

UNIVERSIDADE FEDERAL FLUMINENSE

DIOGO CEDDIA PORTO SILVA

**APLICAÇÕES DO APRENDIZADO DE MÁQUINA  
PARA O APERFEIÇOAMENTO DE TÉCNICAS DE  
MAPEAMENTO SUBMARINO**

Niterói – Rio de Janeiro

2022

DIOGO CEDDIA PORTO SILVA

**APLICAÇÕES DO APRENDIZADO DE MÁQUINA  
PARA O APERFEIÇOAMENTO DE TÉCNICAS DE  
MAPEAMENTO SUBMARINO**

Dissertação apresentada ao curso de Pós-graduação em Dinâmica dos Oceanos e da Terra, na Universidade Federal Fluminense, como requisito para obtenção do título de Mestre em Dinâmica dos Oceanos e da Terra.

Niterói – Rio de Janeiro

2022



## DIOGO CEDDIA PORTO SILVA

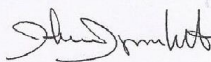
### Aplicações do Aprendizado de Máquina para o Aperfeiçoamento de Técnicas de Mapeamento do Fundo Marinho

Dissertação apresentada ao Programa de Pós-Graduação em Dinâmica dos Oceanos e Terra, da Universidade Federal Fluminense, como requisito parcial para obtenção do grau de Mestre.  
Área de Concentração: Hidrografia

Aprovado em 07/03/2022

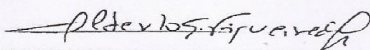
#### BANCA EXAMINADORA

Assinado de forma digital por  
ARTHUR AYRES NETO  
aayres@id.uff.br:78205026734  
Dados: 2022.03.08 16:08:38  
-03'00'



---

Prof. Arthur Ayres Neto, Dr (Orientador)  
Universidade Federal Fluminense



---

Prof. Alberto Figueiredo, Dr.  
Universidade Federal Fluminense



---

Prof. Esteban Walter Gonzalez Clua, Dr.  
Universidade Federal Fluminense

---

Prof. Rodrigo Bijani, Dr.  
Universidade Federal Fluminense



---

Prof. Luciano Fonseca, Dr.  
Universidade Nacional de Brasília

Ficha catalográfica automática - SDC/BIG  
Gerada com informações fornecidas pelo autor

P839a Porto silva, Diogo Ceddia  
APLICAÇÕES DO APRENDIZADO DE MÁQUINA PARA O APERFEIÇOAMENTO  
DE TÉCNICAS DE MAPEAMENTO SUBMARINO / Diogo Ceddia Porto silva  
; Arthur Ayres Neto, orientador. Niterói, 2022.  
99 f. : il.

Dissertação (mestrado)-Universidade Federal Fluminense,  
Niterói, 2022.

DOI: <http://dx.doi.org/10.22409/PPGDOT.2022.m.10465541747>

1. Machine Learning. 2. Marine Geophysics. 3. High-  
Resolution Seismic. 4. Bathymetry. 5. Produção intelectual.  
I. Ayres Neto, Arthur, orientador. II. Universidade Federal  
Fluminense. Instituto de Geociências. III. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

*Dedico ao meu Pai, pelos sólidos exemplos que mostram valor pelo tempo.*

*À minha Mãe, por desde o princípio ser a “melhor mãe do mundo”.*

*Ao meu Avô, por sobreviver e vencer o mais difícil dos tempos e nunca desistir.*

*À minha Avó, pela força e resiliência ímpares que sempre acompanharam um sorriso.*

## **AGRADECIMENTOS**

O primeiro e principal agradecimento vai para a família, a coisa mais importante que temos na vida.

Agradeço em especial ao meu orientador, Arthur Ayres Neto, por todo acompanhamento, conselhos e paciência comigo desde a graduação.

Ao Reinaldo Mozart, por estar sempre à disposição e jamais medir esforços para ajudar no que for possível.

Ao meu grande amigo Felipe Miranda, que me inseriu no mundo da inteligência artificial.

Ao Ministério de Ciência e Tecnologia e Inovação (MCTI)/Conselho Nacional de Pesquisa (CNPq)/Fundo Nacional para o Desenvolvimento da Ciência e Tecnologia (FNDCT), Programa Antártico Brasileiro – PROANTAR Call N° 64/2013, por financiar a pesquisa que forneceu dados para essa dissertação.

Ao Luciano Fonseca, por fornecer dados de backscatter processados e filtrados da Baía do Almirantado.

À todas as pessoas que me auxiliaram em meu percurso.

## **RESUMO**

A presente dissertação apresenta diversas aplicações do aprendizado de máquina para aperfeiçoar técnicas de mapeamento do fundo marinho. Inicialmente, é feita uma revisão das metodologias tradicionais de mapeamento do fundo marinho na Baía Rei George, Ilhas Shetland do Sul, Antártida. Nesta etapa foi mostrado que diferentes informações relativas ao leito marinho (profundidade, backscatter, declividade do fundo, etc) são predominantemente independentes entre si. Tal informação foi base para aplicação do modelo de aprendizado de máquina XGBoost, utilizado para: extrapolar classificação de ecocaráteres, inicialmente disposta em linhas, para uma superfície; demonstrar que o modelo pode oferecer resultados preliminares sobre a ecocaracterização da área, havendo somente interpretação parcial advinda do especialista; e, por último, demonstrar que o modelo é uma alternativa viável para realizar predição em secções de dado difíceis de classificar, seja devido à regiões transicionais de eco-tipos ou limitações da própria aquisição do dado, que dificultam a interpretação. Ao fim, foram evidenciadas aplicações de aprendizado de máquina de baixo custo computacional, aplicabilidade doméstica e acurácias balanceadas de até 99% na aplicação do modelo, demonstrando o potencial imenso que essa ferramenta pode proporcionar.

**Palavras-chave: aprendizado de máquina, geofísica marinha, sísmica de alta resolução, batimetria.**



## **ABSTRACT**

This research project presents several machine-learning applications capable of improving and refining conventional seabed mapping techniques. At first, a review of common mapping methodologies was conducted in King George Bay, South Shetland Islands, Antarctica. It was noticed that variables, which describe the seabed (such as backscatter, depth, slope, etc.), are mostly interdependent. This information was used as baseline for a machine learning approach, using XGBoost model to: (i) extrapolate echo-types – initially represented as lines – along a surface, based on statistic correlations relative to bathymetry, backscatter, slope, aspect and more; (ii) to demonstrate that the model can offer a preliminary result regarding the echo-types distribution along the seismic data, considering a partial interpretation from the specialist; and (iii) to show the model could be used to predict sections of the data which the specialist can not interpret with confidence (transitional echo-characters, data acquisition limitations, anisotropy, etc.). At the end, it was showed that XGBoost is a powerful algorithm to improve SBP interpretation and seafloor mapping, achieving up to 99% of balanced accuracy in several tests, with low computational cost and feasible to be implemented in a domestic environment.

**Keywords: machine learning, marine geophysics, high-resolution seismic, bathymetry.**

## SUMÁRIO

<b>Resumo</b> .....	7
<b>Abstract</b> .....	8
<b>Lista de Figuras</b> .....	10
<b>Lista de Tabelas</b> .....	13
<b>1. Introdução</b> .....	14
<b>2. Definição dos problemas</b> .....	15
<b>3. Estrutura</b> .....	16
<b>1º Artigo: <i>A Comparison of Different Acoustic Methods for Sedimentary Classification of King George Bay, Antarctica</i></b> .....	18
<b>4. Condicionamento dos dados para artigos subsequentes</b> .....	47
<b>2º Artigo: <i>Machine Learning Modeling Applied for Seabed Echo-Characterization in King George Bay, Antarctica</i></b> .....	55
<b>3º Artigo: <i>XGBoost as a Tool to Improve High Resolution Single Channel Seismic Interpretation</i></b> .....	85
<b>5. Conclusão</b> .....	96
<b>6. Códigos</b> .....	97
<b>7. Referências bibliográficas</b> .....	98

## Lista de Figuras

### **1º Artigo:** *A Comparison of Different Acoustic Methods for Sedimentary Classification of King George Bay, Antarctica.*

Figure 1 – Study area, located in King George Bay. Red lines indicate navigation lines of both the MBES and SBP surveys.....	24
Figure 2 – Echo-characters found in King George Bay.....	27
Figure 3 – Bathymetry map overlaid with the echo-character distribution.....	29
Figure 4: Backscatter intensity distribution for each echo-character.....	30
Figure 5: Gradient map overlaid with the echo-character distribution.....	31
Figure 6 – MBES backscatter intensity map overlaid with the echo-character distribution.....	32
Figure 7 – Seismic amplitude map overlaid with the echo-character distribution.....	33
Figure 8 – Backscatter intensity map overlaid with the bathymetric map.....	35
Figure 9 – Backscatter intensity map overlaid with the map of seismic amplitudes.....	36
Figure 10 - Bathymetry map overlaid with the map of seismic amplitudes.....	38
Figure 11 – Relationship between seafloor slopes and seismic amplitudes.....	38
Figura 1: secção do Dataset1, relativo à Baía Rei George.....	47
Figura 2: secção do Dataset2, relativo à Baía Rei George.....	47
Figura 3: secção do Dataset1, relativo à Baía do Almirantado.....	48
Figura 4: secção do Dataset2, relativo à Baía do Almirantado.....	48
Figura 5: esquema mostrando como os Datasets 1 do Almirantado e Rei George foram construídos, segundo critérios espaciais.....	49

Figura 6: esquema mostrando como os Datasets 2 do Almirantado e Rei George foram construídos, segundo critérios espaciais.....	50
Figura 7: da esquerda para a direita: em azul marinho, a desembocadura das baías e, em marrom, a delimitação continental/frente de geleira das baías Almirantado e Rei George, respectivamente. No interior das baías, o DTM (Digital Terrain Model) batimétrico de cada uma das áreas.....	53
Figura 8: Datasets 1 e 2 das Baías Rei George e Almirantado, após procedimentos de concatenação.....	54
<b>2º Artigo: <i>Machine Learning Modeling Applied for Seabed Echo-Characterization in King George Bay, Antarctica.</i></b>	
Figure 1: Area of study.....	60
Figure 2: workflow of the present methodology.....	61
Figure 3: relationship of each feature relative to the target (last rectangle).....	62
Figure 4: data conditioning stage, regarding grid spacing. Feature data represented in black dots, while target data in red dots.....	63
Figure 5: Example of Dataset1. Dataset2 has the same structure, but without the column ‘ECHO’ (no validation data).....	64
Figure 6: Spearman’s Correlation Rank matrix, computed to analyze the monotonic relationship between features.....	69
Figure 7: the PCA result, shown as how each PC explains each feature.....	70
Figure 8: different importance attribution methods result for the Dataset1.....	71
Figure 9: the summary plot of Echo 1.....	74
Figure 10: the summary plot of Echo 2.....	75
Figure 11: the summary plot of Echo 3.....	75

Figure 12: final echo-classification map, showing SBP validation data and uncertainty bias.....76

**3° Artigo: XGBoost as a Tool to Improve High Resolution Single Channel Seismic Interpretation.**

Figure 1: section of the Admiralty Bay dataset, showing its features and size (69905, 10).....89

Figure 2: relationship between BA gains with dataset size used in training for (a) methodology.....90

Figure 3: Relationship between BA gains with number of seismic lines used to train the model (black line). Since each line has a unique size, we plot as blue line the dataset percentage relative to the lines used.....91

Figure 4: seismic lines displayed above bathymetry surface, aiming spatial comprehension of methodology (b) results. Black, blue, and yellow lines are, respectively, Echo 1, 2 and 3, the seismic classes previously classified. In ‘a’, all data available; ‘b’ shows n = 3 lines, which predicted the other 48 with 85.74% of BA; ‘c’ shows n = 7 lines, which predicted the other 44 with 89.95% of BA; ‘d’ shows n = 14 lines, which predicted the other 37 with 95.31% of BA.....92

## Lista de Tabelas

**1º Artigo:** *A Comparison of Different Acoustic Methods for Sedimentary Classification of King George Bay, Antarctica.*

Table 1 – Spearman’s Rank Correlation between methodologies.....34

**2º Artigo:** *Machine Learning Modeling Applied for Seabed Echo-Characterization in King George Bay, Antarctica.*

Table 1 – accuracy and processing time comparison relative to all feature combinations.....73

**3º Artigo:** *XGBoost as a Tool to Improve High Resolution Single Channel Seismic Interpretation.*

Table 1: train and test size, in percentage (%) and absolute values (abs), needed to achieve common thresholds. The absolute values are the number of seismic traces needed.....91

Table 2: the BA and train size (%) achieved for the best combination of n seismic lines.....92

## 1. Introdução

As primeiras sondas, desenvolvidas para operar na investigação do fundo marinho, derivaram de sonares militares e começaram a ser utilizadas com finalidade hidrográfica a partir de meados do século XX (OHI, 2005). Desde então, houve um expressivo desenvolvimento tecnológico e aperfeiçoamento de técnicas e metodologias de investigação do assoalho oceânico. Em 1959, perfiladores sísmicos de sub fundo começaram a ser utilizados para estudar a geologia marinha (Heezen et al., 1959), e em 1977 foi criado um sistema de classificação de resposta acústica, utilizado como referência até hoje (Damuth & Hayes, 1977). Já em 1998, a Organização Hidrográfica Internacional (OHI), na 4ª edição da S-44, considerou que os equipamentos para medir a profundidade haviam chegado no estado da arte – alcançando precisões subdecimétricas com sondadores de feixe simples em profundidades rasas. Mais tarde, no século XXI e principalmente a partir de 2011, a produção científica começou a contemplar, progressivamente e de forma expressiva, algoritmos e aprendizado de máquina (Machine Learning (ML), em inglês) no contexto do mapeamento marinho (Duarte et al., 2020; Leon et al., 2020; Menandro & Bastos, 2020). Dessa forma, entende-se que o desenvolvimento de equipamentos e metodologias culminam na atualidade, com o ápice do desenvolvimento no âmbito computacional. Essa realidade propicia uma busca por novas soluções sob uma nova perspectiva, utilizando ferramentas avançadas de inteligência artificial que, não obstante, possuem implementação simplificada e viável em ambiente doméstico, respaldadas por uma comunidade ativa e dinâmica de desenvolvimento<sup>1</sup>.

---

<sup>1</sup> <https://stackoverflow.com/company> acesso em 23/11/2021.

## 2. Definição dos problemas

A dissertação aqui desenvolvida busca soluções para aprimoramento da interpretação sísmica e mapeamento do fundo marinho. Isso é feito explorando os seguintes problemas:

- (i) Os dados sísmicos monocanal de alta resolução são coletados linearmente, conforme a navegação da embarcação. Em contrapartida, a batimetria multifeixe – que segue a mesma rota de navegação – adquire dados em formato de leque, cobrindo uma área do leito marinho. Notavelmente, há uma diferença inerente significativa nos dados adquiridos, em que um representa linhas enquanto o outro representa áreas. Assim, torna-se difícil uma correlação plena entre esses tipos de dados, a qual objetive uma compreensão da disposição dos ecocaráteres em toda superfície batimétrica. Solução desenvolvida: extrapolação das linhas classificadas em ecocaráteres para toda área abrangida para a batimetria, segundo critérios estatísticos, utilizando modelo construído a partir do ML.
- (ii) Na interpretação sísmica de alta resolução, é comum que o especialista se depare com respostas sísmicas que não se podem classificar em ecocaráteres com tanta certeza. Isso decorre de ecocaráteres transicionais, problemas na aquisição, anisotropia, artefatos (objetos não geológicos-topográficos) e demais subjetividades que podem causar alguma variação na interpretação. Solução desenvolvida: utilizar ML para definir um modelo preditivo, especialmente nos trechos cuja interpretação seja duvidosa.
- (iii) Normalmente, a interpretação sísmica de alta resolução é realizada manualmente. Esse processo é demorado, principalmente em situações que há



um grande volume de dados a serem interpretados. Nesse contexto, é comum que haja a necessidade de um resultado preliminar para suporte ao especialista e à tomada de decisão. Solução desenvolvida: aplicação do ML supervisionado para classificar linhas sísmicas.

### 3. Estrutura

O objetivo primordial da presente dissertação é evidenciar as possibilidades que o aprendizado de máquina (machine learning) pode oferecer para a interpretação sísmica voltada ao mapeamento marinho. A dissertação é composta por três artigos, já submetidos e em análise de revisão.

A dissertação inicia-se com o **1º Artigo: *A Comparison of Different Acoustic Methods for Sedimentary Classification of King George Bay, Antarctica***, que consiste em uma descrição e análise dos tradicionais métodos geofísicos (batimetria, backscatter e sísmica de alta resolução) e outros produtos (declividade do fundo marinho e amplitude sísmica) na Baía Rei George, Antártida. O objetivo final é provar que, devido à diferença de operação e propriedades físicas que regem cada método geofísico, apesar de serem complementares, seus produtos possuem baixa correlação monotônica entre si.

Entre o primeiro artigo e os demais, um capítulo sobre condicionamento dos dados foi inserido. Isso justifica-se pela explicação da estratégia e critérios adotados para montar a matriz de dados necessária para as análises procedentes, assim como para servir como entrada em métodos de ML.

Os conhecimentos do 1º Artigo foram base para aplicação de um método de aprendizado de máquina supervisionado (XGBoost) na mesma localidade, quando foi

percebido que este alcançou elevadas acurácias balanceadas (BA). O resultado motivou o **2º Artigo: *Machine Learning Modeling Applied for Seabed Echo-Characterization in King George Bay, Antarctica***, cujo objetivo é a extrapolação da eco-caracterização sísmica para toda área abrangida pela batimetria, considerando que as linhas sísmicas se situam no interior da superfície batimétrica. Ao resultado da superfície de classes de ecos sísmicos, foi adicionado um fator de confiabilidade para a predição, associado à probabilidade da decisão que o modelo tomou ao atribuir determinada classe para cada ponto de teste. Aqui, é solucionado efetivamente o problema (i) do capítulo 2.

Seguindo a mesma linha, o **3º Artigo: *XGBoost as a Tool to Improve High Resolution Single Channel Seismic Interpretation*** teve como objetivo aplicação do método XGBoost na Baía do Almirantado, Ilhas Shetland do Sul, Antártida. Nesse momento, o foco foi de evidenciar que o modelo XGBoost consegue predizer, com alta BA, tanto pontos aleatórios no dado quanto linhas inteiras. Utilizando apenas 1% dos dados de treino foi possível predizer mais de 99% do dado total com  $BA > 90\%$ . De forma semelhante, menos de 10% das linhas foram capazes de predizer todas as outras com  $BA > 90\%$ . Foi concluído, então, que a abordagem com ML pode fornecer informações úteis e auxiliares para o especialista interpretar o fundo marinho, tais quais: (1) interpretar segmentos duvidosos; e (2) oferecer produto preliminar da interpretação sísmica da área, sugerindo para o especialista uma distribuição de eco caracteres nos dados ainda não interpretados. Aqui, são solucionados efetivamente os problemas (ii) e (iii) do capítulo 2.

# A Comparison of Different Acoustic Methods for Sedimentary Classification of King George Bay, Antarctica (\*).

Diogo Ceddia Porto Silva 1, Arthur Ayres Neto 2 and Rosemary Vieira 3.

Marine Geology Laboratory (LAGEMAR), Department of Geology and Geophysics, Federal Fluminense University, Rio de Janeiro, Brazil.

(\*). Submitted in *Marine Georesources & Geotechnology Journal*.

## ABSTRACT

Multibeam bathymetric and high-resolution seismic are acoustic remote-sensing techniques frequently employed to map marine sediments. Both techniques involve applying sonar theory and typically generate consistent and satisfactory results at a relatively low-cost relative to direct sampling. However, mainly due to their operational peculiarities (frequency range, beam width, etc.), each method generates some discrepancies regarding sedimentary distribution for the same area. Here, we compare three well-established acoustic techniques (MBES backscatter intensity, sub-bottom profiler echo-characterization, and seismic amplitude) applied to characterize seabed geology and assess the possible causes of the discrepancies observed among the methods. The study area is the glaciomarine environment of King George Bay, South Shetland Islands, Antarctica. We demonstrate that each method responds to specific characteristics of the seafloor, mainly because of divergent interactions between the acoustic signal and morphological features at different scales and sedimentological interfaces. Despite some consistencies, considerable mismatches in terms of classification approaches are

apparent. Nevertheless, we show how these divergent classification approaches can be interpreted together to limit ambiguities and enable more precise geological understandings.

**Keywords:** bathymetry; high-resolution seismic; echo-characterization; backscatter; glaciomarine environment; seismic amplitude.

## 1 INTRODUCTION

Here, we compare three acoustic remote-sensing techniques — namely backscatter intensity, seismic amplitude and echo-characterization — to assess inconsistencies between respective datasets for a given locality. We explore the causes of these discrepancies and describe the King George Bay, aiming for enhancing the reliability of seabed acoustic-remote sensing analyses to improve interpretation.

At first, we will briefly describe the glaciomarine environment of our study area and its basic sedimentation characteristics, as well the high-resolution seismic and the MBES capabilities for seafloor classification. Thereon, in our results we compare the occurrence of three echo-characters relative to depth, slope, backscatter and seismic amplitude. In this step, we used geophysics tools to understand the occurrence of echo-characters in the study area. The next step is the statistical analysis, which measured the correlation between depth, slope, backscatter, and seismic amplitude. At this point, we discuss how this relationship eventually correlate, showing overlaid layers aiming for better interpretation.

To inferring sedimentation characteristics of King George Bay based on geophysical methods, a quick understanding of sedimentation in this environment is necessary. Antarctica's continental margin, our study area, is characterized as a glaciomarine environment, where ice and glacial meltwater are the primordial agents of sediment erosion, transport and deposition. In this context, distance relative to the shore is a critical parameter for defining sediment depositional processes. For instance, these processes can be classified as "glaciogenic" when debris-laden glaciers mobilize the sediment directly, "proglacial" when deposition occurs directly in front of the glacier and its meltwater is the main transport agent for the sediment, or "glaciomarine" when the material is deposited in a marine setting through a combination of glacial- and marine-associated activity (Anderson et al. 1983; Anderson 1999; Assine & Vesely 2008).

Classical glaciogenic deposits are coarse and poorly sorted glacial sediments, transported by traction primarily at the glacial front, resulting in a chaotic echo-type without stratification. Proglacial environments are more dynamic, with glacial meltwater and summer rains acting as the main mechanisms for transporting a mixture of pelagic sediments and suspension plumes from meltwater, exhibiting subtle stratification and more pronounced grain sorting. As distance from the glacial front increases, the influence of marine pelagic sedimentation processes also intensifies. In such glaciomarine environments, deposits present clear stratification and sediment grain size becomes finer through increased grain sorting (Magrani & Ayres Neto 2014; Anderson 1999).

