RAFAEL GOMES SIQUEIRA

APRENDIZADO DE MÁQUINAS APLICADO AO MAPEAMENTO DIGITAL DE SOLOS NA ANTÁRTICA: MODELAGEM E PREDIÇÃO ESPACIAL DE CLASSES E ATRIBUTOS DO SOLO

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Solos e Nutrição de Plantas, para obtenção do título de *Doctor Scientiae*.

Orientador: Elpídio Inácio Fernandes Filho

Coorientadores: Carlos Ernesto G. R. Schaefer
José João Lélis Leal de Souza

VIÇOSA - MINAS GERAIS
2023

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
S618a
2023

Siqueira, Rafael Gomes, 1994-

Aprendizado de máquinas aplicado ao mapeamento digital de solos na Antártica: modelagem e predição espacial de classes e atributos do solo / Rafael Gomes Siqueira. – Viçosa, MG, 2023.

1 tese eletrônica (181 f.): il. (algumas color.).

Inclui apêndices.

Orientador: Elpídio Inácio Fernandes Filho.

Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Solos, 2023.

Inclui bibliografia.

DOI: https://doi.org/10.47328/ufvbbt.2023.187

Modo de acesso: World Wide Web.

1. Mapeamento do solo - Antártica, Península. 2. Aprendizado do computador. 3. Mapeamento digital. I. Fernandes Filho, Elpídio Inácio, 1963-. II. Universidade Federal de Viçosa. Departamento de Solos. Programa de Pós-Graduação em Solos e Nutrição de Plantas. III. Título.

CDD 22. ed. 631.47989

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

RAFAEL GOMES SIQUEIRA

APRENDIZADO DE MÁQUINAS APLICADO AO MAPEAMENTO DIGITAL DE SOLOS NA ANTÁRTICA: MODELAGEM E PREDIÇÃO ESPACIAL DE CLASSES E ATRIBUTOS DO SOLO

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Solos e Nutrição de Plantas, para obtenção do título de *Doctor Scientiae*.

APROVADA: 10 de março de 2023.

Assentimento:

Rafael Gomes Siqueira
Autor

Elpídio Inácio Fernandes Filho
Orientador

Aos meus.

AGRADECIMENTOS

Aos meus pais e restante da minha família.

Aos meus professores e orientadores.

Aos meus amigos, colegas e à cada um que de alguma maneira fez parte desta jornada prestes a terminar.

À cidade de Viçosa pelo acolhimento, experiências e amizades feitas ao longo de 11 anos.

À Universidade Federal de Viçosa pela formação, aprendizado e oportunidade de completar meu Doutorado.

Ao Departamento de Solos e ao Programa de Pós-Graduação de Solos e Nutrição de Plantas pela oportunidade de estudar e aprender em um centro de excelência.

Aos professores Carlos Schaefer e Márcio Francelino, pela oportunidade única de conhecer e viver a Antártica.

A todos do Labgeo-UFV pelos conhecimentos que tanto me fizeram crescer. Em especial, ao meu orientador Elpídio pelos conselhos, paciência e ensinamentos.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelas bolsas de estudos ao longo da pós-graduação.

Por fim, ao Projeto Terrantar, que em 2022 completou 20 anos de pesquisas antárticas, sendo não apenas responsável pela compilação da segunda maior base de dados de solos do continente, mas também pela pesquisa de excelência que ao longo de duas décadas contribui para a existência do Brasil na Antártica. Viva a ciência brasileira!

“Ever tried. Ever failed. No matter. Try Again. Fail again. Fail better.” (Samuel Beckett)

RESUMO

SIQUEIRA, Rafael Gomes, D.Sc., Universidade Federal de Viçosa, março de 2023. **Aprendizado de Máquinas aplicado ao Mapeamento Digital de Solos na Antártica: modelagem e predição espacial de classes e atributos do solo**. Orientador: Elpídio Inácio Fernandes-Filho. Coorientadores: Carlos Ernesto Gonçalves Reynaud Schaefer e José João Lélis Leal de Souza.

O mapeamento convencional de solos baseado em modelos mentais é uma abordagem comum para compreensão da distribuição dos solos na Antártica. No entanto, o Mapeamento Digital de Solos baseado em modelos empíricos e quantitativos tem sido escassamente aplicado no continente. As vantagens do mapeamento digital são de grande relevância em paisagens remotas como as áreas livres de gelo da Antártica, onde coletas de solos são muito limitadas. Além disso, existe uma necessidade de dados precisos de solos para a Antártica, especialmente sob as pressões impostas atualmente pelas mudanças climáticas e distúrbios antrópicos no continente. Neste trabalho foram preditas as distribuições espaciais, nas regiões da Antártica Marítima e Península Antártica, de importantes atributos físico-químicos do solo: areia, silte, argila, soma de bases, H+Al, pH, carbono orgânico total, P, Na e P remanescente; e de classes de solos identificadas com os sistemas Soil Taxonomy and World Reference Base – FAO. Para isso, foi utilizada a base de dados de solos do Grupo Terrantar da Universidade Federal de Viçosa - Brasil, compilada a partir de trabalhos de campo realizados durante 20 anos na Antártica. Também foi usado um grande conjunto de covariáveis preditoras representando os fatores do modelo scorpan: atributos do terreno gerados a partir de Modelo Digital de Elevação, dados multiespectrais Sentinel representando covariáveis de vegetação, distâncias euclidianas para geleiras, costa e pinguineiras, geologia, coordenadas e outros atributos do solo. Para a modelagem e mapeamento, técnicas de Aprendizado de Máquinas foram aplicadas e diferentes modelos foram testados, destacando-se o modelo Random Forest. Este estudo mostra o potencial da aplicação da metodologia do Mapeamento Digital de Solos para produção de mapas detalhados na Antártica, um continente geralmente negligenciado para mapeamentos devido à ausência de dados de solos robustos. Com nossos resultados, pretendemos subsidiar a tomada de decisão sobre solos na Antártica, além de fornecer dados pioneiros que podem ser incorporados em modelos ecológicos globais.

Palavras-chave: Antártica. Aprendizado de Máquinas. Mapeamento Digital de Solos. Modelagem. Predição espacial.

ABSTRACT

SIQUEIRA, Rafael Gomes, D.Sc., Universidade Federal de Viçosa, March, 2023. **Machine Learning applied for Soil Digital Mapping in Antarctica: modelling and spatial prediction of soil classes and attributes**. Adviser: Elpídio Inácio Fernandes-Filho. Co-advisers: Carlos Ernesto Gonçalves Reynaud Schaefer and José João Lélis Leal de Souza.

Conventional soil mapping based on mental models is a common approach for understanding the soil distribution since the first soil surveys in Antarctica. However, Digital Soil Mapping based on empirical, quantitative models has rarely been applied in the continent. The Digital Soil Mapping advantages are even more important in remote, harsh landscapes such as the Antarctic ice-free areas, where soil information is very limited. Besides that, nowadays there is a pressing need for accurate soil data for Antarctica, especially under the pressures imposed by climate changes and human disturbances on the continent. This study aimed to predict the spatial distribution of major physico-chemical soil attributes (sand, silt, clay, bases sum, H+Al, pH, total organic carbon, phosphorous, sodium and remaining P; and of soil classes according to the Soil Taxonomy and World Reference Base – FAO systems) in Maritime Antarctica and Antarctic Peninsula regions. For that, we used the Brazilian Terrantar Group soil dataset, compiled from several soil surveys carried out during 20 years in Antarctica. We also used a large set of environmental predictor variables representing the scorpan factors: terrain attributes generated from a Digital Elevation Model, multispectral Sentinel data representing vegetation covariates, Euclidean distances to glaciers, coastline and penguin rookeries representing varied factors of soil development in Antarctica, geology, coordinates, and others soil attributes. For modelling and spatial prediction, we applied the Machine Learning techniques by testing different models, but with special attention to the Random Forest. This study shows the potential of applying the Digital Soil Mapping methodology to produce detailed soil maps in Antarctica, a continent generally neglected so far due to the lack of robust soil datasets. With our results, we aim to subsidize decision-making about soils in Antarctica, besides providing pioneer data that can be incorporated in global ecological models.

Keywords: Antarctica. Machine Learning. Digital Soil Mapping. Modelling. Spatial prediction.

SUMÁRIO

INTRODUÇÃO GERAL

A Antártica possui 90% da água doce no mundo e por causa da sua altitude e posição geográfica exerce uma influência importante nos sistemas atmosféricos e criosféricos globais, com destaque para o hemisfério sul (Ugolini e Bockheim, 2008). O continente antártico compreende 13 x 10^6 km² (incluindo plataformas de gelo), dos quais apenas 0,44% são áreas livres de gelo com potencial para desenvolvimento de solos (Brooks et al., 2019). As áreas livres de gelo são localizadas principalmente na Península Antártica e ilhas do entorno, áreas costeiras no perímetro do continente e ao longo de nunataks e vertentes nas Montanhas Transantárticas (Bockheim, 2015).

Os primeiros estudos de solos na Antártica se iniciaram na segunda metade do século XX, incluindo trabalhos pioneiros de mineralogia, microbiologia, influência da fauna, presença de sais e da relação com processos pedogenéticos, principalmente na Antártica Continental (Blakemore e Swindale, 1958; Claridge, 1965; Claridge e Campbell, 1968; Flint e Stout, 1960; Glazovskaya, 1958; Syroechkovsky, 1959; Ugolini, 1973) e de forma menos intensa na Península Antártica (Everett, 1976; Holdgate et al., 1967; O'Brien et al., 1979).

Desde os primeiros estudos algumas das perguntas cruciais respondidas pelos pedólogos antárticos era se o intemperismo químico realmente ocorreria sob as condições climáticas extremas da Antártica e ainda se o material superficial inconsolidado do continente poderia realmente ser chamado de solo (Ugolini e Bockheim, 2008). Descobertas em torno de processos como mobilização de sais e íons nos perfis, oxidação e formação de horizontes mostraram que os solos da Antártica são suficientemente afetados por processos pedogenéticos e respondem aos fatores de formação como os solos de outras partes do planeta (Bockheim e Ugolini, 1990; Campbell e Claridge, 1987; Tedrow e Ugolini, 1966). Em geral, os solos mais desenvolvidos da Antártica têm sido identificados na região da Antártica Marítima, onde o clima mais quente e úmido em comparação com o restante do continente favorece um intemperismo químico mais pronunciado, acumulação de matéria orgânica, além de transformações físico-químicas associadas à fosfatização (Beyer et al., 2000; Blume et al., 2001; Bolter et al., 1997; Tatur e Myrcha, 1984; Tatur e Myrcha, 1989; Simas et al., 2007, 2008).

Visando aumentar o conhecimento acerca da extensão, diversidade e distribuição dos solos antárticos, o mapeamento e a classificação de solos passaram a ser um dos principais objetivos dos estudos de solos no continente (Balks et al., 2014; Ugolini e Bockheim, 2008). O primeiro mapa de solo da Antártica foi produzido por McCraw (1967). Ainda assim, nas primeiras décadas, as pesquisas de solos na Antártica foram focadas principalmente na descrição da gênese dos solos, sem muitas iniciativas para a compreensão da sua distribuição espacial. A partir dos anos 2000, a disponibilização crescente de mapas geológicos e geomorfológicos, das tecnologias SIG, de imagens de satélite de alta resolução e modelos digitais de elevação possibilitou o desenvolvimento de mapas em diferentes escalas que capturam de forma relativamente detalhada as relações solo-paisagem nos ecossistemas antárticos (Balks et al., 2014).

Ainda assim, poucos mapas de solos têm sido produzidos na Antártica, seja na Antártica continental (Bockheim e McLeod, 2008; McLeod et al., 2008; 2009) ou na Península Antártica (Abakumov et al., 2020; Moura et al., 2012; Rodrigues et al., 2019, Schaefer et al., 2017). Os mapas produzidos também fazem parte do que se pode chamar de mapeamento convencional, baseado em modelos mentais de cada pedólogo, por sua vez, enraizados no modelo conceitual de Dokuchaev/Jenny, onde os solos são produto da ação integrada dos diferentes fatores de formação. No entanto, como o mapeamento convencional é altamente dependente do conhecimento e experiência dos pedólogos em todas as etapas, sua natureza qualitativa pode trazer algumas desvantagens como a falta de reprodutibilidade, dificuldades de validação dos mapas e na representação da real variabilidade dos solos (Ma et al., 2019; Minasny e McBratney, 2016). Além disso, levantamentos convencionais são excessivamente demorados e caros para áreas de grande extensão, e este problema é ainda mais acentuado para ambientes extremos como a Antártica.

Desta maneira, neste estudo buscou-se aplicar de forma inédita os princípios do Mapeamento Digital de Solos (como o modelo scorpan de McBratney, 2003) e técnicas de Aprendizado de Máquinas (*Machine Learning*) para a modelagem e predição de propriedades físico-químicas e classes de solos para regiões de grande valor ecológico na Antártica: as Ilhas Shetland do Sul representando a Antártica Marítima e a área livre de gelo Hope Bay e as Ilhas James Ross representando a Península Antártica. Raras iniciativas têm sido tomadas até então para a aplicação do Mapeamento Digital de Solos na Antártica, envolvendo técnicas de Geostatística (Moraes et al., 2017) e Aprendizado de Máquinas (Roudier et al., 2014; Schunemman et al., 2022); porém, nenhuma com a abrangência espacial abordada neste trabalho. Desta maneira, estes serão os primeiros mapas digitais de solos produzidos para o continente em detalhe em escala regional, uma vez que os projetos globais de mapeamentos de solos não englobam o continente antártico devido à escassez de dados.

REFERÊNCIAS

Balks, M.R., López-Martínez, J., Goryachkin, S. V., Mergelov, N.S., Schaefer, C.E.G.R., Simas, F.N.B., Almond, P.C., Claridge, G.G.C., Mcleod, M., Scarrow, J., 2013. Windows on Antarctic soil-landscape relationships: Comparison across selected regions of Antarctica. Geol. Soc. Spec. Publ. 381, 397–410.

Beyer, L., Pingpank, K., Wriedt, G., Bölter, M., 2000. Soil formation in coastal continental Antarctica (Wilkes Land). Geoderma 95, 283–304.

Blakemore, L.C., Swindale, L.D., 1958. Chemistry and clay mineralogy of a soil sample from Antarctica. Nature 182, 47–48.

Blume, H. P., Kuhn, D. & Bolter, M. 2001. Soils and soilscapes. In: Beyer, L., Bolter, M. (eds) Geoecology of Antarctic Ice-Free Coastal Landscapes. Springer, Berlin, 92–113.

Bockheim, J.G., McLeod, M., 2008. Soil distribution in the McMurdo Dry Valleys, Antarctica. Geoderma 144, 43–49.

Bockheim, J.G., Ugolini, F.C., 1990. A review of pedogenic zonation in well-drained soils of the Southern Circumpolar Region. Quat Res 34, 47–66.

Brooks, S.T., Jabour, J., Van Den Hoff, J., Bergstrom, D.M., 2019. Our footprint on Antarctica competes with nature for rare ice-free land. Nat. Sustainab. 2 (3), 185–190.

Campbell, I.B., Claridge, G.G.C., 1987. Antarctica: Soils, Weathering Processes and Environment. Elsevier, Amsterdam. 368 pp.

Claridge, G.G.C., 1965. The clay mineralogy and chemistry of some soils from the Ross Dependency, Antarctica. N.Z. J. Geol. Geophys. 8, 186–220.

Claridge, G.G.C., Campbell, I.B., 1968a. Origin of nitrate deposits. Nature 217, 428–430.

Bockheim, J.G., 2015. The Soils of Antarctica, 1st ed. Springer International Publishing Switzerland, New York.

Bölter, M., Blume, H.P., Schneider, D., Beyer, L., 1997. Soil properties and distributions of invertebrates and bacteria from King George Island (Arctowski Station), maritime Antarctic. Polar Biol 18, 295–304.

Everett, K.R., 1976. A Survey of Soils in the Region of the South Shetland Islands and Adjacent Parts of the Antarctica Peninsula, vol. 58. Ohio State Univ. Inst. Pol. Stud. Rep, p. 44.

Flint, E.H., Stout, J.D., 1960. Microbiology of some soils from Antarctica. Nature 188, 767–708.

Glazovskaia, M.A., 1958. Weathering and primary soil formation in Antarctica. Scientific Paper of the Institute, Moscow Univiv., Faculty of Geography, 1, 63–76.

Holdgate, M. W., Allen, S. E. & Chambers, M. J. G. 1967. A preliminary investigation of the soils of Signy Island, South Orkney. Islands British Antarctic Survey Bulletin, 12, 53–71.

Lupachev, A. V., Abakumov, E. V., Goryachkin, S. V., Veremeeva, A.A., 2020. Soil cover of the Fildes Peninsula (King George Island, West Antarctica). Catena 193, 104613.

Ma, Y., Minasny, B., Malone, B.P., Mcbratney, A.B., 2019. Pedology and digital soil mapping (DSM). Eur J Soil Sci 70, 216–235.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping, Geoderma.

McCraw, J.D., 1967. Soils of Taylor Dry Valley, Victoria Land, Antarctica, with notes on soils from other localities in Victoria Land. N.Z.J. Geol. Geophys. 10, 498–539.

McLeod, M., Bockheim, J. G., Balks, M. R. 2008. Glacial geomorphology, soil development and permafrost features in central-upper Wright Valley, Antarctica. Geoderma, 144, 93–103.

McLeod, M., Bockheim, J., Balks, M. R., Aislabie, J. 2009. Soils of Western Wright Valley, Antarctica. Antarctic Science, 21, 355–365.

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017.

Moraes, A.G.L., Francelino, M.R., Carvalho Junior, W., Pereira, M.G., Thomazini, A., Schaefer, C.E.G.R., 2012. Environmental Correlation and Spatial Autocorrelation of Soil Properties in Keller Peninsula, Maritime Antarctica. Rev Bras Cienc Solo 2017, 41:e0170021.

Moura, P.A., Francelino, M.R., Schaefer, C.E.G.R., Simas, F.N.B., Mendonça, B.A.F. de, 2012. Distribution and characterization of soils and landform relationships in Byers Peninsula, Livingston Island, Maritime Antarctica. Geomorphology 155–156, 45–54.

O'Brien, R.M.G., Romans, J.C.C, Robertson, L (1979) Three soil profiles from Elephant Island, South Shetland Islands. Br Antarct Surv Bull 47, 1–12.

Rodrigues, W.F., Oliveira, F.S., Schaefer, C.E.G.R., Leite, M.G.P., Gauzzi, T., Bockheim, J.G., Putzke, J., 2019. Soil-landscape interplays at Harmony Point, Nelson Island, Maritime Antarctica: Chemistry, mineralogy, and classification. Geomorphology 336, 77–94.

Roudier, P., McLeod, M., Aislabie, J., Morgan, F., 2014. The why and how of digital soil mapping in Antarctica. In: Arrouays, D., McKenzie, N., Hempel, J., Richer-DeForges, A., Mcbratney, A. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System, pp. 149–153.

Schaefer, C. E. G. R., Costa Pereira, T.T., Ker, J.C., Carreiro Almeida, I.C., Bello Simas, F.N., Soares de Oliveira, F., Corrêa, G.R., Vieira, G., 2015. Soils and Landforms at Hope Bay, Antarctic Peninsula: Formation, Classification, Distribution, and Relationships. Soil Sci. Soc. Am. J. 79, 175–184.

Schünemann, A.L., Almeida, P.H.A., Francelino, M.R., Fernandes-Filho, E.I., Santos, G.R., Daher, M., Schaefer, C.E.G.R., Pereira, A.B., 2022. Soil pedogeochemical attributes prediction by interpolators in ice-free areas of Antarctica. Res. Soc. Dev. 11, 1–15.

Simas, F., Schaefer, C.E.G.R., Melo, V.F., Albuquerque-Filho, M.R., Michel, R.F.M., Pereira, V. V., Gomes, M.R.M., Costa, L.M., 2007. Ornithogenic cryosols from Maritime Antarctica: Phosphatization as a soil forming process. Geoderma 138, 191–203.

Simas, F.N.B., Schaefer, C.E.G.R., Albuquerque-Filho, M.R., Francelino, M.R., Fernandes-Filho, E.I., Costa, L.M. da, 2008. Genesis, properties and classification of Cryosols from Admiralty Bay, maritime Antarctica. Geoderma 144, 116–122.

Syroechkovsky, E.E., 1959. The role of animals in the formation of primary soils under the condition of circumpolar regions of the earth (Antarctica). Zoologicheskiĭ Zhurnal 38, 1770–1775.

Tatur, A.E., Myrcha, A., 1984. Ornithogenic soils on King George lsland, South Shetland lslands (Maritime Antarctic Zone). Pol Polar Res 5, 31–60.

Tatur, A., Myrcha, A., 1989. Soils and Vegetation in Abandoned Penguin Rookeries (Maritime Antarctic). Polar Biol 2, 181–189.

Tedrow, J.C.F., Ugolini, F.C., 1966. Antarctic soils. In: Tedrow, J.C.F. (Ed.), Antarctic Soils and Soil Forming Processes. Antarct. Res. Ser, vol. 8. Am. Geophys. Union, Washington, DC, pp. 161–177.

Ugolini, F.C., Anderson, D.M., 1973. Ionic migration and weathering in frozen Antarctic soils. Soil Sci. 115, 461–470.

Ugolini, F.C., Bockheim, J.G., 2008. Antarctic soils and soil formation in a changing environment: A review. Geoderma 144, 1–8.

CAPÍTULO 1

SPATIAL PREDICTION OF SOIL TEXTURE FOR MARITIME ANTARCTICA AND NORTHERN ANTARCTIC PENINSULA

ABSTRACT

Soil texture is one of the most important soil properties as it drives several physical, chemical, biological, hydrological, and mechanical properties, and processes. In Antarctica the soil texture controls different important ecological processes, such as carbon stocks, nutrient leaching, and toxic metals retention. Thus, there is a pressing need for accurate soil data for the Antarctic ice-free areas, especially under the pressures imposed by climate changes and human disturbances on the continent. In this work, we predicted the distribution of sand, silt, and clay of the main ice-free areas of Maritime Antarctica and Northern Antarctic Peninsula, allowing novel soil texture maps to be produced for Antarctica. We used legacy soil texture data and tested different machine learning models, of which Random Forest presented the best performance to the prediction. The best predictors were associated with climate, topography, and soil development degree, representing the scorpan model factors. The highest accuracy was obtained for clay prediction, mainly in the topsoil (LCCC of 0.49 and R^2 of 0.31). In the mapping, the sand contents presented higher values, with a predominance of sandy loam texture, although clay contents are higher in some spots where chemical weathering is stronger. The final maps presented good spatial consistency, with soil texture distributed according to factors such as geomorphology, parent material and pedogenetic development. With our results, we aim to subsidize decision-making about soils in Antarctica, besides providing pioneer data that can be incorporated in global environmental models.

Keywords: Antarctica. Digital Soil Mapping. Random Forest. Scorpan factors. Soil texture. Spatial modelling.

1. INTRODUCTION

Knowledge about soil properties in Antarctic Peninsula and its offshore islands has increased in the last years as a result of the international effort of scientific groups, leading to production of detailed soil maps (Lupachev et al., 2020; Moura et al., 2012; Rodrigues et al., 2019; Schaefer et al., 2015). Despite the advances, the nature of these maps based on discrete representative profiles is limited from the spatial point of a view, since these conventional approaches assume properties of modal profiles applied to the entire mapping unit, not applicable in Antarctica due to its complexity.

Digital Soil Mapping (DSM) has arisen as an alternative to overcome the limitations of the subjective conventional soil mappings. The DSM principle consists in intersecting soil observation points and layers of secondary environmental data (covariates), so that an empirical quantitate model of some structure is fitted to describe the relationships between soil and environment (Minasny and McBratney, 2016). The *scorpan* model as formulated by McBratney et al., 2003 has gained acceptance as the standard approach for mapping soil, where soil classes or attributes are function of soil (other attributes), climate, organisms, relief, parent material, age, and spatial position, then, fitting well to the DSM scope.

With the advent of the data-driven machine learning (ML) algorithms in soil science, which can build predictive models through data patterns without explicit programming, the prediction of soil attributes in the scope of DSM has progressively increased, with special attention to modelling soil classes (Brungard et al., 2015), soil carbon stocks (Gomes et al., 2019; Xiong et al., 2014) and soil texture. Although the common use of GIS techniques and geostatistics (Akumu et al., 2015; Seyedmohammadi et al., 2019), soil texture prediction has been mainly addressed by the ML modelling, with predictions at regional (Carvalho Junior et al., 2014; Román Dobarco et al., 2017), national (Adhikari et al., 2013; Caubet et al., 2019), continental (Ballabio et al., 2016; Hengl et al., 2015) and global scales (Sousa et al., 2020). Although the current global coverage, the Antarctic environments have not been contemplated until now by any DSM program.

Soil texture is defined as the relative proportion of soil particle size fractions and regulates many other physical, chemical, biological, hydrological and mechanical soil attributes (Hossain et al., 2018). It also affects important processes, such as evapotranspiration, water drainage, nutrient transport, pollutant retention and plant root growth (Li et al., 2014; Mills et al., 2006), thus, directly controlling soil quality and being extremely relevant for management related to environmental resources (Mansuy et al., 2014). Soil texture also plays a key role to understand the soil genesis and distribution, being a property able to give estimations about the pedogenetic degree and weathering rates (Koseva et al., 2010). The vertical and horizontal variability of soil texture is controlled by many processes and factors linked to the pedogenesis, such as illuviation, pedoturbation, surface removal, aggradation and parent material inheritance (Phillips, 2004).

Many ecological processes in the Antarctic ice-free areas are highly dependent on soil properties, of which the soil texture is of key importance. The coastal environments of Antarctica are pools of bioavailable nutrients and carbon due to the long-term sea-land transfers carried out by marine birds, with creation of favorable sites for plant colonization (Bockheim, 2015). The soil retention of these elements is directly dependent on the soil texture, mainly clay content. Indirectly, this also influences other important environmental processes, such as the CO_2–C exchange rates (Thomazini et al., 2016). Soil texture also influences the chemistry of the surface and subsurface waters of the Antarctic ice-free areas, controlling the leaching in the areas where the chemical weathering is more pronounced (Lopes et al., 2021). It is also an important factor to decreasing the reaction of toxic metals in soil, reducing the risk of contamination promoted by the anthropogenic action in Antarctica (Lima Neto et al., 2017).

There are various reasons for accurately predicting the spatial variability of properties such as soil texture in the Antarctic ecosystems: 1) to increase the knowledge about weathering and pedogenic processes in Antarctica, 2) to provide information on the distribution of the main Antarctic soil types, 3) to help environmental decision-making in the Antarctic ice-free areas, mainly under environmental pressures imposed by climate changes and increasing human occupation in the continent, 4) to offer inputs for ecological, hydrological and climatic and other environmental models for the Antarctic ecosystems.

Thus, our main objective was to predict the horizontal and vertical distribution of textural fractions (sand, silt, and clay) of the main ice-free areas of Maritime Antarctica and Northern Antarctic Peninsula, creating novel soil texture maps for Antarctica. We also evaluated the predictive performance of a set of ML algorithms for soil texture prediction, as well as the importance of different environmental predictors to modelling soil texture in Antarctic landscapes.

2. MATERIAL AND METHODS

2.1. Study area

Maritime Antarctica (MA) encompasses the South Shetland Islands (SSI), situated northwest of the Antarctic Peninsula. The Northern Antarctic Peninsula (NAP) is represented here by the coastal area of Hope Bay and the James Ross Islands (JRI), in the Weddell Sea Sector (Fig. 1). In SSI, the ice-free areas represent around 9% of the total area and have a maximum altitude next to 400 m a.s.l. On the other hand, they correspond to 23% of the JRI, and altitudes reach 600 m a.s.l. Deglaciation started between 8 and 6 Ky BP in SSI, whereas in JRI it is dated between 7.4 and 4.7 Ky BP (Ingólfsson and Hjort, 2002).

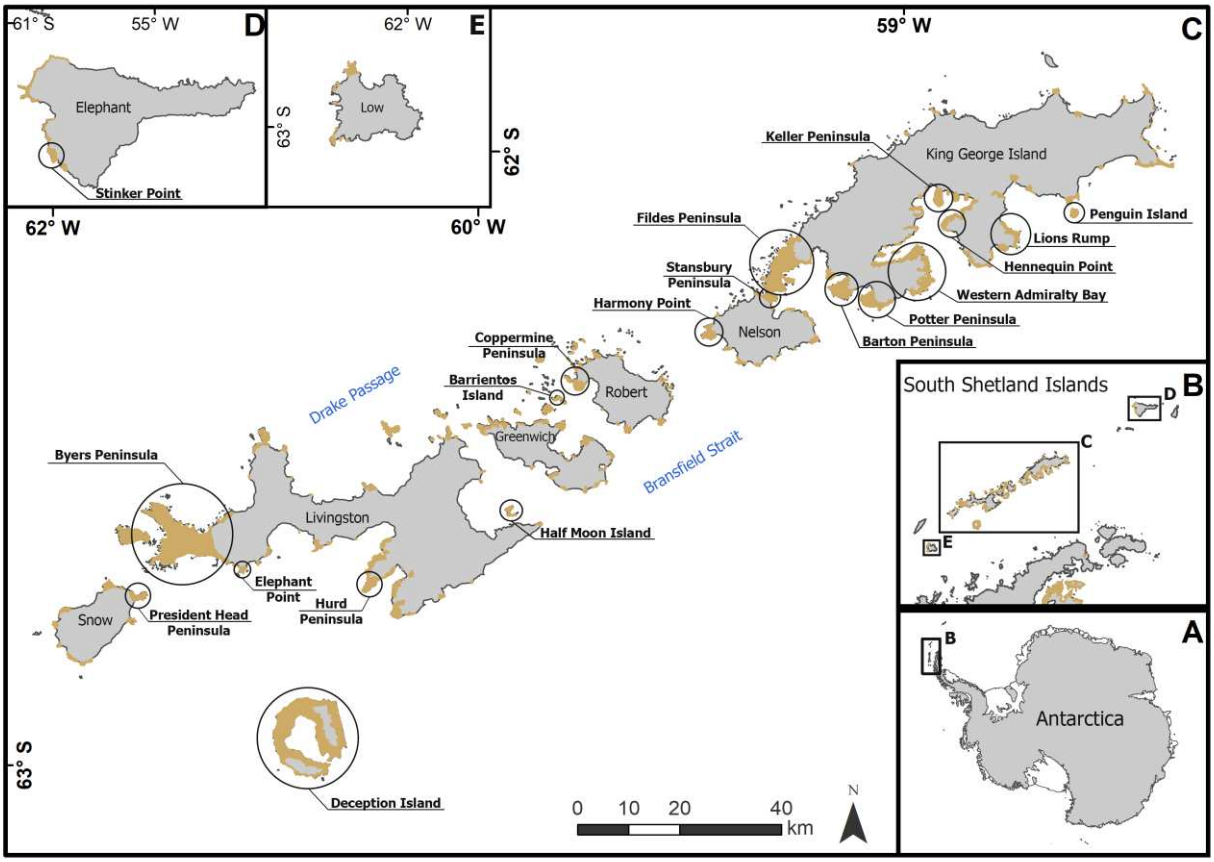


Figure 1. Studied ice-free areas of South Shetland Islands. Circled areas are those with soil samples.

The SSI forms the southern sector of the Scotia Arch and, is composed of marine sandstones and mudstones recording a forearc sedimentation, metamorphic rocks, intrusive bodies, and olivine-basaltic, andesite-basaltic and andesitic lavas associated with tuffs of similar composition (Smellie et al., 2021b, 2021a). The SSI rocks have an overall age varying from Jurassic to Paleogene. The exceptions are the basaltic and andesitic-basaltic lavas and tuffs linked with Late Quaternary to Recent age volcanics from Deception and Penguin islands (Smellie et al., 1984).

Hope Bay area is composed of a series of metasedimentary and volcanic rocks representative of the Antarctic Peninsula geological sequences. In JRI, rocks are part of the back-arc James Ross sedimentary basin, composed by a set of marine sandstones, mudstones and conglomerates predominantly of Cretaceous age (Pirrie, 1994). The Cretaceous sediments are overlain by Neogene olivine-basaltic lavas, hyaloclastic breccia and tuffs (Nelson, 1975) or by Cenozoic sandstones, mudstones and drifts (Zinsmeister, 1982).

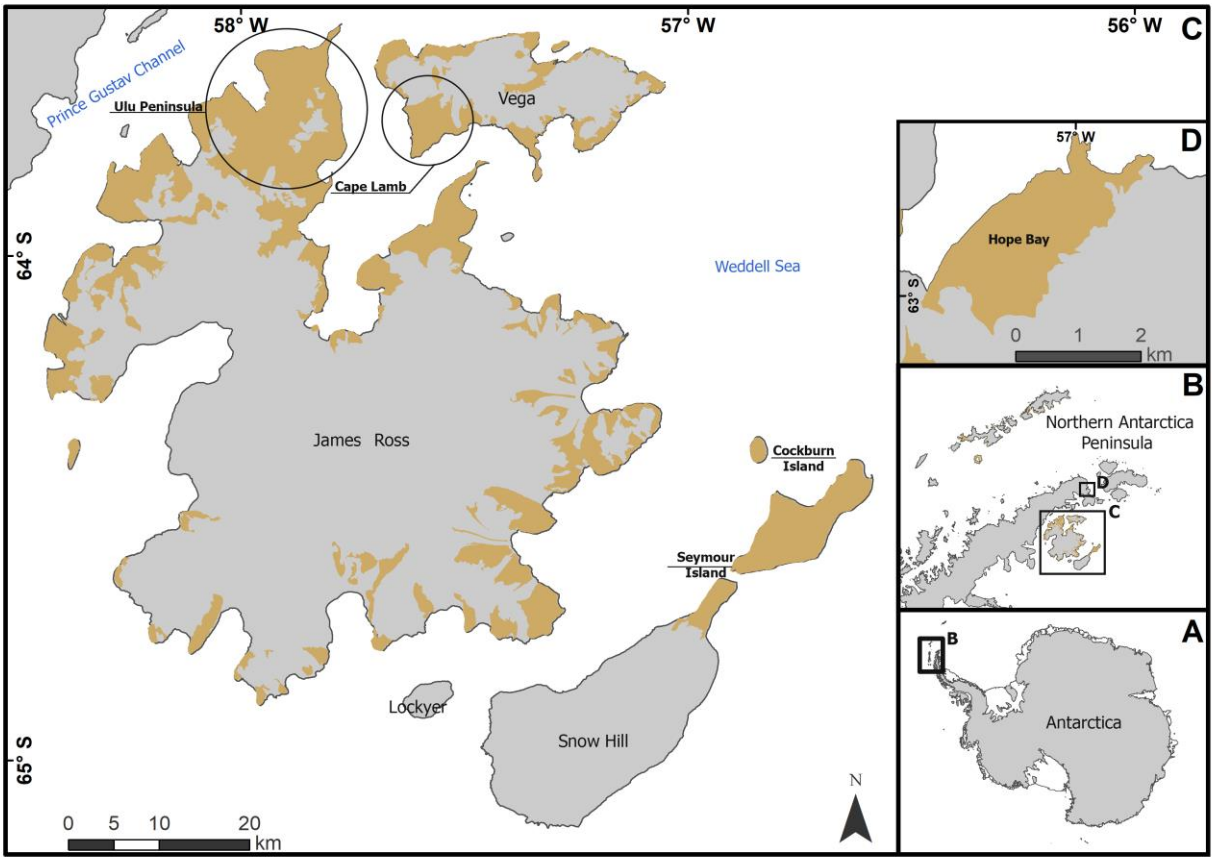


Figure 1 (continued). Studied ice-free areas of Northern Antarctic Peninsula – James Ross Islands and Hope Bay.

The geomorphology of the ice-free areas is marked by a set of erosive and depositional landforms, indicating the complexity of the geomorphic processes (López-Martínez et al., 2012. The SSI landscape is constituted by three distinctive landform groups: moraines deposited after and before the Last Glacial Maximum; raised marine terraces formed by Holocene glacioisostasy; and marine and subaerial platforms, whose the highest ones are also identified as cryoplanation platforms (John and Sugden, 1971). The NAP landforms have strong similarities with the SSI ice-free areas, maintaining marine terraces, moraines and patterned ground surfaces as the main landforms (Schaefer et al., 2015). In the JRI, the topography also shows a strong structural control, with flat-topped plateaus formed by basalt floods (Davies et al., 2013) or resistant sedimentary strata, whereas the less resistant sediments form the lowest erosional surfaces.

The climate and vegetation cover of the ice-free areas varies considerably according to the latitude. The SSI have a sub-Antarctic moist and mild maritime climate with mean annual precipitation ranging from 350 to 800 mm and mean annual air temperature around -2°C (Blumel and Eitel, 1989). Luxuriant carpets and turfs of mosses are widespread in the SSI areas, mainly in areas with higher water availability, whereas lichens dominate on the drier highlands (Schmitz et al., 2020b). Along with lichens and bryophytes two vascular plant species also occur (the grass Deschampsia antarctica and the herb Colobanthus quitensis, Ferrari et al., 2020). Also, macroalgae communities are found next to penguin rookeries (Poelking et al., 2015). In the PA, the climate is somewhat drier, with mean annual temperature of -5°C in Hope Bay and -6°C in JRI, and mean annual precipitation from 300 to 500 mm (van Lipzig et al., 2004). In Hope Bay, vegetation is dominated by macroalgae communities and other lichens/mosses sub-formations. Due to the cold semiarid climate, the vegetation in JRI is scarce, with rare small moss communities.

In MA, the soil types follow an altitudinal gradient. Soils at altitudes > 80 m a.s.l. generally show continuous permafrost and expressive cryoturbation due to the freeze-thaw cycles and are classified as Turbic Gelisols. From 20 to 80 m a.s.l. the permafrost is discontinuous, and Gelisols usually occur associated with Entisols and Inceptisols. Below 20–30 m a.s.l. permafrost is usually absent or sporadic, and most soils key out as Entisols. Important SSI soil features are the thick organic horizons developed under large moss carpets (Michel et al., 2006), and the ornithogenic character promoted by the P-rich avian guano reactions (Rodrigues et al., 2021). The ornithogenesis is also an important process in Hope Bay, with dominance of the ornithogenic Gelisols (Pereira et al., 2013). The main JRI soils are Gelisols due to the widespread permafrost, which can be ice-cemented or dry-frozen (Delpupo et al., 2014). JRI have unique soils with cryoturbic features typical of moister environments mixed with typical desert saline features (Daher et al., 2019a). Acid sulfate soils developed from sulfide-bearing parent materials are also an important pedological type of both region (Lopes et al., 2019, Siqueira et al., 2022a).

2.2. Soil data

We used legacy soil texture data, mainly produced by the Brazilian Terrantar Group, often in intensive soil surveys in the MA and PA during the last two decades. The dataset is composed of complete soil profiles, surface (0 - 10 cm) and pit samples (0 – 30 cm), collected on the main ice-free areas. Additional soil data were obtained from the literature to complement the dataset (Table 1). Overall, the dataset used contained 1035 points, accounting for 2867 samples (Supplementary data, Fig. 1). Most of the samples were in MA, totaling 2578. We used the sand, silt, and clay contents in %. We did not transform the data to meet compositional data-constraints supposedly necessary due to soil texture data being a three-dimensional composition (constant sum in 100%), since there are not always consistent advantages of using data transformations when modelling soil texture, which is dependent on the study area and data distribution as pointed out by Lark and Bishop (2007).

Table 1. Legacy soil texture data of the Maritime Antarctica and Northern Antarctic Peninsula used as input data.

Ice-free areas	Points	Soil profiles	Soil profile samples	Other soil samples	Total samples	Soil Surveys
South Shetland Islands						
King George Island						
Keller Peninsula	145	30	130	266	396	(Francelino et al., 2011; Pires et al., 2017; Schünemann et al., 2022; Thomazini et al., 2016)
Western Admiralty Bay	55	44	164	11	175	(Ferrari et al., 2020; Krauze et al., 2021; Łachacz et al., 2018; Michel et al., 2006; Simas et al., 2008)
Hennequin Point	4	-	-	12	12	(Vieira et al., 2013)
Lions Rump/Low Head	24	24	94	-	94	(Almeida et al., 2021; and unpublished data)
Potter Peninsula	18	18	71	-	71	(Poelking et al., 2015)
Barton Peninsula	41	41	114	-	114	(Lopes et al., 2019; and unpublished data)
Fildes Peninsula	36	36	109	-	109	(Lin et al., 2021; Mendonça et al., 2013; Michel et al., 2014)
Nelson Island						
Stansbury Peninsula	33	23	81	10	91	(Neufeld et al., 2015; Schmitz et al., 2020a; and Meier et al., 2023)
Harmony Point	123	46	180	67	247	(Rodrigues et al., 2019; and unpublished data)
Robert Island						
Coppermine Peninsula	12	12	47	-	47	(Unpublished data)
Livingston Island						
Byers Peninsula	80	80	282	-	282	(Moura et al., 2012; Navas et al., 2005; Navas et al., 2008; Otero et al., 2013; Silva et al., 2022; and unpublished data)
Hurd Peninsula	21	21	76	-	76	(Ganzert et al., 2011; Haus et al., 2016; Ilieva et al., 2021; Navas et al., 2008)
Elephant Point	10	-	-	20	20	(Navas et al., 2017)
Snow Island						
President Head Peninsula	30	30	85	-	85	(Lopes et al., 2022a; Lopes et al., 2022b)
Elephant Island						

Stinker Point	46	31	94	15	109	(Navas et al., 2018; Schmitz et al., 2020b; Schmitz et al., 2022)
Ardley Island	17	17	38	-	38	(Michel et al., 2014)
Deception Island	25	6	12	19	31	(Díaz-Puente et al., 2020; Vlcek et al., 2018; and unpublished data)
Penguin Island	3	3	12	-	12	(Unpublished data)
Half Moon Island	21	21	54	-	54	(Schmitz et al., 2021)
Barrientos Island	5	5	19	-	19	(Daher et al., 2019b)
Northern Antarctic Peninsula						
Hope Bay	21	21	72	-	72	(Pereira et al., 2013; Schaefer et al., 2015)
James Ross Island						
Ulu Peninsula	32	32	106	-	106	(Daher et al., 2019a; Meier et al., 2019; Vlcek et al., 2018)
Vega Island						
Cape Lamb	30	30	86	-	86	(Siqueira et al., 2021)
Seymour Island	44	44	183	-	183	(Delpupo et al., 2014; Gjorup et al., 2020)
Cockburn Island	3	3	6	-	6	(Unpublished data)

2.3. Environmental predictors

To model the spatial variability of soil texture we used a set of environmental predictors related to the *scorpan* factors (McBratney et al., 2003). We generated a set of 61 topographic predictors involving primary and secondary terrain attributes, besides terrain-based attributes classifications (Table 2), derived from the 8-meter spatial resolution Reference Elevation Model of Antarctica (REMA). The REMA mosaic was constructed from individual Digital Elevation Models (DEM) extracted from stereoscopic imagery of the WorldView-1, WorldView-2, WorldView-3, and GeoEye-1 satellites, collected mostly over the austral summer seasons of 2015 and 2016. Each individual DEM was vertically registered to satellite altimetry measurements from Cryosat-2 and ICESat, resulting in absolute uncertainties of less than 1 m over the most area (Howat et al., 2019). Some coastal ice-free areas presented voids due to the proximity to the sea level. These voids were recursively filled with a moving cell window-based interpolation until no empty pixel remained. In the MA ice-free areas covered solely by the 2-meter REMA mosaic, we downscaled the spatial resolution to achieve 8 meters.

We used multispectral data from Sentinel-2 mission satellite imagery as soil predictors, assuming that remote sensing data is able to give detailed information on the soil, since the bare surface reflectance is an intrinsic property of the soil material and may indicate other soil attributes, like texture or mineralogy (McBratney et al., 2003). This assumption is better applied in areas with spatially limited vegetation, as is the case in Antarctica. The images, collected between 2017 and 2021 during the summer seasons, were chosen according to their lowest possible cloud and snow covers in each ice-free area, to maximize the multispectral data capability as soil predictors.

Table 2. Environmental predictors representing the scorpan factors.

Environmental predictors		
r: topography		*c: climate*
Aspect	Slope height	BioClim 1
Convergence index	Slope length	BioClim 2
Cross-secational curvature	Standardized height	BioClim 3
Diurnal anisotropic heating	Surface specific points	BioClim 4
Diffuse insolation	Roughness concentration index	BioClim 5
Direct insolation	Tangential curvature	BioClim 6
Direct to diffuse ratio	Terrain ruggedness index	BioClim 7
Downslope curvature	Terrain surface classification	BioClim 8
Drainage density	Terrain surface convexity	BioClim 9
Effective air flow height	Terrain surface texture	BioClim 10
Elevation	Topographic openess	BioClim 11
Flow line curvature	Topographic position index	BioClim 12
Flow horizontal distance	Topographic wetness index	BioClim 13
Flow vertical distance	Total curvature	BioClim 14
General curvature	Total insolation	BioClim 15
Geomorphons	Upslope curvature	BioClim 16
Gradient	Valley	BioClim 17
Gradient difference	Valley depth	BioClim 18
Hill height	Valley index (0-1)	BioClim 19
Hill index (0-1)	Vector terrain ruggedness	Frost change frequency
Hillslope index (0-1)	Wind exposition	
Hillshade		*o: organisms*
Local curvature		Net primary productivity
Longitudinal curvature	*s: soil*	NDVI
LS factor	Sentinel band 2 (blue)	SAVI
Mass-balance index	Sentinel band 3 (green)	Rookeries (0-1)
Maximum curvature	Sentinel band 4 (red)	Rookeries distance
Mid-slope position	Sentinel band 5 (red edge)	Vegetation (0-1)
Minimum curvature	Sentinel band 6 (red edge)	
Morphometric protection index	Sentinel band 7 (red edge)	*p: parent material*
Multi-resolution ridge top flatness	Sentinel band 8 (NIR)	Geology
Multiresolution index of valley bottom flatness	Sentinel band 8a (red edge)	
Normalized height	Sentinel band 11 (SWIR)	*a: time*
Plan curvature	Sentinel band 12 (SWIR)	Coastline distance
Profile curvature	Sentinel ratio 3/2	Glacier distance
Real surface area	Sentinel ratio 4/2	
Ridge level	Sentinel ratio 8/4	*n: spatial position*
Sky view factor	Sentinel ratio 12/4	Latitude
Slope	Sentinel ratio 12/8	Longitude

We used 6 predictors to represent the biological (organisms) factor: Soil Adjusted Vegetation Index (SAVI) and Normalized Difference Vegetation Index (NDVI) obtained from Sentinel band ratios; a Boolean (0-1) vegetation layer obtained from the NDVI; and the CHELSA net primary productivity (NPP), placed to represent the vegetation influence; and a rookeries Boolean layer and the rookeries distance, representing the fauna (mainly penguins) activity.

A group of 19 bioclimatic covariates available for Antarctica by the CHELSA program were applied as climatic predictors, representing a 30 years' time series (1981-2010). The CHELSA (Climatologies at high resolution for the earth's land surface areas) data consists of temperature and precipitation estimates at a spatial resolution of 30 arc sec, ~1km (Karger et al., 2017), using the ERA5 models as forcing data. Bioclimatic variables provide meaningful biological information, representing annual trends, seasonality and limiting environmental factors, which are often used in ecological modelling. We also used the CHELSA frost change frequency (FCF), which records the number of events in which temperatures vary above or below 0°C. Both the Sentinel and CHELSA variables were resampled using cubic convolution to match the REMA 8-meter spatial resolution.

The distance to glaciers and distance to coastline were used as predictors of the time factor, since it is assumed that there is a time (age) range in the Antarctic ice-free areas, in which areas farthest from the glaciers and from the coastline represent the oldest surfaces, which directly influences soil development. To account for parent material effects, we created a simplified geology layer (e.g., volcanic, sedimentary, plutonic, unconsolidated sediments, etc.) from existing local geological maps (Smellie, 2013; Smellie et al., 2021b, 1984). At last, the spatial position factor was incorporated from the latitude and longitude variables.

2.4. Data preparation

All processing, illustrated in Fig. 2, was done on the free statistical software R (R Core Team, 2023). The majority of the terrain attributes and classifications used as topography covariates were generated with the *RSAGA* package (Brenning, 2008). To model the soil texture, we harmonized the soil data in three standard depths (0-10, 10-30 and 30-100 cm). For this we applied the mass preserving quadratic spline of Bishop et al., 1999 using the *mpspline2* package. For each spline we used a smoothing parameter value of 0.1 leading to values of soil texture for each cm. Next, the spline averaged the soil texture according to the standard depths.

The average step reduced the original number of samples to 2538. In the case of pit samples (0-30), the same value was used to the two depths 0-10 and 10-30.

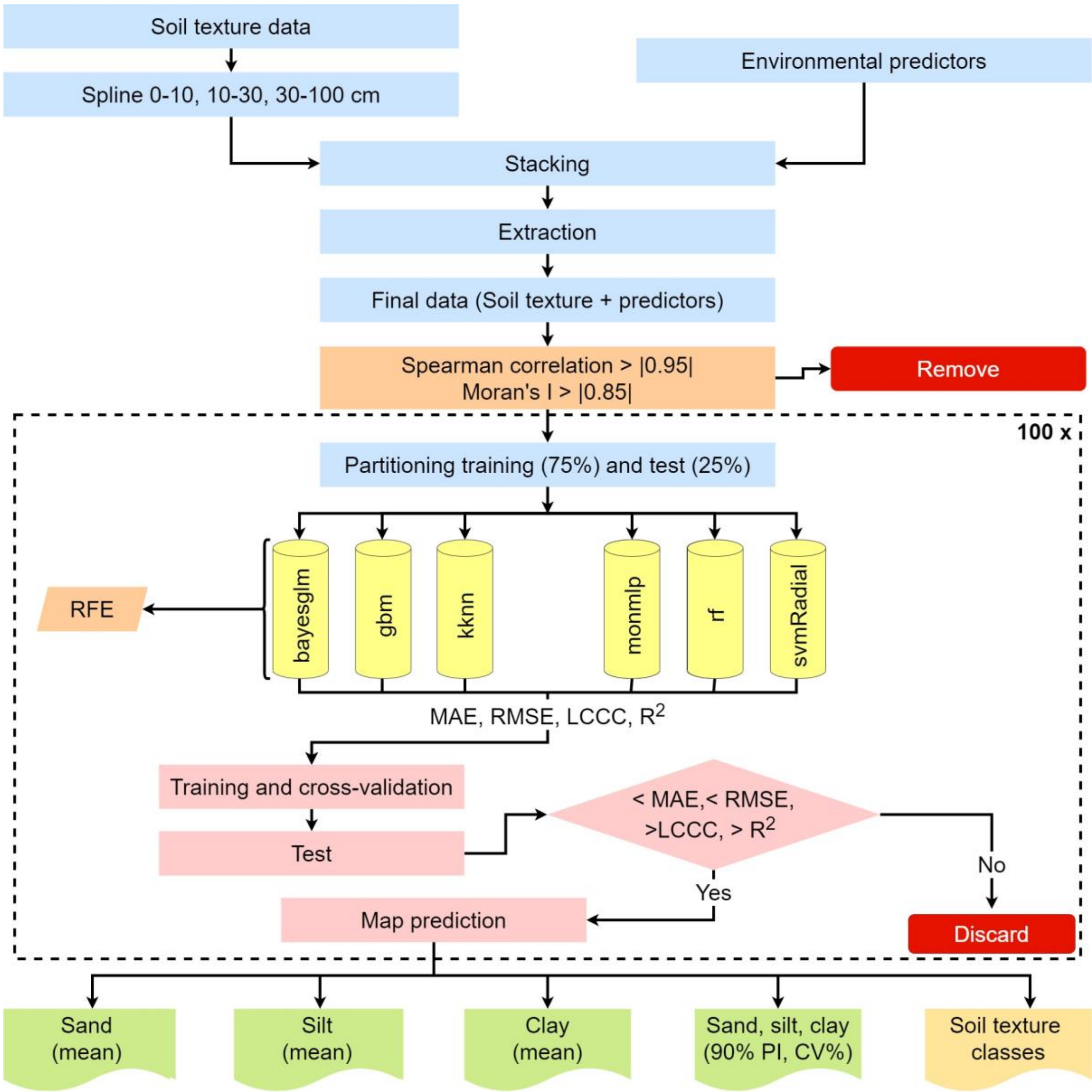


Figure 2. Methodological flowchart containing the steps for the soil texture prediction. To obtain information about the models and metrics acronyms the readers are referred to the topics 2.6. and 2.8., respectively.

The following steps of the processing were performed using the *caret* (Kuhn, 2021) and *mdsFuncs* packages (Fernandes-Filho et al., 2021). The soil texture data were randomly split for each depth into training and test sets in a 3:1 ratio (holdout), with the function *createDataPartition*, which arranges the data into subsets based on percentiles and samples within these subsets. In this way, 75% of the data was included in the training set and 25% in the test set.

2.5. Selection and importance of predictors

To select the most important predictors, we applied a three-step method. Firstly, we removed the highly correlated predictors with Spearman correlation > |0.95|, using the function *findcorrelation*. If two predictors had a high correlation, this function removes the predictor with largest mean correlation with all datasets. We also removed strongly autocorrelated predictors with Moran's I > |0.85|. Then, we applied the Recursive Feature Elimination (RFE) algorithm using the function *rfe* to backward select the optimal set of predictors for each ML model used. RFE starts with the complete set of predictors and calculates the importance of each in a full model, ranking them according to the importance index available for each model. Then it removes the least important predictor and refits the model. This process is repeated several times to determine the best set of predictors, considering the loss of accuracy when removing relevant predictors (Oliveira et al., 2021). RFE was run with the full set of predictors and 130 subsets: first at intervals of 5 predictors until leaving 30 predictors, and then at intervals of 2. RFE was run 100 times, and we chose the optimal subset considering a range of 2% performance loss from the subset with the highest accuracy for the sake of model simplification.

Before the training, we scaled the predictor variables by dividing mean-centered values by their respective standard deviations. This is important to minimize the effect of large variations across the descriptive predictors. Dummy variables were created for the proper handling of categorical predictors with many classes. The relative predictor's importance to the prediction was estimated using the function *varImp*. With this function, values of importance are normalized for a scale of 0 to 100, where the most important predictor is at 100 and the least important at 0. It can use a model-based approach, when the models have specific metrics for calculating importance, or a model-independent approach (Kuhn, 2008). We also evaluated the number of times (frequency) that the predictors were chosen throughout the 100 runs.

2.6. Machine Learning algorithms

We used seven supervised ML algorithms representing the main machine learning families (Table 3). RF is based on 'ensemble learning' (Breiman, 2001), which generates a large set of decision trees and aggregates their results into a single prediction. To build the forest, RF uses the bagging approach (sampling with replacement) to randomly create bootstrap sets from the original training. Each set is then used independently to train a single tree. The training is done by splitting the dataset according to a series of if–then rules, with the most important predictors being placed on the top of trees to create nodes and leaves of higher homogeneity. The outcome is the average prediction of all trees. The main hyperparameter tuned to improve the model performance was the 'mtry', which is the number of predictors used at each tree.

Table 3. Machine learning models used and machine learning families.

Models	R packages	Machine learning families*
Bayesian Generalized Linear Models (GLM)	arm	Probability-based learning
Stochastic Gradient Boosting Models (GBM)	gbm	Information-based learning
Monotone Multi-Layer Perceptron Neural Network (MONMLP)	monmlp	Error-based learning
Radial Support Vector Machine (SVM)	kernlab	Error-based learning
Random Forest (RF)	randomForest	Information-based learning
Weighted k-Nearest Neighbors (k-NN)	kknn	Similarity-based learning

GBM (Friedman, 2002) constructs additive regression trees where subsequent ones are built from the residuals of the previous tree. The model performs a boost by selecting a random subsample at each iteration, without replacement, from the full training data. This subsample is then used to update the model and sequential improvements are obtained with the boosting biasing the new trees to pay more attention to instances that previous models predict incorrectly. The improvement of the model was achieved from the tuning of the hyperparameters 'n. trees', which is the number of adding trees, and 'interaction. depth', which is the number of splits performed on a tree.

SVM (Cortes and Vapnik, 1995) performs the training searching for a hyperplane able to separate the data with the largest possible margin. The margin is constructed as the space between the decision boundary and the first sample points on each side. The Radial SVM uses a radial base kernel function to transform the data into a higher-dimensional space, and then, produce flexible decision boundaries when the data are not easily separated. The algorithm performance depends on a good balance between the 'sigma' and 'cost' hyperparameters. The 'sigma' controls the degree of nonlinearity of the model, while the 'cost' controls the trade-off between margin and training errors.

MONMLP (Lang, 2005) is a multi-layer perceptron neural network developed for multi-dimensional function approximations. Multi-layer perceptron (MLP) consists of a feedforward artificial neural network which contains an input, an output and one or more hidden layers of nonlinearly activating neurons. Each neuron of the hidden layers combines the inputs with back-propagation weights and the output is delivered as the inputs for the neurons of the following layer. The MONMLP has advantages in relation to a normal MLP by incorporating the monotonic constraints between inputs and outputs through the neural network structure. Two hyperparameters were tuned: 'hidden1', which is the number of neurons in the first hidden layer and 'n. ensemble', the number of ensemble members to fit.

The Weighted k-NN (Hechenbichler and Schliep, 2004) performs the prediction of new samples using the k closest samples from the training set. k refers to the number of nearest-neighbor samples used. To avoid overlapping of predictors with large range, all predictors must be transformed before running. Weighted k-NN differs from other k-nearest neighbor algorithms because it uses kernel functions to weigh the neighbors according to their Minkowski distances. The main hyperparameter used for optimization was 'kmax', which is the number of neighbors and must be chosen precisely, since small and large values of k usually overfit and underfit the data, respectively.

GLM (Nelder and Wedderburn, 1972) is a generalized and flexible linear regression which aims to determine the causal relationship between the independent and response variables. The differences of the GLM to the classical linear regressions is that although assuming the predictor has a linear effect, the response is treated with different types of statistical distributions than the normal, besides handling with non-linear relationships. The Bayes GLM incorporates Bayesian techniques for efficiently predicting the unknown parameters of the model and is also considered suitable to avoid overfitting and increase the predictive power. There is no hyperparameter to be tuned in this model.

2.7. Model training

We trained the ML models by extracting and analyzing the predictors' information for each training sample, using the function *train*. We tuned the models using a 10-fold cross-validation and ten possible values of tuning hyperparameters. With this method, the training samples are partitioned into 10 near-equally sized folds. Then, the models are trained by repeatedly leaving out one of the folds and the training performance is determined by predicting on the fold left out (Meyer et al., 2019). To overcome the effect of sampling variability, we ran the models 100 times, creating repeatedly different training and test datasets.

2.8. Model test

To test the models' performance, we applied the fitted models to the test data and expressed the performance of the prediction by five statistical indexes. The mean absolute error (MAE) (Eq. 1) and root mean squared error (RMSE) (Eq. 2) summarize the residuals and describe the absolute accuracy of the models, which means how close the predicted values are to the actual values. The mean bias error (MBE) is calculated as the mean error (Eq. 3). It is not considered a measure of accuracy, since with its positive and negative errors cancel each other, muting the magnitude of errors. Thus, it is a useful indicator of model bias.

The coefficient of determination (R^2) (Eq. 4) indicates the proportion of the variance in the target variable that the model explains and compares how much the model improves prediction over simply using the mean of the observed target variable as the prediction. R^2 varies between 0 and 1. Finally, the Lin's concordance correlation coefficient (LCCC) evaluates the agreement between the observed and predicted values. LCCC assesses both precision and accuracy of the prediction. In this case, precision refers to the spread of points around the regression line (correlation coefficient) for the predicted versus observed. Accuracy refers to the correspondence between that regression line and the perfect line (1:1 line). In general, LCCC can be considered a more appropriate evaluation than R^2 by capturing bias in the model prediction (Khaledian and Miller, 2020). It varies between −1 and 1 and has been assigned the term ρ_c (Eq. 5).

