UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Ciência de dados para determinação da similaridade entre perfis de poços de petróleo: uma abordagem a partir de Mapas Auto-Organizáveis

Cleyton de Carvalho Carneiro

Trabalho de Conclusão de Curso - MBA em Ciência de Dados (CEMEAI)

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Ciência de dados para determinação da similaridade entre perfis de poços de petróleo: uma abordagem a partir de Mapas Auto-Organizáveis

Cleyton de Carvalho Carneiro

USP - São Carlos

2021

CLEYTON DE CARVALHO CARNEIRO

Ciência de dados para determinação da similaridade entre perfis de poços de petróleo: uma abordagem a partir de Mapas Auto-Organizáveis

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciência de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof(a). Dr(a). Afonso Paiva Neto

USP - São Carlos

2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

d278c

de Carvalho Carneiro, Cleyton
Ciência de dados para determinação da similaridade entre perfis de poços de petróleo: uma abordagem a partir de Mapas Auto-Organizáveis / Cleyton de Carvalho Carneiro; orientador Afonso Paiva Neto. -- São Carlos, 2021.
86 p.

Trabalho de conclusão de curso (MBA em Ciência de Dados) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Séries Auto-Organizáveis. 2. Perfilagem de Poços. 3. Similaridade. 4. Sincronização Temporal Dinâmica (DTW). I. Paiva Neto, Afonso, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

DEDICATÓRIA

Ao Mr. Stephen Fraser, pela amizade, e por me apresentar a técnica Self-Organizing Maps.

AGRADECIMENTOS

Agradeço, inicialmente, ao CEMEAI do ICMC da USP São Carlos pelo curso de excelência que nos ofereceu. Sou muito grato pela concessão da bolsa que me proporcionou participar do curso MBA em Ciência de Dados 2021 e, consequentemente, esta desenvolver esta pesquisa. Agradeço a todos os professores e colaboradores do MBA, pelo trabalho desenvolvido com seriedade e profissionalismo.

Agradeço ao Prof. Dr. Afonso Paiva Neto, por toda orientação e pelas relevantes sugestões ao desenvolvimento deste trabalho. Grato pelas conversas e amizade desenvolvidas no decorrer deste curso.

Ao orientando Lucas Blanes Oliveira, que me apresentou uma lacuna da indústria de petróleo. Lacuna esta que desencadeou no tema da pesquisa aqui apresentadas.

Ao doutorando Rodrigo Gouvêa, pelas discussões e pela colaboração científica. Estou certo de que este trabalho terá a continuidade que merece em sua pesquisa de pós-graduação.

A todos os meus orientandos e orientandas que, com seus trabalhos, estabelecem trocas importantes. Estas trocas, sem sombra de dúvidas, me trouxeram reflexões importante sobre o desenvolvimento desta pesquisa.

Ao Prof. Dr. Rafael Gioria, pelas discussões científicas e pela elegante sugestão do algoritmo que nos levou à obtenção da Matriz-P.

A todos da minha família, em especial à minha mãe, D. Cleide Carneiro, minha tia, Clélia Lauande, e irmãos, Ray, Cyntia, Cícero e Celyce, por todo o apoio que sempre me oferecem. Obrigado pelo doce e prazeroso convívio.

Agradeço ao companheiro, Dr. Paulo Afonso Mei, pela força e constante presença.

Aos meus amigos e familiares, pela compreensão quanto à minha ausência nos momentos em que necessitei me dedicar aos estudos deste curso.

Agradeço a Deus pela existência e pela consciência.

EPÍGRAFE

“Nós somos o caminho que escolhemos.
O caminho por onde passo guarda-me
e eu sou o caminho aonde passo e,
embora passe,
eu fico no caminho aonde passo
e vai o caminho comigo
caminho que fica
por onde passo.
E passa.
E passo.
Ficamos”

João de Jesus Paes Loureiro
(Poeta abaetetubense)

RESUMO

CARNEIRO, C. C. **Ciência de dados para determinação da similaridade entre perfis de poços de petróleo:** uma abordagem a partir de Mapas Auto-Organizáveis. 2021. 52 f. Trabalho de conclusão de curso (MBA em Ciência de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

A demanda pela compreensão da similaridade entre poços é recorrente para obtenção de informações estratégicas em pesquisas relacionadas aos reservatórios de petróleo. A formulação de uma medida de similaridade de poços, no entanto, demanda sistemática que compreende mineração e comparação desses poços em amplas bases de dados pré-existente. Os perfis de poços podem ser analisados como séries temporais, substituindo a variável de referência tempo pela profundidade. O crescente desenvolvimento dos algoritmos de ciência de dados e aprendizado de máquina viabilizou a disseminação das aplicações de auto-organização em diversas frentes de pesquisa. No entanto, a auto-organização de séries temporais ainda esbarra em limitações, sobretudo inerentes à topologia dos sistemas de organização. Em 2021 a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) disponibilizou gratuitamente para pesquisas todos os poços das bacias continentais brasileiras. A presente pesquisa visa desenvolver séries auto-organizadas a partir da suíte básica de perfis de poços de petróleo, buscando reduzir a dimensionalidade das séries e compará-las entre si quanto à similaridade. Uma nova abordagem metodológica foi desenvolvida com base no algoritmo *Self-Organizing Maps*, compreendendo a criação de uma matriz de posição, denominada Matriz-P, que foi analisada de forma conjunta à Matriz-D, obtida a partir da redistribuição dos valores de distância da Matriz-U. Além disso, uma nova matriz denominada Matriz-IDP foi desenvolvida com base em um índice que integra distância e posição. As novas matrizes geradas permitiram a reconstrução das séries temporais após o processo de auto-organização, guardando relações topológicas do espaço n-dimensional. Este processo foi aqui denominado de Séries Auto-Organizáveis, em inglês Self-Organizing Series (SOrS). Foram definidas, portanto, as séries baseadas na posição (P-SOrS), na distância (D-SOrS) e no índice de posição e distância (IDP-SOrS). As assinaturas foram comparadas e ordenadas pela similaridade de acordo com o custo de alinhamento não linear calculado pelo algoritmo de Sincronização Temporal Dinâmica, em inglês *Dynamic Time Warping* (DTW). Foram geradas SOrS para 30 poços de 5 bacias sedimentares continentais brasileiras. As séries obtidas para estes poços preservaram as características litoestratigráficas observadas nos perfis compostos, uma vez que as transições litoestratigráficas foram capturadas pelas comparações não lineares do algoritmo de DTW. Os dendrogramas hierárquicos aglomerativos mostraram que poços de uma mesma bacia sedimentar, majoritariamente, formaram grupos entre si. Os gráficos de ranqueamento de similaridade desenvolvidos para cada bacia sedimentar demonstraram similaridades estratigráficas evidentes, tais como nas bacias do Parnaíba e do Solimões. As SOrS, portanto, se mostraram alternativa viável e útil à sistemática da similaridade de poços, mantendo características intrínsecas dos perfis originais. A metodologia se mostra eficiente e escalável à aplicação em outras frentes de pesquisa, bem como para mineração das bases de dados de poços de petróleo na plataforma oceânica.

Palavras-chave: Séries Auto-Organizáveis; Perfilagem de Poços; Similaridade; Sincronização Temporal Dinâmica (DTW).

ABSTRACT

CARNEIRO, C. C. **Data science for determining similarity between oil well logs:** an approach using Self-Organizing Maps. 2021. 52 f. Trabalho de conclusão de curso (MBA em Ciência de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2021.

The demand for understanding the similarity between wells is recurrent to obtain strategic information in research related to oil and gas reservoirs. Calculating the similarity measurement of wells, however, requires an adequate system for mining and comparing these wells against extensive pre-existing databases. Well logs can be analyzed as time series, replacing the reference variable 'time' with 'depth'. The growing development of data science and machine learning algorithms has enabled the spread of self-organizing applications on several research fronts. However, the self-organization of time series still faces limitations, mainly inherent to the topology of the organization systems. In 2021, the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) made all wells in the Brazilian continental basins available free of charge. The present research aims to develop self-organized series from the basic suite of oil well profiles, seeking to reduce the dimensionality of the series and compare them with each other in terms of similarity. A new methodological approach will be presented based on the Self-Organizing Maps algorithm, with the creation of a position matrix, called P-Matrix, which was analyzed together with D-Matrix, obtained by redistributing the distance values from the U-Matrix. In addition, a new matrix called Matrix-IDP was developed based on an index that integrates distance and position. The new generated matrices allowed the reconstruction of the time series after the self-organization process, keeping topological relationships of the n-dimensional space. This process was named here as Self-Organizing Series (SOrS). Therefore, the series based on position (P-SOrS), distance (D-SOrS) and position and distance index (IDP-SOrS) were defined. The signatures were then compared and ordered by similarity according to the nonlinear alignment cost calculated by the Dynamic Time Warping (DTW) algorithm. SOrS were generated for 30 wells from 5 Brazilian continental sedimentary basins. The series obtained for these wells preserved the lithostratigraphic characteristics observed in the composite well log, since the lithostratigraphic transitions were captured by the nonlinear comparisons of the DTW algorithm. The agglomerative hierarchical dendrograms showed that wells from the same sedimentary basin, mostly, formed groups among themselves. The similarity ranking graphs developed for each sedimentary basin showed evident stratigraphic similarities, such as in the Parnaíba and Solimões basins. The SOrS, therefore, proved to be a viable and useful alternative to calculate the well similarity, maintaining intrinsic characteristics of the original well logs. The showed methodology proves to be efficient and scalable for application in other research fronts, as well as for mining the databases of oil and gas wells on the ocean platform.

Keywords: Self-Organizing Series (SOrS); well logs; similarity; Dynamic Time Warping (DTW).

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Apresentação

A ciência de dados tem se demonstrado como eficiente ferramenta integrativa em análise de dados com múltiplas variáveis, e consiste na interseção entre a ciência da computação, estatística e domínios de aplicação singulares (SKIENA, 2017). Tal ciência também foi descrita por (DINOV, 2018) como uma área emergente extremamente transdisciplinar, capaz de lidar com grandes quantidades de dados complexos, incongruentes e dinâmicos de várias fontes, e que visa desenvolver algoritmos, métodos, ferramentas e serviços capazes de receber tais conjuntos de dados e gerar modelos semiautomáticos de suporte a decisões. Em análises de propriedades que influenciam na interação entre rochas e fluidos, onde um extenso número de variáveis estão envolvidas, essa ciência apresenta um significativo potencial integrativo. A partir do ferramental da ciência de dados é possível observar o comportamento conjunto das variáveis, treinar modelos, classificar amostras e, inclusive, prever fenômenos e tendências em amplas bases de dados.

Dentre as inúmeras ferramentas investigativas que adquirem dados auxiliares à avaliação dos reservatórios de petróleo, a Perfilagem de Poços representa fonte de informações com enorme relevância. É a partir dos perfis de poços que são observadas propriedades físicas e químicas das rochas, que possibilitam interpretações acerca da porosidade, permeabilidade, argilosidade, mineralogia ou mesmo saturação por fluidos (Figura 1). Além disso, os perfis de poços representam escala intermediária entre as amostras de rochas e os levantamentos sísmicos. A amarração ou corregistro ente poço e sísmica torna possível o salto de escala das propriedades obtidas na escala do poço até a escala do reservatório.

É possível dividir as etapas de avaliação de um reservatório de hidrocarbonetos em duas fases distintas: exploração e produção. Durante a fase de exploração, o principal objetivo é identificar contextos geológicos e geofísicos em subsuperfície através de mapeamento geológico e/ou sísmica, perfurar poços exploratórios, e realizar uma avaliação preliminar do reservatório utilizando perfis de poços, análises laboratoriais realizadas em amostras de rocha e fluido, e testes de formação. Essa avaliação preliminar visa definir a comercialidade do reservatório, delimitar sua extensão e delinear o planejamento da fase de produção. A depender das dimensões do reservatório e sua complexidade, a fase de exploração pode durar diversos anos e necessitar de dezenas de poços com massivas campanhas de aquisição de dados, se

tornando excessivamente custosa. Após o conhecimento do reservatório atingir um grau de maturidade adequado, é iniciada a fase de produção. Durante essa fase, dezenas a centenas de poços produtores e injetores são perfurados no reservatório, formando uma malha irregular que visa produzir o hidrocarboneto da forma mais otimizada possível. Como a quantidade de poços é superior à da fase de exploração, a aquisição de dados é reduzida, tornando o projeto mais econômico. Entretanto, novos desafios relacionados à fase de produção podem demandar uma nova etapa de aquisição de dados para a atualização do modelo de reservatório, gerando custos não previstos.

Analisando as fases de exploração e produção, as empresas de óleo e gás possuem dois principais objetivos: (i) otimizar ao máximo a fase de exploração; e (ii) evitar a aquisição de novas informações durante a fase de produção. Otimizar a fase de exploração significa definir a economicidade de um reservatório com o mínimo de informações possível. Para isso, os dados sísmicos, de perfis de poços e de análises laboratoriais de amostras de rocha e fluido são comparados com dados de outros reservatórios na busca de análogos. O histórico de produção desses reservatórios análogos irá auxiliar na estimativa de economicidade do reservatório que está sendo explorado. Idealmente, a perfuração de apenas um poço seria suficiente para definir o melhor reservatório análogo, gerando o mínimo de custos. Evitar a aquisição de novas informações na fase de produção significa atualizar os modelos de reservatório sem a aquisição de dados sísmicos, perfis de poços e amostras. Como a não aquisição completa de novas informações é utópica, é esperado que apenas o mínimo de informações seja adquirido à medida que novos poços são perfurados. Caso novos poços apresentem novos desafios que demandem uma compreensão maior das características do poço, é desejado que esses desafios sejam resolvidos através da comparação desses poços com outros que possuem uma quantidade significativa de informações, sejam poços perfurados durante a fase de exploração do mesmo reservatório ou perfurados em outros reservatórios.

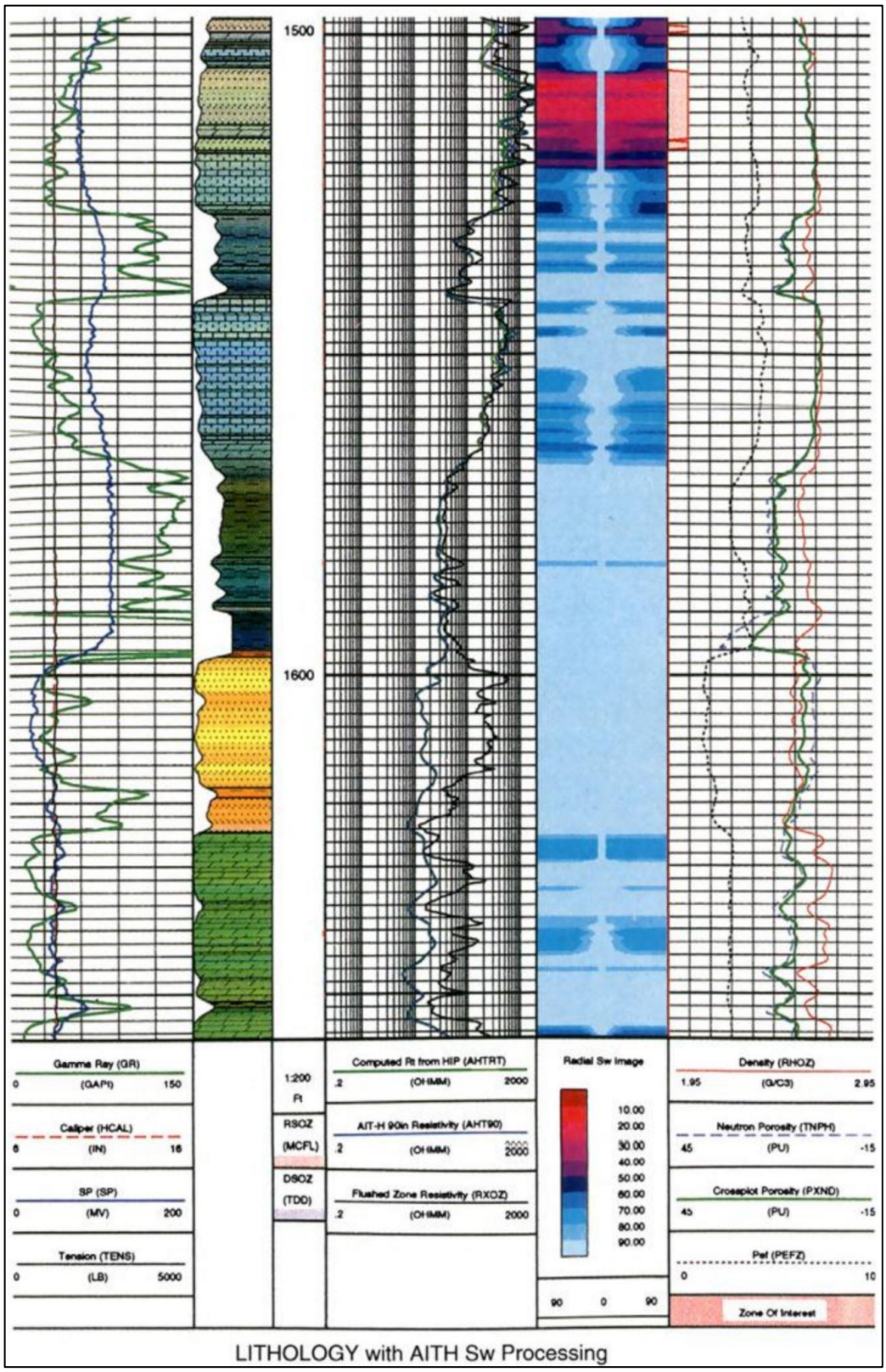
Nesse contexto, algoritmos de inteligência artificial vêm sendo empregados para suprir diversos desafios observados na avaliação de reservatórios, principalmente na fase de produção dos campos. Por necessitar de extensas bases de dados, esses algoritmos são treinados utilizando as informações adquiridas em poços perfurados na fase exploratória e no início da fase de produção, e aplicados nas avaliações dos dados sísmicos, perfis de poços, análises laboratoriais e testes de formação realizados posteriormente.

Com as demandas geradas pelas mudanças climáticas e a necessidade da indústria de óleo e gás de se tornar mais eficiente, as empresas vêm buscando utilizar cada vez mais técnicas

de inteligência artificial para aumentar a economicidade do gerenciamento dos seus reservatórios. No que diz respeito às fases de exploração e produção, se vislumbra que o aprendizado de máquina aplicado às extensas bases de dados existentes de centenas de campos produtores permita uma otimização sem precedentes na aquisição de informações. Entretanto, para que os melhores resultados sejam atingidos, é desejável que esses algoritmos sejam treinados utilizando dados representativos dos novos poços que são perfurados. Modelos de aprendizado de máquinas treinados indiscriminadamente podem gerar resultados pouco representativos dos poços perfurados, principalmente durante a fase exploratória.

Sendo assim, é necessário que seja criada uma etapa anterior à aplicação de algoritmos de aprendizado de máquina aos dados adquiridos em novos poços. Essa etapa consistiria na comparação dos dados adquiridos inicialmente em um novo poço a uma base de dados de poços perfurados em diversos reservatórios, obtendo-se assim uma medida de similaridade entre poços. De posse da medida de similaridade seria possível selecionar da base de dados os poços mais representativos e utilizá-los para treinar modelos que seriam aplicados ao novo poço. A depender do grau de similaridade observado, a aquisição de dados do novo poço poderia ser substituída pelos dados sintéticos, reduzindo custos e tornando as fases exploratórias e de produção mais eficientes. Essa medida de similaridade poderia ser construída, em um futuro, levando em consideração informações como os dados sísmicos e alguns perfis de poços, e o grau de similaridade poderia ser usado para suprimir a aquisição de perfis de poços de mais alto custo, amostras de rocha e fluido, e testes de formação, permitindo a antecipação de informações relativas à produção esperada para o poço utilizando apenas um pequeno conjunto de informações. A abrangência do cálculo da similaridade entre poços permitirá uma comparação inicial de diferentes poços. Os avanços nesse estudo poderiam gerar um aumento da resolução, permitindo que a comparação seja feita entre as diferentes rochas encontradas no poço.

Figura 1– Diferentes tipos de perfis de poços: (1) perfis de raios gama, potencial espontâneo e caliper na coluna esquerda auxiliam na identificação das unidades litológicas; (2) perfis de resistividade na coluna central auxiliam no cálculo da saturação de hidrocarbonetos; (3) perfis de densidade e porosidade na coluna da direita. Obs.: O Reservatório encontra-se na zona de topo da seção destacada em rosa na imagem



Fonte: (SELLEY; SONNENBERG, 2014)

1.2 Objetivos

Como objetivo principal, este trabalho se propõe a analisar a similaridade entre poços de petróleo a partir de uma suíte básica de perfis de poços, composta por cinco variáveis.

A pesquisa tem como objetivos específicos:

1. Desenvolver séries auto-organizáveis (SOrS) capazes de reduzir dimensionalidade da suíte básica de perfis. As análises partirão de uma adaptação inédita da técnica SOM para séries temporais, que obedecerá informações topológicas relacionadas a índices de distância e posição;
2. Analisar a similaridade entre as variáveis de distância e posição a partir de diferentes métricas baseadas na técnica Dynamic Time Warping (DTW), a saber: (i) SOrS-IDP, aplicada ao perfil síntese; (ii) vetorial, aplicada às variáveis de distância e posição com distância Euclidiana; e (iii) vetorial, aplicada às variáveis de distância e posição com distância Manhatan. A análise buscará compreender a métrica que obtêm melhores resultados em relação à similaridade entre os poços;
3. Analisar os agrupamentos obtidos pelas métricas de similaridade, bem como *rankings* de similaridade entre poços com relação a diferentes contextos geológicos e litoestratigráficos.

1.3 Justificativa

O desenvolvimento da medida de similaridade de poços a partir de uma ampla base de dados, a qual o setor de petróleo brasileiro já possui, poderá reverter em produtos tecnológicos capazes de reduzir custos e aumentar a eficácia interpretativa de modelos geológicos de reservatórios. Os custos operacionais de um levantamento de determinados perfis são extremamente elevados, sobretudo quando estes influenciam na interrupção da atividade de perfuração e/ou produção. Nesse sentido, o produto tecnológico capaz de gerar perfis sintéticos, fomentando comparativos para treinamento e aprendizado necessário ao aprendizado de máquina ou mesmo ao aprendizado profundo, é de grande relevância para a indústria que se utiliza de perfis geofísicos de poços para a caracterização e acompanhamento dos reservatórios de petróleo e gás.

Na fase anterior à perfuração de um poço de petróleo é necessária a caracterização dos reservatórios, ou seja, a determinação das propriedades geométricas e estruturais que condicionam um campo de petróleo. Neste processo é necessário combinar conhecimentos de geocientistas e de engenheiros, com o objetivo de construir um modelo onde se possa incorporar todas as informações e dados disponíveis sobre o reservatório. Esses modelos integrados são importantes para se prever, monitorar e otimizar a performance de um campo durante todo o seu ciclo de vida.

Dentre os mais importantes métodos de exploração e caracterização de reservatórios encontram-se a sísmica de reflexão (CORREA, 2003), além de dados de poços, incluindo perfis de raios gama, perfis elétricos (e.g. resistividade), densidade, geoquímico e etc. Os perfis de poços, medem propriedades físico-químicas das rochas, tais como abundância de elementos gamaespectrométricos, resistividade elétrica, densidade, índice de hidrogênio, velocidade de propagação de ondas sonoras etc (SCHLUMBERGER, 2009). Estes perfis são utilizados, entre outras aplicações, para interpretar os tipos de rochas perfuradas pelo poço. A calibração e inspeção desses perfis é feita a partir de descrições de amostras de rochas, realizadas pelo acompanhamento geológico do poço. A litologia interpretada dos poços, é um dado absolutamente fundamental para os trabalhos de exploração de hidrocarbonetos de uma bacia sedimentar. De posse de tais interpretações torna-se possível a elaboração de modelos estratigráficos, constituídos por rochas geradoras, reservatórios e selantes.

As diversas unidades litoestratigráficas que ocorrem nas bacias sedimentares apresentam respostas diferentes em termos das propriedades físico-químicas expostas acima, permitindo que se possa interpretar tais unidades com bases nos dados obtidos pela perfilagem de poços. O processo de interpretação desses dados, no entanto, normalmente é feito de forma visual por geólogos, estando, portanto, sujeito a subjetividades analíticas e interpretativas. Para se minimizar tais subjetividades, tem-se como alternativa a utilização de métodos analíticos semiautomáticos, tais como os propostos nesta pesquisa.

Embora os estudos da similaridade possuam amplas aplicações em diversas áreas de pesquisa, a literatura científica e tecnológica ainda possui uma lacuna em relação aos estudos de similaridade entre os poços de petróleo. A peculiaridade das informações obtidas pelos perfis de poços e a alta quantidade de informação disponível, inclusive publicamente, se somam à vasta aplicação destes perfis na indústria do petróleo. A ciência de dados e a inteligência artificial, a partir de suas abordagens, possibilitam a utilização da informação contida nos perfis de poços buscando, a partir destes dados, compreender relações de similaridade entre poços. A

similaridade de poços, por sua vez, poderá subsidiar a escolha de parâmetros de treinamento para aprendizado de máquina na construção de inúmeros modelos para a indústria do petróleo. Tais modelos orientam a tomada de decisão e minimizam custos operacionais.

2 REVISÃO BIBLIOGRÁFICA

2.1 Ciência de Dados e Inteligência Artificial

Provost e Fawcett (2013) definem ciência de dados como um conjunto de fundamentos capazes de suportar e orientar a extração de informação e conhecimento a partir dos dados. Tais autores reportam também que ciência de dados envolve muito mais do que o uso de algoritmos ou a aplicação da mineração de dados. Para os autores, a ciência de dados estaria baseada em diversos campos de estudos tradicionais, sustentada pela demanda da compreensão relacionada aos princípios fundamentais da análise causal. Os fundamentos da ciência de dados passariam, assim, pelo que tem sido estudado tradicionalmente pela estatística; por métodos de visualização; além das áreas específicas, onde a intuição, a criatividade, o bom senso e o conhecimento de uma aplicação específica devem ser exercitados.

Os trabalhos de (MCCARTHY et al., 1955), inicialmente, descreveram a Inteligência Artificial (IA), como uma área da computação que dispõe de algoritmos (sequência de regras e procedimentos), capazes de simular o raciocínio humano e, essencialmente, perceber informações do entorno, além de tomar decisões e ações que permitam melhorar o sucesso de uma tarefa desempenhada, tudo isso em um formato simplificado de como um sistema biológico, como por exemplo o cérebro, faria. (RUSSELL; NORVIG, 1995) relatam que os conceitos originais de Inteligência Artificial tiveram abordagens diversas ao longo das últimas décadas, mas sempre evocando a capacidade de desenvolver raciocínio ou mesmo ação baseado em aprendizado. O Quadro 1, traduzido de (RUSSELL; NORVIG, 1995), sintetiza as definições de IA em suas principais categorias.