High-resolution seismic analysis (Sub-Bottom Profiler - SBP) is a geophysical approach widely used to map sub-surface structures and to echo-characterize the seafloor, encompassing the set of reflected echo characteristics arising from interactions between seismic signal and seafloor sediments (Damuth & Hayes 1977). Relative to Antarctica's glaciomarine environment, several researches applied the principles of echo-

classification. For example, Solli et al. (2007) used SBP profiles to identify glaciomarine deposits the eastern Antarctic margin. Their study concluded that seismostratigraphic mapping of deep-water depocentres and identification of paleo-channels along a glaciated continental margin may provide important information on long-term ice-sheet dynamics, particularly with regard to patterns of ice drainage represented by coastal ice streams; Magrani & Ayres Neto (2014) also used echo-characters to map sediment distribution in Admiralty Bay, Antarctica. They identified four different sedimentary domains and distinguished proglacial from glaciomarine sediments; Also in Antarctica, Ramalho (2016) noted that, despite sedimentary homogeneity within Port Foster Bay, Deception Island, Antarctica, three echo-types could still be characterized and postulated that transportation and sedimentation processes organized sediment particles differently. Accordingly, disparate sediment frameworks differentially affect acoustic responses and, consequently, seabed seismic characteristics; Yoon et al. (2004) applied High-resolution (3.5 kHz) seismic to reveal sedimentation patterns in the northern South Shetland continental margin and the South Scotia Sea, Antarctica. They discovered eight echo types describing the area.

Besides the echo-classification, seismic amplitude, extracted from the SBP data, is widely studied in seismic explorations. This approach has also been applied to seabed characterization, including in our study area. Several authors used this technique to measure physical properties and classified the sediment types. For example, Brandt et al. (2003) used seismic amplitude to predict the shear strength of marine sediments on the continental slope of the Gulf of Mexico. The authors were able to extrapolate shear strength across a large area with a limited number of sediment cores by inverting the seismic amplitude data. Ayres Neto et al. (2016) applied a similar technique to determine the total organic carbon (TOC) content of coastal sediments on the southeastern Brazilian

continental shelf. Their results indicated an average error of ~5% from using this approach. Goff et al. (2004) combined the reflection intensity of Chirp pulses with backscatter data to measure seabed properties such as grain size, attenuation and porosity. They concluded that vertical-incidence seismic reflection intensity exhibits a stronger correlation with *in situ* velocity measurements, indicating that such data may be more reliable than backscatter in terms of deriving sediment physical properties from acoustic remote-sensing data.

Furthermore, the Multi-Beam Echo Sounder (MBES) technique was applied in the study area for sediment mapping. Also, it was used as a tool to understand the interaction between sediment classification and geomorphology of the seabed. The MBES system is based on transmitting and receiving several acoustic pulses to measure depth with high precision. Data acquisition is achieved using a transducer fitted on a ship's hull, which emits a signal in a fan-shaped geometry of ~120° and that covers a broad expanse of the seafloor. Apart from providing a Digital Terrain Model (DTM) of the seafloor, MBES generates backscatter data (the intensity, in decibels, of the received energy from each depth measurement) (L-3 Communication SeaBeam Instrument, 2000; OHI, 2005; Magrani, 2014). Backscatter intensity is increasingly employed as a tool in seabed classification, including for habitat mapping. For example, Simons & Snellen (2009) used backscatter intensity to generate a sediment classification for an area in the North Sea. They observed that this method displayed a discriminatory performance comparable to that of physically taking geological samples in terms of separating all seafloor types known to occur in the study area; Lark et al. (2015) used swath bathymetry and backscatter data to map seabed sediment texture classes and reported that their geostatistical analysis enabled prediction of coarse and muddy sediments in relatively quiescent, localized deep-water environments; Naudts et al. (2008) integrated different

acoustic techniques to locate gas seeps in the Dnepr paleo-delta of the northwestern Black Sea. Their results showed that the observed backscatter patterns were the result of ongoing methane seepage and precipitation of methane-derived authigenic carbonates on the seafloor. Similarly, Leitão et al. (2016) used backscatter intensity to map the sediment distribution within Port Foster Bay, Deception Island, Antarctica, in the same environment as our study area, noting a very good correlation between backscatter intensity and the sediment type (silt) observed in geological cores taken from the area.

## **2 MATERIALS AND METHODS**

Our study area is located in the interior of King George Bay, King George Island, which is one of the South Shetland Islands in Antarctica (Figure 1). Sub-bottom and MBES data were simultaneously acquired during the XXXIII OPERANTAR Expedition of Summer 2014, from which echo-character, seismic amplitude, backscatter intensity and depth information was gathered. The survey comprised 70 km of seismic and MBES transects covering an area of approximately 10.8 km<sup>2</sup>. The equipment installed onto the hull of the Almirante Maximiano included a Kongsberg SBP-120 system, operating in the 2.5-6.5 kHz frequency range, and a Kongsberg EM-302 echosounder operating at 30 kHz.



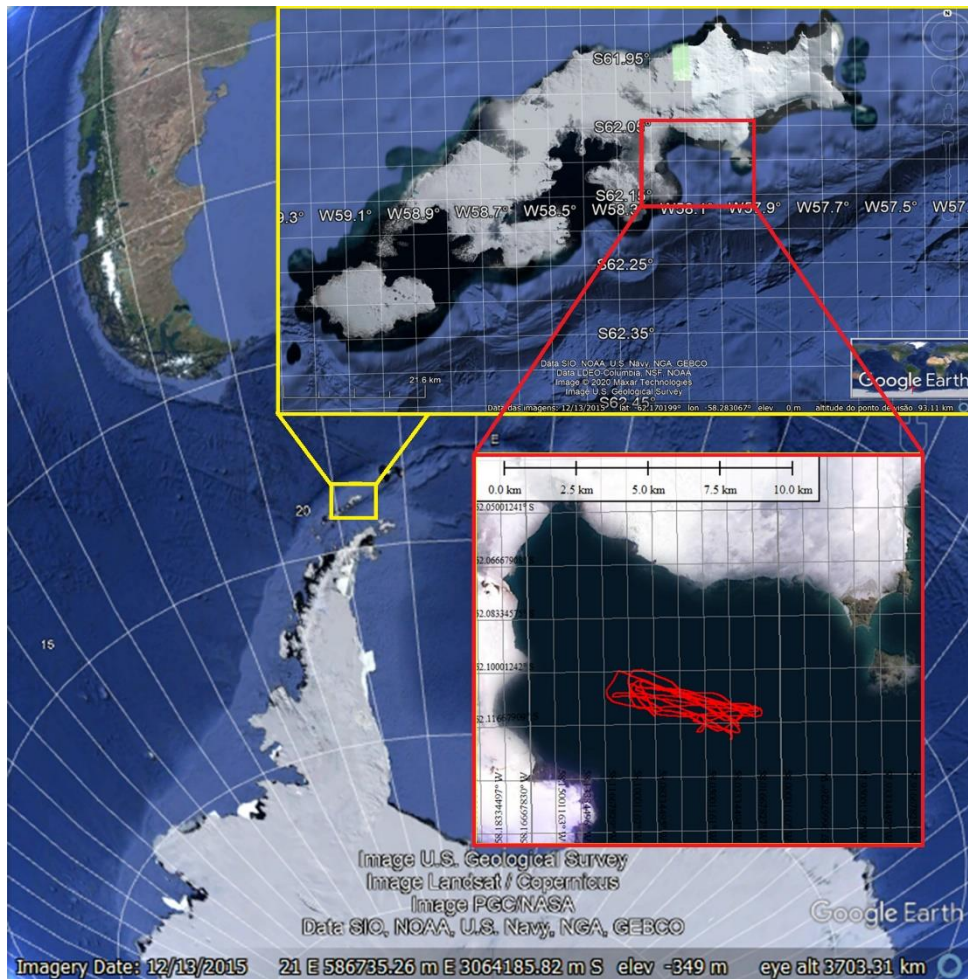


Figure 1 – Study area, located in King George Bay. Navigation lines of both the MBES and SBP surveys are indicated by red lines.

The sub-bottom data was preprocessed during acquisition, where the geometrical spreading was corrected and the frequency spectrum was limited from 2.5 to 6.5 kHz. The processing step used RadExPro 2019.3 software, upon applying the following filters: (1) “notch filter” to remove noisy frequencies from 5840 to 5860 Hz, probably due to ship’s noise or environmental issues; (2) “data envelope” using the Reflection Strength attribute; and (3) “top muting” to remove the water column. The objective of this processing workflow was to preserve the data amplitudes without subjecting the dataset to further modifications. Additionally, the visibility of the surface and subsurface reflectors may enhance with the above mentioned data processing. After this processing

phase, the bottom reflector was selected using a 1-millisecond window to isolate the seafloor amplitude value. The SBP echo-characters were interpreted according to criteria established in Damuth & Hayes (1977).

The MBES (Multi-Beam Echo Sounder) data processing was performed in CARIS HIPS and SIPS 11.2 software, from which DTM – created from CUBE algorithm, with cell size of 3 m – and backscatter – mosaic with the same cell size – information were exported. A manual cleaning procedure was conducted to remove spikes and artifacts. Tides were not corrected, because of the lack of tide data. It is necessary to consider that the region is very remote, and there is not tide monitoring. Moreover, the highest gap between bathymetric lines due to tide variation was lower than 3 meters. Therefore, in a region with depths ranging from 195 to 444 meters, tidal variation is of little importance for assessing the regional morphology of the seabed. The slope information, obtained from the DTM using the Global Mapper 20 software, had few isolated artifacts from 50 to 90 degrees. They also did not influenced the regional analysis neither the statistical, since its occurrence is aleatory along data and correspond to less than 0.5% of the data.

At first, Python was used to manage the exported data (seafloor amplitudes, echo-characters, depths, backscatter and seabed slope) for subsequent analyses. Amplitude values were filtered according to a simple moving average of 11 elements to reduce high-frequency noise following the methodology proposed by Ayres Neto et al. (2016). All data were spatially correlated by means of a Euclidian Distance algorithm using the SciPy library so that all datasets are aligned to the same coordinate points of the study area, enabling quantitative correlations among datasets.

Then, we applied a non-parametric statistical Spearman's Rank Correlation (in the SciPy library) to measure the degree of correlation among variables (depth, backscatter,

echo-characters, amplitude and slope). We adopted this method because none of the data distributions (apart from backscatter) were Gaussian-like.

Furthermore, we plotted maps of echo-characters, backscatter, slope and bathymetry in Global Mapper 20 software in order to check the information obtained using the Spearman's analysis. The Surfer 11 software was used to grid amplitude values according to the krigage method (Meng et al. 2013), which is a more consistent interpolation algorithm that generates fewer artifacts and is more appropriate for irregularly spaced data.

### **3 RESULTS AND DISCUSSIONS**

#### **3.1 Echo-characters**

We identified three distinct echo-characters in the study area from the SBP dataset (Figure 2), which we classified according to the criteria of Damuth & Hayes (1977). Echo 1 represents a well-defined seafloor, with strong and diffuse signal penetration and without subseafloor stratification. It covers approximately 45% of the area, mostly in the shallower northwestern zone (Figure 2). Echo 2 is also shown as a continuous and well-defined seafloor, but with continuous sub-bottom stratification. It corresponds predominantly to the channel talwegs and lies close to the bay's outfall. Echo 2 covers 19% of the study area (Figure 2). Echo 3 is characterized by a weak and somewhat discontinuous seafloor, with sparse erratic stratification and diffuse signal penetration, covering approximately 36% of the study area (Figure 2).

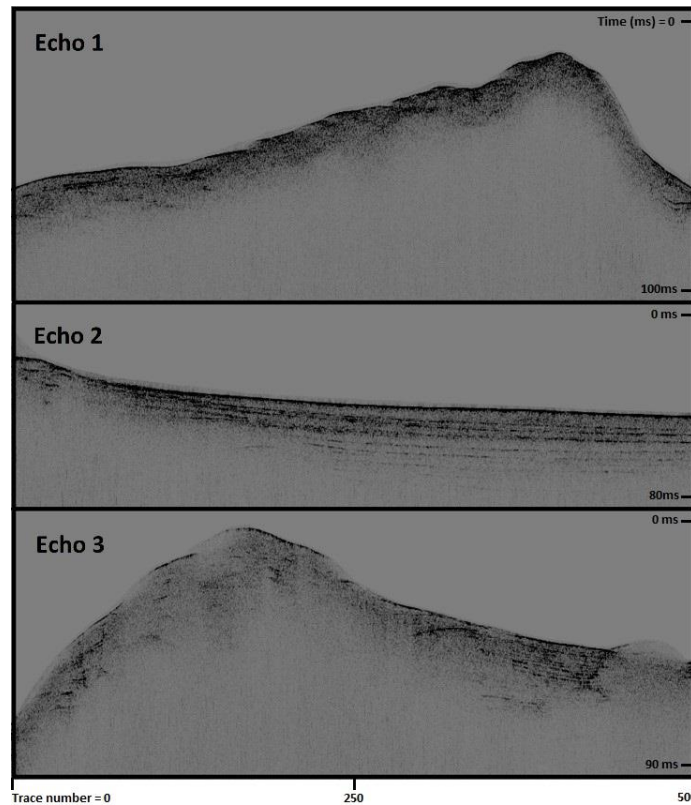


Figure 2 – Echo-characters found in King George Bay.

Magrani & Ayres Neto (2014) previously conducted research on Admiralty Bay, an area west of King George’s Bay with a similar sedimentary environment, and defined three echo-characters (Echos 1, 2 and 4), which exhibit the same characteristics, respectively, as our Echos 3, 2 and 1. Echo 1 in Magrani & Ayres Neto (2014) was mainly associated with shallow areas of Admiralty Bay, i.e., close to shore and incorporating the area directly behind the morainic banks. Geological samples demonstrated that this zone comprises a mixture of coarse material transported by glaciers or their meltwater within a finer matrix of high sand content (between 32 and 55%). This would correspond to our Echo 3. Yoon et al. (2004) also defined a similar echo-character (IA), which was linked to semiconsolidated sediments that had been reworked by bottom currents.

Echo 2 of Magrani & Ayres Neto (2014) and Echo IIA of Yoon et al. (2004) would both appear to correspond to our Echo 2. These echo-characters are characterized by a

well-defined seafloor, with sub-parallel sub-bottom reflectors, and represent sediments deposited by glacial meltwater plumes and hemipelagic sediments. Geological samples from the zone of Echo 2 in Magrani & Ayres Neto (2014) revealed essentially muddy sediments with a sand content of <10%. That echo-character covered a central deep zone (450-550 m) of the Admiralty Bay fjord, i.e., the same depth range observed for Echo 2 in King George Bay.

Our Echo 1, which corresponds to Echos 4 and IIB of Magrani & Ayres Neto (2014) and Yoon et al. (2004) respectively, reflects a well-defined seafloor, with strong and diffuse signal penetration and lacking sub-bottom reflectors. In both King George Bay and Admiralty Bay, these echo-characters cover shallower waters (200-250 m). According to Yoon et al. (2004), the respective echo-character in that study (IIB) represents till deposits on the continental shelf and upper slope.

Ramalho (2016) conducted an echo-characterization of the seafloor of Port Foster Bay, Deception Island, Antarctica, and distinguished three echo-characters. The Echo 2 of that study matches Echo 2 for King George Bay reported herein, and his Echo 3 corresponds to our Echo 1; in both studies, the Damuth & Hayes (1977) criteria were applied. Leitão et al. (2016) reported that the whole of Port Foster Bay is dominated by silty sediments, with backscatter intensity ranging from -22 to -26 dB. They speculated that depositional processes would have influenced how sediment particles were organized within the layers, affecting the acoustic response of SBP signal but without backscatter differentiation. In contrast, in our study, we observed differentiation between echo-characters relative to backscatter intensities (Figure 3). The Gaussian distributions relative to backscatter intensities achieved, respectively for Echo 1, 2 and 3, means of -17.18, -18.87, -19.96, standard deviation of 1.69, 1.80, 1.68, and a Coefficient of Determination ( $R^2$ ) of 0.9973, 0.9970 and 0.9723 for the best Gaussian fit. Therefore,

proving that for each echo-character there is a distinct Gaussian representing the backscatter intensity distribution.

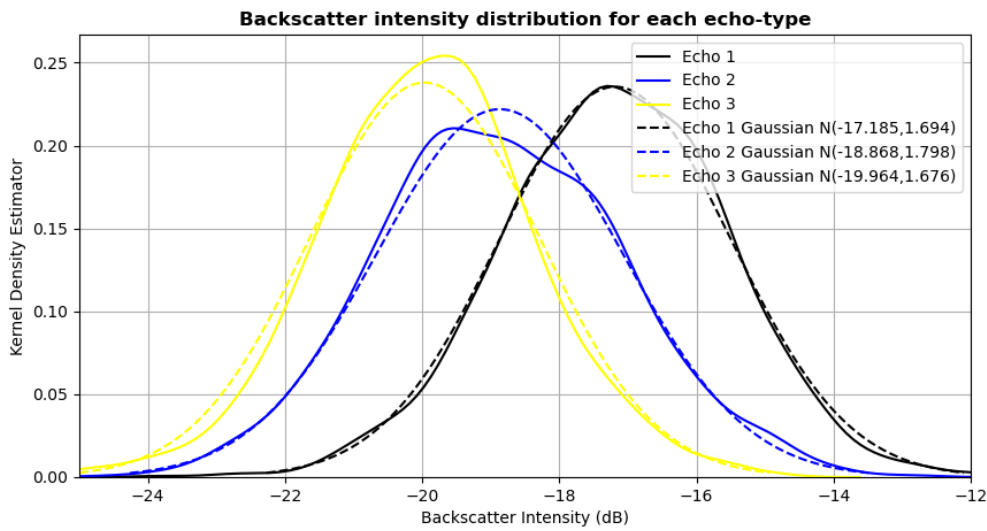


Figure 3 – Backscatter intensity distribution for each echo-character.

Moreover, it is important to reiterate that whereas King George Bay is essentially a glaciomarine environment, the sedimentation characteristics of Port Foster Bay are greatly affected by volcanic processes.

Despite the absence of validating geological samples in our study, a comparison of our echo-characters with those described in other studies for similar areas is possible. It allowed us to infer that our Echo 1 and Echo 2 correspond to coarse and poorly sorted glacial sediments (glaciogenic influenced) to fine well-sorted glaciomarine sediments (proglacial influenced), respectively, with our Echo 3 more reflecting proglacial sedimentation with some influence from glaciomarine processes.

### 3.2 Relationship between echo-characters and bathymetry/slope

In Figure 4, we have overlain the bathymetric map with the distribution of echo-characters to establish if there is any link between bathymetry and the occurrence of different echo-characters. Clearly, Echos 1 and 3 dominate the shallower zones, so they are probably more influenced by the large-grained and poorly-sorted glacial and proglacial sediments, respectively (Figure 4). Echo 2 mainly corresponds to deeper and flatter zones, i.e., within the channel talwegs and closer to the bay's outfall, and probably is more reflective of glaciomarine processes. An overlay of slope data with echo-character distribution (Figure 5) more clearly illustrates the relationship between Echo 2 and the channel talwegs.

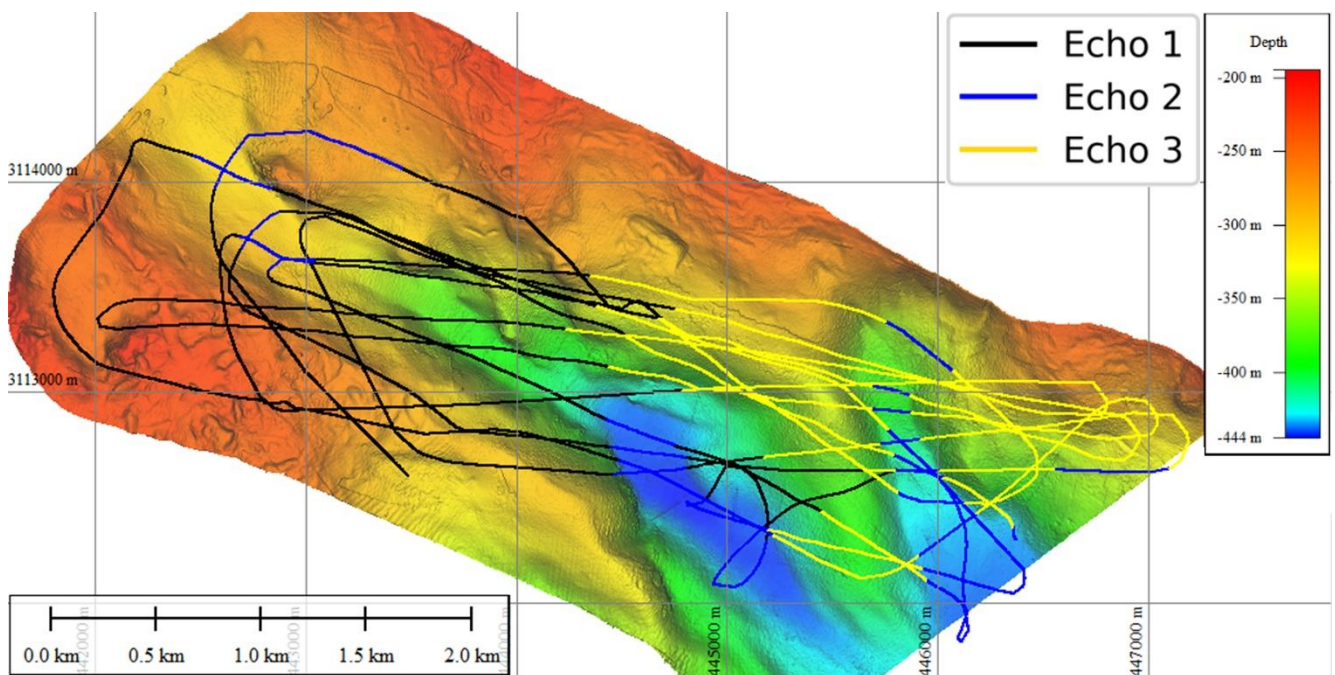


Figure 4 – Bathymetry map overlaid with the echo-character distribution.

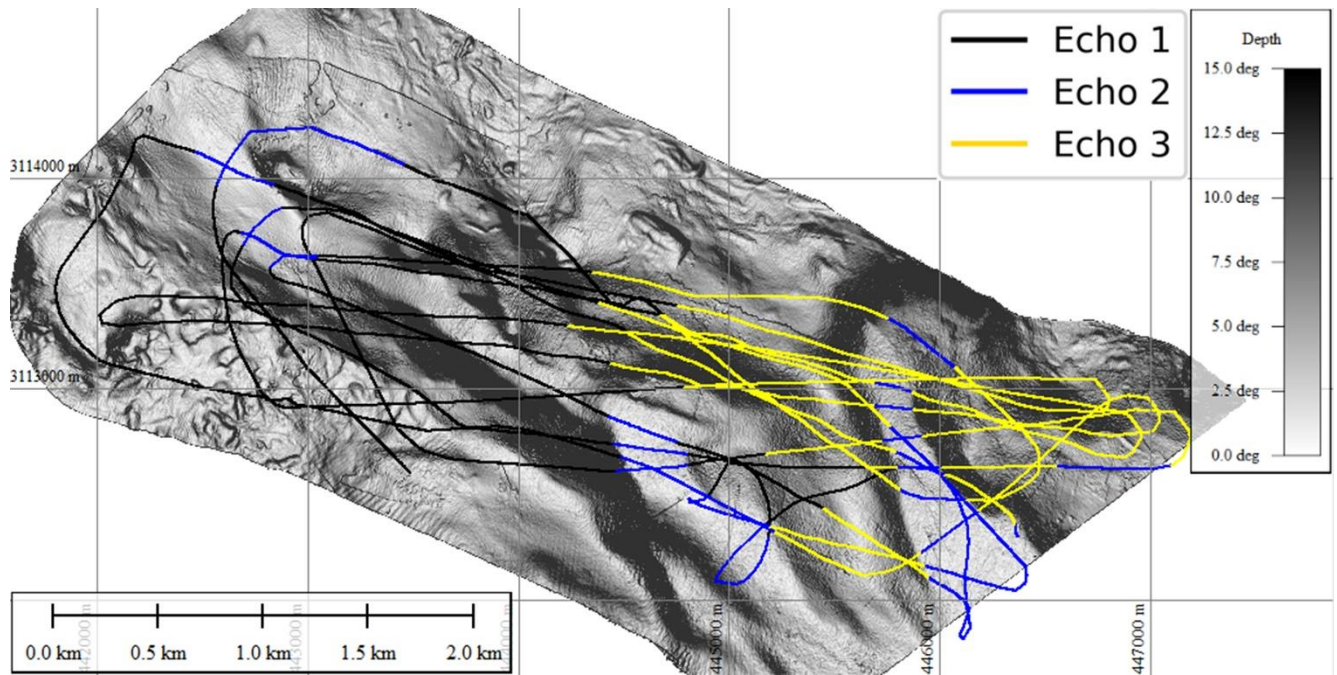


Figure 5: Gradient map overlaid with the echo-character distribution.

