BRUNO MONSERRAT PERILLO

**Imputação e incremento de resolução espacial em dados
geofísicos de bacias sedimentares *onshore*: abordagem a partir
da técnica *Self-Organizing Maps***

São Paulo
2019

BRUNO MONSERRAT PERILLO

**Imputação e incremento de resolução espacial em dados
geofísicos de bacias sedimentares *onshore*: abordagem a partir
da técnica *Self-Organizing Maps***

Dissertação apresentada à Escola
Politécnica da Universidade de São Paulo
para obtenção do título de Mestre em
Ciências.

Área de concentração:
Engenharia Mineral

Orientador:
Prof. Dr. Cleyton de Carvalho Carneiro

Coorientadora:
Dra. Ana Cristina Cancoro de Matos

São Paulo
2019

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 01 de outubro de 2019

Assinatura do autor: Bruno M. Perillo

Assinatura do orientador: Phyllis de C. Carneiro

Catálogo-na-publicação

Perillo, Bruno Monserrat

Imputação e incremento de resolução espacial em dados geofísicos de bacias sedimentares onshore: abordagem a partir da técnica Self-Organizing Maps / B. M. Perillo -- versão corr. -- São Paulo, 2019.
117 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Minas e Petróleo.

1. Self-organizing maps 2. Imputação 3. Incremento de resolução 4. Geofísica I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Minas e Petróleo II. t.

Dissertação de autoria de Bruno Monserrat Perillo, sob o título " **Imputação e incremento de resolução espacial em dados geofísicos de bacias sedimentares onshore: abordagem a partir da técnica Self-Organizing Maps** ", apresentada à Escola Politécnica da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Engenharia Mineral. Aprovado em _____ de _____ de _____ pela comissão julgadora constituída pelos doutores:

_____
Prof. Dr. _____

Instituição: _____

_____
Prof. Dr. _____

Instituição: _____

_____
Prof. Dr. _____

Instituição: _____

Aos meus pais Neusa (in memoriam) e Ronaldo
e a minha filha Laurinha

AGRADECIMENTOS

Ao meu orientador Cleyton de Carvalho Carneiro que neste período sempre esteve disposto a me auxiliar na construção do conhecimento, fazendo o meu aprendizado neste período de pós-graduação mais efetivo e por me apresentar ao estimado mundo dos *self-organizing maps*. À minha coorientadora Ana Cristina Oliveira Cancoro de Matos, também, por sua paciência e disponibilidade em todos momentos de apuros.

Agradeço à Universidade de São Paulo (USP), à Escola Politécnica (EP) e ao Departamento de Engenharia de Minas e Petróleo, pela estrutura. À Fundação de Apoio à Universidade de São Paulo (FUSP), projeto 779, que financiou a minha bolsa de estudos e, com isso, proporcionou o desenvolvimento desta pesquisa. Ao Laboratório de Topografia e Geodesia (LTG), em especial ao professor Denizar Blitzkow por me recebê-lo em seu laboratório e pela concessão dos dados.

Aos professores Ronaldo Carrion, Carina Ulsen, Márcio Sampaio e Patricia Matai por todo suporte e auxílio. Meu muito obrigado ao Eder Molina, Rafael Gioria e Marcelo Guarido que avaliaram diretamente meu trabalho durante a qualificação e defesa, por todas as críticas e sugestões para tornar o desenvolvimento desta pesquisa mais consistente.

Muito obrigado também aos profissionais desta escola: Rosângela Tavares, Marcia Pereira, Eliana Rodrigues, Nice Maciel, Jucelene Fernandes, Olga Mendonça e Neuci Bicov. Sem vocês tudo teria sido mais difícil.

Por fim, àqueles que são fundamentais. Ao meu pai, por me dar o bem maior da humanidade: a educação. Além de me doar o seu tempo, carinho e compreensão nos momentos mais difíceis. À minha mãe, que sempre me mostrou o caminho doce da vida, que sempre me deu amor, que sempre sacrificou o seu próprio tempo e as noites que poderiam ter sido apenas de sono e sonhos para cuidar de mim. Aos meus irmãos: Fábio e Renzo – por serem os meus maiores amigos e por estarem sempre ao meu lado. À minha cunhada e sobrinhos: Cláudia e Bianca e Lorenzo, por me fazer mais feliz. À minha Laurinha, por toda minha ausência só posso dedicar este trabalho a você!

RESUMO

A técnica Mapas Auto-Organizáveis, em inglês *Self-Organizing Maps* (SOM) têm sido amplamente utilizada em segmentação e agrupamento de banco de dados complexos e multivariados, permitindo um melhor entendimento de relações não lineares entre as variáveis. Tal técnica computacional possibilita identificar assinaturas dentro dos mais diversos conjuntos de dados, sejam estes compostos por variáveis categóricas ou contínuas. Diversos autores têm abordado estudos de caso envolvendo o uso de SOM para criar valores sintéticos em regiões com dados faltantes. A escassez de trabalhos sobre incremento de resolução utilizando dados multivariados demonstram as dificuldades encontradas e a necessidade de mais estudos. Este trabalho visa imputar e incrementar resolução espacial em dados, a partir da técnica SOM possibilitando a obtenção de dados com maior resolução a baixo custo, principalmente para dados gravimétricos, cuja aquisição é menos comum e demanda rigor analítico e processual. A primeira abordagem está relacionada à imputação de dados em locais de difícil acesso, tais como a região Amazônica, onde foram utilizados dados de gravimetria terrestre e aérea, distúrbio da gravidade e o modelo GEMMA, que é uma estimativa da espessura da crosta. Os diferentes tamanhos de mapas aplicados nas análises por SOM mostraram alta correlação em valores absolutos entre os dados sintéticos e os dados originais. Em uma segunda abordagem, voltada ao incremento de resolução espacial, foram utilizados dados aerogamaespectrométricos compostos por células originalmente com 125m de resolução espacial. Para fins de comparações a variável tório foi degradada para células com resoluções espaciais de 500 e 1000m. As análises SOM foram desenvolvidas com todo o universo de variáveis gamaespectrométricas incluindo, além da variável tório degradada, os dados de contagem total, urânio e potássio. Os resultados mostraram o incremento da resolução do tório de 500m e 1000m, para a resolução de 125m, que apresentaram, respectivamente, $R^2 = 0,98$ e $0,94$. As análises desenvolvidas demonstram validade à imputação e ao incremento de resolução pela técnica SOM. Os resultados sugerem possibilidades de desenvolvimento de incremento de resolução espacial de variáveis com baixa resolução em áreas compostas por outras variáveis de maior resolução espacial.

Palavra-chave: Mapas Auto-Organizáveis. Imputação. Incremento de resolução. Geofísica. Valores ausentes.

ABSTRACT

The Self-Organizing Maps (SOM) technique has been widely used for segmentation and clustering of complex and multivariate databases, allowing a better understanding of nonlinear relationships between variables. Such computational technique makes it possible to identify signatures within the most diverse data sets, whether these are categorical or continuous variables. Several authors have approached case studies involving the use of SOM to create synthetic values in regions with missing data. The lack of work on resolution increment using multivariate data demonstrates the difficulties encountered and the need for further studies. This work aims to impute and increase spatial resolution in data, using the SOM technique, enabling the obtaining of higher resolution data at low cost, especially for gravimetric data, whose acquisition is less common and demands analytical and procedural rigor. The first approach is related to data imputation in hard to reach places, such as the Amazon region, where terrestrial and air gravity data, gravity disturbance and the GEMMA model, which is an estimate of crust thickness, were used. The different map sizes applied in the SOM analyzes showed a high correlation in absolute values between the synthetic data and the original data. In a second approach, focused on the increment of spatial resolution, aerogamapectrometric data composed by cells originally with 125m of spatial resolution were used. For comparison purposes the thorium variable was degraded for cells with spatial resolutions of 500 and 1000m. The SOM analyzes were developed with the whole universe of gamma-spectrometric variables including, besides the degraded thorium variable, the total count, uranium and potassium data. The results showed an increase in the thorium resolution of 500m and 1000m, for the resolution of 125m, which presented, respectively, $R^2 = 0,98$ e $0,94$. The developed analyzes demonstrate validity to the imputation and the increment of resolution by the SOM technique. The results suggest possibilities of developing spatial resolution increment of low-resolution variables in areas composed by other variables of higher spatial resolution.

Keywords: Self-Organizing Maps. Imputation. Increment of resolution. Geophysics. Missing values.

LISTA DE ILUSTRAÇÕES

FIGURA 1 – LOCALIZAÇÃO DA ÁREA DE ESTUDO VISANDO IMPUTAÇÃO DE DADOS.	18
FIGURA 2 – LOCALIZAÇÃO DA ÁREA DE ESTUDO. POLÍGONO VERMELHO REPRESENTA OS LIMITES DO AEROLEVANTAMENTO 1057 – CPRM.....	19
FIGURA 3 – O PROCESSO DO KDD	22
FIGURA 4 - ESQUEMA DE UM AEROLEVANTAMENTO	24
FIGURA 5 – CUSTOS RELATIVOS APROXIMADOS DE DIFERENTES TIPOS DE LEVANTAMENTOS GEOFÍSICOS EM TERMOS DE ÁREA.	25
FIGURA 6 – ALGUMAS CONFIGURAÇÕES DE AQUISIÇÃO DE DADOS COMUNS EM PESQUISAS GEOFÍSICAS. OS PONTOS REPRESENTAM A AMOSTRAGEM DOS DADOS. (A), TRANSVERSAS REGULARES; (B) TRANSVERSAS IRREGULARES; (C) MALHA QUADRADA; E (D) RANDÔMICA.....	26
FIGURA 8 – SUPERFÍCIES DE REFERÊNCIAS.....	27
FIGURA 9 - PERDA DE INFORMAÇÃO DEVIDO AO LIMITADO NÚMERO DE PONTOS DE MEDIÇÃO.	31
FIGURA 10 – EXEMPLO FALSEAMENTO ESPACIAL EM UM LEVANTAMENTO AEROMAGNÉTICO HIPOTÉTICO. (A) DADOS ORIGINAIS, ESPAÇAMENTO DE LINHAS DE VOO DE 150 M, (B) 300 M E (C) 600 M.	31
FIGURA 11 - ESTRUTURA BÁSICA REDE SOM.....	34
FIGURA 12 – SÍNTESE DAS ETAPAS DE UMA ANÁLISE EM AMBIENTE DE SOM.....	39
FIGURA 13 – SÍNTESE DO PROCESSO DO SOM, APRESENTANDO DADOS DE ENTRADA COM DIFERENTES RESOLUÇÕES.	40
FIGURA 14 – IMAGEM ESPACIALIZADA REPRESENTANDO A AMOSTRAGEM DE UM DADO DE BAIXA RESOLUÇÃO (AZUL) E ALTA RESOLUÇÃO (VERMELHO) COM VALORES AUSENTES EM 3 DIFERENTES ÁREAS DE JANELA.	43
FIGURA 15 - AEROGRAVIMETRIA (AZUL) E GRAVIMETRIA TERRESTRE (VERMELHO) LOCALIZADA NA REGIÃO AMAZÔNICA.....	45
FIGURA 16 – PROCEDIMENTO PARA IMPUTAÇÃO DO SOM.....	49
FIGURA 17 - FLUXOGRAMA DA PREPARAÇÃO DA BASE DE DADOS E PROCESSAMENTO. A: INCREMENTO DE RESOLUÇÃO. B: IMPUTAÇÃO DE DADOS.	51
FIGURA 18 - ANÁLISE EXPLORATÓRIA DA ANOMALIA BOUGUER AÉREA. (A) ANOMALIA INTERPOLADA, (B) PONTOS AMOSTRADOS, (C) HISTOGRAMA E INFORMAÇÕES ESTATÍSTICAS.....	54
FIGURA 19 - ANÁLISE EXPLORATÓRIA DA ANOMALIA BOUGUER TERRESTRE. (A) ANOMALIA INTERPOLADA, (B) PONTOS AMOSTRADOS, (C) HISTOGRAMA E INFORMAÇÕES ESTATÍSTICAS.	55
FIGURA 20 - ANÁLISE EXPLORATÓRIA DA ANOMALIA AR LIVRE AÉREA. (A) ANOMALIA INTERPOLADA, (B) PONTOS AMOSTRADOS, (C) HISTOGRAMA E INFORMAÇÕES ESTATÍSTICAS.....	56
FIGURA 21 - ANÁLISE EXPLORATÓRIA DA ANOMALIA AR LIVRE TERRESTRE. (A) ANOMALIA INTERPOLADA, (B) PONTOS AMOSTRADOS, (C) HISTOGRAMA E INFORMAÇÕES ESTATÍSTICAS.....	57

FIGURA 22 - ANÁLISE EXPLORATÓRIA DO DISTÚRBIO DA GRAVIDADE. (A) ANOMALIA INTERPOLADA, (B) PONTOS AMOSTRADOS, (C) HISTOGRAMA E INFORMAÇÕES ESTATÍSTICAS.....	58
FIGURA 23 - ANÁLISE EXPLORATÓRIA DO MODELO DA MOHO. (A) ANOMALIA INTERPOLADA, (B) PONTOS AMOSTRADOS, (C) HISTOGRAMA E INFORMAÇÕES ESTATÍSTICAS.....	59
FIGURA 24 - ANÁLISE EXPLORATÓRIA DO MODELO DIGITAL DE TERRENO. (A) ANOMALIA INTERPOLADA, (B) PONTOS AMOSTRADOS, (C) HISTOGRAMA E INFORMAÇÕES ESTATÍSTICAS.....	60
FIGURA 25 - ANÁLISE EXPLORATÓRIA DA ALTURA GEOIDAL. (A) ANOMALIA INTERPOLADA, (B) PONTOS AMOSTRADOS, (C) HISTOGRAMA E INFORMAÇÕES ESTATÍSTICAS.	61
FIGURA 26 - ANÁLISE EXPLORATÓRIA DA ANOMALIA MAGNÉTICA DA TERRA. (A) ANOMALIA INTERPOLADA, (B) PONTOS AMOSTRADOS, (C) HISTOGRAMA E INFORMAÇÕES ESTATÍSTICAS.	62
FIGURA 27 - MATRIZ DE CORRELAÇÃO E GRÁFICOS DE DISPERSÃO E HISTOGRAMA DAS VARIÁVEIS ENVOLVIDAS NA ANÁLISE EXPLORATÓRIA.	63
FIGURA 28 - MATRIZ-U E PLANOS DAS COMPONENTES DAS VARIÁVEIS ENVOLVIDAS NA ANÁLISE EXPLORATÓRIA DAS VARIÁVEIS.	64
FIGURA 29 - MATRIZ-U E PLANOS DAS COMPONENTES DAS ANOMALIAS AR-LIVRE E BOUGUER OBTIDAS POR LEVANTAMENTOS GRAVIMÉTRICOS AÉREO E TERRESTRE, DISTÚRBIO DA GRAVIDADE E MODELO DA PROFUNDIDADE DE MOHO. BASEADO NO TAMANHO DE MAPA 14 x 14.	65
FIGURA 30 - MATRIZ-U E PLANOS DAS COMPONENTES DAS ANOMALIAS AR-LIVRE E BOUGUER OBTIDAS POR LEVANTAMENTOS GRAVIMÉTRICOS AÉREO E TERRESTRE, DISTÚRBIO DA GRAVIDADE E MODELO DA PROFUNDIDADE DE MOHO. BASEADO NO TAMANHO DE MAPA 20 x 20.	66
FIGURA 31 - MATRIZ-U E PLANOS DAS COMPONENTES DAS ANOMALIAS AR-LIVRE E BOUGUER OBTIDAS POR LEVANTAMENTOS GRAVIMÉTRICOS AÉREO E TERRESTRE, DISTÚRBIO DA GRAVIDADE E MODELO DA PROFUNDIDADE DE MOHO. BASEADO NO TAMANHO DE MAPA 26 x 26.	67
FIGURA 32 – DUAS JANELAS ONDE FORAM APLICADOS O MÉTODO SOM PARA PREVER ANOMALIA AR LIVRE E ANOMALIA DE BOUGUER.	68
FIGURA 33 – (A) ANOMALIA AR-LIVRE ORIGINAL. (B) IMPUTADO PELO SOM USANDO MS 14 x 14, (D) MS 20 x 20 E (C) 26 x 26.	71
FIGURA 34 – (A) ANOMALIA BOUGUER ORIGINAL. (B) PREDITO PELO SOM USANDO MS 14 x 14, (D) MS 20 x 20 E (C) 26 x 26.	72
FIGURA 35 - (A) ANOMALIA AR-LIVRE ORIGINAL. (B) IMPUTADO PELO SOM USANDO O BMU MÉDIO.	73
FIGURA 36 - (A) ANOMALIA BOUGUER ORIGINAL. (B) IMPUTADO PELO SOM USANDO O BMU MÉDIO.....	73
FIGURA 37 – (A) ANOMALIA AR LIVRE ORIGINAL. A DIFERENÇA ENTRE O ORIGINAL E O PREVISTO PELO SOM É MOSTRADO COM OS SEGUINTE TAMANHOS DE MAPAS (B) MS 14 x 14, (C) MS 20 x 20 E (D) 26 x 26.	74
FIGURA 38 – (A) ANOMALIA BOUGUER ORIGINAL. A DIFERENÇA ENTRE O ORIGINAL E O PREVISTO PELO SOM É MOSTRADO COM OS SEGUINTE TAMANHOS DE MAPAS (B) MS 14 x 14, (C) MS 20 x 20 E (D) 26 x 26.	75