Quadro 1- Definições originais de IA agrupadas em quatro categorias principais

<table>
<tr><td>"O novo e excitante esforço para fazer os computadores pensarem... máquinas com mentes, no sentido pleno e literal" ((KOSKELA et al., 1998)Haugeland, 1985)

"[A automação de] atividades que associamos ao pensamento humano, atividades como tomada de decisões, resolução de problemas, aprendizado ..." (Bellman, 1978)</td><td>"O estudo das faculdades mentais através do uso de modelos computacionais" (Charniak e McDermott, 1985)

"O estudo das computações que possibilitam perceber, raciocinar e agir" (Winston, 1992)</td></tr>
<tr><td>"A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas" (Kurzweil, 1990)

"O estudo de como fazer computadores fazer coisas em que, no momento, as pessoas são melhores" (Rich e Knight, 1991)</td><td>"Um campo de estudo que procura explicar e emular o comportamento inteligente em termos de processos computacionais" (Schalkoff, 1990)

"O ramo da ciência da computação que se preocupa com a automação do comportamento inteligente" (Luger e Stubblefield, 1993)</td></tr>
<tr><td colspan="2">Definições de AI acima, organizados em quatro categorias:

<table><tr><td>Sistemas que pensam como humanos.</td><td>Sistemas que pensam racionalmente.</td></tr><tr><td>Sistemas que agem como seres humanos.</td><td>Sistemas que agem racionalmente.</td></tr></table></td></tr>
</table>

Fonte: traduzido de (RUSSELL; NORVIG, 1995).

Diversos métodos baseados em ciência de dados e inteligência artificial possibilitam o aperfeiçoamento do processo de interpretação, a partir da atribuição de parâmetros sistemáticos e semiautomáticos. São exemplos de técnicas de inteligência artificial os algoritmos de aprendizado de máquina, capazes de prover treinamento e aprendizagem a partir de amplas bases de dados, com a finalidade de segmentar e gerar agrupamentos com base na similaridade entre padrões de amostras, ou mesmo predizer parâmetros em regiões espúrias ou com dados ausentes. Em termos de ciência de dados, os perfis de poços têm seus registros no formato de séries temporais multivariadas. Dentre as técnicas de aprendizado de máquina, as redes neurais artificiais, sejam estas supervisionadas ou não supervisionadas, demonstram amplas

possibilidades de aplicação na para cálculos de similaridade. Nos tópicos seguintes será desenvolvida uma revisão acerca dos perfis de poços enquanto séries temporais multivariadas, bem como das análises para redução da dimensionalidade em busca das medidas de similaridade entre os poços de petróleo.

2.2 Modelos de Bancos de Dados

A denominação ‘banco de dados’ é utilizada para se referir a uma coleção de dados organizada e armazenada que é acessada como parte de um sistema de computador. Ao ser projetado, o banco de dados deve levar em consideração tanto o bom entendimento da estrutura dos dados quanto a demanda do usuário final, que irá armazenar, localizar e gerenciar tais dados (ANGWIN; NELSOR; SYRETT, 1996).

Por muitos anos os bancos de dados operavam dominantemente a partir de bases de dados relacionais. Este formato, também conhecido como Linguagem de Consulta Estruturada ou, do inglês, *Structured Query Language* (SQL), modela os dados de forma que estes sejam percebidos pelos usuários na forma de tabelas. Nas últimas décadas, no entanto, devido ao aumento exponencial da quantidade de dados disponíveis com a ampliação da acessibilidade à internet e aos dispositivos móveis, gerou-se a necessidade de alternativas ao modelo SQL. Este modelo apresenta maior complexidade para gerenciar de forma efetiva grandes quantidade de dados. Como resultado, surgiram os bancos de dados não relacionais, ou não somente SQL, do inglês, *Not Only SQL* (NoSQL) (DAVE, 2012).

A estrutura NoSQL se baseia no modelo *Basically Available, Soft State and Eventually Consistent* (BASE) (JATIN; BATRA, 2016), que apresenta natureza distribuída. Tal modelo garante aos bancos de dados não relacionais escalabilidade, melhor performance e maiores níveis de disponibilidade aos seus usuários. Além disso, a estrutura NoSQL utiliza, de forma geral, menor espaço de armazenamento (FRACZEK; PLECHAWSKA-WOJCIK, 2017). Dentre os diversos tipos de bancos de dados NoSQL, destacam-se: (i) Banco de Documentos, que utiliza o formato *JavaScript Object Notation* (JSON); (ii) Bancos de Dados baseados na estrutura Chave-Valor; (iii) Armazenamento de dados em gráficos, através de nós e bordas; (iv) Armazenamento de Séries Temporais; (v) Armazenamento de Objetos; (vi) Armazenamento por Índice Externo; entre outros.

2.3 Redução de Dimensionalidade em Séries Temporais

As séries temporais são caracterizadas por observações coletadas sequencialmente ao longo do tempo (HAMILTON, 1994; PRIESTLEY, 1983). Quando as observações carregam informações sobre múltiplas variáveis ao longo do tempo, a representação diz respeito a uma série temporal multivariada (WEI, 2019). Podemos tratar dados de perfilagem de poços como séries temporais, onde o tempo é substituído pelas medidas de profundidade em relação a um dado referencial espacial de altitude. Dessa maneira, os diferentes perfis adquiridos pelas ferramentas de perfilagem de poços tratam de diferentes registros lidos em profundidades sequenciais.

Uma observação importante é que a perfilagem de poços, em sua fase de aquisição, atravessa sequências litoestratigráficas, obtendo seus respectivos registros em cada variável analisada. Esses registros, portanto, não deixam de ser relativos a uma série temporal geológica, uma vez que o tempo geológico é reconhecido a partir das rochas. Particularmente nas bacias sedimentares, o tempo geológico se dá em função das sequências deposicionais. Portanto, um perfil de poço pode ser considerado uma série temporal onde a variável ‘tempo’, caracterizada pela profundidade, assume escala do tempo geológico relativo às formações perfiladas. Vale ressaltar que o contexto estrutural, a partir de falhas, fraturas, dobras e/ou outros elementos, podem influenciar e/ou inverter a ordem de determinadas sequências.

Os registros multivariados em perfilagem de poços representam elevada dimensionalidade dado pelo número de variáveis envolvidas nos registros em profundidades. Essa alta dimensionalidade proporciona maior complexidade às aplicação envolvendo mineração de dados e/ou aprendizado de máquina. Isso ocorre devido a diversos desafios associados aos dados multivariados, tais como a ‘maldição da dimensionalidade’ e a significância da medida de similaridade no espaço multidimensional (CASSISI et al., 2012).

A redução da alta dimensionalidade consiste, portanto, em uma das abordagens capazes de proporcionar a análise de similaridade entre os perfis de poços. Para isso, uma das possibilidades para esta redução de dimensionalidade seria o desenvolvimento de vetores essenciais, capazes de proporcionar análises quanto à similaridade.

2.4 Mapas Auto-Organizáveis (SOM)

Os Mapas Auto-Organizáveis, ou do inglês *Self-Organizing Maps* (SOM) consistem em uma técnica baseada em aprendizado não-supervisionado que utiliza a quantização vetorial em

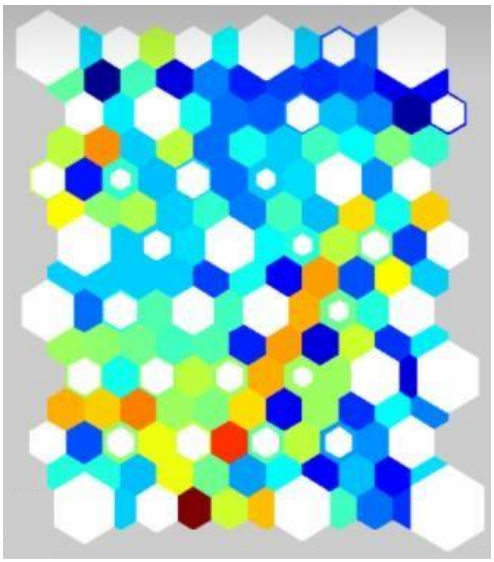
um espaço n-dimesional (KOHONEN, 2000). Os dados estimados pela técnica têm como base as distâncias entre os vetores disponíveis (FESSANT; MIDENET, 2002; KALTEH; HJORTH; BERNDTSSON, 2008).

A partir de etapas competitivas e colaborativas entre neurônios que disputam entre si para melhor representar a distribuição dos dados em um espaço n-dimensional, são eleitas unidades mais ajustadas ou, do inglês, *Best-Matched Units* (BMUs). As informações dos BMUs resultantes são projetadas em uma hipersuperfície que, posteriormente, pode ser planificada em mapas 2D. Relações de topologia e distância relacionadas às amostras originais são preservadas, portanto, nos BMUs. Isso faz com que a técnica seja capaz de representar nos mapas relações de similaridade entre as observações.

2.4.1 Matriz Unificada de Distância (Matriz-U)

Uma das representações mais difundida de um mapa auto-organizado é a Matriz de Distâncias Unificadas (Matriz-U) (Figura 2), do inglês, *Unified Distance Matrix (U-Matrix)* (KASKI; KOHONEN, 1996). Nesta matriz são visualizadas as distâncias entre os neurônios da camada de saída. A partir das distâncias calculadas, os neurônios adjacentes são apresentados em diferentes cores de acordo as relações de proximidade. Usualmente, cores mais quentes indicam uma maior distância, e consequentemente uma maior dissimilaridade entre os *hits*; e cores mais frias indicam uma menor distância ou maior similaridade. Assim, regiões com predominância de cores mais frias da Matriz-U podem ser interpretadas como agrupamento de dados, enquanto que as áreas com cores mais quentes como separadores destes agrupamentos.

Figura 2 - Matriz U com hits representados por hexágonos em branco, e distâncias entre BMUs pela intensidade da cor (cores frias – menor distância, cores quentes – maior distância)



Fonte: Fraser 2007

2.4.2 SOM Recorrente (RSOM)

Os SOM são notórios pela capacidade que possuem em reduzir dimensionalidade dos dados (HAYKIN, 2008; KOHONEN, 2001). Há, no entanto, algumas abordagens que tratam especificamente sobre a utilização de SOM recorrente em séries temporais (KOSKELA et al., 1998; MCQUEEN et al., 2004). Para (KOSKELA et al., 1998), esse tipo de arquitetura é capaz de produzir mapas topológicos contendo diferentes contextos temporais na base de dados. Os trabalhos de (MCQUEEN et al., 2004) mostram que os padrões identificados por RSOM preservam características dos dados de entrada, enquanto uma rede neural supervisionada, por exemplo, utiliza essas representações reduzidas para realizar um mapeamento para um conjunto correspondente de saídas desejadas.

2.5 Similaridade a partir de Vetor Síntese dos Perfis de Poços

A alta dimensionalidade em séries temporais torna a aplicação de métodos de mineração de dados ou de aprendizado de máquina menos eficientes. Isso ocorre, sobretudo, pela sobrecarga computacional em bases de dados muito extensas. Dessa maneira, a utilização de vetores essenciais, que sintetizam as variáveis em cada registro, passa a ser uma forma salutar e recorrente em problemas relacionados às séries temporais. Existem diversas maneiras de se resumir ou extrair a essência relacionada a um conjunto de variáveis observadas em uma mesma posição do tempo, ou do espaço, como passaremos a tratar a informação dos dados de perfilagem de poços. (KRAWCZAK; SZKATUŁA, 2014) tratam da redução de dimensionalidades em séries temporais com uma rotina que eles chamam de *Symbolic Essential Attributes Approximation* (SEAA). Os autores utilizaram redes neurais artificiais *feedforward* para gerar envelopes e produzir uma nova série temporal dada por vetores essenciais que preservavam as características originais dos dados.

Os trabalhos de (DU et al., 2020) abordam a medida de similaridade baseada em características morfológicas na análise de perfis de poços em arenitos. No entanto, o trabalho em questão se utiliza de análises em uma única variável de raios gamma, o que torna a comparação bem mais simples do que na abordagem multivariada.

A similaridade entre perfis de poços também foi descrita por Ali et al., (2021) em uma abordagem para a predição de dados faltantes em perfis sônicos com o uso da similaridade de perfis. A similaridade, neste caso, foi calculada com o uso de redes neurais profundas, do inglês, *Deep Neural Networks* (DNN) em conjunto com métricas das similaridades de Jaccard e Overlap.

A comparação de perfis de poços como vetores essenciais de séries temporais multivariadas torna viável, portanto, o desenvolvimento de análises quanto à similaridade de poços em diferentes contextos. O aprendizado de máquina e a auto-organização de mapas, assim, surge como uma das inúmeras ferramentas capazes de tratar a informação tanto na geração dos vetores essenciais como na análise quantitativa da similaridade.

A similaridade de séries-temporais é usada para medir o grau de similaridade entre subsequências ou a similaridade geral entre diferentes sequências (ESLING; AGON, 2012; MÜLLER, 2007). A determinação da similaridade em séries temporais envolve, portanto, a busca pela série do conjunto analisado que está mais próxima de uma série de referência. Para

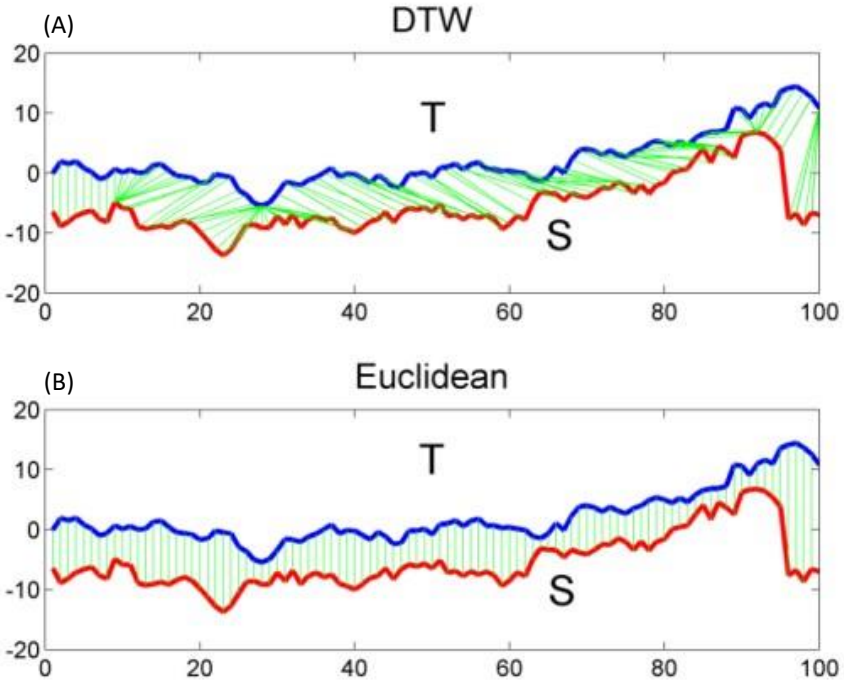
isso, como métrica de similaridade, é utilizada uma medida de distância como, por exemplo, a distância Minkowski ou a similaridade de cossenos (IRANI; PISE; PHATAK, 2016).

Dada a complexidade das series temporais em relação aos dados estáticos, métricas de convencionais de distância podem se mostrar ineficientes para lidar com o dinamismo dos dados. Assim, outras medidas de similaridades surgiram como alternativa, tais como: (i) Sincronização Temporal Dinâmica, do inglês, *Dynamic Time Warping* (DTW); (ii) Distância Editada; (iii) Sincronização Temporal de Distância Editada; (iv) Pulo Mínimo de Custo de Dissimilaridade; e (v) Medida de Correlação Cruzada (DU et al., 2020).

Apesar da distância euclidiana ser surpreendentemente competitiva com outras abordagens mais complexas, ela é pouco apropriada para algumas séries temporais. Como desvantagem para a similaridade a partir da distância euclidiana, destacam-se: a comparação demandar séries do mesmo tamanho; o fato de não lidar bem com ruídos ou *outliers*; e o fato de ser muito sensível às transformações do sinal, deslocamentos, escalamentos da amplitude, etc.

A técnica DTW (KEOGH; RATANAMAHATANA, 2005) traz mais sofisticação às medidas de similaridade, uma vez que permite a comparação de séries temporais de tamanhos diferentes, já que não se utiliza da comparação de pontos ‘um a um’, tal como faz a euclidiana. A técnica DTW se utiliza da comparação ‘muitos para um’, e vice-versa. Isso permite que a métrica reconheça formas similares de séries temporais, mesmo em presença de transformações de sinal, como deslocamentos ou mudança de escala (Figura 3).

Figura 3 - Comparação entre duas métricas de similaridade, a saber: (A) técnica DTW; e (B) distância euclidiana.



Fonte: Cassisi et al. (2012)

2.5.1 Mineração de Dados e Dynamic Time Warping (DTW)

Os conceitos de programação dinâmica e controle adaptativo se popularizaram na década de 1960 e deram origem ao algoritmo de sincronização temporal dinâmica, do inglês *Dynamic Time Warping* (DTW) (BELLMAN; KALABA, 1958). Esse algoritmo se disseminou, principalmente, por envolver aplicação direta em reconhecimento de fala (MYERS; RABINER; ROSENBERG, 1980; SAKOE; CHIBA, 1978), entre outras aplicações, como correspondência de língua escrita e assinatura (EFRAT; FAN; VENKATASUBRAMANIAN, 2007), além de visão computacional (MÜLLER, 2007).

Uma aplicação frequente de DTW está nos processos de mineração e agrupamento de dados em bases compostas por séries temporais. Esse conceito foi introduzido por trabalhos de mineração de dados (BERNDT; CLIFFORD, 1994) e, posteriormente, foi extensivamente utilizado em outros trabalhos (CORRADINI, 2001; KAHVECI; SINGH, 2001; KAHVECI; SINGH; GÜREL, 2002; LI et al., 2020; NIENNATTRAKUL; RATANAMAHATANA, 2007; RATANAMAHATANA; KEOGH, 2005).

Em (NIENNATTRAKUL; RATANAMAHATANA, 2007) foi apontada a diferença entre o uso do resultado do processo de DTW como distância e como métrica de distância. Para se tornar uma métrica, uma distância deve obedecer aos quatro axiomas: (i) simetria; (ii) identidade; (iii) não negatividade; e (iv) desigualdade triangular. A distância DTW não apresenta desigualdade triangular, o que torna impraticável a aplicação de algoritmos como *K-means* nos processos de mineração e classificação de dados, uma vez que as propriedades associativas não são mantidas e dependem da ordenação em que são calculadas. Nesse sentido, os agrupamentos em *k-medoids* e o agrupamento hierárquico aglomerativo são mais adequados e apresentam melhores resultados (KAHVECI; SINGH, 2001; NIENNATTRAKUL; RATANAMAHATANA, 2007; SAMMOUR et al., 2019).

Os trabalhos de (RATANAMAHATANA; KEOGH, 2005) mostram que a capacidade de comparação de séries temporais pelo algoritmo DTW é invariante à diferença de extensão das series temporais, sem a necessidade de interpolação para que as series tenham extensões parecidas. Por outro lado, os autores também apontaram que uma restrição do campo de procura, maior do que os tradicionais 10% da banda de Sakoe-Chiba (SAKOE; CHIBA, 1978), do caminho de alinhamento é fundamental para que algumas aplicações reais de mineração de dados tenham acurácia aceitável.

Segundo a definição formal de (SENIN, 2008), o algoritmo DTW visa encontrar a menor diferença entre duas séries temporais, dada pelo 'custo de alinhamento'. Esse custo é calculado a partir da distância entre um ou mais elementos de uma série para com um ou mais elementos de outra série.

Consideremos duas séries A e B de tamanho *m* e *n*, respectivamente $A = \{a_1, a_2, .., a_m\}$ e $B = \{b_1, b_2, \dots, b_n\}$. É construída, inicialmente, a matriz de distâncias, definidas por D_{base}, derivada das comparações por pares entre os dados das séries temporais A e B. Comumente se utiliza a distância euclidiana, tal como na Equação (1).

$$D_{base} = (A, B) = \ |A[i] - B[j]|, i \in [1{:}N], j \ \in [1{:}M] \quad (1)$$

Essa matriz é conhecida como matriz de custo local. O algoritmo então destaca as áreas de baixo custo da matriz, ou vales, definindo o 'caminho de alinhamento' (em inglês *Warping Path*). Esse caminho de alinhamento $(N \times M)$ é uma sequência $p = (p_1, \dots, p_L)$ com $p_l = (n_l, m_l) \in [1{:}N] \times [1{:}M]$ para $l \in [1{:}L]$ que minimize a distância e satisfaça as três condições:

1- Condições de fronteira: $p_1 = (1,1)$ e $p_L = (N, M)$.

2- Condição de monotonicidade: $n_1 \leq n_2 \leq \cdots \leq n_L$ e $m_1 \leq m_2 \leq \cdots \leq m_L$.

3- Condição de tamanho do passo: $p_{l+1} - p_l \in \{(1,0), (0,1), (1,1)\}\, para\ l \in [1: L-1]$.

A função de custo associada ao caminho de alinhamento $p_{otimizado}$ computado em relação à matriz local de custo *c* , que representa todas as distâncias pareadas, é definida pela Equação (2).

$$c_p(A,B) = \sum_{l=1}^{L} c = (a_{n_l}, b_{m_l}) \tag{2}$$

O caminho que apresenta o custo mínimo de alinhamento é denominado de caminho de alinhamento otimizado $p_{otimizado}$. Esse caminho otimizado é encontrado por um algoritmo ganancioso de computação dinâmica em uma a matriz de custo acumulado D, ou matriz de custo global, definida pelas seguintes regras:

Primeira coluna: $D(1,j) = \sum_{k=1}^{j} c(a_1, b_k), j \in [1, M]$

Primeira linha: $D(i,1) = \sum_{k=1}^{i} c(a_k, b_1), j \in [1, N]$

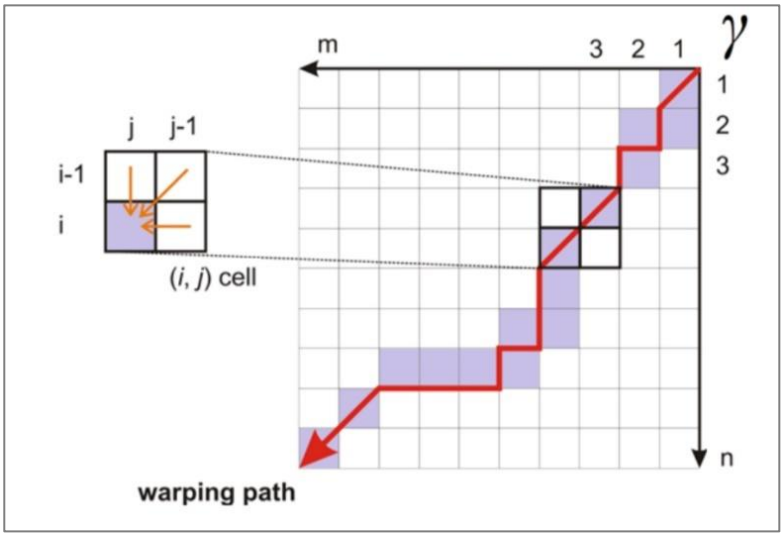
Outros elementos: $D_{base} + \min \begin{cases} D(A_{i-1}, B_{j-1}) & (Correspondência) \\ D(A_{i-1}, B_j) & (Inserção) \\ D(A_i, B_{j-1}) & (Deleção) \end{cases}$

Esse caminho define o alinhamento entre as sequências $A = \{A_1, A_2, .., A_n\}$ e $B = \{B_1, B_2, \dots, B_m\}$ designando cada elemento A_{n_l} de A para cada elemento B_{n_l} de B com base nas seguintes operações:

- Correspondência: quando função de custo se minimiza pela correspondência direta de A_n e B_m (comparação um-para-um). É representado no caminho de alinhamento como um deslocamento diagonal.
- Inserção: quando a função de custo se minimiza com a inserção, ou defasagem um valor na série A com relação a B (comparação muitos-para-um). É representado no caminho de alinhamento como um deslocamento vertical.
- Deleção: quando a função de custo se minimiza com a deleção, defasagem de um valor na série B com relação a A (comparação um-para-muitos). É representado no caminho de alinhamento como um deslocamento horizontal.

A função de distância D_{DTW} é então a função de computação dinâmica que encontra o caminho de alinhamento que minimiza o custo de alinhamento entre as séries temporais, dentre todos os caminhos possíveis (Figura 4).

Figura 4 - Caminho de alinhamento em uma matriz global de custo de comparação de séries temporais



Fonte: Modificado de (CASSISI et al., 2012)

A única restrição para a implementação correta do algoritmo de DTW está no uso de sinais discretos ou, de forma mais generalizada, de séries temporais equidistantes temporalmente (MÜLLER, 2007). E a sua maior limitação consiste no, relativamente, alto custo computacional, que decorre da complexidade dada por O*(mn)* onde *m* e *n* estão relacionados ao tamanho das duas séries temporais comparadas (GU; JIN, 2006). Esse custo tem aumento exponencial com o aumento das séries temporais analisadas, uma vez que envolve o cálculo de matrizes de distâncias entre as séries.

Diversas técnicas são comumente utilizadas para superar a limitação relacionada ao custo computacional. Essas técnicas têm o intuito de extrair aproximações ou amostragens das séries temporais. Como exemplo dessas técnicas, tem-se: (i) simplificação do algoritmo DTW através transformação no domínio das frequências com transformadas discretas de Fourier, do inglês, *Discrete Fourier Transform* (DTF)(AGRAWAL; FALOUTSOS; SWAMI, 1993; PARK et al., 2003); (ii) comparação de histogramas (GU; JIN, 2006); (iii) utilização da aproximação agregada por partes, do inglês, *Piecewise Aggregate Approximation* (PAA)(CHU et al., 2002; KEOGH; PAZZANI, 2000); e (iv) subjanelas deslizantes de procura da correspondência (FALOUTSOS; RANGANATHAN; MANOLOPOULOS, 1994).

3 MATERIAL E MÉTODOS

3.1 Base de Dados

3.1.1 Aquisição da Base de dados de Poços de Exploração e Produção de Petróleo

A Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) disponibilizou no ano de 2021 uma base de dados que contém informações de exploração e produção em bacias sedimentares continentais brasileiras. Tal base de dados apresenta ampla variedade de informações, incluindo levantamentos geofísicos não sísmicos (magnetometria e gravimetria, dentre outros), assim como levantamentos sísmicos e dados de poços. Em relação aos arquivos de poços, a base obtida contém 30.064 poços de 37 bacias sedimentares abrangendo 24 estados brasileiros.