### 3.3 Relationship between echo-characters and backscatter

We overlaid the MBES backscatter intensity map on the echo-character distribution (Figure 6) for cross-referencing, which reveals an apparent relationship between Echo 3 and the -25 to -20 dB intensity range and/or between Echo 2 and the -20 to -15 dB range. Moreover, it is expected for the backscatter intensity to lower in bay's outfall direction (SE), as the glaciomarine influence increase.



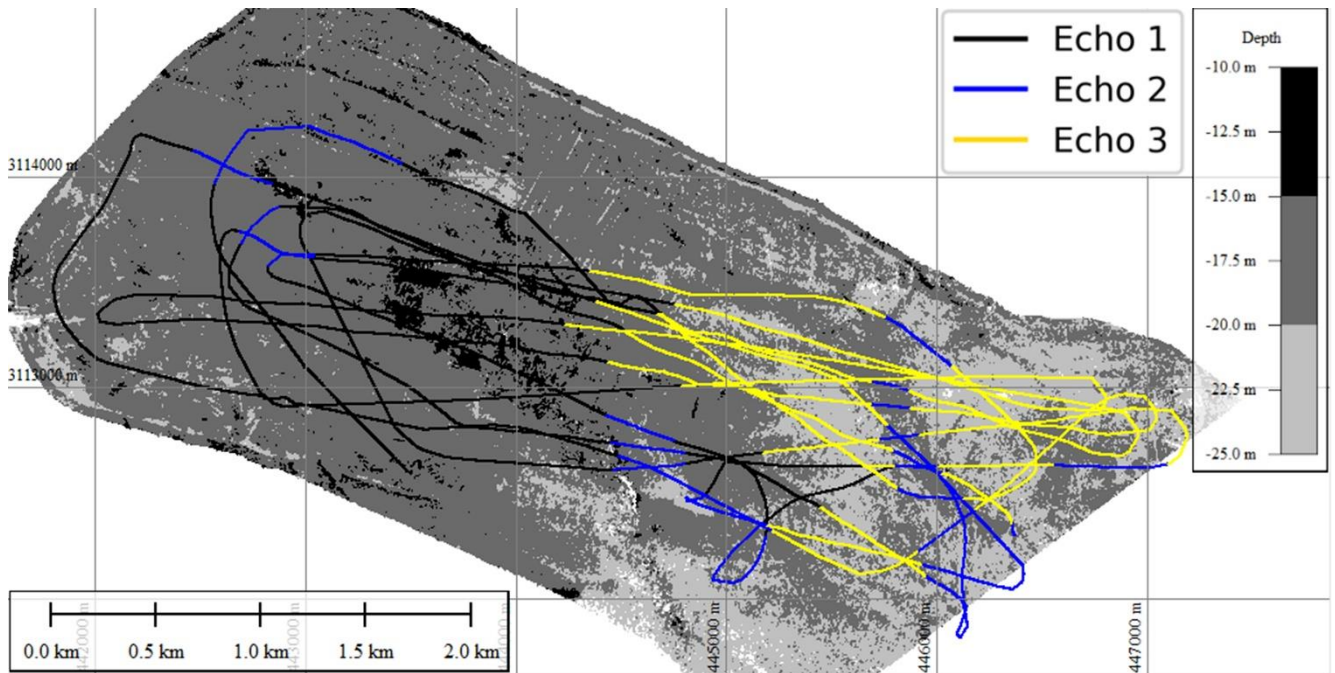


Figure 6 – MBES backscatter intensity map overlaid with the echo-character distribution.

### 3.4 Relationship between echo-characters and seismic amplitude

We also overlaid the seismic amplitudes map on the echo-characters distribution (Figure 7) to evaluate any relationship between these two variables. Overall, all echo-characters exist across the amplitude spectrum. Notably, Echo 2 is the only one that is strongly linked to areas of high seismic amplitude. This outcome contradicts well-established relationships between this kind of echo-character and seismic amplitude (Damuth & Hayes 1977; Solli et al. 2007; Mendoza et al. 2014; Hong & Shen 2020), with the characteristics of Echo 2 typically associated with sediments mainly composed of fine particles (clay and silt) with high water content and low P-wave velocity (Jackson and Richardson 2007). Such sediments are expected to have low impedance contrast with seawater and, consequently, low seismic amplitude values. Magrani & Ayres Neto (2014) observed a link between Echo 2 and median seismic amplitude values along the continental margin of the South Shetland Island.

However, the explanation to this phenomenon is that Echo 2 only occurs inside the thalwegs, where slope and grazing angle of the SBP are near zero. Importantly, the low amplitude values observed in areas of high gradient likely reflect seismic grazing angle. High-resolution seismic equipment, such as SBP, is assumed to operate in a “zero-offset” geometry. Low grazing angles cause energy to be reflected away from the acoustic receiver. It is important to note that not only seabed slope contributes to seismic grazing angle, with roll and pitch oscillations of the ship also playing a role. In our case, the motion reference unit (MRU) of the EM-302 MBES system was connected to the SBP120 system to compensate the seismic data for the ship’s attitude. However, this set-up operates more as a “swell-filter” and has only limited utility to correct seismic grazing angle and amplitude measurement. Moreover, roll and pitch did not reach considerable magnitudes, and were below  $3^\circ$ . Accordingly, we consider our data interpretations to be essentially topographic-dependent. Therefore, the slope controls the amplitude values of the study area.

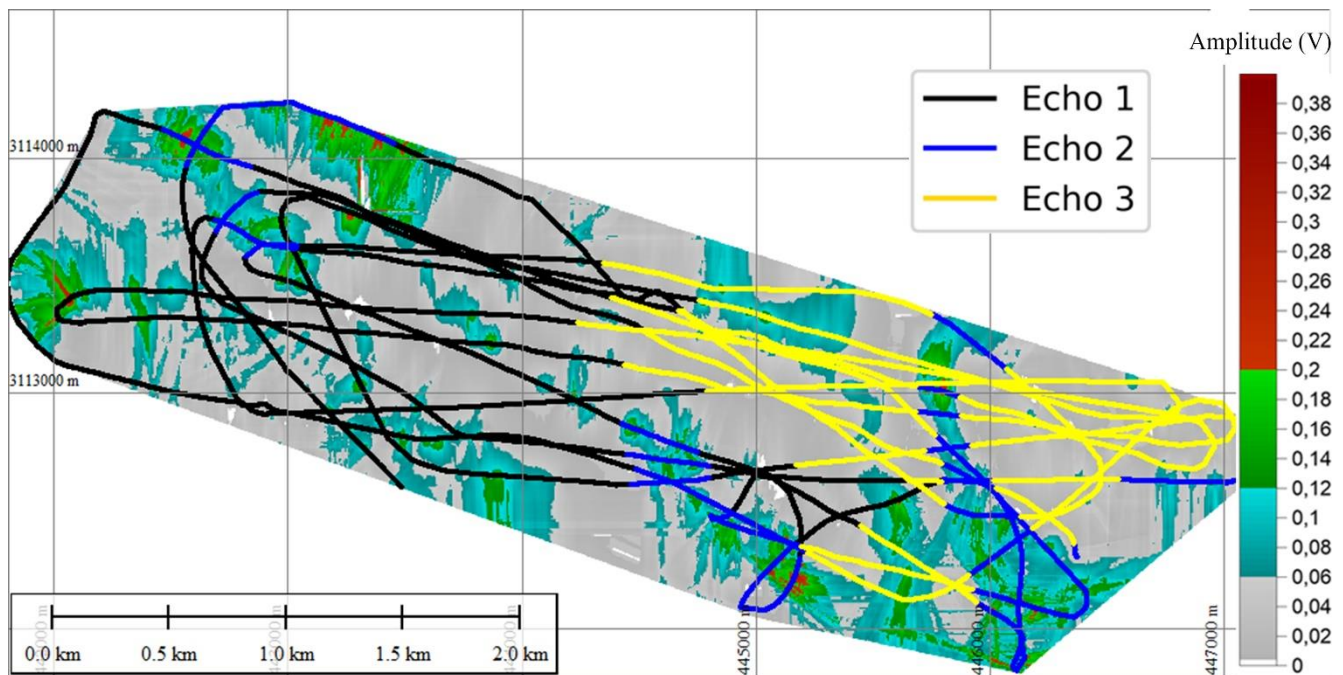


Figure 7 – Seismic amplitude map overlaid with the echo-character distribution.

### 3.5 Spearman’s Rank Correlation

Initially, we performed a Spearman’s Rank correlation analysis. It measures the quality of a monotonic relationship between two independent variables , with values ranging between -1.0 and 1.0 (these extreme values represent perfect negative or positive correlation, respectively). Typically, correlations of between ( $\pm$ ) 0.6 and 1.0 are considered strong, between 0.3 and 0.6 they are "moderate", and between 0.0 and 0.3 they are "weak" (Anderson and Finn 1997; Zhang et al., 2021).

Table 1 – Spearman’s Rank Correlation between methodologies.

	<b>Spearman’s Rank Correlation (<math>\rho</math>)</b>				
	<b>Backscatter <i>versus</i> Depth</b>	<b>Backscatter <i>versus</i> Slope</b>	<b>Backscatter <i>versus</i> Amplitude</b>	<b>Amplitude <i>versus</i> Depth</b>	<b>Amplitude <i>versus</i> Slope</b>
<b>All data</b>	0.268	0.010	-0.019	0.046	-0.478
<b>Echo 1</b>	-0.109	0.127	-0.145	0.196	-0.393
<b>Echo 2</b>	0.446	0.110	0.108	-0.028	-0.271
<b>Echo 3</b>	-0.138	0.070	-0.068	0.215	-0.348

### 3.6 Relationship between backscatter intensity and bathymetry/slope

Upon comparing the backscatter intensity and bathymetry maps (Figure 8), we observed that greater backscatter signal is related to the flanks of the channels and that some bathymetric highs represent harder glacial material. Backscatter intensity largely ranges from -20 to -15 dB within the channels, whereas outside them it mostly ranges

from -25 to -20 dB. The higher backscatter intensities ( $> -18$  dB) are concentrated in depths ranging from -200 m to -325 m, with deeper zones (i.e., below -325 m) having low backscatter intensities. The Spearman's Rank Correlation also shows that backscatter and depth are weak correlated among Echoes 1 and 3 ( $\rho = -0.109$  and  $\rho = -0.138$ , respectively), indicating that the shallower, the greater the backscatter intensity is in mostly coarse and poorly sorted glacial sediments. Otherwise, Echo 2 shows a moderate positive correlation, suggesting that the deeper, the greater the glaciomarine influence is and, therefore, the mostly glaciomarine sediments are more well selected, reflecting more MBES backscatter energy. Moreover, backscatter intensity is not correlated with slope ( $\rho = 0.010$ ), due to MBES corrections.

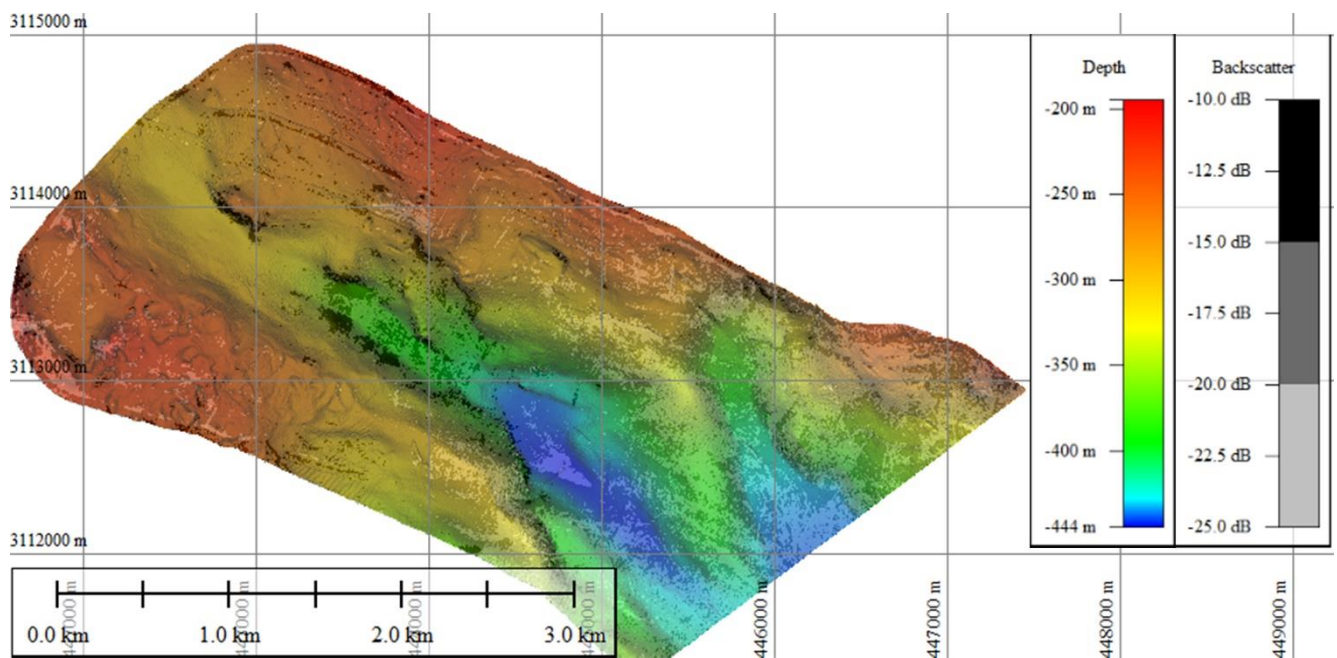


Figure 8 – Backscatter intensity map overlaid with the bathymetric map.

### 3.7 Relationship between backscatter intensities and seismic amplitude

We anticipated that high backscatter intensity values would be correlated with high amplitude values, since both of these parameters are closely related to the acoustic impedance contrast between the seafloor and the overlying water column. However, a

comparison of backscatter intensity and the seismic amplitude data did not reveal any clear visual correlation (Figure 9).

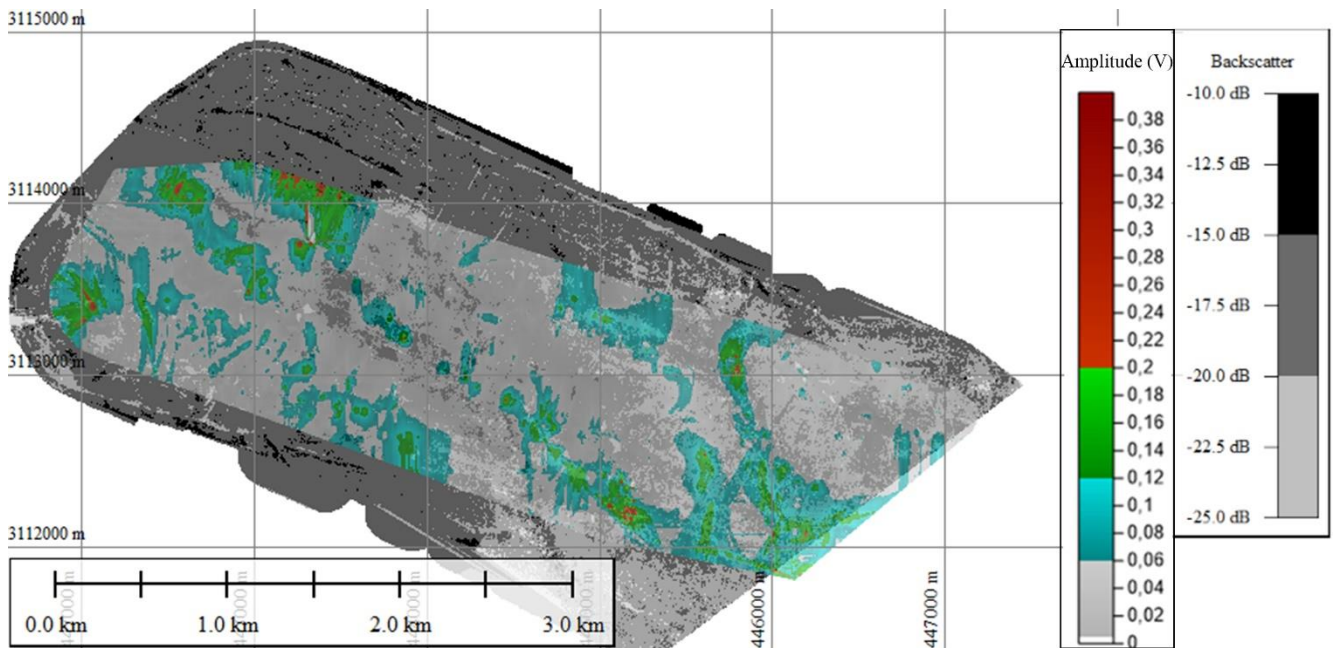


Figure 9 – Backscatter intensity map overlaid with the map of seismic amplitudes.

We also observed approximately none correspondence between backscatter intensity and seismic amplitude in Spearman's Rank correlation ( $\rho = -0.019$ ). However, both properties are primarily dependent on the acoustic impedance contrast between the seafloor and the overlying seawater. Magrani & Ayres Neto (2014) analyzed the acoustic impedance of Admiralty Bay sediments and reported that sediments showing medium-high impedance values (around  $3000 \text{ N.s.m}^{-3}$ ) were mainly associated with glacial/proglacial sediments, whereas glaciomarine sediments presented low impedance values ( $< 2600 \text{ N.s.m}^{-3}$ ). It is important to note that the acoustic equipment we used to obtain our datasets operate at different frequencies; the SPB120 system emits a 2.5-6.5 kHz Chirp Pulse, whereas the EM-302 system has an operational frequency of 30 kHz. Lower frequency signals can better penetrate the seafloor. Fonseca et al. (2017) demonstrated that 30-kHz signal can penetrate up to 30 cm of glaciomarine sediments

and that backscatter intensity readout is susceptible to the volumetric characteristics of the seabed rather than simply water/sediment impedance contrast. In contrast, SBP signals can penetrate several meters below the seabed. As they have much larger wavelengths, small sediment heterogeneities are not detected by SBP signal, so its interaction with the sub-seabed is restricted to larger features.

### **3.8 Relationship between seismic amplitude and bathymetry/slope**

To analyze if the spatial distribution of amplitude values can be linked to topographic features, we overlaid the seismic amplitude map with our bathymetry (Figure 10) and gradient (Figure 11) maps. These overlays demonstrate a link between high amplitude values and shallow areas of smooth gradients, whereas lower amplitudes are associated with higher gradients. This outcome can be explained by the grazing angle of the acoustic footprint being very acute relative to a normal vector, thereby weakening the returning seismic energy (Biffard, 2010). Moreover, amplitudes greater than 0.06 correspond to the bottoms of the channels where the seabed is very flat. Therefore, the amplitude values on the study area are controlled mostly by the slope, because of the near zero grazing angle of the SBP.

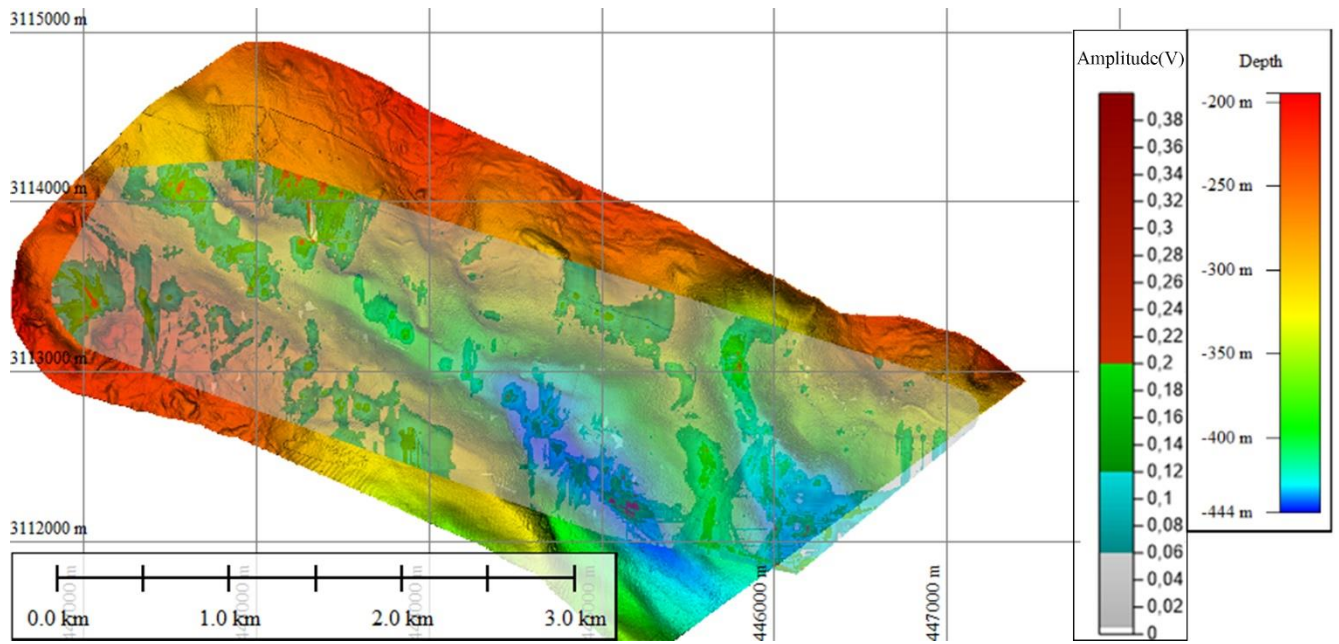


Figure 10 - Bathymetry map overlaid with the map of seismic amplitudes.

