Método de Regressão Linear Múltipla aplicado em dados gravimétricos

Tales Ferraz de Paula*
Alice Marques Pereira Lau**
Walter Filgueira de Azevedo Jr.***

Resumo:

A gravimetria terrestre é um método potencial geofísico que estuda anomalias gravimétricas do campo gravitacional da Terra. Para tal, uma etapa de coleta de dados é necessária para obtenção do valor da gravidade. Uma vez que esse dado em campo tem dependência das propriedades físicas da Terra, latitude, altura em relação a um *datum* e densidade local, foram feitos modelos matemáticos lineares capazes de prevê-lo para uma área de estudo selecionada. Foi utilizado o método de aprendizado de máquina de regressão linear múltipla para construção destes modelos, sendo o valor da gravidade a variável dependente e as propriedades físicas da Terra as independentes. Os dados gravimétricos utilizados foram cedidos pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – ANP, cujas áreas amostradas compreendem as regiões sul, sudeste e centro oeste do Brasil. Os modelos apresentaram coeficientes de correlação muito próximos de 1,0, que indica uma dependência do dado de campo às propriedades físicas aproximadamente linear. Contudo, em regiões com grandes variações de altura, os modelos apresentaram valores de gravidade mais altos que o valor medido em campo, o que indica a limitação destes modelos a regiões com pequenas variações de altura.
Palavras-chave: Geofísica. Aprendizado de máquina. Previsão de gravidade.

Abstract:

Terrestrial gravimetry is a potential geophysical method that studies gravimetric anomalies of Earth's gravitational field. For such, it is necessary a field's data collect for the value of gravity. Since this field's data has dependence from Earth's physical properties, latitude, height in relation to a datum and local density, linear mathematical models capable of predict such data, for a selected study area, were built. The models were made by the machine learning method of multiple linear, were the gravity value was the dependent variable and the Earth's physical properties were the independent ones. The gravimetric data utilized were given by the National Agency for Petroleum, Natural gas and Biofuels – ANP, whose areas comprise the regions from Brazil's south, southeast and Midwest. The models showed correlation coefficients very close to 1.0, which indicates that such dependency from the field's data to these physical properties is approximately linear. However, in regions with large height variations, the models presented gravity values higher than the one measured in field, indicating the limitation of those models to regions with small height variations.
Key-words: Geophysics. Machine Learning. Gravity prediction.

INTRODUÇÃO

O estudo da distribuição do campo gravitacional terrestre é feito por meio de aproximações e correções de modelos matemáticos que consideram uma superfície elipsoide e o geoide, que é o modelo físico da forma da Terra, para simular o planeta e

*Graduando do curso de Bacharelado em Física: linha de formação em Geofísica da Pontifícia Universidade Católica do Rio Grande do Sul. E-mail: tales.paula98@edu.pucrs.br

**Orientadora: Professora adjunta da Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul e doutora em Geologia pela Universidade Federal do Paraná. E-mail: alice.lau@pucrs.br.

***Coorientador: Professor adjunto da Escola de Ciências da Saúde e da Vida e doutor em Física pela Universidade de São Paulo. E-mail: walter@azevedolab.net.

seu campo gravitacional em relação a dimensão e distribuição mássica deste[1]. Estes modelos têm parâmetros físicos, tais como velocidade angular, massa e centro de massa, iguais ou muito próximos aos do planeta Terra e são tratados, à gravimetria, de modo a oferecerem um valor de referência, teórico, às medições da aceleração gravitacional terrestre (valor experimental) para se obter uma anomalia gravimétrica e se realizar as devidas interpretações sobre estas[1,2]. A **fig. 1** apresenta uma comparação da superfície terrestre com os modelos do elipsoide e do geoide. O método geofísico que mede, analisa, corrige e interpreta os valores da gravidade e anomalias gravimétricas é o método da gravimetria.