As an additional validation, we applied the "Null" MAE and "Null RMSE", by calculating the metrics for a null model with the mean observations as predictors. The use of "Null" models is a good strategy to establish thresholds and evaluate the "badness" of the models, since it allows the comparison of the tested models with a model where the parameters (predictors) are 0 (null).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i| \qquad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \qquad (2)$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n} (y_i - x_i) \qquad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (4)$$

$$LCCC = \frac{2\rho\sigma_{x_i}\sigma_{y_i}}{\sigma_{x_i}^2 + \sigma_{y_i}^2 + (\bar{x} - \bar{y})^2} \qquad (5)$$

2.9. Prediction and uncertainty

To build the soil texture predicted maps at different depths, we selected among the models that attained the best performance throughout the test. This resulted in 100 maps for each soil particle size fraction at each depth, and final maps of the sand, silt and clay were the resulting mean value for each pixel. We classified the soil fractions data into soil textural classes applying the package *soiltexture* (Moeys et al., 2022) and using the Soil Taxonomy's textural triangle (Soil Survey Staff, 2014). The classes were validated using the classes of the original dataset and the kappa and global accuracy classification measures.

We also evaluated the local uncertainty (pixel-level uncertainty) following Gomes et al., 2019 proposal. For that, we computed the upper (Q95) and lower (Q5) percentiles for each pixel through the 100 runs. Then, we estimated the uncertainty as the difference between the 95th and the 5th percentiles - 90% Prediction Interval (PI). The higher the PI width for a pixel, the higher the uncertainty. Additionally, we calculated the coefficient of variation (CV % = standard deviation / mean) of the 100 times predictions and made a qualitative evaluation of the spatial patterns of uncertainty across the study area.

3. RESULTS

3.1. Characterization of soil data

In the original data, the sand fraction presented the highest values, with a mean of 63%, followed by silt (mean 23%), and clay fractions (mean 14%). Little changes were observed with depth (Table 4). The soil texture showed predominance of the sandy-loam textural class, but sandy, loamy sand, sandy clay loam and loam classes were also significant in our dataset (Fig. 3).

Table 4. Descriptive statistics of the sand, silt and clay fractions from the soil dataset splined according to the standard depths.

Depth			Min	Max	Mean	Median	SD	CV (%)
0-10 cm	Sand	%	0.02	99.99	63.82	63.80	17.04	26.70
n: 1023	Silt		0.00	75.77	21.90	20.94	12.82	58.55
	Clay		0.00	55.61	14.27	13.92	8.09	56.72
10-30 cm	Sand	%	0.37	99.99	62.53	62.03	17.53	28.03
n: 997	Silt		0.00	94.33	23.42	22.24	13.72	58.59
	Clay		0.11	51.08	14.14	13.55	8.03	56.79

30-100 cm	Sand	%	0.73	99.99	63.07	63.78	18.41	29.19
n: 518	Silt		0.14	77.67	23.13	22.76	13.19	57.00
	Clay		0.04	47.00	13.77	12.73	8.86	64.33
Total	Sand	%	0.02	99.99	63.14	63.20	17.66	27.97
	Silt		0.00	94.33	22.82	21.98	13.24	58.05
	Clay		0.00	55.61	14.06	13.40	8.33	59.28

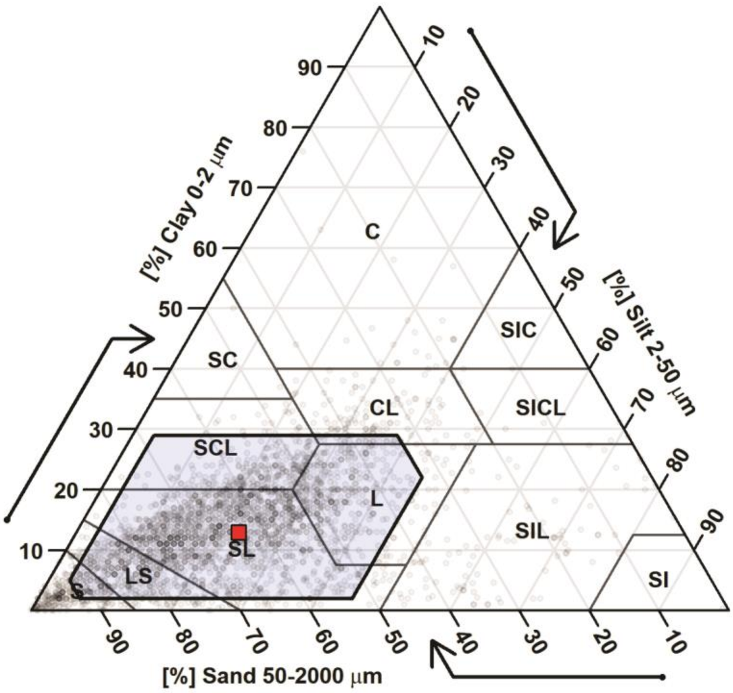


Figure 3. Soil texture triangle with the MA and NAP soil data. Representative value (red point) corresponds to the median and low-high region (polygon) defined by the 5th and 95th marginal percentiles. C: clay, CL: clay loam, L: loam, LS: loamy sand, S: sand, SC: sandy clay, SCL: sandy clay loam, SL: sandy loam, SIC: silty clay, SICL: silty clay loam, SIL: silt loam, SI: silt.

3.2. Selection of predictors

Depending on the depth, 22 to 27 redundant predictors were discarded through the correlation and autocorrelation elimination steps, representing mainly some bioclimatic, spectral bands and indexes, curvature, insolation, terrain attributes linked to the surface slope, and latitude and longitude. The RFE managed the remaining predictors to select the optimal set of predictors. The subsets chosen varied in size according to the models, soil particle size fraction and depth (Fig. 4). Considering the 100 runs, the number of predictors selected for a subset varied from seven to 47 (mean 18) for the clay fraction, seven to 72 (mean 24) for the sand, and five to 58 (mean 15) for the silt, indicating a higher tendency to construct simpler models when predicting the silt fraction. On the other hand, the number of predictors varied from six to 31 (mean 13) for 0-10 cm, nine to 55 (mean 20) for 10-30 cm, and seven to 72 (mean 24) for 30-100, indicating simpler models at the topsoil. Overall, the GLM, GBM and k-NN constructed large subsets for all fractions. The RF selected intermediate-sized subsets of three to 26 predictors (mean 10).

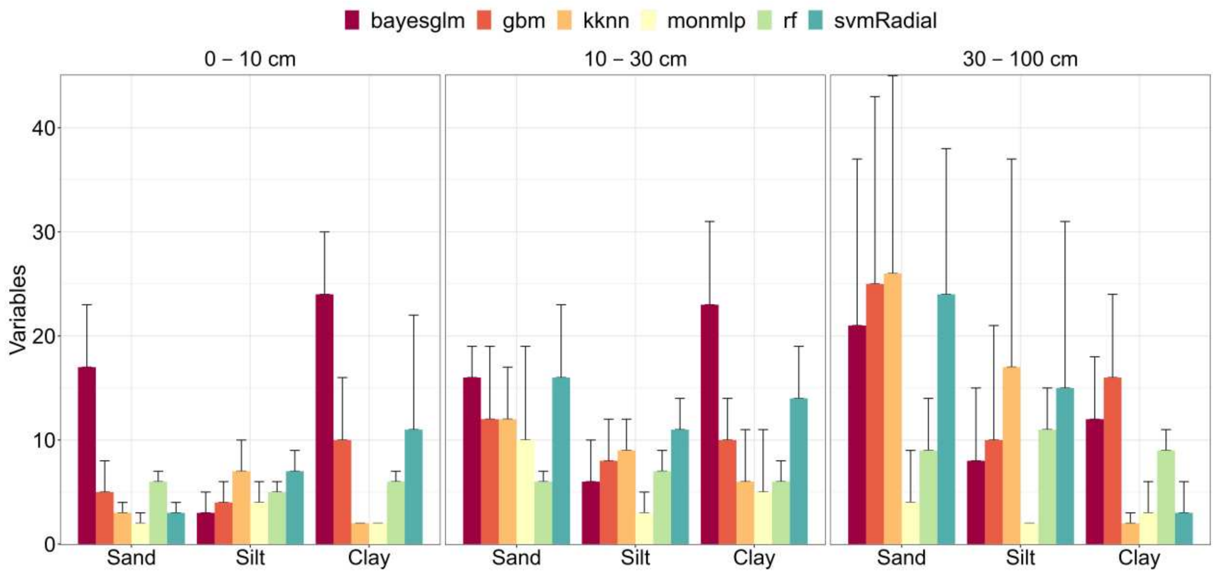


Figure 4. The average number of predictors selected by the models for each soil particle size fraction, throughout the 100 runs. Error bars mean the standard deviation. bayesglm: Bayesian Generalized Linear Models, gbm: Stochastic Gradient Boosting Models, monmlp: Monotone Multi-Layer Perceptron Neural Network, svmRadial: Radial Support Vector Machine, rf: Random Forest, kknn: Weighted k-Nearest Neighbors.

3.3. **Importance of predictors**

In general, the best predictors selected by the RFE were the bioclimatic variables, the Euclidean distances to glaciers and coastline (variables linked to soil development) and a few terrain attributes (Fig. 5). The most important variable for the clay fraction was the temperature seasonality (BioClim 4), achieving 100% of relative importance at the first two depths. It was followed by the glacier distance and isothermality (BioClim 3) at all depths, whereas the annual precipitation (BioClim 12) and precipitation seasonality (BioClim 15) were also important bioclimatics regardless of the depth. In turn, the terrain surface texture and drainage density, variables associated with the surface’s rugosity, were the most relevant terrain attributes for all depths. The Sentinel band 11 and the Sentinel band ratio 8/4 were significant for clay at the depth 0-10, indicating the importance of the multispectral data to prediction at topsoil.

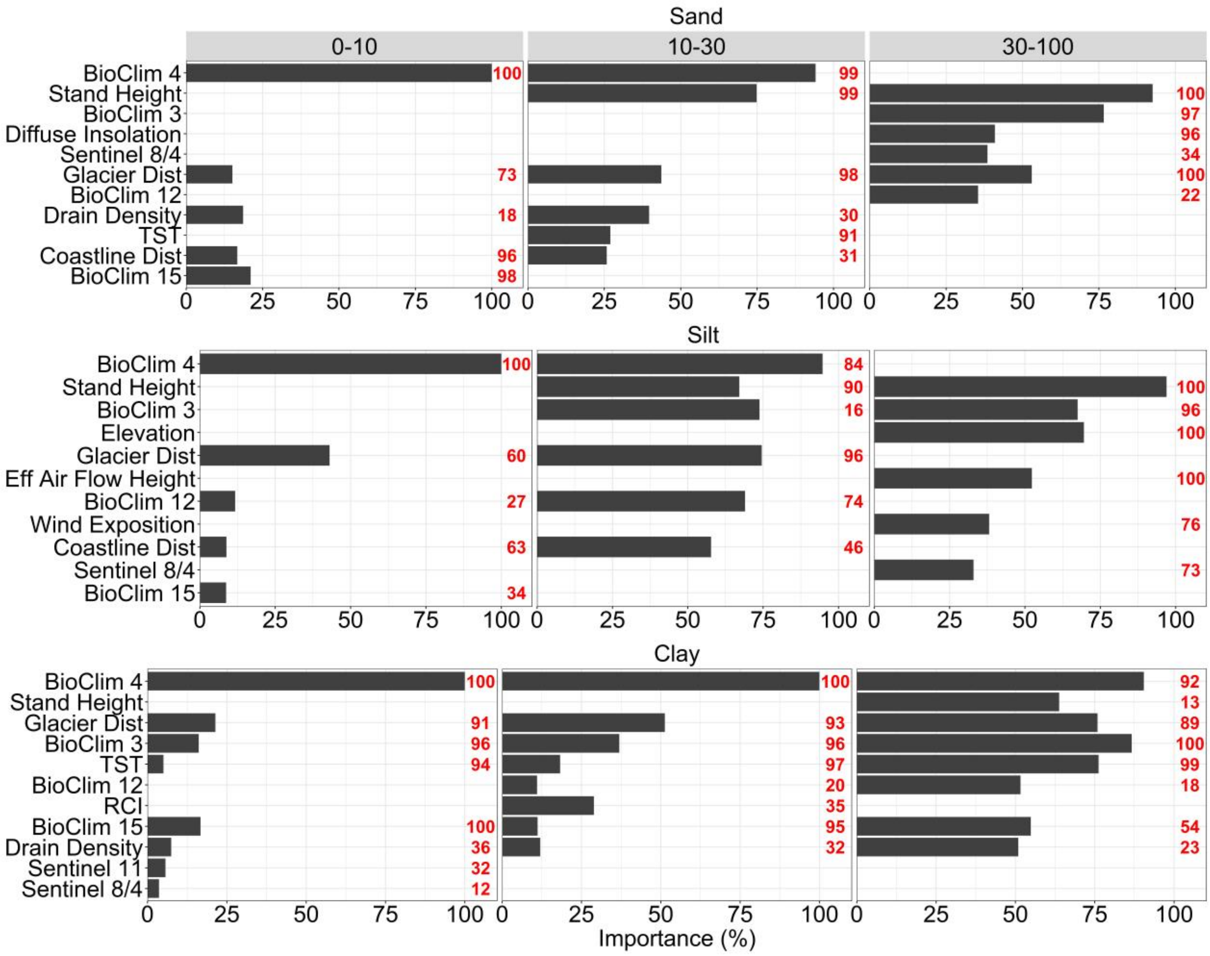


Figure 5. Mean relative importance and frequency of selection during the 100 runs of the RF-predictors for each soil fraction and depth. Depicted data are the top 8 predictors with frequency > 10. The complete data with all predictors chosen in the 100 times are shown in Supplementary data, Fig. 2.

The bioclimatics were also important for the silt and sand fractions, mainly the temperature seasonality which was not the best predictor only for sand at the deepest layer (30-100 cm). The glacier distance and coastline distance were also important for the two fractions, with the first one showing the highest importance for silt at 10-30 cm (75%) and the second for sand at 0-10 cm (25%). For the silt, the terrain attributes were more important when predicting at subsoil, mainly at 30-100 cm depth. The most relevant were those related to the topography's height, such as standardized height, effective air flow heights, elevation, and wind exposition. A similar pattern was observed for the sand fraction. The standardized height was the most important predictor for this fraction at the deepest layer, with a mean relative importance of 95% and 100 times of frequency of selection.

The predictors related to the organism's factor had less importance to the models. The NDVI and SAVI indexes were removed by high correlation with the Sentinel 8/4 predictor, so, we can assume this band ratio as a potential predictor related to the biological factor. Geology had a smaller overall contribution but appeared with some important classes for clay and sand fractions (Supplementary data, Fig. 2). The most important geology classes for clay were the pyritized rocks and marine mudstones and sandstones, parent materials which contribute with high rates of clay to the soils.

3.4. Model performance

The RF presented the best training performance for all fractions and depths, with LCCC values always next to 0.9, and maximum R^2 of 0.32, followed by the GBM, k-NN and SVM (Supplementary data, Table 1). In general, the RF also showed higher test performance, indicating the best agreement and least deviations between the observed and predicted values among the models. It is also followed by the GBM and k-NN models (Fig. 6), whereas the GLM was the worst for almost all fractions and depths (with minimum LCCC and R^2 of 0.04 and 0.02, respectively) (Supplementary data Table 1), which can be expected since the incapacity of this model in extracting the non-linear patterns necessary to modelling the soil texture.

The models had better performance when predicting on clay fraction, with RF showing the highest results at the depth 0-10 cm (LCCC of 0.49 and R^2 of 0.31). As well the other models, the RF generated the worst performance when predicting silt, with mean test LCCC and R^2 as low as 0.17 and 0.29, respectively (Table 5). The bias values were higher at the depth 30-100 for all fractions, indicating the greatest tendency to the errors in the deepest layer. RF was also the model with the lowest MAE and RMSE values (Supplementary data, Table 1) and was the only model which presented better MAE and RMSE than the null models for all soil fractions (Fig. 6), indicating its greatest adequacy for predicting soil texture. The MBE values < 1% indicate insignificant model bias, demonstrating virtual absence of systematical errors. Nonetheless, the negative values indicate a weak tendency of RF to underestimate the observed values.

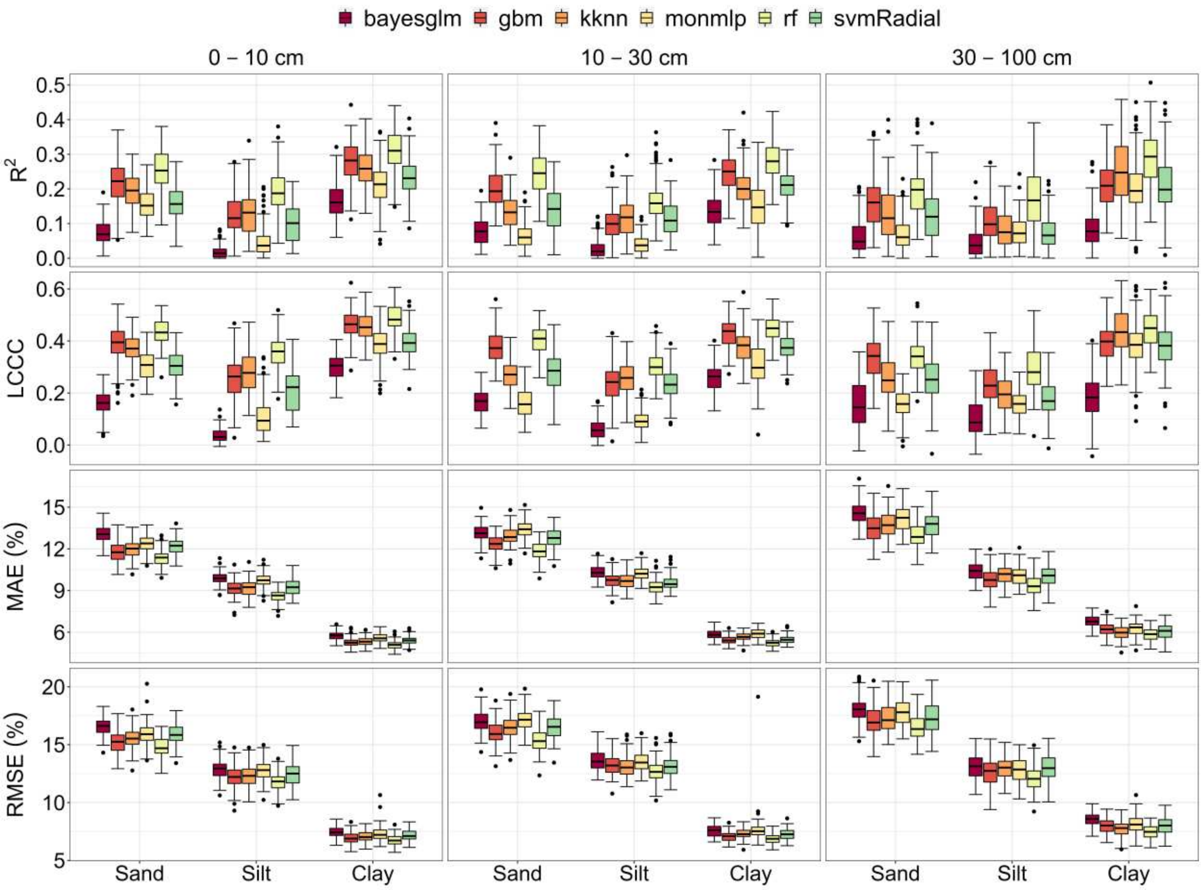


Figure 6. Test performance of the models on soil texture prediction. bayesglm: Bayesian Generalized Linear Models, gbm: Stochastic Gradient Boosting Models, monmlp: Monotone Multi-Layer Perceptron Neural Network, svmRadial: Radial Support Vector Machine, rf: Random Forest, kknn: Weighted k-Nearest Neighbors.

The errors of RF prediction increased at subsoil. The largest decreases were observed for sand, with a fall in LCCC values of 0.43 at 0-10 to 0.34 at 30-100 cm. This decrease is also observed for the silt and clay fractions, although with less intensity (Table 5). Since all measures indicated the RF as the model with the best performance and generalization capacity, we selected it to spatial modelling and producing the soil texture maps.

Table 5. Test performance of the RF for sand, silt, and clay predictions at three depths.

Fractions	Depth	R^2	LCCC	MAE	RMSE	MBE
Sand	0-10	0.26	0.43	11.33	14.82	-0.69
	10-30	0.24	0.40	11.84	15.39	-0.97
	30-100	0.19	0.34	12.98	16.50	-0.96
Silt	0-10	0.19	0.35	8.59	11.73	-0.69
	10-30	0.17	0.30	9.27	12.67	-0.85

	30-100	0.17	0.29	9.31	12.09	-0.88
Clay	0-10	0.31	0.49	5.06	6.77	-0.41
	10-30	0.28	0.45	5.22	6.87	-0.43
	30-100	0.29	0.44	5.85	7.48	-0.44

3.5. Prediction and mapping

According to the statistics of the final RF maps, the sand fraction showed the highest contents at the three depths. The highest sand contents are found at 0-10 cm, with a mean of 66.3% and maximum of 83%. On the other hand, silt contents were higher (mean 24%) in the greatest depth, whereas clay presented the lowest contents between the three fractions, with a general around 13%. The predicted values did not vary significantly through the depths (with a mean difference among depths of just ≤ 3% for all fractions), following the uniformity of the original samples. The silt fraction presented the highest CV%, achieving 20% at the deepest layer and indicating a greater variability throughout the study area. The least variation occurred with the sand fraction, not exceeding 9%. The contents of all fractions are similar between MA and NAP, with a maximum difference of just 5% for the sand fraction (Supplementary data, Table 2). The final maps underestimated and overestimated high and low observed values, respectively. The greatest overestimations occurred for the sand, whereas the highest underestimations were associated with the silt and clay contents (Fig. 7).

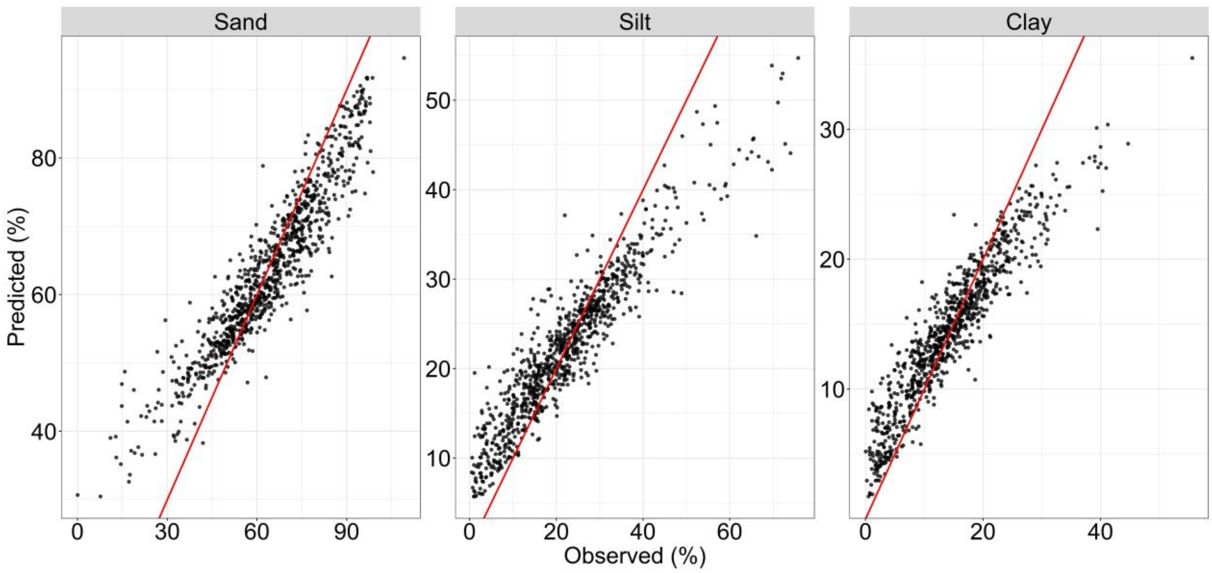


Figure 7. Observed vs mean predicted over the 100 runs data for sand, silt, and clay fractions for the depth 0-10 cm.

Half Moon and Penguin Islands presented the highest sand values, averaging around 76%, followed by Barrientos, Deception, and Livingston islands. Livingston Island presented a great variability of the sand contents, with ice-free areas with mean values of 73% (Hurd Peninsula) and others with values as low as 25% (Byers Peninsula, mean of 60%). The islands with the lowest sand values were Ardley, with 58%, followed by King George, Elephant, and Nelson islands. In NAP, values do not reveal large variations, with mean values ranging between 61 and 65%. The highest silt contents were found in Elephant Island, averaging 26.5% (Supplementary data, Table 3). The highest clay contents are found in Snow, Nelson, King George, and Ardley islands, with mean between 14% and 16%. Western Admiralty Bay in King George Island was the ice-free area with the highest clay contents (19 %), followed by Byers Peninsula (Livingston) and Harmony Point (Nelson) and Keller Peninsula (King George). Vega Island presented the highest values in NAP, averaging 14.5% (Supplementary data, Table 3). The lowest clay contents were found in Half Moon, Penguin, Barrientos, and Elephant islands (Fig. 8,9,10,11 for the mean contents in 0-100 cm; the complete maps for each depth are found in Supplementary data, Fig. 4).

The higher variability of soil texture contents in MA is also expressed for the soil textural classes mapping. The sandy loam texture was the dominant class (Fig. 12), representing almost 100% of the total NPA area. In MA, the loamy sand texture represented 84% of the area, followed by the sandy clay loam, loam and loamy sand, with approximately 5% each one. The clay loam class had an area < 1% and it was found only in Byers Peninsula.

3.6. **Prediction uncertainty**

Considering the local uncertainty, the CV% was higher at the deepest layers, except for the silt fraction. This can be related to a lesser precision when predicting over lesser data, as happened to the depth 30-100 cm. CV% was higher on silt prediction, with mean CV% of 13% and lower for sand (mean 5,5%, Supplementary data, Table 4). Nonetheless, the prediction interval width was larger for the sand fraction (mean PI = 10.9%), indicating higher deviations of the 5th and 95th percentiles from the predicted mean, whereas the lowest values associated to the clay prediction (mean PI = 4.8%) (Supplementary data, Table 5). Considering the depths, the uncertainty addressed by the percentiles follow the trend observed with the CV%, with a general higher and lower uncertainty in the depths 30-100 and 0-10, respectively.

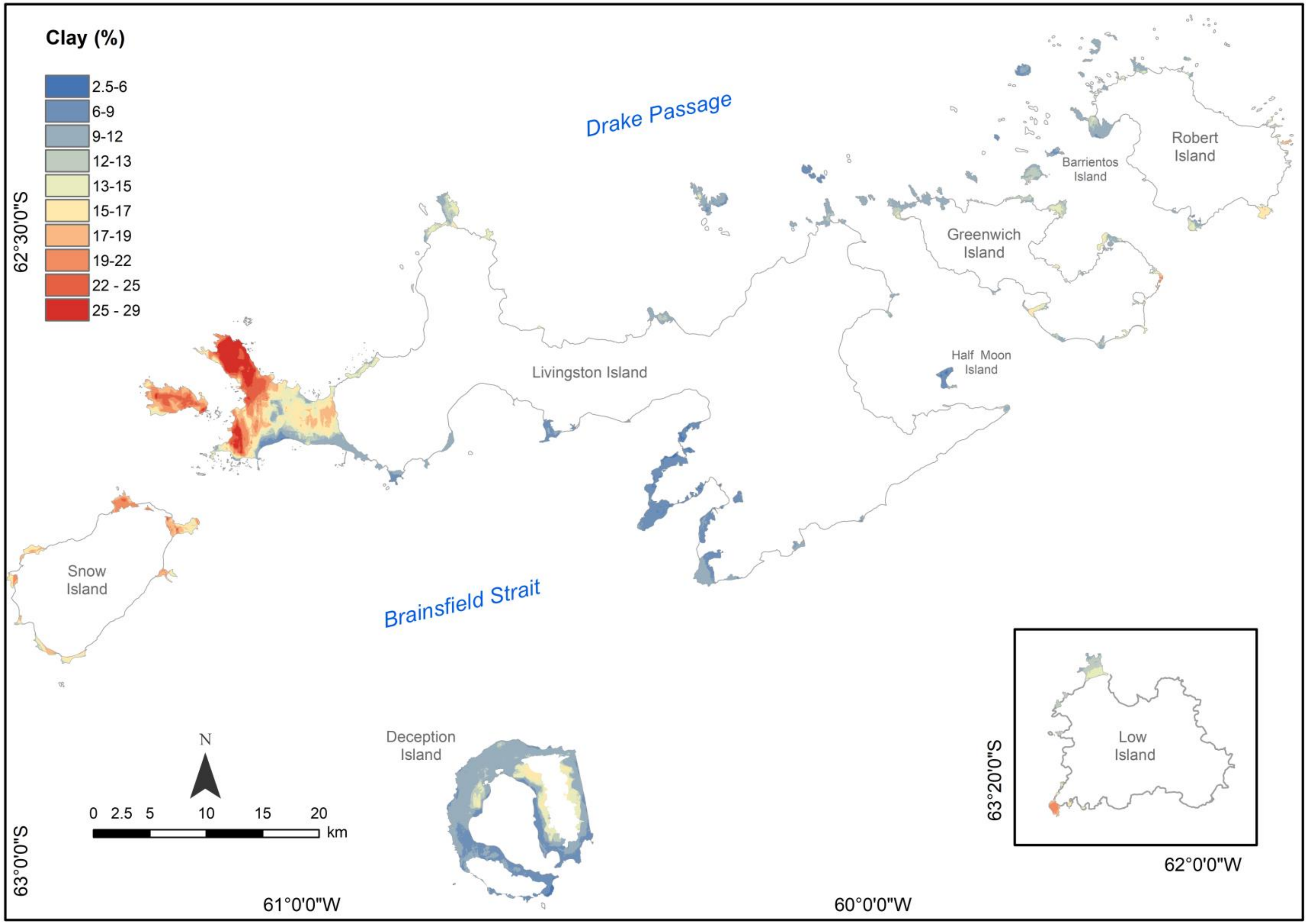


Figure 8. Mean clay distribution for 0-100 cm predicted by RF for Maritime Antarctica.

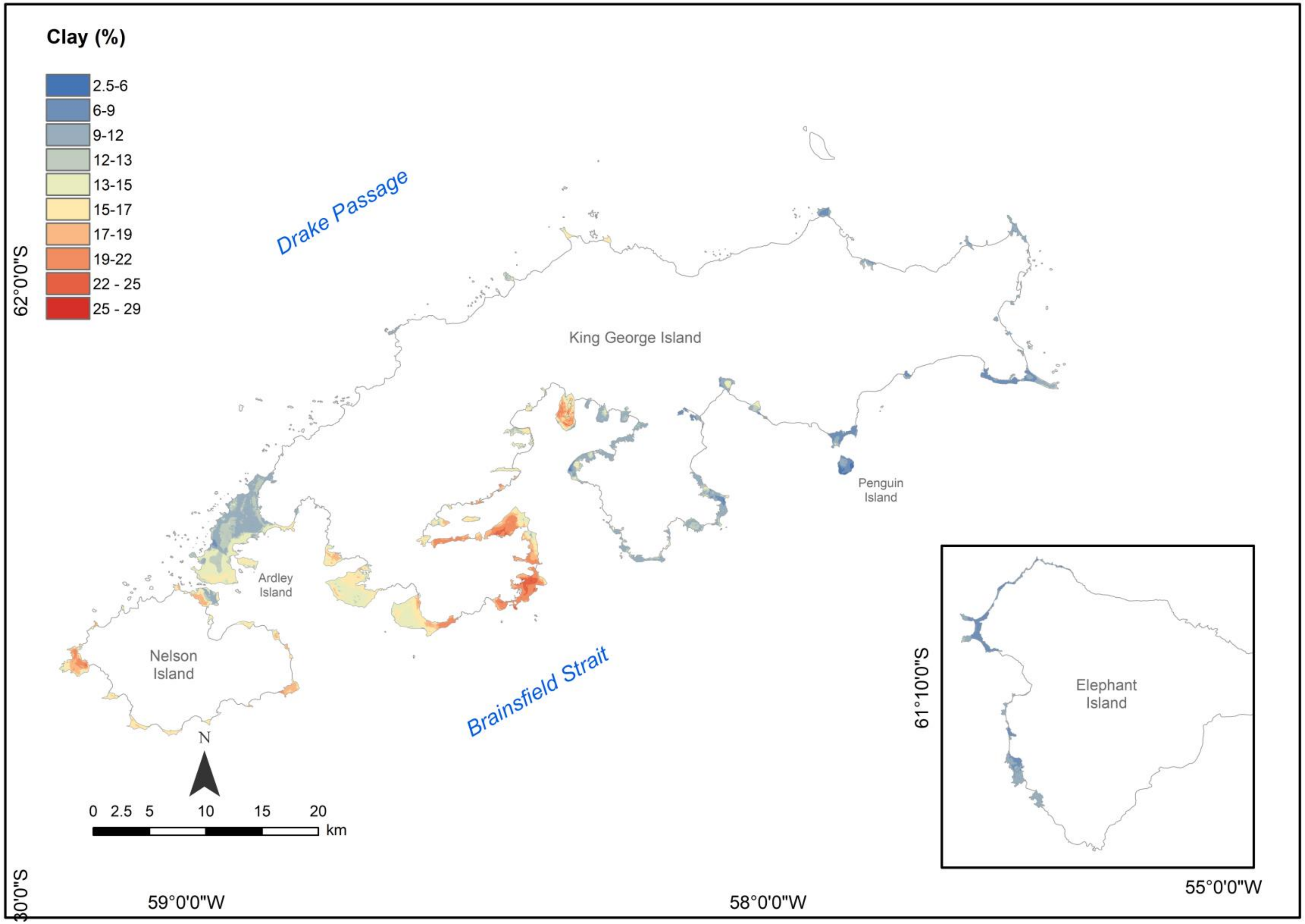


Figure 8. (continued)

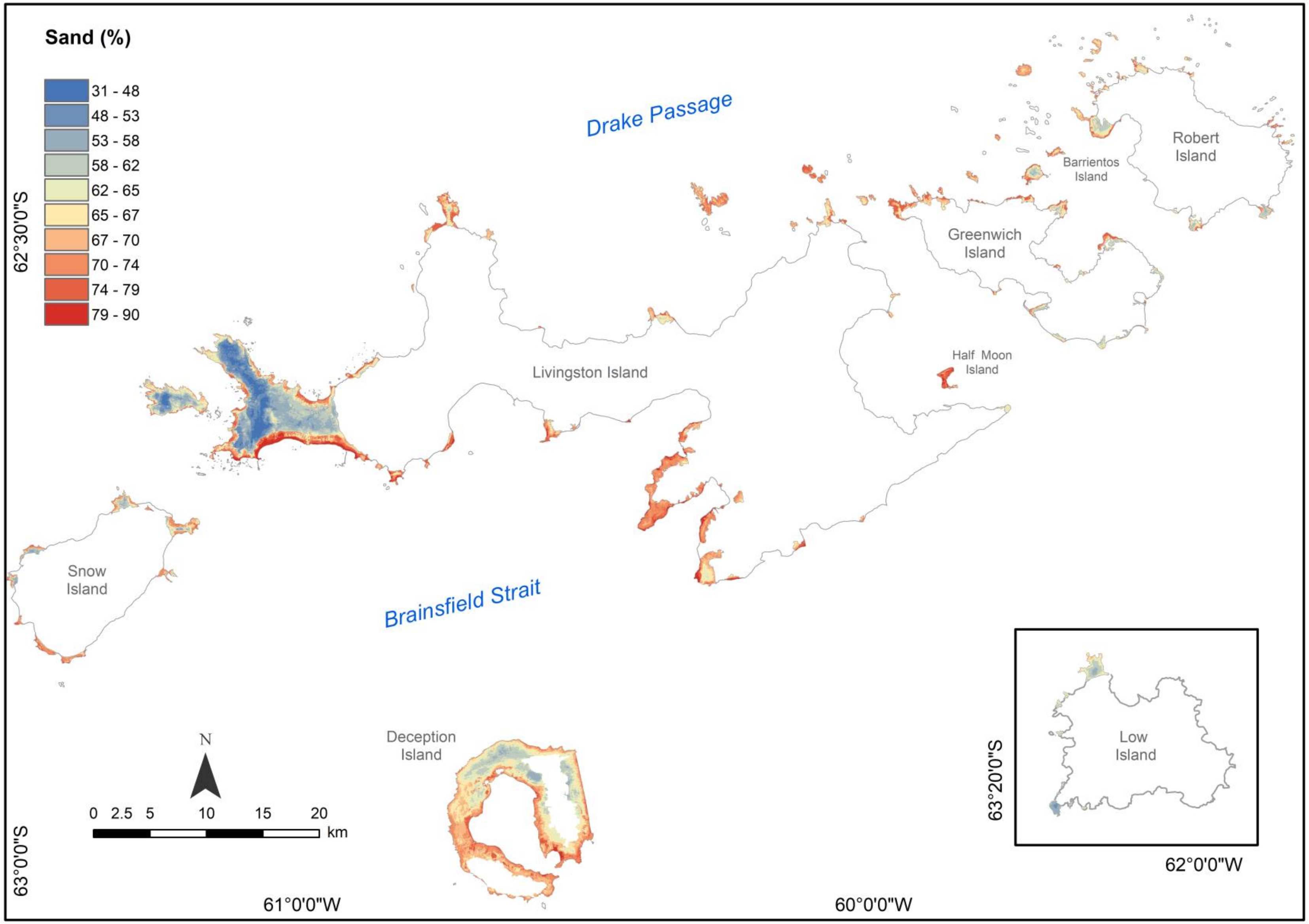


Figure 9. Mean sand distribution for 0-100 cm predicted by RF for Maritime Antarctica.

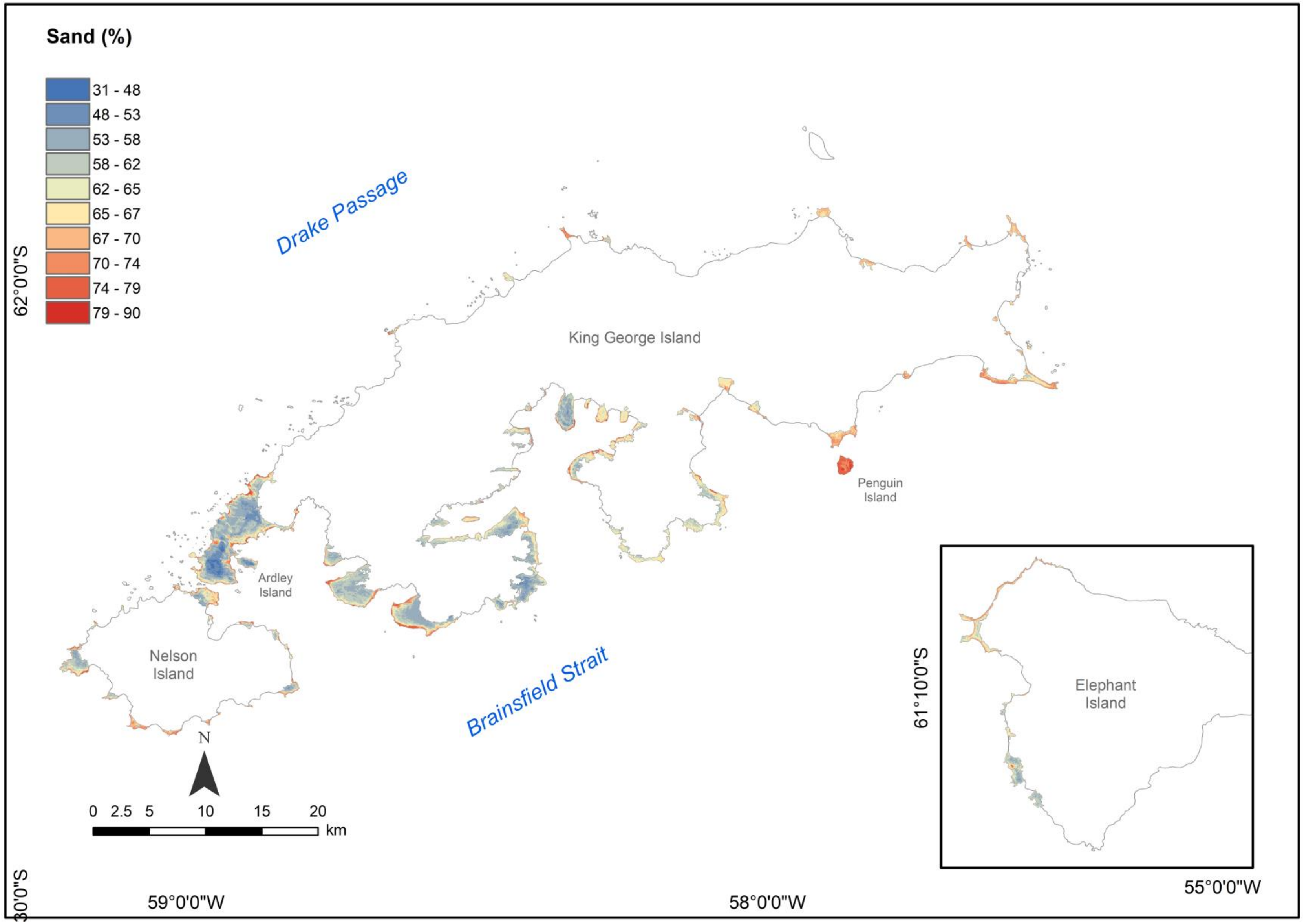


Figure 9. (continued)

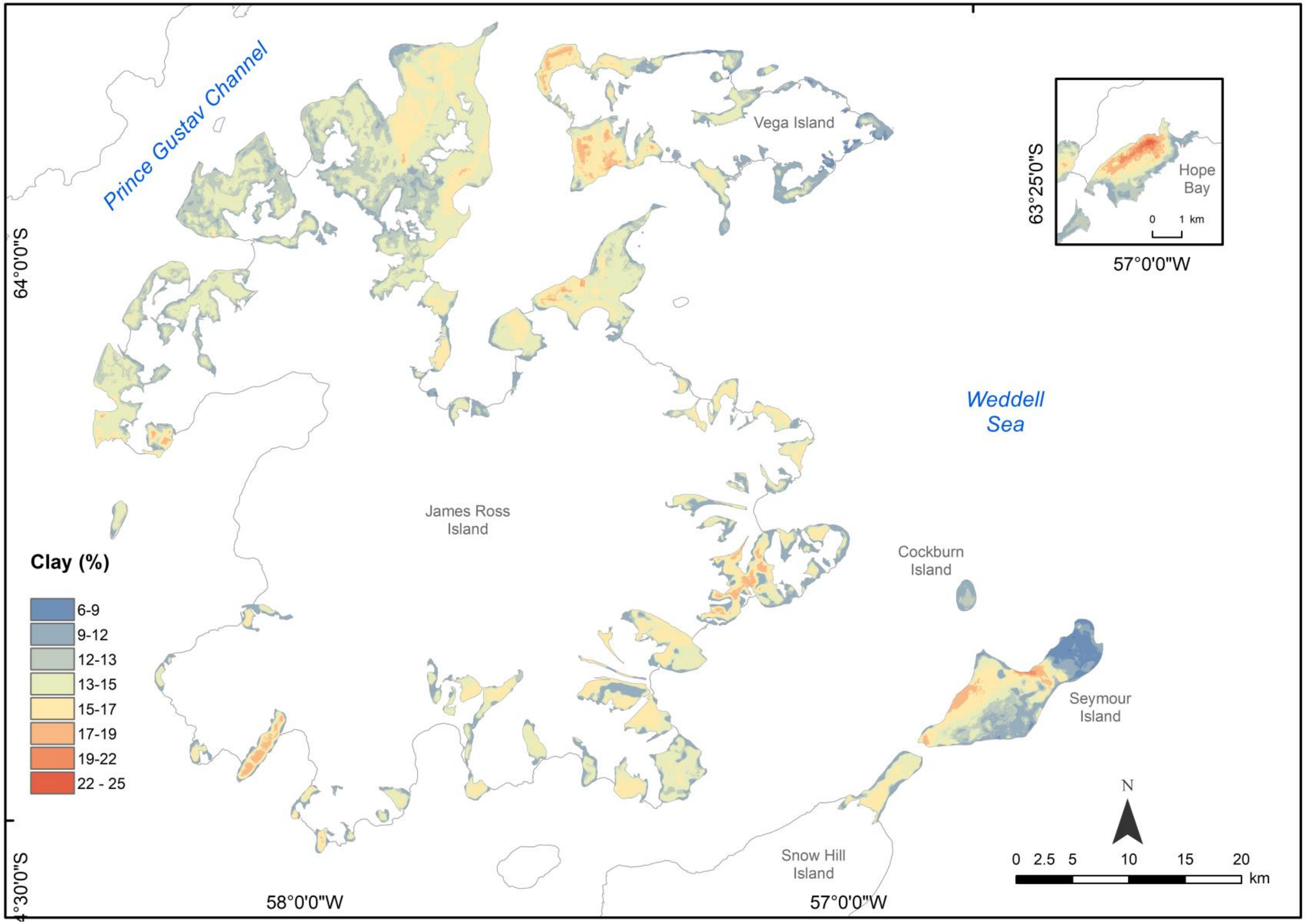


Figure 10. Mean clay distribution for 0-100 cm predicted by RF for Northern Peninsula Antarctica.

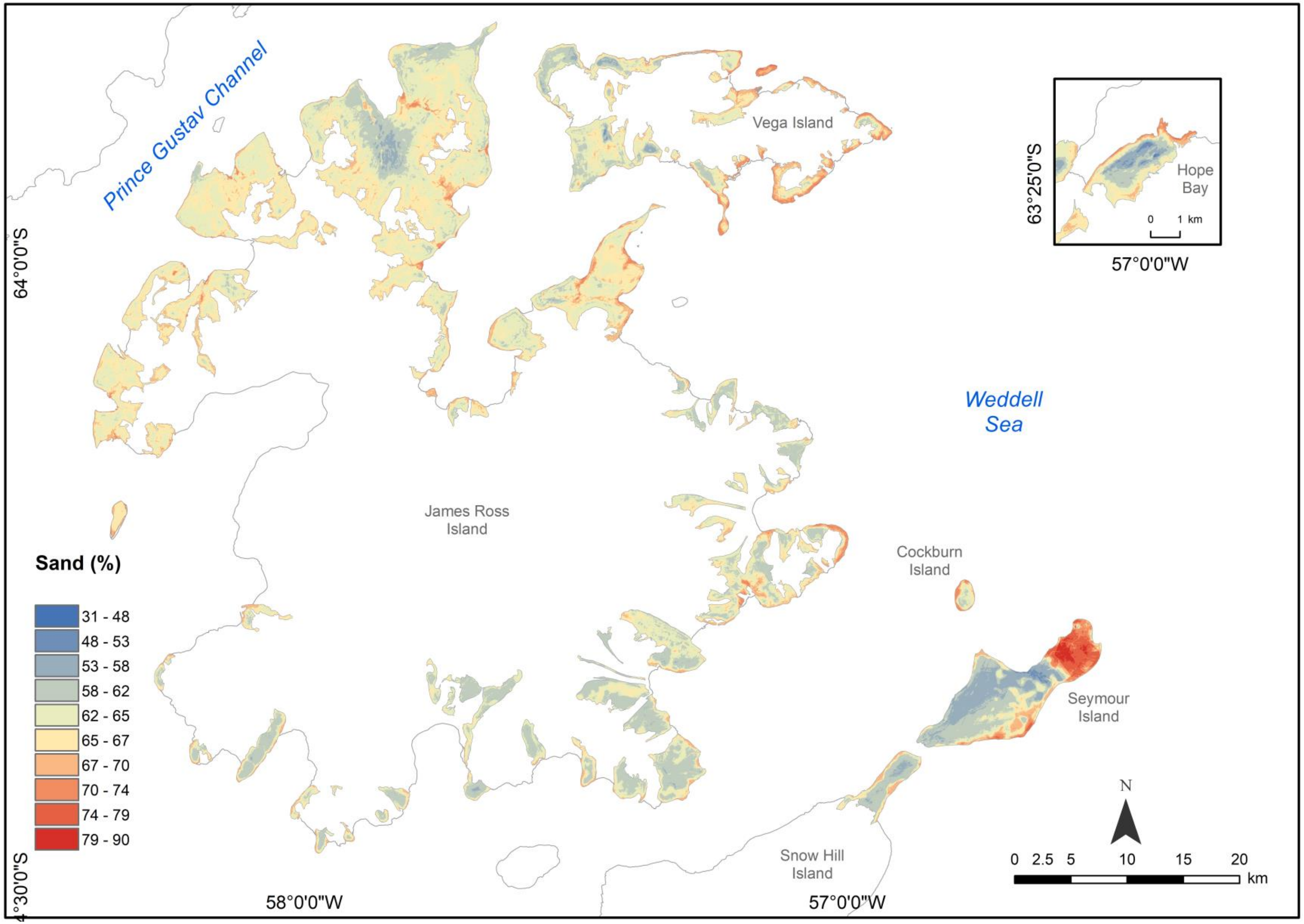


Figure 11. Mean sand distribution for 0-100 cm predicted by RF for Northern Peninsula Antarctica.

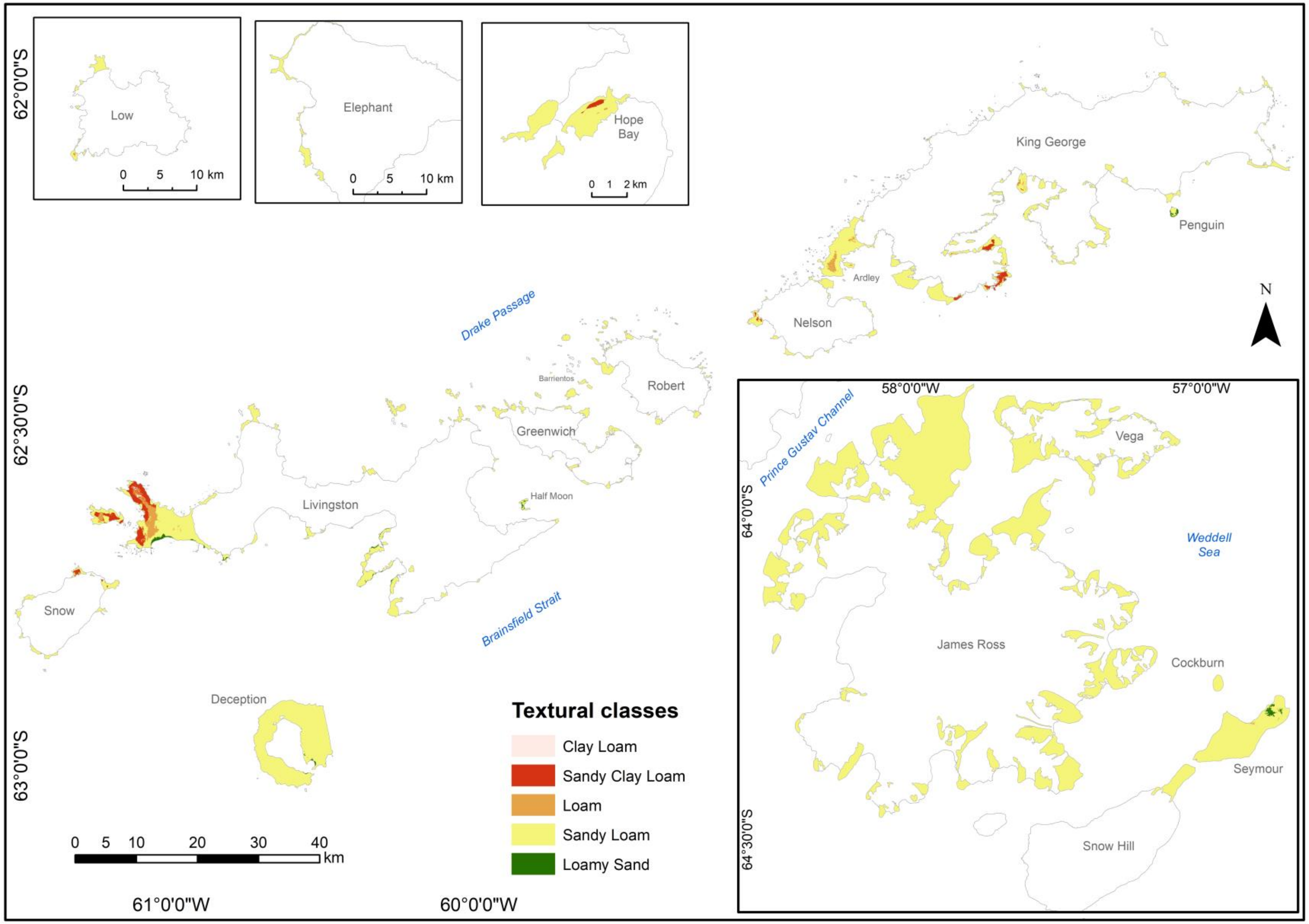


Figure 12. Soil texture classes at 0-100 depth for the Maritime Antarctica and Northern Antarctic Peninsula.

From a look at the Prediction Interval and CV% maps (Supplementary data, Fig. 5 and 6), the uncertainty showed a very clear spatial pattern, mostly controlled by the spatial structure of the main predictor variables and/or by the presence of extreme soil fractions values. The local uncertainty was slightly higher on MA in comparison to NAP, mainly in large areas such as Deception Island, Byers Peninsula and Fildes Peninsula where anomalously high or low contents of sand, clay, and silt, respectively, are common. Differently what was observed by Ballabio et al., 2016 and Gomes et al., 2019 the uncertainty did not increase according to the complexity of terrain and it was not sensible to sample density. So, it is clear the importance of the extreme values for the uncertainty distribution in our study area, indicating the difficulty of our model in dealing with natural outliers associated with the pedological complexity of the Antarctic ice-free areas.

4. DISCUSSION

Our methodological framework was able to optimize the soil texture prediction by applying processes to handle multicollinearity and spatial autocorrelation, to backward selecting the best environmental predictors, to optimize the models hyperparameters, to assess performance using an independent validation dataset and to use repetitions to handle the input data variability. We also tested the prediction accuracy of different ML models, which is an important approach since the performance of the models can largely vary depending on the particularities of each problem and study area.

4.1. RF: optimal model

RF presented the best overall performance when predicting the soil texture. RF has been commonly used for DSM (Akpa et al., 2014; Hengl et al., 2015; Malone and Searle, 2021; Sousa et al., 2020), due to the advantages that this model presents in comparison to other models, which is mainly due to the fact RF is an ensemble model that benefits from a set of decision trees to improve the generalization ability of the model and to determine a more accurate and stable prediction (low variance) (Kelleher et al., 2015).

Besides the predictive power, the RF advantages also include, great efficiency against noises and overfitting, good performance on small datasets, good interpretability, ease of use, optimization of non-additive and non-linear relationships (Biau and Scornet, 2016; Breiman, 2001). RF also largely surpasses individual decision trees by constructing smoother decision surfaces and producing maps more consistent with reality (Chagas et al., 2016).

RF has been widely used in DSM studies relative to other algorithms, such as convolutional neural networks (CNN) (Taghizadeh-Mehrjardi et al., 2020), multiple linear regressions (Chagas et al., 2016; Zeraatpisheh et al., 2019) and other regression trees (Hitziger and Ließ, 2014; Ließ et al., 2012), involving soil texture predictions in different parts of world. In these studies, only the CNN and GBM (Hitziger and Ließ, 2014) models presented prediction performance similar or higher than RF. When comparing ten supervised ML, Caruana and Niculescu-Mizil, 2006 found the GBM and RF, close followed by SVM, as the best learning algorithms, in agreement with the higher performance of these models in our study.

In this study, the best accuracy was found for predictions on clay fraction, whereas the worst was found on silt fraction. Many studies showed the same trend in different regions in the world (Akpa et al., 2014; Mansuy et al., 2014; Poggio and Gimona, 2017), indicating higher correlation of the clay contents with the predictor covariates. However, Román Dobarco et al., 2017 have already found the best accuracy to silt and worse to clay, indicating this is extremely variable according to the region and predictors used. Our results also showed a higher prediction performance in the topsoil and increasing prediction error in the lower depth intervals. This finding is in line with the results of Adhikari et al., 2013 and Taghizadeh-Mehrjardi et al., 2020. In general, the lower accuracy measures found in this study in comparison to other works reveal the complexity of the Antarctic ice-free area's landscapes condition.

4.2. Advances and challenges of our methodology

Soil is a natural three-dimensional body with variability in all three dimensions (Ma et al., 2021). The mass-preserving spline function used in this work is designed to handle this tridimensionality of soil data (Bishop et al., 1999). It is particularly relevant to handle legacy data, making continuous down-profile estimates of attributes and allowing depth standardization of soil observations often obtained over horizons with different depths and thicknesses. Predictions based on the spline function can be defined as a multi-layered pseudo 3D or 2.5D prediction, since it includes independent predictions over depth intervals without influence of the other depths, whereas full 3D modelling approaches are characterized by incorporating depth along with spatial covariates as predictor variables (Ma et al., 2021). However, different studies have pointed out the 3D modelling results in much higher uncertainty and unrealistic predictions relative to the 2.5D approach (Ma et al., 2021; Nauman and Duniway, 2019), which supports maintaining the use of spline in DSM.

Removing predictors with high correlation is also a common method in DSM (Gomes et al., 2019; Mansuy et al., 2014), since the collinearity can drive to problems such as high performance estimates despite few or no significant predictors and predictors estimate instability (Riza et al., 2021). Less addressed in the DSM literature is the need to deal with predictors with strong spatial autocorrelation (SAC). According to Lennon, 2000 predictors with SAC tend to be preferably selected or acquire non-realistic importance due to their spatial structure, mainly if the response variable has also some degree of SAC. This would lead to a biasing in the prediction, causing misinterpretations and preventing spatial models from making meaningful contributions (Meyer et al., 2019). Nonetheless, these steps must be taken with caution (particularly the use of the thresholds) since they can hinder prediction interpretation by eliminating relevant and/or easily interpretable predictors.

The use of variable selection algorithms has been proved an important approach to design parsimonious models, besides increasing the computational efficiency (Gomes et al., 2019). Variables selection is also important when testing different models, as each model can have a particular set of optimal predictors. Variables selection techniques can also enhance the models results (Xiong et al., 2014), although it doesn't happen to all tested datasets (Nussbaum et al., 2018). As a backward process, RFE does not test all predictors combinations, but it is less time-consuming and computationally expensive, being more feasible for larger datasets. In particular, RFE has proven effective to select optimal predictors when applied with Random Forests (Brungard et al., 2015; Hounkpatin et al., 2018). The good fit of the RF with the RFE process is evidenced once the RF-optimal sets almost always present better performance than the full predictors (Supplementary data, Fig. 3), indicating the stronger RF generalization capacity using lesser explanatory variables (Brungard et al., 2015; Gomes et al., 2019).

Some predictors were selected just a few times during the 100 repetitions, conditioned by particularities of the training dataset. Repetitions prevent these predictors retained to solve specific problems in the training dataset from being of great relevance to the overall results. At the same time, combining the frequency of selection along the 100 runs with the relative importance of the predictors to evaluate the final importance of the predictors showed to be of great relevance. Taking our results again, some of the predictors selected sporadically presented high importance scores, such as the drainage density for silt at 0-10 and drainage vertical distance for clay at 30-100, which, although selected only once presented a relative importance higher than 50% (Fig. 5). The use of frequency as a criterion allowed surpassing this problem, avoiding misleading importance evaluation.

