

Universidade de São Paulo
Instituto de Astronomia, Geofísica e Ciências Atmosféricas
Departamento de Geofísica

Ellen Fernandes Marcos

**Quantificação da qualidade da interpolação de
dados de métodos potenciais por fontes equivalentes
utilizando validação cruzada em blocos**

São Paulo

2025

Ellen Fernandes Marcos

**Quantificação da qualidade da interpolação de
dados de métodos potenciais por fontes equivalentes
utilizando validação cruzada em blocos**

Versão corrigida da monografia apresentada ao
Curso de Graduação em Geofísica da Universi-
dade de São Paulo para obtenção do título de
Bacharel em Geofísica.

Orientador: Prof. Dr. Leonardo Uieda

São Paulo

2025

Agradecimentos

Agradeço ao meu orientador Prof. Dr. Leonardo Uieda por me acompanhar durante essa reta final da graduação. Ele me inspirou e guiou, tornando o caminho desta pesquisa possível e confortável. Agradeço também aos encontros com os membros do Compgeolab.

Agradeço especialmente ao Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo (IAG-USP), pelo espaço e pelos encontros com os demais professores que fizeram parte desta jornada.

Agradeço principalmente à minha família, minha mãe Valdilene e meu pai Edson. Vocês me deram estrutura e direção para que eu pudesse trilhar o caminho com meus próprios pés.

*“Não basta abrir a janela
Para ver os campos e o rio.
Não é bastante não ser cego
Para ver as árvores e as flores.
É preciso também não ter filosofia nenhuma.
Com filosofia não há árvores: há ideias apenas.
Há só cada um de nós, como uma cave.
Há só uma janela fechada, e todo o mundo lá fora;
E um sonho do que se poderia ver se a janela se abrisse,
Que nunca é o que se vê quando se abre a janela.”*

Alberto Caeiro

Resumo

Como parte do seu processamento, dados de gravimetria e magnetometria são rotineiramente interpolados em malhas regulares a altitudes constantes. O método das fontes equivalentes é ideal para realizar esse processamento por levar em consideração a física do problema, como a altitude variável dos pontos de observação e o fato dos dados representarem funções harmônicas. A qualidade do resultado obtido varia de acordo com dois parâmetros: a profundidade das fontes e o parâmetro de regularização. Ambos controlam o grau de suavidade da solução e a capacidade do método de prever valores plausíveis onde existem lacunas nos dados observados. Utilizamos o método da validação cruzada em blocos (*Block k-Fold*), de [Roberts et al. \(2017\)](#), para estimar a capacidade das fontes equivalentes de prever dados onde existem lacunas nos levantamentos. Na validação cruzada, os dados são divididos aleatoriamente em dois subconjuntos, um para ajuste (treinamento) do modelo de fontes equivalentes e outro para testar as previsões do modelo. Para dados de métodos potenciais, a divisão aleatória dos dados (método *k-Fold*) não leva em consideração a autocorrelação dos dados no espaço, o que provoca a superestimação da qualidade do ajuste aos dados de teste. A separação dos dados por blocos, ao invés de aleatoriamente, foi introduzida para garantir distância entre os dados que são utilizados para previsão e teste do modelo e sanar o problema. Utilizamos dados sintéticos para avaliar o uso das metodologias "*Block k-Fold*" e "*k-Fold*" para estimar a qualidade da interpolação de dados gravimétricos e magnetométricos por fontes equivalentes. Nossa avaliação incluiu um teste sistemático do efeito da variação do tamanho dos blocos na estimativa da qualidade, tanto para levantamentos terrestres quanto aerolevantamentos. As fontes equivalentes utilizadas foram as de [Soler e Uieda \(2021\)](#) e a validação cruzada por blocos foi implementada pelo pacote Verde ([Uieda, 2018](#)).

Abstract

As part of their processing, gravity and magnetic data are routinely interpolated onto regular grids at constant altitudes. The equivalent source method is ideal for performing this processing as it accounts for the physics of the problem, such as the variable altitude of observation points and the fact that the data represent harmonic functions. The quality of the results depends on two parameters: the depth of the sources and the regularization parameter. Both control the smoothness of the solution and the method's ability to predict plausible values where there are gaps in the observed data. We used blocked cross-validation (Block k-Fold method) by [Roberts et al. \(2017\)](#) to estimate the ability of equivalent sources to predict data in regions with gaps in the surveys. In cross-validation, the data are randomly divided into two subsets, one for fitting (training) the equivalent source model and the other for testing the model's predictions. For potential field data, random division (k-Fold method) does not account for the spatial autocorrelation of the data, which leads to overestimation of the model's predictive quality on test data. Dividing the data into blocks instead of randomly was introduced to ensure a spatial separation between the data used for prediction and testing, addressing this issue. We used synthetic data to evaluate the performance of the "block k-fold" and "k-fold" methodologies for estimating the quality of gravity and magnetic data interpolation using equivalent sources. Our assessment included a systematic test of the effect of block size variation on the quality estimation, both for ground-based and airborne surveys. The equivalent sources used were those from [Soler e Uieda \(2021\)](#), and the block cross-validation implementation was performed using the Verde package ([Uieda, 2018](#)).

Listas de Figuras

2.1	Esquema do método das fontes equivalentes	4
2.2	Separação dos dados pelo método <i>k-Fold</i>	8
2.3	Separação dos dados pelo método <i>Block k-Fold</i>	9
3.1	Pontos de observação dos dados gravimétricos	13
3.2	Campo gravitacional amostrado nos pontos de observação	14
3.3	Interpolação dos dados gravimétricos por fontes equivalentes	14
3.4	Malha verdadeira do modelo gravimétrico	15
3.5	Malha de resíduos do modelo gravimétrico	15
3.6	Variação do parâmetro R^2 com o tamanho do bloco	16
3.7	Levantamento terrestre e interpolação por fontes equivalentes	17
3.8	Malha verdadeira e resíduo do modelo terrestre	17
3.9	Variação do parâmetro R^2 com o tamanho do bloco para levantamento terrestre .	18
3.10	Aerolevantamento 1 e interpolação por fontes equivalentes	19
3.11	Malha verdadeira e resíduo do modelo aéreo 1	19
3.12	Variação do parâmetro R^2 com o tamanho do bloco para aerolevantamento 1 . . .	20
3.13	Aerolevantamento 2 e interpolação por fontes equivalentes	21
3.14	Malha verdadeira e resíduo do modelo aéreo 2	21
3.15	Variação do parâmetro R^2 com o tamanho do bloco para o aerolevantamento 2 . .	22
3.16	Variação do parâmetro R^2 e o espaçamento entre as linhas de voo	23

Sumário

1. <i>Introdução</i>	1
2. <i>Metodologia</i>	3
2.1 Fontes Equivalentes	3
2.2 Validação Cruzada	6
2.2.1 Validação Cruzada Aleatória (k-Fold)	7
2.2.2 Validação Cruzada por Blocos (Block k-Fold)	9
2.3 Procedimentos	10
2.3.1 Malha dos levantamentos sintéticos	10
2.3.2 Interpolação e a validação	11
3. <i>Resultados</i>	13
3.1 Dados sintéticos do campo gravitacional	13
3.2 Dados sintéticos do campo magnético	16
3.2.1 Levantamento terrestre	16
3.2.2 Levantamento aéreo	18
4. <i>Discussão</i>	24
5. <i>Conclusões</i>	26
<i>Referências</i>	27

Capítulo 1

Introdução

O método das fontes equivalentes tem sido utilizado na geofísica para processar os dados de campos potenciais. Proposto por [Dampney \(1969\)](#), ele se baseia em um conjunto discreto de dados observados de campo potencial que pode ser aproximado por outro conjunto discreto de fontes, como massas pontuais. Considerando que o problema inverso é linear, estimamos coeficientes que refletem as propriedades físicas das fontes equivalentes, de forma a ajustar as observações. Em seguida, a distribuição de massa na camada equivalente é utilizada para realizar a transformação desejada do campo potencial.

O custo operacional do método das fontes equivalentes depende diretamente do número de observações, sendo ineficiente no processamento de grandes conjuntos de dados. Algumas técnicas foram propostas na literatura para otimizar o método, como aplicar a decomposição em valores singulares à matriz que contém os dados, reduzindo a dimensão do sistema linear ([Mendonça, 2020](#)). Outra abordagem é a de [Soler e Uieda \(2021\)](#) que combina duas estratégias: uma que reduz o número de parâmetros a serem ajustados, operando em janelas móveis, semelhante ao método de convolução discreta, também utilizado por [Leão e Silva \(1989\)](#); e outra que opera iterativamente em janelas sobrepostas, diminuindo o tamanho do sistema linear. O trabalho de [Oliveira Junior et al. \(2023\)](#) realiza uma comparação sistemática entre esses e outros métodos na literatura utilizados para aumentar a eficiência do método das fontes equivalentes.

Outro desafio particular do método é encontrar os parâmetros apropriados, como a profundidade das fontes e o regularizador (*damping*). Ambos controlam a suavidade da solução e a qualidade de previsão de valores plausíveis onde há lacunas nos levantamentos. Como dados de campo potencial são inherentemente não únicos, existem vários modelos e conjuntos de parâmetros que podem descrever adequadamente os dados. Podemos selecionar automaticamente estes parâmetros, a partir de um conjunto de valores a serem testados, utilizando a validação cruzada.

A validação cruzada é um método estatístico utilizado em aprendizagem de máquinas para selecionar modelos. A abordagem geral do método é dividir todo o conjunto de dados observados em dois subconjuntos, um para ajuste (treinamento) do modelo e outro para avaliação (teste) do modelo. Para dados de campo potencial, a divisão aleatória dos dados não leva em consideração a autocorrelação dos dados no espaço, o que provoca a superestimação da qualidade do ajuste aos dados de teste. Utilizamos como solução o método da validação cruzada em blocos ([Roberts et al., 2017](#)) para estimar a qualidade da interpolação dos dados por fontes equivalentes.

O método da validação cruzada por blocos foi proposto para levar em consideração as estruturas temporais, espaciais, hierárquicas ou filogenéticas dos dados. Idealmente, a validação, a seleção e os erros preditivos do modelo devem ser calculados usando dados independentes. Segundo [Roberts et al. \(2017\)](#), ao violar essa suposição, geramos estimativas de erro que são otimistas e que favorecem a seleção de modelos complexos. No caso da validação cruzada não corrigida (aleatória), por não levar em consideração as estruturas de dependências existentes nos dados, o método favorece modelos ajustados excessivamente (*overfitting*). Diminuímos as estruturas de dependência dos dados ao dividir a região de estudo por blocos de tamanhos iguais. Dessa forma, impomos distância entre os subconjuntos de treinamento e teste do modelo, diminuindo a autocorrelação espacial dos dados.