3.1.2 Seleção dos Poços em relação às Bacias Sedimentares

Foram selecionados 30 poços de 10 campos de produção relativos a 5 bacias sedimentares. Essas informações estão descritas na

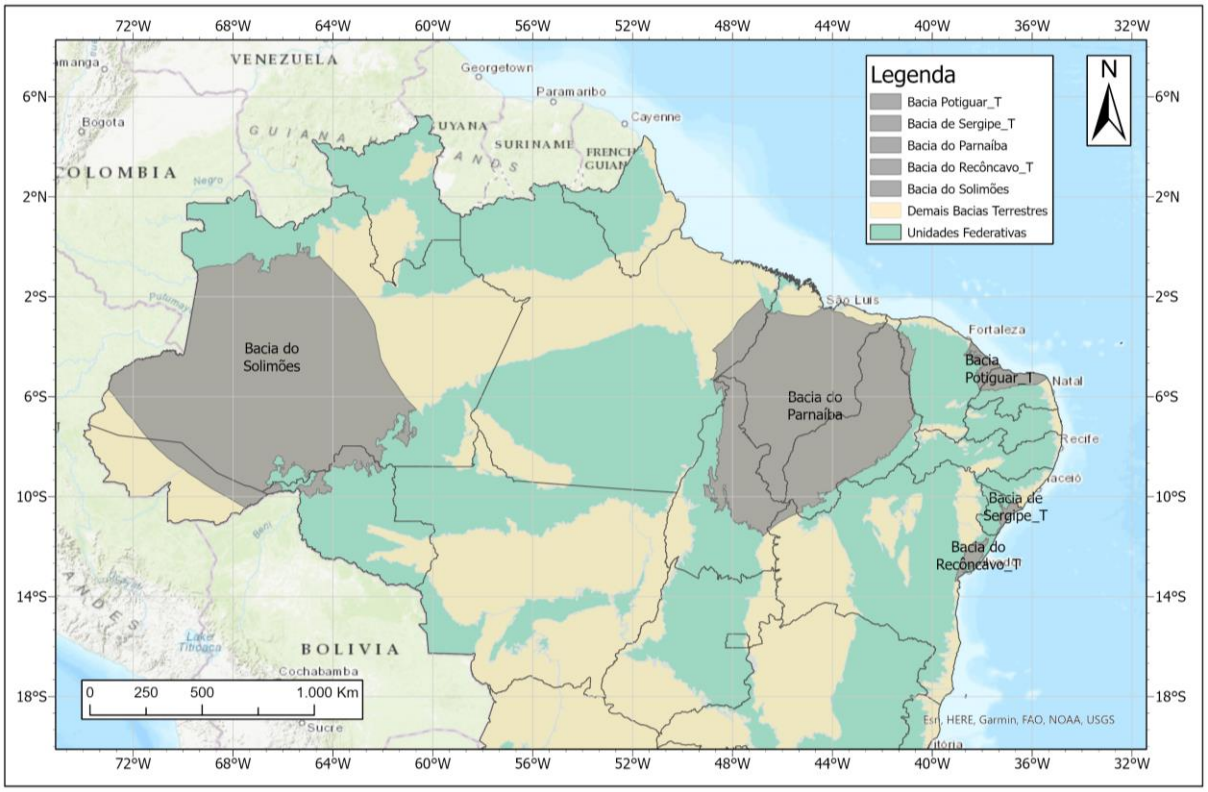
Tabela 1.

A disposição espacial das bacias sedimentares pode ser observada na Figura 5. Na sequência, as figuras Figura 6, Figura 7, Figura 8, Figura 9 e Figura 10 mostram em detalhe cada uma das 5 bacias sedimentares, contendo os poços escolhidos e seus respectivos campos de produção.

Tabela 1 – Poços selecionados para compor a base de dados e seus respectivos campos de produção e bacias sedimentares, bem como a profundidade de aquisição dos perfis

Poço	Campo de Produção	Bacia Sedimentar	Profundidade (m)
3-BRSA-1148-SE	Siririzinho	Sergipe	609
3-BRSA-1098-SE	Siririzinho	Sergipe	760
3-BRSA-1093-SE	Siririzinho	Sergipe	886,7
3-SZ-464-SE	Siririzinho	Sergipe	572,2
3-BRSA-912-SE	Carmópolis	Sergipe	246,3
3-BRSA-1279-SE	Carmópolis	Sergipe	723,5
3-BRSA-1262-SE	Carmópolis	Sergipe	839,5
3-BRSA-1307D-SE	Carmópolis	Sergipe	669,2
6-BRSA-1202D-BA	Dom João	Recôncavo	1427,5
1-BRSA-470D-BA	Guanambi	Recôncavo	921,1
4-BRSA-422-BA	CEXIS	Recôncavo	1810
3-BRSA-349-RN	Canto do Amaro	Potiguar	1794,4
3-BRSA-1295-RN	Canto do Amaro	Potiguar	1065,3
3-BRSA-1193-RN	Canto do Amaro	Potiguar	948,3
3-BRSA-1224-RN	Estreito	Potiguar	437,8
3-BRSA-1210-RN	Estreito	Potiguar	1492,8
3-BRSA-1325-RN	Estreito	Potiguar	492,9
3-BRSA-1320-RN	Alto do Rodrigues	Potiguar	483,3
3-BRSA-1341-RN	Alto do Rodrigues	Potiguar	619,9
3-OGX-95-MA	Gavião Branco	Parnaíba	985,4
4-PGN-25-MA	Gavião Tesoura	Parnaíba	1624,9
1-OGX-117-MA	Gavião Tesoura	Parnaíba	1235,3
4-PGN-19-MA	Gavião Tesoura	Parnaíba	5257,2
7-GVB-8-MA	Gavião Branco	Parnaíba	980,1
7-GVB-6-MA	Gavião Branco	Parnaíba	1035,9
7-RUC-59DP-AM	Rio Urucu	Solimões	922,5
3-BRSA-515-AM	Rio Urucu	Solimões	1809,5
3-BRSA-1057DA-AM	Rio Urucu	Solimões	575,1
3-BRSA-670-AM	Rio Urucu	Solimões	1573,1
3-BRSA-1121-AM	Rio Urucu	Solimões	1427,4

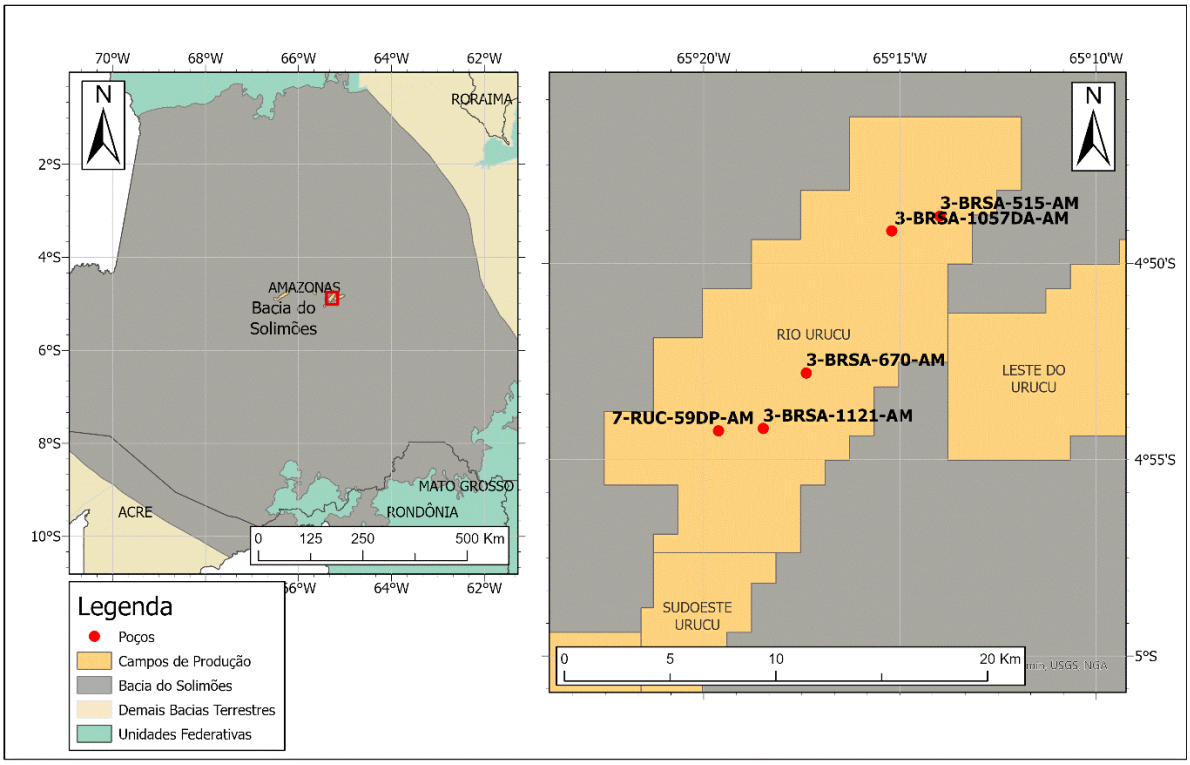
Figura 5 – Bacias sedimentares terrestres selecionadas para as investigações de similaridade entre poços.



Fonte: Autoria própria.

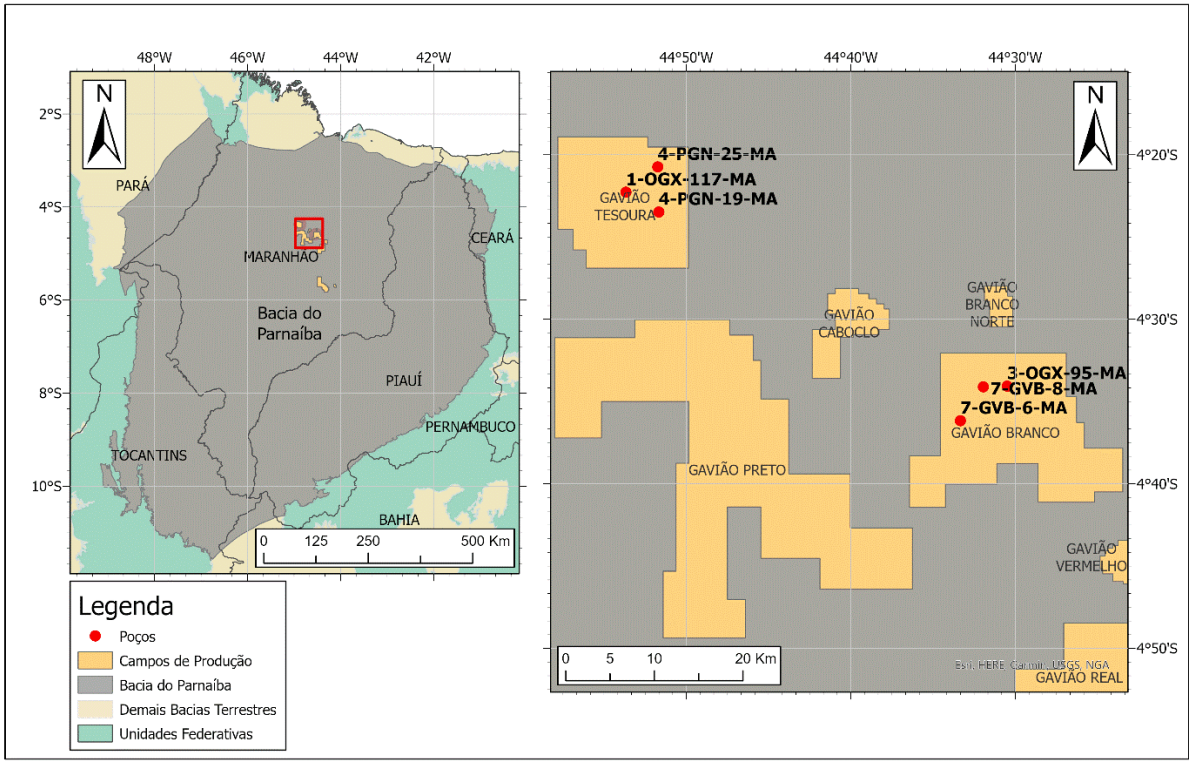
A escolha dos poços levou em consideração, sobretudo, a disposição espacial no ambiente georreferenciado visando abranger tanto poços com contextos geológicos e deposicionais similares quanto distintos entre si. É pressuposto que poços próximos no espaço georreferenciado, sobretudo aqueles encontram-se em um mesmo bloco exploratório ou campo de produção de uma mesma bacia sedimentar, possuem maior similaridade dado ao contexto deposicional e empilhamento estratigráfico parecidos ou idênticos. De mesma forma, poços mais distais, localizados em blocos exploratórios ou campos de produção diferentes em uma mesma bacia sedimentar podem apresentar similaridade ou não, uma vez que os contextos geológicos ainda podem ser parecidos, apesar da relativa distância no espaço georreferenciado. Isso ocorre, sobretudo, porque a bacia sedimentar permanece a mesma, tendo uma identidade deposicional comum. Já em relação aos poços localizados em bacias sedimentares diferentes, consequentemente em blocos exploratórios e campos de produção diversos, a similaridade tende a ser menor, uma vez que há uma brusca variação do contexto geológico e deposicional.

Figura 6 – Bacia do Solimões, com ênfase no campo de produção Rio Urucu, onde foram selecionados 5 poços para as investigações de similaridade.



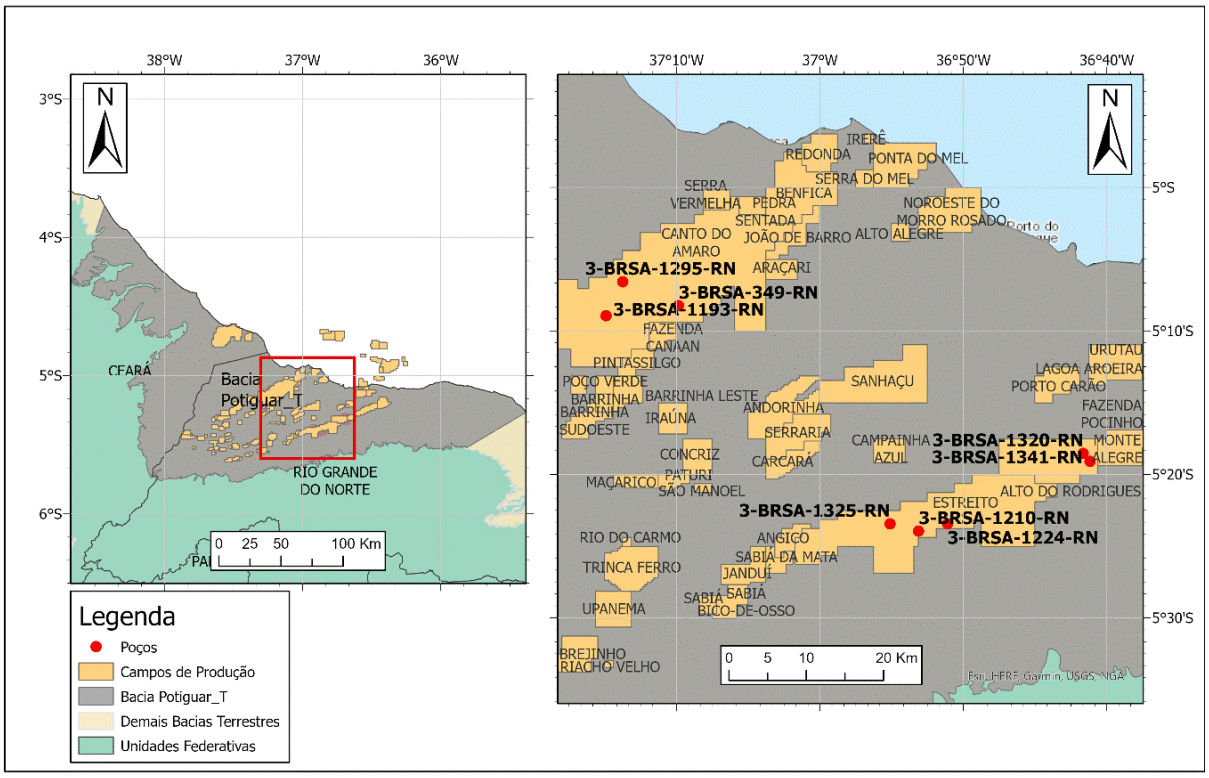
Fonte: Autoria própria.

Figura 7 – Bacia do Parnaíba, com ênfase nos campos de produção Gavião Tesoura e Gavião Branco, onde foram selecionados 6 poços para as investigações de similaridade.



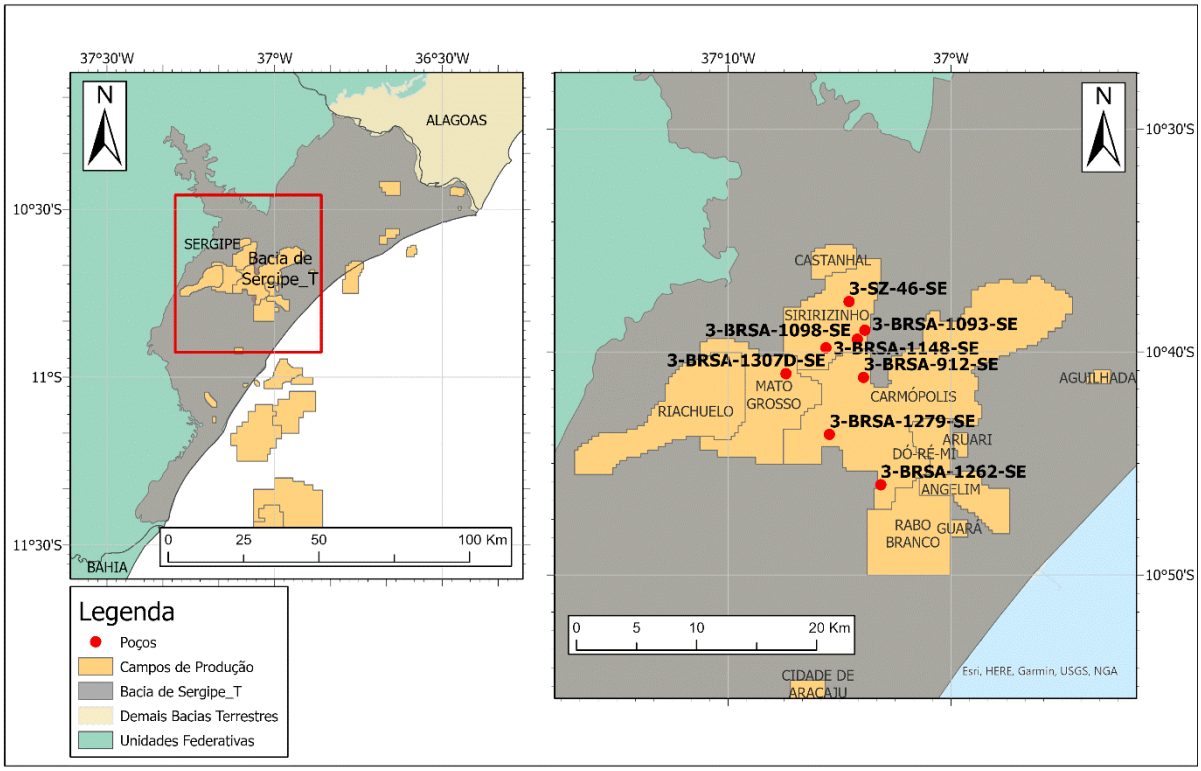
Fonte: Autoria própria.

Figura 8 – Bacia Potiguar, com ênfase nos campos de produção Canto do Amaro, Alto do Rodrigues e Estreito, onde foram selecionados 8 poços para as investigações de similaridade.



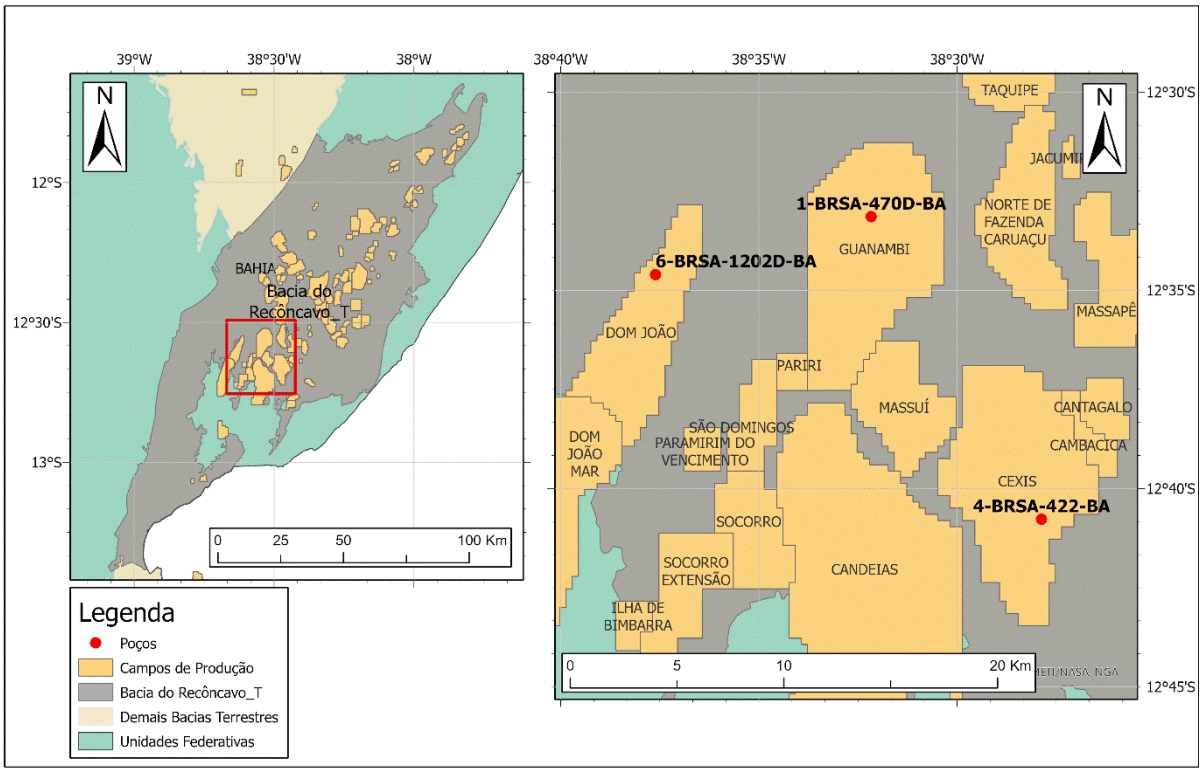
Fonte: Autoria própria.

Figura 9 – Bacia do Sergipe, com ênfase nos campos de produção Siririzinho e Carmópolis, onde foram selecionados 8 poços para as investigações de similaridade.



Fonte: Autoria própria.

Figura 10 – Bacia do Recôncavo, com ênfase nos campos de produção Guanambi, Som João e CEXIS, onde foram selecionados 3 poços para as investigações de similaridade.



Fonte: Autoria própria.

3.1.3 Estruturação da Base de dados para o Processamento em SOM

Os poços selecionados possuem posição espacial e profundidades distintas. Em relação às variáveis de perfilagem de poços medidas, foram selecionadas cinco perfis, a saber: (i) raios gama; (ii) densidade; (iii) nêutrons; (iv) fator fotoelétrico; e (v) sônico. A estruturação da base de dados, portanto, levou em conta tanto os atributos espaciais de posição e profundidade quanto as variáveis geofísicas medidas. Cada profundidade de aquisição em cada um dos poços foi tratada como uma instância na base de dados. Como o intervalo de amostragem para os poços não é homogêneo, todos os poços foram interpolados para um intervalo de aproximadamente 15 cm de espaçamento, comum à maior parte dos poços da base.

A base de dados final reuniu 376.993 instâncias relacionadas aos registros das 5 variáveis de perfilagem em diferentes profundidades dos 30 poços, empilhadas sequencialmente. A estrutura da base de dados relacional resultante pode ser observada na Tabela 2.

Tabela 2 – Estrutura da base de dados SQL mostrando as instâncias dadas pelos poços, suas respectivas profundidades e variáveis

POÇOS	PROF (M)	GR (API)	RHO (G/CC)	PHI (DEC)	PE (B/E)	DT (USPF)
3-BRSA-1148-SE	186,8	51,8284	1,9889	43,8688	50,7869	53,9656
3-BRSA-1148-SE	186,9	58,76	1,9768	42,0174	51,1325	53,2405
3-BRSA-1148-SE	187,0	62,1378	1,9719	44,7107	50,9208	53,2042
3-BRSA-1148-SE	187,1	63,9909	1,9674	45,4104	50,6495	53,3699
3-BRSA-1148-SE	187,2	62,3841	1,9608	43,856	50,1600	53,4589
3-BRSA-1148-SE	187,3	61,1174	1,9605	41,6115	50,4235	53,6301
3-BRSA-1148-SE	187,4	60,1322	1,9648	43,7524	51,2611	53,9355
3-BRSA-1148-SE	187,5	58,9359	1,9799	48,0865	53,7165	54,3635
3-BRSA-1148-SE	187,6	58,7326	1,9876	45,3005	54,6713	54,9808
3-BRSA-1148-SE	187,7	59,2252	1,9889	44,6345	54,9922	55,8016
3-BRSA-1148-SE	187,8	63,0487	1,9681	48,7475	52,8551	56,1582
3-BRSA-1148-SE	187,9	64,1160	1,9500	46,5232	50,8015	55,7979
3-BRSA-1148-SE	188,0	65,4413	1,9327	44,9566	48,5448	55,3311
3-BRSA-1148-SE	188,1	67,6385	1,9276	53,0142	45,2144	55,0467
3-BRSA-1148-SE	188,2	68,9990	1,9376	50,4215	44,4216	55,2301
3-BRSA-1148-SE	188,3	70,0037	1,9517	46,6316	43,9117	55,3509
3-BRSA-1148-SE	188,4	70,5315	1,9696	43,8996	43,6620	55,4903
3-BRSA-1148-SE	188,5	71,5144	1,9929	46,7522	42,4177	56,2143
⋮	⋮	⋮	⋮	⋮	⋮	⋮

3.2 Séries Auto-Organizáveis (SOrS)

O conceito de Séries Auto-Organizáveis é aqui introduzido como uma técnica de auto-organização de séries temporais com base na integração de informações provenientes da posição e da distância obtidos por SOM. A técnica se baseia na junção de informações da Matriz-U, obtida a partir da aplicação da técnica SOM, com uma segunda matriz de valores relacionados às posições dos *hits* da Matriz-U. Essa segunda matriz, aqui denominada de Matriz de Posição (Matriz-P), denota a abordagem inovadora que será apresentada nos tópicos a seguir.