Gravimetria

Consiste na detecção e análise de variações no campo gravitacional terrestre causadas pela diferença de densidade dos corpos em subsuperfície terrestre (**fig. 2**). Comumente, valor da gravidade é medido no local de interesse, gravidade observada (g_{obs}), por meio de um gravímetro, após feitas correções iniciais de maré, relativas ao período do dia, e de instrumento, que são quaisquer instabilidades durante a leitura e *drift* do instrumento[3,4]. O gravímetro é um instrumento extremamente sensível, com precisão, muitas vezes, na ordem de microGals (ver equação 1). Para cálculo da anomalia gravimétrica, são aplicadas, então, as correções: de latitude, que dará a gravidade teórica, g_{teo}, esperada pelo modelo aproximado da Terra, cujo valor é função da latitude e da densidade do modelo assumido; *free air* (g_f) que corrige uma variação no dado causada pela de diferença de altura entre o ponto de medida e o *datum* de referência (uma altura referência selecionada), e ainda a contribuição da massa; da calota de Bouguer (g_c) usada para corrigir efeitos causados pela presença de massa entre o *datum* e o local de aquisição; e de terreno (g_{ter}) muitas vezes não utilizada – por apresentar pouca contribuição ao dado – corrige pequenas variações de topografia do local de aquisição[1,2,3,4]. A correção g_{teo} tem como objetivo corrigir o valor da gravidade para diferentes posições no globo: a gravidade na linha equatorial é menor que nos polos, por conta da rotação planetária. g_f tem como objetivo situar a altura da estação de coleta à altura do *datum* adotado, e g_c adiciona, ou retira, qualquer contribuição mássica que exista entre essa variação de altura. Por fim, após realizadas as correções supracitadas sobre o valor da gravidade observada, tem-se o resultado principal do dado gravimétrico, a anomalia Bouguer ($\Delta B, em\ mGal$), tal que:

$$\Delta B = g_{obs} - g_{teo} \pm g_f \pm g_c + g_{ter}\ [1\ mGal = 1 \times 10^{-5}\ m/s^2] \qquad (1)$$

Os sinais positivo ou negativo serão determinados conforme localização geográfica, ou seja, se a estação está em um nível abaixo ou acima do *datum*[3,4]. As correções citadas visam aproximar o valor experimental de g ao valor teórico, dado pelo modelo de

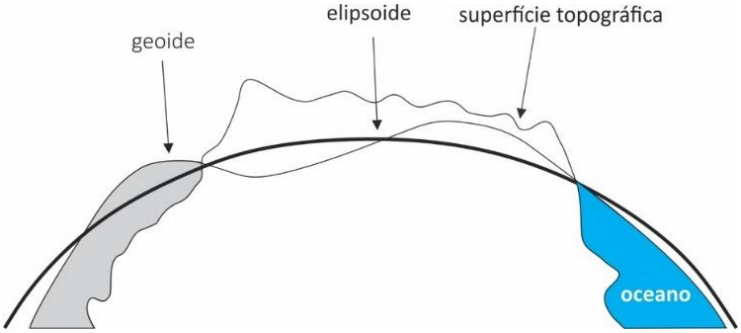

Figura 1: Comparação entre a superfície topográfica terrestre e seus modelos teóricos ao cálculo de gravidade teórica, o geoide e o elipsoide. [1, adaptado]

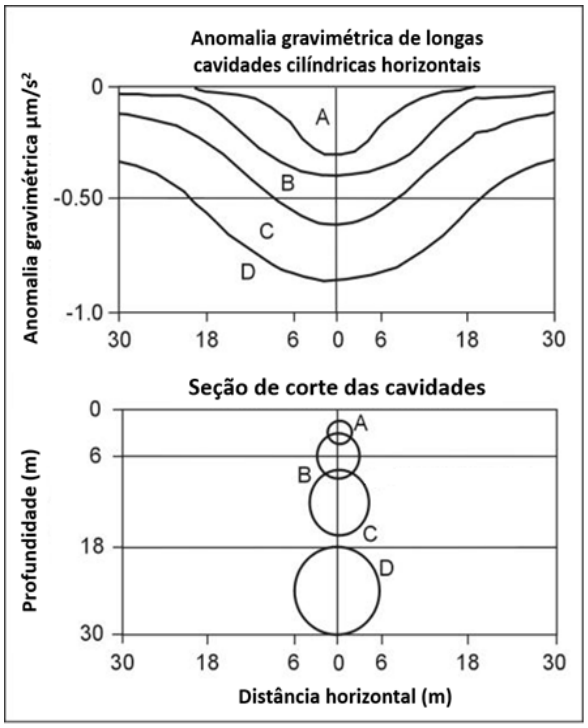

Figura. 2: Exemplo de anomalia gravimétrica causada por cavidades cilíndricas horizontais em uma rocha com densidade uniforme. [6, adaptado]

referência, seja um elipsoide ou geoide, e, a diferença entre estes é a anomalia gravimétrica. Quanto mais precisos os dados de campo, como topografia, latitude, mais preciso é o dado de anomalia gravimétrica. Dependendo, porém, da escala do levantamento gravimétrico, a precisão passa a diminuir – um levantamento local requer precisão centimétrica de altimetria ortométrica/geométrica (em relação ao geoide ou elipsoide, respectivamente), pois as anomalias estão na ordem de poucos mGal, enquanto que em um levantamento regional as anomalias estão na ordem de centenas de mGal, de modo que uma precisão métrica é aceitável para a altimetria[5].