A eficiência do método da validação cruzada por blocos ainda não foi amplamente estudada para dados de campo potencial. Procuramos ocupar esse espaço vazio e tentar responder algumas questões, como o tamanho ideal do bloco e se há alguma relação entre o tamanho do bloco e o espaçamento entre as linhas de voo de um aerolevantamento.

Nas seções seguintes, abordaremos as bases teóricas do método das fontes equivalentes para realizar a interpolação dos dados de campo potencial. Em seguida, exploraremos a validação cruzada na quantificação da qualidade do ajuste do modelo de fontes equivalentes aos dados observados. A partir de dados sintéticos iremos avaliar o uso das metodologias de validação cruzada aleatória (*k-Fold*) e por blocos (*Block k-Fold*). Nossa avaliação inclui um teste sistemático do efeito da variação do tamanho dos blocos na estimativa da qualidade da interpolação, tanto para levantamentos terrestres quanto aerolevantamentos. Utilizamos os pacotes em Python do Verde ([Uieda, 2018](#)) e do Harmonica ([Fatiando a Terra Project et al., 2024](#)) para simular os dados sintéticos do campo potencial. Através dos testes, mostramos que o método *k-Fold* superestima o parâmetro de qualidade do ajuste. Este efeito é maior em modelos ruidosos ou mal amostrados, onde o método *Block k-Fold* fornece estimativas de erro mais confiáveis.

Capítulo 2

Metodologia

Considerando que $f(x, y, z)$ seja uma anomalia de campo potencial, e portanto um campo harmônico, pela Equação de Laplace temos que

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = 0. \quad (2.1)$$

Dessa forma, sem condições de fronteira, dados de campo potencial são inherentemente não únicos. Ao se aproveitar desta propriedade, o método das fontes equivalentes aproxima os dados observados do campo por uma combinação linear de funções harmônicas. Utilizamos essa técnica para interpolar os dados do campo potencial em malhas regulares a altitudes constantes por levar em consideração a física do problema, como a altitude variável dos pontos de observação e o fato dos campos obedecerem a Equação de Laplace (2.1). Realizamos uma inversão de mínimos quadrados com regularização de norma mínima para estimar coeficientes que refletem as propriedades físicas das fontes equivalentes. De posse dos coeficientes obtemos uma malha regular interpolada.

2.1 *Fontes Equivalentes*

O método das fontes equivalentes foi proposto por [Dampney \(1969\)](#) e aproxima a distribuição contínua das anomalias do campo gravitacional pela soma dos efeitos de m fontes pontuais discretas. Dessa forma, o método ajusta o modelo a partir das m fontes localizadas abaixo dos pontos de dados em uma profundidade comum. Portanto, considerando f_i a anomalia de campo gravitacional amostrada no i -ésimo ponto de N dados,

$$f_i = \sum_{j=1}^M g_{ij} p_j, \quad \forall i = 1, 2, \dots, N, \quad (2.2)$$

temos que a equação 2.2 aproxima a anomalia f_i pela soma de m fontes pontuais com densidade p_j . A matriz de sensibilidade ou matriz Jacobiana é definida por

$$g_{ij} = \frac{G}{[(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{3/2}}. \quad (2.3)$$

As coordenadas (x_i, y_i, z_i) correspondem aos pontos de observação, e as coordenadas (x_j, y_j, z_j) correspondem às fontes localizadas abaixo dos pontos de dados. G é a constante gravitacional. A Figura 2.1 representa as observações e as fontes sendo utilizadas para prever os dados em um grid regular.

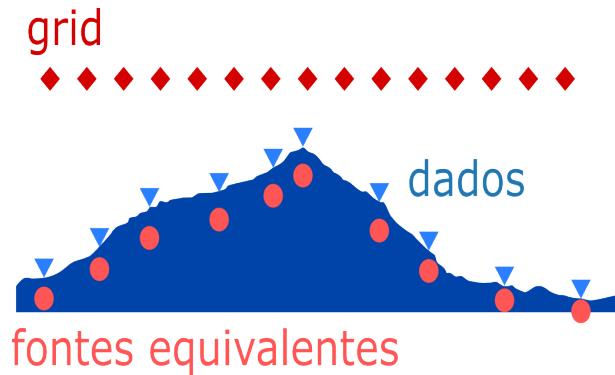


Figura 2.1: Representação das fontes equivalentes (pontos em rosa), localizadas abaixo dos pontos de observação (pontos em azul). A distribuição das fontes é utilizada para prever os dados do campo no grid (pontos em vermelho).

Uma reformulação do problema foi feita por [Cordell \(1992\)](#), conhecida como fontes equivalentes generalizadas. O método considera a distribuição de anomalias de campos potenciais que podem ser modeladas por um dipolo ou por fontes pontuais. Neste trabalho, o método das fontes equivalentes foi implementado através do algoritmo de [Soler e Uieda \(2021\)](#), que segue o proposto por [Cordell \(1992\)](#), assumindo que qualquer função harmônica $d(\bar{p})$ pode ser aproximada pela soma de m efeitos de fontes pontuais discretas,

$$d(\bar{p}) = \sum_{j=1}^M \frac{c_j}{\|\bar{p} - \bar{q}_j\|}. \quad (2.4)$$

Os vetores \bar{p} e \bar{q}_j pertencem aos pontos de observação e às fontes, respectivamente. O coeficiente escalar c_j está relacionado à fonte pontual localizada em \bar{q}_j e $\|\cdot\|$ representa a norma L_2 .

A função harmônica $d(\bar{p})$, avaliada em N pontos discretos, define um conjunto de N equações da forma

$$d(\bar{p}_i) = d_i = \sum_{j=1}^M \frac{c_j}{\|\bar{p}_i - \bar{q}_j\|}, \quad \forall i = 1, 2, \dots, N, \quad (2.5)$$

sendo d_i o valor calculado no ponto \bar{p}_i . A equação 2.5 pode ser reescrita na forma matricial como

$$\bar{d} = \bar{\bar{A}}\bar{c}, \quad (2.6)$$

tal que \bar{d} é o vetor coluna contendo os N valores previstos nos pontos de observação, \bar{c} é o vetor coluna contendo os m coeficientes c_j e $\bar{\bar{A}}$ é a matriz Jacobiana $N \times M$, ou matriz de sensibilidade, com elementos

$$a_{ij} = \frac{1}{||\bar{p}_i - \bar{q}_j||}. \quad (2.7)$$

Para um determinado conjunto de N dados observados, \bar{d}^o , podemos encontrar uma solução de mínimos quadrados para a equação 2.6 e obter os valores de \bar{c} que melhor se ajustam às observações. A partir desses coeficientes podemos prever o valor da função harmônica da equação 2.4 em qualquer ponto do espaço que seja externo às fontes.

A interpolação dos dados de campos potenciais é realizada a partir de uma inversão de mínimos quadrados de norma mínima. Dessa forma, temos um problema sobredeterminado, onde encontramos os melhores valores de \bar{c} que ajustam os dados. De posse dos valores de \bar{c} , prevemos valores que estão nos pontos de uma grade regular, obtendo a malha interpolada.

Podemos encontrar os melhores valores dos coeficientes das fontes, \bar{c} , avaliando aqueles que melhor se ajustam aos valores observados, \bar{d}^o . Esse ajuste é alcançado minimizando a função objetivo

$$\phi(\bar{c}) = [\bar{d}^o - \bar{\bar{A}}\bar{c}]^T \bar{\bar{W}} [\bar{d}^o - \bar{\bar{A}}\bar{c}] + \lambda_d \bar{c}^T \bar{c}, \quad (2.8)$$

em que $\bar{\bar{W}}$ é uma matriz diagonal $N \times N$ de pesos dos dados. O parâmetro de amortecimento λ_d é positivo, possui a mesma quantidade de elementos da matriz Jacobiana $\bar{\bar{A}}$, e define o segundo termo da equação 2.8. Este termo trata-se da regularização de Tikhonov (1977) de ordem zero, conhecida como regularização de mínimos quadrados amortecidos, usada para estabilizar a solução.

O parâmetro de amortecimento λ_d controla o grau de regularização que será aplicado ao modelo. Dessa forma, adicionamos mais um vínculo ao problema de inversão, impondo que as derivadas do dado sejam mínimas. Para valores altos de λ_d produzimos uma solução tão suave que não reproduz as componentes dos dados de alta frequência, falhando em reproduzir as fontes rasas. Para valores baixos de λ_d produzimos uma solução com ajuste excessivo (*overfitting*), tal que a solução pode acabar ajustando as componentes de alta frequência do dado, falhando em

produzir resultados de interpolação realísticos. A faixa de valores razoáveis para λ_d é regulada pela matriz Jacobiana $\bar{\bar{A}}$ e pelos coeficientes, variando entre os conjuntos de dados. Dessa maneira, o método das fontes equivalentes pode ser desafiador quanto à escolha dos melhores parâmetros.

Para resolver o problema de encontrar os parâmetros que melhor se ajustam aos dados, primeiro escalamos a matriz Jacobiana $\bar{\bar{A}}$ para que seus elementos sejam adimensionais e cada coluna tenha variância unitária. Definimos uma matriz diagonal S de dimensões $M \times M$, contendo os valores do desvio padrão das colunas de $\bar{\bar{A}}$. Dessa maneira, obtemos uma matriz $\bar{\bar{B}} = \bar{\bar{A}}S^{-1}$ que é a matriz Jacobiana $\bar{\bar{A}}$ escalada e adimensional. Também definimos $\bar{m} = \bar{\bar{S}}\bar{c}$ como o vetor contendo os coeficientes escalados com a mesma quantidade de elementos que os dados. Assim, reescrevemos o modelo direto da equação 2.6 como

$$\bar{d} = \bar{\bar{A}}\bar{\bar{S}}^{-1}\bar{\bar{S}}\bar{c} = [\bar{\bar{A}}\bar{\bar{S}}^{-1}][\bar{\bar{S}}\bar{c}] = \bar{\bar{B}}\bar{m}. \quad (2.9)$$

E podemos reescrever a função objetivo descrita na equação 2.8 como

$$\phi(\bar{m}) = [\bar{d}^o - \bar{\bar{B}}\bar{m}]^T \bar{\bar{W}} [\bar{d}^o - \bar{\bar{B}}\bar{m}] + \lambda \bar{m}^T \bar{m}, \quad (2.10)$$

em que λ é um parâmetro de amortecimento adimensional. Essa transformação é realizada para que a regularização seja aplicada nos coeficientes escalados \bar{m} em vez de \bar{c} . Dessa forma, reduzimos a faixa de valores de λ que produziriam previsões mais precisas, tornando a escolha de λ independente do conjunto de dados e de suas unidades.