3.2.1 Processamento SOM

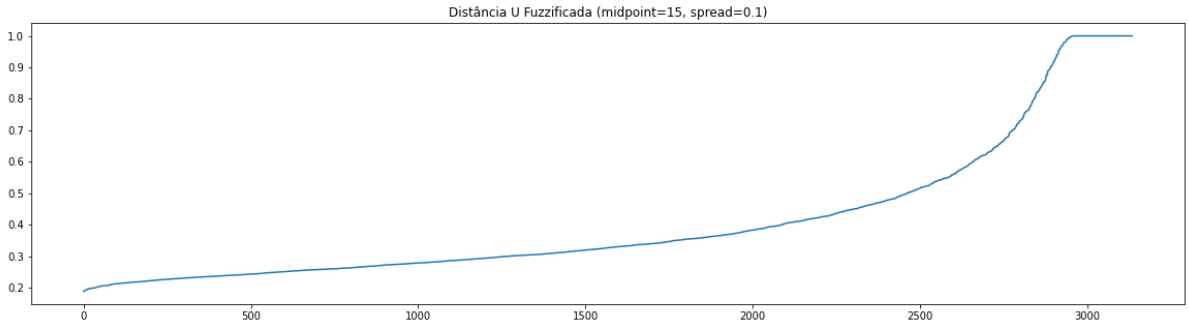
As instâncias (ou amostras) relativas a cada profundidade de poço foram inseridas em um espaço de 5 dimensões. O espaço de dados foi iniciado de forma aleatória. Para a projeção do espaço foi escolhido a superfície de um hipervolume toroidal construído a partir de uma rede hexagonal. O mapa auto-organizado foi configurado com o tamanho de 56 linhas x 56 colunas, totalizando 3.136 neurônios, ou BMU's. O tamanho do mapa segue uma regra heurística (CÉRÉGHINO; PARK, 2009; PARK et al., 2007; VESANTO; ALHONIEMI, 2000) conforme a Equação (3),

$$5\sqrt{N} \tag{3}$$

onde N corresponde ao número total de instâncias.

Após a execução do algoritmo e finalização do mapa auto-organizado, foram desenvolvidas imagens de *component plots* e matriz-U. Dado ao baixo contraste entre as dissimilaridades em função da concentração de um *"hot spot"* de dissimilaridade, a matriz-U foi fuzzificada, produzindo a aqui chamada Matriz-D, para garantir maior distribuição da informação. Para isso, foi utilizada uma função sigmoidal com ponto médio de 50 e espalhamento de 0,02 (Figura 11).

Figura 11 – Função sigmoidal aplicada à matriz-U para proporcionar distribuição mais propícia dos dados



Fonte: Autoria própria.

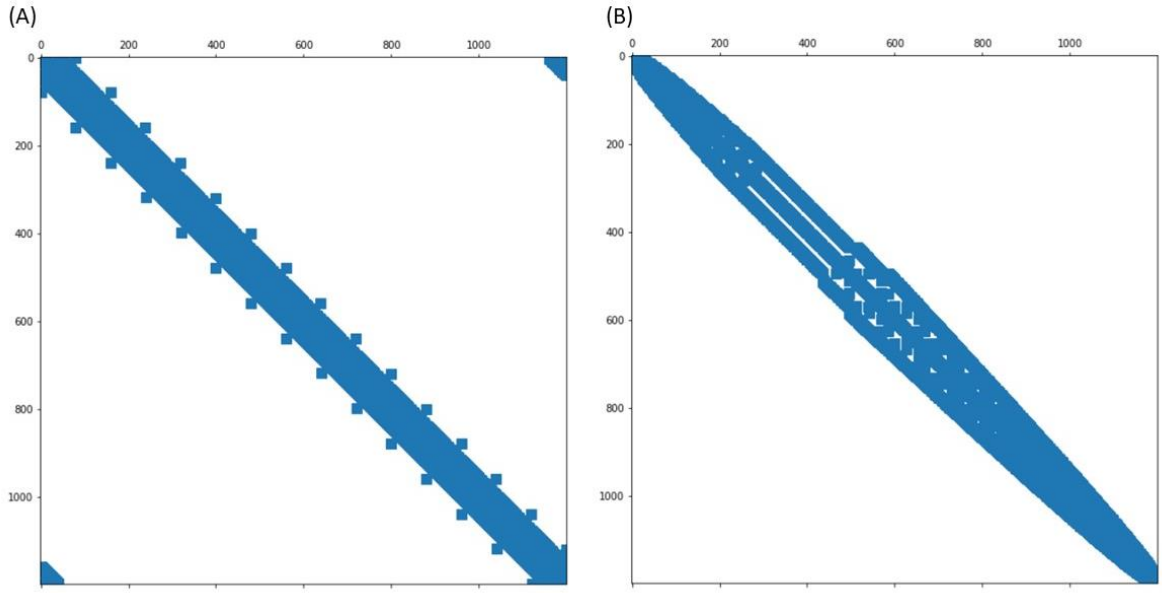
3.2.2 Matriz Integrada de Distância e Posição (Matriz-IDP)

A Matriz-U, ou a sua versão normalizada, aqui chamada de Matriz-D, têm por característica principal a possibilidade de observar relações de distância entre neurônios adjacentes. Dessa maneira, relações obtidas pela dissimilaridade entre neurônios vizinhos são construídas e analisadas com os dados obtidos pela matriz. Visualmente e numericamente nos mapas auto-organizados correspondentes à Matriz-U, a partir da distância euclidiana, torna-se possível observar fronteiras de dissimilaridades que separam grupos de neurônios no espaço das variáveis. No entanto, a Matriz-U pode trazer relações de distância parecidas em regiões diferentes do mapa, separadas por fronteiras de alta dissimilaridade. Ou seja, observando as relações de distância para com a vizinhança, um dado neurônio pode apresentar valores de dissimilaridade equivalentes a um outro neurônio, em posição diferente e distante do mapa. Por isso, torna-se sempre necessário a observação das fronteiras de dissimilaridade para uma análise de agrupamentos, por exemplo, e não apenas a informação numérica da distância. Com isso, pode-se dizer que o valor exclusivo da dissimilaridade entre neurônios vizinhos não é suficiente para determinar o quão similares são os neurônios de todo o mapa sem a informação da posição na matriz.

Dessa maneira, percebeu-se uma demanda pelo desenvolvimento de uma matriz diferente da Matriz-U, capaz de adicionar às distâncias desta, numericamente, as relações de posição. A solução, portanto, passaria pelo desenvolvimento de uma segunda matriz que trouxesse informação relacionada à posição, obedecendo à periodicidade, em se tratando da forma toroidal. Assim, adotou-se como solução a integração de distância (ou dissimilaridade) entre neurônios vizinhos à uma matriz desenvolvida a partir do algoritmo Cuthill-McKee

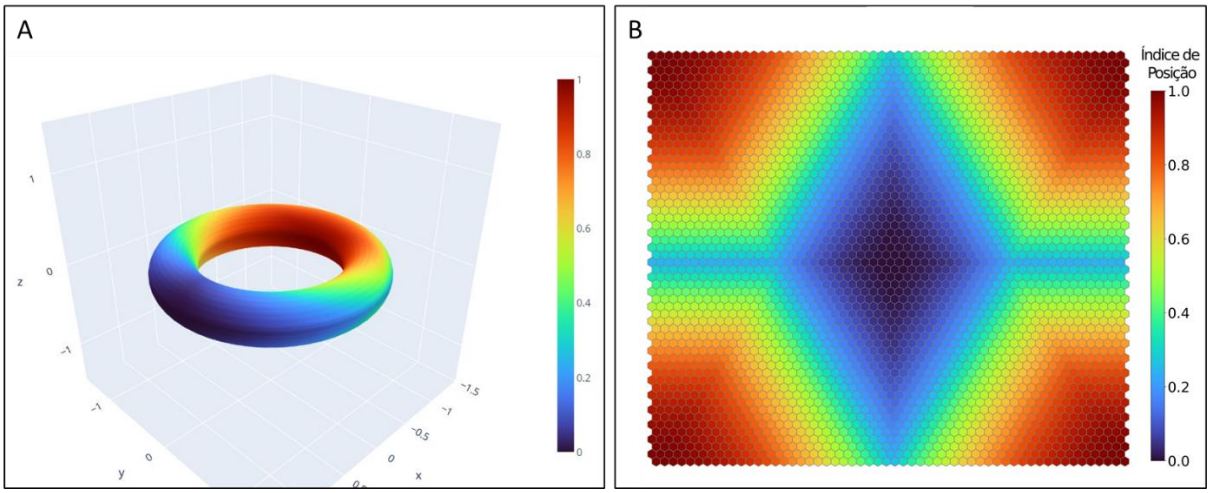
(CUTHILL; MCKEE, 1969). O método consistiu na criação de uma matriz inicialmente enumerando os neurônios do mapa auto-organizado. Paralelamente, desenvolveu-se uma matriz vizinhança esparsa quadrada de dimensão $N_{neurons} \times N_{neurons}$, cuja associação espacial do mapa auto-organizado requer atribuir o valor = 1 para o próprio neurônio, bem como para vizinhos imediatos (Figura 12 A). Esse desenvolvendo possibilita a preservação da periodicidade toroidal. Na sequência, foi aplicado o algoritmo Cuthill-McKee reverso para a permutação da matriz esparsa para uma matriz de banda com banda pequena (Figura 12 B). A renumeração dos neurônios, e sequente normalização [0,1], associada à matriz vizinhança com banda estreita resultante é aqui denominada de Matriz de Posição (Matriz-P). Esta possibilita que os neurônios próximos recebam valores da enumeração próximos também, quantificando suas respectivas posições no toroide (Figura 13).

Figura 12 – (A) Matriz laplaciana esparsa representativa do índice dos neurônios (BMUs); (B) Matriz de índice de neurônios reconfigurada segundo resultado do algoritmo de Cuthill-MacKee



Fonte: Rafael dos Santos Gioria.

Figura 13 – Matriz-P, originada a partir da aplicação do algoritmo Cuthill-McKee em coloração sintética, com periodicidade, aplicada aos mapas SOM em dois formatos de projeção: (A) 3D com topologia toroidal; (B) 2D de (A), em formato de mapa, considerando periodicidade entre as bordas. Cada elemento da matriz apresenta identidade única relacionada à posição

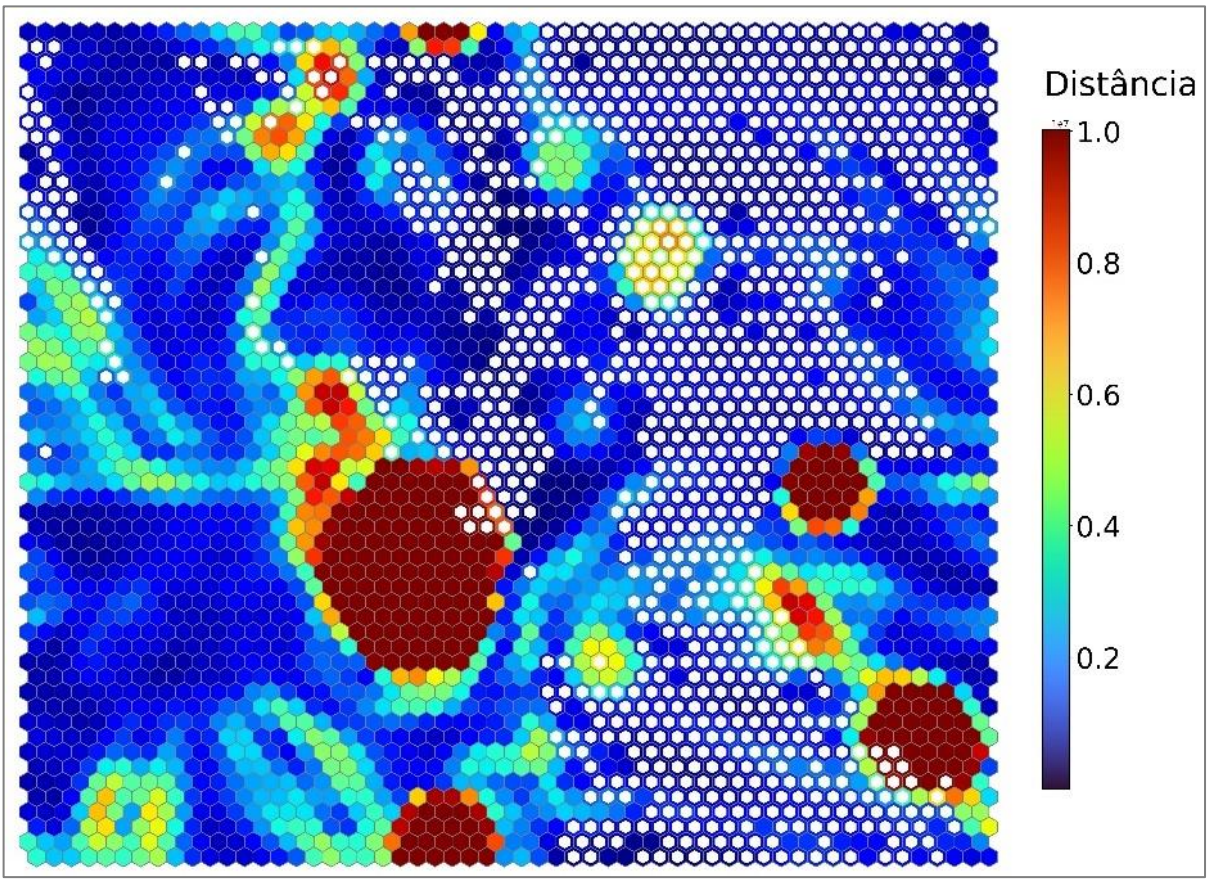


Fonte: Autoria própria.

3.2.3 Ponderações de Posição e Distância para geração de Séries Auto-Organizáveis

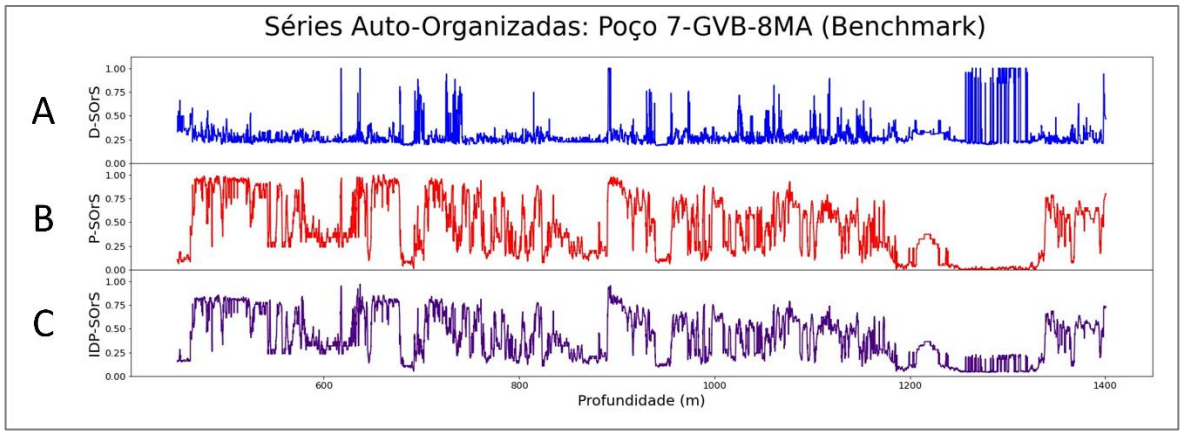
As séries baseadas exclusivamente nos índices distância (Índice-D) ou posição (Índice-P), bem como aquelas geradas pela integração entre posição e distância (IDP), são desenvolvidas com base na posição do BMU de cada uma das instâncias de treinamento, normalizadas para o intervalo de [0,1] (Figura 14). A reconstrução da série temporal, portanto, no Índice-D foi obtido a partir da Matriz-D, construída pela normalização dos valores de dissimilaridade da Matriz-U, enquanto que no Índice-P se baseia na normalização da posição de cada BMU no toroide de treinamento, dada pela Matriz-P. Dessa maneira, é gerada uma assinatura baseada no processo de auto-organização denominado neste trabalho de Séries Auto-Organizáveis, cuja sigla advém do termo aqui designado em inglês como ***S**elf-**Or**ganizing **S**eries* (SOrS). As séries auto-organizadas resultantes são aqui denominadas, respectivamente, de D-SOrS, P-SOrS e IDP-SOrS. Estas séries são apresentadas para o poço 7-GVB-8-MA, da Bacia do Parnaíba, utilizado como *benchmark*, conforme a Figura 15.

Figura 14 – Matriz-U fuzzificada com indicação dos BMUs representativos (hexágonos brancos) das instâncias de treinamento da amostra 7-GVB-8-MA



Fonte: Autoria própria.

Figura 15 – Séries Auto-Organizadas: (A) D-SOrS (azul); (B) P-SOrS (vermelho); e (C) IDP-SOrS (roxo) ponderada pelos índices de posição e distância



Fonte: Autoria própria.

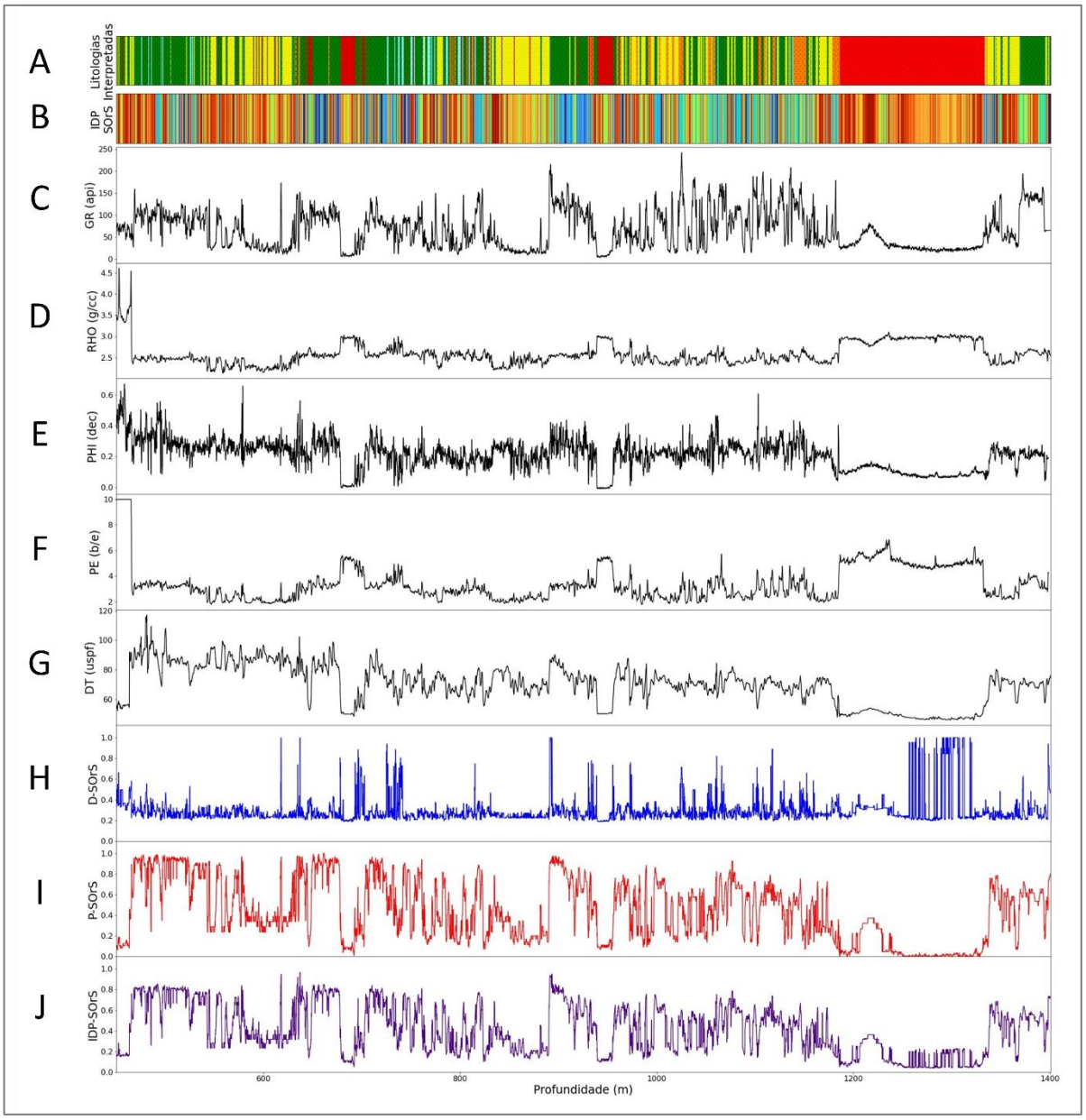
Como o IDP reflete em um índice único contendo informações de posição e distância, para o desenvolvimento de uma média ponderada entre ambas as informações, torna-se necessário definir a proporcionalidade dos índices P e D. Essa proporcionalidade pode variar

com a necessidade ou demanda por cada uma destas duas informações dada a especificidade da aplicação a ser abordada.

Em se tratando da aplicação em perfis de poços, as SOrS refletem características geológicas que variam diferentemente em cada uma das variáveis de entrada, como porosidade, reologia, argilosidade e/ou composição mineralógica. A D-SOrS mostra o nível de dissimilaridade de uma dada instância com relação às vizinhas no mapa de treinamento. Assim, esta série pode refletir variações litológicas mais abruptas, bem como a homogeneidade e/ou heterogeneidade dessas variações. Tal série, portanto, representa a demarcação da variabilidade litológica e não necessariamente das litologias. Já a P-SOrS reflete mudanças nos domínios definidos pela auto-organização dos neurônios de treinamento, os quais podem ser interpretados como litologias ou domínios litológicos. Portanto, esta série trata de um índice mais representativo dos tipos litoestratigráficos, uma vez que unidades litológicas parecidas estarão próximas no espaço n-dimensional projetado no toroide. As séries dos dados de entrada utilizadas no treinamento assim como as SOrS e interpretações litológicas para o poço *benchmark* (7-GVB-8-MA) podem ser observadas na Figura 16.

Após testes com diferentes proporcionalidades, a definição da proporcionalidade de 20% do Índice-D e 80% do Índice-P garantiu ao IDP condições de dar maior ênfase às variedades litoestratigráficas (dadas pela proporção de 80%) em relação às zonas de maior homogeneidade ou heterogeneidade litológica (dadas pela proporção de 20%). Essa proporcionalidade, no entanto, pode ser ajustada ou tratada como um hiperparâmetro, uma vez que dependerá da ênfase demandada.

Figura 16 – Poço *benchmark* 7-GVB-8-MA em diferentes componentes: (A) Perfil composto de interpretação geológica; (B) SOrS-IDP em coloração sintética; (C), (D), (E), (F) e (G) representam, respectivamente, séries temporais das variáveis de entrada GA, RHO, PHI, PE e DT (preto); (H) D-SOrS (azul); (I) P-SOrS (vermelho); e (J) IDP-SOrS (roxo), ponderado a 20%D e 80%P

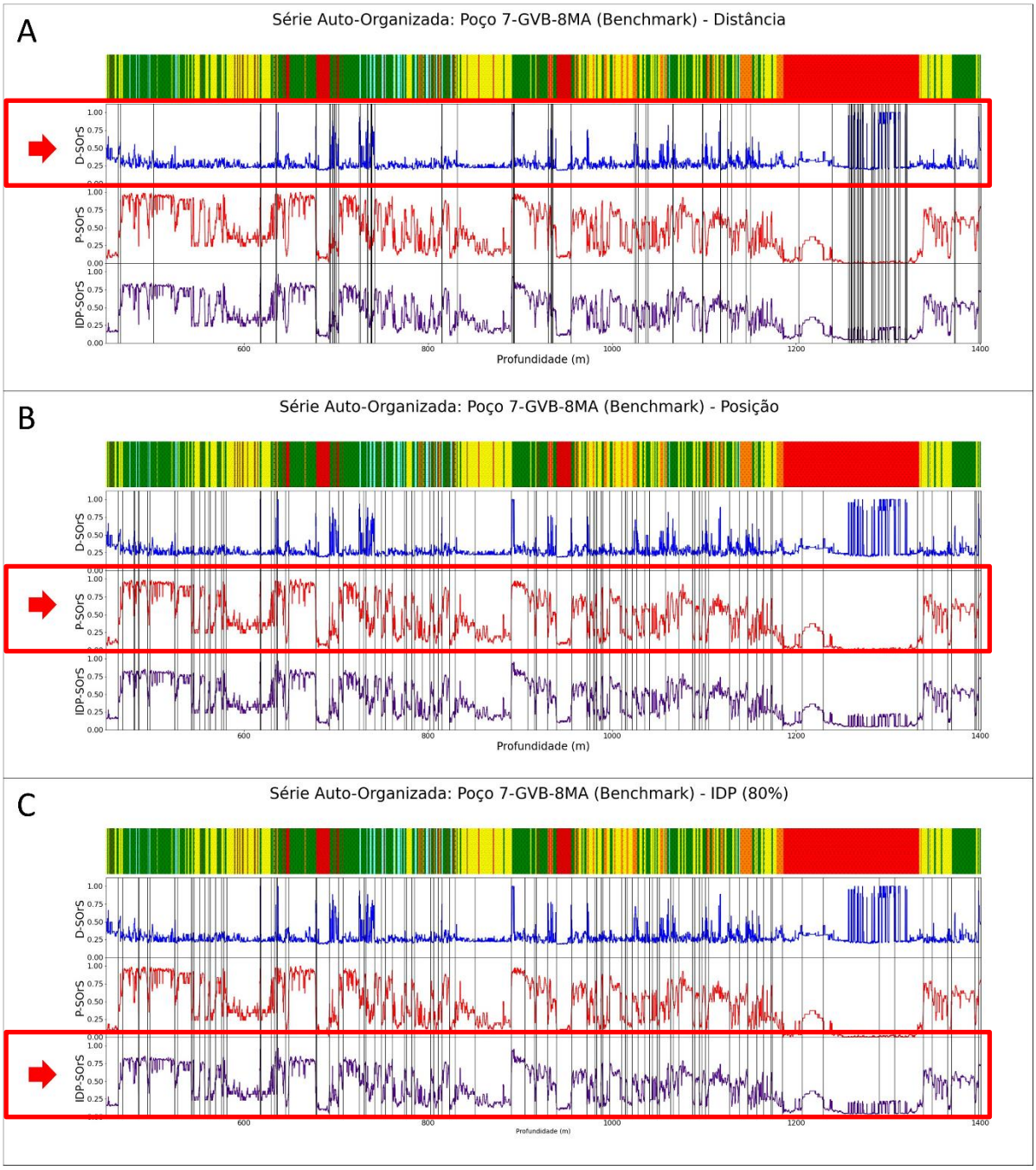


Fonte: Autoria própria.

Para compreender melhor se a proporcionalidade atribuída para os índices D e P no desenvolvimento do IDP eram coerentes com as unidades litoestratigráficas conhecidas para um dado poço, uma análise de detecção de pontos de mudança a partir de *Change Point Detection* (*CPD*) (Killick et al., 2012) foi aplicada nas séries D-SOrS, P-SOrS e IDP-SOrS (Figura 17 A, B e C, respectivamente), tendo como base a biblioteca *Ruptures* do Python® (KILLICK; FEARNHEAD; ECKLEY, 2012; TRUONG; OUDRE; VAYATIS, 2018). A análise separou cada uma das séries em 100 segmentos, minimizando a entropia interna de cada

segmento. Essa análise objetivou averiguar sobre a relevância metodológica de cada uma das SOrS no auxílio à definição de uma proporção não arbitrária para a soma ponderada de cada uma delas na composição da IDP-SOrS.

