CLASSIFICAÇÃO DE PETRÓLEOS

Fernando Guimarães Ferreira

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientador: José Manoel de Seixas

Rio de Janeiro
Março de 2018

CLASSIFICAÇÃO DE PETRÓLEOS

Fernando Guimarães Ferreira

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. José Manoel de Seixas, D.Sc.

Prof. José Gabriel Rodriguez Carneiro Gomes, Ph.D.

Profª. Nadia Nedjah, Ph.D.

Prof. Carmelo Jose Albanez Bastos Filho, D.Sc.

Prof. Alexandre Rodrigues Tôrres, D.Sc.

Prof. Marco Antônio Farah, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2018

Ferreira, Fernando Guimarães

Classificação de Petróleos/Fernando Guimarães Ferreira. – Rio de Janeiro: UFRJ/COPPE, 2018.

XIV, 105 p.: il.; 29, 7cm.

Orientador: José Manoel de Seixas

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2018.

Referências Bibliográficas: p. 92 – 105.

1. Mineração de dados. 2. Classificação de petróleos. 3. Reconhecimento de padrões. 4. Clusterização não-supervisionada. I. Seixas, José Manoel de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

A memória do Vô Abelardo.

Agradecimentos

Chegar a uma defesa de tese de Doutoramento é caminho longo, impossível de ser traçado sozinho. Desde a época de calouro de Engenharia Eletrônica em 2004, foram inúmeras pessoas que fizeram parte dessa trajetória, me incentivando, me desafiando e me dando oportunidades de ir mais longe. É óbvio que é impossível citar todos, mas certamente alguns merecem destaque especial.

Meus pais, Luiza Helena e João Carlos, por todo o apoio. Obrigado por sempre me incentivarem a buscar objetivos, e de terem me ensinado de que nada realmente é inalcançável. Com dedicação e carinho as coisas, cedo ou tarde, acontecem. Sou muito grato pelos conselhos, força e ensinamentos ao longo dessa caminhada.

Agradeço também a minha irmã, Ana Luíza, pelo carinho, companheirismo e confiança. Sem você, seria impossível seguir em frente em tantos desafios. Obrigado por seu empenho e esforço em me ajudar a vencê-los. Nunca esquecerei.

Ao meu orientador José Manoel de Seixas por, ao longo dos anos, ter me apresentado tantas oportunidades, nas mais diferentes áreas. Foram todas instigantes, ao mesmo tempo desafiadoras. A confiança, a amizade e o suporte sempre recebido são indescritíveis. A busca por inovação e multidisciplinaridade guiam muitas das minhas decisões profissionais. Muito obrigado.

Ao grande amigo Thiago Xavier, que me antecedeu no projeto BLEND-BR e, como já havia acontecido no Mestrado, facilitou meu trabalho compartilhando seu conhecimento e experiência. Nesse ponto, não sei mais quantos projetos trabalhamos juntos desde os tempos do CERN. Aprendi em todos.

Ao Gilberto Xavier por todo auxílio ao longo do projeto: desde sua concepção até as discussões frente aos resultados, sempre me recebendo no CENPES. Agradeço também Alexandre Tôrres por ter participado de muitas dessas reuniões, dando *insights* fundamentais para a evolução do trabalho. Em relação ao projeto BLEND-BR, agradeço também Victor Cascão, por ter pavimentado os primeiros passos da pesquisa.

Agradeço aos funcionários da secretaria do PEE. Em especial a Dani e Maurício que muito me ajudaram em todo o percurso de Mestrado e Doutorado.

Aos amigos e sócios da Twist: Laura Moraes e Felipe Grael. Agradeço por compartilharem da mesma motivação em empreender e pelo companheirismo cotidiano.

Não preciso nem citar o quanto me ajudaram nesse trabalho e nem o prazer que é evoluir a empresa junto com vocês.

Ainda sobre a Twist, agradeço a todos que trabalham ou trabalharam com a gente, por sempre terem ajudado tanto nesses 6 anos de existência e contribuírem com meu crescimento. Agradecimento especial ao amigo Luiz Évora, que por seis meses não foi citado no parágrafo anterior.

Agradecimentos também aos amigos Luiz Fernando Frias (que ainda me deu oportunidade de ser co-orientador de projeto de graduação e padrinho de seu casamento) e Breno Arosa pela colaboração e empenho. Vale a menção a Andressa Sivolella que participou ativamente da fundação da empresa.

Agradeço a todos da Incubadora e Parque Tecnológico pelo apoio dado a Twist desde sua concepção.

No trajeto desde a graduação até essa defesa participei de muitos projetos, contudo o período de sete anos trabalhando com a colaboração ATLAS/CERN é especial e vale menção. Às amizades de todos no TileCal/CERN, que desde 2007 fazem de Genebra um lugar especial. Carlos, Giorgi, Brian, James, Heuijin, Keith, Gabriel, Filipe, Luis, Denis Damazio, Rob, Ali, Jalal, Stan, Irene (em memória), Luca, David, Bruno, Bernardo, Torres, Bruno Lange... obrigado pela convivência e histórias decorrentes dela. Muito bom que existam redes sociais por permitirem contato com a maioria. Agradeço a Carmen Maidantchik pela orientação ao longo da minha participação neste projeto. Ainda agradeço Ana Henriques e Bob Stanek que viabilizaram minha estada junto à colaboração *TileCal* e cujo lições de liderança carrego comigo.

Ao amigo Junior Moura pelo apoio, pelas trocas de ideias, pela colaboração com a Twist, pela diversão e por ter causado curto-circuito em dois andares no seu primeiro dia de trabalho no CERN. Até hoje é uma história muito boa, rendendo risadas.

Falando em diversão e risadas, agradeço a tantos amigos pelos momentos de descontração: Zampari, Luiz Guilherme e Amanda, Fabio e Tati, Raphael, Fernanda, Erika e Renato, Felipe e Amanda, Pantoja e Mari, Tepedino e Nanda, Ribeiro e Beatriz, aos demais colegas da graduação e do LPS... Apenas para citar alguns dos responsáveis por deixarem a vida mais leve.

A parte mais legal de escrever essa seção de agradecimento, é que ela poderia ser embaralhada, mas os nomes e parágrafos ainda fariam bastante sentido, pois todos citados tem interseções em diversos momentos, atividades e aspectos da minha vida. Acho isso fantástico.

Por fim, gostaria de agradecer a UFRJ, da qual sou aluno há mais de quatorze anos e onde fica a sede da Twist. Tenho muito orgulho de fazer parte desta instituição e sempre procurarei fazer o meu melhor profissionalmente para fazer jus ao peso dela.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

CLASSIFICAÇÃO DE PETRÓLEOS

Fernando Guimarães Ferreira

Março/2018

Orientador: José Manoel de Seixas

Programa: Engenharia Elétrica

A identificação de padrões em dados de ensaios de óleo bruto fornece informações importantes sobre a estimação das propriedades do petróleo, assim como para a operação e cadeia logística das refinarias. A informação *a priori* sobre as características de determinada amostra de óleo melhora a logística em relação a maneira que as refinarias devem processá-lo, assim como a sua precificação. Essa tese explora técnicas de mineração de dados usando algumas propriedades relevantes para a caracterização dos ensaios, de maneira a agrupar amostras de óleos brutos similares de maneira não-supervisionada. Os resultados mostram que os modelos obtidos são capazes de encontrar padrões ao agrupar as amostras de acordo com essas propriedades. Estes são então comparados à uma classificação comumente usada na indústria, baseada apenas na mensuração da densidade do petróleo.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

OIL CLASSIFICATION

Fernando Guimarães Ferreira

March/2018

Advisor: José Manoel de Seixas

Department: Electrical Engineering

The identification of patterns in the crude oil assay data provides useful information for crude oil properties estimation as well as for the refinery operation and logistics. The *a priori* information about the characteristics of a determined crude improves the logistic concerning which refineries should process it, together with pricing. This work explores data mining techniques over some characterization properties of crude oil assays, in order to group similar crude oils in an unsupervised way. The results show that the derived models are able to find patterns, clustering crudes according these properties. Afterwards, these are compared to a standard classification which is aware only about the oil crude density.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A evolução e o desenvolvimento de técnicas de mineração de dados impulsionam o desenvolvimento de diversas áreas do conhecimento e, por essa razão, a pesquisa em inteligência computacional assume caráter multidisciplinar. Esta característica, somada às inúmeras possibilidades de abordagem a desafios, tornam a área instigante.

Este trabalho acontece no âmbito do reconhecimento de padrões para dados multidimensionais da indústria petrolífera. Nesse setor, devido aos altos valores envolvidos, eficiência e acurácia são extremamente importantes. Pequenas otimizações na operação das empresas de Óleo e Gás geram impactos significativos, e nesses tempos de retomada econômica, com a exploração ocorrendo em regiões cada vez mais inóspitas, inovação é uma estratégia ainda mais importante.

Companhias petrolíferas tem um enorme volume de dados referentes às informações geológicas, aos dados de caracterização dos reservatórios, ao desenvolvimento de testes, às operações logísticas, ao sensoriamento realizado em diversas etapas e à produção propriamente dita. Estima-se que um campo de produção *offshore* produza mais do que 0.75 terabytes de dados semanalmente, que uma refinaria de grande porte gere aproximadamente 1 terabyte de dados brutos diariamente e que as principais empresas do setor armazenam até 2 terabytes de dados todos os dias [1].

Nesse cenário, mineração de dados, aprendizagem de máquina e análises preditivas tornam-se importantes fatores para conduzir processos eficientes e otimizados requeridos por este setor competitivo. Estas ferramentas permitem o acesso rápido

[1] Informações extraídas de [1]

e efetivo a estes imensos bancos de informações e, quando integradas às práticas comuns da engenharia de petróleo, estabelecem novos paradigmas direcionando toda indústria para decisões baseadas em dados.

A crescente adoção de técnicas de mineração de dados também proporcionam revisitar desafios antigos, que passam a ser abordados de uma nova forma. Um ponto enfrentado pelos profissionais da indústria de petróleo se refere à classificação dos diversos tipos de óleos brutos existentes. A cadeia logística e os processos de refino dependem do tipo e das características de sua matéria-prima. De fato, cada refinaria é projetada para processar petróleos com determinadas características. Dependendo das condições, diferentes óleos são misturados para casar com as características toleradas por dada planta.

1.1 Motivação

O petróleo é o *commodity* mais importante do mundo, tanto pelo seu valor de produção, como por sua importância para a economia global. Os derivados do petróleo ainda são os principais combustíveis, somando aproximadamente 33 % do total de energia consumida globalmente. De modo geral, 63 % do consumo de óleo é realizado pelo setor de transporte e sua completa substituição ainda está distante.

O petróleo é um produto natural e o local de sua formação influencia diretamente nas suas características físico-químicas, e por essa razão, uma infinidade de tipos diferentes de óleos podem ser encontrados ao redor do mundo. Trata-se de uma mistura de compostos que consiste predominantemente de hidrocarbonetos e, em menor quantidade, de elementos sulfurados, nitrogenados, oxigenados e metálicos. A sua caracterização é necessária para a otimização processos de produção, refino e usos do petróleo e derivados. Devido à grande complexidade de sua formação, determinar analiticamente a composição de petróleos é uma tarefa complicada, demorada e cara. Em virtude dessa limitação experimental na tarefa de caracterização, utilizam-se métodos alternativos para a análise composicional como, por exemplo, equações empíricas.

Desta maneira, ao longo dos anos, diversos métodos empíricos de classificação de óleo cru foram propostos. Um método simples e direto de classificação é utilizar

as medições de densidade de maneira a indicar o conteúdo das frações leves. A densidade é uma informação muito importante, visto que ela reflete, em termos gerais, as frações de óleos leves e pesados da mistura. Outros métodos utilizam a correlação da densidade com outras propriedades como, por exemplo, viscosidade, também facilmente mensuráveis.

No entanto, é importante notar que ao combinar um subconjunto de propriedades selecionadas, corre-se o risco de limitar a abrangência do modelo de classificação utilizada. Ao descartar certas propriedades, perde-se informações que podem ser úteis quando analisadas a relação entre diferentes amostras. Ainda, ao considerar apenas estatística de primeira ordem, nuanças do comportamento dos compostos não são consideradas, prejudicando não só a modelagem obtida, como sua generalização.

Esses fatos motivam a utilização de técnicas não lineares de aprendizagem de máquina, como a seleção de características e o reconhecimento de padrões. A partir delas, é possível acessar e modelar informações utilizando características intrínsecas do espaço amostral. Tanto a natureza multidimensional da caracterização de petróleos, como o potencial volume de informações são fatores que corroboram para a utilização de algoritmos de inteligência computacional.

1.2 Objetivo

Esse projeto de pesquisa acontece dentro da colaboração entre o Laboratório de Processamentos de Sinais (LPS/COPPE) e a Petrobras. É visado desenvolver um método baseado em técnicas de mineração de dados cegas e de estatísticas de ordem superior para caracterizar amostras de petróleo de acordo com suas propriedades físico-químicas e, assim, propor uma solução mais adequada para o problema de estimação de propriedades citado. O objetivo principal é fornecer uma classificação mais completa e precisa quando comparada à baseada em densidade, sem aumento significativo nos custos de caracterização.

Entre os diversos modelos de aprendizagem de máquina, os algoritmos de clusterização podem ser considerados os mais adequados para problemas nos quais os dados não são rotulados por serem concisos, com muitas aplicações para compactação de dados e classificação. Os modelos aqui desenvolvidos baseiam-se neste paradigma de

aprendizagem não supervisionado e utilizam exclusivamente informações coletadas de esquemas de caracterização simples. Não é necessária uma informação *a priori* sobre as amostras de petróleo bruto.

Dentro desse trabalho, foram avaliados o impacto de cada propriedade na classificação das amostras de óleo bruto, em oposição a utilizar apenas medidas de densidade. Ao utilizar outras propriedades, espera-se que nuanças seja relevadas, possibilitando um modelo de classificação mais acurado. Para isso, adotamos foi realizado uma análise de exploração dos dados, utilizando duas abordagens independentes. Primeiro, cada propriedade foi projetada em um mapa SOM, com o intuito de entender a extensão da influência de cada variável. Em um segundo momento, realizou-se a fatoração dos dados utilizando NMF. Como resultado desta fatoração, são obtidas duas matrizes independentes, uma que correlaciona as propriedades aos agrupamentos desejados, ponderando sua importância e outra, associando cada a mostra a um agrupamento.

Após essa análise e consequente compreensão de como as diferentes propriedades influenciam na caracterização de cada amostra, diferentes métodos de classificação são explorados em conjunto a técnicas de compactação de dados, visto que as amostras são compostas por um número relativamente alto de propriedades. Os resultados são então comparados com a classificação baseada na medida de densidade.

Os agrupamentos obtidos pelas diferentes técnicas são posteriormente validadas por métricas que consideram a compactação intra-cluster e separação entre diferentes agrupamentos, de maneira a balancear as capacidades de generalização e especificidade da representação obtida.

Posteriormente, foram utilizados métodos de otimização baseados em metaheurísticas da natureza consolidados na literatura. Pelo seu sucesso em diversos problemas de reconhecimento de padrões, pesquisadores passaram a desenvolver técnicas de inteligência em enxames para particionar dados. Tais técnicas consistentemente apresentam desempenho superior aos algoritmos clássicos quando aplicados em complexas bases de dados. As figuras de mérito utilizadas na avaliação dos demais métodos serviram como heurísticas para esses algoritmos.

1.3 Organização do documento

Esse documento apresenta-se organizado da seguinte maneira:

- O Capítulo 2 introduzo problema de caracterização de petróleos e sua complexidade. São enumeradas alguns dos índices utilizados atualmente e que servirão como base de comparação no desenvolvimento deste trabalho. Neste capítulo também é apresentada a revisão bibliográfica em relação a utilização de técnicas de mineração de dados relacionadas com a classificação de hidrocarbonetos;

- o Capítulo 3 apresenta as técnicas cegas e de estatísticas de ordem superior utilizadas para esse trabalho. O capítulo tem ênfase nas etapas necessárias para a estimação de agrupamentos não supervisionados. São apresentadas abordagens para seleção de características, algoritmos de clusterização com diferentes heurísticas e as principais figuras de mérito utilizadas ao longo desse trabalho;

- No capítulo 4 é exposto o método proposto para realizar a exploração dos dados, estimar os padrões ocultos e agrupar as amostras de óleo cru disponíveis para essa pesquisa;

- os resultados da mineração de dados e das classificações realizadas são mostrados no Capítulo 5;

- por fim, o capítulo 6 apresenta as conclusões e enumera os desenvolvimentos futuros propostos, assim como potenciais desdobramentos.

Capítulo 2

Qualificação de Petróleos

Petróleo é o *commodity* mais importante do mundo, tanto em relação a sua produção, como por sua importância para a economia global. Trata-se ainda do principal combustível do mundo, somando aproximadamente 33 % do consumo total de energia global. Aproximadamente 63 % do óleo consumido vem do setor de transporte e a substituição por combustíveis alternativos não é eminente, havendo previsão para esse percentual diminuir no máximo 5 % nos próximos 5 anos [2]. A Figura 2.1 apresenta o consumo para as principais fontes de energia primária.

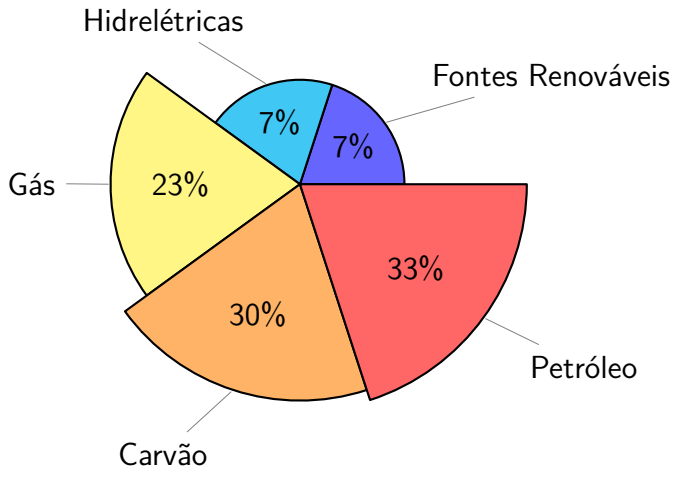


Figura 2.1: Consumo mundial de energia primária. Retirado de [2].

A Agência Internacional de Energia (IEA) prevê que a demanda por petróleo crescerá para um pouco mais do que 100 milhões de barris por dia devido a recuperação econômica global (um acréscimo da demanda de 1,4 milhões bpd em relação a 2017) [3]. A Figura 2.2 mostra os dados históricos de oferta e demanda de petróleo, assim como a projeção da IEA para 2018. No gráfico, verifica-se a tendência de

crescimento tanto na disponibilidade quanto na demanda de petróleo do mercado.

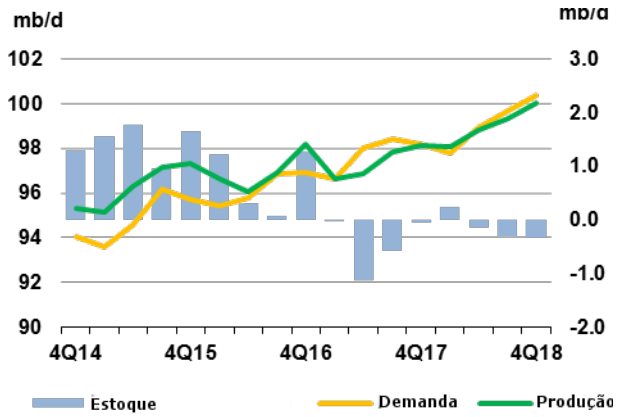


Figura 2.2: Gráfico demostrando a evolução da demanda e produção de petróleo em barris por dia e a previsão até o quarto trimestre de 2018. Dados disponibilizados pela Agência Internacional de Energia (IEA).

A Figura 2.3 apresenta um mapa onde estão destacados os principais países produtores do mundo (membros e não-membros da OPEP). A Arábia Saudita se destaca como maior produtor mundial, com uma produção estimada em 9,5 milhões de barris por dia, e plano de expansão para 12,5 milhões nos próximos 4 anos, segundo dados da Organização para a Cooperação e Desenvolvimento Econômico (OECD) [4].

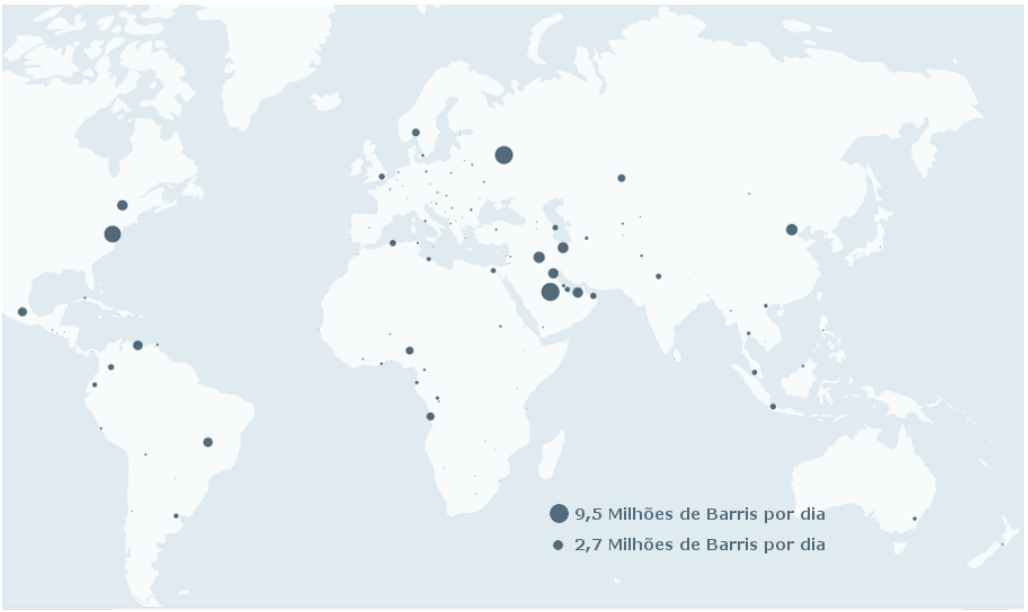


Figura 2.3: Principais produtores de petróleo do mundo. Dados disponibilizados pela Organização para a Cooperação e Desenvolvimento Econômico (OECD).

O Brasil acompanha a tendência global no aumento de produção, impulsionado pela exploração dos campos de pré-sal [5]. O ritmo de descobertas de petróleo no

país a partir dos anos 2000 o posiciona como quarto no mundo que mais aumentou o seu volume de reservas [6]. A Petrobras projeta um aumento na produção de 38 % entre 2018 e 2022 [7].

2.1 Complexidade na composição do petróleo

O petróleo é uma mistura de gases, líquidos e sólidos. É constituído predominantemente por hidrocarbonetos e derivados orgânicos sulfurados, nitrogenados e oxigenados [8]. Os hidrocarbonetos alcançam mais de 90 % da composição. Também encontram-se nessa mistura, compostos organometálicos, enxofre na forma inorgânica (como gás sulfúrico) e metais formadores de sais de ácidos. De forma geral, quanto mais pesado as frações do petróleo, maior o teor de contaminantes [6]. Resumidamente, os constituintes do petróleo podem ser divididos em duas classes: hidrocarbonetos e não hidrocarbonetos. A Figura 2.4 apresenta de maneira esquematizada os constituintes do petróleo e como são classificados.

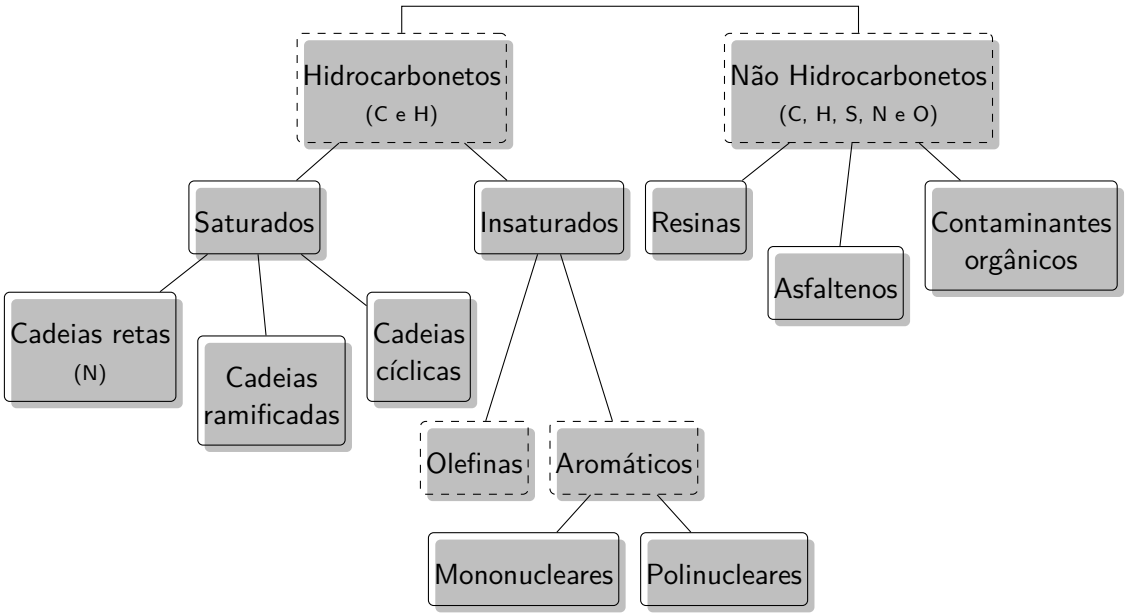


Figura 2.4: Constituição do Petróleo. Adaptado de [6]

Os óleos obtidos de diferentes reservatórios de petróleo possuem características diferentes. Circunstâncias locais durante a formação do óleo são variáveis, o que resulta em uma miríade de diferentes tipos de petróleo bruto com diferentes propriedades químicas e físicas ao redor do mundo. Podem ser pretos, castanhos ou

bastante claros, com densidade e viscosidade variadas, liberando quantidades distintas de gás [9].

As diferentes propriedades afetam significativamente o valor econômico do petróleo bruto e também todos as etapas da sua cadeia de abastecimento, incluindo o *upstream*, a cadeia logística e o *downstream*. A partir destas propriedades é definido o potencial econômico e se é justificável explorar um reservatório de óleo.

A obtenção dos subprodutos do petróleo bruto ocorre através do processo de destilação fracionada, no qual o material é submetido a aquecimento em tanques especializados. Os derivados resultantes são hidrocarbonetos sendo os mais leves compostos por pequenas moléculas e os mais pesados contendo até 70 átomos de carbono. O processo de destilação ocorre justamente pela diferença de tamanho das moléculas, onde quanto menor a molécula de hidrocarboneto, menor é a sua densidade e temperatura de ebulição. Deste modo, cada subproduto é obtido em temperaturas bem específicas. A Tabela 2.1 mostra as frações típicas obtidas do petróleo.

Corte	Temperatura de ebulição (^{o}C)	Composição aproximada	Usos
Gás Residual	—	C_1 — C_2	gás combustível
GLP	Até 40	C_3 — C_4	gás combustível engarrafado, uso doméstico e industrial
Gasolina	40 — 175	C_5 — C_{10}	combustível de automóveis, solvente.
Querosene	175 — 235	C_{11} — C_{12}	iluminação, combustível de aviões a jato
Gasóleo leve	235 — 305	C_{13} — C_{17}	diesel, fornos
Gasóleo pesado	305 — 400	C_{18} — C_{25}	combustível, matéria-prima p/ lubrificantes
Lubrificantes	400 — 510	C_{26} — C_{38}	óleos lubrificantes
Resíduo	Acima de 510	C_{38+}	asfalto, piche, impermeabilizantes

Tabela 2.1: Frações típicas do petróleo. Retirado de [9].

Pela complexidade de sua composição, o petróleo e seus derivados devem ser caracterizados por métodos que permitam conhecer seu comportamento físico e químico, quando submetidos a diferentes condições de operação [10]. Alguns dos objetivos de caracterizar corretamente o petróleo podem ser listados, como por exemplo:

- Refinadores estão interessados na quantidade das sucessivas frações de destilação e na composição química ou propriedades físicas destas frações [11];
- Geólogos procuram identificar e caracterizar os petróleos para relacioná-los à rocha geradora e medir seu grau de evolução [11];
- Durante o transporte e o processamento, conhecer os parâmetros físicos como o ponto de ebulição é crucial para evitar problemas como obstrução ou solidificação [12];
- A operação dos equipamentos da cadeia de processamento dependem das propriedades dos fluídos processados como a densidade, temperatura de ebulição, viscosidade, etc [13];
- Proporcionar um melhor desempenho na extração de subprodutos quando, na fase final de formulação, promove-se a mistura dos componentes de petróleo [10].

Há uma dificuldade prática de se determinar analiticamente a composição de petróleos, pois sua caracterização química só pode ser feita de modo completo apenas para frações leves. Apesar dos grandes avanços das técnicas experimentais nas últimas décadas, uma identificação completa de todas as frações extraídas de um petróleo bruto ainda é muito complicada, demorada e cara [14].

2.2 Ensaios de Petróleo Bruto

A indústria petrolífera costuma realizar medições de propriedades físico-químicas e, assim, estimar qual o tipo de composto predominante numa mistura ou óleo [11]. São os chamados ensaios de petróleo bruto [12].

Convencionalmente, os ensaios de óleo bruto incluem testes relativamente simples para determinar, por exemplo, a densidade, viscosidade e teor total de enxofre, mas também testes bastante complexos, como a caracterização das frações de intervalo de ebulição por meio de destilação física ou artificial.

Tradicionalmente, a maioria desses testes (também conhecidos como via úmida) são realizadas por métodos padrões desenvolvidos pela American Society for Testing

and Materials International [15] e o Instituto de Energia, anteriormente conhecido como Instituto de Petróleo [16]. No entanto, dependendo dos testes laboratoriais, um ensaio de petróleo bruto pode levar de 3 dias a 5 semanas a ser concluído, com custos variando de 2.500 a 30.000 dólares [17, 18].

2.3 Classificação de Petróleos

A tarefa de agrupar óleos cru em uma qualificação padrão é um desafio antigo enfrentado pelos profissionais da indústria do petróleo. Assim, métodos alternativos foram desenvolvidos ao longo dos anos para a determinação da composição química de frações de petróleo, utilizando-se para isso equações empíricas e modelagem composicional [12]. Esses métodos empíricos buscam, através de correlações que utilizam propriedades facilmente medidas tais como, temperatura de ebulição, densidade, viscosidade, obter a composição química das frações [19]. Tais métodos, no entanto, possuem validade em um conjunto limitado de óleos e poucos destes modelos apresentam boa precisão em uma faixa ampla de aplicações [6].

2.3.1 Classificação de petróleos pela densidade

Uma classificação típica para óleos brutos utilizada é relacionada ao *Grau API*, ou *°API*, medida que mede a densidade relativa dos líquidos derivados do petróleo em relação à água [20]. A densidade é uma informação importante, pois reflete, em termos gerais, o conteúdo leve e pesado das frações do petróleo. Através de seus valores podemos classificar o óleo como: leve, médio, pesado e extra-pesado [21].

Um classificação frequentemente referida foi obtida através da análise de 500 amostras de todos os tipos de petróleos processados no Brasil [6]. A Tabela 2.2 apresenta os limiares para cada uma dessas classes de petróleo.

Observa-se que a medida *°API* apresenta uma relação inversamente proporcional à gravidade específica (SG, do inglês *specific gravity*) do óleo. Essa relação pode ser estabelecida como:

$$\mathrm{SG} = \frac{141,5}{^{\circ}\mathrm{API} + 131,5}$$

onde *SG* denota a densidade relativa do óleo em relação à água. Deste modo, as classes podem ser descritas de acordo com a Tabela 2.3.

Classe	Mínimo	Máximo
Leve	31,1	
Médio	22,3	31,1
Pesado	10,0	22,3
Extra-pesado		10,0

Tabela 2.2: Definição de classes considerando o a medida de densidade em °API.

Classe	Mínimo	Máximo
Leve		0,825
Médio	0,825	0,875
Pesado	0,875	1,0
Extra-pesado	1,0	

Tabela 2.3: Definição de classes considerando a medida de gravidade específica.

2.3.2 Classificação por meio de indicador da relação viscosidade-densidade

A constante viscosidade-gravidade (VCG) é bastante utilizada para a classificação de óleos brutos, pois pode ser calculado a partir de propriedades facilmente obtidas, contudo seus valores são apenas válidos para regiões específicas do planeta, pois dependem de temperaturas e viscosidades específicas.