A escolha do parâmetro de regularização (*damping*) e da profundidade das fontes que melhor ajustam os dados observados pode ser feita por meio da validação cruzada. O método estatístico é utilizado na seleção de modelos, permitindo avaliar qual valor de λ , dentro de um conjunto de valores, nos retorna o melhor ajuste aos dados.

2.2 Validação Cruzada

Como visto, o desafio na utilização das fontes equivalentes é a escolha dos parâmetros de amortecimento e da profundidade das fontes. Tais parâmetros controlam o grau de suavidade da solução e a capacidade do modelo de prever dados onde há lacunas nos levantamentos. Podemos definir automaticamente esses parâmetros para maximizar a precisão a partir da validação cruzada.

O método é utilizado para avaliar o desempenho de um modelo, atribuindo pontos de observações a um conjunto de avaliação e calculando métricas de desempenho ao realizar previsões.

O método de validação cruzada divide os dados em dois subconjuntos, um para treinamento e outro para teste do modelo. No método das fontes equivalentes, treinamos o modelo realizando uma inversão para estimar os valores de \bar{c} (equação 2.6) que melhor se ajustam aos dados observados, como visto na seção 2.1. Estes valores são utilizados para avaliação nos dados de teste, geralmente a partir de parâmetros como o Erro Médio Quadrático (MSE) ou o R^2 . Existem inúmeras abordagens de validação cruzada e cada uma difere da outra na maneira com que a divisão dos dados ocorre. Este estudo aborda dois métodos de validação cruzada em que a divisão dos dados ocorre de maneira iterativa. Os dados são divididos k vezes e cada instância da divisão é conhecida como *fold*. Os k *folds*, ou as k dobradas da validação, são subconjuntos do dado que está sendo dividido durante o processo. As divisões são feitas automaticamente, sendo a maior parte do dado utilizada para o treinamento do modelo. Desta forma, a fração dos dados que é separada para treinamento é $\frac{(k-1)}{k}$, e para teste é $\frac{1}{k}$.

Ao avaliar o ajuste de um modelo preditivo levamos em consideração as estruturas internas de dependência existentes nos dados. Tais dependências são comuns em dados geofísicos e devem ser levadas em consideração ao realizar previsões em espaços desconhecidos, segundo [Mahoney et al. \(2023\)](#). Dados de campos potenciais possuem autocorrelação espacial e as métricas de desempenho do modelo podem acabar sendo superajustadas (*overfitting*). Assim, as estimativas de erro acabam sendo excessivamente otimistas e favorecendo a seleção de modelos muito complexos ([Roberts et al., 2017](#)). Dados autocorrelacionados têm a tendência de terem valores semelhantes em pontos de observações retirados de regiões próximas. Levando em consideração essas estruturas podemos aumentar a independência na validação cruzada, dividindo a região de estudo por blocos [Roberts et al. \(2017\)](#).

2.2.1 Validação Cruzada Aleatória (*k-Fold*)

A validação cruzada aleatória (método *k-Fold*), introduzida por [Geisser \(1975\)](#), é a forma mais comum de validação. A maneira com que os dados são divididos está representada na figura 2.2. Atribuímos aleatoriamente cada observação a um dos subconjuntos de treinamento e teste. Sendo k o número de divisões, os modelos são ajustados a cada combinação única de $k - 1$ divisões e avaliados no restante, com as métricas de desempenho estimadas pela média das k iterações ([Stone, 1974](#)). Segundo [Arlot e Celisse \(2010\)](#), a escolha de um valor para k

entre 5 e 10 é suficiente pois a relação entre desempenho e custo computacional não melhora significativamente para valores maiores que 10.

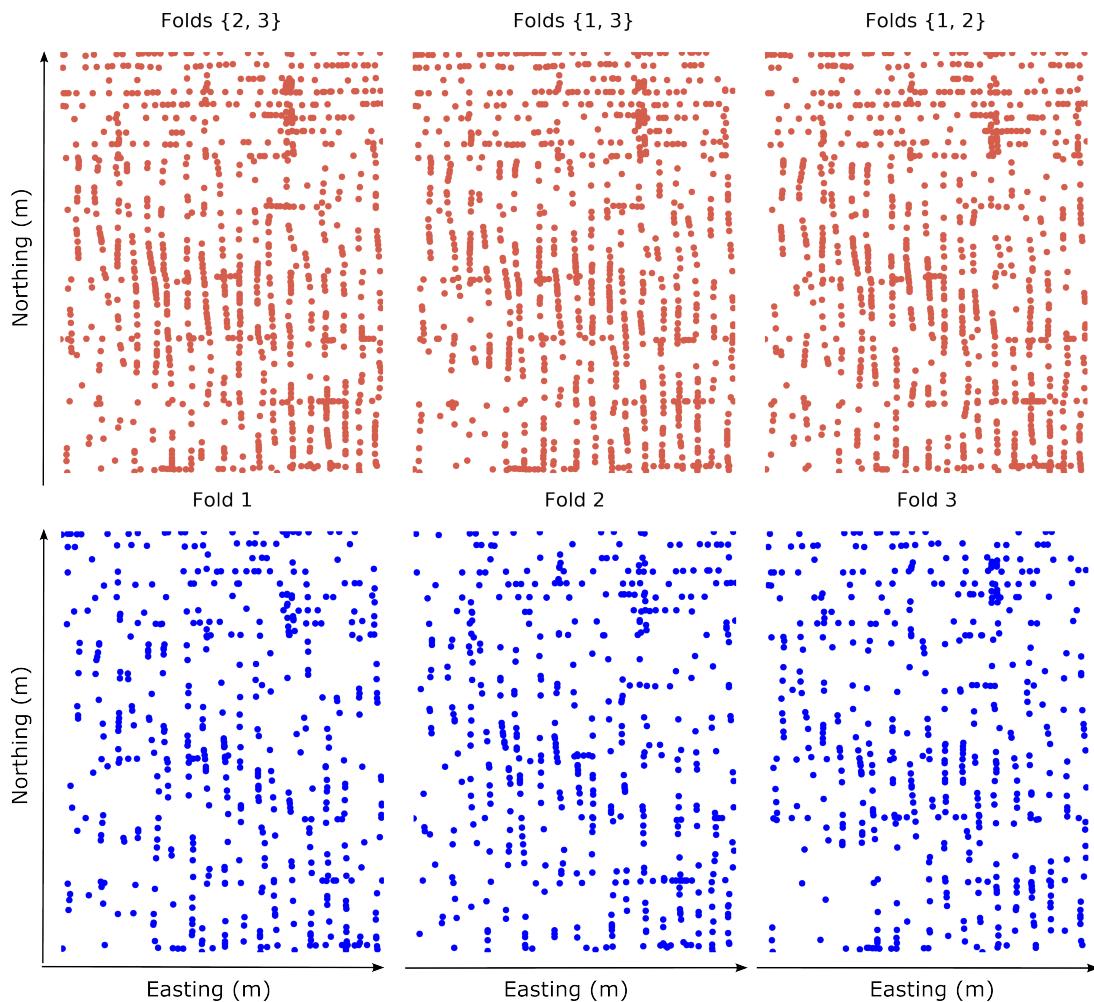


Figura 2.2: Esquema de separação dos dados para treinamento (pontos em vermelho) e teste (pontos em azul) do modelo na validação cruzada aleatória (k -Fold). Os pontos de observação são atribuídos aleatoriamente, sendo a menor fração dos dados retida para avaliação. A divisão dos dados é feita ao longo de 3 *folds*, ou seja, 3 dobradas. O número de pontos retidos nos subconjuntos é definido pela quantidade de *folds* e pelo tamanho do dado.

Ao dividir os dados de maneira aleatória, sem levar em consideração a autocorrelação espacial dos dados de campo potencial, treinamos o modelo em pontos de observação que são vizinhos dos pontos de observação separados para teste. Dessa maneira, proporcionamos resultados de validação que superestimam a capacidade do modelo de prever para novas observações ou para regiões não bem representadas nos dados de treinamento (Bahn e McGill, 2012; Roberts et al., 2017).

2.2.2 Validação Cruzada por Blocos (*Block k-Fold*)

A validação cruzada por blocos (*Block k-Fold*) foi desenvolvida por [Roberts et al. \(2017\)](#) para dados com dependências espaciais, temporais, hierárquicas ou filogenéticas. Dados de campos potenciais estão sujeitos a dependências, estando autocorrelacionados no espaço, como visto anteriormente. Para sanar o problema e tornar as previsões de erros mais realistas, o método divide a região dos dados em um conjunto de blocos uniformes e atribui aleatoriamente cada bloco para treinamento ou teste do modelo (figura 2.3). Dessa forma, impomos uma maior distância entre os pontos dos subconjuntos da validação e reduzimos a dependência entre as amostras.