Figura 17 – Análise CPD para detecção de pontos de mudança, representados pelas linhas em preto, aplicada às séries: (A) D-SOrS, (B) P-SOrS e (C) IDP-SOrS. As linhas pretas correspondem às séries apontadas pelas setas em vermelho. É possível notar que em (A) e (B) as linhas em preto coincidem com variações litoestratigráficas que são contempladas pela IDP-SOrS (C).

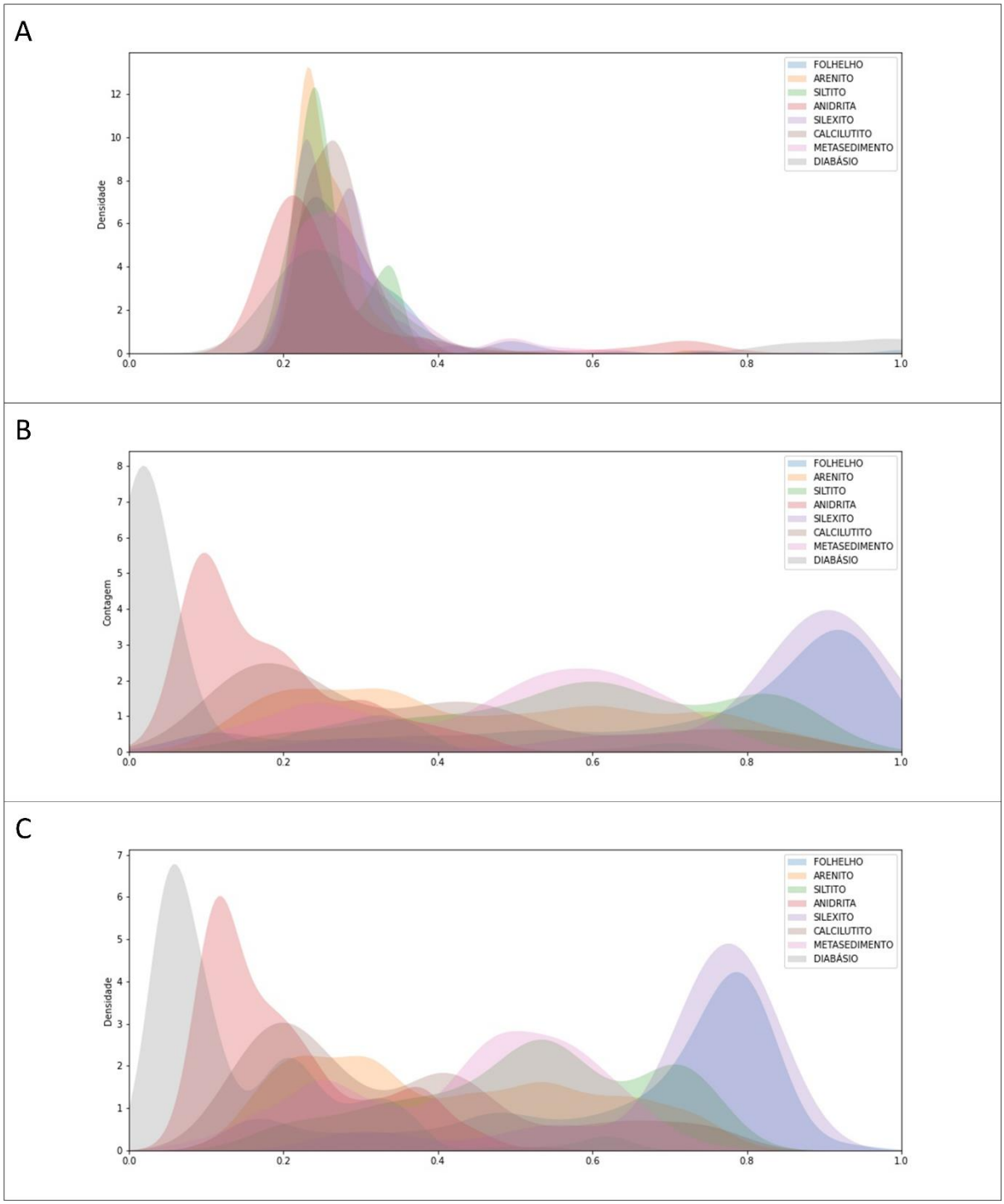


Fonte: Autoria própria.

Quando segmentada, a D-SOrS separou, de forma mais proeminente, domínios de litologias, regiões homogêneas e heterogêneas, bem como transições entre regiões com maior ou menor variabilidade litológica. Já a P-SOrS refletiu de maneira satisfatória a interpretação litológica apresentada, marcando regiões de mesma litologia e suas variações.

No intuito de observar quão descritiva cada uma das SOrS é em relação a cada litologia, uma análise individual destas séries foi performada com relação aos grupos litológicos interpretados para o poço modelo 7-GVB-8-MA. A D-SOrS (Figura 18 A) possui picos altos e de baixa dispersão, mas é pouco discriminativa em relação às litologias. Estes picos estão concentrados sobretudo na faixa de 0.2-0.4. Já a P-SOrS (Figura 18 B) evidencia picos mais baixos e mais dispersos em relação à D-SOrS, mas com clara separação ao longo do intervalo da assinatura, demonstrando que são mais discriminativos com relação às unidades litoestratigráficas. Essa análise corrobora com o esperado em relação às SOrS, sendo a D-SOrS mais relacionada à variabilidade litoestratigráfica e a P-SOrS indicativa de domínios ou fortes transições relacionadas às unidades litoestratigráficas distribuídas no mapa auto-organizado. Do ponto de vista da interpretação geológica, portanto, entende-se que uma abordagem metodológica adequada à ponderação do IDP deve ser majoritariamente influenciada pelo Índice-P, devido ao seu maior potencial discriminativo. Uma terceira análise das curvas cumulativas de densidade de probabilidade foi desenvolvida para a integração IDP-SOrS com proporções de 80% P-SOrS e 20% D-SOrS (Figura 18 C). Esta ponderação preservou as distribuições dos picos associados a cada unidade litoestratigráficas observadas na P-SOrS, porém com menor dispersão ao redor da média. Assim, compreendeu-se que os picos são melhor definidos na abordagem integrada obtida pela IDP-SOrS do que nas séries individuais. A IDP-SOrS, portanto, consistiu em uma das abordagens comparativas para a análise da similaridade entre os perfis de poços. Estas séries tornam possível resumir múltiplos perfis de poços em um único perfil síntese.

Figura 18 – Curvas da estimativa de densidade de probabilidade para: (A) D-SOrS; (B) P-SOrS; e (C) IDP-SOrS com 20% D e 80% P. As curvas da estimativa de densidade de probabilidade foram calculadas para cada uma das unidades litoestratigráficas interpretadas do poço modelo 7-GVB-8-MA



Fonte: Autoria própria.

3.3 Similaridade entre Poços de Petróleo por DTW

A similaridade foi performada a partir da combinação par-a-par entre os 30 poços da base de dados de treinamento. A comparação foi aplicada às séries obtidas pela técnica SOrS em duas abordagens, a saber: (i) perfis síntese gerados pela IDP-SOrS; e (ii) combinação vetorial entre as séries D-SOrS e P-SOrS. O algoritmo aplicado para comparar as séries temporais foi o DTW, abordado no tópico 2.5.1, em razão da sua capacidade de encontrar o alinhamento não linear ótimo entre sequências numéricas que não apresentem necessariamente o mesmo tamanho.

O algoritmo de DTW aplicado nesta presente pesquisa foi modificado do algoritmo de (KAMPER, 2021) e compilado para aceleração do processamento com base na biblioteca Numba (ANACONDA, 2018). Contudo, as performances demandaram alto custo computacional e elevado tempo de processamento, dada a densidade de amostragem das SOrS geradas em alguns dos poços, sobretudo aqueles de maior profundidade. Assim, foi aplicada uma rotina de reamostragem com redução na ordem de 4x em todas as SOrS. A rotina foi aplicada no campo das frequências, com base no método de Fourier, utilizando a função *resample* do módulo *signals* da biblioteca Python SciPy.

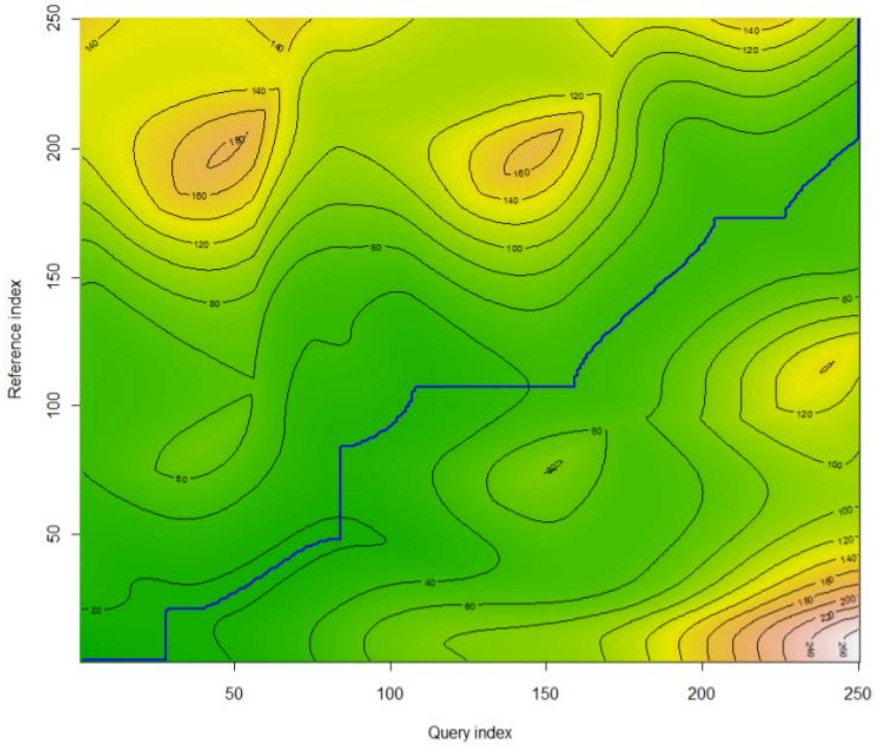
No algoritmo de DTW, inicialmente é calculada a matriz $D(N \times M)$ de distâncias $d(x_i, y_j)$ entre todos os pontos $x = \{x_0, \dots, x_N\}$ e $y = \{x_0, \dots, x_M\}$ das séries temporais de dimensões N e M de entrada. Em seguida, calcula-se o custo de alinhamento, que é representado por uma matriz $D_{custo}(N \times M)$ (Equação (4)) inicializada com a primeira coluna e primeira linha preenchidas por um valor muito alto (infinito). Todas as distâncias de alinhamento representam distâncias euclidianas compreendidas entre os pontos das séries temporais somados a uma penalidade, definida pela menor distância entre o ponto atual, anterior ou próximo da outra série temporal. Tais operações são denominadas, respectivamente, de inserção, correspondência ou deleção.

$$D_{custo} = d(x_i, y_j) + \min \begin{cases} d(x_{i-1}, y_{j-1}) & (Correspondência) \\ d(x_{i-1}, y_j) & (Inserção) \\ d(x_i, y_{j-1}) & (Deleção) \end{cases} para\ i = \{0, \dots, N\}\ e\ j = \{0, \dots, M\} \quad (4)$$

Sincronicamente, calcula-se o caminho de menor custo de alinhamento, definido por uma linha diagonal entre os extremos superior direito e inferior esquerdo, e com descida de gradiente de máximo declive, conforme exemplificado por (SENIN, 2008) (Figura 19). O custo

de alinhamento, ou distância DTW, é então definido como a soma dos custos da matriz $D_{custo}(N \times M)$ nos pontos percorridos pelo caminho de alinhamento otimizado.

Figura 19 – Mapa de calor da matriz de custo com caminho no máximo declive da descida de gradiente



Fonte: Senin (2008).

O custo de alinhamento C_{al} (Equação (5)) é então definido como o valor da somatória dos custos de alinhamento do caminho otimizado $P_{otimizado} = \{d_0, \dots, d_a\}$, onde $a = \max\{N, M\}$ ou, simplesmente, o custo acumulado $d(x_N, y_M)$ da matriz de custo $D_{custo}(N \times M)$, conforme a Equação Y.

$$C_{al} = \sum_{a=0}^{a} d(x_i, y_j)_a \tag{5}$$

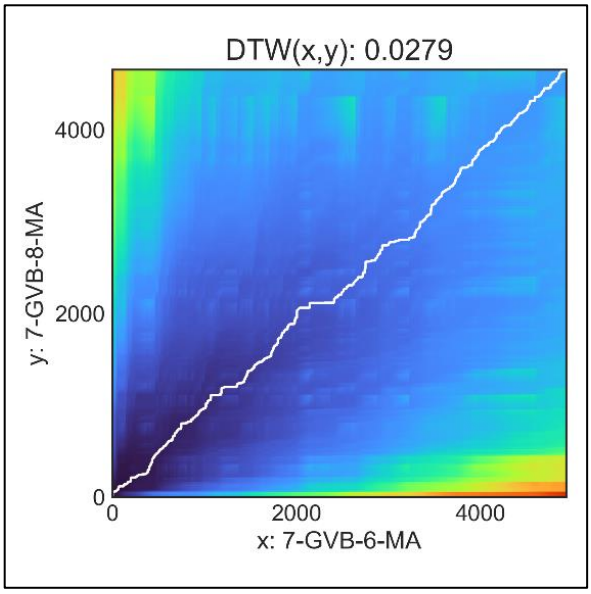
Vale ressaltar que o custo de alinhamento, por ser o resultado de uma somatória de comparações pontuais entre as séries temporais, é uma medida fortemente influenciada pelos tamanhos dos sinais analisados. Assim, essa somatória resulta em valores invariavelmente menores para comparações entre sequências numéricas mais curtas. Portanto, é necessário o cálculo do custo normalizado C_{norm} (Equação (6)) pelas dimensões das séries comparadas entre si. Isso possibilita a obtenção de um valor menos enviesado pelas dimensões das séries.

$$C_{norm} = \frac{\sum_{a=0}^{a} d(x_i, y_j)_a}{(N + M)} \quad (6)$$

3.3.1 Comparação entre Perfis Síntese IDP-SOrS

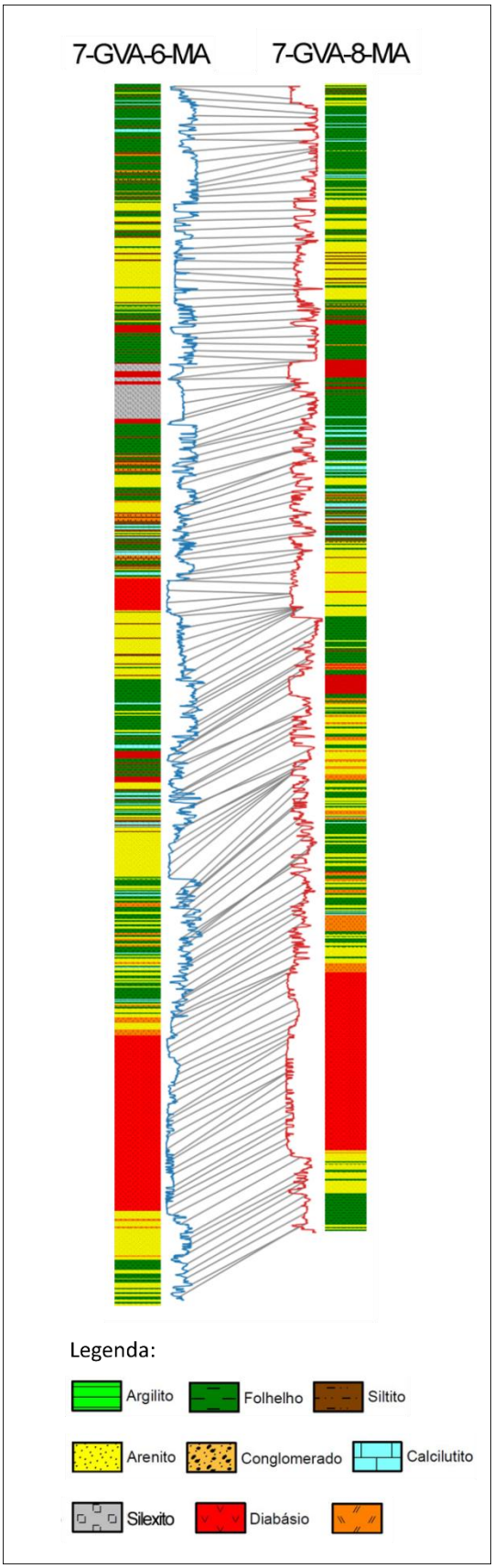
Os perfis síntese IDP-SOrS, conforme mencionado, agregam de forma ponderada os índices de posição e distância. A comparação dos 435 pares de poços, representativos das combinações de 30 poços, viabilizou o desenvolvimento das matrizes de custo. A comparação dos poços 7-GVB-8-MA e 7-GVB-6-MA (Figura 20) e o alinhamento de menor custo das IDP-SOrS, em conjunto com as interpretações litológicas, podem ser observados na Figura 21. Essas comparações foram realizadas para a IDP-SOrS resultante da média ponderada na proporção de 80% P e 20% D, conforme discutido anteriormente.

Figura 20 – Matriz de custo de alinhamento com traçado do caminho de melhor custo para a comparação entre os poços 7-GVB-8-MA e 7-GVB-6-MA



Fonte: Autoria própria.

Figura 21 – Alinhamento de menor custo encontrado pelo algoritmo DTW com interpretação litológica entre os poços 7-GVB-8-MA e 7-GVB-6-MA



Fonte: Autoria própria.

3.3.2 Comparação entre a combinação vetorial de D-SOrS e P-SOrS

A matriz $D(N \times M)$ de distâncias $d(x_i, y_j)$, calculada para duas séries temporais e usada como entrada para o algoritmo de DTW, pode ser calculada para sequências de séries temporais no formato vetorial $d(\overrightarrow{x_i}, \overrightarrow{y_j})$. Para isso, é necessário se estabelecer uma métrica de cálculo da distância, tal como a Euclidiana, Manhatan ou Cosseno, dentre outras.

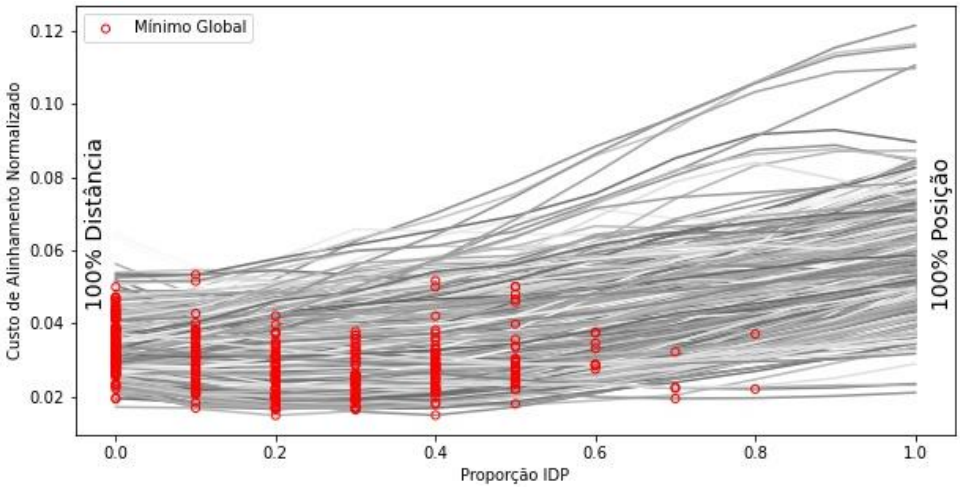
As séries D-SOrS e P-SOrS foram utilizadas como entradas para o algoritmo DTW na forma de uma matriz dessas distâncias vetoriais. Foram avaliadas, portanto, as distâncias Euclidiana e Manhatan na abordagem desta pesquisa.

3.4 Avaliação dos experimentos

3.4.1 Análise do custo mínimo de alinhamento por DTW nas ponderações IDP-SOrS

Entende-se que ambas as séries, D-SOrS e P-SOrS, são relevantes na definição de uma série única, IDP-SOrS, que reflita as litologias e as variabilidades litológicas de um poço de petróleo. Para corroborar esta compreensão, foi desenvolvida uma análise de custo mínimo de alinhamento obtido por DTW para cada proporção entre 0 e 100% variando a cada 10% de D-SOrS e P-SOrS. O custo de alinhamento foi calculado para cada uma das 435 comparações possíveis entre os 30 poços de petróleo (Figura 22). Essa análise mostrou que o custo mínimo de alinhamento varia de 0% a 80% de ponderamento para a P-SOrS, denotando, dessa maneira, a relevância da integração ocasionada pela IDP-SOrS.

Figura 22 – Análise de mínimos de custos de alinhamento DTW para diferentes proporções de soma ponderada de D-SOrS e P-SOrS, aplicada a todas as combinações possíveis entre pares de poços



Fonte: Autoria própria.

3.4.2 Clusterização hierárquica

Com base nas 435 distâncias DTW que representam cada uma das comparações entre as combinações possíveis entre os 30 poços de treinamento, estas distâncias foram organizadas em uma matriz de distâncias [30x30] com valor 0 na diagonal principal. Tal diagonal está relacionada ao custo de alinhamento de uma série temporal consigo mesma. Essa matriz foi utilizada como entrada pré-computada para o cálculo do método aglomerativo hierárquico com a função *linkage* do sub-módulo *hierarchy* do módulo cluster do pacote SciPy. Em seguida foi gerado o dendrograma resultante desse cálculo, e escolhido o *cut-off* capaz de discriminar a quantidade de cinco grupos, relacionados às cinco bacias por estes representados.

Na sequência, foram separados os rótulos associados ao agrupamento gerado pelo *cut-off* escolhido para o cálculo do coeficiente de silhueta. O coeficiente representa uma métrica da distância intra-agrupamento e inter-agrupamento de cada uma das amostras, resultando num valor entre 1 e -1. Valores mais próximos de 1 indicam agrupamentos bem individualizados, enquanto que aqueles mais próximos de 0 indicam sobreposição de agrupamentos e valores negativos, sugerindo erro na classificação de amostras.

4 RESULTADOS

4.1 Séries Auto-Organizáveis

As SOrS obtidas partiram dos resultados de uma análise SOM. Vale ressaltar, no entanto, que essas séries diferem de uma análise exclusivamente SOM, uma vez que são dotadas de informações relacionadas a uma adaptação da Matriz-U, mas também apresentam informações relacionadas à Matriz-P, conforme mencionado no tópico 3.2 deste texto. No entanto, a geração das SOrS depende da aplicação inicial do treinamento da técnica SOM, cujos resultados serão apresentados a seguir.

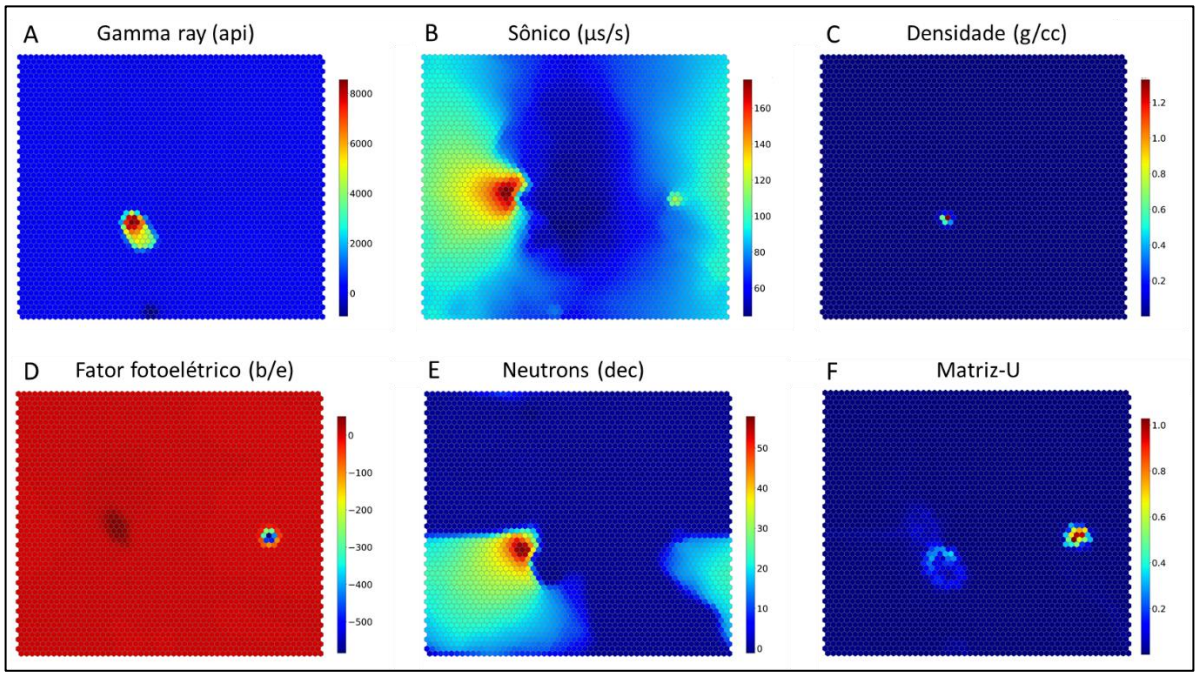
4.1.1 Mapas Auto-Organizáveis

Como resultado da aplicação da técnica SOM, foram geradas imagens das *component plots* e Matriz-U.

4.1.1.1 Component Plots

Os resultados obtidos a partir das *component plots* (Figura 23) mostram a distribuição das variáveis analisadas ante ao mapa auto-organizado produzido. A observação das componentes demonstra uma relação inversa entre os perfis RHO e PE. Os perfis de GR demonstram elevados valores mais próximos às regiões de maiores profundidades (perfil PROF).

Figura 23 – *Component plots* mostrando as variáveis analisadas e suas distribuições de valores em relação ao mapa auto-organizado em tamanho 56 x 56.