Éigenson (1989) estabelece de maneira empírica uma classificação para dividir petróleos em duas classes: parafínicos e naftênicos, usando a viscosidade, na forma da função ($w_\theta = \log\log(v_\theta + 0.8)$). A função w_θ for originalmente proposta por Walther para descrever a relação entre a viscosidade dos líquidos crus e temperaturas, da seguinte maneira [22]:

$$\begin{aligned} w_\theta &= A - B\log(273,2+\theta) \\ w_{\theta_2} &= w_{\theta_1} - B\log(\frac{273,2+\theta_1}{273,2+\theta_2}) \end{aligned} \tag{2.1}$$

onde A e B são coeficientes a serem determinados.

Ao comparar os valores de °API aos da função w_{50}, foi possível estabelecer uma correlação linear entre as duas propriedades, e, a partir daí, projetar um classificador

linear.

2.3.3 Classificação baseada na razão densidade-viscosidade

Farah (2006) [10] propôs um índice novo visando distinguir compostos parafínicos, naftênicos e aromáticos. Utilizou para isso o fator °API / (A/B), onde A e B são os parâmetros da equação de Walther para a viscosidade, característicos de cada substância. Observou-se que este fator apresentou valores bem definidos para as classes de hidrocarbonetos parafínicos, naftênicos, benzeno (e alquia benzenos) e naftaleno (e alquia naftalenos) [19]. A Tabela 2.4 mostra a definição das classes considerando esse índice.

Classe	Mínimo	Máximo
Aromático-asfáltico	4	6
Aromático-naftênico	6	8
Aromático-intermediário	8	10
Naftênico	10	12
Parafínico-naftênico	12	14
Parafínico	14	

Tabela 2.4: Definição de classes considerando o fator °API / (A/B)

Esse índice se relaciona com as categorias baseadas na composição do petróleo que podem ser descritas da seguinte maneira:

Classe parafínica óleos leves, fluidos ou de alto ponto de fluidez, com densidade inferior a 0,85, teor de resina e viscosidade baixa. Apresenta baixo teor de enxofre. A maior parte dos petróleos produzidos no nordeste brasileiro se enquadra nessa classe. São óleos excelentes para a produção de querosene de aviação, diesel, lubrificantes e parafinas [9].

Classe parafínica-naftênica óleos que apresentam teor de resina e asfaltenos um pouco maior que a classe anterior, com baixo teor de enxofre (1%) e teor de naftênicos entre (25 e 40%). A densidade e viscosidade apresentam valores maiores do que os parafínicos, mas ainda são moderados. A maioria dos petróleos oriundos da Bacia de Campos é deste tipo [9].

Classe naftênica óleos com baixo teor de enxofre que se originam da alteração bioquímica de óleos parafínicos e parafínico-naftênicos. Alguns óleos da América do Sul, Rússia e do Mar do Norte pertencem a esta classe que produz frações significativas de gasolina e nafta petroquímica [9].

Classe aromática intermediária óleos frequentemente pesados, contendo 10 a 30% de asfaltenos e resinas e teor de enxofre acima de 1%. A densidade usualmente é maior que 0,85. Esses óleos são encontrados no Oriente Médio, Africa Ocidental, Venezuela, Califórnia e Mediterrâneo (Sicília, Espanha e Grécia).

Classe aromática-naftênica óleos derivados dos parafínicos e parafínico-naftênicos, podendo conter mais de 25% de resinas e asfaltenos, teor de enxofre próximo a 1%. Alguns óleos são encontrados na Africa Ocidental [9].

Classe aromática-asfáltica óleos oriundos de um processo de biodegradação avançada. Podem também se enquadrar alguns poucos óleos verdadeiramente aromáticos não degradados da Venezuela e África Ocidental. Entretanto ela compreende principalmente óleos pesados e viscosos como encontrado no Canadá ocidental, Venezuela e sul da França [9].

2.4 Métodos baseados em reconhecimento de padrões

Apesar de simples de serem aplicados, as classificações apresentadas até aqui, apenas consideram propriedades, como medidas de viscosidade e densidade, descartando estruturas intrínsecas e padrões provenientes de outras. Ao desconsiderar essas informações e não levar em conta importante nuanças, corre-se o risco de desenvolver modelos para estimação de propriedades de óleo cru que tenham baixo desempenho.

Devido quantidade de petróleos existentes, uma abordagem indicada para desenvolver modelos acurados é tentar agruparem um número limitado de grupos suas amostras através das possíveis graus similaridades utilizando técnicas consolidadas de reconhecimento de padrões.

Na literatura podemos encontrar a aplicação destas técnicas principalmente visando a identificação de origem geográfica de derramamentos de óleo cru em costas e oceanos. Alguns autores realizaram análises com componentes principais (PCA) combinando-as com mapas auto-organizáveis (SOM) (os dois algoritmos serão detalhados na Seção 3) para classificar amostras de petróleo em relação a sua origem geográfica a partir de dados coletados com cromatografia gasosa com espectrômetro de massa acoplado (GC-MS) [23].

Fonseca (2006) [24] combinou as duas técnicas para obter a origem de amostras petróleos contidas em duas bases de dados distintas: uma base do Instituto Hidrográfico de Lisboa, contendo 188 amostras, cada uma com 21 dimensões [25] e outra do sistema EUROCRUDE® que conta com 374 amostras, contendo 56 dimensões [26]. Ambas as bases possuem amostras provenientes de 20 origens geográficas distintas. O trabalho conclui que a as redes SOM puderam classificar a origem das amostras, principalmente as regiões bem representadas no banco de dados, chegando a uma acurácia de 96 % para esses casos. Fernández-Varela [27, 28] utiliza a mesma combinação de técnicas para categorizar os tipos de produtos destilados presentes em 45 amostras de poluentes coletados em diversas praias da região da Galícia [29]. Borges (2010) [30] utilizou mapas SOM para classificar 38 amostras de óleo cru, provenientes de 6 regiões geográficas, obtendo aproximadamente 89 % de precisão média. Já Carvalho (2017) [31] desenvolveu uma ferramenta de classificação para hidrocarbonetos como um prelúdio do desenvolvimento de um modelo preditivo de propriedades físico-químicas relevantes para motores com o objetivo final de antever a adequação destes motores a novos biocombustíveis.

Cabe ressaltar que organizações da indústria de Óleo e Gás recorrem cada vez mais a algoritmos de mineração de dados para apoiar sua operação [1]. Detecção de anomalias [32], séries temporais [33], processamento de imagens de satélites [34–36], análises espaciais [37] são algumas das aplicações que já impactam corporações e institutos de pesquisa.

Capítulo 3

Agrupamento de dados

O problema de agrupar dados de acordo com suas características intrínsecas é amplamente abordado na literatura de mineração de dados e aprendizagem de máquinas devido às inúmeras aplicações em sumarização, categorização e segmentação. O problema básico da clusterização de dados pode ser apresentado, em linhas gerais, como: "Dado um conjunto de pontos, particioná-los em conjuntos de elementos mais similares possíveis" [38]. O cálculo da similaridade entre os diferentes grupos encontrados após processo de clusterização pode ser realizado de diversas maneiras [39].

Sob a perspectiva de otimização, o principal objetivo dos algoritmos de clusterização é maximizar tanto a homogeneidade interna de um grupo/cluster como a heterogeneidade entre diferentes grupos.

Os algoritmos de clusterização utilizam dados que, normalmente, não são previamente classificados ou mesmo conhecidos (informações não rotuladas), sendo, por essa razão, classificados como pertencentes à classe de algoritmos não supervisionados. Essas etapas podem ser resumidamente resumidas da seguinte maneira [40]:

HANCER (2017) define as etapas apresentadas na Figura 3.1 como fundamentais para realizar uma análise de agrupamentos.

Seleção de características Essa etapa objetiva eliminar as características que causam efeitos adversos no desempenho do algoritmo e reduzir o números de dimensões envolvidas no problema abordado.

Execução do algoritmo de clusterização O objetivo dessa etapa é determinar o algoritmo mais apto para o problema abordado. Não existe uma técnica

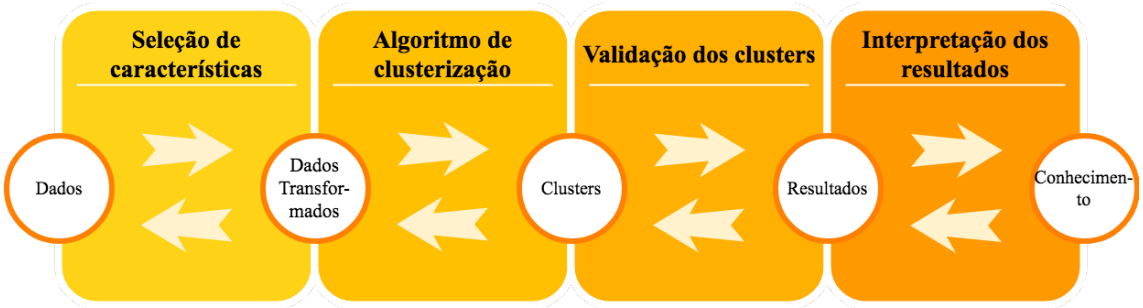


Figura 3.1: Etapas fundamentais em análises com agrupamento de dados. Adaptado de [41].

universal de clusterização consolidada na literatura.

Validação dos agrupamentos obtidos Os clusters obtidos serão avaliados por critérios e índices específicos, que podem também variar conforme o problema.

Interpretação dos resultados os resultados obtidos pelos algoritmos utilizados são processados e analisados para alcançar os resultados esperados pelos *experts*.

Nesse trabalho, foram utilizados diferentes métodos de clusterização, visando objetivos diferentes dentro da pesquisa. Em um primeiro momento, algoritmos exploratórios foram usados com intuito de possibilitar a compreensão da importância das dimensões utilizadas, bem como a influência de cada uma delas na caracterização das amostras disponíveis. Este estudo aprofundado visou criar subsídios para o estudo seguinte, onde um novo processo de clusterização foi realizado a fim de obter uma classificação baseada nas relações intrínsecas dos dados. Com esses objetivos delineados, diferentes configurações foram escolhidas. Este capítulo mostrará as diferentes técnicas de preprocessamento, os algoritmos de clusterização e as medidas de validação dos agrupamentos que foram usadas dentro do método de pesquisa.

3.1 Seleção de Características

A extração de características é uma importante etapa de preprocessamento, necessária para melhorar a qualidade do processo de clusterização. Nem todas as características são igualmente relevantes para a realização de agrupamentos. Para tal, pode-se realizar a seleção de características ou a redução de dimensionalidade.

Enquanto a primeira exclui subconjuntos de características selecionadas, a segunda realiza combinações lineares dessas características usando técnicas como Análise dos Componentes Principais (PCA, do inglês *Principal Component Analysis*) [42, 43] ou a Fatoração de Matrizes Não-Negativas (NMF, do inglês *Non-Negative Matrix Factorization*) [44, 45].

3.1.1 Análise dos Componentes Principais

A PCA é uma abordagem clássica para a redução de dimensionalidade e extração de características [43]. A análise parte do principio que há correlação entre os componentes dos dados originais e, portanto, informação redundante. Essa redundância pode ser removida através de uma transformação linear, que deve ser encontrada. No novo espaço, os componentes são reordenados em ordem decrescente de variância.

A forma fechada de encontrar uma projeção que atenda o critério de descorrelacionamento é através da base formada pelos autovetores da matriz de correlação C_x, onde x são os vetores do sinal original centrados, ou seja, que tiveram a média removida. Esses autovetores serão então os componentes principais, e os autovalores correspondentes darão a variância, ou seja, a energia do sinal correspondente ao componente. Tipicamente usa-se o método de decomposição em valores singulares (SVD, da sigla em inglês para *Singular-value decomposition*) [46].

Na Equação 3.1, se as direções principais são representadas por u_j, podemos definir a transformação linearmente ortogonal dos dados x como:

$$a_j = u_j^T x = x^T u_j, j = 0, \ldots, n-1 \tag{3.1}$$

onde a_j é a projeção dos dados no espaço das componentes principais. A compactação do sinal pode ser feita selecionando-se somente os componentes com maiores autovalores. Analisando-se a soma cumulativa dos autovalores já ordenados, pode-se determinar o número de componentes baseado na quantidade de energia total do sinal original. Em casos práticos, a maior parte da energia se deposita sobre um conjunto pequeno de componentes principais [46]. É importante notar que a PCA apenas considera estatísticas de segunda ordem.

3.1.2 ISOMAP

O espaço dos componentes principais não necessariamente é o espaço que melhor caracteriza a estrutura intrínseca dos dados. O algoritmo ISOMAP [47] realiza a redução de dimensionalidade explorando a distância geodésica entre diferentes pontos num espaço vetorial. A distância geodésica é o menor a caminho entre quaisquer dois pontos num espaço curvo que tenha uma geometria de Riemann [48]. Ele se utiliza da estimativa da distância geodésica entre dois pontos no espaço para encontrar Uma representação de baixa dimensionalidade que mantenha a estrutura do espaço original. Sua grande vantagem em relação a outros métodos, como PCA, é o uso de uma estimativa de distância mínima entre dois pontos mais robusta que a distância euclidiana [47].

O exemplo clássico é a presença de um *manifold* de duas dimensões representado num espaço de três dimensões. A Figura 3.2 mostra a ilustração de um *manifold* 2D representado num espaço de dimensão superior. Assim, apesar de descrito num espaço 3D, sua dimensionalidade intrínseca é 2D.

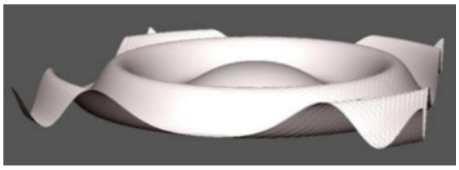


Figura 3.2: Exemplo de uma estrutura de 2D (*manifold*) representada num espaço 3D.

A distância geodésica entre dois pontos é estimada como a soma de inúmeros pequenos segmentos de reta numa vizinhança, onde a distância euclidiana é considerada válida para representar a menor distância entre dois pontos. Essa vizinhança pode ser definida por um raio em relação a uma dada amostra, ou pelos k pontos mais próximos dessa amostra. Determinadas vizinhanças, monta-se um grafo ligando cada ponto com seus vizinhos. Por fim, a distância geodésica entre dois pontos é estimada pelo menor caminho, no grafo, entre esses dois pontos. Nota-se que quanto maior for o número de amostras, melhor será a estimativa desta distância [47].

3.1.3 Fatoração de Matrizes Não-Negativas

NMF e sua extensão NTF (do inglês, *Nonnegative Tensor Factorization*) são apontadas como uma maneira esparsa e eficiente de representar sinais, imagens e dados em geral [44]. Do ponto de vista de processamento de sinais, NMF é muito interessante porque leva em conta correlações espaciais e temporais entre variáveis e frequentemente provê fatores esparsos ou componentes não negativos latentes com significação e interpretação físicos [44].

Seja uma matriz de entrada $X = (x_1, \ldots, x_n)$ que contem uma coleção de n vetores-coluna de dados. Fatora-se X em duas matrizes:

$$X \approx FG^T \tag{3.2}$$

onde, $X \in \mathbb{R}^{p \times n}$, $F \in \mathbb{R}^{p \times k}$ e $G \in \mathbb{R}^{n \times k}$. Para $p < n$ e o *rank* das matrizes F, G é muito menor do que o *rank* de X. F e G são minimizados diminuindo uma função custo. A função custo mais comum é a soma dos erros quadráticos:

$$\min J_{\text{seq}} = ||X - FG^T||^2 \tag{3.3}$$

Uma das propriedades mais interessantes da NMF, é frequentemente é produzir uma representação esparsa dos dados [49]. Em aprendizagem de máquinas, esparsidade é associada à seleção de características, enquanto não negatividade é relacionada com distribuições de probabilidade [44]. Assim, uma representação esparsa e não negativa permite codificar os dados usando poucos componentes ativos, o que torna a codificação simples para ser interpretada [49].

Ao contrário de outros algoritmos, a NMF é um modelo aditivo, o qual não permite subtrações. Desta maneira, é capaz de descrever as partes que compõe a entidade como um todo. Em outras palavras, NMF pode ser considerada como uma representação onde a nulidade representa a ausência e a positividade representa a presença de um evento ou componente [44]. A restrição por valores não negativos torna NMF um problema NP-difícil [50].

Por essas características, diversas aplicações de NMF podem ser encontradas na literatura, como, por exemplo: mineração de texto [51], processamento de imagens [52], separação ceda de fontes [53], análise musical [54], detecção de comunida-

des [55], filtros colaborativos [56] e clusterização [57]. As aplicações relacionadas a tarefa de clusterização serão abordadas mais adiante, ainda nesse capítulo.

A Equação 3.3 assume que a fatoração é usada com dados gaussianos, o que muitas vezes não é verdade [58]. Na prática, diversas funções objetivos podem ser usadas, como por exemplo, a divergência de Kullback-Leibler, comum em aplicações de *text mining* [59] e a distância de Itakura-Saito, indicada em análises com música [60, 61].

Na literatura, existem diversas implementações de algoritmos de NMF [62], onde a maioria utiliza o método em busca coordenada de dois blocos [58]. Esse método atualiza um dos fatores (F ou G) alternadamente fixando os valores do outro. O Algoritmo 1 apresenta como a solução baseada em busca coordenada é estruturada. A escolha do posto r é particularmente desafiadora [63]. Algumas das abordagens utilizadas variam desde tentativa e erro, estimação através de decomposição em valores singulares, uso de conhecimento especialista sobre o problema em questão ou, ainda, o uso da técnica de validação cruzada.

Algoritmo 1: Solução de NMF baseado em busca coordenada

Dados: Matriz não-negativa $X \in \mathbb{R}_+^{p \times n}$ e o posto r.
Resultado: $(F, G) \geq 0$: uma fatoração de posto r, onde $X \approx FG^T$

```
Gerar matrizes iniciais F ≥ 0 e G ≥ 0
enquanto critério de parada não for atingido faça
    F^(t) = atualiza (X, F^(t−1), G^(t−1))
    G^(t) = atualiza (X, F^(t), G^(t−1))
fim
```

Uma maneira popular de realizar a etapa de atualização é através das *atualizações multiplicativas* (MU, do inglês *multiplicative updates*) [64, 65], baseada na estratégia de majoração-minimização (MM) [66]. Dados X, F e G, MU modifica F através da, Equação 3.4:

$$F \leftarrow F \circ \frac{XG^T}{FGG^T} \tag{3.4}$$

Entre as vantagens dessa abordagem estão a sua simples implementação, sua característica escalável e sua aplicabilidade em matrizes esparsas. Um ponto negativo é a sua convergência lenta [67].

NMF com divergência β

A divergência β é uma família de funções de custo parametrizadas por um único parâmetro β que assume a forma da distância Euclideana, a divergência de Kullback-Leibler [68] e a divergência de Itakura-Saito [69] como casos especiais ($\beta = 2, 1, 0$ respectivamente) [70]. Ela foi proposta por Basu et al. (1998) e Eguchi (2001) e pode ser definida como se segue [71, 72]:

$$d_\beta(x|y) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{\beta(\beta+1)}(x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}) & \beta \in \mathbb{R}/\{0,1\} \\ x\log(\frac{x}{y}) - x - y & \beta = 1 \\ \frac{x}{y} - \log(\frac{x}{y}) - 1 & \beta = 0 \end{cases} \tag{3.5}$$

A Figura 3.3 apresenta como a curva se comporta para os diferentes intervalos de β.

A variação da função objetivo em relação ao valor de β, permite que diferentes características do problema sejam acessadas e exploradas. Por exemplo, em Févotte (2009) foi mostrado que fatorizações com valores pequenos de β são relevantes para decomposição de espectros de áudio, pois são tipicamente caracterizados por componentes de baixa potência, normalmente ignorados ou degradados quando usados com divergências KL ou Euclidiana [60].

Por sua vez, a divergência de Kullback-Leibler é oriunda da Teoria da Informação e trata-se de uma medida de entropia cruzada [68, 73]. Ao utilizá-las como função custo da NMF, pode-se mensurar o quanto de informação é perdida no processo de fatoração. Ao mesmo tempo, o resultante passa a ter uma interpretação probabilística, baseada na estimação de máxima verossimilhança entre a distribuição variável-componente e a distribuição componente [74].

3.2 Mapas Auto Organizáveis

O objetivo de um trabalho de exploração é ganhar familiaridade com os dados disponíveis, e, assim, ser capaz de determinar se estes são suficientes e de escolher o preprocessamento adequado, bem como um modelo satisfatório [75].

Uma maneira eficiente de ganhar esta familiaridade é através da visualização das

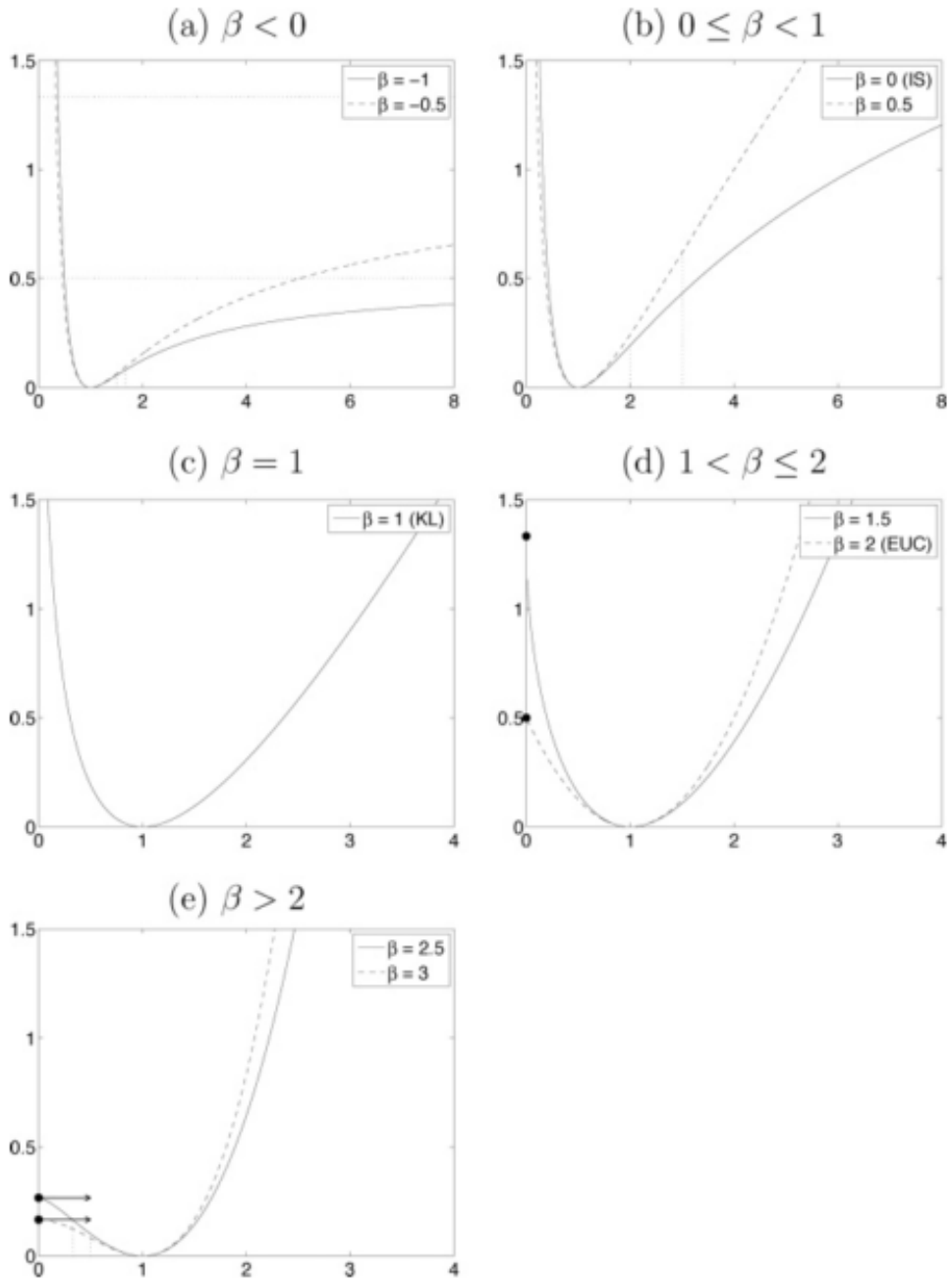


Figura 3.3: Divergências β em função de y. Cada um dos gráficos ilustram os regimes das divergências para os cinco intervalos de valores característicos para β. Retirado de [70].

informações disponíveis. Os Mapas Auto-organizáveis (ou Rede Neural de Kohonen, ou ainda Mapa de Kohonen)<++> [76] são especialmente interessantes para este propósito devido suas propriedades [75].

O SOM é uma rede neural com treinamento não supervisionado, baseado na aprendizagem competitiva, capaz de realizar um organização topológica das entradas fornecidas. Executa um mapeamento não linear dos sinais de um espaço de entrada contínuo de dimensão k para um espaço de características discreto e bidimensional [77]. Na Figura 3.4, pode-se visualizar o diagrama de um mapa auto-organizável bidimensional.

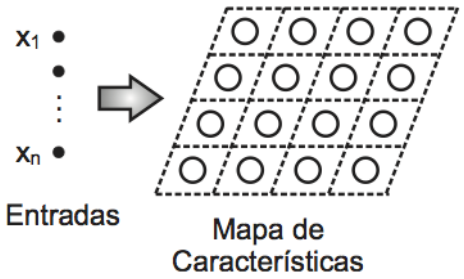


Figura 3.4: Diagrama do SOM. Retirado de [77].

A formação do SOM ocorre através de três processos:

Competição

Para cada vetor de entrada, há apenas um neurônio vencedor. Seja um vetor de entrada $x = x_1, \ldots, x_D$ e as sinapses entre as entradas i e os neurônios j na camada intermediária $w_j = w_{ji} : j = 1, \ldots, N; i = 1, \ldots, D$, onde N é o número de neurônios. Pode-se definir a função discriminante ser a distância Euclidiana quadrática entre o vetor de entrada x e o vetor de pesos wj para cada neurônio j.

$$d_j(x) = \sum_{i=1}^{D} (x_i - w_{ji})^2 \tag{3.6}$$

Em outras palavras, os neurônios cujo o vetor peso estiver mais próximo ao vetor de entrada, ou seja, o mais similar, é declarado vencedor.

Cooperação

Neurônios de regiões adjacentes irão ser mover para uma mesma direção. No

cérebro humano, há evidências da existência de uma interação lateral em um conjunto de neurônios excitados. Em particular, um neurônio que é ativado tende a excitar mais os neurônios que estão imediatamente a sua volta do que os localizados mais distantes [78].

Esse fenômeno inspirou a introduzir no SOM uma vizinhança topológica em torno do neurônio vencedor i e fazê-los decair suavemente de acordo com a distância lateral. Seja:

$$T_{j,I(x)} = \exp(\frac{-S_{j,I(x)}^2}{2\sigma^2}) \tag{3.7}$$

Onde $T_{j,I(x)}$ denota a vizinhança topológica, onde $I(x)$ representa o neurônio vencedor. Essa função tem inúmeras propriedades relevantes: é máxima no neurônio vencedor, é simétrica em torno do neurônio, decresce a zero conforme a distância tende a infinito e é invariante à posição do neurônio vencedor.

Uma característica especial do SOM é que o tamanho da σ vizinhança necessita diminuir com o tempo, normalmente de maneira exponencial.

Adaptativo

Reforça a resposta do neurônio vencedor, e de seus vizinhos, ao padrão de entrada após o ajuste dos pesos sinápticos. É nessa etapa em que as saídas se auto organizam e o mapa de características entre a entrada e a saída se formam.

A motivação em haver uma vizinhança topográfica é que não somente o neurônio vencedor tenha seus pesos atualizados, mas seus vizinhos também. Na prática, a atualização dos pesos ocorre de acordo com a equação 3.8:

$$\Delta w_{ij} = \eta(t)\dot{T}_{j,I(x)}(t)\dot{(}x_i - w_{ji}) \tag{3.8}$$

onde, cada época t é dependente da taxa de aprendizado $\eta(t) = \eta_0 \exp(\frac{-t}{\tau_n})$. O efeito de cada atualização é mover o vetor de pesos w_i do neurônio vencedor e seus vizinhos em direção ao vetor de entrada x.

Desta maneira, o SOM forma um mapa semântico, onde eventos semelhantes

são mapeados conjuntamente e os distintos separados. Esse mapeamento pode ser visualizado através da *U-Matrix* do SOM [79]. Esta utiliza a mesma métrica usada durante o treinamento para calcular distâncias entre pesos de neurônios vizinhos. Como resultado, obtém-se uma matriz que pode ser interpretada como imagem, onde cada coordenada (x, y) são derivadas das coordenadas dos neurônios no *grid* do mapa, e a intensidade corresponde a distância calculada [80]. Esta visualização fornece a possibilidade de uma análise qualitativa dos dados.

Para produzir uma descrição quantitativa dos dados, devem ser selecionados grupos de interesse dentro do mapa gerado. Para tal, utilizam-se as informações geradas pelo SOM em conjunto a outros métodos de agrupamento, como K-médias, por exemplo. Assim, é possível possuir os agrupamentos presentes numa base de dados, rapidamente, de maneira robusta e com uma visualização eficiente.

3.3 Algoritmos Baseados em Distâncias

Métodos baseados em distâncias são frequentemente desejáveis devido sua simplicidade e a facilidade de implementação em uma ampla variedade de cenários. Podem ser separados em basicamente dois tipos: algoritmos planos e hierárquicos.

3.3.1 Algoritmo K-médias

Nos algoritmos planos, os dados são divididos em vários grupos de uma só vez, normalmente com o uso de representantes do particionamento. A escolha da função representativa de particionamento e a de distância são cruciais e regulam o comportamento do algoritmo subjacente. Em cada iteração, os pontos de dados são atribuídos aos seus representantes mais próximos de separação e, em seguida, o representante é ajustado de acordo com os pontos de dados atribuído ao cluster [38].

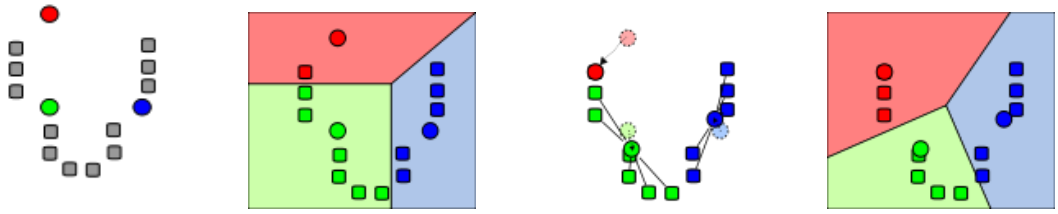
O algoritmo mais utilizado nessa classe é conhecido como K-médias [81], empregando o conceito de centroides. Dados os K centroides espalhados aleatoriamente no espaço de dados, sendo K igual ao número de agrupamentos pré-definidos, o algoritmo agrupa os eventos, de acordo com o distanciamento entre o evento e o centroide. Forma-se então um diagrama de Voronoi. Como métrica, utiliza-se a distância euclidiana quadrática:

$$d_{ki}^2 = \| x_i - c_k \|^2 \tag{3.9}$$

onde x_i são os eventos do conjunto de dados e c_k são os centroides dos agrupamentos.

A seguir, os centroides são recalculados como o baricentro dos eventos associados aos seus agrupamentos, redefinindo o diagrama de Voronoi. Esse processo é repetido até os centroides não se desloquem mais ou quando um determinado número de iterações do algoritmo for realizado.

A Figura 3.5 apresenta graficamente os passos que o algoritmo utiliza.



(a) $K = 3$ são aleatoriamente geradas dentro do espaço dos dados

(b) K agrupamentos são criados.

(c) O centroide de cada agrupamento se torna a nova média

(d) O algoritmo é repetido até convergir.

Figura 3.5: Exemplo do algoritmo de *K-Means*

3.3.2 Algoritmo Hierárquicos

Nesses métodos, os agrupamentos são representados hierarquicamente através de um *dendrograma*, variando os níveis de granularidade. Dependendo se essa representação hierárquica é criada de cima para baixo ou vice-versa, ela pode considerada aglomerativa ou divisiva.

Nos métodos aglomerativos, a abordagem é montar a estrutura de baixo para cima, onde o algoritmo começa a partir dos dados individuais e vai sucessivamente combinando agrupamentos de tal forma que uma estrutura em árvore é montada [82]. Há uma série de opções em como os agrupamentos podem ser aglomerados, balanceando qualidade e eficiência. Alguns exemplos são o agrupamento por vizinhos mais próximo [83] e agrupamento por ligação completa [84].

Nos métodos divisivos, a abordagem é feita de cima para baixo de maneira a sucessivamente dividir as amostras em uma estrutura tipo árvore. Qualquer algoritmo de clusterização plana pode ser utilizado para realizar o particionamento a cada

passo do algoritmo. Esse tipo de algoritmo permite maior flexibilidade em relação à estrutura hierárquica e ao balanceando dos diferentes conglomerados de dados. A Figura 3.6 apresenta uma simples representação dos algoritmos hierárquicos e a orientação de cada uma das abordagens.

