



UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE GEOCIÊNCIAS

RAUL EIGENHEER MELONI

MAPEAMENTO DE POTENCIAL MINERAL PARA DEPÓSITOS DE URÂNIO NA
PROVÍNCIA URANÍFERA DE LAGOA REAL USANDO MACHINE LEARNING

CAMPINAS

2024

RAUL EIGENHEER MELONI

MAPEAMENTO DE POTENCIAL MINERAL PARA DEPÓSITOS DE URÂNIO NA
PROVÍNCIA URANÍFERA DE LAGOA REAL USANDO MACHINE LEARNING

DISSERTAÇÃO APRESENTADA AO INSTITUTO
DE GEOCIÊNCIAS DA UNIVERSIDADE
ESTADUAL DE CAMPINAS PARA A OBTENÇÃO
DO TÍTULO DE MESTRE EM GEOCIÊNCIAS NA
ÁREA DE GEOLOGIA E RECURSOS NATURAIS

Orientador: Prof. Dr. Carlos Roberto de Souza Filho

ESTE EXEMPLAR CORRESPONDE À VERSÃO
FINAL DA DISSERTAÇÃO DEFENDIDA PELO
ALUNO RAUL EIGENHEER MELONI E
ORIENTADA PELO PROF. DR. CARLOS
ROBERTO DE SOUZA FILHO

CAMPINAS

2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Geociências
Marta dos Santos - CRB 8/5892

M492m Meloni, Raul Eigenheer, 1988-
Mapeamento de potencial mineral para depósitos de urânio na Província Uranífera de Lagoa Real usando machine learning / Raul Eigenheer Meloni. – Campinas, SP : [s.n.], 2024.

Orientador: Carlos Roberto de Souza Filho.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Geociências.

1. Aprendizado de máquina. 2. Modelamento de prospectividade mineral. 3. Minérios de urânio. I. Souza Filho, Carlos Roberto de, 1965-. II. Universidade Estadual de Campinas (UNICAMP). Instituto de Geociências. III. Título.

Informações Complementares

Título em outro idioma: Mineral potential mapping for uranium deposits in the Lagoa Real Uranium Province using machine learning

Palavras-chave em inglês:

Machine learning

Mineral potential mapping

Uranium ores

Área de concentração: Geologia e Recursos Naturais

Titulação: Mestre em Geociências

Banca examinadora:

Carlos Roberto de Souza Filho [Orientador]

Emilson Pereira Leite

Adalene Moreira Silva

Data de defesa: 20-08-2024

Programa de Pós-Graduação: Geociências

Identificação e informações acadêmicas do(a) aluno(a)

- ORCID do autor: <https://orcid.org/0000-0002-2025-1363>

- Currículo Lattes do autor: <http://lattes.cnpq.br/6082539385384727>



**UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE GEOCIÊNCIAS**

AUTOR: Raul Eigenheer Meloni

MAPEAMENTO DE POTENCIAL MINERAL PARA DEPÓSITOS DE URÂNIO NA
PROVÍNCIA URANÍFERA DE LAGOA REAL USANDO MACHINE LEARNING

ORIENTADOR: Prof. Dr. Carlos Roberto de Souza Filho

Aprovado em: 20 / 08 / 2024

EXAMINADORES:

Prof. Dr. Carlos Roberto de Souza Filho - Presidente

Prof. Dr. Emilson Pereira Leite

Prof^a. Dr^a. Adalene Moreira Silva

***A Ata de Defesa assinada pelos membros da Comissão Examinadora consta
no processo de vida acadêmica do aluno.***

Campinas, 20 de agosto de 2024.

BIOGRAFIA

Concluiu a graduação em Geologia no ano de 2012, na USP, com trabalho de conclusão de curso na área de metalogenia e prospecção mineral. Trabalhou com prospecção de ouro na indústria por cerca de um ano. Desde 2014, atua no Serviço Geológico do Brasil (SGB) em projetos voltados à avaliação de potencial mineral, envolvendo mapeamento, estudos de metalogênese e modelamento de potencial mineral. Liderou e participou de projetos no Cráton Amazônico por cerca de 7 anos. No SGB, trabalhou principalmente em projetos focados em ouro, cobre e urânio. Nos últimos anos, tem sua atuação voltada à aplicação de técnicas de análise de dados espaciais, incluindo machine learning, para estudos de potencial mineral.

RESUMO

Com o esgotamento progressivo dos depósitos minerais rasos, de mais fácil identificação, o uso de tecnologias de informação tem se tornado fundamental na pesquisa mineral. Métodos de modelamento de potencial mineral (MPM) em ambiente SIG têm sido amplamente aplicados na definição de áreas prospectáveis e alvos exploratórios. Esses métodos permitem integrar uma grande quantidade de dados espaciais e conhecimentos específicos de forma eficiente e dirigida. Entre os métodos de MPM guiados pelos dados, as técnicas de *machine learning* (ML) têm se mostrado promissoras devido à sua capacidade de modelar sistemas complexos, como são os sistemas minerais. Este trabalho teve o objetivo de avaliar a eficácia de diferentes técnicas de ML no modelamento da prospectividade para urânio na Província Uranífera de Lagoa Real (PULR), utilizando diferentes técnicas de ML. Usando uma abordagem de sistemas minerais, foram definidos os componentes críticos para a geração dos depósitos de urânio na PULR, que são do tipo Na-metassomáticos. Com o uso de técnicas de processamento de dados geoespaciais, tais componentes do sistema mineral foram convertidos em mapas evidenciais e então utilizados como variáveis preditivas do modelo de prospectividade. Foram empregadas técnicas de ML não-supervisionadas (*Self-organizing maps*, K-means e agrupamento hierárquico espacialmente restrito) para análise exploratória de dados, e delimitação de domínios geológico-geofísicos. O modelo de prospectividade final foi gerado com o algoritmo de classificação *XGBoost*. No mapa resultante (com scores de prospectividade variando entre 0 e 1), 95,1% das ocorrências conhecidas estão mapeadas em apenas 2,3% da área de estudo, com scores iguais ou superiores a 0,9; e 100% das ocorrências são encontradas dentro de uma fração de 4,7% da área, com scores iguais ou superiores a 0,5. A técnica SHAP (*Shapley additive explanations*) foi utilizada para a avaliação da influência de cada variável nas predições. Usando essa técnica, em conjunto com o conhecimento acerca do sistema mineral da PULR, foi possível propor hipóteses prospectivas para geração de alvos. Os resultados obtidos demonstraram o potencial das técnicas de ML, tanto não-supervisionadas como supervisionadas, para MPM em escala de distrito, com particular relevância para uso em áreas de estudo com mapeamento geológico limitado, como a PULR.

Palavras chave: Aprendizado de máquina; Modelamento de prospectividade mineral; Minérios de urânio.

ABSTRACT

With the progressive depletion of more easily identifiable shallow mineral deposits, the use of information technologies has become essential in mineral exploration. Mineral potential modeling (MPM) methods in a GIS environment have been widely applied in defining prospective areas and exploratory targets. These methods allow for the efficient and directed integration of a large amount of spatial data and specific knowledge. Among the data-driven MPM methods, machine learning (ML) techniques have shown promise due to their ability to model complex systems, such as mineral systems. This study aimed to evaluate the effectiveness of different ML techniques in modeling uranium prospectivity in the Lagoa Real Uranium Province (PULR), employing various ML techniques. Using a mineral systems approach, the critical components for the generation of uranium deposits in the PULR, which are of the Na-metasomatic type, were defined. Through geospatial data processing techniques, these mineral system components were converted into evidential maps and then used as predictive variables in the prospectivity model. Unsupervised ML techniques (self-organizing maps, K-means, and spatially constrained hierarchical clustering) were employed for exploratory data analysis and delineation of geological-geophysical domains. The final prospectivity model was generated using the *XGBoost* classification algorithm. In the resulting map (with prospectivity scores ranging from 0 to 1), 95,1% of the known occurrences are mapped within only 2,3% of the study area, with scores equal to or greater than 0,9; and 100% of the occurrences are found within a fraction of 4,7% of the area, with scores equal to or greater than 0,5. The SHAP technique (Shapley additive explanations) was used to evaluate the influence of each variable on the predictions. Using this technique, in conjunction with knowledge of the PULR mineral system, it was possible to propose prospective hypotheses for target generation. The results obtained demonstrated the potential of ML techniques, both unsupervised and supervised, for MPM at the district scale, with particular relevance for use in study areas with limited geological mapping, such as the PULR.

Keywords: Machine learning; Mineral potential mapping; Uranium ores.

LISTA DE ILUSTRAÇÕES

<i>Figura 1.1: Localização da área de estudo.</i>	20
<i>Figura 1.2: Contexto geológico regional da área de estudo (delimitada aproximadamente pelo retângulo vermelho) (adaptado de Cruz e Alkmim, 2006).</i>	23
<i>Figura 1.3 Mapa geológico da área de estudo. Adaptado de Souza et al. (2003).</i>	25
<i>Figura 1.4: Modelo geológico das mineralizações de U do tipo metassomáticas (adaptado de Dahlkamp, 2009).</i>	29
<i>Figura 2.1: Fluxograma do processamento dos dados aerogeofísicos (CBPM/CPRM, 2006).</i>	33
<i>Figura 2.2: Dados aerogeofísicos e modelo digital de terreno. Dados gamaespectrométricos: contagem total, potássio, urânio equivalente e tório equivalente. Produtos derivados de dados magnetométricos: campo magnético anômalo e gradiente total.</i>	33
<i>Figura 2.3: Fluxograma das filtrações realizadas nos dados aeromagnéticos para a obtenção das primeiras derivadas verticais (1Dz) das continuações ascendentes de: A) 15 km; B) 10 km; C) 5 km; e D) da 1Dz do dado magnetométrico original. UC: Upward continuation (continuação ascendente); CMA: Campo magnético anômalo.</i>	35
<i>Figura 2.4: Histogramas de distribuição dos dados aerogeofísicos (CT, eTh, eU, K e GT) e de MDT não transformados.</i>	36
<i>Figura 2.5: Histogramas de distribuição dos dados aerogeofísicos (CT, eTh, eU, K e GT) normalizados (box-cox) e escalonados (z-score).</i>	37
<i>Figura 2.6: Representação esquemática de um mapa SOM. Modificado de: http://www.statistics4u.com/fundstat_eng/cc_ann_kohonen.html.</i>	38
<i>Figura 2.7: Visualização em 3D de um mapa SOM com topologia toroidal.</i>	38
<i>Figura 2.8: A) Zona de vizinhança centralizada na célula vencedora “c” (BMU). A intensidade da cor laranja indica o quanto cada célula será atualizada na iteração t, e o raio $\sigma(t)$ indica até onde as células do mapa serão influenciadas. Adaptado de Moreira e Costa (1996). B) Função de vizinhança “bolha”. C) Função de vizinhança Gaussiana. Adaptado de Vesanto (1997).</i>	40

<i>Figura 2.9: Exemplo de execução do algoritmo K-means. A cada iteração, os centroides são recalculados (se movimentam), e a atribuição dos pontos aos grupos pode ser modificar. Adaptado de Aggarwal e Reddy (2016).</i>	44
<i>Figura 2.10: Ilustração esquemática do agrupamento do mapa SOM, produzindo 2 níveis de abstração do conjunto e dados. Adaptado de Vesanto e Alhoniemi (2000).</i>	45
<i>Figura 2.11: Exemplo de dendrograma produzido por agrupamento hierárquico. Na base se encontram os pontos de dados individuais (A, B, C, D, E) e, a cada nível da árvore, os pontos, ou grupos, se juntam para formar novos, e maiores, agrupamentos (modificado de Zaki e Meira Jr., 2020).</i>	46
<i>Figura 2.12: Ilustração da matriz de conexões com raio de 708 m utilizada neste trabalho, mostrando o número máximo de pontos candidatos a um agrupamento. A figura da esquerda mostra o exemplo de um ponto localizado em um canto da área, e a da direita, de um ponto localizado a mais do que 750 m das bordas da área de estudo.</i>	48
<i>Figura 2.13: Exemplo do uso de uma medida de validação de grupos (Davies-Bouldin index - DBI) para se descobrir o número ótimo de grupos. No exemplo da figura, o número ótimo de grupos é 6, que minimiza o DBI. Adaptado de Gazley et al. (2021).</i>	49
<i>Figura 2.14: Famílias de lineamentos interpretados na área de trabalho (Barros et al. 2021). As famílias são explicadas na Tabela 2.</i>	52
<i>Figura 2.15: Esquema gráfico da influência (Φ) de cada variável (z) no resultado da predição do modelo ($f(x)$) a partir do valor base $E[fz]$ (modificado de Lundberg e Lee, 2017).</i>	59
<i>Figura 3.1: Camadas de dados de entrada (contagem total, gradiente total, tório, potássio, urânio e modelo digital de terreno) e formato do mapa SOM treinado. Na figura, somente para fins de visualização, o MDT está fundido com as cores do mapa geológico e com a localização das ocorrências de urânio conhecidas.</i>	61
<i>Figura 3.2: Comparação das curvas de distribuição (estimativa de densidade kernel) das variáveis do banco de dados original (transformadas) e das células do mapa SOM.</i>	61

<i>Figura 3.3: A) Mapas das características do modelo SOM: matriz-U, distância média, número de vetores. B) Visualizações dos mapas de ativação dos 6 componentes do modelo SOM: tório, urânio, potássio, contagem total, gradiente total e modelo digital de terreno.</i>	<i>63</i>
<i>Figura 3.4: Mapas de ativação (assinaturas) de características geológicas mapeadas no modelo SOM.</i>	<i>64</i>
<i>Figura 3.5: Esquema ilustrativo do Counter Propagation Network (CPN), aplicado à unidade geológica Complexo Lagoa Real. As setas indicam o mapeamento dos pontos de dados da área de estudo para o modelo SOM, em determinadas células escolhidas como exemplo (imagem superior). O mapeamento reverso projeta a assinatura da unidade modelada no mapa da área (imagem inferior e à direita).</i>	<i>65</i>
<i>Figura 3.6: Esquema gráfico explicativo da análise estatística da variabilidade dos resultados do modelo SOM, e mapas resultantes, com a técnica CPN. O primeiro mapa (da esquerda) é um exemplo dos 100 mapas gerados. Os demais, são os mapas das estatísticas calculadas para os 100 modelos. CV: coeficiente de variação.</i>	<i>67</i>
<i>Figura 3.7: Histograma e função de densidade de probabilidade dos 100 resultados de 1 pixel do mapa. São indicados no histograma os seguintes parâmetros estatísticos: média, desvio padrão, e mediana.</i>	<i>68</i>
<i>Figura 3.8: Diagrama de dispersão, combinado com histogramas, das médias vs. coeficientes de variação dos pixels que apresentaram máximos maiores do que 0 no conjunto dos 100 modelos SOM. A linha vermelha indica o 90º percentil das médias,. A linha azul o valor de cv médio acima do 90º percentil das médias.</i>	<i>69</i>
<i>Figura 3.9: Índice DB vs. número de clusters (k) para os agrupamentos obtidos com SOM + K-means. Os melhores (menores) valores de DB foram alcançados com k = 11 e k = 145.</i>	<i>70</i>
<i>Figura 3.10: Medidas de prospectividade dos grupos com maior número de ocorrências, obtidos de agrupamentos por SOM + K-means, em função de k. São marcados os valores ótimos (k = 9 e k=82) obtidos pela observação dessas medidas e, para comparação, os obtidos com o índice DB (k = 11 e k = 145 - Figura 3.9).</i>	<i>70</i>
<i>Figura 3.11: Agrupamentos das células do mapa SOM, obtidos com k igual a 9, 11, 82 e 145.</i>	<i>71</i>
<i>Figura 3.12: Espacialização dos domínios obtidos pelo agrupamento das células SOM, com k igual a: 9 (A), 11 (B), 82 (C), 145 (D).</i>	<i>72</i>

<i>Figura 3.13: Índice DB vs. número de clusters (k) para os agrupamentos obtidos com AHER. O melhor (menor) valor de DB foi alcançado com k = 4.</i>	73
<i>Figura 3.14: Medidas de prospectividade dos grupos com maior número de ocorrências, obtidos de agrupamentos por AHER, em função de k. São marcados os valores ótimos (k = 7, k = 35 e k = 81) obtidos pela observação dessas medidas e, para comparação, o obtido com o índice DB (k = 4 - Figura 3.13).</i>	74
<i>Figura 3.15: Mapas dos domínios obtidos através do AHER, com valores de k: 4 (A), 7 (B), 35 (C) e 81 (D).</i>	75
<i>Figura 3.16: Comparação gráfica das unidades do Complexo Lagoa Real (barras azuis) com os grupos obtidos por SOM+ K-means, com k = 11 (barras vermelhas). As barras quantificam o valor-t (A) e a densidade de ocorrências (B) e, os pontos, a porcentagem das ocorrências compreendidas em cada grupo/unidade.</i>	77
<i>Figura 3.17: Comparação gráfica das unidades do Complexo Lagoa Real (barras azuis) com os grupos obtidos pelo AHER (barras vermelhas). As barras quantificam o valor-t (A) e a densidade de ocorrências (B) e, os pontos, a porcentagem das ocorrências compreendidas em cada grupo/unidade.</i>	78
<i>Figura 3.18: A) Mapa com as unidades geológicas do Complexo Lagoa Real em destaque, representando a unidade considerada prospectiva (PP4yl1). B) Resultado do agrupamento obtido com 81 grupos e o limite do Complexo Lagoa Real. C) Mapa com os grupos classificados em escala de cor de acordo com sua dissimilaridade (medida em distância euclidiana) com o grupo com o maior número de ocorrências. D) Escala de cor representando o valor-t de cada grupo. O valor-t mede a associação espacial entre as ocorrências e os grupos (conf. item 2.4.3).</i>	79
<i>Figura 3.19: Gráficos log-log da distância vs. densidade de ocorrências, ajustados para o cálculo do índice de singularidade α. As regiões em cinza representam o intervalo de confiança do ajuste da reta, para um nível de significância de 0,95.</i>	81
<i>Figura 3.20: Mapa destacando em cores apenas as famílias identificadas como correlacionadas com as ocorrências ($\alpha < 2,0$).</i>	82
<i>Figura 3.21: Estruturas regionais com expressão em diferentes profundidades, interpretadas a partir da primeira derivada vertical (1Dz) da continuação ascendente do CMA para as altitudes de: A) 15 km; B) 10 km; C) 5 km; e D) da 1Dz do sinal magnetométrico original.</i>	83

<i>Figura 3.22: Reclassificação binária do mapa de dissimilaridade dos grupos mais prospectivos, separando como classe 1 os grupos localizados no 4º quartil, e, como classe 0, os grupos no intervalo entre o 1º e o 3º quartil.....</i>	<i>85</i>
<i>Figura 3.23: Mapas evidenciais que compõem o modelo de prospectividade mineral. A) Distância dos lineamentos NP_DR. B) Distância dos lineamentos NP_Dn+1. C) Distância dos lineamentos APP_Dn_(NP_Dn). D) Distância dos lineamentos profundos. E) Densidade de cruzamentos de lineamentos de famílias com interesse prospectivo. F) Densidade de cruzamentos de lineamentos profundos com lineamentos de famílias com interesse prospectivo. G) Urânio. H) Tório. I) Potássio. J) Contagem total. K) Gradiente total do CMA. L) Clusters com maior semelhança às zonas mineralizadas.....</i>	<i>86</i>
<i>Figura 3.24: Fluxograma da geração do modelo de potencial para urânio, incluindo as etapas de geração de mapas evidenciais, do conjunto de dados de treinamento, otimização e avaliação do modelo final.....</i>	<i>88</i>
<i>Figura 3.25: PCF (Pair Correlation Function) dos pontos com ocorrências de urânio conhecidas. Observa-se que, a partir de 3600 m, o g(r) se estabiliza próximo de 1.</i>	<i>89</i>
<i>Figura 3.26: Pontos de treinamento utilizados no modelo. A) 600 pontos de não ocorrências foram gerados aleatoriamente, respeitando-se a distância mínima de 3600m das ocorrências conhecidas (total de 33). B) Detalhe ilustrando a distância mínima dos pontos de ocorrência (3600 m) e a distância mínima (1800 m) entre os pontos de não ocorrência gerados.....</i>	<i>90</i>
<i>Figura 3.27: Curva ROC e score AUC do modelo preditivo para os conjuntos de dados de treino e de teste. A linha diagonal tracejada representa a classificação de um preditor aleatório.</i>	<i>94</i>
<i>Figura 3.28: Matriz de confusão da predição do conjunto de dados de teste para um limiar de classificação de 0,5.....</i>	<i>94</i>
<i>Figura 3.29: Modelo de prospectividade para urânio gerado com a combinação XGB+SMOTE. Os pontos numerados indicam as amostras explicadas com SHAP. O score corresponde à probabilidade de predição de uma ocorrência pelo modelo (score = 1).</i>	<i>95</i>
<i>Figura 3.30: Gráfico de avaliação de performance do modelo de prospectividade em relação às ocorrências conhecidas. A curva laranja se refere à porcentagem das ocorrências, e a azul à porcentagem da área de trabalho.</i>	<i>96</i>

Figura 3.31: Média global dos valores SHAP de cada variável do modelo preditivo. **eU**: urânio; **scac81_rec**: clusters com maior semelhança às zonas mineralizadas; **eTh**: tório; **APP_Dn_(NP_Dn)_dist**: distância dos lineamentos APP_Dn_(NP_Dn); **NP_Dn+1_dist**: distância dos lineamentos NP_Dn+1; **CT**: contagem total; **ker_prof_all**: densidade de cruzamentos de lineamentos profundos com lineamentos de famílias com interesse prospectivo; **Kperc**: potássio; **NP_DR_dist**: distância dos lineamentos NP_DR; **prof_ED** distância dos lineamentos profundos; **GT**: gradiente total do CMA; **ker_all**: densidade de cruzamentos de lineamentos de famílias com interesse prospectivo.97

Figura 3.32: Waterfall plots com as explicações SHAP dos pontos do modelo indicados na Figura 3.29. Os valores das variáveis na amostra analisada são mostrados no eixo y. As barras vermelhas indicam a intensidade da contribuição (valor SHAP) de cada variável para o aumento do score, enquanto as barras azuis mensuram a contribuição para a diminuição do score no ponto. Neste gráfico os scores são medidos em log-odds. **eU**: urânio; **scac81_rec**: clusters com maior semelhança às zonas mineralizadas; **eTh**: tório; **APP_Dn_(NP_Dn)_dist**: distância dos lineamentos APP_Dn_(NP_Dn); **NP_Dn+1_dist**: distância dos lineamentos NP_Dn+1; **CT**: contagem total; **ker_prof_all**: densidade de cruzamentos de lineamentos profundos com lineamentos de famílias com interesse prospectivo; **Kperc**: potássio; **NP_DR_dist**: distância dos lineamentos NP_DR; **prof_ED** distância dos lineamentos profundos; **GT**: gradiente total do CMA; **ker_all**: densidade de cruzamentos de lineamentos de famílias com interesse prospectivo.....99

Figura 3.33: Gráfico de dependência do tório em relação ao valor SHAP. Cada ponto corresponde a uma amostra do conjunto de dados de treinamento..... 100

Figura 3.34: Exemplo de 5, dos 100, conjuntos de dados de treino, gerados de forma independente, seguindo os critérios apresentados em 3.6.1. 101

Figura 3.35: Mapas resultantes da análise estatística da variabilidade dos resultados do modelo de prospectividade para U para 100 conjuntos de dados de treinamento diferentes. O primeiro mapa (da esquerda) é um exemplo dos 100 mapas gerados. Os demais são os mapas das estatísticas calculadas para os 100 modelos em conjunto..... 102

Figura 3.36: Histograma e função de densidade de probabilidade dos resultados de prospectividade dos 100 modelos em 1 pixel do mapa. São indicados no histograma os seguintes parâmetros estatísticos: média, desvio padrão, e mediana. 103

Figura 3.37: Diagrama de dispersão, combinado com histogramas, das médias vs. coeficientes de variação dos pixels que apresentaram máximos maiores do que 0 no conjunto dos 100 modelos de prospectividade. A linha vermelha indica o 97º percentil das médias. A linha azul o valor de CV médio acima do 97º percentil das médias..... 104

Figura 3.38: Histograma dos scores F_1 , calculados para os conjuntos de teste de cada um dos 100 modelos gerados com os diferentes conjuntos de treinamento. . 104

LISTA DE TABELAS

<i>Tabela 1: Modelo conceitual do sistema mineral de urânio metassomático da PULR, traduzido em vetores prospectivos.....</i>	<i>31</i>
<i>Tabela 2: Idades e fases de deformação associadas a cada família de lineamentos interpretados (Barros et al. 2021). APP_Dn representa a fase de deformação principal que ocorreu entre o Arqueano e o Paleoproterozoico na área; NP_Dn representa a fase principal de deformação neoproterozoica, relacionada ao ciclo brasileiro na área.....</i>	<i>53</i>
<i>Tabela 3: Resultados dos testes Kolmogorov-Smirnov (KS) realizados entre as células do mapa SOM e os dados originais.</i>	<i>62</i>
<i>Tabela 4: Parâmetros dos grupos definidos por SOM + K-means e das unidades geológicas que apresentaram pelo menos uma ocorrência de urânio. W+, W-, contraste e valor-t calculados pelo método “peso de evidências”. Quanto melhor o valor de cada coluna, em relação à prospectividade, mais intensa a cor do preenchimento em vermelho.</i>	<i>76</i>
<i>Tabela 5: Parâmetros dos grupos definidos pelo AHER, com k = 81, e das unidades geológicas, que apresentaram pelo menos uma ocorrência de urânio. W+, W-, contraste e valor-t calculados pelo método “peso de evidências”. Quanto melhor o valor de cada coluna, em relação à prospectividade, mais intensa a cor do preenchimento em vermelho.</i>	<i>78</i>
<i>Tabela 6: Valores do índice de singularidade (α) e do coeficiente de determinação R^2 calculados pelo ajuste da função não-linear, para cada família de lineamentos. .</i>	<i>80</i>
<i>Tabela 7: Avaliação de performance de 6 classificadores diferentes nos conjuntos de dados de teste.....</i>	<i>91</i>
<i>Tabela 8: Hiperparâmetros do XGB+SMOTE otimizados por grid search.....</i>	<i>92</i>
<i>Tabela 9: Métricas de desempenho do modelo XGB+SMOTE com hiperparâmetros otimizados, obtidas com os dados de treino e teste.....</i>	<i>93</i>

SUMÁRIO

1. Introdução	18
1.1. Contexto geológico	22
1.2. Mineralizações de urânio de Lagoa Real.....	26
1.2.1. Depósitos de urânio do tipo metassomático	27
1.2.2. Geologia das mineralizações de urânio da PULR	29
1.2.3. Sistema mineral de urânio metassomático na PULR	30
2. Materiais e Métodos	32
2.1. Dados de aerogeofísica e MDT	32
2.1.1. Processamento dos dados magnetométricos.....	34
2.2. Preparação dos dados.....	35
2.3. <i>Self-Organizing Maps</i> para análise exploratória e preditiva.....	37
2.3.1. <i>Self-Organizing Maps</i>	37
2.3.1.1. Métodos de visualização de SOM.....	40
2.3.1.1.1. <i>Matriz-U</i> , números de vetores e distância média	40
2.3.1.1.2. Planos de componentes	41
2.3.1.2. <i>Counter propagation network</i>	41
2.4. Delimitação de domínios geológico-geofísicos	42
2.4.1. Agrupamento por SOM + K-means.....	42
2.4.2. Agrupamento hierárquico espacialmente restrito (AHER)	45
2.4.3. Seleção do número ótimo de grupos	48
2.5. Determinação de famílias de lineamentos prospectivos através de análise de fractais	50
2.6. Geração do modelo de prospectividade	53
2.6.1. Geração do conjunto de dados de treinamento	53
2.6.2. Seleção do classificador.....	55
2.6.3. Avaliação do modelo com SHAP	59
3. Resultados	60
3.1. Análise exploratória e preditiva com SOM	60
3.2. Delimitação de domínios geológico-geofísicos	69
3.2.1. Agrupamento por SOM + K-means.....	69
3.2.2. Agrupamento por AHER.....	72

3.2.3. Comparação dos resultados	75
3.3. Determinação de famílias de lineamentos prospectivos através de análise de fractais.....	80
3.4. Interpretação de estruturas regionais profundas	82
3.5. Mapas evidenciais	84
3.6. Modelo de prospectividade para U	87
3.6.1. Geração do conjunto de dados de treino	89
3.6.2. Seleção do classificador.....	91
3.6.3. Treinamento, otimização e avaliação do modelo de prospectividade	91
4. Discussões.....	105
5. Conclusões	111
6. Referências	113

1. Introdução

Com o esgotamento progressivo dos depósitos minerais mais rasos, de reconhecimento mais fácil, o emprego de tecnologias informáticas tem tido importância cada vez mais relevante na exploração mineral. Nesse sentido, métodos de modelamento de potencial mineral (MPM) em ambiente SIG têm sido aplicados de forma crescente nas últimas décadas como ferramenta auxiliar na definição de áreas prospectáveis e alvos exploratórios (e.g. [Knox-Robinson e Wyborn 1997](#), [Magalhães e Souza Filho 2012](#), [Porwal e Carranza 2015](#), [Prado et al. 2020](#)).

As técnicas de MPM permitem a integração de uma grande quantidade de dados espaciais, e de conhecimentos específicos, de forma eficiente e dirigida. De modo geral, os métodos de modelamento de dados dividem-se em dois tipos: guiados pelos dados (*data driven*) e guiados pelo conhecimento (*knowledge driven*) ([Bonham-Carter 1994](#)). Nas técnicas guiadas pelo conhecimento, os pesos dos elementos geológicos representados nos mapas são definidos a priori, com base em modelos de depósitos conceituais. Já nas técnicas guiadas pelos dados, os pesos são atribuídos computacionalmente por algoritmos matemáticos. Em geral, as técnicas guiadas pelo conhecimento são mais utilizadas em área pouco exploradas, com menor quantidade de dados disponíveis. Por outro lado, técnicas guiadas pelos dados possuem a vantagem de identificar relações complexas entre os dados, além de minimizar vieses introduzidos pelo modelador. No entanto, essas são mais sensíveis à distribuição e qualidade dos dados utilizados ([Bonham-Carter 1994](#)).

Uma fase fundamental da construção de modelos de prospectividade mineral é a definição dos processos geológicos considerados críticos para a formação do tipo de depósito mineral buscado na área em estudo. Para tanto, tem sido utilizada a abordagem de Sistemas Minerais (*Mineral Systems*, [Wyborn et al. 1994](#), [Hagemann et al. 2016](#)). Essa abordagem de análise possui a vantagem de não se restringir aos critérios estritamente definidos dos modelos genéticos clássicos. Além disso, o entendimento do Sistema Mineral permite sua compartimentação em componentes com expressões bem maiores do que o próprio depósito mineral, passíveis de serem mapeáveis através de elementos espaciais que os representem na escala de modelamento. Dessa forma, os bancos de dados geológicos disponíveis para a área são utilizados com maior eficiência para

elaboração de mapas preditivos/evidenciais que representem aqueles elementos críticos.

Nos últimos anos, tem sido crescente a aplicação de *machine learning* (ML) no modelamento de prospectividade mineral (MPM) guiado pelos dados (Okada 2020, Kreuzer *et al.* 2020, Zuo e Carranza 2023). As técnicas de ML possuem a propriedade de criar modelos matemáticos através de uma espécie de aprendizagem automática e adaptativa, que os permite melhorar com a experiência, isso é, com a progressiva adição de novos dados. Em geral, as técnicas de ML são divididas em 3 grupos principais: supervisionadas, não-supervisionadas e de aprendizado reforçado, sendo os dois primeiros os mais utilizados atualmente para MPM (Kreuzer *et al.* 2020). Nas técnicas supervisionadas, o algoritmo recebe dados classificados com a característica que se deseja modelar (por ex. locais mineralizados ou não mineralizados). Assim o modelo é treinado com base em exemplos prévios. Nas técnicas não-supervisionadas, os algoritmos recebem dados não classificados, de forma que o modelo apenas reconhece padrões e estruturas dentro do conjunto de dados de forma independente, e os classifica de acordo com os grupos reconhecidos. Nas técnicas de aprendizado reforçado, não são carregados dados de entrada, e o aprendizado do modelo se dá através da interação de um agente (por ex. um robô, ou uma simulação) com um ambiente, que geram *feedbacks* positivos ou negativos, dependendo das modificações que essas interações produzem (Alpaydin 2014).

Devido à complexidade dos sistemas minerais, onde inúmeras variáveis geológicas se interrelacionam de forma sutil e diversa (Kastens *et al.* 2009), comportando-se como sistemas críticos auto-organizáveis (Hronsky 2009, McCuaig e Hronsky 2014), a aplicação de técnicas de ML mostra-se interessante para o modelamento desses sistemas. As técnicas a serem utilizadas devem ser escolhidas de acordo com a escala de trabalho, a qualidade, e a distribuição dos dados disponíveis para tanto.

O presente trabalho tem por objetivo avaliar a aplicação de técnicas de aprendizado supervisionado e não-supervisionado em bancos de dados geoespaciais, visando obter *insights* relativos aos guias prospectivos, e ao potencial mineral, de uma área de estudo, e estabelecer hipóteses para o direcionamento de

trabalhos de prospecção mineral. Serão utilizados dados públicos gerados em campanhas de aerogeofísica, sensores remotos e mapas geológicos.

A área escolhida para este estudo é a Província Uranífera de Lagoa Real (PULR, Figura 1.1). Trata-se de uma área com limitada disponibilidade de dados de mapeamento disponíveis, porém conta com levantamentos aerogeofísicos com densidade de dados adequada para um estudo em escala de distrito. Ademais, estudos prévios indicam que pelo menos parte das ocorrências de urânio da PULR podem apresentar assinaturas identificáveis em dados aerogeofísicos regionais (Pires 2012).

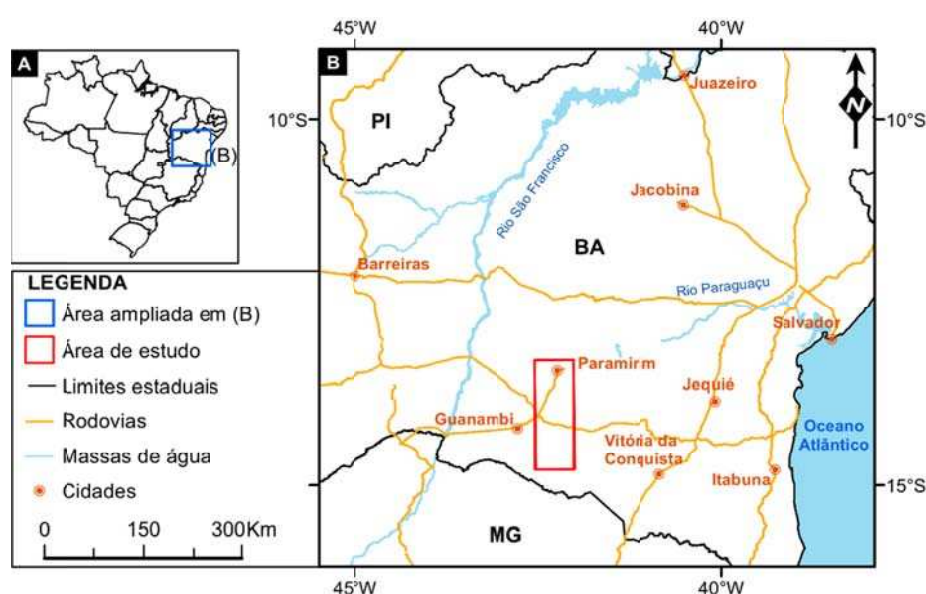


Figura 1.1: Localização da área de estudo.

A primeira técnica de ML aplicada neste trabalho foi o *Self-Organizing Maps* (SOM) (Kohonen 2001), com a finalidade de se realizar uma análise exploratória dos dados geofísicos, em relação às feições geológicas e às mineralizações de urânio da área. Um dos grandes destaques do SOM está em sua capacidade de proporcionar visualizações informativas do espaço de dados, o que o faz uma interessante ferramenta para se explorar grandes conjuntos de dados (Kohonen 2013).

Em seguida, foram utilizadas técnicas de agrupamento (aprendizado não supervisionado) para a delimitação de domínios geológicos a partir de dados gamaespectrometria e magnetometria aéreas, e modelo digital de terreno. As técnicas de agrupamento são adequadas para essa finalidade pois eliminam a

subjetividade inerente à delimitação manual de domínios, que depende do conhecimento e da experiência do interprete (IAEA 2003). Além disso, essas técnicas possuem a capacidade de detectar variações sutis nos padrões de dados, que podem representar anomalias localizadas, e que seriam de difícil identificação de forma manual. No entanto, a utilização de técnicas de agrupamento em dados espacializados pode gerar resultados indesejáveis, uma vez que não leva em conta as suas relações espaciais, tendendo a gerar domínios dispersos, muitas vezes com pouca, ou nenhuma, continuidade espacial, o que geralmente não é desejado quando se trabalha com bancos de dados espacializados. Por esse motivo, nesses casos, em aplicações onde são desejados agrupamentos homogêneos e espacialmente contínuos, podem ser mais adequadas as técnicas de agrupamento espacialmente restritas, denominadas de técnicas de regionalização (Assunção *et al.* 2006). Em experimentos realizados por Lattimer e Lattimer (2022), os autores observaram que a técnica agrupamento aglomerativo espacialmente restrito, quando comparada com outras técnicas de agrupamento regionalizado (*i.e.* AZP - Automatic Zoning Procedure - Openshaw 1977; REDCAP - Guo 2008; Max-P-Regions - Wei *et al.* 2021; SKATER - Spatial' K'luster Analysis by Tree Edge Removal - Assunção *et al.* 2006), mostra um ótimo *trade-off* entre desempenho computacional e qualidade do agrupamento realizado, obtendo resultados de qualidade comparável, com menor demanda computacional, o que é relevante para bancos de dados volumosos (com mais do que 10.000 pontos). Neste trabalho foi aplicado o algoritmo de agrupamento aglomerativo espacialmente restrito, e os resultados obtidos foram avaliados em relação à sua aplicabilidade para análise de prospectividade na área de estudo. Para comparação, foi realizado também o agrupamento dos dados através da combinação de SOM e K-means, a exemplo de Chudasama *et al.* (2022), Leite e Souza Filho (2010), Torppa *et al.* (2019).

Para a elaboração do mapa de potencial para urânio da PULR, elaborou-se um modelo conceitual do sistema mineral gerador das ocorrências com base na identificação/definição dos componentes geológicos fundamentais na geração das mineralizações de urânio. A partir desse modelo conceitual, foram elaborados mapas evidenciais adequados para a vetorização de tais cada componente do sistema mineral. Ao final, foram utilizadas técnicas de aprendizado supervisionado

para a integração dos mapas evidenciais e geração do modelo de potencial para ocorrências de urânio.