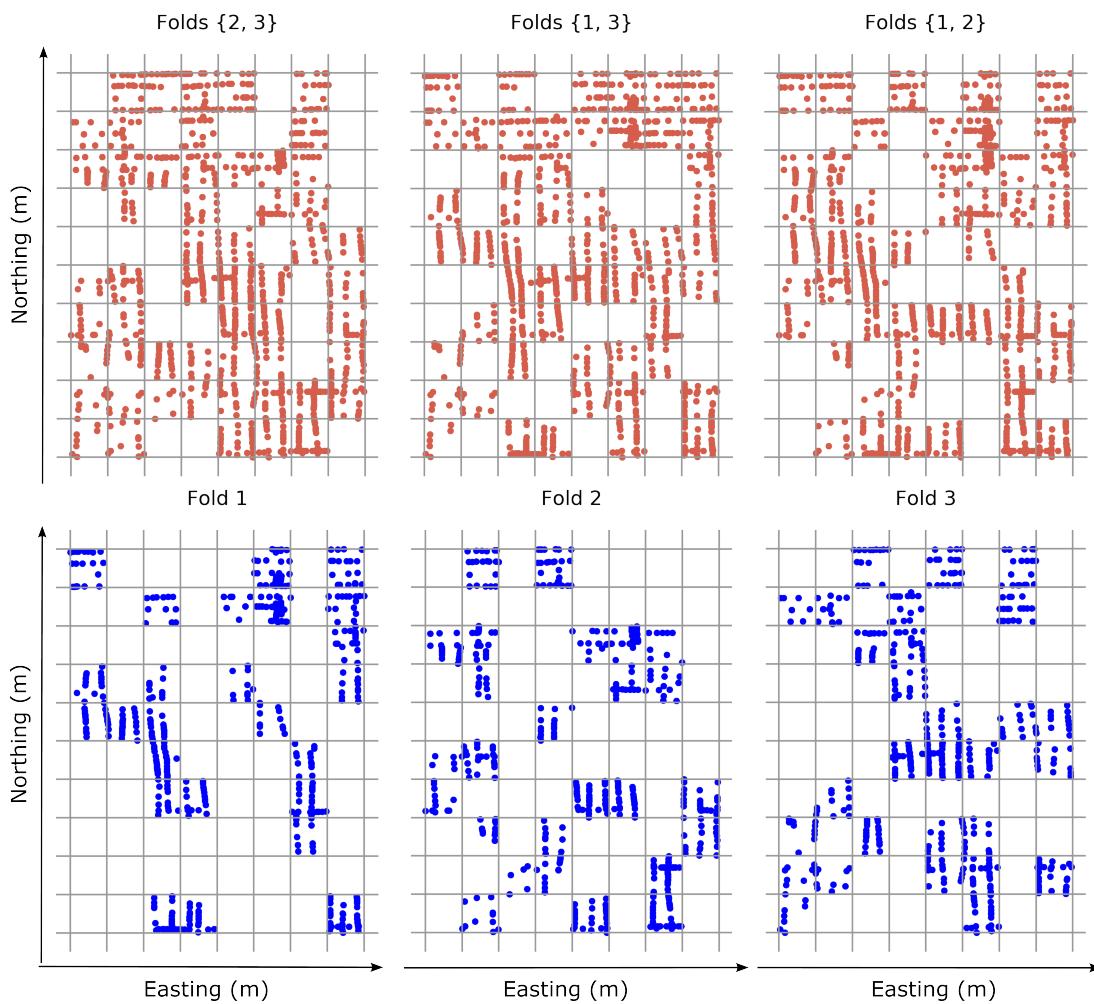


Figura 2.3: Esquema de separação dos dados para treinamento (pontos em vermelho) e teste (pontos em azul) do modelo na validação cruzada por blocos (*Block k-Fold*). A região do levantamento é dividida por blocos, que são atribuídos aleatoriamente aos subconjuntos. A menor fração dos blocos é retida para avaliação. A divisão dos blocos é feita ao longo de 3 *folds*, ou seja, 3 dobradas. O número de blocos retidos para avaliação é definido pela quantidade de *folds* e pelo tamanho do bloco.

2.3 Procedimentos

Como visto, no processamento dos dados de gravimetria e magnetometria, utilizamos o método das fontes equivalentes para interpolar os dados em malhas regulares a altitudes constantes. Por meio da validação cruzada, medimos a capacidade do método de prever dados onde existem lacunas nos levantamentos. Testamos dados sintéticos de fontes magnéticas e gravimétricas, simulando levantamentos terrestres e aerolevantamentos. Os dados dos levantamentos são interpolados utilizando o algoritmo de fontes equivalentes com *Gradient-Boosting* de [Soler e Uieda \(2021\)](#), disponibilizado para uso livre no pacote Harmonica do [Fatiando a Terra Project et al. \(2024\)](#). Medimos a qualidade do ajuste dos dados interpolados a partir da validação cruzada aleatória (*k-Fold* [2.2.1](#)) e da validação cruzada por blocos (*Block k-Fold* [2.2.2](#)), ambas implementadas através do pacote Verde ([Uieda, 2018](#)). Um teste sistemático do efeito da variação do tamanho dos blocos na estimativa da qualidade é feito para o método *Block k-Fold*.

2.3.1 Malha dos levantamentos sintéticos

Na criação da malha dos levantamentos sintéticos utilizamos o pacote Verde ([Uieda, 2018](#)). Definimos uma região com 100 km na direção E-W e 90 km em N-S. Para simular um levantamento terrestre, geramos pontos de maneira semi-aleatória dentro da região definida. Para o campo gravitacional geramos 1000 pontos, com altitude de observação de 1 km. Para o campo magnético 2000 pontos e altitude de 2 km.

Para gerar o aerolevantamento dos dados gravimétricos construímos uma grade com coordenadas espaçadas de maneira regular. Definimos linhas de voo na direção N-S, com espaçamento de 5 km entre elas. Ao longo das linhas de voo o espaçamento entre os pontos é de 500 m. A altitude de voo é constante e equivale a 500 m.

Para simular os dados do levantamento aéreo do campo magnético, concatenamos a grade das linhas de voo com direção N-S com uma grade de linhas na direção E-W com maior espaçamento. O espaçamento entre as linhas de voo E-W é de 20 km, e ao longo das linhas de voo é de 500 m. Variamos o espaçamento entre as linhas de voo na direção N-S, testando espaçamentos de 500 m, 2 km, 4 km e 5 km. Tentamos aproximar a variação da altitude de voo de um cenário real, variando como uma soma de senos e cossenos que oscila ao redor de 500 m. A amplitude máxima da oscilação é de 11 m e a frequência angular é de 0.1 rad/s. Adicionamos ruído às coordenadas da grade, com média 0 e desvio padrão de 150 m. Dessa maneira, as linhas de voo

não ficam perfeitamente alinhadas como no caso do aerolevantamento dos dados gravimétricos.

Amostramos o campo gerado pelas fontes gravimétricas e magnéticas nos pontos das grades dos levantamentos terrestre e aéreo. Utilizamos o pacote Harmonica ([Fatiando a Terra Project et al., 2024](#)) que calcula a aceleração gravitacional e o campo magnético gerados por um conjunto de prismas retangulares em coordenadas cartesianas. Para o modelo gravimétrico, geramos 2 prismas com contraste de densidade entre eles e o meio de 500 kg/m^3 e -300 kg/m^3 . Para o modelo magnético, geramos um conjunto de múltiplas fontes, simulando dois dipolos, um pipe, um dique, uma linha e uma soleira. A inclinação e a declinação magnética das fontes foram definidas como -20° e -15° . Adicionamos um campo regional e um nível de base de 500 mGal aos modelos, além de um ruído de média 0 e desvio padrão de 200 mGal. Os parâmetros de intensidade magnética e profundidade das fontes variam de acordo com os modelos gerados.

Considerando que os campos amostrados nas grades dos levantamentos sejam nossos dados observados, geramos uma malha regular como modelo "verdadeiro" para comparação. Amostramos os campos gerados pelas fontes gravimétricas e magnéticas em uma malha com espaçamento de 2 km e 1 km, respectivamente.

2.3.2 Interpolação e a validação

A partir dos dados sintéticos geramos modelos de levantamentos terrestres e aéreos. Interpolamos os dados utilizando o método das fontes equivalentes (seção [2.1](#)). Como visto, o parâmetro de regularização (*damping*) e a profundidade das fontes controlam o grau de suavidade da solução e sua capacidade de previsão de dados. Definimos um *damping* de 1 em todos os modelos. O parâmetro de profundidade das fontes nos modelos gravitacional e magnético é de 2 km e 1 km, respectivamente. Tendo estimado o coeficiente das fontes, ajustamos o modelo a uma malha regular com espaçamento de 2 km para as fontes gravimétricas e de 1 km para as fontes magnéticas.

O algoritmo de *Gradient-Boosting* aplicado ao método das fontes equivalentes, por [Soler e Uieda \(2021\)](#), trata-se de uma otimização para ajustar modelos paramétricos. Isso permite que solucionemos vários problemas de mínimos quadrados menores em vez de um grande, estimando os parâmetros das fontes em janelas sobrepostas. Dessa maneira, diminuímos o tamanho da memória armazenada pela matriz Jacobiana e aceleramos o processo de inversão.

Utilizamos a validação cruzada (seção [2.2](#)) para medir a qualidade do ajuste por fontes equivalentes calculando coeficientes de determinação R^2 . Os dados são divididos ao longo de

5 folds e cada instância da divisão calcula um parâmetro de determinação R^2 . A média destes valores é comparada com o R^2 verdadeiro, obtido através da equação 2.11,

$$R^2 = 1 - \frac{\sum_i^n r_i^2}{\sum_i^n (d_i - d_m)^2}, \text{ sendo } d_m = \frac{\sum_i^n d_i}{n}. \quad (2.11)$$

O termo d_i se refere ao dado interpolado por fontes equivalentes, com n o seu tamanho e d_m seu valor médio. O resíduo r_i é a diferença entre o modelo de fontes equivalentes e a malha verdadeira.

Variamos o tamanho do bloco utilizado na validação cruzada por blocos (seção 2.2.2), testando blocos de 1 km a 10 km. Comparamos as médias dos valores de R^2 , retornados por cada bloco, com o R^2 verdadeiro. Dessa maneira, podemos determinar qual tamanho de bloco melhor nos aproxima da verdadeira quantificação da qualidade de interpolação. A divisão dos dados na validação cruzada é feita ao longo de *5 folds*, isto é, 5 dobras. Dessa forma, a quantidade de blocos retida para teste é 1/5 da quantidade total de blocos, que varia conforme o tamanho.

Capítulo 3

Resultados

3.1 Dados sintéticos do campo gravitacional

Inicialmente geramos um modelo sintético para dados gravimétricos, simulando um levantamento terrestre e um aéreo. Os procedimentos para criar a grade dos modelos estão descritos na seção 2.3.1. Geramos os pontos de observação dos levantamentos (Figura 3.1) e calculamos a componente vertical da aceleração gravitacional produzida por dois prismas nos pontos de observação (Figura 3.2). Os prismas têm contraste de densidade de 500 kg/m^3 , localizado entre 0 m e 1 km de profundidade, e -300 kg/m^3 , localizado entre 500 m e 2 km de profundidade.