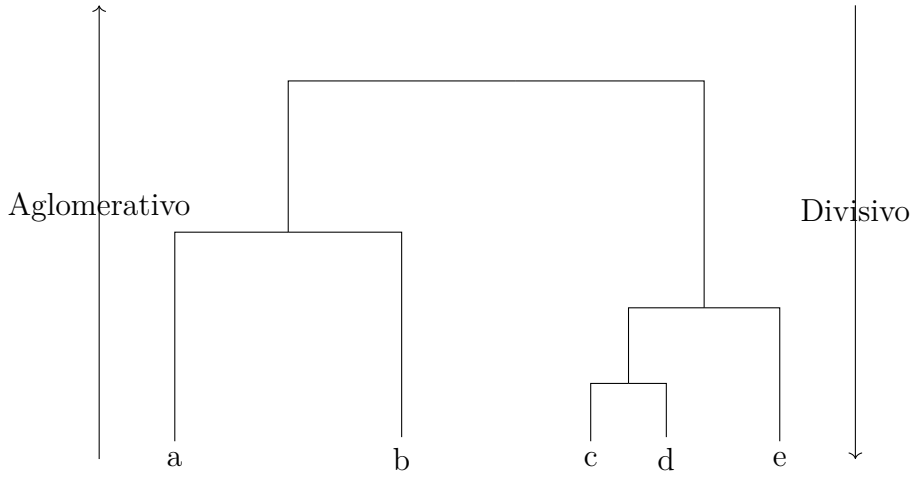


Figura 3.6: Exemplo de dendrograma usado para a clusterização hierárquica.

3.4 NMF como Método de Clusterização

Como explicado anteriormente, o algoritmo de K-médias é um dos mais populares usados para clusterização de dados. Se $X = (x_1, \ldots, x_n)$ corresponder à n pontos, é possível particionar eles em K agrupamentos mutualmente exclusivos. A função objetivo da clusterização por K-médias pode então ser escrita como:

$$\min J_{\text{kmeans}} = \sum_{i=1}^{n} \min ||x_i - x_k||^2 = \sum_{k=1}^{K} \sum_{i} ||x_i - c_k||^2 \tag{3.10}$$

Se considerarmos $C = c_1, \ldots, c_n$ um conjunto de centroides obtidos via clusterização com K-médias. Seja H uma matriz que indica a qual cluster pertence, ou seja, $h_{ki} = 1$ se x_i ao agrupamento c_k e $h_{ki} = 0$ se não for o caso [85]. Deste modo, pode-se escrever a função custo de K-médias como:

$$\min J_{\text{kmeans}} = \sum_{k=1}^{K} \sum_{i=1}^{n} h_{ik} ||x_i - c_k||^2 = ||X - CH^T||^2 \tag{3.11}$$

Ao comparar as equações 3.3 e 3.11, pode-se perceber que elas tratam-se da mesma função objetivo, contudo com restrições diferentes [85]. De fato, a função

objetivo do algoritmo K-médias pode ser expressa de diferentes maneiras: se aplicada a restrição da não negatividade a solução será a NMF [85]; se aplicada a restrição da ortogonalidade a solução encontrada serão os componentes principais [86]; e, por fim, sem quaisquer restrição, obtém-se o conjunto de centroides do espaço amostral.

O elemento de H, H_{ij}, denota uma medida de quanto o evento x_j se relaciona ao cluster k_i [87]. Neste ponto, duas classes distintas de clusterização podem ser definidas:

Hard Clustering Nesse tipo de clusterização, entende-se que o evento pertence a apenas um cluster e a determinação do cluster k_i é realizada a partir da Equação 3.12:

$$k_j = \arg\max_i H_{ij} \tag{3.12}$$

onde k_j é o índice do valor máximo de uma linha de H.

Soft Clustering Nesse método é entendido que o valor da matriz de coeficientes H_{ij} representa a probabilidade de um determinado evento i pertencer ao cluster k_j. Ao contrário de algoritmos do tipo *hard* (como, por exemplo, K-médias), nesse caso é admitido que um evento pertença simultaneamente a um ou mais agrupamentos.

Diferentes aplicações de clusterização baseadas em NMF podem ser encontradas na literatura [88–91], sendo sua utilização em mineração de texto, na tarefa de extração de tópicos, a que talvez mais reverbere [92–94].

Extrair tópicos consiste na tarefa de agrupar diferentes documentos textuais de acordo com sua semântica. É possível representar um documento em formato vetorial utilizando, por exemplo, a medida TF-IDF [95], que representa a importância de uma palavra em relação a uma base textual ou *corpus* linguístico. Como resultado da NMF, dois subprodutos são obtidos: o tópico ao qual o documento pertence e o conjuntos de palavras mais relevantes para para caracterização de tal tópico [96].

3.4.1 Clusterização baseada em Consenso

Existem diversas maneiras de inicializar as matrizes C e H, com vantagens e desvantagens, ligadas principalmente a complexidade do problema [58]. A maneira

mais simples é gerá-las de maneira aleatória inúmeras vezes e manter a melhor solução gerada [97].

Em relação a tarefa de clusterização, a utilização de várias inicializações é um indicativo da estabilidade da solução. Ao comparar os diversos resultados obtidos por diferentes inicializações podemos contabilizar o número de vezes que cada evento foi categorizado no mesmo cluster. Se, ao longo das repetições, os eventos não transitem em diferentes agrupamentos, mais estável é o modelo obtido. Esse processo é denominado "Clusterização baseada em Consenso" (do inglês *Consensus Clustering* (CC)) e é baseado no cálculo da matriz de consenso, que possui dimensão $n \times n$, onde n denota o número de eventos e em que cada elemento representa o número de vezes, em média, que dois eventos foram categorizados pelo mesmo cluster.

3.5 Algoritmos de Clusterização Baseados na Natureza

A inteligência de enxames estuda o comportamento coletivo de sistemas compostos por uma comunidade de indivíduos que interage localmente entre si e com seu ambiente. Enxames usam formas de controle descentralizadas e uma auto-organização para alcançar seus objetivos. Observando o sucesso e a eficiência dos enxames na natureza para resolver problemas complexos, foram desenvolvidos sistemas computacionais que procuram mimetizar seus comportamentos. Assim surgiram técnicas robustas, escaláveis e paralelizadas surgiram para resolver problemas de otimização complexos. É um campo de pesquisa relativamente recente, com os primeiros trabalhos publicados no início da década de 90, porém cada vez mais popular, principalmente por sua característica multidisciplinar [98].

Pelo seu sucesso em diversos problemas de reconhecimento de padrões, pesquisadores passaram a desenvolver técnicas de inteligência em enxames para particionar dados. Tais técnicas consistentemente apresentam desempenho superior aos algoritmos clássicos quando aplicados em complexas bases de dados reais [98]. Nanda (2014) e Kaur (2015) desenvolvem extensas pesquisas em relação à literatura já desenvolvida sobre o tema [99, 100]. Podem-se destacar alguns métodos derivados das primeiras versões dos algoritmos colônia de formigas ou enxames de partículas.

O primeiro algoritmo inspirado no comportamento de animais foi criado no início da década de 1990 por Marco Dorigo com o intuito de resolver uma série de problemas de otimização [101, 102]. O algoritmo de otimização baseado em colônia de formigas (ACO, do inglês *Ant Colony Optimization*) se inspira nos hábitos de forrageamento (busca e a exploração de recursos alimentares) desses insetos na natureza. A ideia principal é mimetizar a comunicação indireta entre as formigas a partir de trilhas de feromônio[1], as quais as permitem à encontrar o caminho mais curto entre o formigueiro e a fonte de comida. Quanto maior o número de formigas transeuntes em um caminho, maior a quantidade de feromônio depositada e, consequentemente, mais atraente a rota se torna para as demais formigas. Desta maneira, as menores rotas são privilegiadas e o caminho mais curto é encontrado [103].

Em relação à tarefa de clusterização, pode-se encontrar algumas abordagens na literatura [104–108]. Uma ideia simples, é imaginar cada amostra como um ponto de um problema de "caixeiro viajante". Este é um problema de otimização amplamente discutido na engenharia e se trata de uma aplicação direta do ACO [109]. Após encontrado um caminho ótimo, ou seja, aquele no qual o caixeiro percorre o menor percurso, recorta-se a trajetória encontrada em subconjuntos baseado em um critério de dissimilaridade entre as amostras, comumente sendo escolhida a distância euclidiana para tal. Cada subconjuntos denotará um cluster [110]. Como pode-se esperar, esse método não é robusto a presença de ruído e possui desempenho limitado, principalmente para problemas de dimensionalidade elevada [111].

Um outro fenômeno biológico das formigas inspirou um novo algoritmo, focado em agrupar dados. Formigas são capazes de se organizar de maneira que conseguem montar estruturas mecânicas como, por exemplo, correntes feitas pelos indivíduos para ligar folhas ou um amontoado de até 40 formigas formando uma espécie de gota. Tais fenômenos coletivos podem ser observados em diversas espécies. As formigas *Linepithema humiles*, encontrada na Argentina, ou as as formigas africanas do tipo *Oecophylla longinoda* são alguns exemplos [106]. A Figura 3.7 apresenta uma observação de auto-organização das formigas argentinas.

O algoritmo AntTree é inspirado nesse processo de auto-organização das formigas e tem como objetivo agrupar dados a partir de uma estrutura de árvore. Nessa

[1]Substância secretada por um animal e reconhecida por animais da mesma espécie na comunicação e no reconhecimento.

Figura 3.7: Formigas formando estruturas complexas a partir da sua auto-organização. Retirado de [112].

representação, cada amostra representa um nó, o que a difere com o dendrograma tradicional, onde apenas as folhas correspondem aos dados [106].

3.5.1 Meta-heurísticas Baseadas em Enxames

Meta-heurísticas são algoritmos estocásticos que contam primeiramente com uma etapa de inicialização aleatória e posteriormente com uma busca local especialista. O processo de inicializações aleatórias cria uma série de soluções arbitrárias, que exploram todo o espaço de busca. Essa exploração é responsável por encontrar o melhor global. A busca local é responsável por determinar a convergência e alcançar bons resultados em uma região específica [113].

Algoritmos inspirados na natureza são baseados em agentes extremamente simples que, por se comportarem coletivamente, são capazes de produzir respostas complexas. Nesses sistemas, o comportamento de cada indivíduo é baseado em impulsos simples como distância entre os vizinhos mais próximos, ou a velocidade do grupo. Três meta-heurísticas baseadas na natureza tem bastante aceitação junto com a comunidade científica, e possuem diversas aplicações nos últimos anos: Enxames de Partículas (PSO, do inglês *Particle Swarm Intelligence*) [114], Colônia de Abe-

lhas (ABC, do inglês *Artificial Bee Colony*) [115] e Cardume de Peixes (FSS, do inglês *Fish-School Search*) [116].

Otimização por Enxames de Partículas

O PSO é um algoritmo estocástico de otimização baseado no comportamento observado em diversos grupos sociais da natureza, como por exemplo, bandos de aves. O enxame de partículas representa as interações entre um número de indivíduos, que não conhecem o objetivo final, mas sim, seu estado atual, seu melhor estado no passado e o melhor desempenho dentre seus vizinhos [117]. Como os indivíduos buscam alcançar o sucesso de seus vizinhos, a população passa a se acumular em regiões ótimas do espaço-busca, descobrindo, assim, boas soluções para problemas complicados [117].

Ao contrário de algoritmos evolutivos [118], PSO não usa nenhuma etapa de seleção de indivíduos. Deste modo, todos os indivíduos sobrevivem durante todo o processo de otimização. As interações sucessivas entre estes indivíduos ajudam a melhorar as soluções do problema ao longo do tempo [117].

As partículas são vetores de D dimensões, localizados dentro de espaço-busca definido, que representam a solução do problema especificado. Ao conjunto de partículas dá-se o nome de enxame. As partículas se movimentam em busca do melhor posicionamento baseado nas interações com seus vizinhos.

A forma mais comum de implementar esse algoritmo define o comportamento das partículas em duas fórmulas. A primeira ajusta a velocidade (ou tamanho do passo) e a segunda movimenta a partícula adicionando a velocidade à sua posição atual.

$$v_{id}^{(t+1)} \leftarrow \alpha v_{id}^{(t)} + U(0,\beta) \times (p_{id} - x_{id}^{(t)}) + U(0,\beta) \times (p_{gd} - x_{id}^{(t)}) \tag{3.13}$$

$$x_{id}^{(t+1)} \leftarrow x_{id}^{(t)} + v_{id}^{(t+1)} \tag{3.14}$$

onde i é índice da partícula, d é a dimensão, x_i é a posição da partícula, v_i é a velocidade, p_i é a melhor posição achada por i, g é o índice do melhor vizinho de i, α é coeficiente de constrição (ou peso inercial), β é a constante de aceleração e $U(0,\beta)$ é um número aleatório gerado a cada movimento. Normalmente usa-se

$\alpha = 0.7298$ e $\beta = \Phi/2$, onde $\Phi = 2.9922$ [119]. Contudo, estes podem ser ajustados com benefícios para alguns problemas específicos [120].

A avaliação do vetor de soluções representado pela partícula i é realizada pela função $f(x)$. Os resultados são comparados e a melhor posição x_i é armazenada. Enquanto o processo de otimização ocorre, cada partícula circula entorno da região centrada nos melhores locais já alcançados p_i e p_g. Conforme as partículas são atualizadas, a suas trajetórias são desviadas para novas regiões do espaço-busca. Deste modo, as partículas encontram ótimos. O Algoritmo 2 apresenta a operação do PSO padrão.

Algoritmo 2: Operação do PSO Padrão

Dados: N: Número de partículas, S: Enxame, P: Melhores posições

1 Inicializa S e $P = S$;
2 Avalia S e P e define g como a melhor posição;
3 **enquanto** *(critério de parada)* **faça**
4 | Atualiza S usando as equações 3.13 e 3.14;
5 | Avalia S
6 | Atualiza P e redefine g
7 **fim**

Pode-se reescrever a equação 3.13 como $v_i^{(t)} = x_i^{(t)} - x_i^{(t-1)}$, ou seja, a velocidade $v_i(t)$ é diferença entre a próxima posição e a anterior. Deste modo, obtemos uma única equação:

$$x_{id}^{(t+1)} \leftarrow x_{id}^{(t)} + \alpha \times (x_{id}^{(t)} - x_{id}^{(t-1)}) + \sum U(0, \frac{\Phi}{2}) \times (p_{gd} - x_{id}^{(t)}) \tag{3.15}$$

Nessa notação podemos interpretar a atualização da posição de uma partícula como a soma de três parcelas: a atual posição, a persistência da partícula e a influência social [117]. Assim, em cada interação, cada partícula começa na última posição alcançada, persiste na última direção com algum peso e faz ajustes conforme a influência social determina e sua própria posição atual.

Uma outra versão do algoritmo foi proposta por Mendes (2004) [121], com uma mudança em relação à topologia da população do enxame. Enquanto na versão padrão do algoritmo a partícula tem consciência de sua melhor posição e a melhor posição do melhor vizinho (onde todas as partículas enxergam todos os vizinhos, ou seja, uma rede completamente conectada), na nova versão é a influência de um

grupo de vizinhos que determina a direção a ser tomada. Essa versão é denominada Enxame de Partículas Completamente Informada (FIPS, do inglês *Fully Informed Particle Swarm*). A FIPS exclui dois aspectos do algoritmo tradicional. Primeiro, a partícula i não tem influência própria direta. Segundo, o melhor vizinho agora é computado com outros e não é mais necessário calcular o sucesso de todos os vizinhos para achar o melhor.

No contexto de clusterização, uma partícula do PSO representa N vetores de centroides [122], onde N é pré-determinado em relação à execução do algoritmo. Pode-se usar o erro de quantização médio [122], índices de validação interna dos agrupamentos [123] ou medidas de distância (euclidiana, manhattan, cosseno, ...) [124, 125] na formulação do problema, como possíveis funções custos a serem otimizadas.

Frequentemente PSO é usado em conjunto com algum algoritmo de clusterização não supervisionado clássico, como, por exemplo K-médias [122, 123, 126, 127]. O algoritmo de K-médias tende a convergir mais rapidamente do que PSO, mas frequentemente realiza uma clusterização menos acurada [122]. Duas abordagens podem ser encontradas na literatura para a utilização das técnicas em conjunto:

1. Executar o algoritmo de K-médias e usar seu resultado para inicializar uma partícula do PSO. As demais partículas serão inicializadas de maneira aleatória [122].

2. Executar o algoritmo PSO e usar seu resultado para inicializar uma execução do algoritmo de K-médias [127]. A ideia por traz dessa abordagem é aproveitar a capacidade de busca global do algoritmo de Enxames e a rápida convergência do K-médias para encontrar ótimos locais [128], já que PSO possui uma convergência muito lenta próxima da solução global [129].

Rana (2012) apresenta estudo com algumas variações do algoritmo de PSO para a tarefa de agrupar os dados de nove bases diferentes. Entre dados artificiais e reais, é de especial interesse desse trabalho a utilização de 56 amostras de petróleo cru. Seus dados possuem cinco atributos baseados na presença de elementos químicos: ferro, vanádio, berílio, hidrocarbonetos saturados e hidrocarbonetos aromáticos. Seus resultados mostram que o algoritmo de K-médias rapidamente caem em mínimos locais, enquanto as variações de PSO convergem para o mínimo glo-

bal do problema, apenas variando o tempo desta convergência. O uso de K-médias com PSO não significou uma melhora significativa no desempenho do processo de descobrir agrupamentos [129].

Otimização por Colônia de Abelhas

Assim como as formigas, as abelhas dedicam grande parte de suas vidas à busca por fontes de comida. Colônias de abelhas possuem um sistema descentralizado de coleta de alimentos e podem ajustar seus padrões de buscas para aumentar a quantidade de néctar adquirido [130].

Abelhas podem estimar a distância da colmeia até as fontes de comida calculando a quantidade de energia gasta enquanto voam, além da direção e a quantidade de alimentos. Essa informação é compartilhada com as demais a partir de uma dança sincronizada e de trofalaxia[2] [130].

Na colmeia, existe uma área onde abelhas forrageadoras (ou operária), aquelas que buscam fontes de alimento, executam danças sincronizadas para atrair novas abelhas para regiões onde há fontes disponíveis. As abelhas que estão na colmeia decidem seguir a abelha que indica o lugar com maior qualidade de alimentos, ou seja, a região que apresenta a maior quantidade de néctar. Essa informação é transmitida pela mudança da intensidade da dança, quanto mais intensa a dança, maior a qualidade da fonte de alimentos. Já as abelhas que decidem explorar novas regiões sem qualquer orientação são chamadas exploratórias.

A ideia por trás dos algoritmos baseados no comportamento das abelha é imaginar que as abelhas possuem uma possível solução para um problema de otimização em sua memória [130]. Essa possível solução corresponde à localização da fonte de comida e tem um fator de qualidade agregado (valor da função custo).

No algoritmo ABC, a colônia de abelhas artificiais contêm três tipos de abelhas:

- Operária
- Observadora
- Exploradora

[2] A trofalaxia é um processo de alimentação em que um indivíduo transfere para outro o alimento que se encontra dentro do seu próprio tubo digestivo por regurgitação.

A abelha observadora espera as operárias desempenharem a dança para decidirem em qual fonte de alimentos irão visitar. A abelha exploratória é responsável pela busca aleatória. Para tais atividades, neste algoritmo, metade da colmeia é composta por abelhas operárias e a outra metade é observadora. Para cada fonte de alimentos, existe apenas uma uma abelha operária. Em outras palavras, o número de operárias é igual ao número de fontes de alimentos em volta da colmeia. As operárias que representam fontes de alimentos que se extinguiram se tornarão exploratórias [115].

No ABC, cada ciclo consiste em três etapas:

1. Mandar as operárias até fontes de alimentos e medir a quantidade de néctar presentes neles;

2. Observadoras escolher a fonte de comida através das informações disponibilizadas pelas operárias;

3. Recrutar abelhas exploradoras para procurar alimento em possíveis localizações.

Na inicialização, um conjunto de fontes de alimentos são aleatoriamente escolhidas por abelhas e a quantidade de néctar é determinada. Então, essas abelhas voltam à colmeia e dividem a informação com as abelhas esperando na área de dança. Na etapa seguinte, após dividir a informação, cada operária vai até a fonte de alimento visitada por ela e escolhe uma nova fonte baseada em informação visual. Na terceira etapa, uma observadora escolhe uma fonte de alimento com maior quantidade de néctar. Após chegar na área, ela escolherá uma nova fonte também baseada em informação visual. Quando a fonte de néctar é abandonada pelas abelhas, uma nova fonte é escolhida aleatoriamente por uma abelha exploradora. Nesse modelo, em cada ciclo pelo menos uma exploradora é recrutada e o número de operárias e observadoras permanece sempre igual [115].

A posição da fonte de alimentos representa uma possível solução do problema de otimização e a quantidade de néctar corresponde a qualidade (*fitness*) da solução associada [131]. O Algoritmo 3 apresenta as etapas necessárias para a movimentação das abelhas e, consequentemente, solução do problema.

Algoritmo 3: ABC

Dados: N: Número de soluções, *limite*: controle de abandono de fonte.

Inicializa população de abelhas x;
Avalia desempenho da população $f(x)$;
enquanto *(critério de parada)* **faça**
    para $i = 1$ *até* $n/2$ **faça**
        `// Fase operária`
        $k \in 1, 2, \ldots, n$, $j \in 1, 2, \ldots, d$, $r \in [0, 1]$
        $v = x_{ij} + r\dot{(}x_{ij} - x_{kj})$;
        Avaliar soluções v e x_i;
        se $f(v)$ *é melhor que* $f(x)$ **então**
            Seleção gananciosa;
        senão
            $\text{contador}_i = \text{contador}_i + 1$;
        fim
    fim
    para $i = 1$ *até* $n/2$ **faça**
        `// Fase observadora`
        Calcular a probabilidade de seleção: $P(x_k) = \frac{f(x_k)}{\sum_k^n f(x_k)}$;
        Selecionar a abelha com a propriedade P;
        Produzir uma nova solução v a partir da abelha selecionada;
        Avaliar soluções v e x_i;
        se $f(v)$ *é melhor que* $f(x)$ **então**
            Seleção gananciosa;
        senão
            $\text{contador}_i = \text{contador}_i + 1$;
        fim
    fim
    para $i = 1$ *até* n **faça**
        `// Fase exploratória`
        se $count_i > limit$ **então**
            x_i recebe nova posição;
        fim
    fim
    Memorizar melhores soluções
fim

Assim como ocorre com os algoritmos de PSO, cada indivíduo do ABC representa um conjunto de centroides para o problema de clusterização. O algoritmo foi testado frente diversas bases de dados reais e apresentou resultados similares ou superiores aos obtidos por PSO [132, 133].

Busca por Cardume de Peixes

O processo de busca com FSS é realizado por uma população de indivíduos com memória limitada, os peixes [116]. Cada peixe representa uma possível solução. Similarmente ao PSO ou ABC, o processo de busca do FÁS é guiado pelo sucesso de alguns indivíduos da população [116].

A principal característica do FÁS é relacionada na memória inata dos indivíduos em relação ao seu sucesso: o peso dos peixes. Ao contrário dos demais algoritmos, não há necessidade de atributos como a melhor posição encontrada, a velocidade, a direção, etc. Apenas o peso do peixe é o suficiente para indicar a melhor posição.

Esse algoritmo é composto por três operadores, responsáveis pelas principais ações durante a sua operação. Sua execução ocorre dentro do "aquário", metáfora para a região delimitada no espaço-busca, onde os peixes podem ser posicionados. A comida é relacionada a função que será otimizada. Os operadores podem ser definidos da seguinte maneira [116]:

Alimentar

Comida é a metáfora para indicar aos peixes as regiões do aquário que eles provavelmente terão os melhores lugares para o processo de busca.

Para encontrar maiores quantidades de comida, os peixes podem executar movimentos independentes e, como resultado, cada peixe pode ganhar ou perder peso, dependendo do sucesso na sua busca por comida. FSS usa um variação de peso proporcional a diferença normalizada entre a avaliação da função de aptidão das últimas duas posições visitadas, como mostrado na Equação 3.16.

$$W_i^{(t+1)} = W_i^{(t)} + \frac{f(x_i^{(t+1)}) - f(x_i^{(t)})}{\max\{|f(x_i^{(t+1)}) - f(x_i^{(t)})|\}} \tag{3.16}$$

onde $W_i^{(t)}$ é o peso do peixe i, $x_i^{(t)}$ é a posição do peixe i e $f(x_i^{(t)})$ é avaliação da função de aptidão em $x_i^{(t)}$. Importante notar que todos os peixes nascem

com o mesmo peso $W_0 = \frac{W_{\max}}{2}$, onde $W_{\max}$ denota uma constante para o peso máximo do peixe.

Nadar

É na realidade uma coleção de operadores responsável por guiar a busca globalmente pelos subespaços do aquário que são coletivamente sentidas por todos os peixeis como as mais promissoras com relação ao processo de busca:

- **Movimento individual**: um peixe nada para uma posição onde a quantidade de comida parece superior em relação a posição atual;
- **Movimento coletivo instintivo**: calcula-se o deslocamento médio de todos os peixes bem-sucedidos para propagar em relação a todos os elementos do cardume;
- **Movimento coletivo volitivo**: último ajuste de posicionamento é realizado baseado no peso médio do cardume. Caso o peso do cardume diminuiu em relação à última etapa, o raio do cardume deveria diminuir, e vice-versa.

Reproduzir

O último operador, que é responsável por refinar a busca realizada. Foi criada para permitir a transição automática entre a exploração e explotação. Quando dois peixes atingem um peso acima de um patamar definido, ou seja, um indicativo que o processo de busca está bem sucedido é alcançado por ambos, estão propensos a procriar. Apenas um herdeiro é criado por cada casal de peixes. O peso do novo peixe será a média entre os dos seus pais e o seu posicionamento será o ponto médio entre os dois. Para manter o tamanho do cardume constante, os menores peixes são eliminados, simulando um processo de seleção natural.

Um ponto fraco do algoritmo FSS é fato de seu desempenho depender nos passos usados pelos movimentos individuais e coletivos [134]. Se o valor desses passos forem altos, o algoritmo apresenta uma pequena capacidade em executar a explotação, devido a pequena granularidade usada na busca. Caso contrário, passos pequenos

deixam a convergência lenta demais, prejudicando o desempenho computacional do algoritmo quando comparado com algoritmos mais simples como o PSO [134].

Para mitigar essa dependência, uma versão otimizada do algoritmo foi proposta por Filho (2013) [134], sendo denominada FSS-II. Nessa versão, não há necessidade de determinar um valor para o passo dos movimentos individuais e o movimento coletivo volitivo. Além do mais, no FSS-II os operadores são combinados em apenas uma equação e a função de aptidão é apenas avaliada uma vez por iteração, por peixe [134]. O Algoritmo 4 apresenta o pseudocódigo dessa versão.

Algoritmo 4: FSS-II

Dados: N: Número de soluções
Inicializa população de peixes x e pesos W;
Aplica uma busca local em cada peixe para definir $x_i^{(1)}$;
Avalia desempenho da população $f(x)$ e determina o peso $W_i^{(1)}$;
enquanto *(critério de parada)* **faça**
    para *cada peixe* **faça**
        `// Operador Alimentação`
        Calcular o deslocamento do peixe;
        Calcular a diferença na avaliação de aptidão;
        Alimentar o peixe usando: $W_i^{(t+1)} = W_i^{(t)} + \Delta f_i(t+1)$;
    fim
    Calcular o baricentro a partir de 3.17;
    para *cada peixe* **faça**
        `// Operador Natação`
        Movimentar de acordo com 3.18
    fim
    Executar operador Reprodução;
fim

Na principal parte do código, primeiro calcula-se o deslocamento individual, que corresponde ao movimento individual na versão original. No FSS-II, esse deslocamento depende da variação da posição em $t-1$ e t. Posteriormente, a variação da aptidão é avaliada e usada para alimentar o peixe. A partir da variação de peso, podemos calcular o baricentro do cardume, como mostrado na Equação 3.17.

$$B(t+1) = \frac{\sum_{i=1}^{N} x_i^{(t)} W_i^{(t+1)}}{\sum_{i=1}^{N} W_i^{(t+1)}} \tag{3.17}$$

Por fim, as posições dos peixes podem ser atualizadas de acordo com a Equação 3.18. Essa atualização é composta por quatro parcelas: a posição atual; o mo-

vimento individual baseado no deslocamento realizado anteriormente; o movimento coletivo instintivo, ponderado por $W_i^{(t+1)}$; e o movimento coletivo volitivo.

$$
\begin{aligned}
x_i^{(t+1)} = x_i^{(t)} &+ \beta \cdot c \cdot \Delta x_i^{(t+1)} \\
&+ c \cdot \text{rand}(0,1) \cdot \frac{\sum_{i=1}^{N} \Delta x_i^{(t+1)} W_i^{(t+1)}}{\sum_{i=1}^{N} W_i^{(t+1)}} \\
&+ c \cdot \text{rand}(0,1) \cdot \text{sign}(\sum_{i=1}^{N} \Delta W_i^{(t+1)}) \cdot (x_i^{(t)} - B(t+1))
\end{aligned} \tag{3.18}
$$

onde, sign$(\cdot)$ retorna o sinal do atributo, definindo a direção do deslocamento em relação ao baricentro. O c é usado para controlar a amplitude do movimento, e β para controlar a contribuição individual de cada peixe durante a movimentação.

Assim como nas demais meta-heurísticas, cada indivíduo do FSS representará uma solução para o problema de dividir N dados em K agrupamentos. Serapião (2016) propõe a utilização de FSS combinado K-Harmónicas (KHM) [135] e K-médias. Para tal, utiliza como função de aptidão, uma medida de dissimilaridade entre os clusters, estabelecendo assim um paralelo entre os algoritmos convencionais e o baseado nos cardumes de peixe. O artigo propõe a mesma inicialização utilizada por van der Merwe [122], onde uma partícula (nesse caso peixe) será inicializada após a execução do algoritmo de K-médias. Os resultados com FSS para o problema de descobrir agrupamentos foram comparados com o K-médias padrão e o PSO. Como função de aptidão foi usado o erro quadrático médio. Novamente, uma série de bases de dados reais foram utilizadas para testar o algoritmo proposto pelo trabalho. A base de petróleos com 56 amostras e 5 dimensões foi uma das utilizadas apresentando resultados idênticos para PSO e FSS. Ambos bem superiores do que algoritmo de K-médias padrão [136].

Como pode-se observar a partir das diferentes proposições de meta-heurísticas aplicadas ao problema de descobrimento de agrupamentos para uma base de dados, existe uma estrutura comum. Primeiramente, uma população de indivíduos é posicionada em posições aleatórias dentro do espaço-busca, onde, cada indivíduo representa um potencial conjunto de centroides. Então, cada amostra do problema é alocada em uma partição (definida por um centroide) determinada pela menor dis-

tância (nesse ponto uma medida de distância deve ser escolhida, como, por exemplo, a distância euclidiana ou de cossenos). Posteriormente, é avaliada a aptidão da solução (indivíduo) a partir da função objetivo (ex: Silhueta, índice de Davies-Bouldin, Erro Médio Quadrático, etc.). O resultado dessa avaliação vai servir como figura de mérito para o algoritmo em questão. No caso do PSO, escolhe-se a partícula mais apta; no ABC, a localidade com mais néctar; e o FSS o peso de cada peixe. Desta maneira, os indivíduos podem se realizar seu próximo movimento e novos centroides serão calculados, num processo cíclico até que as condições de parada sejam atingidas. A Figura 3.8 sintetiza as etapas genéricas para os diferentes meta-heurísticas pesquisadas.

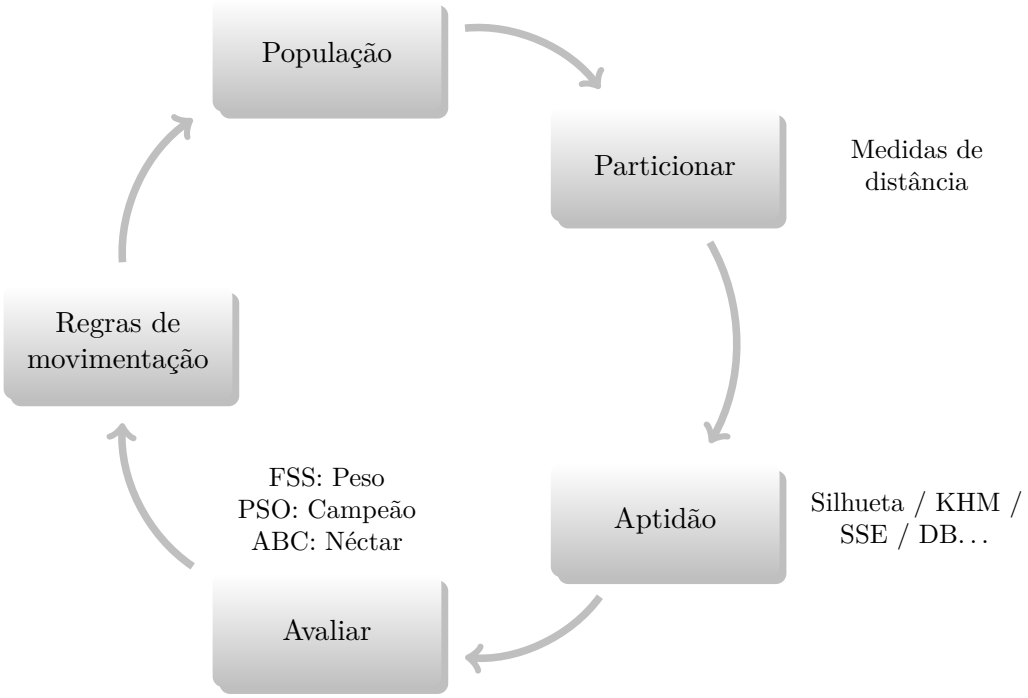


Figura 3.8: Execução de um algoritmo de Inteligência de Enxames integrado com índices de clusterização.