Levando-se em consideração a disponibilidade atual de bancos de dados geoespaciais ao público no Brasil, em particular de dados de aerogeofísica regionais, os resultados deste trabalho poderão servir como exemplo para o uso de novas tecnologias de ML como uma estratégia de baixo custo para a avaliação preliminar de potencial mineral, e geração de alvos exploratórios.

1.1. Contexto geológico

A PULR localiza-se na parte centro-leste do Craton São Francisco ([Almeida 1977](#)), porção norte do Orógeno Araçuaí, mais especificamente na região definida como Aulacógeno do Paramirim ([Pedrosa-Soares et al. 2001](#), [Cruz et al. 2007](#)) (Figura 1.2). Essa área engloba as regiões morfo-estruturais dos vales dos rios Paramirim e São Francisco, em sua porção central, a serra do Espinhaço Setentrional, em sua parte oeste, e a Chapada Diamantina, na parte leste ([Cruz et al. 2008](#)).



Figura 1.2: Contexto geológico regional da área de estudo (delimitada aproximadamente pelo retângulo vermelho) (adaptado de Cruz e Alkmim, 2006).

O embasamento da PULR é composto por rochas arqueanas incluídas no Complexo Paramirim, composto por ortognaisses e migmatitos, e por rochas que fazem parte dos *greenstone belts* Ibitira-Ubiraçaba e Umburana (Cruz et al. 2008). Intrudindo o embasamento, ocorre o Complexo Lagoa Real (~1,7 Ga - Turpin et al. 1988, Cordani et al. 1992).

O Complexo Lagoa Real faz contato com o embasamento por zonas de cisalhamento, sendo seu caráter intrusivo evidenciado pela presença de xenólitos e enclaves do embasamento (Cruz et al. 2008).

Para o estudo realizado neste trabalho, em escala de distrito, adotou-se como base geológica o mapa do estado da Bahia, na escala 1:1.000.000 (Souza et al. 2003) (Figura 1.3). É o único mapa disponível atualmente, com abrangência completa da área de estudo. Para simplificação e tomando-se como referência o Complexo Lagoa Real, as unidades geológicas do mapa podem ser agrupadas em 5 domínios tectono-cronológicos: Embasamento; Complexo Lagoa Real; Coberturas proterozóicas; Coberturas recentes. O embasamento inclui unidades do bloco

Gavião, sobre o qual se desenvolveu o Aulacógeno do Paramirim. Inclui TTG's meso a neoarqueanos, granitoides ricos em K e *greenstonebelts*, além de granitoides e unidades metassedimentares riacianas-orosirianas ([Cruz e Alkmim 2017](#)). O Complexo Lagoa Real inclui as rochas da Suite Intrusiva Lagoa Real, que estão associadas ao magmatismo anorogênico relacionado ao desenvolvimento/abertura do Aulacógeno do Paramirim ([Cruz e Alkmim 2017](#)). Nas Cobertura Proterozoicas, estão incluídas as sequências paleo a mesoproterozoicas (meta)sedimentares e (meta)vulcânicas do Supergrupo Espinhaço e do Grupo Santo Onofre, e as sequências sedimentares neoproterozoicas do Supergrupo São Francisco ([Cruz e Alkmim 2017](#)). As Coberturas recentes incluem os depósitos aluvionares e as coberturas detrito-lateríticas cenozoicas ([Souza et al. 2003](#)).

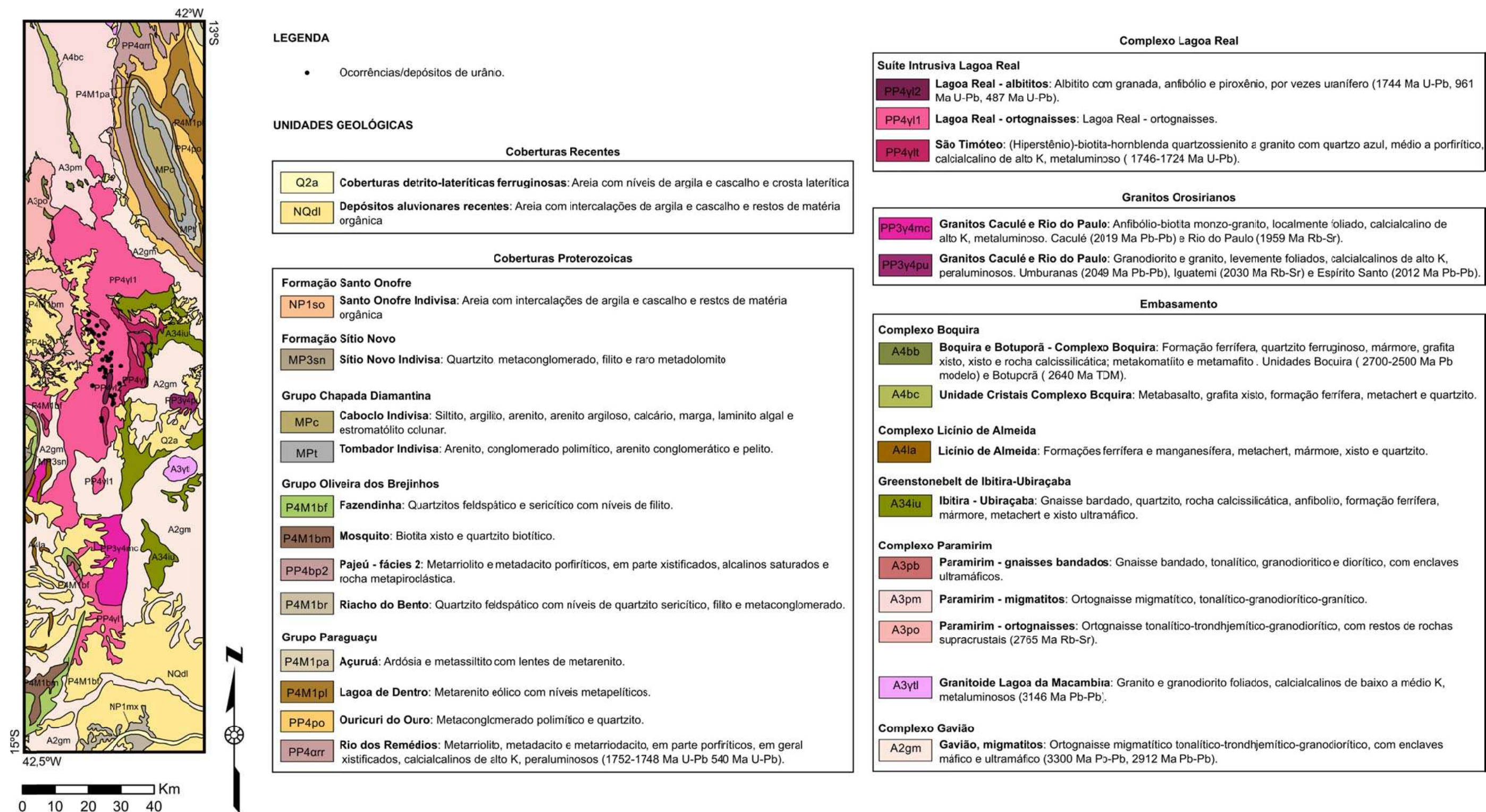


Figura 1.3 Mapa geológico da área de estudo. Adaptado de Souza *et al.* (2003).

1.2. Mineralizações de urânio de Lagoa Real

A prospecção sistemática de minerais radioativos se iniciou no Brasil em 1952, promovida pelo então recém criado Conselho Nacional de Pesquisa, atual CNPq. Em 1955, um acordo de cooperação técnica foi assinado entre Brasil e Estados Unidos para avaliar o potencial de urânio no país. Posteriormente, em 1962, com a criação da Comissão Nacional de Energia Nuclear (CNEN), foi organizado um departamento de exploração mineral, contando com o apoio da Comissão de Energia Atômica da França (CEA). Na década de 1970, a exploração de minerais radioativos pela CNEN foi intensificada com o aumento dos recursos financeiros. Em 1974, foi criada a Nuclebrás, com o propósito de executar a exploração e a produção de urânio no país. Um dos primeiros sucessos foi a descoberta e o desenvolvimento do depósito Osamu Utsumi no planalto de Poços de Caldas. Em 1975, Brasil e Alemanha assinaram um acordo de cooperação para o uso pacífico da energia nuclear, iniciando um ambicioso programa de desenvolvimento nuclear, que resultou na descoberta de áreas com recursos de urânio, incluindo Lagoa Real, Figueira, Quadrilátero Ferrífero, Amarinópolis, Rio Preto/Campos Belos, Itataia e Espinhaços (descoberta e pesquisada pela Nuclam, empresa proveniente de uma parceria entre Brasil e Alemanha) ([NEA/IAEA 2023](#)).

Em 1988, o programa nuclear brasileiro passou por uma reorganização e a exploração de urânio no país ficou a cargo da Indústrias Nucleares do Brasil (INB), que sucedeu a Nuclebrás. A INB descontinuou as atividades de exploração regional de urânio em 1991, focando seus trabalhos, desde então, em melhor definir os recursos na província de Lagoa Real ([NEA/IAEA 2023](#)).

As primeiras mineralizações de urânio na região de Lagoa Real foram identificadas a partir de uma série de levantamentos geofísicos aéreos, realizados na década de 1970. Esses levantamentos iniciais foram focados na prospecção de depósitos de urânio em metaconglomerados, do tipo *Witwatersrand* (ou QPC - *quartz pebble conglomerate*), na Serra do Espinhaço. Após a detecção das primeiras anomalias, novos levantamentos aerogeofísicos foram conduzidos, em 1979, na região de Lagoa Real, no âmbito do projeto São Timóteo, resultando na identificação de 115 anomalias. Em seguida, levantamentos geológicos foram realizados pela Nuclebrás, em parceria com a CBPM, levando à descoberta de diversas ocorrências de urânio ([Pires 2012](#)).

Atualmente, a PULR possui a única mina de urânio em atividade na América Latina, denominada mina do Engenho. Ademais, a mina Cachoeira, que teve seu recurso em *open pit* exaurido, se encontra em processo de licenciamento e redefinição de recursos, para dar início à operação de lavra subterrânea. No total, são conhecidas 38 ocorrências de U na PULR, com recursos minerais estimados em 87 mil toneladas de urânio contido (U_3O_8), distribuídos em 17 depósitos. O complexo minero-industrial da INB, em Caetité, possui potencial de produzir 400 t/ano, podendo aumentar para 800 t/ano, com a atividade da mina do Engenho, a lavra subterrânea da mina Cachoeira, e a expansão da unidade de produção (INB 2024, NEA/IAEA 2023).

1.2.1. Depósitos de urânio do tipo metassomático

Os depósitos e ocorrências de urânio da PULR são classificados como do tipo metassomático, que também são conhecidos como do “tipo albitito” ou “Na-metassomático” (Figura 1.4) (Cuney *et al.* 2012, Wilde 2013, IAEA 2020). A PULR é reconhecida como uma das principais províncias mundiais desse tipo de depósito (Cuney *et al.* 2012). Exemplos de importantes depósitos e distritos similares incluem o Kirovograd-Krivoi Rog na Ucrânia, Central Mineral Belt e Beaverlodge no Canadá, Mount Isa na Austrália e Itataia no Brasil (Wilde 2013).

Devido à escassez de estudos neste tema, os depósitos de urânio metassomáticos são menos conhecidos. Wilde (2013) sumariza algumas das principais características desse grupo de depósitos. Encontram-se quase que exclusivamente em terrenos metamórficos proterozóicos (com exceção do depósito de Coles Hill, nos EUA - Bruneton e Cuney, 2016), principalmente onde predominam rochas orosirianas (1,8 Ga a 2,05 Ga). A mineralização não é associada a uma rocha hospedeira específica, ocorrendo, por exemplo, em granitos/gnaisses, metassedimentos ricos em ferro, metabasaltos e metarriolitos.

Bruneton e Cuney (2016) e Dahlkamp (2009) dividem esse tipo de depósito em 2 subtipos: os “derivados de granitos” e os “derivados de metassedimentos e metavulcânicas”. No primeiro subtipo, ao qual pertence os depósitos conhecidos da PULR, eles ocorrem disseminados em vênulas preenchendo fraturas proximamente espaçadas, falhas e brechas em granitos albitizados e com perda metassomática de quartzo, ao longo de zonas de falhas

formadas antes do metassomatismo, cortando granitos uraníferos. Nesses depósitos as mineralizações podem continuar a mais de 1000 m de profundidade, mas em geral possuem algumas centenas de metros de profundidade. São classificados no segundo subtipo os depósitos da Bacia de Krivoy Rog, na Ucrânia, e os de Central Mineral Belt of Labrador, no Canadá. Na Bacia de Krivoy Rog, os depósitos são hospedados em rochas carbonáticas, quartzito, xisto, formação ferrífera bandada e em microclínio-granitos, com forte alteração sódica, carbonática e silicificação, ao longo de zonas estruturais. Os depósitos de Central Mineral Belt (e.g. Michelin e Jacques Lake) são em grande parte hospedados em rochas metavulcânicas félsicas proterozóicas.

Uma das características principais desses depósitos é a associação das mineralizações econômicas com albititos, o que indica uma relação entre os processos de enriquecimento de urânio e de albitização. Nos distritos com informações existentes sobre a distribuição regional dos albititos (e.g. Lagoa Real e Mount Isa), foi observado que a ocorrência deles é muito mais extensiva do que a da mineralização uranífera, de forma que, aparentemente, a albitização é um processo necessário, porém, não suficiente para a geração das mineralizações. Em Mount Isa, os albititos podem estar associados a mineralizações de outros metais além do urânio, como de cobre e ouro do tipo IOCG ([Wilde 2013](#)).

Em geral, o urânio se encontra disseminado nos albititos, e raramente, forma veios macroscópicos. São exceções a essa característica os depósitos de Jacques Lake e Ace-Fay-Verna, respectivamente localizados nos distritos de Central Mineral Belt e Beaverlodge, no Canadá, onde a maior parte das mineralizações se encontra em veios ([Wilde 2013](#)).

Em grande parte dos depósitos, pode ser observado que o urânio foi introduzido após a albitização da rocha hospedeira, temporalmente próximo ao micro fraturamento ou brechação completa do albitito. Isso, possivelmente junto com a dissolução química, pode ser um fator crítico para a geração de permeabilidade e porosidade nos albititos, que são significativamente maiores do que nas rochas hospedeiras ([Wilde 2013](#)).

Praticamente todos os depósitos de urânio metassomáticos descritos são associados com zonas de milonitização e/ou cataclase. A presença de importantes zonas de milonitização parece ser uma característica fundamental dessa classe. Nos

depósitos que dispõem de dados sobre a morfologia dos corpos de minério, estes costumam ter formas lineares e achatadas, com altos caimentos, dentro de zonas de milonitização (Wilde 2013). Ademais, Dahlkamp (2009) aponta que esses depósitos são associados a rochas metassomatizadas ao longo de sistemas de falhas com raízes profundas.

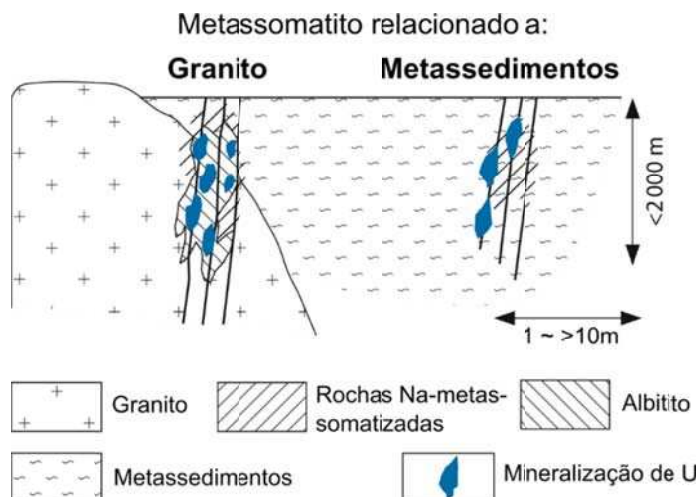


Figura 1.4: Modelo geológico das mineralizações de U do tipo metassomáticas (adaptado de Dahlkamp, 2009).

1.2.2. Geologia das mineralizações de urânio da PULR

As mineralizações uraníferas da PULR ocorrem associadas aos granitos e gnaisses do Complexo Lagoa Real (Granito São Timóteo), disseminadas em rochas ricas em albita ($\pm$ oligoclásio) e albititos, hospedadas em zonas de cisalhamento. A uraninita fina é o principal mineral de minério primário. Embora de forma subordinada, além dos albititos, as mineralizações podem ocorrer também em associação com epidositos (Lobato e Fyfe 1990, Cruz *et al.* 2008). Nos albititos mineralizados, o urânio ocorre como uraninita em cristais microscópicos (5-25 μ m), associados principalmente com andradita (5~20% composição da rocha), mas também podem ocorrer inclusos ou nas bordas de cristais de piroxênio, albita, hornblenda, biotita, calcita e hematita (Lobato e Fyfe 1990). Os albititos mineralogicamente puros (~100% albita) não são mineralizados. Os epidositos são compostos por epidoto (50~80%), além de oligoclásio, aegirina-augita, hornblenda e minerais acessórios. Neles, o epidoto pode conter até 1,2% de UO_2 , e a uraninita costuma não ocorrer (Lobato e Fyfe 1990).

De acordo com [Cruz et al. \(2008\)](#), a sequência de eventos geológicos críticos que levaram à formação dos depósitos de urânio inicia-se com o magmatismo anorogênico do complexo Lagoa Real, associado ao evento de tafrogênese, que resultou na abertura do aulacógeno do Paramirim. Durante esta evolução, os esforços distensivos com direção E-W geraram zonas de fraqueza com direção N-S, as quais o Granito São Timóteo aproveitou durante sua colocação. A cristalização desse granito em cerca de 1,7 Ga resultou em processos de metassomatismo sódico, que promoveram a formação de zonas enriquecidas em urânio associadas a albititos, e regiões marcadas por metassomatismo sódico-cálcico. Essas zonas de metassomatismo foram originadas, possivelmente, a partir da remobilização e concentração de urânio por fluidos tardi-magmáticos ([Maruéjol et al. 1987](#)). Durante o evento Brasiliano, que reativou a estruturação N-S do aulacógeno, se deu a deformação e o metamorfismo desses granitoides, e consequente formação dos gnaisses em fácies anfibolito. Nessa fase, teria ocorrido a individualização dos corpos de albititos através da recristalização dinâmica dos feldspatos, e a concentração do urânio em zonas de cisalhamento dúcteis ([Cruz et al. 2008](#)).

1.2.3. Sistema mineral de urânio metassomático na PULR

Para a construção do mapa de potencial para urânio metassomático da PULR, foi utilizada a metodologia de [McCuaig et al. \(2010\)](#) e [McCuaig e Hronsky \(2014\)](#) de estudo prospectivo baseado no conceito de Sistemas Minerais. Dessa forma, elaborou-se um modelo conceitual do sistema mineral de urânio metassomático da PULR, com base em dados da literatura sobre as mineralizações da área, sua evolução geológica, e sobre o modelo genético de depósitos de urânio do tipo metassomático. A Tabela 1 sumariza este modelo conceitual e como cada um de seus elementos se traduz, ao final, em mapas evidenciais, os quais serão integrados na elaboração de um modelo de prospectividade.

Tabela 1: Modelo conceitual do sistema mineral de urânio metassomático da PULR, traduzido em vetores prospectivos. NP_DR, NP_Dn+1, APP_Dn_(NP_Dn) definidos na Tabela 2.

Componente do sistema mineral	Critérios teóricos	Critérios mapeáveis	Conjuntos de dados	Mapas Evidenciais
Arquitetura	- Falhas crustais profundas/translitosféricas, associadas à abertura do rift que originou o aulacógeno do Paramirim no Estateriano, facilitaram e guiaram a ascensão de magmas do tipo A (Suite Lagoa Real) e fluidos ricos em U; - Estruturas pré a sin-magmáticas focalizaram processos hidrotermais tardi-magmáticos, que localmente geraram enriquecimento primário de U; - Estruturas geradas ou reativas no Neoproterozoico, durante o evento Brasileiro, conduziram fluidos que remobilizaram e concentraram U em níveis econômicos; - Zonas de deformação nos albititos (de microfraturamento a brechação) aumentam a permeabilidade e favorecem a mobilização de fluidos e deposição de minerais de urânio.	- Proximidade, alta densidade e cruzamento de falhas ativas ou geradas durante os eventos de magmatismo tipo A estateriano e de deformação/metamorfismo neoproterozoico; - Falhas profundas.	- Dados vetoriais de estruturas lineares classificadas por evento deformacional; - Dados vetoriais de estruturas lineares, interpretadas da magnetometria, processada para realçar sinais profundos.	Distância dos lineamentos NP_DR.
				Distância dos lineamentos NP_Dn+1.
				Distância dos lineamentos APP_Dn_(NP_Dn).
				Densidade de cruzamentos de lineamentos de famílias associadas às mineralizações.
				Distância dos lineamentos profundos.
				Densidade de cruzamentos de lineamentos profundos com lineamentos de famílias associadas às mineralizações.
Geodinâmica	- Magmatismo tipo A, derivado da abertura do rift/aulacógeno Paramirim no Estateriano, resultando na deposição primária magmática de minerais com U; e posteriormente na mobilização e concentração de U em processos hidrotermais tardi-magmáticos; - Deformação, metamorfismo e hidrotermalismo (focados em ZCs) no Neoproterozoico, durante o evento Brasileiro. Resultou na individualização dos albititos e na mobilização e concentração do U em teores econômicos.	Assinaturas geofísicas-geomorfológicas de regiões que hospedam mineralizações, representando litotipos associados ao magmatismo tipo A estateriano e zonas de alteração metassomática.	Dados gamaespectrométricos, magnetométricos e MDT.	Zonas prospectivas definidas por aprendizado não-supervisionado.
				GT
				eU
				eTh
				K
Fertilidade	- Embasamento litosférico; - Distensão e adelgaçamento litosférico no Estateriano, na litosfera previamente modificada por orogênese e metamorfismo entre o Sideriano e o Orosiriano.	Não se aplicam na escala de trabalho, pois toda a área está no contexto considerado favorável.	--	CT
				--

2. Materiais e Métodos

Neste item são apresentados os dados utilizados como base para o trabalho, sua origem e as etapas de processamento realizadas. Também é apresentada a fundamentação teórica das principais técnicas utilizadas neste trabalho.

2.1. Dados de aerogeofísica e MDT

Os dados de aerogeofísica utilizados nesta pesquisa são provenientes dos levantamentos realizados sob demanda da CBPM (Companhia Baiana de Pesquisa Mineral), nas áreas Campo Alegre de Lourdes-Mortugaba e Ibitiara-Rio de Contas. Os levantamentos são compostos por linhas de voo com espaçamento médio de 500m. Os dados dos levantamentos foram pré-processados (Figura 2.1, [CBPM/CPRM 2006](#)), integrados e disponibilizados pelo SGB/CPRM (Serviço Geológico do Brasil). Foram utilizados, neste estudo, dados de gamaespectrometria (eU, eTh, K e CT) e magnetometria (CMA – campo magnético anômalo, e GT - gradiente total). Os dados de modelo digital de elevação (MDT) são provenientes do sensor ALOS PALSAR, com resolução espacial de 12,5 metros, obtidos a partir do repositório online: <https://search.asf.alaska.edu/#/>. Os mapas de aerogeofísica e de MDT da área de estudo são apresentadas na Figura 2.2.

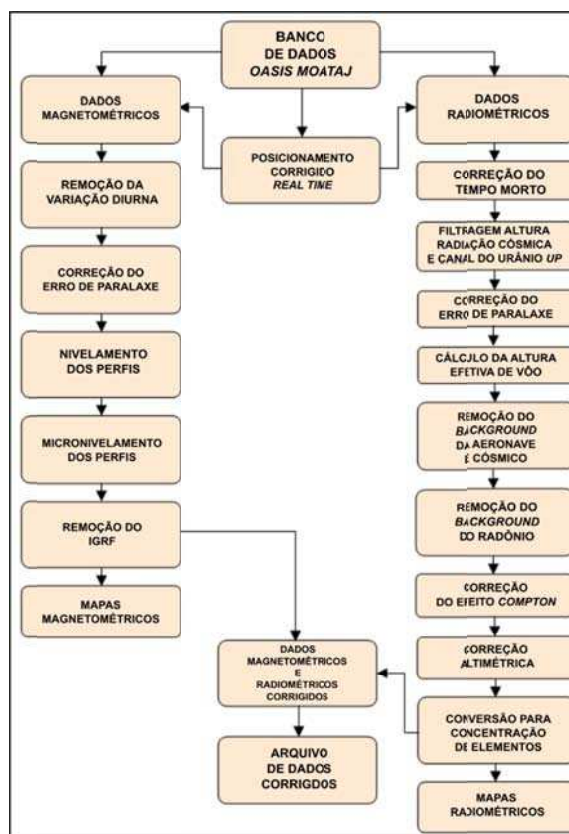


Figura 2.1: Fluxograma do processamento dos dados aerogeofísicos (CBPM/CPRM, 2006).

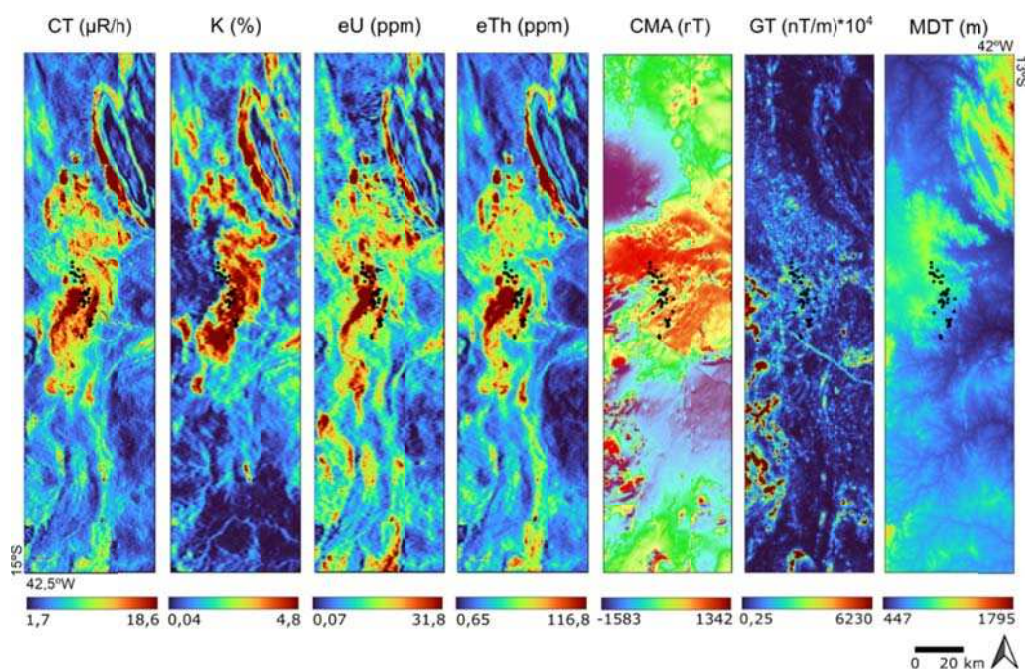


Figura 2.2: Dados aerogeofísicos e modelo digital de terreno. Dados gamaespectrométricos: contagem total, potássio, urânio equivalente e tório equivalente. Produtos derivados de dados magnetométricos: campo magnético anômalo e gradiente total.

2.1.1. Processamento dos dados magnetométricos

A interpretação de estruturas regionais profundas, visando a composição de um dos mapas evidenciais do modelo prospectivo, foi realizada com base em produtos derivados do processamento dos dados aeromagnetométricos para filtragem do sinal com origem em fontes mais rasas e para o realce das feições magnéticas lineares. Para tanto, foi utilizado o método de filtragem “continuada ascendente”, seguido da “derivada vertical”. A continuação ascendente ([Henderson e Zietz, 1949](#)) processa os dados de maneira a simular valores que seriam detectados a uma certa altitude acima do ponto de detecção original, permitindo a filtragem de sinais superficiais e possibilitando a análise dos dados em diferentes profundidades. O método de derivada vertical ([Evjen 1936](#)) foi aplicado aos dados magnéticos para tornar mais visível as feições de alta frequência, destacando, por exemplo, lineamentos e pequenas intrusões, que por vezes podem estar mascarados por feições com sinais de maior amplitude e baixa frequência. Ambas filtragens foram realizadas para altitudes de 5 km, 10 km e 15 km, obtendo-se imagens da derivada vertical de fontes de sinais progressivamente mais profundos. A derivada vertical também foi aplicada ao CMA (Figura 2.3).

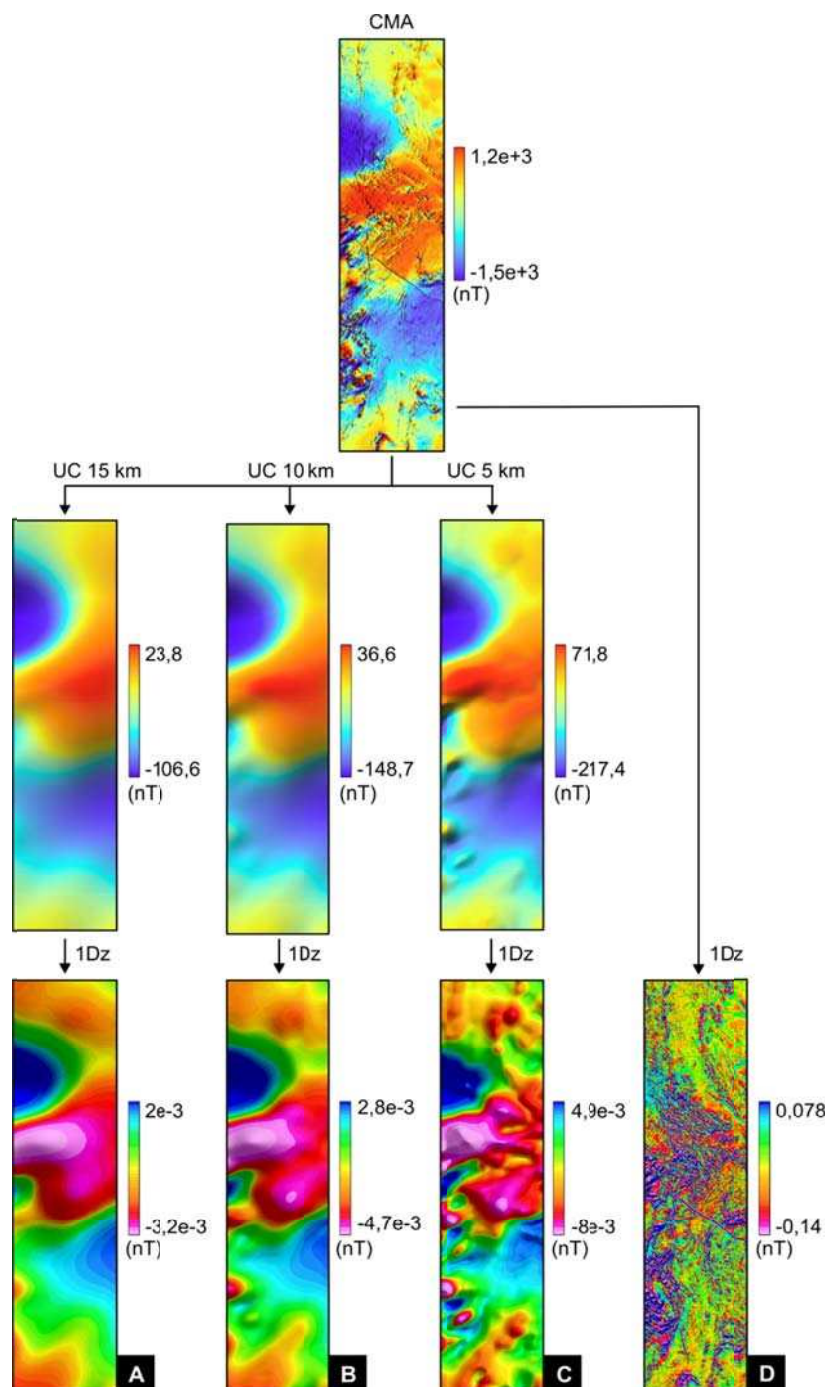


Figura 2.3: Fluxograma das filtragens realizadas nos dados aeromagnéticos para a obtenção das primeiras derivadas verticais (1Dz) das continuações ascendentes de: A) 15 km; B) 10 km; C) 5 km; e D) da 1Dz do dado magnetométrico original. UC: *Upward continuation* (continuação ascendente); CMA: Campo magnético anômalo.

2.2. Preparação dos dados

Para a aplicação dos dados aerogeofísicos e MDT para as análises de SOM (item 2.3) e de delimitação de domínios geológico-geofísicos (item 2.4), os

dados foram pré-processados de forma a melhor adequá-los às análises empregadas, atenuando-se a assimetria de suas distribuições e normalizando-os a uma escala comum (Figura 2.4 e Figura 2.5). Esse procedimento é necessário para análises baseadas em medidas de dissimilaridade por distância (Kaufman e Rousseeuw 2005). Para aproximar os dados aerogeofísicos à distribuição normal, aplicou-se a transformação *box-cox* (Box e Cox 1964). Posteriormente, todos os dados do conjunto foram escalonados ao número de desvios padrões das médias (Z-score).

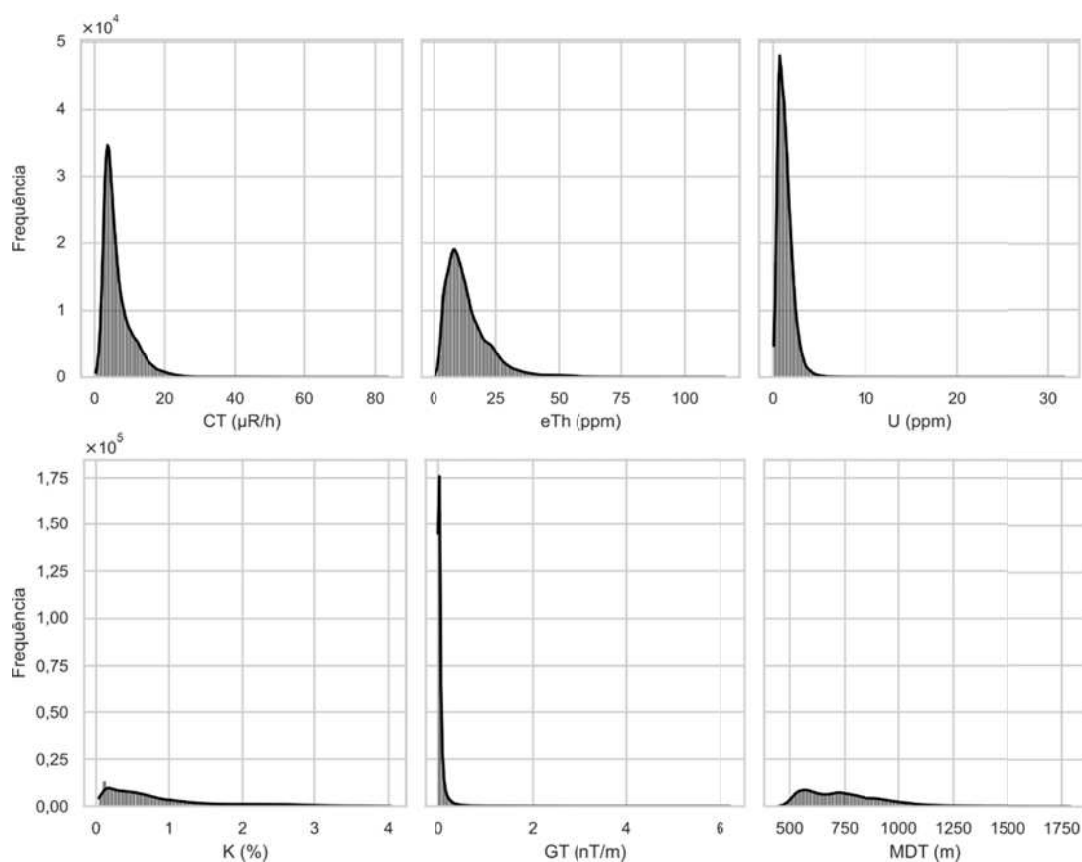


Figura 2.4: Histogramas de distribuição dos dados aerogeofísicos (CT, eTh, eU, K e GT) e de MDT não transformados.

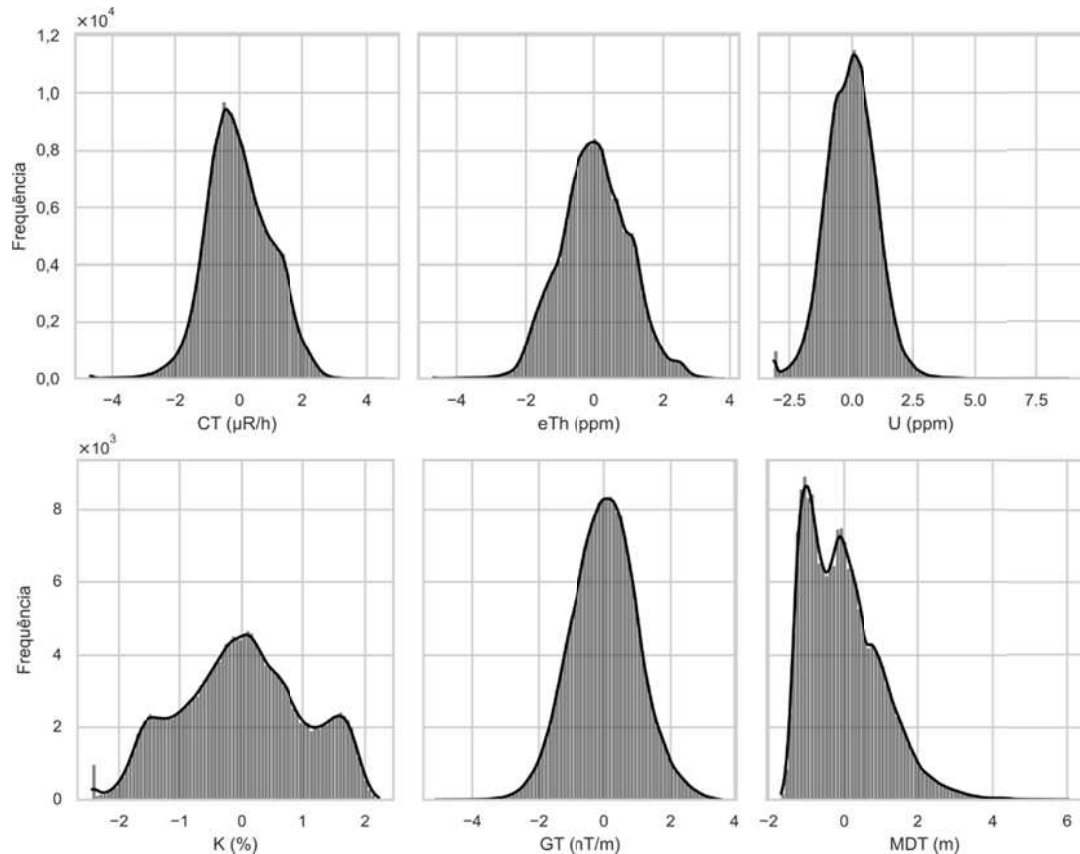


Figura 2.5: Histogramas de distribuição dos dados aerogeofísicos (CT, eTh, eU, K e GT) normalizados (*box-cox*) e escalonados (*z-score*).