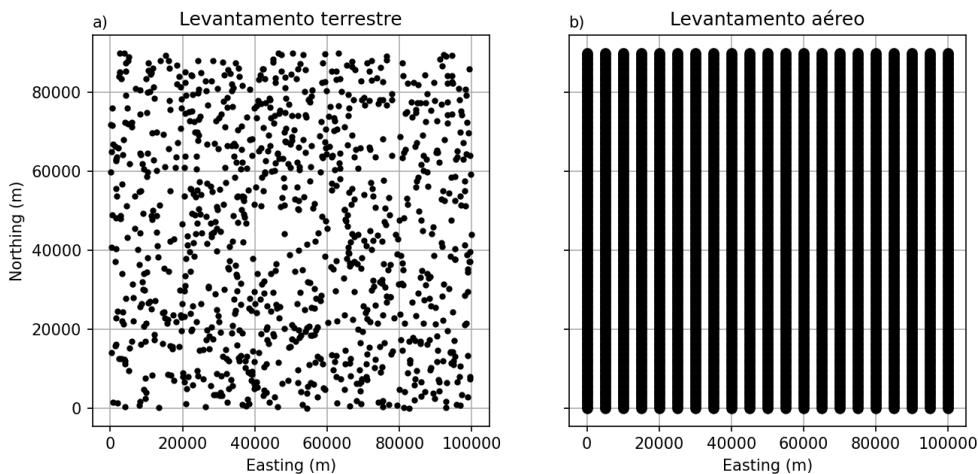


Figura 3.1: Simulações de levantamentos dos dados gravimétricos terrestre (a) e aéreo (b) com espaçamento de 5 km na direção E-W, entre as linhas de voo, e 500 m na direção N-S, ao longo das linhas de voo.

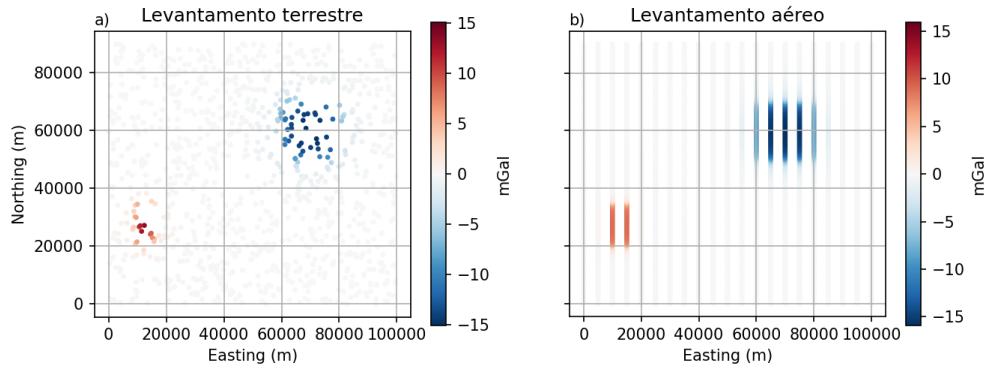


Figura 3.2: Resposta do campo gravimétrico produzido por dois prismas para levantamento terrestre (a) e aéreo (b). Um prisma com contraste de densidade negativo é observado pela cor azul. E um prisma com contraste de densidade positivo pela cor vermelha. Em pontos de observação que não passam sobre os prismas o campo é nulo.

Ajustamos o modelo de fontes equivalentes aos dados com *damping* de 1 e profundidade de 2 km e em seguida utilizamos o modelo para produzir uma malha regular interpolada, com espaçamento de 2 km e altitude de 1 km (Figura 3.3). Definimos uma malha verdadeira amostrando o campo dos corpos numa malha regular (Figura 3.4). Subtraímos a malha regular da malha do modelo de fontes equivalentes e obtemos a malha dos resíduos (Figura 3.5).

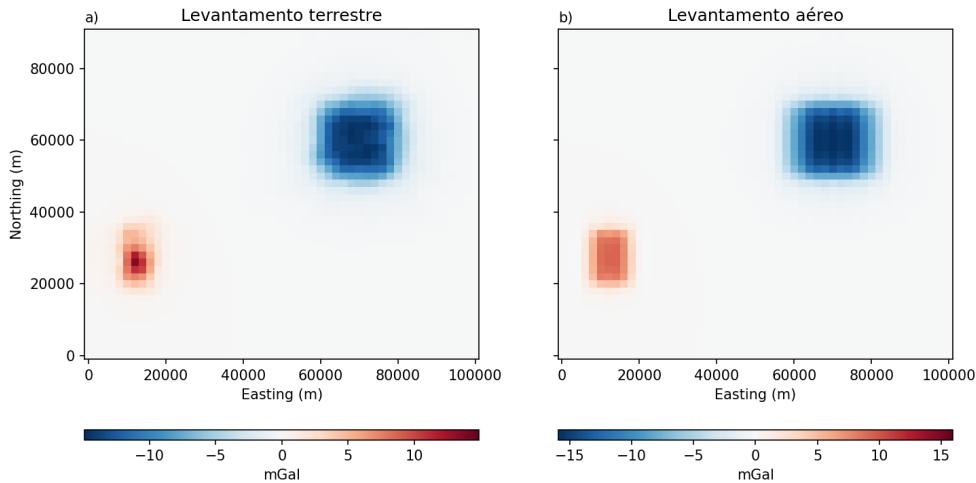


Figura 3.3: Interpolação dos dados gravimétricos por fontes equivalentes com *damping* de 1. A profundidade das fontes e o espaçamento da grade equivale a 2 km.

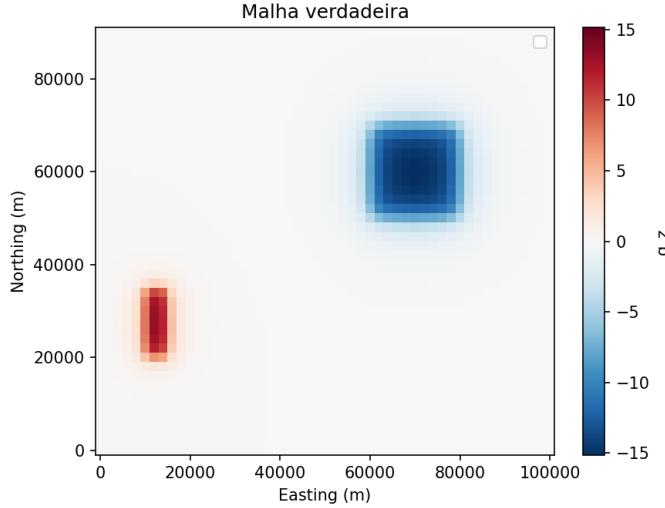


Figura 3.4: Campo gravitacional gerados pelos prismas, amostrados em uma malha regular com espaçamento de 2 km. Chamamos isso de malha verdadeira.

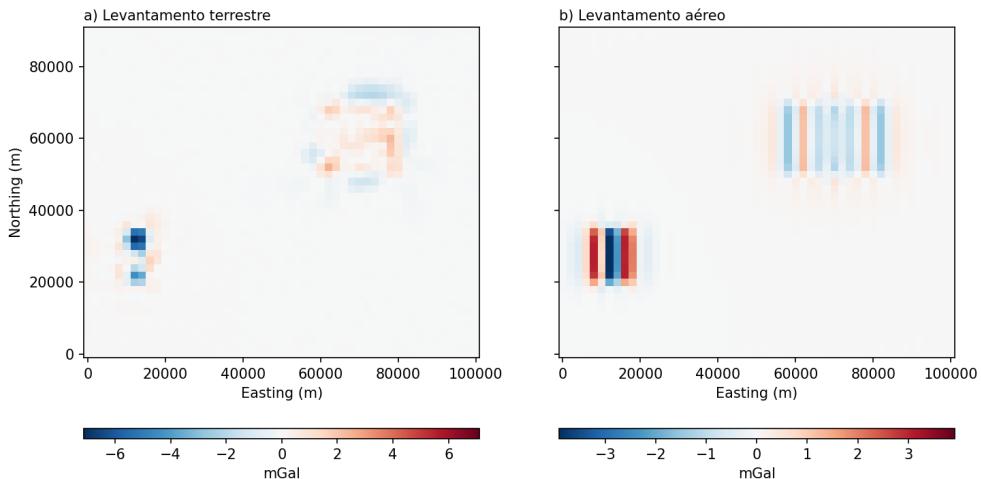


Figura 3.5: Malha de resíduos do modelo gravimétrico, obtido entre o modelo de fontes equivalentes e a malha verdadeira.

O resíduo, calculado a partir da diferença entre o modelo de fontes equivalentes e a malha verdadeira, foi utilizado para estimar o R^2 verdadeiro. O valor foi de aproximadamente 0.97 para ambos os levantamentos.

Realizamos a validação cruzada por blocos do modelo de fontes equivalentes, variando o tamanho do bloco de 1 km a 10 km. Nos gráficos da Figura 3.6 temos o R^2 verdadeiro e a média dos parâmetros R^2 calculados com cada tamanho de bloco. Comparamos estes valores para estimar qual tamanho de bloco é o que mais nos aproxima do valor do R^2 verdadeiro.

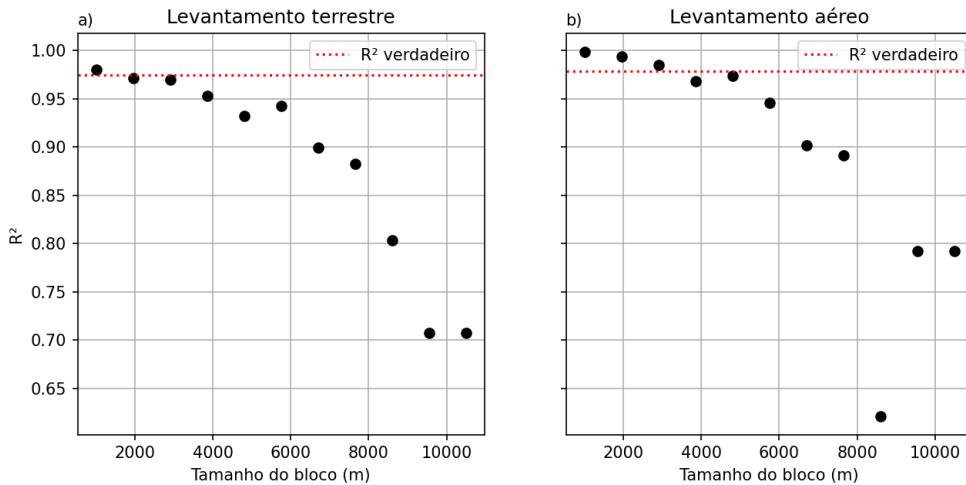


Figura 3.6: Efeito da variação do tamanho do bloco no valor médio do parâmetro R^2 , calculado com a validação cruzada por blocos (*Block k-Fold*). Eixo y indica os valores de R^2 e eixo x indica os valores de blocos testados entre 1 km e 10 km. O R^2 médio calculado para cada tamanho de bloco é representado pelos pontos pretos. Em destaque vermelho é o R^2 verdadeiro obtido pela equação 2.11. Para ambos levantamentos esse valor é de aproximadamente 0.97.

Pela análise dos gráficos da Figura 3.6 temos que o tamanho de bloco que melhor nos aproxima do R^2 verdadeiro é de 3 km para o levantamento terrestre e de 5 km para o aerolevantamento.