FIGURA 39 - IMAGENS ESPACIALIZADAS DA ANOMALIA AR-LIVRE (A, B, C, D E E) E BOUGUER (F, G, H, I E J) NA ÁREA DE IMPUTAÇÃO JANELA 1, DE ACORDO COM A SEGUINTE DISPOSIÇÃO: (A E F) ORIGINAL, (B E G) 14 x 14, (C E H) 20 x 20, (D E I) 26 x 26 E (E E J) MÉDIA DOS BMUs.	76
FIGURA 40 - IMAGENS DA ANOMALIA AR-LIVRE (A, B, C, D E E) E BOUGUER (F, G, H, I E J) NA ÁREA DE IMPUTAÇÃO JANELA 2, DE ACORDO COM A SEGUINTE DISPOSIÇÃO: (A E F) ORIGINAL, (B E G) 14 x 14, (C E H) 20 x 20, (D E I) 26 x 26 E (E E J) MÉDIA DOS BMUs.	77
FIGURA 41 – GRÁFICOS DE CORRELAÇÃO DE AR LIVRE ORIGINAL VS AR LIVRE IMPUTADO. (A) 14x14, (B) 20x20, (C) 26x26 E (D) BMU MÉDIO	79
FIGURA 42 – GRÁFICOS DE CORRELAÇÃO DE BOUGUER ORIGINAL X BOUGUER IMPUTADO. (A) 14x14, (B) 20x20, (C) 26x26 E (D) BMU MÉDIO	80
FIGURA 43 – GRÁFICOS DE CORRELAÇÃO DE AR LIVRE ORIGINAL VS AR LIVRE IMPUTADO NAS ÁREAS DE JANELA. (A) 14x14, (B) 20x20, (C) 26x26 E (D) BMU MÉDIO	81
FIGURA 44 – GRÁFICOS DE CORRELAÇÃO DE BOUGUER ORIGINAL X BOUGUER IMPUTADO NAS ÁREAS DE JANELA. (A) 14x14, (B) 20x20, (C) 26x26 E (D) BMU MÉDIO	82
FIGURA 45 – (A) ANOMALIA AR-LIVRE ORIGINAL. (B) IMPUTADO PELO SOM USANDO MS 14 x 14, (D) MS 20 x 20 E (C) 26 x 26.	83
FIGURA 46 - (A) ANOMALIA BOUGUER ORIGINAL. (B) IMPUTADO PELO SOM USANDO MS 14 x 14, (D) MS 20 x 20 E (C) 26 x 26.	84
FIGURA 47 - (A) ANOMALIA AR-LIVRE ORIGINAL. (B) IMPUTADO PELO SOM USANDO O BMU MÉDIO.	85
FIGURA 48 - (A) ANOMALIA BOUGUER ORIGINAL. (B) IMPUTADO PELO SOM USANDO O BMU MÉDIO.....	85
FIGURA 49 - ANÁLISE EXPLORATÓRIA DE U E K, ANOMALIA INTERPOLADA, HISTOGRAMAS E INFORMAÇÕES ESTATÍSTICAS.	88
FIGURA 50 - ANÁLISE EXPLORATÓRIA DE CT, Th, Th500 E Th1000 ANOMALIA INTERPOLADA, HISTOGRAMAS E INFORMAÇÕES ESTATÍSTICAS.....	89
FIGURA 51 - MATRIZ DE CORRELAÇÃO E GRÁFICOS DE DISPERSÃO E HISTOGRAMA DAS VARIÁVEIS ENVOLVIDAS NA ANÁLISE EXPLORATÓRIA.	90
FIGURA 52 - MATRIZ-U E PLANOS DAS COMPONENTES DAS VARIÁVEIS ENVOLVIDAS NA ANÁLISE EXPLORATÓRIA DAS VARIÁVEIS.	91
FIGURA 53 - MATRIZ-U E PLANOS DAS COMPONENTES DOS CANAIS K, U E Th DEGRADADA EM 500M OBTIDAS POR LEVANTAMENTOS AEROGAMAESPECTROMÉTRICOS.....	92
FIGURA 54 - MATRIZ-U E PLANOS DAS COMPONENTES DOS CANAIS CT, K, U E Th DEGRADADA EM 500M OBTIDAS POR LEVANTAMENTOS AEROGAMAESPECTROMÉTRICOS.....	93
FIGURA 55 - MATRIZ-U E PLANOS DAS COMPONENTES DOS CANAIS K, U E Th DEGRADADA EM 1000M OBTIDAS POR LEVANTAMENTOS AEROGAMAESPECTROMÉTRICOS.....	94

FIGURA 56 - MATRIZ-U E PLANOS DAS COMPONENTES DOS CANAIS CT, K, U E TH DEGRADADA EM 1000M OBTIDAS POR LEVANTAMENTOS AEROGAMAESPECTROMÉTRICOS.....	95
FIGURA 57 - ANOMALIA INTERPOLADA, HISTOGRAMA E ESTATÍSTICA DOS CANAIS TH ORIGINAL, TH DEGRADADA EM 500M COM E SEM CT.	96
FIGURA 58 - ANOMALIA INTERPOLADA, HISTOGRAMA E ESTATÍSTICA DOS CANAIS TH ORIGINAL, TH DEGRADADA EM 1000M COM E SEM CT.	97
FIGURA 59 - (A) TH ORIGINAL – TH DEGRADADO EM 500M. (B) TH ORIGINAL – TH RECUPERADO UTILIZANDO CT. (C) TH ORIGINAL – TH RECUPERADO SEM CT.....	98
FIGURA 60 - (A) TH ORIGINAL – TH DEGRADADO EM 1000M. (B) TH ORIGINAL – TH RECUPERADO UTILIZANDO CT. (C) TH ORIGINAL – TH RECUPERADO SEM CT.....	99
FIGURA 61 – GRÁFICOS DE CORRELAÇÃO. (A) TH DEGRADADO EM 500M VS TH ORIGINAL. (B) TH RECUPERADO UTILIZANDO CT VS TH ORIGINAL. (C) TH RECUPERADO SEM CT VS TH ORIGINAL	100
FIGURA 62 – GRÁFICOS DE CORRELAÇÃO. (A) TH DEGRADADO EM 1000M VS TH ORIGINAL. (B) TH RECUPERADO UTILIZANDO CT VS TH ORIGINAL. (C) TH RECUPERADO SEM CT VS TH ORIGINAL.	101
FIGURA 63 – ANOMALIA INTERPOLADA (A) TH ORIGINAL, (B) TH DEGRADADO EM 500M, (C) TH RECUPERADO COM RESOLUÇÃO DE 125M E (D) TH RECUPERADO UTILIZANDO A VARIÁVEL CT, COM RESOLUÇÃO DE 125M.....	102
FIGURA 64 – ANOMALIA INTERPOLADA (A) TH ORIGINAL, (B) TH DEGRADADO EM 500M, (C) TH RECUPERADO COM RESOLUÇÃO DE 125M E (D) TH RECUPERADO UTILIZANDO A VARIÁVEL CT, COM RESOLUÇÃO DE 125M.....	103
FIGURA 65 – ANOMALIA INTERPOLADA (A) TH ORIGINAL, (B) TH DEGRADADO EM 500M, (C) TH RECUPERADO COM RESOLUÇÃO DE 125M E (D) TH RECUPERADO UTILIZANDO A VARIÁVEL CT, COM RESOLUÇÃO DE 125M.....	104
FIGURA 66 – ANOMALIA INTERPOLADA (A) TH ORIGINAL, (B) TH DEGRADADO EM 500M, (C) TH RECUPERADO COM RESOLUÇÃO DE 125M E (D) TH RECUPERADO UTILIZANDO A VARIÁVEL CT, COM RESOLUÇÃO DE 125M.....	105
FIGURA 67 – ANOMALIA INTERPOLADA (A) TH ORIGINAL, (B) TH DEGRADADO EM 1000M, (C) TH RECUPERADO COM RESOLUÇÃO DE 125M E (D) TH RECUPERADO UTILIZANDO A VARIÁVEL CT, COM RESOLUÇÃO DE 125M.....	106
FIGURA 68 – ANOMALIA INTERPOLADA (A) TH ORIGINAL, (B) TH DEGRADADO EM 1000M, (C) TH RECUPERADO COM RESOLUÇÃO DE 125M E (D) TH RECUPERADO UTILIZANDO A VARIÁVEL CT, COM RESOLUÇÃO DE 125M.....	107
FIGURA 69 – ANOMALIA INTERPOLADA (A) TH ORIGINAL, (B) TH DEGRADADO EM 1000M, (C) TH RECUPERADO COM RESOLUÇÃO DE 125M E (D) TH RECUPERADO UTILIZANDO A VARIÁVEL CT, COM RESOLUÇÃO DE 125M.....	108
FIGURA 70 – ANOMALIA INTERPOLADA (A) TH ORIGINAL, (B) TH DEGRADADO EM 1000M, (C) TH RECUPERADO COM RESOLUÇÃO DE 125M E (D) TH RECUPERADO UTILIZANDO A VARIÁVEL CT, COM RESOLUÇÃO DE 125M.....	109

LISTA DE TABELAS

TABELA 1 – MÉTODOS GEOFÍSICOS E SUAS RESPECTIVAS PROPRIEDADES RELACIONADAS.....	23
TABELA 2 – VISÃO GERAL DAS VARIÁVEIS	46
TABELA 3 – PARÂMETROS DE INICIALIZAÇÃO DO SOM.....	48
TABELA 4 – PARÂMETROS DE INICIALIZAÇÃO DO SOM.....	51
TABELA 5 – COMPARAÇÃO ENTRE DIFERENTES DIMENSÕES DO TAMANHO DE MAPA: (A) MATRIZ 14 x 14; (B) 20 x 20; (C) 26 x 26 E (D) MÉDIA DOS BMU 14 x 14, 20 x 20 E 36 x 26.	70
TABELA 6 – ANÁLISE ESTATÍSTICA DA JANELA 1 DA ANOMALIA AR-LIVRE (A) E BOUGUER (B) PARA DIFERENTES TAMANHOS DE MAPA (MGAL).	78
TABELA 7 – ANÁLISE ESTATÍSTICA DA JANELA 2 DA ANOMALIA AR-LIVRE (A) E BOUGUER (B) PARA DIFERENTES TAMANHOS DE MAPA (MGAL).	78

SUMÁRIO

1	INTRODUÇÃO.....	14
1.1	APRESENTAÇÃO	14
1.2	OBJETIVO	15
1.2.1	<i>Objetivos específicos .....</i>	<i>15</i>
1.3	JUSTIFICATIVA.....	16
1.4	LOCALIZAÇÃO DAS ÁREAS DE ESTUDO.....	17
1.4.1	<i>Áreas de Estudo da Abordagem 1.....</i>	<i>17</i>
1.4.2	<i>Áreas de Estudo da Abordagem 2.....</i>	<i>18</i>
1.5	ESTRUTURA DA DISSERTAÇÃO	19
2	FUNDAMENTAÇÃO TEÓRICA.....	21
2.1	GEOFÍSICA DE EXPLORAÇÃO.....	22
2.1.1	<i>Gravimetria.....</i>	<i>26</i>
2.1.2	<i>Gamaespectrometria.....</i>	<i>28</i>
2.2	RESOLUÇÃO ESPACIAL E AMOSTRAGEM.....	29
2.3	SELF-ORGANIZING MAPS.....	32
2.3.1	<i>Algoritmo.....</i>	<i>35</i>
2.3.2	<i>Projeção e visualização das análises SOM.....</i>	<i>36</i>
2.4	IMPUTAÇÃO DE VALORES AUSENTES.....	37
2.5	VALIDAÇÃO DOS DADOS PARA SELEÇÃO DE VARIÁVEIS E TESTES DE VALIDAÇÃO 40	
3	MATERIAL E MÉTODOS	42
3.1	IMPUTAÇÃO DE DADOS A PARTIR DOS SOM	42
3.2	ABORDAGEM 1 – IMPUTAÇÃO DE DADOS GRAVIMÉTRICOS	43
3.2.1	<i>Aquisição e características dos dados gravimétricos e geodésicos.....</i>	<i>43</i>
3.2.2	<i>Pré-processamento dos dados para obtenção das variáveis.....</i>	<i>46</i>
3.2.3	<i>Processamento SOM.....</i>	<i>47</i>
3.2.4	<i>Síntese do Fluxo Metodológico.....</i>	<i>48</i>
3.3	ABORDAGEM 2 - INCREMENTO DE RESOLUÇÃO EM VARIÁVEIS GEOFÍSICAS GAMAESPECTROMÉTRICAS	50
3.3.1	<i>Aquisição e características dos dados gamaespectrométricos e demais variáveis envolvidas.....</i>	<i>50</i>
3.3.2	<i>Processamento SOM.....</i>	<i>50</i>
4	RESULTADOS	52
4.1	ABORDAGEM 1 – IMPUTAÇÃO DE DADOS GRAVIMÉTRICOS	52

4.1.1	<i>Análise Exploratória das Variáveis</i>	52
4.1.2	<i>Mapas Auto-Organizados produzidos pela análise exploratória das variáveis</i>	64
4.1.3	<i>Análise SOM das variáveis específicas</i>	64
4.1.4	<i>Validação dos valores imputados em relação aos originais</i>	68
4.1.5	<i>Imputação de dados gravimétricos da área total</i>	83
4.1.6	<i>Discussão sobre os resultados imputados</i>	86
4.2	ABORDAGEM 2 – INCREMENTO DE RESOLUÇÃO EM VARIÁVEIS GEOFÍSICAS	
	GAMASPECTROMÉTRICAS	86
4.2.1	<i>Análise Exploratória das Variáveis</i>	87
4.2.2	<i>Mapas Auto-Organizados produzidos pela análise exploratória das variáveis</i>	91
4.2.3	<i>Análise SOM das variáveis específicas</i>	91
4.2.4	<i>Validação dos valores incrementados em relação aos originais</i>	96
4.2.5	<i>Incremento de resolução de dados gamaespectrométricos em detalhes</i>	102
5	CONCLUSÕES	110
	REFERÊNCIAS	112

1 INTRODUÇÃO

1.1 APRESENTAÇÃO

A Geofísica dedica-se ao estudo da composição, da estrutura, das propriedades e da dinâmica da Terra por meios físicos aplicados em superfície ou próximo a ela. Os tipos de rochas podem ser identificados através de variações relacionadas às propriedades físico-químicas dos seus respectivos minerais constituintes. Vários métodos geofísicos podem ser empregados para auxiliar nessa identificação, tais como: gravimétrico, magnético, gamaespectrométrico, sísmico, elétrico etc.

Com base variações de medidas geofísicas torna-se possível a identificação de anomalias e feições geológicas. Essas variações, no entanto, poderão ser melhor exploradas em termos de geometria e intensidade segundo suas resoluções espaciais. Assim, o atributo de resolução espacial passa a ser fator preponderante em estudos exploratórios.

Os dados geofísicos podem ser obtidos a partir de diferentes sistemas de aquisição, sejam estes levantamentos satelitais, aéreos, marinhos ou terrestres. A amostragem atual em levantamentos aéreos ou satelitais, apesar da maior abrangência de varredura, comumente possuem menores resoluções espaciais do que os levantamentos terrestres. No entanto, as limitações de acesso para o desenvolvimento dos levantamentos terrestres podem gerar amostragens irregulares bem como uma cobertura de dados escassa ou inexistente (JAHN; COOK; GRAHAM, 1998). Essas regiões com baixa frequências de amostragens produzem vazios cartográficos, embora estejam em regiões com ampla necessidade de investigações relativas ao substrato geológico.

A região amazônica no Brasil, caracterizada por uma vasta cobertura vegetal e poucas vias de acesso, possui levantamentos geofísicos aéreos e terrestres em diferentes frequências de amostragem. No caso, por exemplo, de levantamentos gravimétricos terrestres, as amostragens são mais comuns nas margens dos rios e seus afluentes. Neste caso permanecem, portanto, extensas áreas sem a devida aquisição de dados.

Existem vários métodos de imputação que podem sinteticamente criar valores em áreas com dados faltantes. No entanto, a maioria desses métodos são estatísticos e processados em uma única variável, como as interpolações baseadas no vizinho mais próximo, que rejeitam muitas informações intrínsecas ao dado (GRUS, 2016). A técnica Mapas Auto-Organizáveis (do inglês *Self-Organizing Maps*, SOM) consiste em um método não-linear e não-estatístico, com base nos princípios de quantização vetorial e nas medidas de similaridade vetorial (KOHONEN, 2001). As aplicações dos SOM podem ser usadas em classificações, previsões, redução de ruído, dentre outros.

Esta dissertação visa produzir valores sintéticos em variáveis gravimétricas na região amazônica, a partir de dados com múltiplas variáveis, contendo amostras de diferentes resoluções espaciais. Para tanto, serão utilizadas unidades de melhor ajuste obtidas a partir de análises por SOM. Com a mesma técnica, porém em uma segunda abordagem, será desenvolvida a imputação e o incremento de resolução espacial utilizando-se variáveis gamaespectrométricas.