2.3. *Self-Organizing Maps* para análise exploratória e preditiva

2.3.1. *Self-Organizing Maps*

O algoritmo de agrupamento *Self-Organizing Maps* (SOM), ou *Kohonen Maps*, é um algoritmo de aprendizado não-supervisionado, que utiliza como base uma arquitetura de redes neurais (Kohonen 2001, Kohonen 1982). No SOM, os dados de entrada *n*-dimensionais são mapeados de maneira não-linear, através de aprendizado competitivo, em células (que podem ser chamadas de neurons ou *codebook-vectors*) em um mapa mono, ou bi-dimensional, chamado de mapa topológico restrito (*constrained topological map*) (Figura 2.6) (Hastie *et al.* 2009, Kaski 2017). As células do mapa SOM são dispostas de forma a preservar a topologia espacial dos dados *n*-dimensionais. Sendo assim, células semelhantes são mapeadas próximas entre si. O algoritmo SOM tem aplicação para agrupamento (Baçã *et al.* 2005), mineração de dados, redução de dimensionalidade, quantização vetorial, classificação e amostragem (Kohonen 2001).

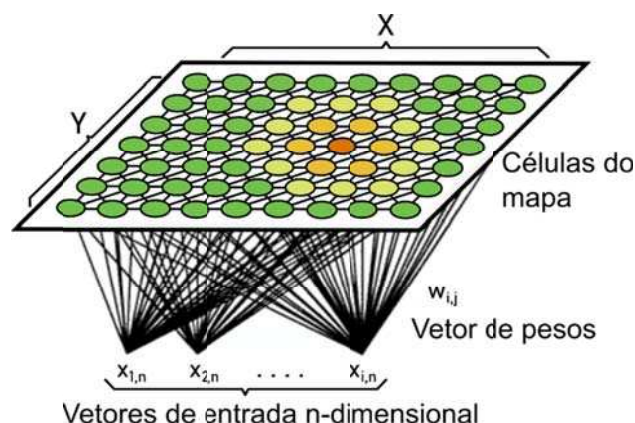


Figura 2.6: Representação esquemática de um mapa SOM. Modificado de: http://www.statistics4u.com/fundstat_eng/cc_ann_kohonen.html.

Os mapas SOM são compostos por um grid regular (que pode possuir uma forma retangular ou toroidal – Figura 2.7 – entre outras) com j número de células dispostas em p linhas x q colunas. As células podem ter formato retangular ou hexagonal. Os dados de entrada l -dimensionais $x = (x_1, x_2, \dots, x_l)$ são mapeados nas células do mapa, que são associadas cada uma a um vetor de referência $m = (m_1, m_2, \dots, m_l)$. A escolha do número ótimo de células de um mapa pode depender do número de amostras. [Vesanto e Alhoniemi \(2000\)](#) sugerem um número de $5 * \sqrt{N}$, sendo N o número de amostras do banco de dados.

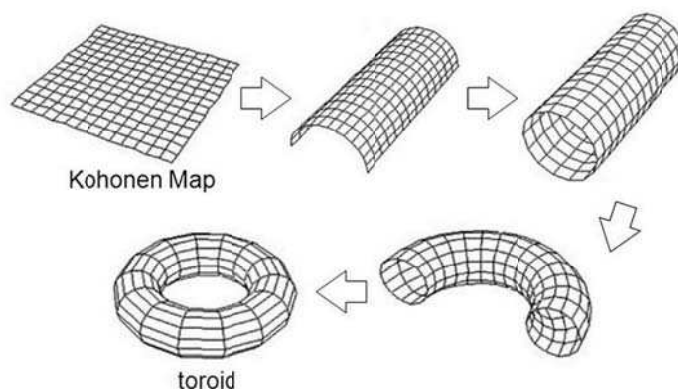


Figura 2.7: Visualização em 3D de um mapa SOM com topologia toroidal.

No treinamento do mapa, cada vetor de referência m_j é inicializado com certos valores. Embora o SOM seja robusto à inicialização, valores iniciais bem definidos podem permitir que o algoritmo chegue a uma boa solução mais rapidamente. Três maneiras usuais de inicialização dos vetores de referência são: 1) valores pequenos escolhidos aleatoriamente; 2) valores de amostras escolhidas

aleatoriamente; 3) de forma linear, onde os vetores m são inicializados em ordem, no espaço linear definido pelos dois principais autovetores dos dados de entrada.

O treinamento se dá iterativamente (t vezes) em dois estágios. No primeiro, denominado de etapa competitiva, uma amostra x , escolhida aleatoriamente, é comparada com os vetores de referência m do mapa. A similaridade é quantificada através de uma determinada medida de distância entre vetores, tipicamente a euclidiana. A célula vencedora c (também chamada de *best-matching unit*, ou BMU) que mais se aproxima de x é:

$$\|x - m_c\| = \min_i \{\|x - m_i\|\} \quad \text{Eq. 1}$$

Onde $\| - \|$ representa a medida de distância utilizada. Em seguida os vetores de referência são atualizados. A regra de atualização do vetor de referência da célula i é:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \quad \text{Eq. 2}$$

Onde $h_{ci}(t)$ é o kernel de vizinhança, centralizada na célula vencedora c . O kernel de vizinhança é formado pela função de vizinhança $h(d, t)$ e pela taxa de aprendizado $\alpha(t)$:

$$h_{ci}(t) = h(\|r_c + r_i\|, t)\alpha(t) \quad \text{Eq. 3}$$

Onde r_c e r_i são as localizações das células c e i , e $\sigma(t)$ é o raio de vizinhança (Figura 2.8.A). A função de vizinhança escolhida é tipicamente a Gaussiana (Figura 2.8.C):

$$\exp\left(-\frac{\|r_c + r_i\|^2}{2\sigma^2(t)}\right) \quad \text{Eq. 4}$$

Outras funções de vizinhança também podem ser utilizadas, como a função “bolha” (Figura 2.8.B), cuja fórmula é:

$$\begin{cases} 1, & \|r_i - r_c\| \leq \sigma(t) \\ 0, & \text{caso contrário} \end{cases} \quad \text{Eq. 5}$$

Os parâmetros $\sigma(t)$ e $\alpha(t)$ diminuem progressivamente (geralmente de forma linear) ao longo das K iterações, e devem ser definidos na configuração do algoritmo (Vesanto 1997, Kaski 2017). Na versão “batch” do algoritmo SOM,

diferente da versão clássica (ou “online”), as células são atualizadas somente após todos os vetores x terem sido apresentados ao mapa. Nessa versão não existe o parâmetro $\alpha(t)$. Quando o raio de vizinhança $\sigma(t)$ diminui a ponto de somente a célula central ser atualizada, o algoritmo passa a funcionar de maneira igual ao K-means (Kohonen 2001).

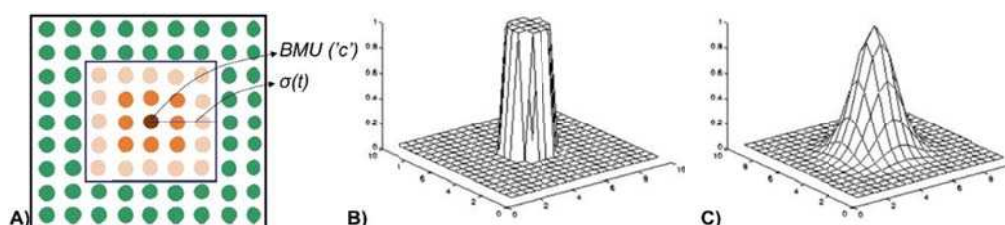


Figura 2.8: A) Zona de vizinhança centralizada na célula vencedora “c” (BMU). A intensidade da cor laranja indica o quanto cada célula será atualizada na iteração t , e o raio $\sigma(t)$ indica até onde as células do mapa serão influenciadas. Adaptado de Moreira e Costa (1996). B) Função de vizinhança “bolha”. C) Função de vizinhança Gaussiana. Adaptado de Vesanto (1997).

2.3.1.1. Métodos de visualização de SOM

As visualizações do mapa SOM permitem a averiguação de sua qualidade em relação ao treinamento do modelo, e de relações existentes entre os dados. Alguns importantes métodos de visualização são: planos de componentes, *U-matrix*, número de vetores, e distância média (erro médio de quantização) (Ultsch e Siemon 1990, Kohonen 2001, Kohonen 2013).

2.3.1.1.1. Matriz-U, números de vetores e distância média

Os mapas *U-matrix* (*unified distance matrix* ou “matriz-U”) , número de vetores e de distância média, são utilizados para avaliar a qualidade do treinamento (Kohonen 2001, Vesanto 1997) e identificar regiões ou células com características distintas, de possível interesse para maior investigação. De modo geral, esses mapas mostram características de homogeneidade interna das células do mapa SOM, e também em relação às suas vizinhanças.

A matriz-U é uma forma de visualização de uma mapa SOM utilizada para avaliar a qualidade topológica do mesmo e o padrão de distribuição dos dados (Ultsch e Siemon 1990). Basicamente, é um mapa que mostra a similaridade entre

as células vizinhas treinadas do mapa. A matriz-U pode ser construída de duas formas. Na primeira, são interpoladas novas células entre cada par de células do mapa. As novas células recebem valores que correspondem à distância (medida com o método de distância empregado no treinamento do mapa – e.g. distância euclidiana) entre as duas células do par e as células originais recebem o valor da média das distâncias com todas as células vizinhas. Na outra forma, as células interpoladas são omitidas, aparecendo somente as células com os valores da distância média.

O mapa do número de vetores mostra quantos vetores foram mapeados em cada célula do mapa SOM. Essa visualização mostra se os vetores de dados foram suficientemente distribuídos pelo mapa durante o treinamento, uma vez que mapas insuficientemente treinados tendem a mostrar uma distribuição muito heterogênea. Em modelos bem treinados, as células e regiões com características mais comuns irão conter um maior número de vetores, e de forma contrária, aquelas com características mais raras apresentarão um número menor de vetores.

O mapa de distância média quantifica a dissimilaridade média entre os vetores e suas respectivas células, de forma que menores valores representam maior homogeneidade. A distância média, também chamada de erro médio de quantização ([Kohonen 2001](#)), é calculada através da métrica de distância escolhida para o modelo SOM (e.g. distância euclidiana).

2.3.1.1.2. Planos de componentes

A visualização de planos de componentes permite a avaliação da contribuição de cada variável do modelo em cada uma das células do mapa SOM, de forma que permite observar a variação dos pesos das variáveis nas diferentes regiões do mapa SOM ([Vesanto 1997](#), [Kohonen 2013](#)).

2.3.1.2. Counter propagation network

A técnica *counter propagation network* (CPN) ([Melssen et al. 2006](#)) é uma forma de se projetar assinaturas “aprendidas” pelo modelo SOM de volta no espaço geográfico, obtendo-se um tipo de modelo preditivo “semi-supervisionado”. Nessa técnica, os valores das células são projetados de volta nas coordenadas espaciais de cada vetor mapeado nessa célula. Para variáveis categóricas, como unidades

geológicas ou ocorrências, as células assumem o valor da proporção de vetores positivos para tal característica em relação ao total de vetores mapeados na célula, recebendo um valor entre 0 e 1. Utilizando-se a técnica CPN é possível visualizar características geológicas escolhidas (unidades geológicas ou mineralizações) no mapa SOM, de forma similar aos mapas de componentes, e avaliar o seu comportamento em relação às variáveis modeladas.

2.4. Delimitação de domínios geológico-geofísicos

A delimitação de domínios geológico-geofísicos foi realizada com duas técnicas de agrupamento (aprendizado não-supervisionado) distintas, para comparação dos resultados. As duas técnicas utilizadas foram a combinação de SOM com K-means, e o agrupamento hierárquico espacialmente restrito (AHER). A primeira técnica é bastante utilizada para o agrupamento de dados geocientíficos espacializados, no entanto, não leva em conta as coordenadas espaciais dos dados. Por esse motivo, essa técnica foi escolhida para comparação com o agrupamento hierárquico espacialmente restrito que, por sua vez, leva em conta a localização dos dados durante a execução do agrupamento.

Para criação dos domínios, foram utilizados dados gamaespectrométricos (CT, K, eU, eTh) e magnetométricos (gradiente total) (Figura 2.2 A a E). Como nesta fase se deseja investigar feições que se refletem na superfície, foram incluídos nesta análise dados de modelo digital de terreno (MDT), obtidos do sensor ALOS PALSAR (Figura 2.2.F).

2.4.1. Agrupamento por SOM + K-means

A combinação de SOM com um algoritmo de agrupamento (e.g. K-means, agrupamento hierárquico) é uma técnica que tem sido utilizada para o agrupamento de dados geocientíficos em estudos de potencial mineral. Por exemplo, [Leite e Souza Filho \(2010\)](#), [Carneiro *et al.* \(2012\)](#) e [Chudasama *et al.* \(2022\)](#) utilizaram K-means para agrupar as células de um mapa SOM e identificar áreas de interesse prospectivo, enquanto [Gazley *et al.* \(2021\)](#) e [Kuhn *et al.* \(2019\)](#) utilizaram o agrupamento hierárquico no lugar do K-means.

O K-means é um algoritmo baseado em dissimilaridade, que é quantificada por medidas de distâncias, como as distâncias euclidiana, Manhattan,

Minkowski e Mahalanobis (Kaufman e Rousseeuw 2005). Esse algoritmo é considerado do tipo representativo, que busca, para cada grupo, um ponto mais representativo possível do mesmo. Geralmente (como no caso do K-means), esse ponto é a média do grupo, caso no qual ele é chamado de “centroide” (Zaki e Meira Jr. 2020, Aggarwal e Reddy 2016).

Formalmente, dado um conjunto de grupos $C = \{C_1, C_2, C_3, \dots, C_k\}$ o algoritmo busca um agrupamento C^* que minimize a soma do erro quadrado (*sum of squared erros* - SSE) do agrupamento:

$$C^* = \min_C \{SSE(C)\}, \quad \text{Eq. 6}$$

$$SSE(C) = \sum_{i=1}^k \sum_{x_j \in C_i} ||x_j - c_i||^2. \quad \text{Eq. 7}$$

Onde c_i é o centroide de um grupo i , calculado como:

$$c_i = \frac{\sum_{x_k \in C_i} x_k}{n_i}, \quad \text{Eq. 8}$$

O algoritmo inicia a execução atribuindo k valores aleatórios (ou valores pré-definidos, ou seguindo algum método de inicialização específico) aos centroides, no espaço dos dados. A cada iteração, k grupos serão formados, atribuindo-se cada ponto ao centroide mais próximo. Em seguida, os centroides são recalculados (Figura 2.9). O processo é repetido até que um critério de parada (e.g. número máximo de iterações), ou de convergência (as iterações produzam mudanças insignificantes ou nulas nos centroides) (Zaki e Meira Jr. 2020, Aggarwal e Reddy 2016).

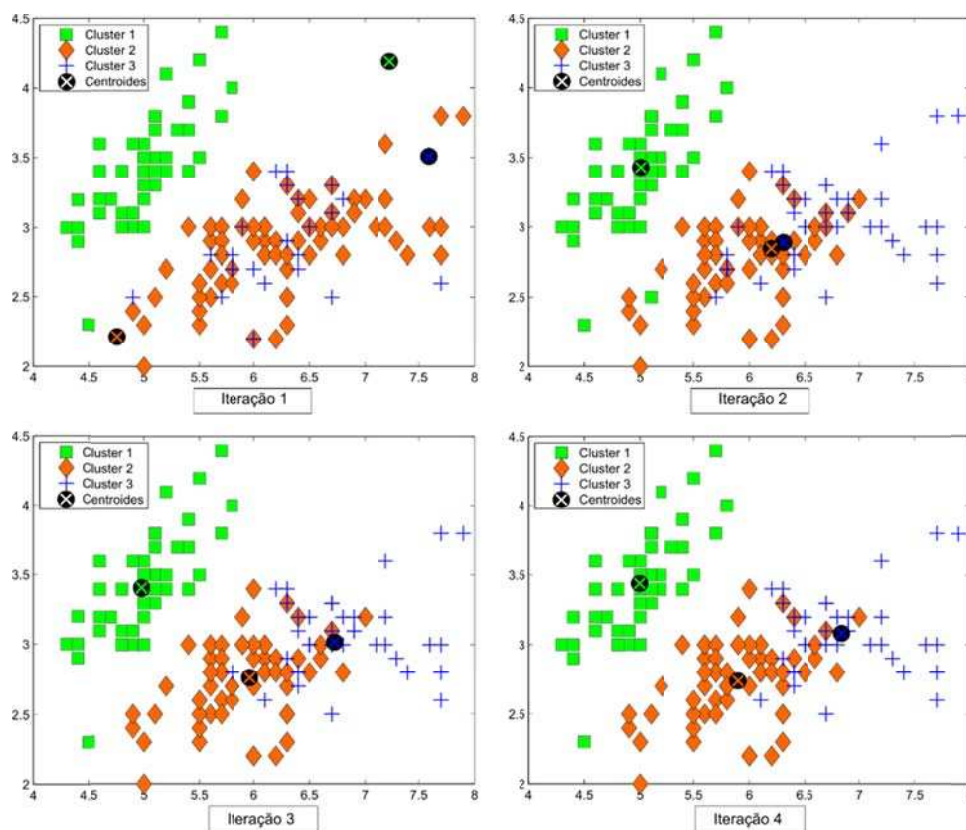


Figura 2.9: Exemplo de execução do algoritmo K-means. A cada iteração, os centroides são recalculados (se movimentam), e a atribuição dos pontos aos grupos pode ser modificar. Adaptado de [Aggarwal e Reddy \(2016\)](#).

Dado um número suficiente de iterações, o SSE decresce de forma monótona até um mínimo local. No entanto, os valores iniciais dos centroides influenciarão no resultado final e no tempo de convergência da execução. Diferentes abordagens foram propostas para se otimizar a escolha dos valores iniciais dos centroides ([Aggarwal e Reddy 2016](#)), e neste trabalho foi utilizado o algoritmo K-means++ ([Arthur e Vassilvitskii 2007](#)). Este algoritmo inicializa os centroides seguindo um critério probabilístico, que otimiza a distribuição dos mesmos dentro do espaço dos dados. Para o primeiro centroide, é escolhido um ponto de forma aleatória. Nos seguintes, são escolhidos, sequencialmente, pontos que estejam o mais distante possível dos centroides já escolhidos. [Arthur e Vassilvitskii \(2007\)](#) demonstram que esse método de inicialização resulta em ganhos em acurácia e velocidade em relação ao K-means original.

Na combinação SOM + K-means, em vez de se agrupar o conjunto de dados diretamente com K-means, primeiramente, é gerado um mapa SOM desse conjunto, e então o K-means é aplicado para agrupar as células do mapa SOM (vide

item 2.3.1). Dentre as vantagens dessa técnica, que trabalha com 2 níveis de abstração (Figura 2.10), em relação ao uso somente de K-means, destacam-se: as capacidades de visualização que o SOM possibilita, de diferentes propriedades e características dos dados; a redução da demanda computacional; suavização de ruídos no banco de dados (Vesanto e Alhoniemi 2000).

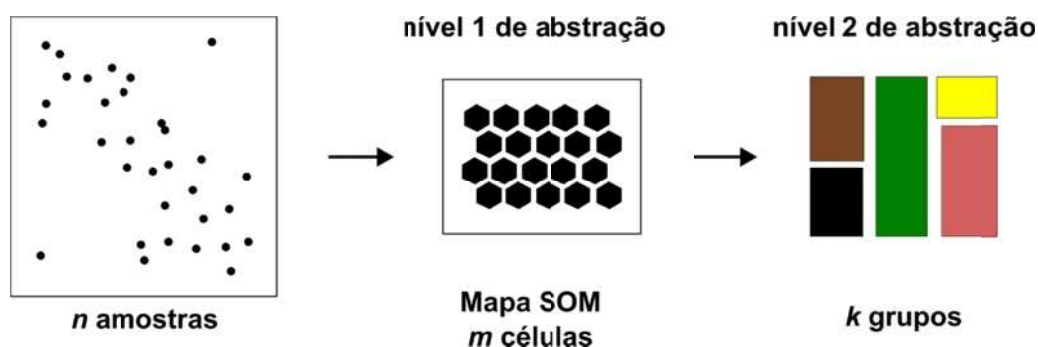


Figura 2.10: Ilustração esquemática do agrupamento do mapa SOM, produzindo 2 níveis de abstração do conjunto e dados. Adaptado de Vesanto e Alhoniemi (2000).

Neste trabalho, o treinamento do mapa SOM foi realizado como descrito no item 2.3.1. Posteriormente, o agrupamento das células do mapa SOM por K-means foi realizado com o pacote *scikit-learn* (Pedregosa et al. 2011) para Python.

2.4.2. Agrupamento hierárquico espacialmente restrito (AHER)

O AHER é uma versão espacialmente restrita do, bem conhecido, algoritmo de Agrupamento Hierárquico (AH) (Hastie et al. 2009, Zaki e Meira Jr. 2020). Assim como o K-means, o AH é um algoritmo baseado em dissimilaridade quantificada por medidas de distâncias. O AH cria uma sequência de partições (*clusters* ou grupos) nos dados, que podem ser visualizados por uma árvore hierárquica chamada de dendrograma, no qual, no nível mais baixo, encontram-se os pontos de dados individuais (as folhas da árvore) e, no topo, o grupo que inclui todos os pontos de dados (Figura 2.11). Nos níveis intermediários, encontram-se os agrupamentos, de forma que, quanto mais alto no dendrograma, menor é k , que representa o número de grupos (Zaki e Meira Jr. 2020).

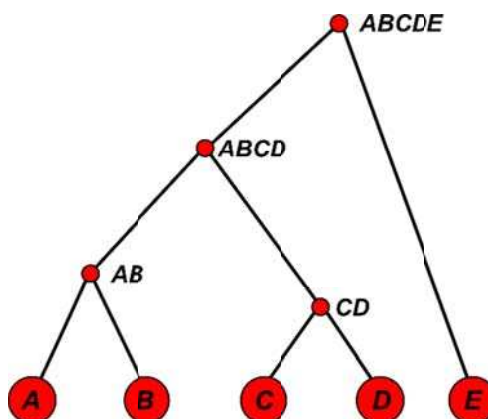


Figura 2.11: Exemplo de dendrograma produzido por agrupamento hierárquico. Na base se encontram os pontos de dados individuais (A, B, C, D, E) e, a cada nível da árvore, os pontos, ou grupos, se juntam para formar novos, e maiores, agrupamentos (modificado de [Zaki e Meira Jr., 2020](#)).

Existem duas abordagens algorítmicas de AH: a divisiva e a aglomerativa. A primeira, inicia no topo do dendrograma (todos os pontos em um grupo) e progressivamente segmenta os grupos até que sobrem os pontos individuais, enquanto a segunda, executa o processo contrário, juntando os dados a partir dos pontos individuais. As técnicas aglomerativas são mais comuns por serem computacionalmente mais eficientes e adequadas na maioria das aplicações ([Kaufman e Rousseeuw 2005](#)). Por esse motivo, esta foi a abordagem aplicada neste trabalho.

No AH aglomerativo, dado um conjunto de grupos $C = \{C_1, C_2, C_3, \dots, C_m\}$, busca-se o par de grupos mais próximos C_i e C_j , os quais são unidos em um novo grupo $C_{ij} = C_i \cup C_j$. Em seguida, atualiza-se o conjunto de grupos C , removendo C_i e C_j , e adicionando C_{ij} , da seguinte forma: $C = (C \setminus \{C_i, C_j\}) \cup \{C_{ij}\}$. O processo é repetido até que C contenha apenas um grupo. Como em cada iteração o número de grupos k é reduzido em um, o resultado é uma sequência de n agrupamentos aninhados, onde n é o número de pontos do conjunto de dados ([Zaki e Meira Jr. 2020](#)).

Diferentes formas de se calcular a proximidade d entre dois grupos, ou de um ponto e um grupo, podem ser utilizadas. Esses critérios são chamados de “ligação” (*linkage*), e os principais tipos são “única” (d_u , Eq. 9), “completa” (d_c , Eq. 10), “média” (d_m , Eq. 11) e “ward” (d_w , Eq. 12) (em inglês, respectivamente: *single linkage*, *complete linkage*, *average linkage* e *ward*), definidas respectivamente como:

$$d_u(C_i, C_j) = \min_{u \in C_i, v \in C_j} \{||u - v||\}, \quad \text{Eq. 9}$$

$$d_c(C_i, C_j) = \max_{u \in C_i, v \in C_j} \{||u - v||\}, \quad \text{Eq. 10}$$

$$d_m(C_i, C_j) = \frac{\sum_{k=1}^{n_i} \sum_{l=1}^{n_j} ||u_k - v_l||}{n_i \cdot n_j}, \quad \text{Eq. 11}$$

$$d_w(C_i, C_j) = \sum_{x_q \in C_i \cup C_j} ||x_q - c_{C_i \cup C_j}||^2 - \sum_{x_q \in C_i} ||x_q - c_{C_i}||^2 - \sum_{x_q \in C_j} ||x_q - c_{C_j}||^2. \quad \text{Eq. 12}$$

onde, c_i representa o centroide de um grupo i , e n_i o número de pontos de um grupo i . A ligação única minimiza a distância entre dois grupos através da menor distância entre dois pontos, um de cada grupo. A completa, é similar à única, mas busca a menor distância entre os pontos mais distantes entre si de diferentes grupos. A ligação média minimiza a distância média entre cada par de pontos de dois grupos. A *ward*, minimiza a soma dos erros ao quadrado dentro dos grupos, minimizando a variância intra-grupo. Cada uma pode ser considerada mais adequada para diferentes aplicações e irão gerar dendogramas (e agrupamentos) distintos para os mesmos conjuntos de dados iniciais ([Hastie et al. 2009](#), [Zaki e Meira Jr. 2020](#)). Neste trabalho, foi utilizada a ligação *ward*, que mostrou desempenho satisfatório em aplicação de regionalização ([Lattimer e Lattimer, 2022](#)).

Para se obter a restrição espacial no AH aglomerativo, é necessária uma matriz de conexões ([Murtagh 1985](#)). Neste trabalho, uma matriz foi gerada a partir das coordenadas (X, Y) dos pontos de dados, onde cada ponto foi conectado com seus vizinhos em um raio de 708 m, permitindo, a depender da localização, um número máximo de 8 a 24 de pontos candidatos a um agrupamento (Figura 2.12). Este comprimento de raio foi escolhido para permitir que cada ponto pudesse se conectar a, além de seus vizinhos imediatos, os vizinhos subsequentes, possibilitando uma “margem de erro” sem prejudicar a continuidade espacial do agrupamento. Para o cálculo de dissimilaridade, foi utilizada a medida de distância Minkowski de ordem $p=2$ (equivalente à distância euclidiana) ([Kaufman e Rousseeuw 2005](#)). O AHER foi implementado neste estudo com o pacote *scikit-learn* ([Pedregosa et al. 2011](#)) para Python.

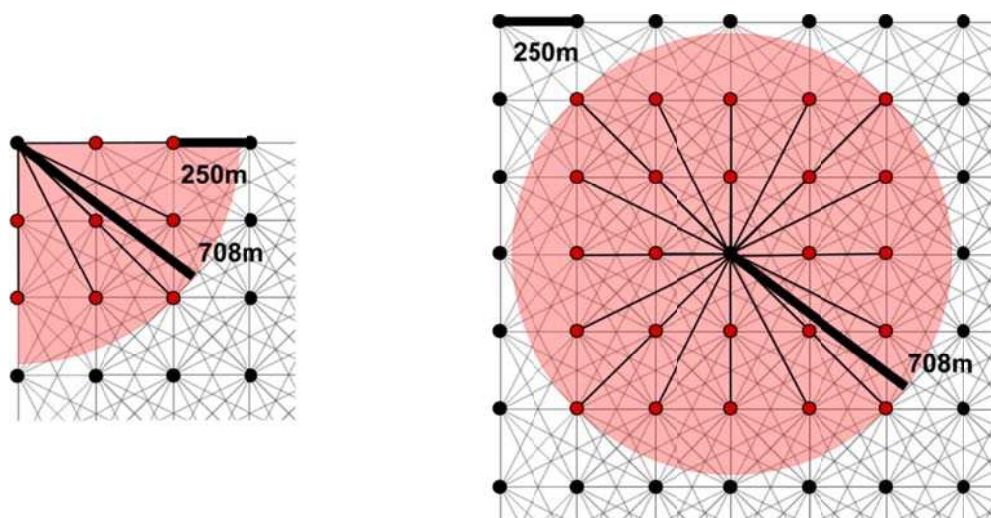


Figura 2.12: Ilustração da matriz de conexões com raio de 708 m utilizada neste trabalho, mostrando o número máximo de pontos candidatos a um agrupamento. A figura da esquerda mostra o exemplo de um ponto localizado em um canto da área, e a da direita, de um ponto localizado a mais do que 750 m das bordas da área de estudo.

2.4.3. Seleção do número ótimo de grupos

Diversas técnicas para avaliação da qualidade interna de agrupamentos existem na literatura. Em geral, os métodos de validação são usados para se averiguar e comparar a qualidade dos resultados de agrupamentos. Podem ser utilizados para se determinar o número ótimo de grupos, funcionando como uma métrica a ser minimizada, testando-se diferentes números de grupos (Figura 2.13). Em geral, essas métricas são baseadas em dois critérios: i) compactação, que mede quão similares são os objetos dentro de um grupo - essa medida é baseada na variância ou distância; ii) separação, que mede o quão separados são os grupos entre si. Em geral, são adotadas medidas de distância entre o centro dos grupos (centroides ou medoides), ou entre os objetos de diferentes grupos (Aggarwal e Reddy 2016).

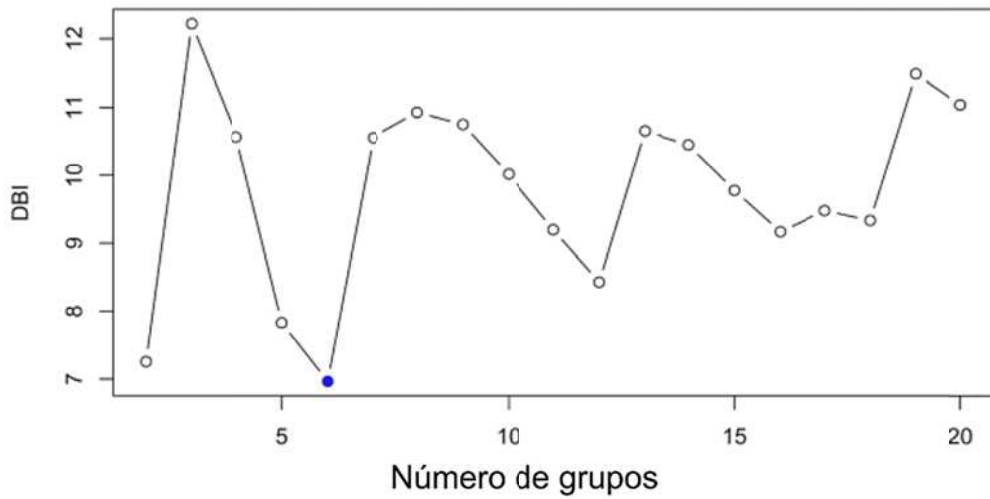


Figura 2.13: Exemplo do uso de uma medida de validação de grupos (Davies-Bouldin index - DBI) para se descobrir o número ótimo de grupos. No exemplo da figura, o número ótimo de grupos é 6, que minimiza o DBI. Adaptado de [Gazley et al. \(2021\)](#).

Um índice muito utilizado para esse fim é o Davies-Bouldin (DB) (e.g. [Leite e Souza Filho 2010](#), [Gazley et al. 2021](#), [Carneiro et al. 2012](#)). Este índice mede o quão compacto os grupos são internamente, e quão separados são entre si, sendo que valores mais baixos representam grupos mais compactos e mais separados. Considerando μ_i a média (centroide) de um grupo:

$$\mu_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j, \quad \text{Eq. 13}$$

e σ_{μ_i} a dispersão dos pontos ao redor de μ_i :

$$\sigma_{\mu_i} = \sqrt{\frac{\sum_{x_j \in C_i} ||x_j - \mu_i||^2}{n_i}} = \sqrt{\text{var}(C_i)}. \quad \text{Eq. 14}$$

o índice DB medido para um par de grupos C_i e C_j é definido como:

$$DB_{ij} = \frac{\sigma_{\mu_i} + \sigma_{\mu_j}}{||\mu_i - \mu_j||}. \quad \text{Eq. 15}$$

Assim, DB_{ij} mede quão compacto os grupos são em relação à separação entre si. Dessa forma, o índice DB para um agrupamento completo é definido como:

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \{DB_{ij}\}. \quad \text{Eq. 16}$$

ou seja, para cada grupo C_i , seleciona-se o grupo C_j que apresenta o maior DB_{ij} (Zaki e Meira Jr. 2020).

Características particulares dos bancos de dados podem afetar a efetividade das métricas de validação de diferentes formas (Aggarwal e Reddy 2016). Considerando que este trabalho compreende um estudo de mapeamento de potencial mineral, cujo objetivo é delimitar áreas que apresentem maior probabilidade de conter ocorrências minerais, aqui foram utilizadas métricas específicas que podem quantificar essas características como critério na seleção do melhor número de grupos k .

Nesse sentido, foram empregados 3 parâmetros: i) valor- t , que mede a associação espacial entre um grupo e as ocorrências; ii) o número máximo de ocorrências presentes dentro de um grupo; iii) a área do cluster com o maior número de ocorrências. O valor- t é calculado pelo método de análise espacial clássico conhecido como Peso de Evidências (*Weights of Evidences*), sendo também chamado de “contraste studentizado” em referência ao valor- t do Teste-T de Student (Bonham-Carter 1994, Agterberg 1989).

Para comparação, a otimização do número de grupos foi realizada tanto com as três métricas citadas, como com o índice DB.

2.5. Determinação de famílias de lineamentos prospectivos através de análise de fractais

Os depósitos minerais hidrotermais têm sua origem relacionada a processos que envolvem fontes de calor, fluidos e condutos por onde esses fluidos aquecidos se movimentam. Estes processos, por sua vez, são controlados por elementos geológicos, como intrusões ígneas, contatos litológicos e/ou falhas. Estudos mostram que esses elementos, assim como diversos outros elementos geológicos, apresentam disposição espacial cuja geometria possui características fractais/multifractais (Roberts *et al.* 1999, Sanderson *et al.* 1994, Carlson 1991, Turcotte 2002, Mandelbrot 1983).

Essa característica pode ser observada também na distribuição de ocorrências minerais hidrotermais (e.g. Wang *et al.* 2015, Wang *et al.* 2012), cuja formação constitui-se em processos singulares, definidos como processos não-lineares, que resultam na descarga de grandes quantidades de energia e massa em um curto tempo (Cheng 2008, Cheng *et al.* 1994, Cheng *et al.* 1996).

Zuo (2016) propôs o uso de tal princípio para a quantificação da associação espacial entre objetos geológicos e mineralizações geneticamente associadas, através do cálculo de um *índice de singularidade* denotado por α , obtido pelo ajuste da função:

$$\rho = C\varepsilon^{\alpha-2} \quad \text{Eq. 17}$$

onde ε é a distância do elemento geológico, ρ é a densidade de depósitos na distância ε , C é uma constante, e 2 representa a dimensão euclidiana de um plano. Valores de $\alpha < 2$ indicam uma relação espacial significativa, que é maior, quanto menor for o valor de α (Zuo 2016).

Como não existem aplicações disponíveis ao público que realizem esta análise, neste trabalho, para se discriminar as famílias de lineamentos prospectivos, o cálculo de α foi realizado em Python, de acordo com os seguintes passos, descritos por Zuo (2016):

- 1) definição empírica de uma distância para o cálculo de ε em intervalos regulares;
- 2) cálculo de ρ_i como:

$$\rho_i = \frac{n_i}{\varepsilon_i}, \quad \text{Eq. 18}$$

onde n_i é o número de ocorrências presentes a uma distância menor que ε_i dos lineamentos de cada família.

- 3) ajuste de uma reta em um gráfico log-log de ρ_i versus ε_i , onde o gradiente da reta será $\alpha - 2$.

O gradiente da reta foi obtido através do ajuste da função exponencial por meio da função *curve_fit* do pacote *SciPy*, de Python.

O conjunto de dados de lineamentos utilizado neste trabalho é proveniente de [Barros *et al.* \(2021\)](#). Foi produzido pela interpretação de imagens derivadas de dados aeromagnetométricos e de modelo digital de terreno (SRTM), levando-se em conta os diferentes eventos deformacionais que ocorreram na área (Figura 2.14). Os autores separam os lineamentos interpretados em 12 famílias, apresentadas na Tabela 2.

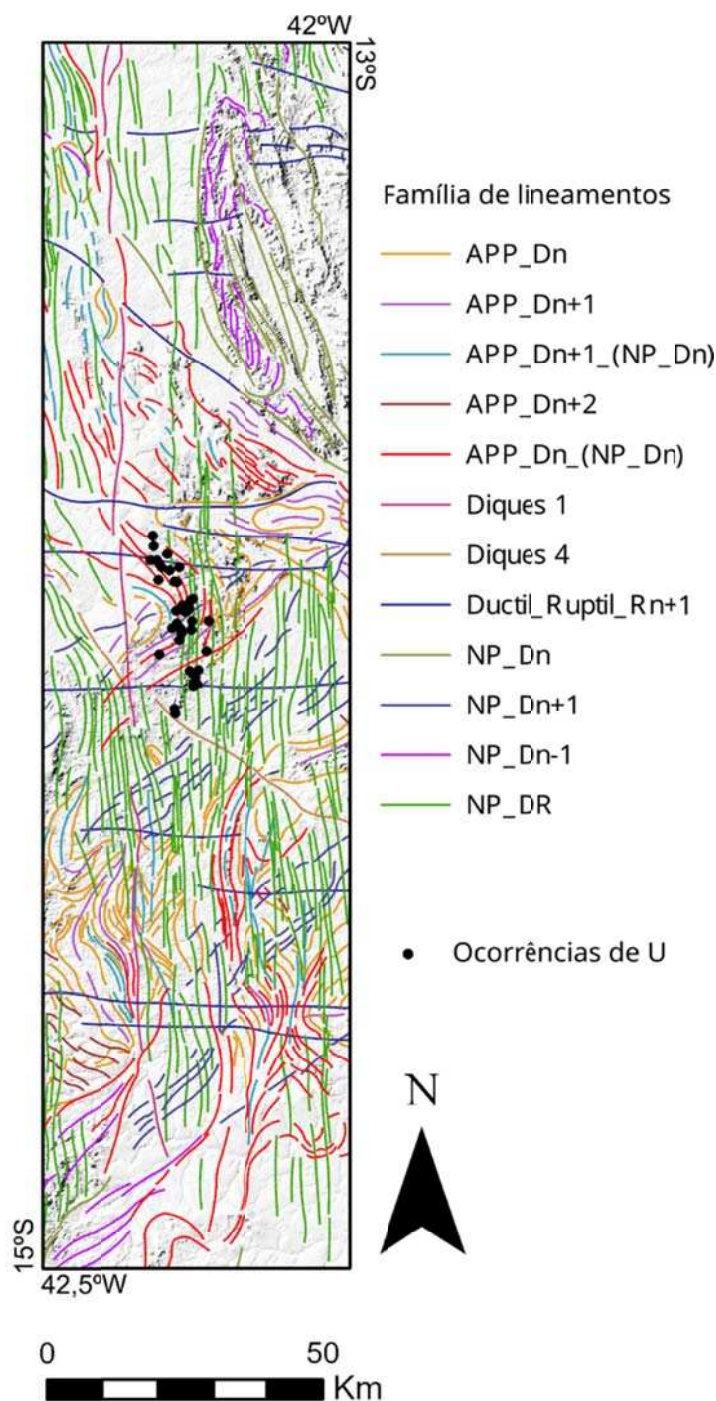


Figura 2.14: Famílias de lineamentos interpretados na área de trabalho ([Barros *et al.* \(2021\)](#)). As famílias são explicadas na Tabela 2.