3.2 Dados sintéticos do campo magnético

Para gerar a grade dos modelos sintéticos do campo magnético seguimos os procedimentos descritos na seção 2.3.1.

3.2.1 Levantamento terrestre

Para simular o levantamento terrestre, amostramos as fontes magnéticas nos pontos de observações gerados (Figura 3.7a). As profundidades das fontes são: 2 km e 4 km para os dipolos, 2 km para o pipe, 1 km para a linha, com direção N-S, e 0.5 km para a soleira, com um dique aflorante na direção E-W. Ajustamos o modelo de fontes equivalentes aos dados com *damping* de 1 e profundidade de 1 km e em seguida utilizamos o modelo para produzir uma malha regular interpolada, com espaçamento de 1 km e altitude de 2 km (Figura 3.7b). Definimos a malha verdadeira e calculamos o resíduo entre o modelo e a malha verdadeira (Figura 3.8). O valor estimado para o R^2 verdadeiro foi de 0.86.

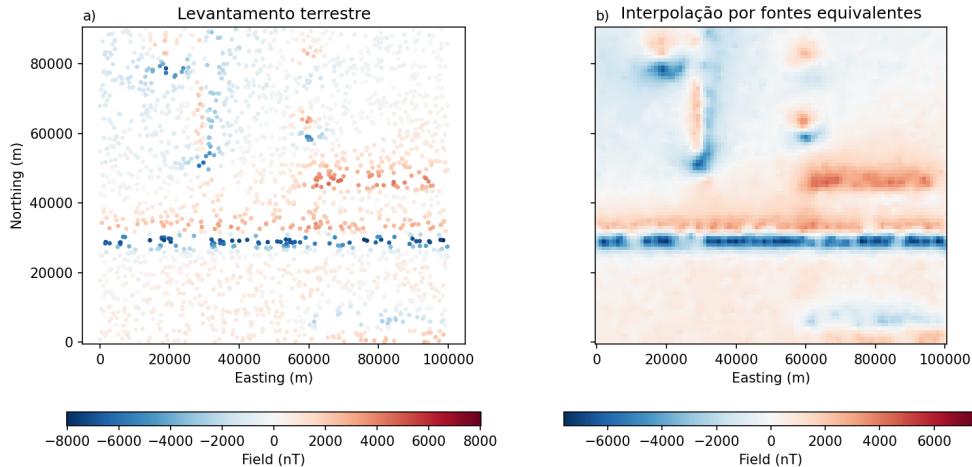


Figura 3.7: a) Resposta do campo magnético gerado pelos corpos. b) Resultado para a interpolação dos dados utilizando o método de fontes equivalentes. É possível notar que em b) as fontes estão melhores demarcadas.

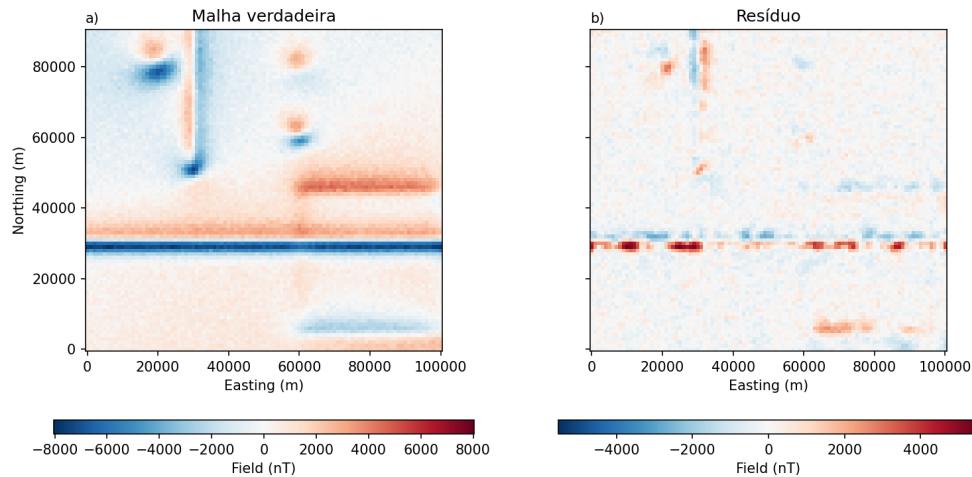


Figura 3.8: Em a) é o resultado do campo magnético gerado pelos corpos geológicos sintéticos, amostrados em uma malha regular, conhecido como malha verdadeira. Em b) é o resíduo entre o modelo interpolado por fontes equivalentes e a malha verdadeira.

Realizamos a validação cruzada por blocos (*Block k-Fold*) e a validação cruzada aleatória (*k-Fold*) do modelo de fontes equivalentes. No gráfico da Figura 3.9 temos o R^2 verdadeiro e a média dos parâmetros R^2 calculados com cada tamanho de bloco. Temos também o R^2 calculado com a validação cruzada aleatória (R^2 aleatório, que equivale a 0.88).

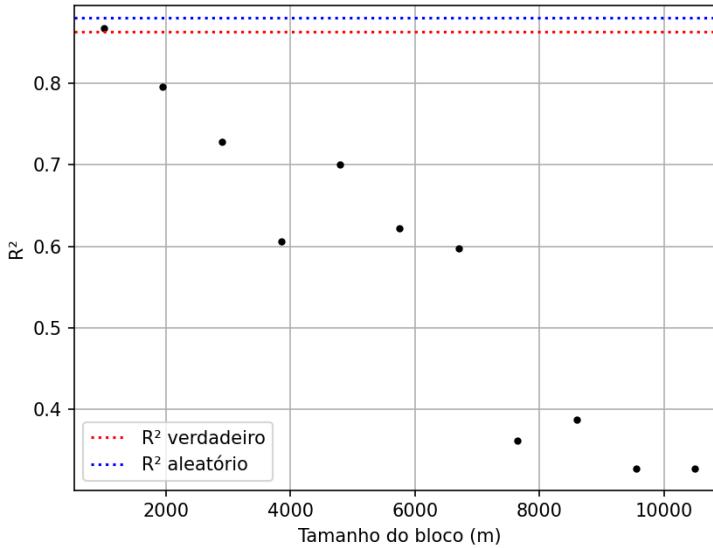


Figura 3.9: Comparação do parâmetro R^2 para cada método utilizado. Os círculos pretos indicam o R^2 médio obtido para cada tamanho de bloco utilizado na validação cruzada por blocos. A linha pontilhada vermelha é o R^2 verdadeiro estimado e em azul o R^2 obtido pela validação cruzada aleatória.

Da análise do gráfico na Figura 3.9, temos que o tamanho de bloco ideal é de 1 km, aproximadamente. Temos também que o R^2 retornado pela validação cruzada aleatória tem um valor próximo do R^2 verdadeiro para este modelo (linhas pontilhadas azul e vermelha, respectivamente).

3.2.2 Levantamento aéreo

Para simular o primeiro modelo do aerolevantamento, amostramos as fontes magnéticas nos pontos de observações gerados (Figura 3.10a). As profundidades das fontes são as mesmas do modelo terrestre anterior. Ajustamos o modelo de fontes equivalentes aos dados com *damping* de 1 e profundidade de 1 km e em seguida utilizamos o modelo para produzir uma malha regular interpolada, com espaçamento de 1 km e altitude de 500 m (Figura 3.10b). Definimos a malha verdadeira e calculamos o resíduo entre o modelo e a malha verdadeira (Figura 3.11). O valor estimado para o R^2 verdadeiro foi de 0.78.

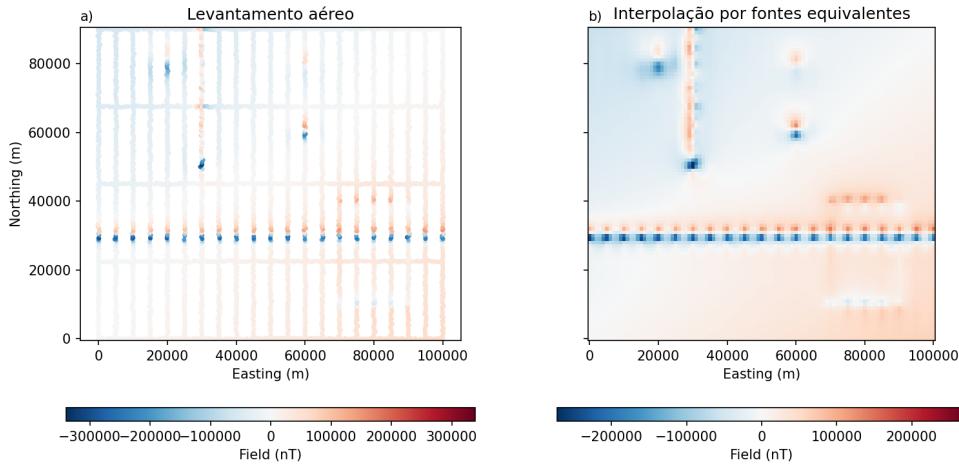


Figura 3.10: a) Levantamento aéreo do campo magnético, com espaçamento de 5 km entre as linhas de voo e 500 m ao longo das linhas de voo. b) Resultado para a interpolação dos dados utilizando o método de fontes equivalentes.

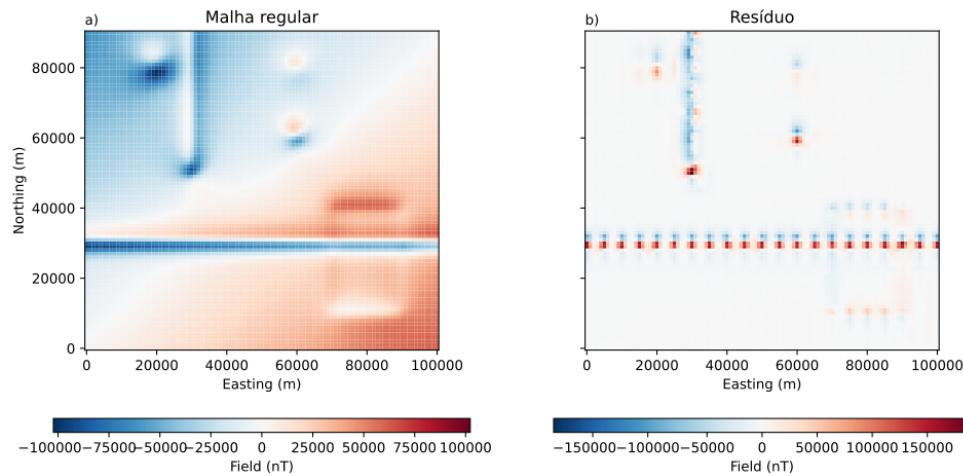


Figura 3.11: Em a) temos o campo magnético gerado pelas múltiplas fontes sintéticas, amostrados em uma malha regular, conhecida como malha verdadeira. Em b) temos o resíduo entre o modelo interpolado por fontes equivalentes e a malha verdadeira.