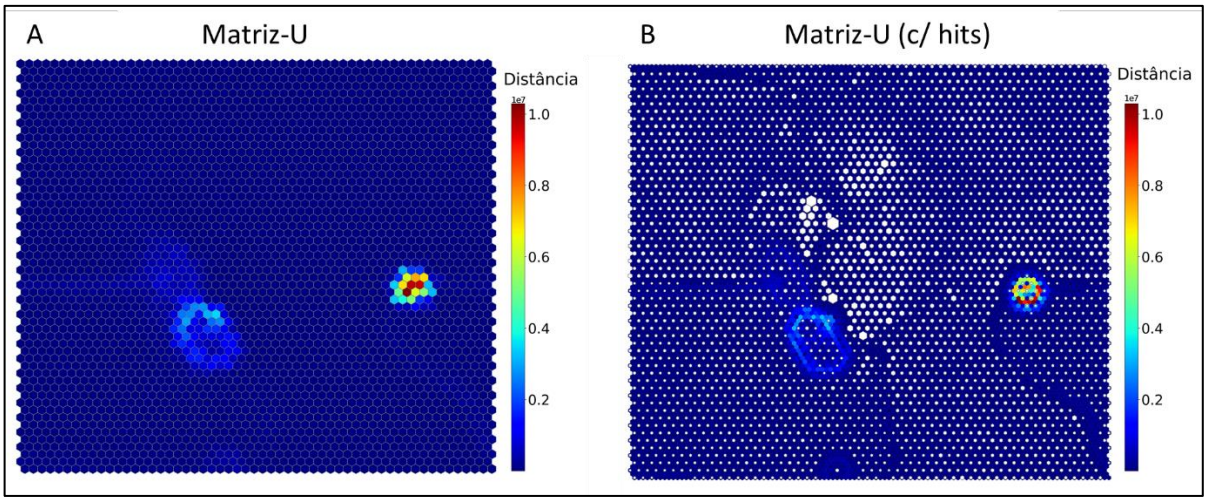


Fonte: Autoria própria.

4.1.1.2 Matriz-U

A partir da matriz-U (Figura 24) foi possível observar relações de dissimilaridade entre neurônios vizinhos. Cores quentes representam maiores dissimilaridades, enquanto que cores frias representam menor dissimilaridade entre neurônios adjacentes. Estas dissimilaridades apresentaram baixo contraste de valores.

Figura 24 – Mapa auto-organizado em tamanho 56 x 56 desenvolvido a partir das instâncias (profundidades) associadas aos 30 poços: (A) Matriz-U; e (B) Matriz-U com *hits* representados por hexágonos em branco, cuja variação de tamanho é proporcional à quantidade de instâncias (profundidades) associadas.

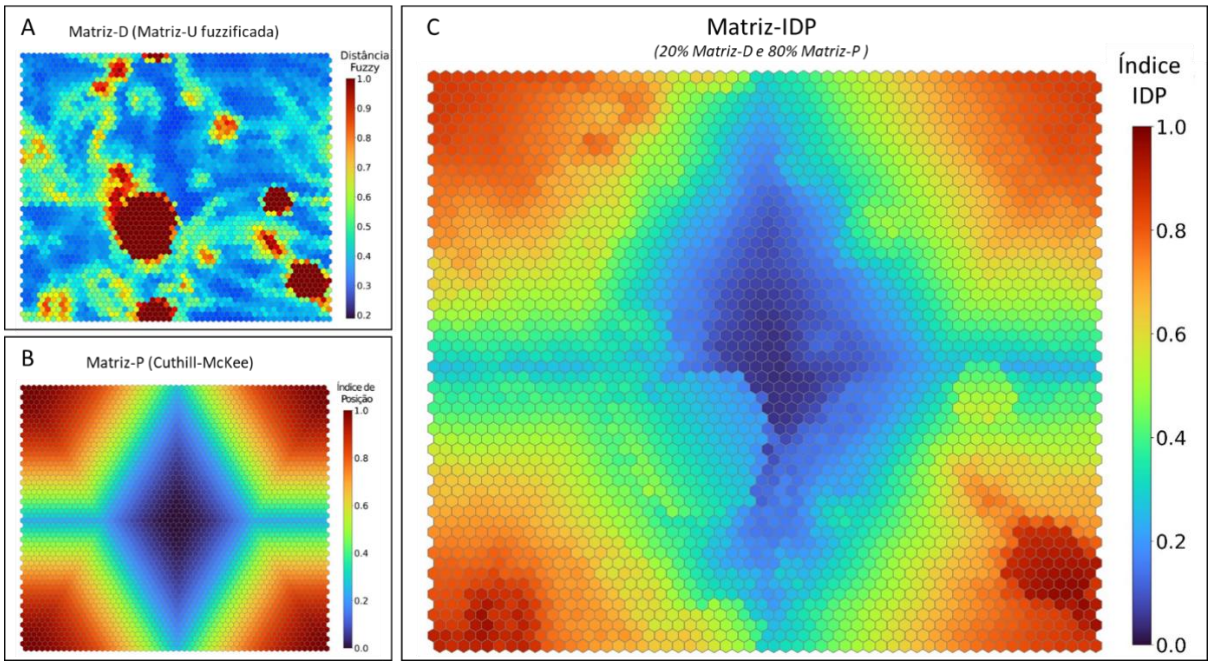


Fonte: Autoria própria.

4.1.2 Matriz Integrada de Distância e Posição (Matriz-IDP)

A Matriz-IDP (Figura 25) associou as informações de distância fuzzificadas, gerada pela Matriz-D, com as relações de posição, dada pela Matriz-P.

Figura 25 – (A) Matriz-D, obtida pela fuzzificação da Matriz-U a partir da função sigmoidal, para propiciar maior contraste dos dados de dissimilaridade; (B) Matriz-P, obtida a partir da aplicação do algoritmo Cuthill-McKee em coloração sintética, com periodicidade preservada; (C) Matriz-IDP, obtida a partir da ponderação da Matriz-D com a Matriz-P.

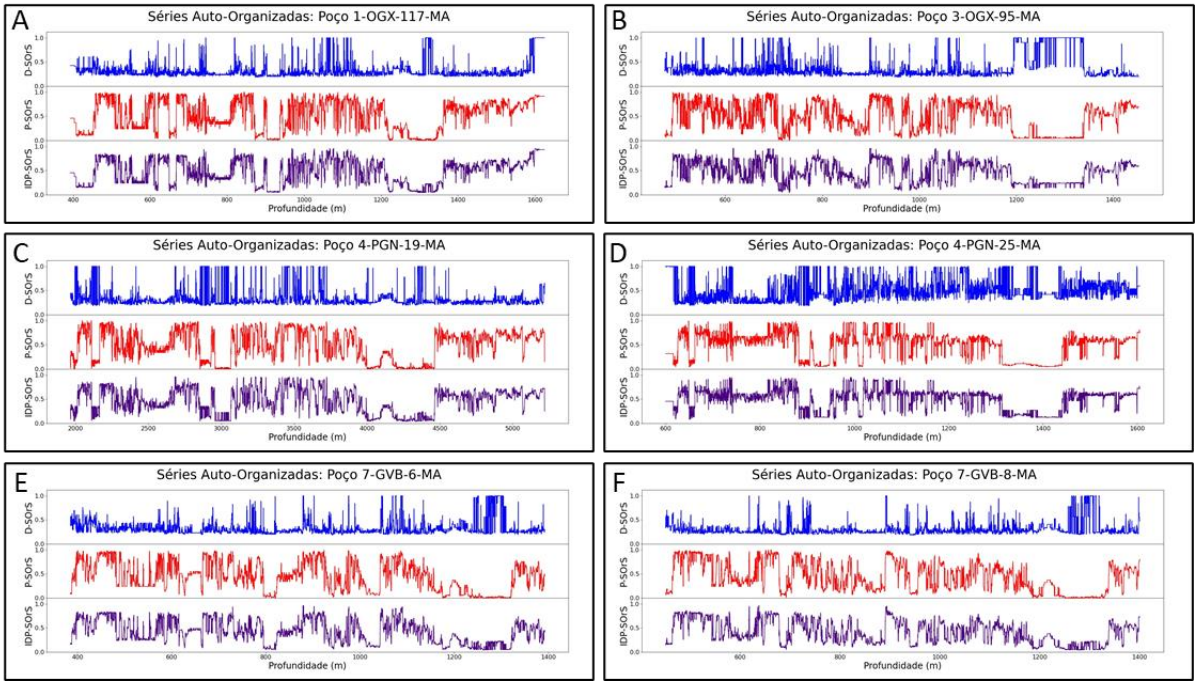


Fonte: Autoria própria.

4.1.3 Séries Auto-Organizáveis

O retorno dos índices gerados pela Matriz-D, Matriz-P e Matriz-IDP ao espaço original, em profundidades, deu origem às séries temporais D-SOrS, P-SOrS e IDP-SOrS. Estas três séries foram performadas para cada um dos 30 poços, e serviram como base para as análises de similaridade sequenciais. A Figura 26 mostra as séries desenvolvidas para a Bacia do Parnaíba.

Figura 26 – Séries Auto-Organizadas D-SOrS (azul), P-SOrS (vermelho); e IDP-SOrS (rôxo) desenvolvidas para os poços da Bacia do Parnaíba: (A) 1-OGX-117-MA; (B) 3-OGX-95-MA; (C) 4-PGN-19-MA; (D) 4-PGN-25-MA; (E) 7-GVB-6-MA; e (F) 7-GVB-8-MA.



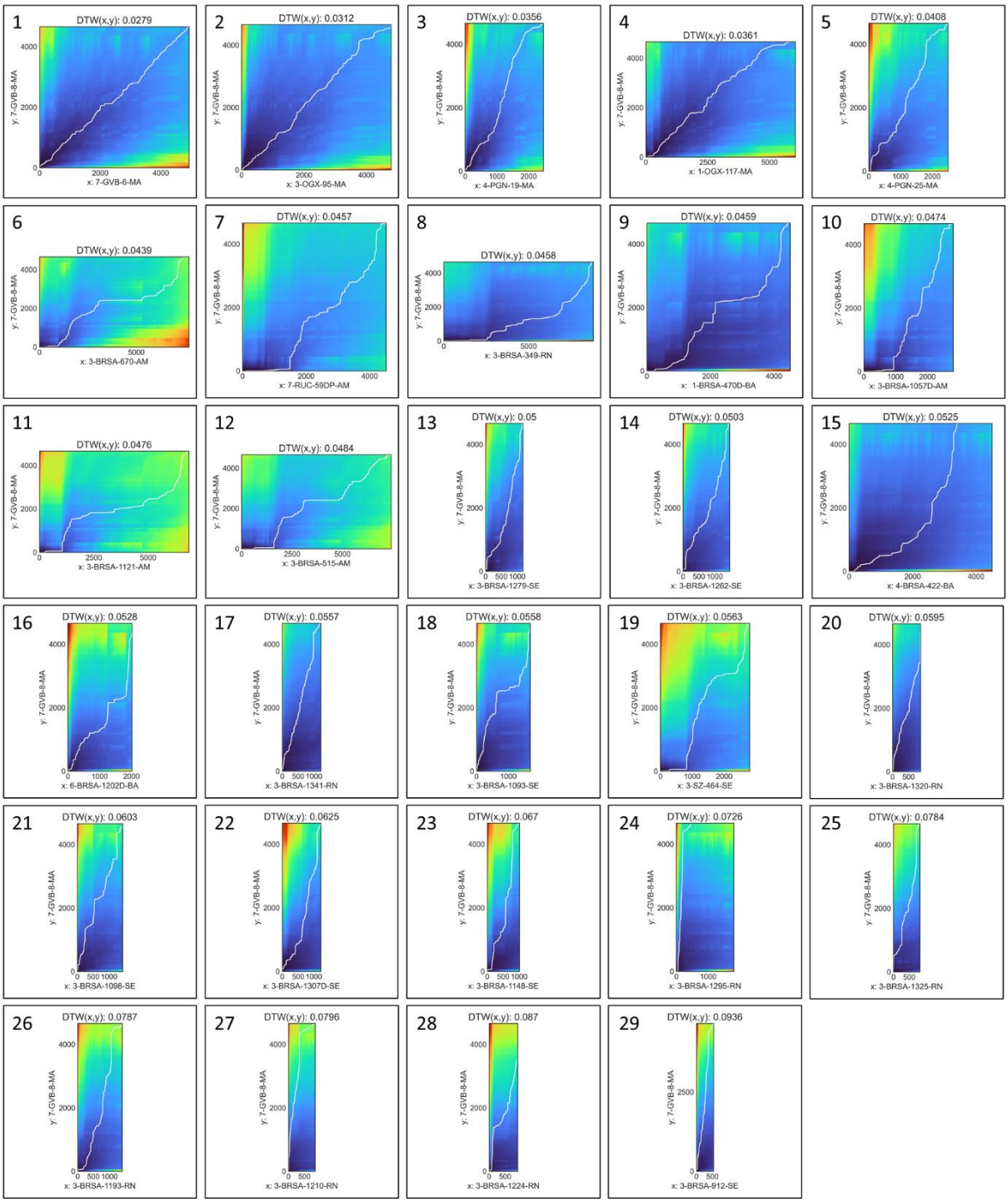
Fonte: Autoria própria.

4.2 Similaridade entre Poços de Petróleo

4.2.1 Análises quanto ao custo de alinhamento DTW

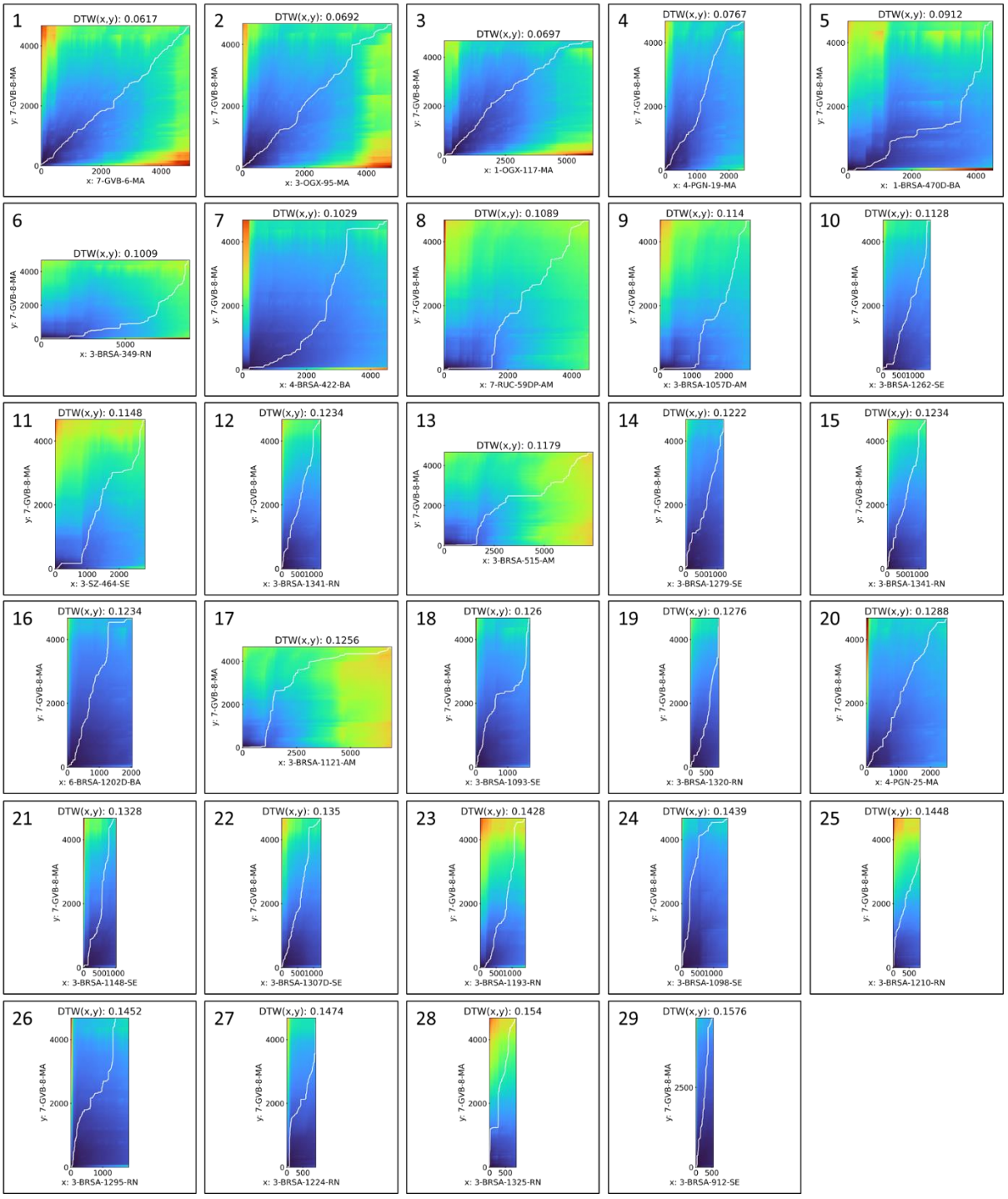
A similaridade entre os poços de petróleo investigados resultou em 435 matrizes de custo de alinhamento DTW para cada uma das três métricas de comparação analisadas, a saber: (i) Comparação entre poços pelo perfil síntese da SOrS-IDP; (ii) Comparação entre poços pelo vetor resultante entre as séries SOrS-D e SOrS-P medida pela distância Euclidiana; (iii) Comparação entre poços pelo vetor resultante entre as séries SOrS-D e SOrS-P medida pela distância Manhatan. Os resultados destas três métricas totalizaram 1.305 matrizes de custo de alinhamento DTW. No entanto, para efeito de compreensão, serão apresentados na Figura 27, Figura 28 e Figura 29, exclusivamente, as matrizes do poço *benchmark* 7-GVB-8-MA em comparação a todos os demais poços analisados.

Figura 27 – Matriz de custo de alinhamento com traçado do caminho de melhor custo para a comparação da IDP-SOrS entre o poço 7-GVB-8-MA e todos os demais poços



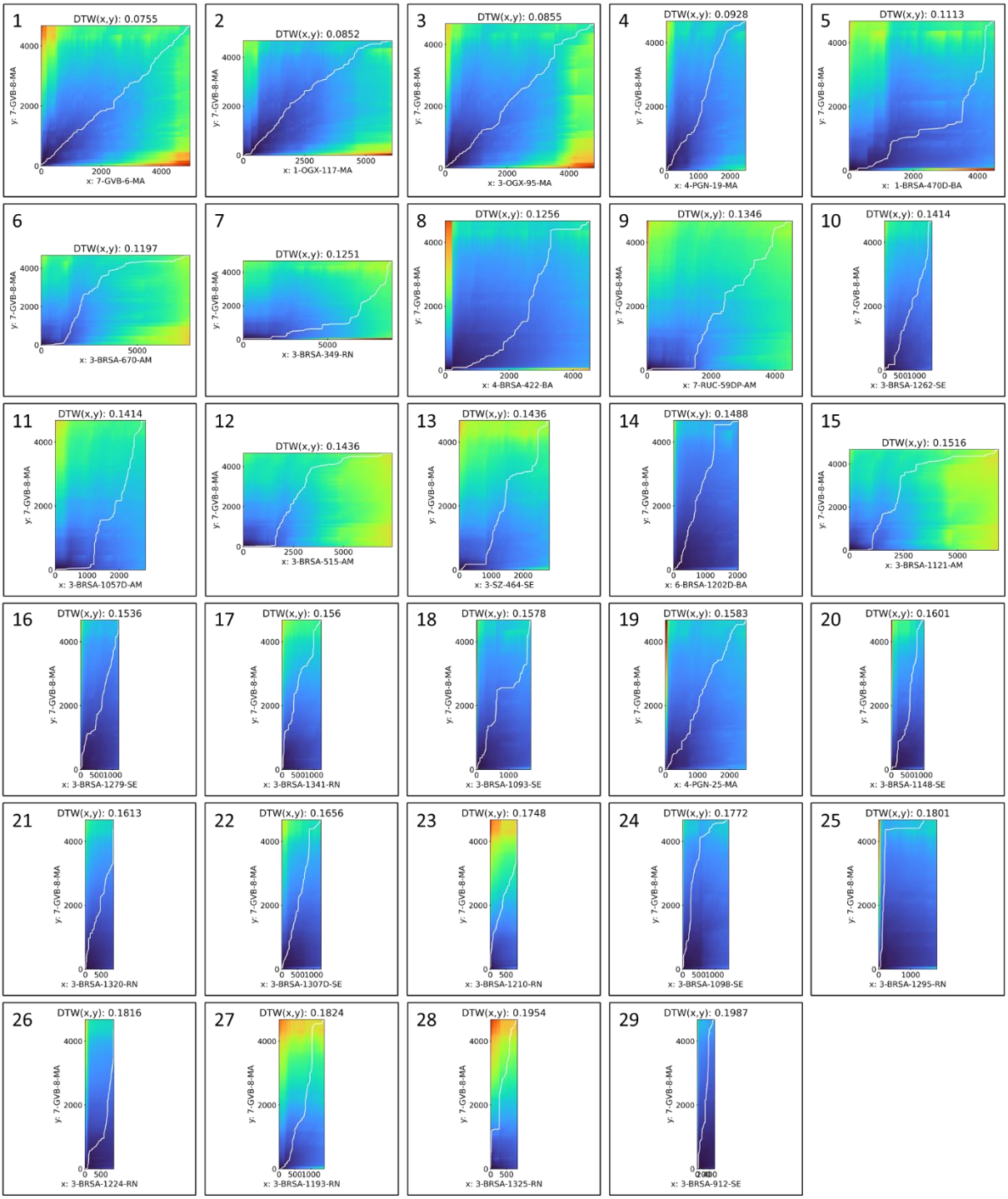
Fonte: Autoria própria.

Figura 28 – Matriz de custo de alinhamento com traçado do caminho de melhor custo para a comparação vetorial das séries D-SOrS e P-SOrS entre o poço 7-GVB-8-MA e todos os demais poços, tendo como base a distância Euclidiana



Fonte: Autoria própria.

Figura 29 – Matriz de custo de alinhamento com traçado do caminho de melhor custo para a comparação vetorial das séries D-SOrS e P-SOrS entre o poço 7-GVB-8-MA e todos os demais poços, tendo como base a distância Manhatan

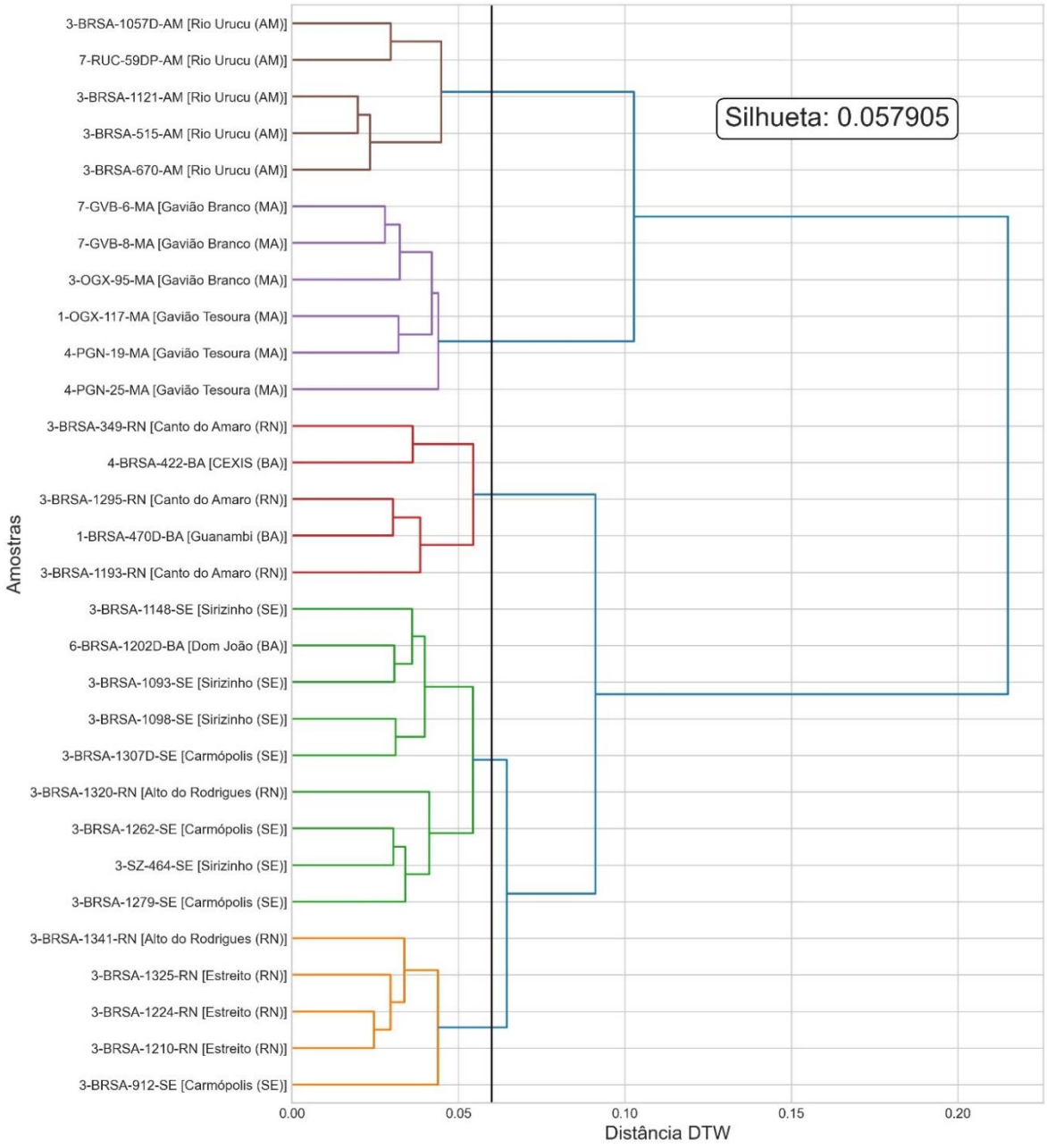


Fonte: Autoria própria.

