 |
| 5 | 1250 | 3801 – 5050 |
| . | . | . |
| . | . | . |
| . | . | . |

Em seguida, foi feita outra tabela contendo todos os casos e o agrupamento (*cluster*) a que cada um deles pertencia, como podemos observar na Tabela 3.

Tabela 3. Associação caso-agrupamento

| Casos correspondentes | Poço | Agrupamento |
|---|---|---|
| 1 – 1250 | 1 | – |
| 1251 – 1575 | 2 | 7 |
| 1576 – 1579 | 3 | 6 |
| 1580 – 1586 | 4 | 7 |
| 1587 – 1592 | 5 | 6 |
| . | . | . |
| . | . | . |
| . | . | . |

Após a interpretação destas últimas tabelas, observou-se em que agrupamento cada poço foi alocado na maioria de seus casos. A partir destas informações, pode-se verificar quais poços possuiam maior número de semelhanças entre si, semelhanças estas referentes ao potencial de produção, visto que foram sugeridas a partir de dados de perfis.

## 8. Conclusão

Em se tratando de realização de análises estatísticas a partir de amostras com grande volume de dados, uma ferramenta muito utilizada, por ser bastante eficaz quando aplicada de maneira correta, é o *K-Means Cluster*.

Na indústria de petróleo e gás não poderia ser diferente. Através da análise dos dados utilizando-se o *K-Means*, foi possível obter um zoneamento dos poços, e com isto destacar e identificar características das formações geológicas de cada grupo, utilizando as características dos poços que o compõem. A eficácia do zoneamento foi comprovada a partir do confronto com informações adicionais, sobre o Campo Escola de Namorado, disponibilizadas pela ANP.

## 9. Agradecimentos

Este trabalho teve o apoio financeiro da ANP (Agência Nacional do Petróleo), através do Programa de Recursos Humanos – PRH (25).

## 10. Referências

[1] BUSSAB, Wilton. O.; MIAZAQUI, Édina S.; ANDRADE, Dalton F.. *Introdução à Análise de Agrupamentos*. Associação Brasileira de Estatística, 9° Simpósio Nacional de Probabilidade e Estatística. São Paulo, 1990.

[2] JOHNSON, Richard A.; WICHERN, Dean W.. Applied Multivariate Statistical Analysis. 3rd ed. Prentice-Hall, 1992.

[3] KOCH, George S. Jr, LINK, Richard F.. *Statistical Analysis of Geological Data* – Vol II, 1971.

[4] KRZANOWSKY, H. J. & MARRIOTT, F. H. C.. *Multivariate Analysis*.Arnold. New York, 1995

[5] LOURENÇO, Alexandre; MATIAS, Rui P.. *Estatística Multivariada*. Instituto Superior de Engenharia do Porto, 2000.

[6] NIE, N. H. et alli. *SPSS – Statistical Package for the Social Sciences*. 2nd Edition. United States of America: McGraw-Hill, 1975.

[7] REIS, Elizabeth. *Estatística Multivariada Aplicada*. Edições Silabo. Lisboa, 1997.

[8] THOMAS, José E. (org.). *Fundamentos de Engenharia de Petróleo*. Rio de Janeiro: Interciência: PETROBRAS, 2001.