Other important approaches in our study are the use of an independent dataset to test (holdout dataset) and of many iterations (repetitions). The use of the holdout dataset is the best approach to really know the predictive power of the fitted models in space, on other words, if the models can predict beyond the spatial location of the observations, and at the same time, avoiding overoptimistic error estimates by testing on observations which were used for fitting the model. Wadoux et al., 2021 also argue that approaches using an external independent dataset (which they call probability sample) similar to the holdout strategy used in this work are the most appropriate method to evaluate the performance of spatial models since they are the best way to deal with the autocorrelation present in most spatial data, far surpassing other approaches of spatial cross-validation (Meyer et al., 2019). The only difference of the probability sampling used by Wadoux et al., 2021 to our holdout strategy is they performed the sampling to obtain the external dataset over the population space, whereas we performed it over the sample space (our sample dataset). However, it is important to remember that having a ready-made population space is mostly viable only in strict methodological procedures, such as in Wadoux et al. 2021, which is not plausible to most modelling works that aim to generate novel data.

In turn, several iterations are important to determine the prediction variability, since different groups of training and test datasets can generate different accuracy results. Repetitions are even more necessary to deal with variations of performance when few sites are used to calibrating models over complex landscapes (Vaysse and Lagacherie, 2017). In our work, the different sets of variables selected in each run show how different training datasets can drive distinct predictors and consequently predicted maps. The repetitions also allow the uncertainty estimation to have a quick idea of mapped areas with lower assurance level (Ballabio et al., 2016; Sousa et al., 2020).

It has already been overly discussed in literature the importance of the evaluation of uncertainties based on 90% PI (Malone et al., 2011). The PI reports the range of values within which the true but unknown value is expected to occur 9 times out of 10, with a 1 out of 20 probabilities for each of the two tails (Arrouays et al., 2014). Its relevance stands out by the fact that a DSM model is considered suitable not only for its ability to accurately predict the value of a soil property (comparing the observed and predicted values) at a given location but also for its ability to predict how uncertain this prediction is (Vaysse and Lagacherie, 2017). A common approach for addressing the uncertainty is to use the quantiles prediction based on the bootstrapped trees from the Quantile Random Forest (Ma et al., 2021; Vaysse and Lagacherie, 2017). However, this method is limited to just one algorithm and alternative approaches have been developed to address the uncertainty for other ML algorithms (Ballabio et al., 2016). In this work, we followed the proposal of Gomes et al., 2019 by using the 100 iterations to construct the 90% PI. This approach not only incorporates the uncertainty present at the trees level into each single iteration, but also permits to address the uncertainty for the wide range of prediction models that can be used in DSM.

Finally, we recognize the limitations of the soil data used in this work. The limited number of soil samples down the soil profiles (largely associated with the shallowness of the Antarctic soils) did not allow the fitting of our data to the six standard depths proposed by the Global Soil Map project (Arrouays et al., 2014), restricting us to only three depths. Another important limitation is the low range of the soil texture data, which resulted in the extensive domain of the sandy-loam soil texture and the homogeneity observed for the two regions mapped. This can also be noted when analyzing the vertical variability of the sand, silt, and clay values, which vary little across the soil depths (Table 4). This could also be potentialized using the RF, which tended to reduce the range of values by smoothing the few extreme values (Fig. 7), a behavior already shown for this algorithm in other works (Yang et al., 2016). Anyway, the homogeneity of the soil texture is a characteristic of the original data, and although the intrinsic difficulties, we assume the results were able to make separations within the limited range.

4.3. Soil texture-environment relationships

The dominance of sand contents is a common feature of the Antarctic soils, with varied proportions of coarse sand and fine sand depending on the parent materials (Siqueira et al., 2021). The predominance of the sandy loam texture on MA and NAP soils has been largely recorded (Daher et al., 2019a; Delpupo et al., 2014; Francelino et al., 2011; Lopes et al., 2019), and has been addressed to the weaker soil development and prevalence of cryogenic weathering processes (Bockheim, 2015), although clay formation by cryoclasty has also been recorded (Simas et al., 2006). Nevertheless, the MA and NAP are the regions with the greatest pedogenic development of Antarctica (Balks et al., 2013). Among other characteristics, this development is evidenced by the greater clay formation conducted by the weathering processes (Haus et al., 2016), conditioned by a warmer and moister climate than in the rest of the continent.

The application of soil texture–environment relationships from the use of different predictors has proved to be an efficient way to model the variability of soil texture fractions (Greve et al., 2012). In general, climatic variables obtained from spatial climate models and topographic variables derived from DEMs are among the most important predictors used for soil texture predictions (Amirian-Chakan et al., 2019; Ließ et al., 2012; Riza et al., 2021).

The bioclimatic variables had higher influence when predicting at the topsoil. These results were consistent with those reported by Ma et al., 2017 and Taghizadeh-Mehrjardi et al., 2020. The higher importance of the bioclimatic predictors at surface shows the close association of the soil texture with the climate, since the greatest exposition to the atmospheric factors favors the surface weathering and clay accumulation, even during limited warmer and wetter conditions (Bishop et al., 2018).

The climate is the main driver of the pedogenesis in the Antarctic continent (Balks et al., 2013). The direct proportional relation of the clay values with the bioclimatics seasonality of temperature (BioClim 4) and precipitation (BioClim15), as well as the negative relation with the isothermality (BioClim 3) (Supplementary data, Fig. 7) shows the importance of the annual climate variability for accumulating clay. Seasonality variables represent the variations of the monthly mean temperatures and monthly precipitation estimates over the year and captures the differences of the seasons in the Antarctic continent. The importance to the soil texture distribution is associated with the fact that they highlight the concentration of optimal climatic conditions which favor the weathering and pedogenetic processes. Whereas in the winter the soil processes are deactivated due to the extremely freezing temperatures, the largest seasonality reflects summer temperatures high enough to fluctuate around the freezing point daily (Hrbáček, et al., 2017; Schaefer et al., 2017), allowing liquid-water availability in soil and freeze-thaw cycles both influencing the clay formation through physical and chemical weathering (Simas et al., 2006).

Although in Antarctica the precipitation rates peak during the equinoxes, the rainfall (liquid-water precipitation) is higher in the summer season (Vignon et al., 2021), what, along with the melting of the snow, contributes to increasing the soil moisture and soil processes during the warmest months. The relevance of the water availability to the soil texture is also evidenced by the annual precipitation (BioClim 12), although without the same importance as the precipitation seasonality by incorporating the water-equivalent snow precipitation.

The predicted soil texture patterns have been largely interpreted as result of catena sequences, with erosion and eluviation of fine particles increasing the sand contents on slopes and clay contents in depositional areas (Amirian-Chakan et al., 2019; Hitziger and Ließ, 2014). This can partly explain the relevance of drainage density for clay prediction, indicating the importance of the soil fluxes in the particle size redistribution (Zhao et al., 2009). At the same time, the higher clay contents in the areas of greater drainage density reveal a greater formation of clay in the areas of a more mature landscape, with a better development of the periglacial drainage network (Siqueira et al., 2021).

The terrain surface texture was the main topographic attribute to predict clay, influencing it negatively (Supplementary data, Fig. 7). In general, the steepest areas of the ice-free areas of Antarctica (e.g., talus-slopes) correspond to younger and unstable paraglacial landscapes modified by a series of erosive agents, such as catena processes, rivers, wind, gravity, etc. (Siqueira et al., 2022b). This leads to the extensive clay removal and concentration of the coarser fractions and evidences the clear confluence between the pedological and geomorphological development in Antarctic ice-free areas (Francelino et al., 2011; Lopes et al., 2022b).

The topographic attributes were more important at the deepest layers. In general, predictors associated with absolute and relative heights have been pointed as the main topographic predictors to modelling soil texture (Carvalho Junior et al., 2014; Hitziger and Ließ, 2014), which corroborate our findings, at least, to the silt and sand predictions. These predictors, coupled with the Euclidean distance variables, also evidence the soil-landscape interplays. The coastline and glacier distances were of great importance for sand and clay/silt contents, respectively (Supplementary data, Fig. 7). The highest sand contents are found in the low coastal domain, mainly in the uplifted marine terraces and beaches where the waves deposition favored the sand grains accumulation and formation of soils with loamy sandy and sandy texture (Francelino et al., 2011; Michel et al., 2014; Schaefer et al., 2015). The importance of the predictor variable standardized height and at lower rates the variables effective air flow heights and elevation, in which the greater contents of sand are directly related to the lowest values of these variables (Supplementary data, Fig. 7), also depicts the dominance of sand on the beaches and terraces raised in the Mid-Holocene at a mean maximum rate of 20 to 30 m a.s.l. (Fretwell et al., 2010).

On the other hand, the highest silt and clay contents are generally found in the highlands over landforms such as moraines and high cryoplanated surfaces (López-Martínez et al., 2012). Moraines are glaciogenic deposits formed after the ice retreat and are composed of non-sorted debris involved by a silt-clayey matrix that commonly originates soils with loamy textures (Francelino et al., 2011). The accumulation of silt and clay is directly related to the glacier's abrasion, which grinds the substrates concentrating the finer particles in the moraine sediments, and the post-glacial cryoclasty which acts on the boulders and minerals left by the glaciers (González-Guzmán et al., 2017). These landforms are generally located closely to the glaciers front (Davies et al., 2013), which explains well the greatest contents of the finer fractions associated to the lowest distances from the glaciers, whereas the opposite is seen to the sand fraction.

The high cryoplanated surfaces are the more stable and older periglacial surfaces from the study area and present highly developed soils with typical presence of finer textures, such as the sandy clay loam common in King George and Nelson islands (Michel et al., 2014; Rodrigues et al., 2019). The interplays of the high contents of fine particles with these high surfaces, whose altitude can reach up to 200-300 m a.s.l, are mainly represented by altitude variables, highlighting the elevation as a great important predictor for the silt fraction.

The high clay content is also closely related to the higher intensity of chemical and physical weathering acting longer in these surfaces. Chemical weathering is mainly associated to the phosphatization process, which has its highest rates over the above-cited islands (Rodrigues et al., 2021; Simas et al., 2007), and conditions intense clay minerals neoformation due to the chemical reactions associated to the input of P-rich birds' guano (Almeida et al., 2021). In turn, the presence of a continuous permafrost over the highlands in MA (Balks et al., 2013) conditions the higher activity of the frost weathering and other periglacial processes in the cryoplanated platforms. Particularly, cryoturbation drives soil particle sorting across the soil profile, resulting in concentration of finer particles in depth and influencing the vertical variability of soil texture in the study area (Siqueira et al., 2021).

Although geology presented a relatively lower importance, the predicted maps show the distribution of the soil texture is closely dependent on the different types of parent materials. Besides the predominance over the unconsolidated sediments of marine terraces, sand contents were also higher in islands with great abundance of tuffs, such as Deception Island, since extremely sandy volcanic soils are formed from this parent material (Vlcek et al., 2018). On the other hand, the most important classes for clay prediction were the marine mudstones and pyritized rocks. The inheritance from mudstones conditions the presence of soils with the highest clay contents found in the study area, located in northwestern Byers Peninsula (Moura et al., 2012). The higher contents of clay in the areas of pyritized rocks evidences the contribution of sulfurization to the clay formation through the extreme weathering associated to pyrite oxidation (Lopes et al., 2019; Siqueira et al., 2022a).

Differently of what was found elsewhere for remote sensing multispectral data (Chagas et al., 2016; Taghizadeh-Mehrjardi et al., 2020), in this work the Sentinel bands had low importance for soil texture prediction. Sentinel bands only had some importance on the clay prediction at the depth 0-10 cm, involving particularly the short-wave-infrared bands – SWIR.

This partially agrees with Gholizadeh et al., 2018 and Taghizadeh-Mehrjardi et al., 2020 who found SWIR bands had a strong influence on the prediction of clay in the topsoil.

In our study, the reflectance of Sentinel 11 band was positively correlated with the clay contents (Supplementary data, Fig. 7), which disagrees with the overall tendency of sandy soils presenting higher reflectance in the shot-wave-infrared. However, this is more valid to light sandy soils with high amounts of quartz (Conforti et al., 2015), as the case of the sandy soils developed from sandstones in NAP. For the predominantly volcanic soils from MA whose sand fraction is dominated by heavy ferromagnesian minerals (black sand), phyllosilicates and light phosphates (clay minerals) (Simas et al., 2006) increase the reflectance at the surface of soils with higher clay contents.

Overall, vegetation indices are the best predictors about the organism factor in soil texture modeling studies (Dharumarajan and Hegde, 2022; Poggio and Gimona, 2017). In this work, the Sentinel ratio 8/4, which, like the vegetation indices, also uses the spectral response of the red and near infrared bands, was the main predictor associated with the vegetation cover. This follows other studies where band ratios have also been successfully used to predict soil texture (Chagas et al., 2016; Gholizadeh et al., 2018). In Antarctica, high contents of clay were also correlated to the areas with extensive plant cover, which can be associated to the larger amounts of organic matter and higher acidity that favors chemical weathering and clay formation in these sites (Otero et al., 2013). Furthermore, many of the present-day vegetated areas correspond to former abandoned penguin rookeries (Tatur and Myrcha, 1989), so bearing greater contents of clay from the phosphatization.

5. CONCLUSIONS

In the present study we created the first digital soil mapping for the ice-free areas of Antarctica, producing a tridimensional soil texture map with high spatial resolution (8 m) for an important region of the continent. To achieve the results, we applied statistical modelling with data-driven ML techniques aiming to predict the contents of sand, silt, and clay for the ice-free areas of MA and NAP. We also used an array of legacy soil texture data produced from pedological fieldworks conducted by the international scientific community, highlighting the Brazilian soil scientist's contribution.

RF was the model which presented the best performance to predict soil texture. The best predictors were associated with climate, topography, and soil development, attending well to the *scorpan* model factors. Higher accuracy was obtained on clay prediction, mainly when predicting in the topsoil. In the maps, sand contents presented higher values, leading to the predominance of the sandy loam texture, although clay contents are higher in some areas where the chemical weathering is stronger. The final maps presented great spatial consistency, with the soil texture distributed according to factors such as geomorphology, parent material and pedogenetic development.

This study is a first contribution to satisfying a crescent need of continuous and accurate soil information for Antarctica, aiming to give subsidies about soil to decision-making toward conservation of the continent. Furthermore, we aim to provide pioneer data that can be incorporated in future environmental models, since the current global models do not incorporate the ice-free areas of Antarctica due to the limited available data and complexity of the continent. This data can be of great value in furnishing supplies to the better understanding of permafrost dynamic in the Antarctic soils, as well as the influence of soil texture on its distribution. This information is even more important in the Antarctic Peninsula and offshore islands, which are the Antarctic zone more threatened by the changes at the global warming context. Our data can also be used as support data in the future scientific expeditions in the region mapped.

REFERENCES

Adhikari, K., Kheir, R.B., Greve, M.B., Bøcher, P.K., Malone, B.P., Minasny, B., McBratney, A.B., Greve, M.H., 2013. High-Resolution 3-D Mapping of Soil Texture in Denmark. Soil Sci. Soc. Am. J. 77, 860–876. https://doi.org/10.2136/sssaj2012.0275

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., 2014. Digital Mapping of Soil Particle-Size Fractions for Nigeria. Soil Sci. Soc. Am. J. 78, 1953–1966. https://doi.org/10.2136/sssaj2014.05.0202

Akumu, C.E., Johnson, J.A., Etheridge, D., Uhlig, P., Woods, M., Pitt, D.G., McMurray, S., 2015. GIS-fuzzy logic based approach in modeling soil texture: Using parts of the Clay Belt and Hornepayne region in Ontario Canada as a case study. Geoderma 239–240, 13–24. https://doi.org/10.1016/j.geoderma.2014.09.021

Almeida, I.C.C., Schaefer, C.E.G.R., Fernandes, R.B.A., Oliveira, F.S., Pereira, T.T.C., 2021. Clay mineralogy and micropedology of phosphate-rich soils from Lions Rump, Maritime Antarctica. J. South Am. Earth Sci. 105. https://doi.org/10.1016/j.jsames.2020.102967

Amirian-Chakan, A., Minasny, B., Taghizadeh-Mehrjardi, R., Akbarifazli, R., Darvishpasand, Z., Khordehbin, S., 2019. Some practical aspects of predicting texture data in digital soil mapping. Soil Tillage Res. 194, 104289. https://doi.org/10.1016/j.still.2019.06.006

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M. d. L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G.L., 2014. Chapter Three: GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. Adv. Agron. 125, 93–134. https://doi.org/10.1016/B978-0-12-800137-0.00003-0

Balks, M.R., López-Martínez, J., Goryachkin, S. V., Mergelov, N.S., Schaefer, C.E.G.R., Simas, F.N.B., Almond, P.C., Claridge, G.G.C., Mcleod, M., Scarrow, J., 2013. Windows on Antarctic soil-landscape relationships: Comparison across selected regions of Antarctica. Geol. Soc. Spec. Publ. 381, 397–410. https://doi.org/10.1144/SP381.9

Ballabio, C., Panagos, P., Monatanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110–123. https://doi.org/10.1016/j.geoderma.2015.07.006

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25, 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modeling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91, 27–45. https://doi.org/10.1016/S0016-7061(99)00003-8

Bishop, J.L., Fairén, A.G., Michalski, J.R., Gago-Duport, L., Baker, L.L., Velbel, M.A., Gross, C., Rampe, E.B., 2018. Surface clay formation during short-term warmer and wetter conditions on a largely cold ancient Mars. Nat. Astron. 2, 206–213. https://doi.org/10.1038/s41550-017-0377-9

Bockheim, J.G., 2015. The soils of Antarctica. Springer New York, New York. https://doi.org/10.1007/978-3-319-05497-1

Blumel, W.D., Eitel B., 1989. Geoecological aspects of maritime-climatic and continental periglacial regions in Antarctica (South Shetlands, Antarctic Peninsula and Victoria Land). Geökodynamic 10: 201-214.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Brenning, A., 2008. Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models, in: SAGA--Seconds Out (Hamburger Beiträge Zur Physischen Geographie Und Landschaftsökologie). pp. 23–32.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. ICML 2006 - Proc. 23rd Int. Conf. Mach. Learn. 2006, 161–168.

Carvalho Junior, W. de, Lagacherie, P., da Silva Chagas, C., Calderano Filho, B., Bhering, S.B., 2014. A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. Geoderma 232–234, 479–486. https://doi.org/10.1016/j.geoderma.2014.06.007

Caubet, M., Román Dobarco, M., Arrouays, D., Minasny, B., Saby, N.P.A., 2019. Merging country, continental and global predictions of soil texture: Lessons from ensemble modelling in France. Geoderma 337, 99–110. https://doi.org/10.1016/j.geoderma.2018.09.007

Chagas, C. da S., de Carvalho Junior, W., Bhering, S.B., Calderano Filho, B., 2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena 139, 232–240. https://doi.org/10.1016/j.catena.2016.01.001

Conforti M, Froio R, Matteucci G, Buttafuoco G. Visible and near infrared spectroscopy for predicting texture in forest soil: An application in southern italy. IForest - Biogeosciences For. 2015;8:339-47. https://doi.org/10.3832/ifor1221-007

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/BF00994018

Daher, M., Schaefer, C.E.G.R., Fernandes Filho, E.I., Francelino, M.R., Senra, E.O., 2019a. Semi-arid soils from a topolithosequence at James Ross Island, Weddell Sea region, Antarctica: Chemistry, mineralogy, genesis and classification. Geomorphology 327, 351–364. https://doi.org/10.1016/j.geomorph.2018.11.003

Daher, M., Schaefer, C.E.G.R., Thomazini, A., Lima Neto, E., Souza, C.D., Lopes, D. do V., 2019b. Ornithogenic soils on basalts from maritime Antarctica. Catena 173, 367–374. https://doi.org/10.1016/j.catena.2018.10.028

Davies, B.J., Glasser, N.F., Carrivick, J.L., Hambrey, M.J., Smellie, J.L., Nývlt, D., 2013. Landscape evolution and ice-sheet behaviour in a semi-arid polar environment: James Ross Island, NE Antarctic Peninsula. Geol. Soc. Spec. Publ. 381, 353–395. https://doi.org/10.1144/SP381.1

Delpupo, C.S., Schaefer, C.E.G.R., Simas, F.N.B., Spinola, D.N., Daher, M., 2014. Soil formation in Seymour Island , Weddell Sea , Antarctica. Geomorphology 255, 87–99. https://doi.org/10.1016/j.geomorph.2014.03.047

Díaz-Puente, F.J., Schmid, T., Pelayo, M., Rodríguez-Rastrero, M., Herraiz, M.J.S., O'Neill, T., López-Martínez, J., 2021. Abiotic factors influencing soil microbial activity in the northern Antarctic Peninsula region. Sci. Total Environ. 750. https://doi.org/10.1016/j.scitotenv.2020.141602

Dharumarajan, S., Hegde, R., 2022. Digital mapping of soil texture classes using Random Forest classification algorithm. Soil Use Manag. 38, 135–149. https://doi.org/10.1111/sum.12668

Fernandes-Filho E., Moquedace C., Veloso G., 2021. mdsFuncs: Auxiliary functions to pre-process raster data and modelling with machine learning models using caret package_. R package version 0.2.10.

Ferrari, F.R., Schaefer, C.E.G.R., Pereira, A.B., Schmitz, D., Francelino, M.R., 2020. Coupled soil-vegetation changes along a topographic gradient on King George Island , maritime Antarctica. Catena 105038. https://doi.org/10.1016/j.catena.2020.105038

Francelino, M.R., Schaefer, C.E.G.R., Simas, F.N.B., Filho, E.I.F., de Souza, J.J.L.L., da Costa, L.M., 2011. Geomorphology and soils distribution under paraglacial conditions in an ice-free area of Admiralty Bay, King George Island, Antarctica. Catena 85, 194–204. https://doi.org/10.1016/j.catena.2010.12.007

Fretwell, P.T., Hodgson, D.A., Watcham, E.P., Bentley, M.J., Roberts, S.J., 2010. Holocene isostatic uplift of the South Shetland Islands, Antarctic Peninsula, modelled from raised beaches. Quat. Sci. Rev. 29, 1880–1893. https://doi.org/10.1016/j.quascirev.2010.04.006

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Ganzert, L., Lipski, A., Hubberten, H.W., Wagner, D., 2011. The impact of different soil parameters on the community structure of dominant bacteria from nine different soils located on Livingston Island, South Shetland Archipelago, Antarctica. FEMS Microbiol. Ecol. 76, 476–491. https://doi.org/10.1111/j.1574-6941.2011.01068.x

Gholizadeh, A., Daniel, Ž., Saberioon, M., 2018. Soil organic carbon and texture retrieving and mapping using proximal , airborne and Sentinel-2 spectral imaging. Remote Sens. Environ. 218, 89–103. https://doi.org/10.1016/j.rse.2018.09.015

Gjorup, D.F., Ernesto, C., Schaefer, G.R., Nogueira, F., Simas, B., Rocha, M., Ferreira, R., Michel, M., Bockheim, J.G., 2020. Sulfurization , Acid-Sulfate soils and active layer monitoring at the semiarid Seymour Island , Antarctica. Geoderma Reg. 22, e00305. https://doi.org/10.1016/j.geodrs.2020.e00305

Gomes, L.C., Faria, R.M., de Souza, E., Veloso, G.V., Schaefer, C.E.G.R., Filho, E.I.F., 2019. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma 340, 337–350. https://doi.org/10.1016/j.geoderma.2019.01.007

González-Guzmán, A., Oliva, M., Souza-Júnior, V.S., Pérez-Alberti, A., Ruiz-Fernández, J., Otero, X.L., 2017. Biota and geomorphic processes as key environmental factors controlling soil formation at Elephant Point, Maritime Antarctica. Geoderma 300, 32–43. https://doi.org/10.1016/j.geoderma.2017.04.001

Greve, M.H., Kheir, R.B., Greve, M.B., Bøcher, P.K., 2012. Quantifying the ability of environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: The case study of Denmark. Ecol. Indic. 18, 1–10. https://doi.org/10.1016/j.ecolind.2011.10.006

Haus, N.W., Wilhelm, K.R., Bockheim, J.G., Fournelle, J., Miller, M., 2016. A case for chemical weathering in soils of Hurd Peninsula, Livingston Island, South Shetland Islands, Antarctica. Geoderma 263, 185–194. https://doi.org/10.1016/j.geoderma.2015.09.019

Hechenbichler, K., Schliep, K., 2004. Weighted k-Nearest-Neighbor Techniques and Ordinal Classification. Mol. Ecol. 399, 17.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., De Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS One 10, 1–26. https://doi.org/10.1371/journal.pone.0125814

Hitziger, M., Ließ, M., 2014. Comparison of three supervised learning methods for digital soil mapping: Application to a complex terrain in the Ecuadorian Andes. Appl. Environ. Soil Sci. 2014. https://doi.org/10.1155/2014/809495

Hossain, M.S., Rahman, G.K.M.M., Alam, M.S., Rahman, M.M., Solaiman, A.R.M., Mia, M.A.B., 2018. Modelling of soil texture and its verification with related soil properties. Soil Res. 56, 421. https://doi.org/10.1071/sr17252

Hounkpatin, O.K.L., Op de Hipt, F., Bossa, A.Y., Welp, G., Amelung, W., 2018. Soil organic carbon stocks and their determining factors in the Dano catchment (Southwest Burkina Faso). Catena 166, 298–309. https://doi.org/10.1016/j.catena.2018.04.013

Howat, I.M., Porter, C., Smith, B.E., Noh, M.J., Morin, P., 2019. The reference elevation model of antarctica. Cryosphere 13, 665–674. https://doi.org/10.5194/tc-13-665-2019

Hrbáček, F., Nývlt, D., Láska, K., 2017. Active layer thermal dynamics at two lithologically different sites on James Ross Island, Eastern Antarctic Peninsula. Catena 149, 592–602. https://doi.org/10.1016/j.catena.2016.06.020

Ilieva, R., Yaneva, R., Zhiyanski, M., Abakumov, E., 2021. Morphological and physico-chemical properties of Cryosoils in the Bulgarian antarctic base on Livingston island, Antarctica. Silva Balc. 22, 57–67. https://doi.org/10.3897/silvabalcanica.22.e66704

Ingólfsson, Ó., Hjort, C., 2002. Glacial history of the Antarctic Peninsula since the Last Glacial Maximum–a synthesis. Polar Res. 21, 227–234. https://doi.org/10.3402/polar.v21i2.6482

John, B.S., Sugden, D.E., 1971. Raised marine features and phases of glaciation in the South Shetland Islands. Br. Antarct. Surv. Bull. 24, 45–111.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2017. Climatologies at high resolution for the earth's land surface areas. Sci. Data 4, 1–20. https://doi.org/10.1038/sdata.2017.122

Kelleher, J.D., MacNamee, J., D'Arcy, A., 2015. Fundamentals of machine learning for predictive data analytics : algorithms, worked examples, and case studies.

Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. Appl. Math. Model. 81, 401–418. https://doi.org/10.1016/j.apm.2019.12.016

Koseva, I.S., Watmough, S.A., Aherne, J., 2010. Estimating base cation weathering rates in Canadian forest soils using a simple texture-based model. Biogeochemistry 101, 183–196. https://doi.org/10.1007/s10533-010-9506-6

Krauze, P., Wagner, D., Yang, S., Spinola, D., Kühn, P., 2021. Influence of prokaryotic microorganisms on initial soil formation along a glacier forefield on King George Island, maritime Antarctica. Sci. Rep. 11, 1–16. https://doi.org/10.1038/s41598-021-92205-z

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., 2021. caret: Classification and Regression Training.

Lark, R.M., Bishop, T.F.A., 2007. Cokriging particle size fractions of the soil. Eur. J. Soil Sci. 58, 763–774. https://doi.org/10.1111/j.1365-2389.2006.00866.x

Łachacz, A., Kalisz, B., Giełwanowska, I., Olech, M., Chwedorzewska, K.J., Kellmann-Sopyła, W., 2018. Nutrient abundance and variability from soils in the coast of king George Island. J. Soil Sci. Plant Nutr. 18, 294–311. https://doi.org/10.4067/S0718-95162018005001101

Lang, B., 2005. Monotonic multi-layer perceptron networks as universal approximators. Lect. Notes Comput. Sci. 3697 LNCS, 31–37. https://doi.org/10.1007/11550907_6

Lennon, J.J., 2000. Red-shifts and red herrings in geographical ecology. Ecography (Cop.). 23, 101–113. https://doi.org/10.1111/j.1600-0587.2000.tb00265.x

Li, X., Liu, J., Zhang, J., Wang, W., Xin, W., 2014. Soil texture distribution simulation and risk assessment using transition probability-based geostatistics. Int. Agrophysics 28, 447–457. https://doi.org/10.2478/intag-2014-0035

Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture. Comparison of regression tree and Random Forest models. Geoderma 170, 70–79. https://doi.org/10.1016/j.geoderma.2011.10.010

Lima Neto, E., Guerra, M.B.B., Thomazini, A., Daher, M., de Andrade, A.M., Schaefer, C.E.G.R., 2017. Soil Contamination by Toxic Metals Near an Antarctic Refuge in Robert Island, Maritime Antarctica: A Monitoring Strategy. Water. Air. Soil Pollut. 228. https://doi.org/10.1007/s11270-017-3245-4

Lin, J., Rayhan, A.B.M.S., Wang, Y., Wu, Z., Lin, Y., Ke, H., Li, T., Chen, K., Cai, M., 2021. Distribution and contamination assessment of heavy metals in soils and sediments from the

Fildes Peninsula and Ardley Island in King George Island, Antarctica. Polar Res. 40. https://doi.org/10.33265/POLAR.V40.5270

Lopes, D. do V., Schaefer, C.E.G.R., Souza, J.J.L.L. de, Oliveira, F.S. de, Simas, F.N.B., Daher, M., Gjorup, D.F., 2019. Concretionary horizons, unusual pedogenetic processes and features of sulfate affected soils from Antarctica. Geoderma 347, 13–24. https://doi.org/10.1016/j.geoderma.2019.03.024

Lopes, D. do V., Souza, J.J.L.L. de, Simas, F.N.B., Oliveira, F.S. de, Schaefer, C.E.G.R., 2021. Hydrogeochemistry and chemical weathering in a periglacial environment of Maritime Antarctica. Catena 197, 104959. https://doi.org/10.1016/j.catena.2020.104959

Lopes, D.D.V., De Oliveira, F.S., De Souza, J.J.L.L., Machado, M.D.R., Schaefer, C.E.G.R., 2022a. Soil pockets phosphatization and chemical weathering of sites affected by flying birds of Maritime Antarctica. An. Acad. Bras. Cienc. 94. https://doi.org/10.1590/0001-3765202220210595

Lopes, D. do V., Oliveira, F.S. de, Pereira, T.T.C., Schaefer, C.E.G.R., 2022b. Pedogeomorphology and weathering at Snow Island, Maritime Antarctica. Catena 217. https://doi.org/10.1016/j.catena.2022.106515

López-Martínez, J., Serrano, E., Schmid, T., Mink, S., Linés, C., 2012. Periglacial processes and landforms in the South Shetland Islands (northern Antarctic Peninsula region). Geomorphology 155–156, 62–79. https://doi.org/10.1016/j.geomorph.2011.12.018

Lupachev, A. V., Abakumov, E. V., Goryachkin, S. V., Veremeeva, A.A., 2020. Soil cover of the Fildes Peninsula (King George Island, West Antarctica). Catena 193, 104613. https://doi.org/10.1016/j.catena.2020.104613

Ma, Y., Minasny, B., Wu, C., 2017. Mapping key soil properties to support agricultural production in Eastern China. Geoderma Reg. 10, 144–153. https://doi.org/10.1016/j.geodrs.2017.06.002

Ma, Y., Minasny, B., McBratney, A., Poggio, L., Fajardo, M., 2021. Predicting soil properties in 3D: Should depth be a covariate? Geoderma 383, 114794. https://doi.org/10.1016/j.geoderma.2020.114794

Malone, B.P., McBratney, A.B., Minasny, B., 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160, 614–626. https://doi.org/10.1016/j.geoderma.2010.11.013

Malone, B., Searle, R., 2021. Updating the Australian digital soil texture mapping (Part 1. Soil Res. 59, 419–434. https://doi.org/10.1071/SR20283

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., Beaudoin, A., 2014. Digital mapping of soil properties in Canadian managed forests at 250m of resolution using the k-nearest neighbor method. Geoderma 235–236, 59–73. https://doi.org/10.1016/j.geoderma.2014.06.032

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping, Geoderma. https://doi.org/10.1016/S0016-7061(03)00223-4

Meier, L.A., Krauze, P., Prater, I., Horn, F., Schaefer, C.E.G.R., Scholten, T., Wagner, D., Mueller, C.W., Kühn, P., 2019. Pedogenic and microbial interrelation in initial soils under semiarid climate on James Ross Island, Antarctic Peninsula region. Biogeosciences 16, 2481–2499. https://doi.org/10.5194/bg-16-2481-2019

Meier, M., Francelino, M.R., Thomazini, A., Pereira, A.B., Krüger, F.L. von, Fernandes-Filho, E.I., Schaefer, C.E.G.R., 2023. Soilscapes and Geoenvironments at Stansbury Peninsula, Nelson Island, Maritime Antarctica. Catena. https://doi.org/10.2139/ssrn.3959649

Mendonça, T., Melo, V.F., Schaefer, C.E.G.R., Simas, F.N.B., Michel, R.F.M., 2013. Clay Mineralogy of Gelic Soils from the Fildes Peninsula, Maritime Antarctica. Soil Science Society of America Journal 77, 1842–1851. https://doi.org/10.2136/sssaj2012.0135

Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. Ecol. Modell. 411, 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815

Michel, R.F.M., Schaefer, C.E.G.R., Dias, L.E., Simas, F.N.B., Benites, Vinicius de Melo Mendonça, E. de S., 2006. Ornithogenic Gelisols (Cryosols) from Maritime Antarctica: Pedogenesis, Vegetation, and Carbon Studies. Soil Sci. Soc. Am. 70, 1370–1376. https://doi.org/10.2136/sssaj2005.0178

Michel, R.F.M., Schaefer, C.E.G.R., López-Martínez, J., Simas, F.N.B., Haus, N.W., Serrano, E., Bockheim, J.G., 2014. Soils and landforms from Fildes Peninsula and Ardley Island, Maritime Antarctica. Geomorphology 225, 76–86. https://doi.org/10.1016/j.geomorph.2014.03.041

Mills, A.J., Fey, M. V., Gröngröft, A., Petersen, A., Medinski, T. V., 2006. Unravelling the effects of soil properties on water infiltration: Segmented quantile regression on a large data set from arid south-west Africa. Aust. J. Soil Res. 44, 783–797. https://doi.org/10.1071/SR05180

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Moeys, J., Shangguan, W., Petzold, R., Minasny, B., Rosca, B., Jelinski, N., Zelazny, W., Souza, R.M.S., Safanelli, J.L., ten Caten, A., 2022. Functions for Soil Texture Plot, Classification and Transformation. R package version 1.5.1.

Moura, P.A., Francelino, M.R., Schaefer, C.E.G.R., Simas, F.N.B., Mendonça, B.A.F. de, 2012. Distribution and characterization of soils and landform relationships in Byers Peninsula, Livingston Island, Maritime Antarctica. Geomorphology 155–156, 45–54. https://doi.org/10.1016/j.geomorph.2011.12.011

Nauman, T.W., Duniway, M.C., 2019. Relative prediction intervals reveal larger uncertainty in 3D approaches to predictive digital soil mapping of soil properties with legacy data. Geoderma 347, 170–184. https://doi.org/10.1016/j.geoderma.2019.03.037

Navas, A., Soto, J., López-Martínez, J., 2005. Radionuclides in soils of Byers Peninsula, South Shetland Islands, Western Antarctica. Appl. Radiat. Isot. 62, 809–816. https://doi.org/10.1016/j.apradiso.2004.11.007

Navas, A., López-Martínez, J., Casas, J., Machín, J., Durán, J.J., Serrano, E., Cuchi, J.A., Mink, S., 2008. Soil characteristics on varying lithological substrates in the South Shetland Islands, maritime Antarctica. Geoderma 144, 123–139. https://doi.org/10.1016/j.geoderma.2007.10.011

Navas, A., Oliva, M., Ruiz-Fernández, J., Gaspar, L., Quijano, L., Lizaga, I., 2017. Radionuclides and soil properties as indicators of glacier retreat in a recently deglaciated permafrost environment of the Maritime Antarctica. Sci. Total Environ. 609, 192–204. https://doi.org/10.1016/j.scitotenv.2017.07.115

Navas, A., Serrano, E., López-Martínez, J., Gaspar, L., Lizaga, I., 2018. Interpreting environmental changes from radionuclides and soil characteristics in different landform contexts of Elephant Island (maritime Antarctica). L. Degrad. Dev. 29, 3141–3158. https://doi.org/10.1002/ldr.2987

Nelder, A., Wedderburn, R., 1972. Generalized Linear Models. J. R. Stat. Soc. 135, 370–384.

Nelson, P.H.H., 1975. The James Ross Island Volcanic Group of North-East Graham Land, British Antarctic Survey Scientific Reports.

Neufeld, Â.D.H., de Godoi, S.G., Pereira, A.B., Bayer, C., Schünemann, A.L., Victoria, F. de C., de Albuquerque, M.P., Camargo, E., Vieira, F.C.B., 2015. Methane and nitrous oxide fluxes in relation to vegetation covers and bird activity in ice-free soils of Rip Point, Nelson Island, Antarctica. Polar Res. 34, 1–12. https://doi.org/10.3402/polar.v34.23584

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil 4, 1–22. https://doi.org/10.5194/soil-4-1-2018

Oliveira, G. de C., Arruda, D.M., Fernandes Filho, E.I., Veloso, G.V., Francelino, M.R., Schaefer, C.E.G.R., 2021. Soil predictors are crucial for modelling vegetation distribution and its responses to climate change. Sci. Total Environ. 780, 146680. https://doi.org/10.1016/j.scitotenv.2021.146680

Otero, X.L., Fernández, S., de Pablo Hernandez, M.A., Nizoli, E.C., Quesada, A., 2013. Plant communities as a key factor in biogeochemical processes involving micronutrients (Fe, Mn, Co, and Cu) in Antarctic soils (Byers Peninsula, maritime Antarctica). Geoderma 195–196, 145–154. https://doi.org/10.1016/j.geoderma.2012.11.018

Pereira, T.T.C., Schaefer, C.E.G.R., Ker, J.C., Almeida, C.C., Almeida, I.C.C., Pereira, A.Ô.B., 2013. Genesis, mineralogy and ecological significance of ornithogenic soils from a semi-desert polar landscape at Hope Bay, Antarctic Peninsula. Geoderma 209–210, 98–109. https://doi.org/10.1016/j.geoderma.2013.06.012

Phillips, J.D., 2004. Geogenesis, pedogenesis, and multiple causality in the formation of texture-contrast soils. Catena 58, 275–295. https://doi.org/10.1016/j.catena.2004.04.002

Pires, C. V., Schaefer, C.E.R.G., Hashigushi, A.K., Thomazini, A., Filho, E.I.F., Mendonça, E.S., 2017. Soil organic carbon and nitrogen pools drive soil C-CO2 emissions from selected soils in Maritime Antarctica. Sci. Total Environ. 596–597, 124–135. https://doi.org/10.1016/j.scitotenv.2017.03.144

Pirrie, D., 1994. Petrography and provenance of the Marambio Group, Vega Island, Antarctica. Antarct. Sci. 6, 517–527. https://doi.org/10.1017/S0954102094000775

Poelking, E.L., Schaefer, C.E.R., Fernandes Filho, E.I., Andrade, A.M., Spielmann, A.A., 2015. Soil – landform – plant-community relationships of a periglacial landscape on Potter Peninsula, maritime Antarctica. Solid Earth 583–594. https://doi.org/10.5194/se-6-583-2015

Poggio, L., Gimona, A., 2017. 3D mapping of soil texture in Scotland. Geoderma Reg. 9, 5–16. https://doi.org/10.1016/j.geodrs.2016.11.003

R Core Team, 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Riza, S., Sekine, M., Kanno, A., Yamamoto, K., Imai, T., Higuchi, T., 2021. Modeling soil landscapes and soil textures using hyperscale terrain attributes. Geoderma 402, 115177. https://doi.org/10.1016/j.geoderma.2021.115177

Rodrigues, W.F., Oliveira, F.S., Schaefer, C.E.G.R., Leite, M.G.P., Gauzzi, T., Bockheim, J.G., Putzke, J., 2019. Soil-landscape interplays at Harmony Point, Nelson Island, Maritime Antarctica: Chemistry, mineralogy and classification. Geomorphology 336, 77–94. https://doi.org/10.1016/j.geomorph.2019.03.030

Rodrigues, W.F., Soares, F. de O., Schaefer, C.E.G.R., Leite, M.G.P., Pavinato, P.S., 2021. Phosphatization under birds' activity: Ornithogenesis at different scales on Antarctic Soilscapes. Geoderma 391. https://doi.org/10.1016/j.geoderma.2021.114950

Román Dobarco, M., Arrouays, D., Lagacherie, P., Ciampalini, R., Saby, N.P.A., 2017. Prediction of topsoil texture for Region Centre (France) applying model ensemble methods. Geoderma 298, 67–77. https://doi.org/10.1016/j.geoderma.2017.03.015

Schaefer, C. E. G. R., Costa Pereira, T.T., Ker, J.C., Carreiro Almeida, I.C., Bello Simas, F.N., Soares de Oliveira, F., Corrêa, G.R., Vieira, G., 2015. Soils and Landforms at Hope Bay, Antarctic Peninsula: Formation, Classification, Distribution, and Relationships. Soil Sci. Soc. Am. J. 79, 175–184. https://doi.org/10.2136/sssaj2014.06.0266

Schaefer, C. E. G. R., Michel, R.F.M., Delpupo, C., Senra, E.O., Bremer, U.F., Bockheim, J.G., 2017. Active layer thermal monitoring of a Dry Valley of the Ellsworth Mountains, Continental Antarctica. Catena 149, Part 2, 603-615. https://doi.org/10.1016/j.catena.2016.07.020

Schmitz, D., Schaefer, C.E.R.G., Putzke, J., Francelino, M.R., Ferrari, F.R., Corrêa, G.R., Villa, P.M., 2020a. How does the pedoenvironmental gradient shape non-vascular species assemblages and community structures in Maritime Antarctica? Ecol. Indic. 108, 105726. https://doi.org/10.1016/j.ecolind.2019.105726

Schmitz, D., Villa, P.M., Putzke, J., Michel, R.F.M., Campos, P.V., Neto, J.A.A.M., Schaefer, C.E.G.R., 2020b. Diversity and species associations in cryptogam communities along a pedoenvironmental gradient on Elephant Island, Maritime Antarctica. Folia Geobot. 55, 211–224. https://doi.org/10.1007/s12224-020-09376-2

Schmitz, D., Villa, P.M., Michel, R.F.M., Putzke, J., Pereira, A.B., Schaefer, C.E.G.R., 2021. Species composition, diversity and coverage pattern of associated communities of mosses-lichens along a pedoenvironmental gradient in Maritime Antarctica. An. Acad. Bras. Cienc. 94, 1–17. https://doi.org/10.1590/0001-3765202120200094

Schmitz, D., Michel, R.F.M., Ferrari, F.R., Villa, P.M., Francelino, M.R., Putzke, J., López-Martínez, J., Schaefer, C.E.G.R., 2022. Soil-landform-vegetation interplays at Stinker Point, Elephant Island, Antarctica. An. Acad. Bras. Cienc. 94, 1–24. https://doi.org/10.1590/0001-3765202220210676

Schünemann, A.L., Almeida, P.H.A., Francelino, M.R., Fernandes-Filho, E.I., Santos, G.R. dos, Daher, M., Schaefer, C.E.G.R., Pereira, A.B., 2022. Soil pedogeochemical attributes prediction by interpolators in ice-free areas of Antarctica. Res. Soc. Dev. 11, 1–15.

Seyedmohammadi, J., Navidi, M.N., Esmaeelnejad, L., 2019. Geospatial modeling of surface soil texture of agricultural land using fuzzy logic, geostatistics and GIS techniques. Commun. Soil Sci. Plant Anal. 50, 1452–1464. https://doi.org/10.1080/00103624.2019.1626870

Silva, J.P., Souza, J.J.L.L. de, Soares, E.M.B., Schaefer, C.E.G.R., 2022. Soil organic matter accumulation before, during, and after the last glacial maximum in Byers Peninsula, Maritime Antarctica. Geoderma 428, 116221. https://doi.org/10.1016/j.geoderma.2022.116221

Simas, F., Schaefer, C.E.G.R., Melo, V.F., Albuquerque-Filho, M.R., Michel, R.F.M., Pereira, V. V., Gomes, M.R.M., Costa, L.M., 2007. Ornithogenic cryosols from Maritime Antarctica: Phosphatization as a soil forming process. Geoderma 138, 191–203.

Simas, F.N.B., Schaefer, C.E.G.R., Melo, V.F., Guerra, M.B.B., Saunders, M., Gilkes, R.J., 2006. Clay-sized minerals in permafrost-affected soils (Cryosols) from King George Island, Antarctica. Clays Clay Miner. 54, 721–736. https://doi.org/10.1346/CCMN.2006.0540607

Simas, F.N.B., Schaefer, C.E.G.R., Albuquerque-Filho, M.R., Francelino, M.R., Fernandes-Filho, E.I., Costa, L.M. da, 2008. Genesis, properties and classification of Cryosols from Admiralty Bay, maritime Antarctica. Geoderma 144, 116–122. https://doi.org/10.1016/j.geoderma.2007.10.019

Siqueira, R.G., Schaefer, C.E.G.R., Fernandes Filho, E.I., Corrêa, G.R., Francelino, M.R., Souza, J.J.L.L. de, Rocha, P. de A., 2021. Weathering and pedogenesis of sediments and basaltic rocks on Vega Island , Antarctic Peninsula. Geoderma 382, 114707. https://doi.org/10.1016/j.geoderma.2020.114707

Siqueira, R.G., Lopes, D. V., De Souza, J.J.L.L., Schaefer, C.E.G.R., Souza, C.D., De Oliveira, F.S., Fernandes Filho, E.I., 2022a. Acid sulfate soils from Antarctica: Genesis and properties along a climatic gradient. An. Acad. Bras. Cienc. 94, 1–20. https://doi.org/10.1590/0001-3765202120210625

Siqueira, R.G., Veloso, G. V., Fernandes-Filho, E.I., Francelino, M.R., Schaefer, C.E.G.R., Corrêa, G.R., 2022b. Evaluation of machine learning algorithms to classify and map landforms in Antarctica. Earth Surf. Process. Landforms 47, 367–382. https://doi.org/10.1002/esp.5253

Smellie, J.L., 2013. Geological Map of James Ross Island 1 . James Ross Island Volcanic Group. BAS GEOMAP 2 Ser. Sheet 5, Br. Ant− Arct. Surv. Cambridge. 110. https://doi.org/10.1029/2006GC001450.20.

Smellie, J.L., Hunt, R.J., McIntosh, W.C., Esser, R.P., 2021a. Lithostratigraphy, age and distribution of Eocene volcanic sequences on eastern King George Island, South Shetland Islands, Antarctica. Antarct. Sci. 33, 373–401. https://doi.org/10.1017/S0954102021000213

Smellie, J.L., McIntosh, W.C., Whittle, R., Troedson, A., Hunt, R.J., 2021b. A lithostratigraphical and chronological study of Oligocene-Miocene sequences on eastern King George Island, South Shetland Islands (Antarctica), and correlation of glacial episodes with global isotope events. Antarct. Sci. 33, 502–532. https://doi.org/10.1017/S095410202100033X

Smellie, J.L., Pankhurst, R.J., Thomson, M.R.A., Davies, R.E.S., 1984. The geology of the South Shetland Islands: VI. Stratigraphy, geochemistry and evolution., Scientific Reports - British Antarctic Survey.

Soil Survey Staff, 2014. Keys to Soil Taxonomy, 12th ed.

Sousa, L., Poggio, L., Batjes, N., Heuvelink, G., Kempen, B., Riberio, E., Rossiter, D., 2020. SoilGrids 2.0: producing quality-assessed soil information for the globe. SOIL Discuss. 1–37.

Taghizadeh-Mehrjardi, R., Mahdianpari, M., Mohammadimanesh, F., Behrens, T., Toomanian, N., Scholten, T., Schmidt, K., 2020. Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran. Geoderma 376, 114552. https://doi.org/10.1016/j.geoderma.2020.114552

Tatur A, Myrcha A. Soils and Vegetation in Abandoned Penguin Rookeries (Maritime Antarctic). Polar Biol. 1989;2:181-9

Thomazini, A., Francelino, M.R., Pereira, A.B., Schünemann, A.L., Mendonça, E.S., Almeida, P.H.A., Schaefer, C.E.G.R., 2016. Geospatial variability of soil CO2-C exchange in the main terrestrial ecosystems of Keller Peninsula, Maritime Antarctica. Sci. Total Environ. 562, 802–811. https://doi.org/10.1016/j.scitotenv.2016.04.043

van Lipzig, N.P.M., King, J.C., Lachlan-Cope, T.A., van den Broeke, M.R., 2004. Precipitation, sublimation, and snow drift in the Antarctic Peninsula region from a regional atmospheric model. J. Geophys. Res. 109, 1–16. https://doi.org/10.1029/2004JD004701

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64. https://doi.org/10.1016/j.geoderma.2016.12.017

Vignon, E., Roussel, M.-L., Gorodetskaya, I.V., Genthon C., Berne, A. 2021. Present and Future of Rainfall in Antarctica. Geophys. Res. Lett. https://doi.org/10.1029/2020GL092281.

Vieira, F.C.B., Pereira, A.B., Bayer, C., Schünemann, A.L., Victoria, F. de C., de Albuquerque, M.P., de Oliveira, C.S., 2013. In situ methane and nitrous oxide fluxes in soil from a transect in Hennequin Point, King George Island, Antarctic. Chemosphere 90, 497–504. https://doi.org/10.1016/j.chemosphere.2012.08.013

Vlcek, V., Pospíšilová, L., Uhlík, P., 2018. Mineralogy and chemical composition of cryosols and andosols in Antarctica. Soil Water Res. 13, 61–73. https://doi.org/10.17221/231/2016-SWR

Wadoux, A.M.J.C., Heuvelink, G.B.M., de Bruin, S., Brus, D.J., 2021. Spatial cross-validation is not the right way to evaluate map accuracy. Ecol. Modell. 457. https://doi.org/10.1016/j.ecolmodel.2021.109692

Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2014. Holistic environmental soil-landscape modeling of soil organic carbon. Environ. Model. Softw. 57, 202–215. https://doi.org/10.1016/j.envsoft.2014.03.004

Yang, R.M., Zhang, G.L., Liu, F., Lu, Y.Y., Yang, Fan, Yang, Fei, Yang, M., Zhao, Y.G., Li, D.C., 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecol. Indic. 60, 870–878. https://doi.org/10.1016/j.ecolind.2015.08.036

Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., Finke, P., 2019. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. Geoderma 338, 445–452. https://doi.org/10.1016/j.geoderma.2018.09.006

Zhao, Z., Chow, T.L., Rees, H.W., Yang, Q., Xing, Z., Meng, F.R., 2009. Predict soil texture distributions using an artificial neural network model. Comput. Electron. Agric. 65, 36–48. https://doi.org/10.1016/j.compag.2008.07.008

Zinsmeister, W.J., 1982. Review of the Upper Cretaceous-Lower Tertiary sequence on Seymour Island, Antarctica. J. Geol. Soc. London. 139, 779–785.

CAPÍTULO 2

MODELLING AND SPATIAL PREDICTION OF MAJOR SOIL CHEMICAL PROPERTIES IN ANTARCTICA USING RANDOM FOREST

ABSTRACT

The Digital Soil Mapping advantages, including quantitative estimations of soil attributes in areas with little or no soil information, are even more important in remote, harsh landscapes such as the Antarctic ice-free areas. Our objective in this work was to predict the spatial distribution of major soil chemical attributes in the Northern Antarctic Peninsula and Maritime Antarctic regions; as well as to identify the main environmental drivers of these attributes. For this, we compiled one of the largest soil databases of the Antarctic continent and applied the machine learning algorithm Random Forest. We predicted seven soil attributes, and they can be listed as follows from the highest performance to the lowest: bases sum (mean R^2 of 0.41 and maximum of 0.48), H+Al, pH, total organic carbon, remaining P (P-rem), Na, and available P (mean R^2 of 0.25). The pH and P-rem presented the lowest prediction uncertainty (mean PIR of 0.07 and 0.13, respectively), whereas the Na and P presented the highest (mean PIR of 0.85 and 1.24). Both error and uncertainty indicators show the difficulties of the Random Forest in handling natural outliers. In general, the spatial variability of the soil attributes were mainly related to the distribution of the mean annual temperature and annual precipitation climatic covariates, vegetation indexes and multispectral bands related to the vegetation response, the rookeries distance expressing the birds' activity, and topographic attributes associated to the altitude. Overall, the soil attributes showed a climatic gradient, with higher values of bases sum, pH, and Na in Northern Antarctic Peninsula, while the highest values of TOC, H+Al and P were present in Maritime Antarctica. This clearly shows the climate covariates influence the soil distribution in a macroscale, whereas the terrain predictors control the soil variability at local scales through strong soil-landscape relationships. This study shows the potential of applying the Digital Soil Mapping methodology to produce detailed soil maps in Antarctica, aiming to improve soil information as a support for environmental policy in the continent.

Keywords: Antarctic Peninsula. Digital Soil Mapping. Machine Learning. Maritime Antarctica. Soil attributes.

1. INTRODUCTION

Digital Soil Mapping (DSM) has shifted from an emerging sub-discipline to one of the most relevant fields of soil science (Ma et al., 2019; Minasny et al., 2014). DSM growth has been prompted by advances in computation, population of new data produced by remote and on-the-ground sensors and the growing global demand for soil data (McBratney et al., 2000; Nelson et al., 2011). According to Lagacherie and McBratney, 2006, DSM can be defined as the creation of soil spatial information systems coupling field and laboratory observational methods with spatial, computational inference systems. So, the DSM aims to model spatial or spatial-temporal soil distribution from the use of quantitative techniques, making estimations even in areas with little or no soil information (McBratney et al., 2003; Minasny and McBratney, 2016).

The DSM can be subdivided into five basic steps, which are important to describe its nature: 1) the input of soil observations in the form of soil legacy data or new samples; 2) the selection of spatial, environmental covariates to represent the scorpan factors (McBratney et al., 2003); c) the building of statistical or mathematical models to quantitatively relate the soil observations and covariates; d) the model assessment through performance and uncertainty issues; and e) the use of the fitted models to predict in space the soil distribution as a function of the covariates (Ma et al., 2019; Minasny et al., 2014; Thompson et al., 2012).

Many computational techniques have been developed to map soils inside the scope of DSM, including knowledge-based models, soil geostatistics and the most common spatially factorial models, also known as environmental correlation approaches (Minasny and McBratney, 2016). In environmental correlation, many works apply Machine Learning (ML) techniques (Khaledian and Miller, 2020), as powerful tools to implement the DSM principles by algorithms which independently fit to the data to extract the best function that underlies the predictive relationships between soil observations and predictor variables (Brungard et al., 2015; Wadoux et al., 2021). In this way, the computational ML algorithms have been proved to be the most successfully techniques applied to digital mapping soil types and attributes (Brungard et al., 2015; Camera et al., 2017; Dharumarajan and Hegde, 2022; Guo et al., 2015; Hengl et al., 2015; Ma et al., 2017; Meier et al., 2018; Riza et al., 2021; Román Dobarco et al., 2017).

Algorithms from different ML families have been tested in DSM, such as probability-based (Bayesian models) (Taalab et al., 2015), similarity-based (k-nearest neighbors) (Mansuy et al., 2014), error-based (artificial neural networks and multiple regressions) (Ballabio et al., 2019; Kalambukattu et al., 2018) and information-based (decision trees and random forests) learners (Xiong et al., 2014; Yang et al., 2016). Random Forest is one of the most used ML algorithms due to the advantages it presents in comparison to other models, including greater efficiency against noise and overfitting, good performance on small datasets, ease of use and interpretation and production of more consistent maps (Greve et al., 2012; Hengl et al., 2018). Many studies have pointed out RF presenting greater performance for modelling different soil characteristics, such as soil carbon stocks (Gomes et al., 2019; Hounkpatin et al., 2018), soil texture (Chagas et al., 2016), soil types (Chagas et al., 2017; Taghizadeh-Mehrjardi et al., 2020) and soi chemistry (Kumaraperumal et al., 2022).

Maybe the greatest advances achieved with DSM are related to the mapping of soil attributes continuously distributed on space, since the conventional surveys are not able to constrain the real variability of these attributes (Kumaraperumal et al., 2022). DSM achievements are even more important in remote, harsh environments where obtaining continuous soil data at great resolutions and extents is very restricted due to difficulties of accessibility, cost, and time, as the case of the Antarctic ice-free areas (Siqueira et al., 2022).

Probably, the main importance of obtaining soil information in Antarctica is the extreme vulnerability of its environment to the global warming (Turner et al., 2014). The most part of the pedosphere in the continent is composed of permafrost-affected soils (Bockheim and Hall, 2002), in which the effects of the rising temperatures (e.g., permafrost melting and changes in soil hydrology) are crucial in the global changes' scenario. Other ecological shifts are changes on the vegetation cover (Convey and Peck, 2019), changes in distribution and size of flying birds' populations (Barbraud et al., 2011), intensification of geomorphic processes, and exposure of new substrates, which are factors that are proven to affect pedogenic development and the physicochemical character of the soils in Antarctica (Francelino et al., 2011; González-Guzmán et al., 2017; Lopes et al., 2019; Otero et al., 2013; Rodrigues et al., 2021; Simas et al., 2007). This condition is even more critical in the Antarctic Peninsula and Maritime Antarctica, regions where the temperature rising rates are the greatest in the continent (Vaughan et al., 2003).

Understanding the current state of the soils as to their chemical variability is a fundamental step to monitoring the future responses of the Antarctic ecosystems to climate changes, and detailed maps produced by DSM in Antarctica can subsidize a wide range of research communities with spatial soil information able to be used as support for ecological models, environmental management, and future scientific expeditions over the continent. In view of this, we aimed to apply the Random Forest algorithm to modelling and map the spatial distribution of seven soil chemical attributes: bases sum (BS), total organic carbon (TOC), potential acidity (H + Al), available Na, available P, pH in water, and remaining-P (P-rem), in the Northern Antarctic Peninsula and Maritime Antarctic regions; as well as to identify and understand the main environmental drivers controlling the spatial distribution of these key attributes.

2. MATERIAL AND METHODS

2.1. Study area

Maritime Antarctica (MA) encompasses the South Shetland Islands (SSI) (Fig. 1). The SSI are composed of marine sediments, metamorphic rocks, intrusive bodies, tuffs and basaltic to andesitic lavas, of age varying from Jurassic to Late Quaternary (Smellie et al., 2021b, 2021a). The geomorphology is marked by a set of erosive and depositional landforms typical of paraglacial and periglacial environments, including landforms such as moraines, raised beaches, outwash plains, high platforms, blockfields and talus and scree slopes (Francelino et al., 2011; López-Martínez et al., 2012). The SSI have a maritime climate with mean annual precipitation (MAP) ranging from 350 to 800 mm and mean annual air temperature (MAAT) around -2°C (Blumel and Eitel, 1989). Mosses-lichens communities form the main vegetation, whereas flowering plants are also common in some areas (Ferrari et al., 2020). The soils, in turn, follows an altitudinal gradient, with Gelisols dominating in the high parts due to the presence of a continuous permafrost, and the gradual increase of other classes, such as Inceptisols and Entisols as altitude decreases (Balks et al., 2013).

The Northern Antarctic Peninsula (NAP) is represented by the James Ross Islands (JRI) and the Antarctic Peninsula coastal area Hope Bay (Fig. 1). The JRI are mainly composed of marine sandstones, mudstones, and conglomerates predominantly of Cretaceous age overlain by Neogene basaltic lavas and breccia (Pirrie, 1994). The JRI geomorphology is like the SSI, with topography also showing a strong structural control due to sedimentary and volcanic rocks (Jennings et al., 2021). The climate is drier, with MAAT varying from -5°C to -7°C and MAP from 300 to 500 mm (van Lipzig et al., 2004). Due to the cold semi-arid climate, the vegetation is scarce. The main JRI soils are Gelisols due to the widespread permafrost, which can be ice-cemented or dry-frozen, with sporadic shallow Entisols (Daher et al., 2019a; Delpupo et al., 2014).