Realizamos a validação cruzada por blocos (*Block k-Fold*) e a validação cruzada aleatória (*k-Fold*) do modelo de fontes equivalentes. No gráfico da Figura 3.12 temos o R^2 verdadeiro e a média dos parâmetros R^2 calculados com cada tamanho de bloco. Temos também o R^2 calculado com a validação cruzada aleatória, que equivale a 0.96.

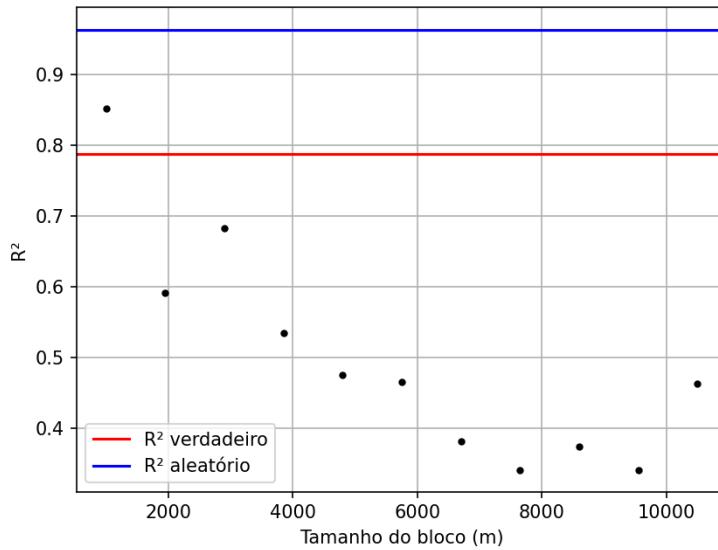


Figura 3.12: Comparação do parâmetro R^2 para cada método utilizado. Os círculos pretos indicam o R^2 médio obtido para cada tamanho de bloco utilizado na validação cruzada por blocos. A linha pontilhada vermelha é o R^2 verdadeiro estimado e em azul o R^2 obtido pela validação cruzada aleatória.

Da análise do gráfico na Figura 3.12, temos que o tamanho de bloco ideal varia entre 1 km e 3 km, aproximadamente. Podemos também observar que a validação cruzada aleatória superestima o valor do R^2 (distância entre a reta vermelha e a reta azul). Essa diferença é considerável, diferentemente do modelo terrestre anterior.

No segundo modelo do aerolevantamento, diminuímos a intensidade do sinal das fontes e alteramos as profundidades dos dipolos, tornando-os mais rasos, a 1.5 km e a 2 km. Alteramos também a profundidade da linha (com direção N-S) para 1.5 km, tornando-a mais profunda. Dessa maneira, diminuímos a amplitude do sinal do campo em algumas ordens de grandeza. O nível de ruído permaneceu constante.

O modelo de fontes equivalentes foi ajustado aos dados da mesma maneira que o aerolevantamento 1 (Figura 3.13b). A malha verdadeira e o resíduo estão ilustrados na Figura 3.14. O valor estimado para o R^2 verdadeiro foi de 0.24. Observa-se uma significativa diminuição do parâmetro R^2 , comparado ao modelo anterior, devido a diminuição da amplitude do sinal das fontes magnéticas sintéticas.

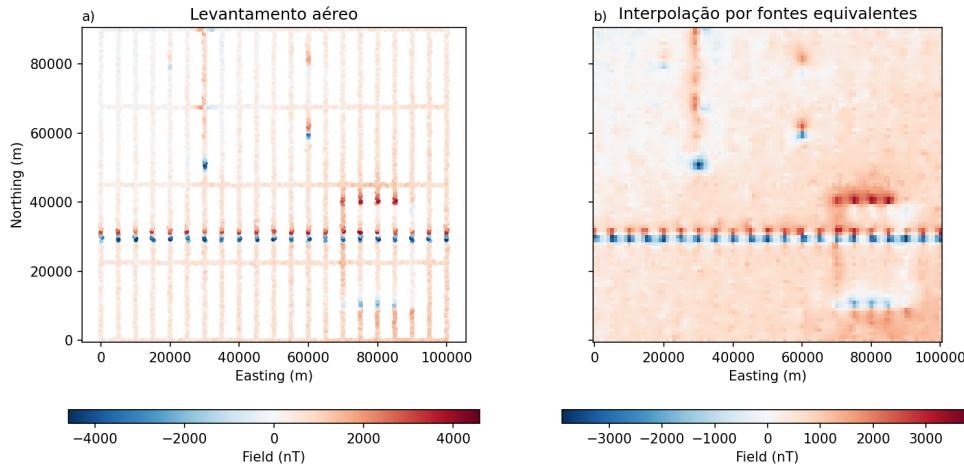


Figura 3.13: a) Resposta do campo magnético gerado pelos corpos. b) Resultado para a interpolação dos dados utilizando o método de fontes equivalentes. É possível notar que em b) as fontes estão melhores demarcadas.

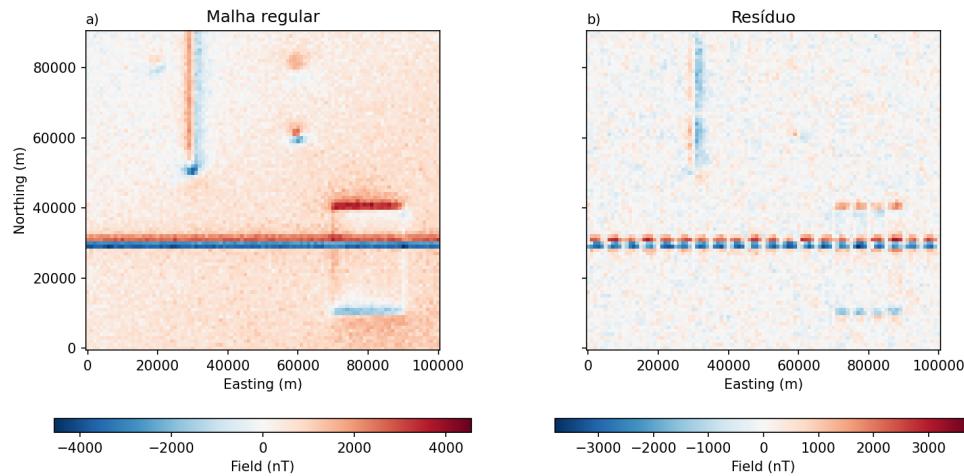


Figura 3.14: Em a) é o resultado do campo magnético gerado pelos corpos geológicos sintéticos, amostrados em uma malha regular, conhecido como malha verdadeira. Em b) é o resíduo entre o modelo interpolado por fontes equivalentes e a malha verdadeira.

Realizamos a validação cruzada por blocos (*Block k-Fold*) e a validação cruzada aleatória (*k-Fold*) do modelo de fontes equivalentes. No gráfico da Figura 3.15 temos o R^2 verdadeiro e a média dos parâmetros R^2 calculados com cada tamanho de bloco. Temos também o R^2 calculado com a validação cruzada aleatória, que equivale a 0.76.

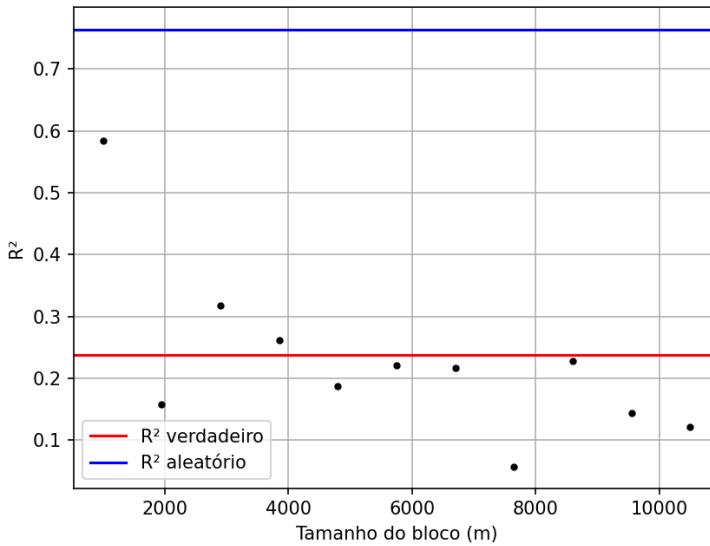


Figura 3.15: Comparação do parâmetro R^2 para cada método utilizado. Os círculos pretos indicam o R^2 médio obtido para cada tamanho de bloco utilizado na validação cruzada por blocos. A linha pontilhada vermelha é o R^2 verdadeiro estimado e em azul o R^2 obtido pela validação cruzada aleatória.

Da análise do gráfico da Figura 3.15, temos que o tamanho de bloco ideal varia entre 2 km e 9 km, aproximadamente. Podemos observar que o efeito de superestimação do parâmetro R^2 da validação cruzada aleatória é maior neste modelo que no anterior. Enquanto que a utilização dos blocos na validação cruzada aproxima o valor do R^2 do verdadeiro (observado nos pontos em preto oscilando ao redor da reta vermelha).

A partir do modelo descrito anteriormente, estudamos o efeito da variação do espaçamento entre as linhas de voo na qualidade da interpolação por Fontes Equivalentes. Testamos espaçamentos de 500 m, 2 km e 4 km. Combinamos os resultados com o resultado do modelo anterior, com espaçamento de 5 km entre as linhas. Na Figura 3.16 temos as curvas para cada modelo de espaçamento entre as linhas de voo (curvas contínuas). Tais curvas representam a relação do parâmetro R^2 com o tamanho de bloco. Temos também representado o parâmetro R^2 verdadeiro para cada modelo (linhas pontilhadas). Observamos que conforme o espaçamento entre as linhas de voo aumenta (curva pontilhada vermelha) o R^2 verdadeiro diminui, pois temos mais lacunas nos dados que necessitam de previsão.