1.2 OBJETIVO

O objetivo deste trabalho é imputar valores faltantes e incrementar a resolução espacial de dados geofísicos multivariados, tendo como base a dependência espacial de variáveis regionalizadas no espaço n -dimensional, a partir de análises SOM. O enfoque analítico será desenvolvido a partir de estudos sistemáticos em bacias sedimentares *onshore*.

1.2.1 Objetivos específicos

Como objetivos específicos, este trabalho visa:

- Alcançar uma sistemática capaz de desenvolver maiores resoluções espaciais em dados multivariados;
- Compreender a relevância quantitativa de variáveis em análises de incremento de resolução;
- Investigar a influência amostral em diferentes resoluções e suas relações com o incremento de resolução.

1.3 JUSTIFICATIVA

Na região amazônica existem dados geofísicos aéreos e terrestres disponíveis em amplas bases de dados. Essas informações podem ser utilizadas tanto para produzir informação em áreas de vazios cartográficos quanto para validar a precisão dos valores sintéticos obtidos. A reserva de parte dos dados terrestres existentes para posterior comparação com os valores sintéticos gerados, conforme Schafer e Graham (2002), fornece a avaliação quantitativa e acurácia desses resultados.

As bacias sedimentares intracratônicas apresentam relevância em termos de exploração e produção de recursos energéticos, sobretudo óleo e gás natural. O mapeamento litoestratigráfico, bem como o desenvolvimento de interpretações relativas ao arcabouço tectônico e estrutural dessas bacias depende diretamente de dados geofísicos, sejam estes aéreos ou terrestres. De todo modo, a imputação e/ou o incremento de resolução espacial, favorecem o desenvolvimento de informação adicional capaz de auxiliar no processo de elaboração de mapas.

A fase exploratória de bacias sedimentares normalmente é iniciada com levantamentos de dados por sensores aerotransportados (JAHN; COOK; GRAHAM, 1998), que podem ser desenvolvidos por agentes ou agências (locais, empresas, órgãos governamentais etc.). Tais levantamentos, no entanto, envolvem dados de diferentes resoluções espaciais. Nesta fase, tempo e orçamento são cruciais, fazendo com seja necessária e cabível a aplicação de algoritmos que possibilitem a maior exploração da base de dados no intuito de gerar produtos que melhor descrevam às regiões de interesse.

Dentre os algoritmos de *machine learning* existentes, os SOM apresentam vasta possibilidade de aplicações, dentre elas a imputação de dados, que também pode ser aplicada para incremento de resolução espacial. Isso ocorre porque em uma base de dados com amostras contendo múltiplas variáveis, o desenvolvimento de análises SOM fornece o tratamento de cada amostra como uma unidade vetorial. Inseridos em um espaço n-dimensional, onde n é o número de variáveis envolvidas na análise, neurônios aleatórios escolhem uma unidade de melhor ajuste (do inglês, *Best Matching Unit*, ou BMU), baseados em etapas de disputa e cooperação para representar as amostras originais. Desse modo, cada amostra original é representada por um vetor BMU, que apresenta um valor correspondente para cada variável inserida

na base de dados (FRIEDEL, 2011). Assim, em uma base de dados multivariada, torna-se possível gerar valores sintéticos de amostras com valores faltantes a partir das suas respectivas BMUs. As pesquisas apresentadas em Samad & Harp (1992) e Folguera et al. (2015) mostram estudos de caso envolvendo o uso de BMUs para criar valores sintéticos em regiões com dados faltantes.

A escassez de trabalhos sobre incremento de resolução utilizando dados multivariados demonstra as dificuldades encontradas e a necessidade de mais estudos. Este trabalho utilizará o SOM como estimador para gerar incremento de resolução, possibilitando a obtenção de dados com maior resolução e baixo custo, principalmente para dados gravimétricos, cuja aquisição é menos comum e demanda rigor analítico e processual.

1.4 LOCALIZAÇÃO DAS ÁREAS DE ESTUDO

Este trabalho utiliza duas abordagens, compreendidas em áreas de estudos distintas, ambas localizadas na Amazônia. As áreas serão descritas nos subtópicos a seguir.

1.4.1 Áreas de Estudo da Abordagem 1

A primeira área de estudo está localizada entre 2°N e 5°S em latitude e 44°W e 58°W em longitude (Figura 1). Os pontos em vermelhos representam um levantamento gravimétrico terrestre da década 1950, obtido pela empresa PETROBRAS S.A. Os pontos em azul claro retratam um levantamento gravimétrico aéreo executado no período de abril de 2007 a março de 2008 para Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP).

A diferença da anomalia ar livre obtida dos levantamentos terrestre e aéreo apresentam média e desvio padrão iguais a -13,6 mGal e 11,5 mGal, respectivamente. Para estes dados, a máxima diferença positiva e negativa são de 77,1 mGal e -145,8 mGal, respectivamente. Devido a esses resultados, tornou-se inviável a utilização direta do levantamento aéreo para aplicações geodesicas. Neste estudo serão imputadas anomalias ar livre nas regiões onde os levantamentos gravimétricos terrestres são inexistentes. Para tanto, criou-se uma grade regular com 5' de resolução espacial para as células dos levantamentos terrestre e aéreo, do modelo geopotencial global e do modelo digital de terreno.

Figura 1 – Localização da área de estudo visando imputação de dados.

Fonte: Bruno Monserrat Perillo, 2019.

1.4.2 Áreas de Estudo da Abordagem 2

A segunda área de estudo compreende a RESERVA NACIONAL DO COBRE E ASSOCIADOS – RENCA, (Figura 2). Esta reserva possui relevante valor econômico (OLIVEIRA, 2010), além de se demonstrar como uma ampla relevância em relação aos estudos acadêmicos. A região também é reconhecida por se tratar de uma das fronteiras geológicas do país.

A RENCA possui uma área de aproximadamente 47 mil km², delimitada entre os paralelos de 01°00'00" de latitude norte e 00°40'00" de latitude sul e os meridianos 52°00'00" e 54°18'00" de longitude, abrangendo parte dos estados do Pará e Amapá (Figura 2). A RENCA foi constituída através do Decreto Federal nº 89.404, de 24 de fevereiro de 1984, e alterado pelo Decreto Federal nº 92.107, de 1º de dezembro de 1985, o qual, no seu Artigo 2º, dispõe que todos os trabalhos de pesquisa destinados à determinação de ocorrências de cobre e seus associados são de responsabilidade exclusiva da Companhia de Pesquisa de Recursos Minerais (CPRM) – Serviço Geológico do Brasil.

A área pré-cambriana da RENCA situa-se entre o Distrito Manganésífero da Serra do Navio (DMSN-AP) e a Província Mineral de Carajás (PMC-PA). Pelo histórico de alta favorabilidade metalogenética tanto do DMSN quanto da PMC, e como essa região apresenta precárias condições de acesso, devido à extensa cobertura de

floresta equatorial e espesso manto de intemperismo, houve a proposição de levantamento aerogeofísico de semidetalhe (magnetometria e gamaespectrometria), para melhorar o conhecimento geológico-estrutural da reserva e apontar para alvos anômalos que mereçam detalhamento (CPRM, 1999).

Figura 2 – Localização da área de estudo. Polígono vermelho representa os limites do aerolevantamento 1057 – CPRM.

Fonte: Bruno Monserrat Perillo, 2019.

1.5 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação é composta por cinco capítulos.

No Capítulo 1 o tema é introduzido em seus múltiplos aspectos. Os objetivos a serem alcançados são elencados, bem como a justificativa para o desenvolvimento da pesquisa. As áreas de estudo são elencadas e mostradas em bases georreferenciadas.

O Capítulo 2 é constituído pela fundamentação teórica dos principais temas abordados para o desenvolvimento da dissertação: (i) geofísica de exploração, (ii) resolução espacial e amostragem, (iii) *Self-Organizing Maps*, (iv) imputação e valores ausentes, (v) técnicas estatísticas para validação dos resultados, (vi) projeção e visualização SOM e (vii) imputação e incremento de resolução.

A descrição metodológica é apresentada no Capítulo 3, onde são relatadas informações sobre o processo de aquisição, processamento dos dados geofísicos bem como o processamento SOM. O capítulo é estruturado a partir de duas abordagens, uma delas relativa à imputação de dados e a outra relacionada ao incremento de resolução espacial.

O Capítulo 4 apresenta os resultados e discussões relacionados às duas abordagens mencionadas no Capítulo 3.

As conclusões sobre os resultados atingidos e sobre recomendações para possíveis trabalhos futuros estão resumidas no Capítulo 5, destacando os ganhos encontrados no trabalho, bem como as principais dificuldades observadas. Além disso, são elencados possíveis temas a serem abordados em investigações posteriores.

2 FUNDAMENTAÇÃO TEÓRICA

A mineração de dados, do inglês *data mining* (DM) é uma etapa do processo de descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases* – KDD). Seu propósito é buscar padrões relevantes entre variáveis e, dessa forma, extrair informação e conhecimento a partir dos dados de forma automatizada ou semiautomatizada (HAN; KAMBER; PEI, 2012; SILVA; PERES; BOSCARIOLI, 2017).

O KDD é constituído por uma série de processos, que incluem o conhecimento do domínio do dado, a obtenção dos dados, pré-processamento, mineração de dados e pós-processamento (Figura 3). Essa estrutura pode ser sintetizada nos tópicos abaixo:

- Obtenção dos dados e organização destes na devida área de interesse;
- Pré-processamento, etapa de identificação e/ou remoção de dados espúrios e inconsistentes, seleção de atributos relevantes, reorganização do dado para o processo de DM;
- Mineração de dados, que consiste na extração de informação a partir de um modelo que represente o conjunto de dados, tal como: agrupamentos, associações e predição;
- Pós-processamento, representado pela validação, visualização e interpretação da descoberta de conhecimento propiciado pelo DM.

Diversas técnicas podem ser utilizadas para DM, dentre as quais podemos mencionar os algoritmos genéticos, as redes neurais artificiais, os classificadores Bayesianos, as árvores de decisão, dentre outras. Tais técnicas são empregadas para solucionar problemas comumente encontrados em experimentos científicos, tais como classificação, agrupamentos e predição

A classificação e agrupamentos (*clustering*) são frequentemente confundidos. Eles refletem propriedades comuns (similaridades) do conjunto de dados. Na classificação, o modelo ou função classifica um item do dado em uma das classes disponíveis. No agrupamento não há classes pré-definidas, dessa forma, o objetivo é encontrar grupos com diferentes características.

Embora todas as classes de conhecimento obtidas pela DM envolvam de alguma maneira previsões, a imputação e o incremento de resolução, os quais referem-se aos objetivos desta pesquisa, tangem à predição de valores baseados em informações de outros valores existentes.

Figura 3 – O processo do KDD

Fonte: Bruno Monserrat Perillo, 2019.

Para a realização da descoberta de conhecimento, é essencial a compreensão do domínio de aplicação. Nesta dissertação, os conceitos de predição serão aplicados aos dados geofísicos e geodesicos, a fim de buscar padrões válidos e úteis para a imputação e o incremento de resolução espacial.

2.1 GEOFÍSICA DE EXPLORAÇÃO

A exploração de recursos naturais ocorre através da avaliação de técnicas empregadas na investigação de depósitos minerais ou reservatórios de bens diversos, e na possível avaliação do volume da reserva a ser explorada. Diversos fatores podem influenciar nas escolhas dessas técnicas, tais como as variações das propriedades físicas a que estas técnicas representam ou ressaltam nas suas diferentes profundidades. Tais fatores podem ser determinantes à construção de um modelo geológico de maior correspondência para com a realidade. Outro fator relevante para ser avaliado é o custo relativamente baixo dos métodos indiretos ante ao amplo benefício para a tomada de decisão nas fases de exploração e produção.

Os métodos geofísicos podem ser empregados em estudos teóricos e aplicados. A fonte geradora para os diferentes registros pode ser ativa (tais como explosivos, injeção de corrente elétrica etc.) ou passiva, emitida a partir de fontes

naturais (terremoto, campo gravitacional, radiação cósmica, gamaespectrométricas etc).

Nesta dissertação, serão desenvolvidas análises de DM aplicada aos métodos gravimétrico e gamaespectrométricos, comumente empregados em análises exploratórias de áreas de interesses. Tais métodos possibilitam estudos regionais de bacias sedimentares e/ou ambientes cratônicos. Muitas vezes os dados são adquiridos simultaneamente através de aerolevantamentos, para diminuir custos envolvidos na aquisição.

Cada método empregado na investigação é capaz de representar diferentes informações físico-químicas em diferentes níveis de profundidades. Esses métodos, portanto, possibilitam o reconhecimento de diferentes tipos de rochas através de variações relacionadas às propriedades físicas. Com base nos registros medidos, torna-se possível a identificação de anomalias e feições geológicas. Neste sentido, vários métodos são utilizados para que se complementem e, dessa forma, diminuam a incerteza e eliminem problemas de ambiguidade (problema inverso).

Uma síntese dos métodos geofísicos é mostrada na Tabela 1:

Tabela 1 – Métodos Geofísicos e suas respectivas propriedades relacionadas

Método geofísico	Propriedade física
Gravimétrico	Densidade
Gamaespectrométrico	radioatividade
Magnético	Suscetibilidade magnética
Sísmico	Densidade, módulo elástico e impedância acústica
Elétricos - Resistividade	Condutividade e capacitância elétrica
Eletromagnético	Condutividade e indutância elétrica constante dielétrica e resistividade

Fonte: Bruno Monserrat Perillo, 2019.

As aquisições de dados geofísico são desenvolvidas a partir de levantamentos satelitais, aéreos, marinhos ou terrestres.

No levantamento aéreo são sobrevoados linhas de medições paralelas e, comumente, equidistantes. Linhas de controle são voadas com maior espaçamento perpendicularmente às linhas de aquisição (Figura 4).

Figura 4 - Esquema de um aerolevantamento

Fonte: Ribeiro, Mantovani e Louro, 2014.

O baixo custo, a rapidez, a possibilidade de fazer medidas em locais de difícil acesso logístico, simultaneamente, com mais de um método geofísico, faz com que este tipo de levantamento seja priorizado em estudos regionais ou de reconhecimento. Os métodos comumente utilizados neste tipo de levantamento são: gravimetria, magnetometria e gamaespectrometria, que podem ser precedidas pelo prefixo aero, para sintetizar o método empregado e a forma de aquisição.

A Figura 5 exibe uma aproximação relativa de custo para 3 exemplos de levantamentos. O mesmo montante financeiro é possível cobrir uma área de 160 km² com um aerolevamento magnetométrico e gamaespectrométrico ao invés de uma cobertura de 3,6 km² de gravimetria terrestre. Segundo Dentith e Mudge (2014) fatores como condições econômicas, tempo, áreas de difíceis acessos e montanhosas podem influenciar diretamente no custo de aquisição.

Os levantamentos aéreos, normalmente, adquirem dados de resolução espacial limitada. Essa limitação de resolução pode ocasionar em restrições ao processo de interpretação, sobretudo em escalas onde há necessidade da compreensão da informação em escalas de maior detalhe. Este tipo de levantamento, no entanto, é utilizado em trabalhos que envolvem as escalas regional, reconhecimento e detalhamento. O espaçamento entre linhas define o objetivo de investigação. Levantamentos de detalhe podem utilizar espaçamentos de 50 m entre linhas de aquisição, ao passo que os levantamentos regionais tendem a espaçamentos de 250 a 500 m. Outros fatores operacionais importantes para este tipo de aquisição são a altura de voo e a frequência de amostragem. Tais informações

poderão refletir na qualidade da informação e nos tamanhos das células geradas em etapas de interpolação subsequentes.

Figura 5 – Custos relativos aproximados de diferentes tipos de levantamentos geofísicos em termos de área.

Fonte: Adaptado de Dentith e Mudge, 2014.

Levantamentos terrestres podem contribuir de diversas maneiras para o detalhamento da investigação em superfície e subsuperfície, seja pela facilidade de adensar as medições em uma área específica, ou mesmo pela praticidade e simplicidade para refazer medidas em um determinado ponto cuja informação coletada se demonstrou espúria ou duvidosa. No entanto, a dificuldade de acesso em áreas de interesse pode gerar amostragens irregulares ou cobertura de dados escassos ou inexistentes (Figura 6). Tais dificuldades podem ser causadas por vegetação densa, terrenos acidentados, extensa área inundada, falta de permissão etc. Nestas situações, muitas vezes é mais simples que a coleta dos dados seja desenvolvida nas margens de lagos, rios, córregos ou outros canais de drenagens.

A Figura 6 mostra diferentes configurações de amostragem nos levantamentos aéreos e terrestre. Em (a) e (c) o espaçamento entre os pontos de medições são regulares à distância constante. Entretanto, outros problemas práticos podem interferir nestas situações (b). (d) representa uma típica aquisição terrestre onde os pontos são aleatórios, embora muitas vezes é possível medir com pequenos espaçamentos ao longo de estradas.