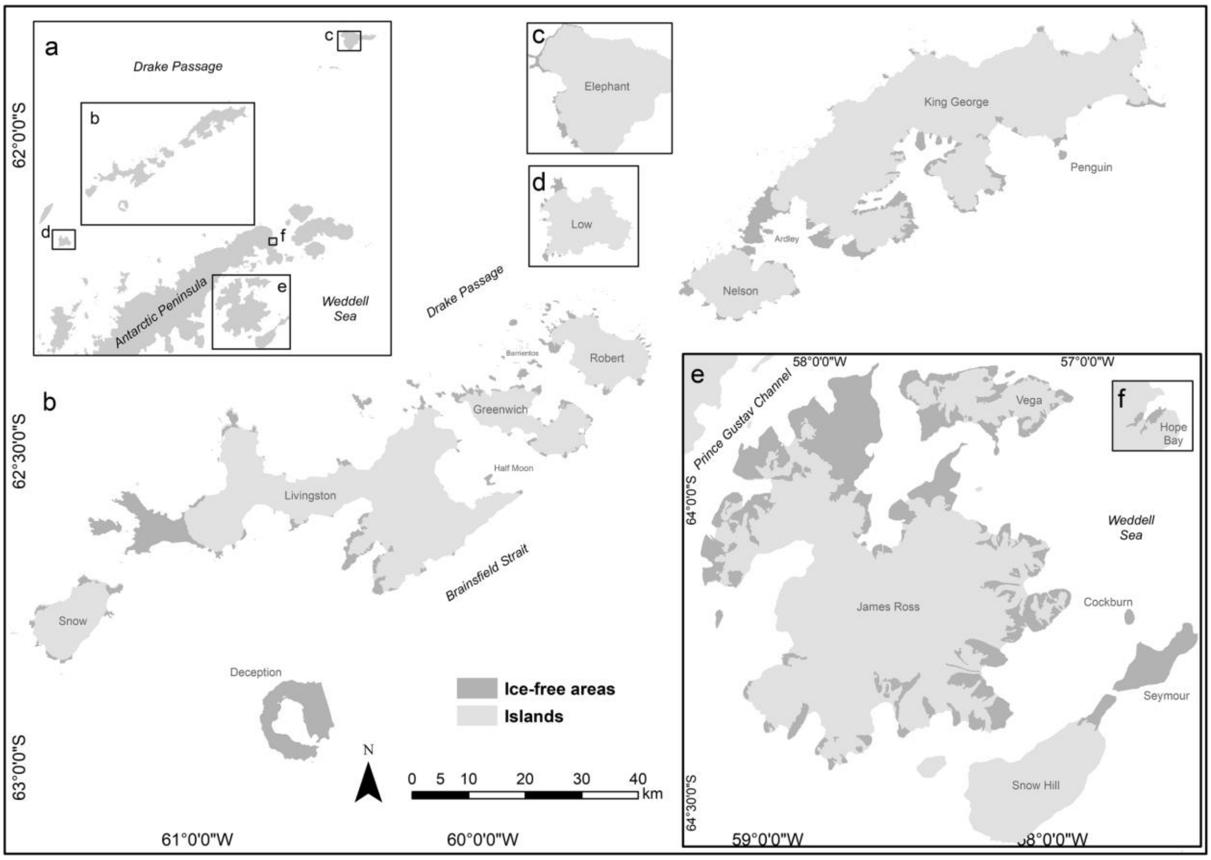


Figure 1. a) The location of the mapped areas: Maritime Antarctica encompassed by the islands of the South Shetland Islands (b – most part of the archipelago, c – Elephant Island, d – Low Island) and Northern Antarctic Peninsula represented by the James Ross Islands (e) and Hope Bay (f).

2.2. Soil data and environmental covariates

We used legacy soil data produced by the Brazilian Terrantar Group in intensive soil surveys in Antarctica. The dataset is composed of complete soil profiles, surface (0 - 10 cm) and pit samples (0 – 30 cm). Additional soil data were obtained from the literature to complement the dataset (Table 1). The total number of soil samples varied for each of the chemical attributes modelled, ranging from 2207 for P-rem to 3140 for TOC (Table 2).

Table 1. Legacy data of the Maritime Antarctica and Northern Antarctic Peninsula containing the soil chemical data used in this work.

Ice-free areas	Points	Soil profiles	Soil profile samples	Other soil samples	Total samples	Soil Surveys
South Shetland Islands						
King George Island						
Keller Peninsula	162	30	130	286	416	(Carvalho et al., 2013; Francelino et al., 2011; Pires et al., 2017; Poggere et al., 2016; Roesch et al., 2012; Schünemann et al., 2022; Thomazini et al., 2016; and unpublished data)
Western Admiralty Bay	74	66	248	17	265	(Bölter et al., 1997; Ferrari et al., 2020; Kozeretska et al., 2010; Krauze et al., 2021; Łachacz et al., 2018; Michel et al., 2006; Simas et al., 2008; Tatur and Myrcha, 1984)
Hennequin Point	4	-	-	12	12	(Vieira et al., 2013)
Lions Rump/Low Head	24	24	94	-	94	(Almeida et al., 2021; and unpublished data)
Potter Peninsula	18	18	71	-	71	(Poelking et al., 2015)
Barton Peninsula	72	41	114	31	145	(Lee et al., 2004; Lopes et al., 2019)
Fildes Peninsula	47	47	141	43	184	(Abakumov, 2018; Abakumov et al., 2021; Alekseev and Abakumov, 2020; Boy et al., 2016; Mendonça et al., 2013; Michel et al., 2014)
Nelson Island						
Stansbury Peninsula	33	23	81	10	91	(Meier et al., 2023; Schmitz et al., 2020a)
Harmony Point	123	46	180	67	247	(Rodrigues et al., 2019; and unpublished data)
Robert Island						
Coppermine Peninsula	14	12	47	4	51	(Thomazini et al., 2020; and unpublished data)
Livingston Island						
Byers Peninsula	98	80	282	37	319	(Silva et al., 2022; Moura et al., 2012; Navas et al., 2008, 2005; Otero et al., 2013)
Hurd Peninsula	21	21	76	12	88	(Ganzert et al., 2011; Haus et al., 2016; Ilieva et al., 2021; Jordanova et al., 2022; Navas et al., 2008)
Cape Shirreff	6	-	-	6	6	(Ramírez-Fernández et al., 2019)
Elephant Point	10	-	-	20	20	(Navas et al., 2017)
Snow Island						
President Head Peninsula	30	30	85	-	85	(Lopes et al., 2022a, b)
Elephant Island						
Stinker Point	46	31	94	15	109	(Navas et al., 2018; Schmitz et al., 2022, 2020b)
Ardley Island	45	23	62	22	84	(Boy et al., 2016; Guo et al., 2018; Michel et al., 2014; Perfetti-Bolaño et al., 2018)
Deception Island	25	5	11	39	50	(Díaz-Puente et al., 2021; Fermani et al., 2007; Vlcek et al., 2018; and unpublished data)
Penguin Island	3	3	12	-	12	(Unpublished data)
Half Moon Island	21	21	54	-	54	(Schmitz et al., 2021)
Barrientos Island	5	5	19	-	19	(Daher et al., 2019b)
Northern Antarctic Peninsula						
Hope Bay	21	21	72	-	72	(Pereira et al., 2013; Schaefer et al., 2015)
James Ross Island						
Ulu Peninsula	40	38	131	2	133	(Daher et al., 2019a; Meier et al., 2019; Prater et al., 2021; Prietzel et al., 2019; Vlcek et al., 2018)
Vega Island						
Cape Lamb	30	30	86	14	100	(Moreno et al., 2012; Siqueira et al., 2021)
Seymour Island	44	44	183	-	183	(Delpupo et al., 2014; Gjorup et al., 2020)
Cockburn Island	3	3	6	-	6	(Unpublished data)

As environmental covariates (Table 3), we used a set of primary and secondary terrain attributes and terrain-based classifications derived from the 8-meter spatial resolution Reference Elevation Model of Antarctica (REMA) (Howat et al., 2019). The terrain attributes were generated using the R package *RSAGA* (Brenning, 2008). We also used multispectral bands (varying from the visible to the short-wave-infrared) from Sentinel-2 mission satellite imagery. For representing the biological influence, we used the Soil Adjusted Vegetation Index (SAVI) and Normalized Difference Vegetation Index (NDVI), besides the rookeries Euclidean distance representing the fauna activity. We also used the bioclimatic covariates available for Antarctica by the Climatologies at high resolution for the earth's land surface areas (CHELSA) program (Karger et al., 2017). The Euclidean distances to glaciers and to coastline were used as predictors of the time factor whereas to account for parent material effects, we used a simplified geology extracted from local geological maps. The features used for the creation of the Euclidean distance covariates (rookeries, coastline, and glacier limits) were delimited by photointerpretation over the Sentinel images cited before. All the covariates were resampled to a 32-meter spatial resolution. For more information about the relation of the covariates used in this work with the scorpan-factors (McBratney et al., 2003), their sources, generation and preparation, the readers are referred to Chapter 1.

Table 2. The number of samples and statistics for each attribute from the soil observations and according to the depth increments. BS: bases sum, TOC: total organic carbon, H+Al: potential acidity, Na: available Na, P: available P, pH: pH in water, P-rem: remaining P.

Attribute	Total	0-10 cm	10-30 cm	30-100 cm		0-10 cm			10-30 cm			30-100 cm		
			n			Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
BS	2505	859	849	475	**$cmol_c\ kg^{-1}$**	0.5	17.1	89.1	0.6	17.4	139	0	20.7	80.7
TOC	3140	1187	1114	547	**$dag\ kg^{-1}$**	0	3.3	46.6	0	2.7	40.6	0	1.7	25.3
H+Al	2479	853	854	470	**$cmol_c\ kg^{-1}$**	**0**	7.8	62	0	8	50.8	0	7.9	49.1
Na	2351	802	781	470	**$mg\ kg^{-1}$**	**0**	489	9177	0	427	5830	0	470	6243
P	2591	903	885	482	**$mg\ kg^{-1}$**	**0**	547	11911	0	565	9998	0	476	10602
pH	3056	1117	1053	537	-	**2.8**	6.15	10.3	2.8	6.2	9.6	2.4	6.3	9.7
P-rem	2207	768	762	451	**$mg\ L^{-1}$**	**0.1**	29.9	60	0.2	30	60	0.8	32.9	60

All data processing, modelling and prediction were carried out using the statistical software R (R Core Team, 2023). To fit the originally discrete data in standardized soil depths we applied a mass-preserving or equal-area quadratic spline using the R package *mpspline2* (O'Brien et al., 2022), which models a continuous depth function of attributes in soil profiles

(Bishop et al., 1999). An equal-area spline consists of a series of local quadratic polynomials that join at 'knots' located at the soil horizon boundaries. It passes through each soil horizon, maintaining the average value of the soil attribute, and is linear between and quadratic within the horizons giving a linear-quadratic smoothing spline. The areas above and below the fitted spline in any horizon are equal—hence the name equal-area (Adhikari et al., 2013). For the spline function, we used a smoothing parameter value of 0.1, leading to values of soil attributes for each cm. After that, the continuous values of the spline were averaged within three depth increments, 0 to 10, 10 to 30 and 30 to 100 cm. The new values for the standard depths acted as new point data at different depths, and before the regression modelling, they were split into training and test sets in a 3:1 ratio (holdout), with the function *createDataPartition* of the *caret* package (Kuhn, 2008). In this way, 75% of the data was included in the training set and 25% in the test set.

Table 3. Environmental covariates organized according to their types and sources.

Data source	n	Types of covariates
Soil		
Sentinel 2 dataset	15	Multispectral bands of visible, near infrared and short-wave infrared, besides bands ratios.
Climate		
Climatic data from CHELSA	19	Bioclimatic data containing MAAT, MAP, precipitation seasonality, isothermality, etc.
Organisms		
Sentinel 2 dataset and CHELSA	6	Vegetation indexes such as NDVI and SAVI, distance to rookeries, and net primary productivity.
Relief		
Digital Elevation Model REMA	61	Terrain attributes of height, slopeness, flatness, hills, curvature, aspect and hillshade, surface gradient, drainage density, insolation, texture, valley, wind exposition, wetness, geomorphons, landforms, etc.
Parent materials		
Geology map	1	Layer of generalized geological units, containing volcanic, sedimentary, metamorphic rocks, unconsolidated sediments, pyritized rocks,
Time (soil development)		
Euclidean distances	2	Euclidean distances to glaciers and coastline

2.3. Covariates selection

To select the most important covariates, we applied a two-step procedure according to Poggio et al., 2021, using the *caret* package. Firstly, we conducted a de-correlation to select only covariates that had a Spearman pairwise correlation coefficient < |0.95| with all other covariates. For each pair of covariates correlated above this threshold, the predictors with the greatest mean correlation with all set of predictors were removed.

As the second step, we applied the Recursive Feature Elimination (RFE) algorithm (Guyon et al., 2002) to backward selecting the best predictors for each soil attribute model. The RFE is a methodology that has proven very effective to select an optimal subset of covariates for the Random Forest model (Gomes et al., 2019; Hounkpatin et al., 2018). The algorithm starts by fitting a model using all covariates, assessing its performance, and ranking covariate importance. The least important covariates are then removed from the pool, and again the model is fitted, assessed, and the least important covariates are removed. The procedure repeats down to a pool between 2 (in our case) and N covariates. Here, we backwardly refitted the model at every 5 covariates until 30 of them last, and then at intervals of 2, since removing just one predictor at each fit is largely time-consuming. The RFE was run 50 times and the subsets conducting to models with the lowest Mean Absolute Error (MAE) were chosen as the optimal subsets for the training and prediction steps. This two-step procedure have as main goals to minimize the redundancy between covariates, and obtain a more parsimonious, interpretable, and computationally feasible model (Poggio et al., 2021).

2.4. Random Forest training

The Random Forest (RF) is a ML algorithm developed to be used for both regression and classification modelling (Breiman, 2001). RF is a method of 'ensemble learning' that bases its result on the aggregation of the single results of individual Classification and Regression Trees (CART) (Breiman, 2001). To build the decision trees, RF uses the bagging to randomly select bootstrapped samples from the original training set. The bagging is a mechanism where the selection of samples is done with replacement - where individual samples can be chosen more than once. About one-third of the observations is left out (out of bag - OOB), which is used to test the performance of the trees as well as the importance of the covariates (Liaw and Wiener, 2002).

Additionally, RF also randomly selects covariates subsets, aiming to create de-correlated trees (Nussbaum et al., 2018). These covariates are used for setting the if-then rules at the tree nodes, splitting the data in the training process or establishing the paths of the new observations during the prediction. The definition of the rules over the predictor space as well as the choice of the covariates for each node are made recursively to create intermediate and leaf nodes with the lowest possible variance. In regression tasks, this is made through the combination of covariate/rules that guarantees the least mean square error (MSE) in the resulting nodes (Kelleher et al., 2015). In this process, the best covariate available for the tree is placed in the root node, to optimize the tree growth. The predicted values of new observations will be the values of the response variable in one leaf node, defined as the mean value of the training samples sent into this node. In RF, the outcome for each point is the median of the prediction in all trees, resulting in a lower bias and variance (Hengl et al., 2018).

Three parameters must be defined for RF training: the number of trees in the forest (ntree = default 500), the minimum number of data points in each terminal node, which also sets the tree depth (nodesize = default 5), and the number of covariates selected for each tree (mtry = default p/3). The ntree and nodesize were used as default, whereas the mtry values were chosen through model optimization, since it is the only RF tuning parameter that requires special judgment (Nussbaum et al., 2018). For defining the optimized 'mtry' value, we applied a repeated 10-fold cross-validation during the training, using the *caret* package. With this method, the training samples are partitioned into 10 nearly-equally sized folds. Then, the models are trained by repeatedly leaving out one of the folds and the training performance is determined by predicting on the fold left out. To overcome the effect of sampling variability, we ran the models 50 times, creating repeatedly different training and test datasets.

The relative covariate's importance (RI) was estimated on caret package using the function varImp. For the RF, this function uses the OOB data to compute the MSE for each tree, before and after permuting each predictor variable. The mean differences between the two accuracies over all trees gives us the mean decrease in accuracy index, which measures the decrease in accuracy that results from the exclusion of covariates from the regression. The higher the decrease, the higher the relative predictor importance. With varImp, the values of importance are normalized for a scale of 0 to 100 (Kuhn, 2008). We also constructed partial dependence plots with the *randomForest* package (Liaw and Wiener, 2022) and a Principal Component Analysis (PCA) using the *FactoMineR* package (Husson et al., 2023) to better understand the relationships of the soil attributes with the main environmental covariates. Partial dependence is an analysis which aim to obtain the functional form of the association between the soil attributes and the individual covariates (Wadoux et al., 2023).

2.5. Random Forest test, prediction, and uncertainty

To test RF performance, we applied the independent test data through statistical indexes. The MAE (Eq. 1) (Eq. 1) and root mean squared error (RMSE) (Eq. 2) summarize the residuals and describe the absolute accuracy of the models, which means how close the predicted values are to the actual values. The NRMSE and NMAE are the versions of these indexes normalized by the average value of the observations, which allow the comparison among models of different units. The coefficient of determination (R^2) (Eq. 3) indicates the proportion of the variance in the target variable that the models explains, and the Lin's concordance correlation coefficient (LCCC) evaluates the agreement between the observed and predicted values. LCCC (Eq. 4) assesses both precision and accuracy of the prediction (Khaledian and Miller, 2020). As an additional validation, we applied the "Null MAE" and "Null RMSE", by calculating the metrics for a null model with the mean observations as predictors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (3)$$

$$LCCC = \frac{2\rho\sigma_{x_i}\sigma_{y_i}}{\sigma_{x_i}^2 + \sigma_{y_i}^2 + (\bar{x} - \bar{y})^2} \quad (4)$$

Where x_i represents soil texture observations, $\bar{x}$ is the average of these values, y_i represents the soil texture values predicted and $\bar{y}$ is the average of these values. $\sigma_{x_i}^2$ and $\sigma_{y_i}^2$ are the respective variances and ρ is the Pearson correlation coefficient. The indices were obtained for all 50 runs and compared by the mean values.

The final maps for each chemical attribute at each increment depth was obtained through averaging the spatial prediction of the 50 runs. Besides the predicted mean, we also computed the 95th and the 5th percentiles for each pixel through the runs. Then, we estimated the prediction pixel-level uncertainty by calculating the 90% Prediction Interval (PI, Eq. 5) and the Prediction Interval Ratio (PIR, Eq. 6) according to proposal of Poggio et al. (2021).

$$PI = Q95 - Q5 \quad (5)$$

$$PIR = \frac{PI}{Q50} \quad (6)$$

Where Q95 is the 95th percentile, Q5 is the 5th percentile and Q50 is the median. With the PIR, we can evaluate comparatively the uncertainty for soil attributes with different units.

3. RESULTS

3.1. RF performance

In training, the repeated-cross validation R^2 varied from 0.28 to 0.48 (mean values for the 50 runs), whereas the LCCC results presented more optmistic values, varying from 0.94 to 0.90. Accuracy in test accounted for a mean R^2 very similar to training, with a range from 0.24 to 0.47. On the other hand, the LCCC values were substantially lower than in training, although keeping a score higher than the R^2, varying from 0.37 to 0.64 (Supplementary data, Table 1). The BS presented the best performance according to the R^2 and LCCC on both training and test. In test, the accuracy of the BS model is significantly higher than for the other attributes at the depth 0-10 cm. At the depth 10-30 cm, the differences are not signficant in relation to H+Al, pH and TOC when considering the R^2. The lowest accuracy are generally associated to the attributes P and Na, with the exception to the 30-100 cm, where the P-rem retained the lowest R^2 and LCCC. The P and Na performances also presented the largest variations along the 50 runs (e.g., widest boxplots), indicating a higher instability of the results (Fig. 2).

When we consider the normalized error metrics, we find a pattern slightly different. The pH presented the lowest NMAE and NRMSE at the three depths, followed by P-rem and BS. In turn, the highest errors were found for the P, TOC and Na (Fig. 2). The worst results for the P and Na prediction are due to errors linked to the presence of extreme values (natural outliers). This, in turn, is associated to the large range of these attributes, with values varying in mg kg^{-1} scale from 0 to 1191 and 0 to 9177 for the P and Na, respectively, whereas their means are only 529 and 462 mg.kg (Table 2). This is corroborated when we look to the highest P and Na values for the RMSE/MAE ratio. On the other hand, the lowest pH ratio indicates the outliers affect more the error metrics than the accuracy ones.

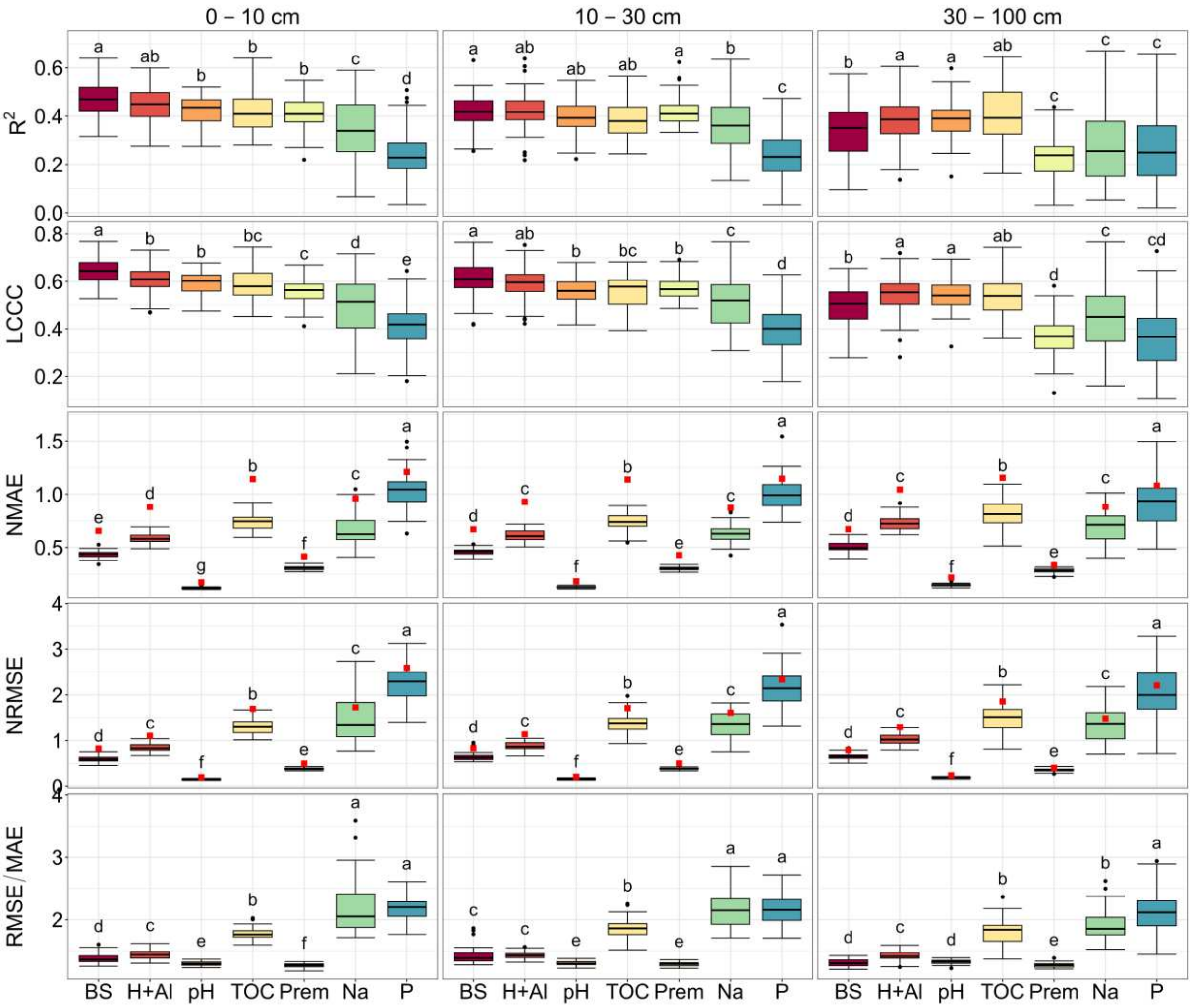


Figure 2. Test performance of the attribute models accounting the accuracy and error metrics. The median of the NRMSE and NMAE of the Null models are represented by the red square plotted along with their respective models. The letters represent the significance of the differences at 95% of confidence.

The difficulties of the RF in predicting natural outliers is an issue which provenly affect the model performance, and it has already been addressed in other studies (Yang et al., 2016). Although this effect is more pronounced for the P and Na prediction, it can be seen in different degrees for all others attributtes, with the MAE increasing significantly for the last part of the data distribution (Fig. 3).

In general, the largest errors are associated to the underestimation of the extreme maximum values, although for the pH and P-rem the greatest MAE values represent the overestimations of the extreme minimum values. Overall, the lowest accuracy are found in the 30-100 depth, where the smallest number of samples hinders the modelling (Table 2). Nonetheless, for many of the attributes modelled, the differences between the three depths are not statistically significant, indicanting a vertical homogeneity of performance. All combinations of attributes and depths presented RMSE and MAE lower than the Null models, indicating the adequacy of the RF for modelling soil attributes.

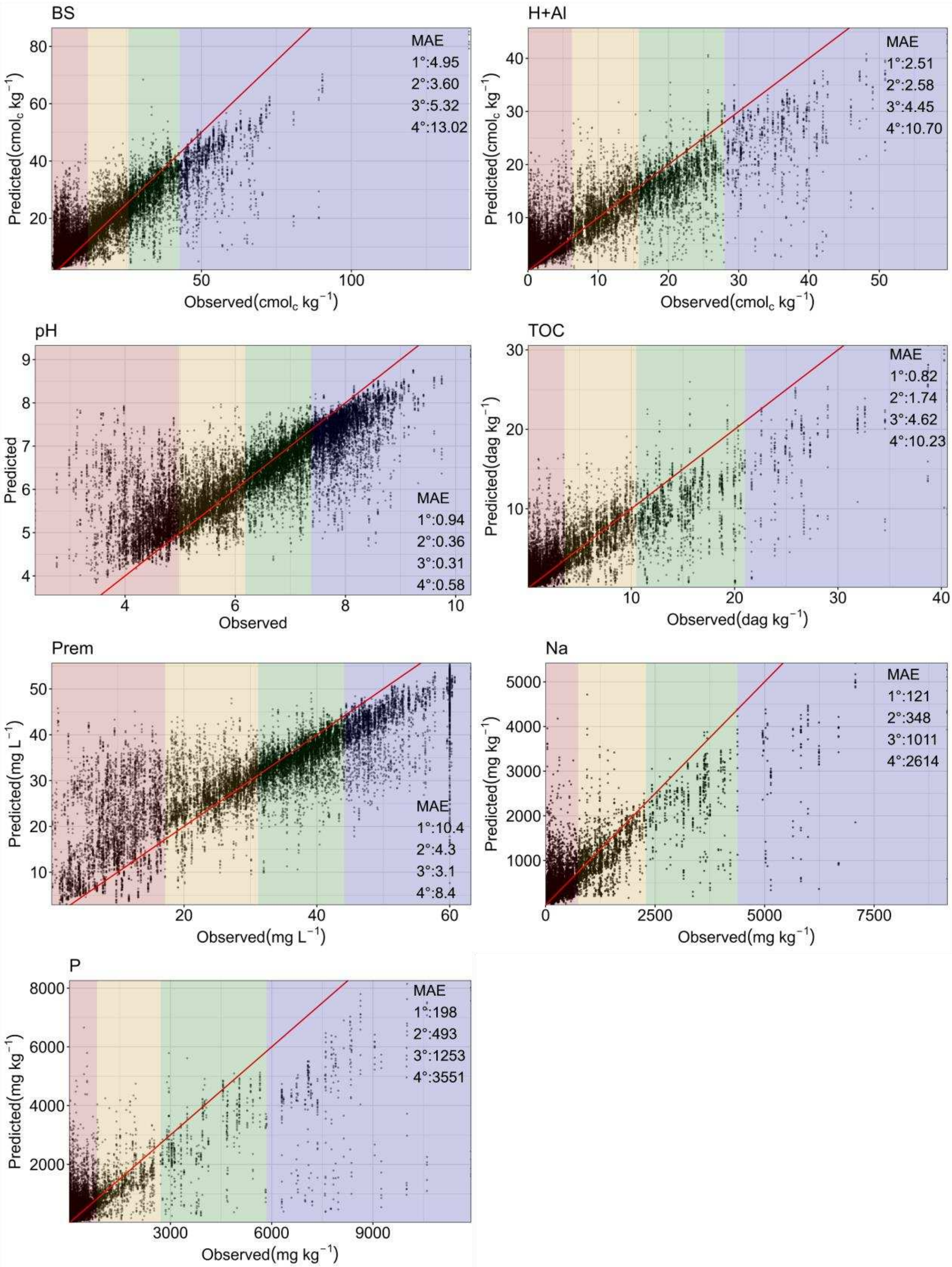


Figure 3. Distribution of the predicted vs observed data for each attribute accounting all test samples from the three depths and the 50 runs; and the MAE calculated for classes of the data, represented by the different colours. The data classification was done using the Jenk's natural breaks method. Colors: red = 1° MAE; yellow = 2° MAE; green = 3° MAE; blue = 4° MAE.

There were large variations of prediction errors at the islands level. The prediction in Nelson Island showed one of the best results in terms of performance, with the LCCC test being higher than 0.75 for all attributes and the accuracy being the highest one for the BS, pH and P-rem. Snow Island and Hope Bay also presented very sactisfatory performances (although with higher LCCC variations), occupying the first position for the remaining soil attributes. At last, the greatest prediction errors were obtained for Barrientos and Deception islands, which commonly presented LCCC lower than 0.10 and negative values at some cases (Fig. 4).

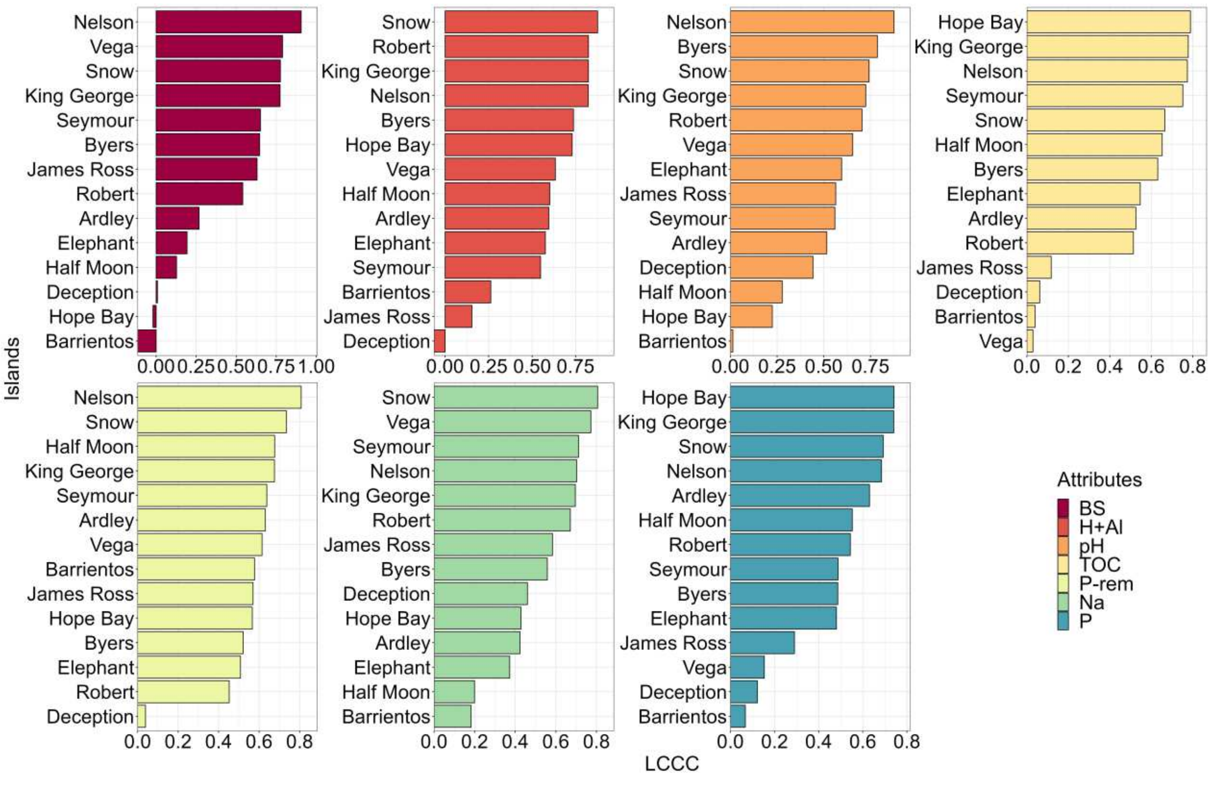


Figure 4. Performance of the prediction for each attribute and island mapped. The LCCC values are the mean for the three depths along the 50 runs.

3.2.Covariates selection and importance

The size of subsets selected by the RFE varyed largely for each soil attribute along the 50 runs. The smallest sets retained only 2 variables, whereas the largest ones presented up to 110 variables, indicating the importance of many repetitions associated to different training data to obtain an optimal subset of covariates. The attributes with the smallest sets of covariates were the BS, TOC and Na, with median values between 4 and 8 covariates (Fig. 5), showing the greater RF capacity in extracting the best relationships of these attributes when dealing with just a few predictors, although this is not reflected in similar performances.

H+Al and pH presented intermediate-sized subsets, varying from 10 to 14 covariates. On the other hand, the largest subsets were obtained for the P-rem and P, with maximum median values of 34 and 25 predictors, respectively (Fig. 5). These attributes also presented the largest variations of subsets sizes along the 50 runs, which reflects the lack of certainty when handling different training sets. The models with the largest number of RFE-covariates also have higher probability of incorporating poorly- important predictors whose patterns extracted are more related to the presence of artificial noises rather than a real spatially referenced environmental pattern. This can induce to a significant lost of performance due to overfitting, which in part, can explain the worst performance of the P modelling.

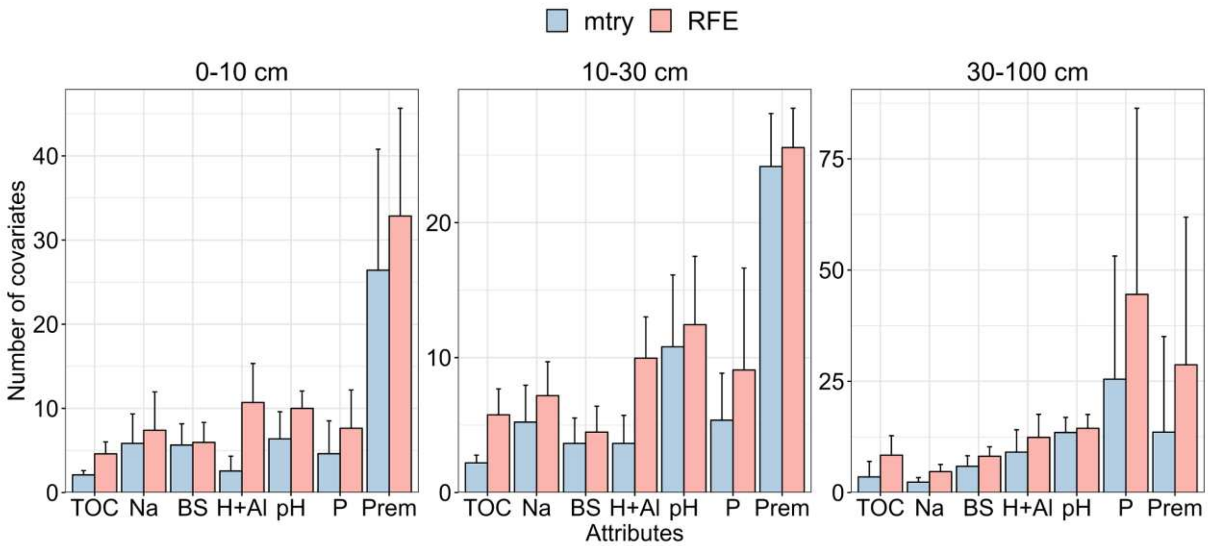


Figure 5. The average number of predictors selected by the RFE and chosen for the trees building (mtry) for each attribute and depth. Error bars mean the standard deviation.

The size of the RFE-subsets influences directly the number of covariates chosen for the individual decision trees in RF. The lowest mtry values were obtained for the TOC at 0-10 cm (median of 2), whereas the highest ones were found for the P-rem at the same depth (median of

26 and maximum of 110), indicating the construction of trees more complex and not strongly de-correlated, according to Nussbaum et al. (2018). With a few exceptions, the largest numbers of RFE covariates were found for the depth 30-100 cm, of least number of samples (Fig. 5).

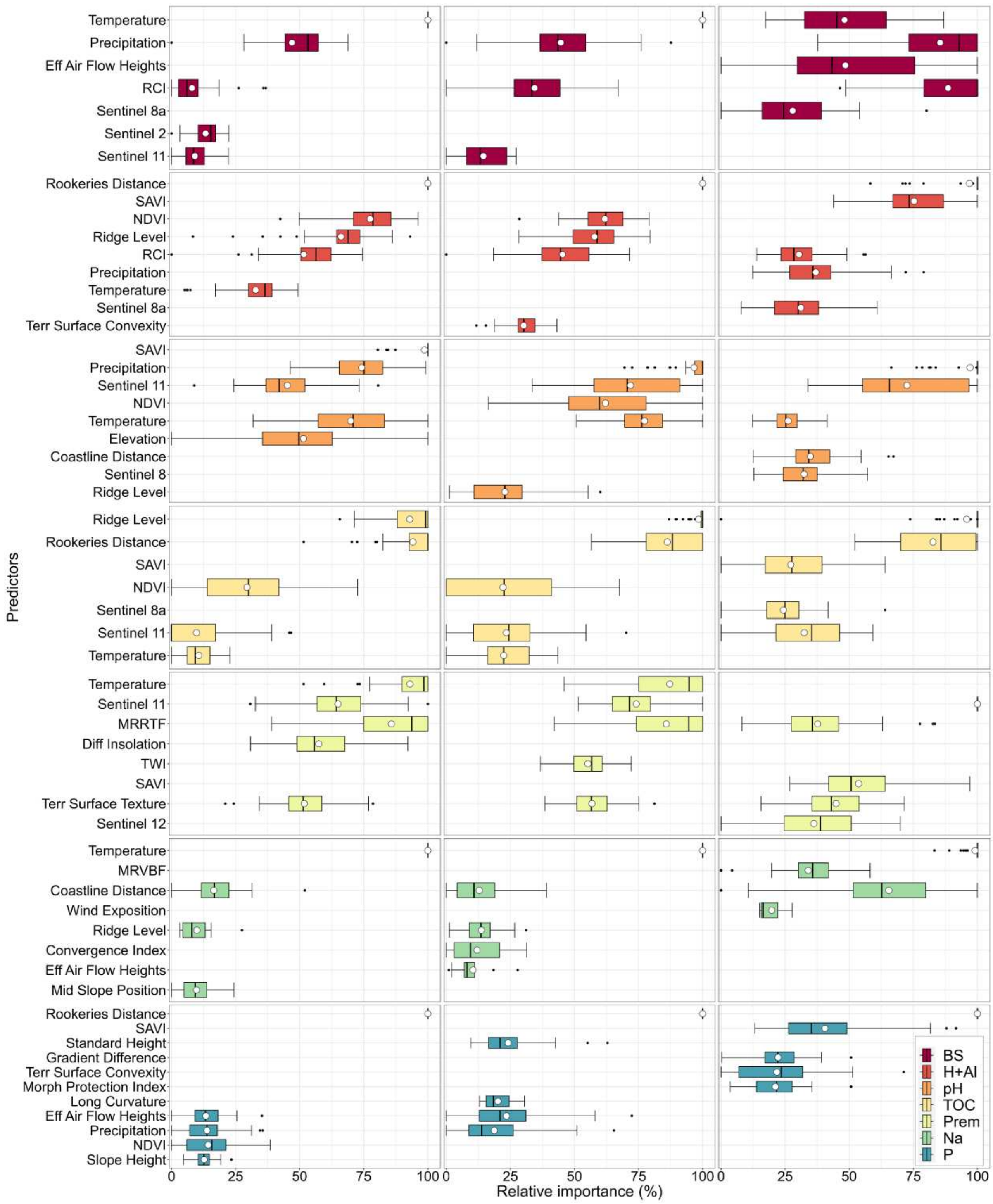


Figure 6. Relative importance of the best 5 covariates for each soil attribute and depth. Boxplots record the variability of importance along the 50 runs.

Each soil attribute model also differed on the most important covariates controlling the prediction. The most important climatic predictors were the Mean Annual Temperature and the Annual Precipitation. Temperature was the most important covariate for the Na and P-rem prediction, whereas it was among the most important predictors for the TOC. In turn, Precipitation was more important for the P and pH prediction, being the best covariate for the pH in the most depths. Temperature was the most important covariate for the BS in the first two depths, whereas the Precipitation was more important in the last one. Although less relevant in general, Temperature and Precipitation were also more important for the H+Al prediction at surface and subsurface, respectively (Fig. 6).

Rookeries Distance was one of the most important organisms covariates, representing the birds activity (ornithogenesis) in the Antarctic environments. It was calculated as the euclidean distance for the major penguin rookeries. It was the most important covariate for the P and H+Al prediction, whereas it was the second more important for the TOC. The vegetation indexes SAVI and NDVI were of great importance for the TOC, H+Al and pH, being the most important covariate for the last attribute in the depth 0-10 cm. The Sentinel bands 11 (short-wave infra-red), 8 (near infra-red) and 8a (red-edge) also reflected in the most the vegetation signal, and were important at different degrees for the BS, H+Al pH, P-rem and TOC (Fig. 6).

The topographic attributes selected were mainly related to altitude, slope and surface dissecation. RCI (roughness concentration index) was the most important covariate for BS prediction at the depth 30-100 cm. Ridge Level (height associated to the basin dividers and in the Antarctic ice-free areas directly related to the absolute elevation) was the most important covariate for the TOC. These two topographic attributes were also relevant for the H+Al prediction. A higher number of terrain attributes were selected for the P-rem, Na and P prediction. For the Na, predictors associated to water accumulation and the Coastline Distance were also important (Fig. 6).

The most important covariates have a well-defined pattern of their contribution when the partial dependence is plotted (Fig. 7). BS, Na and pH contents sharply decrease around -6°C and it marks the transition between the colder NAP and the warmer MA. The TOC contents do not respond to the temperature before -2.5°C, and after a small fall, they tend to a rapid increase following the temperature evolution, showing the temperature is a relevant driver of the TOC only in the warmest MA. The precipitation also influences in different degrees the variables. The pH drops around 500 mm yr-1, whereas the P content starts to respond positively to the precipitation from 450 mm yr-1. These transitions also marks the differences between MA and NAP.

The TOC and H+Al contents respond to the NDVI up to 0.75. In turn, the pH has an opposite pattern with values decreasing as the SAVI values progress, with the highest response range between 0 and 0.5. Rookeries Distance has different ranges of contribution for each attribute influenced. The highest influence on the P prediction occurs up to 500 m, from where the drop in the values persists in slow rates until approximately 2000 m. Therefore, the closer to the rookeries, the higher the P contents. A similar behaviour was found for TOC and H+Al. The TOC contents tend to decrease to approximately 80 m of altitude, whereas the H+Al response lasts longer, up to 150 m. On the other hand, BS and pH increase up to around 80 m. The Na have a singular behaviour in space. Is values decrease abruptly from the coastline to approximately 120 m of distance, and as a response to the topography, they go back up from 100 m to 175 m of altitude, when finally the curve flats out (Fig. 7).

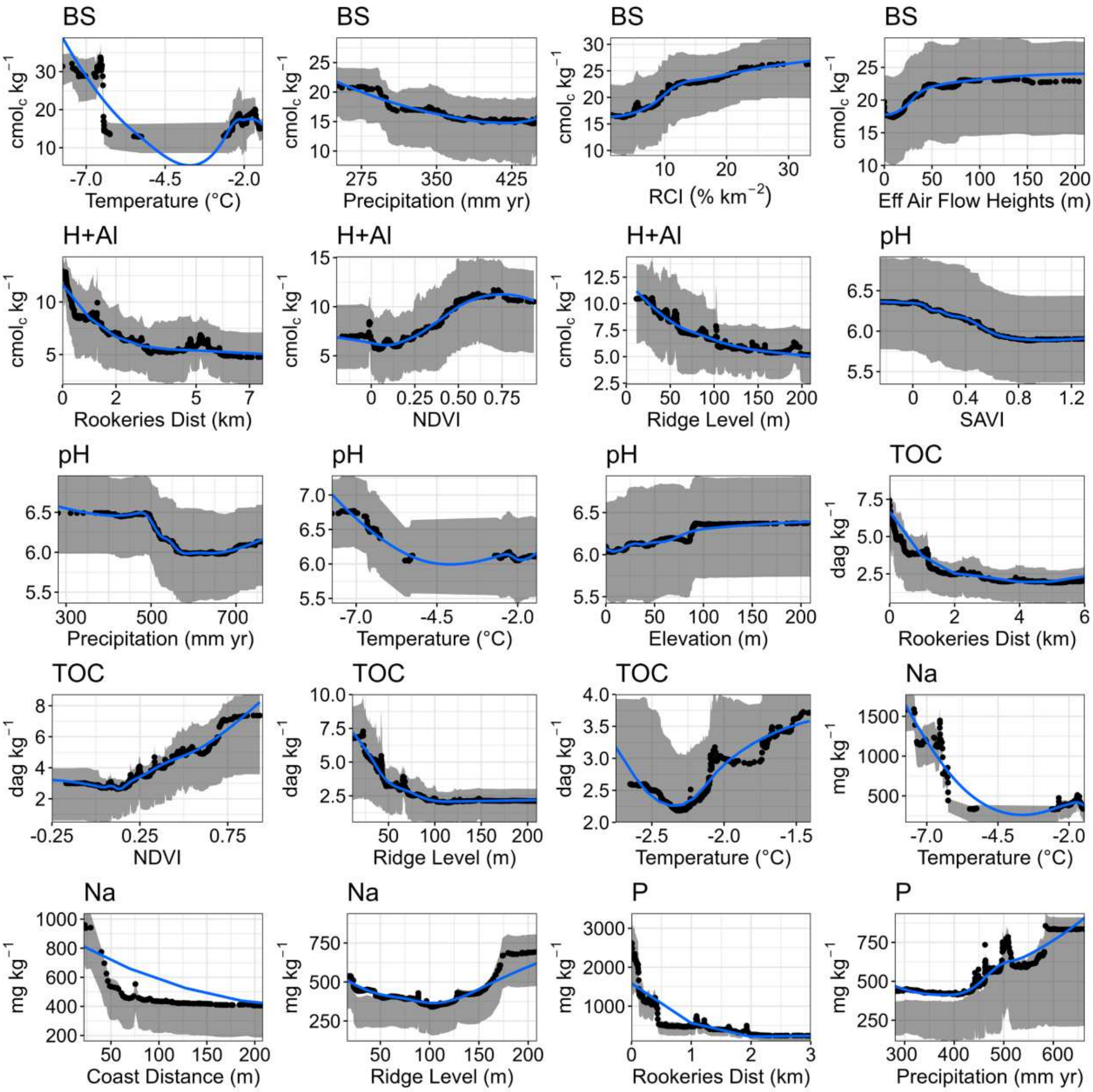


Figure 7. Partial dependence plots relating the soil attributes predicted values (constructed over the training dataset) and the marginalized influence of important environmental covariates.

3.3. Soil attributes distribution

Observing a PCA with the soil attributes modelled, it is possible to distinguish two groups whose final distribution is controlled by the main environmental covariates selected in the modelling process (Fig. 8). The two groups of attributes also represent the remarkable pedological differences between MA and NAP regions and their distribution is clearly distinguishable in the final predicted maps (Fig. 9). The first group of attributes includes the TOC, H+Al and P, and is more associated to the MA soils, whereas the second group of attributes is composed of BS, pH and Na, and its highest values are concentrated in the NAP soils.

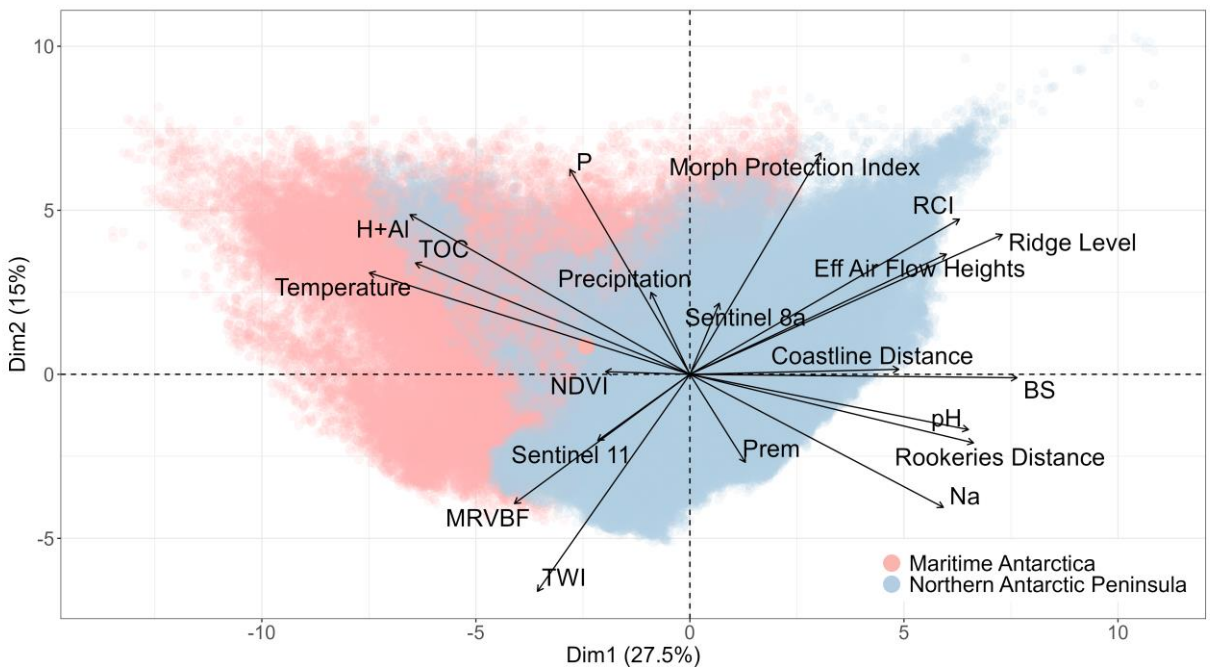


Figure 8. Principal Component Analysis with the soil attributes modelled and some of the most important covariates, selected for representing the combined action of climate, organisms and topography.

The strongest positive correlation was found for the H+Al with the TOC ($r = 0.77^*$) and P ($r = 0.66^*$), whereas the correlations are moderate in the second group, with maximum of

0.58* between BS and Na (Table 4). On the other hand, the strongest negative correlations were found for the pH with the H+Al (r = - 0.67*) and TOC (r = - 0.54*), indicating an severe spatial opposition of these attributes, and consequently, of the covariates contribution, as also seen in the partial dependence (Fig. 7). The P-rem was the attribute with the weaker correlation, not being incorporated in any of the two groups.

Table 4. Pearson correlation coefficient for the soil attributes predicted.

Attribute	TOC	H+Al	P	BS	Na	pH	P-rem
TOC	**1.00**						
H+Al	0.78*	**1.00**					
P	0.57*	0.66*	**1.00**				
BS	-0.39*	-0.43*	-0.15*	**1.00**			
Na	-0.30*	-0.36*	-0.16*	0.58*	**1.00**		
pH	-0.54*	-0.67*	-0.37*	0.56*	0.23*	**1.00**	
P-rem	-0.28*	-0.24*	-0.04	0.04	0.16*	0.17*	**1.00**

*$p \leq 0.05$.

The BS contents varied in the final maps from 3 to 64 $cmol_c$ kg^{-1}, with small differences between the three depths. The higher contents are present in NAP, with overall mean of 31 against 21 $cmol_c$ kg^{-1} in MA (Fig. 10). In NAP, the greatest values are shown in James Ross Island, whereas in MA the highest values are found in Deception Island (Fig. 11). The pH varied from 4.3 to 8.7, and there was no significant differences in depth. In both MA and NAP the mean pH was close to neutrality (Fig. 10). Nonetheless, MA presented a greater variability, with Deception recording the highest and Ardley the lowest pH (Fig. 11).

The Na contents varied from 11 to 4409 mg kg^{-1} and the differences between MA and NAP were much bigger than for the previous attributes. In the first region, the mean was 392 mg kg^{-1} and in the second it was 1672 mg kg^{-1} (Fig. 10). As for the BS and pH, the highest Na contents are found in surface and subsurface for NAP and MA, respectivelly, indicating the different pedogenesis mechanisms acting on each region. In NAP, the Na contents followed an aridity gradient, with the highest values in the drier Seymour Island and the lowest in Hope Bay (Fig. 11).

The TOC distribution varied from 0.16 to 24 dag kg^{-1}. Although the great variation in surface, the TOC contents are significantly higher in MA. As expected, the TOC decreases gradually from surface to the deepest layer (Fig. 10). The highest contents are found in Ardley and Barrientos islands, whereas the lowest values are concentrated in NAP Vega and James

Ross (Fig. 11). The H+Al contents follow a similar spatial pattern to the TOC, but increasing at the deepest layer from a mean of 3.6 $cmol_c$ kg^{-1} at surface to 4.8 $cmol_c$ kg^{-1}, indicating the increase of acidic conditions in depth, mainly in NAP (Fig. 10).

The P distribution follows a similar distribution to the TOC and H+Al with clear dominance in MA, although the peninsular ice-free area Hope Bay presented the highest mean values due to the strong biological influence. In turn, the lowest contents are found in Seymour Island (Fig. 11), indicating the influence of the rookerie located in the island is spatially restricted due to the severe climate and/or absence large abandoned sites. The P contents tend to be higher with depth, indicating the importance of the birds guano leaching to the P profiles enrichment and deepening of ornithogenesis. In both regions, the mean P-rem values stay around 30 to 35 mg L^{-1} and do not vary according to depth (Fig. 10).

3.4. Prediction uncertainty

According to the PI values, there is no pattern between the prediction uncertainty and the standirzed soils depths. Whereas the PI range was higher for BS, TOC and Na in the depth 0-10, it was higher for depth 30-100 for H+Al, pH, P-rem and P. It disagrees with the results of Siqueira et al. (2023), which found higher uncertainties in the last depth when predicting soil texture in Antarctica, supposedly due to the smaller number of samples (Supplementary data, Table 2). Taking into account the PIR values, the pH and P-rem were the soil attributes with the least uncertainty, whereas the Na and P presented the worst results. This clearly indicates that, such as for the prediction errors, the uncertainty indicators of the RF models are also higher for the soil attributes with higher ranges and extreme values.

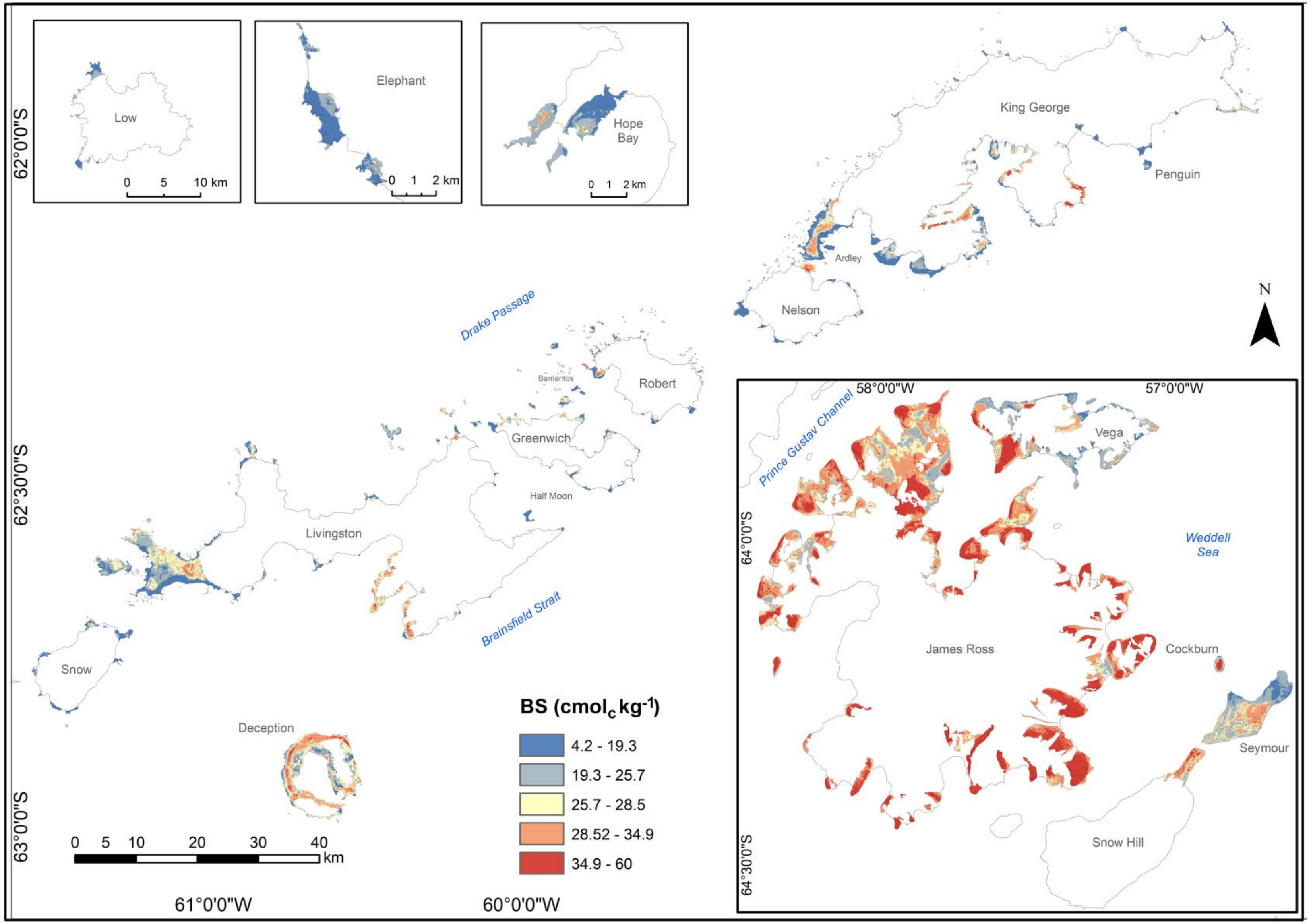


Figure 9. Predicted mean BS contents for the soil depth interval 0-100 cm.