A interpretação sobre o dado da anomalia gravimétrica permite uma estimativa de profundidade e forma do corpo causador desta por meio de inversão do dado. Todavia, a interpretação de anomalias de campos potenciais é ambígua: diferentes corpos causadores podem causar a mesma anomalia. Cabe, então, a utilização de dados externos disponíveis como conhecimento *a priori* do local de estudo, dados de poços, ou até mesmo de outro método geofísico – para diminuição desta ambiguidade[7].

Uma vez que o valor experimental da gravidade g_{obs} tem dependência implícita da rotação planetária, da altura na qual se faz o levantamento gravimétrico e da densidade do meio que exerce essa atração gravitacional[2,5], deste modo, quanto maior a massa, maior a atração, buscou-se encontrar, como primeira aproximação, uma relação matemática linear entre g_{obs} e a posição geográfica (latitude, θ), altura em relação a um *datum*, h, e densidade, ρ, utilizando o método de aprendizado de máquina de regressão linear para previsão dos dados de g_{obs}. Com o modelo matemático proposto, pode-se analisar uma possível supressão das etapas de campo de um levantamento gravimétrico, levando em conta a precisão do modelo, fazendo-se com que seja necessário somente dados de h, ρ e θ para cálculo de ΔB – o que pode reduzir significativamente tempo e custo de um levantamento de dados gravimétricos, pois h e θ podem facilmente ser obtidos com precisão aceitável com base em cartas geográficas. Além disso, é comum adotar-se um valor fixo de ρ para cálculo das correções[1,3,4,5,8], de $2{,}67 - 2{,}7\ g\ cm^{-3}$,

mas, para maior precisão, pode-se utilizar valores com base em mapas geológicos. Neste trabalho, foram utilizados valores de ρ fornecidos pelos dados. Descrições pormenorizadas sobre o método da gravimetria e teoria gravimétrica podem ser encontradas em [7,9].

Machine Learning

O método de regressão linear é um dos mais simples e comumente aplicados métodos na aprendizagem de máquina para determinação de uma função linear – a equação de uma reta – entre duas ou mais variáveis[10,11,12] com intuito de previsão de dados. Para verificação da precisão do modelo matemático gerado, comumente usa-se o coeficiente de correlação (R^2) como uma estimativa da qualidade do modelo de previsão[11,12] – quanto mais próximo de 1, menor a diferença entre os valores previstos e os reais.

Na geofísica, o uso da aprendizagem de máquina ganhou espaço nos últimos anos, desde integração de atributos geológicos de dados de poços geofísicos para estimação de permeabilidade[13], integração de dados de diferentes métodos geofísicos para exploração e mitigação de riscos ambientais[14], até identificação e classificação de rochas reservatório[15], entre outras aplicações[16,17]. Cada um destes autores trabalhou com métodos diferentes, mas a metodologia é a mesma: treinamento de máquina com dados de modo a desenvolver modelo matemático para reconhecimento de padrões e previsão de dados com certa acurácia. Maiores detalhes sobre *machine learning* e regressão linear podem ser encontradas em [11].

Dados e modelagem matemática

Os dados gravimétricos analisados e empregados para treinamento e determinação do modelo matemático foram disponibilizados pela ANP, intitulados de 0401_GRAV_DESUDNEXPAR_PARANA, que consiste de um levantamento gravimétrico de 20806 pontos. A região do levantamento corresponde às regiões sul,