Figura 6 – Algumas configurações de aquisição de dados comuns em pesquisas geofísicas. Os pontos representam a amostragem dos dados. (a), transversas regulares; (b) transversas irregulares; (c) malha quadrada; e (d) randômica.

Fonte: Adaptado de Dentith e Mudge, 2014.

2.1.1 Gravimetria

A gravimetria estuda as variações da aceleração de gravidade da Terra, dando informações referentes às variações laterais de densidade. O método gravimétrico fornece não apenas a avaliação de concentração de minerais como minérios, petróleo etc., mas também informações quanto à forma e às dimensões da Terra.

Medições de aceleração da gravidade são obtidas através de equipamentos denominados gravímetros, que podem ser classificados em dois tipos: (i) absoluto, empregado na determinação da magnitude absoluta da aceleração da gravidade e (ii) diferencial, que afere variações da aceleração da gravidade de um lugar para outro. Segundo Lowrie (2007), na prospecção, são obtidos gradientes gravimétricos da ordem de 0,01 mGal ($1,0 \times 10^{-5} \text{ m/s}^2$).

Em estudos gravimétricos são utilizados 3 diferentes tipos de superfícies (Figura 7), a saber:

- (i) Superfície física da Terra, onde são realizadas as medições gravimétricas;
- (ii) Elipsoide de revolução, que é uma superfície de referência, convencionalmente adotada para facilitar cálculos geodesicos;
- (iii) Geoide, que consiste em uma superfície equipotencial coincidente com o nível médio não perturbado dos mares.

O posicionamento altimétrico de um ponto na superfície topográfica é dado por sua distância ao geóide contada na vertical - altitude ortométrica (H). Por outro lado, o posicionamento planimétrico está vinculado à superfície do elipsoide de revolução, referencial adotado internacionalmente como aproximação da Terra.

A distância entre as superfícies geoidal e elipsoidal é chamada altura ou ondulação geoidal (N).

Figura 7 – Superfícies de referências

Fonte: Arana, 2009.

A forma da Terra é aproximada por um esferoide achatado nos pólos no qual representa uma superfície equipotencial coincidente com o nível médio dos mares. Diversas fórmulas foram propostas para determinar o valor da aceleração da gravidade em qualquer ponto do elipsoide, como a Fórmula Internacional da Gravidade adotada em 1980, conforme a Equação 1:

$$g_{teor} = 978032,7 \cdot (1 + 0,0053024 + \sin^2 \varphi - 0,0000058 \cdot \sin^2 2\varphi) \quad (1)$$

onde g_{teor} é a aceleração da gravidade teórica, em mGal, e φ é a latitude geográfica.

Anomalia gravimétrica Δg é a diferença entre os valores de aceleração da gravidade medido e o valor teórico fornecido pela Equação 2:

$$\Delta g = g_{med} - g_{teor} \quad (2)$$

Alguns fatores podem influenciar as medições do campo gravitacional como a elevação, topografia, marés, latitude, erros instrumentais e as variações de densidade em subsuperfície. A fim de isolar efeitos indesejados são aplicadas algumas correções. Dentre eles, Bouguer e ar-livre.

A correção ar-livre visa compensar os efeitos da variação de altitude dos pontos de medida na superfície física em relação ao geoide, ou seja, redução ao nível do mar. A correção ar-livre (Equação 3), em mGal, é dada por:

$$C_{al} = 0,3086h \quad (3)$$

onde h é a altitude ortométrica, altitude da medida acima do geoide no ponto.

Especificamente, a anomalia ar-livre, em mGal, é obtida pela Equação 4.

$$\Delta g_{al} = g_{med} + C_{al} - g_{teor} \quad (4)$$

A anomalia Bouguer é dada pela Equação 5, obtida após a aplicação das correções ar-livre e Bouguer (Equação 6). A correção Bouguer, em mGal, propicia a remoção das massas topográficas.

$$\Delta g_B = g_{med} + C_{al} - C_B - g_{teor} \quad (5)$$

$$C_B = 0,0419\rho h \quad (6)$$

Onde ρ é a densidade do material crustal e usualmente é utilizado $\rho = 2,67 \text{ g/cm}^3$ quando a densidade for desconhecida, que é a média para a crosta superior.

2.1.2 Gamaespectrometria

O método gamaespectrométrico monitora a abundância relativa da radiação gamaespectrométrica proveniente das rochas. Tal método pode ser aplicado através de aerolevantamentos, em poços, ou mesmo por levantamentos terrestres. As medidas são obtidas na forma de radiação eletromagnética dentro do espectro dos

raios gama. Os elementos comumente utilizados em exploração de recursos naturais são o urânio (^{238}U), o tório (^{232}Th) e o potássio (^{40}K).

Aerolevantamentos gamaespectrométricos são frequentemente utilizados por se tratar de dados, relativamente, de alta resolução e baixo custo. Mesmo quando a aquisição é compartilhada com outros métodos, a frequência amostral da gamaespectrometria é maior.

Nos levantamentos gamaespectrométricos, além dos registros da radiação de interesse, existe uma série de fenômenos ou efeitos indesejáveis que também são captados pelos sensores. Tais efeitos são ocasionados por ruídos de fundo, como a radioatividade da aeronave, radiação cósmica entre outros (RIBEIRO; MANTOVANI; LOURO, 2014). Para atenuar esses efeitos é necessário o desenvolvimento de uma série de etapas de pré-processamento nos dados brutos, tornando-os adequados para utilização nas etapas seguintes.

Diversas aplicações do método são catalogados na literatura, como exploração de petróleo (SAUNDERS, 1994), prospecção mineral, mapeamento geológico (RIBEIRO; MANTOVANI; LOURO, 2014), cratera de impacto (VASCONCELOS; LEITE; CRÓSTA, 2012) etc.

2.2 RESOLUÇÃO ESPACIAL E AMOSTRAGEM

O conceito de resolução espacial em processamento digital de imagens diz respeito à quantidade de informação que um dado apresenta por unidade de área, ou seja, o grau de definição de uma imagem. Em outras palavras, é o tamanho que um pixel apresenta em cada uma de suas arestas. Quanto menor for o pixel maior será a resolução espacial (DENTITH; MUDGE, 2014). Cada ponto de amostragem georreferenciada, estruturado em grade regular, pode ser representado em imagem, como um pixel colorido. Desta forma, facilitar a identificação e extração da informação contida nas imagens, para posterior interpretação (CRÓSTA, 1992).

Uma das principais questões desenvolvidas durante o planejamento de um levantamento geofísico refere-se à frequência de medições. Essa informação estará diretamente relacionada ao custo, ao tamanho da anomalia, ao objetivo, ao prazo, dentro outros. A frequência de medições de uma determinada propriedade ao longo de uma linha ou grade possibilita que feições medidas sejam espacialmente mapeadas com maior ou menor detalhe (JAHN; COOK; GRAHAM, 1998)

Segundo Hood, Holroyd e Mcgrath (1979); Jahn, Cook e Graham (1998); Reynolds (2011), deve-se atentar para o falseamento espacial (Figuras 8 e 9), ou seja, a perda de detalhes da informação resultante de coleta de dados em pontos de medição insuficientes. Por se tratar de medições pontuais, uma prática comum é a utilização de interpoladores, atribuindo desta forma, a continuidade desejada. Consequentemente, o falseamento espacial fica evidente em mapas interpolados (Figura 9). Entre outros fatores, a eficiência da interpolação depende da quantidade de informação amostrada.

A Figura 9 apresenta dados de um aerolevante magnético, com espaçamento de 150 m (A). Em (B) e (C), o dado foi degradado em 2 e 4 vezes, simulando o espaçamento de 300 m e 600 m, respectivamente. A taxa de amostragem é a quantidade de medições em uma unidade de tempo ou distância. A taxa mínima pode ser computada por meio da Frequência de Nyquist (Equação 7). Supondo uma amostragem a cada 2 ms, a frequência de amostragem é de 500 Hz, serão preservados. A linha azul representa a frequência fictícia gerada por falseamento e a linha vermelha, mostra a frequência de onda senoidal maior que a frequência de Nyquist. O falseamento ocorre quando a taxa de amostragem for menor que f_N

$$f_N = \frac{1}{2\Delta t} \quad (7)$$

onde Δt é o intervalo de amostragem.

Figura 8 - Perda de informação devido ao limitado número de pontos de medição.

Fonte: Bruno Monserrat Perillo, 2019

Figura 9 – Exemplo falseamento espacial em um levantamento aeromagnético hipotético. (A) Dados originais, espaçamento de linhas de voo de 150 m, (B) 300 m e (C) 600 m.

Fonte: Adaptado de Hood, Holroyd e Mcgrath, 1979.

2.3 SELF-ORGANIZING MAPS

Diversas redes de auto-organização têm sido exploradas nas últimas décadas, tendo o aprendizado competitivo como fundamento básico, tal como apresentado por Rumelhart e Zipser (1985). Algumas redes deste tipo são conhecidas como *Adaptive Resonance Theory – ART* (CARPENTER; GROSSBERG, 1988), *Fukushima's cognitron* (FUKUSHIMA, 1975) e *SOM*, que, frequentemente, são referenciados como redes de Kohonen ou mapa de conservação de topologia (KOHONEN, 2001).

No início dos anos 80, Kohonen introduziu um novo mapa de projeção não-linear, cuja tarefa é projetar dados multidimensionais complexos e não lineares em um espaço de baixa dimensão com relações geométricas simples (KOHONEN, 2001). Seu objetivo é descobrir estruturas estatísticas ou classes no dado de entrada, sem a necessidade de nenhuma intervenção do usuário.

Este método tem sido utilizado para análise e detecção de padrões em dados multidimensionais, tais como multiespectrais (SAVELIEV; DOBRININ, 1999); hiperspectrais (PENN, 2002); geoquímicos (PENN, 2005); geofísicos (KLOSE, 2006; CARNEIRO et al., 2012), entre outros.

Algumas finalidades destes tipos de problemas podem ser descritas como:

- Agrupamento: no resultado do algoritmo SOM, é obtido a rotulagem dos agrupamentos no dado de entrada;
- Redução de dimensionalidade: os dados de entrada são representados por um subespaço de dimensão inferior à dimensionalidade do dado, objetivando reduzir o número de características que as representa;
- Extração de características: extrai características dos dados de entrada, similar à redução de dimensionalidade;
- Quantização vetorial: o algoritmo busca a discretização ótima do espaço de entrada.

Em análises por SOM, cada amostra de entrada é tratada como um vetor em um espaço n -dimensional definido pelas quantidades de atributos analisados. As amostras originais são inseridas em um espaço n -dimensional construído a partir das variáveis envolvidas; um tamanho de mapa com topologia definida é produzido gerando, assim, uma determinada quantidade de neurônios semeados

aleatoriamente. A quantidade de neurônios é, portanto, dependente do tamanho de mapa escolhido. Os neurônios irão competir e cooperar para, finalmente, elegerem os BMU's que serão projetados nos mapas com a topologia previamente definida. A distância euclidiana (topológica e topográfica) entre os BMU's adjacentes serão representadas no mapa auto-organizado.

As unidades SOM, também chamadas de neurônios, estão conectadas às suas estruturas vizinhas (Figura 10), formando mapas, tipicamente hexagonais ou retangulares, que determinam sua vizinhança. Este mapa é uma grade flexível que se molda na similaridade dos dados.

Com base nessa lógica, os dados próximos ao espaço de entrada terão a representação próxima ao espaço de saída, pertencendo ao mesmo cluster ou a clusters próximos. Portanto, análises SOM preservam as relações métricas e topológicas do espaço de entrada em uma rede bidimensional que pode ser usada como uma ferramenta de visualização, com a sua possível organização em clusters (KURODA et al., 2012).

Os vetores resultantes do treinamento, BMUs (Best Matching Unit), representam a melhor distribuição de um grupo com amostras similares (CARNEIRO, 2010). Estes vetores são projetados numa hipersuperfície cujas visualizações podem ser obtidas através da matriz da distância unificada (U-Matrix) e das parcelas componentes (Component Plots). Por preservar as relações topológicas do espaço amostral, baseada nas similaridades, esta técnica pode ser utilizada para agrupar e prever dados.

O algoritmo é baseado em um processo de aprendizagem competitivo, no qual apenas uma saída de unidade computacional, também denominada unidade mais representativa (BMU) ou grupo local de unidades computacionais. A intensidade de atração do vetor de entrada é controlada pela aprendizagem média, e o número de vetores atraídos pelo BMU vencedor é controlado pelo raio de vizinhança.

O número de unidades no mapa (número de BMUs) é escolhido a priori e determina a eficácia e generalização do modelo, pois durante o treinamento o método forma uma rede elástica que se adapta aos dados de entrada. Vesanto e Alhoniemi (2000) sugerem a Equação 8, para escolher o tamanho da unidade de saída onde m é tamanho da matriz e n é o número de amostras treinadas.

$$m = 5 \cdot \sqrt{n} \quad (8)$$

A proximidade de pontos num mapa treinado tende a representar a similaridade dos objetos correspondentes no espaço original. Para efeito de visualização, Ultsch (2003) descreveu a U-Matrix que contém resultado do cálculo das distâncias entre neurônios adjacentes no mapa treinado. A projeção do conteúdo da U-Matrix no espaço de saída permite uma imagem representando as regiões que concentram neurônios similares, separados por regiões de fronteira com baixa densidade de neurônios.

Figura 10 - Estrutura Básica rede SOM

Fonte: Ghaseminezhad e Karami, 2011.

Os estágios do algoritmo SOM podem ser sintetizados em 4 etapas como é descrito:

- **Inicialização:** Um peso inicial aleatório é atribuído a todas as conexões.
- **Competição:** Busca pelo BMU, através de medida de similaridade, usualmente utilizada a distância Euclidiana representada pela Equação 9, onde é calculado a distância entre o vetor de entrada e o vetor de pesos. O neurônio cuja distância é mínima é então denominado BMU.

$$BMU = \underset{j}{\operatorname{argmin}} \{ \|x_i - w_j\| \} \quad (9)$$

x_i é o vetor de entrada, onde o vetor x_i é a i -ésima entrada escolhida no tempo t ; w_j é o valor dos pesos.

- **Cooperação:** O BMU encontrado na fase competitiva determina o centro e sua vizinhança topológica por meio da interatividade entre os

neurônios. A função gaussiana utilizada, Equação 10, mostra a proximidade do neurônio analisado com o BMU.

$$h_{ci}(t) = e^{-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}} \quad (10)$$

$h_{ci}(t)$ define a região de influência que a amostra de entrada tem sobre o SOM; t e T representam a quantidade de iterações e o número máximo de iterações.

A função de onda de Ricker também pode ser utilizada como função de vizinhança nesta etapa.

- **Adaptação Sináptica (Aprendizagem):** A cada iteração, os pesos de cada neurônio da rede são atualizados de acordo com uma função de treinamento, Equação 11. Além disso, todos os neurônios serão atualizados de acordo com uma taxa de aprendizado, Equação 12.

$$w_j(t + 1) = w_j(t) + \alpha [x_i - w_j(t)] \quad (11)$$

$$\alpha(t) = \alpha_0 \left(1 - \frac{t}{T}\right) \quad (12)$$

α_0 é a razão de aprendizado inicial.

2.3.1 Algoritmo

O algoritmo de treinamento do SOM preserva as relações topológicas baseado nas relações de similaridades do banco de dados. Um esquema de algoritmo SOM é descrito.

passo 1	Defina o formato da rede (uni, bi, tridimensional). Defina a geometria de vizinhança entre os neurônios (quadrada, hexagonal). Defina a forma da hipersuperfície (planar, cilíndrico ou toroidal). Defina o tamanho do mapa (número de vetores semente). Inicialização dos pesos w_j . Defina parâmetros de taxa de aprendizado. Defina o número de iterações T
passo 2	Enquanto $t < T$, faça Passo 3 a 7.
passo 3	Para cada vetor de entrada x , faça Passos 4 e 5.
passo 4	Para cada j , encontre: $BMU = \underset{j}{\operatorname{argmin}} \{ \ x_i - w_j\ \}$
passo 5	Atualize o peso do BMU $w_j(t+1) = w_j(t) + \alpha [x_i - w_j(t)]$
passo 6	Atualize a taxa de aprendizagem. $\alpha(t) = \alpha_0 \left(1 - \frac{t}{T}\right)$
passo 7	Reduza o raio da vizinhança topológica.

Fonte: Bruno Monserrat Perillo, 2019
ci

2.3.2 Projeção e visualização das análises SOM

Na U-Matrix (veja Figura 11), cores tendendo ao azul indicam neurônios próximos, enquanto aquelas que tendem ao vermelho neurônios distantes bem como regiões de fronteiras. As cores estão associadas à distância euclidiana medida entre neurônios adjacentes.

O intuito da exploração de dados em espaços n -dimensionais é reduzir (d) a dimensionalidade ou o número de instâncias (n). O comportamento e espalhamento das amostras no espaço (de entrada – vetorial – espaço dos dados) permite a classificação ou categorização de amostras.

As parcelas componentes possibilitam visualizar e quantificar a contribuição dos valores relativos às variáveis analisadas para cada neurônio resultante posicionado no mapa auto-organizável, sendo possível verificar as relações entre as respostas das várias componentes. A matriz-U possibilita, então, a classificação dos dados relacionada ao vetor similaridade construído a partir dessas amostras.