4.2.2 Clusterização hierárquica do custo de alinhamento DTW

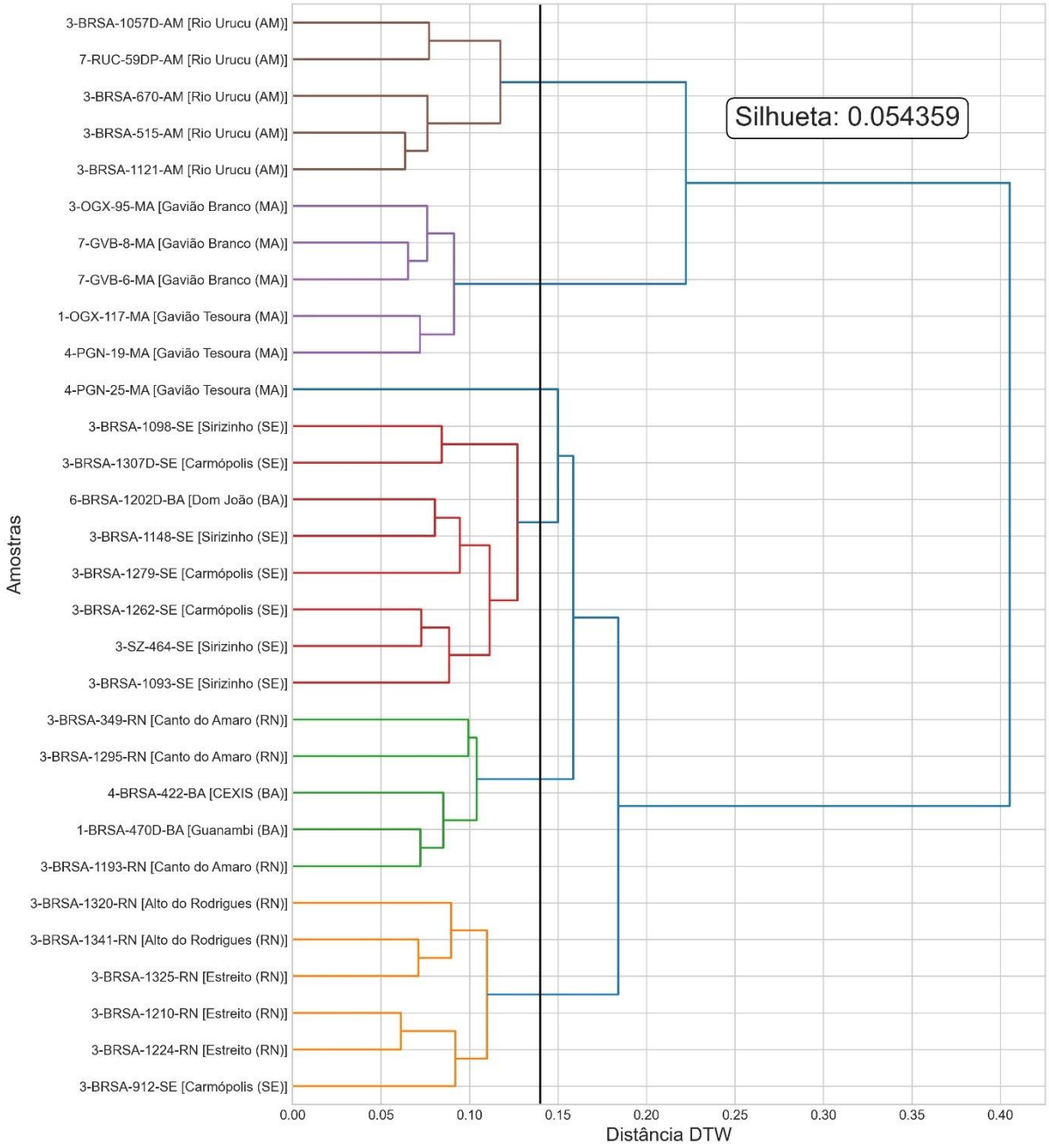
A avaliação do custo de alinhamento DTW por clusterização hierárquica resultou em três dendrogramas, tendo como base as diferentes medidas aplicadas às SOrS. Na Figura 30 observa-se o dendrograma obtido a partir da IDP-SOrS ponderada com 20% D e 80% P. Já a Figura 31 apresenta o dendrograma resultante da análise vetorial entre a D-SOrS e a P-SOrS medida com a distância Euclidiana. Outro dendrograma está apresentado na Figura 32, obtida pela análise vetorial entre a D-SOrS e a P-SOrS medidos com a distância Manhatan

Figura 30 – Clusterização hierárquica dos poços desenvolvida a partir dos custos de alinhamento obtidos pela IDP-SOrS ponderada a 20% D e 80% P.



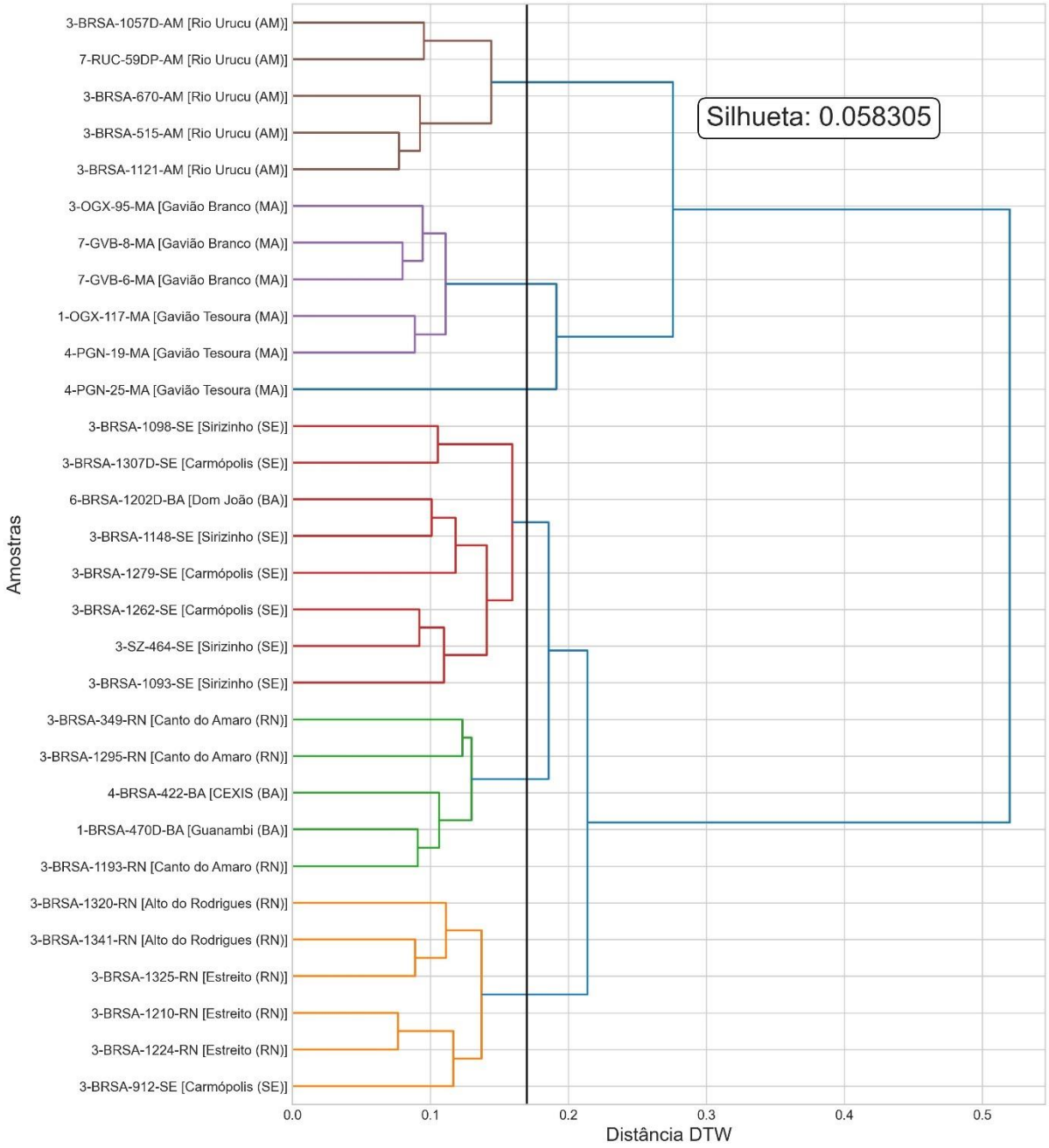
Fonte: Autoria própria.

Figura 31 – Clusterização hierárquica dos poços desenvolvida a partir dos custos de alinhamento obtidos pela análise vetorial entre a D-SOrS e a P-SOrS medida com a distância Euclidiana.



Fonte: Autoria própria.

Figura 32 – Clusterização hierárquica dos poços desenvolvida a partir dos custos de alinhamento obtidos pela análise vetorial entre a D-SOrS e a P-SOrS medida com a distância Manhatan.



Fonte: Autoria própria.

Ao observarmos os valores de silhueta, compreendemos que os três resultados são muito próximos entre si. A silhueta de 0,058 obtida pela clusterização hierárquica aplicada às análises vetoriais da P-SOrS e D-SOrS com distância Manhatan mostraram-se levemente mais adequadas. Essa observação ocorre em função de ser, dentre as três silhuetas analisadas, aquela que está mais próxima ao valor 1. No entanto, a silhueta de 0,057 obtida pela IDP-SOrS denota uma proximidade grande do resultado numericamente mais aderente.

Nas três formas de clusterização, os grupos são, majoritariamente, formados por poços de uma mesma bacia sedimentar. Os poços das bacias do Solimões (AM) e Parnaíba (MA) foram totalmente agrupados de forma correta. Vale ressaltar que a clusterização obtida a partir da IDP-SOrS conseguiu agregar os poços de um mesmo campo de produção em ramificações hierárquicas específicas para ambas as bacias, o que reflete uma coerência dos resultados.

Em relação aos poços da Bacia Potiguar, estes ficaram em clusters mais próximos quando estudados a partir da análise vetorial DTW entre SOrS-D e SOrS-P com distância Manhatan. Embora seja uma observação sutil, esta foi uma diferença diagnóstica para a escolha da medida a ser utilizada nas análises de similaridade de poços que serão desenvolvidas a seguir.

Sete dos oito poços da bacias do Sergipe (SE) formaram um mesmo grupo nas três abordagens analisadas. Apenas um poço (3-BRSA-912-SE) não permaneceu neste mesmo grupo, denotando maior dissimilaridade nas três abordagens.

Dois dos três poços da Bacia do Recôncavo (BA) foram reunidos em um mesmo grupo nos três processos de clusterização. Esses dois poços, no entanto, formaram clusters com outros grupos das bacias Potiguar e Sergipe.

4.2.3 *Rankings* de similaridade entre perfis de poços

A análise dos *rankings* de similaridade entre poços aqui desenvolvida parte do princípio da correlação espacial, que pressupõe que amostras mais próximas entre si no espaço georreferenciado são mais similares do que aquelas que se encontram mais distantes. Vale ressaltar que, nas aplicações de perfilagem de poços, há de se considerar, ainda, a variabilidade do empilhamento litoestratigráfico, o contexto deposicional, além das diferenças de profundidade entre os poços. De forma simplificada, neste trabalho, consideraremos os contextos deposicionais e litoestratigráficos como relacionados às bacias sedimentares e campos de produção relacionados aos poços analisados. É importante ressaltar que bacias localizadas às margens do continente possuem contextos geológicos e empilhamentos estratigráficos parecidos. Portanto, de antemão, é de se esperar que os poços de das bacias Potiguar, do Sergipe e do Recôncavo apresentem contextos litoestratigráficos mais similares entre si do que as bacias intracontinentais do Solimões e Parnaíba. Esse entendimento deverá ser considerado na análise dos *rankings* desenvolvida a seguir.

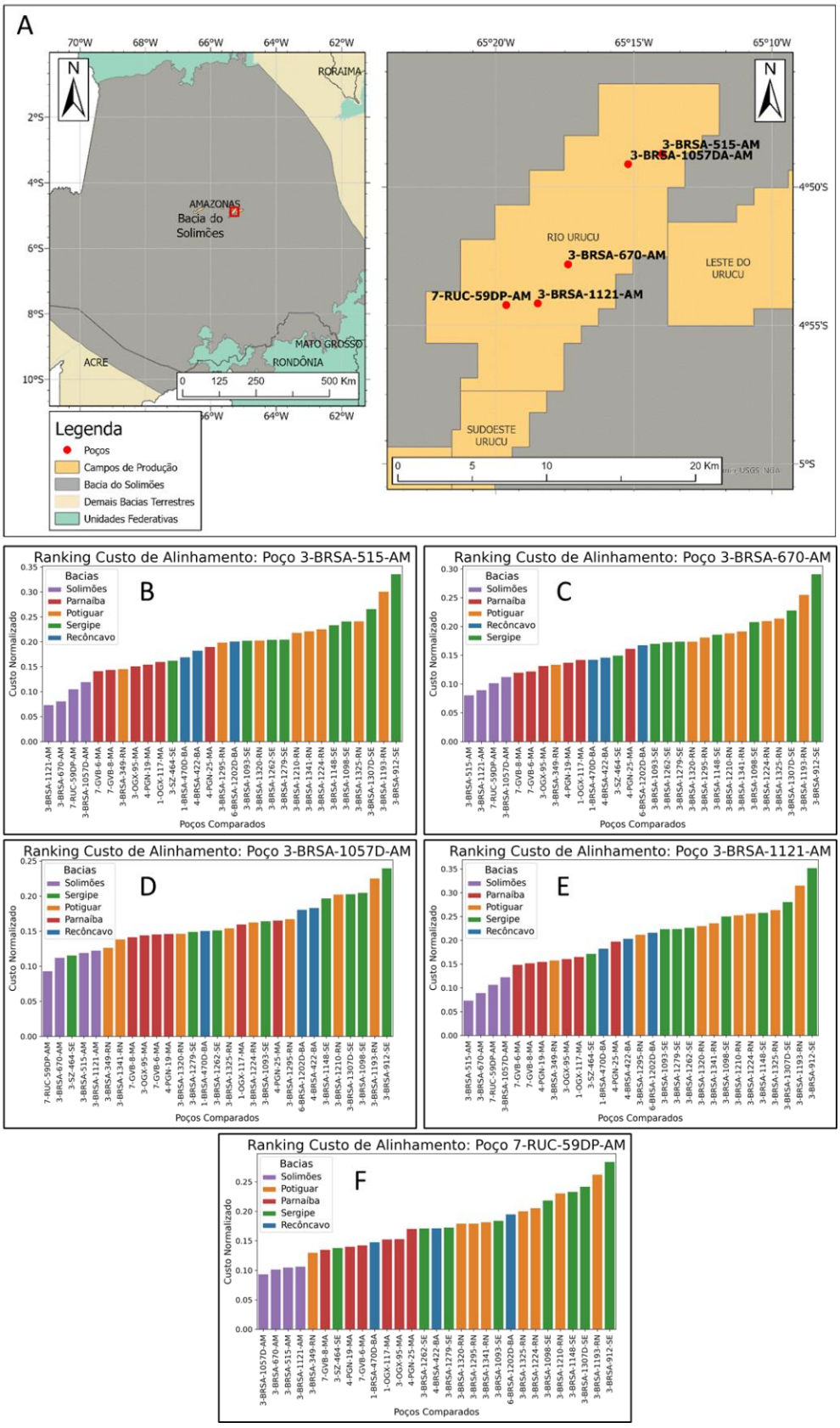
4.2.3.1 Bacia do Solimões

Os poços da Bacia do Solimões apresentaram baixo custo normalizado de alinhamento DTW entre si. Todos estes poços estão em um mesmo campo de produção (Rio Urucu), conforme a Figura 33A. O primeiro poço analisado, 3-BRSA-515-AM (1809,5 m), Figura 33B, apresenta maior similaridade, respectivamente, aos poços 3-BRSA-1121-AM (1427,4 m) e 3-BRSA-670-AM (1573,1 m). É importante ressaltar que o poço mais proximal a este primeiro poço seria o 3-BRSA-1057DA-AM (571,1m). No entanto, a profundidade rasa deste poço em relação aos demais fez com que o custo de alinhamento fosse mais elevado do que os demais poços da mesma bacia, com profundidades maiores. Notou-se, portanto, que profundidades mais próximas entre si possibilitaram menor custo de alinhamento. Isso ocorre, provavelmente, pelo aumento de correspondências entre sequências deposicionais atravessadas pelas informações de perfilagem dos poços.

Ao observarmos o *ranking* relacionado ao poço de menor profundidade desta bacia, 3-BRSA-1057DA-AM (575,1 m), Figura 33D, notamos que ele apresentou o menor custo de alinhamento relacionado ao poço com a segunda menor profundidade da mesma bacia, 7-RUC-59DP-AM (922,5 m). Isso reforça a compreensão de que a profundidade, ou o intervalo de tempo abrangido pela série temporal, influencia diretamente no cálculo do custo de alinhamento DTW. Este mesmo poço apresentou como terceiro menor custo de alinhamento um poço da Bacia de Sergipe, 3-SZ-464-SE (572,2 m) com contexto deposicional diferente. No entanto, compreende-se que o custo de alinhamento relativamente baixo ocorreu, sobretudo, em função das profundidades equivalentes.

É importante ressaltar que os poços mais profundos, 3-BRSA-515-AM (1809,5 m) e BRSA-670-AM (1573,1 m), após os poços da própria Bacia do Solimões, apresentaram maior similaridade em relação aos poços da Bacia do Parnaíba. Ambas as bacias são intracontinentais e, portanto, apresentam contextos geológicos mais similares. Isso denota, portanto, a coerência das informações obtidas.

Figura 33 – (A) Disposição espacial dos poços da Bacia do Solimões; Rankings de similaridade baseado no custo de alinhamento para os poços: (B) 3-BRSA-515-AM; (C) 3-BRSA-670-AM; (D) 3-BRSA-1057DA-AM; (E) 3-BRSA-1121-AM; e (F) 7-RUC-59DP-AM



Fonte: Autoria própria.

4.2.3.2 Bacia do Parnaíba (MA)

Os poços analisados na Bacia do Parnaíba (Figura 34) apresentam profundidades variando entre 5254,2 a 980,1 m. Exceto pelo poço 4-PGN-25-MA (1624,9 m), todos os demais poços apresentaram menores custos de alinhamento DTW relacionado aos demais poços da mesma bacia.

Quando analisados quanto às similaridades dentro de um mesmo campo de produção, os poços apresentaram como menor custo de alinhamento, dado pela primeira posição no *ranking*, um dos poços imediatamente vizinhos, do mesmo campo de produção. A exceção, novamente, ocorreu pelo poço 4-PGN-25-MA (1624,9 m).

Em relação às correspondências dos poços em bacias externas, não foi observado um padrão comum. Isso pode acontecer pelo fato da Bacia do Parnaíba representar um contexto intracratônico com intensos episódios de vulcanismo. A presença de rochas vulcânicas afere variações abruptas nas unidades litoestratigráficas, caracterizadas por composições mineralógicas bem distintas, que rompem a sequência deposicional. Embora as rochas vulcânicas sejam bem marcadas pelas SOrS, sua presença pode implicar em custos de alinhamento DTW maiores, sobretudo em contextos geológicos diferentes.

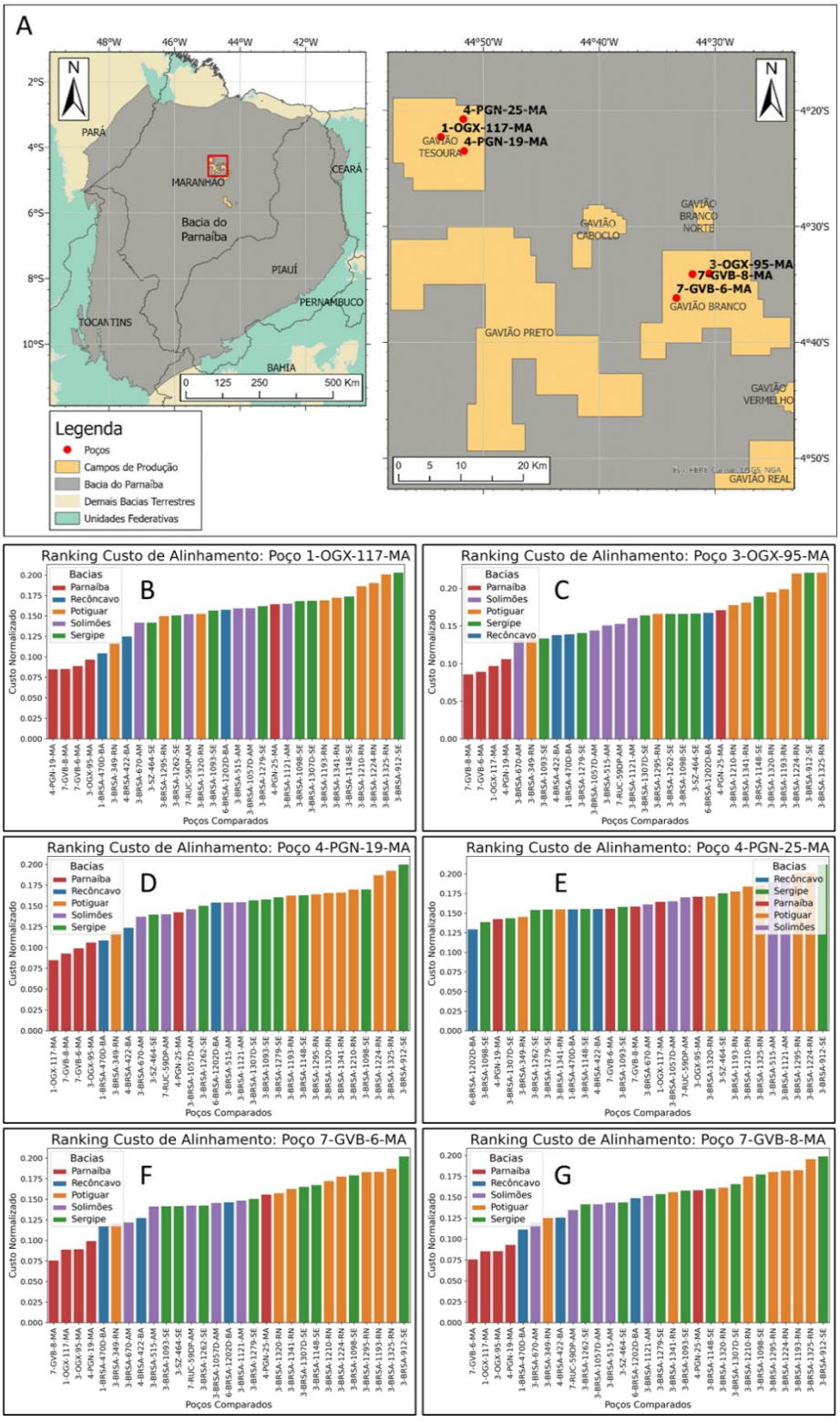
4.2.3.3 Bacia Potiguar

Dos oito poços analisados na Bacia Potiguar quatro destes, a saber: 3-BRSA-1210-RN (1492,8 m), 3-BRSA-1224-RN (437,8 m), 3-BRSA-1325-RN (492,7 m), 3-BRSA-1341-RN (619,9 m) apresentaram menor custo de alinhamento DTW com poços da mesma bacia. Os demais poços apresentaram menor custo ranqueado para com poços da Bacia do Recôncavo.

Quanto à similaridade em um mesmo campo de produção, os três poços localizados no campo Estreito apresentaram baixos custos de alinhamento entre si, indicando maior similaridade. Os poços localizados no campo Alto do Rodrigues também apresentaram similaridades para com os poços do mesmo campo, bem como para com poços do campo Estreito. É importante ressaltar, ao se observar a Figura 35 A, que ambos os campos de produção se encontram em um mesmo lineamento estrutural.

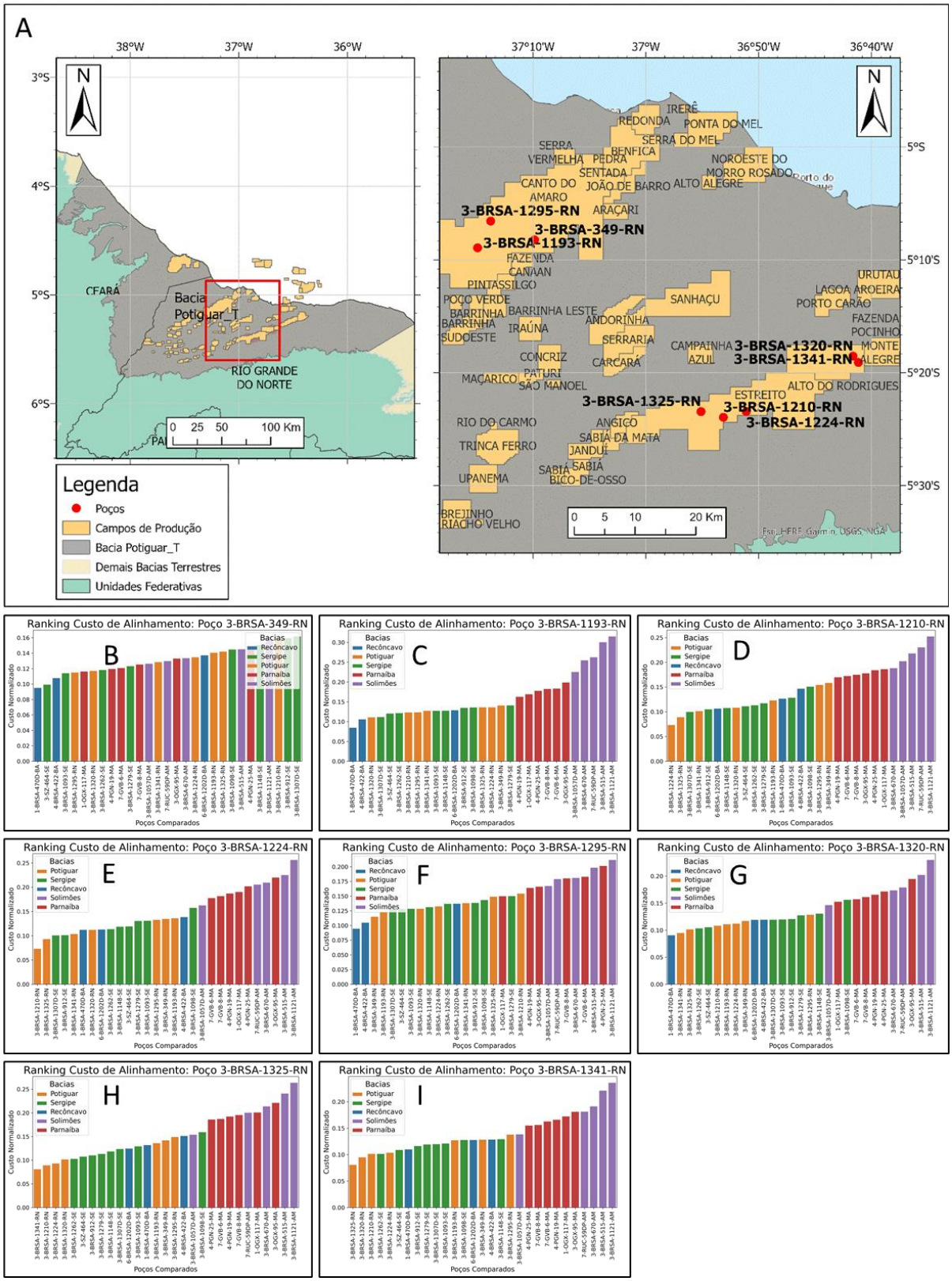
O poço que obteve os maiores custos de alinhamento, portanto menores similaridades aos poços da mesma bacia, foi o 3-BRSA-349-RN (1794,4 m). A profundidade deste poço, portanto, está bem acima da profundidade média dos poços desta bacia.

Figura 34 – (A) Disposição espacial dos poços da Bacia do Parnaíba; Rankings de similaridade baseado no custo de alinhamento para os poços: (B) 1-OGX-117-MA; (C) 3-OGX-95-MA; (D) 4-PGN-19-MA; (E) 4-PGN-25-MA; (F) 7-GVB-6-MA; e (G) 7-GVB-8-MA



Fonte: Autoria própria.

Figura 35 – (A) Disposição espacial dos poços da Bacia Potiguar; Rankings de similaridade baseado no custo de alinhamento para os poços: (B) 3-BRSA-349-RN; (C) 3-BRSA-1193-RN; (D) 3-BRSA-1210-RN; (E) 3-BRSA-1224-RN; (F) 3-BRSA-1295-RN; (G) 3-BRSA-1320-RN; (H) 3-BRSA-1325-RN; e (I) 3-BRSA-1341-RN



Fonte: Autoria própria.