3.6 Índices de Validação de Clusterização

Após o processo de clusterização dos dados, necessita-se de um meio para avaliar seu desempenho a fim de avaliar e criar uma base de comparação entre os diferentes resultados obtidos. Validar agrupamentos, através da avaliação da qualidade dos resultados de uma clusterização, é reconhecido como uma das etapas fundamentais para o sucesso de aplicações com essa proposta [39].

As duas principais categorias de validação de agrupamento são a validação de clusterização interna e a validação de clusterização externa. A principal diferença entre elas é se há informação externa para ser usada no processo de validação, como, por exemplo, classes rotuladas [137]. Assim, problemas que utilizam a validação externa já sabem *a priori* o número correto de conglomerados. Por outro lado, quando apenas a validação interna for possível, por não ter nenhuma informação além das próprias amostras, pode ser usada para, além de qualificar o resultado entre diferente algoritmos, determinar o número ótimo de agrupamentos [39]. Vale lembrar que, como regra geral, não há disponível quaisquer informações sobre os rótulos das classes. Assim, a validação interna é frequentemente a única opção.

Diversos índices de validação de clusterização internas foram propostas na literatura, como, por exemplo, índice de Silhueta [138], índice de Davies-Bouldin [139] e índice de desvio padrão [140]. Xiong et al. (2013) faz um extenso estudo sobre a monotonicidade e alguns diferentes aspectos (ruído, densidade, formato dos dados, etc.) de cada uma das funções presentes na Tabela 3.1, onde estão alguns dos principais índices [39].

Todos os índices apresentados apresentam particularidades e podem ser utilizados em diferentes cenários. Em especial, para esse projeto foi optada a utilização do índice de Silhueta. Apesar de demandar maior esforço computacional, este índice permite a compreensão da situação de cada uma das amostras utilizadas no processo de agrupamento. Em trabalhos dedicados à comparação dos diferentes métodos, o índice de Silhueta possui os melhores resultados [140, 141].

3.6.1 Índice de Silhueta

O índice de Silhueta [138] é definido como:

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}} \tag{3.19}$$

onde $a(i)$ é a dissimilaridade média entre a a i-ésima amostra e todas as outras amostras do seu mesmo agrupamento e $b(i)$ é a menor dissimilaridade média entre a i-ésima amostra e a as amostras dos outros agrupamentos. Pode-se reescrever a definição de silhueta como:

$$s(i) = \begin{cases} 1 - a(i)b(i), & \text{se } a(i) < b(i) \\ 0, & \text{se } a(i) = b(i) \\ a(i)/b(i) - 1, & \text{se } a(i) > b(i) \end{cases} \tag{3.20}$$

Pode-se perceber que $-1 < s(i) < 1$. Quanto mais a i-ésima amostra estiver bem condicionada no seu agrupamento (menor dissimilaridade média $a(i)$), mas $a(i) < b(i)$ e, logo, mais o índice de silhueta se aproxima de 1. Da mesma forma, quanto pior condicionada no seu agrupamento, mais o índice de silhueta se aproxima de -1. Um índice de silhueta nulo indica que a amostra i está na fronteira entre duas ou mais classes. Note que o índice de silhueta é definido para cada amostra de dados. Para a configuração total dos agrupamentos, utiliza-se o índice médio de silhueta, considerando todas as amostras de dados, o que resulta na equação observada na Tabela 3.1. Novamente, quanto mais próximo da unidade, melhor é a configuração.

3.6.2 Índice de Validação de Clusterização Baseado nos Vizinhos mais Próximos

No entanto, o índice de Silhueta compartilha as mesmas desvantagens das demais métricas de validação: a sensibilidade à presença de dados ruidosos e a necessidade de agrupamentos com formato esférico [39]. A presença de ruído entre as amostras impacta significadamente os resultados obtidos, principalmente quando buscam-se mínimos e máximos [39]. Já quando o agrupamento não possui formato esférico, o desempenho dos índices se torna imprevisível, pois são baseados em medidas de distância.

Xiong propõe uma nova métrica capaz avaliar agrupamentos em diferentes formatos e que seja mais robusta à presença de dados ruidosos. Esse índice de validação externo é baseado no algoritmo de vizinhos mais próximos [142] e na complementação de dois conceitos importantes: a *separação inter-cluster* e *compactação intra-cluster*.

Separação Inter-cluster Baseada em Vizinhos mais Próximos

Na literatura, alguns pesquisadores acreditam que a separação entre diferentes agrupamentos deveria ser mais importante que a compactação interna de cada

Índice	Notação	Definição	Valor ótimo
Valor Eficaz	RMSSTD	$\frac{\sum_i \sum_{x \in C_i} \|\|x-c_i\|\|^2}{P \sum_i (n_i - 1)}$	Joelho
R Quadrado	R^2	$\frac{\sum_{x \in D} \|\|x-c\|\|^2 - \sum_i \sum_{x \in C_i} \|\|x-c_i\|\|^2}{\sum_{x \in D} \|\|x-c\|\|^2}$	Joelho
Hubert Modificado Γ	Γ	$\frac{2}{n(n-1)} \sum_{x \in D} \sum_{y \in D} d(x,y) d_{x \in C_i, y \in C_j}(c_i, c_j)$	Joelho
Calinski-Harabasz	CH	$\frac{\sum_i n_i d^2(c_i,c)/(NC-1)}{\sum_i \sum_{x \in C_i} d^2(x,c_i)/(n-NC)}$	Máximo
I	I	$\left(\frac{\max_{i,j} d(c_i,c_j) \sum_{x \in D} d(x,c)}{NC \sum_i \sum_{x \in C_i} d(x,c_i)}\right)^p$	Máximo
Dunn's	D	$\min_i \{\min_j (\frac{\min_{x \in C_i, y \in C_j} d(x,y)}{\max_k \{\max_{x,y \in C_k} d(x,y)\}})\}$	Máximo
Silhueta	S	$\frac{1}{NC} \sum_i \{\frac{1}{ni} \sum_{x \in C_i} \frac{b(x)-a(x)}{\max(b(x),a(x))}\}$	Máximo
Davies-Bouldin	DB	$\frac{1}{NC} \sum_i \max_{j, j \neq i} \{\frac{n_j \sum_{x \in C_i} d(x,c_i) + n_i \sum_{x \in C_k} d(x,c_j)}{n_i n_j d(c_i,c_j)}\}$	Mínimo
Xie-Beni	XB	$\frac{\sum_i \sum_{x \in C_i} d^2(x,c_i)}{n \min_{i,j \neq i} d^2(c_i,c_j)}$	Mínimo

D: Conjunto de dados; n: número de amostras; c: centro de D; P: número de atributos em D; NC: número de clusters; C_i: agrupamento i; n_i: número de amostras em C_i, c_i: centro de C_i; $d(x,y)$: distância entre x e y

Tabela 3.1: Índices para Validação Interna de Agrupamentos. Adaptado de [39]

.

conjunto de dados quando se tratando da avaliação de desempenho [39]. Normalmente, essa tarefa é realizada ao medir a distância entre pontos específicos que caracterizam o agrupamento, como por exemplo, o centroide. Um único representante não consegue carregar toda a informação geométrica do cluster, e por essa razão as métricas mais comuns dependem que os agrupamentos tenham uma formação esférica.

Xiong propõe uma abordagem onde os elementos escolhidos para a validação sejam aqueles que carregam a informação geométrica do cluster. Ou seja, múltiplos objetos são selecionados de maneira que eles representem de forma efetiva o cluster como um todo e a partir de deles calcula-se um grau de afastamento para os demais clusters.

Para selecionar tais elementos, é importante ter conhecimento do conceito de *consistência dos vizinhos próximos em um agrupamento*. Para qualquer amostra dentro de um agrupamento, seus vizinhos mais próximo devem pertencer ao mesmo agrupamento [86]. Seguindo essa lógica, se um objeto está no centro de um cluster e está cercado por elementos do mesmo cluster, ele não carrega informação relevante sobre a separação entre agrupamentos. Já um elemento no limite entre dois agrupamentos pode ser vizinho por diversas amostras de ambos conjuntos. A proporção de elementos vizinhos pertencentes ao próprio agrupamento e a outros é determinante do quão relevante essa amostra é em relação à separação dos agrupamentos em questão. Nesta linha de raciocínio, a seguinte formulação para a separação inter-cluster (Sep) é usada:

$$Sep = \max_{i=1,2,\ldots,NC}\left(\frac{\sum_{j=1,2,\ldots,n_i}(q_j)}{kn_i}\right) \tag{3.21}$$

onde, NC é o número de clusters, k é o número de vizinhos mais próximos, n_i é o número de amostras que estão no agrupamento C_i e q_j é o número de vizinhos que não pertencem ao cluster C_i. Nota-se que quanto menor o valor de Sep, melhor a separação inter-cluster.

Compactação Interna dos Agrupamentos

A compactação interna de um cluster é uma parte fundamental do processo de validação. Para a métrica proposta por Xiong et al., a seguinte medida é

usada:

$$Com = \sum_{i}[\frac{2\sum_{x,y\in C_i} d(x,y)}{(n_i(n_i-1))}] \tag{3.22}$$

onde, NC é o número de clusters, n_i é o número de amostras que estão no agrupamento C_i, x e y são amostras de C_i. Como a Equação 3.22 mostra, a compactação adotada por Xiong et al. consiste em computar a média de todas as distâncias ponto a ponto entre as amostras pertencentes ao cluster C_i (onde $i = 1, \ldots, CN$) e somar os resultados calculados. Quanto mais essa soma tende a 0, maior a compactação dos agrupamentos obtidos.

Definidos como calcular a separação entre os agrupamentos e a compactação interna dos agrupamentos em questão, resta determinar um índice único capaz de carregar ambas as informações:

$$CVNN = Sep_{\text{norm}} + Com_{\text{norm}} \tag{3.23}$$

onde Sep_{norm} refere-se à medida de separação intercluster normalizada para diferentes números de cluster, ou seja, varia-se o número de agrupamentos e calcula-se o Sep para cada configuração. O maior valor obtido será o coeficiente de normalização. O mesmo vale para Com_{norm}. É simples observar que agrupamentos bem definidos são designados pelo índice $CVNN$ tendendo à zero [39].

Os resultados apresentados por Xiong (2014) são bastante animadores em relação a acurácia do índice em apresentar corretamente o número ótimo de agrupamentos para problemas reais, tendo desempenho bastante superior aos demais índices [39].

Contudo, existem alguns pontos que podem ser levantados em relação a composição do índice $CVNN$:

- o valor de Com_{norm} tende a crescer proporcionalmente conforme o número de agrupamentos aumenta, exceto no caso onde cada amostra denota um agrupamento distinto. Essa relação fica clara quando observamos que o número de parcelas aumenta em razão ao número de agrupamentos. É fácil perceber que essa relação ocorre devido ao aumento de parcelas na computação de Com. A Figura 3.9 demostra o crescimento de Com, conforme aumenta-se o número de agrupamentos.

- ao somar as duas componentes (Sep e Com), perde-se a noção de quanto cada uma contribui para o índice final. Por exemplo, se um conjunto de dados A está separado de tal maneira que todos os agrupamentos são muito compactos, porém a separação inter-clusters é elevada; e outro conjunto B possui valores medianos para ambos subíndices, pode-se obter valores semelhantes para CVNN. Contudo, a interpretação de cada um das configurações deveria ser diferente.

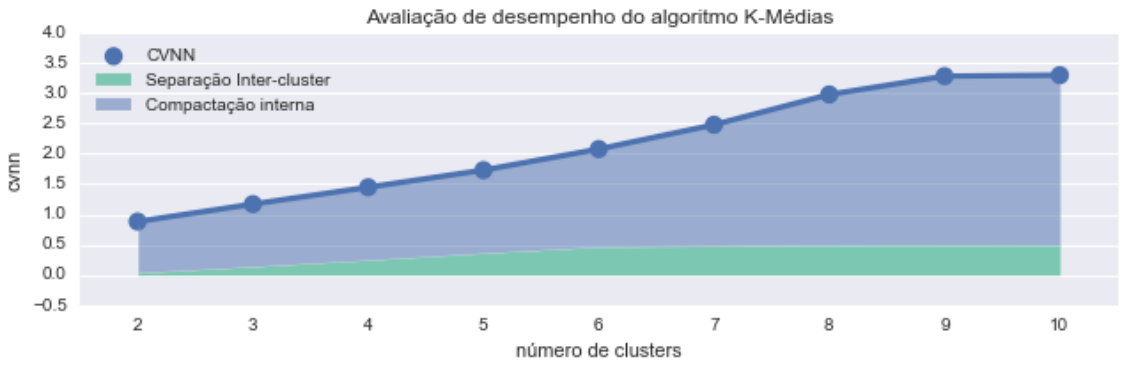


Figura 3.9: Exemplo de composição do índice $CVNN$. Observa-se que o a compactação interna cresce conforme o número de agrupamentos cresce.

Duas adaptações são então sugeridas em relação ao índice original para contornar as limitações citadas:

1. Normalizar Com em relação ao número de agrupamentos.

2. Considerar a composição dos subíndices como um ponto em plano bidimensional, onde a configuração de agrupamentos "perfeitos" é denotada pelo ponto $(0, 0)$. Deste modo, um novo índice é formado como medida de distância entre o ponto $(0, 0)$ e (Sep, Com). Quanto menor esse valor, melhor a clusterização.

A Figura 3.10 apresenta o mesmo exemplo apresentado anteriormente, porém com uma regra de composição diferente. Nela, a configuração com 3 ou 4 agrupamentos seriam as melhores opções. Assim, o novo índice pode ser definido como:

$$CVNN_{sq} = \sqrt{Sep^2 + \frac{Com^2}{NC}} \tag{3.24}$$

onde NC denota o número de agrupamentos.

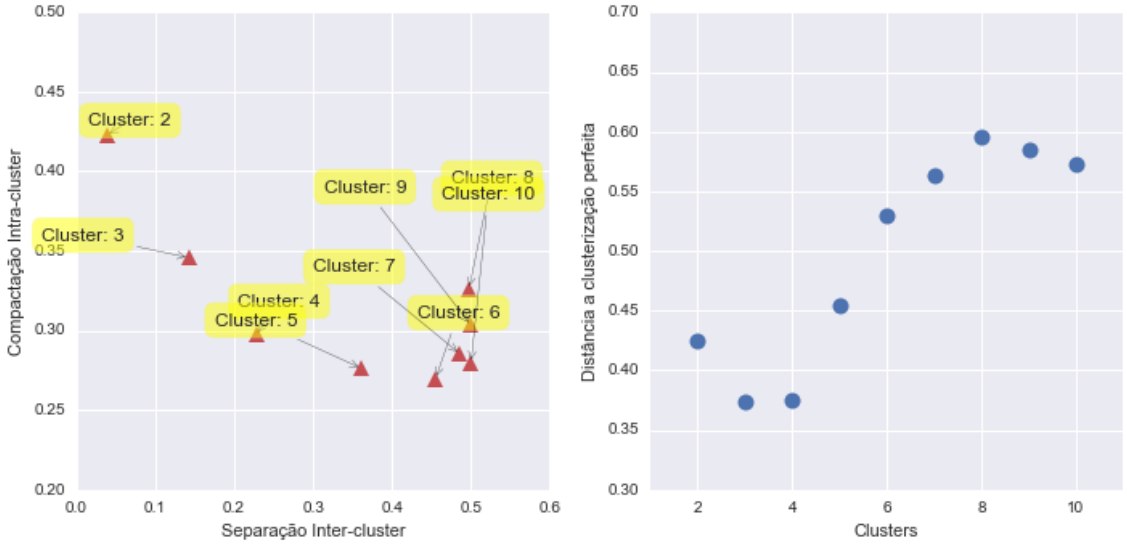


Figura 3.10: Para o exemplo mostrado anteriormente, agora cria-se um plano onde a configuração desejada está o mais próximo possível do ponto $(0, 0)$. No caso, as configurações com 3 e 4 agrupamentos são as melhores opções.

Capítulo 4

Método

Os ensaios de óleo bruto usados nessa pesquisa são formados por 19 propriedades físico-químicas (ver Tabela 4.1). Essas propriedades consistem em medidas como, por exemplo, densidade, viscosidades, presença de metais, enxofre, todas comumente mensuradas em estratégias de ensaios simplificados de petróleo brutos. A base de dados consiste em 49 amostras.

Antes de criar qualquer modelo, os dados foram normalizados, com intuito de evitar a dominância de qualquer propriedade sobre as outras devido a diferenças em suas faixas dinâmicas. Por essa mesma razão, a função $\log_{10}$ foi aplicada sobre as viscosidades já que é uma variável que apresenta distribuição de calda longa, com valores que variam até a a ordem de 10^6. Em seguida, todas as variáveis foram subtraídas dos valores de média e divididas pelo seu desvio médio padrão, para que suas distribuições passem a ser centradas em zero e com variância unitária.

Com os dados normalizados, foi seguida o método de trabalho adotado é esquematizado na Figura 4.1.

A elaboração do modelo de classificação contou com duas etapas:

Exploração dos Dados Visa a compreensão de como as propriedades se relacionam e a influência que exercem na caracterização das amostras de petróleo. Para tal, lançou-se mão de duas abordagens distintas e independentes: (i) projeção dos dados em um mapa de Kohonen para posterior avaliação visual através dos planos de componentes [143] e (ii) a análise da matriz F da Equação 3.2, chamada de *matriz mistura*. Nessa análise pode-se mensurar a importância de cada uma das variáveis para cada componente da NMF [144, 145].

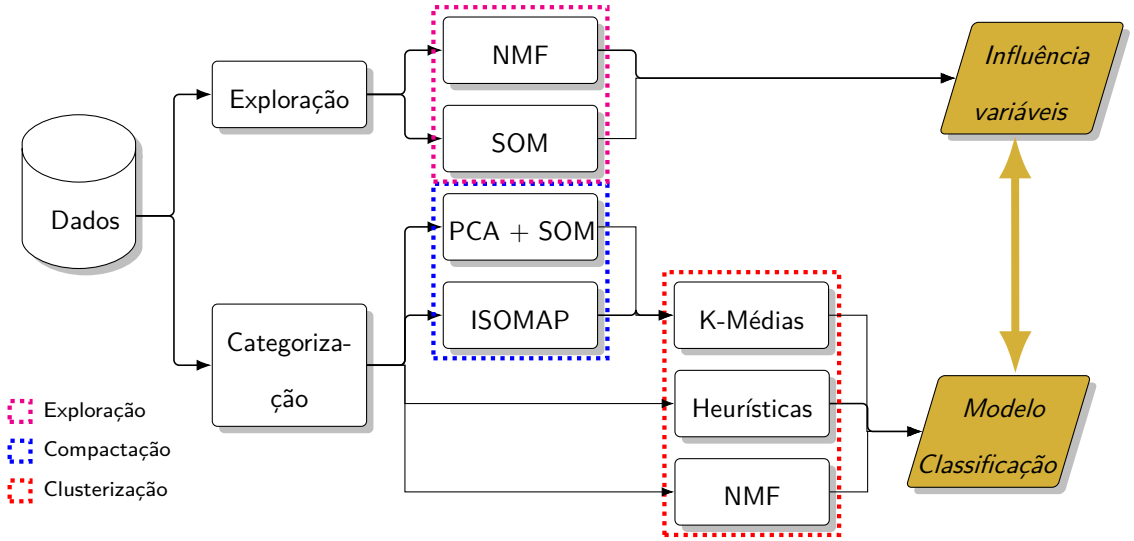


Figura 4.1: Esquema do método utilizado pesquisa.

Modelos de classificação Após a compreensão e caracterização das propriedades, foram criados os modelos de classificação baseados em algoritmos de clusterização. Como explicado no Capítulo 3, cada uma das técnicas utilizadas explora um aspecto diferente em relação a tarefa. O número de clusters foi considerado igual ao número de classes baseadas na classificação por densidade representadas na base de dados (contabilizando três clusters, visto que apenas as classes Leve, Médio e Pesados estão representadas).

O algoritmo de K-médias foi combinado com dois preprocessamentos diferentes. Primeiro, compactação de dados através de PCA com um mapa SOM. Compactar os dados é uma abordagem indicada, visto que os dados são compostos por relativamente alta dimensionalidade. Também foi utilizado ISOMAP, como preprocessamento, com o intuito de acessar informação de ordem superior (o que não é possível com PCA, devido sua natureza linear).

Com o intuito de obter agrupamentos melhor definidos, ou seja, com melhor desempenho em relação a figura de mérito adotado, foram utilizados algoritmos de otimização baseados na natureza, fazendo uma função objetivo fundamentada no índice de silhueta. Foram escolhidos três algoritmos amplamente utilizados na bibliografia, com aplicações para clusterização.

Os métodos de clusterização utilizados até aqui ocorrem no espaço euclidiano. É interessante utilizar outras distâncias e divergências, de modo a explorar no-

vas interpretações do espaço característico que possam colaborar para a tarefa de clusterização. O algoritmo NMF permite que sejam utilizadas diferentes divergências e como mostrado no capitulo anterior, é bastante utilizado para particionar conjuntos de dados. É válido citar novamente que a NMF agrupa os dados ao mesmo tempo que mostra a relevância das propriedades em relação aos clusters.

Cada uma das etapas contou com reuniões com os especialistas da Petrobras para validação das hipóteses decorrentes dos resultados obtidos, principalmente no que se refere às propriedades físico-químicas e às características das amostras estudadas. Os modelos foram construídos iterativamente, onde as observações realizadas em relação a etapa de exploração de dados, serviram de insumo para a interpretação dos agrupamentos obtidos, e vice-versa. As seções a seguir descrevem as particularidades aplicadas a cada um dos algoritmos.

4.1 SOM

Como dito anteriormente, a primeira abordagem para explorar os dados foi projetar sobre um mapa de Kohonen os dados normalizados. A granularidade do mapa é ajustada de forma a ser o maior possível, mas minimizando a quantidade de neurônios não ativados [148] e o formato do mapa respeita a proporção entre as variâncias do primeiro e do segundo componente principal, conforme mostra Equação 4.1 [143].

$$\begin{aligned} n_x n_y &= 5\sqrt{n_d} \\ \frac{n_y}{n_x} &= \sqrt{0.75}\frac{\gamma_1}{\gamma_2} \end{aligned} \tag{4.1}$$

Onde n_x é o número de colunas e n_y o de linhas na grade, n_d o número de amostras disponíveis para treinamento e γ_1 e γ_2 os dois maiores autovalores da análise PCA para esse conjunto de dados.

Dois parâmetros são importantes para o treinamento do mapa: o coeficiente de aprendizagem que decresce uniformemente de $0,05$ até $0,01$ em 1000 épocas; e o raio de vizinhança escolhido de maneira que $\frac{2}{3}$ de todas as distâncias são cobertas no início do processo e diminui de forma constante até apenas o neurônio vencedor seja atualizado. A grade usada foi a hexagonal e o formato do mapa toroidal.

A inspeção visual do mapa foi realizada através de dois tipos de gráficos:

Matriz de distâncias (ou *D-Matrix*) a representação gráfica onde cada neurônio é representado por um hexágono (quando a grade hexagonal é usada) e eles são posicionados lado a lado, respeitando o formato do mapa ($n_x \times n_y$). Um

Nome	Descrição
Densidade	A massa (ou peso) de uma unidade de volume de qualquer substância a uma determinada temperatura. Mede o quão pesado ou quão leve um petróleo líquidoé, quando comparado com a água
Asfaltenos	A quantidade de fração de asfalteno insolúvel em heptano determinada pelo método de fracionamento SARA. Proporções muito maiores em óleos pesados, do que os óleos médios ou de API leve [146]
Enxofre	Total de enxofre na amostra
NAT	Número total de ácido
Nitrogênio	Quantidade de nitrogênio básico e normal
Níquel	Quantidade de níquel. Metais normalmente não são usados para a classificação do óleo bruto. Mas a medida de dessas substâncias podem ser relevantes para esse propósito [147]
Vanádio	Quantidade de vanádio
Carbono	Quantidade de resíduo de carbono
Ponto de escoamento	A menor temperatura na qual o petróleo irá escorrer ou fluir quando é esfriadosem nenhum distúrbio nas condições especificadas no método do teste
Viscosidade	Uma medida da habilidade de um líquido em fluir ou a medida de sua resistênciaem fluir em temperaturas que variam entre 10°C à 50°C (nesse trabalho denotado como V@10°C, V@20°C, V@30°C, V@40°C e V@50°C)
Ponto de ebulição verdadeiro	O ponto de ebulição verdade se apresenta como percentagem por peso quando calculado à condições normais detemperatura e pressão. É medido em 10 wt%, 30 wt%, 50 wt%, 70 wt% e 90 wt%

Tabela 4.1: O ensaio simplificado do óleo cru. Descrição das variáveis usadas neste trabalho.

mapa de calor é projetado na visualização, representando a distância média entre os neurônios. Se um determinado neurônio está longe de sua vizinhança, significando que o mapa teve que se esticar para representar corretamente as amostras, então existe uma região de esparsidade de dados, o que pode representar espaço entre diferentes agrupamentos de dados.

Plano de componentes cada linha do *code-book* (representando uma variável do espaço dos dados) é projetada na representação do mapa. As regiões mais influenciadas pela variável apresentarão valores mais altos (destacados por um mapa de calor).

4.2 ISOMAP

Em relação a compactação de dados promovida através de ISOMAP, a Figura 4.2 mostra o erro de reconstrução em função do número de dimensões intrínsecas. Pode ser visto que cinco dimensões são suficientes para representar os dados originais já que o erro de reconstrução não apresenta redução drástica usando mais componentes. Para estimar a topologias (do inglês, *manifolds*) intrínsecas dos dados, o grafo de vizinhança foi construído com cinco vizinhos, o corresponde à 10 % do conjunto de dados. É importante notar que este algoritmo se torna menos eficiente a medida que o tamanho da vizinhança da topologia aumenta [149].

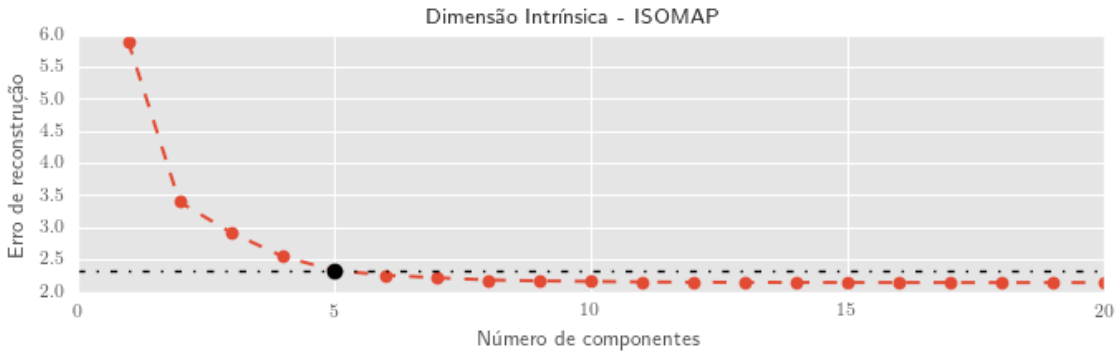


Figura 4.2: Erro de reconstrução para os dados de entrada em função do número de dimensões intrínsecas do ISOMAP. A linha tracejada vermelha representa o erro. A assíntota pontilhada provê um ponto de referência para comparação visual em relação ao posicionamento do "joelho" da função.

4.3 K-médias

Como dito anteriormente, foram assumidos que a análise teria três agrupamentos devido a correspondência com a classificação baseada em densidade. Para certificarmos que três agrupamentos é uma boa solução para o problema, foram testadas configurações com 2 a 10 agrupamentos. Cada configuração foi repetida 50 vezes, com inicialização dos vetores solução feita a partir do algoritmo *kmeans++* [150].

Assim como todos os métodos de clusterização utilizados nesse trabalho, a principal figura de mérito considerada foi o índice de silhueta. Será mostrado também o índice CVNN.

4.4 Algoritmos baseados na natureza

Três classes de algoritmos baseados na natureza foram utilizados nessa tese: PSO, ABC e FSS-II. O espaço busca em todos os algoritmos foi definido como um conjunto de três vetores centroides, inicializados aleatoriamente. A função objetivo $J(X)$ foi definida como:

$$J(x) = 1 - \text{sillhouette index} + 10P \tag{4.2}$$

onde P é um inteiro não-negativo representando o número de clusters vazios encontrados. Como o índice de silhueta varia de -1 a 1, a função objetivo está limitada ao intervalo $[0, 30]$.

Os algoritmos foram implementados dentro do *framework pygmo*, desenvolvido e disponibilizado pela Agência Espacial Europeia [151].

4.4.1 PSO

Quatro variações do algoritmo de PSO descritas no artigo original [114]a e duas utilizadas no trabalho de Blackwell (2007) [152] foram utilizadas:

1. PSO canônico com peso inercial ($\omega = 0,7298$) e componentes cognitivo e social iguais ($\eta_1 = \eta_2 = 2,05$).

2. PSO canônico com peso inercial e componentes cognitivos e sociais iguais e aleatórios, dentro da faixa de números reais entre 0 e 4.

3. PSO com valores de peso inercial, componentes cognitivos e sociais iguais e aleatórios.

4. PSO com os mesmos valores para todos os componentes.

5. PSO canônico com um componente de constrição ($\omega = 0,7298$), para controlar a convergência das propriedades das partículas [153].

6. PSO plenamente informada (FIPS). São utilizadas todas as informações da vizinhança para fazer a atualização dos parâmetros. Utiliza peso inercial ($\omega = 0,7298$) com componentes cognitivos e sociais iguais ($\eta_1 = \eta_2 = 2,05$).

Em todas as configurações foram usadas 100 partículas, com 100 iterações.

4.4.2 ABC

Foram utilizadas 50 abelhas observadoras e 50 abelhas operárias, para 100 iterações. Também foi estabelecido em vinte tentativas o limite de vezes que uma fonte de alimentação é descartada.

4.4.3 FSS-II

Para o algoritmo de cardume de peixes, foram usados 100 indivíduos. Os parâmetros configurados foram: o peso inicial escolhido aleatoriamente dentro de um intervalo entre 300 e 600, $\alpha = 0,1$ e $\beta = 0,1$.

4.5 NMF

Como visto no Capítulo 3, a NMF é bastante utilizada para a tarefa de clusterização. Foi mostrado também, que no seu modo padrão, ela pode ser apresentada com a mesma formulação do algoritmo de K-médias, apenas com a restrição de valores não negativos. Nesse projeto utilizou-se a função objetivo mostrada na Equação 4.3.

$$0.5 * ||X - FG^T||_\beta \tag{4.3}$$

onde $||A||_\beta$ denota uma medida de divergência baseada na distribuição β. Com diferentes divergências, espera-se que determinadas características dos dados sejam realçadas, dependendo da interpretação da função de perda. Para tal, experimentou-se diferentes valores de β (onde $\beta \in$ [0,0; 0,1; 0,5; 1,0; 1,5; 2,0]). Optou-se pela implementação de atualizações multiplicativas como algoritmo de otimização. Foram usadas um número máximo de 400 iterações, com ponto de parada configurado para $||A||_\beta < 10^{-4}$. Esse valor de tolerância foi escolhido baixo o suficiente de modo a estressar o modelo e entender o efeito em diferentes métricas. Cada configuração foi inicializada de maneira aleatória, 200 vezes.

O primeiro passo foi compreender a estabilidade dos modelos obtidos com diferentes divergências. O modelo mais estável e com melhor desempenho na tarefa de clusterização será o escolhido para o restante da análise. Algumas métricas foram observadas (i) A convergência do erro de reconstrução (ii) O ponto de parada do algoritmo (iii) A evolução do índice de silhueta para cada iteração e; (iv) a construção e visualização das Matrizes de Consenso.