As análises SOM dependem da escolha de uma malha como formato de visualização. Para a maximização da vizinhança, diversos autores têm mencionado a utilização de malhas hexagonais. Uma superfície de um hipervolume deve ser utilizada para a projeção dos neurônios ou BMU's. O hipervolume toroidal, dada a sua periodicidade preservada, possibilita constatada aderência a manutenção das características topológicas. (MOUNT; WEAVER, 2011). Para a definição do tamanho de mapa auto-organizado resultante, será utilizada a Equação 1 proposta por Vesanto e Alhoniemi (2000). Dessa maneira, um tamanho de mapa será escolhido como adequado para este estudo exploratório. Com a geração do mapa auto-organizado, é possível produzir imagens da Matriz-U e parcelas componentes.

2.4 IMPUTAÇÃO DE VALORES AUSENTES

A ausência de dados é uma questão recorrente em investigações científicas, sobretudo em dados geofísicos. Tais ausências acontecem tanto por dificuldades logísticas e operacionais inerentes à aquisição de dados em locais de difícil acesso, quanto pela baixa resolução espacial, falhas nos sensores, ruídos etc. Dados ausentes podem ocasionar no aumento de incertezas quanto à estimativa dos parâmetros de interesse do estudo, bem como levar a conclusões equivocadas.

Uma forma de minimizar o efeito de áreas sem levantamento geofísico seria usar métodos como interpolação ou imputação. Em estatística, imputação é definida como a predição de dados ausentes com base em algum procedimento, geralmente usando um modelo matemático em combinação com informações disponíveis (DICKSON; GIBLIN, 2007).

As técnicas mais simples de imputação substituem os valores através da média, da mediana, por interpolação ou regressão linear. Dentre tais técnicas destacam-se a krigagem e vizinho mais próximo. Embora sejam de simples implementação, existem algumas desvantagens nessas técnicas de imputação, como por exemplo a subestimação da variabilidade, ou mesmo a dificuldade de utilização de outras variáveis do próprio conjunto de dados para prover melhores resultados.

A imputação de dados utilizando a técnica SOM já foi aplicada por Samad & Harp (1992) e Wang (2003), usando exclusivamente a correlação espacial para prever os valores ausentes pelos valores denominados BMUs. Fessant & Midenet (2002) propuseram a utilização do SOM para imputação e correção de dados errôneos

aplicados a levantamentos em geral. Merlin et al. (2010) utilizaram dupla imputação através de adaptações do SOM para correlação temporal e para predição espacial. Folguera et al. (2015) mostraram a eficácia do SOM em dados quimiométricos, considerando a imputação de diferentes variáveis simultaneamente.

A técnica SOM possibilita o agrupamento não-supervisionada a partir da quantização vetorial em um espaço n-dimensional. Os dados estimados pela técnica têm como base as distâncias entre os vetores disponíveis. Para os dados de menor resolução espacial, o processo de estimativa tradicional é dado por substituição, os valores são produzidos a partir dos vetores BMU's. Muitas vezes os conjuntos de dados resultarão em previsões tendenciosas (DICKSON; GIBLIN, 2007).

A auto-organização de mapas é uma técnica de agrupamento não-supervisionada que utiliza a quantização vetorial em um espaço n-dimensional (KOHONEN, 2001). Os dados estimados pela técnica têm como base as distâncias entre os vetores disponíveis (FESSANT; MIDENET, 2002; WANG, 2003; JUNNINEN et al., 2004; KALTEH; HJORTH; BERNDTSSON, 2008; KALTEH; HJORTH, 2009). Para os dados com menor resolução espacial, o processo de estimativa tradicional é dado por substituição, onde os valores são produzidos a partir dos vetores das Best Matching Units (BMU's). Muitas vezes os conjuntos de dados resultarão em previsões tendenciosas (DICKSON; GIBLIN, 2007; MALEK et al., 2008), o que faz necessário a utilização de técnicas como *bootstrap* (BREIMAN, 1996), conjunto média (RALLO et al., 2002), ou do vizinho mais próximo (MALEK et al., 2008).

Nesse sentido, o espaço de dados é introduzido de forma aleatória. Como variáveis de entrada, são utilizadas as variáveis geofísicas (gravimetria e gamaespectrométricos) e modelo digital de terreno.

Cada amostra apresenta um valor para cada uma das variáveis inseridas nas análises. Como resultados, teremos os BMU's para cada amostra e variável analisada, bem como BMU's para esses mesmos parâmetros nas amostras com análises incompletas. Todo o processo é exemplificado na Figura 11. Além disso, a clusterização realizada a partir do mapa de similaridade entre as amostras permitirá a separação de amostras conforme suas características de similaridade natural.

Os valores resultantes para o incremento da resolução são determinados, portanto, a partir dos BMU's para cada amostra original. Assim, os pixels originais,

com baixa resolução espacial (Figura 12), receberão novos valores otimizados, tendo como base a resolução espacial da variável de menor tamanho de pixel. Desse modo, os valores resultantes terão incremento de resolução quando comparado aos originais.

Figura 11 – Síntese das etapas de uma análise em ambiente de SOM

Fonte: CARNEIRO, 2010.

Figura 12 – Síntese do processo do SOM, apresentando dados de entrada com diferentes resoluções.

Fonte: Bruno Monserrat Perillo, 2019.

A Figura 12 exemplifica todo o processo de incremento de resolução. Como entrada há 3 diferentes imagens que possuem resoluções diferentes. A de maior resolução possui 1 pixel que será relacionado com 4 pixels da imagem de menor resolução.

2.5 VALIDAÇÃO DOS DADOS PARA SELEÇÃO DE VARIÁVEIS E TESTES DE VALIDAÇÃO

Técnicas de visualização são comumente utilizadas no início ou fim do processo de mineração de dados, tendo como principal finalidade a representação de imagens gráficas que possam facilitar o entendimento pelos seres humanos (RABELO, 2007). Dentre as técnicas mais comuns, podemos destacar a matriz de correlação e dispersão e a visualização SOM.

A correlação das variáveis trata da medida da direção e do grau de relacionamento entre duas variáveis (DESHPANDE; VANDERSLUIS; MYERS, 2013; ALTMAN; KRZYWINSKI, 2015). O coeficiente amostral pode ser chamado de coeficiente de Pearson, coeficiente de correlação Pearson ou simplesmente de coeficiente de correlação, que é a medida mais conhecida de interdependência entre atributos da base de dados.

Para uma série de n medições de X e Y , x_i e y_i para $i = 1, 2, \dots, n$, o coeficiente de correlação da amostra pode ser usado para estimar o coeficiente de correlação de Pearson da população r entre X e Y . Então, o coeficiente de correlação da amostra é escrito como (Equação 13):

$$r = \frac{\sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \pm \sqrt{R^2} \quad (13)$$

em que $\bar{x}$ e $\bar{y}$ são as médias amostrais de X e Y .

Os valores de correlação se limitam entre -1 e 1. Os valores indicam a intensidade da correlação e o sinal representa a direção, negativa ou positiva.

<u>Valor absoluto de r</u>	<u>Força de correlação</u>
$r > 0,7$	forte
$0,5 < r > 0,7$	moderada
$0,3 < r > 0,5$	fraca
$r < 0,3$	desprezível

Para quantificar a eficácia na predição é utilizado o coeficiente de determinação (R^2) determinada pela Equação 14, que é indicado para explicar em termos de proporções o quanto a variável predita é igual a variável original. É uma medida em termos de proporções das variabilidades de duas variáveis, ou simplesmente, o quanto a variável predita é “igual” a variável original.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad 0 \leq R^2 \leq 1 \quad (14)$$

y_i = valor da observação individual

$\bar{y}$ = média de todas observações

$\hat{y}_i$ = valor previsto da observação y_i

Valores de R^2 que tendem a 1 ou 100% indicam que o modelo é o que melhor se ajusta à amostra.

3 MATERIAL E MÉTODOS

3.1 IMPUTAÇÃO DE DADOS A PARTIR DOS SOM

Nesta pesquisa serão investigadas diversas variáveis aleatórias que possam contribuir com a imputação e o incremento de resolução em análises SOM.

Como resultados, teremos os BMU's para cada amostra e variável analisada, bem como para os mesmos parâmetros nas amostras com análises incompletas. Além disso, o agrupamento realizado a partir do mapa de similaridade entre as amostras permitirá a separação de amostras conforme suas características de similaridade natural.

Os valores resultantes para o incremento da resolução serão determinados, portanto, a partir dos BMU's para cada amostra original. Assim, os pixels originais, com baixa resolução espacial, receberão novos valores mais ajustados, tendo como base a resolução espacial da variável de menor tamanho de pixel. Desse modo, os valores resultantes terão incremento de resolução quando comparados aos originais.

Aqui faz-se a distinção entre imputação e incremento de resolução. Embora conceitualmente ambas tratem da recuperação de valores ausentes na base de dados, o incremento de resolução é tratado aqui como a imputação de valores, de forma periódica, com amostragem igual à do atributo de maior resolução contido na base de dados. Já a imputação trata da recuperação de valores ausentes, em intervalo limitado e não-periódico, a partir das relações não-lineares dos atributos da base de dados que possuem amostras no mesmo intervalo. A representação tanto da imputação quanto o incremento de resolução é exemplificado na Figura 13. Pontos azuis representam os pontos de amostragem de um dado de baixa resolução com ampla cobertura. Os pontos vermelhos, dado de alta resolução com falta de informação em algumas áreas. Desta forma, a imputação será a criação de valores sintéticos da variável representada pelos pontos vermelhos, recuperando a informação nestas janelas. O incremento de resolução possibilita então que a variável em azul possua a mesma resolução da variável em vermelho.

Figura 13 – Imagem espacializada representando a amostragem de um dado de baixa resolução (azul) e alta resolução (vermelho) com valores ausentes em 3 diferentes áreas de janela.

Fonte: Bruno Monserrat Perillo, 2019.

A escassez de trabalhos sobre incremento de resolução utilizando dados gravimétricos compostos por múltiplas variáveis demonstra as dificuldades encontradas e a necessidade de mais estudos sobre o tema. Neste contexto, propomos a aplicação dos SOM para imputar dados de gravimetria terrestre na região amazônica. Nesta região há pouca disponibilidade de levantamentos de gravimetria terrestre devido à grande dificuldade de obtenção de tais dados, embora haja abundância em outras variáveis obtidas por fontes aerotransportadas e satelitais.

3.2 ABORDAGEM 1 – IMPUTAÇÃO DE DADOS GRAVIMÉTRICOS

3.2.1 Aquisição e características dos dados gravimétricos e geodésicos

Visando a imputação de dados gravimétricos, foi escolhido uma área de estudo (Figura 1) localizada na região norte do Brasil. Esta área é compreendida por densa cobertura vegetal que impossibilita o acesso para levantamentos terrestres.

Um levantamento da gravimetria terrestre foi realizado pela Petrobras na década de 1950, próximo ao rio Amazonas e aos afluentes. Um total de 43.605 pontos com diferentes padrões de espaçamento foram coletados. Por se tratar de um levantamento complexo e com possibilidade de múltiplos erros, foi realizado o controle de qualidade deste levantamento pela equipe do Laboratório de Topografia e Geodesia (LTG) da Universidade de São Paulo.

O levantamento aerogravimétrico executado no período de abril de 2007 a março de 2008 pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), cobriu toda a área analisada nesta pesquisa (Figura 14). Consistindo em linhas gravimétricas espaçadas de 6 e 24km (ANDRADE; KONZEN, 2008).

O LTG forneceu a grade, com resolução de 5', de anomalias ar livre e Bouguer médias (aérea e terrestre) para essa análise.

A diferença da anomalia ar livre obtida dos levantamentos terrestre e aéreo apresentam média e desvio padrão iguais a -13,6 mGal e 11,5 mGal, respectivamente. As máximas diferenças positiva e negativa são de 77,1 mGal e -145,8 mGal, respectivamente. Devido a esses resultados, tornou-se inviável usar diretamente o levantamento aéreo para aplicações geodésicas. As células de 5' de resolução foram desenvolvidas em levantamentos terrestre e aéreo, do modelo geopotencial global e do modelo digital de terreno.

O Centro Alemão de Pesquisa em Geociências (GFZ) e o Grupo de Pesquisa de Geodesia Espacial (GRGS) / CNES, produziram o modelo de geopotencial global *European Improved Gravity model of the Earth by New techniques* (EIGEN-6C4), grau e ordem 2190, que é um modelo de campo de gravidade global (FÖRSTE et al., 2014; SHAKO et al., 2014). Este modelo é baseado no campo gravitacional e nos dados de estado estacionário do satélite *Ocean Circulation Explorer* (GOCE) e do satélite *Gravity Recovery and Climate Experiment* (GRACE).

O projeto *GOCE Exploitation for Moho Modeling and Applications* (GEMMA), financiado pela Agência Espacial Europeia e pelo Politécnico de Milão, teve como objetivo estimar a profundidade da superfície de Mohorovičić (Moho) através dos dados do GOCE. Moho é uma descontinuidade que delimita as diferenças de propriedades físicas entre a crosta e o manto superior da Terra (FOWLER, 2004). Outro modelo global difundido para estimar a espessura da crosta é o modelo chamado CRUST2.0 (BASSIN; LASKE; MASTERS, 2000), baseado somente em dados sísmicos. Entretanto, o modelo GEMMA contempla o CRUST2.0, contemplando diversos dados considerados confiáveis do ponto de vista técnico e científico. Segundo Reguzzoni e Sampietro (2015) o modelo GEMMA aparenta ser uma melhoria do CRUST2.0.

O *Earth Magnetic Anomaly Grid 2* (EMAG2) é uma malha de resolução espacial de 2" do arco da anomalia magnética de Terra e altitude de 4 km acima do geoide. A malha foi compilado a partir de medições magnéticas aéreas, marinhas e por satélites (MAUS et al., 2009). Este tipo de modelo fornece informações da estrutura e composição da crosta.

Figura 14 - Aerogravimetria (azul) e gravimetria terrestre (vermelho) localizada na região amazônica.

Fonte: Bruno Monserrat Perillo, 2019.

O modelo digital de terreno (MDT) usado neste trabalho (SAM3s_v2), possui uma resolução de 3" de arco, compreendendo um total de 1983 células de grade de 1° x 1° (MATOS et al. 2008). Este modelo é baseado na Shuttle Radar Topography Mission (SRTM) e está vinculado ao modelo geopotencial global EIGEN-GL04C (FORSTE et al., 2008). Esse MDT apresentou melhor consistência nos resultados do modelo geoidal para a América do Sul do que (BLITZKOW et al., 2015) do SRTM original, onde utilizou-se o modelo geopotencial global EGM96 (LEMOINE et al., 1998). A Tabela 2 apresenta informações referente a cada uma das variáveis, como unidade e resolução original.

Tabela 2 – Visão geral das variáveis

	Variáveis	Patrocinador Agencia de Fomento	Unidade	Resolução Espacial do Dado Processado	Resolução Original
1	Ar-livre e Bouguer terrestre	PETROBRAS	mGal	5'	levantamento
2	Ar-livre e Bouguer aéreo	Agência Nacional de Petroleo	mGal	5'	levantamento
3	Modelo Geopotencial Global (EIGEN6C4)	GFZ Potsdam, CNES/GRGS e Kiel University	mGal	5'	Grau e ordem igual 2190
4	GOCE Explotation for Moho Modeling and Aplications (GEMMA)	Agencia Espacial Europeia e Politécnico de Milão	metros	5'	1°
5	Modelo digital de terreno (SAM3s_v2)	Laboratório de Topografia e Geodesia- Universidade de São Paulo	metros	5'	3"
6	Grade de anomalia magnética da Terra (EMAG2)	Comissão do mapa geológico mundial	nT	5'	2'

Fonte: Bruno Monserrat Perillo, 2019.

3.2.2 Pré-processamento dos dados para obtenção das variáveis

3.2.2.1 Ar-livre e Bouguer

Primeiramente, a anomalia ar-livre terrestre pontual foi calculada usando os dados adquiridos dos levantamentos gravimétricos terrestres, sendo estes: coordenadas, altura ortométrica e aceleração da gravidade terrestre. Com esta informação e a correção do terreno, foram obtidas as anomalias Bouguer completa. Posteriormente, uma grade de resolução de 5' foi gerada com coordenadas no centro das células. As médias das anomalias Bouguer foram obtidas pela média simples, utilizando os pontos pertencentes a cada malha. Para este processo, foi utilizado pacote de computacional Generic Mapping Tools (GMT - WESSEL et al., 2013). Posteriormente, as anomalias ar livre médias (terrestre) nas células da malha de 5' foram reconstruídas, usando o SAM3s_v2.

As grades com resolução de 5' das anomalias ar livre e Bouguer, dos dados aerogravimétricos, foram obtidas utilizando spline de curvatura contínua de tensão ajustável do pacote computacional GMT.