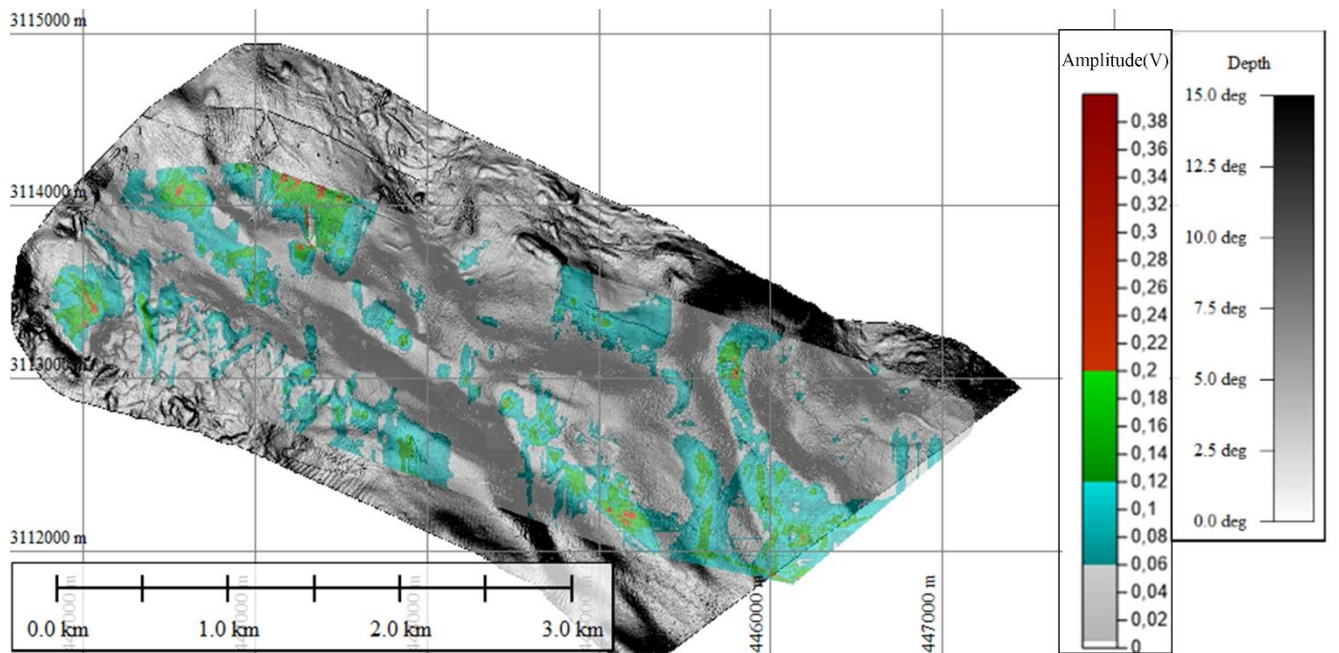


Figure 11 – Relationship between seafloor slopes and seismic amplitudes

## 4 CONCLUSIONS

Our findings reveal that the acoustic methods we considered for seafloor classification (multibeam backscatter intensity, seismic amplitude and echo-characterization) are mostly not correlated with each other and provide distinct information according to the physical principles of how they are derived. Nevertheless, through comparative and detailed analyses of these methods, ambiguities can be reduced or removed, thereby improving data interpretations.

We found that King George Bay has a complex morphology, with three echo-types strongly similar to those found by Leitão et al. (2016), Ramalho (2016), Magrani & Ayres Neto (2014) and Yoon et al. (2004), all classified according to Damuth & Hayes (1977). Relative to the backscatter signal, Ramalho (2016) argued that all echo-types found in Port Foster Bay, glaciomarine bay in Antarctica, had the same intensity. However, in this research we discovered that King George Bay's echo-types have a differentiation (Figure 12), despite the apparent homogeneous distribution of backscatter intensity along study area (Figure 6). However, Port Foster Bay is greatly affected by volcanic processes in comparison with King George bay, which could explain the difference of Ramalho (2016) and our findings.

We have shown that the sedimentary nature and roughness of the seafloor, as well as the grazing angle of the seismic signal, are the main factors controlling acoustic responses for these three methods. However, there is a frequency dependency that affects signal penetration depth, which determines the impact of the volumetric characteristics of sediment for low-frequency signals. Such effects are theoretically known, but were not directly evaluated in our comparison of the methods.



We understand that echochacterization of the seafloor is a very subjective approach and depends greatly on the expertise of the individual conducting the analysis. However, since the variables (depth, backscatter intensity, amplitude, and slope) are largely independent, an artificial intelligence algorithm could be trained to predict echo-characters, based on classification supervised-learning. An implementation of a semi-automatic procedure to map the sedimentary distribution of a region of seafloor would have significantly positive temporal and cost implications, enabling a more efficient generation of reliable results and predictions. This solution is an ongoing research, which required prior analysis and understanding of the disposition and correlation of the variables depth, backscatter, seismic amplitude, echo-characterization and slope of the study area.

## **5 ACKNOWLEDGEMENTS**

The research was funded by the Ministry of Science and Technology and Innovation (MCTI)/Research National Council (CNPq)/National Fund for the Development of Science and Technology (FNDCT), Brazilian Antarctic Program – PROANTAR Call N° 64/2013.

## **6 REFERENCES**

Anderson, J.B; Brake, C.; Domack, E.; Myers, N.; Wright, R. (1983) - Development of a Polar Glacial-Marine Sedimentation Model from Antarctic Quaternary Deposits and Glaciological Information. IN: Glacial-marine Sedimentation. p. 233–264. Springer US.

Anderson, J. B. (1999). Antarctic marine geology. John B. Anderson. Cambridge: Cambridge University Press, vii + 289p, illustrated, hard cover. ISBN 0-521-59317-4. Polar Record.

37. 163. 10.1017/S0032247400027054.

Anderson, T. W. & Finn, J. D. (1997). The New Statistical Analysis of Data. Springer-Verlag, New York. 712 pp.

Assine, Mario & Vesely, Fernando. (2008). Ambientes Glaciais. In: PEDREIRA DA SILVA, A.J.; ARAGÃO, A.N.F.; MAGALHÃES A.J.C. (Eds.). Ambientes de Sedimentação Siliciclástica no Brasil. Ed. Beca. p. 24-51. São Paulo. 2008.

Ayres Neto, A., B. B. Mota, A. L. Belem, A. L. Albuquerque, and R. Capilla. (2016) - Seismic Peak Amplitude as a Predictor of TOC Content in Shallow Marine Sediments. Geo-Marine Letters 36 (5): 395–403. doi.org/10.1007/s00367-016-0449-3

Biffard, B. R., Bloomer, S. F., Chapman, N. R. and Preston, J. M. (2010) - The role of echo duration in acoustic seabed classification and characterization, Proceedings of IEEE OCEANS, 8 p., 2010

Borisov, D., Frey, D., Levchenko, O. (2020) - Sediment waves on the Santa Catarina Plateau (western South Atlantic). Journal of South American Earth Sciences 102.102698

Brand, J. R., D. L. Lanier, W. J. Berger Iii, V. R. Kasch, and A. G. Young. (2003) - Relationship Between Near Seafloor Seismic Amplitude, Impedance, and Soil Shear

Strength Properties and Use in Prediction of Shallow Seated Slope Failure. In Proceedings of the 35th Offshore Technology Conference, 1–17, Houston, TX.

Damuth, J. E., & Hayes, D. E. (1977). ECHO CHARACTER OF THE EAST BRAZILIAN CONTINENTAL MARGIN AND ITS RELATIONSHIP TO SEDIMENTARY PROCESSES. *Marine Geology*. Vol. 24(2). p. 73-96. 1977.

Duck, R.W. and Herbert, R.A. (2006) - High-resolution shallow seismic identification of gas escape features in the sediments of Loch Tay, Scotland: tectonic and microbiological associations. *Sedimentology* 53, 481–493

Fonseca, L. and Mayer, L. (2007) - Remote estimation of surficial seafloor properties through the application angular range analysis to multibeam sonar data. *Mar Geophys Res* 28(2):119–126

Fonseca, L., Hung, E.M., Ayres Neto, A., Magrani, F.J.G. (2017) - Waterfall notch-filtering for restoration of acoustic backscatter records from Admiralty Bay, Antarctica. *Marine Geophysical Research* 39.139–149

George, R.A., Gee, L., Hill, A.W., Thomson, J.A and Jeanjean, P. (2002) - High-Resolution AUV Surveys of the Eastern Sigsbee Escarpment. In Proceedings of the 34th Offshore technology Conference, paper 14139, Houston.

Goff, J.A., Kraft, B.J., Mayer, L.A., Schock, S.G., Sommerfield, C. K., Olson, H.C., Gulick, S.P.S., Nordfjord, S. (2004) - Seabed characterization on the New Jersey middle and outer shelf: correlatability and spatial variability of seafloor sediment properties. *Marine Geology* 209. 147–172

Hong, E. and Chen, I.S. (2000) - Echo characters and sedimentary processes along a rifting continental margin, northeast of Taiwan. *Continental Shelf Research* 20: 599-617

Jackson DR, Richardson MD (2007) - High-frequency seafloor acoustics. Springer, New York

L-3 Communication SeaBeam Instrument. (2000). *Multibeam Sonar Theory of Operation* (pp. 2–13).  
<https://www.ldeo.columbia.edu/res/pi/MBsystem/sonarfunction/SeaBeamMultibeamTheoryOperation.pdf>

Lark, R.M., Marchant, B.P., Dove, D., Green, S.L., Stewart, H., Diesing, M. (2015) - Combining observations with acoustic swath bathymetry and backscatter to map seabed sediment texture classes: The empirical best linear unbiased predictor. *Sedimentary Geology* 328. 17–32

Leitão, F. J., Ayres Neto, A., & Vieira, R. (2016). Morphological and sedimentary characterization through analysis of multibeam data at deception Island, Antarctic. *Brazilian Journal of Geophysics*, 34(2), 1–10. <https://doi.org/10.22564/rbgf.v34i2.792>

Llave, E., Schönfeld, J., Hernández-Molina, F.J., Mulder, T., Somoza, L., Díaz del Río, V., Sánchez-Almazo, I. (2006) - High-resolution stratigraphy of the Mediterranean outflow contourite system in the Gulf of Cadiz during the late Pleistocene: The impact of Heinrich events. *Marine Geology* 227. 241– 262

Magrani, F. J. G. and Ayres Neto (2014). Seismic characteristics and sedimentary distribution on the South Shetland Island continental margin, Antarctica. *Brazilian Journal of Geophysics* 32 (3). 549-560

Medwin, H. & Clay, C.S. (1998) – *Fundamentals of Acoustic Oceanography*. Academic Press Limited, London. 718pp

Mendoza, U., Ayres Neto, A., Abuchacra, R.C., Barbosa, C.F., Figueiredo Jr, A.G., Gomes, M.C., Belem, A.L., Capilla, R., Albuquerque, A.L.S. (2014) - Geoacoustic character, sedimentology and chronology of a crossshelf Holocene sediment deposit off Cabo Frio, Brazil (southwest Atlantic Ocean). *Geo-Mar Lett* 34:297–314.

Meng, Q., Liu, Z., Borders, B.E. (2013) - Assessment of regression kriging for spatial interpolation – comparisons of seven GIS interpolation methods. *Cartography and Geographic Information Science*, 40 (1). 28-39

Naudts, L., Greinert, J., Artemov, Y., Beaubien, S.E., Borowski, C., De Batist, M. (2008) - Anomalous sea-floor backscatter patterns in methane venting areas, Dnepr paleo-delta, NW Black Sea. *Marine Geology* 251: 253–267

OHI. (2005). MEDIÇÃO DA PROFUNDIDADE. In Manual de Hidrografia, pub. C-13 (pp. 115–190).

Ramalho, D. C. S. (2016). Characterization of vulcanoglaciomarine sediments using echo-characters in the Port Foster Bay, Depetion Island, Antarctica (In Portuguese).. MSc. Dissertation. Fluminense Federal University.

Ropper, C.N. and Zimmerman, M. (2007) - A bottom-up methodology for integrating underwater video and acoustic mapping for seafloor substrate classification. *Continental Shelf Research* 27. 947–957

Simons, D.G., Snellen, M. (2009) - A Bayesian approach to seafloor classification using multi-beam echo-sounder backscatter data. *Applied Acoustics* 70. 1258–1268

Solli, K., Kuvaas, B., Kristoffersen, Y., Leitchenkov, G., Guseva, J., Gandjukhin, V. (2007) - Seismic morphology and distribution of inferred glaciomarine deposits along the East Antarctic continental margin, 20°E–60°E. *Marine Geology* 237. 207–223

van der Meer, J.J.M., Menzies, J. and Rose, J. (2003) - Subglacial till: The deforming glacier bed. *Quaternary Science Reviews* 22: 1659–1685

Yoon, S.H., Yoon H.I. & Kang, C.Y. (2004) - Late Quaternary Sedimentary Processing Northern Continental Margin of the South Shetland Islands, Antarctica, *The Sea. Journal of the Korean Society of Oceanography*, 9: 1–12.

Zhang, Wengang; Wu, Chongzhi; Zhong, Haiyi; Li, Yongqin; Wang, Lin (2021) -  
Prediction of undrained shear strength using extreme gradient boosting and random  
forest based on Bayesian optimization.

#### 4. Condicionamento dos dados

A etapa de condicionamento dos dados é a que organiza e estrutura os dados para serem inseridos em determinada função/modelo. Ao todo, para atender todas as finalidades deste trabalho, foram gerados dois datasets para cada uma das baías, com as respectivas características:

- Dataset1 da Baía Rei George, o qual contém coordenadas UTM (EPSG 32732; WGS84; hemisfério sul) (X, Y) em metros, dados de ecocaracterização da área (ECO) classificados em classe 1 (Eco1), classe 2 (Eco2) e classe 3 (Eco3), com correspondentes atributos: profundidade (Z) em metros, backscatter (BS) em decibéis, declividade do fundo (SLOPE) em graus, distância para a desembocadura da baía (distOUTFALL) em metros e distância para o continente/frente de geleira (distGLACIER) em metros (Figura 1);

King George Bay Dataset1									
X	Y	ECO	Z	BS	SLOPE	distOUTFALL	distGLACIER		
441898.07	3113241.21	1.00	-244.76	-19.87	2.15	2294.68	3140.29		
445117.28	3112370.01	2.00	-439.11	-18.98	2.73	441.07	5109.43		
441896.93	3113672.92	1.00	-246.93	-16.64	0.77	2639.19	2977.85		
444497.46	3113395.49	1.00	-335.37	-15.72	6.42	827.24	4111.09		
446452.83	3113048.39	3.00	-282.62	-21.56	11.10	602.42	4349.05		
...	...	...	...	...	...	...	...		
								(14857,9)	

Figura 1: secção do Dataset1, relativo à Baía Rei George.

- Dataset2 da Baía Rei George, o qual contém todos os dados da superfície batimétrica, incluindo os atributos X, Y, Z, BS, SLOPE, distOUTFALL e distGLACIER (Figura 2);

King George Bay Dataset2							
X	Y	BS	Z	SLOPE	distGLACIER	distOUTFALL	
442440.00	3112740.00	-16.59	-238.51	5.24	3851.90	1564.34	
442445.00	3112740.00	-17.34	-239.18	2.01	3856.28	1561.33	
442450.00	3112740.00	-17.55	-239.11	3.27	3860.66	1558.33	
442420.00	3112745.00	-15.30	-237.41	12.28	3831.98	1580.45	
442425.00	3112745.00	-15.93	-237.99	2.46	3836.35	1577.41	
...	...	...	...	...	...	...	
							(427576,11)



Figura 2: secção do Dataset2, relativo à Baía Rei George.

- Dataset1 da Baía do Almirantado, o qual contém os atributos X, Y, ECO, Z, BS, SLOPE, distOUTFALL, distGLACIER, ASPECT e nome da linha sísmica (line) (Figura 3);

Admiralty Bay Dataset1										
X	Y	ECO	Z	BS	SLOPE	distOUTFALL	distGLACIER	ASPECT	line	
429900.33	3099573.04	2.00	-489.69	-32.00	14.05	2656.25	4058.12	340.00	Alm - L2b.sgy	
431343.47	3097163.57	3.00	-593.14	-30.00	3.63	5408.21	6473.68	22.00	Alm - L2b.sgy.001	
424919.96	3110278.89	1.00	-425.24	-28.00	8.45	9073.63	2015.43	103.00	OP32 - Almirantado 3	
429296.31	3105660.51	3.00	-394.37	-28.00	4.36	3203.25	2195.83	313.00	OP32 - Almirantado 3	
425873.87	3108882.88	1.00	-423.19	-27.00	44.57	7437.34	2315.26	190.00	OP32 - Almirantado 3	
...	...	...	...	...	...	...	...	...	...	

(69905, 10)

Figura 3: secção do Dataset1, relativo à Baía do Almirantado.

- Dataset2 da Baía do Almirantado, o qual contém todos os dados da superfície batimétrica, incluindo os atributos X, Y, Z, BS, SLOPE, distOUTFALL, distGLACIER e ASPECT (4);

Admiralty Bay Dataset2							
X	Y	Z	BS	SLOPE	distOUTFALL	distGLACIER	ASPECT
429464.00	3095264.00	-506.99	-25.00	5.87	6586.71	6405.53	27.00
429464.00	3095268.00	-506.99	-26.00	7.69	6582.90	6402.58	27.00
429468.00	3095268.00	-507.30	-27.00	8.45	6584.10	6405.27	27.00
429472.00	3095268.00	-507.44	-26.00	6.23	6585.31	6407.98	152.00
429472.00	3095272.00	-507.50	-26.00	9.56	6581.50	6405.03	152.00
...	...	...	...	...	...	...	...

(5420085, 8)

Figura 4: secção do Dataset2, relativo à Baía do Almirantado.

Os atributos dos Datasets1 e 2 diferem no sentido que os Datasets1 contém o atributo ECO; e os Datasets da Almirantado diferem da Rei George por aqueles conterem o atributo ASPECT. A escolha de suprimir o atributo ASPECT dos dados provenientes da Rei George justifica-se por ter ocasionado ruídos e redução da acurácia do modelo testado. Portanto, foi completamente descartável para a análise e optou-se pela remoção completa; entretanto o mesmo não foi observado no Almirantado, o qual obteve papel

relevante para aumento da precisão do modelo. Além disso, o Dataset2 do Almirantado contém um atributo exclusivo, referente ao nome das linhas sísmicas classificadas. Isso se dá por um requisito necessário para aplicação no **3º Artigo: XGBoost as a Tool to Improve High Resolution Single Channel Seismic Interpretation**. Abaixo, segue descrição da metodologia desenvolvida para criação dos datasets.

Inicialmente, os grids profundidade (Z) em metros, backscatter (BS) em decibéis, declividade do fundo (SLOPE) em graus e direção da declividade do fundo (ASPECT) em graus estão separados, apesar de geograficamente sobrepostos. Cada um deles corresponde a uma tabela de três colunas, nomeadas 'X', 'Y' e respectivo atributo (Z, BS, SLOPE, ASPECT). Para correto condicionamento, é necessário que todos esses grids estejam concatenados em uma só tabela.

Para realizar a concatenação dos Datasets1 do Almirantado e Rei George, foi fixado determinada coordenada geográfica da linha sísmica (X, Y, ECO) e, posteriormente, atribuído a ela a medição mais próxima (vizinho) de Z, BS e SLOPE, no caso da Rei George, ou Z, BS, SLOPE e ASPECT, no caso do Almirantado, desde que estejam espaçados em até duas vezes o espaçamento do grid da batimetria (Figura X). Para a Baía Rei George, esse espaçamento do grid foi de 3 metros, enquanto na Baía do Almirantado foi de 4 metros (Figura 5).

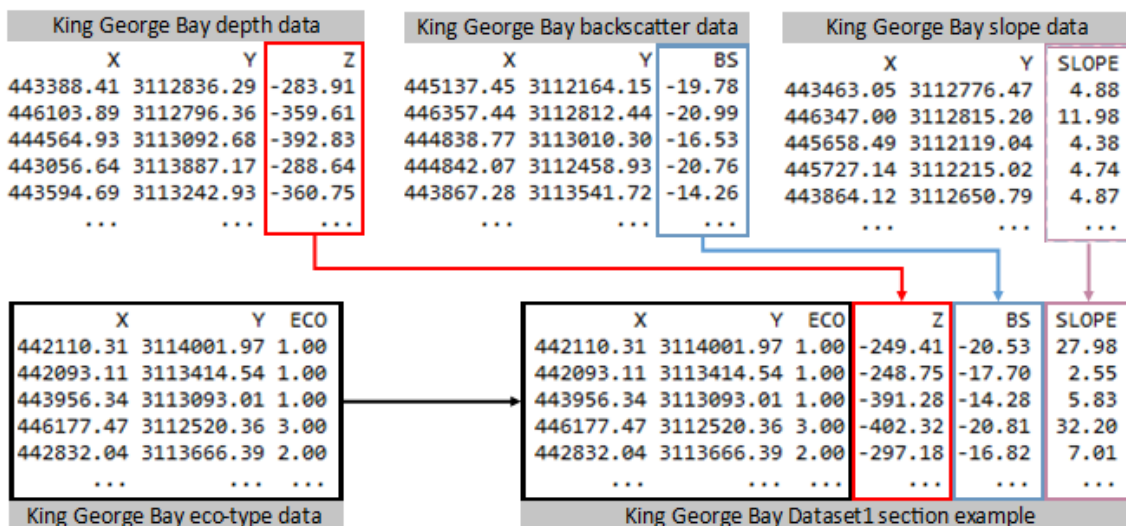


Figura 5: esquema mostrando como os Datasets 1 do Almirantado e Rei George foram construídos, segundo critérios espaciais.

Para realizar a concatenação dos Datasets2 do Almirantado e Rei George, foi fixado determinada coordenada geográfica da batimetria (X, Y, Z) e, posteriormente, atribuído a ela a medição mais próxima (vizinho) de BS e SLOPE, no caso da Rei George, ou BS, SLOPE e ASPECT, no caso do Almirantado, respeitando a mesma regra de espaçamento dos Datasets1 (Figura 6).

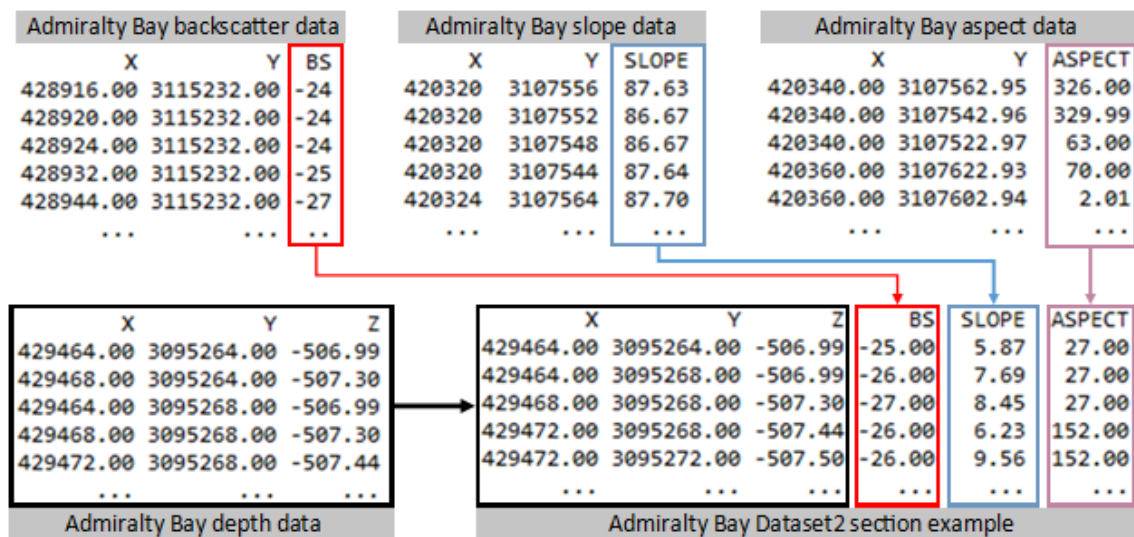


Figura 6: esquema mostrando como os Datasets 2 do Almirantado e Rei George foram construídos, segundo critérios espaciais.