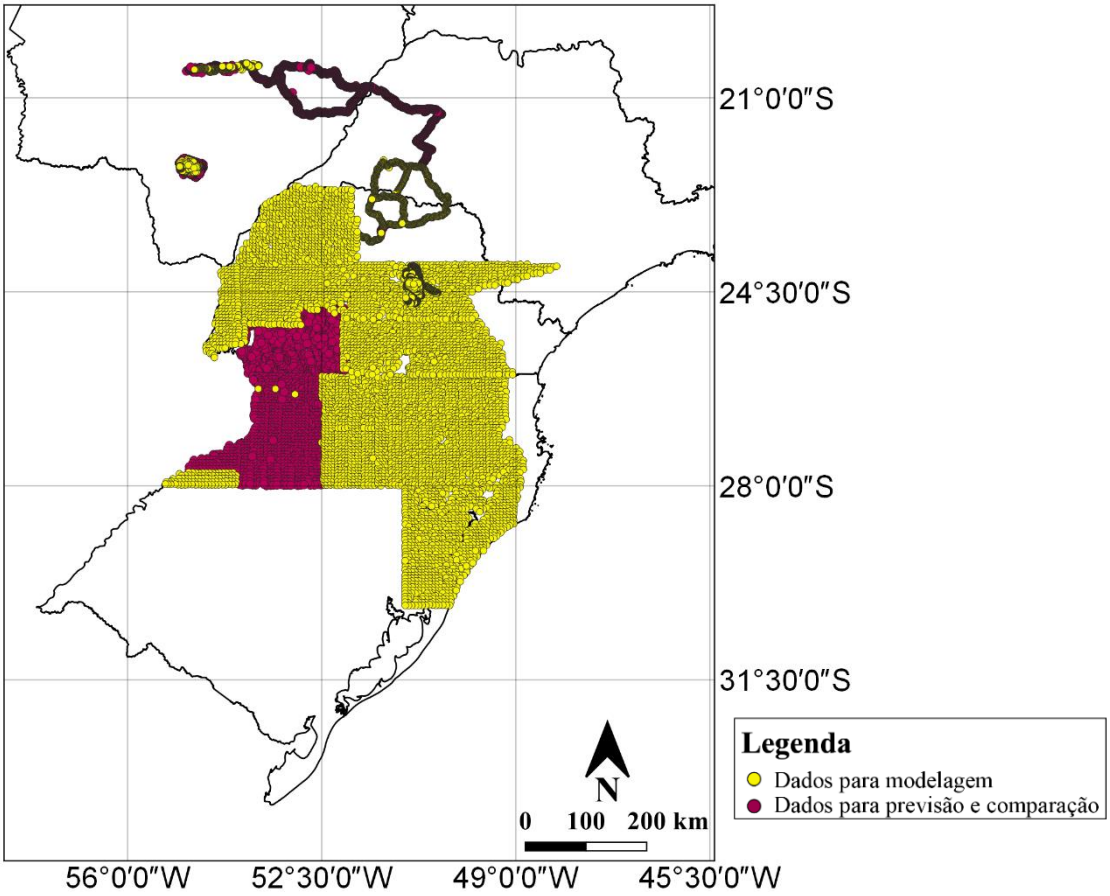

Figura. 3: Distribuição geográfica dos dados gravimétricos e dados selecionados para MLR_1