3.2.2.2 Organização dos dados

Por se tratar de bases de dados distintos, foi necessário o agrupamento para que todas as variáveis fossem reamostradas a uma mesma resolução. Para tanto,

criou-se uma malha com células de 5' de resolução espacial dos levantamentos terrestre e aéreo, do modelo geopotencial global e do modelo digital de terreno. Não houve necessidade de normalizar os dados por serem dados de mesma ordem de grandeza. Este conjunto de dados totalizou 6912 amostras.

Tendo em vista a validação dos resultados imputados, foram criadas duas áreas com exclusão parcial de dados (janela - *gaps*), permitindo assim a comparação dos dados terrestres existentes para com os valores sintéticos gerados. Tal como descrito em Schafer e Graham (2002), esse formato fornece a avaliação quantitativa e acurácia dos resultados. Além disso, como as variáveis de interesse possuem valores ausentes, as amostras relativas a tais pontos foram descartadas para que possa ser avaliado e quantificado a imputação nas áreas de janelas. Dessa forma, o segundo conjunto de dados passou a ter 5185 amostras.

3.2.3 Processamento SOM

A análise SOM foi realizada através do software SiroSOM (FRASER; DICKSON, 2007). Este software é uma interface gráfica baseada no SOM toolbox escrita em Matlab por Vesanto et al. (2000).

Os parâmetros iniciais escolhidos para a configuração do SOM foram, inicialização randômica, grade hexagonal e hipervolume toroidal para a projeção dos neurônios BMU's.

O treinamento SOM foi desenvolvido em três tamanhos diferentes de mapas: (A) 14 x 14; (B) 20 x 20; e (C) 26 x 26. Foram definidos três experimentos para verificar a influência do parâmetro do tamanho do mapa (MS). A análise foi desenvolvida por variações da Equação 16 proposta por Vesanto e Alhoniemi (2000). $N = 5185$, é a quantidade de dados. Nesse sentido, foram testados MS's com tamanho correspondente a Equação 16 dividido por 2, representado na Equação 15, e Equação 16 multiplicado por 2 representado na Equação 17, conforme descrito a seguir:

$$MS = \frac{5\sqrt{N}}{2} \cong 14 \times 14 \quad (15)$$

$$MS = 5\sqrt{N} \cong 20 \times 20 \quad (16)$$

$$MS = 10\sqrt{N} \cong 26 \times 26 \quad (17)$$

Os tamanhos de mapas foram aproximadamente escolhidos de forma que formassem uma matriz quadrada e a escolha por números pares por se tratar de uma projeção toroidal. A seguir (Tabela 3) são apresentados os parâmetros de treinamentos com os respectivos erros.

Tabela 3 – Parâmetros de inicialização do SOM

Dimensão	Inicialização	Grade	Projeção	Erro	
				Quantização	Topográfico
(A) 14 x14	Randômica	hexagonal	toroidal	0,467	0,191
(B) 20x20				0,378	0,15
(C) 26x26				0,322	0,12

Fonte: Bruno Monserrat Perillo, 2019.

Os resultados foram comparados e interpretados usando avaliações estatísticas, especificamente, matrizes de correlação.

3.2.4 Síntese do Fluxo Metodológico

Uma síntese da sequência metodológica pode ser vista na Figura 15. As variáveis foram inicialmente analisadas para observar correlações que podem gerar tendências espúrias aos resultados. Assim, foram excluídas as variáveis com correlações próximas de -1, 0 e 1. Foram excluídas as variáveis EMAG2, altura geoidal e modelo digital de terreno, que apresentaram correlação predominantemente baixa em relação às demais.

As demais variáveis foram ajustadas para garantir a correspondência espacial entre as amostras. Em seguida, foram criadas janelas (*gap*) nos dados originais, para posterior comparação com os valores sintéticos.

Figura 15 – Procedimento para imputação do SOM

Fonte: Bruno Monserrat Perillo, 2019.

3.3 ABORDAGEM 2 - INCREMENTO DE RESOLUÇÃO EM VARIÁVEIS GEOFÍSICAS GAMAESPECTROMÉTRICAS

3.3.1 Aquisição e características dos dados gamaespectrométricos e demais variáveis envolvidas

Para possibilitar a avaliação da eficácia do método SOM no incremento de resolução, foi necessária a preparação de duas bases de dados distintas simulando problemas comumente encontrados na prospecção geofísica – baixa resolução associada a limitações nas taxas de amostragem. As bases de dados foram compostas por mais de 1.600.000 amostras para as 4 variáveis de interesse (contagem total, potássio, tório e urânio).

Com o objetivo de estudar o incremento de resolução, a abordagem escolhida neste trabalho foi preparar uma base de dados que mantivesse a resolução original (125 m) para as variáveis CT, K e U, mas na qual a variável original Th fosse reamostrada com uma resolução decimada de quatro vezes em relação à original, resultando em uma grade com amostras a cada 500m. Um segundo conjunto de dados decimado em oito vezes também foi preparado, totalizando uma degradação de 1000m.

Ao todo foram realizados quatro processamentos SOM. Foram processados os dois conjuntos acima citados, incluindo ou não a variável CT afim de investigar a influência de variáveis de alta correlação, pela alta correlação entre CT e Th.

A síntese da preparação das bases de dados tanto para o estudo do incremento de resolução quanto para o estudo da imputação pode ser observada no fluxograma de processamento simplificado, contido na Figura 16.

3.3.2 Processamento SOM

Os parâmetros iniciais escolhidos para a configuração do SOM foram, inicialização randômica, grade hexagonal e hipervolume toroidal para a projeção dos neurônios BMU's (Tabela 4).

O mapa SOM utilizado foi de 80 x 80, definido pela expressão proposta por Vesanto e Alhoniemi (2000). Os resultados foram comparados e interpretados usando avaliações estatísticas.

Tabela 4 – Parâmetros de inicialização do SOM

	Th500	Th500ct	Th1000	Th1000ct
Variáveis	K e U	K, U e CT	K e U	K, U e CT
Resolução das variáveis de entrada	125m			
Resolução de Th	500m	500m	1000m	1000m
Tamanho do mapa	80 x 80			
Inicialização	Randômica			
Forma	Toroidal			
Grade	Hexagonal			
Vizinhança	Gaussiano			
qe	0,119	0,153	0,121	0,194
te	0,063	0,0683	0,0608	0,0596

Fonte: Bruno Monserrat Perillo, 2019.

Figura 16 - Fluxograma da preparação da base de dados e processamento.

A: Incremento de Resolução. B: Imputação de Dados.

Fonte: Bruno Monserrat Perillo, 2019.

4 RESULTADOS

4.1 ABORDAGEM 1 – IMPUTAÇÃO DE DADOS GRAVIMÉTRICOS

4.1.1 Análise Exploratória das Variáveis

Este subcapítulo apresenta a imputação de dados gravimétricos terrestres utilizando outras variáveis geofísicas/geodesicas que, a partir das similaridades destas variáveis no espaço n-dimensional, possam imputar valores plausíveis para a variável desejada. É também foco, discutir a influência do tamanho do mapa, utilizando três diferentes tamanhos de mapas a priori. Na pesquisa foram utilizados diferentes tamanhos de mapas com a finalidade de investigar a influência deste parâmetro nos resultados.

Com uma abordagem exploratória dos dados, foram plotados as Figura 17 a Figura 25 que apresentam as anomalias interpoladas, pontos amostrados, histogramas e a estatística básica de cada variável. Tais variáveis, de modo geral, apresentam feições similares apesar de diferir as propriedades geofísicas, magnitudes e unidades medidas.

Duas questões cruciais para trabalhos que envolvem mineração de dados dizem respeito à redução de dimensionalidade e à seleção de atributos, pois ambas dependem da estrutura da base de dados. Estas questões estão intimamente ligadas à quantidade de amostras e variáveis. As anomalias Bouguer e ar livre terrestres, representados pelas Figuras 18 e 20, evidenciam a falta de informações em vários pontos das imagens. Essa ausência de dados enfatiza, portando, a justificativa desta pesquisa.

Tendo em vista a identificação de variáveis pouco correlacionadas que poderiam induzir a erros de classificação no espaço n-dimensional, foram desenvolvidas matrizes de correlação. Em uma primeira análise, a matriz apresentada na Figura 26 levou em conta todas as variáveis utilizadas nas análises SOM.

A matriz de correlação apresenta as correlações entre as variáveis. Numa abordagem visual, a diagonal principal é rotulada e mostra o histograma de cada variável. Os coeficientes de correlação Pearson e valor-p encontram-se acima da diagonal principal. Na parte inferior, gráficos de dispersão com linha de tendência

média em vermelho. A probabilidade de significância (valor-p) é representado por '***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$ e em branco para $p > 0,5$.

Foram identificadas três variáveis com baixa correlação. Nesse sentido, as variáveis EMAG2, altura geoidal e modelo de terreno digital não foram consideradas, já que obtiveram predominantemente correlações inferiores a 0,005 (Figura 26). As exclusões de variáveis com baixas correlações não foram meramente empíricas. Após sucessivos processamentos SOM, testando a exclusão de variável por variável, admitiu-se a retirada de variáveis com baixíssimas correlações.

A definição da eliminação de variáveis com baixas correlações deve ser cautelosa, pois mesmo baixas correlações (relativas) podem contribuir significativamente para o resultado. A seguir, serão apresentadas as análises exploratórias de todas variáveis envolvidas inicialmente e já reamostradas com resolução espacial de 5'.

Figura 17 - Análise exploratória da anomalia Bouguer aérea. (A) Anomalia interpolada, (B) pontos amostrados, (C) histograma e informações estatísticas.

Anomalia Bouguer – gravimetria aérea (mGal)

Fonte: Bruno Monserrat Perillo, 2019.

Figura 18 - Análise exploratória da anomalia Bouguer terrestre. (A) Anomalia interpolada, (B) pontos amostrados, (C) histograma e informações estatísticas.

Anomalia Bouguer – gravimetria terrestre (mGal)

Fonte: Bruno Monserrat Perillo, 2019.

Figura 19 - Análise exploratória da anomalia ar livre aérea. (A) Anomalia interpolada, (B) pontos amostrados, (C) histograma e informações estatísticas.

Anomalia ar livre – gravimetria aérea (mGal)

Fonte: Bruno Monserrat Perillo, 2019.

Figura 20 - Análise exploratória da anomalia ar livre terrestre. (A) Anomalia interpolada, (B) pontos amostrados, (C) histograma e informações estatísticas.

Anomalia ar livre – gravimetria terrestre (mGal)

Fonte: Bruno Monserrat Perillo, 2019.

Figura 21 - Análise exploratória do distúrbio da gravidade. (A) Anomalia interpolada, (B) pontos amostrados, (C) histograma e informações estatísticas.

Distúrbio da gravidade (mGal)

Fonte: Bruno Monserrat Perillo, 2019.

Figura 22 - Análise exploratória do modelo da Moho. (A) Anomalia interpolada, (B) pontos amostrados, (C) histograma e informações estatísticas.

Profundidade Moho (m)

Fonte: Bruno Monserrat Perillo, 2019.

Figura 23 - Análise exploratória do modelo digital de terreno. (A) Anomalia interpolada, (B) pontos amostrados, (C) histograma e informações estatísticas.

Modelo Digital de Terreno (m)

Fonte: Bruno Monserrat Perillo, 2019.

Figura 24 - Análise exploratória da altura geoidal. (A) Anomalia interpolada, (B) pontos amostrados, (C) histograma e informações estatísticas.

Altura geoidal (m)

Fonte: Bruno Monserrat Perillo, 2019.

Figura 25 - Análise exploratória da anomalia magnética da Terra. (A) Anomalia interpolada, (B) pontos amostrados, (C) histograma e informações estatísticas.

Anomalia Magnética da Terra (mT)

Fonte: Bruno Monserrat Perillo, 2019.

Figura 26 - Matriz de correlação e gráficos de dispersão e histograma das variáveis envolvidas na análise exploratória.

Fonte: Bruno Monserrat Perillo, 2019.

4.1.2 Mapas Auto-Organizados produzidos pela análise exploratória das variáveis

A representação da Matriz-U é uma exibição destinada a resumir a estrutura e ilustrar a complexidade inerente a um conjunto de dados (Figura 27). Os nós na representação da Matriz-U são coloridos usando uma escala de temperatura de cor para indicar o grau de similaridade. Cores azuis indicam neurônios próximos, vermelho neurônios distantes bem como amarelo regiões de fronteiras, indicando, portanto, uma similaridade média.

Na Figura 27 as variáveis MDT, EMAG2 e Altura Geoidal apresentam valores com distribuição diferente das demais variáveis segundo a disposição no mapa auto-organizado.

Figura 27 - Matriz-U e planos das componentes das variáveis envolvidas na análise exploratória das variáveis.

Fonte: Bruno Monserrat Perillo, 2019.

4.1.3 Análise SOM das variáveis específicas

Os planos de componentes de três tamanhos de mapas (Figuras 28, 29, 30) demonstram uma forte correlação das variáveis, independentemente do tamanho do

mapa. As diferenças nas posições de anomalia do mapa estão relacionadas, exclusivamente, à inicialização da projeção 2D do toróide.

Figura 28 - Matriz-U e planos das componentes das anomalias Ar-livre e Bouguer obtidas por levantamentos gravimétricos aéreo e terrestre, distúrbio da gravidade e modelo da profundidade de Moho. Baseado no tamanho de mapa 14 x 14.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 29 - Matriz-U e planos das componentes das anomalias Ar-livre e Bouguer obtidas por levantamentos gravimétricos aéreo e terrestre, distúrbio da gravidade e modelo da profundidade de Moho. Baseado no tamanho de mapa 20 x 20.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 30 - Matriz-U e planos das componentes das anomalias Ar-livre e Bouguer obtidas por levantamentos gravimétricos aéreo e terrestre, distúrbio da gravidade e modelo da profundidade de Moho. Baseado no tamanho de mapa 26 x 26.

Fonte: Bruno Monserrat Perillo, 2019.

4.1.4 Validação dos valores imputados em relação aos originais

O método SOM foi aplicado para prever valores ausentes nas áreas de ausências de dados (ou lacunas). Em seguida, os resultados foram comparados com os valores conhecidos. As informações foram omitidas em duas áreas (Figura 31) dentro das coordenadas 2°S-3°S / 55°W-56°W e 1°S-2°S / 46°W-47°W (6% da área total). Os valores da média, desvio padrão, máximo, mínimo e coeficiente de determinação (R^2) da diferença entre o conjunto de dados e os resultados sintéticos do método SOM usando um conjunto de dados com valores omissos são apresentados na Tabela 5.

Figura 31 – Duas janelas onde foram aplicados o método SOM para prever anomalia Ar livre e anomalia de Bouguer.

Fonte: Bruno Monserrat Perillo, 2019.

Nas áreas de janela, após análise de MS com 14 x 14, os valores de R^2 apresentaram 92% para a anomalia ar livre e 90% para anomalia Bouguer. Utilizando um MS 20 x 20, o coeficiente de determinação apontou 92% e 91%, para anomalia ar livre e Bouguer, respectivamente. O terceiro experimento, MS 26 x 26, o R^2 apresentou 92% de anomalia ar livre e 91% de anomalia de Bouguer. A média dos BMUs referente a cada tamanho de mapa apresentou 94% e 93% para anomalia ar livre e anomalia Bouguer, respectivamente (Tabela 5).

As Figuras 32 e 33 mostram as anomalias ar livre e Bouguer de acordo com os mapas originais e imputados, respectivamente. As Figuras 34 e 35, apresentam os

dados originais e os obtidos pela média dos BMUs para ambas anomalias. Outros mapas são apresentados como a diferença entre os valores originais e os imputados da anomalia ar livre e Bouguer para cada tamanho de mapa (Figuras 36 e 37), respectivamente. Adiante, serão apresentados nas Figuras 38 e 39, a predição em detalhe nas duas áreas de janelas, janela 1 e janela 2, para ambas anomalias.

Os coeficientes de determinação e histogramas (Figuras 40 e 41) entre os dados originais e os dados obtidos através da imputação em diferentes tamanhos de mapa de SOM (14 x 14, 20 x 20 e 26 x 26) e nas regiões de janelas (Figuras 42 e 43). As Tabelas 6 e 7 demonstram os valores estatísticos dos dados obtidos das variáveis analisadas nos diferentes tamanhos de mapas, bem como os coeficientes de correlação e a média das BMUs.

As Figuras 44, 45, 46 e 47 mostram as regiões de imputação das variáveis ar livre e Bouguer terrestre em diferentes tamanhos de mapa, completando todas as células ausentes (janelas) nos dados originais. Visualmente, o mapa de maior tamanho apresentou o melhor resultado por preservar as magnitudes do dado original (Figuras 44c e 45c).

Tabela 5 – Comparação entre diferentes dimensões do tamanho de mapa: (A) Matriz 14 x 14; (B) 20 x 20; (C) 26 x 26 e (D)

Média dos BMU 14 x 14, 20 x 20 e 36 x 26.