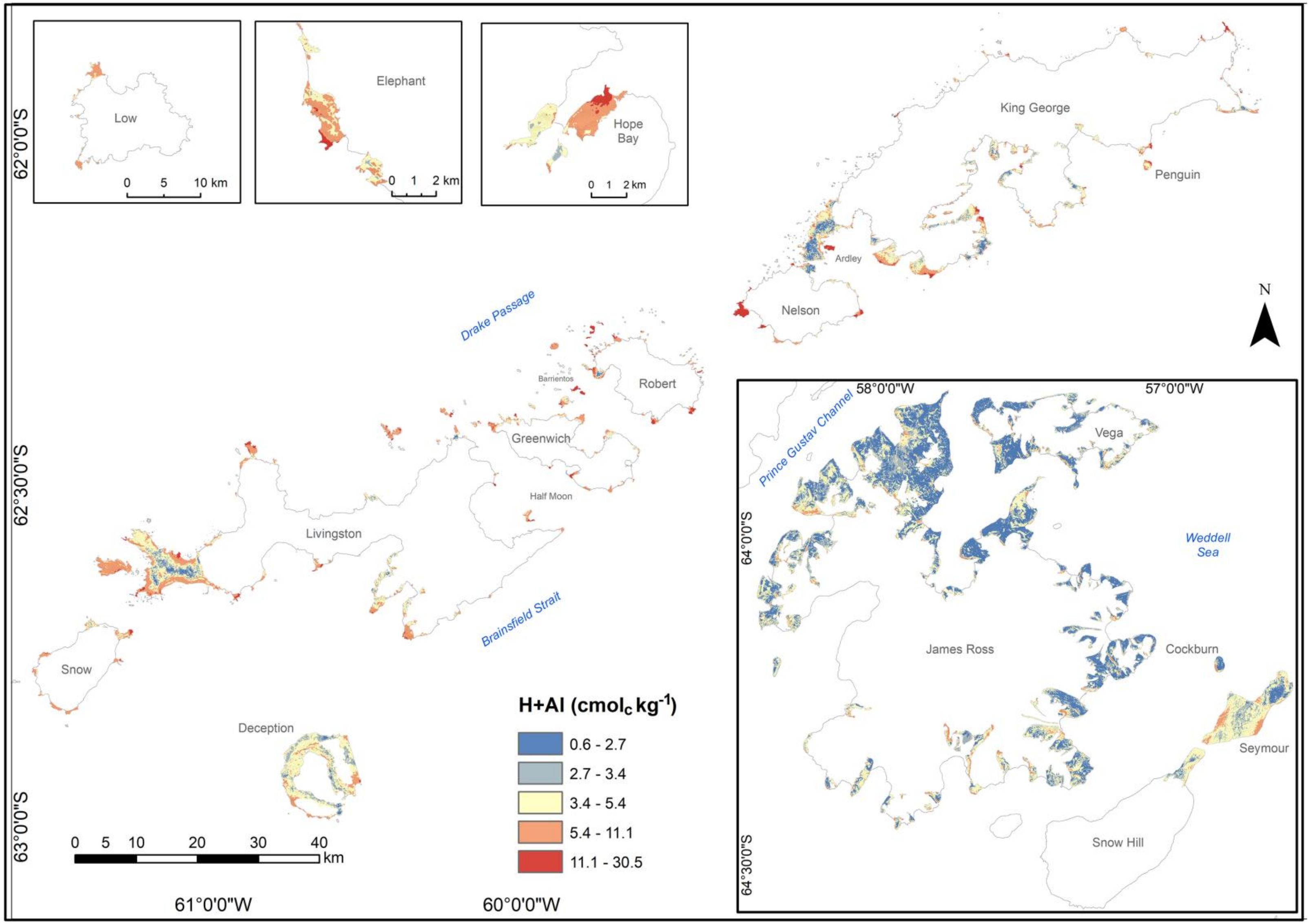


Figure 9 (continued). Predicted mean H+Al contents for the soil depth interval 0-100 cm.

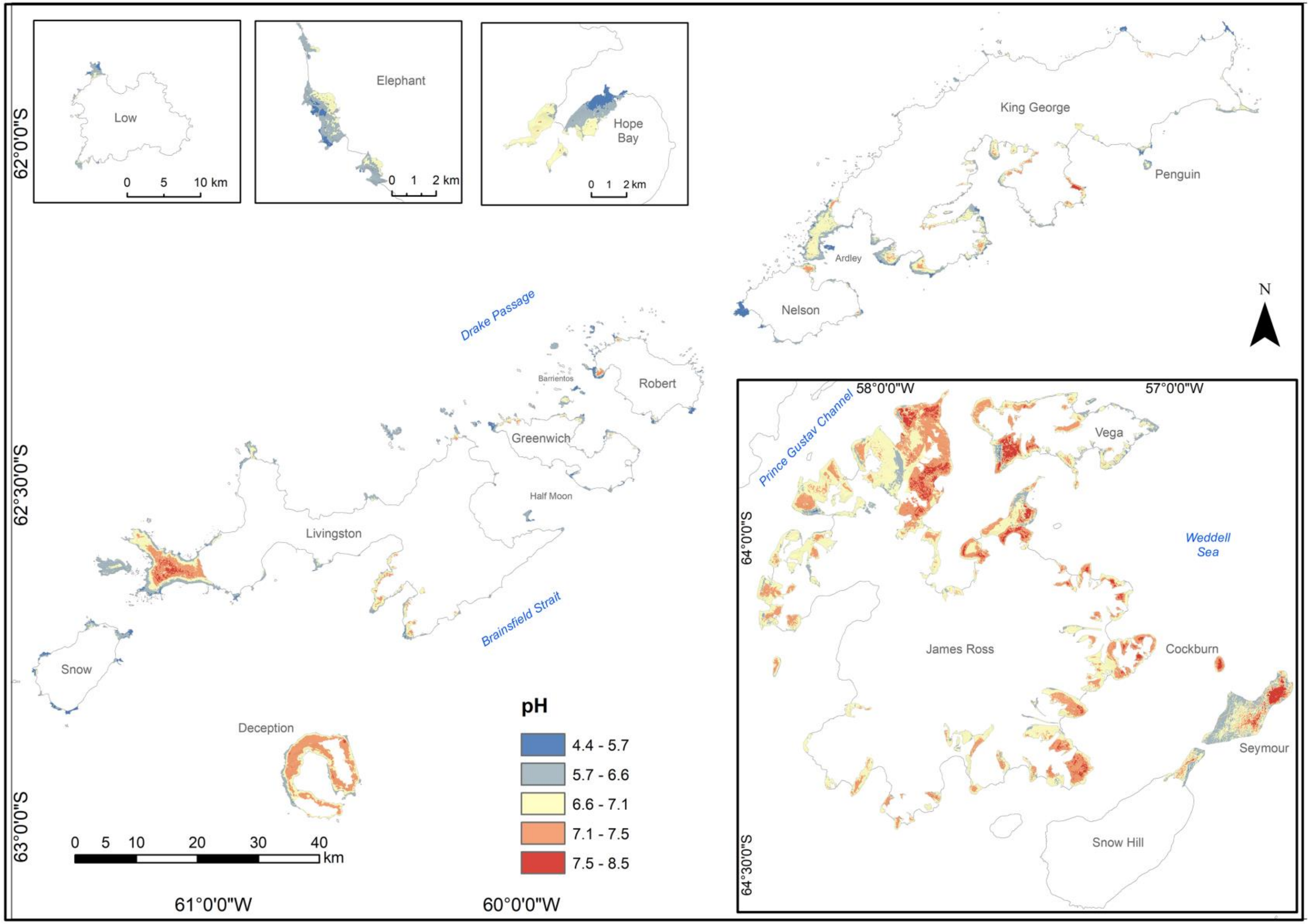


Figure 9 (continued). Predicted mean pH for the soil depth interval 0-100 cm.

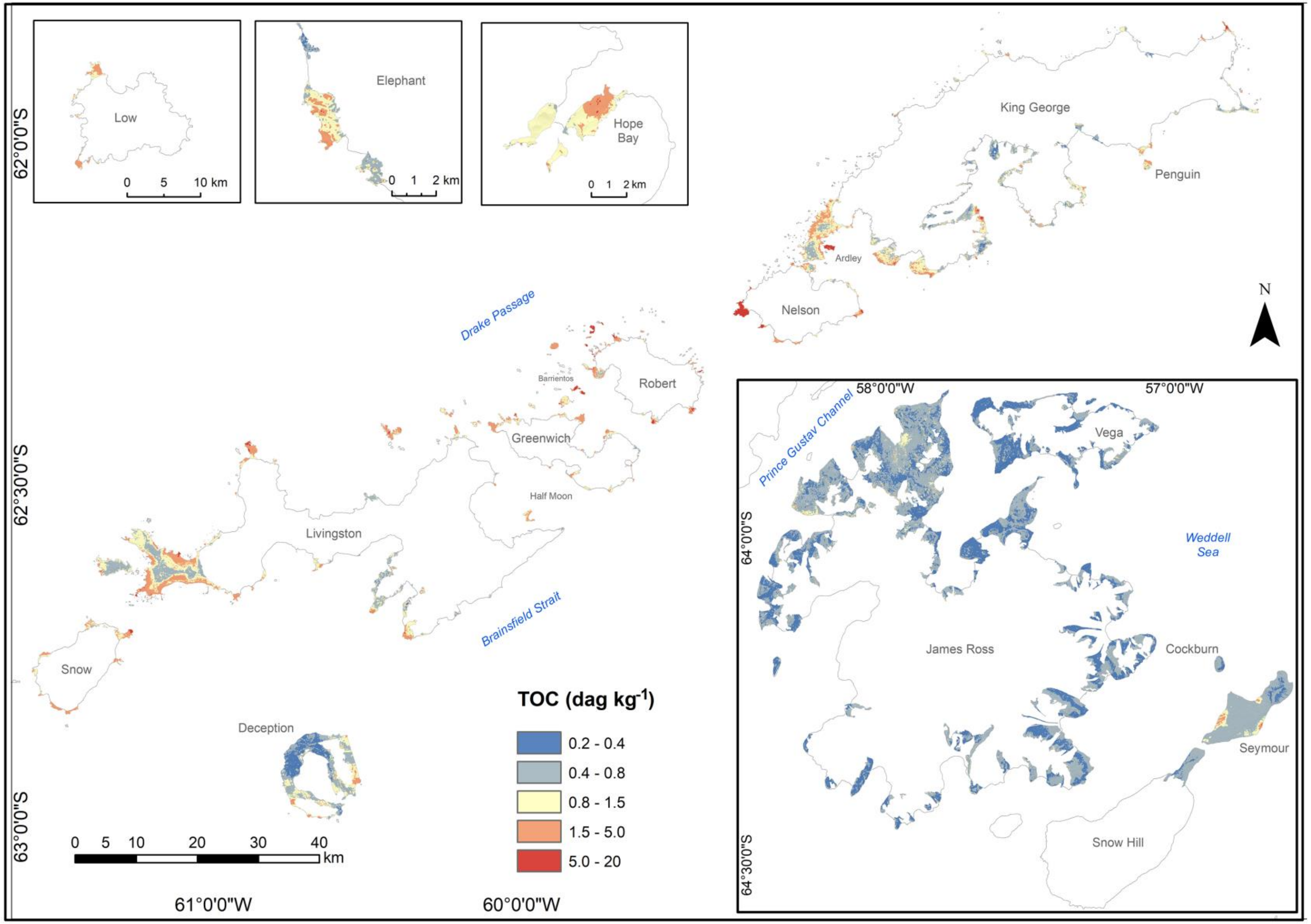


Figure 9 (continued). Predicted mean TOC contents for the soil depth interval 0-100 cm.

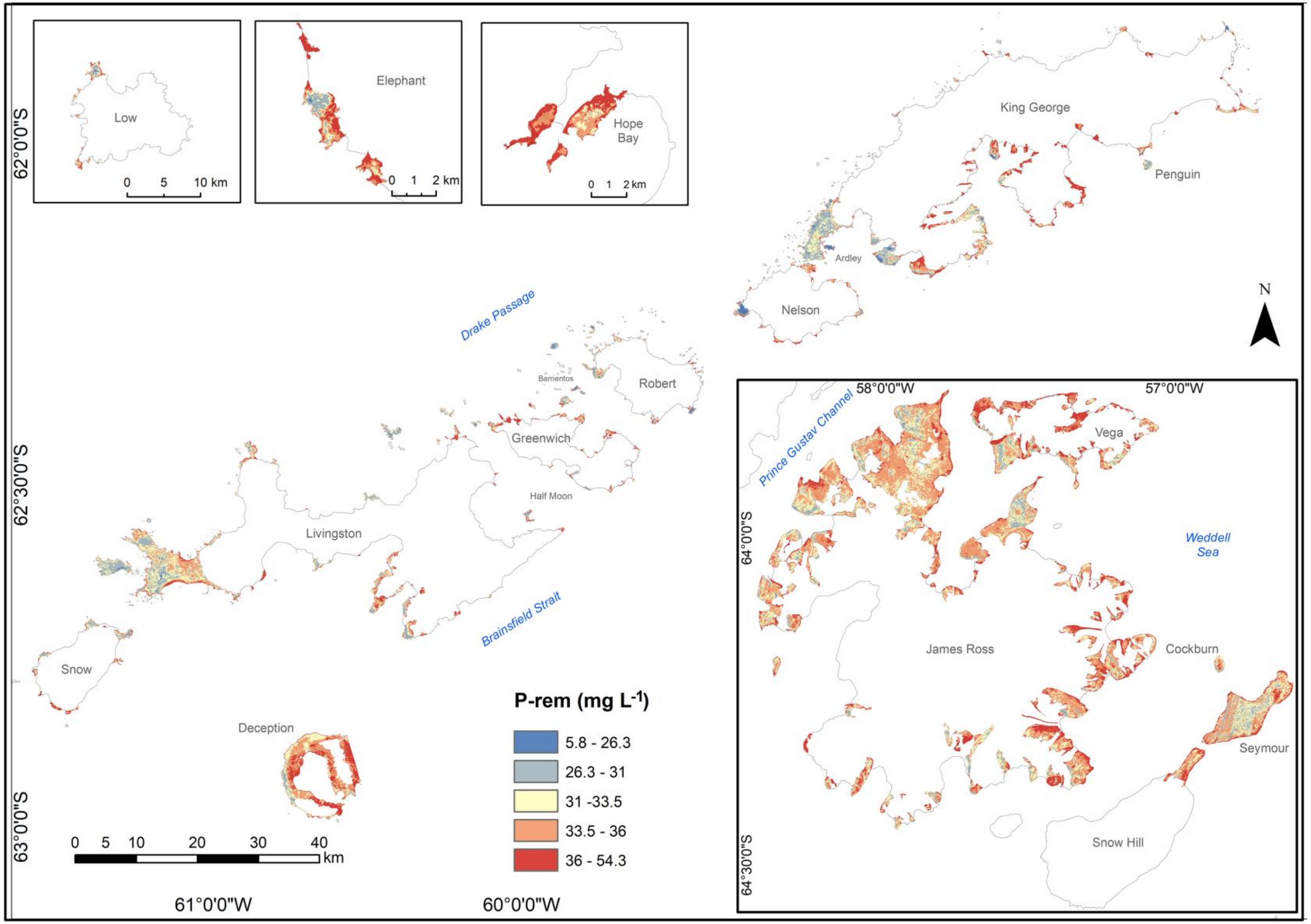


Figure 9 (continued). Predicted mean P-rem contents for the soil depth interval 0-100 cm.

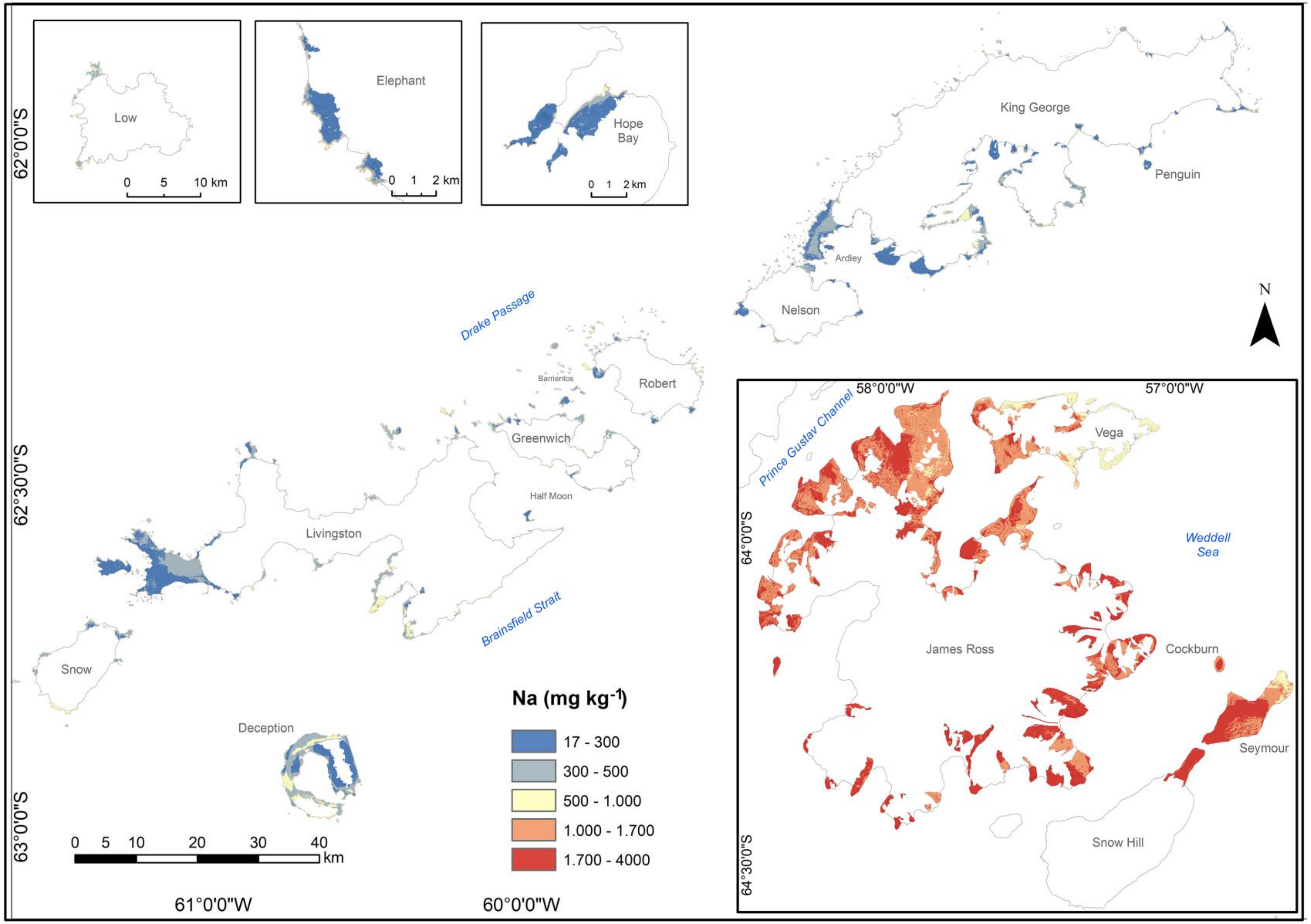


Figure 9 (continued). Predicted mean Na contents for the soil depth interval 0-100 cm.

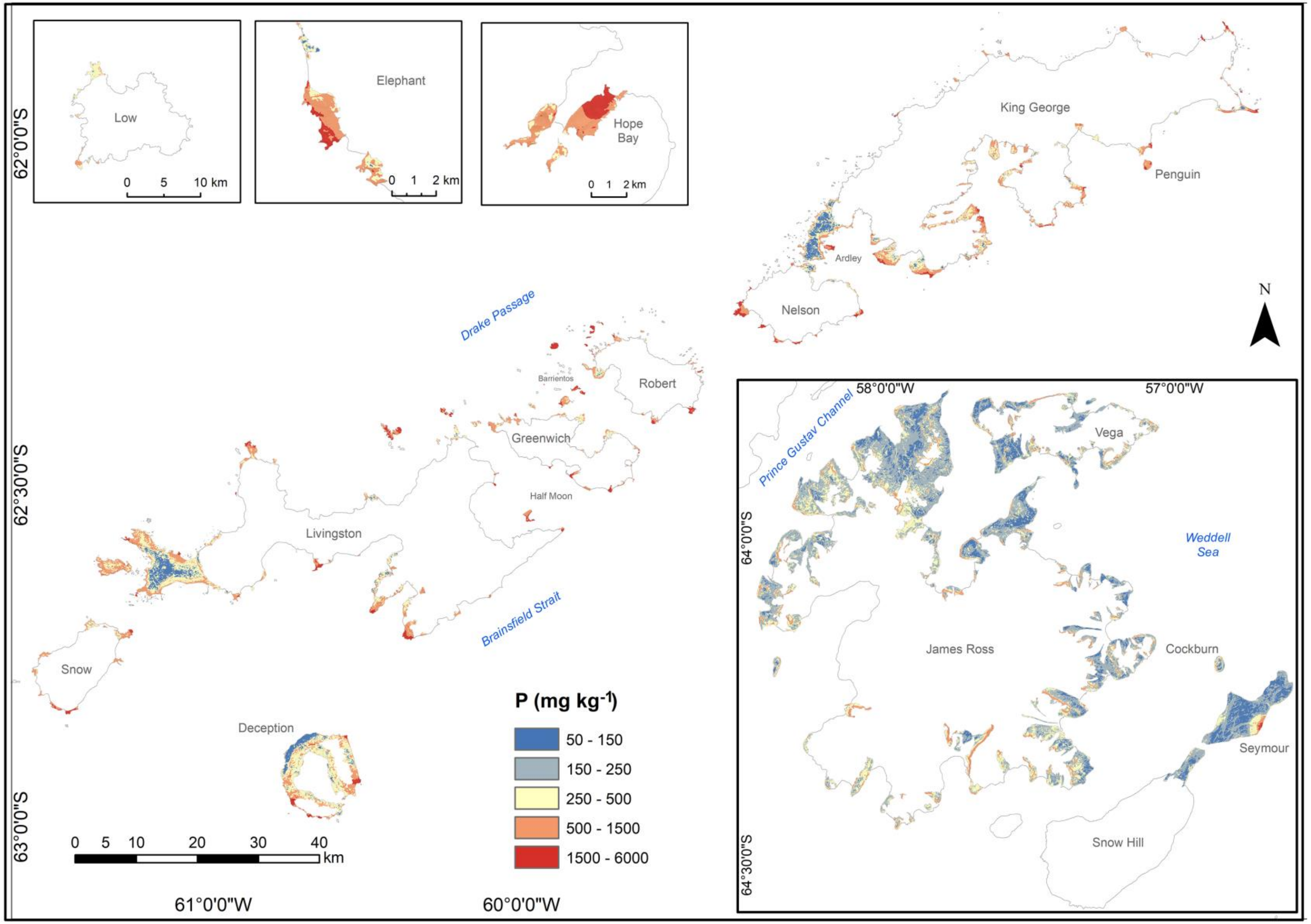


Figure 9 (continued). Predicted mean P contents for the soil depth interval 0-100 cm.

4. DISCUSSION

4.1. Soil variability and soil-environment relationships

In general, the best covariates selected by the RF model are predictors representing three basic environmental factors: the climate, the organisms and the topography. Then, we can assume these are the main soil forming factors (or scorpan factors for the DSM) influencing the chemical character and differences of the soils from the regions mapped. Besides contributing to increasing the predictive power of the models, these covariates also express the diverse environmental factors driving the weathering and ecological processes acting on the Antarctic soils. Due to their environmental significance, these covariates have been successfully applied to modelling major soil chemical attributes elsewhere, including different landscapes in Europe, Africa and Asia (Ballabio et al., 2019; Dharumarajan et al., 2017; Hengl et al., 2015; Kumaraperumal et al., 2022; Ma et al., 2017).

The climate, expressed here by the Temperature and Precipitation covariates, is considered the main driver of soil differentiation in Antarctica and influence the soil characteristics in a number of ways (Balks et al., 2013). A clear climatic gradient led by the temperature and precipitation rates controlls the differences in the spatial distribution of the soil attributes between the moister and warmer MA and the semi-arid NAP. MA is considered one of the Antarctic regions with the greatest soil development due to its septentrional position and the maritime climate (Bockheim and Ugolini, 1990; Bockheim, 2015). The relatively high water availability and air temperatures during the summer season in MA favours the chemical weathering and percolation of solutes in the soils, which is evidenced by the lower BS and higher H+Al contents.

The climate also favours the biological activity and the vegetation development in MA, which affects soil development by way of organic C inputs and biogeochemical nutrient cycling (Abakumov, 2018; Otero et al., 2013). In Antarctica, the fauna influence is of great importance as an organism soil forming factor and happens mainly through the pedogenic process known as ornithogenesis (Simas et al., 2007; Tatur and Myrcha, 1984). The ornithogenesis, represented by the Rookeries Distance covariate, consists in the influence of seabirds (mainly penguins) in soil genesis by the deposition and microbial decomposition of guano, which enriches the soil through a sea-land transfer of nutrients (Abakumov et al., 2021; Bokhorst et al., 2007), notably P (which explains the great importance for this attribute), and promotes intense geochemical changes through the interaction of the guano leachates with the mineral substrate (Almeida et al., 2021; Pereira et al., 2013; Rodrigues et al., 2021; Tatur and Barczuk, 1985). The ornithogenesis is more effective in MA due to the large presence of rookeries, which explains the highest contents of P predicted in this region, with the already cited exception of Hope Bay.

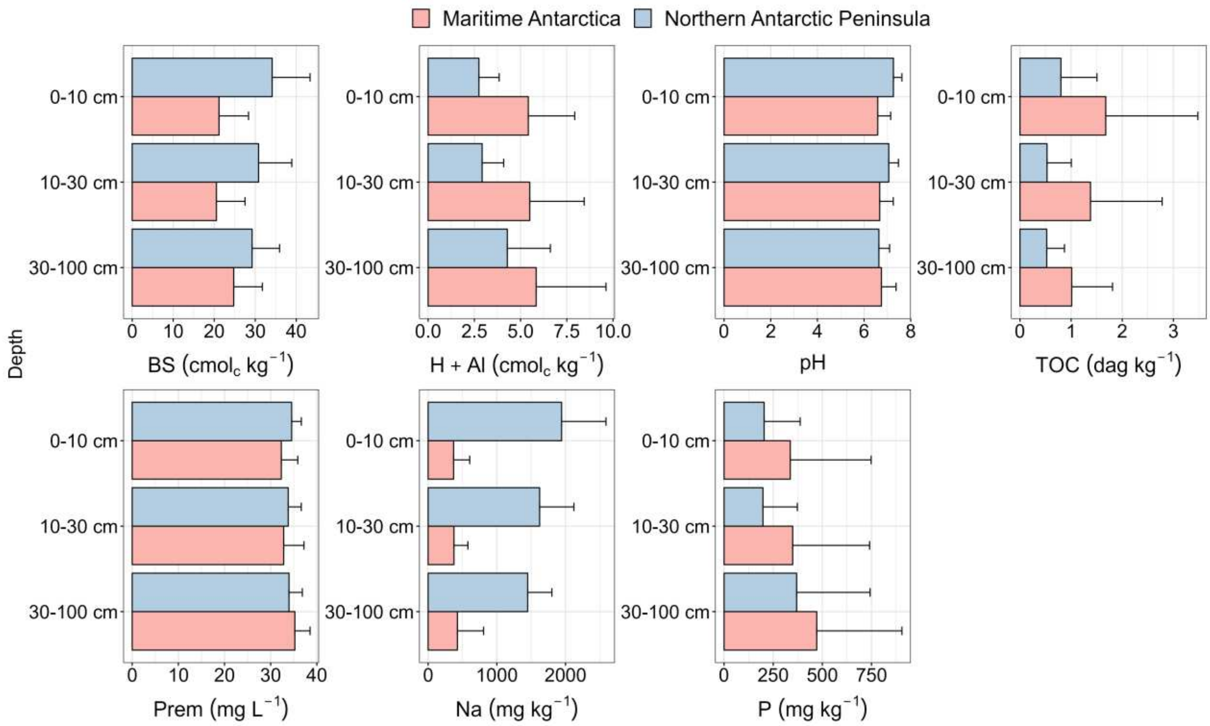


Figure 10. Mean contents of the predicted attributes in the final maps for the Maritime Antarctica and Northern Antarctic Peninsula regions. The values are the mean accounting the three depths along the 50 runs. Error bars mean the standard deviation.

The vegetation influence were also very important to differentiate the pedoenvironments of the mapped study area. The multispectral data associated to the vegetation cover used in this work have been successfully applied to map the spatial-temporal distribution of plant communities in Antarctica (Miranda et al., 2020; Victoria et al., 2013). The greater vegetation abundance, coupled with hydromorphic conditions (Thomazini et al., 2016), favours the organic matter accumulation in the MA soils (Michel et al., 2006; Rodrigues et al., 2019). This leads to the formation of carbon-rich soils that date back to more than 5 Ky BP (Björck et al., 1991) and make the MA ecossystems as one of the most important pools of carbon in Antarctica. This is evidenced by the greatest TOC contents in MA soils, which reach the maximum values in the areas where thick organic soils are common, whereas under the semi-arid climate of NAP, the extreme low temperatures and liquid water conditions the formation of ahumic soils (Daher et al., 2019a).

Although the present-day penguin rookeries are marked by the absence of vegetation, the ornithogenesis is an important factor to understand the TOC variability since the abandoned rookeries constitute high fertility sites enriched by the former birds activity, which favour plant colonization and the consequent input of organic matter into soil (Simas et al., 2007; Tatur and Myrcha, 1989). This relation also explains the importance of the vegetation indexes to the P prediction, showing the relevance of rookeries abandonment following the landscape evolution to spatially expand the ornithogenesis in Hope Bay and MA. Besides the present-day rookeries, the areas in Antarctica with the greatest P contents are also marked by the presence of several former rookeries abandoned after displacement of the birds to new sites next to the coast, after their older nesting sites have been uplifited by glacio-isostasy following deglaciation (Tatur et al., 1997; Tatur, 2002).

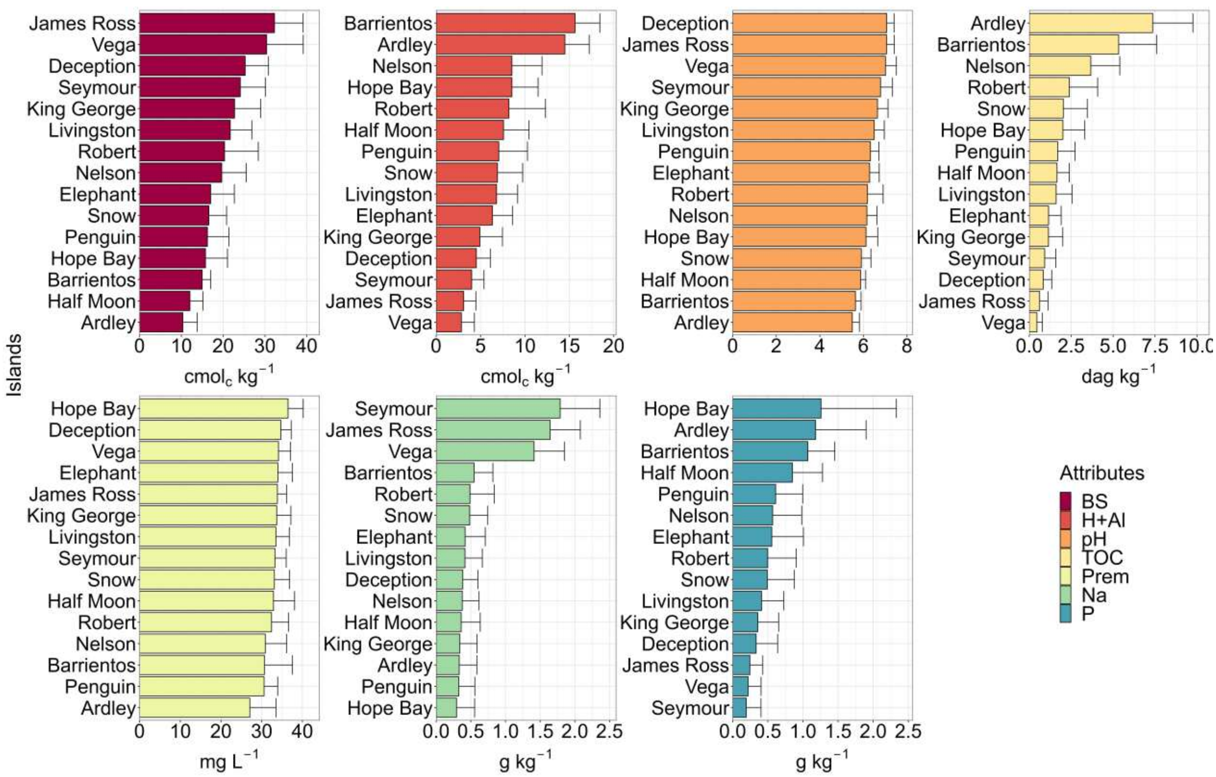


Figure 11. Mean contents of the predicted attributes in the final maps for the islands of the study area. The values are the mean accounting the three depths along the 50 runs. Error bars mean the standard deviation.

On the other hand, low P and TOC contents in islands with presence of penguin rookeries are related to local contexts where abandoned sites are not common, such as low rates of uplift in Seymour Island (the only island in NAP without a local glacier) or very recent and localized fauna occupation due to volcanic activities, such as in Deception Island. The great spatial associations of the H+Al with the TOC, and consequently with the environmental covariates which explains the carbon accumulation, such as the rookeries proximity and the vegetation indexes, is related to the abundance of superficial charges of the organic matter, which favours the concentration of the potential acidity in soils.

Considering the main covariates selected to the prediction of the soil pH values, it is possible to assume the leaching removing cationic bases and concentrating aluminium in the moister environments, and the vegetation influence, were the main factors driving the acidification in a wider scale in Antarctica. The release into soil of organic acids by the plants is a process very significant to the chemistry of the Antarctic soils (Otero et al., 2013), which is corroborated by the vegetation covariates as the most important to the pH prediction at the topsoil. The ornithogenesis is another factor very relevant to the soil reaction (Daher et al., 2019b), although it was not of great importance in the prediction. In part, this may be associated to the fact that the ornithogenic soils have a large pH range from alkaline to acidic conditions depending on the time of the guano within the soil, which can have hindered the recognition of a clear spatial pattern of the ornithogenesis with the pH.

Whereas many MA soils have features related to the moister climate and the organisms influence, the NAP environments, with exception of Hope Bay, are markedly characterized by the semi-desertic conditions of the region. Its first characteristic is the lower soil weathering degree in comparison to MA (Siqueira et al., 2022). Although drier climates tend to restrict the acidification due to mechanisms such as carbonate buffering, the differences of pH between the two regions are reduced due to the occurrence of the sulfurization, a pedogenic process which leads to the formation of acid sulfate soils through the pyrite oxidation (Lopes et al., 2019; Simas et al., 2006). The sulfurization is a typical process in NAP due to the presence of very common pyritized rocks and contributes to the lowering of the soil pH mainly in the deepest layers, where the contact with the dry atmosphere is reduced (Delpupo et al., 2014; Siqueira et al., 2021). The lowest BS and highest H+Al contents in the last layer indicate the greatest weathering rates in depth of the NAP soils, whereas in MA the highest weathering is observed in surface.

Another important feature of NAP soils is the extremely high Na contents associated with the salinization, typical process of the dry environments of the region (Delpupo et al., 2014). Salinization is the process of accumulation of soluble salts in soils, and is evidenced by the high concentration of Na in the topsoil, due to salts migration by the strong evaporation-activated capillarity (Bockheim and Ugolini, 1990). The Na contents are also higher next to the sea due to inputs by the marine spray, which is evidenced by the importance of the coastline distance predictor. Terrain attributes also highlights the importance of the drainage to the Na redistribution. Despite the local controls related to the coast and topography, the Na is the soil attribute which best responded to the climatic gradient between MA and NAP. On the other hand, the P-rem is the attribute modelled whose distribution demonstrated the higher azonal behaviour. The P-rem represents the P adsorption capacity of the soils and is a proxy of the weathering degree by indicating the formation of free Fe-oxides. In Antarctica, the areas with the least P-rem contents are those under the sulfurization (Francelino et al., 2011; Lopes et al., 2019), however, as this process occurs very locally in the ice-free areas without a spatial pattern, the final P-rem distribution was not related to any clear pedogenic process.

The relevance of the climate in the mapped areas is even more evidenced taking into account the parent material influence. Overall, considering only the geology contribution, we could expect the NAP soils would have a higher development degree due to the strongly pre-weathered sedimentary rocks which most of the soils are originated from, whereas the MA soils are developed mostly from notably bases-richer andesitic-basaltic rocks (Siqueira et al., 2022). The opposite of this pattern, i.e. NAP soils being richer and less weathered, depicts the importance of the climate and the climate-induced biogenic activity to confer the greater pedogenesis of the MA soils. This may in part explain the lower importance of the geological information for the spatial predictions in this work. However, this not mean the geology can not be an important factor at detailed scales (Formation level). For example, in NAP, more specifically in JRI, the sedimentary rocks bearing the highest Na contents are those whose marine deposition succeded in deltaic to shallow water environments (Pirrie, 1994), which is printed over the soils (Siqueira et al., 2021). In particular, this can be seen in Seymour Island, with the southern part of the island showing notably more Na than the northern part.

It was possible to observe that all major soil attributes predicted in our study tended to follow a topographic gradient in the ice-free areas mapped, evidenced by the varied importance of a series of terrain attributes. Whereas the climatic covariates controlled the distribution of the soil attributes in a macroscale, the topographic predictors tended to control the soil variability at local scales through strong soil-landscape relationships (Fig. 7), which match the findings of Balks et al. (2013). Overall, the highest parts of the landscape presented soils with higher pH and BS, indicating a lower weathering degree, whereas the low lying areas near the coastline presented higher TOC, H+Al and P contents, evidencing a higher biomass production and fauna activity influencing the chemical characteristics of the soils. This explains Ardley and Barrientos islands presenting the highest contents of TOC and P in MA, since they are small islands composed majoritaly of low marine terraces where penguin accessibility is facilitated due to the proximity to the sea.

In turn, the higher contents of BS and pH in the upper parts is also directly associated with the geomorphology of the Antarctic landscapes. Whereas the lowlands are marked by the conspicuous presence of uplifited, sandy marine terraces where the biological activity pioneering acts (Rodrigues et al., 2022), the highlands are rich in poorly weathered, silty-clayey glaciogenic sediments (moraines), whose grinding promoted originally by the glacial abrasion and then by periglacial processes tends to elevate the pH by releasing OH groups and bases (González-Guzmán et al., 2017), which are retained in the clay surfaces due to the low chemical weathering and leaching. In NAP, the higher eutrophy and alkalinity of the highlands also reflect the lithological diversity, with bases-rich basaltic soils predominating in the upper parts of the landscape (Daher et al., 2019a). The Na, in turn, have a non-monotonic behaviour, with high values next to the coast and, after a decrease, increasing again in the higher parts of the landscape.

4.2. Limitations of the DSM in Antarctica

It is important to recognize the limitations of our work as one of the first attempts to produce detailled digital soil maps in a very important region of Antarctica. One of the major limitations of the legacy data used in this study was the sources variability, with samples being collected from different soil surveys through decades of fieldwork in Antarctica, which may lead to uncertainties associated with the pedologist’s expertise, sampling instruments (e.g., GNSS devices with different positional errors) or different laboratory analyzes. Besides that, the samples used were collected under the regime of qualitative pedology, and directed by pedologists in the field to meet their different objectives (e.g., identification of representative profiles, collections for microbiology studies, etc.), which is also a limiting factor in the application of the DSM.

Although the importance proved in the modelling, it was noted the climatic and the distance variables with a typical “gradient pattern” entailed local spatial artefacts, even with the final values being consistent. This indicates a certain difficulty of our model in constructing smooth decision surface in some areas where the number of samples was not enough to detect the landscape variability. The RF model also had difficulties to predict extreme values, which was potentialized in our study due to the anoumalous characteristics of the so unique Antarctic soils. This problem also varied with the regions mapped. For example, while attributes such as

TOC and P were underestimated in MA where high contents of these attributes are relatively common, some pixels with anoumalous high values were created in NAP. Nonetheless, these restrictions are mainly associated to the intrisinc difficulties of obtaining soil data in Antarctica, which itself attests to the relevance of this type of work for the knowledge of the Antarctic soils.

Furthermore, the constraints cited here are gaps which must be adressed as new soil data is gradually added for the improvement of the Antarctic digital soil maps presented here, mainly in the ice-free areas where the sampling density is reduced and/or the error and the uncertainty metrics were higher. Besides improving the use of sampling strategies more adapted to the DSM (e.g., that covers more the environmental variability of the covariates), other options to maximize the new samplings in Antarctica can involve the use of NIR reflectance spectroscopy or even the adoption of a polar version of the homosoil concept, as proposed by Roudier et al. (2014), mainly for populating soil information in areas where soil data are sparser.

5. CONCLUSIONS

This work offers an advance in DSM by the compilation of one of the largest Antarctic soil databases for the production of digital soil maps on the only continent not yet covered by global soil mapping projects. In the present study we evaluated the performance of the machine learning algorithm Random Forest for the prediction of important soil chemical attributes in Antarctica, providing information about soil quality, soil genesis, pedogenic processes, soil ecological functions, and interplays with the Antarctic ecosystem.

Overall, the soil attributes showed a latitudinal/climatic gradient, with the BS, pH and sodium values being higher in the polar semideserts of the Northern Antarctic Peninsula, while the TOC, H+Al and P contents being greater in the warmer and moister Maritime Antarctica. The P-rem was the only attribute which did not present a behavior according to the latitude. The environmental covariates selected for the prediction represented important formation factors of the Antarctic soils, such as climatic drivers controlling the broad soil distribution trends, the organisms (vegetation and ornithogenesis) influencing in the coastal environments, and the topography driving soil variability at local scales.

Our results proved very satisfying in terms of environmental significance and realistic in terms of performance, and we hope the data generated here can be used to support scientific expeditions, decision-making about conservation of the Antarctic soils and terrestrial environments. Also, it can be useful for ecological, hydrological, and climatic models in Antarctica.

REFERENCES

Abakumov, E., 2018. Content of available forms of nitrogen, potassium, and phosphorus in ornithogenic and other soils of the Fildes Peninsula (King George Island, Western Antarctica). Biological Communications 63, 109–116. https://doi.org/10.21638/spbu03.2018.203

Abakumov, E., Lupachev, A., Andreev, M., Wang, W., Ji, X., 2021. The influence of brown and south polar skua on the content of plant nutrient in the soils from the Fildes Peninsula (King George Island, West Antarctica). Chemistry and Ecology 37, 185–199. https://doi.org/10.1080/02757540.2020.1839435

Adhikari, K., Kheir, R.B., Greve, M.B., Bøcher, P.K., Malone, B.P., Minasny, B., McBratney, A.B., Greve, M.H., 2013. High-Resolution 3-D Mapping of Soil Texture in Denmark. Soil Science Society of America Journal 77, 860–876. https://doi.org/10.2136/sssaj2012.0275

Alekseev, I., Abakumov, E., 2020. The content and distribution of trace elements and polycyclic aromatic hydrocarbons in soils of Maritime Antarctica. Environ Monit Assess 192. https://doi.org/10.1007/s10661-020-08618-2

Almeida, I.C.C., Schaefer, C.E.G.R., Fernandes, R.B.A., Oliveira, F.S., Pereira, T.T.C., 2021. Clay mineralogy and micropedology of phosphate-rich soils from Lions Rump, Maritime Antarctica. J South Am Earth Sci 105. https://doi.org/10.1016/j.jsames.2020.102967

Balks, M.R., López-Martínez, J., Goryachkin, S. v., Mergelov, N.S., Schaefer, C.E.G.R., Simas, F.N.B., Almond, P.C., Claridge, G.G.C., Mcleod, M., Scarrow, J., 2013. Windows on Antarctic soil-landscape relationships: Comparison across selected regions of Antarctica. Geol Soc Spec Publ 381, 397–410. https://doi.org/10.1144/SP381.9

Ballabio, C., Lugato, E., Fernández-Ugalde, O., Orgiazzi, A., Jones, A., Borrelli, P., Montanarella, L., Panagos, P., 2019. Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. Geoderma 355, 113912. https://doi.org/10.1016/j.geoderma.2019.113912

Barbraud, C., Gavrilo, M., Mizin, Y., Weimerskirch, H., 2011. Comparison of emperor penguin declines between Pointe Géologie and Haswell Island over the past 50 years. Antarct Sci 23, 461–468. https://doi.org/10.1017/S0954102011000356

Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modeling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91, 27–45. https://doi.org/10.1016/S0016-7061(99)00003-8

Björck, S., Malmer, N., Hjort, C., Sandgren, P., Ingólfsson, Ó., Wallén, B., Lewis Smith, R.I., Jónsson, B.L., 1991. Stratigraphic and Paleoclimatic Studies of a 5500-Year-Old Moss Bank on Elephant Island, Antarctica. Arctic & Alpine Research 23, 361–374.

Blumel, W.D., Eitel B., 1989. Geoecological aspects of maritime-climatic and continental periglacial regions in Antarctica (South Shetlands, Antarctic Peninsula and Victoria Land). Geökodynamic 10: 201-214.

Bockheim, J.G., Ugolini, F.C., 1990. A review of pedogenic zonation in well-drained soils of the Southern Circumpolar Region. Quat Res 34, 47–66. https://doi.org/10.1016/0033-5894(90)90072-S

Bockheim, J.G., 2015. The Soils of Antarctica, 1st ed. Springer International Publishing Switzerland, New York. https://doi.org/10.1007/978-3-319-05497-1_13

Bokhorst, S., Huiskes, A., Convey, P., Aerts, R., 2007. External nutrient inputs into terrestrial ecosystems of the Falkland Islands and the Maritime Antarctic region. Polar Biol 30, 1315–1321. https://doi.org/10.1007/s00300-007-0292-0

Bölter, M., Blume, H.P., Schneider, D., Beyer, L., 1997. Soil properties and distributions of invertebrates and bacteria from King George Island (Arctowski Station), maritime Antarctic. Polar Biol 18, 295–304. https://doi.org/10.1007/s003000050191

Boy, J., Godoy, R., Shibistova, O., Boy, D., McCulloch, R., de La Fuente, A.A., Morales, M.A., Mikutta, R., Guggenberger, G., 2016. Successional patterns along soil development gradients formed by glacier retreat in the Maritime Antarctic, King George Island. Revista Chilena de Historia Natural 89, 1–17. https://doi.org/10.1186/s40693-016-0056-8

Breiman, L., 2001. Random forests. Mach Learn 45, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Brenning, A., 2008. Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models, in: SAGA--Seconds Out (Hamburger Beiträge Zur Physischen Geographie Und Landschaftsökologie). pp. 23–32.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019

Camera, C., Zomeni, Z., Noller, J.S., Zissimos, A.M., Christoforou, I.C., Bruggeman, A., 2017. A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. Geoderma 285, 35–49. https://doi.org/10.1016/j.geoderma.2016.09.019

Carvalho, J.V.D.S., Sá Mendonça, E., la Scala, N., Reis, C., Reis, E.L., Schaefer, C.E.G.R., 2013. CO2-C losses and carbon quality of selected Maritime Antarctic soils. Antarct Sci 25, 11–18. https://doi.org/10.1017/S0954102012000648

Chagas, C. da S., de Carvalho Junior, W., Bhering, S.B., Calderano Filho, B., 2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena (Amst) 139, 232–240. https://doi.org/10.1016/j.catena.2016.01.001

Chagas, C. da S., Pinheiro, H.S.K., Carvalho Junior, W., Anjos, L.H.C., Pereira, N.R., Bhering, S.B., 2017. Data mining methods applied to map soil units on tropical hillslopes in Rio de Janeiro, Brazil. Geoderma Regional 9, 47–55. https://doi.org/10.1016/j.geodrs.2017.03.004

Convey, P., Peck, L.S., 2019. Antarctic environmental change and biological responses. Sci Adv 5. https://doi.org/10.1126/sciadv.aaz0888

Daher, M., Schaefer, C.E.G.R., Fernandes Filho, E.I., Francelino, M.R., Senra, E.O., 2019a. Semi-arid soils from a topolithosequence at James Ross Island, Weddell Sea region, Antarctica: Chemistry, mineralogy, genesis and classification. Geomorphology 327, 351–364. https://doi.org/10.1016/j.geomorph.2018.11.003

Daher, M., Schaefer, C.E.G.R., Thomazini, A., Lima Neto, E., Souza, C.D., Lopes, D. do V., 2019b. Ornithogenic soils on basalts from maritime Antarctica. Catena (Amst) 173, 367–374. https://doi.org/10.1016/j.catena.2018.10.028

Delpupo, C.S., Schaefer, C.E.G.R., Simas, F.N.B., Spinola, D.N., Daher, M., 2014. Soil formation in Seymour Island, Weddell Sea, Antarctica. Geomorphology 255, 87–99. https://doi.org/10.1016/j.geomorph.2014.03.047

Dharumarajan, S., Hegde, R., 2022. Digital mapping of soil texture classes using Random Forest classification algorithm. Soil Use Manag 38, 135–149. https://doi.org/10.1111/sum.12668

Dharumarajan, S., Hegde, R., Singh, S.K., 2017. Spatial prediction of major soil properties using Random Forest techniques - A case study in semi-arid tropics of South India. Geoderma Regional 10, 154–162. https://doi.org/10.1016/j.geodrs.2017.07.005

Díaz-Puente, F.J., Schmid, T., Pelayo, M., Rodríguez-Rastrero, M., Herraiz, M.J.S., O'Neill, T., López-Martínez, J., 2021. Abiotic factors influencing soil microbial activity in the northern Antarctic Peninsula region. Science of the Total Environment 750. https://doi.org/10.1016/j.scitotenv.2020.141602

Fermani, P., Mataloni, G., van de Vijver, B., 2007. Soil microalgal communities on an antarctic active volcano (Deception Island, South Shetlands). Polar Biol 30, 1381–1393. https://doi.org/10.1007/s00300-007-0299-6

Ferrari, F.R., Schaefer, C.E.G.R., Pereira, A.B., Schmitz, D., Francelino, M.R., 2020. Coupled soil-vegetation changes along a topographic gradient on King George Island, maritime Antarctica. Catena (Amst) 105038. https://doi.org/10.1016/j.catena.2020.105038

Francelino, M.R., Schaefer, C.E.G.R., Simas, F.N.B., Filho, E.I.F., de Souza, J.J.L.L., da Costa, L.M., 2011. Geomorphology and soils distribution under paraglacial conditions in an ice-free area of Admiralty Bay, King George Island, Antarctica. Catena (Amst) 85, 194–204. https://doi.org/10.1016/j.catena.2010.12.007

Ganzert, L., Lipski, A., Hubberten, H.W., Wagner, D., 2011. The impact of different soil parameters on the community structure of dominant bacteria from nine different soils located on Livingston Island, South Shetland Archipelago, Antarctica. FEMS Microbiol Ecol 76, 476–491. https://doi.org/10.1111/j.1574-6941.2011.01068.x

Gjorup, D.F., Ernesto, C., Schaefer, G.R., Nogueira, F., Simas, B., Rocha, M., Ferreira, R., Michel, M., Bockheim, J.G., 2020. Sulfurization, Acid-Sulfate soils and active layer monitoring at the semiarid Seymour Island, Antarctica. Geoderma Regional 22, e00305. https://doi.org/10.1016/j.geodrs.2020.e00305

Gomes, L.C., Faria, R.M., de Souza, E., Veloso, G.V., Schaefer, C.E.G.R., Filho, E.I.F., 2019. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma 340, 337–350. https://doi.org/10.1016/j.geoderma.2019.01.007

González-Guzmán, A., Oliva, M., Souza-Júnior, V.S., Pérez-Alberti, A., Ruiz-Fernández, J., Otero, X.L., 2017. Biota and geomorphic processes as key environmental factors controlling soil formation at Elephant Point, Maritime Antarctica. Geoderma 300, 32–43. https://doi.org/10.1016/j.geoderma.2017.04.001

Greve, M.H., Kheir, R.B., Greve, M.B., Bøcher, P.K., 2012. Quantifying the ability of environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: The case study of Denmark. Ecol Indic 18, 1–10. https://doi.org/10.1016/j.ecolind.2011.10.006

Guo, P.T., Li, M.F., Luo, W., Tang, Q.F., Liu, Z.W., Lin, Z.M., 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. Geoderma 237–238, 49–59. https://doi.org/10.1016/j.geoderma.2014.08.009

Guo Y., Wang N., Li G., Rosas G., Zang J., Ma Y., Liu J., Han W., Cao H., 2018. Direct and indirect effects of penguin feces on microbiomes in Antarctic ornithogenic soils. Front Microbiol 9. https://doi.org/10.3389/fmicb.2018.00552

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach Learn 46, 389–422. https://doi.org/10.1023/A:1012487302797

Haus, N.W., Wilhelm, K.R., Bockheim, J.G., Fournelle, J., Miller, M., 2016. A case for chemical weathering in soils of Hurd Peninsula, Livingston Island, South Shetland Islands, Antarctica. Geoderma 263, 185–194. https://doi.org/10.1016/j.geoderma.2015.09.019

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., de Jesus, J.M., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS One 10, 1–26. https://doi.org/10.1371/journal.pone.0125814

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ Prepr. https://doi.org/10.7717/peerj.5518

Hounkpatin, O.K.L., op de Hipt, F., Bossa, A.Y., Welp, G., Amelung, W., 2018. Soil organic carbon stocks and their determining factors in the Dano catchment (Southwest Burkina Faso). Catena (Amst) 166, 298–309. https://doi.org/10.1016/j.catena.2018.04.013

Howat, I.M., Porter, C., Smith, B.E., Noh, M.J., Morin, P., 2019. The reference elevation model of antarctica. Cryosphere 13, 665–674. https://doi.org/10.5194/tc-13-665-2019

Husson, F., Josse, J., Le, S., Mazet, J., 2023. Multivariate Exploratory Data Analysis and Data Mining. R package version 2.8.

Ilieva, R., Yaneva, R., Zhiyanski, M., Abakumov, E., 2021. Morphological and physico-chemical properties of Cryosoils in the Bulgarian antarctic base on Livingston island, Antarctica. Silva Balcanica 22, 57–67. https://doi.org/10.3897/silvabalcanica.22.e66704

Jennings, S.J.A., Davies, B.J., Nývlt, D., Glasser, N.F., Engel, Z., Hrbáček, F., Carrivick, J.L., Mlčoch, B., Hambrey, M.J., 2021. Geomorphology of Ulu Peninsula, James Ross Island, Antarctica. J Maps 17, 125–139. https://doi.org/10.1080/17445647.2021.1893232

Jordanova, D., Jordanova, N., Dimov, D., Georgieva, B., Ishlyamski, D., 2022. The role of tephra additions on development of incipient soils from Livingston Island (Antarctic Peninsula) revealed by environmental magnetism. Catena (Amst) 212, 106103. https://doi.org/10.1016/j.catena.2022.106103

Kalambukattu, J.G., Kumar, S., Arya Raj, R., 2018. Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. Environ Earth Sci 77, 1–14. https://doi.org/10.1007/s12665-018-7367-9

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2017. Climatologies at high resolution for the earth's land surface areas. Sci Data 4, 1–20. https://doi.org/10.1038/sdata.2017.122

Kelleher, J.D., MacNamee, J., D'Arcy, A., 2015. Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies.

Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. Appl Math Model 81, 401–418. https://doi.org/10.1016/j.apm.2019.12.016

Kozeretska, I.A., Parnikoza, I.Y., Mustafa, O., Tyschenko, O. v., Korsun, S.G., Convey, P., 2010. Development of Antarctic herb tundra vegetation near Arctowski station, King George Island. Polar Sci 3, 254–261. https://doi.org/10.1016/j.polar.2009.10.001

Krauze, P., Wagner, D., Yang, S., Spinola, D., Kühn, P., 2021. Influence of prokaryotic microorganisms on initial soil formation along a glacier forefield on King George Island, maritime Antarctica. Sci Rep 11, 1–16. https://doi.org/10.1038/s41598-021-92205-z

Kuhn, M., 2008. Building predictive models in R using the caret package. J Stat Softw 28, 1–26. https://doi.org/10.18637/jss.v028.i05

Kumaraperumal, R., Pazhanivelan, S., Geethalakshmi, V., Raj, M.N., Muthumanickam, D., Kaliaperumal, R., Shankar, V., Nair, A.M., Yadav, M.K., Vendan, T., Kshatriya, T., 2022. Comparison of Machine Learning-Based Prediction of Qualitative and Quantitative Digital Soil-Mapping Approaches. Land (Basel) 11. https://doi.org/10.3390/land11122279

Łachacz, A., Kalisz, B., Giełwanowska, I., Olech, M., Chwedorzewska, K.J., Kellmann-Sopyła, W., 2018. Nutrient abundance and variability from soils in the coast of king George Island. J Soil Sci Plant Nutr 18, 294–311. https://doi.org/10.4067/S0718-95162018005001101

Lagacherie, P., McBratney, A.B., 2006. Chapter 1 Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping. Developments in Soil Science 31, 3–22. https://doi.org/10.1016/S0166-2481(06)31001-X

Lee, Y. il, Lim, H.S., Yoon, H. il, 2004. Geochemistry of soils of King George Island, South Shetland Islands, West Antarctica: Implications for pedogenesis in cold polar regions. Geochim Cosmochim Acta 68, 4319–4333. https://doi.org/10.1016/j.gca.2004.01.020

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R news 2/3, 18–22.

Liaw, A., Wiener, M., 2022. Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.7-1.1.