Para tal, algumas metodologias foram implementadas e testadas, e a complexidade dos algoritmos foi avaliada. A complexidade do algoritmo ( $O$ ) é uma medida da quantidade de operações fundamentais/primitivas dentro de uma função com entrada de tamanho  $n$ , a fim de avaliar seu custo computacional. Dessa forma, permite comparar o desempenho de diferentes funções entre si, avaliando suas taxas de crescimento de complexidade/custo computacional. As abordagens avaliadas foram:

- (1) Fixar um determinado ponto  $(P_x, P_y)$  geográfico da malha de profundidade (X, Y, Z) e calcular via distância euclidiana o vizinho  $(Q_x, Q_y)$  mais próximo com medição de BS, SLOPE e ASPECT. A distância Euclidiana (D) é definida por:

$$D = \sqrt{(P_x - Q_x)^2 + (P_y - Q_y)^2}.$$

Nessa abordagem, foi utilizada a função `scipy.spatial.distance.cdist`<sup>2</sup>. Considerando como exemplo o dataset do Almirantado, que contém aproximadamente 5.4 milhões de medições ( $n = 5,4 * 10^6$ ) de Z, de BS, de SLOPE e de ASPECT, a quantidade de iterações do algoritmo é na ordem de  $3 * (5.4 * 10^6)^2 \cong 8,7 * 10^{13}$ . Como cada iteração leva em torno de  $2,34 * 10^{-8}$  segundos para ocorrer, tem-se o tempo estimado médio de  $2 * 10^6$  segundos para calcular a concatenação, ou aproximadamente 23 dias e 4 horas. Essa abordagem é a mais simples para o problema, entretanto devido ao elevado tamanho do dataset se torna impraticável pelo tempo computacional exigido, na medida em que possui crescimento quadrático – a complexidade desse algoritmo é  $O(n^2)$ .

- (2) Cálculo de distância Euclidiana da abordagem (1), entretanto utilizando GPU oferecida pela plataforma da Google Colab<sup>3</sup>. Para implementação foi utilizada a API da RapidsAi<sup>4</sup> que, instalada em máquina virtual e configurada no Colab, integrou as bibliotecas CuPy<sup>5</sup> e CuDf<sup>6</sup>, designadas para operar na GPU oferecida gratuitamente pela Google. Nesse caso, foi utilizada a GPU Tesla T4<sup>7</sup> em detrimento da CPU CORE i7-7500 2.70GHz utilizada nas abordagens (1) e (3).

---

<sup>2</sup> <https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.cdist.html> SciPy v1.7.1.

<sup>3</sup> <https://colab.research.google.com/> acesso em 23/11/2021.

<sup>4</sup> <https://rapids.ai/start.html> acesso em 23/11/2021.

<sup>5</sup> <https://github.com/cupy/cupy> acesso em 23/11/2021.

<sup>6</sup> <https://github.com/rapidsai/cudf> acesso em 23/11/2021.

<sup>7</sup> <https://www.nvidia.com/pt-br/data-center/tesla-t4/> acesso em 23/11/2021.

Dessa forma, a velocidade de processamento foi aumentada, reduzindo o tempo total de concatenação para aproximadamente 66 minutos.

- (3) Cálculo do vizinho mais próximo de determinado ponto utilizando a ferramenta *K-D-Tree*. Essa metodologia, implementada através da biblioteca *pykdtree*<sup>8</sup>, é por definição uma árvore de busca binária que contém  $k$  valores em cada nóculo da árvore. No referido caso de aplicação, temos  $k$  correspondente à quantidade de variáveis descritivas do fundo. O tempo computacional gasto foi de  $4,5 * 10^{-2}$  segundos para concatenar Z, BS, SLOPE e ASPECT no dataset do Almirantado. Para essa abordagem, a complexidade da *K-D-Tree* é  $O(n)$ .

Avaliando as complexidades de algoritmos das abordagens (1), (2) e (3), observamos uma discrepância enorme em termos de complexidade, visto que (1) e (2) crescem quadraticamente e (3) cresce linearmente. Nesse cenário, para valores muito grandes de  $n$ , algoritmos de complexidade quadrática tendem a se tornar inviáveis computacionalmente. Por outro lado, a GPU Tesla T4 possui 320 núcleos para paralelizar o processamento, ocasionando um tempo de processamento cerca de 505 vezes menor para a complexidade  $O(n^2)$ . Visto o exposto, conclui-se que a implementação *K-D-Tree* é a mais simples e rápida computacionalmente quando se objetiva encontrar o vizinho mais próximo, podendo ser empregada em computadores domésticos com facilidade. Alternativamente, quando o alto custo computacional for inevitável, a GPU torna-se a saída mais eficaz.

O próximo passo é incluir os atributos distância até desembocadura da baía (distOUTFALL) e distância até o continente, ou a frente de geleira (distGLACIER). Para tal, resta somente as soluções (1) e (2), que calculam a distância euclidiana entre as

---

<sup>8</sup> <https://pypi.org/project/pykdtree/> acesso em 23/11/2021.

coordenadas da tabela montada nas Figuras 5 e 6 e a desembocadura da baía e continente/frente de geleira. Para esta etapa, foram necessários cerca de 150 pontos definidos manualmente para delimitar a desembocadura e continente das baías Rei George e Almirantado, implicando em um custo computacional estimado de 20 segundos utilizando a abordagem (1) (Figura 7).

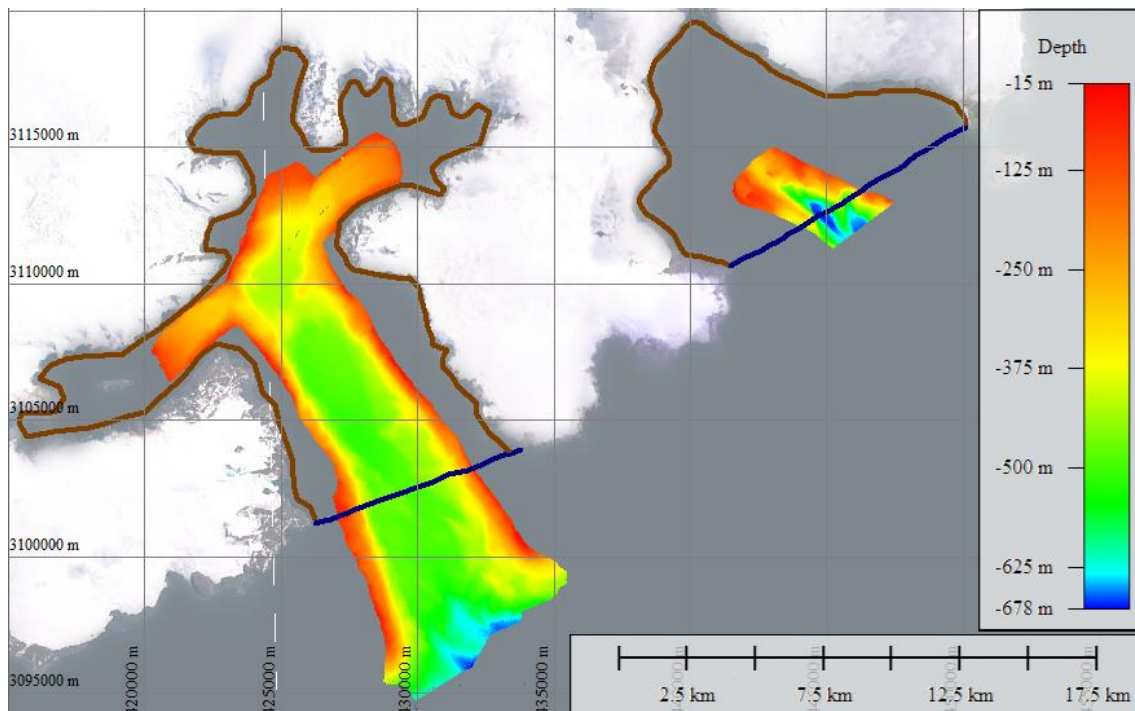


Figura 7: da esquerda para a direita: em azul marinho, a desembocadura das baías e, em marrom, a delimitação continental/frente de geleira das baías Almirantado e Rei George, respectivamente. No interior das baías, o DTM (Digital Terrain Model) batimétrico de cada uma das áreas.

Após concatenação de todas os atributos, tem-se que todos os datasets empregados nesse trabalho estão finalizados (Figura 8).

King George Bay Dataset1								
X	Y	ECO	Z	BS	SLOPE	distOUTFALL	distGLACIER	
441898.07	3113241.21	1.00	-244.76	-19.87	2.15	2294.68	3140.29	
445117.28	3112370.01	2.00	-439.11	-18.98	2.73	441.07	5109.43	
441896.93	3113672.92	1.00	-246.93	-16.64	0.77	2639.19	2977.85	
444497.46	3113395.49	1.00	-335.37	-15.72	6.42	827.24	4111.09	
446452.83	3113048.39	3.00	-282.62	-21.56	11.10	602.42	4349.05	
...	...	...	...	...	...	...	...	(14857, 9)

King George Bay Dataset2							
X	Y	BS	Z	SLOPE	distGLACIER	distOUTFALL	
442440.00	3112740.00	-16.59	-238.51	5.24	3851.90	1564.34	
442445.00	3112740.00	-17.34	-239.18	2.01	3856.28	1561.33	
442450.00	3112740.00	-17.55	-239.11	3.27	3860.66	1558.33	
442420.00	3112745.00	-15.30	-237.41	12.28	3831.98	1580.45	
442425.00	3112745.00	-15.93	-237.99	2.46	3836.35	1577.41	
...	...	...	...	...	...	...	(427576,11)

Admiralty Bay Dataset1										
X	Y	ECO	Z	BS	SLOPE	distOUTFALL	distGLACIER	ASPECT	line	
429900.33	3099573.04	2.00	-489.69	-32.00	14.05	2656.25	4058.12	340.00	Alm - L2b.sgy	
431343.47	3097163.57	3.00	-593.14	-30.00	3.63	5408.21	6473.68	22.00	Alm - L2b.sgy.001	
424919.96	3110278.89	1.00	-425.24	-28.00	8.45	9073.63	2015.43	103.00	OP32 - Almirantado 3	
429296.31	3105660.51	3.00	-394.37	-28.00	4.36	3203.25	2195.83	313.00	OP32 - Almirantado 3	
425873.87	3108882.88	1.00	-423.19	-27.00	44.57	7437.34	2315.26	190.00	OP32 - Almirantado 3	
...	...	...	...	...	...	...	...	...	...	(69905, 10)

Admiralty Bay Dataset2									
X	Y	Z	BS	SLOPE	distOUTFALL	distGLACIER	ASPECT		
429464.00	3095264.00	-506.99	-25.00	5.87	6586.71	6405.53	27.00		
429464.00	3095268.00	-506.99	-26.00	7.69	6582.90	6402.58	27.00		
429468.00	3095268.00	-507.30	-27.00	8.45	6584.10	6405.27	27.00		
429472.00	3095268.00	-507.44	-26.00	6.23	6585.31	6407.98	152.00		
429472.00	3095272.00	-507.50	-26.00	9.56	6581.50	6405.03	152.00		
...	...	...	...	...	...	...	...	...	(5420085, 8)

Figura 8: Datasets 1 e 2 das Baías Rei George e Almirantado, após procedimentos de concatenação.

# Machine Learning Modeling Applied for Seabed Echo-Characterization in King George Bay, Antarctica (\*\*).

Diogo Ceddia Porto Silva 1<sup>a</sup>, Reinaldo Mozart da Gama e Silva 2<sup>b</sup> and Arthur Ayres Neto 3<sup>c</sup>

<sup>a</sup> Marine Geology Laboratory (LAGEMAR), Department of Geology and Geophysics, Federal Fluminense University, Rio de Janeiro, Brazil, ORCID(s): 0000-0002-6278-2250

<sup>b</sup> Independent Consultant.

<sup>c</sup> Marine Geology Laboratory (LAGEMAR), Department of Geology and Geophysics, Federal Fluminense University, Rio de Janeiro, Brazil, ORCID(s): 0000-0002-2982-245X

(\*\*) **Submitted In *Computers & Geosciences Journal*.**

## ABSTRACT

The combination of multibeam bathymetry, backscatter and high-resolution seismic is a well-known method for seafloor mapping. In this context, we highlight the difference between bathymetry – that covers the totality of the survey area – and high-resolution seismic data, restricted to the survey lines. This discrepancy interferes negatively in the interpretation of the seabed since correlation between bathymetry and seismic becomes limited. In this research, we applied machine learning modeling to extrapolate the echo characterization (obtained from seismic interpretation) to the area covered by the multibeam bathymetry. The approach was developed and explored, obtaining outcomes around 98% of balanced accuracy using eXtreme Gradient Boosting (XGBoost). Our results encompass redundancy analysis, data explainability, feature selection, prediction for data extrapolation and further analysis. In conclusion, we established a reliable workflow, which achieved an auxiliary tool for seafloor characterization. Our research provides a useful alternative approach for mapping



geospatial data which may have acquisition limitations (cost, high complexity, etc.), since one has enough gridded data available from an area of interest and liable to be related.

**Keywords:** data processing, echo-characterization, seismic, bathymetry, machine-learning.

## INTRODUCTION

Many marine activities (marine geology, commercial fishing, offshore oil exploration and production, cable and pipeline maintenance, underwater warfare, engineering works, wind farms, and dredging operations) need tools and methods to characterize and understand the seafloor. Multibeam Echo Sounders (MBES) are well designed for this task: usually installed under a ship's hull, it transmits a sound pulse (ping) with a wide across-track and narrow along-track angular sector. Those fan-shaped beams can acquire high-density data from large areas of the seabed. Moreover, MBES systems have quickly developed over the last 60 years, being a reliable way to characterize the seafloor. They are considered one of the main mapping systems, due to the ability to provide both bathymetric and backscatter image of the surveyed area. The bathymetric data outputs an accurately georeferenced digital terrain model (DTM), while the backscatter image is a mosaic of echo amplitudes containing information about the nature and geoacoustic properties of the seafloor (Hellequin et al., 2003; Xinghua and Yongqi, 2004; Zhou et al., 2020). Another important tool is the high-resolution seismic, acquired by sub-bottom profiler (SBP) systems. The equipment is also usually mounted in the ship's hull, emitting a vertically directed acoustic pulse that interacts with the seafloor. When it operates in frequencies below 10 kHz, the seismic signal can penetrate

deep in the seabed, revealing inner sedimentary settings. Those acoustic responses – known as echo-types – are visible due to impedance contrast between different layers defined by transport and nature of the sediments. The SBP data can be interpreted according to echo-types, whereas each echo-type can be correlated with seafloor composition, if physical sediment samples are obtained. However, the approach to collect physical samples is very expensive and limited. Similarly SBP data only acquire linearly, along ship's navigation track (Saleh and Rabah, 2016; Wang et al., 2021), and can be considered a limited method as well when compared to MBES area data acquisition property.

There is a huge discrepancy in term of amount of information in this scenario, since bathymetry is a surface arrangement and seismic is settled as navigation lines. Since echo characterization is confined to seismic lines, its interpretation is barely related with other attributes, such depth, backscatter intensity, seafloor slope and more. This article aims to offer a way to understand echo characterization as a coverage layer, a surface that can overlay any other; built based on correlations with geophysical, topographic or any other griddable data. To achieve our goal, we used eXtreme Gradient Boosting (XGBoost) algorithm to predict echo-types in the area covered by the bathymetry, using the SBP data for training and evaluation of the model. In other words, our research intends to overcome the acquisition limitation of SBP data, offering an auxiliary way to interpret the seabed.

Several authors, as we propose, used machine-learning techniques to predict missing data and improve mapping, estimations, interpretations, and data acquisition limitations in remote sensing field. Li et al. (2020), for example, used multiple environmental variables (griddable data from area of interest) to map soil thickness, a difficult and expensive data to acquire. They found that XGBoost outperformed other

models, predicting a map with limited validation data. Sahin, 2020 built a map of landslide susceptibility based on few landslide occurrences. He constructed his dataset based on existing maps, digital terrain models and satellite images. He also found that XGBoost had better performance in ensemble models. Shendryk et al. (2020) also applied XGBoost for mapping an invasive grass in tropical savannas of northern Australia. They reached up to 91% of accuracy using satellite images, and collecting field data for supervised learning. Arabameri et al. (2021) did a spatial modeling of gully erosion using ensemble algorithms, where GE-XGBoost got the best performance, with 89.56% of accuracy. They used spatial data and database containing recorded instances of erosion, arguing that their model is a promising method for large-scale mapping of gully erosion susceptibility. Emadi et al. (2020) predicted and mapped the content of Soil Organic Carbon (SOC), in northern Iran, using several AI algorithms, including XGBoost. They used soil samples, satellite images, terrain attributes, climatic data, and other data to extrapolate the SOC content in a bigger area. They also stated that their predicted map could be used as a base line for further studies, both locally and in a worldwide scale.

Considering these researches, this paper presents such a similar methodology for the prediction of seismic echo-characterization. Our dataset includes well-known techniques for seafloor mapping, such as bathymetry, backscatter and high-resolution seismic. Many authors used those tools to classify both morphologically and sedimentologically the seabed. Early 2000's, Hellequin et al. (2003) and Xinghua and Yongqi (2004) demonstrated the importance of bathymetry and backscatter processing for seafloor characterization. More recently researches point to the same direction. Magrani (2014), for example, used bathymetry and echo-characterization to map the morphology, sediment distribution and sedimentary thickness of Admiralty Bay, South Shetland Islands, Antarctica. Saleh and Rabah (2016) used Sub-Bottom Profilers (SBP)

for sediment classification, estimating and analyzing reflection and sediment absorption coefficients. Ji et al. (2020) used several features, including backscatter data, for sediment classification of the seabed. They achieved accuracies over 90%, and extrapolated the classification of few samples to the whole surveyed area. Wang et al. (2021) classified seabed sediments based on particle size parameters, using XGBoost. They suggested that the model could be used as an auxiliary or alternative approach for sediment texture mapping, as well as supplementary to the analysis of sedimentary environment. Finally, Marsh and Brown (2009) used multibeam bathymetry and backscatter data for seabed classification, through neural network. They stated the need to develop automated computational methods that transform large areas of spatially-complex ('high dimensional') bathymetric and backscatter data into simpler, easily-visualization ('low dimensional') maps that, in some way, characterize the seafloor. We could not agree more, since our goal is to overcome a data acquisition limitation and predict echo-types within a bigger area, transforming an overlay of grids (high dimensionality) into a simpler one – the echo-character map.

The study area is the King George Bay, located in King George Island, South Shetland Islands, Antarctica. This area was surveyed with sub-bottom profiler and multibeam bathymetry concomitantly. The data consisted of a bathymetric DTM of 10.8 Km<sup>2</sup>, and approximately 58 Km of echo-classified high-resolution seismic lines (Figure 1).

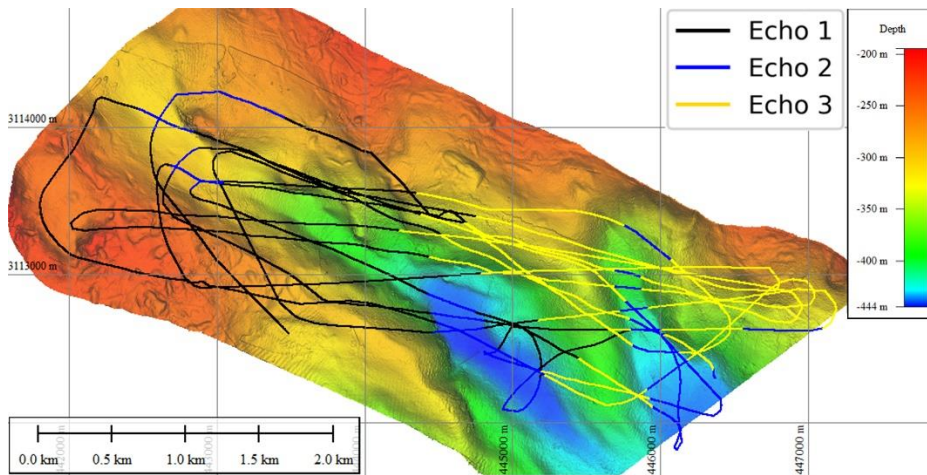


Figure 1: King George Bay’s data, regarding SBP echo-characters and bathymetric coverage.

## MATERIALS AND METHODS

The fundamental stages of the adopted methodology comprises the following steps: (1) preprocessing, involving the data filtering and conditioning; (2) data analysis, redundancy analysis, explainability, and feature selection; and (3) prediction, relative to the echo-classification extrapolation and posterior analysis, which contemplates geological coherence, computational cost and uncertainty map (Figure 2).

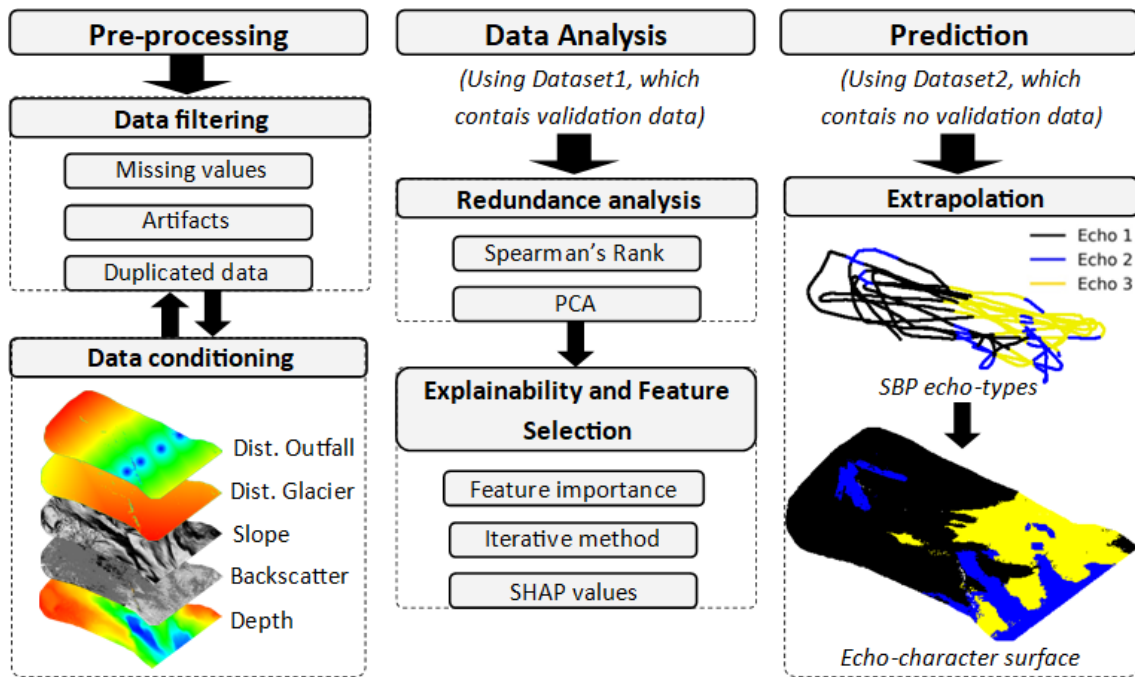


Figure 2: workflow of the present methodology.

Initially, the data was submitted to carefully preprocessing steps. This stage is fundamental for ML models, since it guarantees the best quality of the data used for training (Huang, Li and Xie, 2015). First, outliers, inconsistent data and artifacts (non-geological-topographical structures) were all removed from the training dataset. We mostly removed anomalies caused by the absence of tide correction, which resulted in slopes ranging from 50 to 90 degrees, and backscatter inconsistencies in the seafloor. Moreover, this stage filtered missing values and categorical variables. Fortunately, only 0.5% of the entire dataset was removed.