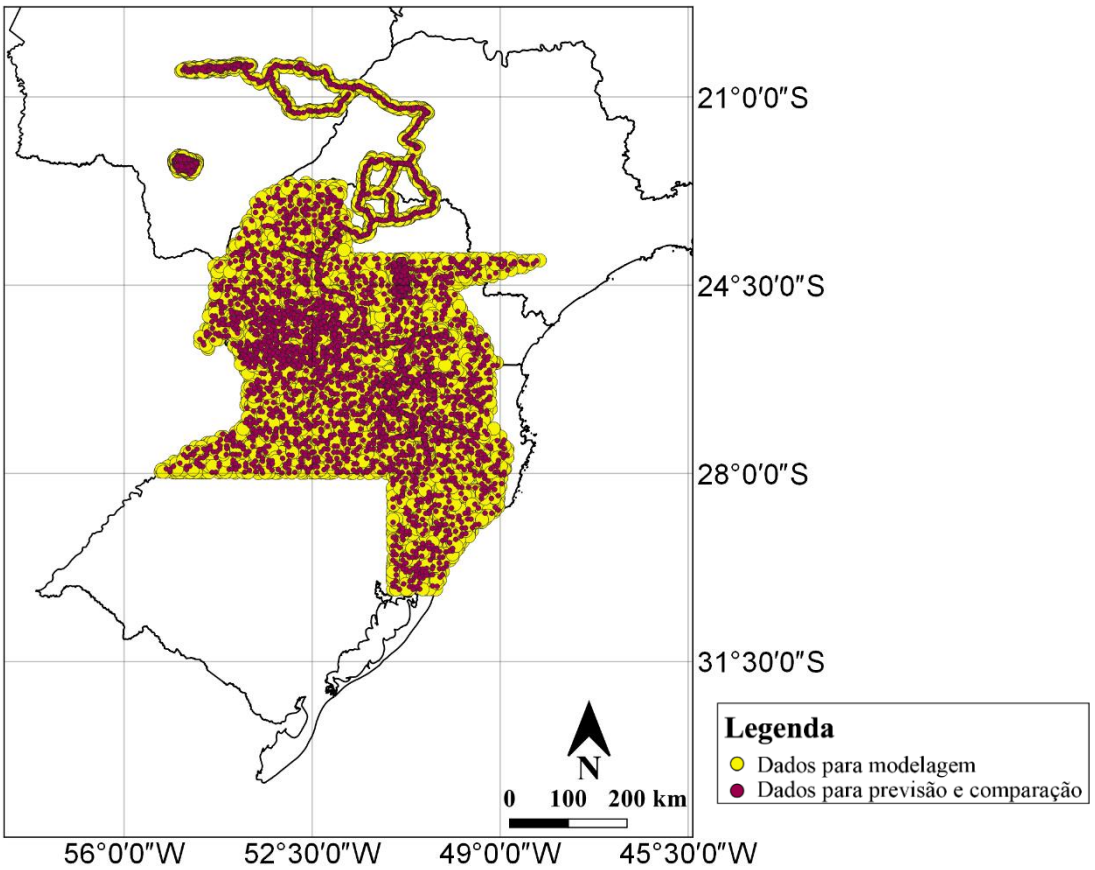

Figura. 4: Distribuição geográfica dos dados gravimétricos e exemplo de modelo aleatório utilizado para MLR_2.

sudeste e centro oeste do Brasil, englobando principalmente os estados de Rio Grande do Sul, Santa Catarina e Paraná, além dos estados do Mato Grosso do Sul e São Paulo (**fig. 3**).Foram feitos cinco modelos matemáticos com a utilização do *software Molegro Data Modeller 2.6.0. – software* com diversas funções para análise de dados[18]; dentre elas, criação de regressões lineares. Estes modelos são equações de 1° grau, cujas variáveis independentes são θ, ρ e h, enquanto a variável dependente é g_{obs}. Para realização do primeiro modelo, *Multiple Linear Regression* 1, MLR_1, os dados foram agrupados conforme valor de ρ, que variou nos valores de 2,1; 2,2; 2,67 e 2,7 $g\,cm^{-3}$ e foram selecionados 70% de cada valor de ρ para modelagem, totalizando 14567 dados (*training set*), de modo a obter-se assim um modelo mais abrangente e para avaliação da influência da distribuição dos dados na modelagem[11]. O restante, 6239 (30%) dados, foi utilizado para comparação com a predição – *test set* (**fig. 3**). Para realização dos outros quatro modelos, os dados foram selecionados de maneira aleatória (**fig. 4**), mantendo, porém, a mesma proporção de dados para *training* e *test set* do primeiro modelo. O valor máximo dos dados foi de 9327,15 $mGal$, e o mínimo de 8480,80 $mGal$.

RESULTADOS E DISCUSSÕES

O valor médio de R^2 dos *test sets* foi de 0,98555, sendo que todos os cinco modelos apresentam R^2 muito próximos entre si, conforme a tabela 1, que indica que os valores previstos com o modelo de regressão têm alta correlação com os valores experimentais. MLR_M foi obtido fazendo-se a média entre cada coeficiente, uma vez que o modelo é linear. A tabela 2 apresenta a relevância de cada coeficiente, mostrando que ρ, para esta área de estudo, apresenta a menor influência no dado gravimétrico, enquanto a influência da rotação da Terra, implícita em θ, possui a maior. O coeficiente independente (de valor médio de 7311.85 $mGal$) apresenta grande influência para os dados, uma vez que

g_{obs} médio de todos os dados tem valor de $8737.56\ mGal$. Já a baixa influência do valor de ρ pode ser dada justamente por este ser o parâmetro que menos varia, e é um valor tido como média local à correção g_c.

É possível observar, porém, que MLR_1, apresenta maior variação de cada parâmetro em relação aos demais e o menor R^2 dos modelos. Uma possível explicação para isso é que para este modelo os dados foram selecionados com uma metodologia diferente das demais, com maior foco em ρ, como pode ser verificado na **Fig. 3**, de modo que os dados utilizados para modelagem ficaram mais concentrados e os valores mais altos de g_{obs} foram todos utilizados para modelagem, o que explica a falta destes em MLR_1, diferentemente dos outros modelos (**fig. 5**) - é notável que o termo independente apresenta menor valor neste modelo, demonstrando essa variação.