Tabela 2: Idades e fases de deformação associadas a cada família de lineamentos interpretados (Barros *et al.* 2021). APP_Dn representa a fase de deformação principal que ocorreu entre o Arqueano e o Paleoproterozoico na área; NP_Dn representa a fase principal de deformação neoproterozoica, relacionada ao ciclo brasileiro na área.

Família	Idade	Fase	Obs
NP_Dn-1	Neoproterozoico	Dn-1	
NP_Dn	Neoproterozoico	Dn	
NP_Dn+1	Neoproterozoico	Dn+1	
APP_Dn_(NP_Dn)	Neoproterozoico	Dn	Reativação de APP_Dn
APP_Dn+1	Arqueano a Paleoproterozoico	Dn+1	
APP_Dn+2	Arqueano a Paleoproterozoico	Dn+2	
Diques 4	Pré NP_Dn	-	Enxame Chapada Diamantina e Curaçá
DRn+1	-	-	Estruturas dúcteis-rúpteis
APP_Dn	Arqueano a Paleoproterozoico	Dn	
NP_DR	Neoproterozoico	pós-Dn	Estruturas dúcteis-rúpteis
APP_Dn+1_(NP_Dn)	Neoproterozoico	Dn	Reativação de APP_Dn+1
Diques 1	Fanerozoico	-	Enxame Serra do mar

2.6. Geração do modelo de prospectividade

A integração dos mapas de evidência e geração do modelo de prospectividade foi realizada através de técnica de classificação por aprendizado supervisionado. Para a geração do mapa de prospectividade, primeiro foi gerado o conjunto de dados de treinamento, utilizado para ajuste, avaliação e otimização do modelo. Em seguida, foi selecionada, dentre 6 técnicas de classificação, aquela que apresentou o melhor desempenho para os dados da área, a qual foi utilizada para a geração do modelo preditivo. Ao final, utilizou-se a técnica SHAP para a avaliação da influência de cada variável nas previsões do modelo. Essas três etapas são descritas nos itens a seguir.

2.6.1. Geração do conjunto de dados de treinamento

Um dos grandes desafios da aplicação de métodos de ML no MPM incide no desbalanceamento das classes (pontos mineralizados vs. não mineralizados), que constituem os dados de treinamento do modelo. Este tipo de problema também

é comum a outras aplicações de ML, como por exemplo aquelas que envolvem a detecção de anomalias, onde a classe majoritária (dados não anômalos) compreendem mais que 99% dos dados. Isso ocorre, por exemplo, em aplicações de detecção de transações financeiras fraudulentas (Dal Pozzolo *et al.* 2014) e em sistemas de monitoramento (Russo *et al.* 2020).

Uma abordagem utilizada em MPM para o treinamento de modelos supervisionados desbalanceados é a subamostragem da classe majoritária (loais não-mineralizados), e/ou a sobreamostragem da classe minoritária (loais mineralizados), de forma a diminuir a desproporção entre as classes.

Para se aumentar o conjunto de dados da classe positiva (loais mineralizados), uma estratégia é a sobreamostragem por criação de amostras sintéticas baseadas nas amostras conhecidas. Por exemplo, Brown *et al.* (2003) avaliaram a sobreamostragem dos loais mineralizados através da criação de pontos sintéticos via adição de ruído. Os melhores resultados foram obtidos com um aumento de 46 para 1058 pontos mineralizados, com aplicação de 40% a 80% de ruído uniforme aleatório. Pontos não-mineralizados foram subamostrados aleatoriamente para que a proporção entre as classes se aproximasse de 1:1.

Outra ferramenta que pode ser utilizada para sobreamostragem, pela criação de amostras sintéticas, é o algoritmo *synthetic minority oversampling* (SMOTE) (Chawla *et al.* 2002). Esta técnica foi aplicada por Li *et al.* (2020) e Prado *et al.* (2020) em modelos de RF e SVM, respectivamente. Os autores obtiveram melhores resultados com taxas de sobreamostragem de pontos mineralizados de 900% e 2000% respectivamente, combinado com subamostragem dos pontos sem mineralização, produzindo balanceamento das classes com proporções de aproximadamente 1:1. Similarmente, Hariharan *et al.* (2017) utilizaram SMOTE com taxa de sobreamostragem das ocorrências de 500% em um modelo RF, produzindo uma proporção de ocorrências para não ocorrências de aproximadamente 1,2.

Neste trabalho, utilizou-se uma variação do algoritmo SMOTE, denominada *BorderlineSMOTE* (Han *et al.* 2005) para a sobreamostragem da classe minoritária (ocorrências de U). O *BorderlineSMOTE* é uma opção interessante para o caso em questão, pois busca gerar amostras sintéticas com características que se apresentam no limite entre aquelas das amostras negativas e positivas, que são as mais difíceis de serem classificadas. Por esse motivo, é uma ferramenta interessante

para conjuntos de dados que, além de desbalanceados, possuem uma distribuição complexa no espaço dos dados. O algoritmo *BorderlineSMOTE* foi implementado neste trabalho por meio do pacote de Python *imbalanced-learn* (Lemaître et al. 2017).

Na geração do conjunto de dados de treinamento, para a classe negativa, em geral, selecionam-se aleatoriamente, ou seguindo-se critérios específicos, pontos sem mineralizações conhecidas (e.g. Prado et al. 2020, Rodriguez-Galiano et al. 2015, Brown et al. 2003). Busca-se selecionar um número de pontos que permita uma distribuição pela área de estudo de forma suficiente para se obter uma boa representatividade dos padrões geológicos presentes. Porém, para se evitar a desproporção dos dados, o número de pontos não-mineralizados não deve ser muito superior ao de pontos mineralizados.

Para se diminuir a chance de seleção acidental de um ponto que contenha uma ocorrência não conhecida, Prado et al. (2020) utilizaram, como critério restritivo, um raio de distância mínima das ocorrências conhecidas, definido através de análise de padrões de pontos pela função chamada de *Pair Correlation Function* (PCF). A função PCF é simbolizada por $g(r)$, onde r é a distância entre duas posições no espaço. Esta função avalia a propriedade de segunda ordem do padrão de distribuição dos pontos, ou seja, quantifica como a presença de um ponto influencia na ocorrência de outros pontos no espaço (Stoyan e Stoyan 1996, Agterberg 2013, Ripley 1988, Illian et al. 2007). Valores de $g(r)$ maiores do que 1 indicam uma tendência de aglomeração de pontos (a presença de um ponto aumenta a chance de ocorrência de outro ponto em um raio r). Valores iguais a 1 indicam uma distribuição aleatória dos pontos (processo de Poisson uniforme), e valores de $g(r)$ menores que 1 indicam a tendência de dispersão dos pontos (a presença de um ponto diminui a chance de ocorrência de outro ponto em um raio r). Neste trabalho, a função PCF foi calculada por meio do pacote de R *spatstat* (Baddeley et al. 2015).

2.6.2. Seleção do classificador

Para a integração dos mapas de evidência e geração do modelo de prospectividade, foram avaliadas seis técnicas distintas de classificação, com o objetivo de se identificar aquela com o melhor desempenho para o conjunto de

dados deste estudo. As técnicas testadas foram: *random forests* (RF), *XGBoost* (XGB), *gradient boosting* (GB), regressão logística (RL), *support vector machine* (SVM) e *naive bayes* (NB). Três dessas técnicas são do tipo *ensemble* (RF, XGB e GB), uma classe de algoritmos de classificação que têm obtido excelentes resultados em MPM em trabalhos recentes (e.g. [Fan et al. 2022](#), [Lachaud et al. 2023](#), [Rodriguez-Galiano et al. 2015](#), [Parsa 2021](#), [Yin e Li 2022](#), [Keykhay-Hosseinpour et al. 2020](#)). As outras 3, são técnicas clássicas, baseadas em abordagens distintas (SVM, NB e RL), que serão descritas abaixo.

As técnicas de aprendizado baseadas em algoritmos do tipo *ensemble* (em assembleia) se baseiam na observação de que diversos modelos “fracos”, ou simples, quando combinados, podem apresentar resultados preditivos robustos, com baixos viés e variância ([Zaki e Meira Jr. 2020](#), [Hastie et al. 2009](#)). Esse tipo de algoritmo funciona como uma espécie de “assembleia de classificadores”. Duas das principais técnicas de *ensemble* são: *boosting* (e.g. GB e XGB) e *bootstrap aggregation* (abreviado como *bagging*) (e.g. RF). Na técnica *bagging*, diversos modelos (e.g. árvores de decisão simples) são executados em paralelo, em amostras do tipo *bootstrap* do conjunto de dados de treinamento ([Hastie et al. 2009](#)). Ao final, cada modelo apresenta um voto (sua predição individual), e a classe que recebeu o maior número de votos é apresentada como resultado final. Na técnica *boosting*, diversos modelos “fracos”, de rápida execução, são combinados, de forma que, a cada execução, um modelo informe o seguinte quais classes foram pior identificadas, para então serem priorizadas na execução seguinte ([Kubat 2015](#), [Brown 2017](#)).

O SVM é um método baseado no princípio de discriminantes lineares de maior margem. Para realizar a classificação, o algoritmo busca um hiperplano que maximize a margem entre as classes. Para soluções de problemas não lineares, emprega-se o “truque do kernel” (*kernel trick*), para se encontrar o intervalo de decisão não-linear ótimo entre as classes, que corresponde a um hiperplano em um espaço multidimensional não-linear ([Zaki e Meira Jr. 2020](#)). O “truque do kernel” é uma função de transformação chamada *kernel*, que mapeia as dimensões do espaço do conjunto de dados de entrada para um espaço com mais dimensões, onde as classes possam ser melhor separadas linearmente. Algumas das funções

de *kernel* mais utilizadas são: linear, polinomial, sigmoidal e função de base radial (*radial basis function* - RBF) (Mello e Ponti 2018).

O NB é uma técnica, do tipo probabilística, que realiza a classificação a partir da estimativa da probabilidade de uma amostra pertencer a cada classe. O algoritmo do NB utiliza o teorema de Bayes para determinar qual classe maximiza a probabilidade posterior da amostra. Para tanto, realiza uma estimativa da função de densidade de probabilidade de cada classe modelada, através de uma distribuição normal multivariada. Esse algoritmo tem como característica a propriedade de assumir a independência das variáveis, o que contorna a necessidade de se estimar uma quantidade muito grande de variáveis, no caso de conjuntos de dados com muitas variáveis (Zaki e Meira Jr. 2020).

A RL é uma técnica que calcula a probabilidade posterior das classes através de funções lineares (Hastie *et al.* 2009).

Para se avaliar a performance de cada um dos 6 métodos, foram empregadas 4 métricas de desempenho de classificação: acurácia, precisão, *recall*, F_1 e ROC-AUC (Kubat 2015, Pedregosa *et al.* 2011).

A acurácia (*Acc*) mede a frequência de classificações corretas em relação ao total de classificações realizadas. É calculada como:

$$Acc = \frac{N_{TP} + N_{TN}}{N_{FP} + N_{FN} + N_{TP} + N_{TN}} \quad \text{Eq. 19}$$

onde N representa o número de classificações, TP (*true positive*) quantifica as classificações positivas corretas, TN (*true negative*) as negativas corretas, FP (*false positive*) as falsas positivas, e FN (*false negative*) as falsas negativas.

A precisão (*Pr*), mede a porcentagem de positivos corretamente classificados em relação ao total de classificações positivas:

$$Pr = \frac{N_{TP}}{N_{FP} + N_{TP}} \quad \text{Eq. 20}$$

O *recall* (*Re*) mede a probabilidade de que um ponto da classe positiva tenha sido corretamente classificado:

$$Re = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad \text{Eq. 21}$$

O F_1 é calculado como a média harmônica entre precisão e *recall*:

$$F_1 = \frac{2}{\frac{1}{Re} + \frac{1}{Pr}} = 2 \frac{Pr \cdot Re}{Pr + Re} = \frac{2N_{TP}}{2N_{TP} + N_{FP} + N_{FN}} \quad \text{Eq. 22}$$

A métrica ROC-AUC (abreviação de *área under the receiver operator characteristics curve*) é calculada como a área sob a curva ROC. A curva ROC é construída pela fração de positivos reais para o total de positivos (taxa de positivos reais) vs. fração de falsos positivos para o total de positivos. Ou seja, pode-se dizer que é a sensibilidade (ou *recall*) vs. 1 – especificidade, essa última calculada como:

$$\text{especificidade} = \frac{N_{TN}}{N_{TN} + N_{FP}} \quad \text{Eq. 23}$$

Embora não se possa afirmar, com certeza, que todas as amostras negativas são de fato locais sem ocorrências, uma vez que foram selecionadas de forma semi-aleatória, é esperado que o número de ocorrências seja muito menor que o de não-ocorrências, de tal forma que as métricas de desempenho preditivo, apesar de não serem uma medida perfeita, permanecem úteis para fins de comparação e de avaliação do ajuste dos modelos às características das ocorrências.

Para a avaliação dos modelos, o conjunto de dados (com 638 pontos) foi separado em um conjunto de treino, composto por 75% dos pontos (478), e um conjunto de teste, com os 25% restantes (160). A separação foi realizada de forma estratificada, ou seja, com proporções iguais de pontos mineralizados e não mineralizados nos dois conjuntos. As métricas de desempenho foram calculadas sobre o conjunto de teste, que são os dados não vistos pelos modelos durante o treinamento. Tanto o treinamento dos modelos como a avaliação de desempenho foram realizados com o pacote *scikit-learn* para Python ([Pedregosa et al. 2011](#)), com exceção do treinamento do modelo XB, que foi realizado com o pacote *XGBoost* também para Python ([Chen e Guestrin 2016](#)).

2.6.3. Avaliação do modelo com SHAP

Para se entender melhor o comportamento do modelo em relação aos mapas de evidências, realizou-se análises do modelo com a técnica SHAP (*Shapley additive explanations*). Esse método, que é baseado na Teoria dos Jogos, é fundamentado na técnica proposta por [Shapley \(1953\)](#) para se distribuir, de maneira justa, uma recompensa de um jogo cooperativo entre os membros de um time. Com base nessa técnica, [Lundberg e Lee \(2017\)](#) propuseram um método para explicação de modelos preditivos, o que é particularmente útil para os modelos complexos, como os modelos de assembleia e as redes neurais profundas. No método proposto, pressupõe-se que cada variável do modelo é um membro de um time (ou um "jogador"), e as previsões do modelo são as recompensas do jogo. Dessa forma, os valores de Shapley (valores SHAP) indicam as contribuições das variáveis de entrada para o resultado do modelo em cada amostra. Um valor SHAP é proporcional à contribuição de uma variável para a mudança no valor esperado da predição do modelo (o valor médio de predição do modelo para as amostras de treinamento). Assim, um valor SHAP positivo indica um aumento no valor de saída do modelo, e um valor negativo indica uma diminuição (Figura 2.15). Neste trabalho, a análise SHAP foi realizada com o pacote *shap* para Python.

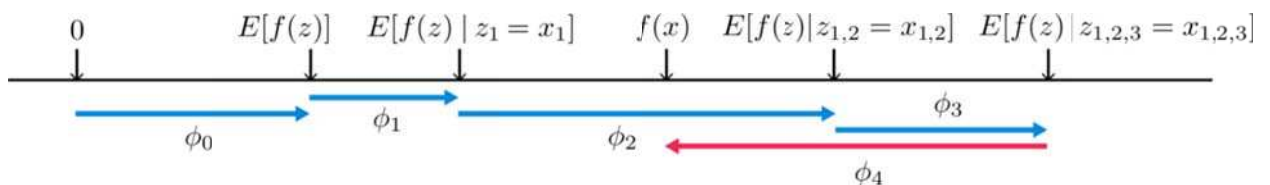


Figura 2.15: Esquema gráfico da influência (Φ) de cada variável (z) no resultado da predição do modelo ($f(x)$) a partir do valor base $E[f(z)]$ (modificado de [Lundberg e Lee, 2017](#)).

3. Resultados

3.1. Análise exploratória e preditiva com SOM

Os mapas SOM foram treinados com grid hexagonal de 28 x 39 (total de 1092 células) (Figura 3.1), função de vizinhança do tipo “bolha”, raio inicial de 26 células, e 300 períodos de treinamento, com o algoritmo *batch*. Não existem regras gerais para a definição desses hiperparâmetros, de tal forma que eles devem ser escolhidos empiricamente, levando-se em conta as características do banco de dados (e.g. número de pontos e de variáveis), e o objetivo do estudo. No entanto, existem alguns indicativos de que o mapa SOM não foi adequadamente treinado, como a presença de células vazias, e uma heterogeneidade muito grande na distribuição dos vetores pelas células do mapa (Kohonen 2001, Kohonen 2013, Vesanto 1997, Vesanto 2002). Outra forma proposta de se avaliar o treinamento do mapa SOM é através da observação da distribuição (em um histograma ou uma curva de estimativa de densidade kernel) das células do mapa, em comparação com a distribuição dos dados originais. Yin e Allinson (1995) propõem que as células de um mapa adequadamente treinado possuem uma distribuição semelhante à de uma amostra retirada do banco de dados original (ou seja, duas amostras de uma mesma população). Nesse sentido, Ponmalai e Kamath (2019) sugerem o uso do teste Kolmogorov-Smirnov (KS) biamostrual para se avaliar quão semelhante são as duas distribuições. Esse teste avalia a hipótese de que duas amostras foram retiradas de uma mesma população (H_0). Portanto, um valor- p menor do que o nível de significância α escolhido (e.g. $\alpha=0,05$) rejeita tal hipótese. Tanto a observação dos mapas de características do modelo SOM obtido (Figura 3.3.A), como as distribuições das variáveis em relação às das células SOM (Figura 3.2), e os testes KS (Tabela 3), indicam que o mapa SOM foi adequadamente treinado.

Neste trabalho, a técnica SOM foi implementada por meio do pacote *kohonen* 3.0.11 para plataforma R (Wehrens e Buydens 2007, Wehrens e Kruisselbrink 2018).

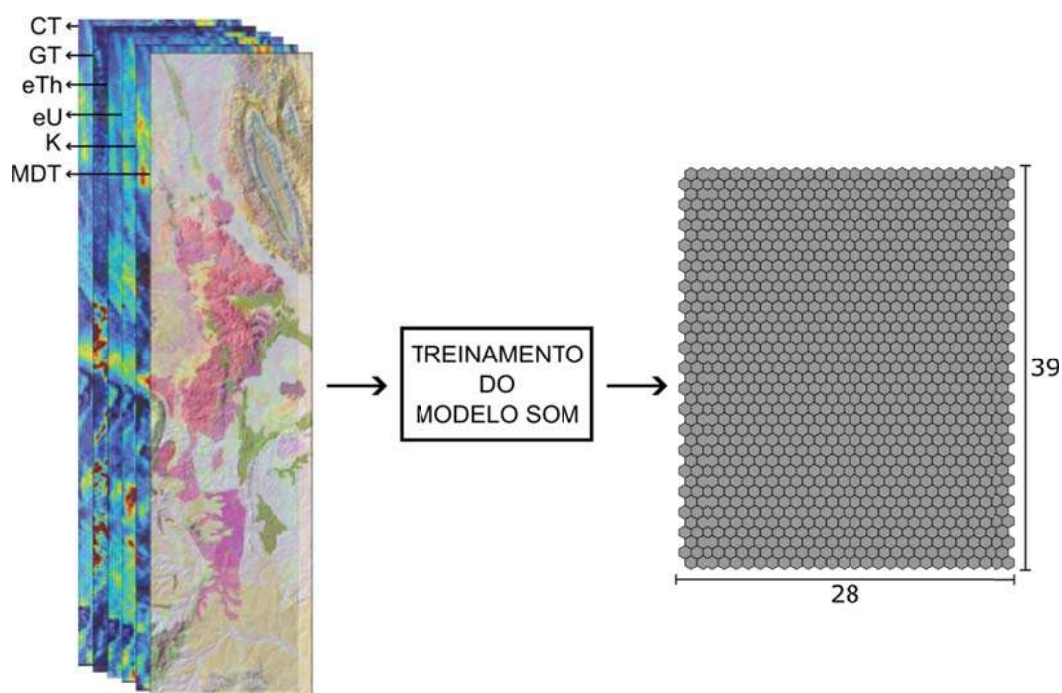


Figura 3.1: Camadas de dados de entrada (contagem total, gradiente total, tório, potássio, urânio e modelo digital de terreno) e formato do mapa SOM treinado. Na figura, somente para fins de visualização, o MDT está fundido com as cores do mapa geológico e com a localização das ocorrências de urânio conhecidas.

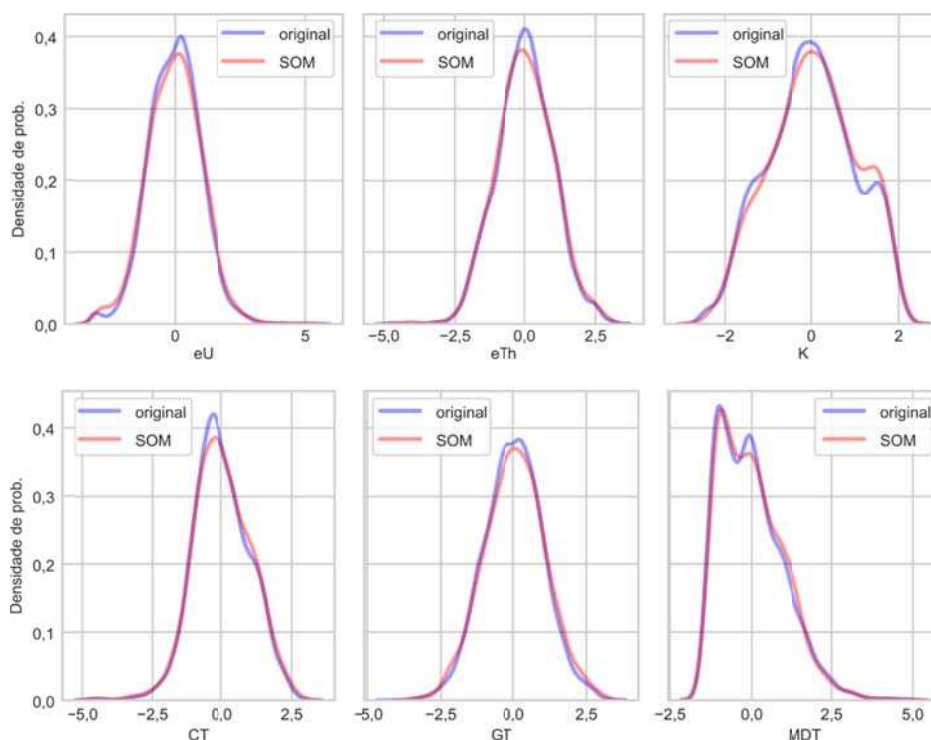


Figura 3.2: Comparação das curvas de distribuição (estimativa de densidade kernel) das variáveis do banco de dados original (transformadas) e das células do mapa SOM.

Tabela 3: Resultados dos testes Kolmogorov-Smirnov (KS) realizados entre as células do mapa SOM e os dados originais.

	KS	Valor-<i>p</i>
eTh	0,019	0,88
eU	0,02	0,84
K	0,032	0,28
GT	0,027	0,47
CT	0,024	0,63
MDT	0,033	0,26

Após o treinamento do modelo, foram produzidos mapas das seguintes propriedades: números de vetores por células; matriz-U; distância média (Figura 3.3.A). Ademais, foram plotados mapas de intensidade de cada uma das dimensões (ou variáveis) mapeadas (Figura 3.3.B).

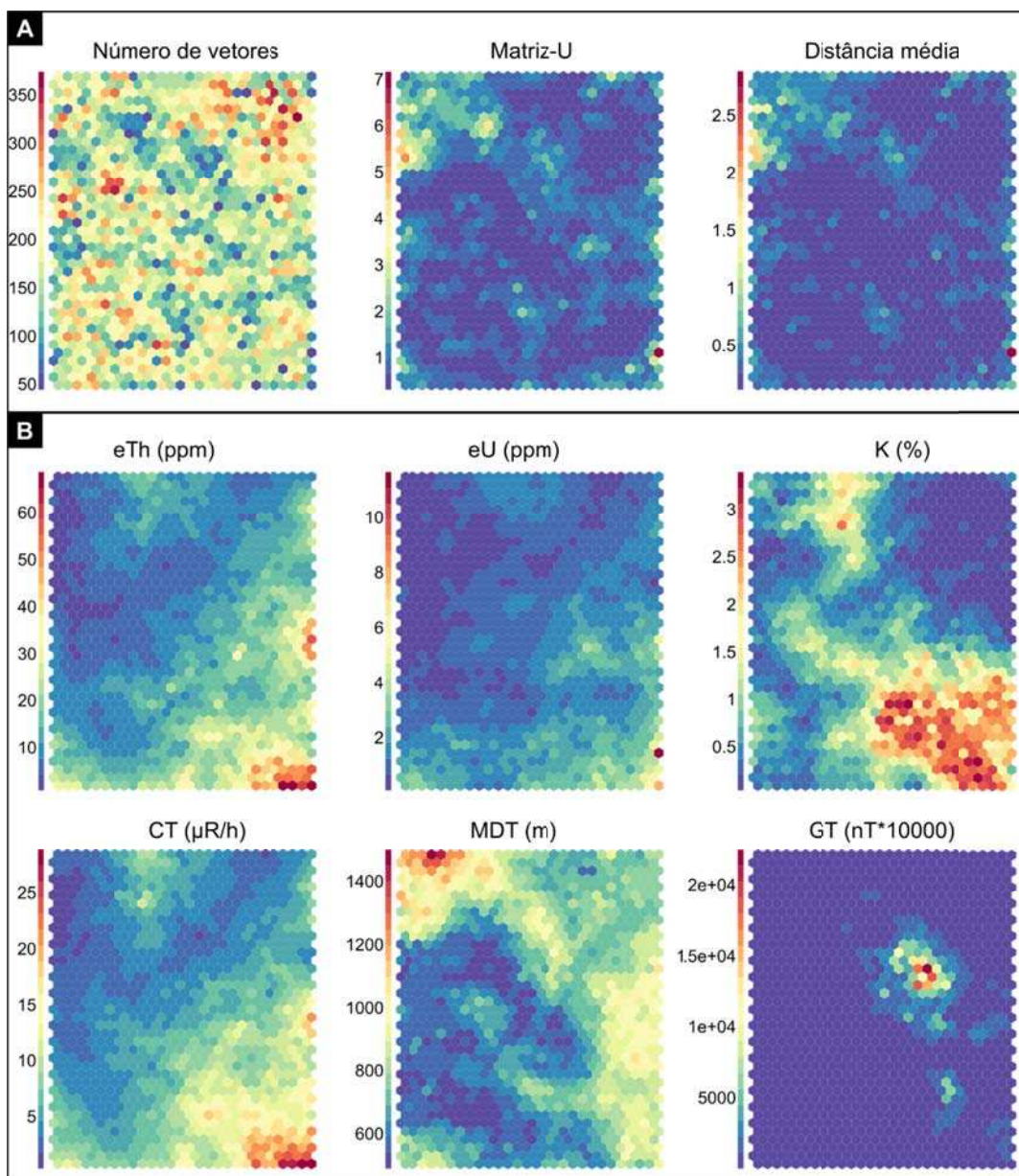


Figura 3.3: A) Mapas das características do modelo SOM: matriz-U, distância média, número de vetores. B) Visualizações dos mapas de ativação dos 6 componentes do modelo SOM: tório, urânio, potássio, contagem total, gradiente total e modelo digital de terreno.

Posteriormente, utilizou-se a técnica CPN para se projetar 6 características geológicas de interesse no mapa SOM bidimensional, obtendo-se um mapa da assinatura de cada uma dessas características. As características projetadas foram: Formação Açuruá, Complexo Gavião (migmatitos), Complexo Licínio de Almeida, Complexo Lagoa Real, Depósitos Detrito-lateríticos e ocorrências de urânio (Figura 3.4).

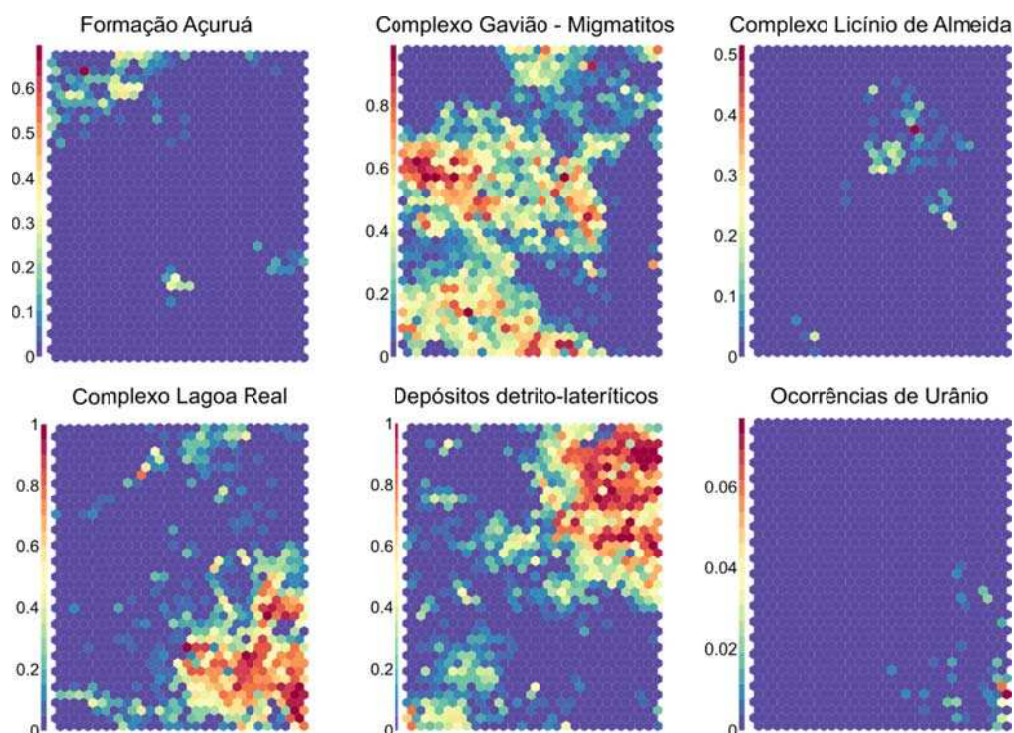


Figura 3.4: Mapas de ativação (assinaturas) de características geológicas mapeadas no modelo SOM.

Para se visualizar a distribuição da assinatura obtida para o Complexo Lagoa Real (Figura 3.5) e para as ocorrências de urânio (Figura 3.6), estas feições foram projetadas do mapa SOM de volta nas dimensões geográficas da área de estudo, onde cada célula do mapa SOM está associada a uma ou mais posições da área, e seu valor é replicado no mapa. Dessa forma, as regiões com valores mais elevados representam locais com assinaturas mais semelhantes àsquelas “aprendidas” pelo modelo SOM.

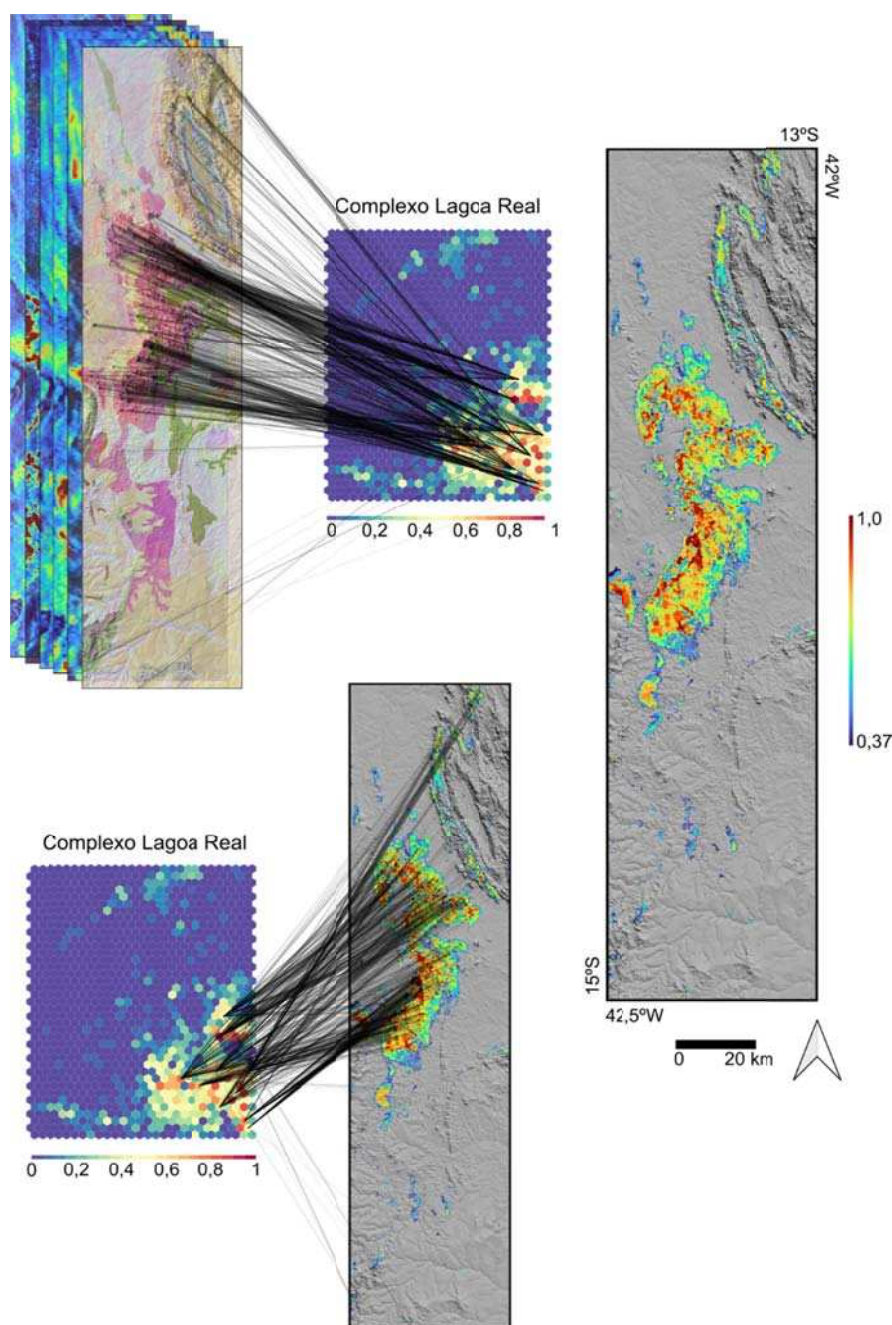


Figura 3.5: Esquema ilustrativo do *Counter Propagation Network* (CPN), aplicado à unidade geológica Complexo Lagoa Real. As setas indicam o mapeamento dos pontos de dados da área de estudo para o modelo SOM, em determinadas células escolhidas como exemplo (imagem superior). O mapeamento reverso projeta a assinatura da unidade modelada no mapa da área (imagem inferior e à direita).

Levando-se em consideração que o algoritmo de treinamento dos mapas SOM possui componentes de aleatoriedade (provenientes da inicialização do mapa e da ordem de atualização das células), realizou-se uma avaliação quantitativa da variabilidade dos resultados obtidos. Para tanto, foram treinados, sequencialmente,

100 mapas SOM, sempre com os mesmos hiperparâmetros. Utilizando-se a técnica CPN, projetou-se na área de estudo a assinatura das ocorrências de urânio conhecidas, obtidas de cada um dos mapas SOM treinados, de forma que cada célula do grid da área possuísse 100 valores de intensidade de assinaturas diferentes. Para se visualizar a variabilidade dos resultados, obtidos a partir de tal conjunto de dados, foram compostos grids dos seguintes parâmetros estatísticos: valores máximos, valores mínimos, médias e coeficientes de variação (CV - calculado pela razão entre desvio padrão/média) (Figura 3.6). Os resultados obtidos foram: máximos entre 0,01 e 0,2; mínimos entre 0,01 e 0,05; CV entre 0,01 e 10; e médias entre 0,001 e 0,08.