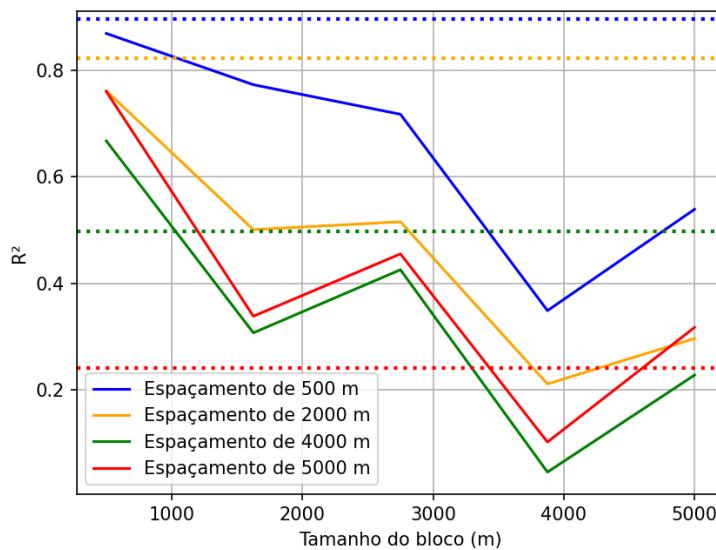


Figura 3.16: Variação do R^2 verdadeiro e o espaçamento entre as linhas de voo. Para cada espaçamento entre as linhas de voo foi calculado um R^2 médio para diversos tamanhos de bloco utilizado na validação cruzada. A linha pontilhada indica o valor do parâmetro R^2 verdadeiro com seus respectivos espaçamentos.

Capítulo 4

Discussão

Iniciamos a construção dos dados sintéticos do modelo gravimétrico simulando uma grade para o levantamento terrestre e outra para o levantamento aéreo. Este modelo inicial se trata de um caso simples, onde não adicionamos ruído aos dados. A altitude de observação não varia de ponto a ponto. As linhas de voo se estendem ao longo de apenas uma direção e são perfeitamente uniformes (Figura 3.1), descrevendo o caso mais ideal trazido por este trabalho. Ao interpolar os dados por fontes equivalentes, observamos que não há diferença nítida entre os modelos do levantamento terrestre e aéreo na Figura 3.3. No entanto, ao calcular a diferença entre estes levantamentos e a malha verdadeira (Figura 3.4), obtemos o resíduo (Figura 3.5), onde é possível notar as discrepâncias nos levantamentos.

Na Figura 3.5b observamos o padrão do levantamento aéreo, devido ao dado ser super amostrado ao longo das linhas de voo. A presença de lacunas nos levantamentos (dados faltantes entre as linhas de voo) motivou a investigação da dependência entre o espaçamento das linhas de voo e o tamanho de bloco ideal utilizado na validação cruzada por blocos. Da análise dos gráficos na Figura 3.6, temos que o bloco ideal para o levantamento terrestre foi de $3\ km$, pela proximidade do R^2 calculado com o R^2 verdadeiro. Para o aerolevantamento, tivemos um bloco ideal de $5\ km$, que equivale ao espaçamento entre as linhas de voo. Dessa maneira, para este modelo, parece haver uma dependência entre o padrão das linhas de voo e o tamanho de bloco utilizado na validação cruzada por blocos. Todavia, esta dependência não foi observada nos demais modelos onde trabalhamos com o campo magnético.

Ao estudar o campo magnético, construímos modelos mais complexos, com múltiplas fontes e adição de ruído. No primeiro modelo sintético do campo magnético, simulamos um levantamento terrestre que teve como parâmetro R^2 verdadeiro estimado em 0.86. Levando em consideração que um modelo perfeito teria este valor igual a 1, temos que a interpolação está consideravelmente

próxima do dado verdadeiro. Da análise do gráfico na Figura 3.6 vemos que a estimativa do parâmetro R^2 pela validação cruzada aleatória (no valor de 0.88) não está demasiadamente distante do valor verdadeiro e a utilização de um bloco de 1 km foi ideal para alcançar este valor.

Para simular os aerolevantamentos sintéticos dos dados do campo magnético, utilizamos dois modelos que diferenciam na amplitude do sinal das fontes magnéticas. O primeiro modelo apresentado tem a amplitude do sinal de 100000 nT (Figura 3.11a) e o segundo de 4000 nT (Figura 3.14a). Como mantemos constante o nível de ruído, o modelo com sinal mais fraco é mais suscetível às perturbações, fazendo com que este tenha a menor estimativa do parâmetro R^2 verdadeiro, num valor de 0.24. Enquanto que para o modelo de sinal mais forte temos uma estimativa de 0.78. A discrepância destes valores diz respeito à qualidade do sinal medido, tendo em vista que todos os outros parâmetros como espaçamento da grade e quantidade de pontos permaneceram constantes. O modelo com sinal mais fraco é o mais próximo de medidas realizadas no mundo real, onde o sinal das fontes gera perturbações relativamente pequenas no campo magnético observado. Da análise dos gráficos das Figuras 3.12 e 3.15, vemos que a superestimação do valor do parâmetro R^2 , calculado pela validação cruzada aleatória, é maior no modelo com sinal mais fraco. Este valor está representado pelas linhas azuis dos gráficos. Neste modelo do gráfico da Figura 3.15 temos uma estimativa de 0.24 para o R^2 verdadeiro, e uma estimativa de 0.76 feita pela validação cruzada aleatória. Enquanto isso, a utilização de um bloco com tamanho de 2 km até 9 km foi o que aproximou melhor o parâmetro do valor verdadeiro. Desta análise podemos inferir que, por não levar em consideração a autocorrelação dos dados discutida na seção 2.2, a validação cruzada aleatória superestima o parâmetro de qualidade do modelo. Este efeito é consideravelmente maior quando a qualidade dos dados é pior, neste caso com o ruído mais acentuado.

Estudamos a variação do espaçamento entre as linhas de voo do aerolevantamento na Figura 3.16. Da análise do gráfico temos que os modelos com espaçamento de 4 km (curva verde) e 5 km (curva vermelha) têm o tamanho de bloco ideal em aproximadamente 1.5 km. Os modelos com espaçamento menor (curvas amarela e azul) não produziram valores próximos do R^2 verdadeiro para nenhum tamanho de bloco testado. Ou seja, em levantamentos bem amostrados, a utilização de blocos na validação cruzada não aproximou o valor do parâmetro de determinação do verdadeiro. Já nos levantamentos mal amostrados, como nos casos das curvas verde e vermelha, onde temos muitas lacunas entre os dados, a utilização dos blocos nos aproximou do valor verdadeiro.

Capítulo 5

Conclusões

Ao refinar os modelos preditivos obtidos a partir da inversão por fontes equivalentes, estudamos a variação do tamanho do bloco na validação cruzada por blocos. A partir de modelos sintéticos mais complexos, vimos que não parece haver uma relação entre o espaçamento das linhas de voo e o tamanho do bloco utilizado na validação. No entanto, o espaçamento entre as linhas de voo dita a quantidade de lacunas vazias entre os dados. Quanto mais mal amostrado for o levantamento, ou quanto pior for a razão sinal/ruído, mais necessário se faz a divisão por blocos para estimar a qualidade do ajuste verdadeiro do modelo de fontes equivalentes.

Temos que na validação cruzada aleatória, as estruturas de dependência dos dados não são levadas em consideração, gerando *overfitting* e fornecendo parâmetros de qualidade de ajuste otimistas. Neste estudo vimos que a superestimação feita pela validação cruzada aleatória pode chegar a ser superior a 3 vezes o valor verdadeiro (Figura 3.15). A utilização da validação cruzada por blocos se fez indispensável na determinação do parâmetro de qualidade do ajuste nestes casos.

Tendo em vista que parece existir um padrão dos valores do parâmetro R^2 oscilarem conforme aumentamos o tamanho dos blocos (Figuras 3.9, 3.12, 3.15), deve-se levar em consideração diferentes configurações de modelos sintéticos, variando o sinal das fontes. Dessa forma, futuros trabalhos deverão investigar se há uma relação entre o comprimento de onda emitida pelas fontes e o tamanho de bloco ideal utilizado na validação. Para conjuntos de dados grandes seria interessante investigar a aplicação da validação cruzada em uma parcela dos dados e a viabilidade da extração dos resultados para o resto do levantamento. Por fim, os resultados poderão futuramente impactar na produção de malhas que combinam diversos levantamentos para grandes territórios.

Referências Bibliográficas

Arlot S., Celisse A., A survey of cross-validation procedures for model selection, *Statistics Surveys*, 2010, vol. 4

Bahn V., McGill B. J., Testing the predictive performance of distribution models, *Oikos*, 2012, vol. 122, p. 321–331

Cordell L., A scattered equivalent-source method for interpolation and gridding of potential-field data in three dimensions, *Geophysics*, 1992, vol. 57, p. 629

Dampney C., The equivalent source technique, *Geophysics*, 1969, vol. 34, p. 39

Fatiando a Terra Project Castro Y. M., Esteban F. D., Li L., Oliveira Jr V. C., Pesce A., Shea N., Soler S. R., Souza-Junior G. F., Tankersley M., Uieda L., Uppal I., , 2024 Harmonica v0.7.0: Forward modeling, inversion, and processing gravity and magnetic data

Geisser S., The Predictive Sample Reuse Method with Applications, *Journal of the American Statistical Association*, 1975, vol. 70, p. 320–328

Leão J. W. D., Silva J. B. C., Discrete linear transformations of potential field data, *GEOPHYSICS*, 1989, vol. 54, p. 497–507

Mahoney M. J., Johnson L. K., Silge J., Frick H., Kuhn M., Beier C. M., , 2023 Assessing the performance of spatial cross-validation approaches for models of spatially structured data

Mendonça C. A., Subspace method for solving large-scale equivalent layer and density mapping problems, *GEOPHYSICS*, 2020, vol. 85, p. G57–G68

Oliveira Junior V. C., Takahashi D., Reis A. L. A., Barbosa V. C. F., Computational aspects of the equivalent-layer technique: review, *Frontiers in Earth Science*, 2023, vol. 11

Roberts D. R., Bahn V., Ciuti S., Boyce M. S., Elith J., Guillera-Arroita G., Hauenstein S., Lahoz-Monfort J. J., Schröder B., Thuiller W., Warton D. I., Wintle B. A., Hartig F., Dormann C. F., Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, *Ecography*, 2017, vol. 40, p. 913–929

Soler S. R., Uieda L., Gradient-boosted equivalent sources, *Geophysical Journal International*, 2021, vol. 227, p. 1768–1783

Stone M., Cross-Validatory Choice and Assessment of Statistical Predictions, *Journal of the Royal Statistical Society: Series B (Methodological)*, 1974, vol. 36, p. 111

Tikhonov A., *Solutions of Ill-Posed Problems*. V.H. Winston, 1977

Uieda L., Verde: Processing and gridding spatial data using Green's functions, *Journal of Open Source Software*, 2018, vol. 3, p. 957