A **fig. 5** apresenta as curvas de dispersão de cada modelo gerado. Nota-se que todas apresentam até quatro pontos distantes do restante. Estes pontos são comuns entre alguns modelos, totalizando 7 pontos nos quais a modelagem apresenta grande diferença do dado experimental (de um total de 20806 pontos). Uma análise detalhada revela que todos apresentam algo em comum: uma variação de h (Δh) muito grande com os pontos ao seu entorno. Esta é a única grande diferença entre os pontos discrepantes e o restante, uma vez que ρ destes é constante e θ varia normalmente – e é a essa variação que atribui-se tal discrepância. Assim, pode-se concluir que o modelo não é adequado para mudanças bruscas de h, ou seja, Δh grandes . A tabela 3 apresenta alguns pontos discrepantes (*) e seus parâmetros, além do dado previsto médio, bem como seus vizinhos mais próximos, para destacar Δh.

O R^2 alto, tanto o médio quanto o de cada modelo, no entanto, mostra que g_{obs} possui uma relação aproximadamente linear entre os valores de $h, \rho\ e\ \theta$ da área de estudo, o que permite uma previsão satisfatória deste valor. Para este levantamento gravimétrico de escala regional, este modelo matemático atende a uma precisão aceitável do valor de g_{obs}, conforme[5], pois a divergência dos valores está na dezena.

Tabela 1: Modelos gerados e modelo médio (MLR_M).

Nome	*Modelo*	R^2
MLR_1	$g_{obs} = -71.0036\,\theta - 53.228\,\rho - 0.234075\,h + 7282.12$	0.98254
MLR_2	$g_{obs} = -68.4751\,\theta - 44.798\,\rho - 0.230713\,h + 7320.55$	0.98693
MLR_3	$g_{obs} = -68.5248\,\theta - 44.878\,\rho - 0.231369\,h + 7319.93$	0.98725
MLR_4	$g_{obs} = -68.5865\,\theta - 45.018\,\rho - 0.231585\,h + 7319.01$	0.98509
MLR_5	$g_{obs} = -68.6345\,\theta - 44.874\,\rho - 0.231773\,h + 7317.65$	0.98593
MLR_M	$g_{obs} = -69.0449\,\theta - 46.559\,\rho - 0.231903\,h + 7311.85$	0.98555

Tabela 2: Relevância (peso) de cada coeficiente para cada modelo.

Nome	Relevância do coeficiente		
	θ	$\rho(g\ cm^{-3})$	$h\ (m)$
MLR_1	1.14590	0.08192	0.44734
MLR_2	1.14209	0.08155	0.44635
MLR_3	1.14120	0.08191	0.44730
MLR_4	1.13652	0.08090	0.44248
MLR_5	1.13891	0.08494	0.45281
MLR_M	1.12885	0.09842	0.48060

.