Lopes, D. do V., Oliveira, F.S. de, Pereira, T.T.C., Schaefer, C.E.G.R., 2022. Pedogeomorphology and weathering at Snow Island, Maritime Antarctica. Catena (Amst) 217. https://doi.org/10.1016/j.catena.2022.106515

Lopes, D. do V., Schaefer, C.E.G.R., Souza, J.J.L.L. de, Oliveira, F.S., Simas, F.N.B., Daher, M., Gjorup, D.F., 2019. Concretionary horizons, unusual pedogenetic processes and features of sulfate affected soils from Antarctica. Geoderma 347, 13–24. https://doi.org/10.1016/j.geoderma.2019.03.024

Lopes, D.D.V., de Oliveira, F.S., de Souza, J.J.L.L., Machado, M.D.R., Schaefer, C.E.G.R., 2022. Soil pockets phosphatization and chemical weathering of sites affected by flying birds of Maritime Antarctica. An Acad Bras Cienc 94. https://doi.org/10.1590/0001-3765202220210595

López-Martínez, J., Serrano, E., Schmid, T., Mink, S., Linés, C., 2012. Periglacial processes and landforms in the South Shetland Islands (northern Antarctic Peninsula region). Geomorphology 155–156, 62–79. https://doi.org/10.1016/j.geomorph.2011.12.018

Ma, Y., Minasny, B., Malone, B.P., Mcbratney, A.B., 2019. Pedology and digital soil mapping (DSM). Eur J Soil Sci 70, 216–235. https://doi.org/10.1111/ejss.12790

Ma, Y., Minasny, B., Wu, C., 2017. Mapping key soil properties to support agricultural production in Eastern China. Geoderma Regional 10, 144–153. https://doi.org/10.1016/j.geodrs.2017.06.002

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., Beaudoin, A., 2014. Digital mapping of soil properties in Canadian managed forests at 250m of resolution using the k-nearest neighbor method. Geoderma 235–236, 59–73. https://doi.org/10.1016/j.geoderma.2014.06.032

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping, Geoderma. https://doi.org/10.1016/S0016-7061(03)00223-4

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey, Geoderma. https://doi.org/10.1016/S0016-7061(00)00043-4

Meier, L.A., Krauze, P., Prater, I., Horn, F., Schaefer, C.E.G.R., Scholten, T., Wagner, D., Mueller, C.W., Kühn, P., 2019. Pedogenic and microbial interrelation in initial soils under semiarid climate on James Ross Island, Antarctic Peninsula region. Biogeosciences 16, 2481–2499. https://doi.org/10.5194/bg-16-2481-2019

Meier, M., de Souza, E., Francelino, M.R., Fernandes Filho, E.I., Schaefer, C.E.G.R., 2018. Digital soil mapping using machine learning algorithms in a tropical mountainous area. Rev Bras Cienc Solo 42, 1–22. https://doi.org/10.1590/18069657rbcs20170421

Meier, M., Francelino, M.R., Thomazini, A., Pereira, A.B., Krüger, F.L. von, Fernandes-Filho, E.I., Schaefer, C.E.G.R., 2023. Soilscapes and Geoenvironments at Stansbury Peninsula, Nelson Island, Maritime Antarctica. Catena (Amst). https://doi.org/10.2139/ssrn.3959649

Mendonça, T., Melo, V.F., Schaefer, C.E.G.R., Simas, F.N.B., Michel, R.F.M., 2013. Clay Mineralogy of Gelic Soils from the Fildes Peninsula, Maritime Antarctica. Soil Science Society of America Journal 77, 1842–1851. https://doi.org/10.2136/sssaj2012.0135

Michel, R.F.M., Schaefer, C.E.G.R., Dias, L.E., Simas, F.N.B., Benites, Vinicius de Melo Mendonça, E. de S., 2006. Ornithogenic Gelisols (Cryosols) from Maritime Antarctica: Pedogenesis, Vegetation, and Carbon Studies. Soil Science Society of America 70, 1370–1376. https://doi.org/10.2136/sssaj2005.0178

Michel, R.F.M., Schaefer, C.E.G.R., López-Martínez, J., Simas, F.N.B., Haus, N.W., Serrano, E., Bockheim, J.G., 2014. Soils and landforms from Fildes Peninsula and Ardley Island, Maritime Antarctica. Geomorphology 225, 76–86. https://doi.org/10.1016/j.geomorph.2014.03.041

Minasny, B., Malone, B., Stockmann, U., Odgers, N., McBratney, A.B.., 2014. Pedometrics. Reference Module in Earth Systems and Environmental Sciences 1–10. https://doi.org/10.1016/b978-0-12-409548-9.09163-6

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Miranda, V., Pina, P., Heleno, S., Vieira, G., Mora, C., E.G.R. Schaefer, C., 2020. Monitoring recent changes of vegetation in Fildes Peninsula (King George Island, Antarctica) through satellite imagery guided by UAV surveys. Science of the Total Environment 704, 135295. https://doi.org/10.1016/j.scitotenv.2019.135295

Moreno, L., Silva-Busso, A., López-Martínez, J., Durán-Valsero, J.J., Martínez-Navarrete, C., Cuchí, J.A., Ermolin, E., 2012. Hydrogeochemical characteristics at Cape Lamb, Vega Island, Antarctic Peninsula. Antarct Sci 24, 591–607. https://doi.org/10.1017/S0954102012000478

Moura, P.A., Francelino, M.R., Schaefer, C.E.G.R., Simas, F.N.B., Mendonça, B.A.F. de, 2012. Distribution and characterization of soils and landform relationships in Byers Peninsula, Livingston Island, Maritime Antarctica. Geomorphology 155–156, 45–54. https://doi.org/10.1016/j.geomorph.2011.12.011

Navas, A., López-Martínez, J., Casas, J., Machín, J., Durán, J.J., Serrano, E., Cuchi, J.A., Mink, S., 2008. Soil characteristics on varying lithological substrates in the South Shetland Islands, maritime Antarctica. Geoderma 144, 123–139. https://doi.org/10.1016/j.geoderma.2007.10.011

Navas, A., Oliva, M., Ruiz-Fernández, J., Gaspar, L., Quijano, L., Lizaga, I., 2017. Radionuclides and soil properties as indicators of glacier retreat in a recently deglaciated permafrost environment of the Maritime Antarctica. Science of the Total Environment 609, 192–204. https://doi.org/10.1016/j.scitotenv.2017.07.115

Navas, A., Serrano, E., López-Martínez, J., Gaspar, L., Lizaga, I., 2018. Interpreting environmental changes from radionuclides and soil characteristics in different landform contexts of Elephant Island (maritime Antarctica). Land Degrad Dev 29, 3141–3158. https://doi.org/10.1002/ldr.2987

Navas, A., Soto, J., López-Martínez, J., 2005. Radionuclides in soils of Byers Peninsula, South Shetland Islands, Western Antarctica. Applied Radiation and Isotopes 62, 809–816. https://doi.org/10.1016/j.apradiso.2004.11.007

Nelson, M.A., Bishop, T.F.A., Triantafilis, J., Odeh, I.O.A., 2011. An error budget for different sources of error in digital soil mapping. Eur J Soil Sci 62, 417–430. https://doi.org/10.1111/j.1365-2389.2011.01365.x

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil 4, 1–22. https://doi.org/10.5194/soil-4-1-2018

O'Brien, L., Malone, B., Hengl, T., Bishop, T., Rossiter, D., Beaudette, D., Brown, A., 2022. Mass-Preserving Spline Functions for Soil Data. R package version 0.1.6.

Otero, X.L., Fernández, S., de Pablo Hernandez, M.A., Nizoli, E.C., Quesada, A., 2013. Plant communities as a key factor in biogeochemical processes involving micronutrients (Fe, Mn, Co, and Cu) in Antarctic soils (Byers Peninsula, maritime Antarctica). Geoderma 195–196, 145–154. https://doi.org/10.1016/j.geoderma.2012.11.018

Pereira, Thiago Torres C., Schaefer, C.E.G.R., Ker, J.C., Almeida, C.C., Almeida, I.C.C., Pereira, A.Ô.B., 2013. Genesis, mineralogy, and ecological significance of ornithogenic soils from a semi-desert polar landscape at Hope Bay, Antarctic Peninsula. Geoderma 209–210, 98–109. https://doi.org/10.1016/j.geoderma.2013.06.012

Perfetti-Bolaño, A., Moreno, L., Urrutia, R., Araneda, A., Barra, R., 2018. Influence of Pygoscelis Penguin Colonies on Cu and Pb Concentrations in Soils on the Ardley Peninsula, Maritime Antarctica. Water Air Soil Pollut 229. https://doi.org/10.1007/s11270-018-4042-4

Pires, C. V., Schaefer, C.E.R.G., Hashigushi, A.K., Thomazini, A., Filho, E.I.F., Mendonça, E.S., 2017. Soil organic carbon and nitrogen pools drive soil C-CO2 emissions from selected soils in Maritime Antarctica. Science of the Total Environment 596–597, 124–135. https://doi.org/10.1016/j.scitotenv.2017.03.144

Pirrie, D., 1994. Petrography and provenance of the Marambio Group, Vega Island, Antarctica. Antarct Sci 6, 517–527. https://doi.org/10.1017/S0954102094000775

Poelking, E.L., Schaefer, C.E.R., Fernandes Filho, E.I., Andrade, A.M., Spielmann, A.A., 2015. Soil – landform – plant-community relationships of a periglacial landscape on Potter Peninsula, maritime Antarctica. Solid Earth 583–594. https://doi.org/10.5194/se-6-583-2015

Poggere, G.C., Melo, V.F., Francelino, M.R., Schaefer, C.E., Simas, F.N., 2016. Characterization of products of the early stages of pedogenesis in ornithogenic soil from Maritime Antarctica. Eur J Soil Sci 67, 70–78. https://doi.org/10.1111/ejss.12307

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. Soil 7, 217–240. https://doi.org/10.5194/soil-7-217-2021

Prater, I., Hrbáček, F., Braun, C., Vidal, A., Meier, L.A., Nývlt, D., Mueller, C.W., 2021. How vegetation patches drive soil development and organic matter formation on polar islands. Geoderma Regional 27. https://doi.org/10.1016/j.geodrs.2021.e00429

Prietzel, J., Prater, I., Colocho Hurtarte, L.C., Hrbáček, F., Klysubun, W., Mueller, C.W., 2019. Site conditions and vegetation determine phosphorus and sulfur speciation in soils of Antarctica. Geochim Cosmochim Acta 246, 339–362. https://doi.org/10.1016/j.gca.2018.12.001

R Core Team, 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project. org/.

Ramírez-Fernández, L., Trefault, N., Carú, M., Orlando, J., 2019. Seabird and pinniped shape soil bacterial communities of their settlements in Cape Shirreff, Antarctica. PLoS One 14, 1–22. https://doi.org/10.1371/journal.pone.0209887

Riza, S., Sekine, M., Kanno, A., Yamamoto, K., Imai, T., Higuchi, T., 2021. Modeling soil landscapes and soil textures using hyperscale terrain attributes. Geoderma 402, 115177. https://doi.org/10.1016/j.geoderma.2021.115177

Rodrigues, W.F., Oliveira, F.S., Schaefer, C.E.G.R., Leite, M.G.P., Gauzzi, T., Bockheim, J.G., Putzke, J., 2019. Soil-landscape interplays at Harmony Point, Nelson Island, Maritime Antarctica: Chemistry, mineralogy, and classification. Geomorphology 336, 77–94. https://doi.org/10.1016/j.geomorph.2019.03.030

Rodrigues, W.F., Soares, F. O., Schaefer, C.E.G.R., Leite, M.G.P., Pavinato, P.S., 2021. Phosphatization under birds' activity: Ornithogenesis at different scales on Antarctic Soilscapes. Geoderma 391. https://doi.org/10.1016/j.geoderma.2021.114950

Rodrigues, W.F., Machado, M. R., Oliveira, F.S., Schaefer, C.E.G.R., Leite, M.G.P., Michel, R.F.M., Gauzzi, T., 2022. Soil-chronosequence and Quaternary landscape evolution at the marine terraces of Harmony Point, Nelson Island, Maritime Antarctica. An Acad Bras Cienc 94, 1–20. https://doi.org/10.1590/0001-3765202220201141

Roesch, L.F.W., Fulthorpe, R.R., Pereira, A.B., Pereira, C.K., Lemos, L.N., Barbosa, A.D., Suleiman, A.K.A., Gerber, A.L., Pereira, M.G., Loss, A., da Costa, E.M., 2012. Soil bacterial community abundance and diversity in ice-free areas of Keller Peninsula, Antarctica. Applied Soil Ecology 61, 7–15. https://doi.org/10.1016/j.apsoil.2012.04.009

Román Dobarco, M., Arrouays, D., Lagacherie, P., Ciampalini, R., Saby, N.P.A., 2017. Prediction of topsoil texture for Region Centre (France) applying model ensemble methods. Geoderma 298, 67–77. https://doi.org/10.1016/j.geoderma.2017.03.015

Schaefer, C.E.G.R., Pereira, T.T.C., Ker, J.C., Almeida, I.C.C., Simas, F.N.B., Oliveira, F.S., Correa, G.R., Vieira, G., 2015. Soils and Landforms at Hope Bay, Antarctic Peninsula: Formation, Classification, Distribution, and Relationships. Soil Science Society of America Journal 79, 175–184. https://doi.org/10.2136/sssaj2014.06.0266

Schmitz, D., Michel, R.F.M., Ferrari, F.R., Villa, P.M., Francelino, M.R., Putzke, J., López-Martínez, J., Schaefer, C.E.G.R., 2022. Soil-landform-vegetation interplays at Stinker Point, Elephant Island, Antarctica. An Acad Bras Cienc 94. https://doi.org/10.1590/0001-3765202220210676

Schmitz, D., Villa, P.M., Michel, R.F.M., Putzke, J., Pereira, A.B., Schaefer, C.E.G.R., 2021. Species composition, diversity and coverage pattern of associated communities of mosses-lichens along a pedoenvironmental gradient in Maritime Antarctica. An Acad Bras Cienc 94, 1–17. https://doi.org/10.1590/0001-3765202120200094

Schmitz, D., Schaefer, C.E.R.G., Putzke, J., Francelino, M.R., Ferrari, F.R., Corrêa, G.R., Villa, P.M., 2020a. How does the pedoenvironmental gradient shape non-vascular species assemblages and community structures in Maritime Antarctica? Ecol Indic 108, 105726. https://doi.org/10.1016/j.ecolind.2019.105726

Schmitz, D., Villa, P.M., Putzke, J., Michel, R.F.M., Campos, P.V., Neto, J.A.A.M., Schaefer, C.E.G.R., 2020b. Diversity and species associations in cryptogam communities along a pedoenvironmental gradient on Elephant Island, Maritime Antarctica. Folia Geobot 55, 211–224. https://doi.org/10.1007/s12224-020-09376-2

Schünemann, A.L., Almeida, P.H.A., Francelino, M.R., Fernandes-Filho, E.I., Santos, G.R., Daher, M., Schaefer, C.E.G.R., Pereira, A.B., 2022. Soil pedogeochemical attributes prediction by interpolators in ice-free areas of Antarctica. Research, Society and Development 11, 1–15.

Silva, J.P., Souza, J.J.L.L. de, Soares, E.M.B., Schaefer, C.E.G.R., 2022. Soil organic matter accumulation before, during, and after the last glacial maximum in Byers Peninsula, Maritime Antarctica. Geoderma 428, 116221. https://doi.org/10.1016/j.geoderma.2022.116221

Simas, F., Schaefer, C.E.G.R., Melo, V.F., Albuquerque-Filho, M.R., Michel, R.F.M., Pereira, V. v., Gomes, M.R.M., Costa, L.M., 2007. Ornithogenic cryosols from Maritime Antarctica: Phosphatization as a soil forming process. Geoderma 138, 191–203.

Simas, F.N.B., Schaefer, C.E.G.R., Albuquerque-Filho, M.R., Francelino, M.R., Fernandes-Filho, E.I., Costa, L.M. da, 2008. Genesis, properties and classification of Cryosols from Admiralty Bay, maritime Antarctica. Geoderma 144, 116–122. https://doi.org/10.1016/j.geoderma.2007.10.019

Simas, F.N.B., Schaefer, C.E.G.R., Melo, V.F., Guerra, M.B.B., Saunders, M., Gilkes, R.J., 2006. Clay-sized minerals in permafrost-affected soils (Cryosols) from King George Island, Antarctica. Clays Clay Miner 54, 721–736. https://doi.org/10.1346/CCMN.2006.0540607

Siqueira, R.G., Lopes, D. v., de Souza, J.J.L.L., Schaefer, C.E.G.R., Souza, C.D., de Oliveira, F.S., Fernandes Filho, E.I., 2022. Acid sulfate soils from Antarctica: Genesis and properties along a climatic gradient. An Acad Bras Cienc 94, 1–20. https://doi.org/10.1590/0001-3765202120210625

Siqueira, R.G., Schaefer, C.E.G.R., Fernandes Filho, E.I., Corrêa, G.R., Francelino, M.R., Souza, J.J.L.L. de, Rocha, P. de A., 2021. Weathering and pedogenesis of sediments and basaltic rocks on Vega Island, Antarctic Peninsula. Geoderma 382, 114707. https://doi.org/10.1016/j.geoderma.2020.114707

Smellie, J.L., Hunt, R.J., McIntosh, W.C., Esser, R.P., 2021a. Lithostratigraphy, age and distribution of Eocene volcanic sequences on eastern King George Island, South Shetland Islands, Antarctica. Antarct Sci 33, 373–401. https://doi.org/10.1017/S0954102021000213

Smellie, J.L., McIntosh, W.C., Whittle, R., Troedson, A., Hunt, R.J., 2021b. A lithostratigraphical and chronological study of Oligocene-Miocene sequences on eastern King George Island, South Shetland Islands (Antarctica), and correlation of glacial episodes with global isotope events. Antarct Sci 33, 502–532. https://doi.org/10.1017/S095410202100033X

Taalab, K., Corstanje, R., Zawadzka, J., Mayr, T., Whelan, M.J., Hannam, J.A., Creamer, R., 2015. On the application of Bayesian Networks in Digital Soil Mapping. Geoderma 259–260, 134–148. https://doi.org/10.1016/j.geoderma.2015.05.014

Taghizadeh-Mehrjardi, R., Schmidt, K., Eftekhari, K., Behrens, T., Jamshidi, M., Davatgar, N., Toomanian, N., Scholten, T., 2020. Synthetic resampling strategies and machine learning for digital soil mapping in Iran. Eur J Soil Sci 71, 352–368. https://doi.org/10.1111/ejss.12893

Tatur, A., 2002. Ornithogenic Ecosystems in the Maritime Antarctic — Formation, Development and Disintegration. In: Beyer, L., Bölter, M. (eds) Geoecology of Antarctic Ice-Free Coastal Landscapes. Ecological Studies, vol 154. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-56318-8_10

Tatur, A., Barczuk, A., 1985. Ornithogenic phosphates on King George Island in the maritime Antarctic. Antarctic nutrient cycles and food webs 163–168. https://doi.org/10.1007/978-3-642-82275-9_22

Tatur, A.E., Myrcha, A., 1984. Ornithogenic soils on King George lsland, South Shetland lslands (Maritime Antarctic Zone). Pol Polar Res 5, 31–60.

Tatur, A., Myrcha, A., 1989. Soils and Vegetation in Abandoned Penguin Rookeries (Maritime Antarctic). Polar Biol 2, 181–189.

Tatur, A., Myrcha, A., Niedgodzisz, J., 1997. Formation of abandoned penguin rookery ecosystems in the maritime Antarctic. Polar Biol 17, 405-417.

Thomazini, A., Francelino, M.R., Pereira, A.B., Schünemann, A.L., Mendonça, E. de S., Michel, R.F.M., Schaefer, C.E.G.R., 2020. The current response of soil thermal regime and carbon exchange of a paraglacial coastal land system in maritime Antarctica. Land Degrad Dev 31, 655–666. https://doi.org/10.1002/ldr.3479

Thomazini, A., Francelino, M.R., Pereira, A.B., Schünemann, A.L., Mendonça, E.S., Almeida, P.H.A., Schaefer, C.E.G.R., 2016. Geospatial variability of soil CO2-C exchange in the main terrestrial ecosystems of Keller Peninsula, Maritime Antarctica. Science of the Total Environment 562, 802–811. https://doi.org/10.1016/j.scitotenv.2016.04.043

Thompson, J.A., Roecker, S., Grunwald, S., Owens, P.R., 2012. Digital soil mapping: Interactions with and applications for hydropedology. Hydropedology 1, 665–709. https://doi.org/10.1016/B978-0-12-386941-8.00021-6

Turner, J., Barrand, N.E., Bracegirdle, T.J., Convey, P., Hodgson, D.A., Jarvis, M., Jenkins, A., Marshall, G., Meredith, M.P., Roscoe, H., Shanklin, J., French, J., Goosse, H., Guglielmin, M., Gutt, J., Jacobs, S., Kennicutt, M.C., Masson-Delmotte, V., Mayewski, P., Navarro, F., Robinson, S., Scambos, T., Sparrow, M., Summerhayes, C., Speer, K., Klepikov, A., 2014. Antarctic climate change and the environment: An update. Polar Record 50, 237–259. https://doi.org/10.1017/S0032247413000296

van Lipzig, N.P.M., King, J.C., Lachlan-Cope, T.A., van den Broeke, M.R., 2004. Precipitation, sublimation, and snow drift in the Antarctic Peninsula region from a regional atmospheric model. J Geophys Res 109, 1–16. https://doi.org/10.1029/2004JD004701

Vaughan, D.G., Marshall, G.G., Connoley, W.M., Parkinson, C., Mulvaney, R., Hodgson, D.A., King, J.C., Pudsey, C.J., Turner, J., 2003. Recent rapid regional climate warming on the Antarctic Peninsula. Clim Change 60, 243–274. https://doi.org/10.1023/A

Victoria, F. de C., de Albuquerque, M.P., Pereira, A.B., Simas, F.N.B., Spielmann, A.A., Schaefer, C.E.G.R., 2013. Characterization and mapping of plant communities at Hennequin Point, King George Island, Antarctica. Polar Res 32, 0–10. https://doi.org/10.3402/polar.v32i0.19261

Vieira, F.C.B., Pereira, A.B., Bayer, C., Schünemann, A.L., Victoria, F. de C., de Albuquerque, M.P., de Oliveira, C.S., 2013. In situ methane and nitrous oxide fluxes in soil from a transect in Hennequin Point, King George Island, Antarctic. Chemosphere 90, 497–504. https://doi.org/10.1016/j.chemosphere.2012.08.013

Vlcek, V., Pospíšilová, L., Uhlík, P., 2018. Mineralogy and chemical composition of cryosols and andosols in Antarctica. Soil and Water Research 13, 61–73. https://doi.org/10.17221/231/2016-SWR

Wadoux, A.M.J.C., Román-Dobarco, M., McBratney, A.B., 2021. Perspectives on data-driven soil research. Eur J Soil Sci 72, 1675–1689. https://doi.org/10.1111/ejss.13071

Wadoux, A.M.J.-C., Saby, N.P.A., Martin, M.P., 2023. Shapley values reveal the drivers of soil organic carbon stock prediction. Soil 9, 21–38. https://doi.org/10.5194/soil-9-21-2023

Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Comerford, N.B., 2014. Holistic environmental soil-landscape modeling of soil organic carbon. Environmental Modelling and Software 57, 202–215. https://doi.org/10.1016/j.envsoft.2014.03.004

Yang, R.M., Zhang, G.L., Liu, F., Lu, Y.Y., Yang, Fan, Yang, Fei, Yang, M., Zhao, Y.G., Li, D.C., 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecol Indic 60, 870–878. https://doi.org/10.1016/j.ecolind.2015.08.036

CAPÍTULO 3

MACHINE LEARNING FOR MAPPING SOIL CLASSES IN ANTARCTICA

ABSTRACT

Conventional soil mapping based on mental models is a common approach for understanding the soil distribution since the first soil surveys in Antarctica. However, Digital Soil Mapping based on empirical, quantitative models has been scarcely applied in the continent. So, in this work our objective was to digital map the soil classes distribution of Maritime Antarctica and Northern Antarctic Peninsula using the Machine Learning model Random Forest. We combined 820 soil profiles classified according to Soil Taxonomy and WRB/FAO systems with environmental covariates from different sources. Here, we used soil physical and chemical attributes modelled in the chapters before as predictor variables, along with the other most common predictors. The best performance was obtained for the classification at the first level of the Soil Taxonomy system (Kappa was 0.28, Accuracy 0.62 and F1-Score 0.51), whereas the worst performance was obtained for the WRB system. The best selected covariates involved terrain and soil attributes reflecting the strong control of the soil-landscape relationships on the soil classes distribution. Gelisols (in particular, Turbels) were more common in high parts of the landscape due to the presence of the permafrost, which is well expressed by the Elevation as the most important covariate. In turn, Orthels were more common in semi-arid conditions represented by the high Sodium contents. Entisols (in particular, Orthents) dominate the low coastal areas with high H+Al contents, whereas Inceptisols were characterized by topographic control coupled with higher weathering rates evidenced by soil attributes such as the H+Al and remaining P. Although outlining well the general character of the soil distribution in the regions mapped, our results were limited by the unbalanced distribution of the soil classes, identifying a lower pedodiversity than probably exists in Antarctica.

Keywords: Antarctic soils. Classification. Cryosols. Digital Soil Mapping. Random Forest. Soil maps.

1. INTRODUCTION

Mapping soil classes has been performed in Antarctica since the early soil surveys (Ugolini and Bockheim, 2008). Since then, conventional soil maps have proliferated, from the East Antarctica polar deserts (Bockheim and McLeod, 2008; McLeod et al., 2008) to the wettest coastal areas of the Antarctic Peninsula (Moura et al., 2012; Rodrigues et al., 2019; Schaefer et al., 2015). Currently, two main taxonomic systems are used for mapping Antarctic soils: WRB/FAO (IUSS/WRB, 2014) and Soil Taxonomy (Soil Survey Staff, 2014), although taxonomic limitations have been pointed out over the years (Simas et al., 2008). Regardless of this, such systems have been sufficient so far to characterize the main types of soils in Antarctica, important by playing fundamental ecological roles and being proxies to understand the paleoclimatic evolution of the continent (Bockheim, 1997; Haus et al., 2016; Michel et al., 2014; Siqueira et al., 2021).

Soil maps are graphic representations of soil's spatial distribution (Ma et al., 2019). Historically, soil mapping has been based on mental models in the process known as conventional soil mapping (Minasny and McBratney, 2016). As main reference, usually it is used the qualitative version of the Jenny's state-factor model, also known as clorpt, which is a conceptual model that formalizes the idea of soil as a product of the integrated action of the soil forming factors (Jenny, 1941). For the process of mapping itself, this model principle is applied from the use of remote sensing imagery and digital elevation models as supporting data to identify landscape features, which are then integrated into field observations to establish the boundaries of soil units to be drawn on soil maps (Meier et al., 2018; Minasny and McBratney, 2016).

However, as conventional mapping is highly dependent on the knowledge and expertise of soil surveyors in all mapping stages, its qualitative nature may yield many drawbacks, such as high uncertainties, lack of reproducibility, difficulties in representing soil complexity and lack of means to assess the maps quality (Ma et al., 2019; Minasny et al., 2014). Furthermore, conventional surveys are excessively time-consuming and expensive for large areas. To overcome these limitations, the Digital Soil Mapping (DSM) field has evolved in the last decades along with the computational power and statistical algorithms to describe the spatial distribution patterns of soils on a mathematical basis, aiming to feed appropriately the growing global demand for accurate soil spatial information (Lagacherie and McBratney, 2006; McBratney et al., 2000).

DSM uses the scorpan factor model, which is an empirical and spatial model developed as a conceptual system for soil mapping under the DSM scope and is a revised formulation of the clorpt model (Ma et al., 2019; Thompson et al., 2012). Its general formula is written as S (c, a) = f (s, c, o, r, p, a, n), where S (c, a) are soil classes or soil attributes as a function of soil (other attributes), climate, organisms, relief, parent material, age, and spatial position (McBratney et al., 2003). Such as the conventional surveys, the scorpan has as scientific rationale the relationships between soil and soil forming factors but focusing in describing these relationships through explicit statistical or rule-based models that supersede the mental models in many ways (Ma et al., 2019). Another difference is that whereas clorpt model was formulated to explain soil formation, scorpan was designed as a pragmatic model for predicting soil properties and classes (Minasny and McBratney, 2016; Thompson et al., 2012). Nevertheless, pedological knowledge can still be extracted from the predicted maps, making use of quantitative relations to identify the main factors underlying the soil distribution (Ma et al., 2019).

Machine learning (ML), developed from pattern recognition and computational learning in artificial intelligence, is the main statistical technique applied along with the scorpan model (Gomes et al., 2019; Kumaraperumal et al., 2022; Meier et al., 2018; Sena et al., 2020, Siqueira et al., 2023). The ML algorithms were designed to simulate the cognitive capability of the human brain and in soil science they are applied to mimic the reasoning of pedologists to make soil predictions (Minasny and McBratney, 2016).

One of the most common applications of ML algorithms in soil science are the supervised classification tasks of soil types (Brungard et al., 2015; Camera et al., 2017; Carvalho Junior et al., 2014; Taalab et al., 2015; Taghizadeh-Mehrjardi et al., 2020). Mapping the spatial distribution of soil taxonomic classes is an important key for guiding decision making in soil resource assessments, since they group the basic soil information in an environmental viewpoint (Lagacherie, 2005). One of the biggest contributions of the automatic classification of soil types is the more precise discretization of the soil units. Whereas in conventional soil mapping, the soil classes boundaries are based on natural boundaries identified by photointerpretation (Thompson et al., 2012), ML algorithms draw decision boundaries over the landscape based on quantitative rules.

Therefore, coupling legacy soil data characterized by these taxonomy systems and the new DSM techniques can be a good strategy to produce more accurate soil maps at higher resolutions and extents for the remote ice-free areas of Antarctica. Our main objective in this work was to produce the first digital maps of soil classes at different taxonomic levels for the Maritime Antarctic and Northern Antarctic Peninsula regions, besides evaluating the performance and environmental relations found by the ML algorithm Random Forest.

2. MATERIAL AND METHODS

2.1.Study area

Maritime Antarctica (MA) encompasses the South Shetland Islands (SSI) (Fig. 1). The SSI are composed of marine sediments, metamorphic rocks, intrusive bodies, tuffs and basaltic to andesitic lavas, of age varying from Jurassic to Late Quaternary (Smellie et al., 2021b, 2021a). The geomorphology is marked by a set of erosive and depositional landforms typical of paraglacial and periglacial environments, including landforms such as moraines, raised beaches, outwash plains, high platforms, blockfields and talus and scree slopes (Francelino et al., 2011; López-Martínez et al., 2012). The SSI have a maritime climate with mean annual precipitation (MAP) ranging from 350 to 800 mm and mean annual air temperature (MAAT) around -2°C (Blumel and Eitel, 1989). Mosses-lichens communities form the main vegetation, whereas flowering plants are also common in some areas (Ferrari et al., 2020).

The Northern Antarctic Peninsula (NAP) is represented by the James Ross Islands (JRI) and the Antarctic Peninsula coastal area Hope Bay (Fig. 1). The JRI are mainly composed of marine sandstones, mudstones, and conglomerates predominantly of Cretaceous age overlain by Neogene basaltic lavas and breccia (Pirrie, 1994). The JRI geomorphology is like the SSI, with topography also showing a strong structural control due to sedimentary and volcanic rocks (Jennings et al., 2021). The climate is drier, with MAAT varying from -5°C to -7°C and MAP from 300 to 500 mm (van Lipzig et al., 2004). Due to the cold semi-arid climate, the vegetation is scarce.

2.2. Soil profiles and environmental covariates

As training data, we used information about soil classes of spatially referenced profiles described by the Brazilian Terrantar Group in intensive soil surveys in Antarctica. Additional soil profiles were obtained from the literature to complement the dataset, totalizing 820 complete soil profiles characterized and classified according to the WRB/FAO and Soil Taxonomy systems. In the Soil Taxonomy, the first three categorical levels were used for prediction (Table 1). For the modelling, the soil profiles were randomly split into training and test sets in a 4:1 ratio (holdout), with the function *createDataPartition* from the R package caret. In this way, 80% of the data was included in the training set and 20% in the test set.

As environmental covariates, we used a set of primary and secondary terrain attributes and terrain-based classifications derived from the 8-meter spatial resolution Reference Elevation Model of Antarctica (REMA) (Howat et al., 2019). The terrain attributes were generated using the R package RSAGA (Brenning, 2008). We also used multispectral bands from Sentinel-2 mission satellite imagery. From Sentinel data, we also used the vegetation indexes Soil Adjusted Vegetation Index (SAVI) and Normalized Difference Vegetation Index (NDVI) indexes, and the soil texture index Grain Size Index (GSI). As the main covariates representing the soil factor, we used the soil attributes (0-100 cm) predicted in Chapters 1 and 2, as follows: clay, silt, sand, bases sum, H+Al, pH, TOC, Na, P and P-rem. To account for parent material effects, we used simplified geology extracted from local geological maps. The final predicted maps were generated with spatial resolution of 32 meters.

Table 1. Soil classes of the Antarctic soils according to WRB/FAO and Soil Taxonomy (first, second and third levels).

WRB		**ST 1st level**		**ST 2nd level**		**ST 3rd level**			
Andosols	6	*Andisols*	6	*Aquents*	22	*Anhyorthels*	6	*Haplohemists/Humigelods*	3
Arenosols	35	*Entisols*	304	*Aquepts*	10	*Aquiturbels*	24	*Haplorthels*	37
Cambisols	67	*Gelisols*	427	*Fibrists*	**7**	*Aquorthels*	11	*Haploturbels*	206
Cryosols	426	*Histosols/Spodosols*	7	*Fluvents*	21	*Cryofibrists*	4	*Histoturbels*	4
Fluvisols/Gleysols	27	*Inceptisols*	75	*Gelands*	6	*Cryopsamments*	43	*Humigelepts*	13
Histosols/Podzols	8			*Gelepts*	64	*Dystrogelepts*	11	*Petraquepts*	5
Leptosols	139			*Histels*	34	*Eutrogelepts*	4	*Psammaquents*	3
Regosols	111			*Orthels*	80	*Fibristels*	29	*Psammorthels*	21
				Orthents	219	*Gelaquents*	20	*Psammoturbels*	47
				Psamments	43	*Gelaquepts*	5	*Sapristels*	6
				Turbels	313	*Gelifluvents*	22	*Umbriturbels*	32
						Gelorthents	217	*Umbrorthels*	5
						Haplogelepts	35	*Vitrigelands*	6

The covariates set were passed by the two-step selection procedure used in the first two chapters. So, the covariates with correlation coefficient > |0.95| were eliminated (function findcorrelation), aiming to minimize the collinearity effects and the remaining covariates were treated with the Recursive Feature Elimination (RFE) algorithm to selection of the most important predictors (function rfe). The RFE was run 50 times and the Kappa obtained by cross-validation was the index used for the RFE optimization.

The relative covariate's importance (RI) was estimated on caret package using the function varImp. For classification in Random Forest, the function uses the mean decrease of the Gini index, which measures how each covariate contributes to the homogeneity of the nodes and leaves in the trees and is directly proportional to the predictor's importance.

2.3.Training

For the modelling of soil classes, the Random Forest ML algorithm (Breiman, 2001), which consists of many Classification and Regression Trees (CART) whose individual outputs are grouped to obtain the result. This makes Random Forest have advantages over the single CARTs, such as higher predictive power, efficiency against noises and overfitting, and lower bias and variance (Hengl et al., 2018).We trained the Random Forest model with the function train of the package caret. We tuned the model using a 10-fold cross validation and ten possible values of the hyperparameter 'mtry' with the function tuneLength. To overcome the effect of sampling variability, we ran the models 50 times, creating repeatedly different training and test datasets.

2.4.Test and prediction

We tested the RF model performance by applying the independent test data through statistical indexes derived from the confusion matrix of the classification. Accuracy (Eq. 1) provides an agreement rate of classification, by dividing the number of corrected classified samples by the number of all samples. The Kappa Index (Eq. 2) is a measure of the real agreement, minus the agreement by chance; in other words, it is a measure of how much the classification agrees with the reference data (Congalton, 1991). However, both Accuracy and Kappa are not efficient when dealing with imbalanced classes (classes with a very disproportionate number of samples). For solving this problem, two other indexes were used. Balanced Accuracy (Eq. 3) is the arithmetic mean of the Sensitivity (Eq. 4) and Specificity (Eq. 5) indexes, which can also be used as individual metrics and represent the omission and commission errors, respectively. The F1-score (Eq. 6), in turn, also combine two other indices, being the harmonic mean of the Precision (Eq. 7) and of the Sensibility (termed as Recall in the F1-score concepts).

$$\mathrm{Accuracy} = \sum_{i=1}^{m} \frac{n_{ii}}{n} \tag{1}$$

$$\mathrm{Kappa} = \frac{n \sum_{i=1}^{m} n_{ii} - \sum_{i=1}^{m} (n_{i.} n_{.i})}{n^2 - \sum_{i=1}^{m} (n_{i.} n_{.i})} \tag{2}$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} - \text{Specificity}}{2} \tag{3}$$

$$\text{Sensitivity or Recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FN}} \tag{4}$$

$$\text{Specificity} = \frac{\sum \text{TN}}{\sum \text{TN} + \text{FP}} \tag{5}$$

$$\text{F1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

$$\text{Precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \text{FP}} \tag{7}$$

where n is the total number of samples, m is the total number of classes, n_{ii} is the number of correctly predicted samples, $n_{i.}$ is the sum of row i, $n_{.i}$ is the sum of column i of the confusion matrix, TP is the true positives, FN is the false negatives, TN is the true negatives and FP is the false positives. The indexes were obtained for all 50 runs and compared by the mean values. The final maps for each chemical attribute at each increment depth was obtained through averaging the spatial prediction of the 50 runs. We also ran two uncertainty analysis for classification tasks: the variety map (number of different classes that each pixel of the map presented) and the mode frequency (number of times that the most frequent class in each pixel was chosen).

3. RESULTS AND DISCUSSION

3.1. Classification performance

The model ST1 (Soil Taxonomy 1st level) presented the best performance in the training, with the highest Kappa, Accuracy, Balanced Accuracy and F1-Score (Fig. 1). The average ST1 Kappa and Accuracy were 0.31 and 0.64, whereas values of the other metrics were always next the maximum (1). These overestimated values are directly related to the cross-validation method used in the training, which tends to overvalue the performance in relation to the holdout method with independent data used in the test.

In the test, the average ST1 Kappa was 0.28, Accuracy 0.62 and F1-Score 0.51. Overall, the ST1 model is followed by the models ST2 and ST3 (2^{nd} and 3^{hd}, respectively), indicating a gradual loss of performance as the number of soil classes modeled increases. This trend is not only observed for Balanced Accuracy, where the ST3 model presents the best performance, (average 0.58), indicating the suitability of this index to deal with unbalanced classifications, as is the case of ST3. However, the trend cited above is not observed for all set of models, since the WRB model (simpler than the ST2 and ST3) presented, in general, the lowest performance in the test, with Kappa of 0.17 and F1-Score of 0.34. Considering that the set of top covariates selected for each model do not show large differences and the soil locations used were the same, we could assume the WRB soil keys would be more unrelated to the Antarctic landscape patterns, hardening the extraction of environmental relationships, at least with the soil and topographic predictors used in this study.

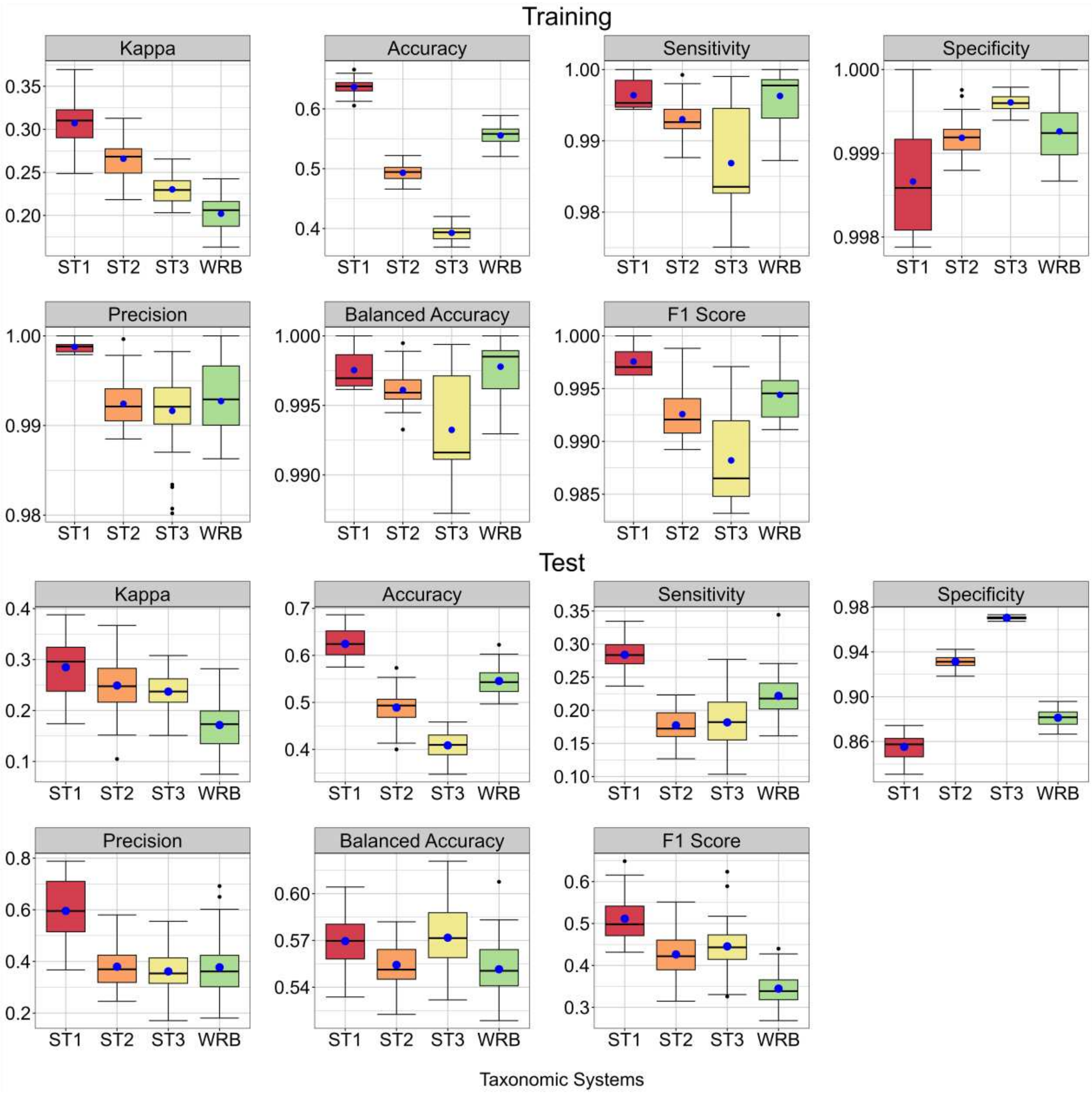


Figure 1. Performance of the training and test of the four classifications performed in this work. ST1: Soil Taxonomy 1st level; ST2: Soil Taxonomy 2nd level; ST3: Soil Taxonomy 3^{hd} level; WRB: WRB/FAO system.

In general, the Sensitivity presented much lower values than the Specificity for all models, indicating the errors of omission (associated with the misclassification of positive values as negative values) were more important for the loss of performance than the commission ones (associated with the misclassification of negative values as positive). The models also varied in terms of these two metrics, with ST1 having the best sensitivity and the worst specificity, which indicates a greater probability of presenting commission errors compared to the other models. The precision, in turn, is linked to the relation between the positive values classified correctly and those classified wrongly, and the ST1 also presented the highest values (Fig. 1).

Analyzing the confusion matrixes, we can see all predictions follow a similar tendency. In general, the greatest rates of agreement are associated with the soil classes with the highest number of samples, highlighting the Turbels (main Gelisols in Antarctica) and Orthents (main Entisols) in the ST2 (Fig. 2) and the Cryosols in the WRB maps (Fig. 3). This indicates the greater easer of the RF model to identify the most correct patterns when there is a greater number of soil classes' samples for interpretation of the soil-environment interplays. Although there was great agreement for the most abundant classes, we also found classes of low abundance showing agreement rates, such as the Histosols/Spodosols and Cambisols equalizing or even surpassing those of the more widespread Regosols and Leptosols classes in the WRB confusion matrix (Fig. 3).

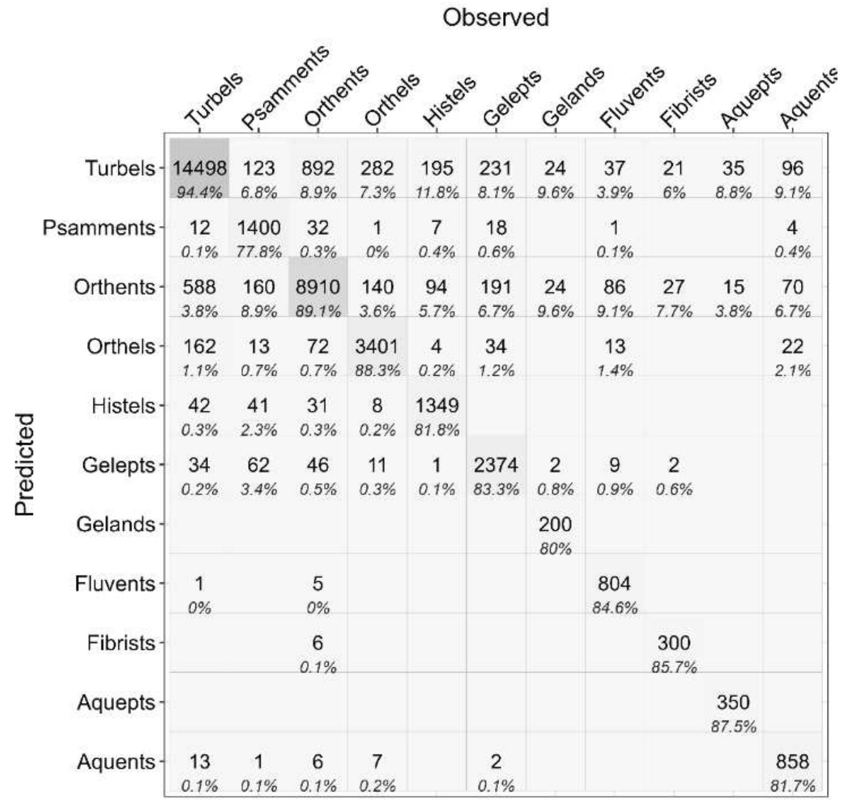


Figure 2. Confusion matrix of the ST2 model classification. The observed vs predicted data includes the prediction done in the training and test throughout the 50 runs.

This can be related to two facts: the spatial proximity of the few samples, facilitating the model's interpolative capacity, or the environmental particularities of these classes, which, despite being rare, have outstandingly discernible patterns in the landscape. In general, most of the classes showed the greatest confusions with the most abundant classes, regardless of pedological similarities, indicating the tendency already mentioned of the RF models in favoring the classes of the highest occurrence.

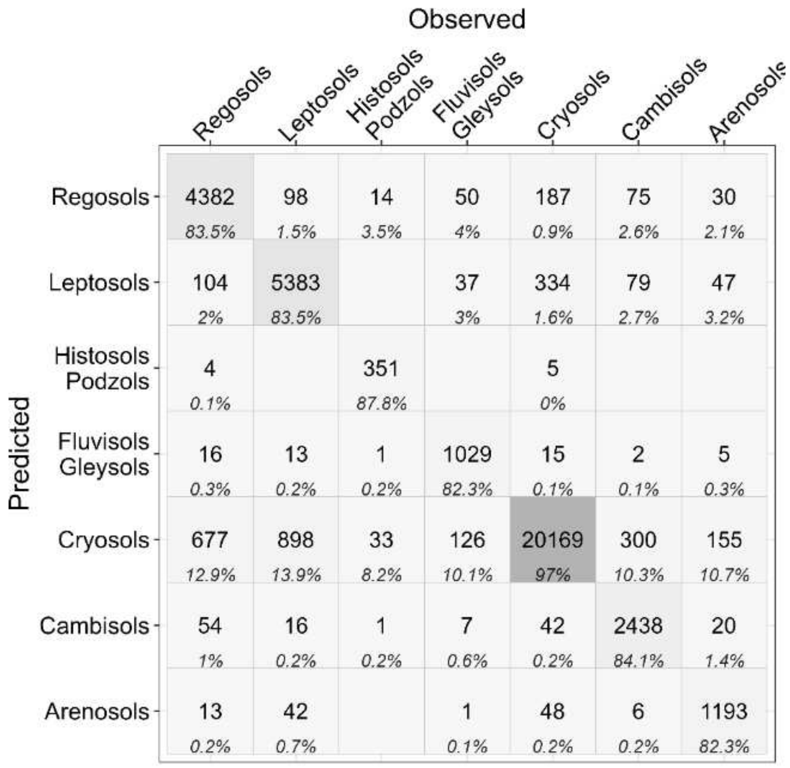


Figure 3. Confusion matrix of the WRB model classification. The observed vs predicted data includes the prediction done in the training and test throughout the 50 runs.

3.2. Covariates importance

Unlike what was observed in the first two chapters, the number of covariates selected by the RFE increased according to the performance of the classifications. The ST1 model presented the largest set of covariates, with an average of 42 (throughout 50 runs). It was followed by ST2 (average 26), ST3 (average 18) and WRB (average 17) models. However, the ST3 presented the lowest range, with maximum number of covariates of just 30, whereas the other models presented the total number of covariates as maximum (120). Overall, the topographic and soil attributes were the main covariates for all classifications, sharing the top ten most important covariates scores (Figure 4). The elevation was the main covariate for ST1 and WRB models, whereas the Sodium was the most important for ST2 and ST3.

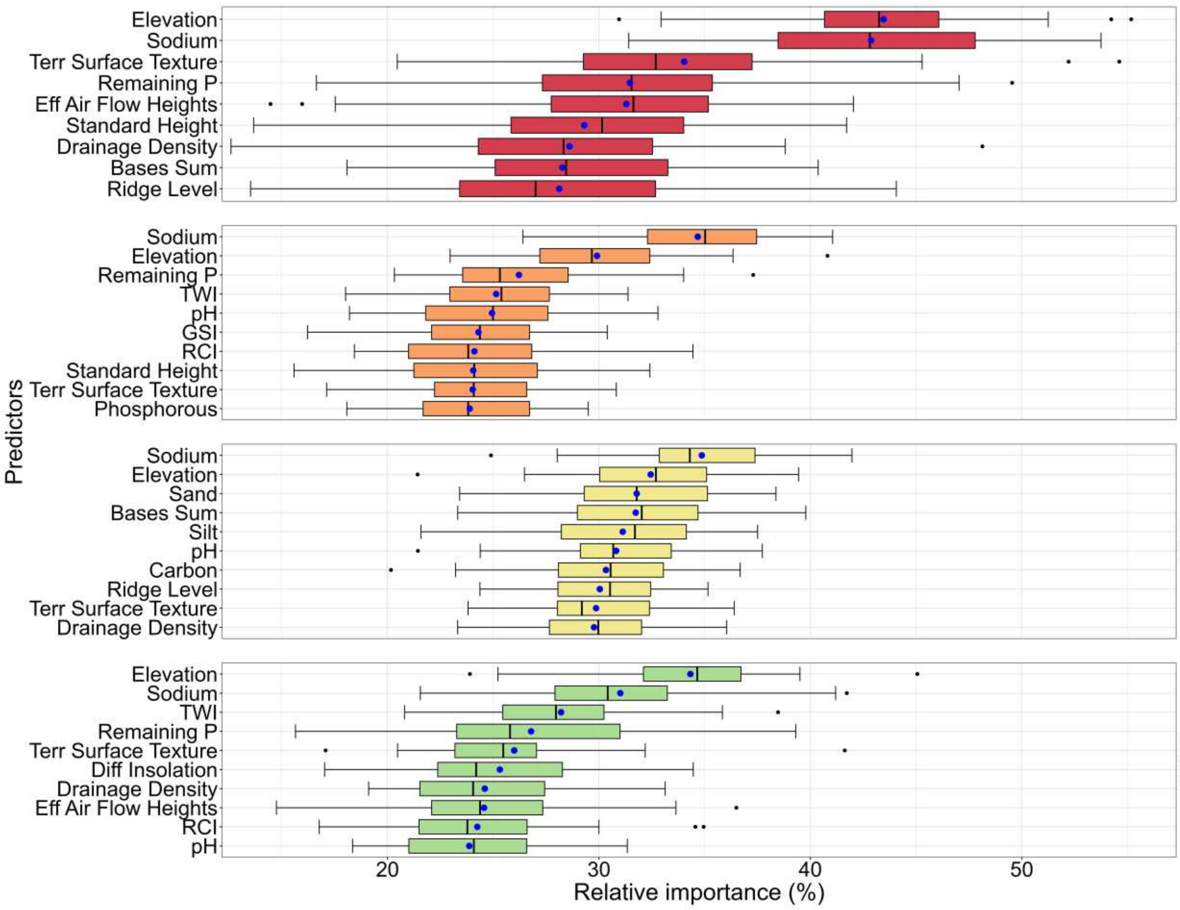


Figure 4. Relative importance of the best 10 covariates for RF model. Only the covariates selected more than 10 times were considered relevant for the models and represented. Boxplots record the variability of importance along the 50 runs.

The main topographic attributes selected were mainly related to altitude, surface dissection, inclination, and surface texture. Besides elevation, others important altitude covariates were the Effective Air Flow Heights, Ridge Level and Standardized Height. The main predictors associated with the inclination and surface texture were Terrain Surface Texture, Roughness Concentration Index, and the Diffuse Insolation (areas with higher slope tend to have lower incoming solar radiation due to the incidence angle). Drainage Density and Terrain Wetness Index were the main covariates linked to the surface dissection, and directly express the fluvial erosive action as an important control of the soil classes distribution in Antarctica.

Both physical and chemical attributes were important. Besides Sodium, Remaining P, pH, and Bases Sum were others important chemical attributes, appearing in different degrees among the most important covariates for all models. On the other hand, the physical attributes were selected among the top ten predictors only for the ST3 model, highlighting the Sand and Silt variables. The terrain attributes related to altitude show that soil classes distribution in Antarctica follows a topographic gradient, whereas the soil attributes importance indicate that soil classes have a clear variability related to the weathering and pedogenic processes that affect the soil physico-chemical characteristics. In particular, the Sodium importance is related to the latitudinal and topographic gradients which mainly control the distribution of the Gelisols/Cryosols.

The bioclimatic variables used in the first two chapters were not included here for two reasons: first, we considered the predicted soil attributes reflect the climate trends in the regions studied, so the soil attributes incorporate in some way the bioclimatics information; second, preliminary tests (not discussed here) showed us the use of bioclimatics induced to the creation of spatial artefacts on the soil classes distribution, decreasing the consistency of the final maps, although not inducing significant changes on performance. We also consider the use of bioclimatics in the context of the Antarctic regions mapped tended to hinder the identification of the soil classes on a local scale, whereas the combined use of the soil and topographic attributes expressed better the local soil-landscape relationships and improved the delimitation of the soil units in our maps.

3.3. Predicted maps

The final predicted maps show the clear dominance of the Gelisols/Cryosols in all maps. In the ST1 map, the Gelisols class presented 89.5% of the total area, Entisols presented 10.5% and Inceptisols presented only 0.1%. Although they appeared in some maps out of 50, the classes Histosols/Spodosols and Andisols did not appear in the final modal map due to their low occurrence in the study area. In the ST2 map, Turbels dominated with 71.2% of the total area, followed by the Orthels (14.4%) and Orthents (13.8%), whereas the other classes did not present area largest than 0.5%. In turn, the Haploturbels dominated the ST3 map, with 78.1% of the total area. The Gelorthents were second most abundant class, with 17.1%, followed by the Haplorthels (3.1%), Psammorthels (0.8%). The other classes did not exceed 0.5%. The WRB map followed the same pattern, with Cryosols corresponding to 94.3% of the total, followed by the Leptosols with 2.9% and Regosols, with 2.1%. Again, the remaining classes had area lower than 0.5%.

This result reflects two contrasting, important facts. First, it expresses the most abundant classes of the Antarctic continent (e.g., Gelisols and Entisols according to the Soil Taxonomy System and Cryosols, Leptosols and Regosols in the WRB/FAO) (Campbell and Claridge, 1987; Bockheim et al., 2015). Nonetheless, the excessive domain in terms of total area of the Cryosols/Gelisols probably reflects a systematic problem of the RF model: the suppression of the rare classes. RF is a generalist model, which uses in classification the mode of training data at two levels: at the level of leaf or terminal nodes and at the level of decision trees, which disfavor the classes of lowest occurrence and make their distribution in the final maps very limited. This can be further potentialized by our methodology that applies multiple repetitions of the entire process and obtain the final maps from the mode of the individual predicted maps. All of that indicates the difficulties of the RF model for unbalanced classifications, although it maintains better performance than most of the other models.

Despite this, our maps were still able to extract important environmental patterns about the distribution of soil classes in the mapped ice-free areas. For further analysis, we will consider the ST2 results, which was the model that presented the greatest pedodiversity on the final map (Fig. 5). Although the dominance of the Turbels in both regions, MA clearly presented a higher soil variability than NAP, due to the higher occurrence of classes such as Orthents, Orthels, Gelepts, and in a lesser degree, Psamments. In MA, Turbels predominated in the highest parts of the landscape, such as in the high cryoplanated platforms and in the sectors closest to the glaciers (Michel et al., 2014; Rodrigues et al., 2019). Turbels are the most common Gelisols (permafrost-affected soils) in MA, due to the high-water availability and soil temperature fluctuations around 0°C that activate freeze-thaw cycles and soil cryoturbation (Simas et al., 2008).

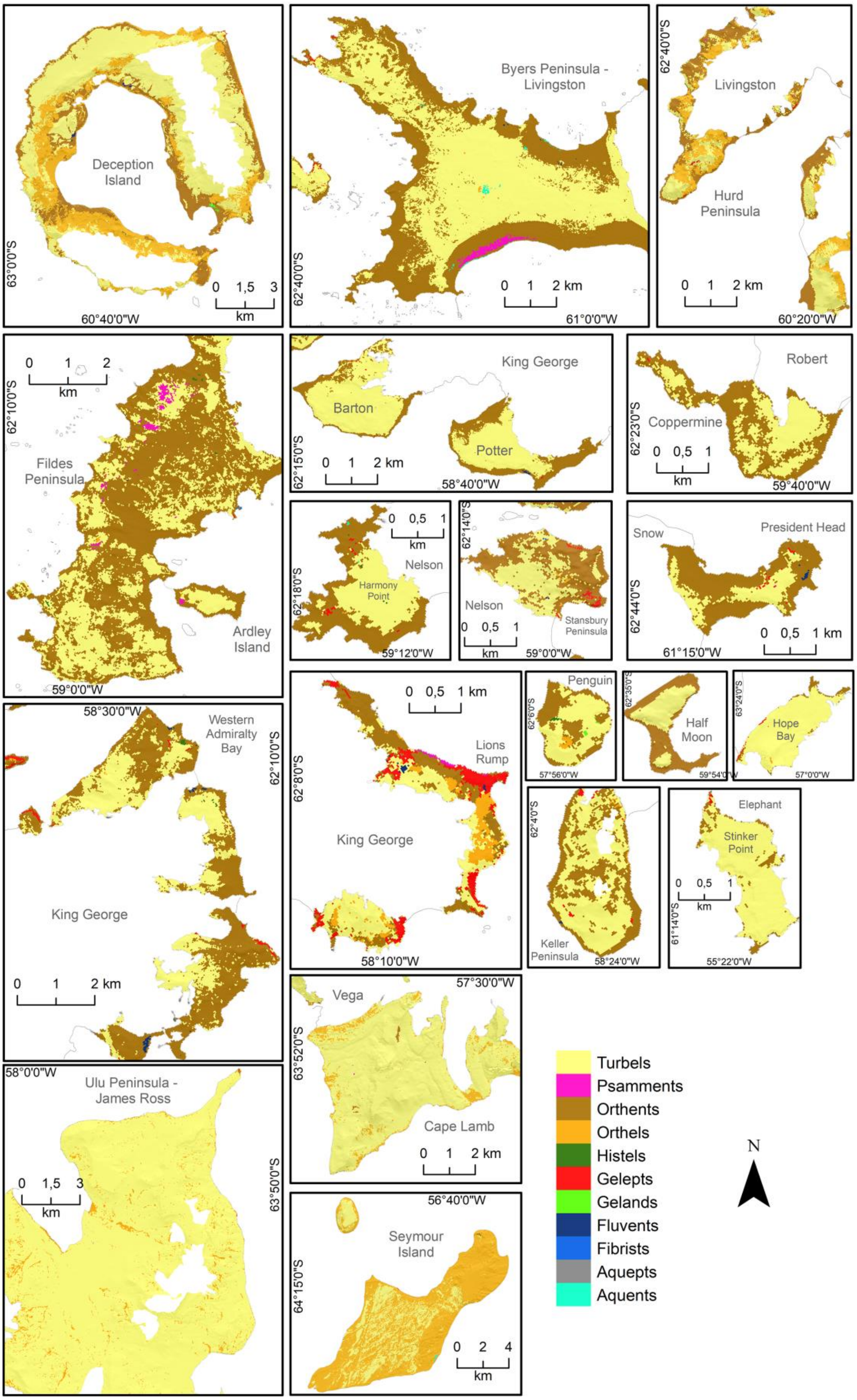


Figure 5. Soil Taxonomy 2nd level predicted soil classes map, highlighting the main ice-free areas of Maritime Antarctica and Northern Antarctic Peninsula.

Orthels were also mapped expressively in some areas such as Deception Island and Hurd Peninsula – Livingston Island, where the extremely sandy soils restrict the ice-induced soil movements, even with presence of permafrost next to the surface. The dominance of the Gelisols in the highlands is well expressed by the importance of the Elevation covariate in all taxonomic levels (Fig. 6) and fits appropriately to the conceptual model existing in MA about the distribution of permafrost (López-Martínez et al., 2012). According to this model, the Gelisols dominate mainly at altitudes above 80 m a.s.l., where permafrost is continuous (Balks et al., 2013).

On the other hand, Gelisols are ubiquitous in NAP due to the colder climate, with the widespread permafrost being identified less than 100 cm deep even in the lowlands next to the coast (Siqueira et al., 2021). In NAP the permafrost can be ice-cemented or dry-frozen (with temperatures below 0°C but without ice), depending on the soil aridity (Delpupo et al., 2014). Islands such as James Ross and Vega have soils with sufficient moisture for the development of ice-cemented permafrost and activation of cryoturbation processes in the active layer, which reflects the dominance of the Turbels in these islands (Daher et al., 2019; Siqueira et al., 2021). However, the same is not seen in Seymour Island, where on our final map the Orthels with dry-frozen permafrost spatially surpassed the Turbels due to the extreme soil aridic conditions. Although Sodium was a soil attribute important for all the Gelisols classes, it was more relevant for the Orthels (Fig. 6), indicating the close relationship between this class and the semi-arid conditions of Seymour Island favoring the salinization process (Gjorup et al., 2020).