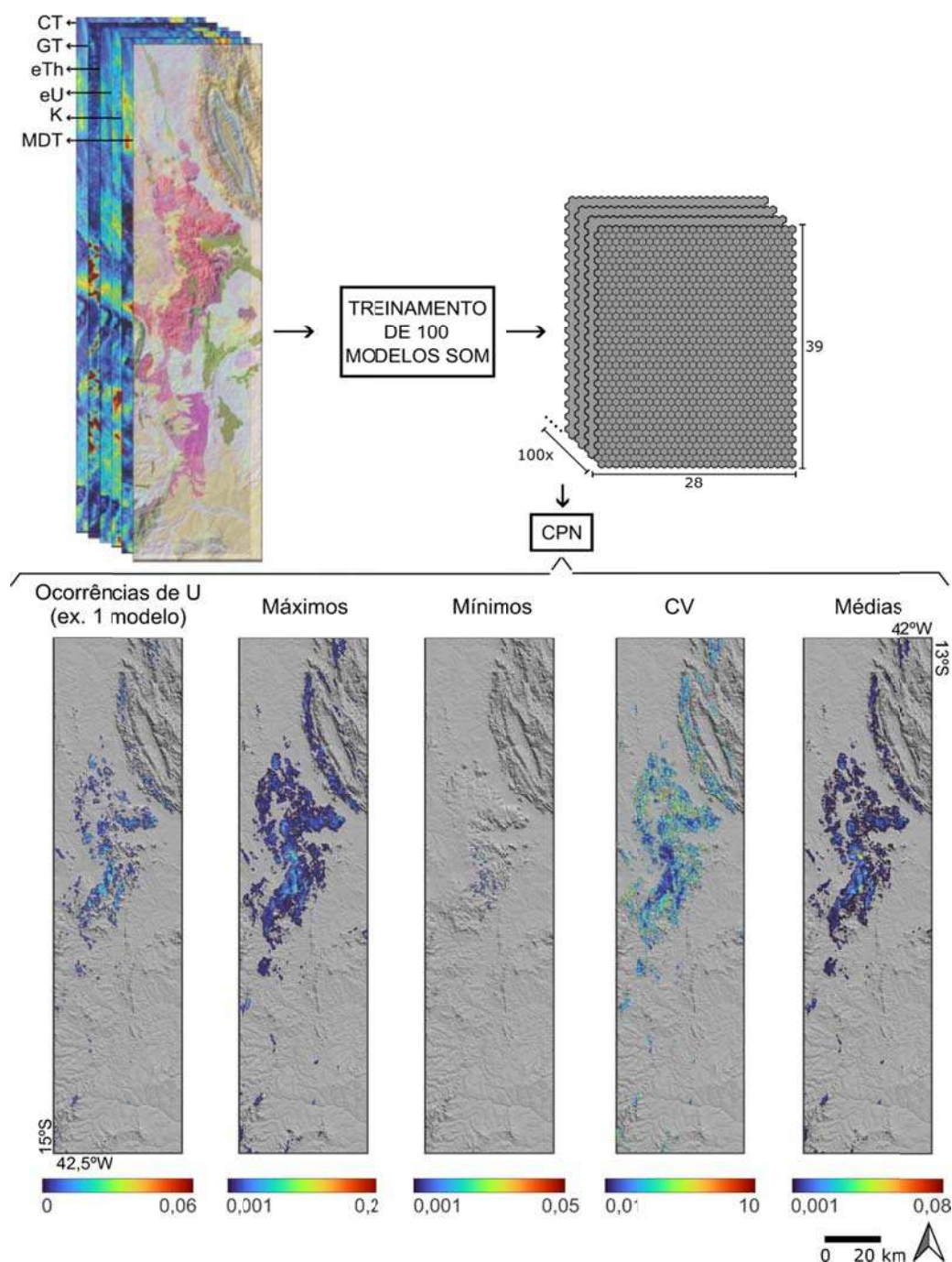


Figura 3.6: Esquema gráfico explicativo da análise estatística da variabilidade dos resultados do modelo SOM, e mapas resultantes, com a técnica CPN. O primeiro mapa (da esquerda) é um exemplo dos 100 mapas gerados. Os demais, são os mapas das estatísticas calculadas para os 100 modelos. CV: coeficiente de variação.

Para uma avaliação pontual, elaborou-se um gráfico de histograma dos resultados em um pixel selecionado (coordenadas X: 791062; Y: 8452839), o qual, por sua vez, é associado a uma célula de cada mapa SOM (Figura 3.7). Neste

ponto, a média observada foi próxima à mediana, respectivamente 0,03 e 0,031, e o desvio padrão foi de 0,015.

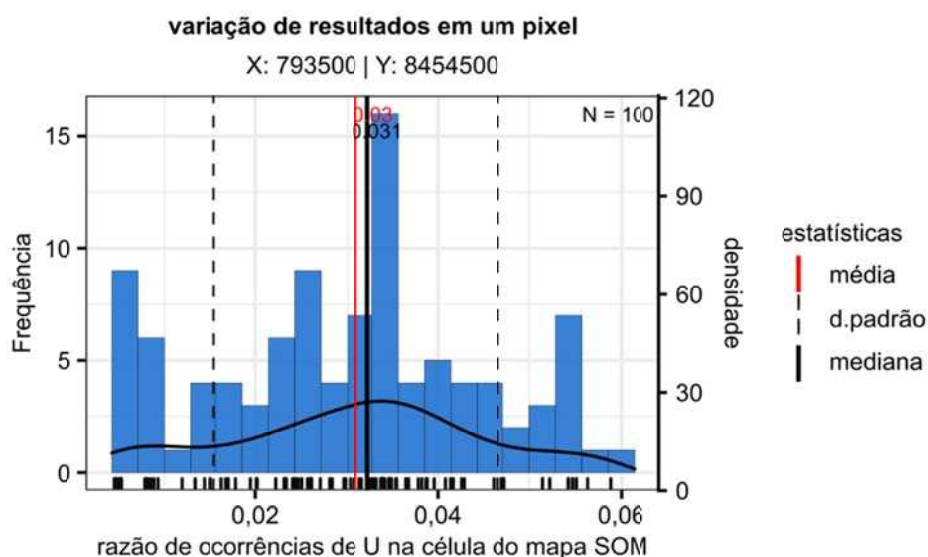


Figura 3.7: Histograma e função de densidade de probabilidade dos 100 resultados de 1 pixel do mapa. São indicados no histograma os seguintes parâmetros estatísticos: média, desvio padrão, e mediana.

Para se observar como o resultado se comporta em relação aos resultados médios, elaborou-se um diagrama de dispersão, combinado com histogramas, das médias vs. coeficientes de variação, excluindo-se os pixels que apresentaram máximos igual a 0, ou seja, aqueles que obtiveram resultados igual a 0 em todos os modelos gerados (Figura 3.8). Os valores de CV variaram entre 0,01 e 10. A partir do 90° das médias, o CV médio foi de 0,6.

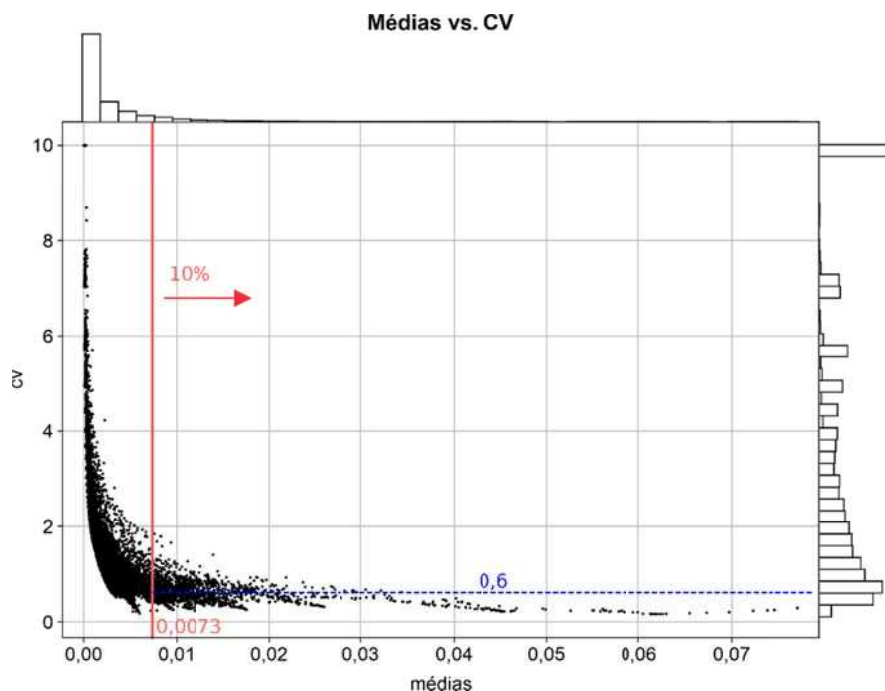


Figura 3.8: Diagrama de dispersão, combinado com histogramas, das médias vs. coeficientes de variação dos pixels que apresentaram máximos maiores do que 0 no conjunto dos 100 modelos SOM. A linha vermelha indica o 90º percentil das médias. A linha azul o valor de cv médio acima do 90º percentil das médias.

3.2. Delimitação de domínios geológico-geofísicos

3.2.1. Agrupamento por SOM + K-means

Para se obter os domínios geológico-geofísicos por aprendizado não-supervisionado, primeiramente, utilizou-se a técnica combinada de SOM + K-means. Para tanto, utilizou-se o mapa SOM treinado previamente (item 3.1), a partir do qual, as 1092 células foram agrupadas com K-means. Para se determinar o melhor número de grupos (k) para o estudo, foram realizados agrupamentos com k variando de 1 a 150 e, para cada um, calculou-se o índice DB (Figura 3.9) e os 3 parâmetros de prospectividade descritos em 2.4.3 (Figura 3.10).

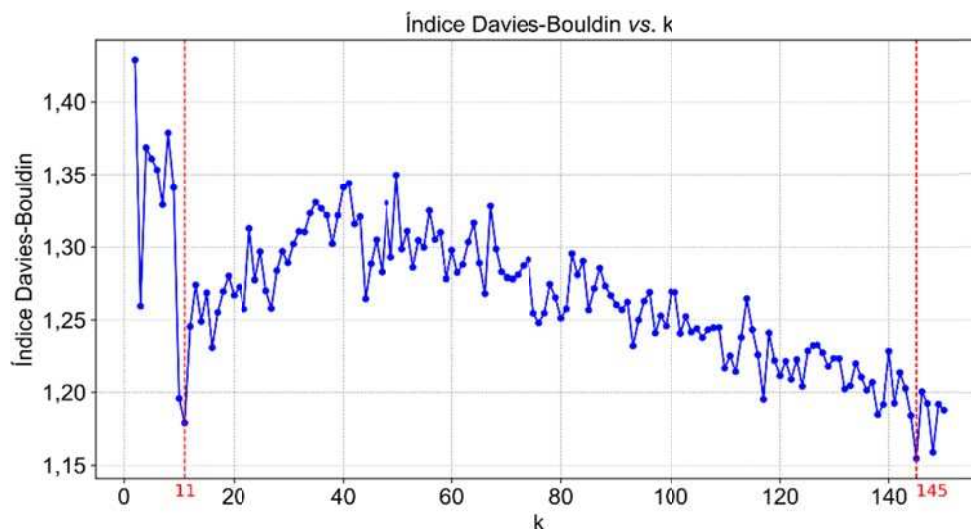


Figura 3.9: Índice DB vs. número de clusters (k) para os agrupamentos obtidos com SOM + K-means. Os melhores (menores) valores de DB foram alcançados com $k = 11$ e $k = 145$.

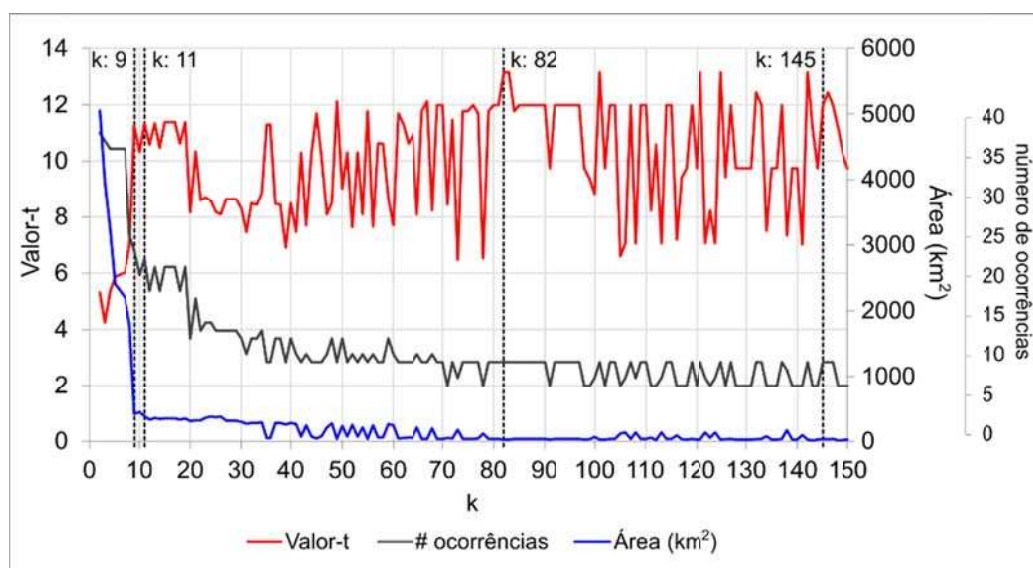


Figura 3.10: Medidas de prospectividade dos grupos com maior número de ocorrências, obtidos de agrupamentos por SOM + K-means, em função de k . São marcados os valores ótimos ($k = 9$ e $k = 82$) obtidos pela observação dessas medidas e, para comparação, os obtidos com o índice DB ($k = 11$ e $k = 145$ - Figura 3.9).

Os menores valores do índice DB foram atingidos com $k = 11$ ($DB = 1,18$) e $k = 145$ ($DB = 1,16$). Já no gráfico de k vs. parâmetros de prospectividade, os melhores valores de k foram $k = 9$ (*ocorrências*: 23; *área*: 422,1 km²; *valor-t*: 11,26), $k = 82$ (*ocorrências*: 9; *área*: 24 km²; *valor-t*: 13,15). Nesse caso, a seleção dos melhores k foi subjetiva, e seguiu a lógica de que, utilizar um número maior de k ,

somente é vantajoso se este for acompanhado de uma diminuição da área do grupo com maior número de ocorrências, de um aumento do valor-t, e ainda preservar um número significativo de ocorrências. Para fins de comparação, foram indicados os k com os menores índices DB: k = 11 (*ocorrências*: 22; *área*: 383,6 km²; *valor-t*: 11,34) e k= 145 (*ocorrências*: 9; *área*: 37,7 km²; *valor-t*: 11,99).

Para se visualizar a distribuição dos grupos com os melhores k no modelo SOM, foram gerados mapas dos agrupamentos com os 4 valores de k (9, 11, 82 e 145) (Figura 3.11).

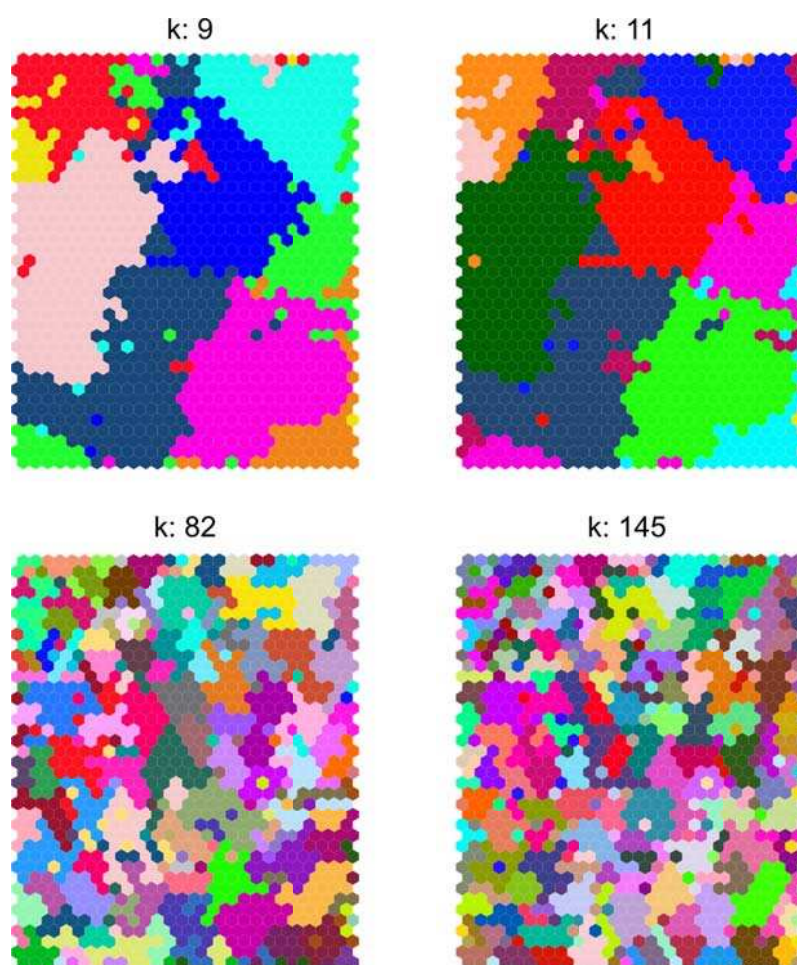


Figura 3.11: Agrupamentos das células do mapa SOM, obtidos com k igual a 9, 11, 82 e 145.

Após a seleção dos melhores valores de k (9, 11, 82 e 145), a partir da observação do índice DB, e dos três parâmetros (número ocorrências, área e valor-t), para k entre 1 e 150, procedeu-se com a espacialização em mapas da área de estudo, dos domínios obtidos pelo agrupamento das células do mapa SOM (Figura 3.12). Como o modelo SOM não possui restrição em relação às coordenadas

espaciais da área, os modelos com k mais elevados (82 e 145) mostraram uma fragmentação espacial excessiva. Portanto, o modelo gerado com $k = 11$ (Figura 3.12.B) foi considerado o mais adequado no balanceamento entre os parâmetros prospectivos e a coerência espacial e geológica dos domínios.

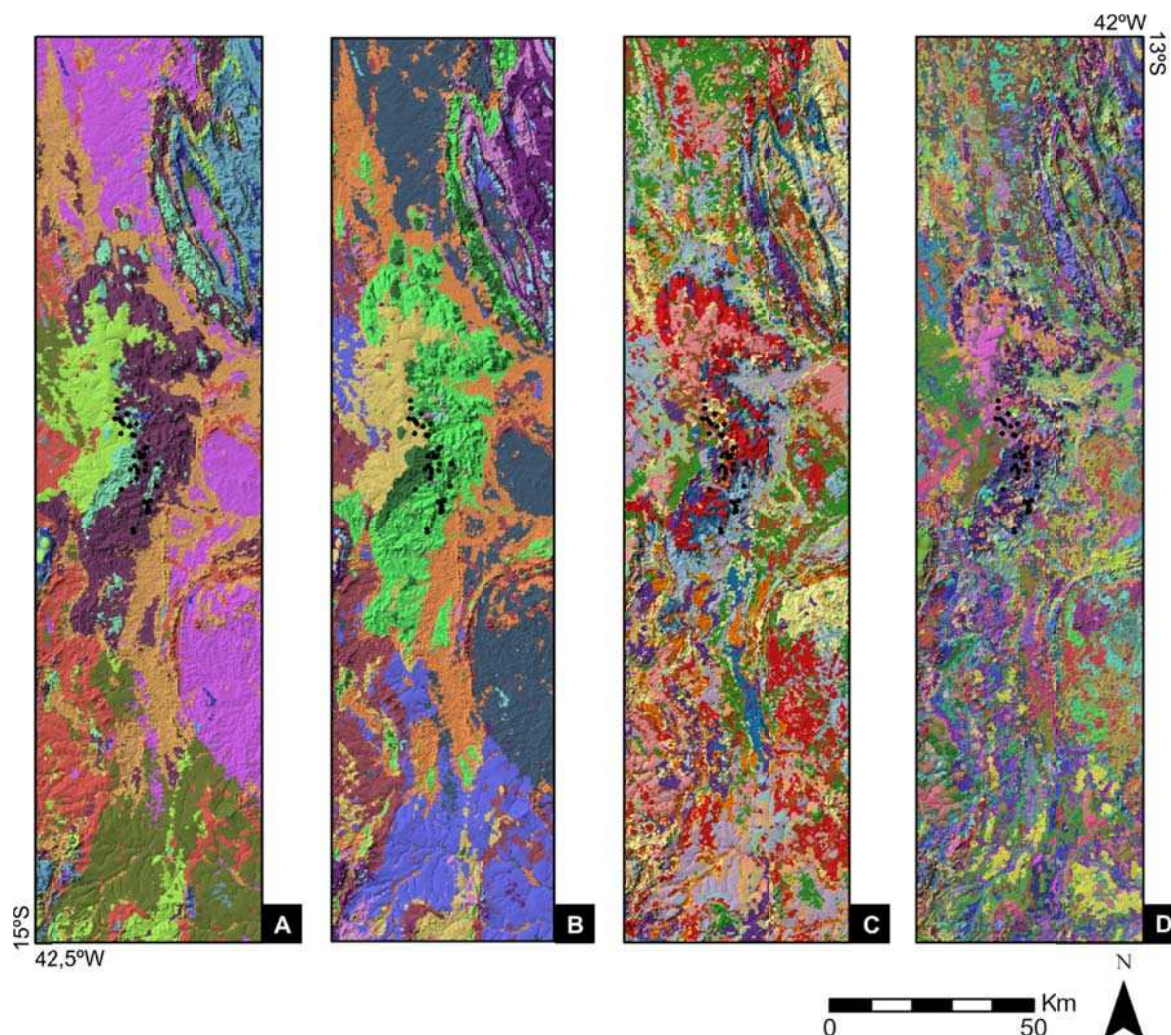


Figura 3.12: Espacialização dos domínios obtidos pelo agrupamento das células SOM, com k igual a: 9 (A), 11 (B), 82 (C), 145 (D).

3.2.2. Agrupamento por AHER

A geração dos domínios geológico-geofísicos com AHER foi realizada com a mesma metodologia de escolha dos melhores valores de k utilizada no tópico 3.2.1. Primeiramente, calculou-se o índice DB (Figura 3.13) e os 3 parâmetros de prospectividade (Figura 3.14) para k variando de 1 a 150.

O menor valor do índice DB foi obtido com $k = 4$ (DB = 1,89). Já no gráfico de k vs. parâmetros de prospectividade, os melhores valores de k foram $k = 7$

(ocorrências: 34; área: 768,3 km²; valor-t: 9,1), k = 35 (ocorrências: 29; área: 680,4 km²; valor-t: 10,4) e k = 81 (ocorrências: 17; área: 108,6 km²; valor-t: 13,7). Nesse caso, a seleção dos melhores k foi subjetiva. Adotou-se a lógica de que, utilizar um número maior de k, somente é vantajoso se este for acompanhado de uma diminuição da área do grupo com maior número de ocorrências, de um aumento do valor-t, e ainda preservar um número significativo de ocorrências. Para fins de comparação, foi também indicado o k com o menor índice DB: k = 4 (ocorrências: 35; área: 4433,3 km²; valor-t: 5).

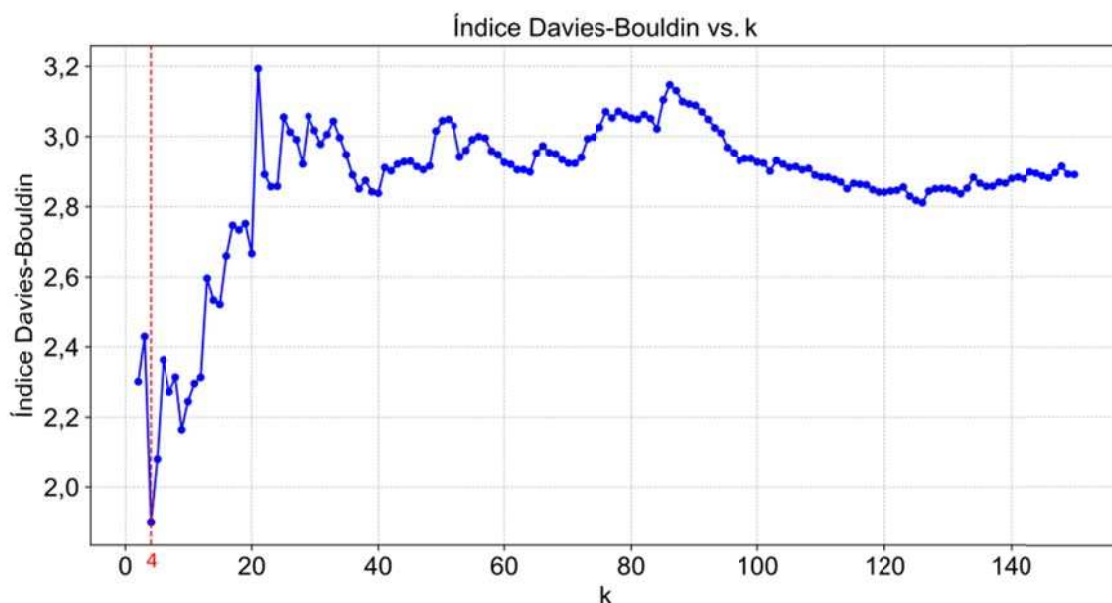


Figura 3.13: Índice DB vs. número de clusters (k) para os agrupamentos obtidos com AHER. O melhor (menor) valor de DB foi alcançado com k = 4.

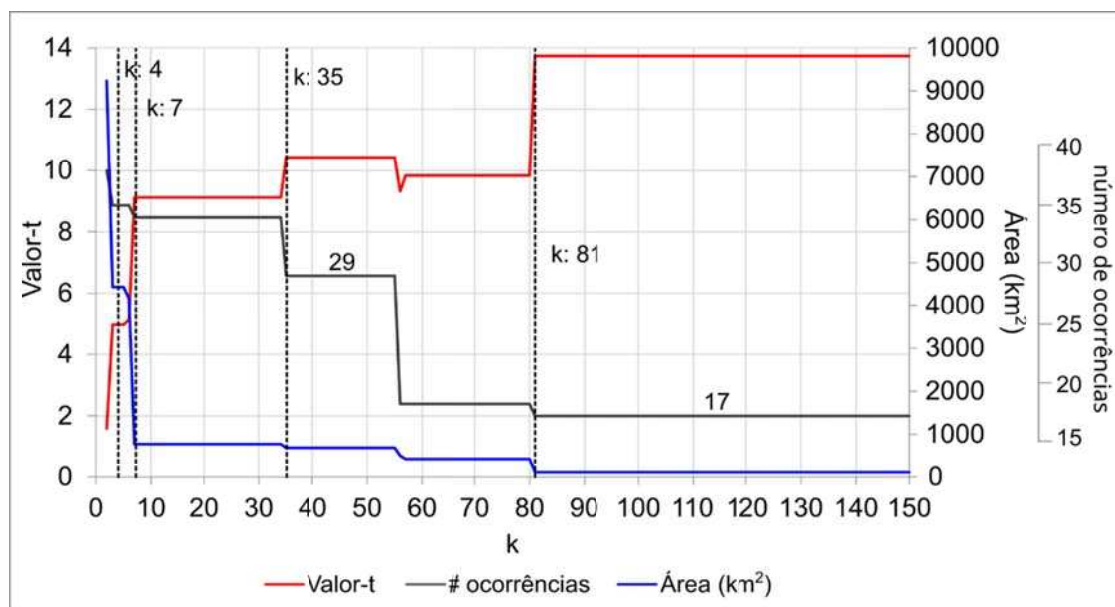


Figura 3.14: Medidas de prospectividade dos grupos com maior número de ocorrências, obtidos de agrupamentos por AHER, em função de k. São marcados os valores ótimos ($k = 7$, $k = 35$ e $k = 81$) obtidos pela observação dessas medidas e, para comparação, o obtido com o índice DB ($k = 4$ - Figura 3.13).

Após a seleção dos melhores valores de k (4, 7, 35 e 81), a partir da observação do índice DB, e dos três parâmetros (número ocorrências, área e valor-t), para k entre 1 e 150, procedeu-se com a espacialização dos domínios na área de estudo (Figura 3.15). A restrição espacial do AHER resultou em domínios espacialmente coerentes nos 4 valores de k. Dessa forma, o modelo obtido com $k = 81$ foi considerado mais adequado, uma vez que apresentou os melhores parâmetros prospectivos, com uma definição de domínios com coerência espacial e geológica.

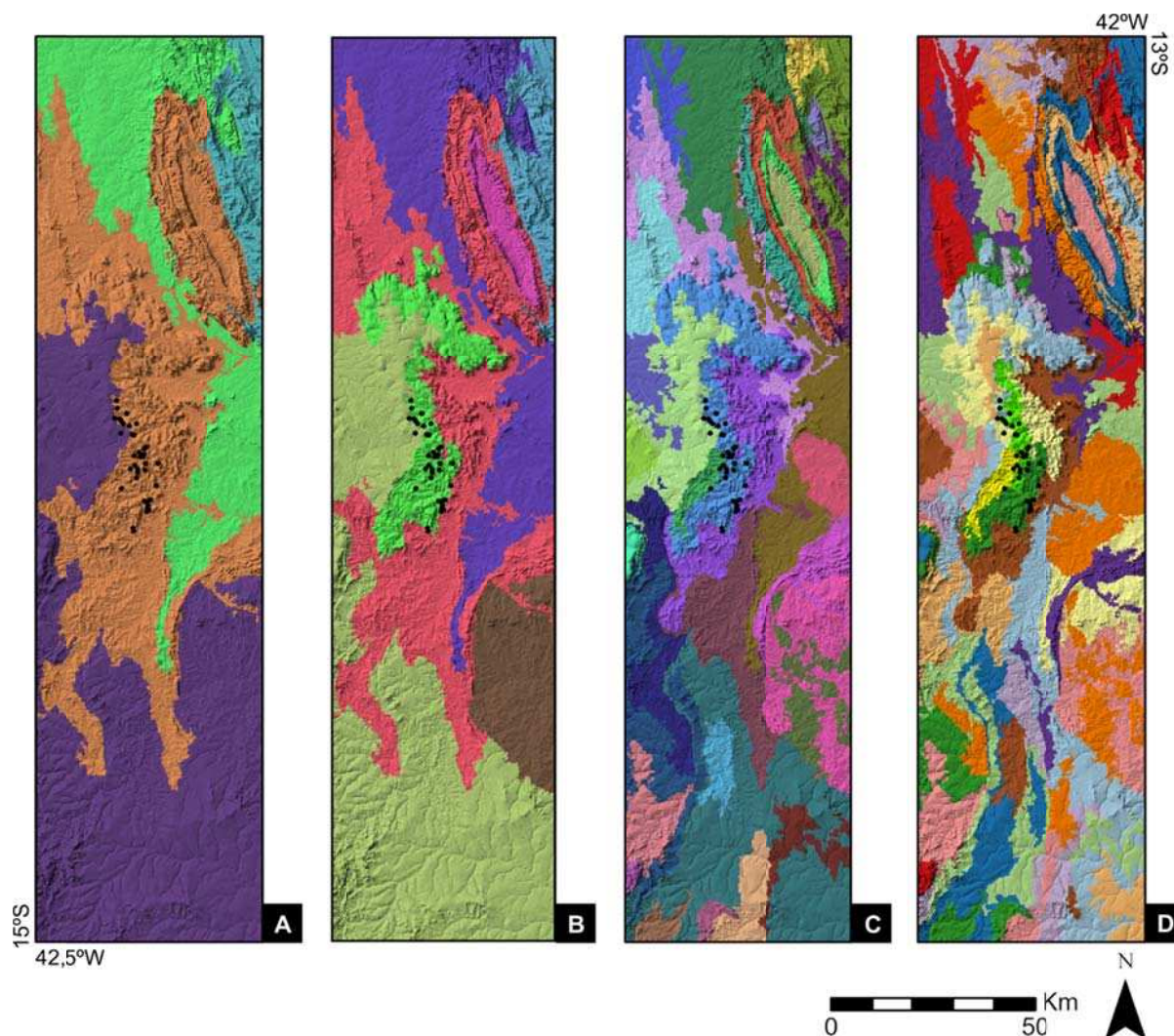


Figura 3.15: Mapas dos domínios obtidos através do AHER, com valores de k : 4 (A), 7 (B), 35 (C) e 81 (D).

3.2.3. Comparação dos resultados

Para se avaliar a potencial vantagem prospectiva do uso dos domínios geológico-geofísicos, em relação às unidades geológicas, os domínios obtidos com ambos métodos foram comparados frente aos 3 parâmetros otimizados na seleção do k , além da densidade de ocorrências (Tabela 4 e Tabela 5, Figura 3.16 e Figura 3.17).

No mapa geológico, todas as ocorrências estão localizadas dentro de duas unidades que fazem parte do Complexo Lagoa Real, sendo que 97,4% (37 ocorrências) encontram-se na unidade “Lagoa Real – ortognaisses” (PP4yl1), e somente 2,6% (1 ocorrência) na unidade São Timóteo. A primeira compreende a maior parte do Complexo Lagoa Real no mapa, delimitando uma área de 1656,56

km², com um valor-t de 5,37. A segunda, representa lentes delimitadas dentro dos ortognaisses Lagoa Real, compreendendo uma área de 154,56 km², com um valor-t de 0,71.

Dos domínios geológico-geofísicos obtidos com SOM + K-means, somente 1 (grupo 5) apresentou valor-t (11,34) mais elevado que a unidade PP4yl1, porém, com um ganho significativo (2,1x maior) em relação à unidade geológica, contendo 57,9% das ocorrências (22) dentro de uma área com 383,56 km². No entanto, os grupos obtidos por esse método são espacialmente fragmentados e ocorrem mais ou menos dispersos por toda a área de trabalho, o que não ocorreu com o AHER.

Tabela 4: Parâmetros dos grupos definidos por SOM + K-means e das unidades geológicas que apresentaram pelo menos uma ocorrência de urânio. W+, W-, contraste e valor-t calculados pelo método “peso de evidências”. Quanto melhor o valor de cada coluna, em relação à prospectividade, mais intensa a cor do preenchimento em vermelho.

Grupo	Area(km ²)	# Ocorr.	% Ocorr.	Ocorr./km ²	W+	W-	Contraste	Valor-t
5	383,56	22	57,9%	0,057	2,90	-0,83	3,73	11,34
1	1780,88	12	31,6%	0,007	0,75	-0,22	0,97	2,78
4	1041,88	3	7,9%	0,003	-0,10	0,01	-0,11	-0,18
6	3,13	1	2,6%	0,320	4,63	-0,03	4,66	4,55
PP4yl1	1656,56	37	97,4%	0,022	1,95	-3,49	5,44	5,37
PP4ylt	154,56	1	2,6%	0,006	0,71	-0,01	0,72	0,71

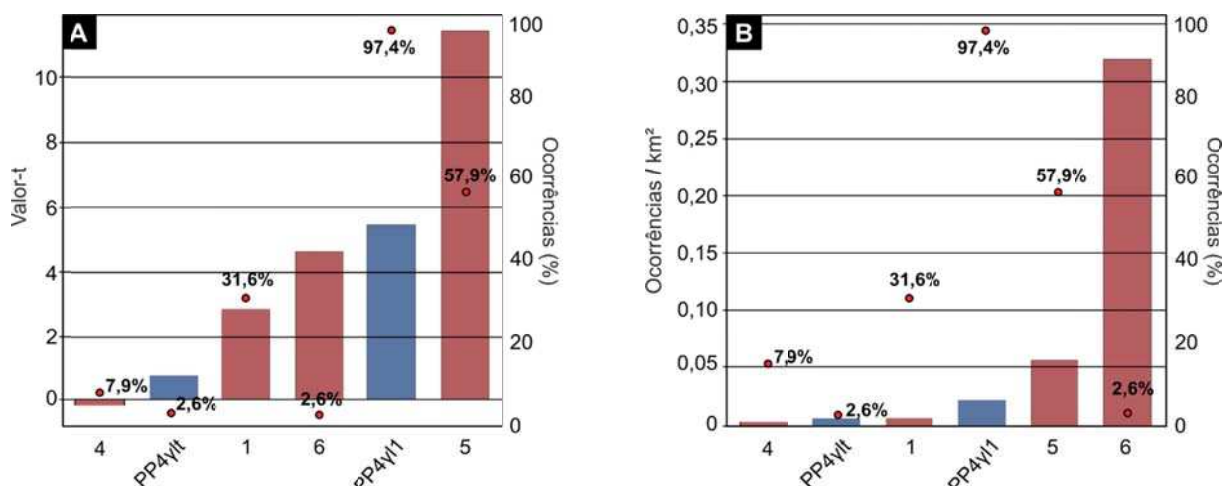


Figura 3.16: Comparação gráfica das unidades do Complexo Lagoa Real (barras azuis) com os grupos obtidos por SOM+ K-means, com $k = 11$ (barras vermelhas). As barras quantificam o valor-t (A) e a densidade de ocorrências (B) e, os pontos, a porcentagem das ocorrências compreendidas em cada grupo/unidade.

Nos domínios definidos com AHER, o melhor (grupo 39) contém 44,7% (17) das ocorrências, em uma área com 108,25 km², com um valor-t de 13,74, o maior valor obtido dentre os agrupamentos e as unidades geológicas. Além desse, outros 2 domínios apresentaram valor-t maior que o da unidade PP4yl1: grupo 27, com 28,9% (11) das ocorrências em uma área de 182 km², e valor-t de 9,14; grupo 34, com 13,2% (5) das ocorrências, em 87,81 km², com valor-t de 6,29. Ou seja, ao se considerar os 3 melhores grupos, 87% (33) das ocorrências se localizam em domínios com valor-t maiores que o da unidade PP4yl1, em uma área expressivamente menor, somando 377,35 km² (Figura 3.18 - A e B). Ademais, os domínios definidos pelo AHER são espacialmente contínuos.

Tabela 5: Parâmetros dos grupos definidos pelo AHER, com $k = 81$, e das unidades geológicas, que apresentaram pelo menos uma ocorrência de urânio. W+, W-, contraste e valor-t calculados pelo método “peso de evidências”. Quanto melhor o valor de cada coluna, em relação à prospectividade, mais intensa a cor do preenchimento em vermelho.

Grupo	Area(km ²)	# Ocorr.	% Ocorr.	Ocorr./km ²	W+	W-	Contraste	Valor-t
39	108,25	17	44,7%	0,157	3,91	-0,58	4,49	13,74
27	182	11	28,9%	0,060	2,95	-0,33	3,28	9,14
34	87,81	5	13,2%	0,057	2,89	-0,13	3,02	6,29
22	254	3	7,9%	0,012	1,31	-0,06	1,37	2,28
48	316,88	1	2,6%	0,003	-0,01	0	-0,01	-0,01
11	480,38	1	2,6%	0,002	-0,42	0,01	-0,44	-0,43
PP4yI1	1656,56	37	97,4%	0,022	1,95	-3,49	5,44	5,37
PP4yIt	154,56	1	2,6%	0,006	0,71	-0,01	0,72	0,71

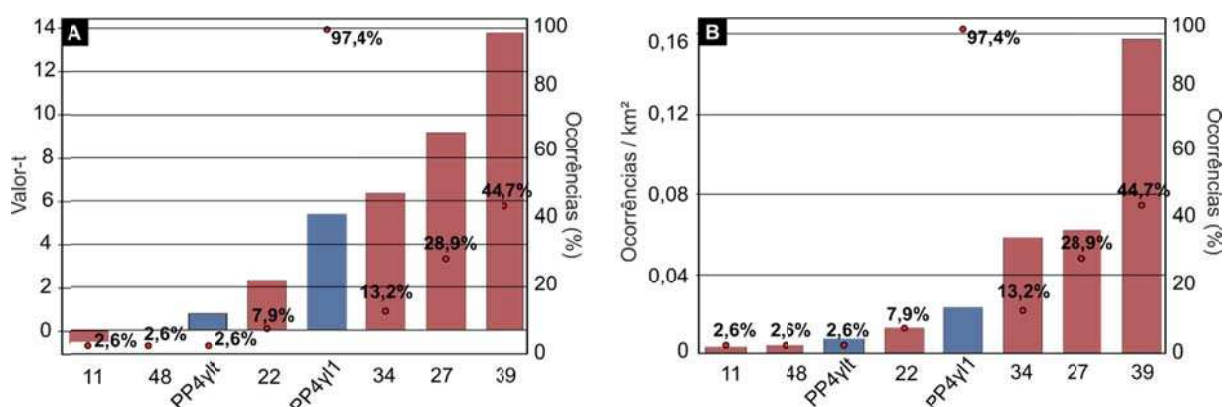


Figura 3.17: Comparação gráfica das unidades do Complexo Lagoa Real (barras azuis) com os grupos obtidos pelo AHER (barras vermelhas). As barras quantificam o valor-t (A) e a densidade de ocorrências (B) e, os pontos, a porcentagem das ocorrências compreendidas em cada grupo/unidade.