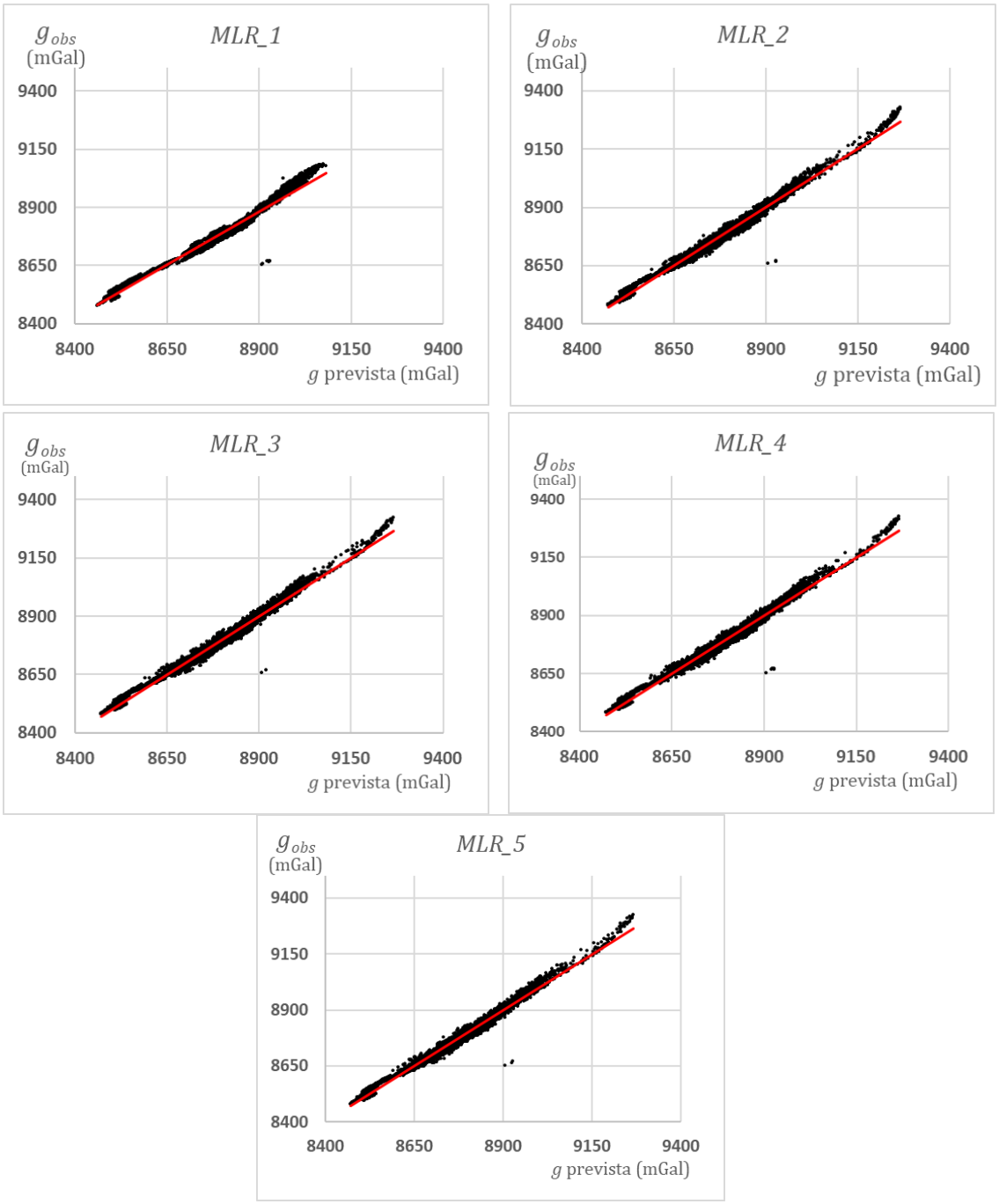

Figura 5: Curvas de dispersão de cada modelo, que comparam o dado experimental (g_{obs}) com o dado do modelo ($g_{prevista}$).

Tabela 3: Exemplo de dados discrepantes destacados e seus vizinhos próximos.

Ponto	θ	$\rho(g\ cm^{-3})$	$h\ (m)$	$g_{obs}\ (mGal)$	$g\ prevista\ (mGal)$
13218	-25.43619	2.67	694.69	8754.05	8783.45
13219	-25.33711	2.67	729.53	8728.63	8767.59
13220	-25.34926	2.67	57.29*	8663.22*	8926.55*
13221	-25.32722	2.67	37.59*	8669.14*	8926.61*
13222	-25.31427	2.67	33.48*	8667.27*	8929.57*
13223	-25.31421	2.67	849.59	8706.99	8738.53
13224	-25.29187	2.67	903.02	8697.09	8724.44
13245	-25.11167	2.67	966.95	8675.71	8711.39
13246	-25.11342	2.67	954.72	8677.32	8750.76
13247	-25.10511	2.67	4.61*	8667.03*	8919.62*
13248	-25.13299	2.67	835.81	8698.97	8703.69
13249	-25.15332	2.67	907.78	8687.82	8696.68

***** Pontos com valores de h baixos quando comparados aos dados próximos, resultando em um dado de $g\ prevista$ com grande variação do dado experimental. g_{obs}.

CONCLUSÃO

Como esta é uma primeira aproximação de um modelo matemático para algo até então novo na aplicação em dados gravimétricos, buscou-se uma modelagem mais simples, para apenas uma área de estudo. Os modelos gerados apresentaram alto coeficiente de correlação, indicando uma relação quase linear[10,12] para os parâmetros selecionados para a área de estudo – salvo nas regiões com grandes variações de altura, conforme sugerem as curvas de dispersão. Valores de ρ mais precisos e variados (ao invés de um valor médio), podem oferecer um aumento da relevância deste parâmetro aos modelos matemáticos gerados. Contudo, para elaboração de um modelo mais completo, e com maior aplicação, ou para elaboração de uma metodologia para geração de modelos gravimétricos matemáticos, uma quantidade maior de testes se faz necessária, e, para isso, mais dados são necessários. Para qualquer um dos casos, o objetivo é o mesmo: tendo-se os dados de $h,\ \rho\ e\ \theta$, que podem ser obtidos com base em mapas geológicos e de altimetria, pode-se fazer a previsão de g_{obs} da área de estudo, realizar-se as correções da equação 1, e, por fim, calcular-se ΔB.