A construção da Matriz de Consenso ocorre da seguinte maneira [154]:

1. Para cada execução da NMF, é computada uma matriz $N \times N$, onde N é o número de amostras, em que a posição (i, j) é igual à 1 se as amostras i e j pertencem ao mesmo cluster e 0 caso contrário. A matriz resultante é chamada Matriz de Conectividade;

2. Calcula-se a média de todas as matrizes de conectividade. O resultado será uma matriz de dimensão $N \times N$, onde $0,0 \leq m_{ij} \leq 1,0$, onde quanto menor o valor de m_{ij}, menor a possibilidade das amostras i e j terem sido agrupadas em alguma inicialização. Essa é a Matriz de Consenso;

3. As colunas e linhas da matriz são reordenadas de tal maneira a grupar amostras que aparecem mais frequentemente nos mesmos clusters, lado a lado. Ao realizar a reordenação, a matriz assume formato onde a sua diagonal é composta por submatrizes com valores próximos a 1,0 e a parte externa dessas

submatrizes são quase que predominantemente com valores próximos a nulidade. O número de submatrizes presentes na diagonal é o posto da NMF (ou número de clusters). A Figura 4.3 mostra uma matriz de consenso reordenada como um exemplo arbitrário, onde as aproximadamente 250 amostras estão organizadas em 5 clusters.

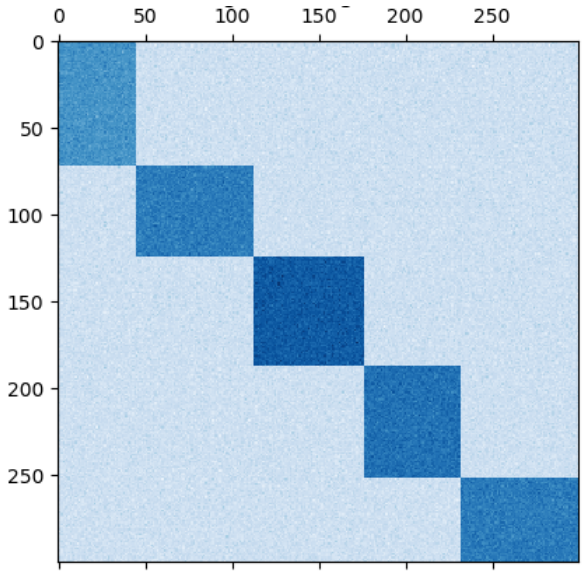


Figura 4.3: Exemplo de matriz de consenso.

Nota-se que uma matriz de consenso extraída a partir de um processo estável possui uma grande concentração de valores $1,0$ (dentro das submatrizes) e $0,0$ (nas regiões externas às submatrizes). Essa noção origina dois índices de estabilidade:

- A média de todos os valores internos às submatrizes tem que ser o mais próximo de $1,0$ e a média dos valores externos a ela tem que ser o mais próximo de $0,0$ possível. Podemos então definir um índice tal que:

$$I = \frac{\sqrt{\text{média}(m_{ij})\text{^}2 + (1 - \text{média}(m_{kl}))\text{^}2}}{2} \tag{4.4}$$

  onde $0 < I < 1$, i e j são índices pertencentes às submatrizes, enquanto k e l não. Quanto mais próximo de $1,0$, mais estável a configuração escolhida.

- Brunet (2004) mostrou que o coeficiente de correlação cofenética [155] representa uma medida de dispersão da matriz de consenso [154].

Nesse ponto, se não for possível definir o melhor valor de β através desse índice, verifica-se para qual distribuição o índice de silhueta teve melhor valor.

Definido o a melhor configuração, realiza-se um *hard clustering* no vetor de componentes obtidos. O resultado da clusterização será avaliado pelas mesmas Figuras de mérito usadas com os métodos anteriores.

A fatoração ainda permite verificarmos, simultaneamente, a influência de cada uma das propriedades através da verificação do vetor mistura.

Capítulo 5

Resultados

Neste capítulo, são apresentados os resultados experimentais obtidos ao longo do desenvolvimento do trabalho. Primeiramente, é exposta a análise de exploração dos dados utilizando SOM. Posteriormente, são apresentados os resultados da criação dos modelos de classificação usando o algoritmo de clusterização K-médias (com diferentes preprocessamentos). Seguindo a ordem apresentada no Capítulo 4, os modelos obtidos com NMF e com os algoritmos baseados na natureza são mostrados. Por fim, todos os resultados são confrontados e analisados.

5.1 Exploração dos dados

A Figura 5.1 mostra a distribuição cumulativa de acordo com a medida de densidade do petróleo. A função-degrau (em vermelho) mostra a função de distribuição empírica. Os marcadores pretos representam as observações das amostras e linha contínua é a curva distribuição cumulativa estimada por uma KDE gaussiana.

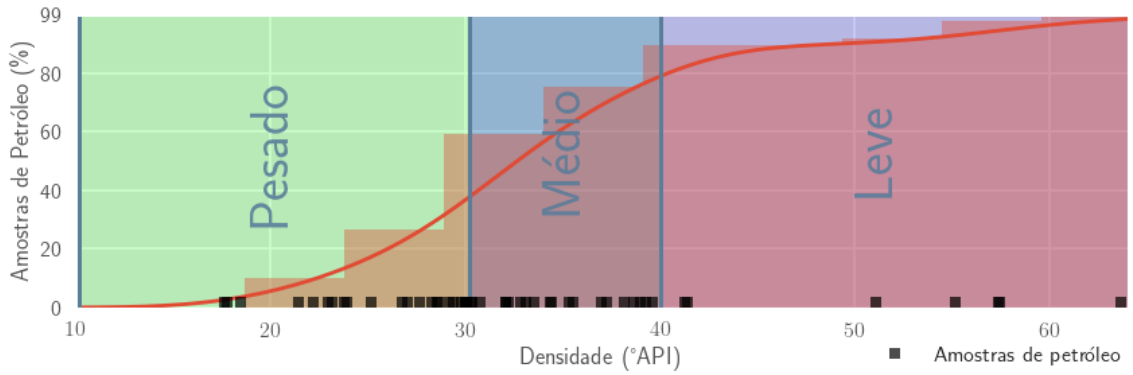


Figura 5.1: As 49 amostras estão distribuídas entre as classes: leve, médio, pesado.

Pode-se observar, que a partir da classificação baseada na densidade, as amostras utilizadas nessa tese podem ser classificadas como amostras *leves* (20 %), *médias* (40 %) e *pesadas* (40 %). Como dito anteriormente, todos os procedimentos de clusterização focarão em três agrupamentos.

A Figura 5.2 mostra as distribuições das propriedades dos dados das amostras dos 49 petróleos, bem como a correlação de uma em função da outra. A diagonal da matriz mostra a função de densidade de probabilidade de cada propriedade, também estimado utilizando kernel gaussiano.

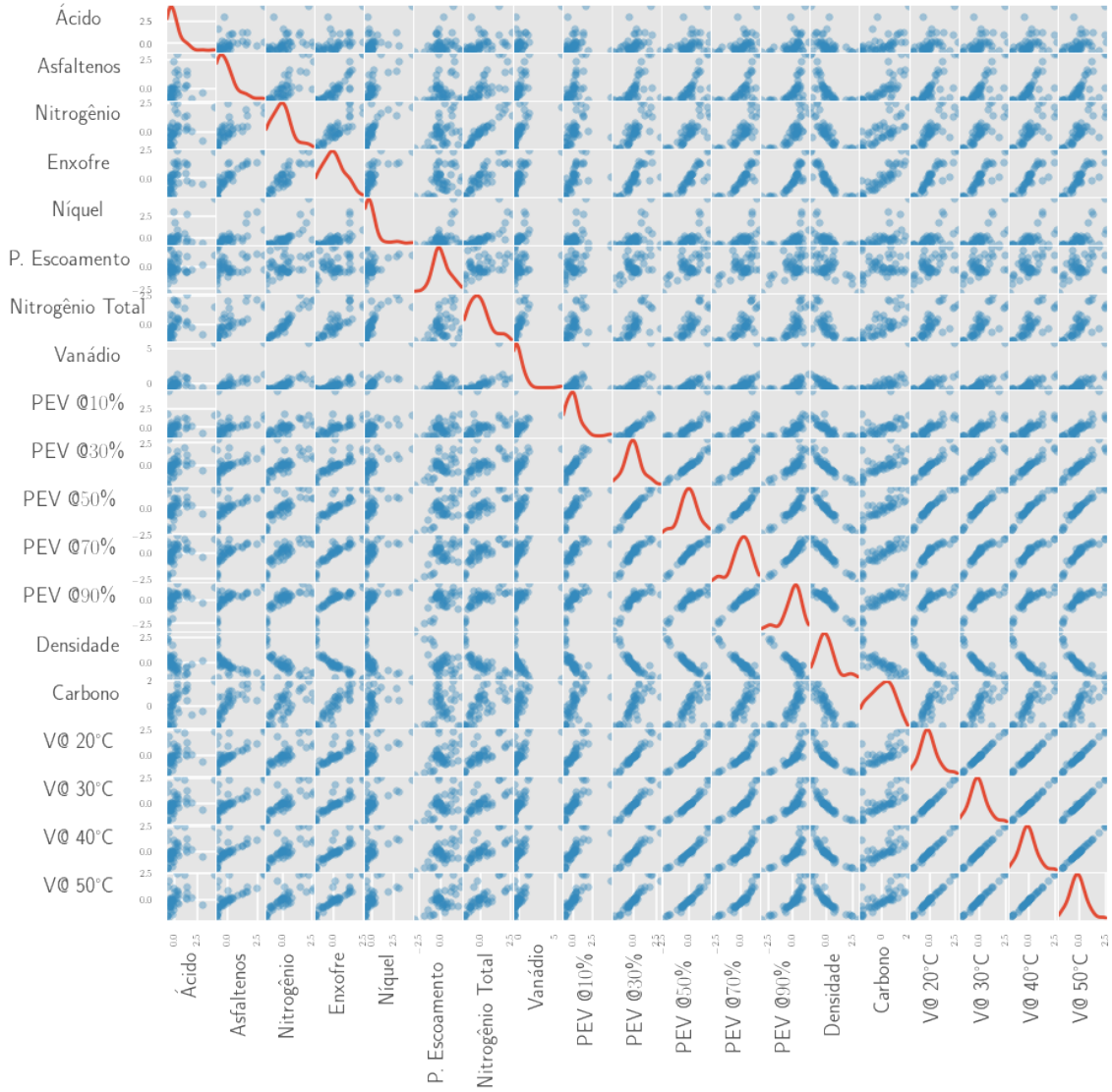


Figura 5.2: Correlação entre as propriedades das amostras de petróleo. A diagonal representa a função de densidade de probabilidade da propriedade, considerando um estimador por kernel Gaussiano.

Pode-se observar que algumas propriedades são altamente correlacionadas entre si, o que é esperado, já que muitas se relacionam através de formulações químí-

cas. Por exemplo, não é esperado que viscosidade e densidade sejam independentes. Seguindo o mesmo raciocínio, o ponto de escoamento é associado com conteúdo parafínico elevado, tipicamente achado em amostras mais leves, devido as características de suas ligações de carbono [156].

Como pode-se esperar pela presenta de dados correlacionados, a utilização da análise de componentes principais ocasional um grande fator de compactação dos dados. Quando analisamos a curva de variância acumulada (Figura 5.3), pode-se observar que a primeira componente representa mais do que 60 % da energia do sinal. Ainda, quase toda a variância (99,38 %) está acumulada nos 10 primeiros componentes, o que produz uma redução de aproximadamente 50 % da dimensionalidade.

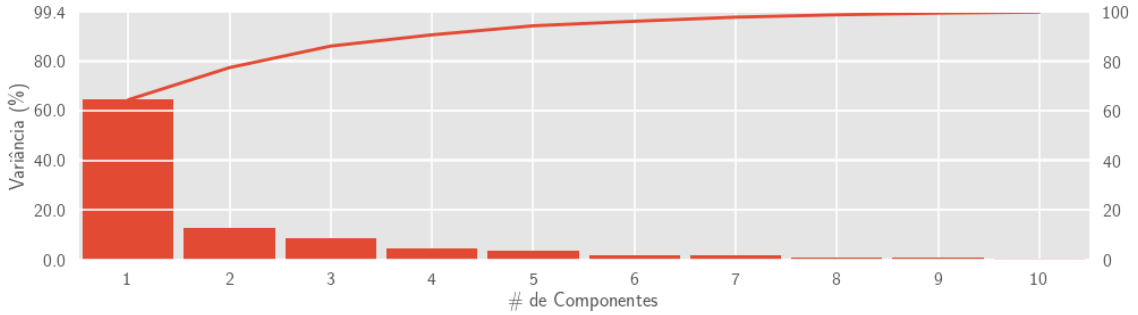
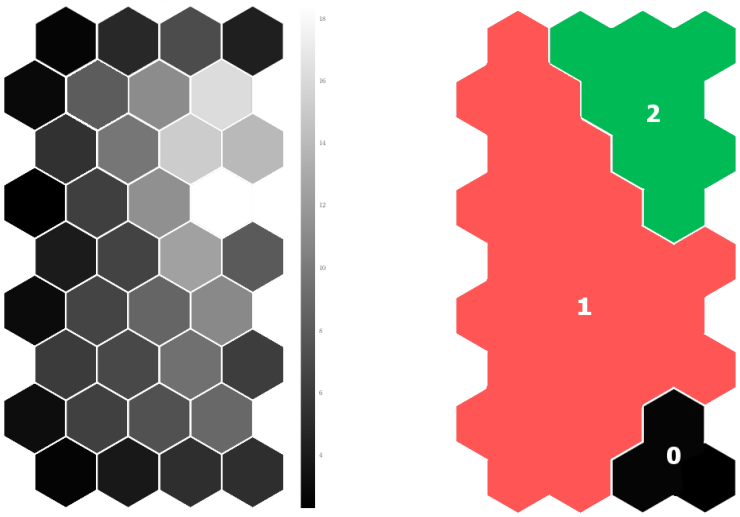


Figura 5.3: Curva da variância acumulada para PCA sobre as propriedades das amostras de óleo cru.

A partir da relação entre as duas primeiras componentes da PCA, ficou estabelecido que o formato do mapa SOM é 9×4 (ver Equação 4.1), totalizando 36 neurônios, usando topologia toroidal. A Figura 5.4a apresenta a *D-Matrix* (matriz de distâncias) do mapa criado. Pode-se notar que não há evidências claras à existência de clusters através da inspeção visual. Assim, o *code-book* da grade do SOM foi clusterizado usando o algoritmo K-médias. A Figura 5.4b mostra a melhor configuração de três clusters obtidas.

A partir do mapa SOM, podemos estimar a contribuição de cada variável de entrada para a composição do modelo através de cada uma de suas projeções. A Figura 5.5 mostra os planos de componentes, uma para cada propriedade de petróleo. Alguns detalhes interessantes pode ser destacados:

- A densidade realmente influência a região da parte inferior do mapa, à direita. Essa é a mesma região que concentrou as amostras de óleo bruto com



(a) Visualização da *D-Matrix* dos mapas SOM de duas dimensões usando as projeções dos dados normalizados.

(b) Clusters calculados sobre o SOM *codebook*.

Figura 5.4: Mineração de dados através da exploração de dados por SOM.

menor densidade como mostrado na Figura 5.6. Nenhuma outra propriedade influenciou essa região de maneira tão proeminente.

- Por outro lado, as variáveis ligadas a viscosidade são responsáveis por ativar os neurônios localizados na parte superior-esquerda. Assim, essa região concentra as amostras com os maiores valores de viscosidade (Figura 5.6).

- Medidas de alguns componentes como nitrogênio e níquel também influenciam mais a parte superior do mapa. Valores altos para essas variáveis indicam petróleos mais pesados. Essa região parece representar os óleos pesados mais viscosos.

- É possível observar que a influência da fluidez do óleo (ponto de escoamento) está espalhada pelos neurônios centrais do mapa, onde essa nem a densidade ou nenhuma das viscosidades apresentaram predominância. De fato, essa propriedade foi a única a ativar os neurônios localizados no lado esquerdo inferior do mapa. Essa área do mapa é parte do agrupamento rotulado como $C1$ (ver Figura 5.5). PVE e a quantidade de carbono também tem forte incidência sobre esse cluster, já que eles ativam a parte superior esquerda do mapa.

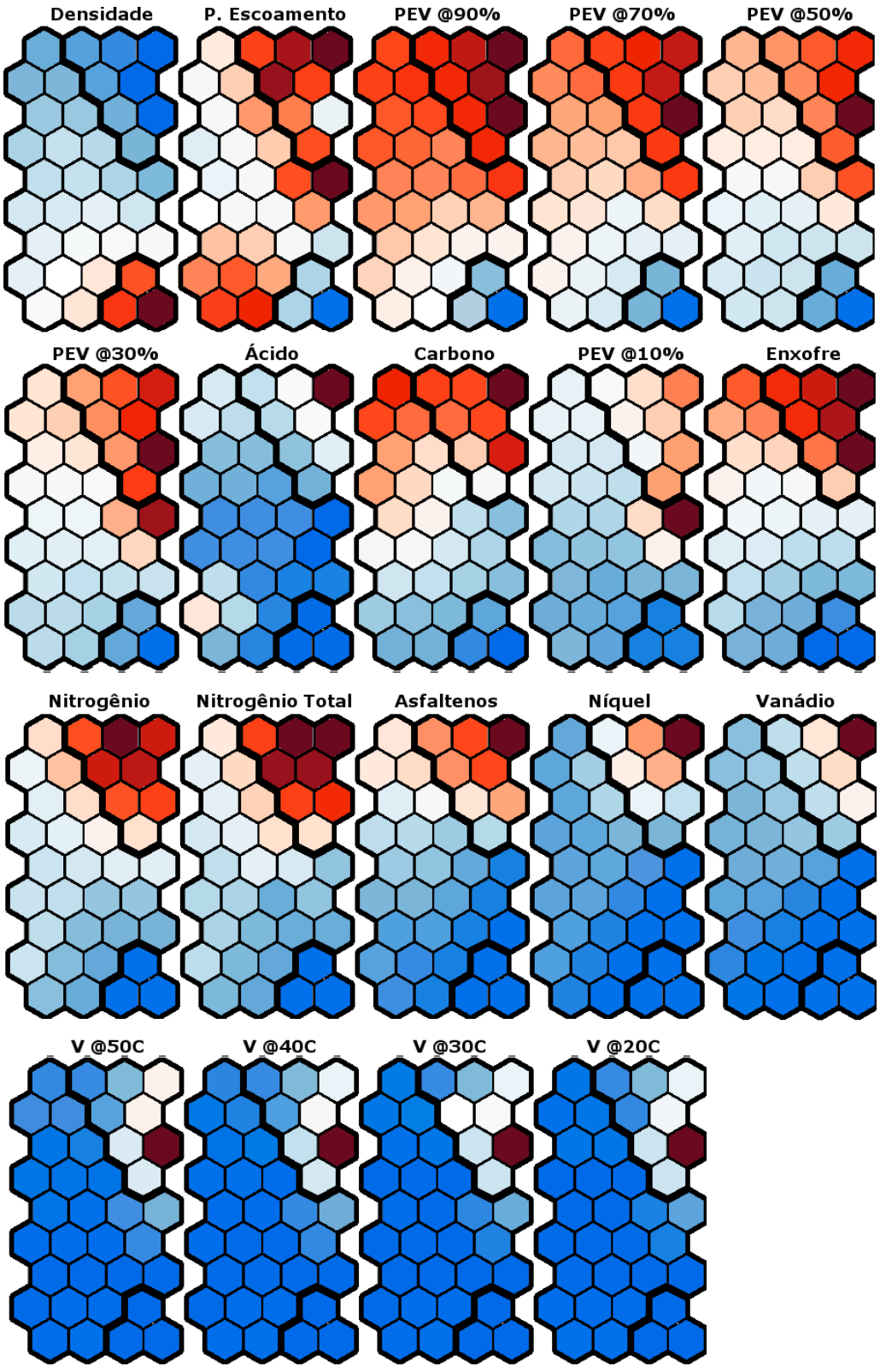


Figura 5.5: A contribuição de cada uma das propriedades do petróleo ao SOM. Quanto mais excitado um neurônio, mais vermelho é a representação. Ao contrário, as células azuis indicam que o neurônio correspondente não está sendo ativado. Os clusters do *code-book* da grade do SOM usando K-médias é destacada.

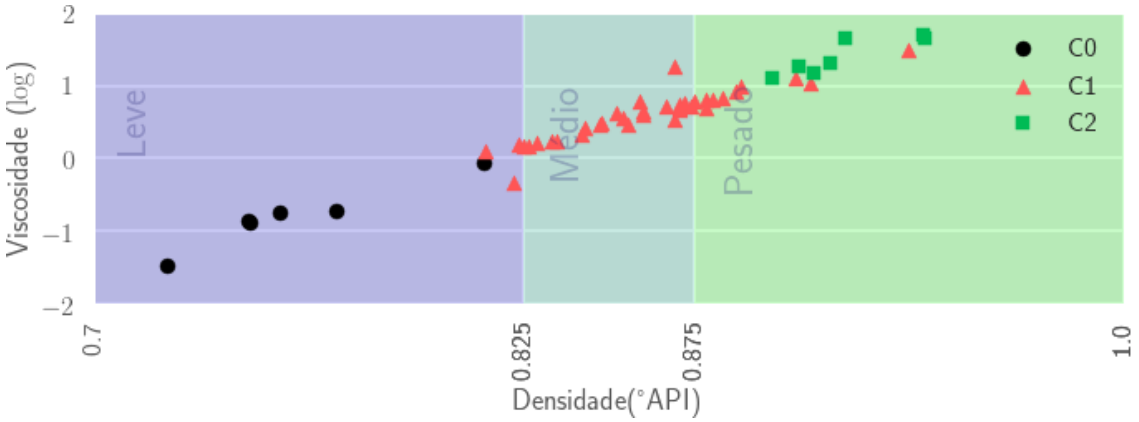


Figura 5.6: Viscosidade e densidade para as amostras de óleo bruto. Os petróleos estão marcados de acordo com os três clusters: $C1$, $C2$ e $C3$.

Ao realizar essa inspeção nos planos de componentes e entendendo como o algoritmo SOM organiza as diferentes amostras, é possível entender padrões aprendidos por esse método não supervisionado. O Cluster $C2$ está localizado na parte superior direita do mapa, região fortemente influenciada por medidas de viscosidade e elementos metálicos. Por outro lado, o Cluster $C0$ está localizado na parte inferior direita, onde a medida de densidade é mais relevante, já que os neurônios localizados nessa região são bastante sensíveis a essa propriedade. Essa análise sugere que mesmo que a densidade e viscosidade tenham papeis primordiais na classificação de petróleos, outras propriedades, como a fluidez, carregam informação suficiente para impactar a maneira que as amostras se agrupam. Assim, essa tarefa de exploração de dados aponta que explorar nuanças provenientes de diferentes propriedades tem um impacto positivo numa melhor categorização dos óleos brutos, o que nesse trabalho realizamos através dos métodos de clusterização.

5.2 Classificação dos petróleos através de K-médias

Para criar o modelo de classificação utilizando o algoritmo de K-médias foram utilizados dois mecanismos distintos de compactação. Primeiro, aplicando PCA e mantendo as 10 primeiras componentes e projetando elas em um mapa SOM. A segunda abordagem foi utilizar 5 dimensões intrínsecas do ISOMAP (ver Figuras 5.3 e 4.2 respectivamente).

A Figura 5.7 mostra o desempenho para a tarefa de clusterização de acordo com as figuras de mérito utilizadas em função do número de clusters usados no algoritmo de K-médias, com as diferentes compactações utilizadas. No gráfico é mostrado valor médio e o desvio-padrão de 100 diferentes inicializações do algoritmo. A configuração com dois clusters apresenta os melhores valores de índices. Contudo, como observado anteriormente, foi escolhida a utilização de três agrupamentos, visando a comparação direta com o método padrão de classificação. Pode ser notado, que ambos os índices indicam que essa configuração é plausível, já que se encontram dentro da mesma barra de erro. Como esperado, o desempenho da clusterização é melhorado após a compactação de dados, ambos preprocessamentos, PCA e ISOMAP, tem medidas melhores que os obtidos quando apenas foi usado SOM.

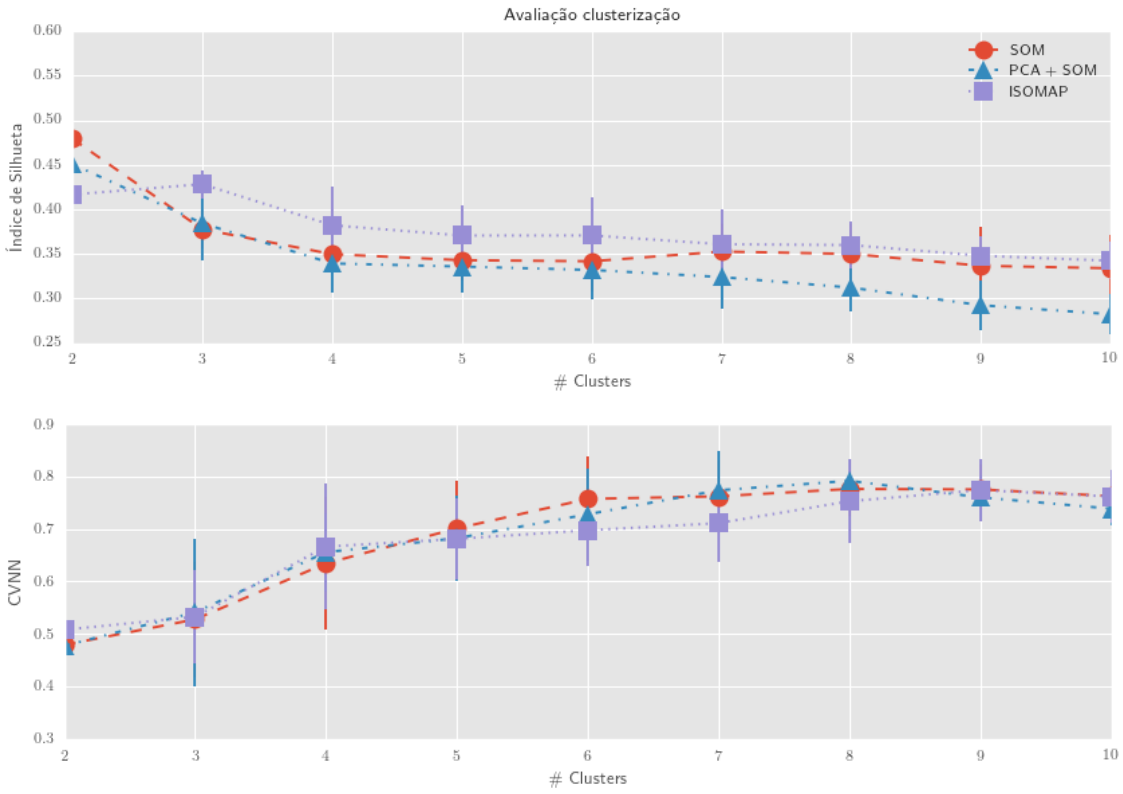


Figura 5.7: Avaliação da clusterização utilizando os índices CVNN considerando SOM sobre os dados normalizados e depois da compactação de dados: transformação por PCA e projeção por ISOMAP.

O índice de silhueta pode ser analisada considerando cada amostra disponível (ver Equação 3.19). Essa análise revela quais amostras estão bem representadas pelo seus respectivos clusters. A Figura 5.8 apresenta uma forma de avaliação visual muito utilizada em relação ao índice de silhueta. Nela, cada amostra é representada por

um ponto em seu respectivo cluster. Amostras com coeficiente de silhueta próximos a $1,0$ indicam que a a amostra se encontra longe dos clusters vizinhos, estando assim bem condicionada (ver Equação 3.20). O gráfico também permite entender o quão equilibrados os clusters são em relação ao número de elementos que eles representam.

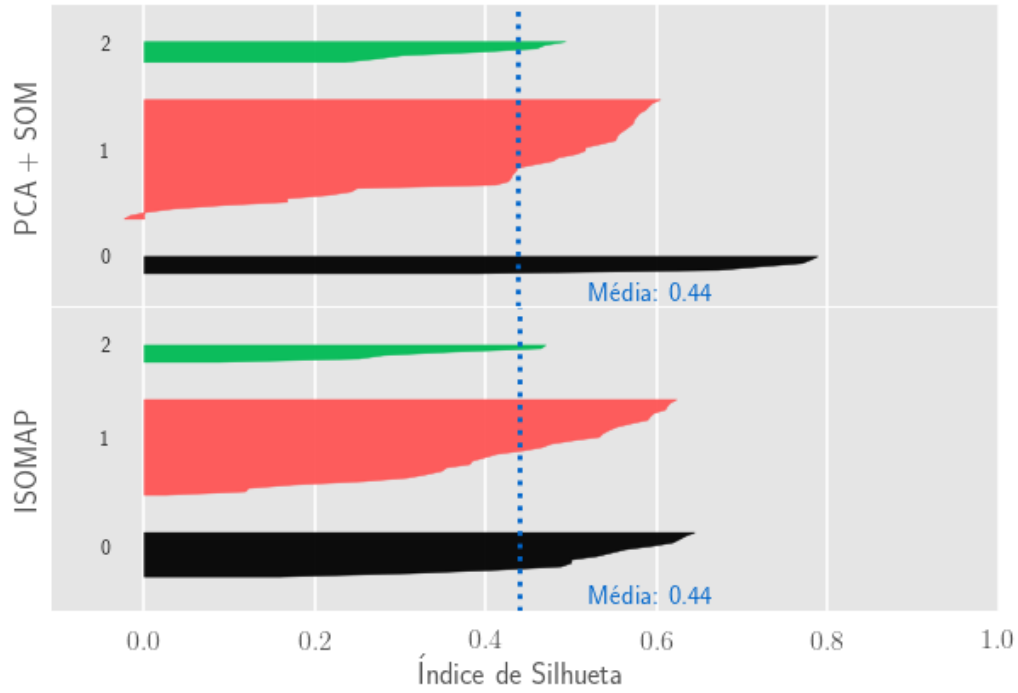


Figura 5.8: Índice de silhueta para cada amostra, considerando a clusterização realizada com compactação de dados: usando PCA e SOM em conjunto ou as dimensões intrínsecas do ISOMAP.

É possível verificar que em ambos as abordagens as amostras estão bem condicionadas, onde todos os agrupamentos possuem pontos com índice de silhueta acima da média geral. Contudo, podemos dizer que desempenhou melhor na representação dos dados devido a menor flutuação no tamanho dos três clusters comparado aos clusters obtidos quando combinamos PCA e SOM.

A Figura 5.9 mostra a função da distribuição acumulada (CDF) para os clusters de cada abordagem mostrada até aqui, considerando o valor da densidade, sobreposta pela classificação típica. As curvas CDF foram estimadas usando a densidade de *kernel* (KDE) com base gaussiana, para qual a largura da banda foi determinada de acordo com a Regra de Scott [157]. As curvas indicam o quanto cada agrupamento obtido avança sobre uma ou mais classes da classificação típica. Alguns dos agrupamentos invadem os limites em percentagem relevantes dos dados (cada valor é estimado usando a CDF). A frequência que amostras são classificadas de maneira

diferente são realçadas no gráfico pelas áreas hachuradas.