In the data conditioning stage, we used a Eucliden Distance algorithm to connect all data grids and assemble the datasets. Each grid relates to a different feature, which are: depth (Z), in meters, set as negative below sea level; seafloor slope (SLOPE), in decimal degrees, extracted from the bathymetric DTM using Global Mapper 20 software; backscatter intensity (BS), in decibels, that correspond to the intensity of the returned

bathymetric signal; distance relative to the glacier front (distGLACIER), in meters, which controls the content of proglacial material in the sedimentary deposit; and the distance relative to the bay's outfall (distOUTFALL), in meters, which controls the glaciomarine influence in the sedimentary deposit (Anderson, 1999; Anderson *et al*, 1983; Magrani, 2014). We created the distOUTFALL and distGLACIER features, motivated by previous studies (Anderson, 1999; Anderson *et al*, 1983; Leitão, 2015; Leitão *et al.*, 2016; Magrani, 2014; Viana, 2014). These authors claimed that distances relative to the glacier front and to the bay's outfall influences the sedimentary setting. Our prediction target are the echo-types, previously classified as Echo 1, Echo 2 and Echo 3 and according to Damuth, J. E., & Hayes, D. E. (1977) criteria (Figure 3).

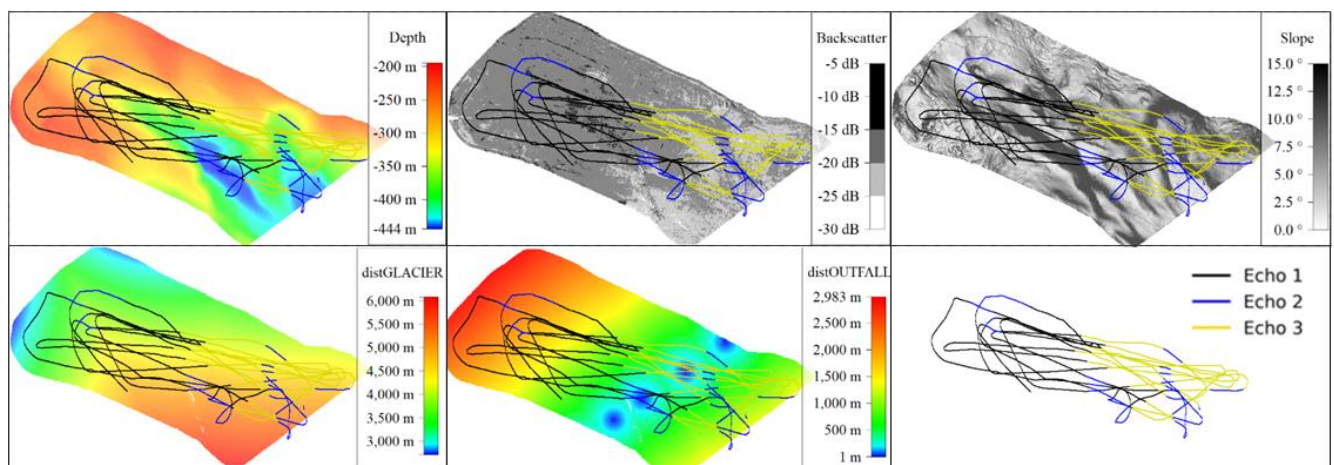


Figure 3: relationship of each feature relative to the target (last rectangle).

The grid spacing was set to 5 m (Figure 4, represented by black dots), despite the SBP ping rate results in approximately 1 m distancing between seismic traces (Figure 4, represented by red dots). The Euclidean distance algorithm associate one red dot to the closest black dot, with a threshold of 10 m – two times the feature grid spacing. In this context, several SBP samples (echo-types) registered the same Z, BS, and SLOPE

information, which imply in duplicated information in the dataset for the same target classification. Consequently, this data redundancy added an unreal and random weighing into the ML model used in our approach. To solve this problem, we deleted all duplicated rows with the same value of ECHO, Z, BS, and SLOPE. This redundancy filtering reduced the validation data in 74.4%, with an accuracy loss of 1% using eXtreme Gradient Booster (XGB) method. The accuracy loss is interpreted as an uncertainty of the model, caused by incorrect weighing. Moreover, the dataset reduction improved a lot the computational cost, with an average processing time reduction of 70% for model fitting and testing.

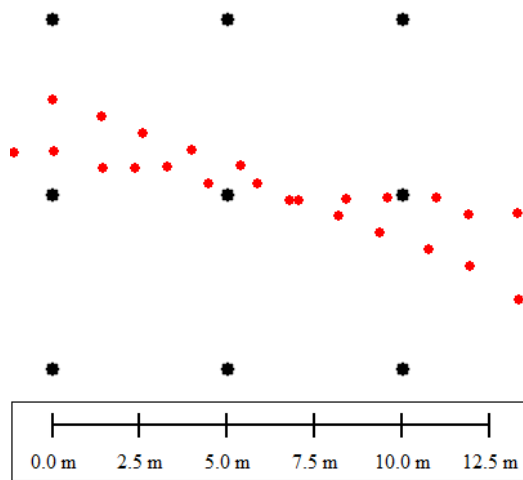


Figure 4: data conditioning stage, regarding grid spacing. Feature data represented in black dots, while target data in red dots.

The data were splitted into two sets: Dataset1 (14,857 rows), which contains only data with validation – therefore, each echo-type measurement with its respectively closest features –, and Dataset2 (427,576 rows), which has no validation data – corresponding to the whole area without echo-classification. The Dataset1 is unbalanced: 48.3% of the data corresponds to Echo1, 14.9% to Echo 2 and 36.8% to Echo 3 (Figure 5).



X	Y	ECHO	BS	Z	SLOPE	distOUTFALL	distGLACIER
443453.095568	3.112635e+06	1.0	-16.1266	-304.582	5.216	854.088927	4806.278085
445142.692872	3.112172e+06	2.0	-18.2358	-440.435	3.640	588.635288	5301.259000
443482.211989	3.114048e+06	1.0	-16.9652	-284.788	9.306	1916.574193	3556.628357
443003.384917	3.113908e+06	1.0	-14.7571	-268.944	4.776	2084.756726	3831.229380
447018.952637	3.112807e+06	3.0	-21.1238	-315.077	12.217	1192.165632	4521.974287
...	...	...	...	...	...	...	...

Figure 5: Example of Dataset1. Dataset2 has the same structure, but without the column ‘ECHO’ (no validation data).

Proceeding the data analysis stage, we chosen the eXtreme Gradient Boosting (XGBoost) algorithm to perform our supervisionated modeling. Based on the boosting strategy, XGBoost can obtain a strong learner from weak learners. The XGBoost algorithm can improve computing speed by parallel learning, prevent over-fitting and improve performance. Also, it is widely used by data scientists to achieve state-of-the-art results on many machine learning challenges, such as Kaggle (17 of 29 winning solutions in 2015) and KDDCup 2015 (all 10 winning teams used XGB) (Ariza-Garzon et al., 2020; Chen and Guestrin, 2016; Li et al., 2020). Moreover, several authors applied this algorithm for mapping in geoscience and remote sensing fields (Arabameri et al., 2021; Emadi et al., 2020; Ji et al., 2020; Li et al., 2020; Sahin, 2020; Shendryk et al., 2020; Wang et al., 2021), achieving consistent results. Before the training stage, Dataset1 was normalized (\*1) and randomly splitted without repeated elements, 80% for training and 20% for validation. Every data-splitting procedure, including this one, took into consideration the 0.8-0.2 ratio and the proportionally stratified sampling (\*2), due to the imbalanced dataset. We used the balanced accuracy score as metric of our model (\*3). Then, a preliminary training was done as basis for further data analysis.

The data analysis stage embraces redundancy analysis, data explainability and feature selection techniques, aiming for data/model comprehension and feature selection. Understanding why a model makes a prediction is important for trust, actionability, accountability, debugging, and many other common tasks. We established a guideline to define what is and what is not important for the prediction, seeking to avoid redundancy and overfitting. Those tools provide the necessary information for a possible dimensionality reduction, therefore improving computational cost (Ariza-Garzon et al., 2020; Coyle and Weller, 2020; Lundberg and Lee, 2017a; Raihan-Al-Masud and Mondal, 2020; Roscher et al., 2020). Moreover, this stage provides necessary information to understand the decisive variables that control class occurrence in this environment. For redundancy analysis, explainability and feature selection, several tools were tested: Spearman's Rank Correlation (\*4) and Principal Component Analysis (\*5), to check for correlations between features; a comparison between XGB feature importance (\*6), SelectKBest function (\*7) and SHAP importance (\*8), in order to rank the most relevant features; SHAP values, to comprehend how the model makes decisions; and an iterative approach, which provides information of every combination of features.

An important step for dataset improvement is the redundancy analysis. To solve this issue, Spearman's Rank Correlation matrix evaluate the monotonic degree of relationship between features. A monotonic relationship measures if a variable 'A' is directly or inversely correlated to another variable 'B'. We chose Spearman's instead of Pearson's because our data is mostly non-parametrized. High correlations ( $>0.9$ ) could imply in redundant data and, therefore, excessive processing time. We set correlation values based on Anderson and Finn (1997): weak correlation from 0 to 0.3; medium correlation from 0.3 to 0.6; strong correlation from 0.6 to 0.9; and perfect correlation from 0.9 and 1.

Aiming an in-depth understanding of the possible redundancy between features, a Principal Component Analysis (PCA) were conducted. This is a technique for reducing the dimensionality of datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximizes variance. The same features inputted in the XGB model constitute the new variables, or Principal Components (PC's). Each PC explain a percentage of the variance of the data, making possible to evaluate the redundancy of the features: if the features are redundant, they will express themselves in the same proportions in the PC's. Otherwise, they will compose the PC's in some unique way. (Jolliffe et al., 2016; Maćkiewicz and Ratajczak, 1993; Shlens et al., 2014).

Later, we compared several methods that attribute feature importance, due to inconsistent results: they often assign higher importance to features with lower impact on the model's output (Lundberg and Lee, 2017a). Our guideline is to address functions with increasing computational cost, comparing the cost-benefit of each approach.

At first, we computed the XGB feature importance. Because it is an inner output from XGBoost model, it has no additional computational cost. Five importance types can be calculated (\*6):

- Weight, defined as the number of times a feature is used to split the data across all trees.
- Gain, defined as the average gain across all splits the feature is used in. It implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. A higher gain means the feature is more important for generating a prediction.

- Cover, defined as the average coverage across all splits the feature is used in.

This metric means the relative number of observations related to this feature, reflecting the proportion of leaf nodes each feature decides.

- Total\_gain, defined as the total gain across all splits the feature is used in;
- Total\_cover, defined as the total coverage across all splits the feature is used in.

SelectKBest function, on the other hand, assigns importance scores based on univariate statistical tests, computing the ANOVA F-value variance (\*7).

Finally, we computed the SHAP (SHapley Additive exPlanations) feature importance (\*8), considered the only consistent feature attribution method by Lundberg and Lee (2017a). This tool provides a better interpretability for the model, assigning to each feature an importance value for a particular prediction. SHAP values have three properties: local accuracy, which determines that the explanation of the model should match the original model; missingness, if the simplified inputs represent feature presence, then missingness requires features missing in the original input to have no impact; and consistency, which states that if a model changes so that some simplified input's contribution increases or stays the same regardless of other inputs, that input's attribution should not decrease (Ariza-Garzon et al., 2020; Lundberg and Lee, 2017b, 2017a; Lundberg, S.M., Erion, G., Chen, H. et al., 2020). However, since SHAP measures the feature importance for each class, we did a weighted average (respecting the data imbalance) for all classes and defined the overall importance for every feature. Later in explainability stage, SHAP values were computed also aiming understand how the model take decisions.

As the true evidence for evaluate previous redundancy, explainability and feature selection analysis, we implemented an iterative approach that computes the accuracy and

time performances of all possible combinations of  $n$  features. This methodology was used to find and select key feature combinations among a dataset. This tool is considered the ultimate validation for feature selection/elimination, because it iteratively computes the model with different feature combination setups (Mosavi et al., 2020). This test intended to be the last performance validation of feature importance methods, clarify possible redundancy between features and to, finally, define the best feature combinations for  $n$  features. This methodology is the most expensive regarding processing time.

The prediction stage trained the model with the entire Dataset1 and realized a prediction with Dataset2. The result was an echo-classification coverage layer, along all area covered by the bathymetry. In this step, we compared the layer with and without optimization, always aiming for geological coherence. An uncertainty map was also used, built based on the `XGBClassifier.predict_proba()` (\*6). The function outputs the probability, for each measurement, of being Echo 1, Echo 2 or Echo 3. We established a threshold of uncertainty of 5% and created a map of predictions with less than 95% of confidence.

The usage of XGB model, as well all the analysis and studies, were done using the same machine, with a CORE i7-7500 CPU 2.70GHz processor, and 8Gb RAM memory.

### **3 RESULTS AND DISCUSSIONS**

#### **Spearman's Rank Correlation**

As a preliminary step, we verify the redundancy of features in Dataset1. The Spearman's Rank Correlation showed expected medium and strong correlations: when depth (Z) increases, distance to the glacier increases, and distance to the bay's outfall and backscatter decreases. Moreover, according to this test, since there are no high (>0.9) correlations levels, there would be no redundant features in the dataset (Figure 6).

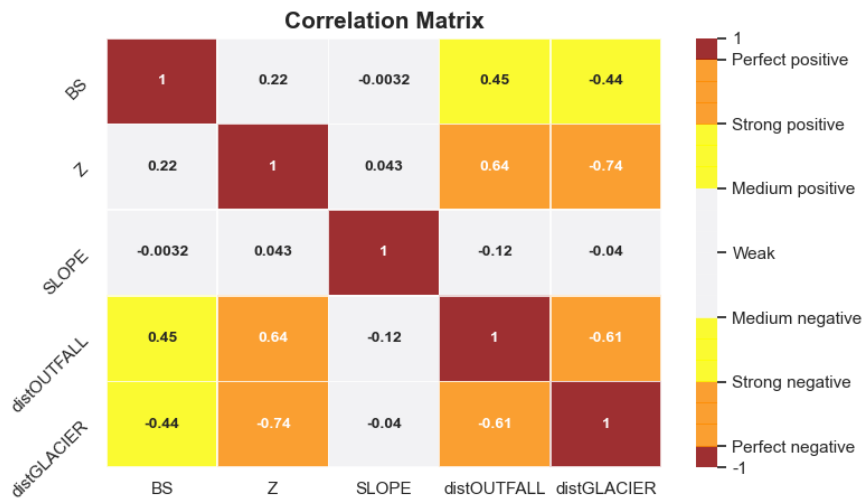


Figure 6: Spearman's Correlation Rank matrix, computed to analyze the monotonic relationship between features.

### Principal Component Analysis (PCA)

Another tool that can be used before ML application is the Principal Component Analysis (PCA). When two features are perfectly correlated, they are expressed by the same number of Principal Components (PC's) and shares uniquely one PC. The results showed that features Z and distGLACIER have similar PC distribution, therefore explaining data variance in a suck like way and being potentially redundant (Figure 7). In addition, PC5 is mostly distributed in the Z and distGLACIER features (85%), which corroborate this idea. However, since Spearman's Rank test measured a correlation coefficient of -0.74 (therefore not perfect-correlated) between these two features, we

decided to keep them in the dataset. It is important to reiterate that Spearman’s Rank and PCA are basis for feature selection/elimination, whereas reduce the computational cost before inputting data into the model.

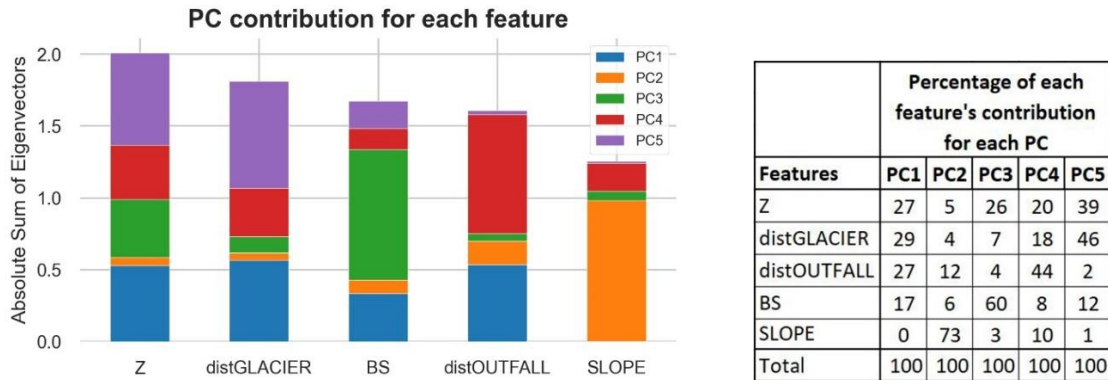


Figure 7: the PCA result, showing how each PC explains each feature.

### Feature importance

In this step, we compared different methods for feature importance attribution, with increasingly processing time. XGB feature importance, as an inner product of XGBoost model output, does not have any additional processing cost. It can be calculated based on ‘weight’, ‘gain’, ‘total\_gain’, and ‘total\_cover’, which are countable characteristics of gbtrees. SelectKBest had 0.003 s of processing time, and SHAP importance took 5.6 s. Comparing each attribution method (Figure 8), we observe that different methods for the same dataset has inconsistent results, as stated by Lundberg and Lee (2017a). Therefore, picking one method solely as a reliable importance rank between features can imply in a non-optimal feature selection procedure. Moreover, it can lead to a misunderstanding about the key features that control the echo occurrence in the area. On the other hand, a reliable way to measure importance of features can clarify and

quantify what the literature considers as key factors that control sedimentary settings in glaciomarine environments.

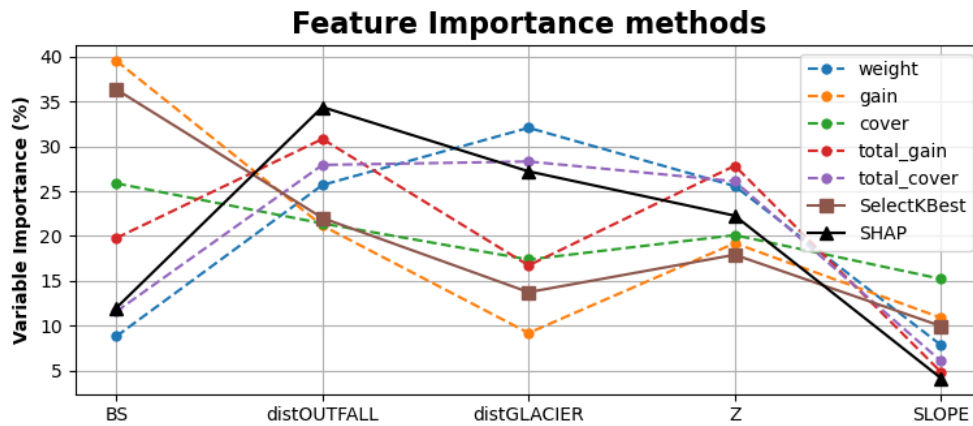


Figure 8: different importance attribution methods result for the Dataset1.

### Iterative approach

Since importance attribution methods have inconsistent results, we iterate every combination of feature and evaluate its time and accuracy (Table 1). In this final analysis, the true importance is highlighted by accuracy results and can be compared with other importance attribution methods. This step is the most computationally expensive, with an average processing time of 50 seconds for 100 iterations. In an overall analysis, several distinct combinations of features (15 of 31) resulted in great accuracy marks (>90%), strongly suggesting that our features, in general, correlate well with our targets. The results also point that when Z, distOUTFALL and distGLACIER are present, the accuracy is around 98% (first four combinations in Table 1). Without any of those features, the accuracy drops to less than 96%. The best combination of 3 features is ‘Z, distOUTFALL, distGLACIER’; the best combination of 2 is ‘distOUTFALL, distGLACIER’, followed by ‘Z, distGLACIER’ and ‘Z, distOUTFALL’; and the best solo feature for prediction is ‘distGLACIER’, followed by ‘Z’ and ‘distOUTFALL’. Therefore, analyzing the table we



can conclude these are the most relevant features for predictions, while BS and SLOPE are less important. Comparing with the feature importance methods, we can state that only ‘weight’, ‘total\_cover’ and SHAP had ranked Z, distOUTFALL and distGLACIER above SLOPE and BS; while ‘gain’ and SelectKBest ranked BS as the most important feature; and ‘cover’ and ‘total\_gain’ ranked BS higher than distGLACIER. That said, we can also agree with Scott M. Lundberg and Lee (2017a), which considered SHAP as the only consistent method for feature attribution, insofar as SHAP performed really well estimating the most important features for model’s decision. Another interesting fact is that ‘gain’ is the standard importance type of XGBoost in Python (\*6) and did not performed well – it defined BS as the most important feature. That said, we can state that the inconstancy of importance attribution algorithms is really something to consider, and default XGBoost importance type can mislead to wrong interpretations. However, SHAP importance is a consistent method that led us to the assertive solution. Relative to the variables that control sedimentary settings in King George Bay, we can state that among our features the most important ones are distance to the glacier and to the bay’s outfall. Those findings confirm what Anderson et al. (1983), Anderson (1999) and Assine & Vesely (2008) define as main sedimentation control factors in glaciomarine environments.

Considering the possible redundancy between Z and distGLACIER highlighted by Spearman’s Rank Correlation and PCA analysis, the iterative approach revealed that both features describe differently the target. This assumption is because both features are necessarily together for the best accuracy scores of Table 1. Those variables, together, have an accuracy of 93.116% and are the second-best combination of two features. Therefore, all analysis done until now reveal no need to drop any features justified by redundancy level.

Table 1: accuracy and processing time comparison relative to all feature combinations.

RFE					
Feature combination	Accuracy (%)	Time (s)	Feature combination	Accuracy (%)	Time (s)
BS, Z, SLOPE, distOUTFALL, distGLACIER	98.154	2.309	BS, SLOPE, distGLACIER	81.657	1.638
Z, distOUTFALL, distGLACIER	98.053	1.569	BS, Z, SLOPE	81.490	1.609
BS, Z, distOUTFALL, distGLACIER	97.949	1.445	BS, SLOPE, distOUTFALL	78.266	1.636
Z, SLOPE, distOUTFALL, distGLACIER	97.919	1.430	BS, Z	76.019	1.471
BS, SLOPE, distOUTFALL, distGLACIER	96.155	1.481	BS, distGLACIER	73.526	1.478
BS, distOUTFALL, distGLACIER	95.427	1.630	SLOPE, distGLACIER	73.153	1.459
BS, Z, distGLACIER	95.402	1.601	Z, SLOPE	66.319	1.437
BS, Z, SLOPE, distGLACIER	95.296	1.476	SLOPE, distOUTFALL	64.904	1.478
SLOPE, distOUTFALL, distGLACIER	94.952	1.633	BS, distOUTFALL	64.830	1.484
BS, Z, SLOPE, distOUTFALL	94.902	1.475	BS, SLOPE	64.650	1.492
Z, SLOPE, distGLACIER	94.530	1.609	distGLACIER	61.650	1.452
distOUTFALL, distGLACIER	93.706	1.441	Z	59.959	1.420
BS, Z, distOUTFALL	93.159	1.606	distOUTFALL	54.363	1.452
Z, distGLACIER	93.116	1.428	BS	53.552	1.475
Z, SLOPE, distOUTFALL	91.474	1.611	SLOPE	44.812	1.397
Z, distOUTFALL	88.063	1.433			

## SHAP Values

SHAP Values was the algorithm chosen to understand how the model makes decisions. It is more robust and computationally expensive, taking 5.6 seconds to finish the analysis (the same for SHAP importance). The results show the impact of the input features on each individual prediction, as shown in the summary plots below (Figures 9, 10 and 11) (\*8). The color bar represents a gradation between low (blue) and high (red) values of the normalized feature. The X-axis corresponds to how a feature influences the model output, either positively or negatively.