AGRADECIMENTOS

Os autores agradecem à Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – ANP, pelo fornecimento dos dados, à Pontifícia Universidade Católica do Rio Grande do Sul – PUCRS, pela oportunidade de realização deste trabalho, e o autor de Paula agradece ao Instituto do Petróleo e dos Recursos Naturais – IPR, pela experiência vivida neste e ao Ministério da Educação – MEC, pela bolsa de estudos fornecida.

REFERENCIAL BIBLIOGRÁFICO

1 LI, X.; GÖETZE, H. Ellipsoid, geoid, gravity, geodesy, and geophysics. **Geophysics**, Tulsa, v. 66, n. 6, p. 1660-1668, novembro-dezembro de 2001;

2 VAN DER HILST. The Earth's Gravitational field. *In:* **Essentials of Geophysics**. Cambridge: MIT OpenCourseWare, 2004. Cap. 2, p. 31-77. 2004;

3 ARIFFIN, K. D. **Survei Graviti**. Geofizik Carigali, 1. ed., Bangi, 2004;

4 ARIFFIN, K. D. **Geophysical Surveying Using Gravity**. Gravity Methods, 1. ed., Bangi, 2004;

5 MOLINA, E. **Gravimetria: fundamentos e aplicações**. São Paulo: ANP. 2006;

6 WIGHTMAN, W. E.; JALINOOS, F.; SIRLES, P.; HANNA, K. **Application of Geophysical methods to Highway Related Problems.** Federal Highway Administration, FHWA-IF-04-021, 2003;

7 KEAREY, P.; BROOKS, M.; HILL, I. **An Introduction to Geophysical Exploration**. 3° ed. Oxford: Blackwell Science, 2002;

8 HAMMER, S. **Density determinations by underground gravity measurements**. Annual Meeting of Society of Exploration Geophysicists. Chicago, 1950;

9 PARASNIS, D. S. **Principles of Applied Geophysics**. 5° ed. Londres: Springer, 1996;

10 BITENCOURT-FERREIRA, G.; SILVA, A. D.; AZEVEDO JR., W. F. Application of Machine Learning Techniques to Predict Binding Affinity for Drug Targets. A Study of Cyclin-Dependent Kinase 2. **Current Medicinal Chemistry**, Sharjah.v. 26, n. 0, 1-11, 2019;

11 COELHO, L. P.; RICHERT, W. **Building Machine Learning Systems with Python**. 2° ed. Birmingham: Packt Publishing Ltd., 2015;

12 CASTELLARO, S.; MULARGIA, F.; KAGAN, Y. Y. Regression problems for magnitudes. **Geophysical Journal International**, Oxford, v. 165, n.3, 913-930, junho de 2006.

13 ALMEIDA, P.; CARRASQUILLA, A. **Integrating Geological Attributes with a multiple Linear Regression of Geophysical Well Logs to Estimate the Permeability of Carbonate Reservoirs in Campos Basin, Southeastern Brazil**. Barcelona: AAPG/SEG International Conference and Exhibition. Abril de 2017.

14 DELL'AVERSANA, P.; CIURLO, B.; COLOMBO, S. **Integrated Geophysics and Machine Learning for Risk Mitigation in Exploration Geosciences**. Copenhaga: 80th EAGE Conference & Exhibition. Junho de 2018;

15 KURODA, M. C. **Técnicas de aprendizagem de máquina bio-inspiradas aplicadas ao estudo de rochas reservatório**. Campinas, 2016;

16 BOUGHER, B. B. **Machine learning applications to geophysical data analysis.** Vancouver. Agosto de 2016;

17 QADROUH, A. N.; CARCIONE, J. M.; ALAJMI, M.; ALYOUSIF, M. M. A tutorial on machine learning with geophysical applications. **Bollettino di Geofisica Teorica ed Applicata**. Trieste, v. 60, n. 3, 375-402. 2019;

18 THOMSEM, R.; CHRISTENSEN, M. H. MolDock: a new technique for high-accuracy molecular docking. **J. Med. Chem**., v. 49(11), 3315-3321, 2006;