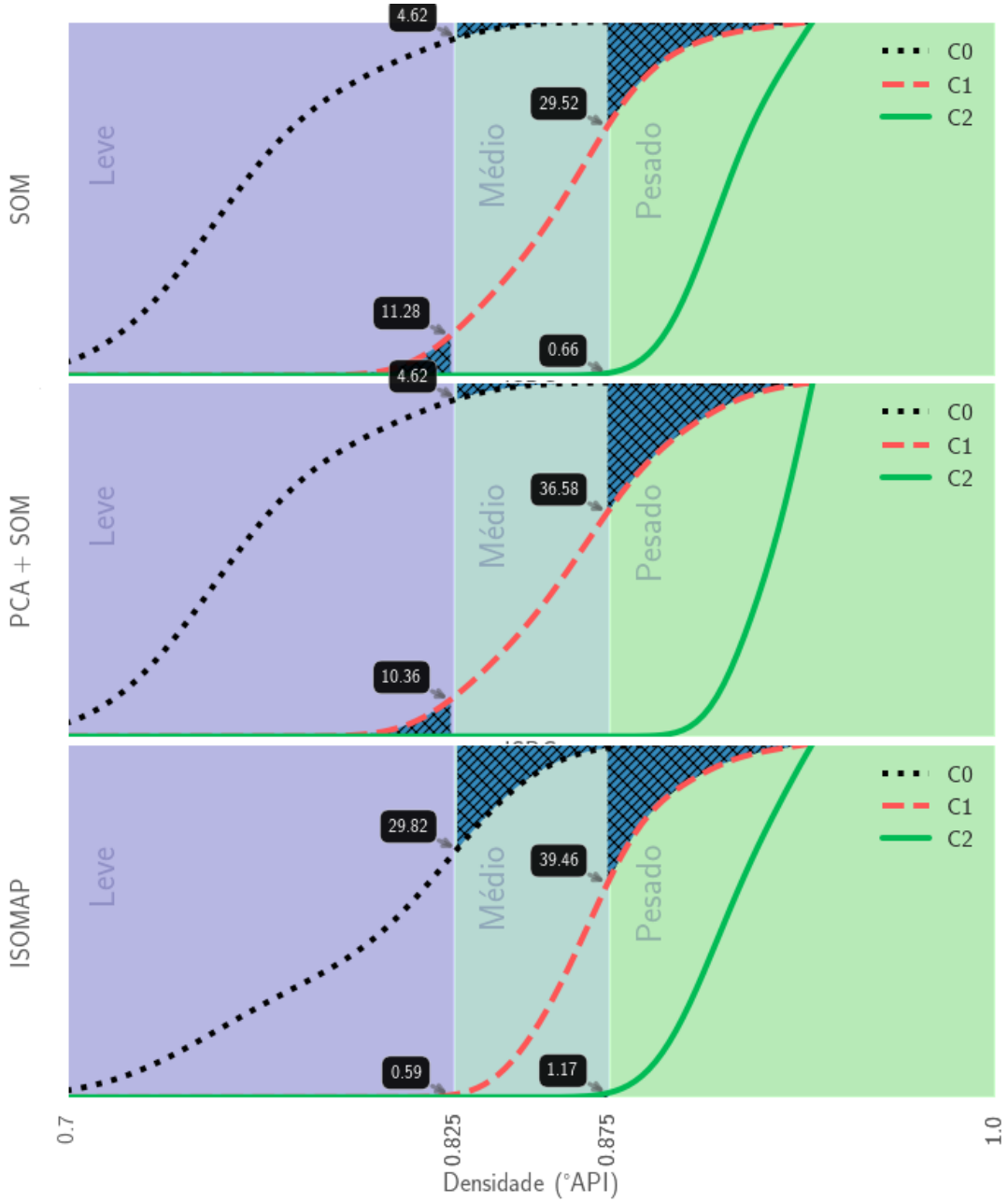


Figura 5.9: Curvas CDF para cada um dos clusters em comparação com a classificação típica (usando densidade).

Ao realizar a análise em relação a essas áreas hachuradas (onde os clusters cruzam os limites da classificação padrão), pode-se notar que o algoritmo ISOMAP apresenta as maiores áreas de interseção. Para esse algoritmo o cluster $C0$ compreende amostras das classes Leve e Médias. Aproximadamente 30 % das amostras deste cluster seria classificadas como Médias pelo critério de densidade. Contudo, a análise multivariada encontrou similaridades que vão além. Esse fato sugere que padrões

intrínsecos são deixados de lado ao considerar apenas a informação de densidade.

A Figura 5.10 apresenta cada uma das distribuições de propriedades divididas entre os clusters computados.

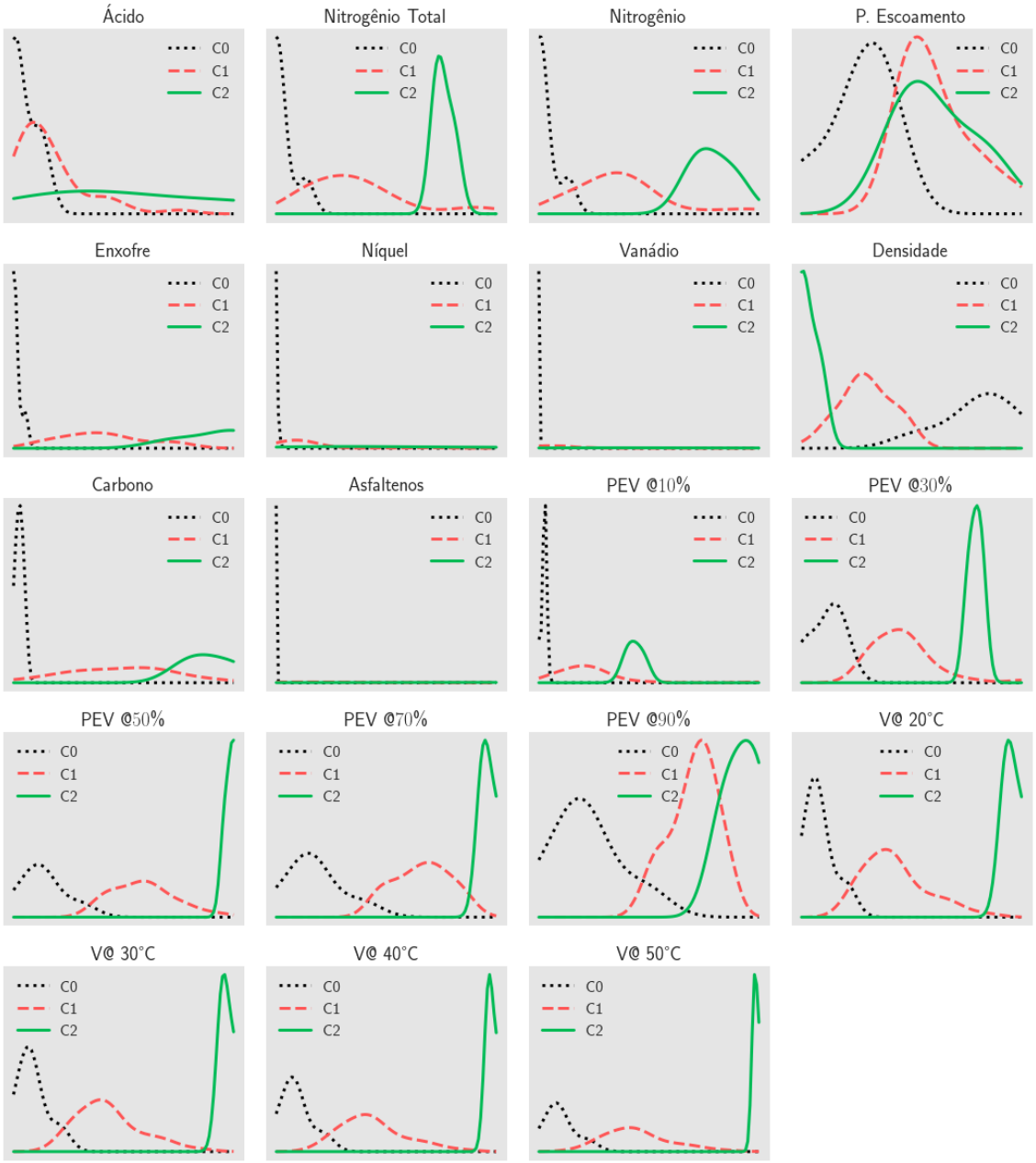


Figura 5.10: Distribuição das propriedades, considerando os três agrupamentos estimados pelo K-médias sobre a projeção dos dados nos dez primeiros componentes principais.

As PDFs foram estimadas novamente utilizando a estimativa de densidade *kernel* usando base gaussiana. Podemos ver que o cluster $C2$ pode ser separado observando apenas as propriedades de viscosidade, o que condiz com a conclusão anterior que essas variáveis são determinantes para agrupar petróleos pertencentes ao cluster de

petróleos mais pesados. As propriedades de fluidez (especialmente as variáveis PEV @10 % e PEV @30 %) aparentam ser relevantes para a determinação do cluster $C1$, mas não são sozinhas responsáveis pela separação, pois há pequena interseção entre as distribuições.

A Figura 5.11 mostra a mesma visualização de dados, mas para os modelos obtidos com ISOMAP. Como a separação ocorre no espaço da transformação (não-linear), é esperado que nenhuma variável sozinha seja responsável pela separação.

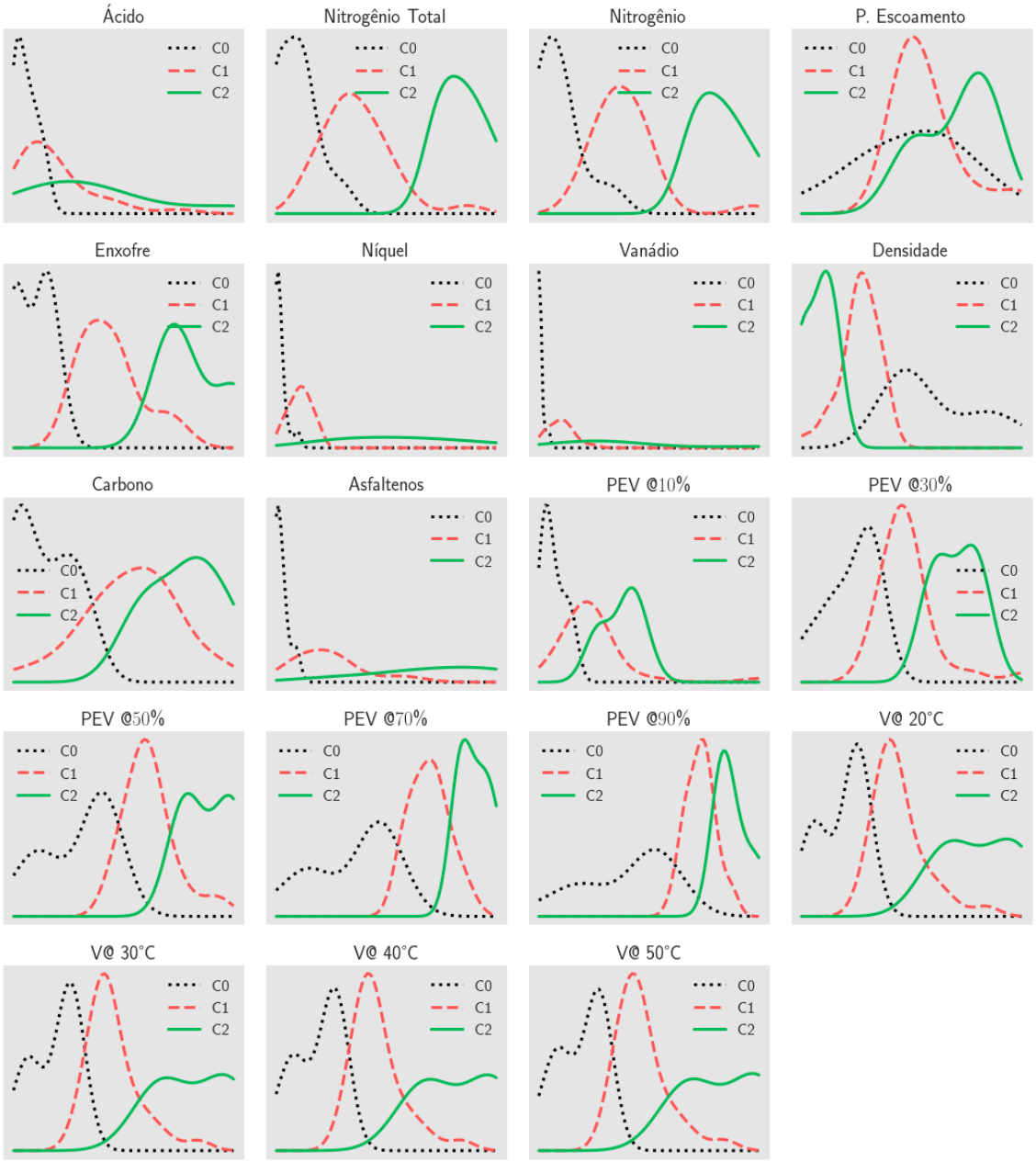


Figura 5.11: Distribuição das propriedades, considerando os três agrupamentos estimados pelo K-médias sobre a projeção dos dados nos dez primeiros componentes principais.

5.3 Classificação dos petróleos através de heurísticas baseadas na natureza

Com o algoritmo de K-médias, a clusterização é realizada a partir da otimização da distância entre cada amostra e o centroide do cluster ao qual ela pertence. Nessa seção, são apresentados os resultados utilizando os algoritmos de otimização baseados na natureza, mais especificadamente PSO, ABC e FSS. Como mostrado na Seção 3.5, tais algoritmos permitem um refinamento na função custo, permitindo que esta apresente uma composição mais complexa. Nesta pesquisa, como explicado na Seção 4.4, utilizou o índice de silhueta como base para a uma nova função objetivo, não só preocupada na composição de um cluster, mas também na relação dele com os demais encontrados. Foram considerados modelos compostos por três centroides.

A Figura 5.12 mostra a convergência para todas as variações utilizadas dos algoritmos. No gráfico, os pontos mostram a média do *fitness* para todas as inicializações e a barra de erro representa o desvio padrão. Para facilitar a visualização, são mostrados pontos a cada 10 passos. Em geral, os algoritmos tiveram uma convergência mais rápida, contudo, com 40 movimentos, todos os algoritmos já haviam convergido.

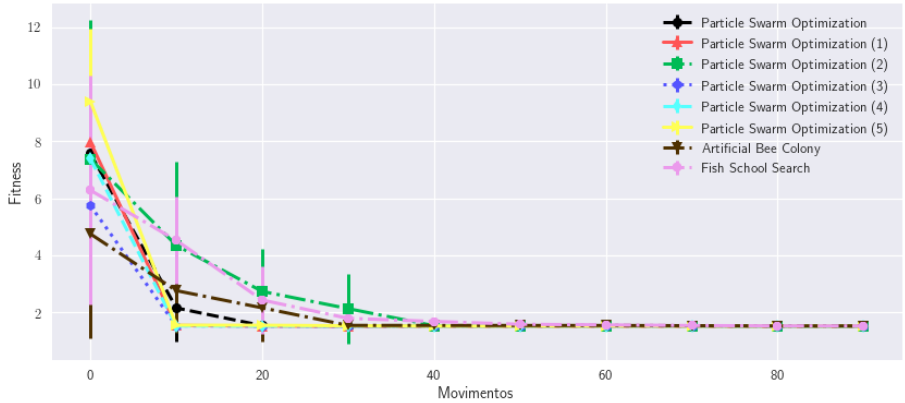


Figura 5.12: Convergência dos algoritmos de otimização baseados na natureza para a aplicação de clusterização.

A Figura 5.13 mostra o desempenho do melhor modelo obtido por cada algoritmo. No caso da PSO, a variação implementada com peso inercial Ω (ou fator de constrição) obteve o melhor resultado. O resultado do índice de silhueta para essa

configuração foi $0,44$ é semelhante ao resultado alcançado com K-médias associado com um processo de preprocessamento (PCA em conjunto com SOM ou ISOMAP).

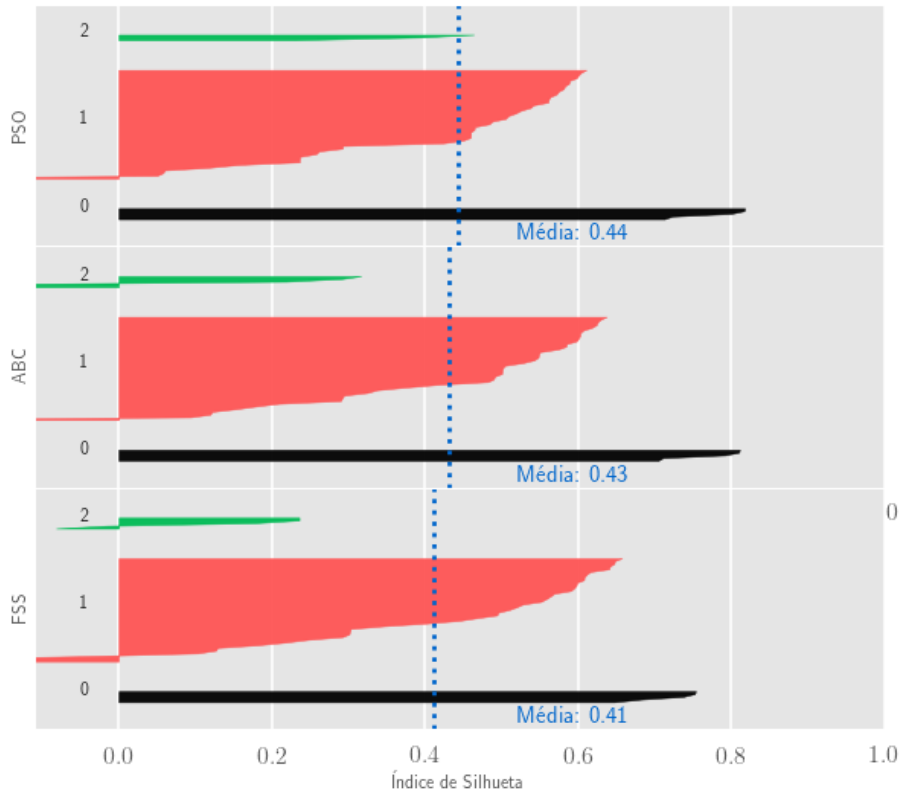


Figura 5.13: Curvas CDF para cada um dos clusters computados com algoritmos baseados na natureza em comparação com a classificação típica (usando densidade.)

Contudo, esta análise não conta com nenhum tipo de preprocessamento. A Figura 5.14 é mostrada apenas a mérito de comparação. Aqui, K-médias foi inicializado 1000 vezes, com inicialização usando o algoritmo de *kmeans++* [150]. O melhor modelo obtido desempenhou com índice de silhueta igual a $0,37$, inferior às três heurísticas.

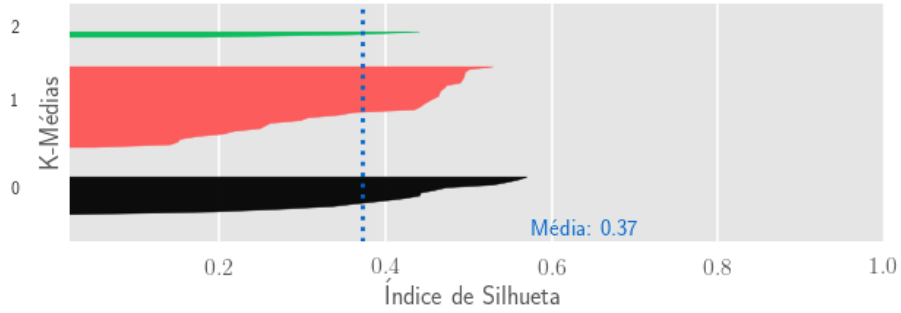


Figura 5.14: Curvas CDF para cada um dos clusters computados com K-médias.

Em termos de desempenho, os clusters obtidos através de PSO, apresentam-se

melhor conformados, visto que todos os agrupamentos possuem representantes acima da média e existem menos amostras com valores negativos. Todos os algoritmos captaram as mesmas características gerais, o que reflete a quantidade de amostras em cada cluster. Essa conclusão também pode ser tirada analisando o gráfico mostrado na Figura 5.15, pois as interseções entre os clusters obtidos e as classes típicas são similares também.

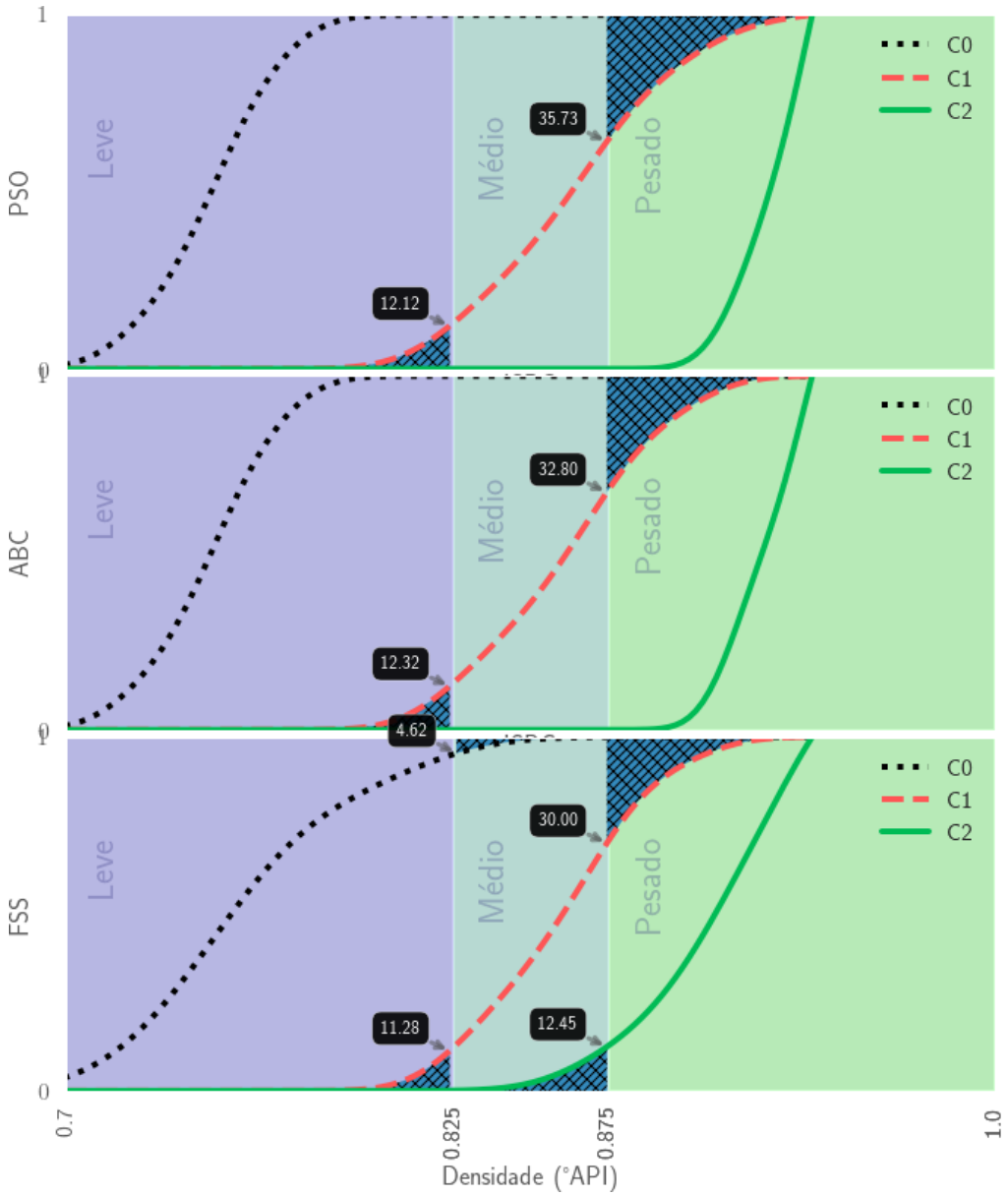


Figura 5.15: Curva CDF para cada um dos clusters encontrados com os algoritmos de otimização baseados na natureza em comparação com a classificação típica (usando densidade).

Em todos os casos, o cluster $C1$ invadiu tanto os clusters de óleos classificados como leves (PSO: $12,12$ %, ABC: $12,32$ % e FSS: $11,28$ %) como os pesados

(PSO:35, 73 %, ABC: 32, 80 % FSS: 30, 00 %) em patamares similares. Um diferença relevante, contudo, é o algoritmo FSS ter agrupado uma amostra classificada como óleo médio associada no cluster caracterizado por petróleos mais denso ($C2$). Na Figura 5.16, as amostras de óleo bruto são organizadas em um plano cartesiano que projeta a viscosidade em função da densidade. Cada amostra é destacada por um marcador diferente. A amostra que tipicamente seria classificada como um óleo médio possui medida de viscosidade muito maior do que óleos com densidade similar. Portanto, como caracterizado na análise realizada na Seção 5.1, FSS também captou a viscosidade como característica relevante para a composição do cluster $C2$.

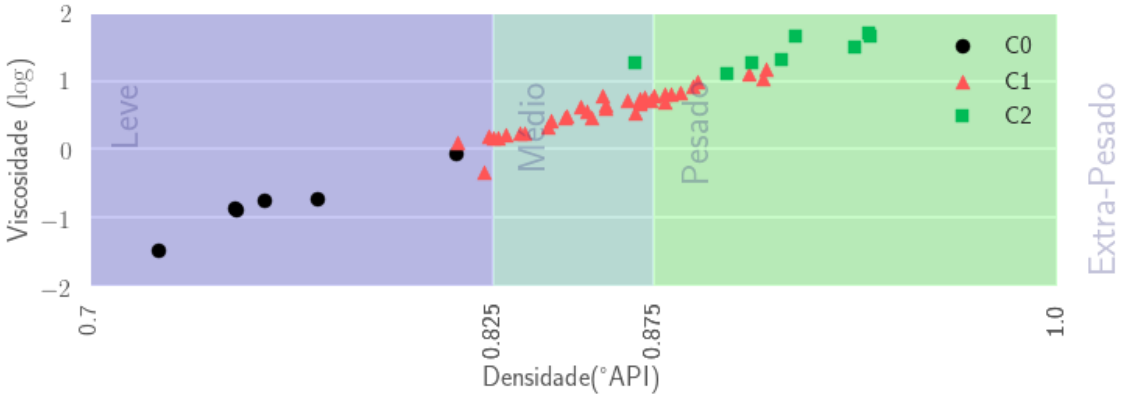


Figura 5.16: Amostras de Petróleo organizadas de acordo com a densidade e viscosidade. Cada amostra é destacada de acordo com o cluster que foi categorizada.

5.4 Classificação dos petróleos através de NMF

Até agora, as amostras (ou suas projeções) foram agrupadas em função da distância euclidiana. Como foi exposto no Capítulo 3, o algoritmo da NMF permite realizar a tarefa utilizando outras distâncias ou divergências. Como dito na Seção 4.5, iremos explorar uma família de divergências conhecida com divergência β. Ao longo dos próximos parágrafos serão mostrados como foi escolhido o valor de β e consequentemente o modelo mais adequado, sua resposta em relação ao índice de silhueta e de que maneira cada propriedade influenciou nessa categorização. As análises focarão na obtenção de três clusters, como realizado na análise de SOM e ISOMAP.

Escolha do valor de β

Duas informações são determinantes para a melhor escolha do valor da constante β: a convergência do treinamento e a estabilidade frente às diferentes inicializações. Como trata-se de uma tarefa de clusterização iremos avaliar também o índice de silhueta. Usaremos seis valores distintos de β, inclusive para os casos especiais onde a divergência assume o comportamento da divergência de Itakura-Sato ($\beta = 0$), divergência de Kullback-Leibler ($\beta = 1$) e distância Euclidiana ($\beta = 2$).

Pela Figura 5.17, nota-se que para todos os casos a fatoração convergiu assintoticamente. Cada gráfico representa uma configuração diferente, a linha representa a média de todas as 200 inicializações e a barra de erro é formada pelo valor do desvio-padrão. Por se tratarem de divergências distintas, não podemos comparar os patamares atingidos.

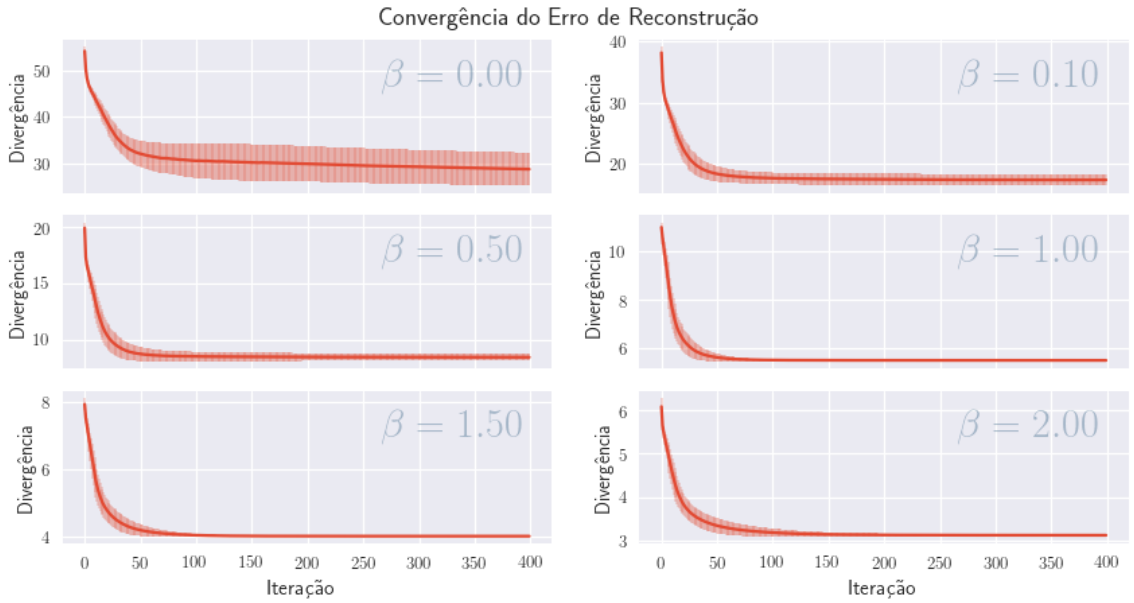


Figura 5.17: Convergência para o processo de fatoração para os diferentes valores de β.

Para estabelecer um patamar de comparação podemos recorrer ao erro médio quadrático na reconstrução, mostrado na Figura 5.18. A linha tracejada em cada gráfico representa o valor da média para as 200 inicializações e, novamente, a barra de erro é calculada a partir do desvio padrão. A linha pontilhada vertical indica o ponto de parada do algoritmo. Pode-se verificar que o modelo que utiliza Itakura-Sato diverge nessa avaliação. Ao contrário, todos outros modelos convergiram, embora notamos que quanto menor o valor de β, menor foi o valor alcançado. É impor-

tante ressaltar, que essa análise sozinha não é suficiente para escolher ou mesmo descartar alguma configuração, visto que a medida de erro quadrático favorece a distância euclidiana, por sua similaridade matemática. No entanto, no caso da não convergência, trata-se de um indicativo negativo.

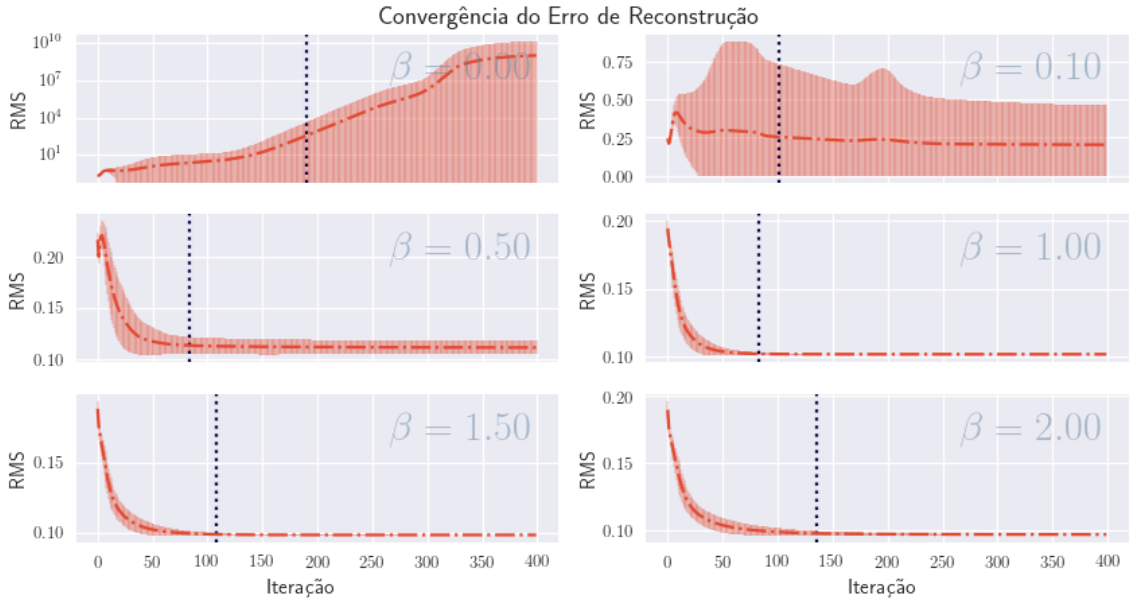
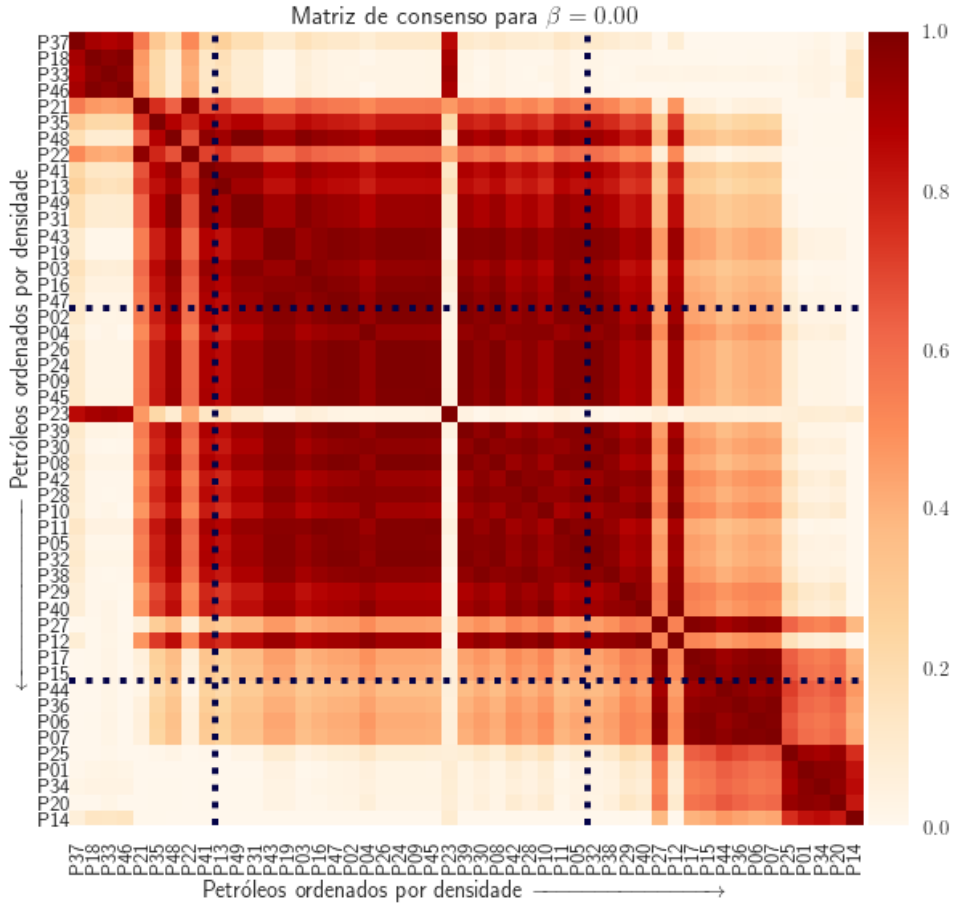


Figura 5.18: Erro médio Quadrático (RMS) para os diferentes valores de β.