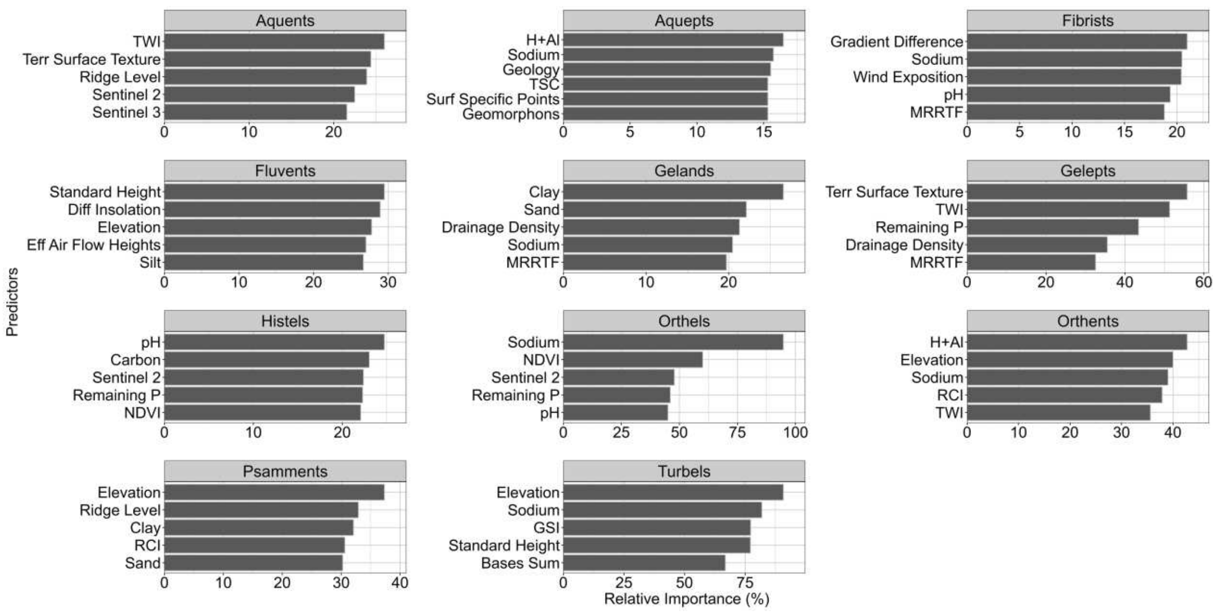


Figure 6. Average relative importance of the environmental covariates for each soil class of the ST2 model.

The high importance of the Elevation for the Entisols classes, highlighting the Orthents (Fig. 6), follows an inverse relation of those found for the Gelisols. Whereas the Gelisols are more common in the highlands, the Entisols dominate in the lowlands where the permafrost is discontinuous, sporadic, or absent (Balks et al., 2013). This pattern can be easily seen in areas such as Byers Peninsula – Livingston Island, Harmony Point – Nelson Island, Keller Peninsula – King George Island, Half Moon Island, among others (Fig. 5). The higher importance of the H+Al contents (Fig. 6) show the relevance for the Orthents prediction of the soil acidification associated to the biota activity in the coastal lands (González-Guzmán et al., 2017). Although the most common presence in the lowlands, Orthents are also found in highlands where the lithic character impedes the permafrost identification. This is the case for areas such as Fildes Peninsula, where Orthents are commonly seen in the central part of the area (which corresponds to Leptosols in the WRB map) (Fig. 7).

Although clearly suppressed in some areas where its occurrence is supposedly more common (such as in Barton Peninsula – King George Island), spots of Inceptisols (Cambisols in the WRB) can be seen in some areas, highlighting the Gelepts found around Lions Rump – King George (Fig. 5). The Inceptisols represent the most developed soils in the Antarctic regions mapped and their formation are generally related to ornithogenesis (as the case of the Gelepts found in Lions Rump) and to sulfurization (Almeida et al., 2021, Lopes et al., 2019). Besides the topographic control expressed by the terrain attributes, the importance of soil attributes directly related to the weathering degree, such as H+Al (Aquepts) and Remaining P (Gelepts), indicate the significance of the soil development to the Inceptisols characterization. The importance of Geology for the Aquepts prediction indicates the relevance of the parent material to sulfurization (Francelino et al., 2011, Lopes et al., 2019). Other relationships found in our work show the importance of the soil attributes for the soil classes prediction, such as: the pH and Carbon attributes selected for the Histels (organic soils), and the Clay and Sand selected for the Psamments (arenic Entisols) and Gelands (Andisols notably sandy) prediction.

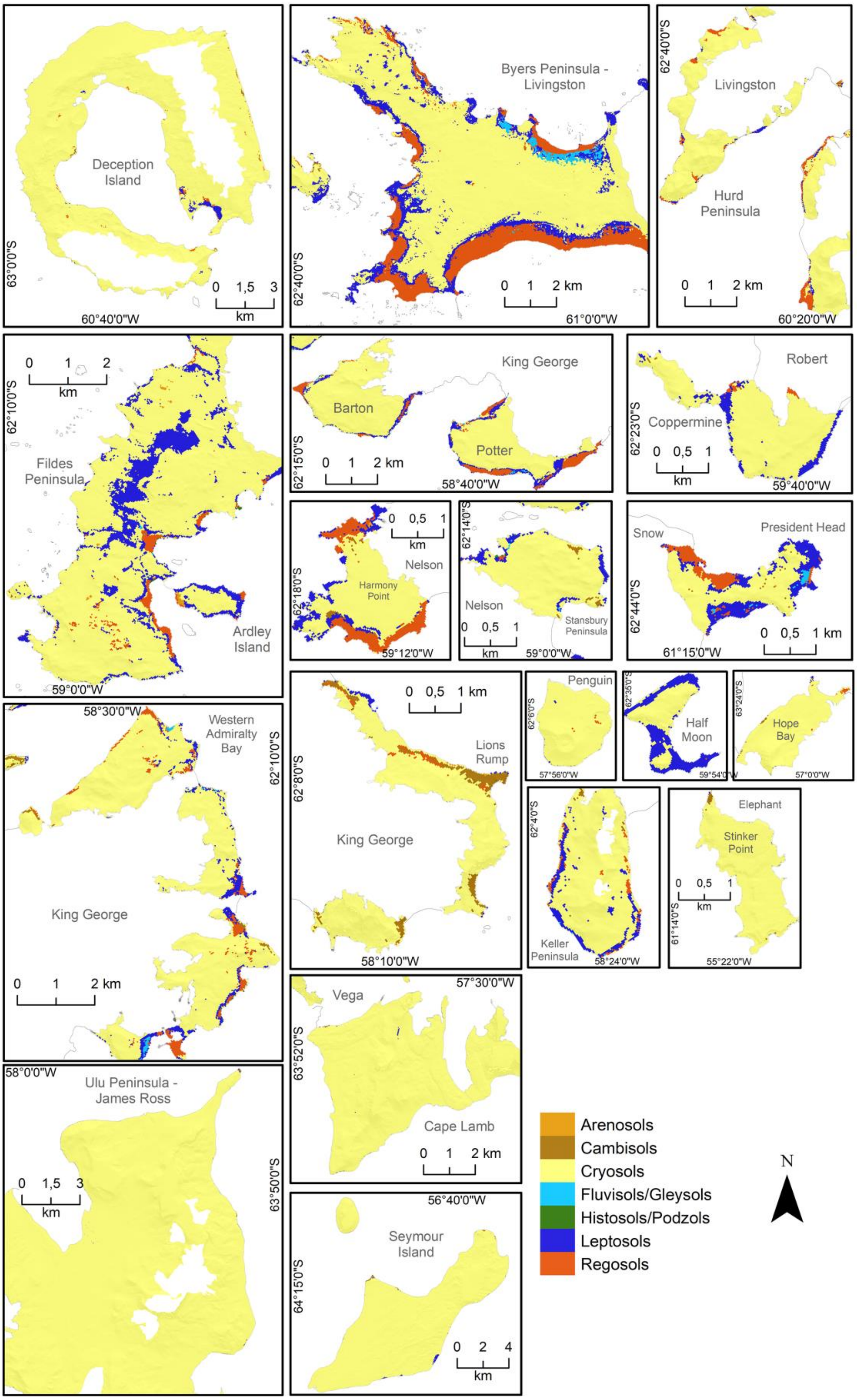


Figure 7. WRB predicted soil classes map, highlighting the main ice-free areas of Maritime Antarctica and Northern Antarctic Peninsula.

3.4. Uncertainty maps

Considering the variety indicator, it is possible to see the ST1 and ST3 maps presenting a higher number of pixels covered by class 1 (pixels that during the 50 runs showed only the mode value). On the other hand, the ST2 and WRB models presented most pixels covered by class 2 (during the 50 runs, these pixels shared two classes, the mode and another). Now, considering the mode frequency (the frequency of times the mode was chosen), the ST2 presented values significantly lower than the other maps, which, along with the variety results (Table 2), shows the ST2 model presented the highest uncertainty in this work. This is an unexpected result, considering that one could expect the greatest uncertainty associated with the ST3 model, which has a higher number of soil classes. However, the suppression of rare classes already discussed above was more intense in the ST3 map than in ST2, showing a greater spatial heterogeneity in the last, which also reflects the highest uncertainty. The uncertainty of the final maps was higher in the MA, precisely because of the greater pedodiversity of this region (Fig. 8).

Table 2. Variety classes of the predicted soil classes maps.

ST 1st level		**ST 2nd level**		**ST 3rd level**		**WRB**	
Variety Classes	Pixels Count	Variety Classes	Pixels Count	Variety Classes	Pixels Count	Variety Classes	Pixels Count
1	**751243**	*1*	427155	*1*	**577970**	*1*	457581
2	455006	*2*	**597148**	*2*	378622	*2*	**558440**
3	29554	*3*	167946	*3*	206581	*3*	147953
4	4	*4*	39324	*4*	54207	*4*	57748
		5	3698	*5*	14206	*5*	12979
		6	497	*6*	3284	*6*	1104
		7	39	*7*	815	*7*	2
				8	113		
				9	8		
				10	1		

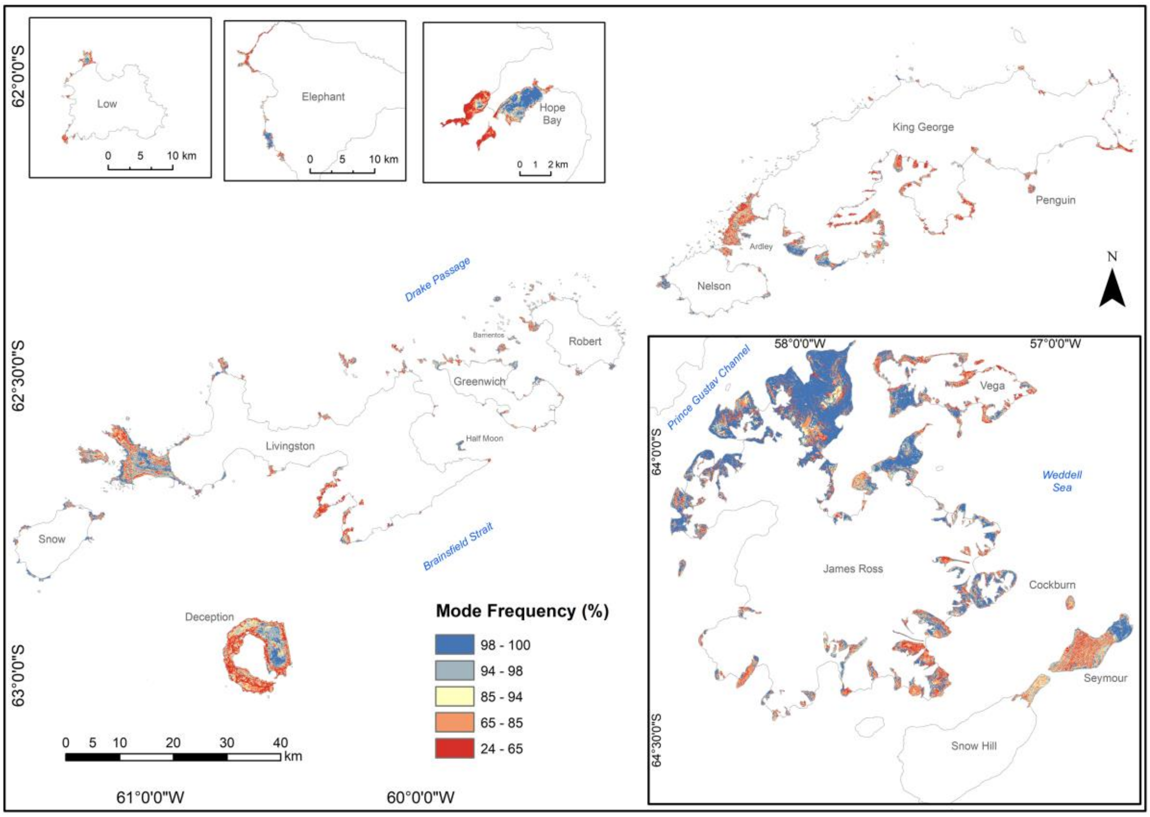


Figure 8. Mode frequency of the ST2 predicted soil class map.

4. CONCLUSIONS

This study was the first attempt to produce a digital map of soil classes for the Antarctic continent, involving two important regions, the Maritime Antarctica, and the Northern Antarctic Peninsula. The use of soil attributes as environmental covariates proved to be an approach with great potential for mapping soil classes, since the variability of soil physico-chemical attributes in Antarctica are directly related to soil formation processes, which improves the interpretation capability of the ML models. In general, our results expressed the most abundant soil classes of the Antarctic regions mapped, outlining environmental relationships of the soil classes with the climate, the topography and the pedogenesis. However, our results were limited by the unbalanced distribution of the soil classes, which, despite representing in part the natural variability of the Antarctic soils, with a predominance of Gelisols and Entisols, contributed to the suppression of rare classes, reducing the pedodiversity of the Antarctic environments. Possibilities for overcoming this problem are the use of object-oriented classification as an alternative to the pixel-based approach used here or even the application of ensemble models that join the performance of the RF with other ML models that deal better with the mapping of rare classes.

REFERENCES

Almeida, I.C.C., Schaefer, C.E.G.R., Fernandes, R.B.A., Oliveira, F.S., Pereira, T.T.C., 2021. Clay mineralogy and micropedology of phosphate-rich soils from Lions Rump, Maritime Antarctica. J. South Am. Earth Sci. 105. https://doi.org/10.1016/j.jsames.2020.102967

Balks, M.R., López-Martínez, J., Goryachkin, S.V., Mergelov, N.S., Schaefer, C.E.G.R., Simas, F.N.B., Almond, P.C., Claridge, G.G.C., Mcleod, M., Scarrow, J., 2013. Windows on Antarctic soil-landscape relationships: Comparison across selected regions of Antarctica. Geol. Soc. Spec. Publ. 381, 397–410. https://doi.org/10.1144/SP381.9

Bockheim, J.G., 1997. Properties and Classification of Cold Desert Soils from Antarctica. Soil Science Society of America Journal 61, 224–231. https://doi.org/10.2136/sssaj1997.03615995006100010031x

Bockheim, J.G., McLeod, M., 2008. Soil distribution in the McMurdo Dry Valleys, Antarctica. Geoderma 144, 43–49. https://doi.org/10.1016/j.geoderma.2007.10.015

Bockheim, J.G., 2015. The Soils of Antarctica, 1st ed. Springer International Publishing Switzerland, New York. https://doi.org/10.1007/978-3-319-05497-1_13

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019

Camera, C., Zomeni, Z., Noller, J.S., Zissimos, A.M., Christoforou, I.C., Bruggeman, A., 2017. A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. Geoderma 285, 35–49. https://doi.org/10.1016/j.geoderma.2016.09.019

Campbell, I.B., Claridge, G.G.C., 1987. Antarctica: Soils, weathering processes and environment. Elsevier, Amsterdam, pp. 368p.

Carvalho Junior, W., Lagacherie, P., da Silva Chagas, C., Calderano Filho, B., Bhering, S.B., 2014. A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. Geoderma 232–234, 479–486. https://doi.org/10.1016/j.geoderma.2014.06.007

Daher, M., Schaefer, C.E.G.R., Fernandes Filho, E.I., Francelino, M.R., Senra, E.O., 2019a. Semiarid soils from a topolithosequence at James Ross Island, Weddell Sea region, Antarctica: Chemistry, mineralogy, genesis and classification. Geomorphology 327, 351–364. https://doi.org/10.1016/j.geomorph.2018.11.003

Delpupo, C.S., Schaefer, C.E.G.R., Simas, F.N.B., Spinola, D.N., Daher, M., 2014. Soil formation in Seymour Island, Weddell Sea , Antarctica. Geomorphology 255, 87–99. https://doi.org/10.1016/j.geomorph.2014.03.047

Francelino, M.R., Schaefer, C.E.G.R., Simas, F.N.B., Filho, E.I.F., de Souza, J.J.L.L., da Costa, L.M., 2011. Geomorphology and soils distribution under paraglacial conditions in an ice-free area of Admiralty Bay, King George Island, Antarctica. Catena 85, 194–204. https://doi.org/10.1016/j.catena.2010.12.007

Gjorup, D.F., Ernesto, C., Schaefer, G.R., Nogueira, F., Simas, B., Rocha, M., Ferreira, R., Michel, M., Bockheim, J.G., 2020. Sulfurization, Acid-Sulfate soils and active layer monitoring at the semiarid Seymour Island, Antarctica. Geoderma Regional 22, e00305. https://doi.org/10.1016/j.geodrs.2020.e00305

Gomes, L.C., Faria, R.M., de Souza, E., Veloso, G.V., Schaefer, C.E.G.R., Filho, E.I.F., 2019. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma 340, 337–350. https://doi.org/10.1016/j.geoderma.2019.01.007

González-Guzmán, A., Oliva, M., Souza-Júnior, V.S., Pérez-Alberti, A., Ruiz-Fernández, J., Otero, X.L., 2017. Biota and geomorphic processes as key environmental factors controlling soil formation at Elephant Point, Maritime Antarctica. Geoderma 300, 32–43. https://doi.org/10.1016/j.geoderma.2017.04.001

Haus, N.W., Wilhelm, K.R., Bockheim, J.G., Fournelle, J., Miller, M., 2016. A case for chemical weathering in soils of Hurd Peninsula, Livingston Island, South Shetland Islands, Antarctica. Geoderma 263, 185–194. https://doi.org/10.1016/j.geoderma.2015.09.019

IUSS/WRB, 2014. World Reference Base for Soil Resources. International soil classification system for naming soils and creating legends for soil maps., World Soil Resources Reports No. 103. World Soil Resources Reports No. 106. FAO, Rome.

Kumaraperumal, R., Pazhanivelan, S., Geethalakshmi, V., Raj, M.N., Muthumanickam, D., Kaliaperumal, R., Shankar, V., Nair, A.M., Yadav, M.K., Vendan, T., Kshatriya, T., 2022. Comparison of Machine Learning-Based Prediction of Qualitative and Quantitative Digital Soil-Mapping Approaches. Land (Basel) 11. https://doi.org/10.3390/land11122279

Lagacherie, P., 2005. An algorithm for fuzzy pattern matching to allocate soil individuals to pre-existing soil classes. Geoderma 128, 274–288. https://doi.org/10.1016/j.geoderma.2005.04.009

Lagacherie, P., McBratney, A.B., 2006. Chapter 1 Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping. Developments in Soil Science 31, 3–22. https://doi.org/10.1016/S0166-2481(06)31001-X

Lopes, D. do V., Schaefer, C.E.G.R., Souza, J.J.L.L. de, Oliveira, F.S., Simas, F.N.B., Daher, M., Gjorup, D.F., 2019. Concretionary horizons, unusual pedogenetic processes and features of sulfate affected soils from Antarctica. Geoderma 347, 13–24. https://doi.org/10.1016/j.geoderma.2019.03.024.

López-Martínez, J., Serrano, E., Schmid, T., Mink, S., Linés, C., 2012. Periglacial processes and landforms in the South Shetland Islands (northern Antarctic Peninsula region). Geomorphology 155–156, 62–79. https://doi.org/10.1016/j.geomorph.2011.12.018

Ma, Y., Minasny, B., Malone, B.P., Mcbratney, A.B., 2019. Pedology and digital soil mapping (DSM). Eur J Soil Sci 70, 216–235. https://doi.org/10.1111/ejss.12790

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping, Geoderma. https://doi.org/10.1016/S0016-7061(03)00223-4

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey, Geoderma. https://doi.org/10.1016/S0016-7061(00)00043-4

McLeod, M., Bockheim, J.G., Balks, M.R., 2008. Glacial geomorphology, soil development and permafrost features in central-upper Wright Valley, Antarctica. Geoderma 144, 93–103. https://doi.org/10.1016/j.geoderma.2007.10.017

Meier, M., de Souza, E., Francelino, M.R., Fernandes Filho, E.I., Schaefer, C.E.G.R., 2018. Digital soil mapping using machine learning algorithms in a tropical mountainous area. Rev Bras Cienc Solo 42, 1–22. https://doi.org/10.1590/18069657rbcs20170421

Michel, R.F.M., Schaefer, C.E.G.R., López-Martínez, J., Simas, F.N.B., Haus, N.W., Serrano, E., Bockheim, J.G., 2014. Soils and landforms from Fildes Peninsula and Ardley Island, Maritime Antarctica. Geomorphology 225, 76–86. https://doi.org/10.1016/j.geomorph.2014.03.041

Michel, R.F.M., Schaefer, C.E.G.R., Poelking, E.L., Simas, F.N.B., Fernandes Filho, E.I., Bockheim, J.G., 2012. Active layer temperature in two Cryosols from King George Island, Maritime Antarctica. Geomorphology 155–156, 12–19. https://doi.org/10.1016/j.geomorph.2011.12.013

Minasny, B., Malone, B., Stockmann, U., Odgers, N., McBratney, A.B.., 2014. Pedometrics. Reference Module in Earth Systems and Environmental Sciences 1–10. https://doi.org/10.1016/b978-0-12-409548-9.09163-6

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Moura, P.A., Francelino, M.R., Schaefer, C.E.G.R., Simas, F.N.B., Mendonça, B.A.F. de, 2012. Distribution and characterization of soils and landform relationships in Byers Peninsula, Livingston Island, Maritime Antarctica. Geomorphology 155–156, 45–54. https://doi.org/10.1016/j.geomorph.2011.12.011

Rodrigues, W.F., Oliveira, F.S., Schaefer, C.E.G.R., Leite, M.G.P., Gauzzi, T., Bockheim, J.G., Putzke, J., 2019. Soil-landscape interplays at Harmony Point, Nelson Island, Maritime Antarctica: Chemistry, mineralogy and classification. Geomorphology 336, 77–94. https://doi.org/10.1016/j.geomorph.2019.03.030

Schaefer, C.E.G.R., Pereira, T.T.C., Ker, J.C.J.C., Almeida, I.C.C., Simas, F.N.B., Oliveira, F.S., Correa, G.R., Vieira, G.G., Costa Pereira, T.T., Ker, J.C.J.C., Carreiro Almeida, I.C., Bello Simas, F.N., Soares de Oliveira, F., Corrêa, G.R., Vieira, G.G., 2015. Soils and Landforms at Hope Bay, Antarctic Peninsula: Formation, Classification, Distribution, and Relationships. Soil Science Society of America Journal 79, 175–184. https://doi.org/10.2136/sssaj2014.06.0266

Sena, N.C., Veloso, G.V., Fernandes-Filho, E.I., Francelino, M.R., Schaefer, C.E.G.R., 2020. Analysis of terrain attributes in different spatial resolutions for digital soil mapping application in southeastern Brazil. Geoderma Regional 21, 1–12. https://doi.org/10.1016/j.geodrs.2020.e00268

Simas, F.N.B., Schaefer, C.E.G.R., Albuquerque-Filho, M.R., Francelino, M.R., Fernandes-Filho, E.I., Costa, L.M. da, 2008. Genesis, properties and classification of Cryosols from Admiralty Bay, maritime Antarctica. Geoderma 144, 116–122. https://doi.org/10.1016/j.geoderma.2007.10.019

Siqueira, R.G., Schaefer, C.E.G.R., Fernandes Filho, E.I., Corrêa, G.R., Francelino, M.R., Souza, J.J.L.L. de, Rocha, P. de A., 2021. Weathering and pedogenesis of sediments and basaltic rocks on Vega Island, Antarctic Peninsula. Geoderma 382, 114707. https://doi.org/10.1016/j.geoderma.2020.114707

Siqueira, R.G., Veloso, G. v., Fernandes-Filho, E.I., Francelino, M.R., Schaefer, C.E.G.R., Corrêa, G.R., 2022. Evaluation of machine learning algorithms to classify and map landforms in Antarctica. Earth Surf Process Landf 47, 367–382. https://doi.org/10.1002/esp.5253

Soil Survey Staff, 2014. Keys to Soil Taxonomy, 12th ed., Soil Conservation Service. Rome.

Taalab, K., Corstanje, R., Zawadzka, J., Mayr, T., Whelan, M.J., Hannam, J.A., Creamer, R., 2015. On the application of Bayesian Networks in Digital Soil Mapping. Geoderma 259–260, 134–148. https://doi.org/10.1016/j.geoderma.2015.05.014

Taghizadeh-Mehrjardi, R., Schmidt, K., Eftekhari, K., Behrens, T., Jamshidi, M., Davatgar, N., Toomanian, N., Scholten, T., 2020. Synthetic resampling strategies and machine learning for digital soil mapping in Iran. Eur J Soil Sci 71, 352–368. https://doi.org/10.1111/ejss.12893

Thompson, J.A., Roecker, S., Grunwald, S., Owens, P.R., 2012. Digital soil mapping: Interactions with and applications for hydropedology. Hydropedology 1, 665–709. https://doi.org/10.1016/B978-0-12-386941-8.00021-6

Ugolini, F.C., Bockheim, J.G., 2008. Antarctic soils and soil formation in a changing environment: A review. Geoderma 144, 1–8. https://doi.org/10.1016/j.geoderma.2007.10.005

CONCLUSÃO GERAL

Os resultados finais deste trabalho envolveram a elaboração de uma série de mapas de atributos físicos, químicos e classes de solos produzidos em três profundidades distintas (0-10 cm, 10-30 cm e 30-100 cm) para duas regiões de fundamental importância do continente antártico: a Antártica Marítima, representada pelas Ilhas Shetland do Sul, e a Península Antártica, representada pelas Ilhas James Ross, localizadas no Mar de Weddell, e a área livre de gelo conhecida como Hope Bay, localizada na região costeira da península. Este trabalho consiste na integração do grande número de dados e informações gerados pelo Grupo Terrantar ao longo de mais de 20 anos de pesquisa em solos na Antártica com as técnicas e conhecimentos recentes produzidos no âmbito do Mapeamento Digital de Solos.

Para a aplicação do mapeamento digital dos solos antárticos foram usadas técnicas de Aprendizado de Máquinas (Machine Learning) envolvendo passos como tratamento dos dados, seleção das covariáveis ambientais mais importantes, treinamento ou calibração dos modelos, testes estatísticos usando conjuntos de dados independente para validação dos modelos criados, e a predição espacial ou mapeamento, onde os atributos do solo são extrapolados para áreas sem informação em função da distribuição das covariáveis ambientais. A qualidade dos nossos modelos foi avaliada em termos de acurácia e incerteza. Uma das grandes vantagens da metodologia utilizada neste trabalho foi o uso de múltiplas repetições (e.g., 100 repetições) para se lidar com a variabilidade no processo de separação dos dados de treinamento e teste.

Buscou-se também avançar em um dos grandes desafios do Mapeamento Digital de Solos atualmente que é a conciliação dos resultados das predições com interpretações claras e objetivas com base nos fatores de formação do solo. As técnicas de seleção de variáveis, como o Recursive Feature Elimination (RFE), são uma implementação fundamental para a interpretação dos resultados, pois contribuem para a construção de modelos mais simples com manutenção do poder preditivo, a partir da seleção das variáveis mais importantes. Não obstante, o conhecimento pedológico foi fundamental para a explicação dos nossos modelos, uma vez que as covariáveis ambientais selecionadas representaram de forma satisfatória os fatores de formação responsáveis pela distribuição dos solos na Antártica.

Diferentes modelos de Aprendizado de Máquinas foram testados para a modelagem e predição da textura do solo (areia, silte e argila), e o Random Forest apresentou a melhor performance. Entre as frações do solo, a fração argila apresentou os maiores valores de performance em todas as profundidades, com destaque para a profundidade 0-10 cm (LCCC de 0,49 e R^2 de 0,31). As melhores covariáveis foram associadas ao clima, topografia e desenvolvimento dos solos. As variáveis climáticas mais importantes foram aquelas relacionadas à sazonalidade da temperatura e precipitação, mostrando que a distribuição da textura do solo na Antártica é função da variabilidade climática anual no continente Antártico. As covariáveis mais importantes associadas ao relevo variaram de acordo com a fração do solo mapeada. Para a fração argila, variáveis de dissecação do relevo, declividade e drenagem foram mais importantes, indicando que a distribuição desta fração na Antártica é fortemente condicionada pelos processos de catena. Variáveis de altitude foram mais importantes para a fração areia, indicando maior concentração desta fração nas partes mais baixas da paisagem e próximas à costa (terraços marinhos) e maiores conteúdos de silte foram observados nos setores mais altos próximos às geleiras. Este padrão de distribuição espacial também pode ser detectado pela seleção das covariáveis de desenvolvimento do solo: distância à costa e distância à geleira. Em geral, as variáveis climáticas foram mais importantes em superfície, enquanto as covariáveis do relevo adquiriram maior importância para a predição nas camadas mais profundas do solo.

Quanto à modelagem dos atributos químicos, o Random Forest foi aplicado para a predição de sete atributos do solo de extrema relevância para os ecossistemas antárticos: pH, Soma de bases, Carbono Orgânico Total, H+Al (acidez potencial), P disponível, Na disponível e P remanescente (P-rem). A Soma de Bases foi o atributo com melhor performance (LCCC de 0,58 e $R2^2$ de 0,41), seguido por H+Al, pH, Carbono, P-rem, Na e P. Considerando a incerteza, o pH e o P-rem apresentaram as maiores precisões (PIR médio de 0,07 e 0,13, respectivamente), enquanto o Na e P apresentaram as piores (PIR médio de 0,85 e 1,24). Ambos os indicadores de acurácia e incerteza mostraram a dificuldade do Random Forest no tratamento de solos com características raras da Antártica, expressadas na ocorrência de outliers naturais (valores extremos mínimos e máximos).

Em geral, os atributos químicos do solo apresentaram um gradiente latitudinal/climático bem expresso, com os valores Soma de Bases, pH e Na sendo maiores nas áreas da Península Antártica, enquanto os teores de Carbono, H+Al e P foram maiores na Antártica Marítima. O P-rem foi o único atributo que não apresentou um comportamento espacial variando de acordo com a latitude. As covariáveis ambientais selecionadas para a predição não apenas representaram importantes fatores de formação dos solos como também expressaram as diferentes escalas espaciais em que esses fatores influenciam a variabilidade dos solos na Antártica. Os três principais fatores de formação representados foram o clima, relevo e organismos. O clima, expresso pelas covariáveis de temperatura e precipitação, controlou a distribuição dos atributos em macroescala, contribuindo para as distinções pedológicas entre as duas regiões mapeadas, com maiores valores de Soma de Bases, pH e Na nos ambientes mais frios e secos (semidesertos polares da Península Antártica) e maiores valores de Carbono, H+Al e P nos ambientes mais úmidos e quentes da Antártica Marítima. Os organismos e o relevo são determinantes em escala mais local. A influência biogênica foi importante para atributos como o Carbono, H+Al, pH e P, e foi expressa principalmente pelos índices de vegetação (NDVI e SAVI) e a distância às pinguineiras. A última variável representou a influência da ornitogênese, importante processo de gênese dos solos conduzido pelas aves nos ecossistemas antárticos. Por fim, a influência do relevo foi expressa principalmente pelas variáveis de altitude, com as partes mais altas da paisagem apresentando solos menos intemperizados e com maiores valores de Soma de Bases e pH e as partes mais baixas apresentando maiores conteúdos de Carbono, P e H+Al, indicando solos com maior desenvolvimento pedogenético e atividade biológica.

Por fim, o terceiro capítulo foi destinado para a modelagem e predição das classes de solos da Antártica, também utilizando-se o Random Forest. Para este trabalho, foram utilizados os dois principais sistemas taxonômicos de classificação dos solos da Antártica: Soil Taxonomy e WRB/FAO. Por sua vez, três mapeamentos foram feitos no âmbito da Soil Taxonomy, contemplando os três primeiros níveis taxonômicos. A utilização de atributos do solo como covariáveis mostrou-se uma abordagem com grande potencial para o mapeamento de classes de solos, uma vez que a variabilidade dos atributos físico-químicos do solo na Antártica está diretamente relacionada aos processos de formação, o que melhora a capacidade de interpretação dos modelos de Aprendizado de Máquinas. De maneira geral, os resultados expressaram as classes de solos mais abundantes das regiões antárticas mapeadas, traçando relações ambientais das classes de solos com o clima, relevo e pedogênese. No entanto, os resultados foram limitados pela distribuição desbalanceada das classes de solos, que, apesar de representarem em parte a variabilidade natural dos solos antárticos (com predomínio dos Gelisols e Entisols), contribuíram para a supressão de classes raras, reduzindo a pedodiversidade dos mapeamentos.

Este estudo é uma contribuição inicial para se atender à necessidade crescente de informações contínuas e precisas sobre os solos da Antártica, visando fornecer subsídios sobre solos para a tomada de decisões visando a conservação do continente. Além disso, estes são dados pioneiros que podem ser incorporados em futuros modelos ambientais, uma vez que os modelos globais atuais não incorporam as áreas livres de gelo da Antártica devido as limitações de disponibilidade de dados.

APÊNDICES
CAPÍTULO 1

Supplementary Table 1. Part1. Accuracy metrics of training for all models.

Metrics	Depth	Bayes GLM	GBM	MONMLP	Radial SVM	RF	Weighted k-NN	Null
					Sand			
R^2	**0-10**	0.08	0.23	0.16	0.18	**0.26**	0.20	-
	10-30	0.08	0.23	0.08	0.14	**0.24**	0.15	-
	30-100	0.10	0.23	0.09	0.15	**0.24**	0.16	-
					Silt			
	0-10	0.03	0.16	0.06	0.11	**0.20**	0.14	-
	10-30	0.05	0.15	0.06	0.13	**0.19**	0.14	-
	30-100	0.08	0.18	0.10	0.11	**0.21**	0.12	-
					Clay			
	0-10	0.17	0.29	0.21	0.24	**0.31**	0.26	-
	10-30	0.15	0.26	0.16	0.22	**0.29**	0.21	-
	30-100	0.13	0.26	0.22	0.24	**0.32**	0.27	-
					Sand			
LCCC	**0-10**	0.20	0.64	0.39	0.42	**0.90**	0.56	-
	10-30	0.22	0.75	0.25	0.52	**0.90**	0.51	-
	30-100	0.28	0.83	0.24	0.51	**0.90**	0.52	-
					Silt			
	0-10	0.06	0.52	0.17	0.39	**0.89**	0.58	-
	10-30	0.10	0.57	0.14	0.43	**0.89**	0.52	-
	30-100	0.19	0.56	0.21	0.38	**0.90**	0.45	-
					Clay			
	0-10	0.34	0.74	0.46	0.52	**0.90**	0.63	-
	10-30	0.31	0.76	0.42	0.57	**0.90**	0.63	-
	30-100	0.27	0.84	0.50	0.50	**0.91**	0.67	-
					Sand			
MAE	**0-10**	13.01	11.63	12.29	11.97	**11.31**	11.87	13.57
	10-30	13.10	12.02	13.26	12.68	**11.75**	12.77	13.73
	30-100	14.45	12.98	14.24	13.70	**12.76**	13.69	14.81
					Silt			
	0-10	9.72	8.95	9.57	9.08	**8.60**	9.05	9.76
	10-30	10.14	9.47	10.08	9.46	**9.15**	9.57	10.31
	30-100	10.17	9.41	9.90	9.78	**9.10**	9.79	10.42
					Clay			
	0-10	5.72	5.18	5.59	5.38	**5.06**	5.34	6.41
	10-30	5.81	5.40	5.87	5.41	**5.24**	5.63	6.44
	30-100	6.62	6.05	6.26	5.96	**5.78**	5.95	7.10
					Sand			
RMSE	**0-10**	16.49	15.09	15.87	15.58	**14.77**	15.40	17.03
	10-30	16.91	15.55	17.00	16.44	**15.28**	16.30	17.50
	30-100	17.85	16.34	17.85	17.23	**16.12**	17.12	18.38
					Silt			
	0-10	12.59	11.83	12.45	12.21	**11.56**	12.00	12.76
	10-30	13.40	12.70	13.32	12.91	**12.35**	12.82	13.66
	30-100	12.80	12.04	12.56	12.54	**11.76**	12.48	13.15
					Clay			
	0-10	7.41	6.84	7.28	7.13	**6.75**	7.08	8.09

10-30	7.46	6.98	7.51	7.15	**6.82**	7.24	8.03
30-100	8.36	7.70	8.05	7.88	**7.36**	7.71	8.86

Supplementary Table 1. Part2. Accuracy metrics of test for all models.

Metrics	Depth	Bayes GLM	GBM	MONMLP	Radial SVM	RF	Weighted k-NN	Null
					Sand			
R^2	**0-10**	0.07	0.22	0.16	0.16	**0.26**	0.19	-
	10-30	0.08	0.20	0.07	0.14	**0.24**	0.13	-
	30-100	0.06	0.16	0.07	0.13	**0.19**	0.13	-
					Silt			
	0-10	0.02	0.13	0.05	0.10	**0.19**	0.13	-
	10-30	0.03	0.10	0.04	0.12	**0.17**	0.12	-
	30-100	0.05	0.11	0.08	0.08	**0.17**	0.08	-
					Clay			
	0-10	0.17	0.28	0.21	0.23	**0.31**	0.26	-
	10-30	0.13	0.25	0.15	0.21	**0.28**	0.20	-
	30-100	0.09	0.21	0.20	0.21	**0.29**	0.25	-
					Sand			
LCCC	**0-10**	0.16	0.39	0.31	0.31	**0.43**	0.37	-
	10-30	0.17	0.38	0.16	0.28	**0.40**	0.27	-
	30-100	0.16	**0.34**	0.16	0.25	**0.34**	0.26	-
					Silt			
	0-10	0.04	0.26	0.11	0.21	**0.35**	0.27	-
	10-30	0.06	0.23	0.09	0.24	**0.30**	0.25	-
	30-100	0.10	0.24	0.16	0.18	**0.29**	0.19	-
					Clay			
	0-10	0.30	0.46	0.39	0.40	**0.49**	0.46	-
	10-30	0.26	0.43	0.31	0.37	**0.45**	0.38	-
	30-100	0.19	0.39	0.38	0.38	**0.45**	0.44	-
					Sand			
MAE	**0-10**	13.05	11.74	12.37	12.19	**11.33**	11.97	13.61
	10-30	13.19	12.33	13.42	12.76	**11.84**	12.94	13.86
	30-100	14.59	13.60	14.21	13.75	**12.99**	13.79	14.69
					Silt			
	0-10	9.88	9.13	9.74	9.25	**8.59**	9.16	9.88
	10-30	10.31	9.76	10.22	9.56	**9.26**	9.71	10.38
	30-100	10.41	9.81	10.08	10.06	**9.31**	10.13	10.50
					Clay			
	0-10	5.73	5.26	5.60	5.40	**5.06**	5.33	6.48
	10-30	5.85	5.43	5.91	5.48	**5.22**	5.68	6.46
	30-100	6.77	6.24	6.27	6.04	**5.85**	5.96	7.09
					Sand			
RMSE	**0-10**	16.56	15.25	15.94	15.87	**14.82**	15.56	17.09
	10-30	17.04	15.96	17.21	16.53	**15.39**	16.54	17.60
	30-100	18.04	17.13	17.85	17.35	**16.49**	17.31	18.19
					Silt			

0-10	12.88	12.23	12.74	12.49	**11.73**	12.26	12.92
10-30	13.69	13.25	13.58	13.17	**12.66**	13.15	13.78
30-100	13.12	12.67	12.83	12.97	**12.09**	12.92	13.20
				Clay			
0-10	7.45	6.94	7.33	7.16	**6.77**	7.09	8.11
10-30	7.56	7.05	7.69	7.24	**6.87**	7.31	8.05
30-100	8.54	7.99	8.10	7.95	**7.48**	7.75	8.80

Supplementary Table 2. Statistics of sand, silt, and clay contents for the entire area and for the MA and NAP.

	Depth		Min	Max	Mean	SD	CV (%)
Sand	0-10	%	49.29	83.06	66.30	5.01	7.56
	10-30		45.77	83.58	64.95	5.50	8.47
	30-100		42.80	83.43	63.48	6.52	10.27
Silt	0-10	%	8.94	33.91	20.78	3.63	17.47
	10-30		9.66	38.97	23.33	3.88	16.63
	30-100		9.22	40.81	23.78	4.81	20.23
Clay	0-10	%	6.88	22.98	13.01	2.08	15.99
	10-30		6.93	20.93	12.89	2.01	15.59
	30-100		6.79	21.35	13.13	2.22	16.91
Maritime Antarctica							
Fractions	Depth		Min	Max	Mean	SD	CV (%)
Sand	0-10	%	50.43	83.06	67.02	5.09	7.59
	10-30		48.28	83.58	65.45	5.62	8.59
	30-100		42.80	83.43	63.16	7.24	11.46
Silt	0-10	%	8.94	33.91	20.66	3.85	18.64
	10-30		9.66	38.62	23.16	4.09	17.66
	30-100		9.22	40.81	23.65	5.21	22.03
Clay	0-10	%	6.88	20.54	12.62	2.04	16.16
	10-30		6.93	20.93	12.79	2.05	16.03
	30-100		7.18	21.35	13.33	2.30	17.25
Northern Antarctica Peninsula							
Fractions	Depth		Min	Max	Mean	SD	CV (%)
Sand	0-10	%	49.29	77.29	62.90	4.60	7.31
	10-30		45.77	78.55	62.57	4.97	7.94
	30-100		48.75	80.60	65.00	3.11	4.78
Silt	0-10	%	12.39	31.93	21.35	2.58	12.08
	10-30		14.15	38.97	24.14	2.87	11.89
	30-100		12.66	40.11	24.41	2.93	12.00
Clay	0-10	%	8.78	22.98	14.85	2.25	15.15
	10-30		7.90	19.80	13.33	1.86	13.95
	30-100		6.79	18.40	12.17	1.82	14.95

Supplementary Table 3. Statistics of sand, silt and clay contents for the islands and each ice-free area.

Ice-free areas	Sand				Silt				Clay			
	Mean	Min	Max	SD	Mean	Min	Max	SD	Mean	Min	Max	SD
South Shetland Islands												
King George Island	**61.25**				**24.94**				**14.46**			
Keller Peninsula	58.99	41.47	82.25	5.49	26.24	9.99	39.31	4.19	17.06	9.30	27.28	2.48
Western Shore of Admiralty Bay	59.42	42.53	74.83	4.69	24.01	11.53	33.63	3.20	19.17	10.36	29.32	2.44
Hennequin Point	64.43	51.22	83.77	5.68	24.33	10.90	33.00	4.46	11.23	4.05	17.58	1.79
Cape Lions Rump	63.52	46.99	81.54	4.35	25.52	11.42	40.19	3.48	10.86	4.15	16.90	1.71
Potter Peninsula	61.42	50.60	80.67	6.37	23.07	9.14	33.19	5.08	15.70	9.03	21.63	1.98
Barton Peninsula	60.41	50.93	79.08	4.62	25.40	10.15	36.65	3.89	14.75	11.23	21.99	1.07
Fildes Peninsula	59.01	41.05	82.31	6.61	26.92	9.52	44.76	5.52	12.57	6.36	19.64	1.43
Other areas	62.78	41.47	85.41	5.87	23.99	9.09	40.19	4.15	14.33	4.05	29.32	3.97
Nelson Island	**61.94**				**22.93**				**15.56**			
Stansbury Peninsula	63.10	48.99	81.42	5.69	23.20	9.65	33.81	4.02	14.21	6.56	19.85	2.58
Harmony Point	61.84	46.03	79.84	5.87	21.37	8.61	30.22	3.56	17.65	11.88	26.45	1.94
Other areas	60.88	41.25	81.42	7.63	24.23	9.65	44.17	6.47	14.81	6.56	26.45	1.97
Robert Island	**66.48**				**21.59**				**12.02**			
Coppermine Peninsula	65.83	56.21	80.48	5.32	22.07	11.12	31.00	3.58	11.32	7.96	14.45	0.96
Other areas	67.12	53.62	80.48	5.73	21.10	9.81	32.04	4.76	12.72	7.51	19.37	2.10
Livingston Island	**69.64**				**20.52**				**10.81**			
Byers Peninsula	59.99	24.56	90.27	9.72	24.01	5.42	52.97	6.52	17.87	4.79	30.51	5.24
Hurd Peninsula	72.90	58.53	88.55	3.65	20.54	7.11	32.42	3.07	7.10	4.48	10.38	0.84
Elephant Point	74.11	61.48	84.70	5.32	17.63	9.63	27.83	4.48	8.84	5.93	11.64	1.13
Other areas	71.57	57.94	88.55	4.65	19.91	7.11	32.42	3.65	9.44	4.01	16.46	2.28
Snow Island	**66.98**				**18.97**				**17.43**			
President Head Peninsula	67.05	45.36	80.50	5.20	18.53	7.56	29.37	3.88	17.29	12.62	29.81	1.98
Other areas	66.90	51.48	77.57	6.42	19.41	11.07	32.32	4.63	17.57	11.91	24.53	2.50
Elephant Island	**61.79**				**26.55**				**9.43**			
Stinker Point	60.13	46.78	85.84	3.82	27.45	8.01	39.34	2.97	9.61	5.12	15.58	1.07
Other areas	63.46	46.78	85.84	4.73	25.64	8.01	39.34	2.98	9.25	5.12	15.58	1.17
Ardley Island	**57.90**	41.05	73.58	6.55	**26.08**	13.97	40.77	5.29	**14.58**	10.94	19.64	1.04
Deception Island	**65.57**	53.25	83.53	4.83	**23.61**	11.29	35.96	2.99	**10.89**	4.71	17.67	2.39
Penguin Island	**75.81**	67.80	85.41	2.47	**16.54**	9.09	21.89	1.73	**7.43**	4.48	11.19	1.55
Half Moon Island	**76.16**	59.80	86.36	3.58	**18.31**	8.89	28.42	3.57	**7.76**	4.01	13.95	1.68
Barrientos Island	**71.24**	63.07	76.59	3.32	**18.88**	13.67	26.69	2.33	**9.56**	7.54	11.78	0.87
Greenwich Island	**68.45**	55.07	79.21	5.66	**20.38**	9.21	33.34	4.71	**12.02**	7.50	21.63	1.78
Low Island	**60.66**	51.26	73.90	3.55	**26.21**	19.11	33.15	1.57	**14.10**	10.54	20.59	2.61
Other smaller islands	**67.33**	42.25	80.48	3.91	**20.86**	9.63	36.28	3.00	**13.03**	6.60	26.44	1.13
Northern Antarctic Peninsula												
Hope Bay	61.98	39.67	79.93	4.80	24.18	13.01	38.81	2.83	14.34	7.27	27.89	3.59
James Ross Island	**64.81**				**23.04**				**12.52**			
Ulu Peninsula	63.54	53.90	75.25	2.40	24.11	12.81	32.92	2.27	13.62	9.43	19.44	1.19
Other areas	63.59	46.26	77.63	2.59	23.74	12.81	35.92	2.03	13.71	8.38	20.96	1.55
Vega Island	**64.23**				**22.09**				**14.45**			
Cape Lamb	62.96	46.26	74.68	2.32	22.27	13.86	35.92	2.04	15.22	10.70	20.96	1.76

Other areas	64.51	46.26	78.89	3.53	22.72	13.86	35.92	2.03	13.44	6.38	20.96	2.39
Seymour Island	**63.37**	44.80	86.46	7.26	**23.05**	9.11	33.95	3.22	**12.72**	5.17	21.52	2.97
Cockburn Island	**65.16**	59.29	76.31	3.73	**22.64**	16.79	30.27	1.73	**11.46**	8.39	13.76	1.06
Snow Hill Island	**61.31**	55.23	72.46	3.51	**24.57**	16.96	30.41	1.38	**14.00**	9.74	16.64	1.54

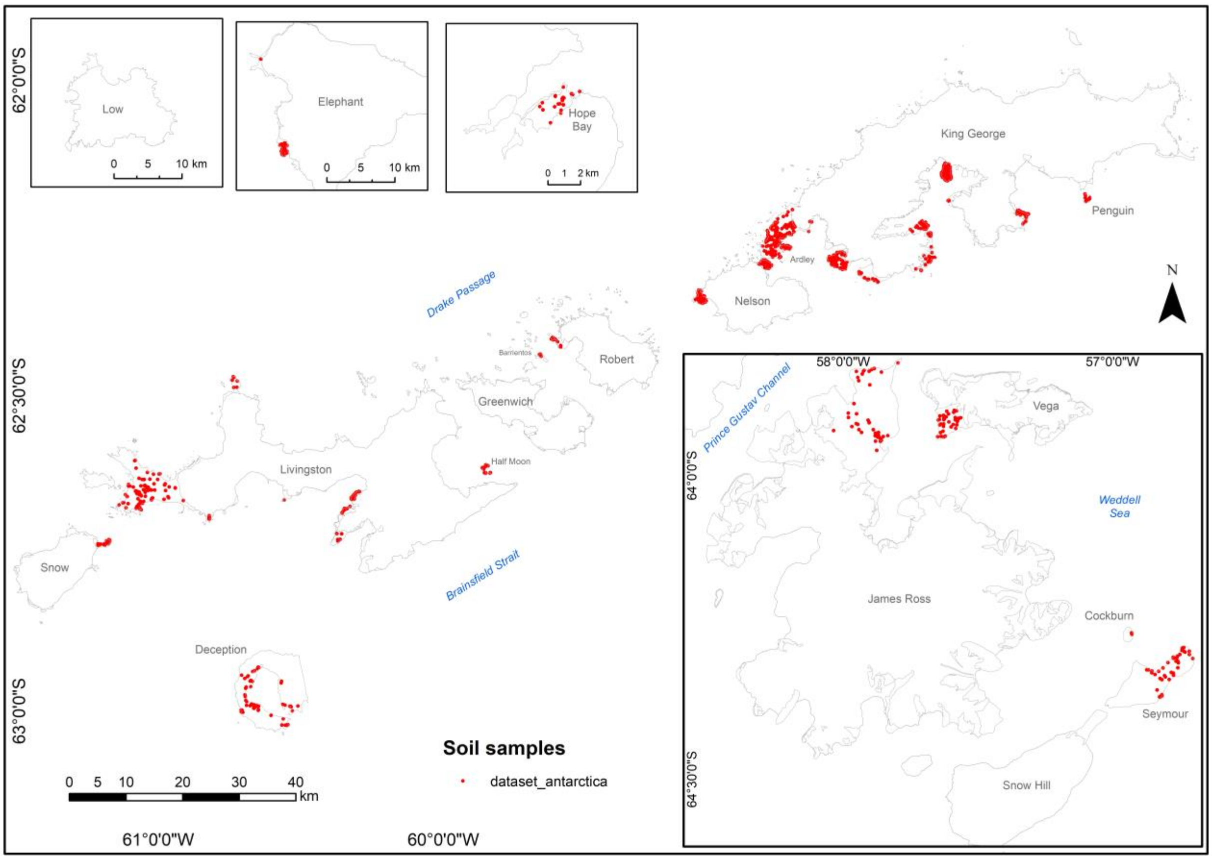
Supplementary Table 4. Coefficient of variation according to the soil fractions and depth.

Fractions	Depth		Min	Max	Mean	SD
Sand	0-10	%	0.93	34.55	4.98	1.52
	10-30		0.88	35.7	5.11	1.47
	30-100		1.09	48.95	6.51	1.36
Silt	0-10	%	1.91	64.95	14.7	4.74
	10-30		1.42	55.55	13.23	4
	30-100		2.3	44.62	10.77	3.52
Clay	0-10	%	1.47	63.59	10.44	3.21
	10-30		1.37	51.43	11.32	3.18
	30-100		3.03	50.71	13.96	3.56

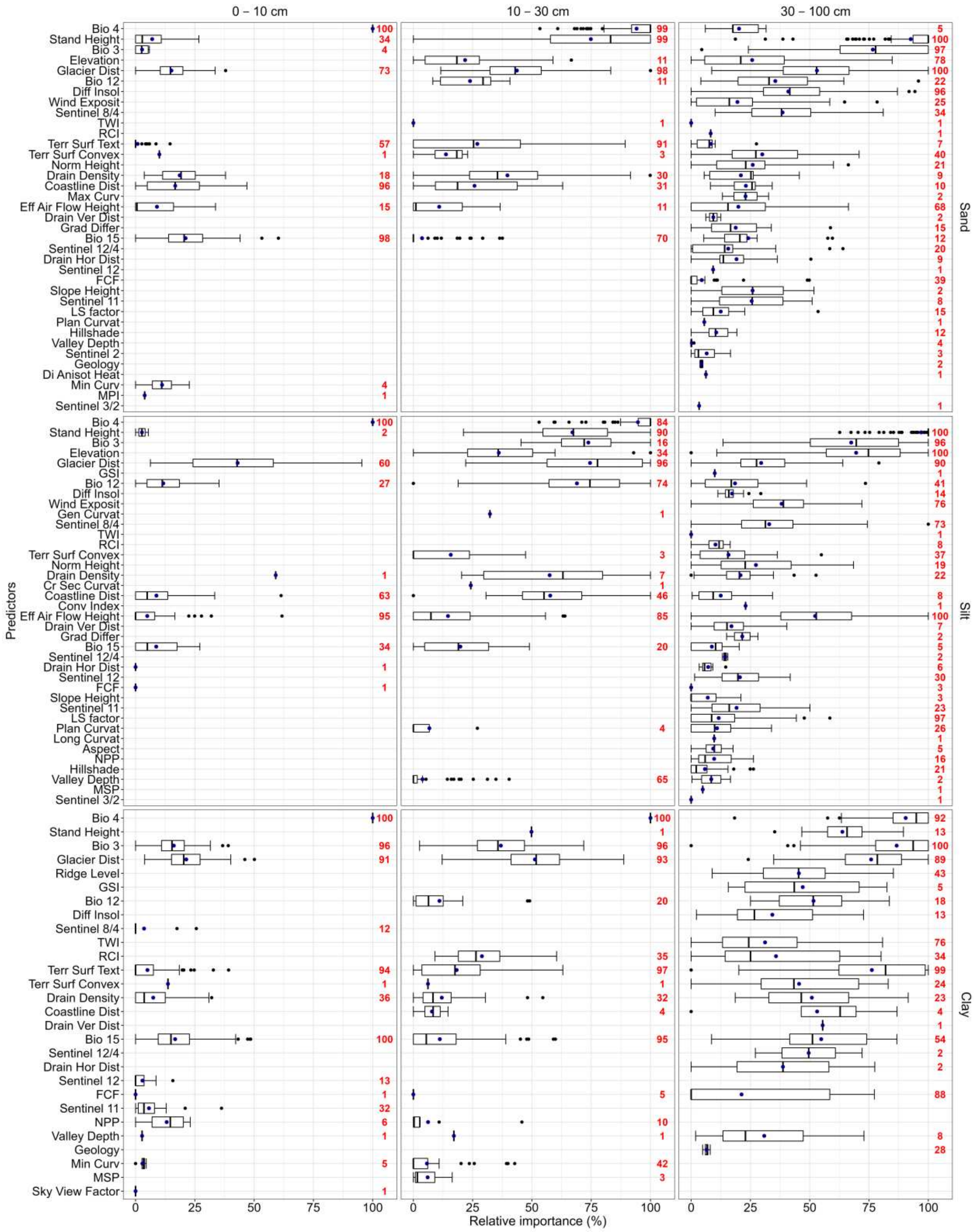
Supplementary Table 5. Percentiles and prediction interval addressing uncertainty according to the soil fractions and depth.