Como possíveis formas de classificar os domínios definidos pelo AHER, em relação aos seus potenciais prospectivos, foram gerados 2 mapas dos 81 domínios, reclassificados da seguinte forma: um mapa onde os grupos são classificados em escala de cor de acordo com seu valor-t (Figura 3.18.D); e um mapa similarmente classificado, porém, cuja escala de cor quantifica a dissimilaridade (medida por distância euclidiana) de cada grupo com aquele que

incluiu o maior número de ocorrências, e o maior valor-t (Figura 3.18.C). O último foi posteriormente utilizado como base para a elaboração de um mapa evidencial do modelo de prospectividade (item 3.5).

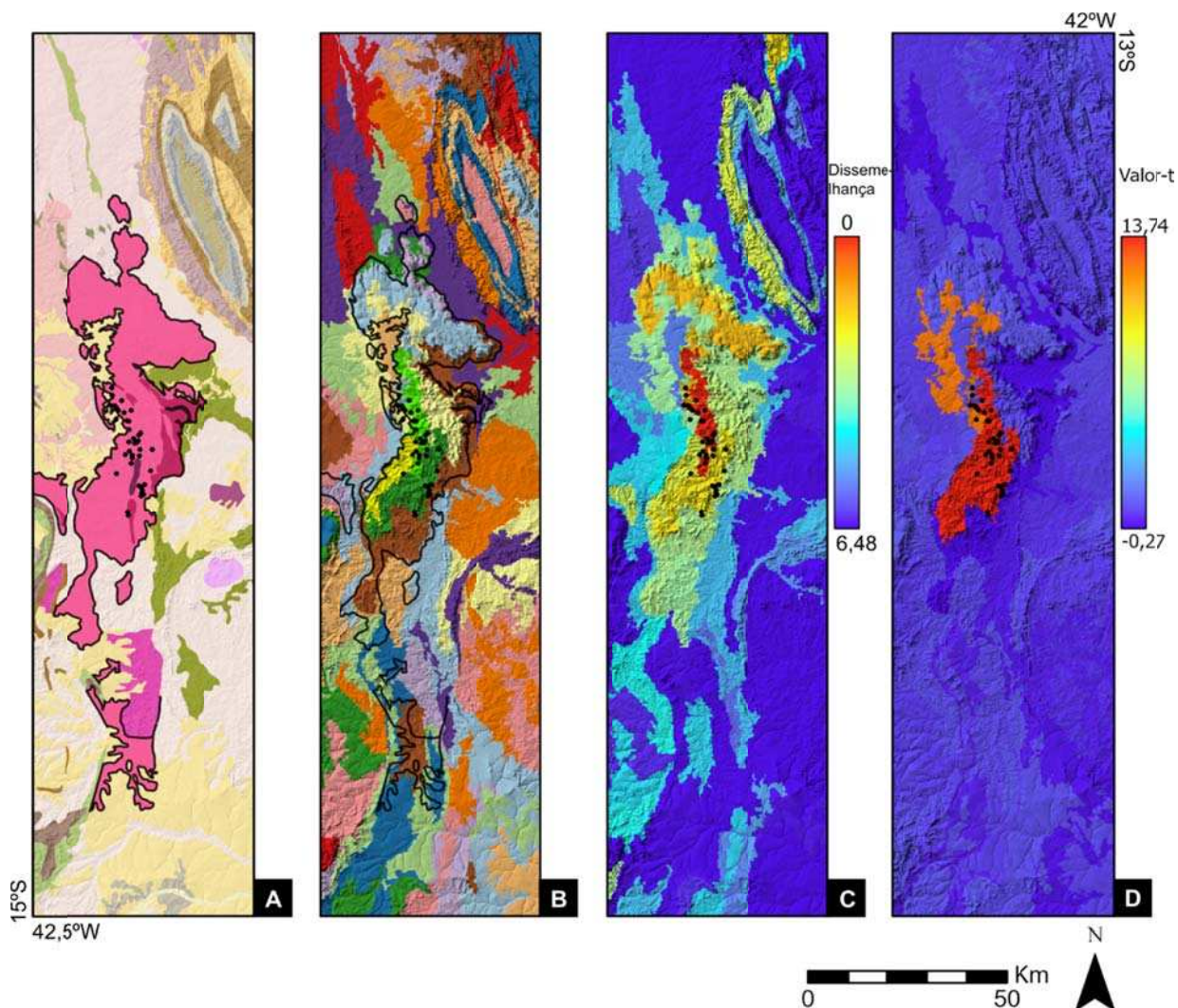


Figura 3.18: A) Mapa com as unidades geológicas do Complexo Lagoa Real em destaque, representando a unidade considerada prospectiva (PP4yl1). B) Resultado do agrupamento obtido com 81 grupos e o limite do Complexo Lago Real. C) Mapa com os grupos classificados em escala de cor de acordo com sua dissimilaridade (medida em distância euclidiana) com o grupo com o maior número de ocorrências. D) Escala de cor representando o valor-t de cada grupo. O valor-t mede a associação espacial entre as ocorrências e os grupos (conf. item 2.4.3).

3.3. Determinação de famílias de lineamentos prospectivos através de análise de fractais

O ajuste das funções controladoras não-lineares permitiu a estimativa do índice de singularidade (α) e do coeficiente de determinação (R^2) associado ao ajuste da reta para 8 das 12 famílias de lineamentos, sendo que para 4 não foi possível se obter um ajuste adequado (Tabela 6). Para as famílias NP_Dn+1, APP_Dn_(NP_Dn) e NP_DR foram obtidos valores de alfa menores do que 2, o que indica relação espacial positiva entre as ocorrências de U e essas famílias de lineamentos (Figura 3.20). Para as famílias DRn+1 e Diques 4, foram obtidos valores pouco abaixo e acima de 2 (aproximadamente 1,99 e 2,05 respectivamente), porém com R^2 quase nulo, mostrando que os ajustes de reta obtidos não foram satisfatórios. Para as famílias APP_Dn+1, APP_Dn+1_(NP_Dn) e APP_Dn, os valores de α obtidos foram maiores do que 2, indicando uma possível relação inversa das ocorrências com os lineamentos dessas famílias. Para as demais famílias, não foi possível se obter um ajuste de reta, indicando total falta de relação com as ocorrências. A Figura 3.19 mostra os gráficos log-log das funções que puderam ser ajustadas.

Tabela 6: Valores do índice de singularidade (α) e do coeficiente de determinação R^2 calculados pelo ajuste da função não-linear, para cada família de lineamentos.

Família	α	R^2
NP_Dn+1	1,55	0,86
APP_Dn_(NP_Dn)	1,60	0,92
NP_DR	1,74	0,67
DRn+1	1,99	0,00
Diques 4	2,05	0,01
APP_Dn+1	2,31	0,62
APP_Dn+1_(NP_Dn)	2,70	0,85
APP_Dn	2,97	0,82
NP_Dn	-	-
NP_Dn-1	-	-
APP_Dn+2	-	-
Diques 1	-	-

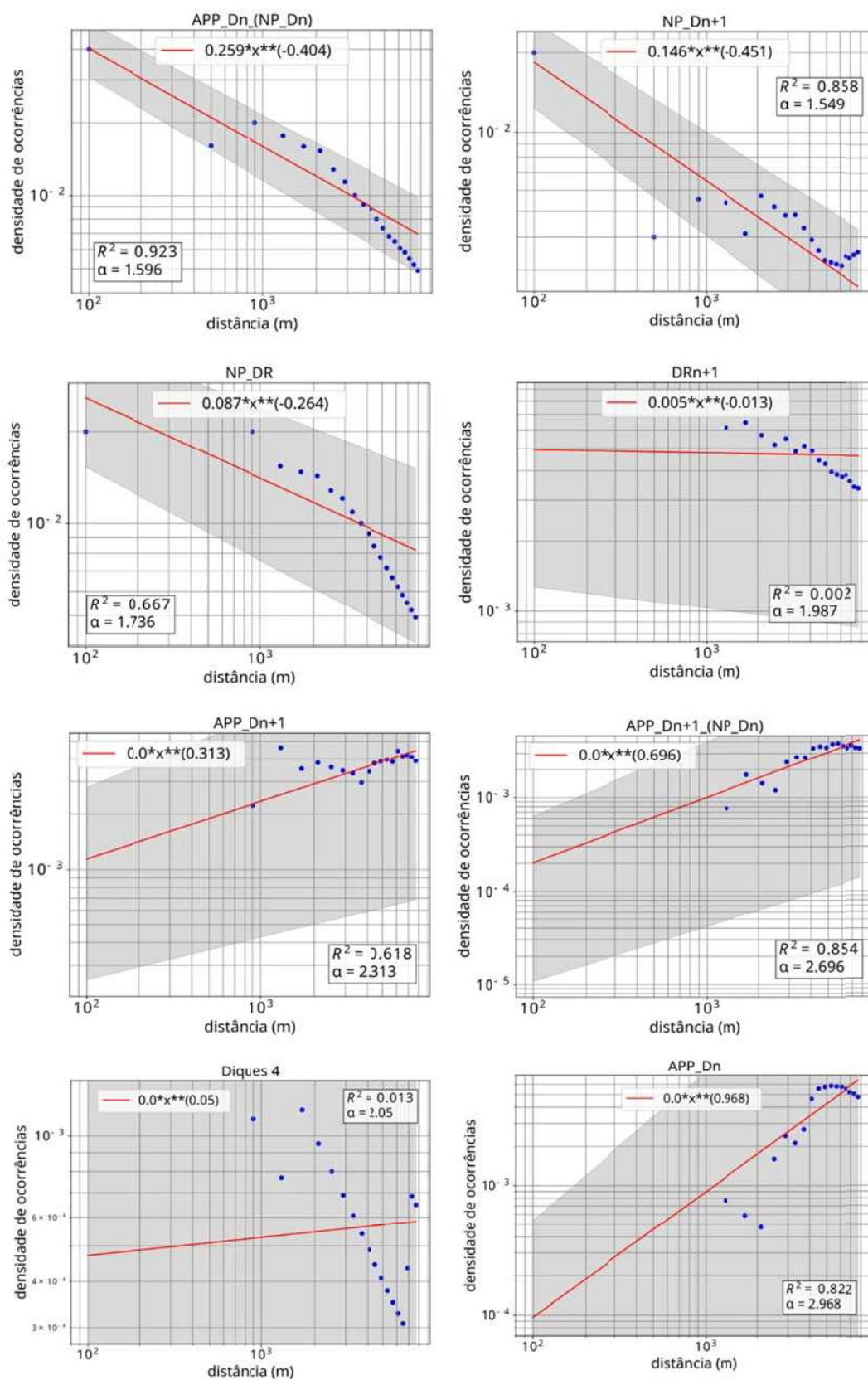


Figura 3.19: Gráficos log-log da distância vs. densidade de ocorrências, ajustados para o cálculo do índice de singularidade α . As regiões em cinza representam o intervalo de confiança do ajuste da reta, para um nível de significância de 0,95.

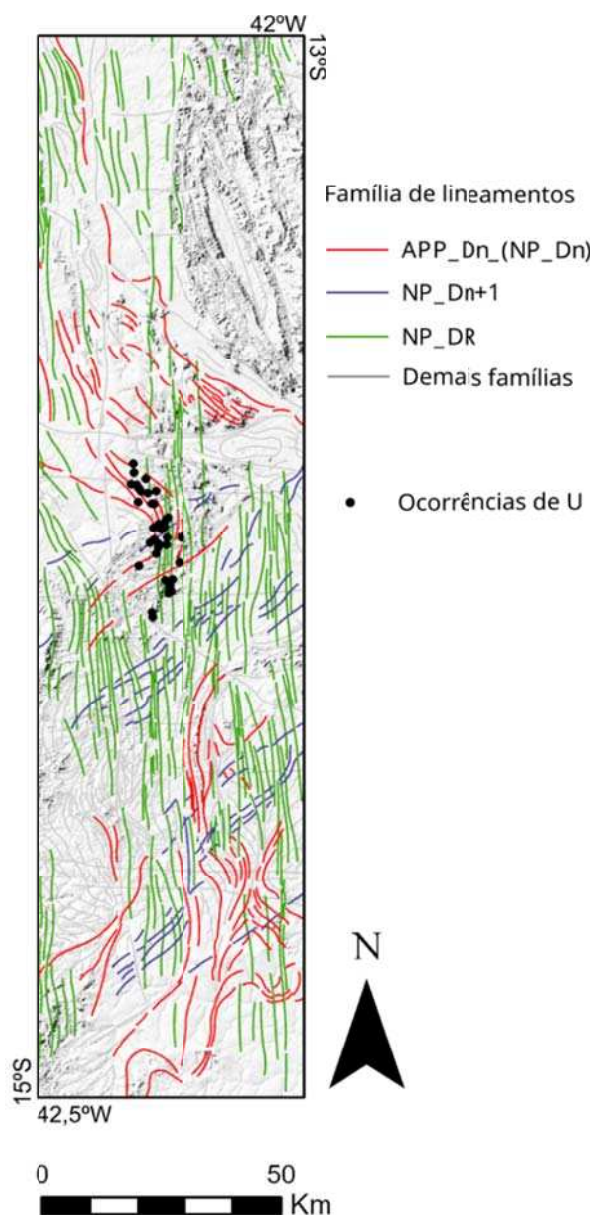


Figura 3.20: Mapa destacando em cores apenas as famílias identificadas como correlacionadas com as ocorrências ($\alpha < 2,0$).

3.4. Interpretação de estruturas regionais profundas

Através da interpretação visual dos produtos derivados de dados magnetométricos, foi possível traçar estruturas lineares com comprimentos que apresentam resposta magnética a até aproximadamente 15 km de profundidade (Figura 3.3).

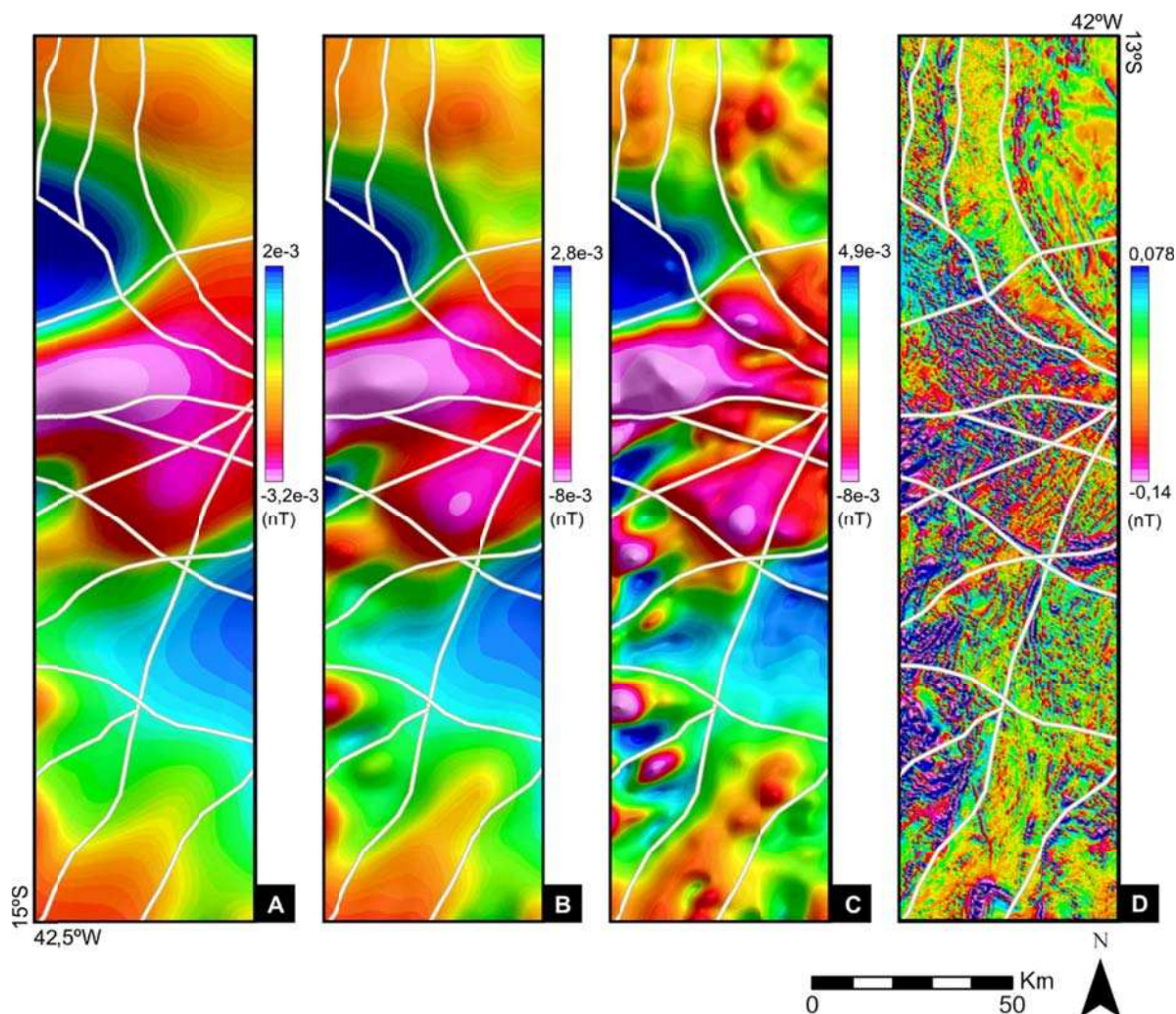


Figura 3.21: Estruturas regionais com expressão em diferentes profundidades, interpretadas a partir da primeira derivada vertical (1Dz) da continuação ascendente do CMA para as altitudes de: A) 15 km; B) 10 km; C) 5 km; e D) da 1Dz do sinal magnetométrico original.

Estruturas com direções variando entre N, NW e NE, com comprimentos na ordem de dezenas de quilômetros (até 60 km), predominam na parte N e S da área. São marcadas por faixas com menor sinal magnético e/ou quebras no padrão do relevo magnético. Estas estruturas podem estar associadas aos esforços distensionais que ocorreram durante a instalação do aulacógeno do Paramirim, cuja evolução se iniciou no Paleoproterozoico. Na parte central da área foram traçadas estruturas com direções variando entre E e NE. Estas estruturas estão associadas com os limites de uma zona com resposta magnética mais elevada, aproximadamente na região de ocorrência do Complexo Lagoa Real. Possivelmente, esta zona está relacionada com as raízes magmáticas do Complexo Lagoa real, que

teve sua colocação favorecida pelas estruturas geradas durante a formação do aulacógeno.

3.5. Mapas evidenciais

Para a composição do modelo de prospectividade para urânio foram elaborados 12 mapas evidenciais, que representam características favoráveis para a ocorrência de depósitos de urânio, seguindo o modelo conceitual do sistema mineral da área (Tabela 1).

Para compor o componente do sistema mineral que mapeia a arquitetura litosférica favorável, foram gerados 6 mapas evidenciais, sendo 4 derivados dos lineamentos das famílias com interesse prospectivo, definidas através da análise por índice de singularidade fractal (3.3), um derivado de lineamentos profundos (3.4), e um derivado do cruzamento de ambos.

Os mapas de distâncias dos lineamentos de interesse prospectivo (Figura 3.23 – A a C) e dos lineamentos profundos (Figura 3.23 - D) foram elaborados com a ferramenta “Distance Accumulation”, do software ArcGis Pro, que calcula a distância linear a partir do centro das linhas. O mapa de densidade de cruzamento de lineamentos prospectivos (Figura 3.23 - E) foi elaborado com a ferramenta “Kernel Density”, do software ArcGis Pro, a partir dos pontos de intersecção entre as linhas das famílias de lineamentos prospectivos. O mapa de densidade de intersecções entre os lineamentos profundos e os prospectivos (Figura 3.23 - F) foi produzido da mesma forma, no entanto, utilizando-se os pontos de intersecção com os lineamentos profundos.

O componente do sistema mineral que mapeia o momento de geodinâmica favorável foi composto por 6 mapas evidenciais, sendo 4 derivados diretamente dos canais da aerogamaespectrometria (potássio, tório, urânio e contagem total - Figura 3.23 – G a J), 1 derivado da aeromagnetometria (gradiente total do campo magnético anômalo - Figura 3.23 – K), e 1 derivado dos domínios geológicos-geofísicos prospectivos obtidos por aprendizado não-supervisionado (Figura 3.23 – L).

Os mapas derivados de dados de gamaespectrometria e magnetometria aéreos foram utilizados no modelo diretamente com os valores dos bancos de dados de origem. O mapa de zonas prospectivas foi elaborado a partir do mapa de

semelhança, calculado do agrupamento otimizado, obtido pela técnica AHER (item 3.2.3). Uma reclassificação binária foi realizada, definindo-se um limite no valor do 3º quartil (2,03) dos valores de dissemelhança, de forma que os grupos com valores inferiores a 2,03 foram considerados prospectivos (classe 1) e os grupos com valores maiores do que 2,03 foram considerados não-prospectivos (classe 0) (Figura 3.22).

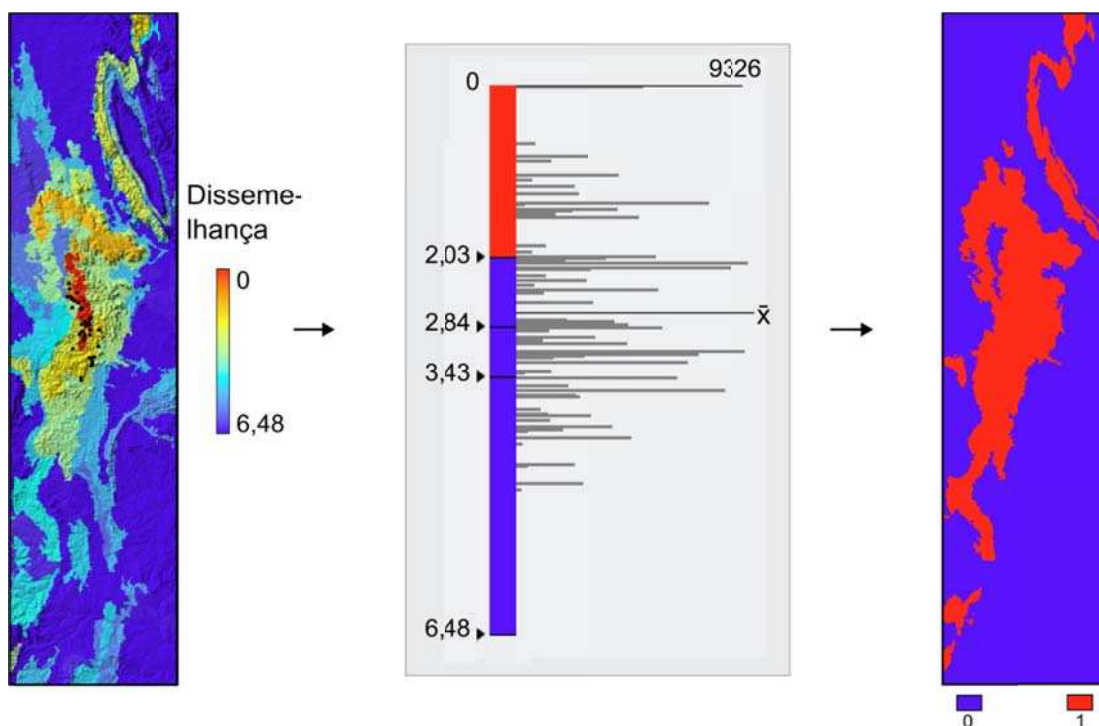


Figura 3.22: Reclassificação binária do mapa de dissemelhança dos grupos mais prospectivos, separando como classe 1 os grupos localizados no 4º quartil, e, como classe 0, os grupos no intervalo entre o 1º e o 3º quartil.

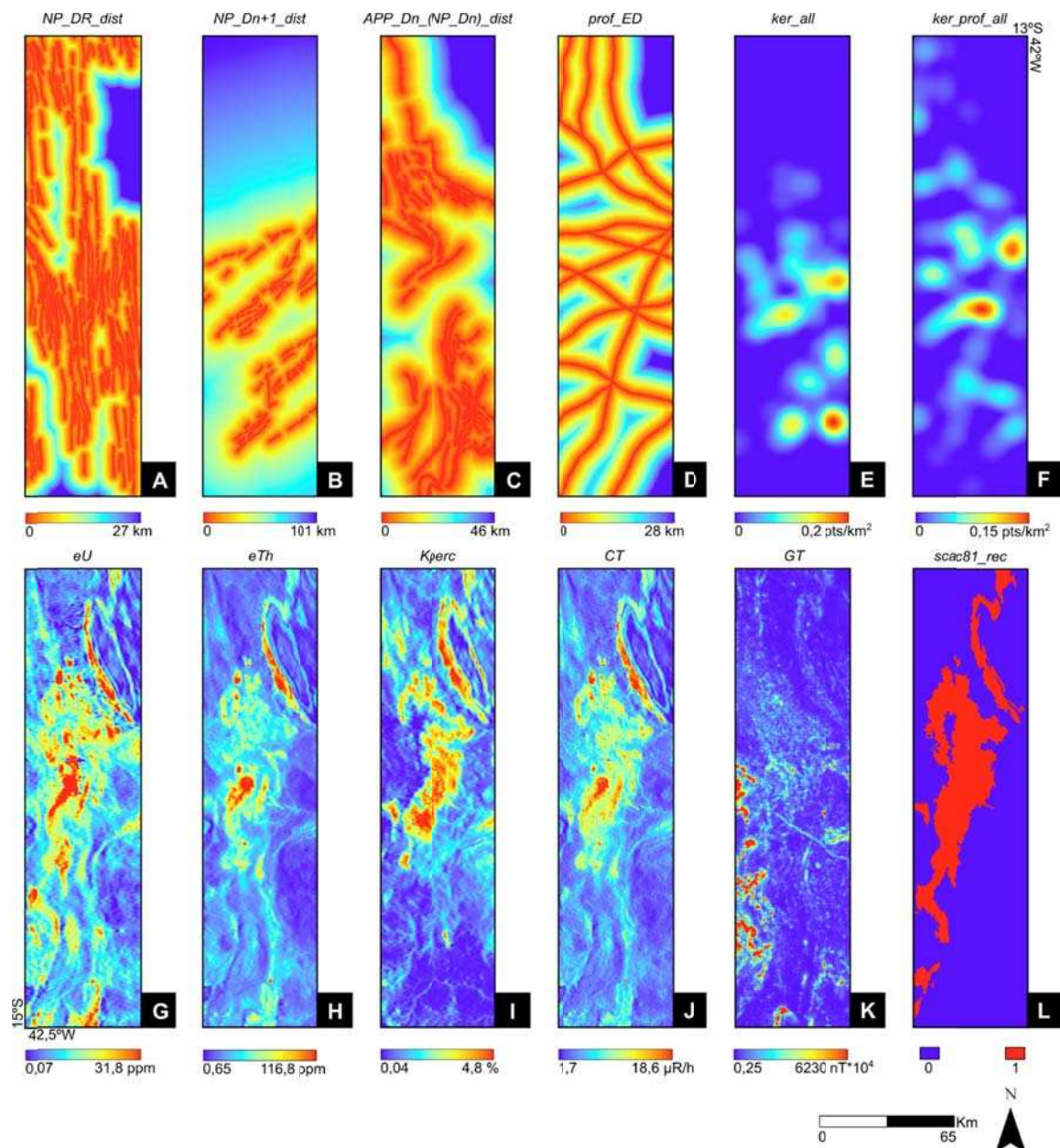


Figura 3.23: Mapas evidenciais que compõem o modelo de prospectividade mineral. A) Distância dos lineamentos NP_DR . B) Distância dos lineamentos NP_Dn+1 . C) Distância dos lineamentos $APP_Dn_ (NP_Dn)$. D) Distância dos lineamentos profundos. E) Densidade de cruzamentos de lineamentos de famílias com interesse prospectivo. F) Densidade de cruzamentos de lineamentos profundos com lineamentos de famílias com interesse prospectivo. G) Urânio. H) Tório. I) Potássio. J) Contagem total. K) Gradiente total do CMA. L) Clusters com maior semelhança às zonas mineralizadas.

3.6. Modelo de prospectividade para U

Neste item são apresentados os resultados de cada etapa da geração do modelo de prospectividade para urânio, que foram as seguintes: 1) geração do conjunto de dados de treinamento; 2) seleção do algoritmo de classificação; 3) treinamento, otimização e avaliação do modelo obtido. Ademais, após a geração do modelo de prospectividade, realizou-se uma avaliação quantitativa do impacto do conjunto de treinamento no modelo obtido, que é apresentada ao final do item. Essas 3 etapas, além das etapas anteriores relacionadas à geração dos mapas evidenciais, são sumarizadas no fluxograma da Figura 3.24.

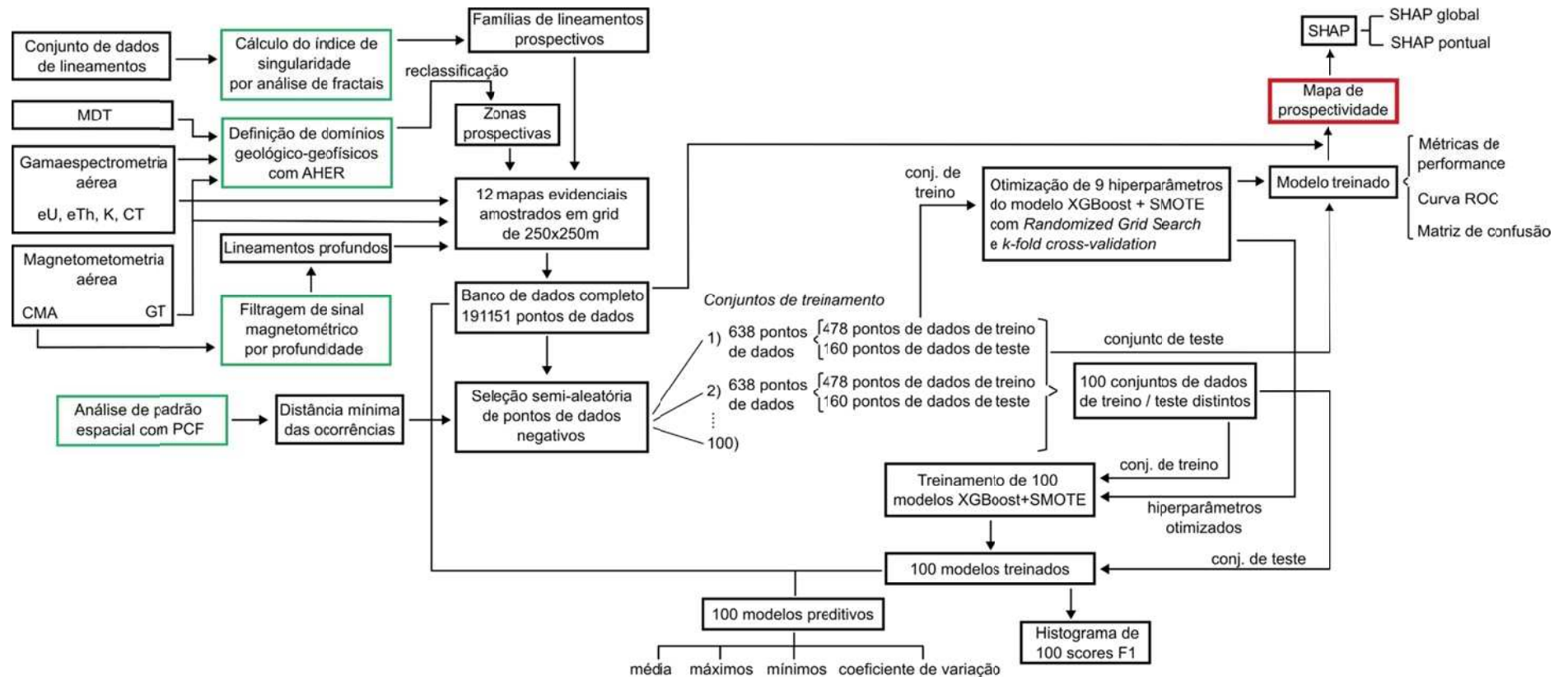


Figura 3.24: Fluxograma da geração do modelo de potencial para urânio, incluindo as etapas de geração de mapas evidenciais, do conjunto de dados de treinamento, otimização e avaliação do modelo final.

3.6.1. Geração do conjunto de dados de treino

Através da análise da função $g(r)$, determinou-se um raio mínimo de 3600 m ao redor das ocorrências conhecidas para seleção de pontos de treinamento negativos. Esse raio foi escolhido pois, a partir de 3600 m, a função $g(r)$ se estabiliza próximo de 1, diminuindo-se a chance de se selecionar um local com ocorrência não conhecida (Figura 3.25).

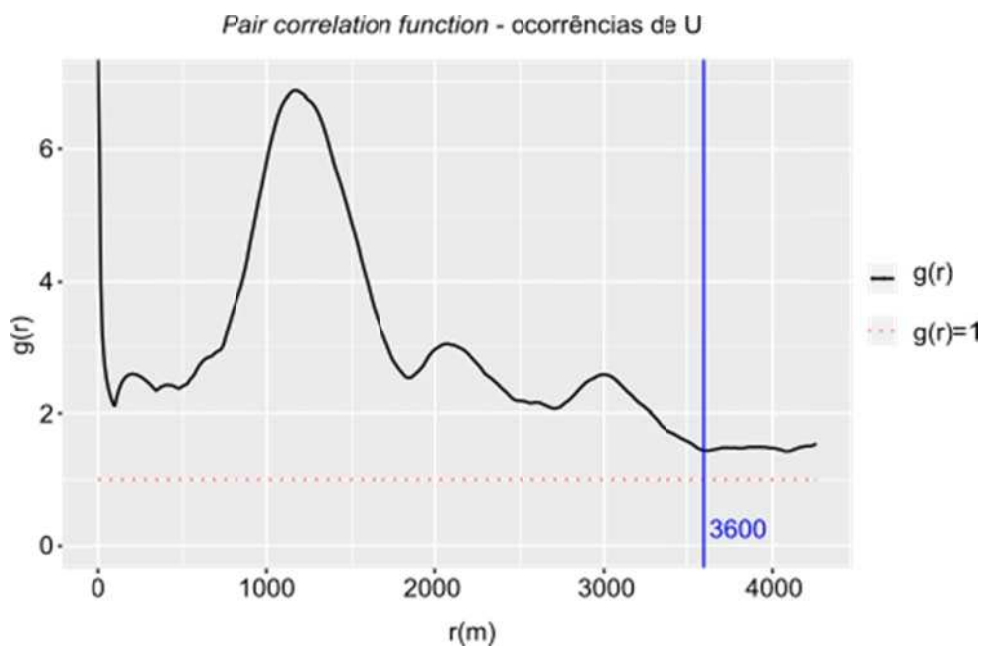


Figura 3.25: PCF (*Pair Correlation Function*) dos pontos com ocorrências de urânio conhecidas. Observa-se que, a partir de 3600 m, o $g(r)$ se estabiliza próximo de 1.

Para a geração do conjunto de dados de treinamento foram selecionados aleatoriamente 600 pontos em locais sem ocorrências de urânio conhecidas. Para se obter uma melhor distribuição areal dos pontos determinou-se, através de tentativa e erro, uma distância mínima de 1800 m entre cada ponto de treinamento negativo (Figura 3.26).

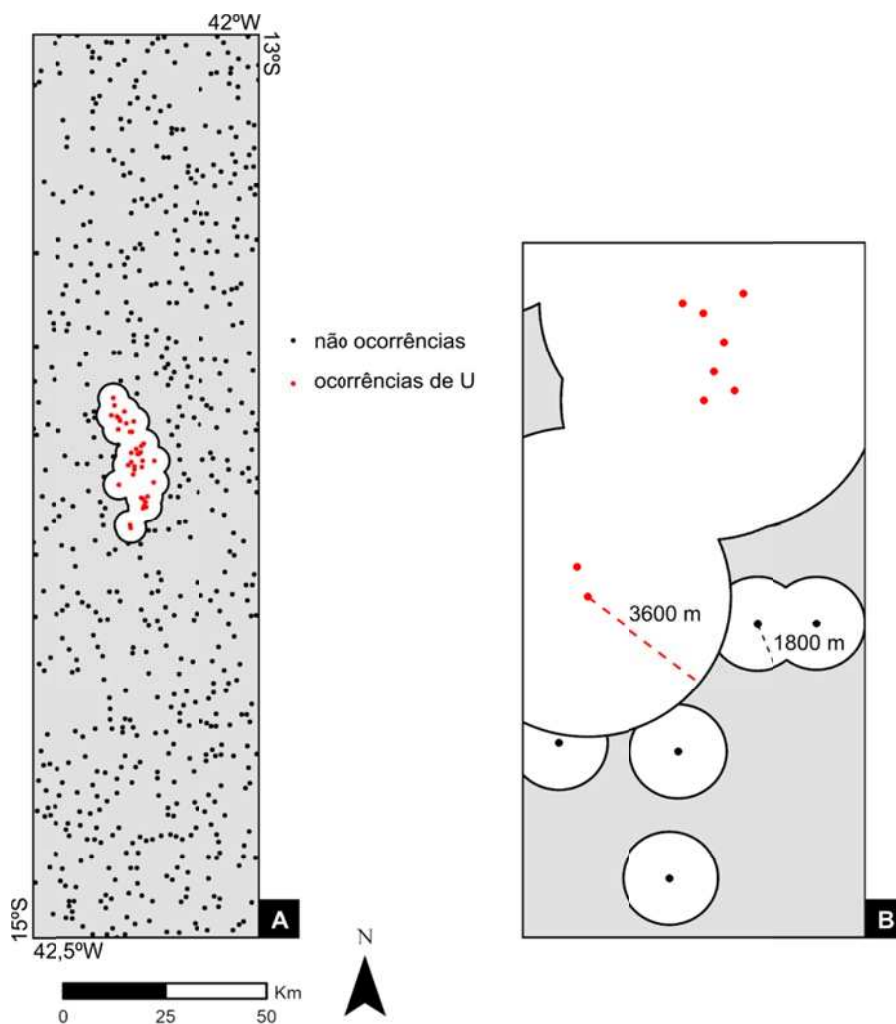


Figura 3.26: Pontos de treinamento utilizados no modelo. A) 600 pontos de não ocorrências foram gerados aleatoriamente, respeitando-se a distância mínima de 3600m das ocorrências conhecidas (total de 33). B) Detalhe ilustrando a distância mínima dos pontos de ocorrência (3600 m) e a distância mínima (1800 m) entre os pontos de não ocorrência gerados.

Dessa forma, o conjunto de dados de entrada do modelo é composto de um total de 638 pontos, sendo 38 ocorrências e 600 não ocorrências, portanto, com uma proporção de amostras da classe minoritária para a majoritária de 0,064.