(A) Análise A MS 14 x 14	Resolução Espacial da Área Total (valores originais 5' x BMU X')						Resolução Espacial do GAP	
	Ar-livre aéreo	Bouguer aéreo	Disturb Grav	Moho Gemma	Ar-livre Ter	Bouguer Ter	Ar-livre Ter	Bouguer Ter
Média (dif)	-0,081	-0,085	-0,019	0,000	0,128	0,285	-2,760	-1,758
Desv pad (dif)	6,089	6,045	5,722	1,206	7,615	6,369	7,339	6,989
Máx (dif)	32,499	31,145	25,187	4,594	74,153	71,360	21,942	26,662
Mín (dif)	-41,308	-41,791	-29,398	-5,883	-61,511	-40,598	-27,944	-23,528
R ²	0,943	0,950	0,927	0,944	0,923	0,947	0,924	0,908
(B) Análise B MS 20 x 20	Resolução Espacial da Área Total (valores originais 5' x BMU X')						Resolução Espacial do GAP	
	Ar-livre aéreo	Bouguer aéreo	Disturb Grav	Moho Gemma	Ar-livre Ter	Bouguer Ter	Ar-livre Ter	Bouguer Ter
Média (dif)	-0,032	-0,061	-0,009	-0,007	0,085	0,008	-1,975	-1,295
Desv pad (dif)	4,964	4,652	4,682	0,965	6,072	5,414	7,793	7,065
Máx (dif)	32,280	32,853	26,474	5,911	56,119	62,656	22,163	25,733
Mín (dif)	-37,041	-29,894	-25,151	-5,312	-40,803	-38,756	-23,061	-22,453
R ²	0,961	0,969	0,950	0,963	0,952	0,964	0,922	0,910
(C) Análise C MS 26 x 26	Resolução Espacial da Área Total (valores originais 5' x BMU X')						Resolução Espacial do GAP	
	Ar-livre aéreo	Bouguer aéreo	Disturb Grav	Moho Gemma	Ar-livre Ter	Bouguer Ter	Ar-livre Ter	Bouguer Ter
Média (dif)	0,000	-0,001	-0,007	0,001	0,037	-0,028	-2,234	-1,703
Desv pad (dif)	4,040	3,903	3,947	0,826	5,194	4,700	8,069	7,061
Máx (dif)	26,531	27,426	21,126	4,132	54,193	62,312	20,485	27,761
Mín (dif)	-35,877	-28,516	-23,076	-5,094	-39,287	-29,474	-21,892	-21,761
R ²	0,974	0,978	0,964	0,972	0,964	0,972	0,920	0,913
(D) Análise D MS médio	Resolução Espacial da Área Total (valores originais 5' x BMU X')						Resolução Espacial do GAP	
	Ar-livre aéreo	Bouguer aéreo	Disturb Grav	Moho Gemma	Ar-livre Ter	Bouguer Ter	Ar-livre Ter	Bouguer Ter
Média (dif)	-0,037	-0,049	-0,012	-0,002	0,049	-0,006	-2,323	-1,585
Desv pad (dif)	4,660	4,468	4,288	0,860	5,917	5,248	7,237	6,456
Máx (dif)	28,012	26,609	23,797	3,714	58,237	65,443	21,530	26,719
Mín (dif)	-34,216	-28,518	-25,173	-4,719	-46,790	-34,442	-23,497	-21,260
R ²	0,967	0,973	0,960	0,973	0,955	0,967	0,939	0,927

Figura 32 – (A) Anomalia ar-livre original. (B) Imputado pelo SOM usando MS 14 x 14, (D) MS 20 x 20 e (C) 26 x 26.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 33 – (A) Anomalia Bouguer original. (B) Predito pelo SOM usando MS 14 x 14, (D) MS 20 x 20 e (C) 26 x 26.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 34 - (A) Anomalia ar-livre original. (B) Imputado pelo SOM usando o BMU médio.

Figura 35 - (A) Anomalia Bouguer original. (B) Imputado pelo SOM usando o BMU médio.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 36 – (A) Anomalia ar livre original. A diferença entre o original e o previsto pelo SOM é mostrado com os seguintes tamanhos de mapas (B) MS 14 x 14, (C) MS 20 x 20 e (D) 26 x 26.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 37 – (A) Anomalia Bouguer original. A diferença entre o original e o previsto pelo SOM é mostrado com os seguintes tamanhos de mapas (B) MS 14 x 14, (C) MS 20 x 20 e (D) 26 x 26.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 38 - Imagens espacializadas da anomalia ar-livre (A, B, C, D e E) e Bouguer (F, G, H, I e J) na área de imputação janela 1, de acordo com a seguintes disposições: (A e F) Original, (B e G) 14 x 14, (C e H) 20 x 20, (D e I) 26 x 26 e (E e J) Média dos BMUs.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 39 - Imagens da anomalia ar-livre (A, B, C, D e E) e Bouguer (F, G, H, I e J) na área de imputação janela 2, de acordo com a seguintes disposições: (A e F) Original, (B e G) 14 x 14, (C e H) 20 x 20, (D e I) 26 x 26 e (E e J) Média dos BMUs.

Fonte: Bruno Monserrat Perillo, 2019.

Tabela 6 – Análise estatística da janela 1 da anomalia ar-livre (A) e Bouguer (B) para diferentes tamanhos de mapa (mGal).

(A)	Ar-Livre Terrestre			
	14x14	20x20	26x26	Média
Média (dif)	0,483	1,292	1,476	1,083
Desv pad (dif)	6,519	7,328	7,364	6,661
Máx (dif)	21,942	22,163	20,485	21,530
Mín (dif)	-17,075	-18,752	-16,773	-14,701
R²	0,941	0,935	0,938	0,949

(B)	Bouguer Terrestre			
	14x14	20x20	26x26	Média
Média (dif)	-0,174	0,381	0,517	0,241
Desv pad (dif)	6,175	6,648	6,329	5,881
Máx (dif)	26,662	25,733	27,761	26,719
Mín (dif)	-14,79	-19,138	-16,034	-15,525
R²	0,923	0,911	0,920	0,930

Fonte: Bruno Monserrat Perillo, 2019.

Tabela 7 – Análise estatística da janela 2 da anomalia ar-livre (A) e Bouguer (B) para diferentes tamanhos de mapa (mGal).

(A)	Ar-Livre Terrestre			
	14x14	20x20	26x26	Média
Média (dif)	-6,098	-5,338	-6,054	-5,830
Desv pad (dif)	6,621	6,779	6,916	6,047
Máx (dif)	9,109	12,543	7,728	9,793
Mín (dif)	-27,944	-23,061	-21,892	-23,497
R²	0,869	0,866	0,861	0,899

(B)	Bouguer Terrestre			
	14x14	20x20	26x26	Média
Média (dif)	-3,388	-3,021	-3,988	-3,465
Desv pad (dif)	7,413	7,089	7,068	6,503
Máx (dif)	14,431	15,703	12,036	12,327
Mín (dif)	-23,528	-22,453	-21,761	-21,260
R²	0,834	0,853	0,858	0,884

Fonte: Bruno Monserrat Perillo, 2019.

Figura 40 – Gráficos de correlação de ar livre original vs ar livre imputado. (A) 14x14, (B) 20x20, (C) 26x26 e (D) BMU médio

Fonte: Bruno Monserrat Perillo, 2019.

Figura 41 – Gráficos de correlação de Bouguer original x Bouguer imputado. (A) 14x14, (B) 20x20, (C) 26x26 e (D) BMU médio

Fonte: Bruno Monserrat Perillo, 2019.

Figura 42 – Gráficos de correlação de ar livre original vs ar livre imputado nas áreas de janela. (A) 14x14, (B) 20x20, (C) 26x26 e (D) BMU médio

Fonte: Bruno Monserrat Perillo, 2019.

Figura 43 – Gráficos de correlação de Bouguer original x Bouguer imputado nas áreas de janela. (A) 14x14, (B) 20x20, (C) 26x26 e (D) BMU médio

Fonte: Bruno Monserrat Perillo, 2019.

4.1.5 Imputação de dados gravimétricos da área total

Figura 44 – (A) Anomalia ar-livre original. (B) Imputado pelo SOM usando MS 14 x 14, (D) MS 20 x 20 e (C) 26 x 26.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 45 - (A) Anomalia Bouguer original. (B) Imputado pelo SOM usando MS 14 x 14, (D) MS 20 x 20 e (C) 26 x 26.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 46 - (A) Anomalia ar-livre original. (B) Imputado pelo SOM usando o BMU médio.

Figura 47 - (A) Anomalia Bouguer original. (B) Imputado pelo SOM usando o BMU médio.

Fonte: Bruno Monserrat Perillo, 2019.

4.1.6 Discussão sobre os resultados imputados

Os resultados apresentados nas Figuras 42 e 43 demonstram que as imputações obtidas na janela apresentam boa correspondência com os dados originais, atingindo coeficientes com R^2 iguais a 94 e 93% de, para Ar Livre e para Anomalias Bouguer, respectivamente.

Em todas as situações, o histograma aponta para uma clara correspondência entre a distribuição dos valores gerados através da imputação e os valores originais. Os histogramas apresentados nas Figuras 40, 41, 42 e 43, no entanto, não obedecem a um padrão específico relacionado aos tamanhos dos mapas

A análise de parâmetros estatísticos (Tabelas 5, 6 e 7) aponta para uma melhor correlação quando os valores médios dos BMUs são usados em diferentes tamanhos de mapa.

Visualmente, as anomalias apresentadas por imputação demonstram correspondência aos dados originais, com leves divergências. Os resultados obtidos para tamanhos de mapas de maiores (26 x 26) apresentaram maior correspondência visual aos dados originais, corroborando com os indicadores de correlação.

4.2 ABORDAGEM 2 – INCREMENTO DE RESOLUÇÃO EM VARIÁVEIS GEOFÍSICAS GAMAESPECTROMÉTRICAS

Este subcapítulo apresenta os resultados da segunda abordagem desta dissertação. Os SOM, até este ponto, foram utilizados exclusivamente com o intuito de imputação de valores ausentes. Entretanto, entendemos o incremento de resolução em uma imagem como uma técnica que envolve imputação regular. Neste sentido, o capítulo irá abordar o incremento de dados gamaespectrométricos, em específico a variável Th, que por ora possui resolução de 125m juntamente com todos os outros canais, CT, U e K. Para avaliar a eficácia do SOM como incrementador a variável Th fora degradada em 500 e 1000m, ou seja, para cada valor de Th possui 4 ou 8 valores possivelmente distintos, respectivamente.

A validação dos resultados obtidos com os mapas auto-organizáveis foi realizada em 3 análises:

1. comparação visual entre o mapa original e com resolução espacial incrementada;
2. mapa da diferença entre a variável original e a com resolução espacial incrementada.
3. Análise estatística.

4.2.1 Análise Exploratória das Variáveis

Nas Figuras 48 e 49 são apresentadas as análises exploratórias dos dados utilizados nesta abordagem, que incluem mapas no espaço geográfico, histogramas e valores estatísticos de cada variável.

Observa-se que os mapas possuem uma feição geral coerente entre si, entretanto, há detalhes específicos em algumas variáveis, em função da técnica de investigação geofísica em resposta à natureza geológica. O intuito deste trabalho é exatamente a possibilidade de usar variáveis diferentes que sejam espacialmente corregistradas, para o incremento de resolução espacial.

A variável CT possui visualmente uma imagem extremamente semelhante ao Th. Entretanto, possuem unidades físicas e valores estatísticos distintos. As variáveis Th500 e Th1000 (Figura 49) foram originadas a partir da variável Th, que teve sua resolução espacial diminuída ou degradada. A degradação mantém o padrão e os valores semelhantes ao dado original de Th. No mapa da variável Th1000 é possível observar a perda da resolução.

A matriz de correlação (Figura 50) representa três diferentes informações: (i) histogramas na diagonal principal; (ii) acima, os coeficientes de correlação e valor-p; e (iii) abaixo, os gráficos de dispersão. A variável AS foi excluída da análise devido à baixa correlação entre as outras variáveis envolvidas. A variável Th está representada na matriz, mas não foi utilizada em nenhum processamento, justamente por ser o alvo de investigação.

Decidiu-se inicialmente fazer dois diferentes processamentos, para investigar a funcionalidade e eficiência para diferentes degradações, utilizando CT, U, K e a variável Th degradada (500 ou 1000m). Entretanto, devido a altíssima correlação entre as variáveis CT e Th, optou-se em refazer estes processamentos, dessa vez excluindo a variável CT.

Figura 48 - Análise exploratória de U e K, Anomalia interpolada, histogramas e informações estatísticas.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 49 - Análise exploratória de CT, Th, Th500 e Th1000 Anomalia interpolada, histogramas e informações estatísticas.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 50 - Matriz de correlação e gráficos de dispersão e histograma das variáveis envolvidas na análise exploratória.

Fonte: Bruno Monserrat Perillo, 2019.

4.2.2 Mapas Auto-Organizados produzidos pela análise exploratória das variáveis

Como análise exploratória também foram desenvolvidos os planos das componentes das variáveis envolvidas a fim de investigar a similaridades entre tais variáveis. Na matriz-U (Figura 51) a variável AS (amplitude do sinal analítico) apresenta alta dissimilaridade com as demais variáveis. Sendo, portanto, excluída do trabalho.

Figura 51 - Matriz-U e planos das componentes das variáveis envolvidas na análise exploratória das variáveis.

Fonte: Bruno Monserrat Perillo, 2019.

4.2.3 Análise SOM das variáveis específicas

A Matriz-U obtida é representada nas Figuras 52, 53, 54 e 55 de modo que a escala de cores indica as semelhanças entre nós adjacentes no mapa da seguinte maneira: as cores frias (azul) indicam grande similaridade entre os nós adjacentes e as cores quentes (vermelho) menor similaridade.

Cada variável influi para a geração desta Matriz-U já que, para cada par de amostras, tem-se variáveis semelhantes ou distintas. Como cada nó é um vetor no espaço definido por cada variável é possível visualizar a contribuição de cada variável através da mesma escala de temperatura de cores da Matriz-U. Essa influência é representada por meio de *component plots*.

Figura 52 - Matriz-U e planos das componentes dos canais K, U e Th degradada em 500m obtidas por levantamentos aerogamaespectrométricos.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 53 - Matriz-U e planos das componentes dos canais CT, K, U e Th degradada em 500m obtidas por levantamentos aerogamaespectrométricos.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 54 - Matriz-U e planos das componentes dos canais K, U e Th degradada em 1000m obtidas por levantamentos aerogamaespectrométricos.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 55 - Matriz-U e planos das componentes dos canais CT, K, U e Th degradada em 1000m obtidas por levantamentos aerogamaespectrométricos.

Fonte: Bruno Monserrat Perillo, 2019.

4.2.4 Validação dos valores incrementados em relação aos originais

O método SOM foi aplicado para incrementar a resolução espacial da variável Th que foi inicialmente degradada, simulado uma variável de baixa resolução. A degradação ocorreu em duas etapas, 500m (Figura 56) e 1000m (Figura 57). Além disso, foi processado simulando a exclusão da variável CT afim de verificar a influência da alta correlação entre as variáveis CT e Th.

Figura 56 - Anomalia interpolada, histograma e estatística dos canais Th original, Th degradada em 500m com e sem CT.

Fonte: Bruno Monserrat Perillo, 2019.

Observa-se que visualmente (Figuras 56 e 57) considerando ou não a variável CT, aparentam ter a mesma feição. Além disso, não há discrepância nos parâmetros estatísticos.

Figura 57 - Anomalia interpolada, histograma e estatística dos canais Th original, Th degradada em 1000m com e sem CT.

Fonte: Bruno Monserrat Perillo, 2019.

As Figuras 58 e 59, mostram as diferenças entre os valores originais e os incrementados em cada degradação, apresentando valores homogêneos (próximos a zero). Adiante, são apresentados nas Figuras 60 e 61, mostram a baixa dispersão entre os dados originais e imputados.

Figura 58 - (A) Th original – Th degradado em 500m. (B) Th original – Th recuperado utilizando CT. (C) Th original – Th recuperado sem CT.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 59 - (A) Th original – Th degradado em 1000m. (B) Th original – Th recuperado utilizando CT. (C) Th original – Th recuperado sem CT.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 60 – Gráficos de correlação. (A) Th degradado em 500m vs Th original. (B) Th recuperado utilizando CT vs Th original. (C) Th recuperado sem CT vs Th original

Fonte: Bruno Monserrat Perillo, 2019.

Figura 61 – Gráficos de correlação. (A) Th degradado em 1000m vs Th original. (B) Th recuperado utilizando CT vs Th original. (C) Th recuperado sem CT vs Th original.

Fonte: Bruno Monserrat Perillo, 2019.

4.2.5 Incremento de resolução de dados gamaespectrométricos em detalhes

Nas Figura 62 a Figura 69, são apresentados detalhes (zoom) em diferentes áreas para comparar a eficácia do métodos SOM em cada um dos processamentos em relação ao original. Chama-se a atenção os processamentos realizados considerando a variável CT pela eficiente recuperação.