In an overall analysis, Z, distGLACIER and distOUTFALL are the most important features for the model to decide for any echo-type. Some unique differences between echoes and geological relationships are:

- In Echo 1, shallower depths ( $Z$ ) impact positively in the model's output. Moreover, backscatter is a more decisive feature for this echo-type, assuming higher values. In fact, this class occurs in shallower areas, where backscatter signal is stronger. Echo 1 also occurs closer to the glacier front – most important feature – and further to the bay's outfall.

- In Echo 2,  $Z$  is the most important feature, with a high impact in the model's output for greater depths. Here, slope is more decisive relative to other echoes, assuming low values. This is explained by the occurrence of the echo-type inside the channel's thalwegs (Figure 1).

- In Echo 3, the class occurrence is closer to the bay's outfall, which is the most decisive feature for the model's output.

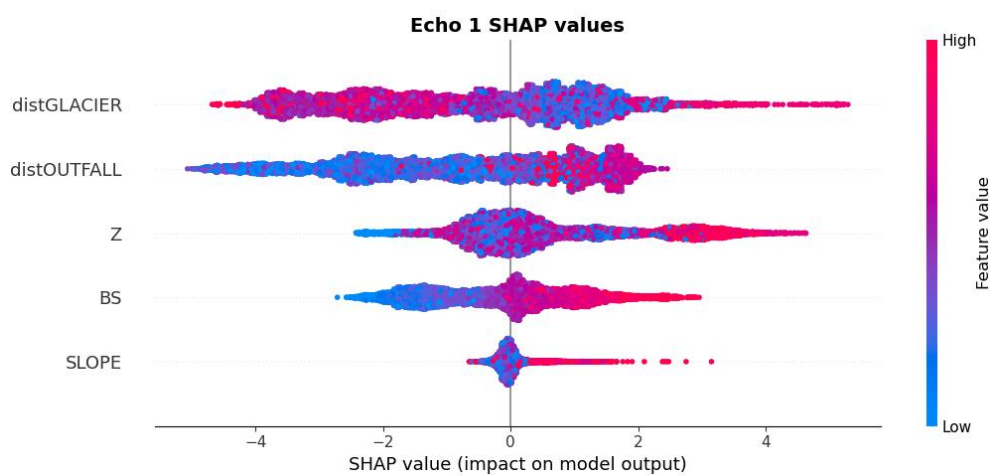


Figure 9: the summary plot of Echo 1.

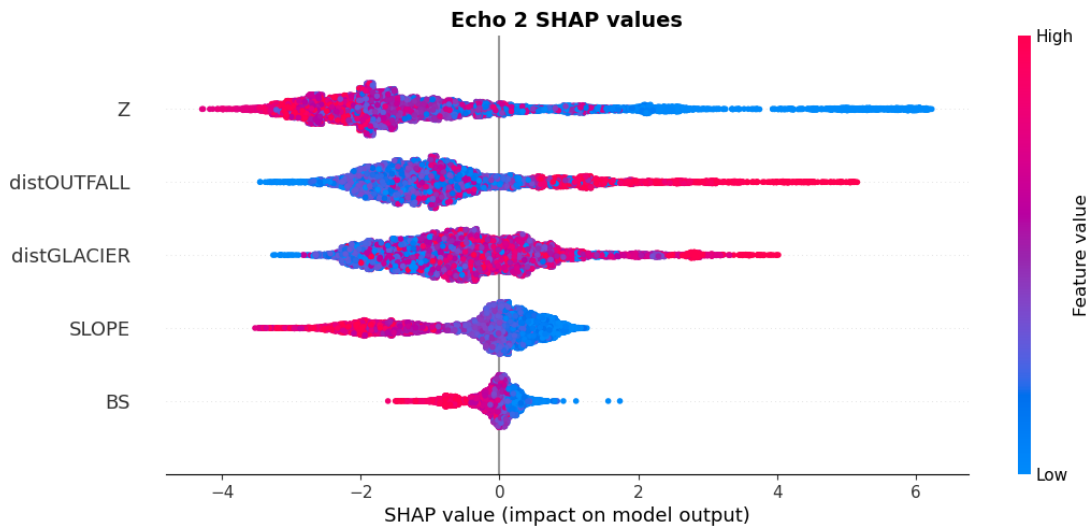


Figure 10: the summary plot of Echo 2.

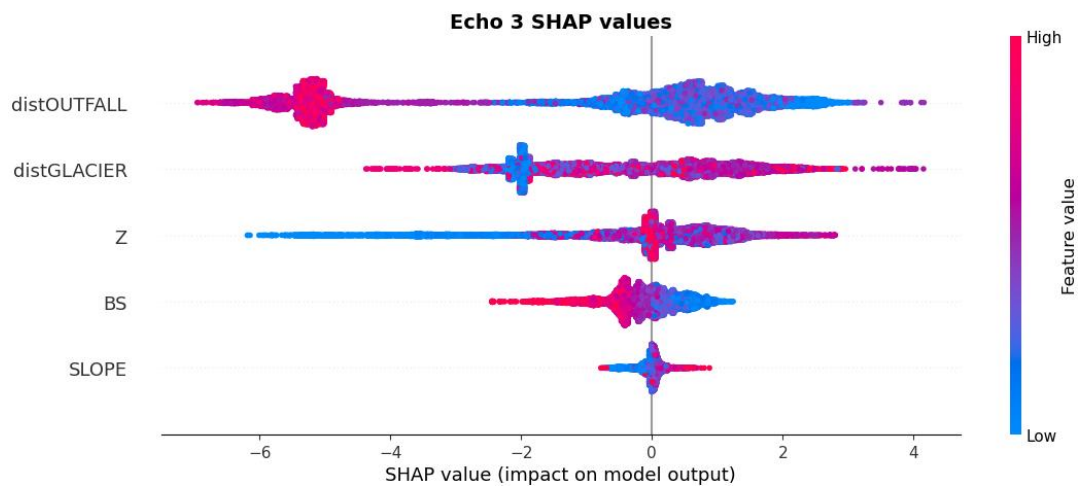


Figure 11: the summary plot of Echo 3.

**Prediction stage: echo-character map**

The model was trained using Dataset1 and predicted occurrences of echo-types for Dataset2. The result was a coverage layer, in black for Echo 1, blue for Echo 2, and Yellow for Echo 3. The validation information was displayed as lines (Figure 12). In a red palette, the uncertainty of the predictions was overlaid. A threshold of 95% of

certainty was established: less than that, prediction was considered unreliable (marked in red). In this sense, it is notable that most of the uncertainty occurs between different echo-types, which can be interpreted as transitional sedimentary settings and/or incorrect human seafloor-classification. The echo-classification from SBP data is done by hand and, therefore, bound to human subjectivity for interpreting seismic images and classifying them in echo-types. In this context, there is an inherent human difficulty in classifying transitional sedimentological settings in echo-types.

This final map integrates predicted data, validation data and predicted confidence threshold, therefore can be considered relevant auxiliary data for seafloor classification. Moreover, inherent SBP limitations would be minimized.

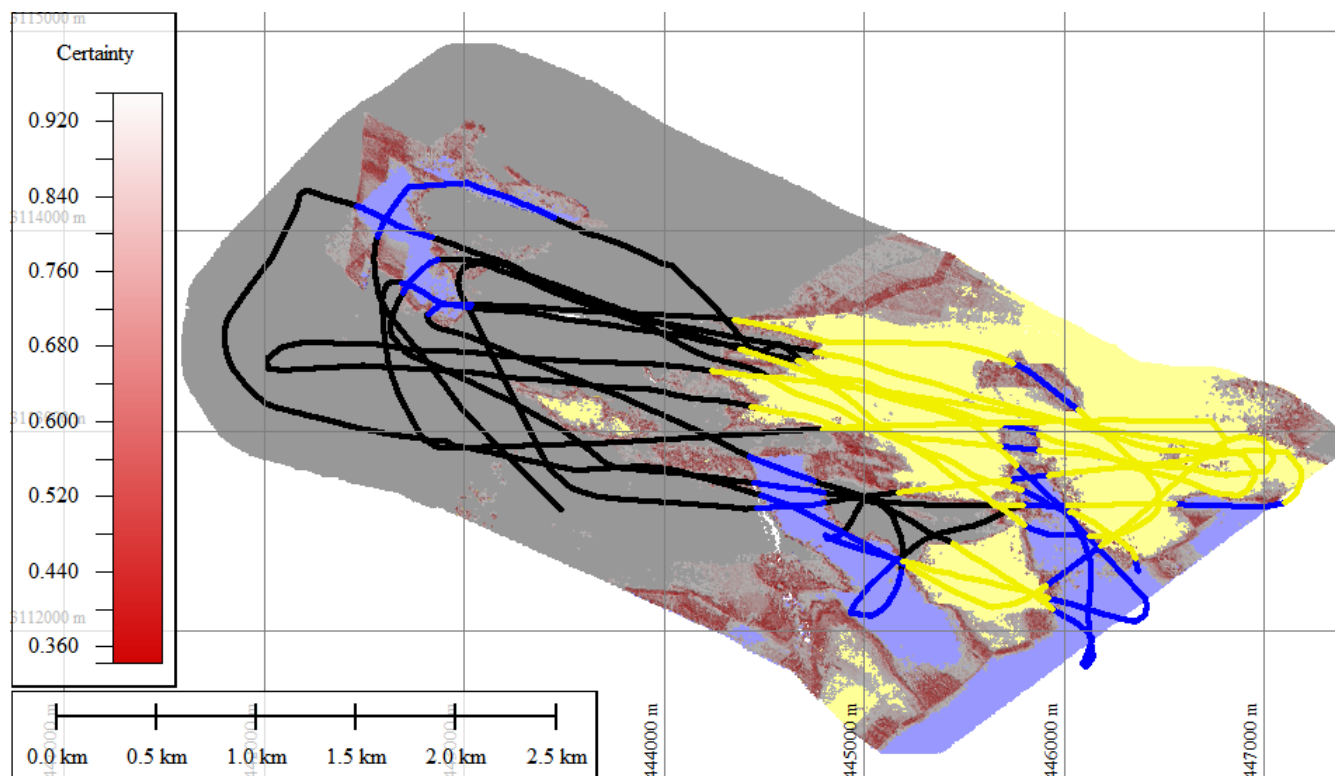


Figure 12: final echo-classification map, showing SBP validation data and uncertainty bias.

## 4 CONCLUSIONS

In conclusion, we presented a reliable workflow for alternative and auxiliary seafloor mapping technique. Regarding our results, four observations can be done:

- Explainability methods were used to analyze the data and understand everything about. Feature importance attribution approaches are – besides SHAP – inconsistent. However, SHAP, ‘weight’ and ‘total\_cover’ lead us to a correct importance rank, using an iterative method as the ground truth. Moreover, SHAP values provided reliable information about how the model makes decisions relative to each predicted class. Using this tool, we were able to measure how each geological, geophysical, and morphological characteristics impact and define echo-types, even according to visual correlations in Figure 1. Although some authors (Anderson et al., 1983; Anderson, 1999; Assine & Vesely, 2008) point the variables that control some phenomenon, IA strategies can show how variables interact in specific environments, quantifying its importance in the definition of the phenomena. In this sense, our research concludes that in the glaciomarine environment observed in King George Bay, distance to the glacier front and to the bay’s outfall are the most important variables regarding echo-types definition, whereas each echo-type had a different setting of control features.

- Predictions will never substitute real data; however, it can provide relevant information when there are lack of data and/or acquisition limitations.

- There is an inherent limitation of our method, according to the nature of features used: backscatter, depth, slope and distances to the glacier and bay’s outfall are relative. Backscatter intensity varies from area to area, and it is often associated with seafloor hardness, and not sediment type – even though in specific areas you can associate hardness to a specific sediment. Thus, the same backscatter intensity range can represent

very distinct seafloor compounds in different areas (Ji et al., 2020). Depth, slope and distances to the glacier and bay's outfall are also exceptional relative to each location. Since each environment is unique, each area needs specific AI modeling regarding similar features. Nonetheless, even though the model developed in our research is specific for the King George Bay glaciomarine environment, our approach shows the necessary procedure to create similar models in any area, if there is griddable data reliable to be correlated.

- Predicting echo-character maps can be challenging because the validation data is human-interpreted by hand. In this sense, the prediction veracity also reflects the ability to correct classify echo-types. In addition, most of the prediction uncertainty occurs between echo classes, in transitory sedimentary settings since the precise limit between different echo-types are the most subjective areas for manual classification.

## 5 References

Anderson, J.B; Brake, C.; Domack, E.; Myers, N.; Wright, R. (1983) - Development of a Polar Glacial-Marine Sedimentation Model from Antarctic Quaternary Deposits and Glaciological Information. IN: Glacial-marine Sedimentation. p. 233–264. Springer US.

Anderson, J. B. (1999). Antarctic marine geology. John B. Anderson. Cambridge: Cambridge University Press, vii + 289p, illustrated, hard cover. ISBN 0-521-59317-4. Polar Record.

37. 163. 10.1017/S0032247400027054.

Anderson, T. W. & Finn, J. D. (1997). *The New Statistical Analysis of Data*. Springer-Verlag, New York. 712 pp.

Andrzej Maćkiewicz, Waldemar Ratajczak (1993). Principal components analysis (PCA). *Computers & Geosciences*, Volume 19, Issue 3, Pages 303-342, ISSN 0098-3004, [https://doi.org/10.1016/0098-3004\(93\)90090-R](https://doi.org/10.1016/0098-3004(93)90090-R).

Arabameri, A., Chandra Pal, S., Costache, R., Saha, A., Rezaie, F., Seyed Danesh, A., Pradhan, B., Lee, S., Hoang, N.-D., 2021. Perdition of gully erosion susceptibility mapping using novel ensemble machine learning algorithms. *Geomatics, Nat. Hazards Risk* 12, 469–498. <https://doi.org/10.1080/19475705.2021.1880977>

Ariza-Garzon, M.J., Arroyo, J., Caparrini, A., Segovia-Vargas, M.J., 2020. Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending. *IEEE Access* 8, 64873–64890. <https://doi.org/10.1109/ACCESS.2020.2984412>

Assine, Mario & Vesely, Fernando. (2008). Ambientes Glaciais. In: PEDREIRA DA SILVA, A.J.; ARAGÃO, A.N.F.; MAGALHÃES A.J.C. (Eds.). *Ambientes de Sedimentação Siliciclástica no Brasil*. Ed. Beca. p. 24-51. São Paulo. 2008.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 13-17-Aug, 785–794. <https://doi.org/10.1145/2939672.2939785>

Coyle, D., Weller, A., 2020. “Explaining” machine learning reveals policy challenges.



Science (80-. ). 368, 1433 LP – 1434. <https://doi.org/10.1126/science.aba9647>

Damuth, J. E., & Hayes, D. E. (1977). ECHO CHARACTER OF THE EAST BRAZILIAN CONTINENTAL MARGIN AND ITS RELATIONSHIP TO SEDIMENTARY PROCESSES. *Marine Geology*. Vol. 24(2). p. 73-96. 1977.

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., Scholten, T., 2020. Predicting and mapping of soil organic carbon using machine learning algorithms in northern iran. *arXiv*.

Fengfan Wang, Jia Yu, Zhijie Liu, Min Kong, Yunfan Wu. Study on offshore seabed sediment classification based on particle size parameters using XGBoost algorithm. *Computers & Geosciences*, Volume 149, 2021, 104713, ISSN 0098-3004, <https://doi.org/10.1016/j.cageo.2021.104713>.

Hellequin, L., Boucher, J.M., Lurton, X., 2003. Processing of high-frequency multibeam echo sounder data for seafloor characterization. *IEEE J. Ocean. Eng.* 28, 78–89. <https://doi.org/10.1109/JOE.2002.808205>

Ji, X., Yang, B., Tang, Q., 2020. Seabed sediment classification using multibeam backscatter data based on the selecting optimal random forest model. *Appl. Acoust.* 167, 107387. <https://doi.org/10.1016/j.apacoust.2020.107387>

Jianglin Huang, Yan-Fu Li, Min Xie, (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and Software Technology*. Volume 67, pages 108-127, ISSN 0950-5849, <https://doi.org/10.1016/j.infsof.2015.07.004>.

Jolliffe, I.T., Cadima, J., Cadima, J., 2016. Principal component analysis : a review and recent developments Subject Areas : Author for correspondence :

Leitão, F.J., 2015. Caracterização Morfológica e Sedimentar a Partir de Dados de Multifeixe na Baía Foster , Ilha Deception , Antártica Caracterização Morfológica e Sedimentar a Partir de Dados de Multifeixe na Baía Foster , Ilha Deception , Antártica.

Leitão, F.J., Ayres Neto, A., Vieira, R., 2016. Morphological and sedimentary characterization through analysis of multibeam data at deception Island, Antarctic. Rev. Bras. Geofis. 34, 1–10. <https://doi.org/10.22564/rbgf.v34i2.792>

Li, X., Luo, J., Jin, X., He, Q., Niu, Y., 2020. Improving soil thickness estimations based on multiple environmental variables with stacking ensemble methods. Remote Sens. 12, 1–21. <https://doi.org/10.3390/rs12213609>

Lundberg, S.M., Lee, S.I., 2017a. Consistent feature attribution for tree ensembles. arXiv.

Lundberg, S.M., Lee, S.I., 2017b. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 2017-Decem, 4766–4775.

Lundberg, S.M., Erion, G., Chen, H. et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2, 56–67 (2020). <https://doi.org/10.1038/s42256-019-0138-9>

Magrani, F.J.G., 2014. CARACTERIZAÇÃO SEDIMENTAR GLACIOMARINHA DA DEGLACIAÇÃO DA BAÍA DO ALMIRANTADO DESDE O ÚLTIMO

MÁXIMO GLACIAL, ARQUIPÉLAGO DAS SHETLAND DO SUL,  
ANTÁRTICA.

Marsh, I., Brown, C., 2009. Neural network classification of multibeam backscatter and bathymetry data from Stanton Bank (Area IV). *Appl. Acoust.* 70, 1269–1276.  
<https://doi.org/10.1016/j.apacoust.2008.07.012>

Mosavi, A., Hosseini, F.S., Choubin, B., Goodarzi, M., Dineva, A.A., 2020. Groundwater Salinity Susceptibility Mapping Using Classifier Ensemble and Bayesian Machine Learning Models. *IEEE Access* 8, 145564–145576.  
<https://doi.org/10.1109/ACCESS.2020.3014908>

Raihan-Al-Masud M, Mondal MRH (2020). Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. *PLOS ONE* 15(2): e0228422. <https://doi.org/10.1371/journal.pone.0228422>

Roscher, R., Bohn, B., Duarte, M.F., Garcke, J., 2020. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access* 8, 42200–42216.  
<https://doi.org/10.1109/ACCESS.2020.2976199>

Sahin, E.K., 2020. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Appl. Sci.* 2, 1–17. <https://doi.org/10.1007/s42452-020-3060-1>

Saleh, M., Rabah, M., 2016. Seabed sub-bottom sediment classification using parametric sub-bottom profiler. *NRIAG J. Astron. Geophys.* 5, 87–95.  
<https://doi.org/10.1016/j.nrjag.2016.01.004>

Shendryk, Y., Rossiter-Rachor, N.A., Setterfield, S.A., Levick, S.R., 2020. Leveraging

High-Resolution Satellite Imagery and Gradient Boosting for Invasive Weed Mapping. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 13, 4443–4450.  
<https://doi.org/10.1109/JSTARS.2020.3013663>

Shlens, J., View, M., Introduction, I., 2014. A Tutorial on Principal Component Analysis.

Viana, F., 2014. Caracterização sedimentar glaciomarinha da deglaciação da baía do almirantado desde o último máximo glacial, arquipélago das shetland do sul, antártica.

Wang, F., Yu, J., Liu, Z., Kong, M., Wu, Y., 2021. Study on offshore seabed sediment classification based on particle size parameters using XGBoost algorithm. *Comput. Geosci.* 149, 104713. <https://doi.org/10.1016/j.cageo.2021.104713>

Xinghua, Z., Yongqi, C., 2004. Seafloor sediment classification based on multibeam sonar data. *Geo-spatial Inf. Sci.* 7, 290–296. <https://doi.org/10.1007/BF02828555>

Zhou, P., Chen, G., Wang, M., Chen, J., Li, Y., 2020. Sediment classification of acoustic backscatter image based on stacked denoising autoencoder and modified extreme learning machine. *Remote Sens.* 12, 1–18.  
<https://doi.org/10.3390/rs12223762>

#### **Documentation references:**

\*1 <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>  
sci-kit-learn 0.24

\*2 [http://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.train\\_test\\_split.html](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html)  
sci-kit-learn 0.24

\*3 [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced\\_accuracy\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html)  
sci-kit-learn 0.24

\*4 <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html>

Pandas 1.3.0

\*5 <https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>  
sci-kit-learn 0.24

\*6 [https://xgboost.readthedocs.io/en/latest/python/python\\_api.html](https://xgboost.readthedocs.io/en/latest/python/python_api.html)

Python 3.7

<https://github.com/dmlc/xgboost/blob/master/python-package/xgboost/core.py#L953>[[1]

xgboost 1.5.0

\*7 [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.SelectKBest.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html)

sci-kit-learn 0.24

\*8 <https://shap.readthedocs.io/en/latest>

# XGBoost as a Tool to Improve High-Resolution Seismic Interpretation (\*\*\*)

Diogo Ceddia Porto Silva 1<sup>a</sup>, Reinaldo Mozart da Gama e Silva 2<sup>b</sup> and Arthur Ayres Neto 3<sup>c</sup>

<sup>a</sup> Marine Geology Laboratory (LAGEMAR), Department of Geology and Geophysics, Federal Fluminense University, Rio de Janeiro, Brazil, ORCID(s): 0000-0002-6278-2250

<sup>b</sup> Independent Consultant.

<sup>c</sup> Marine Geology Laboratory (LAGEMAR), Department of Geology and Geophysics, Federal Fluminense University, Rio de Janeiro, Brazil, ORCID(s): 0000-0002-2982-245X

**(\*\*\*) Submitted In *Computers & Geosciences Journal*.**

## **Abstract**

The interpretation of high-resolution seismic data can be a difficult task. Transitional echo characters, anisotropy, data acquisition problems that interferes in seismic image, large volume of data, lack of human resources, short deadlines, and urgency of preliminary results to support decision making are often a reality that puts pressure on the specialist to provide a timely and quality interpretation. Considering this scenario, we used artificial intelligence approaches to prove that: (a) correct modeling can fulfill gaps in interpretation – when lines are difficult or tricky to interpret and the specialist leave them uninterpreted or can not classify it with certainty; and (b) it is possible to predict echo-types of seismic lines which were not human-interpreted by hand, providing quick and preliminary result to support decision making. Our results showed that XGBoost algorithm needed randomly 1% to predict 99% of the dataset with 90% of balanced accuracy (BA) (Brodersen et al., 2010), and 4% to predict 96% of the dataset with 95% of BA. This demonstrates that the expert could let the algorithm fulfill

uninterpreted gaps with confidence, solving the first problem. In addition, we reduce the dimensionality of our dataset removing seismic lines entirely, reaching up to 82.7% of BA with only 4 out of the 51 lines available, whose corresponds to 10.7% of total dataset – as so being able to achieve the second goal (2).