A partir do conjunto de 638 pontos, 25% dos dados foram separados em um conjunto de teste, de forma estratificada (com a mesma proporção de ocorrências para não ocorrências do conjunto original). O conjunto de treinamento (os 75% restante dos dados) foi utilizado para o ajuste dos classificadores. O conjunto de teste foi utilizado para a avaliação de desempenho dos modelos ajustados.

3.6.2. Seleção do classificador

O conjunto de dados de treinamento foi utilizado para o ajuste de 6 classificadores diferentes, que tiveram seus desempenhos avaliados com o conjunto de dados de teste (Tabela 7).

Tabela 7: Avaliação de performance de 6 classificadores diferentes nos conjuntos de dados de teste.

Classificador	Acurácia	Precisão	Recall	F ₁	ROC-AUC
XGBoost	0,994	0,909	1,000	0,952	1,000
Random Forest	0,987	0,900	0,900	0,900	0,997
Gradient Boosting	0,981	0,818	0,900	0,857	0,999
Logistic Regression	0,975	0,750	0,900	0,818	0,993
SVM	0,950	0,563	0,900	0,692	0,973
Naive Bayes	0,937	0,500	0,900	0,643	0,986

Os resultados mostram que os algoritmos RF e XGB apresentaram os melhores desempenho. Em estudo comparativo realizado por [Parsa \(2021\)](#), ambas técnicas apresentaram resultados comparáveis, com desempenho ligeiramente melhor do XGB. Como observado por [Prado et al. \(2020\)](#), a métrica F₁ é mais relevante para MPM, uma vez que leva em conta quantas amostras positivas (ocorrências) foram classificadas corretamente, assim como o desempenho do modelo em não gerar falsos positivos, isso é, classificar amostras negativas como positivas. Portanto, devido ao score F₁ maior do XGB, decidiu-se proceder à elaboração do mapa de potencial mineral com esse classificador.

3.6.3. Treinamento, otimização e avaliação do modelo de prospectividade

O XGB tem bom desempenho em uma grande variedade de aplicações diferentes ([Chen e Guestrin 2016](#)), como apontado por [Parsa \(2021\)](#). O método possui uma grande quantidade de hiperparâmetros que devem ser ajustados para se obter um melhor desempenho na classificação e para se evitar o sobre-ajuste (*overfitting*) do modelo aos dados de treino. Para tanto, se utilizou o algoritmo *randomized grid search*, disponível no pacote *scikit-learn*, que realiza uma busca aleatória dentro de um determinado espaço de combinações de parâmetros possíveis. A lógica da utilização da busca aleatória baseia-se na possibilidade de busca em um espaço grande de combinações de hiperparâmetros, sem que o tempo

de processamento se torne excessivo, o que poderia ocorrer no caso de uma busca exaustiva de todas as combinações dentro do mesmo espaço de possibilidades.

Os parâmetros ajustados para o XGB+SMOTE foram (Tabela 8): *learning_rate* - taxa de encolhimento dos pesos dos estimadores. Afeta a velocidade de ajuste dos erros para cada novo estimador, utilizado para evitar o sobre-ajuste do modelo; *n_estimators* – número de estimadores (árvores de decisão); *max_depth*: número máximo de níveis (profundidade) de cada estimador; *min_child_weight* – número mínimo de amostras em cada nóduo resultante de uma possível divisão; *subsample* – proporção de amostras selecionadas para o subconjunto de treinamento na geração de uma nova árvore; *colsample_bytree* - proporção de variáveis selecionadas para a geração de uma nova árvore; *lambda* – regularização L2; *alpha* – regularização L1; *sampling_strategy* – quantidade de sobreamostragem realizada pelo SMOTE, representada pela proporção resultante de amostras da classe minoritária para a classe majoritária.

Tabela 8: Hiperparâmetros do XGB+SMOTE otimizados por *grid search*.

Parâmetro	Valores testados	Melhor valor
<i>learning_rate</i>	0,01; 0,05; 0,1	0,1
<i>n_estimators</i>	200, 300, 400	200
<i>max_depth</i>	3; 5; 7	5
<i>min_child_weight</i>	3; 5; 10	3
<i>subsample</i>	0,5; 0,7	0,7
<i>colsample_bytree</i>	0,5; 0,7	0,5
<i>lambda</i>	1; 10; 100	1
<i>alpha</i>	0; 0,1; 1; 10	0,1
<i>sampling_strategy</i>	0,5; 1	0,5

Após a definição dos melhores parâmetros com base no *randomized grid search*, o modelo XGB+SMOTE foi ajustado aos dados de treinamento. A partir da avaliação realizada com o conjunto de dados de teste, foram calculadas as métricas de performance apresentadas na Tabela 9.

Tabela 9: Métricas de desempenho do modelo XGB+SMOTE com hiperparâmetros otimizados, obtidas com os dados de treino e teste.

Score	Treino	Teste
Acurácia	0,996	0,987
Precisão	0,970	0,920
Recall	1,000	0,990
F ₁	0,966	0,909
ROC-AUC	1,000	0,990

As métricas de desempenho mostram que foi possível se obter um ajuste muito bom, com scores próximos do máximo. Como é de se esperar, os scores do conjunto de teste foram ligeiramente menores do que os do conjunto de treino, uma vez que os dados do primeiro não são vistos pelo modelo durante o treinamento. A pequena diferença entre os scores obtidos nos conjuntos de treino e teste mostra que não houve sobreajuste ao conjunto de treino.

Ademais, foi elaborada uma curva ROC para a avaliação de desempenho do modelo em diferentes limiares de probabilidade de classificação (Figura 3.27), e uma matriz de confusão, mostrando o resultado da classificação com um limiar de probabilidade de 0,5 (Figura 3.28). Quanto mais próximas as curvas estiverem do canto superior esquerdo, melhor o desempenho do classificador. Observa-se que o modelo obteve excelente desempenho nos dados de treino e teste, com score ROC-AUC arredondado para 1 no conjunto de treino e 0,99 no de teste. Na matriz de confusão, observa-se que todos 148 pontos de não ocorrências e 9 de ocorrências foram classificados corretamente. Somente 3 pontos de não ocorrências foram classificados equivocadamente como ocorrências.

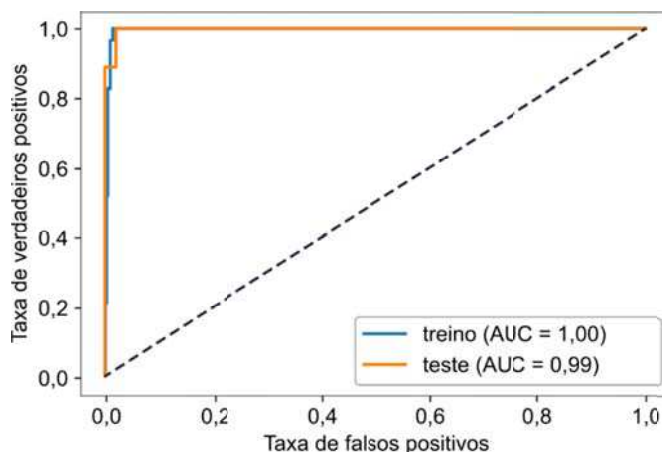


Figura 3.27: Curva ROC e score AUC do modelo preditivo para os conjuntos de dados de treino e de teste. A linha diagonal tracejada representa a classificação de um preditor aleatório.

Real	Não ocorrência	148 (Negativos reais)	3 (Falsos positivos)
	Ocorrência	0 (Falsos negativos)	9 (Positivos reais)
		Não ocorrência	Ocorrência
		Predito	

Figura 3.28: Matriz de confusão da predição do conjunto de dados de teste para um limiar de classificação de 0,5.

Em seguida, o modelo treinado foi utilizado para prever a probabilidade de ocorrências (ou *prospectividade*) no conjunto total de amostras da área de estudo, gerando-se o mapa de prospectividade para urânio (Figura 3.29).

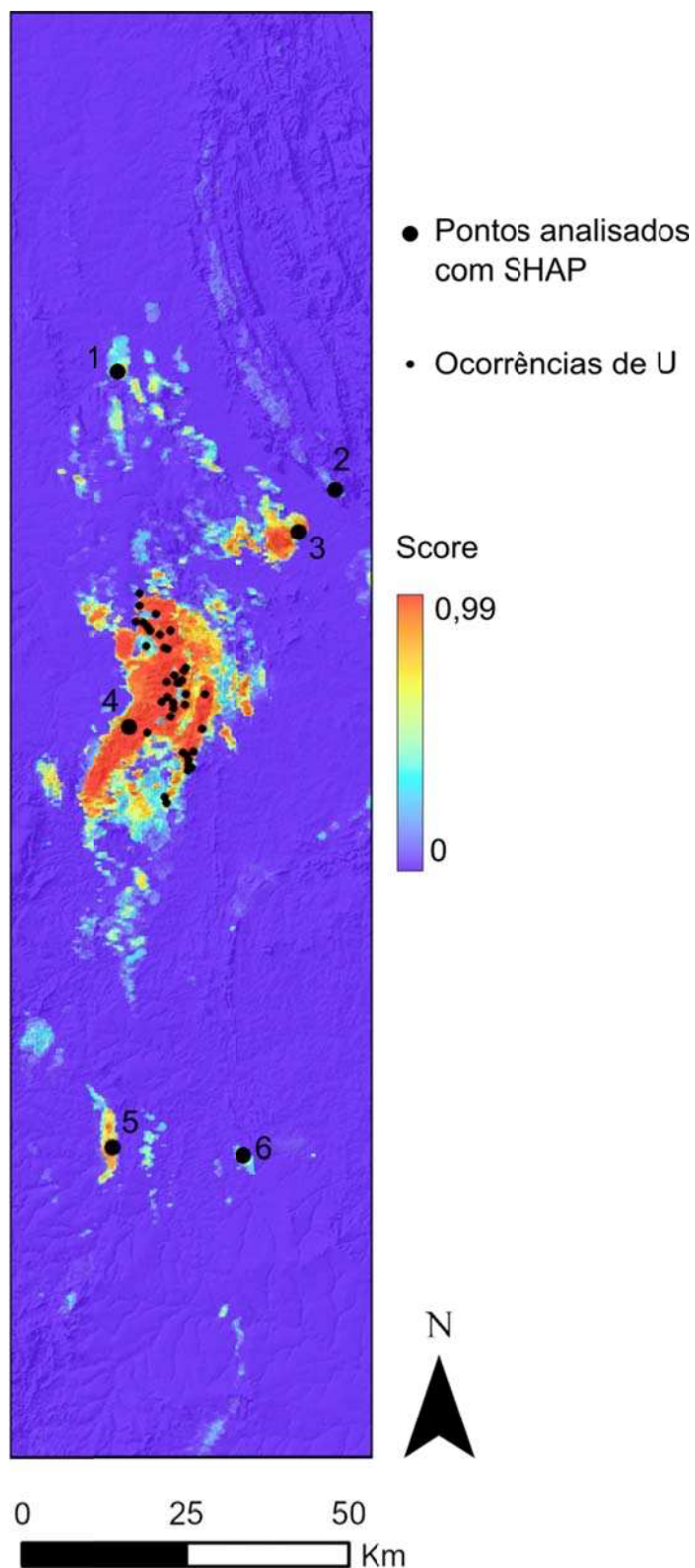


Figura 3.29: Modelo de prospectividade para urânio gerado com a combinação XGB+SMOTE. Os pontos numerados indicam as amostras explicadas com SHAP. O score corresponde à probabilidade de predição de uma ocorrência pelo modelo (score = 1).

O mapa de prospectividade foi avaliado em termos da porcentagem dos depósitos conhecidos incluídos, em relação à porcentagem de área mapeada em determinado score de prospectividade (Figura 3.30). Destaca-se que 95,1% das ocorrências conhecidas encontram-se mapeadas dentro de apenas 2,3% das áreas com scores de prospectividade igual ou maior do que 0,9, e 100% das ocorrências se encontram em uma fração da área de 4,7%, com scores de prospectividade igual ou maior do que 0,5.

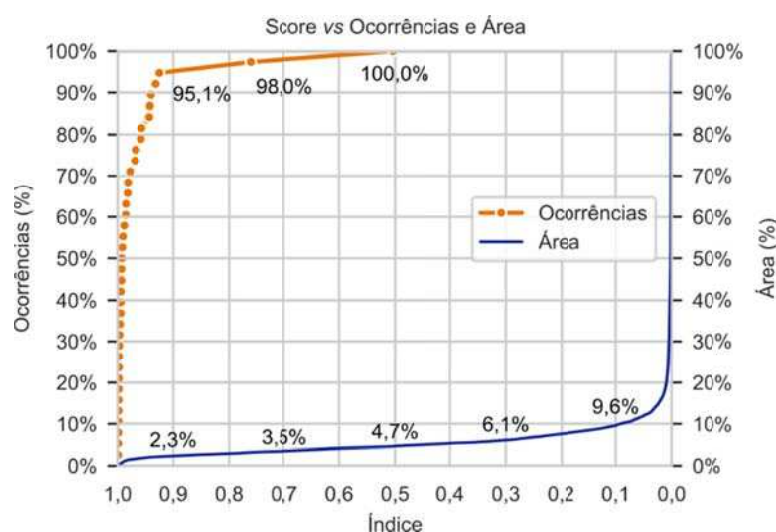


Figura 3.30: Gráfico de avaliação de performance do modelo de prospectividade em relação às ocorrências conhecidas. A curva laranja se refere à porcentagem das ocorrências, e a azul à porcentagem da área de trabalho.

Para se entender a contribuição das variáveis no resultado do modelo, realizou-se um gráfico de contribuição global das variáveis com a técnica SHAP. Este gráfico mostra a média da contribuição, em valor SHAP, de cada variável, em toda a área do modelo gerado (Figura 3.31).

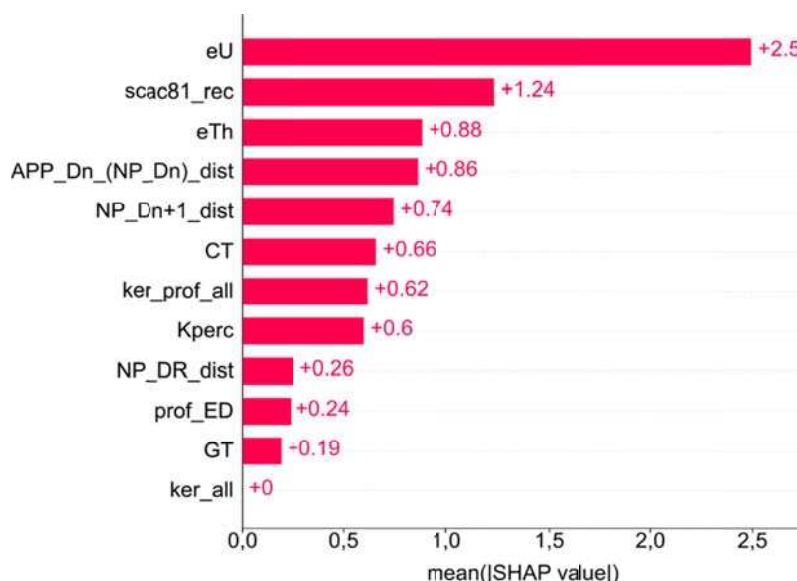


Figura 3.31: Média global dos valores SHAP de cada variável do modelo preditivo. **eU**: urânio; **scac81_rec**: clusters com maior semelhança às zonas mineralizadas; **eTh**: tório; **APP_Dn_(NP_Dn)_dist**: distância dos lineamentos *APP_Dn_(NP_Dn)*; **NP_Dn+1_dist**: distância dos lineamentos *NP_Dn+1*; **CT**: contagem total; **ker_prof_all**: densidade de cruzamentos de lineamentos profundos com lineamentos de famílias com interesse prospectivo; **Kperc**: potássio; **NP_DR_dist**: distância dos lineamentos *NP_DR*; **prof_ED** distância dos lineamentos profundos; **GT**: gradiente total do CMA; **ker_all**: densidade de cruzamentos de lineamentos de famílias com interesse prospectivo.

A variável com a maior contribuição nos scores de prospectividade foi o **eU**, com valor SHAP médio de 2,5. Em segundo lugar, com praticamente a metade do primeiro, ficou o mapa de zonas prospectivas (**scac81_rec**) obtido da reclassificação dos domínios geológico-geofísicos gerado com o AHER. As próximas 6 variáveis, com valores entre 0,88 e 0,6, foram **eTh**, **APP_Dn_(NP_Dn)_dist**, **NP_Dn+1_dist**, **CT**, **ker_prof_all**, **Kperc**. Com os menores valores, as variáveis **NP_DR_dist**, **prof_ED** e **GT** apresentaram 0,26 e 0,19, respectivamente. Por fim, a variável **ker_all** teve valor SHAP médio 0, mostrando que ela não teve contribuição no aumento dos scores de prospectividade.

Para se avaliar como as variáveis contribuíram de formas diversas para a prospectividade em diferentes regiões da área de estudo, foram selecionadas 6 áreas que apresentaram maiores scores de prospectividade, que foram avaliadas com SHAP aplicado a amostras pontuais. Em cada uma dessas regiões selecionou-

se um ponto com prospectividade elevada para ser analisado (Figura 3.29). Os resultados são apresentados na forma de *waterfall plots* (“diagramas cachoeira”) apresentados na figura Figura 3.32.

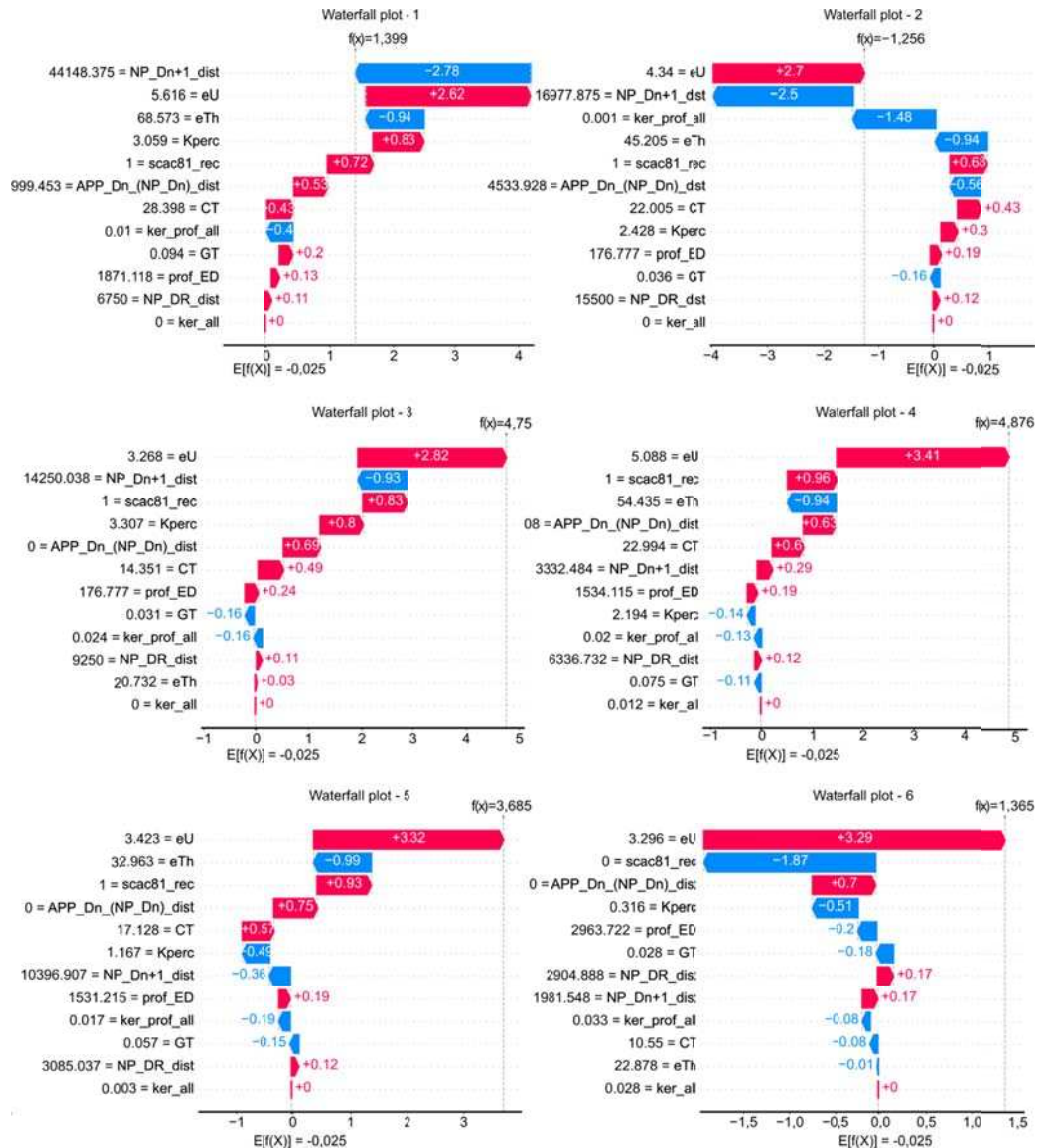


Figura 3.32: *Waterfall plots* com as explicações SHAP dos pontos do modelo indicados na Figura 3.29. Os valores das variáveis na amostra analisada são mostrados no eixo y. As barras vermelhas indicam a intensidade da contribuição (valor SHAP) de cada variável para o aumento do score, enquanto as barras azuis mensuram a contribuição para a diminuição do score no ponto. Neste gráfico os scores são medidos em *log-odds*. **eU**: urânio; **scac81_rec**: clusters com maior semelhança às zonas mineralizadas; **eTh**: tório; **APP_Dn_(NP_Dn)_dist**: distância dos lineamentos *APP_Dn_(NP_Dn)*; **NP_Dn+1_dist**: distância dos lineamentos *NP_Dn+1*; **CT**: contagem total; **ker_prof_all**: densidade de cruzamentos de lineamentos profundos com lineamentos de famílias com interesse prospectivo; **Kperc**: potássio; **NP_DR_dist**: distância dos lineamentos *NP_DR*; **prof_ED** distância dos lineamentos profundos; **GT**: gradiente total do CMA; **ker_all**: densidade de cruzamentos de lineamentos de famílias com interesse prospectivo.

Os resultados dos diagramas cachoeira dos valores SHAP mostram que as amostras 1, 2 e 3 tiveram em comum contribuições positivas principalmente das variáveis *eU*, *scac81_rec*, *Kperc*, *CT* e *APP_Dn_(NP_Dn)_dist*; e negativas de *NP_Dn+1_dist*. Já as amostras 4, 5 e 6 tiveram em comum contribuições positivas principalmente das variáveis *eU*, *APP_Dn_(NP_Dn)_dist*; e negativas de *eTh* e *Kperc*. A variável *scac81_rec* foi positiva para todas as amostras, com exceção da 6, que se encontra fora da área favorável definida pela mesma.

Observa-se que o resultado de prospectividade em resposta à algumas das variáveis não é monotônico, ou seja, não acompanha necessariamente o aumento ou a diminuição de seu valor, de maneira uniforme. Um exemplo é o tório, cuja contribuição diminui, em geral, com o aumento de seu valor. No gráfico de dependência do tório em relação ao valor SHAP (Figura 3.33), é possível observar que seu comportamento se estrutura distintamente, de maneira geral, em 3 intervalos: i) máximos valores SHAP entre 0 e cerca de 18,5 ppm, com o máximo ocorrendo perto de 18 ppm; ii) entre cerca de 18,5 e 24 ppm, valores ligeiramente negativos a neutros; iii) menores valores SHAP a partir de 24 ppm, com o mínimo ocorrendo em cerca de 26 ppm.

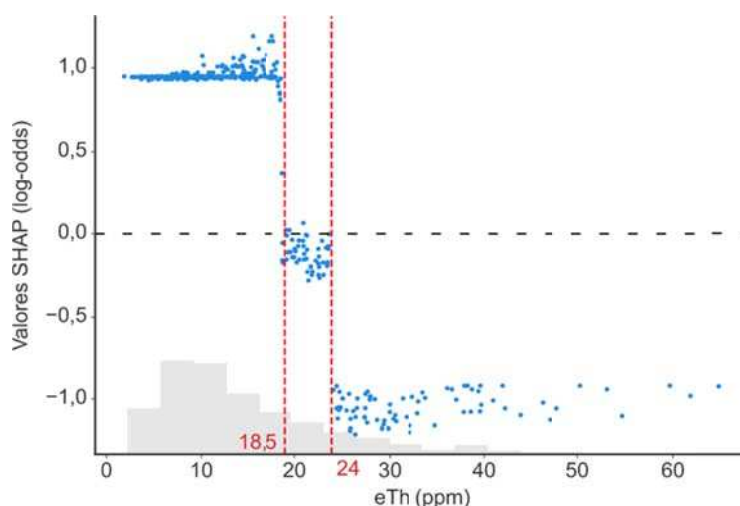


Figura 3.33: Gráfico de dependência do tório em relação ao valor SHAP. Cada ponto corresponde a uma amostra do conjunto de dados de treinamento.

Levando-se em consideração o caráter semi-aleatório contido no processo de subamostragem da classe majoritária (pontos sem ocorrências) para a geração do conjunto de dados de treinamento, e que tal conjunto é composto de, somente, cerca de 0,33% do total do banco de dados completo, procedeu-se à uma

análise quantitativa da influência da variação do conjunto de dados de treinamento no modelo de prospectividade final obtido. Para tanto, foram gerados 100 conjuntos de treino e teste (Figura 3.34), que foram utilizados para o ajuste de 100 modelos com os mesmos hiperparâmetros (Figura 3.24). Para se visualizar a variabilidade dos resultados, obtidos a partir dos diferentes conjuntos, foram compostos mapas dos seguintes parâmetros estatísticos: valores máximos, valores mínimos, médias e coeficientes de variação Figura 3.35.

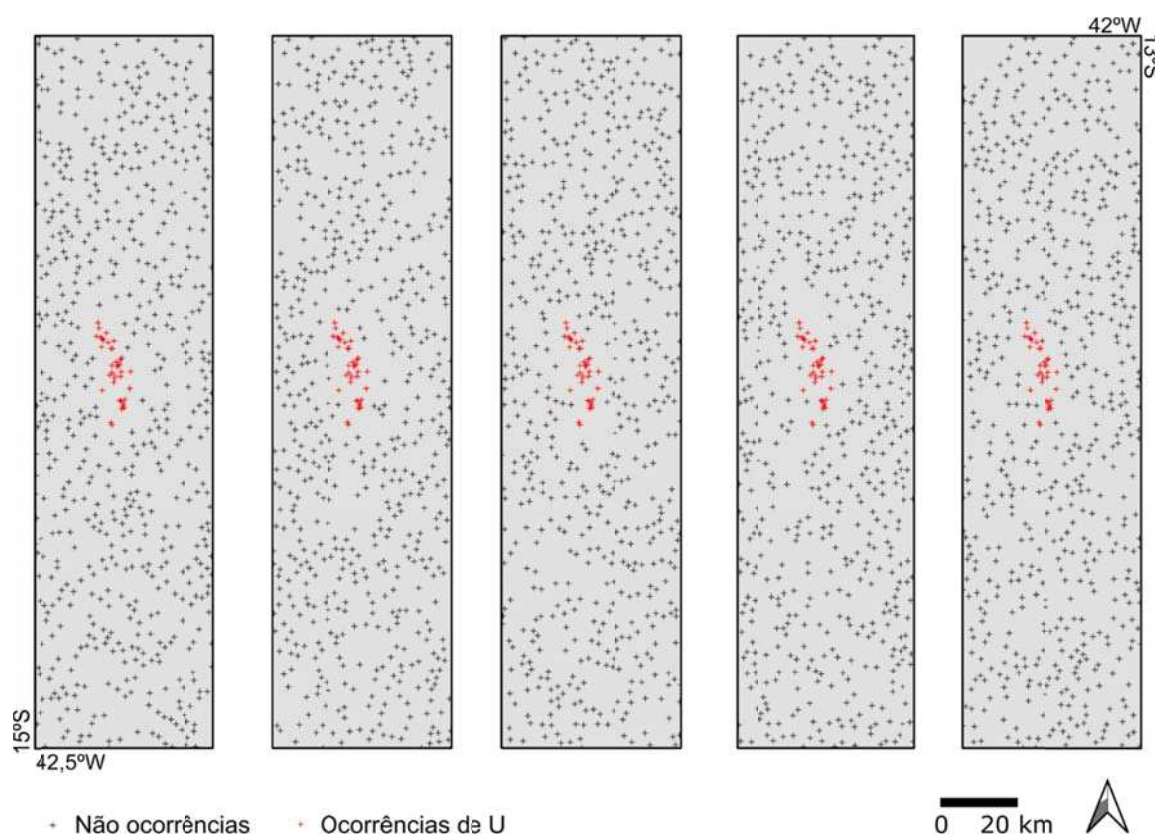


Figura 3.34: Exemplo de 5, dos 100, conjuntos de dados de treino, gerados de forma independente, seguindo os critérios apresentados em 3.6.1.

O mapa de valores máximos apresenta scores de prospectividade mais elevada com uma abrangência maior do que o modelo individual, porém, sem alterações significativas em relação às regiões onde elas ocorrem. Já no mapa de mínimos, diversas regiões que aparecem no modelo individual apresentam score nulo, o que faz o mapa ter uma abrangência de prospectividade bem mais restrita às regiões próximas das ocorrências conhecidas. O mapa de CV mostra que há uma tendência geral de menor variabilidade dos resultados nas áreas com scores médios mais elevados, o que pode ser verificado no diagrama de dispersão da Figura 3.37.

A partir do 97° percentil das médias (0,217), que corresponde aos 3% dos pontos com resultados médios mais altos, o CV médio é de 0,21.

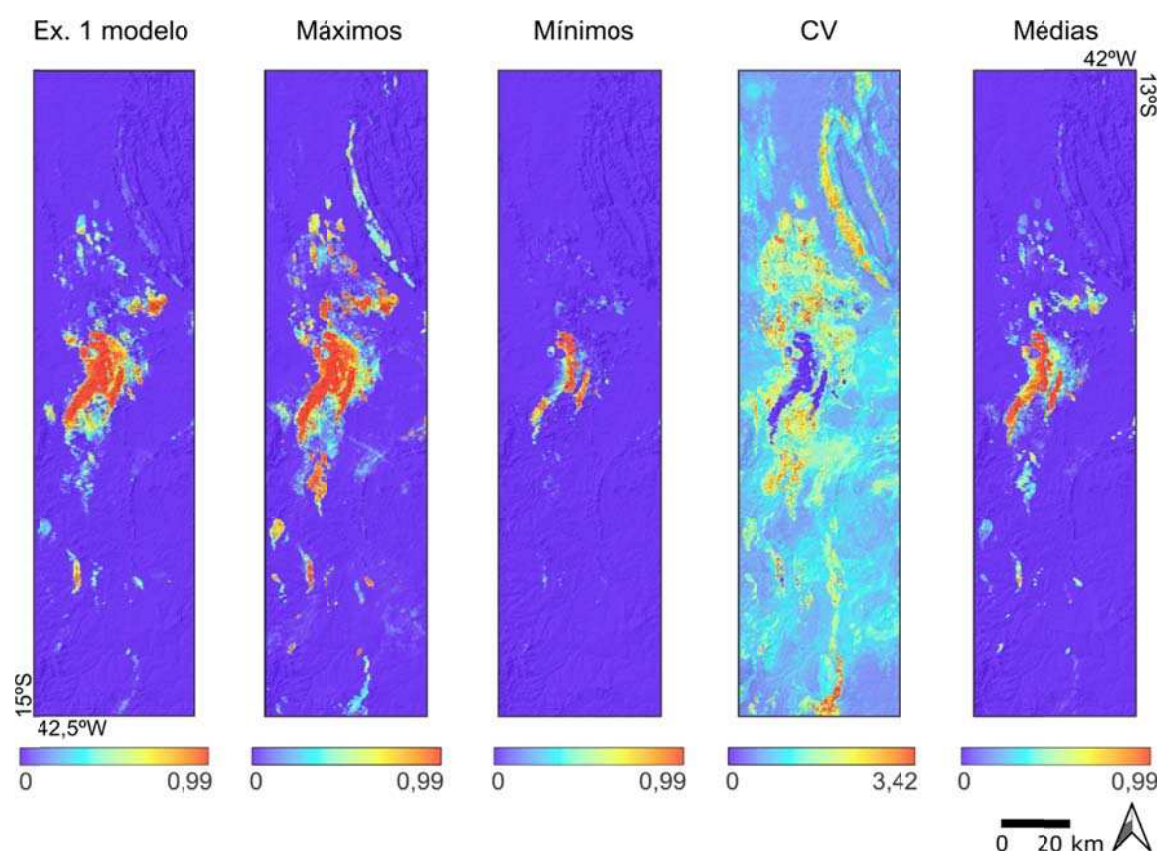


Figura 3.35: Mapas resultantes da análise estatística da variabilidade dos resultados do modelo de prospectividade para U para 100 conjuntos de dados de treinamento diferentes. O primeiro mapa (da esquerda) é um exemplo dos 100 mapas gerados. Os demais são os mapas das estatísticas calculadas para os 100 modelos em conjunto.

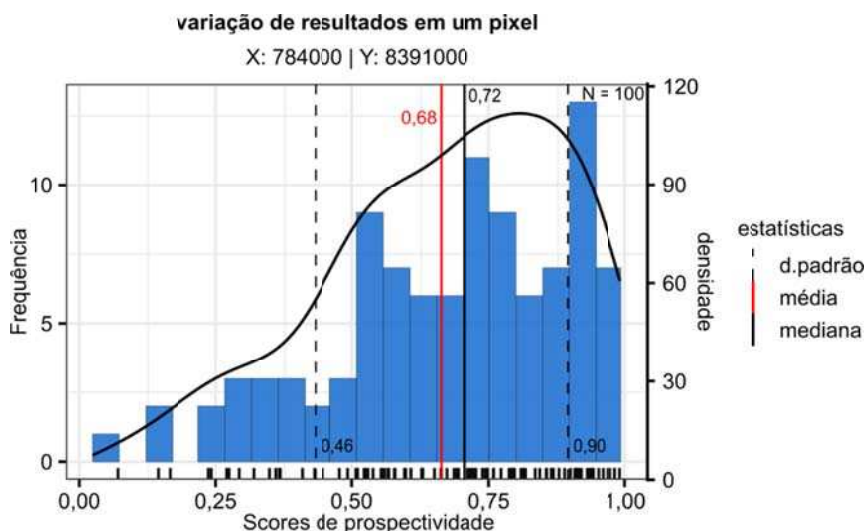


Figura 3.36: Histograma e função de densidade de probabilidade dos resultados de prospectividade dos 100 modelos em 1 pixel do mapa. São indicados no histograma os seguintes parâmetros estatísticos: média, desvio padrão, e mediana.

Na avaliação individual de um pixel (1 local do modelo), observa-se que os resultados apresentam uma distribuição assimétrica à esquerda, com mediana (0,72) ligeiramente maior do que a média (0,66), e desvio padrão de 0,22 (Figura 3.36).

O histograma dos scores F_1 calculados para os conjuntos de teste dos 100 modelos mostra uma distribuição próxima à normal, com a média e a mediana muito próximas entre si (0,77 e 0,76 respectivamente), máximo de 0,89, mínimo de 0,64, e desvio padrão de 0,05 (Figura 3.38).

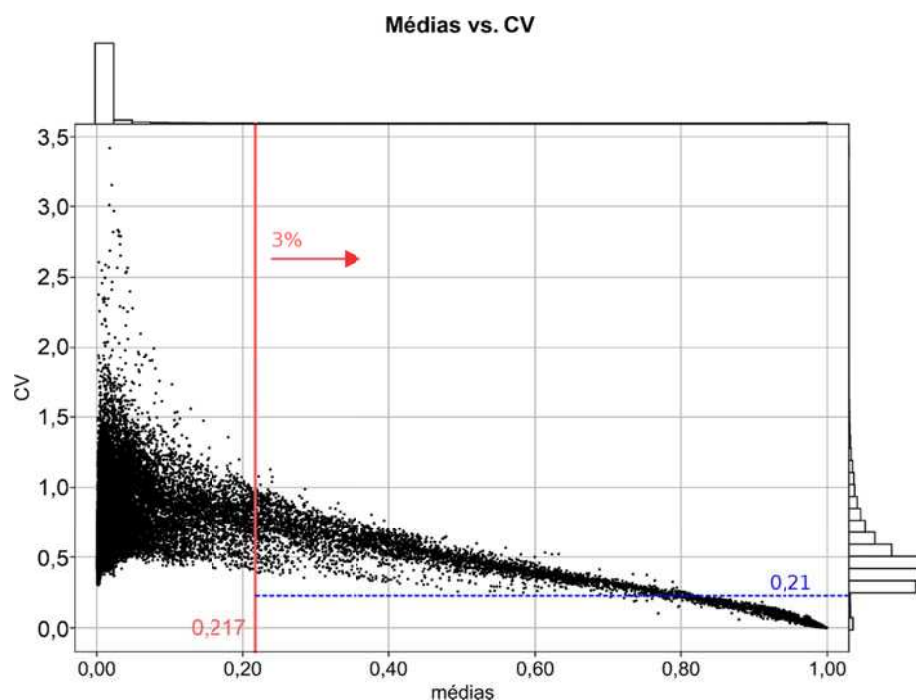


Figura 3.37: Diagrama de dispersão, combinado com histogramas, das médias vs. coeficientes de variação dos pixels que apresentaram máximos maiores do que 0 no conjunto dos 100 modelos de prospectividade. A linha vermelha indica o 97º percentil das médias. A linha azul o valor de CV médio acima do 97º percentil das médias.

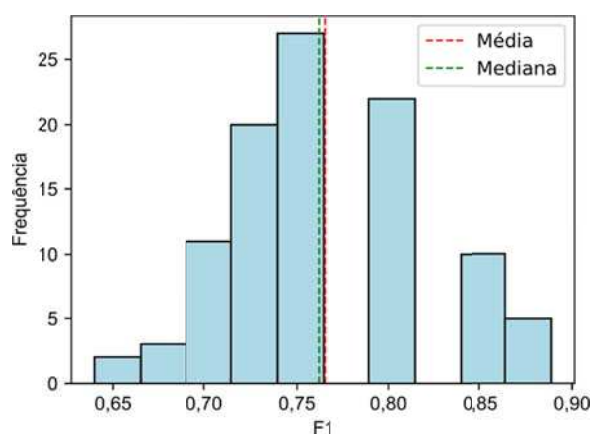


Figura 3.38: Histograma dos scores F_1 , calculados para os conjuntos de teste de cada um dos 100 modelos gerados com os diferentes conjuntos de treinamento.