Figura 62 – Anomalia interpolada (A) Th original, (B) Th degradado em 500m, (C) Th recuperado com resolução de 125m e (D) Th recuperado utilizando a variável CT, com resolução de 125m.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 63 – Anomalia interpolada (A) Th original, (B) Th degradado em 500m, (C) Th recuperado com resolução de 125m e (D) Th recuperado utilizando a variável CT, com resolução de 125m.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 64 – Anomalia interpolada (A) Th original, (B) Th degradado em 500m, (C) Th recuperado com resolução de 125m e (D) Th recuperado utilizando a variável CT, com resolução de 125m.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 65 – Anomalia interpolada (A) Th original, (B) Th degradado em 500m, (C) Th recuperado com resolução de 125m e (D) Th recuperado utilizando a variável CT, com resolução de 125m.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 66 – Anomalia interpolada (A) Th original, (B) Th degradado em 1000m, (C) Th recuperado com resolução de 125m e (D) Th recuperado utilizando a variável CT, com resolução de 125m.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 67 – Anomalia interpolada (A) Th original, (B) Th degradado em 1000m, (C) Th recuperado com resolução de 125m e (D) Th recuperado utilizando a variável CT, com resolução de 125m.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 68 – Anomalia interpolada (A) Th original, (B) Th degradado em 1000m, (C) Th recuperado com resolução de 125m e (D) Th recuperado utilizando a variável CT, com resolução de 125m.

Fonte: Bruno Monserrat Perillo, 2019.

Figura 69 – Anomalia interpolada (A) Th original, (B) Th degradado em 1000m, (C) Th recuperado com resolução de 125m e (D) Th recuperado utilizando a variável CT, com resolução de 125m.

Fonte: Bruno Monserrat Perillo, 2019.

5 CONCLUSÕES

Esta dissertação está inserida no contexto de mineração de dados utilizando-se aprendizado de máquina não supervisionado. A técnica SOM tem sido amplamente utilizada em agrupamento de banco de dados complexos e multivariados, permitindo um melhor entendimento das interdependências não-lineares entre as variáveis.

Com o intuito de entender como o incremento de resolução se comporta, a técnica SOM foi aplicada aos dados geofísicos. Uma vez que o SOM permite imputar valores ausentes em intervalos sem leituras para determinada variável e consequentemente incrementar a resolução de variáveis amostradas com baixa resolução tendo como base a dependência espacial de variáveis regionalizadas no espaço n-dimensional.

Numa primeira abordagem, que utilizou exclusivamente a imputação de dados em local de difícil acesso como a região Norte do Brasil, foram utilizados dados de gravimetria terrestre e aérea, distúrbio da gravidade e o modelo GEMMA, que é uma estimativa da espessura da crosta. Os diferentes tamanhos de mapas aplicados nas análises de SOM mostram pouca discrepância nos valores absolutos gerados entre a correlação dos dados sintéticos com os dados originais. Entretanto, quando utilizada a média dos BMUs, composto pelos BMUs dos três mapas citados, os resultados apresentaram sutil melhora em relação aos mapas com diferentes tamanhos. Analisando o conjunto de dados, é possível perceber que cada tamanho de mapa possui determinados valores *outliers*. Neste sentido, uma filtragem dos dados através da média dos BMUs, minimizaram máximos e mínimos em cada amostra.

Uma alta correlação foi observada entre os dados originais e sintéticos produzidos nos BMUs. Os valores sintéticos médios da anomalia ar livre e Bouguer terrestre apresentaram uma correlação de 94% e 93% com os originais, respectivamente.

Na segunda abordagem, em que foram utilizados dados aerogamaespectrométricos com amostragem a cada 125m, na área da RENCA, A variável tório foi degradada para 500 e 100m. Tendo como base a arquitetura das variáveis conjuntas no espaço n-dimensional com eixos formados pela contagem total, urânio e potássio, foi possível incrementar a resolução do tório, obtendo $R^2 = 0,98$ e $R^2 = 0,94$ nas degradações de 500 e 1000m, respectivamente. O mapa da diferença

apresenta tendência central próxima à zero e distribuição normal. Essas análises validam o incremento de resolução pela técnica Self-Organizing Maps, gerando possibilidades de desenvolvimento de incremento de resolução de variáveis com baixa resolução em áreas compostas por outras variáveis de maior resolução espacial.

Verificou-se que não houve vazamento de dados causados por alta correlação entre as variáveis CT e Th. Dessa forma, verificamos que quando a base de dados é constituída de variáveis que possuem regimes de correlação suficientes, o incremento de resolução pode se tornar vantajoso, possibilitando informação incremental dadas as relações de interdependência não-lineares entre estas.

As análises de SOM podem ser amplamente utilizadas para a geração de dados sintéticos em levantamentos geofísicos, especialmente em regiões de difícil acesso, como a região amazônica. Entretanto, estudos mais aprofundados sobre as variáveis envolvidas, bem como as melhores rotinas de pré-processamento ainda não são bem definidas na literatura, ficando de motivação para trabalhos futuros com esta metodologia, igualmente, uma investigação mais detalhada da influência do tamanho de mapa com dimensões extrapoladas.

REFERÊNCIAS

- ABDUL MALEK, M. et al. Imputation of Time Series Data via Kohonen Self Organizing Maps in the Presence of Missing Data. **Engineering and Technology**, v. 41, 2008.
- ALTMAN, N.; KRZYWINSKI, M. Points of Significance: Association, correlation and causation. **Nature Methods**, v. 12, n. 10, p. 899–900, 2015. Disponível em: <<http://dx.doi.org/10.1038/nmeth.3587>>.
- ANDRADE, F., & KONZEN, L. **Levantamento Aerogeofísico Para A Identificação De Áreas Com Ocorrência Potencial De Petróleo E Gás Nas Bacias Do Amazonas, Marajó E Adjacências**. [s.l: s.n.].
- ARANA, J. M. **Introdução à Geodésia Física**. [s.l: s.n.]. Disponível em: <<http://www2.fct.unesp.br/docentes/cartto/arana/Apgfis.pdf>>.
- BASSIN, C.; LASKE, G.; MASTERS, G. The Current Limits of Resolution for Surface Wave Tomography in North America. **Eos**, v. 81, 2000.
- BLITZKOW, D.; MATOS, A. C. O. C. . G. et al. Estudios de gravedad y el modelo cuasi-geoide para la América del Sur. In: Simposio Sirgas 2015, Santo Domingo. **Anais**. Santo Domingo: 2015. Disponível em: <http://www.sirgas.org/fileadmin/docs/Boletines/Bol20/37_Blitzkow_et_al_2015_Gravedad_y_cuasi-geoide_Suramerica.pdf>.
- BREIMAN, L. Bagging Predictors. **Machine Learning**, v. 24, n. 421, p. 123–140, 1996.
- CARNEIRO, C. de C. et al. Semiautomated geologic mapping using self-organizing maps and airborne geophysics in the Brazilian Amazon. v. 77, n. 4, 2012.
- CARNEIRO, C. de C. Modelagem espacial de dados aerogeofísicos e interpretação de imagens SAR aplicados ao mapeamento geológico e metalogenético da região central do Domínio Bacajá, Estado do Pará. 2010.
- CARPENTER, G.; GROSSBERG, S. The ART of Adaptive Pattern Recognition by a Network. p. 11, 1988.
- CPRM, C. B. de P. de R. M. Relatório Final de Pesquisa: Projeto RENCA-1057. **Relatório Técnico**, 1999.

CRÓSTA, A. P. Processamento digital de imagens de sensoriamento remoto. **1992**, p. 170, 1992. Disponível em: <[http://www.ige.unicamp.br/~beto/GA235_2013/PDI/Livro Alvaro.pdf](http://www.ige.unicamp.br/~beto/GA235_2013/PDI/Livro%20Alvaro.pdf)>.

DENTITH, M.; MUDGE, S. T. **Geophysics for the Mineral Exploration Geoscientist**. [s.l.] Cambridge University Press, 2014.

DESHPANDE, R.; VANDERSLUIS, B.; MYERS, C. L. Comparison of Profile Similarity Measures for Genetic Interaction Networks. **PLoS ONE**, v. 8, n. 7, 2013.

DICKSON, B. L.; GIBLIN, a. M. An evaluation of methods for imputation of missing trace element data in groundwaters. **Geochemistry: Exploration, Environment, Analysis**, v. 7, n. 2, p. 173–178, 2007.

FESSANT, F.; MIDENET, S. Self-organising map for data imputation and correction in surveys. **Neural Computing and Applications**, v. 10, n. 4, p. 300–310, 2002.

FOLGUERA, L. et al. Self-organizing maps for imputation of missing data in incomplete data matrices. **Chemometrics and Intelligent Laboratory Systems**, v. 143, p. 146–151, 2015. Disponível em: <<http://dx.doi.org/10.1016/j.chemolab.2015.03.002>>.

FÖRSTE, C.; BRUINSMA, S. L.; ABRIKOSOV, O.; LEMOINE, J-M.; SCHALLER, T.; GÖTZE, H. J.; EBBING, J.; MARTY, J-C.; FLECHTNER, F.; BALMINO, G. B. EIGEN-6C4: The latest combined global gravity field model including GOCE data up to degree and order 2190 of GFZ Potsdam and GRGS Toulouse. In: **Anais**. 2014.

FORSTE, C. et al. EIGEN-GL05C - A new global combined high-resolution GRACE-based gravity field model of the GFZ-GRGS cooperation. **Geophysical Research Abstracts**, v. 10, n. April 2006, p. Abstract No. EGU2008-A-06944, 2008.

FOWLER, C. M. R. **The Solid Earth: An Introduction to Global Geophysics**. Cambridge UnivPress, 2004.

FRASER, S. J.; DICKSON, B. L. A new method for data integration and integrated data interpretation: self-organizing maps. v. 7, p. 907–910, 2007. Disponível em: <<http://www.dmec.ca/ex07-dvd/E07/pdfs/67.pdf>>.

FRIEDEL, M. J. Modeling hydrologic and geomorphic hazards across post-fire

landscapes using a self-organizing map approach. **Environmental Modelling and Software**, v. 26, n. 12, p. 1660–1674, 2011. Disponível em: <<http://dx.doi.org/10.1016/j.envsoft.2011.07.001>>.

FUKUSHIMA, K. Cognitron: A self-organizing multilayered neural network. **Biological Cybernetics**, v. 20, n. 3–4, p. 121–136, 1975.

GHASEMINEZHAD, M. H.; KARAMI, A. A novel self-organizing map (SOM) neural network for discrete groups of data clustering. **Applied Soft Computing Journal**, v. 11, n. 4, p. 3771–3778, 2011.

GRUS, J. **Data Science do zero: Primeiras regras com o Python, Manuais, Projetos, Pesquisas de Programação para Java**. [s.l.] Alta Books, 2016.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. [s.l.: s.n.]

HOOD, P. J.; HOLROYD, M. T.; MCGRATH, P. H. Magnetic methods applied to base metal exploration. In: HOOD, P. J. (Ed.). **Geophysics and Geochemistry in the Search for Metallic Ores**. [s.l.] Geological Survey of Canada Economic Geology Report, 1979. p. 77–104.

JAHN, F.; COOK, M.; GRAHAM, M. **Hydrocarbon exploration and production**. [s.l.] Elsevier, 1998.

JUNNINEN, H. et al. Methods for imputation of missing values in air quality data sets. **Atmospheric Environment**, v. 38, n. 18, p. 2895–2907, 2004.

KALTEH, A. M.; HJORTH, P. Imputation of missing values in a precipitation-runoff process database. **Hydrology Research**, v. 40, n. 4, p. 420–432, 2009.

KALTEH, A. M.; HJORTH, P.; BERNDTSSON, R. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. **Environmental Modelling & Software**, v. 23, n. 7, p. 835–845, 2008. Disponível em: <<http://linkinghub.elsevier.com/retrieve/pii/S1364815207001879>>.

KLOSE, C. D. Self-organizing maps for geoscientific data analysis: Geological interpretation of multidimensional geophysical data. **Computational Geosciences**, v. 10, n. 3, p. 265–277, 2006.

KOHONEN, T. **The self-organizing map** *Neurocomputing*, 2001.

KURODA, M. C. et al. Electrofacies characterization using self-organizing maps. **Revista Brasileira de Geofísica**, v. 30, n. 3, p. 287–299, 2012.

LEMOINE, F. G. et al. The Development of the Joint NASA GSFC and the National Imagery and Mapping Agency (NIMA) Geopotential Model EGM96. **NASA Goddard Space Flight Center; Greenbelt, MD United States.**, n. July, p. 461–469, 1998. Disponível em: <http://link.springer.com/10.1007/978-3-662-03482-8_62>.

LOWRIE, W. **Fundamentals of geophysics**. [s.l.] Cambridge University Press, 2007.

MATOS, A.; BLITZKOW, D. **Modelagem Digital de Terrenos (MDT) de 3" para a América do Sul (projeto de Pos-Doc)**. 2008. Escola Politécnica, Universidade de São Paulo, São Paulo, 2008.

MAUS, S. et al. EMAG2: A 2-arc min resolution Earth Magnetic Anomaly Grid compiled from satellite, airborne, and marine magnetic measurements. **Geochemistry, Geophysics, Geosystems**, v. 10, n. 8, 2009.

MERLIN, P. et al. X-SOM and L-SOM: A double classification approach for missing value imputation. **Neurocomputing**, v. 73, n. 7–9, p. 1103–1108, 2010. Disponível em: <<http://dx.doi.org/10.1016/j.neucom.2009.11.019>>.

MOUNT, N. J.; WEAVER, D. Self-organizing maps and boundary effects: Quantifying the benefits of torus wrapping for mapping SOM trajectories. **Pattern Analysis and Applications**, v. 14, n. 2, p. 139–148, 2011.

OLIVEIRA, M. J. de. Diagnóstico do setor mineral do Estado do Amapá. p. 148, 2010.

PENN, B. S. Using self-organizing maps for anomaly detection in hyperspectral imagery. **IEEE Aerospace Conference Proceedings**, v. 3, p. 1531–1535, 2002.

PENN, B. S. Using self-organizing maps to visualize high-dimensional data. **Computers and Geosciences**, v. 31, n. 5, p. 531–544, 2005.

RABELO, E. Avaliação de técnicas de visualização para mineração de dados. **Universidade Estadual de Maringá**, p. 103, 2007. Disponível em: <<http://www.din.uem.br/~mestrado/diss/2007/rabelo.pdf>>.

RALLO, R. et al. Neural virtual sensor for the inferential prediction of product

quality from process variables. **Computers and Chemical Engineering**, v. 26, n. 12, p. 1735–1754, 2002.

REGUZZONI, M.; SAMPIETRO, D. GEMMA: An Earth crustal model based on GOCE satellite data. **International Journal of Applied Earth Observation and Geoinformation**, v. 35, n. PA, p. 31–43, 2015. Disponível em: <<http://dx.doi.org/10.1016/j.jag.2014.04.002>>.

REYNOLDS, J. M. **An Introduction to Applied and Environmental Geophysics**. [s.l.] Wiley, 2011.

RIBEIRO, V. B.; MANTOVANI, M.; LOURO, V. H. A. Aerogamaespectrometria e suas aplicações no mapeamento geológico. **Terrae Didactica**, v. 10, n. 1, p. 29, 2014.

RUMELHART, D. E.; ZIPSER, D. Feature discovery by competitive learning. **Cognitive Science**, v. 9, n. 1, p. 75–112, 1985.

SAMAD, T.; HARP, S. A. Self-organization with partial data. v. 3, p. 205–212, 1992.

SAUNDERS, D. F. Tests of Australian aerial radiometric data for use in petroleum reconnaissance. **Geophysics**, v. 59, n. 3, p. 411, 1994.

SAVELIEV, A. A.; DOBRININ, D. V. Hierarchical multispectral image classification based on self organized maps. **International Geoscience and Remote Sensing Symposium (IGARSS)**, v. 5, p. 2510–2512, 1999.

SCHAFER, J. L.; GRAHAM, J. W. Missing Data : Our View of the State of the Art. v. 7, n. 2, p. 147–177, 2002.

SHAKO, R. et al. EIGEN-6C: A High-Resolution Global Gravity Combination Model Including GOCE Data. In: FLECHTNER, F.; SNEEUW, N.; SCHUH, W.-D. (Ed.). **Observation of the System Earth from Space - CHAMP, GRACE, GOCE and future missions: GEOTECHNOLOGIEN Science Report No. 20**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. p. 155–161.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados: com aplicações em R**. [s.l.] Elsevier Brasil, 2017.

ULTSCH, A. U*-Matrix: a tool to visualize clusters in high dimensional data. n. 36, p. 1–12, 2003.

VASCONCELOS, M. A. R.; LEITE, E. P.; CRÓSTA, A. P. Contributions of gamma-ray spectrometry to terrestrial impact crater studies: The example of Serra da Cangalha, northeastern Brazil. **Geophysical Research Letters**, v. 39, n. 4, p. 1–7, 2012.

VESANTO, J. et al. SOM Toolbox for Matlab 5. **Technical Report A57**, v. 2, n. 0, p. 59, 2000.

VESANTO, J.; ALHONIEMI, E. Publication 7 Clustering of the Self-Organizing Map Clustering of the Self-Organizing Map. **IEEE Transactions on Neural Networks**, v. 11, n. 3, p. 586–600, 2000.

WANG, S. Application of self-organising maps for data mining with incomplete data sets. p. 42–48, 2003.

WESSEL, P. et al. Generic Mapping Tools: Improved Version Released. **Eos, Transactions American Geophysical Union**, v. 94, n. 45, p. 409–410, 2013. Disponível em: <<http://doi.wiley.com/10.1002/2013EO450001>>.