4.2.3.4 Bacia do Sergipe

Cinco dos oito poços analisados na Bacia do Sergipe (Figura 36) apresentaram menor custo de alinhamento DTW associado a poços da mesma bacia sedimentar: 3-BRSA-1093-SE (886,7 m), 3-BRSA-1098-SE (760 m), 3-BRSA-1262-SE (839,5 m), 3-BRSA-1279-SE (723,5 m) e 3-SZ-464-SE. Dos três demais poços analisados, dois registraram menores custos associados a poços da Bacia do Recôncavo, 3-BRSA-912-SE (246,3 m) e 3-BRSA-1148-SE (609 m), e um à Bacia do Potiguar, 3-BRSA-1307D-SE (669,2 m).

Nesta bacia os campos de produção estão muito próximos entre si e não foi possível observar tendência relacionada à similaridade de poços em um mesmo campo de produção.

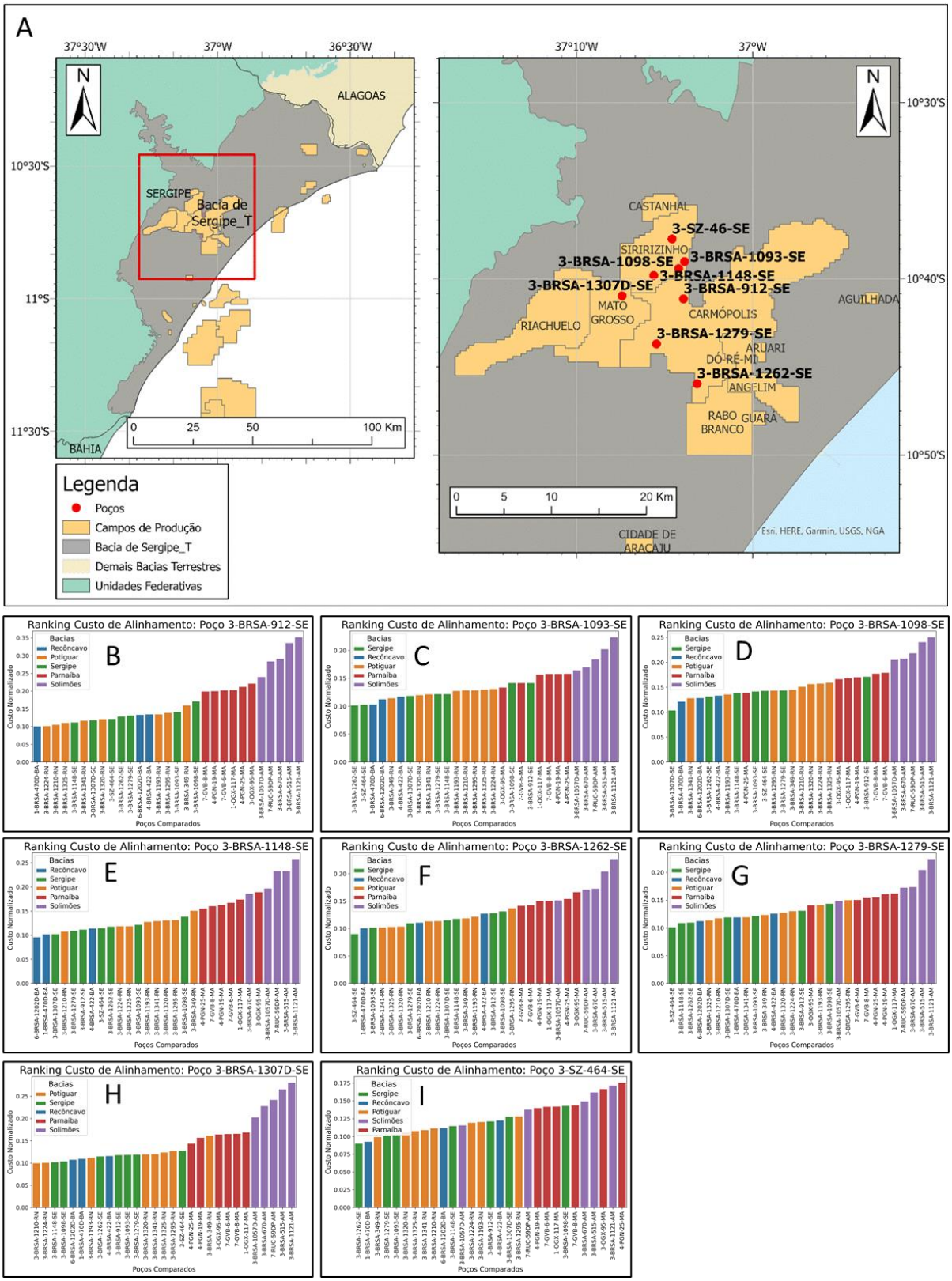
Uma observação relevante é que os maiores custos de alinhamento DTW para sete dos oito poços desta bacia estão relacionados à Bacia do Solimões. Para um dos poços os custos mais elevados estão relacionados a poços da Bacia do Parnaíba seguido pela Bacia do Solimões.

4.2.3.5 Bacia do Recôncavo

Os três poços analisados desta bacia pertencem a campos de produção distintos, conforme a Figura 37, e possuem profundidades diferentes entre si, a saber: 1-BRSA-470D-BA (921,1 m), 4-BRSA-422-BA (1810 m) e 6-BRSA-1202D-BA (1427,5 m). Dos três, apenas o poço 4-BRSA-422-BA (1810 m) obteve menor custo de alinhamento DTW, ou maior similaridade, associado a um poço da mesma bacia. o 1-BRSA-470D-BA (921,1 m). O poço 6-BRSA-1202D-BA (1427,5 m) obteve maior similaridade com cinco dos poços da Bacia do Sergipe e com um dos poços da Bacia Potiguar.

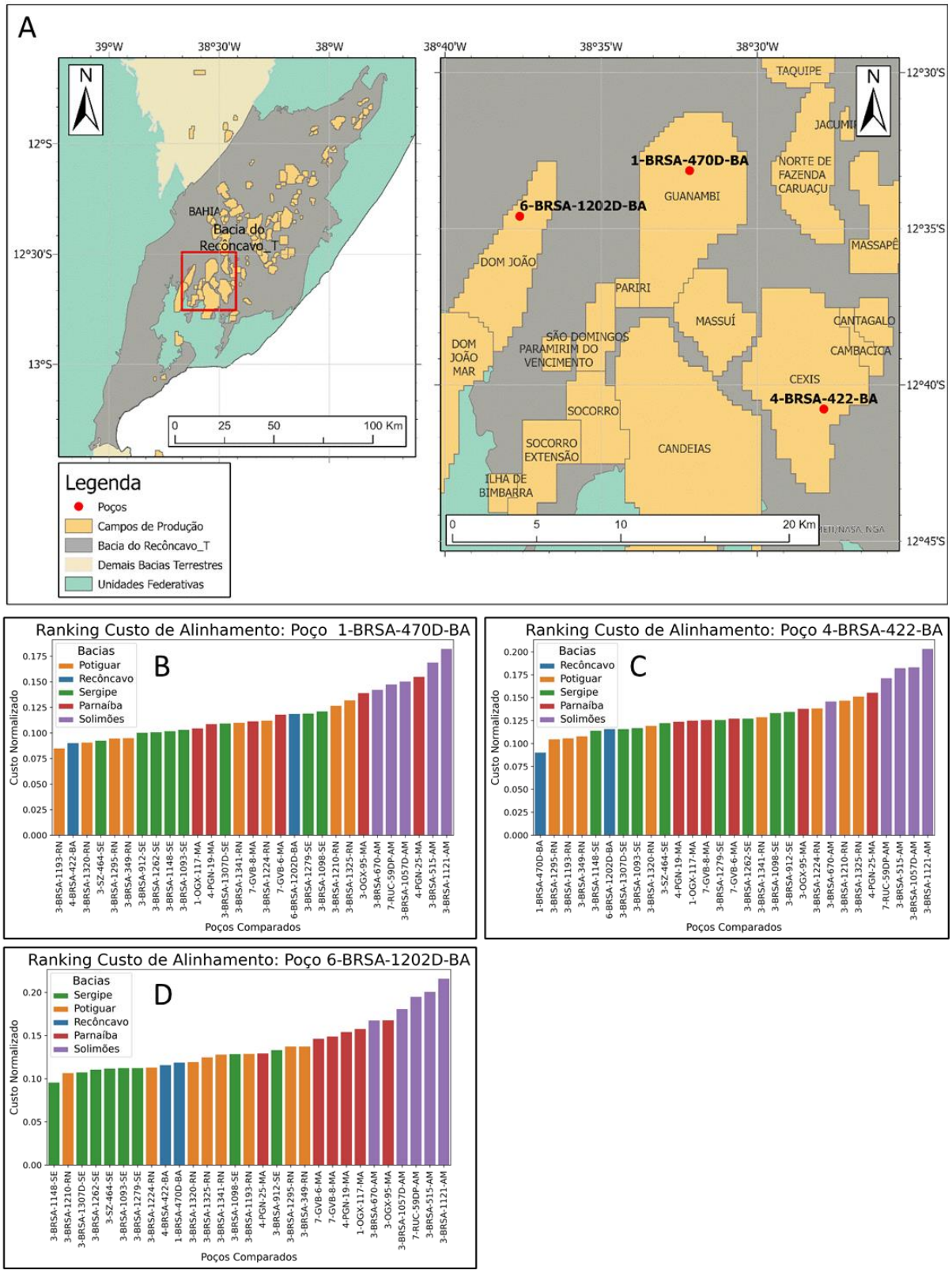
O ranqueamento dos poços evidenciou uma grande similaridade dos três poços para com as bacias do Sergipe e Potiguar, em contraste com os maiores custos de alinhamento para as bacias do Solimões e Parnaíba.

Figura 36 – (A) Disposição espacial dos poços da Bacia do Sergipe; Rankings de similaridade baseado no custo de alinhamento para os poços: (B) 3-BRSA-912-SE; (C) 3-BRSA-1093-SE; (D) 3-BRSA-1098-SE; (E) 3-BRSA-1148-SE; (F) 3-BRSA-1262-SE; (G) 3-BRSA-1279-SE; (H) 3-BRSA-1307D-SE; e (I) 3-SZ-464-SE



Fonte: Autoria própria.

Figura 37 – (A) Disposição espacial dos poços da Bacia do Recôncavo; Rankings de similaridade baseado no custo de alinhamento para os poços: (B) 1-BRSA-470D-BA; (C) 4-BRSA-422-BA; e (D) 6-BRSA-1202D-BA



Fonte: Autoria própria.

4.3 Discussões sobre Similaridade entre Poços e Bacias Sedimentares

A similaridade entre poços de petróleo, nesta pesquisa, resultou da comparação entre os perfis síntese obtidos a partir das SOrS. Vale ressaltar que as SOrS partiram da auto-organização desenvolvida a partir de uma suíte básica de perfis de poços, tratados aqui como séries temporais. Dessa maneira, há de se compreender que as variáveis inseridas como entrada ao sistema de auto-organização poderão ser cruciais às ferramentas de comparação. A comparação partindo da técnica DTW pode ser adotada com base em um único perfil, como por exemplo, o perfil de GR. No entanto, compreende-se que essa abordagem seria limitada, uma vez que as informações do poço viriam de uma única variável. Isso faz com que as SOrS sejam vantajosas em relação às comparações de similaridade a partir de uma única variável.

As observações relacionadas à aglomeração hierárquica produziram resultados muito próximos entre si. A escolha entre uma das três abordagens para a análise do ranqueamento foi tomada a partir da silhueta obtida pelos dendrogramas. Vale ressaltar, no entanto, que o custo de alinhamento normalizado para a IDP-SOrS apresentaram-se menores quando relacionados aos custos da análise vetorial entre D-SOrS e P-SOrS obtidas pela distância Euclidiana e Manhatan. Nota-se que, salvo pequenas variações, os resultados são muito parecidos entre si.

O contexto tectônico e litoestratigráfico das bacias sedimentares mais próximas à margem continental por vezes são semelhantes entre si. Com uma quantidade de amostras de 30 poços, é possível que estas bacias forneçam séries correspondentes entre si, o que traria limitações às métricas baseadas no custo de alinhamento normalizado. É muito provável que, em uma base de dados com maior quantidade de poços, os ranqueamentos de similaridade obedecessem à uma lógica com resultados mais aderentes.

5 CONCLUSÕES

A busca por medidas de similaridade entre poços de petróleo representa uma demanda recorrente do setor de óleo e gás. Entretanto, as soluções de abordagem para a problemática necessitavam de uma metodologia capaz de preservar as características intrínsecas dos perfis. Os resultados aqui apresentados proporcionaram a geração de um perfil síntese, eficiente em questão de processamento, e escalabilidade para a busca em bancos de dados. A presença dessas duas características, eficiência e escalabilidade, é de fundamental relevância estratégica nas atividades de exploração e produção de hidrocarbonetos, especialmente nos estágios iniciais que envolvem a pesquisa e exploração de um potencial campo petrolífero.

A sistemática de criação das SOrS, propostas nesta pesquisa, permitiram a análise de perfis de poços, aqui tratados como séries temporais multivariadas, com base na auto-organização SOM. A geração da D-SOrS, P-SOrS e IDP-SOrS representando, respectivamente, índices de distância, posição, e uma ponderação entre distância e posição, propiciou a consequente reconstrução dessas séries auto-organizadas no domínio das profundidades originais relacionadas aos perfis. A partir destas séries auto-organizadas foi possível observar as similaridades entre um sinal representativo de todas as variáveis de treinamento. Contudo, essa comparação é fortemente dependente da capacidade do algoritmo SOM discretizar, no mapa auto-organizado, os dados de entrada. À vista disso a avaliação dos impactos dos parâmetros de treinamento SOM na assinatura unificada, como o tamanho do mapa de treinamento, ainda necessitam ser melhor explorados.

As SOrS viabilizaram assinaturas unificadas e comparáveis, com resultados coerentes e satisfatórios para os 30 poços selecionados analisados. As séries geradas preservaram a similaridade litoestratigráfica observada nestes poços. Além disso, as séries foram capazes de discretizar grupos de similaridade com significado geológico, conforme a medida do custo de alinhamento provida pelo algoritmo DTW. Vale ressaltar que a interpretabilidade das SOrS é fortemente dependente das variáveis escolhidas para o treinamento, uma vez que diferentes variáveis geram relações diferentes de posição e distância no processo de auto-organização.

A similaridade entre poços demonstrou uma clara separação entre bacias intracratônicas daquelas de margem continental. Os poços da Bacia do Solimões e Parnaíba foram muito bem agrupados e possibilitaram ranques de poços mais similares aos demais poços da própria bacia sedimentar. Já os poços das bacias de margem continental, geralmente associadas a sistemas de riftes, por vezes tiveram suas similaridades confundidas entre si, porém com maioria dos poços

registrando menor custo de alinhamento DTW aos poços da própria bacia ou, por vezes, do mesmo campo de produção.

O desenvolvimento das SOrS resultaram também em três matrizes para representação dos resultados dos Mapas Auto-Organizáveis, denominadas de Matriz-D, Matriz P e Matriz-IDP. A matriz-D trata de uma simples redistribuição dos valores de distância da Matriz-U dos mapas auto-organizáveis. Já a matriz-P levou em consideração a posição das instâncias no toroide de treinamento. Isso tornou possível a captura e preservação das variações relacionadas à organização espacial das instâncias de treinamento com relação à similaridade. Portanto, as SOrS foram capazes de captar, numericamente, as modulações relacionadas ao arranjo dos dados na matriz de neurônios de saída do algoritmo SOM. Essa abordagem inédita proporcionou um resultado menos disruptivo do que uma análise por agrupamentos da Matriz-U, tradicionalmente aplicada em mapas auto-organizáveis, e que se baseia unicamente nas informações de distância unificada. Neste cenário comumente aplicado, a análise por agrupamentos pode gerar fronteiras que não apresentem, necessariamente, um significado interpretativo. Dessa maneira, o método aqui desenvolvido possui potencial de transcender a aplicação apresentada no presente trabalho.

Diversos outros pontos devem ser explorados em futuros trabalhos, sobretudo com o intuito de tornar a medida de comparação de poços uma métrica. Entre esses pontos, temos a possibilidade de avaliar assinaturas parciais de poços em janelas deslizantes, com o objetivo de gerar correlações segmentadas dos poços. Considerando que o algoritmo DTW é unidirecional, o desenvolvimento de uma métrica também deve passar pela avaliação das comparações entre assinaturas de poços nos dois sentidos.

6 REFERÊNCIAS

AGRAWAL, R.; FALOUTSOS, C.; SWAMI, A. **Efficient similarity search in sequence databases**. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). **Anais**...Springer, Berlin, Heidelberg, out. 1993. Disponível em: <https://link.springer.com/chapter/10.1007/3-540-57301-1_5>

ALI, M. et al. Machine learning - A novel approach of well logs similarity based on synchronization measures to predict shear sonic logs. **Journal of Petroleum Science and Engineering**, v. 203, p. 108602, ago. 2021.

ANACONDA. **Numba: A High Performance Python Compiler**. Disponível em: <https://numba.pydata.org/>. Acesso em: 5 jan. 2022.

ANGWIN, M.; NELSOR, J. L.; SYRETT, B. C. **Technical database design**. NACE - International Corrosion Conference Series. **Anais**...1996.

BELLMAN, R.; KALABA, R. On adaptive control processes. **IRE Transactions on Automatic Control**, v. 4, n. 2, p. 1–9, 1958.

BERNDT, D.; CLIFFORD, J. Using dynamic time warping to find patterns in time series. **Workshop on Knowledge Knowledge Discovery in Databases**, v. 398, p. 359–370, 1994.

CASSISI, C. et al. Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining. In: **Advances in Data Mining Knowledge Discovery and Applications**. [s.l.] InTech, 2012.

CÉRÉGHINO, R.; PARK, Y.-S. Review of the Self-Organizing Map (SOM) approach in water resources: Commentary. **Environmental Modelling & Software**, v. 24, n. 8, p. 945–947, ago. 2009.

CHU, S. et al. **Iterative Deepening Dynamic Time Warping for Time Series**. Proceedings. **Anais**...Society for Industrial & Applied Mathematics (SIAM), abr. 2002. Disponível em: <https://epubs.siam.org/page/terms>

CORRADINI, A. **Dynamic time warping for off-line recognition of a small gesture vocabulary**. Proceedings of the International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. **Anais**...Institute of Electrical and Electronics Engineers Inc., 2001.

CORREA, O. L. S. **Petróleo. Noções Sobre Exploração, Perfuração, Produção e Microbiologia. 1a Reimpressão Revista**. Engenharia edição ed. Rio de Janeiro (RJ): Interciência, 2003.

CUTHILL, E.; MCKEE, J. **Reducing the bandwidth of sparse symmetric matrices**. Proceedings of the 1969 24th national conference. **Anais**...: ACM '69.New York, NY, USA: Association for Computing Machinery, 26 ago. 1969. Disponível em: <https://doi.org/10.1145/800195.805928>. Acesso em: 13 jan. 2022

DAVE, M. SQL and NoSQL Databases. **International Journal of Advanced Research in Computer Science and Software Engineering**, ago. 2012.

DINOV, I. D. **Data Science and Predictive Analytics: Biomedical and Health Applications using R**. Cham: Springer International Publishing, 2018.

DU, R. et al. A similarity measure recognized by morphological characteristics analysis of well logging curves: application to the knowledge domain of sandstone reservoir. **Arabian Journal of Geosciences**, v. 13, n. 18, p. 947, set. 2020.

EFRAT, A.; FAN, Q.; VENKATASUBRAMANIAN, S. Curve matching, time warping, and light fields: New algorithms for computing similarity between curves. **Journal of Mathematical Imaging and Vision**, v. 27, n. 3, p. 203–216, nov. 2007.

ESLING, P.; AGON, C. Time-series data mining. **ACM Computing Surveys**, v. 45, n. 1, p. 12:1-12:34, 7 dez. 2012.

FALOUTSOS, C.; RANGANATHAN, M.; MANOLOPOULOS, Y. **Fast subsequence matching in time-series databases**. Association for Computing Machinery (ACM), 1994. Disponível em: <https://www.researchgate.net/publication/2456367_Fast_Subsequence_Matching_in_Time-Series_Databases>

FESSANT, F.; MIDENET, S. Self-Organising Map for Data Imputation and Correction in Surveys. **Neural Computing & Applications**, v. 10, n. 4, p. 300–310, 1 abr. 2002.

FRACZEK, K.; PLECHAWSKA-WOJCIK, M. **Comparative Analysis of Relational and Non-relational Databases in the Context of Performance in Web Applications**. [s.l: s.n.].

GU, J.; JIN, X. **A Simple Approximation for Dynamic Time Warping Search in Large Time Series Database**. (E. Corchado et al., Eds.)Intelligent Data Engineering and Automated Learning – IDEAL 2006. **Anais**...: Lecture Notes in Computer Science.Berlin, Heidelberg: Springer, 2006.

HAMILTON, J. **Time Series Analysis**. Princeton, N.J: Princeton University Press, 1994.

HAYKIN, S. **Neural Networks and Learning Machines**. 3ª Edição ed. New York: Prentice Hall, 2008.

JATIN; BATRA, S. MONGODB Versus SQL: A Case Study on Electricity Data. In: [s.l: s.n.]. p. 297–308.

KAHVECI, T.; SINGH, A. **Variable length queries for time series data**. Proceedings - International Conference on Data Engineering. **Anais**...Springer, Berlin, Heidelberg, 2001. Disponível em: <https://link.springer.com/chapter/10.1007/11875581_101>

KAHVECI, T.; SINGH, A.; GÜREL, A. **Similarity searching for multi-attribute sequences**. Proceedings of the International Conference on Scientific and Statistical Database Management, SSDBM. **Anais**...IEEE Computer Society, 2002.

KALTEH, A. M.; HJORTH, P.; BERNDTSSON, R. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. **Environmental Modelling & Software**, v. 23, n. 7, p. 835–845, 1 jul. 2008.

KAMPER, H. **Dynamic Time Warping Application**, 2021.

KEOGH, E. J.; PAZZANI, M. J. **Scaling up dynamic time warping for datamining applications**. Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. **Anais**...: KDD '00.New York, NY, USA: Association for Computing Machinery, 1 ago. 2000. Disponível em: <https://doi.org/10.1145/347090.347153>. Acesso em: 13 jan. 2022

KILLICK, R.; FEARNHEAD, P.; ECKLEY, I. A. Optimal detection of changepoints with a linear computational cost. **Journal of the American Statistical Association**, v. 107, n. 500, p. 1590–1598, 7 jan. 2012.

KOHONEN, T. **Self-Organizing Maps: 30**. 3rd 2001 ed. edição ed. Berlin ; New York: Springer, 2000.

KOHONEN, T. **Self-Organizing Maps**. 3. ed. Berlin Heidelberg: Springer-Verlag, 2001.

KOSKELA, T. et al. **Temporal sequence processing using recurrent SOM**. 1998 Second International Conference. Knowledge-Based Intelligent Electronic Systems. Proceedings KES'98 (Cat. No.98EX111). **Anais**... In: 1998 SECOND INTERNATIONAL CONFERENCE. KNOWLEDGE-BASED INTELLIGENT ELECTRONIC SYSTEMS. PROCEEDINGS KES'98 (CAT. NO.98EX111). abr. 1998.

KRAWCZAK, M.; SZKATUŁA, G. An approach to dimensionality reduction in time series. **Information Sciences**, v. 260, p. 15–36, mar. 2014.

LI, H. et al. Adaptively constrained dynamic time warping for time series classification and clustering. **Information Sciences**, v. 534, p. 97–116, set. 2020.

MCCARTHY, J. et al. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. p. 13, 1955.

MCQUEEN, T. A. et al. **A Recurrent Self-Organizing Map for Temporal Sequence Processing**. (A. Lotfi, J. M. Garibaldi, Eds.)Applications and Science in Soft Computing. **Anais**...: Advances in Soft Computing.Berlin, Heidelberg: Springer, 2004.

MÜLLER, M. DTW-Based Motion Comparison and Retrieval. In: **Information Retrieval for Music and Motion**. [s.l.] Elsevier, 2007. v. 534p. 211–226.

MYERS, C.; RABINER, L. R.; ROSENBERG, A. E. Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, v. 28, n. 6, p. 623–635, 1980.

NIENNATTRAKUL, V.; RATANAMAHATANA, C. A. On clustering multimedia time series data using k-means and dynamic time warping. **Proceedings - 2007 International Conference on Multimedia and Ubiquitous Engineering, MUE 2007**, p. 733–738, 2007.

PARK, S. et al. Similarity search of time-warped subsequences via a suffix tree. **Information Systems**, v. 28, n. 7, p. 867–883, 2003.

PARK, Y.-S. et al. Community patterns of benthic macroinvertebrates collected on the national scale in Korea. **Ecological Modelling**, v. 203, n. 1–2, p. 26–33, abr. 2007.

PRIESTLEY, M. B. **Spectral Analysis and Time Series, Two-Volume Set: Volumes I and II**. Amsterdam Heidelberg: Academic Press, 1983.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking**. 1ª edição ed. Beijing Köln: O'Reilly Media, 2013.

RATANAMAHATANA, C. A.; KEOGH, E. **Three myths about dynamic time warping data mining**. Proceedings of the 2005 SIAM International Conference on Data Mining, SDM 2005. **Anais**...Society for Industrial and Applied Mathematics Publications, 2005. Disponível em: <https://epubs.siam.org/page/terms>

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. Englewood Cliffs, N.J: Prentice Hall, 1995.

SAKOE, H.; CHIBA, S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, v. 26, n. 1, p. 43–49, 1978.

SAMMOUR, M. et al. Modified dynamic time warping for hierarchical clustering. **International Journal on Advanced Science, Engineering and Information Technology**, v. 9, n. 5, p. 1481–1487, 2019.

SCHLUMBERGER. **Schlumberger Book: Log Interpretation Charts 2009**. [s.l: s.n.].

SELLEY, R. C.; SONNENBERG, S. A. **Elements of Petroleum Geology**. 3ª edição ed. [s.l.] Academic Press, 2014.

SENIN, P. Dynamic time warping algorithm review. **Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA**, v. 855, n. 1–23, p. 40, 2008.

SKIENA, S. S. **The Data Science Design Manual**. Cham: Springer International Publishing, 2017.

TRUONG, C.; OUDRE, L.; VAYATIS, N. Selective review of offline change point detection methods. **Signal Processing**, v. 167, 2 jan. 2018.

VESANTO, J.; ALHONIEMI, E. Clustering of the self-organizing map. **IEEE Transactions on Neural Networks**, v. 11, n. 3, p. 586–600, maio 2000.

WEI, W. W. S. **Multivariate Time Series Analysis and Applications**. 1ª edição ed. Hoboken, NJ: Wiley, 2019.