4. Discussões

O uso do aprendizado não-supervisionado para análise do banco de dados de aerogeofísica e MDT permitiu uma análise exploratória aprofundada desses dados. O uso do SOM proporcionou a visualização de padrões de distribuição conjunta desses dados e de como os mesmos se relacionam com as diferentes feições geológicas presentes na área de estudo. Com a técnica CPN foi possível visualizar características geológicas selecionadas (unidades geológicas ou mineralizações) no mapa SOM e avaliar o seu comportamento em relação às variáveis modeladas. Nota-se que quando se projeta as unidades geológicas no mapa SOM, enquanto algumas unidades mostram um padrão de ativação em regiões bem definidas, outras mostram ativações dispersas. É possível se observar, por exemplo, que a assinatura produzida pela Formação Açuruá é consideravelmente mais restrita do que aquela produzida pelo Complexo Gavião – Migmatitos (Figura 3.4), sugerindo que, a variação litológica incluída na primeira, é menor do que na segunda. Isso pode ter diversos motivos, como a variação composicional própria da unidade, ou um baixo detalhamento no mapeamento geológico da mesma, que resultaria no agrupamento de litotipos com características variadas e, portanto, ativações dispersas no mapa SOM. A projeção da assinatura obtida pelo SOM, de forma reversa ao mapa da área, produz um mapa de intensidade dessa assinatura mapeada na área, onde os maiores valores representem regiões com assinaturas típicas da característica geológica modelada. Esse mapa pode ser utilizado, por exemplo, para se identificar regiões com assinaturas típicas da unidade modelada, mas que foram classificadas em outras unidades. No mapa de assinatura, produzido com essa técnica, do Complexo Lagoa Real (Figura 3.5), se destaca uma faixa com direção aproximada NW-SE, na região NE da área, com altos valores de assinatura dessa unidade. No entanto, a cartografia geológica do mapa 1:1.000.000 mapeia essa faixa como Formação Rio dos Remédios. Não obstante, a semelhança da assinatura dessa unidade com aquela do Complexo Lagoa Real pode ser considerada coerente, tendo em vista que estudos mais detalhados nessa região definem essa unidade como Grupo Rio dos Remédios, contendo em sua base a formação Novo Horizonte, que é constituída por uma sequência com 600 m de espessura de rochas vulcânicas ácidas, correlacionadas, temporal e geneticamente, com os granitoides félsicos do

Complexo Lagoa Real ([Cruz e Alkmim 2017](#), [Danderfer et al. 2009](#), [Guimarães et al. 2008](#)).

A análise dos resultados obtidos a partir de modelos SOM, utilizados com a técnica CPN, mostrou uma forte variabilidade quando avaliados de forma quantitativa, com 56% dos pixels atingindo coeficientes de variação acima de 2. No entanto, é relevante destacar que a variabilidade diminui com o aumento das médias dos modelos, sendo que os dados correspondentes às médias maiores do que o 90º percentil (médias maiores do que 0,0073), possuem um coeficiente de variação médio de 0,6 (Figura 3.8). Não obstante, [Wehrens e Buydens \(2007\)](#) ressaltam que apesar da variabilidade considerável dos modelos SOM, as interpretações qualitativas derivadas dos mesmos são coerentes entre si. Nesse sentido, a geração de múltiplos modelos, seguida da geração de mapas dos máximos, mínimos, coeficientes de variação e médias, mostra-se uma alternativa interessante para uma interpretação mais robusta dos resultados.

O agrupamento espacialmente restrito dos dados de aerogeofísica e de MDT, com a técnica AHER, resultou em uma delimitação de grupos com continuidade espacial mais coerente do que o realizado com SOM + K-means, considerando-se que estão sendo agrupadas propriedades derivadas das litologias em subsuperfície que possuem continuidade espacial (Figura 3.12 e Figura 3.15). As unidades delimitadas por SOM + K-means conformam domínios espacialmente fragmentados, aumentando a chance de estarem agrupando elementos de diferentes unidades e contextos geológicos. Na sobreposição dos grupos ao modelo digital de relevo, observa-se que os limites dos mesmos são congruentes com os diferentes padrões de relevo. No modelo obtido por AHER, com 81 grupos, é possível se observar uma segmentação significativamente maior em comparação ao mapa geológico na escala 1:1.000.000. Em relação à delimitação de áreas prospectivas, ao se comparar as unidades obtidas nos agrupamentos com as unidades geológicas, considerando aquelas que apresentam ao menos uma ocorrência de urânio, ambas técnicas (SOM + K-means e AHER) mostraram potencial vantagem prospectiva em relação às unidades da cartografia geológica (Tabela 4, Tabela 5; Figura 3.16, Figura 3.17). No entanto, as unidades delimitadas por AHER tiveram melhores resultados em parâmetros prospectivos, além de mostrarem uma continuidade espacial geologicamente mais coerente. Observa-se

que, os três grupos definidos por AHER com mais ocorrências, que juntos compreendem cerca de 87% das ocorrências, apresentam densidades de ocorrências de 7 a 2,5 vezes maior do que a unidade geológica mais prospectiva, e valores-t de 2,6 a 1,2 vezes maiores. A classificação dos grupos em escalas de cor, que representam a dissimilaridade de cada grupo com aquele com mais ocorrências (Figura 3.18-C), ressaltou principalmente grupos na região do Complexo Lagoa Real, além de uma faixa NW-SE, na região NE da área, assim como ocorreu no modelo gerado com o SOM (Figura 3.6). Já a classificação baseada no valor-t (Figura 3.18-D) resalta menos grupos, uma vez que o cálculo desse valor, baseado na estatística bayesiana de probabilidade condicional, leva em conta a presença/ausência de ocorrências em determinada extensão de área de busca.

A modelagem preditiva com aprendizado supervisionado realizada com o algoritmo de classificação XGB, implementado em combinação com SMOTE para o balanceamento do número de amostras positivas/negativas, mostrou uma excelente performance no banco de dados de teste, medida pelos scores acurácia, precisão, recall, F_1 e ROC-AUC (Tabela 9). A análise global de valores SHAP (Figura 3.31) mostrou que todas as variáveis foram relevantes para o modelo preditivo, com exceção do mapa de densidade de cruzamentos de lineamentos de famílias com interesse prospectivo (*ker_all*).

A análise de valores SHAP pontuais das 6 amostras (Figura 3.29 e Figura 3.32) mostrou que diferentes fatores influenciaram o score de prospectividade em diferentes áreas da PULR. O ponto 2 localiza-se em uma faixa com score médio a baixo, com direção aproximadamente NW-SE, no NE da área de estudo. Essa mesma faixa, que é a área de ocorrência da Formação Rio dos Remédios/Formação Novo Horizonte, também apresentou assinatura positiva nos modelos obtidos com SOM (Figura 3.5 e Figura 3.6). No entanto, o modelo de prospectividade XGB+SMOTE apresentou um score relativamente mais baixo nessa região, uma vez que, apesar da assinatura geofísica semelhante à do Complexo Lagoa Real, ele buscou outros fatores de prospectividade associados às ocorrências conhecidas, que não estão presentes ali. Enquanto o urânio (*eU*) teve uma forte contribuição positiva no score desse ponto, este foi diminuído devido, principalmente, às variáveis relacionadas aos fatores estruturais. Considerando-se a assinatura semelhante ao Complexo Lagoa Real nessa área, seu alto sinal de urânio e a sua provável relação

genética com as rochas vulcânicas ácidas a intermediárias da Formação Rio dos Remédios/Formação Novo Horizonte (Cruz e Alkmim 2017), essa faixa, que contém o ponto 2, pode ser considerada como potencial para a ocorrência de depósitos metassomáticos em (meta)vulcânicas, caso ali sejam identificadas zonas com um arcabouço estrutural favorável.

O ponto 1, localizado pouco a N do Complexo Lagoa Real, mostrou um score moderadamente elevado, principalmente devido ao urânio e ao potássio (K_{perc}). Diferente dos demais pontos analisados, este teve uma ligeira contribuição positiva do gradiente total (GT). Litologias ricas em magnetita, que teriam uma resposta de GT mais elevada, poderiam ter favorecido a precipitação de U nessa área, tendo em vista que Lobato e Fyfe (1990) propuseram, como mecanismo de precipitação de U, a redução do fluido hidrotermal por oxidação da magnetita. Em concordância, Marques *et al.* (2021) observaram uma associação espacial entre magnetita parcialmente hematitizada e uraninita. O principal fator que diminui o score na região do ponto 1 foi a elevada distância às estruturas NP_Dn+1, que sugerem que essa região tenha sido menos afetada pela deformação brasileira associada ao orógeno Araçuai que, segundo Cruz *et al.* (2008), teria provocado a individualização dos albitos e concentração de urânio em zonas de cisalhamento.

O ponto 3 apresentou o segundo melhor score dos pontos analisados. Este ponto se localiza dentro da unidade Complexo Lagoa Real, porém em uma região sem ocorrências conhecidas, a NE dos depósitos de U. Dentre as 12 variáveis, 9 contribuíram positivamente neste local, sendo que a principal contribuição negativa foi a distância das estruturas NP_Dn+1 e, em menor grau, o gradiente total e a densidade de intersecções entre os lineamentos profundos e os prospectivos (ker_{prof_all}). Dessa forma, observa-se que as características estruturais deste local foram o principal diferenciador em relação às ocorrências conhecidas.

O ponto 4 é o que se localiza mais próximo da região das ocorrências e depósitos (a 2,5 km da ocorrência mais próxima) e apresentou o maior score de prospectividade. O principal diferenciador foi a contribuição negativa do tório (eTh), devido ao seu valor elevado. Valores gamaespectrométricos elevados de tório na região estão associados com coberturas residuais mais bem desenvolvidas, o que,

por outro lado, pode sugerir um potencial para ocorrências não aflorantes, recobertas por tais coberturas.

Os pontos 5 e 6 localizam-se em manchas isoladas de prospectividade moderada, na parte sul da área de estudo, ambos fora da unidade Complexo Lagoa Real. Os dois pontos tiveram forte contribuição positiva do urânio. No ponto 5, as principais contribuições negativas foram do tório e do potássio, o que sugere uma possível variação litológica, ou de alteração metassomática, em relação às ocorrências conhecidas. No ponto 6, a variável que provocou a maior diminuição de score foi a zona prospectiva (*scac81_rec*), uma vez que esse ponto se encontra fora da mesma. Em segundo foi o potássio, assim como no ponto 5. Ambos pontos mostram, de uma maneira geral, uma contribuição positiva do conjunto estrutural. Dado o alto score do ponto 5, e que ele se localiza dentro da zona prospectiva, a qual está associada, em grande parte, ao Complexo Lagoa Real e à Formação Rio dos Remédios, considera-se possível que este ponto esteja em uma ocorrência não mapeada de uma dessas unidades litológicas, com o potencial para a existência de albititos. O ponto 6, embora se localize fora da zona do *scac81_rec*, pode também estar em um local de ocorrência não mapeada do Complexo Lagoa Real ou da Formação Rio dos Remédios, porém ressalta-se aqui a possibilidade de uma ocorrência não aflorante, uma vez que as variáveis relacionadas às feições de subsuperfície apresentaram contribuições positivas.

A análise de valores SHAP do modelo preditivo, através do gráfico de dependência, permite a localização de intervalos prospectivos para os valores das variáveis preditivas (Figura 3.33). Na análise realizada com tório, observou-se que valores menores do que cerca de 18,5 ppm apresentaram contribuições positivas no score prospectivo do modelo, o que define um limite de favorabilidade para essa variável.

A análise da influência do conjunto de treinamento no resultado do modelo XGB+SMOTE mostrou que o mesmo exerce uma influência significativa. No entanto, observa-se que as regiões com os maiores scores de prospectividade, que apresentaram valores mais altos no mapa dos mínimos (Figura 3.35), são as que apresentam menor CV. Essa tendência é visível também no diagrama de dispersão das médias vs. CVs (Figura 3.37), no qual observa-se uma tendência de queda diminuição dos CVs à medida em que as médias aumentam, que se torna mais

acentuada próximo do 95º percentil em diante. Essa constatação é relevante, uma vez que os scores mais altos são os mais importantes.

5. Conclusões

As análises realizadas neste trabalho demonstraram a utilidade de técnicas de machine learning, tanto não-supervisionadas, como supervisionadas, no mapeamento de prospectividade mineral em escala de distrito, com particular relevância para uso em áreas de estudo com mapeamento geológico limitado, como a PULR.

- O SOM se mostrou uma importante ferramenta para a análise exploratória inicial do banco de dados geoespaciais. Com o SOM, foi possível gerar visualizações que permitem o levantamento de *insights* sobre relações entre as variáveis e as características geológicas da área. Seu uso também permitiu a identificação de assinaturas típicas de unidades geológicas e das mineralizações de urânio na área de estudo. Apesar de mostrar importante variabilidade dos resultados devido ao componente de aleatoriedade do algoritmo, a geração de múltiplos modelos viabiliza uma análise quantitativa mais robusta.
- Os resultados da aplicação das técnicas de agrupamento SOM + K-means e AHER demonstraram que ambos possuem potencial para delimitar áreas de interesse, com potencial ganho prospectivo em relação às unidades geológicas da cartografia disponível.
- A técnica AHER produziu resultados mais coerentes geologicamente em comparação com a combinação de SOM + K-means, que não considera a posição espacial dos dados. A seleção do número de grupos (k) a partir da otimização dos 3 parâmetros de prospectividade (valor-t, área e número de ocorrências) do melhor grupo, mostrou-se uma alternativa eficiente para a definição de domínios de interesse prospectivo.
- O uso de XGB+SMOTE foi capaz de gerar um modelo preditivo com altos scores de desempenho. A técnica SHAP permitiu a verificação da relevância de cada variável preditiva do modelo como um todo, além da explicação do resultado do modelo para pontos específicos. Dessa forma, foi possível verificar quais fatores geológicos foram mais relevantes em diferentes áreas do mapa preditivo, o que, em combinação com o

conhecimento geológico do sistema mineral da PULR, permitiu a elaboração de hipóteses prospectivas para a geração de novos alvos exploratórios.

A abordagem adotada neste trabalho para a PULR pode servir como um exemplo para outras regiões, contribuindo para a exploração mineral de forma mais direcionada e informada. A utilização de dados públicos, em combinação com técnicas de ML, pode ser uma estratégia eficiente para a priorização de áreas para prospecção mineral, maximizando o uso dos recursos disponíveis e aumentando as chances de sucesso na identificação de novos depósitos minerais.

6. Referências

- Aggarwal C. C., Reddy C. K. (orgs.) 2016. *Data Clustering: Algorithms and Applications*. New York, Chapman and Hall/CRC, 652 p. doi: 10.1201/9781315373515.
- Agterberg F. P. 1989. Computer Programs for Mineral Exploration. *Science*, **245**(4913):76–81. doi: 10.1126/science.245.4913.76.
- Agterberg F. P. 2013. Fractals and spatial statistics of point patterns. *Journal of Earth Science*, **24**(1):1–11. doi: 10.1007/s12583-013-0305-6.
- Almeida F. F. M. 1977. O Craton Do São Francisco. *Revista Brasileira de Geociências*, **7**(1):349–364.
- Alpaydin E. 2014. *Introduction to machine learning*. (Third edition) Cambridge, Massachusetts, The MIT Press, 613 p.
- Arthur D., Vassilvitskii S. 2007. K-means++: the advantages of careful seeding. Em: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. USA, Society for Industrial and Applied Mathematics, p. 1027–1035.
- Assunção R. M., Neves M. C., Câmara G., Da Costa Freitas C. 2006. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, **20**(7):797–811. doi: 10.1080/13658810600665111.
- Baço F., Lobo V., Painho M. 2005. Self-organizing Maps as Substitutes for K-means Clustering. Em: Sunderam V.S., van Albada G.D., Sloot P.M.A., Dongarra J. (orgs.). *Computational Science – ICCS 2005*. Berlin, Heidelberg, Springer Berlin Heidelberg, p. 476–483. doi: 10.1007/11428862_65.
- Baddeley A., Rubak E., Turner R. 2015. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press, 828 p.
- Barros R. A., Silva G. F., Santos R. S. V., Lopes C. G. 2021. Interpretação Geofísica-Estrutural da Área do Vale do Paramirim. Relatório Interno. Salvador-BA, SGB-CPRM. 32 p.
- Bonham-Carter G. 1994. *Geographic information systems for geoscientists: modelling with GIS*. (1st ed.) Oxford; New York, Pergamon, 398 p.
- Box G. E. P., Cox D. R. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, **26**(2):211–243. doi: 10.1111/j.2517-6161.1964.tb00553.x.

- Brown G. 2017. Ensemble Learning. Em: Sammut C., Webb G.I. (orgs.). *Encyclopedia of Machine Learning and Data Mining*. Boston, MA, Springer US, p. 393–402.
- Brown W. M., Gedeon T. D., Groves D. I. 2003. Use of noise to augment training data: A neural network method of mineral-potential mapping in regions of limited known deposit examples. *Natural Resources Research*, **12**(2):141–152. doi: 10.1023/A:1024218913435.
- Bruneton P., Cuney M. 2016. Geology of uranium deposits. Em: *Uranium for Nuclear Power*. Elsevier, p. 11–52. doi: 10.1016/B978-0-08-100307-7.00002-8.
- Carlson C. A. 1991. Spatial distribution of ore deposits. *Geology*, **19**(2):111. doi: 10.1130/0091-7613(1991)019<0111:SDOOD>2.3.CO;2.
- Carneiro C. de C., Fraser S. J., Crósta A. P., Silva A. M., Barros C. E. de M. 2012. Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the Brazilian Amazon. *GEOPHYSICS*, **77**(4):K17–K24. doi: 10.1190/geo2011-0302.1.
- Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, **16**:321–357. doi: 10.1613/jair.953.
- Chen T., Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. Em: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. p. 785–794. doi: 10.1145/2939672.2939785.
- Cheng Q. 2008. Non-Linear Theory and Power-Law Models for Information Integration and Mineral Resources Quantitative Assessments. *Mathematical Geosciences*, **40**(5):503–532. doi: 10.1007/s11004-008-9172-6.
- Cheng Q., Agterberg F. P., Ballantyne S. B. 1994. The separation of geochemical anomalies from background by fractal methods. *Journal of Geochemical Exploration*, **51**(2):109–130. doi: 10.1016/0375-6742(94)90013-2.
- Cheng Q., Agterberg F. P., Bonham-Carter G. F. 1996. Fractal pattern integration for mineral potential estimation. *Nonrenewable Resources*, **5**(2):117–130. doi: 10.1007/BF02257585.

- Chudasama B., Torppa J., Nykänen V., Kinnunen J. 2022. Target-scale prospectivity modeling for gold mineralization within the Rajapalot Au-Co project area in northern Fennoscandian Shield, Finland. Part 2: Application of self-organizing maps and artificial neural networks for exploration targeting. *Ore Geology Reviews*, **147**:104936. doi: 10.1016/j.oregeorev.2022.104936.
- Cordani U. G., Iyer S. S., Taylor P. N., Kawashita K., Kei S., McReath I. 1992. Pb-Pb, Rb-Sr, and K-Ar systematics of the Lagoa Real uranium province (south-central Bahia, Brazil) and the Espinhaço cycle (ca. 1.5-1.0 Ga). *Journal of South American Earth Sciences*, **5**(1):33–46. doi: 10.1016/0895-9811(92)90058-7.
- Cruz S. C. P., Alkmim F. F. 2017. The Paramirim Aulacogen. Em: Heilbron M., Cordani U.G., Alkmim F.F. (orgs.). *São Francisco Craton, Eastern Brazil*. Cham, Springer International Publishing, p. 97–115. doi: 10.1007/978-3-319-01715-0_6.
- Cruz S. C. P., Alkmim F. F. 2006. The Tectonic interaction between the Paramirim Aulacogen and the Araçuaí Belt, São Francisco craton region, Eastern Brazil. *Anais da Academia Brasileira de Ciências*, **78**(1):151–173. doi: 10.1590/S0001-37652006000100014.
- Cruz S. C. P., Alkmim F. F., Leite C. M. M., Jordt-Evangelista H., Cunha J. C., Matos E. C., Noce C. M., Marinho M. M. 2007. Geologia e arcabouço estrutural do Complexo Lagoa Real, vale do Paramirim, Centro-Oeste da Bahia. *Revista Brasileira de Geociências*, **37**(4S):128–146. doi: 10.25249/0375-7536.200737s4128146.
- Cruz S. C. P., Miranda L. L. F., Veiga P. M. de O. 2008. *Província uranífera de Lagoa Real, Bahia*. Salvador, Companhia Baiana de Pesquisa Mineral-CBPM, 68 p.
- Cuney M., Emetz A., Mercadier J., Mykchaylov V., Shunko V., Yuslenko A. 2012. Uranium deposits associated with Na-metasomatism from central Ukraine: A review of some of the major deposits and genetic constraints. *Ore Geology Reviews*, **44**:82–106. doi: 10.1016/j.oregeorev.2011.09.007.
- CBPM/CPRM. 2006. *Projeto Aerogeofísico Campo Alegre de Lourdes Mortugaba*: relatório final do levantamento e processamento dos dados magnetométricos e gamaespectrométricos. Rio de Janeiro: CBPM; CPRM Texto e anexos (mapas).
- Dahlkamp F. J. 2009. *Uranium Deposits of the World: Asia*. Berlin Heidelberg, Springer-Verlag, 508 p.
- Dal Pozzolo A., Caelen O., Le Borgne Y.-A., Waterschoot S., Bontempi G. 2014. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, **41**(10):4915–4928. doi: 10.1016/j.eswa.2014.02.026.

- Danderfer A., De Waele B., Pedreira A. J., Nalini H. A. 2009. New geochronological constraints on the geological evolution of Espinhaço basin within the São Francisco Craton—Brazil. *Precambrian Research*, **170**(1–2):116–128. doi: 10.1016/j.precamres.2009.01.002.
- Evjen H. M. 1936. The place of the vertical gradient in gravitational interpretations. *GEOPHYSICS*, **1**(1):127–136. doi: 10.1190/1.1437067.
- Fan M., Xiao K., Sun L., Zhang S., Xu Y. 2022. Automated Hyperparameter Optimization of Gradient Boosting Decision Tree Approach for Gold Mineral Prospectivity Mapping in the Xiong’ershan Area. *Minerals*, **12**(12):1621. doi: 10.3390/min12121621.
- Gazley M., Hood S. B., Cracknell M. J. 2021. Soil-sample geochemistry normalised by class membership from machine-learned clusters of satellite and geophysics data. *Ore Geology Reviews*, **139**:104442. doi: 10.1016/j.oregeorev.2021.104442.
- Guimarães J.T.; Santos R.A.S.; Melo R.C. (orgs.). 2008. Geologia da Chapada Diamantina Ocidental. In: Guimarães J.T., Santos R.A., Melo R.C. *Projeto Ibitiara – Rio de Contas*. Companhia Baiana de Pesquisa Mineral-CBPM, Salvador, Arquivos Abertos 31, 64 p.
- Guo D. 2008. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, **22**(7):801–823. doi: 10.1080/13658810701674970.
- Hagemann S. G., Lisitsin V. A., Huston D. L. 2016. Mineral system analysis: Quo vadis. *Ore Geology Reviews*, **76**:504–522. doi: 10.1016/j.oregeorev.2015.12.012.
- Han H., Wang W.-Y., Mao B.-H. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Em: Huang D.-S., Zhang X.-P., Huang G.-B. (orgs.). *Advances in Intelligent Computing*. Berlin, Heidelberg, Springer Berlin Heidelberg, p. 878–887.
- Hariharan S., Tirodkar S., Porwal A., Bhattacharya A., Joly A. 2017. Random Forest-Based Prospectivity Modelling of Greenfield Terrains Using Sparse Deposit Data: An Example from the Tanami Region, Western Australia. *Natural Resources Research*, **26**(4):489–507. doi: 10.1007/s11053-017-9335-6.
- Hastie T., Tibshirani R., Friedman J. 2009. *The Elements of Statistical Learning*. New York, NY, Springer New York, 745 p.
- Henderson R. G., Zietz I. 1949. The upward continuation of anomalies in total magnetic intensity fields. *GEOPHYSICS*, **14**(4):517–534. doi: 10.1190/1.1437560.

- Hronsky J. M. A. 2009. Self-Organized Critical Systems and Ore Formation: The Key to Spatial Targeting?
- International Atomic Energy Agency (IAEA) 2003. *Guidelines for Radioelement Mapping Using Gamma Ray Spectrometric Data*. Vienna, Austria, International Atomic Energy Agency, 173 p.
- International Atomic Energy Agency (IAEA). 2020. *Descriptive Uranium Deposit and Mineral System Models*. Vienna, Austria, International Atomic Energy Agency, 328 p.
- Indústrias Nucleares do Brasil (INB). 2024. Produção. Disponível em: <<https://www.inb.gov.br/pt-br/Nossas-Atividades/Ur%C3%A2nio/Produ%C3%A7%C3%A3o>>. Acesso em: 20 mai. 2024.
- Illian J., Penttinen A., Stoyan H., Stoyan D. 2007. *Statistical Analysis and Modelling of Spatial Point Patterns*. (1. Edição) John Wiley & Sons,. doi: 10.1002/9780470725160.
- Kaski S. 2017. Self-Organizing Maps. Em: Sammut C., Webb G.I. (orgs.). *Encyclopedia of Machine Learning and Data Mining*. Boston, MA, Springer US, p. 1129–1132.
- Kastens K. A., Manduca C. A., Cervato C., Frodeman R., Goodwin C., Liben L. S., Mogk D. W., Spangler T. C., Stillings N. A., Titus S. 2009. How Geoscientists Think and Learn. *Eos, Transactions American Geophysical Union*, **90**(31):265–266. doi: 10.1029/2009EO310001.
- Kaufman L., Rousseeuw P. J. 2005. *Finding groups in data: an introduction to cluster analysis*. Hoboken, N.J, John Wiley & Sons, 342 p.
- Keykhay-Hosseinpour M., Kohsary A. H., Hossein-Morshedy A., Porwal A. 2020. A machine learning-based approach to exploration targeting of porphyry Cu-Au deposits in the Dehsalm district, eastern Iran. *Ore Geology Reviews*, **116**(November). doi: 10.1016/j.oregeorev.2019.103234.
- Knox-Robinson C. M., Wyborn L. A. I. 1997. Towards a holistic exploration strategy: Using Geographic Information Systems as a tool to enhance exploration. *Australian Journal of Earth Sciences*, **44**(4):453–463. doi: 10.1080/08120099708728326.
- Kohonen T. 2013. Essentials of the self-organizing map. *Neural Networks*, **37**:52–65. doi: 10.1016/j.neunet.2012.09.018.
- Kohonen T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, **43**(1):59–69. doi: 10.1007/BF00337288.
- Kohonen T. 2001. *Self-organizing maps*. (3rd ed.) Heidelberg, Springer Berlin, 502 p.

- Kreuzer O. P., Yousefi M., Nykänen V. 2020. Introduction to the special issue on spatial modelling and analysis of ore-forming processes in mineral exploration targeting. *Ore Geology Reviews*, **119**:103391. doi: 10.1016/j.oregeorev.2020.103391.
- Kubat M. 2015. *An Introduction to Machine Learning*. Cham, Springer International Publishing, 291 p. doi: 10.1007/978-3-319-20010-1.
- Kuhn S., Cracknell M. J., Reading A. M. 2019. Lithological mapping in the Central African Copper Belt using Random Forests and clustering: Strategies for optimised results. *Ore Geology Reviews*, **112**:103015. doi: 10.1016/j.oregeorev.2019.103015.
- Lachaud A., Adam M., Mišković I. 2023. Comparative Study of Random Forest and Support Vector Machine Algorithms in Mineral Prospectivity Mapping with Limited Training Data. *Minerals*, **13**(8):1073. doi: 10.3390/min13081073.
- Lattimer B., Lattimer A. 2022. Creating Compact Regions of Social Determinants of Health. doi: 10.48550/arXiv.2209.11836.
- Leite E. P., Souza Filho C. R. de 2010. Mapas auto-organizáveis aplicados ao mapeamento do potencial mineral na região de serra leste, província mineral de Carajás, Pará. *Revista Brasileira de Geofísica*, **28**(3):397–409. doi: 10.1590/S0102-261X2010000300006.
- Lemaître G., Nogueira F., Aridas C. K. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, **18**(17):1–5.
- Li T., Xia Q., Zhao M., Gui Z., Leng S. 2020. Prospectivity Mapping for Tungsten Polymetallic Mineral Resources, Nanling Metallogenic Belt, South China: Use of Random Forest Algorithm from a Perspective of Data Imbalance. *Natural Resources Research*, **29**(1):203–227. doi: 10.1007/s11053-019-09564-8.
- Lobato L. M., Fyfe W. S. 1990. Metamorphism, metasomatism, and mineralization at Lagoa Real, Bahia, Brazil. *Economic Geology*, **85**(5):968–989. doi: 10.2113/gsecongeo.85.5.968.
- Lundberg S. M., Lee S. 2017. A Unified Approach to Interpreting Model Predictions. Em: Guyon I., Luxburg U.V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R. (orgs.). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., p. 4765–4774.

- Magalhães L. A., Souza Filho C. R. 2012. Targeting of Gold Deposits in Amazonian Exploration Frontiers using Knowledge- and Data-Driven Spatial Modeling of Geophysical, Geochemical, and Geological Data. *Surveys in Geophysics*, **33**(2):211–241. doi: 10.1007/s10712-011-9151-1.
- Mandelbrot B. B. 1983. *The fractal geometry of nature*. (Updated and augmented) New York, W.H. Freeman,.
- Marques C., Cabral A. R., Rios F. J. 2021. Whole-rock chemistry of the Gameleira I uranium deposit, Lagoa Real, Brazil. *Geochemistry*, **81**(1):125677. doi: 10.1016/j.chemer.2020.125677.
- Maruéjol P., Cuney M., Fuzikawa K., Netto A. M., Poty B. 1987. The Lagoa Real Subalkaline Granitic Complex (South Bahia, Brazil): A Source for Uranium Mineralizations Associated with Na-Ca Metasomatism. *Revista Brasileira de Geociências*, **17**(4):578–594.
- McCuaig T. C., Beresford S., Hronsky J. 2010. Translating the mineral systems approach into an effective exploration targeting system. *Ore Geology Reviews*, **38**(3):128–138. doi: 10.1016/j.oregeorev.2010.05.008.
- McCuaig T. C., Hronsky J. M. A. 2014. The Mineral System Concept The Key to Exploration Targeting. Em: *Building Exploration Capability for the 21st Century*. Society of Economic Geologists,. doi: 10.5382/SP.18.08.
- Mello R. F., Ponti M. A. 2018. *Machine Learning: A Practical Approach on the Statistical Learning Theory*. Cham, Springer International Publishing,. doi: 10.1007/978-3-319-94989-5.
- Melssen W., Wehrens R., Buydens L. 2006. Supervised Kohonen networks for classification problems. *Chemometrics and Intelligent Laboratory Systems*, **83**(2):99–113. doi: 10.1016/j.chemolab.2006.02.003.
- Moreira J., Costa L. D. F. 1996. Neural-based color image segmentation and classification using self-organizing maps. Em: *Anais do IX SIBGRAPI*. Caxambú, MG p. 47–54.
- Murtagh F. 1985. A Survey of Algorithms for Contiguity-constrained Clustering and Related Problems. *The Computer Journal*, **28**(1):82–88. doi: 10.1093/comjnl/28.1.82.
- Nuclear Energy Agency (NEA) / International Atomic Energy Agency (IAEA). 2023. *Uranium 2022: Resources, Production and Demand*. OECD Publishing, Paris. doi: 10.1787/2c4e111b-en.

- Okada K. 2020. A Historical Overview of the Past Three Decades of Mineral Exploration Technology. *Natural Resources Research*. doi: 10.1007/s11053-020-09721-4.
- Openshaw S. 1977. A Geographical Solution to Scale and Aggregation Problems in Region-Building, Partitioning and Spatial Modelling. *Transactions of the Institute of British Geographers*, **2**(4):459–472. doi: 10.2307/622300.
- Parsa M. 2021. A data augmentation approach to XGboost-based mineral potential mapping: An example of carbonate-hosted Zn Pb mineral systems of Western Iran. *Journal of Geochemical Exploration*, **228**:106811. doi: 10.1016/j.gexplo.2021.106811.
- Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É. 2011. *Scikit-learn*: Machine Learning in Python. *Journal of Machine Learning Research*, **12**(85):2825–2830.
- Pedrosa-Soares A. C., Noce C. M., Wiedemann C. M., Pinto C. P. 2001. The Araçuaí-West-Congo Orogen in Brazil: An overview of a confined orogen formed during Gondwanaland assembly. *Precambrian Research*, **110**(1–4):307–323. doi: 10.1016/S0301-9268(01)00174-7.
- Pires F. R. M. 2012. *Urânio no Brasil: geologia, jazidas e ocorrências*. Rio de Janeiro, Vitruva, 264 p.
- Ponmalai R., Kamath C. 2019. *Self-Organizing Maps and Their Applications to Data Analysis*. United States, Lawrence Livermore National Laboratory (LLNL). doi: 10.2172/1566795.
- Porwal A., Carranza E. J. M. 2015. Introduction to the Special Issue: GIS-based mineral potential modelling and geological data analyses for mineral exploration. *Ore Geology Reviews*, **71**:477–483. doi: 10.1016/j.oregeorev.2015.04.017.
- Prado E. M. G., Souza Filho C. R., Carranza E. J. M., Motta J. G. 2020. Modeling of Cu-Au prospectivity in the Carajás mineral province (Brazil) through machine learning: Dealing with imbalanced training data. *Ore Geology Reviews*, **124**(June):103611. doi: 10.1016/j.oregeorev.2020.103611.
- Ripley B. D. 1988. *Statistical Inference for Spatial Processes*. Cambridge University Press,. doi: 10.1017/CBO9780511624131.

- Roberts S., Sanderson D. J., Gumiel P. 1999. Fractal analysis and percolation properties of veins. *Geological Society, London, Special Publications*, **155**(1):7–16. doi: 10.1144/GSL.SP.1999.155.01.03.
- Rodriguez-Galiano V. F., Sanchez-Castillo M., Chica-Olmo M., Chica-Rivas M. 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*, **71**:804–818. doi: 10.1016/j.oregeorev.2015.01.001.
- Russo S., Disch A., Blumensaat F., Villez K. 2020. Anomaly Detection using Deep Autoencoders for in-situ Wastewater Systems Monitoring Data. *ArXiv*, **abs/2002.0**.
- Sanderson D. J., Roberts S., Gumiel P. 1994. A fractal relationship between vein thickness and gold grade in drill core from La Codosera, Spain. *Economic Geology*, **89**(1):168–173. doi: 10.2113/gsecongeo.89.1.168.
- Souza, J. D.; Melo, R. C.; Kosin, M. Mapa geológico do estado da Bahia. Versão 1.1. Salvador: CPRM, 2003. Escala 1:1.000.000.
- Shapley L. S. 1953. 17. A Value for n-Person Games. Em: Kuhn H.W., Tucker A.W. (orgs.). *Contributions to the Theory of Games (AM-28), Volume II*. Princeton, Princeton University Press, p. 307–318. doi: doi:10.1515/9781400881970-018.
- Stoyan D., Stoyan H. 1996. Estimating Pair Correlation Functions of Planar Cluster Processes. *Biometrical Journal*, **38**(3):259–271. doi: 10.1002/bimj.4710380302.
- Torppa J., Nykänen V., Molnár F. 2019. Unsupervised clustering and empirical fuzzy memberships for mineral prospectivity modelling. *Ore Geology Reviews*, **107**:58–71. doi: 10.1016/j.oregeorev.2019.02.007.
- Turcotte D. L. 2002. Fractals in petrology. *Lithos*, **65**(3–4):261–271. doi: 10.1016/S0024-4937(02)00194-9.
- Turpin L., Maruejol P., Cuney M. 1988. U-Pb, Rb-Sr and Sm-Nd chronology of granitic basement, hydrothermal albitites and uranium mineralization (Lagoa Real, South-Bahia, Brazil). *Contributions to Mineralogy and Petrology*, **98**(2):139–147. doi: 10.1007/BF00402107.
- Utsch A., Siemon H. P. 1990. Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. Em: Widrow B., Angeniol B. (orgs.). *Proceedings of the International Neural Network Conference (INNC-90), Paris, France, July 9–13, 1990 1. Dordrecht, Netherlands*. Dordrecht, Netherlands, Kluwer Academic Press, p. 305–308.

- Vesanto J. 2002. Data Exploration Process Based on the Self-Organizing Map. *Acta Polytechnica Scandinavica*, (115):96.
- Vesanto J. 1997. Data Mining Techniques Based on the Self-Organizing Map. Helsinki University of Technology,.
- Vesanto J., Alhoniemi E. 2000. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, **11**(3):586–600. doi: 10.1109/72.846731.
- Wang G., Carranza E. J. M., Zuo R., Hao Y., Du Y., Pang Z., Sun Y., Qu J. 2012. Mapping of district-scale potential targets using fractal models. *Journal of Geochemical Exploration*, **122**:34–46. doi: 10.1016/j.gexplo.2012.06.013.
- Wang Z., Zuo R., Zhang Z. 2015. Spatial analysis of Fe deposits in Fujian Province, China: Implications for mineral exploration. *Journal of Earth Science*, **26**(6):813–820. doi: 10.1007/s12583-015-0597-9.
- Wehrens R., Buydens L. M. C. 2007. Self- and Super-organizing Maps in R: The kohonen Package. *Journal of Statistical Software*, **21**(5):1–19. doi: 10.18637/jss.v021.i05.
- Wehrens R., Kruisselbrink J. 2018. Flexible Self-Organizing Maps in **kohonen** 3.0. *Journal of Statistical Software*, **87**(7). doi: 10.18637/jss.v087.i07.
- Wei R., Rey S., Knaap E. 2021. Efficient regionalization for spatially explicit neighborhood delineation. *International Journal of Geographical Information Science*, **35**(1):135–151. doi: 10.1080/13658816.2020.1759806.
- Wilde A. 2013. Towards a Model for Albitite-Type Uranium. *Minerals*, **3**(1):36–48. doi: 10.3390/min3010036.
- Wyborn L. A. I., Heinrich C. A., Jaques A. L. 1994. Australian Proterozoic mineral systems: essential ingredients and mappable criteria. *Australian Institute of Mining and Metallurgy Publication Series*, (94):109–115.
- Yin H., Allinson N. M. 1995. On the Distribution and Convergence of Feature Space in Self-Organizing Maps. *Neural Computation*, **7**(6):1178–1187. doi: 10.1162/neco.1995.7.6.1178.
- Yin J., Li N. 2022. Ensemble learning models with a Bayesian optimization algorithm for mineral prospectivity mapping. *Ore Geology Reviews*, **145**:104916. doi: 10.1016/j.oregeorev.2022.104916.
- Zaki M. J., Meira Jr. W. 2020. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. (2nd Edição) Cambridge University Press, 776 p. doi: 10.1017/9781108564175.

- Zuo R. 2016. A nonlinear controlling function of geological features on magmatic–hydrothermal mineralization. *Scientific Reports*, **6**(1):27127. doi: 10.1038/srep27127.
- Zuo R., Carranza E. J. M. 2023. Machine Learning-Based Mapping for Mineral Exploration. *Mathematical Geosciences*, **55**(7):891–895. doi: 10.1007/s11004-023-10097-3.