Continuando a análise sobre a estabilidade dos modelos, foram estabelecidas as matrizes de consenso contendo uma visualização ligeiramente modificada em relação ao que foi proposto em Brunet (2004), como pode ser visto na Figura 5.19.

Ao invés de ordenar as amostras a partir da frequência de coocorrência nos mesmos agrupamentos, a visualização apresenta os petróleos ordenados pela medida de densidade, medida em gravidade específica (SG). Para facilitar a comparação com a classificação realizada a partir dessa propriedade, duas linhas pontilhadas foram traçadas em cada um dos eixos, mostrando quais petróleos seriam classificados como Pesado, Médio e Leve (seguindo aa direção do eixo x). Apenas pela averiguação visual, pode-se notar que quando é utilizada a divergência de Itakura-Sato há uma maior dispersão nos valores da matriz. Quando observa-se a escala de cor, verifica-se diversos valores no entre $0,2$ e $0,8$. Com isso, os três clusters pretendidos não bem destacados (os petróleos leves poderiam ser divididos em dois agrupamentos distintos).

Ainda no primeiro gráfico, pode-se perceber um petróleo (P23) destacado migrando dos petróleos médios para o cluster de petróleos mais densos. De fato, esse



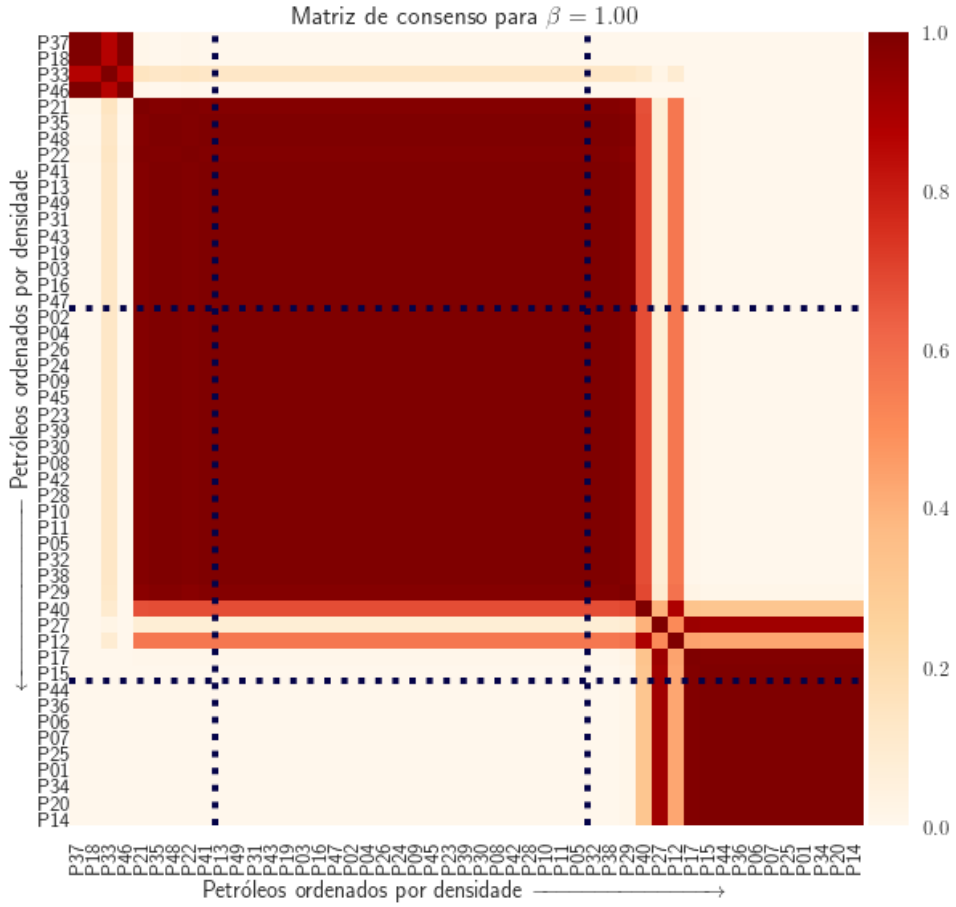


Figura 5.19: Matriz de consenso para dois valores distintos de β: Itakura-Sato e Kullback-Leibler.

petróleo é mais viscoso do que os petróleos que estão na faixa de densidade dele. Como visto na análise com SOM, a viscosidade é determinante para a criação do cluster $C2$. Esse resultado indica que o mesmo pode ser verdade para os modelos criados com NMF. Verificaremos essa hipótese mais adiante.

Analisando o segundo gráfico, observamos agrupamentos mais consistentes. A matriz possui poucos valores entre $0,2$ e $0,4$, o que indica que mesmo em 200 inicializações, os modelos criados são mais estáveis. Ao calcular os índices de estabilidade, as observações realizadas mostram-se corretas. A Tabela 5.1 mostra resumidamente o desempenho para cada uma das configurações testadas, e a partir dela, seleciona-se $\beta = 0.5$ e $\beta = 1.0$

β	I	Correlação Cofenética	β	I	Correlação Cofenética
0.0	0.58	0.95	1.0	0.68	1.00
0.1	0.67	1.00	1.5	0.67	0.99
0.5	0.68	1.00	2.0	0.64	0.98

Tabela 5.1: Índices de estabilidades baseados na matriz de consenso.

Finalmente, o valor de β pode ser definido, após verificar a variação do índice de silhueta para cada uma das inicializações. A Figura 5.20 mostra um *boxplot* com a mediana e a variação (através dos primeiro e terceiro quartis) dos índices de silhueta. A mediana para a configuração com $\beta = 1.0$ é um pouco mais alta que as demais distribuições. Já a dispersão dos valores de silhueta é menor para essa configuração. Deste modo, a divergência escolhida para o restante da análise será Kullback-Leibler.

Influência das propriedades

Como explicado anteriormente, através da fatoração da NMF, pode-se realizar duas atividades distintas de maneira simultânea. Para tal, basta analisar cada uma das matrizes fatoradas separadamente. Nesta análise, a matriz de mistura tem o formato 19×3, onde cada coluna representa uma dimensão do espaço original. Vimos também, que pela propriedade aditiva da NMF, podemos interpretar os valores dessa

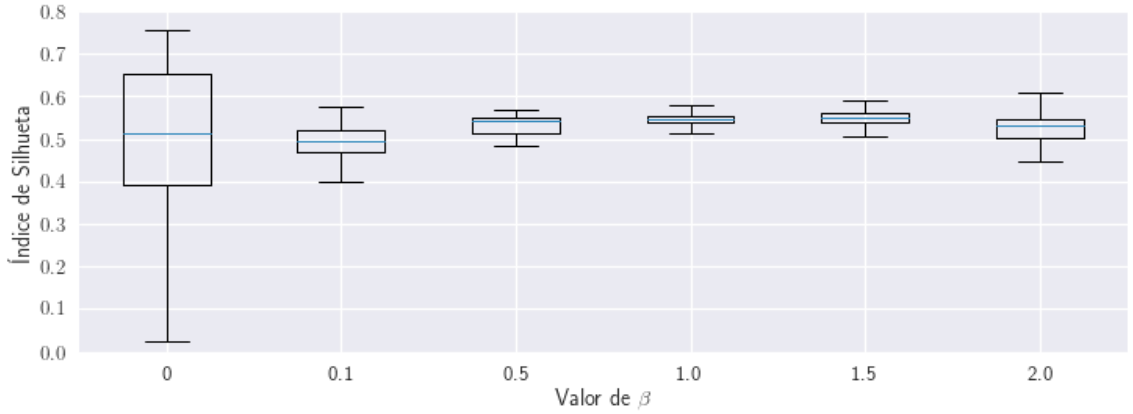


Figura 5.20: Representação em *boxplot* mostrando os valores do índice de silhueta para diferentes configurações.

matriz como a contribuição de cada dimensão para um determinado cluster.

A Figura 5.21 mostra um gráfico formado a partir da matriz de mistura da NMF. Nele, cada coluna representa uma propriedade distinta e as linhas suas projeções sobre os clusters, onde quanto mais escuro, maior sua importância. As propriedades são ordenadas a partir de uma clusterização hierárquica realizada com os vetores-coluna como entrada. O objetivo é reunir as propriedades que influenciam um determinado cluster e facilitar a compreensão das regras de composição dos clusters.

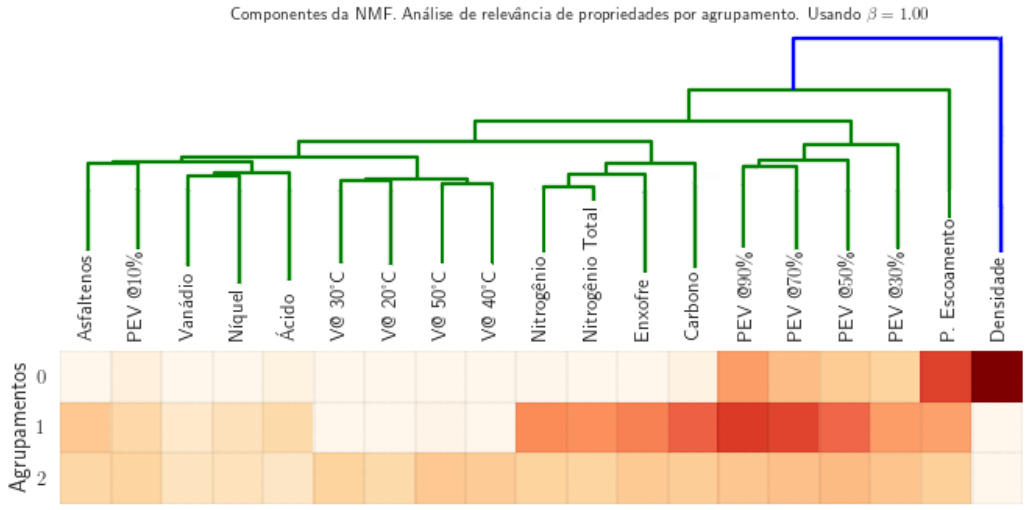


Figura 5.21: Influência de cada propriedade sobre os clusters. A matriz mistura da NMF é reordenada de acordo com o resultado clusterização hierárquica realizada em cima dos vetores colunas.

Pode-se perceber a propriedade densidade influência o cluster $C0$[1]. Ainda sobre

[1]Os clusters foram nomeados de maneira que concordem com as características dos clusters obtidos anteriormente, assim $C0$ é um agrupamento caracterizado por petróleos menos densos.

esse cluster, verifica-se que elementos químicos como metais, enxofre e nitrogênios não apresentam coeficientes elevados. Esse comportamento é esperado, visto que petróleos menos densos possuem menor concentração de elementos que não seja hidrocarbonetos.

Já as variáveis relacionadas ao ponto de ebulição, apesar de contribuírem com todos os clusters, possuem maior impacto no cluster $C1$, que também é caracterizado pelas contribuições dos elementos químicos Nitrogênio, Carbono e Enxofre.

Em relação ao cluster $C2$, nota-se que nenhuma propriedade se sobressai em termos da magnitude do coeficiente. Contudo, é importante perceber que as propriedades de viscosidade não apresentam valores de coeficientes relevantes para os clusters $C0$ e $C1$, indicando que petróleos mais viscosos tendem a ser agrupados juntos neste cluster $C2$.

Essa análise quando comparada com os resultados expostos na Seção 5.1 apresenta diversas similaridades:

- Petróleos menos densos caracterizam o cluster $C0$;
- Metais e demais elementos químicos não manifestam influência no cluster $C0$;
- Petróleos mais viscosos caracterizam o cluster $C2$;
- No SOM, as variáveis ponto de ebulição verdadeiro ativam a região central do mapa, onde estão os neurônios agrupados como $C1$, assim como na NMF contribuem mais para a segunda componente.

O fato das duas análises concordarem em diversos aspectos corrobora para a afirmação que outras propriedades contribuem para a tarefa de clusterização do petróleo.

É interessante pontuar como a clusterização hierárquica organizou as informações. O algoritmo permite, associado à sua visualização, agrupar as variáveis em diferentes níveis. Por exemplo, a densidade encontra-se destacada em relação às demais propriedades, demonstrando sua relevância na tarefa de categorização, principalmente para a classe de petróleos menos densos. Como esperado, diferentes mensurações para uma propriedade foram reunidas, vide as medidas de viscosidade e ponto de ebulição verdadeiro. Nitrogênio, enxofre e carbono formam um grupo de propriedades que é relevante para a identificação dos petróleos no cluster $C1$.

Modelo de classificação dos petróleos

A escolha de como as amostras se organizam em clusters é realizada através da matriz de componentes. A Figura 5.22 mostra os petróleos ordenados pelo valor de densidade. Duas linhas verticais denotam a região de fronteira para a classificação baseada nesta propriedade.

Os petróleos menos densos estão concentrados no cluster $C0$. Para o estabelecimento da classificação foi escolhido um procedimento de clusterização *hard*, onde o maior coeficiente do vetor-linha indica a qual cluster a amostra pertence.

Contudo, ao avaliar as nuanças do mapa de calor, verifica-se que em alguns casos, o limiar de escolha é próximo. Por exemplo, o petróleo P23 possui coeficientes em C1 e C2 similares. Como dito anteriormente, esse petróleo possui viscosidade alta quando comparado com petróleos com densidades similares. Desta maneira, apesar de, nesse caso, o petróleo pertencer ao cluster $C1$, ele possui características comuns aos óleos agrupados no cluster $C2$.

A visualização da Figura 5.22 também mostra que o cluster $C0$ se estende em petróleos das classes Leve e Médio da classificação típica. Já o cluster $C1$, reúne petróleos classificados como Médios e Pesados.

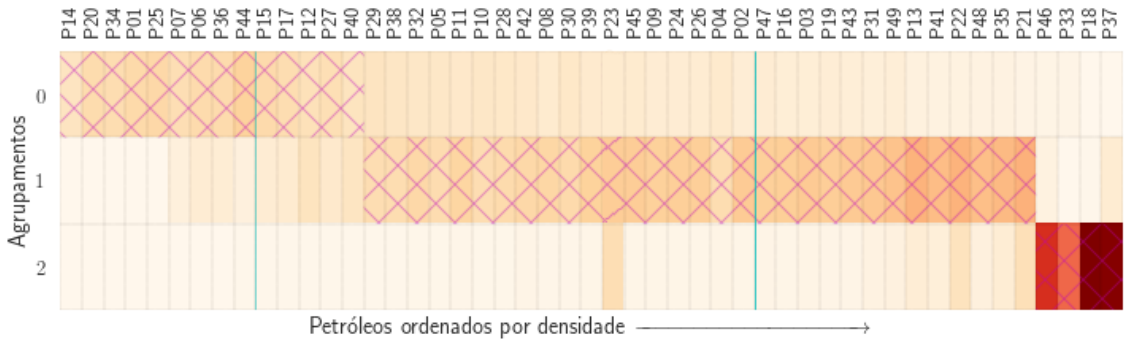


Figura 5.22: Coeficientes dos petróleos, ordenadas por densidade, mostrados através de mapa de calor. Cada linha representa um cluster e a escolha do cluster é feita a partir da hachura.

A Figura 5.23 observam-se as áreas de interseção entre os clusters encontrados e a classificação baseada em petróleo. O comportamento da invasão é similar ao que acontece no algoritmo de ISOMAP (ver Seção 5.2), onde:

- as amostras categorizadas no cluster $C0$ se estendem entre as classificações baseadas no °API “Leve” e “Médio” com aproximadamente $29,82$ % de suas

amostras sendo pertencentes à última;

- $43,65$ % das amostras agrupadas em $C1$ são originalmente classificadas como petróleos pesados.

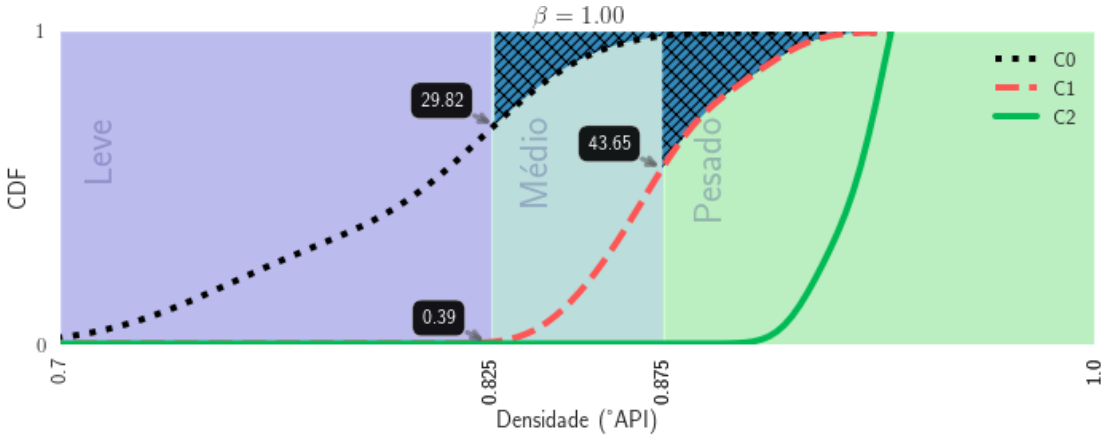


Figura 5.23: Curva CDF para cada um dos clusters encontrados com NMF em comparação com a classificação típica (usando densidade).

A Figura 5.24 mostra que as amostras estão bem condicionadas, onde todos os agrupamentos possuem pontos com índice de silhueta acima da média geral. O resultado da silhueta média apresenta-se superior aos encontrados nos algoritmos usados até aqui.

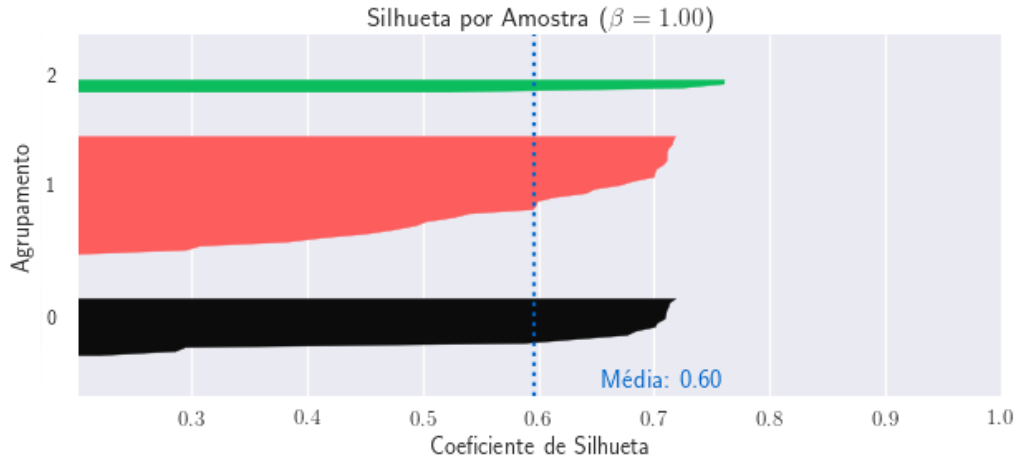


Figura 5.24: Índice de silhueta para cada amostra, considerando a clusterização realizada com NMF

5.5 Comparação do desempenho dos algoritmos

A Tabela 5.2 resume todos os resultados encontrados com as diferentes abordagens propostas, mostrando o melhor desempenho para cada uma. Em relação às figuras de mérito, conclui-se que a clusterização através da fatoração de matrizes teve os melhores resultados para ambos índices: silhueta e CVNN. O algoritmo PSO é o que teve maior cluster $C1$. O algoritmo FSS teve a pior performance em relação ao índice silhueta, mas teve o valor de CVNN entre os melhores desempenhos. Quando os dois índices não concordam, há indica que o cluster encontrado não é esférico (e então a distância euclidiana não é a melhor medida, o que prejudica o índice de silhueta).

Algoritmos	Leve-Médio	Médio-Leve	Médio-Pesado	Pesado-Médio	Silhueta	CVNN
PCA + SOM	4,6%	10,4%	29,5%	0,0%	0,44	0,44
ISOMAP	29,8%	0,6%	39,5%	1,17%	0,44[2]	0,35
NMF	29,8%	0,0%	43,7%	0,0%	**0,60**[3]	**0,34**
PSO	0,0%	12,1%	35,7%	0,0%	0,44	0,46
ABC	0,0%	12,3%	32,8%	0,0%	0,43	0,40
FSS	4,6%	11,3%	30,0%	12,5%	0,41	0,36

Tabela 5.2: Resultados obtidos com os diferentes algoritmos.

A Figura 5.25 dois gráficos distintos. O gráfico de cima apresenta a qual cluster as amostras de petróleo pertencem de acordo com cada um dos algoritmos. As amostras são ordenadas ordem crescente de densidade. Duas linhas verticais grossas demarcam a qual classe típica o petróleo pertence. Com base nessa visualização, pode-se destacar:

- Os algoritmos sempre agruparam os petróleos menos densos como cluster $C0$. A quantidade de óleos variou e tanto NMF como ISOMAP tiveram maior penetração em petróleos com densidades maiores.

- PSO classificou 41 agrupamentos como cluster $C1$.

[2]Calculado no espaço de busca do ISOMAP
[3]Calculado no espaço dos componentes da NMF

- ISOMAP apresentou as classes mais balanceadas.
- FSS foi o único algoritmo que apresentou um petróleo médio presente no cluster $C2$.

O gráfico de baixo é exatamente igual ao primeira, apenas reordenado por viscosidade (V @50). Os rótulos do eixo x, são marcados de acordo com a classificação baseada na densidade. Nessa visualização, vemos como se justifica agrupar o petróleo P23 com petróleos mais densos. Por esse óleo ser bastante viscoso, ele é atraído para o cluster $C2$, fortemente influenciado por essa propriedade.

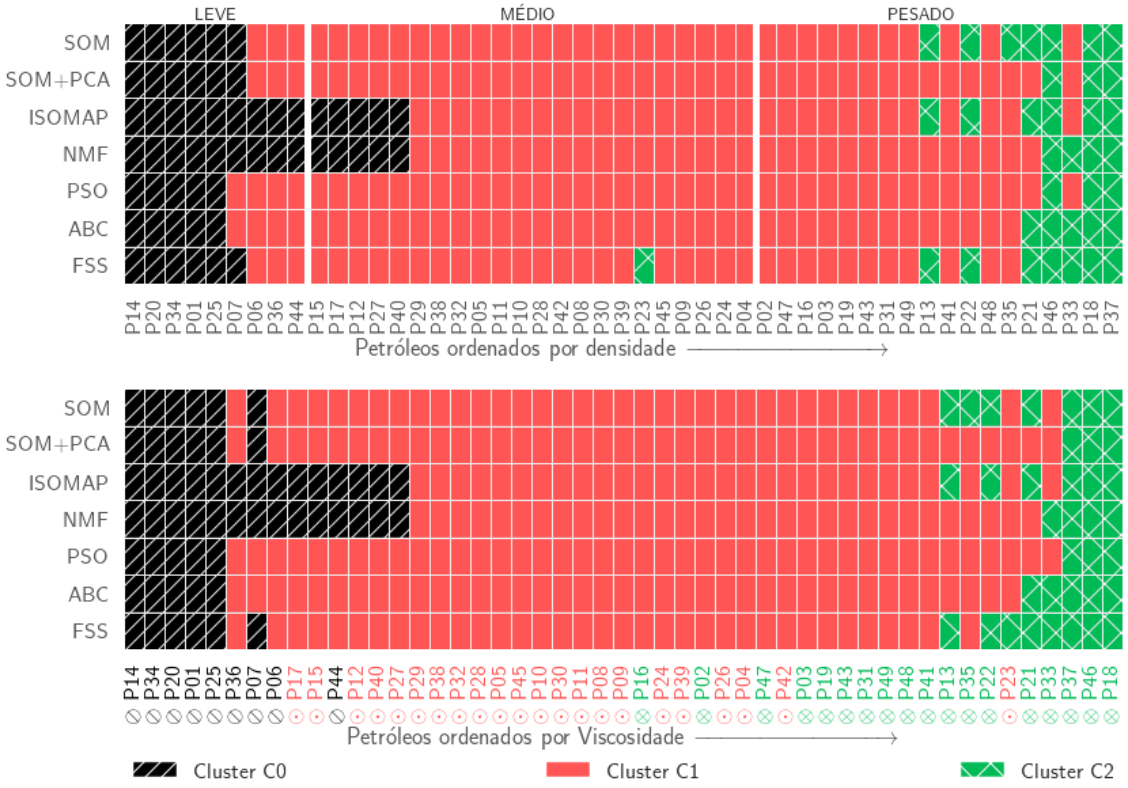


Figura 5.25: Clusterização dos petróleos.

5.6 Extensões do Modelo

Nessa sessão, serão mostradas duas extensões dos modelos. O primeiro é a comparação com o índice $°API/(A/B)$, proposto por Farah. Esse é um índice mais granular do que a escala baseada puramente baseada na densidade (°API).

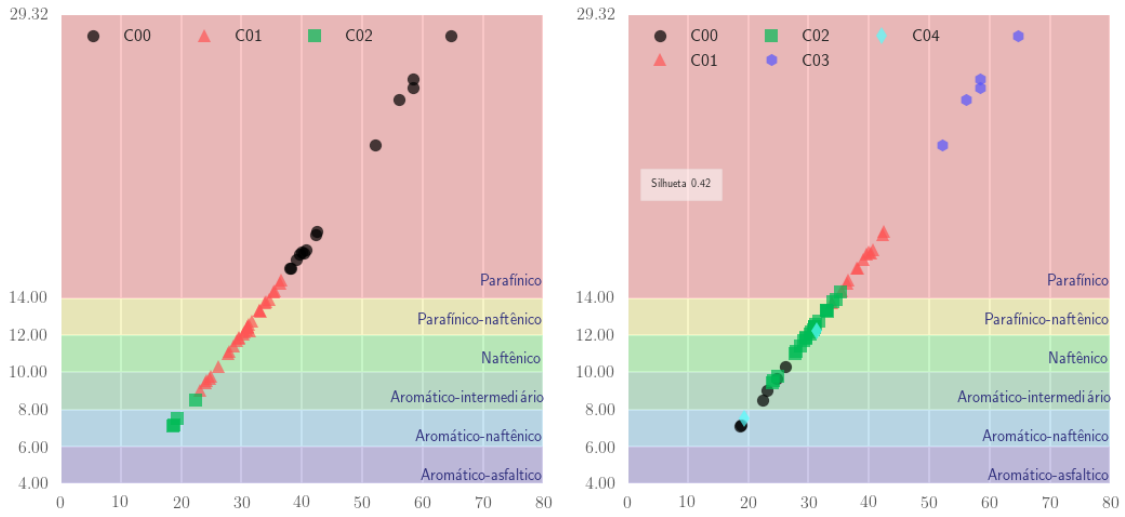
Na segunda extensão, é verificado como os modelos obtidos até aqui, obtidos a

partir de uma base de apenas 49 petróleos, se comportam quando usados com uma amostragem maior de petróleos.

5.6.1 Comparação com o Índice de Farah

Como exposto na Seção 2.3.3, Farah (2006) propôs um índice que relaciona as classes Aromático-asfáltico, Aromático-naftênico, Aromático-intermediário Naftênico, Parafínico-naftênico e Parafínico a razão entre a medida de densidade. Assim como a classificação feita a partir da densidade, esse índice provê uma maneira simples de classificar o óleo bruto com poucos parâmetros. Mas torna-se mais completa, pois leva em consideração a informação referente a viscosidade de determinada amostra.

A Figura 5.26a apresenta os 49 petróleos projetados no plano Índice de Farah × Densidade. Cada amostra está rotulada pelo cluster que foi encontrado a partir do modelo de NMF utilizando a divergência de Kullback-Leibler. Esse modelo foi escolhido por ser o melhor resultado e por facilitar a interpretação sobre a influência das variáveis



(a) Modelo NMF com três clusters (b) Modelo ISOMAP com cinco clusters

Figura 5.26: Amostras de petróleo representadas em um gráfico, onde a densidade é projetada no eixo x e o índice de Farah no eixo y. As classes típicas segundo essa medida estão destacadas, e a classificação obtida em relação ao modelo do ISOMAP para 5 clusters com as cores diferentes.

Pode-se observar que o cluster $C0$ está todo contido e se estende por toda a classe dos óleos Parafínicos. Essa classe é definida por óleos com baixa densidade

e viscosidade. Ao contrário da classificação por densidade, onde esse cluster se localiza em duas classes, aqui ele caracteriza bem essa classe. O cluster $C1$ está localizado por toda a extensão das classes Parafínico-naftênico e Naftênico, com alguma interseção com a classe anterior e com a posterior, Aromático-Intermediário. São classes com densidade um pouco maior. Como vimos na análise com NMF, a propriedades ligadas a volatilidade tem e ponto de ebulição tem bastante influência nesse cluster. As propriedades Asfaltenos e Enxofre influenciam nesse cluster. Já o cluster $C2$ mostra petróleos na classe aromática-naftênica, óleos com bastante enxofre e asfaltenos. São petróleos bastante viscosos.

Já a Figura 5.26b apresenta uma classificação realizada com ISOMAP com 5 clusters. Revisitando a Figura 5.7 verificamos que o desempenho desse modelo, medido através do índice de silhueta, foi satisfatório quando comparados com outros modelos (mesmo com três clusters), totalizando $0,42$ de silhueta média.

Ao contrário da comparação com três clusters, nenhum agrupamento pareceu descrever inteiramente uma classe. As propriedades $C01$ e $C03$ caracterizam petróleos parafínicos. O restante dos clusters estão difusos entre as classes, sem uma clara distinção. É possível, e até esperado, que modelos com mais clusters fiquem degradados devido ao baixo número de amostras que foram usados, tornando-se, assim, uma limitação dos modelos obtidos.

5.6.2 Extrapolação para bases maiores

Os modelos foram criados a partir de uma base pública composta por 49 petróleos. Uma questão a ser entendida é a capacidade de extrapolação dos modelos em bases maiores. Com acesso a base de homologação da PETROBRAS, foi possível projetar 4359 óleos brutos no modelo obtido com preprocessamento de PCA e SOM e clusterização K-médias. Por questões de sigilo comercial, a análise completa não será exposta. Contudo, na Tabela 5.3, são mostrados os resultados parciais.

Pode-se notar que existem diferenças, porém a base de homologação possui aproximadamente 100 vezes mais dados. Devido a diferença no volume de informações, é possível afirmar que o modelo possui valores similares para as duas bases.

No que se refere às figuras de mérito, houve uma deterioração no índice de silhueta, principalmente, devido a saturação do mapa SOM que contou com apenas

Base	Leve-Médio	Médio-Leve	Médio-Pesado	Pesado-Médio	Silhueta	CVNN
49	4,6%	10,4%	29,5%	0,0%	0,44	0,44
4359	0,91%	20,1%	24,0%	0,0%	0,34	0,48

Tabela 5.3: Comparação da execução do Algoritmo SOM na base pública e base de homologação da PETROBRAS.

36 neurônios. O cluster que mais degradou o índice foi o $C2$, que na base pública, possuía apenas 3 amostras e portanto não está bem representada.