Fractions	Depth	Min	Max	Mean	Predicted mean	Deviations from the mean*
5th percentile						
Sand	0-10	22.5	87.58	58.96	62.9	-3.94
	10-30	15.77	86.85	57.66	62.57	-4.91
	30-100	8.43	87.27	58.07	65	-6.93
Silt	0-10	2.45	48.8	16.89	21.35	-4.46
	10-30	3.62	43.86	20.21	24.14	-3.93
	30-100	3.54	50.9	20.34	24.41	-4.07
Clay	0-10	1.15	25.42	11.75	14.85	-3.1
	10-30	2	23.18	11.18	13.33	-2.15
	30-100	1.84	27.44	10.46	12.17	-1.71
95th percetile						
Sand	0-10	41.7	98.98	69.55	62.9	6.65
	10-30	35.24	100	67.64	62.57	5.07
	30-100	22.06	93.28	70.04	65	5.04
Silt	0-10	7.75	60	26.14	21.35	4.79
	10-30	7.35	67.07	29.71	24.14	5.57
	30-100	7.41	63.78	28.8	24.41	4.39
Clay	0-10	2.65	40.49	16.04	14.85	1.19
	10-30	4.9	33.74	16.22	13.33	2.89
	30-100	3.54	36.23	15.67	12.17	3.5
Prediction interval (90%)						
Sand	0-10	1.56	37.15	10.59		
	10-30	1.67	42.25	9.98		
	30-100	2.14	42.04	11.97		
Silt	0-10	1.44	38.17	9.26		
	10-30	0.89	40	9.5		
	30-100	1.36	32.52	8.46		
Clay	0-10	0.75	22.02	4.28		
	10-30	0.76	18.65	5.04		
	30-100	1.27	19.7	5.21		

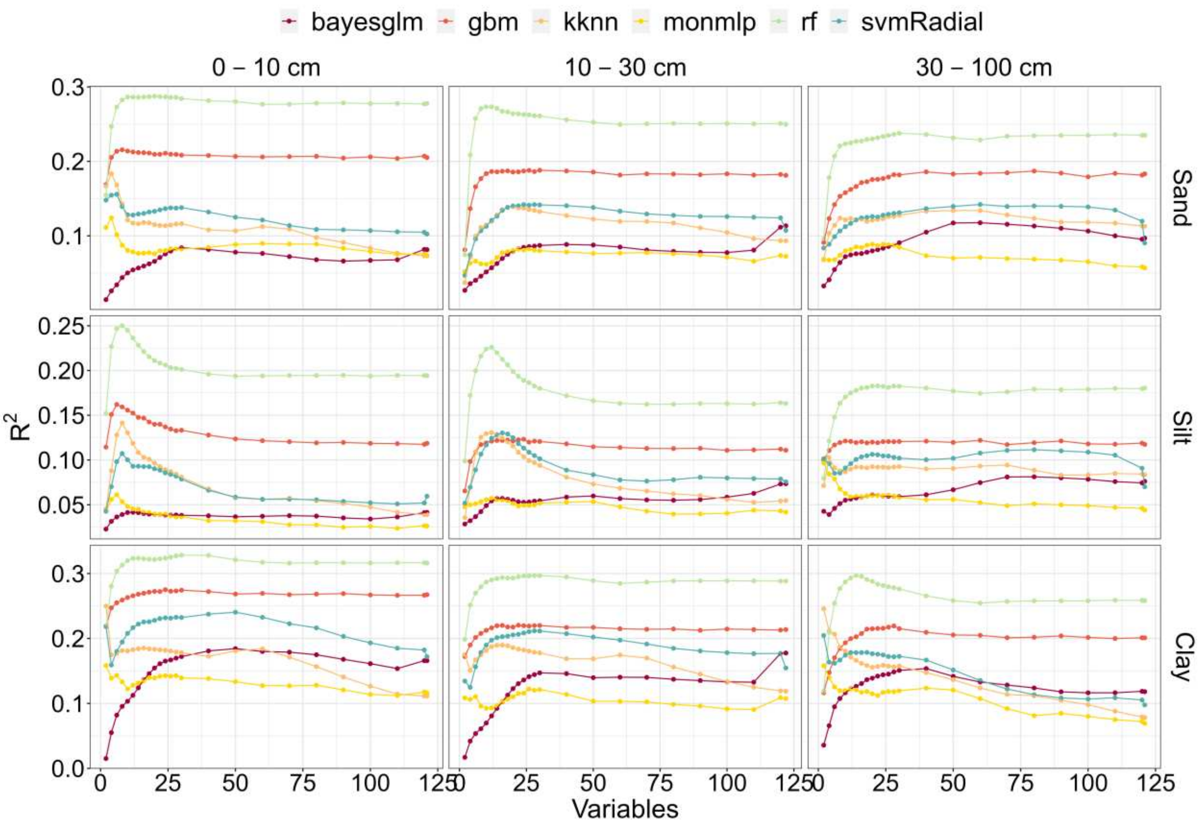
* Mean percentiles – predicted mean



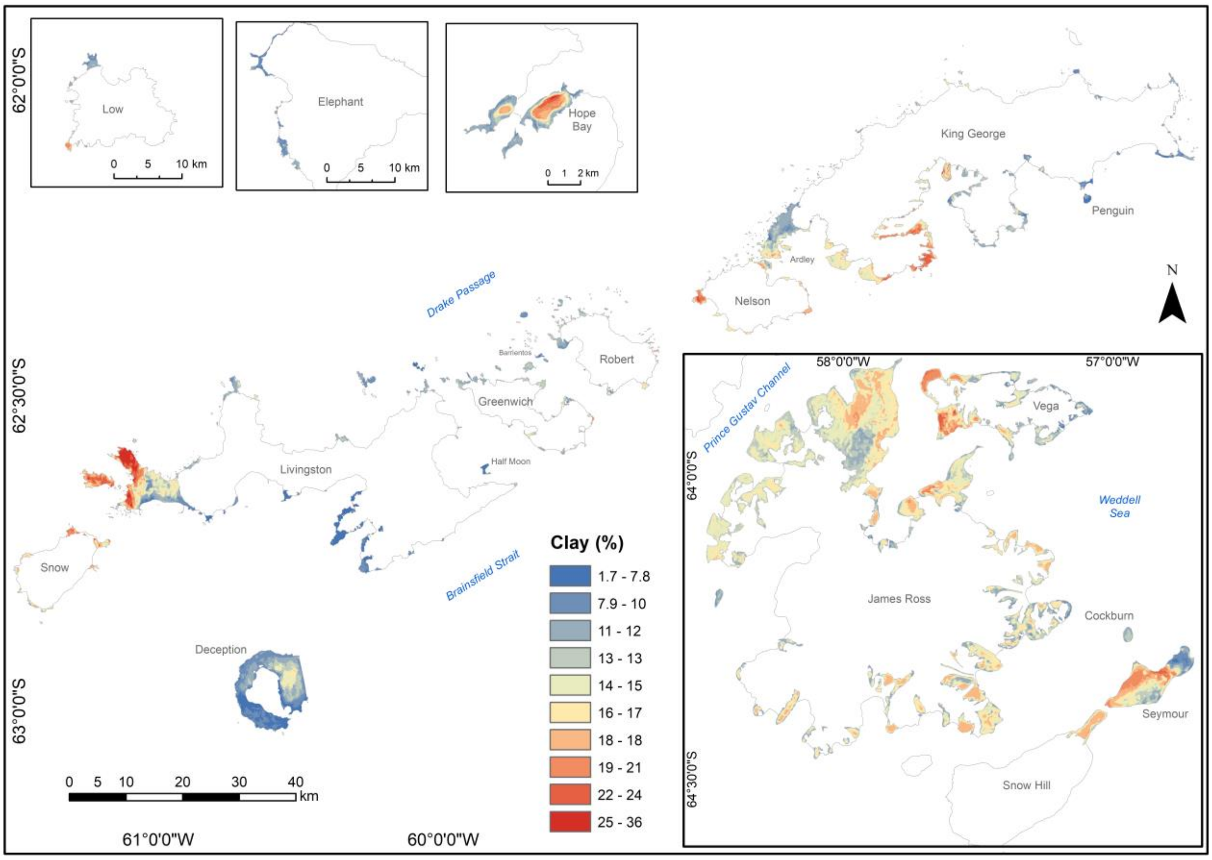
Supplementary Figure 1. Spatial distribution of the soil samples used in this work.



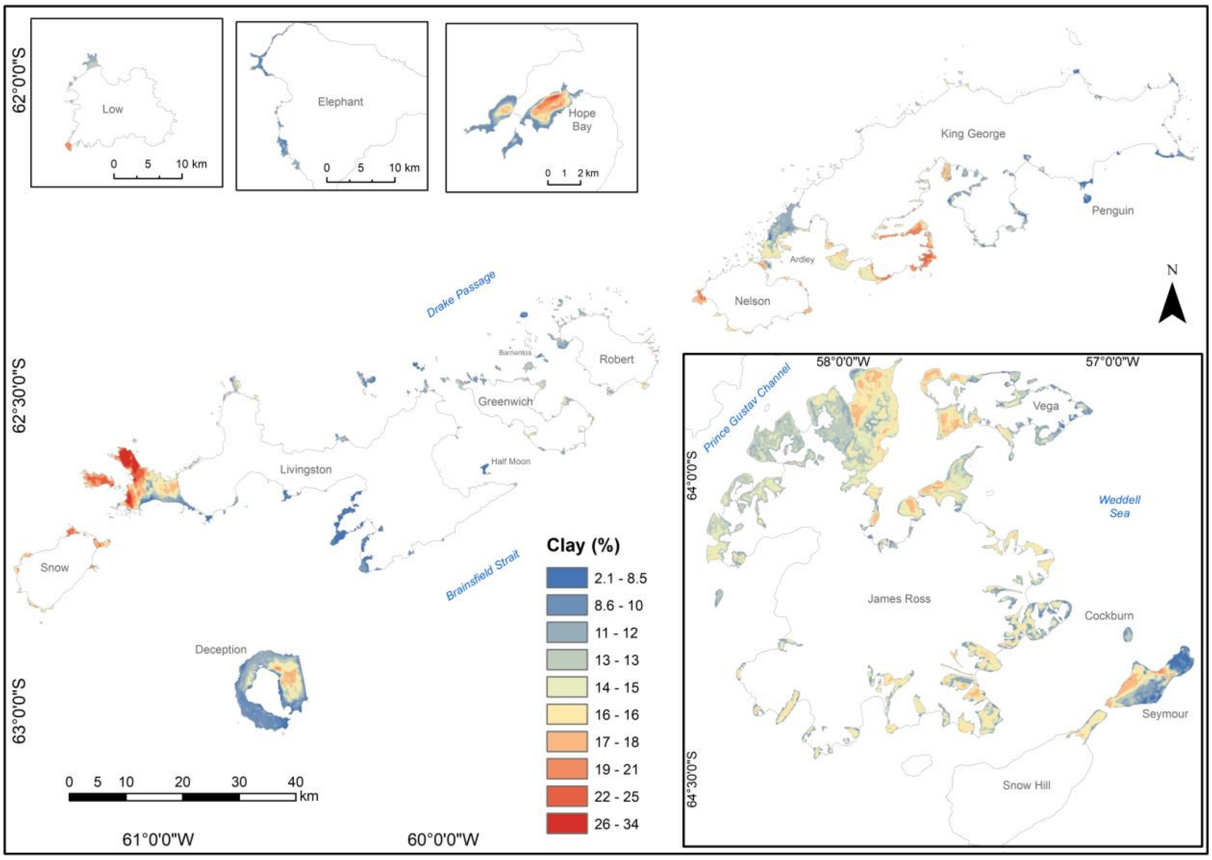
Supplementary Figure 2. Relative importance and frequency of selection of the RF-predictors for each soil fraction and depth, with all predictor variables chosen during the 100 runs.



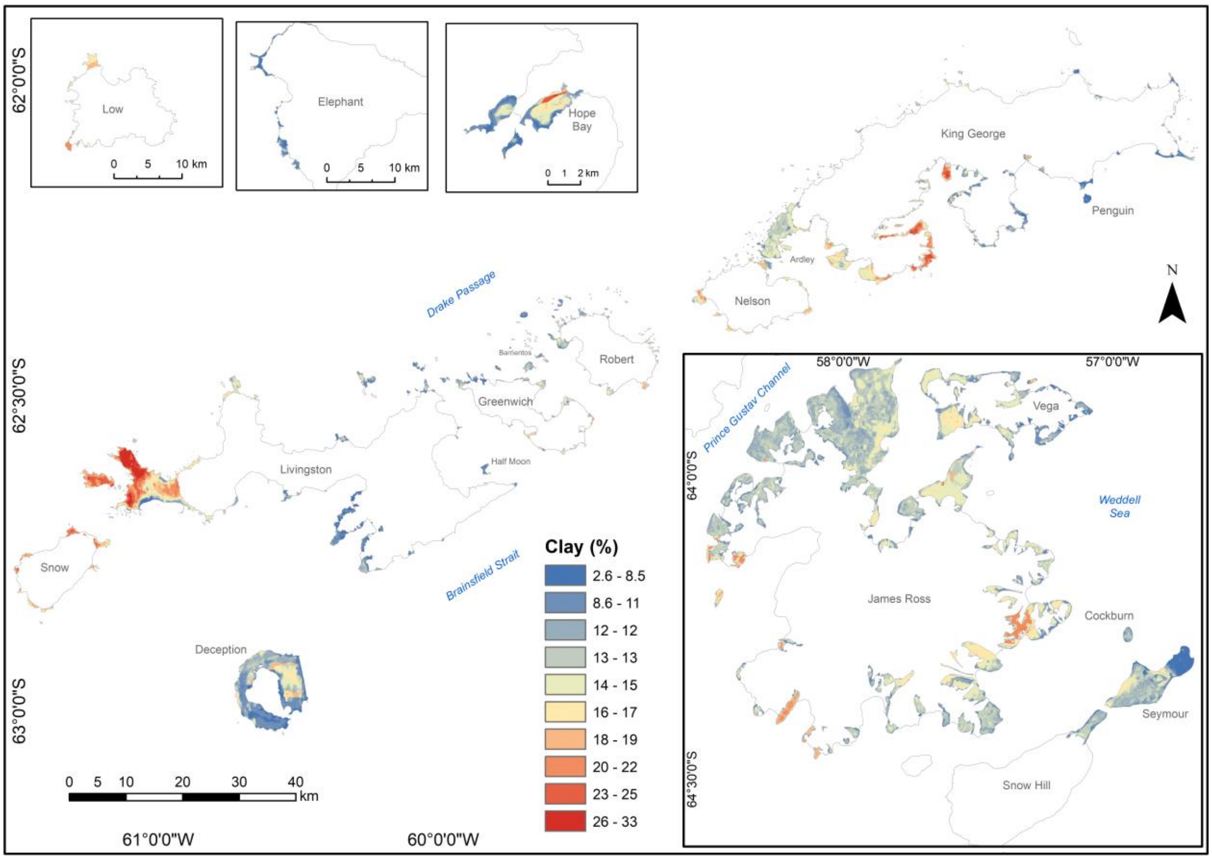
Supplementary Figure 3. The model's performance behavior in the process of selecting predictors using the RFE for each fraction and depth.



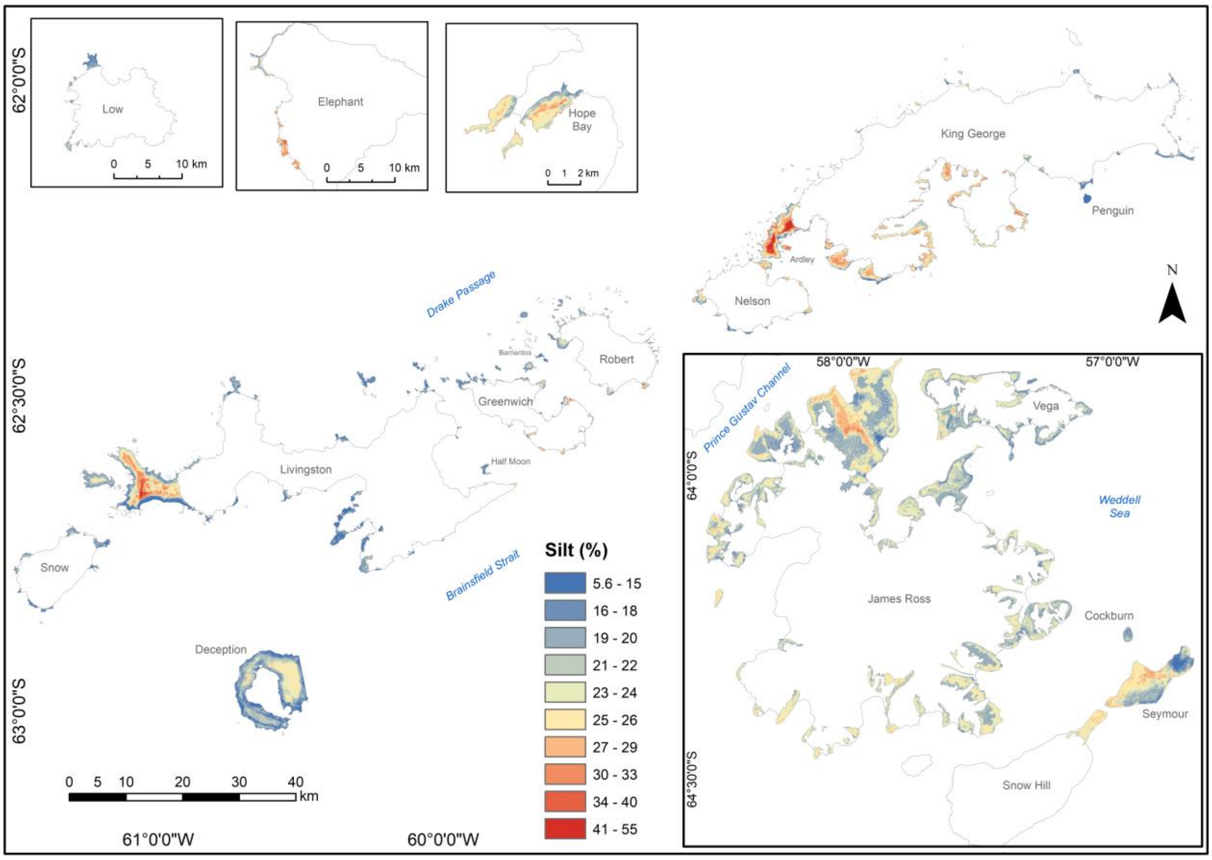
Supplementary Figure 4. Predicted clay map for depth 0-10 cm.



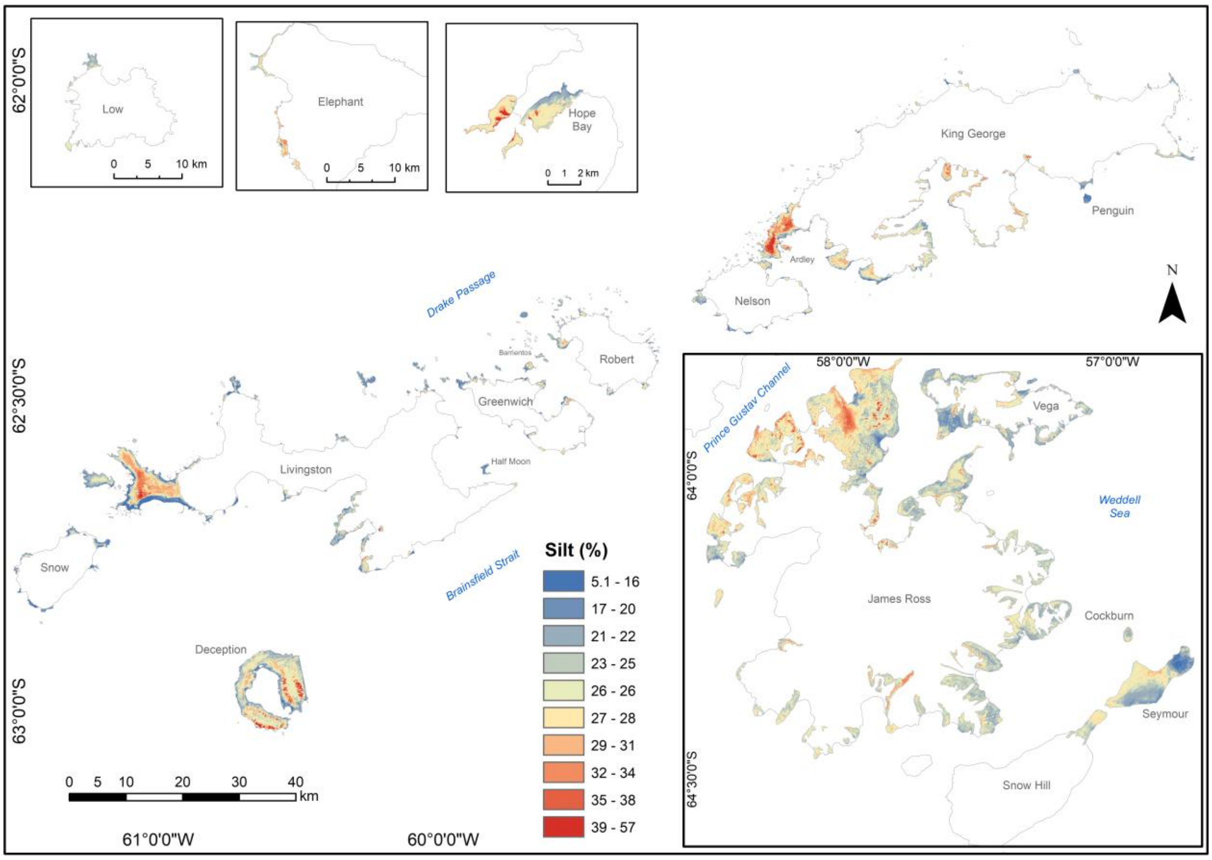
Continuation. Supplementary Figure 4. Predicted clay map for depth 10-30 cm.



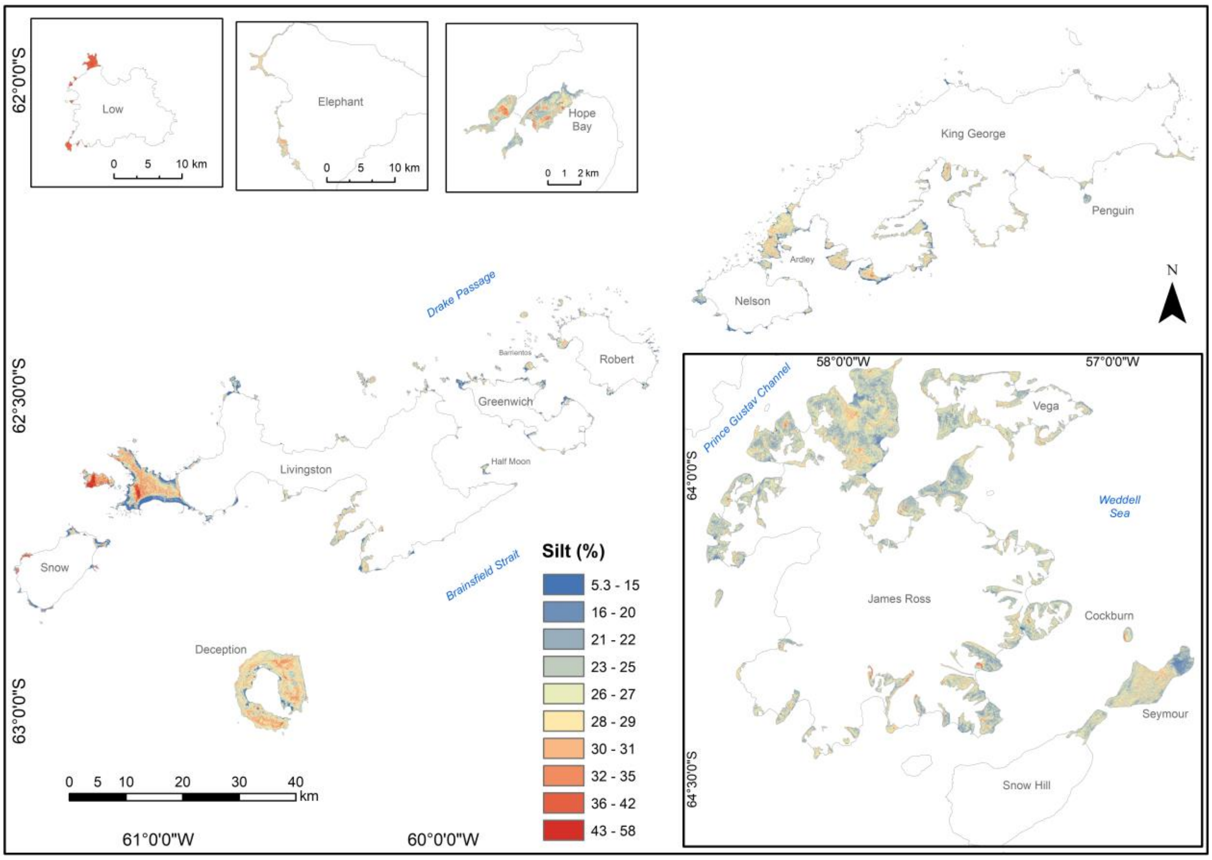
Continuation. Supplementary Figure 4. Predicted clay map for depth 30-100 cm.



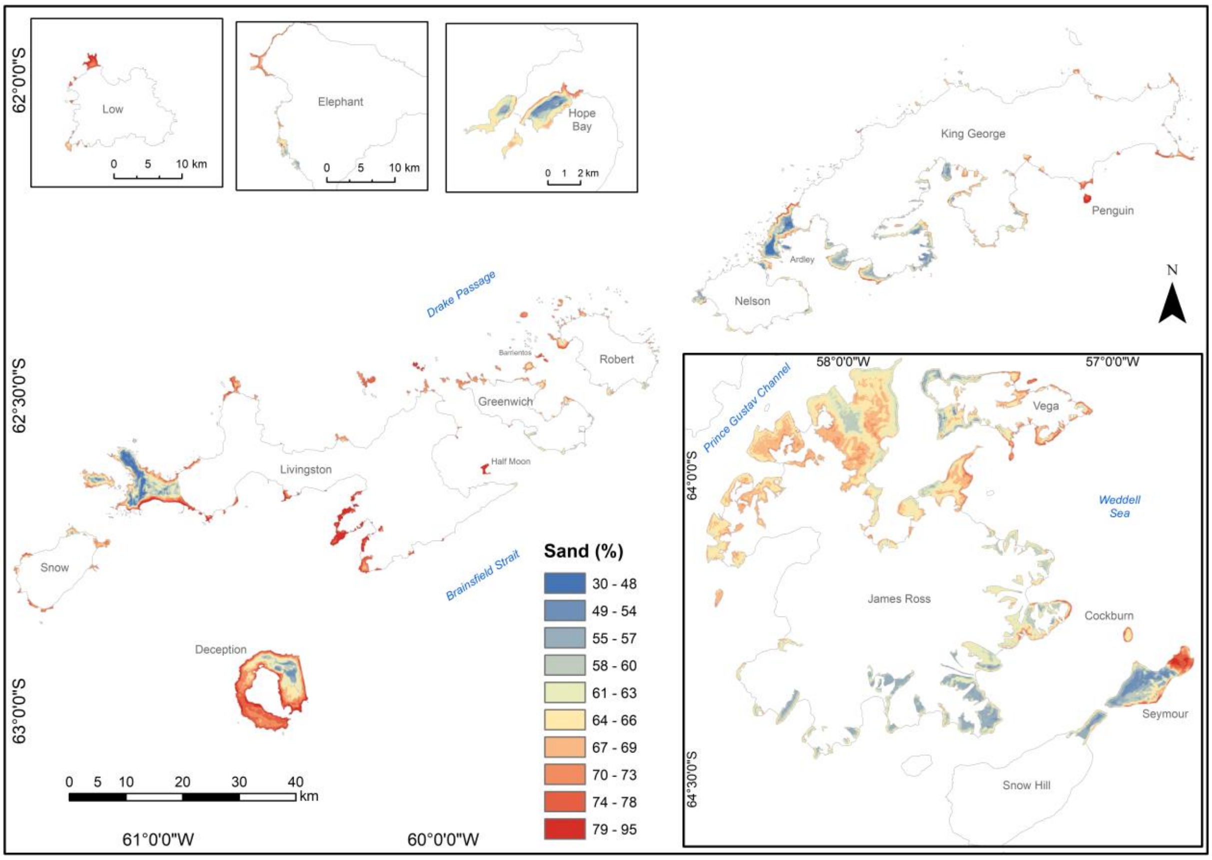
Continuation. Supplementary Figure 4. Predicted silt map for depth 0-10 cm.



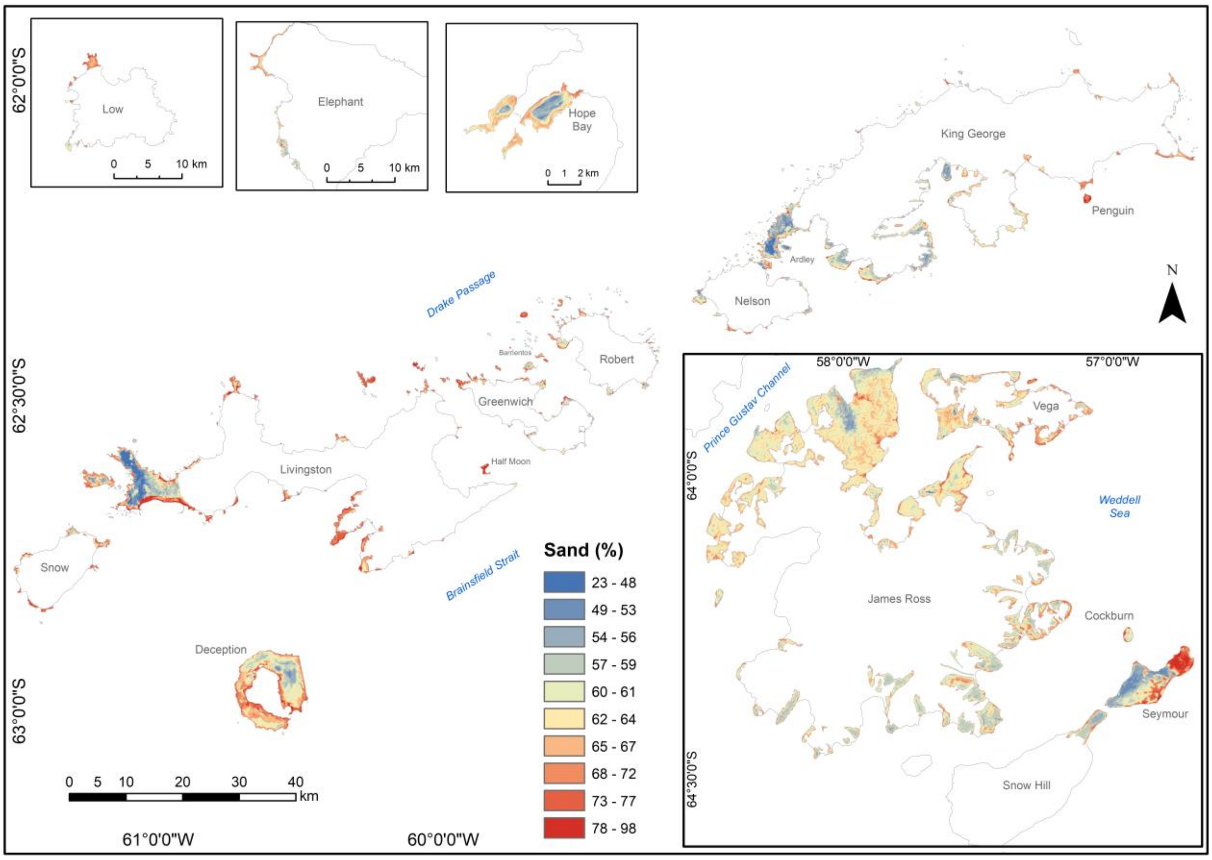
Continuation. Supplementary Figure 4. Predicted silt map for depth 10-30 cm.



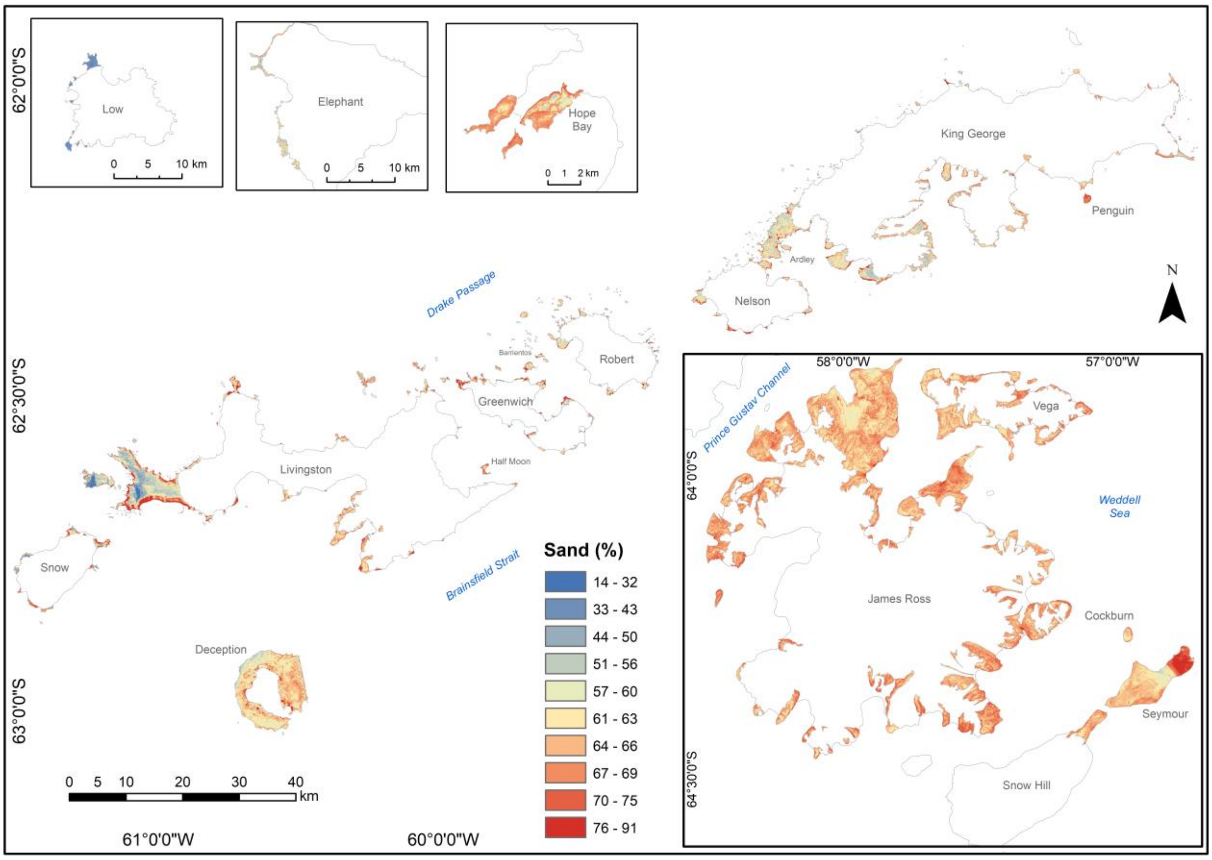
Continuation. Supplementary Figure 4. Predicted silt map for depth 30-100 cm.



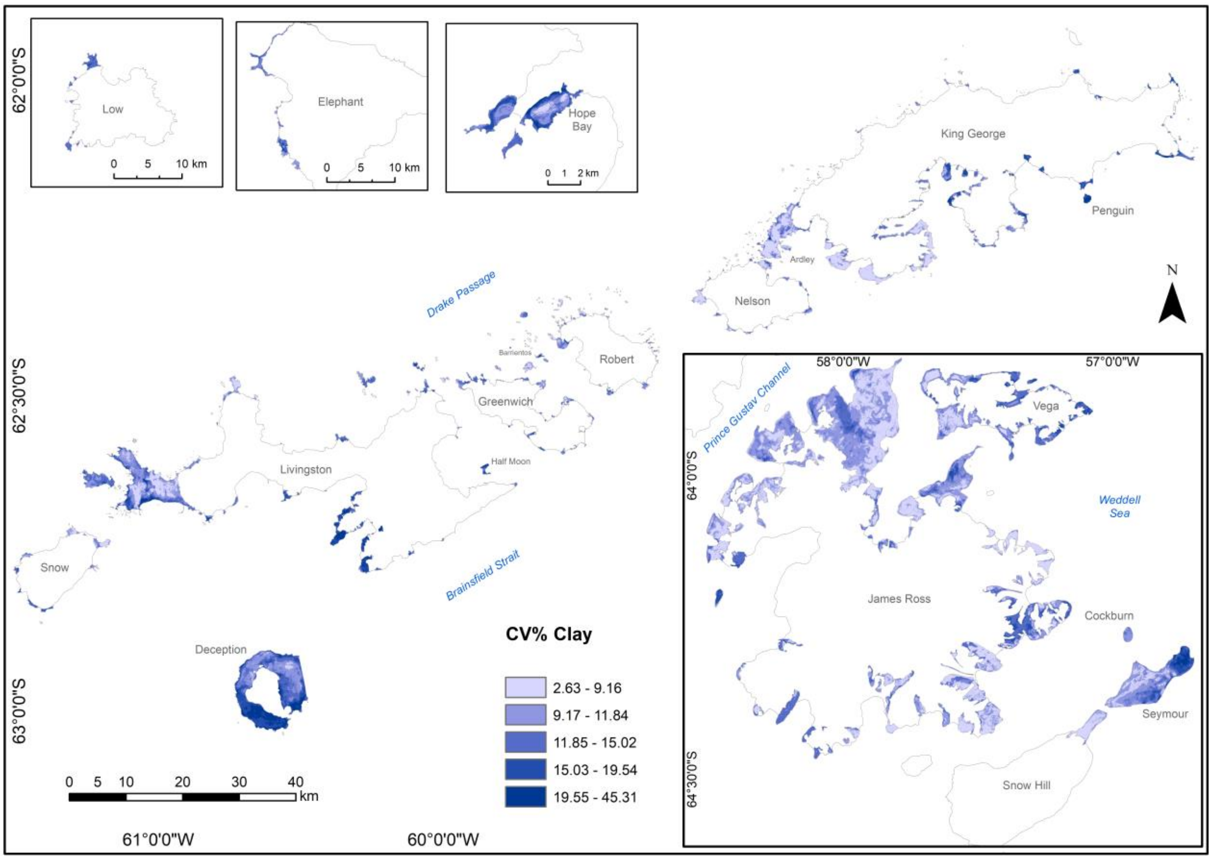
Continuation. Supplementary Figure 4. Predicted sand map for depth 0-10 cm.



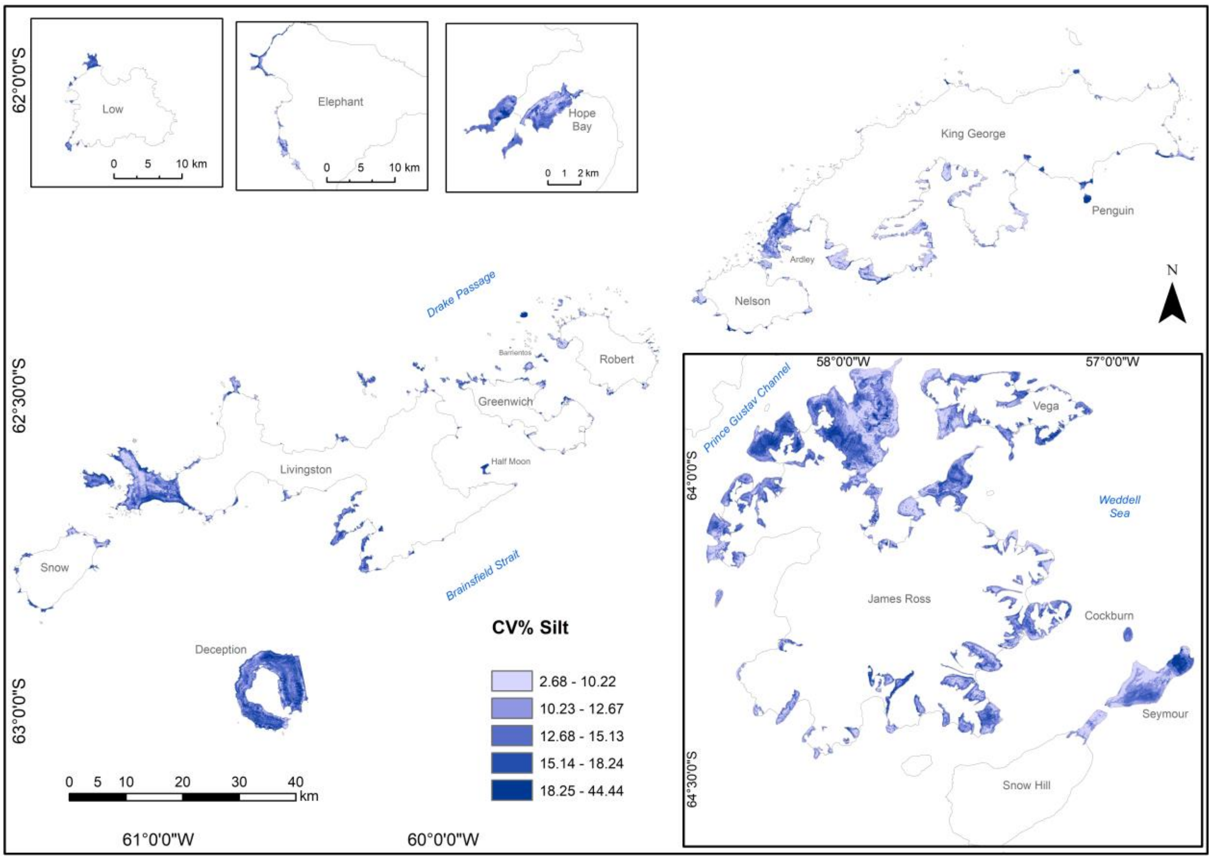
Continuation. Supplementary Figure 4. Predicted sand map for depth 10-30 cm.



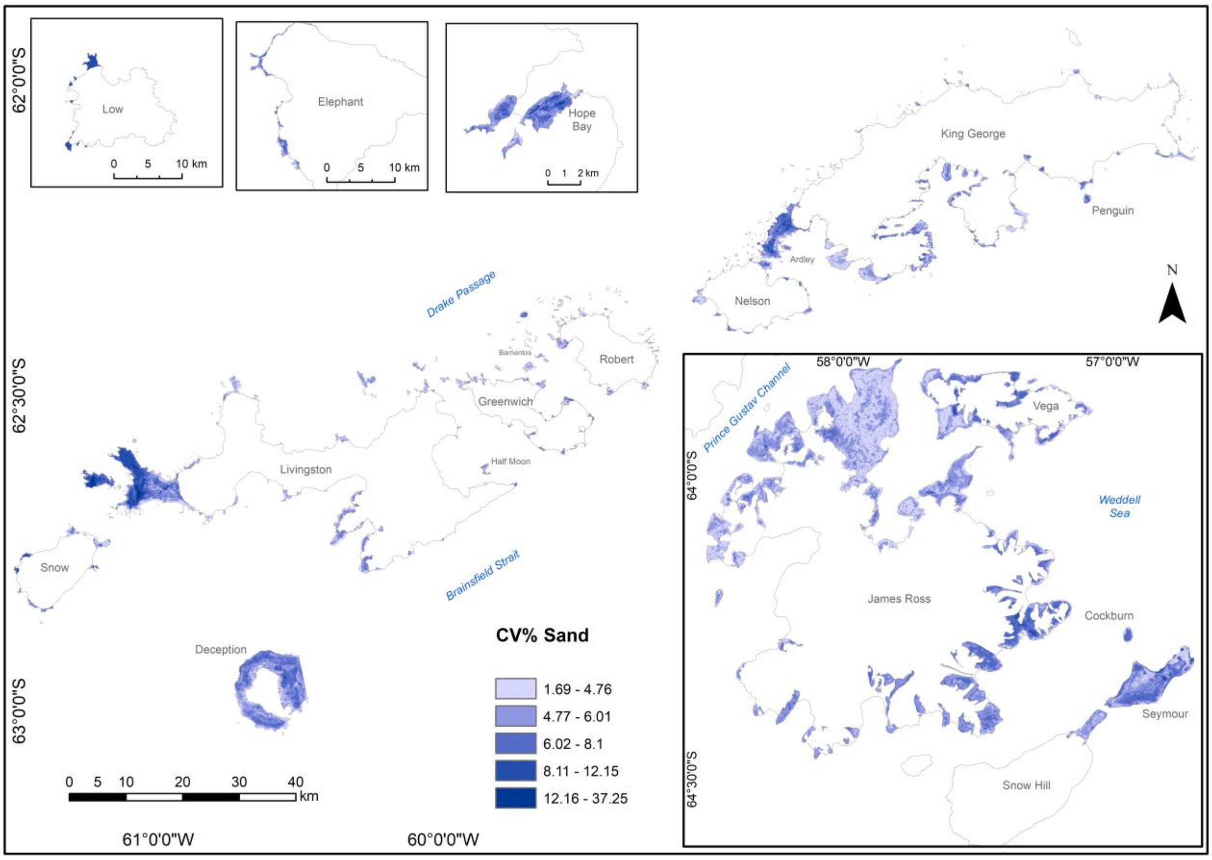
Continuation. Supplementary Figure 4. Predicted sand map for depth 30-100 cm.



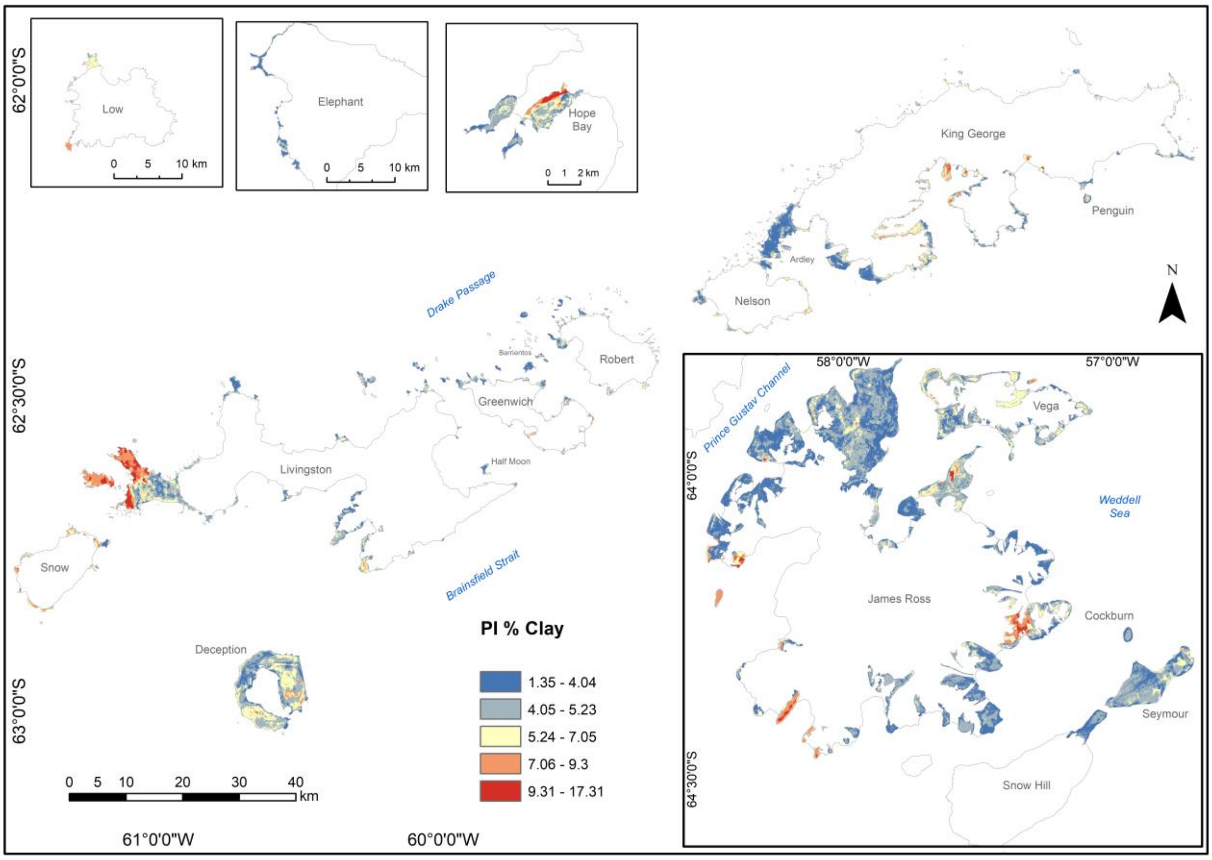
Supplementary Figure 5. Coefficient of variation for clay in 0-100 cm.



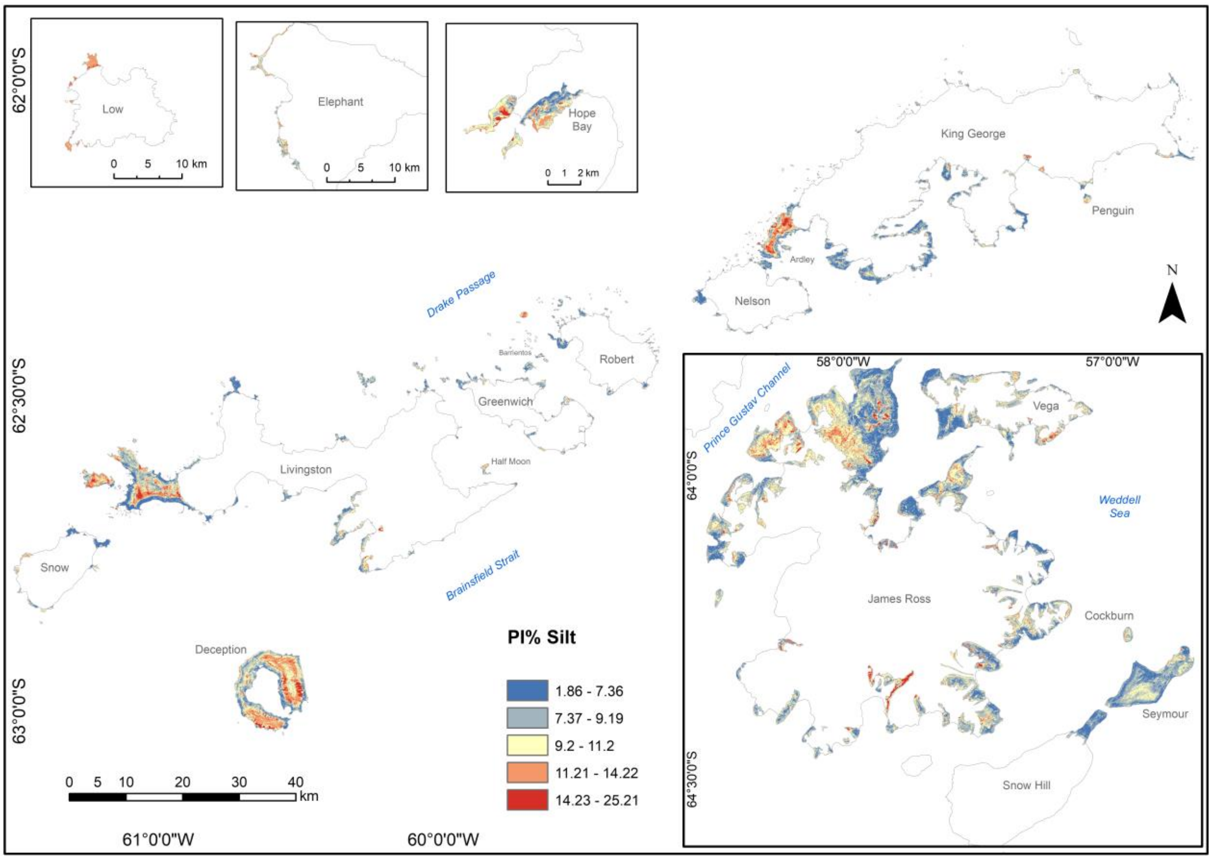
Continuation. Supplementary Figure 5. Coefficient of variation for silt in 0-100 cm.



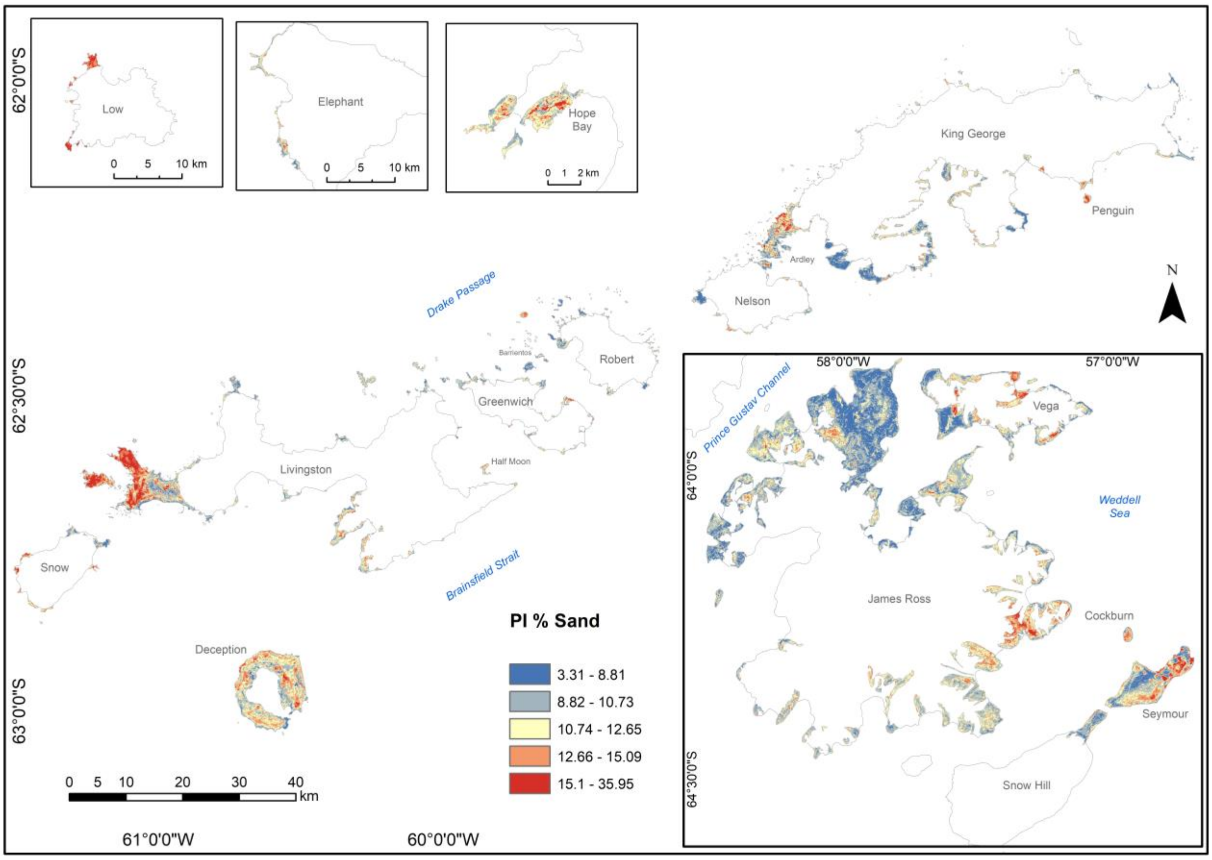
Continuation. Supplementary Figure 5. Coefficient of variation for sand in 0-100 cm.



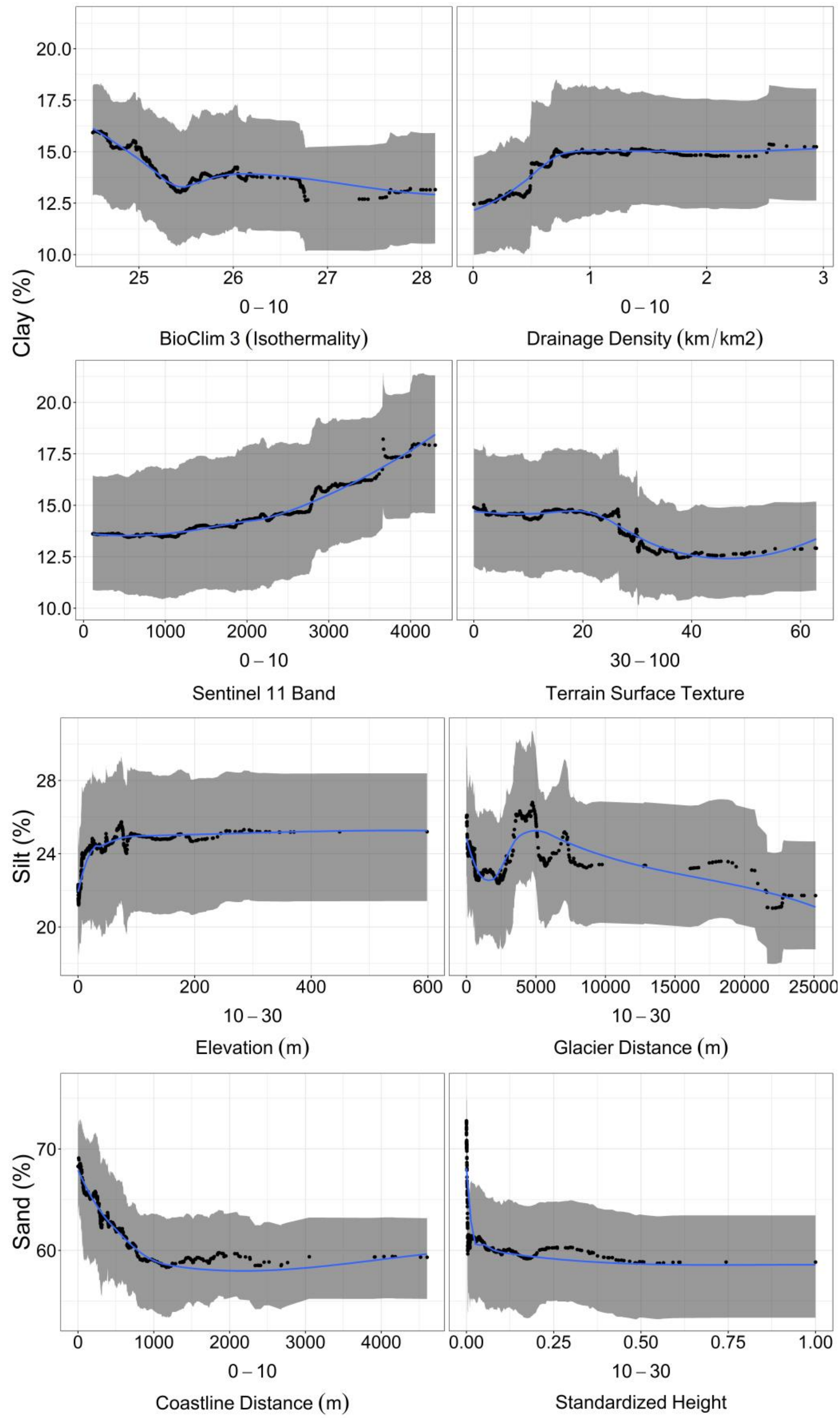
Supplementary Figure 6. Prediction Interval for clay in 0-100 cm.



Continuation. Supplementary Figure 6. Prediction Interval for silt in 0-100 cm.



Continuation. Supplementary Figure 6. Prediction Interval for sand in 0-100 cm.



Supplementary Figure 7. Mean partial dependence (for the 100 runs) of the sand, silt, and clay fractions with some predictor variables.

CAPÍTULO 2

Table 1. Mean accuracy and error metrics of the RF models in the training and test.

Metrics	Depth	BS	TOC	H+Al	Na	P	pH	P-rem
Training								
R^2	0-10	0.48	0.43	0.47	0.42	0.34	0.43	0.42
	10-30	0.44	0.40	0.45	0.41	0.34	0.41	0.43
	30-100	0.35	0.44	0.43	0.34	0.35	0.37	0.28
LCCC	0-10	0.94	0.92	0.93	0.91	0.91	0.94	0.94
	10-30	0.93	0.91	0.93	0.92	0.90	0.94	0.94
	30-100	0.92	0.91	0.93	0.88	0.90	0.93	0.91
MAE	0-10	7.31	2.41	4.50	285.30	470.40	0.74	8.87
	10-30	7.78	1.99	4.87	242.00	478.60	0.79	8.82
	30-100	10.19	1.36	5.56	327.60	438.80	0.95	9.10
MAE Null	0-10	11.13	3.69	6.89	419.40	592.90	1.06	12.17
	10-30	11.64	3.08	7.50	335.60	601.20	1.13	12.48
	30-100	13.68	2.00	8.21	417.40	549.70	1.37	10.93
RMSE	0-10	10.11	4.22	6.43	609.40	1010.80	0.95	11.29
	10-30	11.27	3.72	6.90	488.00	1002.40	1.02	11.29
	30-100	13.33	2.44	7.86	621.70	885.00	1.24	11.55
RMSE Null	0-10	13.93	5.56	8.80	797.40	1254.10	1.26	14.70
	10-30	14.90	4.79	9.24	638.60	1236.50	1.32	14.85
	30-100	16.42	3.23	10.26	760.60	1122.10	1.55	13.40
Test								
R^2	0-10	0.47	0.42	0.44	0.34	0.25	0.42	0.41
	10-30	0.42	0.38	0.42	0.36	0.24	0.40	0.42
	30-100	0.34	0.41	0.38	0.28	0.26	0.39	0.24
LCCC	0-10	0.64	0.59	0.61	0.51	0.41	0.59	0.56
	10-30	0.61	0.56	0.59	0.51	0.40	0.56	0.57
	30-100	0.49	0.54	0.54	0.45	0.37	0.54	0.37
MAE	0-10	7.31	2.43	4.49	307.20	481.50	0.73	8.95
	10-30	7.81	2.00	4.89	255.90	492.90	0.79	8.86
	30-100	10.13	1.42	5.78	328.10	455.50	0.93	9.26
MAE Null	0-10	11.06	3.77	6.79	444.86	552.50	1.06	12.17
	10-30	11.57	3.08	7.42	365.23	577.50	1.13	12.58
	30-100	13.51	2.03	8.27	413.81	547.80	1.38	10.90
RMSE	0-10	10.12	4.30	6.46	686.70	1052.30	0.95	11.32
	10-30	11.16	3.75	6.98	553.70	1065.60	1.02	11.38
	30-100	13.28	2.56	8.21	635.30	981.00	1.23	11.78
RMSE Null	0-10	13.82	5.59	8.57	795.78	1155.80	1.24	14.65
	10-30	14.51	4.70	9.09	670.72	1178.50	1.32	14.87
	30-100	16.11	3.22	10.31	707.87	1067.50	1.56	13.36

Table 2. Average values of percentiles and uncertainty indicators of RF soil attribute models.

		BS	H+Al	pH	TOC	P-rem	Na	P
5th Perc	0-10	23.10	2.42	6.87	0.75	31.79	843.21	131.33
	10-30	22.78	2.55	6.68	0.56	31.04	756.47	125.32
	30-100	23.33	2.80	6.38	0.40	31.83	703.73	170.50
95th Perc	0-10	36.44	5.11	7.24	1.46	35.61	2177.50	390.21
	10-30	34.20	4.99	7.20	1.11	35.53	1752.18	439.53
	30-100	32.27	6.84	7.00	1.06	36.59	1527.05	607.60
PI (90%)	0-10	13.36	2.69	0.37	0.71	3.81	1334.29	258.88
	10-30	11.42	2.43	0.52	0.55	4.49	995.71	314.21
	30-100	8.93	4.04	0.62	0.66	4.75	823.32	437.10
PIR	0-10	0.46	0.90	0.05	0.71	0.11	0.94	1.14
	10-30	0.44	0.73	0.08	0.72	0.14	0.85	1.51
	30-100	0.33	0.91	0.09	1.05	0.14	0.75	1.07