**Keywords: High-resolution seismic; sub-bottom profiler; machine learning.**

## **Introduction**

The interest in mapping sub-structures of the seabed is present in many kinds of marine activities (marine and geological mapping, engineering works, cable and pipeline maintenance and installation, offshore windfarms, dredging operations), which require detailed information about seafloor composition, topography, and inner sedimentary settings. The traditional solution is to collect physical sediment samples. However, for large areas, this procedure is expensive and time consuming (Hellequin et al., 2003; Saleh and Rabah, 2016). An alternative approach is to use seismic remote sense techniques, which can acquire faster and bigger volume of data with smaller relative budget. In this context, Sub-Bottom Profilers (SBP) are well designed: usually installed under a ship's hull, it transmits a single-channel high-frequency pulse into the seafloor, revealing physical properties of target's surface and sub-surface. Each pulse (or ping) penetrates up to several tens of meters (depending on the kind of sediment), reaching different settings of layers. When the pulse reaches a layer, it will reflect and refract, depending on its acoustic impedance ( $Z$ ) – the 'resistance' that some material imposes on the propagation of the wave (Tang et al., 2005; Xinghua and Yongqi, 2004). Considering that, the echo-signal is registered measuring the elapsed time between the transmission and reception of

the acoustic energy when it encounters boundaries of different sediment layers (Müller and Wunderlich, 2003; Saleh and Rabah, 2016; Zhao and Liu, 2020).

After data acquisition, it is necessary to analyze, interpret and classify seismic facies. Since 1950's the SBP technique is used to study marine geology (Heezen et al., 1959), but only in 1970's J. E. Damuth proposed the first classifications of acoustic response (Damuth and Hayes, 1977), which are still used as base-line by contemporary researchers. Since then, many tested the technique and classification method to characterize and map the seafloor, according to Maestro et al. (2018). Moreover, one of the most popular ways today to visualize and interpret SBP data is using the Reflection Strength attribute, also known as envelope or instantaneous amplitude ( $a(t)$ ). It is defined as  $a(t) = |z(t)| = \sqrt{z_r^2(t) + z_i^2(t)}$ , where  $z$  is the analitic signal expressed in complex plane. This attribute consists of the absolute sum of both polarities, highlighting important seismic features by total amplitude (Cunha, 2020; Koson et al., 2014).

High-resolution seismic data is, since J. E. Damuth, commonly interpreted by hand and, therefore, bound to human subjectivity. In this context, echo-classification of seismic facies can usually have considerable variations from specialist to specialist, due to transitional echo characters, data acquisition problems and artifacts (non-geological-topographic features). Another problem is that some lines can be tricky and ambiguous to interpret, commonly caused by anisotropy. All these issues together can be, from the specialist's perspective, interpretations with some level of uncertainty. Considering this situation, our approach shows the power of ML to predict echo-types, creating the possibility of statistical-based auxiliary information that helps echo interpretation.

## **Methodology**



## The dataset

The dataset used in this article is from Admiralty Bay, South Shetland Islands, Antarctica. It is a result of the interpretation of 51 seismic lines, containing the following features (Figure 1):

- (1) X, relative to longitudinal south UTM coordinates from EPSG 32721, datum WGS 84, extracted from each seismic trace.
- (2) Y, relative to latitudinal south UTM coordinates from EPSG 32721, datum WGS 84, extracted from each seismic trace.
- (3) ECO, relative to the classes from man-made interpretation of seismic echo-types, named as Echo 1, 2 and 3.
- (4) Z, relative to the depth in meters extracted from a bathymetric DTM.
- (5) BS, relative to the backscatter intensity in decibels of the bathymetric signal.
- (6) SLOPE, relative to the seabed slope, in degrees, of the bathymetric DTM.
- (7) distOUTFALL, relative to the distance in meters between a given grid point and the Admiralty Bay's outfall (delimited by a line connecting the two most extreme points of the bay, for each side).
- (8) distGLACIER, relative to the distance in meters between a given grid point and the Admiralty Bay's margin/ glacier front.
- (9) ASPECT, relative to the slope direction in degrees of the bathymetric DTM.
- (10) Line, relative to the name of the seismic line which a given coordinate belongs.

The (4), (5), (6) and (9) features are relative to the same bathymetric DTM, overlaid in the seismic lines. The DTM and the seismic lines were correlated based on the nearest neighbor, according to a threshold established as two times the bathymetric grid space – in this case, 8 meters. The algorithm used to achieve this was the K-D-Tree, a

binary search tree that can support  $k$  values for each node (Bentley, 1975). It was the best choice considering computational time, when compared to brute force Euclidean distance algorithms, whose computational complexity growth is  $n^2$ , while K-D-Tree is  $n$  – considering that  $n$  equals to the input data size. The (7) and (8) features were computed using Euclidean distance.

Admiralty Bay Dataset										
X	Y	ECO	Z	BS	SLOPE	distOUTFALL	distGLACIER	ASPECT	line	
429900.33	3099573.04	2.00	-489.69	-32.00	14.05	2656.25	4058.12	340.00	Alm - L2b.sgy	
431343.47	3097163.57	3.00	-593.14	-30.00	3.63	5408.21	6473.68	22.00	Alm - L2b.sgy.001	
424919.96	3110278.89	1.00	-425.24	-28.00	8.45	9073.63	2015.43	103.00	OP32 - Almirantado 3	
429296.31	3105660.51	3.00	-394.37	-28.00	4.36	3203.25	2195.83	313.00	OP32 - Almirantado 3	
425873.87	3108882.88	1.00	-423.19	-27.00	44.57	7437.34	2315.26	190.00	OP32 - Almirantado 3	
...	...	...	...	...	...	...	...	...	...	
										(69905, 10)

Figure 1: section of the Admiralty Bay dataset, showing its features and size (69905, 10).

### Data manipulation

We did two types of data dimensionality reduction: (a) by randomly dropping rows in the dataset, affecting all seismic lines equally; and (b) by dropping entire seismic lines, selecting the best lines to input in the machine learning model (XGBoost (Chen and Guestrin, 2016)). Solution (a) is meant to prove that interpretation gaps can be easily solved; and solution (b) proves that the expert do not need to human-interpret by hand all seismic lines to offer a quick preliminary result of the entire data classification.

Methodology (a) was achieved splitting randomly the data in stratified proportional samples (according to Echo 1, 2 and 3). We did 0.1% for training and 99.9% for testing; then 0.2% for training and 99.8% for testing; and so on, until 99.9% for training and 0.1% for testing (Figure 2).

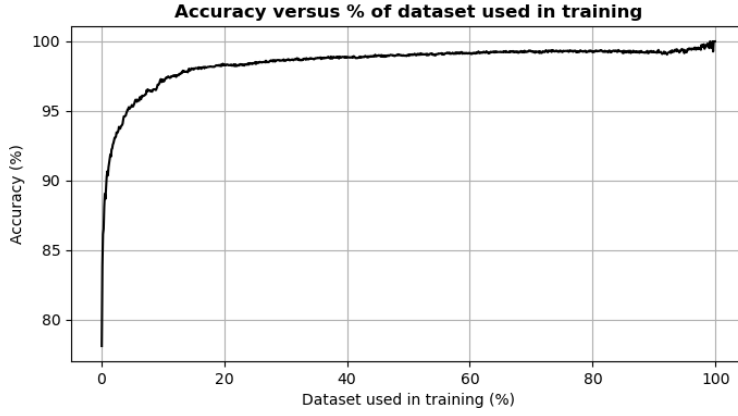


Figure 2: relationship between BA gains with dataset size used in training for (a) methodology.

For methodology (b), it was necessary to pick  $n$  lines and test its BA. All possible combinations of 51 seismic lines are:

$$\sum_{p=0}^n C_{n,p} = \sum_{p=0}^n \left[ \frac{n!}{(p! * (n-p)!)} \right] = C_{51,1} + C_{51,2} + \dots + C_{51,51} \approx 2.25 * 10^{15}.$$

For each combination, we test the lines into the model and measure its BA. Since fitting and predicting a model  $2.25 * 10^{15}$  times is impracticable, we chose a different approach. At first, we trained and evaluated the model with each line individually, using the remaining lines as test data for BA metrics (one for training and 50 for testing). In this step, we fit and test 51 times. Then, we picked the best line – which had higher BA predicting the others – and selected it. In the next step, we trained the model with two lines: the selected one and another random line, in order to get the pair with higher BA. So, in this step, we fitted and tested 50 times, measured the highest BA, and selected the second-best line. As this strategy goes on, it requires to fit and test  $\sum_{n=1}^{51} n = 1275$  times to define the best  $n$  combinations of lines. Clearly, our solution does not show the optimal best combinations of  $n$  lines. However, it is an easy and efficient way to overcome an

expensive computational cost, since fitting and predicting 1275 times is way more viable (Figure 3).

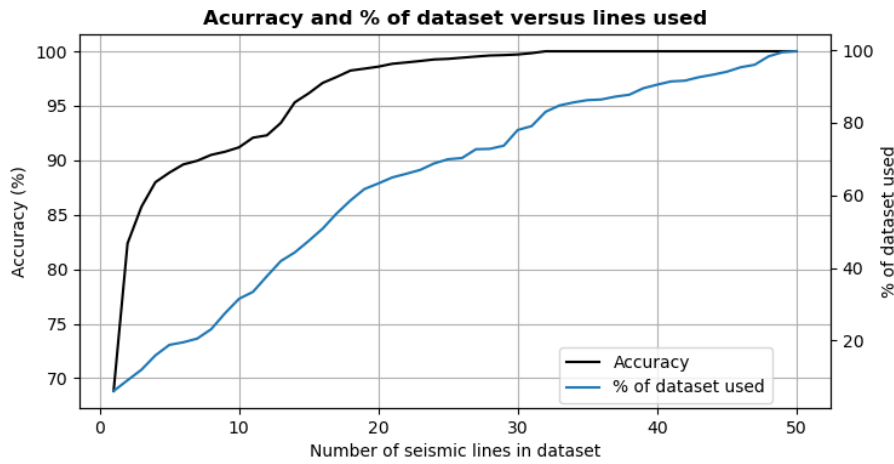


Figure 3: Relationship between BA gain with number of seismic lines used to train the model (black line). Since each line has a unique size, we plot as blue line the dataset percentage relative to the lines used.

## Results

For the methodology (a) – which goal was to greatly decrease the training size – we found that less than 1% of the dataset is necessary to train the model with outcomes greater than 90% of BA. For some thresholds, see Table 1.

Table 1: train and test size, in percentage (%) and absolute values (abs), needed to achieve common thresholds. The absolute values are the number of seismic traces needed.

BA (%)	train size (%)	test size (%)	train size (abs)	test size (abs)
86.14	0.3	99.7	210	69695
90.26	0.9	99.1	629	69276
94.95	4.2	95.8	2936	66969
98.04	14.3	85.7	9996	59909
99.01	45.2	54.8	31597	38308
Dataset size = 69905				

For methodology (b) – which goal was to find the best combinations of  $n$  lines – we found that few lines could predict the remaining data with BA greater than 85% (Table 2).

Table 2: the BA and train size (%) achieved for the best combination of  $n$  seismic lines.

$n$	BA (%)	training size (%)	$n$	BA (%)	training size (%)	$n$	BA (%)	training size (%)	$n$	BA (%)	training size (%)
1	68.82	6.09	9	90.78	27.55	17	97.64	54.99	25	99.3	69.96
2	82.37	9.07	10	91.18	31.5	18	98.23	58.62	26	99.41	70.28
3	85.74	11.97	11	92.07	33.41	19	98.4	61.75	27	99.51	72.71
4	87.98	15.92	12	92.28	37.73	20	98.59	63.29	28	99.61	72.82
5	88.86	18.81	13	93.42	41.9	21	98.85	64.94	29	99.65	73.68
6	89.62	19.52	14	95.31	44.28	22	98.97	65.94	30	99.71	78.01
7	89.95	20.54	15	96.14	47.46	23	99.11	67.02	31	99.82	79.09
8	90.49	23.13	16	97.09	50.83	24	99.25	68.8	32	100	83.02

For each combination of  $n$  lines, we decide to visualize them in Admiralty Bay to check for any spatial pattern (Figure 4). The results suggested that lines computationally selected to train the model try to reflect most anisotropy and variability as possible among the area.

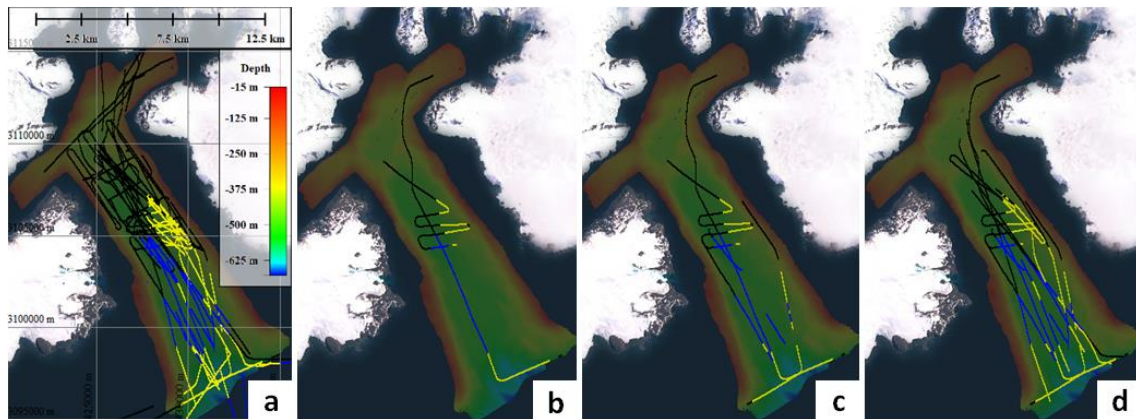


Figure 4: seismic lines displayed above bathymetry surface, aiming spatial comprehension of methodology (b) results. Black, blue, and yellow lines are, respectively, Echo 1, 2 and 3, the seismic classes previously classified. In *a*, all data available; *b* shows  $n = 3$  lines, which predicted the other 48 with 85.74% of BA; *c* shows

$n = 7$  lines, which predicted the other 44 with 89.95% of BA;  $d$  shows  $n = 14$  lines, which predicted the other 37 with 95.31% of BA.

## **Conclusion**

Considering the high BA reached in our approach (>90% with <1% stratified proportional sample of the dataset in methodology (a); and >85% for using at least 3 out of 51 lines to predict the rest in methodology (b)), we conclude that machine learning algorithms (as XGBoost) can be used as an auxiliary tool for seafloor interpretation. This is said assuming that echo-types are related with other spatial data, all represented as overlaid grids.

We can summarize stating that:

- Machine learning algorithms can predict missing data. It can predict entire lines or small gaps with high BA.
- It is possible to choose some spatially distributed lines, interpret and input them as training in a machine learning algorithm. Then, it can be used to predict the other lines, offering a preliminary result of the interpretation. The output information is meant to provide a preliminary result and suggest, for the specialist, the class distribution among study area.

## **References**

Bentley, J.L., 1975. Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM* 18, 509–517.

- Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. Proc. 20th Int. Conf. Pattern Recognit. 3121–24.
- Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. <https://doi.org/10.1145/2939672.2939785>
- Cunha, J.B.F.A., 2020. APRIMORAMENTO DOS DADOS DE SÍSMICA DE ALTA-RESOLUÇÃO DA BACIA DE BRANSFIELD (ANTÁRTICA) UTILIZANDO APRENDIZADO DE MÁQUINA PROFUNDO (DEEP- LEARNING).
- Damuth, J.E., Hayes, D.E., 1977. ECHO CHARACTER OF THE EAST BRAZILIAN CONTINENTAL MARGIN AND ITS RELATIONSHIP TO SEDIMENTARY PROCESSES.
- Heezen, B.C., Tharp, M., Ewing, M., 1959. The Floors of the Oceans: I. The North Atlantic, in: Heezen, B.C., Tharp, M., Ewing, M. (Eds.), The Floors of the Oceans: I. The North Atlantic. Geological Society of America, p. 0. <https://doi.org/10.1130/SPE65-p1>
- Hellequin, L., Boucher, J.M., Lurton, X., 2003. Processing of high-frequency multibeam echo sounder data for seafloor characterization. IEEE J. Ocean. Eng. 28, 78–89. <https://doi.org/10.1109/JOE.2002.808205>
- Koson, S., Chenrai, P., Choowong, M., 2014. Seismic Attributes and Their Applications in Seismic Geomorphology. Bull. Earth Sci. Thail. 6, 1–9.
- Maestro, A., Jané, G., Fernández-Saéz, F., Llave, E., Bohoyo, F., Navas, J., Mink, S., Gómez-Ballesteros, M., Martín-Dávila, J., Catalán, M., 2018. Echo-character of the NW iberian continental margin and the adjacent abyssal plains. J. Maps 14, 56–

67. <https://doi.org/10.1080/17445647.2018.1424653>

Müller, S., Wunderlich, J., 2003. Detection of embedded objects using parametric sub-bottom profilers. *Int. Hydrogr. Rev.* 4, 76.

Saleh, M., Rabah, M., 2016. Seabed sub-bottom sediment classification using parametric sub-bottom profiler. *NRIAG J. Astron. Geophys.* 5, 87–95.  
<https://doi.org/10.1016/j.nrjag.2016.01.004>

Tang, Q.H., Zhou, X.H., Liu, Z.C., Du, D.W., 2005. Processing multibeam backscatter data. *Mar. Geod.* 28, 251–258. <https://doi.org/10.1080/01490410500204595>

Xinghua, Z., Yongqi, C., 2004. Seafloor sediment classification based on multibeam sonar data. *Geo-spatial Inf. Sci.* 7, 290–296. <https://doi.org/10.1007/BF02828555>

Zhao, D., Liu, Z., 2020. Chapter 4 Side-scan Sonar and Sub-bottom Profiler Surveying.



## 5. Conclusão

A dissertação apresenta uma linha lógica, inicialmente fazendo a revisão da aplicação dos métodos geofísicos na Baía Rei George (**1º Artigo: *A Comparison of Different Acoustic Methods for Sedimentary Classification of King George Bay, Antarctica***) e mostrando a independência das variáveis profundidade, backscatter, declividade de fundo, e amplitude sísmica. As metodologias utilizadas para compreender o fundo marinho mostraram-se majoritariamente independentes entre si, consoante os diferentes princípios de operação e propriedades físicas de cada método (frequência etc.). Esse conhecimento favorece a utilização de abordagens de inteligência artificial em geral, como aplicação do modelo XGBoost (escolhido nessa dissertação). A performance do modelo apresentou acurácias balanceadas superiores a 99%, e serviu como base para os subsequentes **2º Artigo: *Machine Learning Modeling Applied for Seabed Echo-Characterization in King George Bay, Antarctica*** e **3º Artigo: *XGBoost as a Tool to Improve High Resolution Single Channel Seismic Interpretation***. No 2º Artigo, o modelo foi utilizado para extrapolação da ecocharacterização de dados dispostos em linhas (ecocaracterização sísmica) para superfícies inteiras – coincidindo com a cobertura batimétrica da área. O produto gerado forneceu, além do mapa de ecocaráteres, um registro de confiabilidade da predição. Notou-se que os trechos de menor confiabilidade do mapa correspondem às regiões transicionais entre os distintos ecos. Esse fenômeno é justificado pelas incertezas e subjetividades da ecocharacterização feita à mão pelo especialista, que frequentemente possui dificuldades em delimitar regiões transicionais de domínios sedimentares distintos, traduzidos em eco respostas diferentes. Além disso, os domínios das classes para cada baía e entre Datasets1 e 2, pela ótica do modelo estatístico, são muito próximos entre si. O 3º Artigo objetivou evidenciar que a alta

acurácia balanceada é alcançada não somente retirando dados aleatoriamente do dataset, mas sim também retirando linhas sísmicas por completo. Esse resultado mostrou o poder das ferramentas de aprendizado para auxiliar o especialista, como: usar predição para gerar resultados preliminares acerca da distribuição de classes de ecocaráteres no dado; e usar a predição para prever trechos que haja incerteza da interpretação pelo especialista. É entendido que as incertezas são causadas por ecocaráteres transicionais, problemas na aquisição do dado, anisotropia, artefatos etc.

Em síntese geral, XGBoost mostrou ser uma ferramenta poderosa capaz de solucionar problemas reais. A maior contribuição dessa dissertação é a concepção de criar superfícies de ecocaráteres de forma simples, integrando um fator de confiabilidade na predição. Em uma visão mais ampla, o presente trabalho mostra os passos para a construção de superfícies de atributos de distribuição espacial limitada, seja por causa de uma limitação de aquisição/logística ou financeira. A segunda maior contribuição é evidenciar a habilidade do XGBoost em realizar predições em trechos duvidosos de interpretar no dado sísmico. Isso pode ser implementado em softwares de interpretação sísmica, com duas vantagens: servir de base a auxiliar o especialista; e realizar a interpretação via predição de trechos duvidosos do dado. Dessa forma, seria eximido do especialista “chutar” a classificação sísmica onde não há suficiente certeza do próprio, dando espaço para uma alternativa de classificação baseada em relações estatísticas.

## **6. Códigos**

Todos os códigos desenvolvidos nessa dissertação estão disponíveis no endereço <https://github.com/dceddiaps?tab=repositories>. Caso hajam quaisquer questionamentos ou problemas de disponibilidade/transparência, contatar [diogoceddia@id.uff.br](mailto:diogoceddia@id.uff.br).

## 7. Referências Bibliográficas<sup>9</sup>

- Damuth, J. E., & Hayes, D. E. (1977). *ECHO CHARACTER OF THE EAST BRAZILIAN CONTINENTAL MARGIN AND ITS RELATIONSHIP TO SEDIMENTARY PROCESSES*.
- Duarte, C. M., Agusti, S., Barbier, E., Britten, G. L., Castilla, J. C., Gattuso, J.-P., Fulweiler, R. W., Hughes, T. P., Knowlton, N., Lovelock, C. E., Lotze, H. K., Predragovic, M., Poloczanska, E., Roberts, C., & Worm, B. (2020). Rebuilding marine life. *Nature*, 580(7801), 39–51. <https://doi.org/10.1038/s41586-020-2146-7>
- Heezen, B. C., Tharp, M., & Ewing, M. (1959). The Floors of the Oceans: I. The North Atlantic. In B. C. Heezen, M. Tharp, & M. Ewing (Eds.), *The Floors of the Oceans: I. The North Atlantic* (Vol. 65, p. 0). Geological Society of America. <https://doi.org/10.1130/SPE65-p1>
- Leon, A. Z., Huvenne, V. A. I., Benoist, N. M. A., Ferguson, M., Bett, B. J., & Wynn, R. B. (2020). Assessing the Repeatability of Automated Seafloor Classification Algorithms, with Application in Marine Protected Area Monitoring. In *Remote Sensing*.
- Menandro, P. S., & Bastos, A. C. (2020). Seabed mapping: A brief history from meaningful words. *Geosciences (Switzerland)*, 10(7), 1–17. <https://doi.org/10.3390/geosciences10070273>
- OHI. (2005). MEDIÇÃO DA PROFUNDIDADE. In *Manual de Hidrografia, pub. C-13* (pp. 115–190).

---

<sup>9</sup> Não incluídas as referências do corpo dos artigos. Essa seção contém somente as referências bibliográficas do corpo da dissertação.