Capítulo 6

Conclusões

Esta tese de doutorado visou o desenvolvimento de um método baseado em técnicas de mineração de dados cegas e de estatísticas de ordem superior para caracterizar amostras de petróleo de acordo com suas propriedades físico-químicas.

Atualmente, a indústria utiliza métodos empíricos através da utilização de limites determinísticos de uma ou duas variáveis. Enquanto eficientes e simples, tais métodos podem limitar a abrangência da classificação. Ao descartar certas propriedades, perdem-se informações que podem ser úteis quando analisadas na relação entre diferentes amostras. Assim, nuanças são desconsideradas, prejudicando não só a modelagem obtida, como sua generalização.

As técnicas de seleção de características, projeção dos dados e clusterização utilizadas, permitiram acessar as informações intrínsecas do espaço amostral. Ao realizar a exploração dos dados, compreendeu-se a influência das propriedades na maneira que as amostras de petróleo se correlacionam. A densidade e a viscosidade são realmente propriedades marcantes e importantes para a caracterização, mas outras variáveis atuam, principalmente nas amostras com valores de densidade próximas aos limites estabelecidos pelos métodos típicos. Os modelos de clusterização aplicados foram escolhidos por abordarem a tarefa sob aspectos diferentes. Foram estabelecidas técnicas de compactação de dados, a escolha da função de otimização focada no desempenho de clusterização e a utilização de diferentes divergências. Apesar de tais métodos ocasionarem pequenas diferenças nos resultados obtidos, evidenciaram a existência de tais nuanças. Pelo menos 50% das amostras migraram para grupos diferentes dos que estavam acomodados quando apenas a densidade é considerada.

Aplicando o algoritmo ISOMAP mais amostras migram de uma classificação para outra, aumentando essa diferença para até 71 % das amostras.

Em relação às figuras de mérito, o melhor desempenho obtido foi alcançado com a NMF, levando em conta o uso da divergência de Kullback-Leibler. Esse algoritmo ainda permite a interpretação dos coeficientes do espaço latente como uma ponderação, devido a sua natureza não-negativa e como uma relação de influência das propriedades frente aos clusters obtidos.

Empresas petrolíferas acessam milhares de amostras de óleo bruto em sua operação. As nuanças encontradas na classificações pelo método proposto podem afetar a logística e a cadeia de produção e ocasionar, por fim, a otimização de processos e mitigação de despesas. Essas empresas podem, a partir desses resultados, avaliar o impacto de adotar a classificação proposta em suas operações.

6.1 Trabalhos Futuros

Prevê-se como desdobramento direto dessa pesquisa, a aplicação do método apresentado em uma base de dados maior, a fim de obter modelos mais robustos. Como foi mostrado, ao comparar com classificações mais segmentadas, a clusterização ficou limitada em desempenho, degradando a qualidade dos modelos.

Contudo, foi mostrado que o modelo, obtido com 49 petróleos, comportou-se de maneira similar ao categorizar uma base cem vezes maior. Essa informação permite vislumbrar a criação de um modelo completo utilizando, para isso, apenas um recorte dos dados. Será necessário estabelecer uma maneira de selecionar os óleos brutos mais importantes para tal tarefa.

Em relação aos algoritmos, NMF, e sua utilização para clusterização, está cada vez mais popular. Diversos implementações surgem toda hora, como novos funções de atualização e interpretações, que podem impactar positivamente os resultados.

A motivação por algoritmos baseados na natureza se mantêm como uma perspectiva interessante. A utilização deles combinados com funções de custo mais eficientes, e até com a incorporação de conhecimento especialista, é uma área de desenvolvimento que pode representar ganhos nos resultados.

Por fim, a criação de um *framework* para a elaboração de misturas de óleo bruto,

baseado na utilização desta categorização, com a intensão de impactar na operação de refinarias e plantas industriais.

Referências Bibliográficas

[1] ABOU-SAYED, A., OTHERS. "Data mining applications in the oil and gas industry", *Journal of Petroleum Technology*, v. 64, n. 10, pp. 88–95, 2012.

[2] WORLD ENERGY COUNCIL. "World Energy Resources 2016 - Executive Summary". `http://www.wec-france.org/DocumentsPDF/Etudes_CME/2016-WER-Synthese-ENG.pdf`, 2017. Available at December 2017.

[3] INTERNATIONAL ENERGY AGENCY. *Oil Market Report*. Relatório técnico, Fevereiro 2018.

[4] ORGANIZAÇÃO DE COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO. "Crude oil production Statistics Home". `https://data.oecd.org/energy/crude-oil-production.htm`, 2018. disponível em Março de 2018.

[5] RICCOMINI, C., SANT, L. G., TASSINARI, C. C. G., et al. "Pré-sal: geologia e exploração", *Revista USP*, , n. 95, pp. 33–42, 2012.

[6] FARAH, M. *Petróleo e seus derivados*. LTC, 2012. ISBN: 9788521620525.

[7] PETROBRAS. *Plano de Negócios e Gestão 2018-2022.* Relatório técnico, Novembro 2017.

[8] "Terminology Relating to Petroleum, Petroleum Products, and Lubricants". 2010. Disponível em: <`https://doi.org/10.1520/d4175-09a`>.

[9] THOMAS, J., TRIGGIA, A., CORREIA, C., et al. *Fundamentos de engenharia de petróleo*. PETROBRAS e Interciência, 2004.

[10] FARAH, M. *Caracterização de Frações de petróleo pela viscosidade. 2006, 271f.* Tese de Doutorado, Tese, 2006.

[11] ZILIO, E., U., P. "Identificação e distribuição dos principais grupos de compostos presentes nos petróleos brasileiros", *Bol. Téc. PETROBRAS*, v. 45, n. 1, pp. 21–25, 2002.

[12] SPEIGHT, J. G. *The chemistry and technology of petroleum*. CRC press, 2014.

[13] RIAZI, M. "Characterization and properties of petroleum fractions", 2005.

[14] XAVIER, G. M., MIZUTANI, F. T., DE MELLO, L. F., et al. "Avaliação de correlações para estimativa de teores de hidrocarbonetos pna em frações de petróleo", *4º Congresso Brasileiro de Pesquisa e Desenvolvimento em Petróleo e Gás*.

[15] ASTM INTERNATIONAL. "ASTM International Home". `https://www.astm.org/`, 2017.

[16] ENERGY INSTITUTE. "Petroleum Institute Home". `https://www.energyinst.org/home`, 2017.

[17] CARLISLE, B. "Crude Oil Assay For the Non - Technical". 2017. Available at December 2017.

[18] SPEIGHT, J. G. "Crude Oil – Assay". In: *Rules of Thumb for Petroleum Engineers*, John Wiley & Sons, Inc., pp. 179–180, 2017.

[19] ANDRADE, B. F., MENEZES, B. C., SAKAI, P. N., et al. "Utilização de viscosidade, densidade e o unifac-visco para a caracterização de frações de petróleo", *4º Congresso Brasileiro de Pesquisa e Desenvolvimento em Petróleo e Gás*.

[20] ERNEST, L., JAMES, J., MORAN, D., et al. "Measurement problems in the instrument and laboratory apparatus fields, in systems of units", *National and International Aspects, American Association for the Advancement of Science, Washington, DC*, 1959.

[21] WAUQUIER, J.-P. *Petroleum Refining: Crude oil, petroleum products, process flowsheets*, v. 1. Éditions Technip, 1995.

[22] ÉIGENSON, A. "Correlation of viscosity, density, and chemical characteristics of crude oils and petroleum products", *Chemistry and Technology of Fuels and Oils*, v. 25, n. 2, pp. 104–109, 1989.

[23] HITES, R. A. "Gas chromatography mass spectrometry", *Handbook of instrumental techniques for analytical chemistry*, pp. 609–626, 1997.

[24] FONSECA, A. M., BISCAYA, J. L., AIRES-DE SOUSA, J., et al. "Geographical classification of crude oils by Kohonen self-organizing maps", *Analytica Chimica Acta*, v. 556, n. 2, pp. 374–382, 2006. ISSN: 00032670. doi: 10.1016/j.aca.2005.09.062.

[25] JANEIRO, J., MARTINS, F., RELVAS, P. "Towards the development of an operational tool for oil spills management in the Algarve coast", *Journal of coastal conservation*, v. 16, n. 4, pp. 449–460, 2012.

[26] WRANG, P. "European Crude Oil Identification System (Project EUROCRUDE)", 1995.

[27] FERNÁNDEZ-VARELA, R., ANDRADE, J. M., MUNIATEGUI, S., et al. "Identification of fuel samples from the Prestige wreckage by pattern recognition methods", *Marine Pollution Bulletin*, 2008. ISSN: 0025326X. doi: 10.1016/j.marpolbul.2007.10.025.

[28] FERNÁNDEZ-VARELA, R., GÓMEZ-CARRACEDO, M., BALLABIO, D., et al. "The use of diagnostic ratios, biomarkers and 3-way Kohonen neural networks to monitor the temporal evolution of oil spills", *Marine pollution bulletin*, v. 96, n. 1-2, pp. 313–320, 2015.

[29] ALBAIGÉS, J., MORALES-NIN, B., VILAS, F. "The Prestige oil spill: a scientific response". 2006.

[30] BORGES, C., GÓMEZ-CARRACEDO, M. P., ANDRADE, J. M., et al. "Geographical classification of weathered crude oil samples with unsupervised self-organizing maps and a consensus criterion", *Chemometrics and Intelligent Laboratory Systems*, v. 101, n. 1, pp. 43–55, 2010. ISSN: 01697439. doi: 10.1016/j.chemolab.2010.01.001.

[31] CARVALHO ROCHA, W. F., SCHANTZ, M. M., SHEEN, D. A., et al. "Unsupervised classification of petroleum Certified Reference Materials and other fuels by chemometric analysis of gas chromatography-mass spectrometry data", *Fuel*, v. 197, pp. 248–258, 2017.

[32] CHAMKALANI, A. "A novel technique for screening of asphaltene deposition by the pattern recognition method", *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, v. 38, n. 3, pp. 450–457, 2016.

[33] ZHANG, J.-L., ZHANG, Y.-J., ZHANG, L. "A novel hybrid method for crude oil price forecasting", *Energy Economics*, v. 49, pp. 649–659, 2015.

[34] MURUGAN, J. S., PARTHASARATHY, V. "An automatic pattern matching approach for oil spill detection in SAR images", *Advances in Environmental Biology*, v. 10, n. 11, pp. 143–151, 2016.

[35] LI, Y., CUI, C., LIU, Z., et al. "Detection and monitoring of oil spills using moderate/high-resolution remote sensing images", *Archives of environmental contamination and toxicology*, v. 73, n. 1, pp. 154–169, 2017.

[36] ZHANG, K., SHEN, C., WANG, H., et al. "Research on Marine Oil Spill Pollution Detection Based on Image Recognition Algorithm", *DEStech Transactions on Computer Science and Engineering*, , n. iceiti, 2017.

[37] ENGLANDER, J., BRODRICK, P., BRANDT, A. "Monitoring Oilfield Operations and GHG Emissions Sources Using Object-based Image Analysis of High Resolution Spatial Imagery". In: *AGU Fall Meeting Abstracts*, 2015.

[38] Kumar, V. (Ed.). *Data Clustering Algorithms and Applications*. Minneapolis, Minnesota, E.U.A, Taylor & Francis Group, LLC, 2014.

[39] XIONG, H., LI, Z. "Clustering Validation Measures", *Aggarwal, C. C., & Reddy, C. K. (Eds.). (2014). Data Clustering: Algorithms and Applications. Boca Raton, FL: CRC.*, v. 43, n. 3, pp. 571–605, 2014.

[40] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. "Unsupervised learning". In: *The elements of statistical learning*, Springer, pp. 485–585, 2009.

[41] HANCER, E., KARABOGA, D. "A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number", *Swarm and Evolutionary Computation*, v. 32, pp. 49–67, 2017.

[42] JOLLIFFE, I. *Principal component analysis*. Wiley Online Library, 2002.

[43] HYVÄRINEN, A., KARHUNEN, J., OJA, E. *Independent component analysis*, v. 46. John Wiley & Sons, 2004.

[44] CICHOCKI, A., ANH-HUY, P. "Fast local algorithms for large scale nonnegative matrix and tensor factorizations", *IEICE transactions on fundamentals of electronics, communications and computer sciences*, v. 92, n. 3, pp. 708–721, 2009.

[45] HUANG, K., SIDIROPOULOS, N., SWAMI, A. "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition", *Signal Processing, IEEE Transactions on*, v. 62, n. 1, pp. 211–224, 2014.

[46] GRAEL, F. *ATLASom: Processamento de Texto para a Gerência de uma Colaboração Científica de Grande Porte*. Tese de Mestrado, COPPE/UFRJ, Rio de Janeiro, 2013.

[47] TENENBAUM, J. B., DE SILVA, V., LANGFORD, J. C. "A Global Geometric Framework for Nonlinear Dimensionality Reduction", *Science*, v. 290, n. 5500, pp. 2319–2323, dec 2000.

[48] BESSON, G., LOVRIC, M., MIN-OO, M., et al. *Riemannian geometry*, v. 4. American Mathematical Soc., 1996.

[49] GUPTA, M. D., XIAO, J. "Non-negative matrix factorization as a feature selection tool for maximum margin classifiers". In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 2841–2848. IEEE, 2011.

[50] VAVASIS, S. A. "On the complexity of nonnegative matrix factorization", *SIAM Journal on Optimization*, v. 20, n. 3, pp. 1364–1377, 2009.

[51] PAUCA, V. P., SHAHNAZ, F., BERRY, M. W., et al. "Text mining using non-negative matrix factorizations". In: *Proceedings of the 2004 SIAM International Conference on Data Mining*, pp. 452–456. SIAM, 2004.

[52] HOYER, P. O. "Non-negative matrix factorization with sparseness constraints", *Journal of machine learning research*, v. 5, n. Nov, pp. 1457–1469, 2004.

[53] CICHOCKI, A., ZDUNEK, R., AMARI, S.-I. "New algorithms for non-negative matrix factorization in applications to blind source separation". In: *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, v. 5, pp. V–V. IEEE, 2006.

[54] FÉVOTTE, C., VINCENT, E., OZEROV, A. "Single-channel audio source separation with NMF: divergences, constraints and algorithms". 2017.

[55] YANG, L., CAO, X., JIN, D., et al. "A unified semi-supervised community detection framework using latent space graph regularization", *IEEE transactions on cybernetics*, v. 45, n. 11, pp. 2585–2598, 2015.

[56] LUO, X., ZHOU, M., XIA, Y., et al. "An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems", *IEEE Transactions on Industrial Informatics*, v. 10, n. 2, pp. 1273–1284, 2014.

[57] WANG, D., NIE, F., HUANG, H. "Fast robust non-negative matrix factorization for large-scale data clustering". In: *25th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2104–2110.

[58] GILLIS, N. "The why and how of nonnegative matrix factorization", *Regularization, Optimization, Kernels, and Support Vector Machines*, v. 12, n. 257, 2014.

[59] YANG, Z., ZHANG, H., YUAN, Z., et al. "Kullback-Leibler divergence for nonnegative matrix factorization". In: *International Conference on Artificial Neural Networks*, pp. 250–257. Springer, 2011.

[60] FÉVOTTE, C., BERTIN, N., DURRIEU, J.-L. "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis", *Neural computation*, v. 21, n. 3, pp. 793–830, 2009.

[61] FÉVOTTE, C. "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization". In: *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 1980–1983. IEEE, 2011.

[62] WANG, Y.-X., ZHANG, Y.-J. "Nonnegative matrix factorization: A comprehensive review", *IEEE Transactions on Knowledge and Data Engineering*, v. 25, n. 6, pp. 1336–1353, 2013.

[63] KANAGAL, B., SINDHWANI, V. "Rank selection in low-rank matrix approximations: A study of cross-validation for NMFs". In: *Proc Conf Adv Neural Inf Process*, v. 1, pp. 10–15, 2010.

[64] LEE, D. D., SEUNG, H. S. "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems*, pp. 556–562, 2001.

[65] LIN, C.-J. "Projected gradient methods for nonnegative matrix factorization", *Neural computation*, v. 19, n. 10, pp. 2756–2779, 2007.

[66] SUN, Y., BABU, P., PALOMAR, D. P. "Majorization-minimization algorithms in signal processing, communications, and machine learning", *IEEE Transactions on Signal Processing*, v. 65, n. 3, pp. 794–816, 2017.

[67] HAN, J., HAN, L., NEUMANN, M., et al. "On the rate of convergence of the image space reconstruction algorithm", *Operators and matrices*, v. 3, n. 1, pp. 41–58, 2009.

[68] BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA, Springer-Verlag New York, Inc., 2006. ISBN: 0387310738.

[69] ITAKURA, F. "Minimum prediction residual principle applied to speech recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 23, n. 1, pp. 67–72, 1975.

[70] FÉVOTTE, C., IDIER, J. "Algorithms for Nonnegative Matrix Factorization with the β-Divergence", *Neural Computation*, v. 23, n. 9, pp. 2421–2456, 2011. ISSN: 0899-7667. doi: 10.1162/NECO_a_00168. Disponível em: <http://www.mitpressjournals.org/doi/abs/10.1162/NECO{_}a{_}00168>.

[71] BASU, A., HARRIS, I. R., HJORT, N. L., et al. "Robust and efficient estimation by minimising a density power divergence", *Biometrika*, v. 85, n. 3, pp. 549–559, 1998.

[72] EGUCHI, S., KANO, Y. "Robustifing maximum likelihood estimation by psi-divergence", *ISM Research Memorandam*, v. 802, 2001.

[73] PAUCA, V. P., PIPER, J., PLEMMONS, R. J. "Nonnegative matrix factorization for spectral data analysis", *Linear Algebra and its Applications*, v. 416, n. 1, pp. 29–47, jul 2006. doi: 10.1016/j.laa.2005.06.025. Disponível em: <https://doi.org/10.1016/j.laa.2005.06.025>.

[74] LUO, M., NIE, F., CHANG, X., et al. "Probabilistic Non-Negative Matrix Factorization and Its Robust Extensions for Topic Modeling." In: *AAAI*, pp. 2308–2314, 2017.

[75] VESANTO, J., ALHONIEMI, E. "Clustering of the self-organizing map", *Neural Networks, IEEE Transactions on*, v. 11, n. 3, pp. 586–600, 2000.

[76] KOHONEN, T. "The self-organizing map", *Proceedings of the IEEE*, v. 78, n. 9, pp. 1464–1480, 1990.

[77] SIMAS FILHO, E. *Análise Não-Linear de Componentes Independentes para uma Filtragem Online Baseada em Calorimetria de Alta Energia e com Fina Segmentação*. Tese de Doutorado, COPPE/UFRJ, Rio de Janeiro, 2010.

[78] HAYKIN, S. S., HAYKIN, S. S., HAYKIN, S. S., et al. *Neural networks and learning machines*, v. 3. Pearson Education Upper Saddle River, 2009.

[79] ULTSCH, A. *Self-Organizing Neural Networks for Visualization and Classification*. London, UK, Springer, 1993.

[80] CASCÃO, V. *Modelos de Inteligência Computacional para apoio à triagem de pacientes e diagnótico clínico de tuberculose pulmonar*. Tese de Mestrado, COPPE/UFRJ, Rio de Janeiro, 2011.

[81] TAN, P.-N., STEINBACH, M., KUMAR, V. "Data mining cluster analysis: Basic concepts and algorithms". 2013.

[82] MÜLLNER, D. "fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python", *Journal of Statistical Software*, v. 53, n. 9, pp. 1–18, 2013.

[83] STUETZLE, W., NUGENT, R. "A generalized single linkage method for estimating the cluster tree of a density", *Journal of Computational and Graphical Statistics*, 2012.

[84] BERKHIN, P. "Grouping Multidimensional Data: Recent Advances in Clustering". cap. A Survey of Clustering Data Mining Techniques, pp. 25–71, Berlin, Heidelberg, Springer Berlin Heidelberg, 2006.

[85] LI, T., DING, C. H. "Nonnegative Matrix Factorizations for Clustering: A Survey." 2013.

[86] DING, C., HE, X. "K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization". In: *Proceedings of the 2004 ACM symposium on Applied computing*, pp. 584–589. ACM, 2004.

[87] YOKOI, T. "Topic Number Estimation by Consensus Soft Clustering with NMF". In: *International Conference on Future Generation Information Technology*, pp. 63–73. Springer, 2010.

[88] BUCAK, S. S., GUNSEL, B. "Online video scene clustering by competitive incremental NMF", *Signal, Image and Video Processing*, v. 7, n. 4, pp. 723–739, 2013.

[89] BLUMENSATH, T. "Sparse matrix decompositions for clustering". In: *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pp. 1163–1167. IEEE, 2014.

[90] LEE, J.-H., PARK, S. "Document Clustering using Term reweighting based on NMF", *Journal of the Korea Society of Computer and Information*, v. 13, n. 4, pp. 11–18, 2008.

[91] GREENE, D., CAGNEY, G., KROGAN, N., et al. "Ensemble non-negative matrix factorization methods for clustering protein–protein interactions", *Bioinformatics*, v. 24, n. 15, pp. 1722–1728, 2008.

[92] ARORA, S., GE, R., MOITRA, A. "Learning topic models–going beyond SVD". In: *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pp. 1–10. IEEE, 2012.

[93] KUANG, D., CHOO, J., PARK, H. "Nonnegative matrix factorization for interactive topic modeling and document clustering". In: *Partitional Clustering Algorithms*, Springer, pp. 215–243, 2015.

[94] BELFORD, M., MAC NAMEE, B., GREENE, D. "Stability of topic modeling via matrix factorization", *Expert Systems with Applications*, v. 91, pp. 159–169, 2018.

[95] AIZAWA, A. "An information-theoretic perspective of tf–idf measures", *Information Processing & Management*, v. 39, n. 1, pp. 45–65, 2003.

[96] GUILLE, A., SORIANO-MORALES, E.-P. "TOM: A library for topic modeling and browsing." In: *EGC*, pp. 451–456, 2016.

[97] CICHOCKI, A., ZDUNEK, R., PHAN, A. H., et al. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[98] ABRAHAM, A., DAS, S., ROY, S. "Swarm intelligence algorithms for data clustering", *Soft Computing for Knowledge Discovery and Data Mining*, pp. 279–313, 2008.

[99] NANDA, S. J., PANDA, G. "A survey on nature inspired metaheuristic algorithms for partitional clustering", *Swarm and Evolutionary Computation*, v. 16, pp. 1–18, 2014.

[100] KAUR, P. "Applications of Swarm Intelligence in Data Clustering", v. 3, n. 4, pp. 85–95, 2015.

[101] DORIGO, M. "Optimization, learning and natural algorithms", *Ph. D. Thesis, Politecnico di Milano, Italy*, 1992.

[102] MOHAN, B. C., BASKARAN, R. "A survey: Ant Colony Optimization based recent research and implementation on several engineering domain", *Expert Systems with Applications*, v. 39, n. 4, pp. 4618–4627, 2012.

[103] DORIGO, M., BIRATTARI, M., ST, T. "Ant Colony Optimization", , n. November, pp. 96–99, 2006.

[104] JAFAR, O. A. M., SIVAKUMAR, R. "Ant-based Clustering Algorithms : A Brief Survey", *International Journal*, v. 2, n. 5, pp. 787–796, 2010.

[105] HANDL, J., MEYER, B. "Improved ant-based clustering and sorting in a document retrieval interface", *Parallel Problem Solving from Nature—PPSN VII*, pp. 913–923, 2002.

[106] AZZAG, H., MONMARCHE, N., SLIMANE, M., et al. "AntTree: A new model for clustering with artificial ants", *2003 Congress on Evolutionary Computation, CEC 2003 - Proceedings*, v. 4, pp. 2642–2647, 2003.

[107] HANDL, J., KNOWLES, J., DORIGO, M. "Ant-based clustering: a comparative study of its relative performance with respect to k-means, average link and 1d-som", *Design and application of hybrid ...*, , n. i, pp. 1–16, 2003.

[108] HANDL, J., MEYER, B. "Ant-based and swarm-based clustering", *Swarm Intelligence*, v. 1, n. 2, pp. 95–113, 2007.

[109] STÜTZLE, T., DORIGO, M. "ACO algorithms for the traveling salesman problem", *Evolutionary Algorithms in Engineering and Computer Science*, pp. 163–183, 1999.

[110] BRUGGEMAN, C.-M., LIU, W., MITCHELL, M., et al. "Creating Clusters : An Analysis of Traveling Salesman as a Clustering Mechanism", pp. 1–9, 2010.

[111] KEOGH, E., MUEEN, A. "Encyclopedia of Machine Learning". cap. Curse of Dimensionality, pp. 257–258, Boston, MA, Springer US, 2010.

[112] THERAULAZ, G., BONABEAU, E., SAUWENS, C., et al. "Model of droplet dynamics in the Argentine ant Linepithema humile (Mayr)", *Bulletin of mathematical biology*, v. 63, n. 6, pp. 1079–1093, 2001.

[113] YANG, X.-S. *Nature-inspired metaheuristic algorithms*. Luniver press, 2010.

[114] KENNEDY, J., EBERHART, R. "Particle Swarm Optimization", *Engineering and Technology*, pp. 1942–1948, 1995.

[115] KARABOGA, D., BASTURK, B. "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm", *Journal of Global Optimization*, v. 39, n. 3, pp. 459–471, 2007.

[116] FILHO, C. J. A. B., NETO, F. B. D. L., LINS, A. J. C. C., et al. "A novel search algorithm based on fish school behavior", *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, pp. 2646–2651, 2008.

[117] KENNEDY, JAMES. "Encyclopedia of Machine Learning". cap. Particle Swarm Optimization, pp. 760–766, Boston, MA, Springer US, 2010.

[118] FREITAS, A. A. *Data mining and knowledge discovery with evolutionary algorithms*. Springer Science & Business Media, 2013.

[119] CLERC, M., KENNEDY, J. "The particle swarm-explosion, stability, and convergence in a multidimensional complex space", *Evolutionary Computation, IEEE Transactions on*, v. 6, n. 1, pp. 58–73, 2002.

[120] ERIK, M., PEDERSEN, H., PEDERSEN, M. E. H. "Good parameters for particle swarm optimization", *Technical Report HL1001, Hvass Laboratories*, v. HL1001, pp. 1–12, 2010.

[121] MENDES, R. *Population topologies and their influence in particle swarm performance*. Tese de Doutorado, Universidade do Minho, 2004.

[122] VAN DER MERWE, D., ENGELBRECHT, A. "Data Clustering using Particle Swarm Optimization", *IEEE*, pp. 215–220, 2003.

[123] AHMADYFARD, A., MODARES, H. "Combining PSO and k-means to enhance data clustering". In: *Telecommunications, 2008. IST 2008. International Symposium on*, pp. 688–691. IEEE, 2008.

[124] KARTHI, R., ARUMUGAM, S., RAMESHKUMAR, K. "Comparative evaluation of Particle Swarm Optimization Algorithms for Data Clustering using real world data sets." v. 8, n. 1, 2008.

[125] EL-TARABILY M., ABDEL-KADER R., MARIE M. ABDEL-AZEEM G. "A PSO-Based Subtractive Data Clustering Algorithm", v. 3, n. 5, pp. 13–17, 2013.

[126] HU, G., CHEN, M.-Y., HE, W., et al. "Clustering-based particle swarm optimization for electrical impedance imaging". In: *Advances in Swarm Intelligence*, Springer, pp. 165–171, 2011.

[127] CUI, X., POTOK, T. E. "Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm", *Engineering*, pp. 185 – 191, 2005.

[128] CUI, X., POTOK, T., PALATHINGAL, P. "Document clustering using particle swarm optimization", *Intelligence Symposium*, pp. 185–191, 2005.

[129] KAO, Y.-T., ZAHARA, E., KAO, I.-W. "A hybridized approach to data clustering", *Expert Systems with Applications*, v. 34, n. 3, pp. 1754–1762, 2008.

[130] PARPINELLI, R., LOPES, H. "New inspirations in swarm intelligence: a survey", *International Journal of Bio-Inspired Computation*, v. 3, n. 1, pp. 1, 2011.

[131] KARABOGA, D., AKAY, B. "A survey: algorithms simulating bee swarm intelligence", *Artificial Intelligence Review*, v. 31, n. 1-4, pp. 61–85, 2009.

[132] ZHANG, C., OUYANG, D., NING, J. "An artificial bee colony approach for clustering", *Expert Systems with Applications*, v. 37, n. 7, pp. 4761–4767, 2010.

[133] KARABOGA, D., OZTURK, C. "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", *Applied Soft Computing*, v. 11, n. 1, pp. 652–657, 2011.

[134] BASTOS-FILHO, C. J. A., NASCIMENTO, D. O. "An enhanced fish school search algorithm", *Proceedings - 1st BRICS Countries Congress on Computational Intelligence, BRICS-CCI 2013*, , n. 2, pp. 152–157, 2013.

[135] ZHANG, B., HSU, M., DAYAL, U. "K-harmonic means-a data clustering algorithm", *Hewlett-Packard Labs Technical Report HPL-1999-124*, 1999.

[136] SERAPIÃO, A. B., CORRÊA, G. S., GONÇALVES, F. B., et al. "Combining K-Means and K-Harmonic with Fish School Search Algorithm for data clustering task on graphics processing units", *Applied Soft Computing*, v. 41, pp. 290–304, 2016.

[137] FÄRBER, I., GÜNNEMANN, S., KRIEGEL, H.-P., et al. "On Using Class-Labels in Evaluation of Clusterings", *Proc. 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings (MultiClust 2010) in conjunction with 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010), Washington, DC, USA*, p. 9, 2010.

[138] ROUSSEEUW, P. J. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, v. 20, pp. 53–65, 1987.

[139] DAVIES, D. L., BOULDIN, D. W. "A cluster separation measure." *IEEE transactions on pattern analysis and machine intelligence*, v. 1, n. 2, pp. 224–227, 1979.

[140] WANG, K., WANG, B., PENG, L. "CVAP: validation for cluster analyses", *Data Science Journal*, v. 8, pp. 88–93, 2009.

[141] PETROVIC, S. "A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters". In: *Proceedings of the 11th Nordic Workshop of Secure IT Systems*, pp. 53–64, 2006.

[142] ANDONI, A., INDYK, P. "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions". In: *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pp. 459–468. IEEE, 2006.

[143] VESANTO, J. "SOM-based data visualization methods", *Intelligent data analysis*, v. 3, n. 2, pp. 111–126, 1999.

[144] CHUANG, J., MANNING, C. D., HEER, J. "Termite: Visualization techniques for assessing textual topic models". In: *Proceedings of the international working conference on advanced visual interfaces*, pp. 74–77. ACM, 2012.

[145] GAUJOUX, R. "Generating heatmaps for Nonnegative Matrix Factorization", *R Foundation for Statistical Computing, Vienna, Austria*, 2014.

[146] MULLINS, O. C., SHEU, E. Y., HAMMAMI, A., et al. *Asphaltenes, heavy oils, and petroleomics*. Springer Science & Business Media, 2007.

[147] BARWISE, A. "Role of nickel and vanadium in petroleum classification", *Energy & Fuels*, v. 4, n. 6, pp. 647–652, 1990.

[148] KOHONEN, T. *Self-Organizing Maps*. Springer Berlin Heidelberg, 2001. doi: 10.1007/978-3-642-56927-2. Disponível em: <`https://doi.org/10.1007/978-3-642-56927-2`>.

[149] BALASUBRAMANIAN, M., SCHWARTZ, E. L. "The isomap algorithm and topological stability", *Science*, v. 295, n. 5552, pp. 7–7, 2002.

[150] ARTHUR, D., VASSILVITSKII, S. "k-means++: The advantages of careful seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[151] IZZO, D. "Pygmo and pykep: Open source tools for massively parallel optimization in astrodynamics (the case of interplanetary trajectory optimization)". In: *Proceedings of the Fifth International Conference on Astrodynamics Tools and Techniques, ICATT*, 2012.

[152] BLACKWELL, T. M., KENNEDY, J., POLI, R., et al. "Particle swarm optimization", *Swarm intelligence*, v. 1, n. 1, pp. 33–57, 2007.

[153] BANSAL, J. C., SINGH, P., SARASWAT, M., et al. "Inertia weight strategies in particle swarm optimization". In: *Nature and Biologically Inspired Computing (NaBIC), 2011 Third World Congress on*, pp. 633–640. IEEE, 2011.

[154] BRUNET, J.-P., TAMAYO, P., GOLUB, T. R., et al. "Metagenes and molecular pattern discovery using matrix factorization", *Proceedings of the national academy of sciences*, v. 101, n. 12, pp. 4164–4169, 2004.

[155] FARRIS, J. S. "On the cophenetic correlation coefficient", *Systematic Zoology*, v. 18, n. 3, pp. 279–285, 1969.

[156] MCCAIN, W. D. *The properties of petroleum fluids*. PennWell Books, 1990.

[157] SCOTT, D. W. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.